THE UNIVERSITY *of* EDINBURGH

GENETIC PLEIOTROPY IN DISEASE AS A SOURCE OF DRUG TARGET DISCOVERY

Marie Zechner

THE UNIVERSITY
of EDINBURGH

Doctor of Philosophy
The University of Edinburgh

2023

Abstract

The genetic basis of most diseases is complex. They are the result of many genetic variants which each confer part of the disease risk. A rapidly expanding number of genome-wide association studies (GWAS) offers a wealth of information on disease genetics and promises to advance our understanding of disease. However, translating this knowledge to novel clinical insights remains a fundamental challenge. In an attempt to address this issue, I exploit the fact that pleiotropy – the phenomenon whereby a single genetic variant can affect multiple seemingly unrelated traits – is widespread throughout the human genome. This can be used to elucidate disease mechanisms by highlighting which diseases are likely to share common molecular architecture, which can improve our understanding of less well-known diseases and reveal opportunities for drug repurposing. Additionally, combining data from multiple diseases which share an associated variant improves statistical power for the genomic region, enabling discoveries that would not have been detected in one GWAS alone.

In the first part of this thesis, I utilised this approach to understand the molecular underpinnings of pathology in critical Covid-19. The Covid-19 pandemic was an unprecedented healthcare challenge and in spite of rapid scientific progress, death and severe illness from Covid-19 is still a worldwide threat to human health. In an effort to improve treatment options, I aimed to obtain further insight into the functional mechanisms behind critical Covid-19 by studying where it genetically intersects with other diseases.

I explored pleiotropy between critical Covid-19 and a large range of diseases on a genome-wide and localised scale by analysing a curated dataset of 228 disease GWAS. Calculating their overall genetic correlation using high-definition likelihood inference analysis revealed multiple disease connections with critical Covid-19, suggesting causal genetic overlap with other diseases. I then performed a genome-wide multi-trait colocalization analysis using the Bayesian algorithm HyPrColoc (Hypothesis Prioritisation for multi-trait Colocalization) to isolate shared causal variants. Seven pleiotropic variants were detected as shared between critical Covid-19 and other diseases, including idiopathic pulmonary fibrosis, asthma, Crohn's disease, systemic lupus erythematosus, psoriasis, rheumatoid arthritis, allergic disease, hypothyroidism and

hypertension. A novel Covid-19 association with a variant in the gene *SLC39A8*, which is known to be widely pleiotropic, suggests pathways involving the divalent metal ion transporter ZIP8 as potential therapeutic targets. Pathway enrichment analysis of genes with expression identified to be affected by the pleiotropic variants pointed to the cellular checkpoint PD-1 as a possible drug target. This protein controls multiple downstream pathways that have been implicated in Covid-19.

In the second part of this thesis, I investigated whether shared genetic variants can pinpoint drug repurposing opportunities, as common underlying molecular mechanisms between diseases may indicate that a drug used for one disease could successfully treat another. Pleiotropic variants between the 228 disease GWAS were identified via genome-wide colocalization analysis using HyPrColoc. Potential repurposing candidates were selected by finding drugs currently used or in clinical trials for one of two diseases sharing a variant and testing if the drug targets matched genes functionally connected to the shared variant. This method identified 3625 drug repurpose candidates, suggesting an untapped potential in drug repurposing opportunities. Notably, there was particularly strong support for the use of TYK2 inhibitors in the treatment of autoimmune hypothyroidism.

In this thesis I highlight functional mechanisms shared between critical Covid-19 and other diseases, identifying molecular pathways for potential therapeutic interventions. My results indicate that pleiotropic genetic variants are a valuable source of information for drug repurposing studies.

Lay Abstract

Human disease is influenced by both genetics and the environment. The genetics of such diseases are complicated, with many of the small variations that naturally differ in the DNA of individuals contributing to the risk. These genetic variants impact diseases by affecting molecular functions in our cells. Identifying them helps improve our understanding of what mechanisms are involved in the disease, enabling us to design drugs that compensate for deficient cellular processes. We have identified many of these detrimental variants by studying the genetics of people with diseases through genome-wide association studies (GWAS). However, translating this knowledge to novel clinical insights remains a fundamental challenge. I aimed to address this issue by studying multiple diseases simultaneously. This utilises a genetic phenomenon called pleiotropy, which results in a single genetic variant impacting multiple observable traits in an individual that on the surface seem unrelated. By finding variants that affect multiple diseases, we can gain understanding about the molecular mechanisms of less well understood diseases and potentially repurpose drugs from one disease to another. Additionally, combining data from multiple diseases that are all similarly affected by a variant improves the power of a study for that region of DNA, which enables discoveries that would otherwise require testing the genetics of a much larger number of individuals.

In the first part of this thesis, I focused on gaining a better understanding of the molecular underpinnings of severe forms of Covid-19. In spite of rapid scientific progress, death and severe illness from Covid-19 is still a worldwide threat to human health. A better understanding of the disease is crucial in order to improve treatment options. I compared the genetics of critical Covid-19 with those of a large range of diseases using a dataset of 228 genetic studies. First, I examined their overall genetic similarities and found that there are many connections between critical Covid-19 and other diseases. I then identified which variants are shared between Covid-19 and other diseases. My results indicate that certain variants that affect critical Covid-19 are also connected with a variety of other diseases, including idiopathic pulmonary fibrosis, asthma, Crohn's disease, systemic lupus erythematosus, psoriasis, rheumatoid arthritis, allergic disease, hypothyroidism and hypertension. By exploring which cellular mechanisms are affected by these variants I identified potential therapeutic targets for

Covid-19, including zinc and manganese import into cells and a molecular checkpoint that controls multiple important cellular functions that have been previously connected to the immune response in individuals with Covid-19.

In the second part of this thesis, I explored whether we can use shared genetic variation to discover drugs that could be repurposed from one disease to another. I investigated the similarities between 228 genetic disease studies and selected diseases with a shared genetic variation. For each disease pair, I determined whether drugs that are currently used or in clinical trials for one of the diseases have targets that are functionally related to the effects of their shared genetic variation. I found a large number of potential candidates, the most promising of which could be further explored by future experimental tests in cells. There was particularly strong evidence that two drugs that are currently used to treat psoriasis and rheumatoid arthritis respectively could also be used to treat immune-system related thyroid dysfunction.

In this thesis I highlight several molecular mechanisms that could be targets for new treatments of severe forms of Covid-19. My results indicate that genetic similarities between diseases are a valuable source of information for drug repurposing studies.

Acknowledgements

First, I would like to thank my supervisors Prof. Kenny Baillie, Dr. Erola Pairo-Castineira and Prof. Chris Haley for guiding me through this project. Thank you to Kenny, for always inspiring me and reminding me why we do the work we do. Thank you to Erola, who saved my whole life and PhD project many times over. Thank you to Prof. Chris Haley for lending his immense genetics expertise to this project. I would also like to express my gratitude to the additional members of my doctoral committee – Thank you to Dr. Nicola Pirastu for his advice on the measures of pleiotropy and to Prof. Ruth King for making sure our meetings were delightful.

I want to thank my fellow Baillie lab researchers who have truly been one of the best parts of this endeavour. Thank you to Maaike, who has been with me all the way and supported me through all my ups and downs. Thank you to Konrad for his great humour and analytical insight. Thank you to Sara for welcoming me into the lab and helping me find my feet whenever I needed help. Thank you to Johnny for his medical expertise and advice. Thank you to Evangelos for introducing me to genetic correlation analysis. Thank you to Nikos, Bo, Josh, Melissa, Nick, James, Ana, Dominique, Suzanne, Clark, Eamon, Max, Natasha, Steven, Nelly and Akira for all our entertaining and interesting discussions, as well as our shared commiserations and celebrations over the years.

My thanks to the administrative team. Our wonderful Precision Medicine DTP programme admins, Kate, Susan and Maree have been incredibly helpful at every turn. I am also very grateful to Jen and her tireless efforts in making sure meetings are able to happen in spite of busy schedules.

A big thank you to my friends Victor, Michelle, Giulia, Sarah, Raven, Matt and Narcís who have made this journey a joyous one.

Thank you to Thomas and Corinna of the online emotional support team.

Thank you to Speck for being the best boy and making sure I go outside every day.

I am so grateful to my family for their love and support through all these years. Thank you to my parents, Irene and Konrad, for their encouragement and unwavering belief in my capabilities. Thank you to my brother Roland for being the funniest person I know. Thank you to my grandmother Hannelore for all our cheerful conversations. Thank you to Andreas and Edda for their support throughout my academic undertakings.

Lastly and most of all, I would like to thank my partner Constantinos for having my back through all of this. His love, humour, support and encouragement have made a challenging time all the brighter. His keen scientific, statistical and programming insight has been invaluable in helping me craft this project.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

Marie Zechner

Edinburgh, 01.12.2023

Table of contents

Chapter 1
Introduction

1.1 Challenges in disease genetics

For centuries biologists have debated questions of nature versus nurture. We now know that most human traits are influenced by both genetic and environmental factors. This is also true for diseases, which all to a varying amount have a genetic component. Though rare diseases are often caused by mutations within a single gene, most common diseases are polygenic, occurring due to the complex interaction of multiple genetic and environmental influences[1]. Such multifactorial diseases can be impacted by hundreds to thousands of genetic variants[2]. The impact of each variant can be very small, but through interplay with multiple other variants and the presence of certain environmental context, may lead to an increased risk of susceptibility to a disease or its severity.

We have identified many of these disease-associated variants through genome-wide association studies (GWAS). GWAS are performed by genotyping hundreds to millions of individuals to statistically link single base positions in the DNA (single nucleotide polymorphisms, SNPs) to disease-related-phenotypes[3]. The relationship between sequence variation and disease provides a means of exploring molecular pathology by identifying changes caused in the cell due to the disease-associated SNPs. The aim of such studies is to gain a better understanding of the disease and thus highlight therapeutic strategies. GWAS findings have had an increasing translational impact, prioritising among known treatment options and identifying novel drug targets[4 7]. However, there is still a large discrepancy between the tens of thousands of discovered genetic associations[8] and the functional insight we have gained. The translation of this knowledge into novel clinical understanding remains a fundamental challenge to this approach.

Multiple factors have made it difficult to convert genetic signals into functional understanding of the biology underlying diseases. The first challenge is identifying which variant is causing a GWAS signal. Pinpointing the relevant variant within a locus is difficult due to the haplotype structure of the human genome[9]. The variant conferring

the disease risk is often in strong linkage disequilibrium (LD) with a subset of other variants because they are commonly co-inherited within a population. SNPs that are in strong LD can have statistically indistinguishable connections with the disease, even though only one variant may be actually causal[10]. Additionally, LD structures are population dependent which necessitates separate analyses for genetic data from different ancestries[11]. Discerning the causal variant is further complicated by the fact that many GWAS are based on incomplete DNA microarray (chip array) analysis rather than whole-genome sequencing data[12], which can result in the causal variant being either unavailable or badly imputed from a reference panel.

Once a lead variant is correctly identified, the next challenge is ascertaining its molecular consequence. This is relatively straightforward if the variant is in the coding sequence of a gene, particularly if the result is a missense mutation changing the protein amino acid sequence, or an early stop codon truncating the protein. However, most disease-associated SNPs are not within coding regions, and are often distant from genes[13], complicating the identification of impacted genes. Such intronic and intergenic variants are predicted to affect cell function by altering gene expression levels, for example by altering transcription factor binding sites or allelic chromatin states[14], or through changes in mRNA splicing, stability, localisation and translation[15–17]. These effects are often highly specific to certain tissues or cell types[14], and can also depend on the environmental context and cell state[18], rendering their identification even more complex.

The uncertainty in decoding GWAS signals inevitably necessitates experimental follow up. However, few variants have been thoroughly explored through *in vivo* studies[19], as such studies are costly and time intensive. In the face of a deluge of uncovered disease associations, prioritisation of candidate variants is therefore crucial. In this thesis I aim to advance the functional mechanistic insight gained from disease-associated variants and prioritise candidates for future experimentation by studying the genetic overlap between diseases.

1.2 Pleiotropy

Pleiotropy is the phenomenon whereby a single gene or genetic variant can affect multiple seemingly unrelated phenotypes[20]. It is widespread throughout the human genome and common among polygenic diseases[21,22]. Though pleiotropy at the level of loci or genes is more common, many genetic variants have also been identified as pleiotropic[21]. This effect can be utilised to elucidate disease mechanisms by highlighting

which diseases are likely to share common molecular architecture, which can improve our understanding of less well-known diseases and reveal opportunities for drug repurposing. Combining data from multiple diseases sharing an associated variant improves statistical power for the genomic region, enabling discoveries that would not have been detected in any one GWAS alone[23,24].

Of the several different underlying causes that can lead to observed pleiotropy, only one is biological in nature[25,26] (Figure 1). Horizontal or biological pleiotropy occurs when one variant or gene affects multiple phenotypes, either directly or via a common intermediary. There are many potential mechanisms by which one variant or a single gene may affect multiple traits; alternate transcripts can be created from the same locus through alternate start and stop codons or alternative mRNA splicing[27] and RNA processing can differ depending on tissue and cell state[28]. Proteins may have multiple functions through differing interaction partners, or due to cell state or tissue dependent effects[29]. An example of such multifactorial impact is the missense mutation 1858C/T in the gene *PTPN22* which has been described to have context-dependent consequences. The variant leads to deficient T cell and B cell response to immune system stimulation[30] but also impairs the removal of autoreactive B cells[31]. By contrast, vertical or mediated pleiotropy arises when a variant or gene affects one phenotype and this phenotype in turn affects another. For example, genetic variants have been identified as associated with both low-density lipoprotein (LDL) levels and risk of myocardial infarction[7], while LDL levels themselves are a risk factor for myocardial infarction[32]. While horizontal pleiotropy is informative in terms of understanding shared pathology between diseases, vertical pleiotropy may be useful in identifying strategies for disease prevention. Finally, spurious pleiotropy refers to a mistaken assumption of underlying pleiotropy. Such misinterpretations can arise in situations where a GWAS signal caused by two different variants is erroneously identified as the same signal due to high LD between them, or due to bias within experimental studies such as a non-random sample selection or phenotypic misclassification.

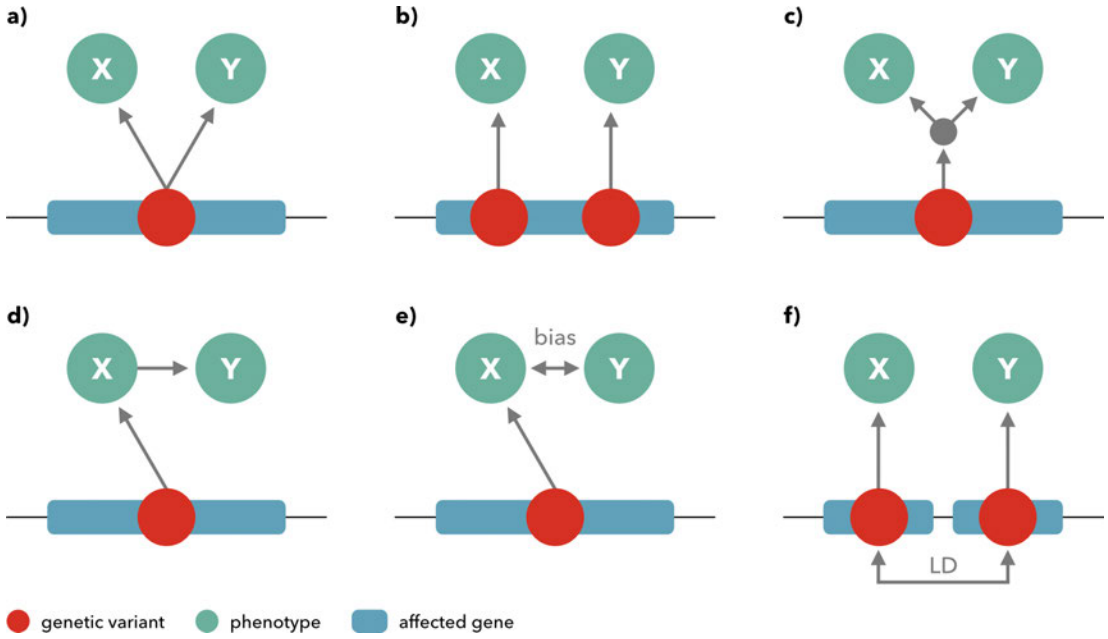


Figure 1: The different types of mechanism of pleiotropy. Pleiotropy is considered biological or horizontal when a) a genetic variant affects two phenotypes, b) a gene affects two phenotypes, even through two different variants or c) a genetic variant affects two phenotypes through a common intermediary. d) Mediated or vertical pleiotropy occurs when a genetic variant affects one phenotype which then in turn affects another. Spurious pleiotropy is a mistaken assumption of pleiotropy, either e) due to bias in the experimental study, such as phenotype misclassification, or f) due to high LD between two different genetic variants – affecting different genes – making the GWAS signal appear as the same for both variants. Figure adapted from Solovieff et al.[25] and van Rheenen et al[26].

Pleiotropic effects are not necessarily uniform in their direction of influence, as what is detrimental to one disease may be protective against another. Such antagonistic pleiotropy may be an evolutionary explanation for why certain variants are carried forward in spite of their associated disease risk, as they may also confer selection-relevant protection depending on factors such as an individual's age and environmental context[33]. The classical example of this is a point mutation in β-globin, which causes sickle cell disease but is protective against malaria[34,35]. The presence of antagonistic pleiotropic effects precludes solely relying on genome-wide techniques for estimating shared genetics between diseases, as widely varying underlying local architecture can create the same result due to opposing effects in different loci[26]. To address this issue, I explore disease similarity on a genome-wide as well as localised level.

1.3 Genome-wide association studies and their interpretation

Genome-wide association studies aim to uncover how our genetics influence specific phenotypes by screening the genome of large cohorts of people. Over the last two decades, GWAS have revolutionised our understanding of disease genetics by identifying thousands of genetic associations. Although the specific techniques have evolved over the years, the workflow for conducting a GWAS generally involves the following steps[36]. DNA samples, phenotypic and demographic data are collected from a cohort of interest, such as individuals suffering from a disease, as well as from a healthy control group. The genetic samples are then genotyped to determine their DNA sequence, either through the use of microarrays, which assess common variants, or through whole-genome sequencing methods. Quality control of the data, such as the analysis of population stratification, typically through covariates such as ancestry, age and sex, as well as the exclusion of badly genotyped SNPs or samples, is necessary. If a limited number of SNPs are genotyped, it is possible to impute information on further variants from a matched reference population dataset. Statistical tests are then performed for each variant to test for association to determine whether the allele has a negative or positive impact on the likelihood of the trait. Typically, associations with binary traits such as whether or not an individual has a disease are assessed through logistic regressions, while continuous traits are assessed with linear regression models.

There are multiple limitations to consider when conducting a genome-wide association study. Unknown or unaccounted for differences between the study and control cohorts can confound the analysis. These arise through non-random sampling from the population, which can be caused by participation bias or unmatched sociodemographic factors[37,38]. Another issue is posed by the high number of association tests required to run a GWAS necessitating a stringent multiple-testing threshold to avoid false positives, which is likely to lead to false negatives[39]. The standard GWAS threshold of P value $< 5 \times 10^{-8}$, which is based on the Bonferroni correction for one million independent tests, may be too conservative as genetic variants are highly correlated throughout the genome and therefore not truly independent[40,41]. One way to circumvent this issue is by acquiring very large sample sizes, but that may not always be feasible depending on the phenotype of interest. Another strategy is to assess genetic associations through alternate statistical frameworks, such as KnockoffGWAS which analyse all genetic variants simultaneously in a manner that accounts for population structures and controls for false

discoveries[42]. Bayesian methods, which are probability based rather than frequentist offer another approach to manage the high load of multiple hypothesis testing, in particular when it comes to fine mapping of the associated variants or recovery of false negatives through post GWAS meta analyses[24,43 45].

Another challenge to the field of genetic research is that complex disease variants identified by GWAS only explain a small part of the estimated trait heritability that was established through family studies, known as "missing heritability"[46]. This could be due to large effects of rare variants that have not been adequately typed in the GWAS, which may in particular be the case for diseases with large fitness consequences[47,48]. Gene-environment interactions and epigenetic variance, which are not captured by GWAS, may also make up part of the missing heritability[49]. Another potential explanation is offered by the omnigenic model, which poses that gene regulation and protein networks are so highly interconnected that all genes expressed in a cell involved in a disease affect the function of disease-relevant pathways[50]. This hypothesis is bolstered by studies showing that common variants with individual effect sizes that are too small to pass GWAS thresholds strongly contribute to the heritability of many traits[51,52]. It is important to note that under this model not all variants or genes are expected to contribute equally to each phenotype, so it remains relevant to prioritise candidates to further biological understanding and uncover opportunities for drug discovery.

Disease GWAS aim to find information about the aetiology of a disease by finding variants with functional consequences to infer underlying molecular mechanisms. However, GWAS test statistical associations, which are a measure of correlation rather than causation. Non-causal SNPs can be associated with a trait because both SNP and trait are correlated with an unobserved variable, such as geographic location[53]. Vertical pleiotropy, where a SNP affects one trait which then affects another trait, can also lead to associations without direct biological link[25,26]. Epistasis caused by nonlinear interactions between causal factors can obscure which variants are biologically causal[54]. Additionally, associations can be indicative of a number of factors other than a causal biological relationship. Variants are frequently identified as associations because they are in linkage disequilibrium with actual trait-associated variants if they are commonly co-inherited[10]. Stochastic noise can also result in associations that are not present at larger population sample sizes. Careful follow up exploration of GWAS results is therefore crucial for their interpretation.

Triangulating evidence with follow up analyses that incorporate information from additional data sources can help address multiple challenges posed by GWAS results. Linking genetic variants of interest to potential functional consequences can assist in narrowing down which variant in a locus is the most likely source of a GWAS signal and offers further support of a variant being truly biologically causal. Moreover, testing which SNPs are most strongly supported by additional evidence helps prioritise variants for functional experiments and can highlight which molecular mechanism and which tissues or cell types may be affected. One way to gather further information about genetic variants is to integrate GWAS results with functional genomics data. Although some GWAS associations are located in protein coding sequences, which enables a more direct study of their likely effect, many more associated variants fall into intergenic regions[13]. As GWAS signals are enriched in accessible chromatin and gene regulatory elements, it is thought that these variants take effect by altering mRNA expression or processing[14,55,56,56,57]. By integrating functional genomics data, we can determine whether a GWAS variant is also associated with effects on regulating cellular mechanisms, such as levels of gene or protein expression, RNA splicing, chromatin accessibility or DNA methylation. These effects are measured in studies of molecular quantitative trait loci (molQTL), which combine genome sequencing and molecular assays in order to test for genetic associations with the quantified molecular features[58]. So far, effects on gene expression have explained more GWAS associations than other regulatory mechanisms such as splicing or polyadenylation[59,60]; as such, studies have primarily focused on testing GWAS overlap with expression quantitative trait loci (eQTL).

One statistical method that aims to use GWAS data for causal inference is Mendelian randomization[61]. In this type of analysis, genetic variants in association with an exposure function as proxy measures in order to determine the relationship between the exposure and an outcome of interest. This can be used integrate GWAS with functional genetic data by designating for example eQTL to be the exposure with the GWAS being the measure of outcome, thereby testing if the genetic variants impact the disease phenotype through their influence on gene expression. As genetic variants are randomly distributed from parental alleles, this enables study structures akin to randomized control trials in situations where groups cannot be assigned to undergo the exposure for ethical or practical reasons. Randomizing test groups results in confounders having an equal probability of affecting test and control group. Using genetic variants as proxies also avoids reverse causation, as SNPs are already present at birth and uninfluenced by developing conditions or environmental circumstances in later life.

Although Mendelian randomization is a powerful tool for causal inference, it is limited by the strong assumptions it necessitates. Firstly, the genetic variant has to be associated with the exposure. This assumption can be directly verified, whereas the others can only be disproven or assessed through sensitivity analyses[62,63]. Secondly, the genetic variant cannot be associated with the outcome through any means other than the exposure. As we have discussed, pleiotropy is widespread throughout the genome, with many if not all genetic variants affecting multiple phenotypes. It is therefore likely that the second assumption always introduces some measure of bias that should be carefully considered. The third assumption is that there are no causes of the outcome that also influence the genetic variant. Although genetic variants are fixed and not subject to outside influences, confounding between SNPs and the studied outcome can still occur due to factors such as population stratification, linkage disequilibrium, pleiotropy or assortative mating introducing spurious associations[64–66]. To minimise bias introduced by these assumptions, Mendelian randomization is best deployed in combination with other complementary methods[61,67].

Colocalization analysis offers a means to ascertain the likelihood that a GWAS lead variant shares an association with a locus affect gene expression, and therefore pinpoint a potential molecular mechanism through which the GWAS variant impacts the disease phenotype. This is a Bayesian method that tests for a shared associated variant between datasets by calculating support for five different hypotheses: That there are no associated variants in either dataset, that there is only association with trait 1, that there is only association with trait 2, that both traits are associated but have different lead variants or that both traits are associated with the same variant[44]. Another method that seeks to leverage eQTL data for post GWAS interpretation is transcriptome-wide association study (TWAS)[68,69]. TWAS perform association tests between genes and phenotypes based on reference data from GWAS and eQTL associations statistics, and are primarily used to prioritise target genes. To date, a wide array of studies has used the overlap of eQTL mapping and GWAS signals to gain insight into the potential functional mechanisms underlying disease-associated variants[44,59,68,70–72].

However, in spite of widespread efforts to interpret the functional effects of GWAS associations through changes in gene expression, it has become clear that a large percentage of GWAS signals do not correspond with eQTLs[73–76]. One explanation could be that current gene expression studies lack sufficient power, or due to challenges in the overlap analyses[77,78], though this has also been argued to the contrary[74]. It may also be that GWAS variants are likely to at least partially only function as eQTL in specific cell

types or within specific cellular context, such as during an immune system response[79 83]. Regulation of processes other than gene expression, such as splicing or polyadenylation[59, 60], may also explain some of the functional impact of GWAS variants. Additionally, studies are currently mainly able to detect cis-eQTLs, which act upon nearby genes, rather than trans-eQTLs, which affect distant genes[50,84,85]. Trans-eQTLs could mediate the effect of GWAS variants, however they are thought to be themselves mediated through cis-eQTL[74,86]. Recently, Mostafavi et al. have shown that GWAS and eQTL signals differ systematically[76]. In this study, eQTLs but not GWAS signals were found to cluster near transcription start sites, while GWAS signals were enriched near transcription factor genes. Genes in proximity with GWAS signals were found to be evolutionarily constrained and high in functional annotations with varied regulation across different cell types, while genes near eQTL signals were more likely to be under relaxed evolutionary constraint and had fewer functional annotations with simpler regulation processes across tissues. They argue that these studies are biased to detect different types of variants. GWAS look for genetic variants that have a measurable impact on a phenotype, biasing them towards genes with high functional importance. Studies mapping eQTL are powered to identify variants with a big impact on gene expression, which biases them toward promoter regions. They are also more likely to be detected if they are at high frequencies, which they are more likely to drift to at genes with limited selective constraint. Ultimately, a combination of multiple different approaches is likely needed to close the gap between GWAS associations and their functional interpretation. Larger sample sizes of functional genomics datasets, as well as the investigation of specific cell types and different cellular contexts will uncover additional information. Additionally complimentary techniques, such as predictions for regulatory elements based on DNA sequence[87,88] and functional assays such as CRISPR screens[89] should be helpful in further elucidating these connections.

Another benefit of integrating GWAS results with functional genomics data is the potential to identify disease-relevant cell types. Complex diseases often involve many interacting cell types, and it can be difficult to discern which cells are actually causal to the development of the disease, rather than just involved in the course of the disease. Uncovering which cells drive a disease cannot only further our understanding of it but also pinpoint targets for treatment, prevention and drug validation[90]. While GWAS associations do not necessarily constitute a causal relationship, they are an indication of potential causality that can help prioritise cell types for further evaluation.

GWAS results can point to disease-relevant cell types if the predicted effect of the associated variants is cell-type specific or if they are enriched in DNA regions that are particularly active in specific cell types[91]. Elucidating this relationship requires statistical methods that integrate GWAS results with transcription data or functional genomic annotations. Colocalization of GWAS variants with eQTL data, as outlined above, can point toward target cells if the queried eQTL data is cell-type specific. Other expression-based approaches include SNPsea, which checks for cell type specific expression levels of genes in or near disease loci[92], and RolyPoly, a regression model integrating GWAS effect sizes around genes and cell type specific gene expression[93]. The SNPsea method is limited by how genes are selected, as GWAS variants do not necessarily impact nearby genes. Additionally, expression-based methods generally assume that cells with high expression levels of a gene will experience the highest impact of genetic variants affecting the gene, which may not always be the case. Alternatively, analyses can test for GWAS variant enrichment in genomic areas with functional annotations, such as open chromatin regions, histone modifications or DNA methylation. Open chromatin is a sign that a DNA section is transcriptionally active within a cell and can be assessed through ATAC-seq (Assay for Transposase Accessible Chromatin using sequencing)[94,95] or DNase hypersensitivity assays[96]. Epigenetic chromatin markers such as DNA methylation and modifications on histones, which are the DNA's scaffold proteins, regulate chromatin structure and can be measured via ChIP(chromatin immunoprecipitation)- or nanopore sequencing[97,98]. Big initiatives such as ENCODE (Encyclopedia of DNA Elements)[99] and Roadmap Epigenomics[100] have provided databases of epigenetic markers across tissues. Many analysis methods seek to interpret GWAS results with chromatin profiles, for example EpiGWAS[101], GREGOR[102] and CHEERS[103], which use chromatin mark data, while GoShifter[104], fGWAS[105] and GARFIELD[106], which focus on functional or chromatin annotations. The latter two use full GWAS summary statistics rather than index results of the lead GWAS variants in order to improve detection power[105,106]. Although many different disease-cell type connections have been identified[91], it is highly likely that the increased availability of higher resolution context specific single-cell datasets will further improve the accuracy of disease-cell prediction over the coming years, advancing our understanding of how cells contribute to complex diseases.

Although testing if a genetic variant is causal or aetiologically relevant for a disease ultimately requires *in vitro* and *in vivo* experiments – such as CRISPR perturbation of GWAS loci, massively parallel reporter assays or molecular activity and animal models – post GWAS analyses can prioritise variants and offer hypotheses for potential functional

molecular mechanisms. In this thesis, I study the genetic overlap of disease GWAS to prioritise associated variants and further explore their potential molecular effects through colocalization analysis with tissue and single cell specific eQTL.

1.4 Critical Covid-19

In 2020, the Covid-19 (coronavirus disease 2019) pandemic created an urgent healthcare challenge. Considerable research efforts accelerated our knowledge of the disease as well as strategies for clinical intervention. In spite of rapid scientific progress, death and severe illness from Covid-19 is still a worldwide threat to human health[107]. Furthering our understanding of the underlying disease mechanisms could be instructive in pinpointing novel therapeutic targets.

Covid-19 is caused by infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It presents with a wide spectrum of clinical manifestations that can vary from asymptomatic to severe and fatal. Covid-19 is considered a systemic disease with common symptoms including coughing, fatigue, gastrointestinal distress, fever, myalgias, weakness and a loss of smell and taste[108]. More severe complications include pneumonia, acute respiratory distress syndrome, cardiac injury, thrombosis, sepsis and cardiac, liver, renal and neurologic injury[109]. Mortality rates vary strongly by age, with risks of below 0.5% for infected individuals under 50, approximately 8% for individuals aged 80 and 20% at 90 years of age[110]. Critical Covid-19 describes a subgroup of patients suffering from hypoxaemic respiratory failure[111] with acute lung injury[112]. This clinical phenotype is thought to be caused by immune-mediated inflammation within the lung and lung blood vessels[112].

Severe forms of Covid-19 unfold within the lung in two phases[113,114]. The first phase of early viral illness is marked by viral replication and virus-mediated tissue damage. SARS-CoV-2 infects the lung directly[115,116] and is also disseminated from the upper respiratory tract[117] where viral replication is the highest[118]. Patients with an impaired anti-viral response are at greater risk of subsequently suffering from critical illness[119]. In particular, a deficient type I interferon response[120,121] and dysfunctional adaptive immune response featuring an imbalance of T and B cells[122 124] have both been linked to problems with SARS-CoV-2 clearance. Although these factors have been associated with severe disease progression, autopsies of deceased Covid-19 patients have reported that in many cases inflammation and tissue damage occurs disconnected from viral presence and distribution[112,125]. This deterioration is attributed to the second phase of illness, which is

characterised by immune-mediated inflammatory lung injury. A dysfunctional innate immune response leads to excessive and uncontrolled inflammation[126]. Macrophages and monocytes have been identified to likely be important cell types in driving this immune system reaction[127] [129]. This profound tissue damage can result in acute respiratory distress syndrome and hypoxaemic respiratory failure[130].

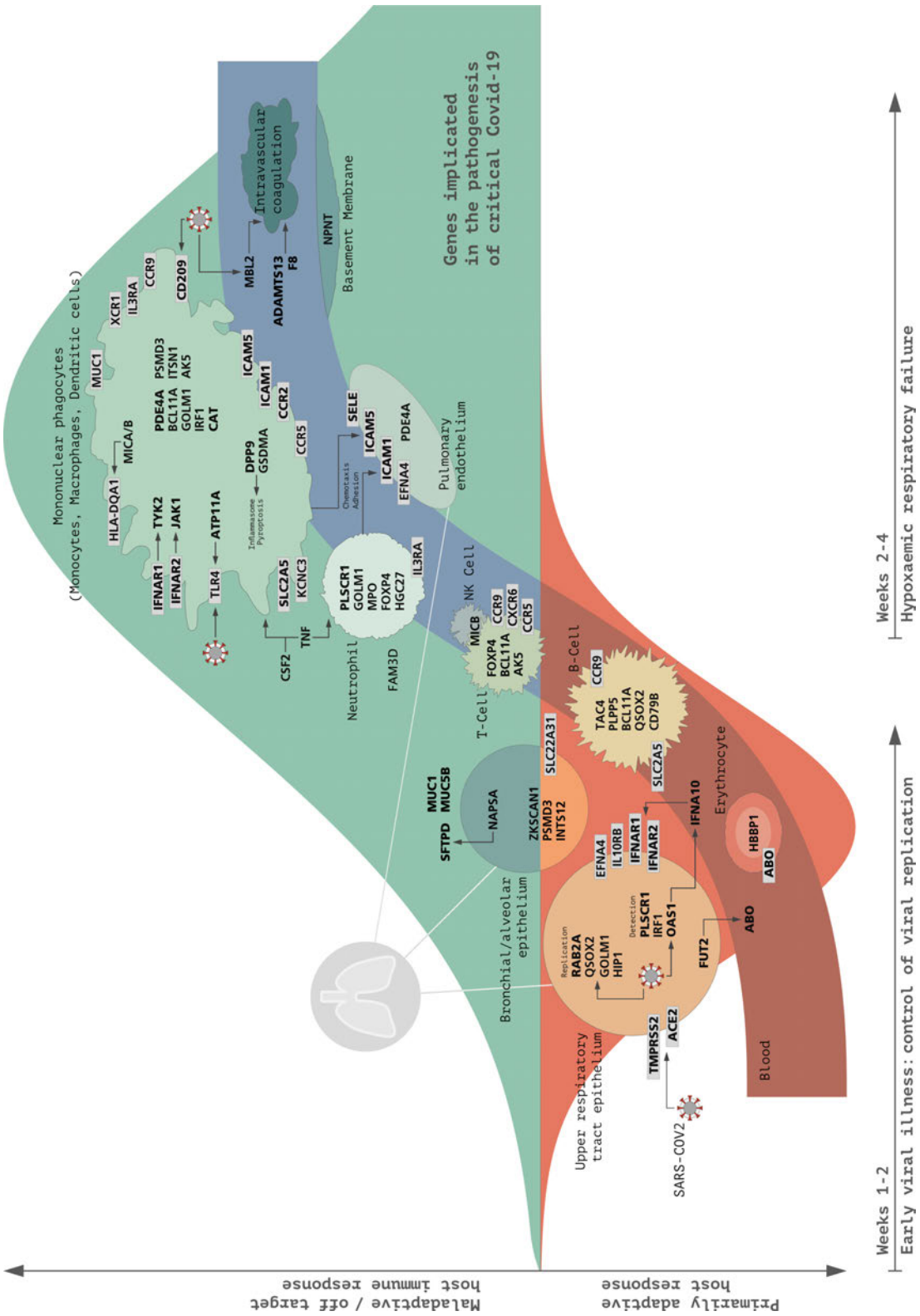


Figure 2: Genes implicated in the pathogenesis of critical Covid-19. Predicted consequences for genetic variants associated with critical Covid-19 as identified by GWAS[131]. This simplified illustration shows the hypothesised impact of the variant-associated proteins in the pathogenesis of Covid-19 over time. Predicted host immune response processes are divided into an adaptive response in controlling the

early viral replication phase (orange), and a maladaptive response contributing to hypoxaemic respiratory failure at a later disease stage (green)[132]. Bold font names are used for a higher level of confidence in the gene identification and suggested biological role. This figure was published in the paper "GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19"[131].

There is a big genetic component to risk of death from infectious disease[133]. Covid-19 provides a particularly strong example for the power of genetic studies to help gain biological insight into infectious diseases and result in clinical impact. GWAS determined many genetic variants that confer susceptibility to Covid-19 infection or increase risk of disease severity[5,131,134,135] (Figure 2). This directly led to the trial prioritisation of the drug baricitinib, which subsequently showed therapeutic benefits to hospitalised Covid-19 patients[5,136]. As the inflammatory response to SARS-CoV-2 infection has a major impact on Covid-19 pathogenesis[137,138] and there is pervasive pleiotropy among immune system diseases[139], it is likely that Covid-19 shares genetic associations with other diseases. In the first part of the thesis, I explore pleiotropy between critical Covid-19 and a large range of diseases in an effort to further understand the molecular underpinnings of its pathology and pinpoint potential drug targets.

1.5 Drug repurposing

The discovery of novel drugs is a process that is expensive, time intensive and fraught with failure at all stages[140–142]. Only approximately 10% of trialed drugs are approved for treatment[143]. Therefore, finding new applications for existing therapeutics is an attractive opportunity, offering a shortened time-frame as well as reduced cost and safety concerns. Famously successful applications of drug repurposing include the angina medication sildenafil, now used to treat erectile dysfunction and pulmonary hypertension[144], and thalidomide, which was repurposed from treating morning sickness to targeting multiple myeloma[145]. The unprecedented availability of large-scale "omics" and other biomedical data now offers the potential to predict many more repurposing candidates through computational approaches.

GWAS data is a valuable source of information for drug repurposing efforts. Genetically supported targets increase the likelihood of successful drug application by more than two-fold if the associated gene can be clearly determined[4,146]. GWAS are generally used in drug repurposing to link diseases to potential drug targets; this is often done in combination with other biological data, and intended to be validated by subsequent preclinical animal studies. The specific strategies of these repurposing approaches can

vary, encompassing candidate gene studies, pathway mapping, expression pattern-based approaches, biological networks, machine learning and disease similarity analyses[147,148].

Candidate gene repurposing studies aim to link GWAS associations to top relevant genes, using functional annotation tools[19] or by identifying related changes in gene expression[149], and then ascertain whether an identified gene encodes a known drug target. However, not all gene products are currently druggable. Pathway-based approaches seek to widen the potential of identifying a relevant drug by considering interaction partners and protein cascades related to the disease-associated genes[150]. An increasing number of computational drug repurposing tools are utilising network biology. This type of analysis more closely considers the highly interconnected nature of cellular processes[151] and allows for the integration of many different data sources, such as databases on disease genes, protein-protein interactions, protein pathways, drugs and drug targets[148,152–154].

In the latter part of this thesis I explore the identification of shared associated variants as a means of pinpointing drug repurposing opportunities. Studying genetic links between diseases at the level of the variant can reveal connections that otherwise might stay hidden due to erroneous assumptions about the affected genes in the disease GWAS, as matching GWAS signals to target genes is fraught with difficulties. Li et al. suggested this approach in their molecular network analysis of comorbidities[155]. However, they define shared SNPs based on lead variants in the original GWAS rather than calculating the likelihood of the GWAS signals originating from the same variant in both GWAS. This method is not robust as the GWAS lead variant may not be the true source of the association. Another benefit of integrating information on shared disease genetics into drug repurposing approaches is that it opens up the possibility of detecting new disease connections. By exploring multiple GWAS in combination, the genetics of similar traits can add power to the original study, with the potential to detect additional associated variants that are not identifiable in one GWAS alone. The stringent multiple comparison corrections and limited sample sizes of GWAS likely lead to a number of false negative results[39]. Previous attempts have been made to rescue these associations by looking at all nominally significant variants and testing whether their associated genes are statistically significantly enriched within pathways[156]. However, such an approach runs a much higher risk of mistakenly focusing on the effects of variants that are not truly associated with the disease.

1.6 Thesis outline

My overarching goal is to identify pleiotropic genetic variants shared between diseases in order to pinpoint potential therapeutic targets for critical Covid-19 for future functional validation and find opportunities for drug repurposing.

Chapter 3 explores pleiotropy between critical Covid-19 and other diseases at a genome-wide level, analysing genetic correlations between the traits using high-definition likelihood inference. Genetic correlation quantitatively describes the relationship between two traits based on common genetic architecture, which are averaged across the genome. While genetic correlations with critical Covid-19 had yet to be studied when I embarked on this analysis, they have now been widely reported in the literature[160]. I show genetic correlations between critical Covid-19 and 56 other diseases. Though standard errors for these connections were large, my findings matched previously published results, in particular with regard to the link between critical Covid-19 and cardiovascular disease as well as idiopathic pulmonary fibrosis.

In chapter 4, I investigate regional pleiotropy between critical Covid-19 and other diseases, using colocalization analysis to detect shared associated variants. I established colocalization via HyPrColoc, a method based on a Bayesian algorithm which calculates whether two or more traits share the same associated variant in a genomic region. Seven pleiotropic variants were detected as shared between critical Covid-19 and other diseases, including idiopathic pulmonary fibrosis, asthma, Crohn's disease, systemic lupus erythematosus, psoriasis, rheumatoid arthritis, allergic disease, hypothyroidism and hypertension. Among these results was a novel critical Covid-19 association with a variant in the gene *SLC39A8*, which suggests pathways involving the divalent metal ion transporter ZIP8 as potential therapeutic targets.

Chapter 5 further evaluates the pleiotropic variants identified in the previous chapter by exploring their effects on gene expression in order to establish their potential biological consequence. Genes with impacted expression were identified through colocalization with tissue-specific expression quantitative trait locus data, which measures how genetic variation influences RNA expression levels. This method uncovered 55 genes with altered expression in whole blood, lung, transverse colon or certain subsets of peripheral blood mononuclear cells. Follow up pathway enrichment analysis of these genes pointed to the cellular checkpoint PD-1 as a possible drug target. This protein controls multiple downstream pathways that have been implicated in Covid-19.

Finally, chapter 6 explores whether shared genetic variants can pinpoint drug repurposing opportunities, as common underlying molecular mechanisms between diseases may indicate that a drug used for one disease could successfully treat another. Pleiotropic variants between 228 disease GWAS were identified via genome-wide colocalization analysis using HyPrColoc. Potential repurposing candidates were selected by finding drugs currently used or in clinical trials for one of two diseases sharing a variant and testing if the drug targets matched genes functionally connected to the shared variant. This method identified 3625 drug repurposing candidates, suggesting an untapped potential in drug repurposing opportunities. Notably, there was particularly strong support for the use of TYK2 inhibitors in the treatment of autoimmune hypothyroidism.

Chapter 2
Methods and Data

2.1 Datasets

The following datasets were utilised over the course of my doctoral work:

Genome-Wide Association Studies (GWAS)

2.1.1 Critical Covid-19

For critical Covid-19 I used our GenOMICC (Genetics of susceptibility and mortality in critical care) consortium whole genome sequencing GWAS[134]. The phenotype for this GWAS was selected using the clinical definition of critical Covid-19, focusing on intensive care patients with hypoxaemic respiratory failure with acute lung injury. Cases were recruited by GenOMICC, controls consisted of mild Covid-19 cases recruited by GenOMICC and individuals from the 100,000 Genomes project[161]. For my work I used the European ancestry version, consisting of 5,989 cases and 42,891 controls. The GWAS was lifted over to genome build GRCh37/HG19 using the Im lab's Harmonization and Imputation tools[162], built on UCSC liftOver[163].

2.1.2 Diseases of the immune system

I obtained a dataset of 54 immune system disease GWAS to explore their connections with critical Covid-19, as the inflammatory response to SARS-CoV-2 infection has a major impact on COVID-19 pathogenesis[137,138]. GWAS were identified using the MRC IEU GWAS database[165] and downloaded directly from the OpenGWAS server. The GWAS VCF file format was converted to TSV format for easier use with other GWAS files and the analysis software. For UK Biobank search results, I downloaded the GWAS directly from the Neale lab's own website instead[166]. Additionally, I acquired GWAS for two traits of interest that were not available via the described databases separately: Two GWAS for Churg-Strauss syndrome /eosinophilic granulomatosis with polyangiitis

(EGPA) by Lyons et al.[167] were downloaded via the NHGRI-EBI GWAS Catalog[168], under accession numbers GCST009248 and GCST009249 and a meta-analysis of 3 GWAS of idiopathic pulmonary fibrosis by Allen et al.[169] was kindly shared with me by the authors.

The following search terms were used to select relevant GWAS on the MRCIEU OpenGWAS database:
achalasia, Addison's disease, adult Still's disease, agammaglobulinemia, allergic disease, allergy, alopecia areata, amyloidosis, ankylosing spondylitis, anti-GBM nephritis, anti-TBM nephritis, antiphospholipid syndrome, arthritis, asthma, autoimmune angioedema, autoimmune dysautonomia, autoimmune encephalomyelitis, autoimmune hepatitis, autoimmune inner ear disease, autoimmune myocarditis, autoimmune oophoritis, autoimmune orchitis, autoimmune pancreatitis, autoimmune retinopathy, autoimmune urticaria, axonal & neuronal neuropathy, bacterial, Baló disease, Behcet's disease, benign mucosal pemphigoid, bronchitis, bullous pemphigoid, Castleman disease, celiac disease, Chagas disease, chickenpox, chronic bronchitis, emphysema, chronic inflammatory demyelinating polyneuropathy, chronic recurrent multifocal osteomyelitis, Churg-Strauss syndrome, eosinophilic granulomatosis, cicatricial pemphigoid, Cogan's syndrome, cold agglutinin disease, coxsackie myocarditis, CREST syndrome, Crohn's disease, dermatitis, dermatitis herpetiformis, dermatomyositis, Devic's disease, neuromyelitis optica, discoid lupus, diverticulitis, dolitis, Dressler's syndrome, eczema, endometriosis, eosinophilic esophagitis, eosinophilic fasciitis, erythema nodosum, Escherichia coli, essential mixed cryoglobulinemia, Evans syndrome, fibromyalgia, fibrosing alveolitis, gastritis, giant cell arteritis, temporal arteritis, giant cell myocarditis, glomerulonephritis, Goodpasture's syndrome, granulomatosis with polyangiitis, Graves' disease, Guillain-Barre syndrome, Hashimoto's thyroiditis, hemolytic anemia, Henoch-Schonlein purpura, herpes gestationis, pemphigoid gestationis, hidradenitis suppurativa, acne inversa, hypogammalglobulinemia, idiopathic pulmonary fibrosis, IgA nephropathy, IgG4-related sclerosing disease, immune thrombocytopenic purpura , inclusion body myositis, infection, interstitial cystitis, juvenile arthritis, juvenile diabetes, juvenile myositis, Kawasaki disease, Lambert-Eaton syndrome, leukocytoclastic vasculitis, lichen planus, lichen sclerosus, ligneous conjunctivitis, linear IgA disease, lyme disease chronic, Meniere's disease, meningitis, microscopic polyangiitis, mixed connective tissue disease, Mooren's ulcer, Mucha-Habermann disease, multifocal motor neuropathy, multiple sclerosis, myasthenia gravis, myelin oligodendrocyte glycoprotein antibody disorder, myositis, narcolepsy, neonatal lupus, neuromyelitis optica, neutropenia, ocular cicatricial

pemphigoid, oesophagitis, optic neuritis, palindromic rheumatism, pancreatitis, PANDAS (pediatric autoimmune neuropsychiatric disorders, associated with streptococcal infections), paraneoplastic cerebellar degeneration, paroxysmal nocturnal hemoglobinuria, Parry Romberg syndrome, pars planitis, peripheral uveitis, Parsonage-Turner syndrome, pemphigus, pericarditis, peripheral neuropathy, perivenous encephalomyelitis, pernicious anemia, pneumonia, Poems syndrome, polyarteritis nodosa, polyglandular syndromes, polymyalgia rheumatica, polymyositis, postmyocardial infarction syndrome, postpericardiotomy syndrome, primary biliary cholangitis, primary biliary cirrhosis, primary sclerosing cholangitis, progesterone dermatitis, psoriasis, psoriatic arthritis, pure red cell aplasia, pyoderma gangrenosum, Raynaud's phenomenon, reactive arthritis, reflex sympathetic dystrophy, relapsing polychondritis, restless legs syndrome, retroperitoneal fibrosis, rheumatic fever, rheumatoid arthritis, sarcoidosis, Schmidt syndrome, scleritis, scleroderma, sepsis, Sjögren's syndrome, sperm autoimmunity, testicular autoimmunity, staphylococcus, stiff person syndrome, streptococcus, subacute bacterial endocarditis, Susac's syndrome, sympathetic ophthalmia, systemic lupus erythematosus, Takayasu's arteritis, temporal arteritis, giant cell arteritis, thrombocytopenic purpura, thyroid eye disease, Tolosa-Hunt syndrome, transverse myelitis, tuberculosis, type 1 diabetes, ulcerative colitis, undifferentiated connective tissue disease, uveitis, vasculitis, viral, viral hepatitis, virus, vitiligo, Vogt-Koyanagi-harada disease

Due to differences in linkage disequilibrium (LD) between populations of different ancestries, a cross-ancestry investigation would have been impossible with the employed analyses, which make heavy use of local LD structures. The GWAS were therefore limited to European ancestry, which is the most widely available across traits. All GWAS were mapped to the human genome assembly GRCh37/HG19. Where they were not included in the original file, variant rsIDs were added either from the accompanying variants.tsv.bgz file provided by the Neale lab (for Neale lab UK Biobank GWAS) or using a script kindly provided by Dr. Andy Law. To achieve sufficient power to study shared genetics between the diseases, only traits with a permissive minimum effective sample size of 4500 were selected. Effective sample sizes were calculated using the following formula:

$$Neff = \frac{4 \times number\ of\ cases \times number\ of\ controls}{number\ of\ cases + number\ of\ controls}$$

As HyPrColoc requires genetic variants to be present in all analysed GWAS datasets, I only selected GWAS with a minimum overlap of 75% of variants with all other GWAS

in the dataset. Table 1 shows the final selection of the immune system related disease dataset.

This dataset previously contained two additional GWAS: Childhood onset asthma (ebi-a-GCST007800) and adult onset asthma (ebi-a-GCST007799) obtained via the MRC IEU GWAS database using data by by Ferreira et al.[170] The removal of these traits due to flaws in the downloaded data is discussed in chapter 4.

Table 1: Immune system disease GWAS.

Trait	MRCIEU, GWAS Catalog or UKB Neale lab key	Neff	Citation
Eczema	ieu-a-996	31,752	EAGLE consortium[171]
Ulcerative colitis (de Lange)	ebi-a-GCST004133	36,160	de Lange et al.[172]
Inflammatory bowel disease	ieu-a-31	32,372	Liu et al.[173]
Crohn's disease	ieu-a-30	17,029	Liu et al.[173]
Ulcerative colitis (Liu)	ieu-a-32	20,792	Liu et al.[173]
Churg-Strauss syndrome (ANCA-negative)	GCST009249	total N: 7,040	Lyons et al.[167]
Churg-Strauss syndrome (ANCA-positive)	GCST009248	total N: 6,847	Lyons et al.[167]
Allergic disease (asthma, hay fever or eczema)	ebi-a-GCST005038	360,837	Ferreira et al.[174]
Idiopathic pulmonary fibrosis	NaN	13,729	Allen et al.[169]
Systemic lupus erythematosus	ebi-a-GCST003156	13,220	Bentham et al.[175]
Asthma, doctor diagnosed	22127.gwas.imputed_v3.both_sexes	40,885	Neale lab[166]
Hayfever or allergic rhinitis (diagnosed by doctor)	22126.gwas.imputed_v3.both_sexes	64,573	Neale lab[166]
Emphysema/chronic bronchitis (diagnosed by doctor)	6152_6.gwas.imputed_v3.both_sexes	24,134	Neale lab[166]

Trait	MRCIEU, GWAS Catalog or UKB Neale lab key	Neff	Citation
Asthma (diagnosed by doctor)	6152_8.gwas.imputed_v3.both_sexes	147,301	Neale lab[166]
Hayfever, allergic rhinitis or eczema (diagnosed by doctor)	6152_9.gwas.imputed_v3.both_sexes	256,444	Neale lab[166]
Daytime dozing / sleeping (narcolepsy)	1220.gwas.imputed_v3.both_sexes	total N: 359,752	Neale lab[166]
Asthma (self-reported)	20002_1111.gwas.imputed_v3.both_sexes	148,259	Neale lab[166]
Emphysema/chronic bronchitis (self-reported)	20002_1113.gwas.imputed_v3.both_sexes	19,844	Neale lab[166]
Meningitis (self-reported)	20002_1247.gwas.imputed_v3.both_sexes	6,035	Neale lab[166]
Multiple sclerosis (self-reported)	20002_1261.gwas.imputed_v3.both_sexes	5,285	Neale lab[166]
Hayfever/allergic rhinitis (self-reported)	20002_1387.gwas.imputed_v3.both_sexes	77,937	Neale lab[166]
Pneumonia (self-reported)	20002_1398.gwas.imputed_v3.both_sexes	20,431	Neale lab[166]
Bronchitis (self-reported)	20002_1412.gwas.imputed_v3.both_sexes	10,644	Neale lab[166]
Tuberculosis (self-reported)	20002_1440.gwas.imputed_v3.both_sexes	7,037	Neale lab[166]
Eczema/dermatitis (self-reported)	20002_1452.gwas.imputed_v3.both_sexes	36,322	Neale lab[166]
Psoriasis (self-reported)	20002_1453.gwas.imputed_v3.both_sexes	16,573	Neale lab[166]
Malabsorption/coeliac disease (self-reported)	20002_1456.gwas.imputed_v3.both_sexes	6,320	Neale lab[166]
Diverticular disease/diverticulitis (self-reported)	20002_1458.gwas.imputed_v3.both_sexes	16,308	Neale lab[166]
Ulcerative colitis (self-reported)	20002_1463.gwas.imputed_v3.both_sexes	7,623	Neale lab[166]

Trait	MRCIEU, GWAS Catalog or UKB Neale lab key	Neff	Citation
Rheumatoid arthritis (self-reported)	20002_1464.gwas.imputed_v3.both_sexes	15,889	Neale lab[166]
Measles / morbillivirus (self-reported)	20002_1568.gwas.imputed_v3.both_sexes	5,010	Neale lab[166]
Chickenpox (self-reported)	20002_1571.gwas.imputed_v3.both_sexes	5,737	Neale lab[166]
Intestinal infectious diseases	AB1_INTESTINAL_INFECTIONS.gwas.imputed_v3.both_sexes	17,011	Neale lab[166]
Childhood asthma (age<16)	ASTHMA_CHILD.gwas.imputed_v3.both_sexes	7,928	Neale lab[166]
Suggestive for eosinophilic asthma	ASTHMA_EOSINOPHIL_SUGG.gwas.imputed_v3.both_sexes	9,149	Neale lab[166]
Asthma-related pneumonia	ASTHMA_PNEUMONIA.gwas.imputed_v3.both_sexes	23,215	Neale lab[166]
Bronchitis	BRONCHITIS.gwas.imputed_v3.both_sexes	16,929	Neale lab[166]
Noninfectious colitis	COLITNONINFNAS.gwas.imputed_v3.both_sexes	34,894	Neale lab[166]
Pneumonia, organism unspecified	J18.gwas.imputed_v3.both_sexes	18,283	Neale lab[166]
Unspecified acute lower respiratory infection	J22.gwas.imputed_v3.both_sexes	12,451	Neale lab[166]
Chronic gastritis	K11_CHRONGASTR.gwas.imputed_v3.both_sexes	7,125	Neale lab[166]
Other gastritis (incl. Duodenitis)	K11_OTHGASTR.gwas.imputed_v3.both_sexes	40,847	Neale lab[166]
Oesophagitis	K20.gwas.imputed_v3.both_sexes	18,941	Neale lab[166]
Gastritis and duodenitis	K29.gwas.imputed_v3.both_sexes	48,932	Neale lab[166]

Trait	MRCIEU, GWAS Catalog or UKB Neale lab key	Neff	Citation
Ulcerative colitis	K51.gwas.imputed_v3.both_sexes	8,521	Neale lab[166]
Other non-infective gastro-enteritis and colitis	K52.gwas.imputed_v3.both_sexes	34,179	Neale lab[166]
Acute pancreatitis	K85.gwas.imputed_v3.both_sexes	5,150	Neale lab[166]
Other rheumatoid arthritis	M06.gwas.imputed_v3.both_sexes	5,582	Neale lab[166]
Rheumatoid arthritis	M13_RHEUMA.gwas.imputed_v3.both_sexes	6,391	Neale lab[166]
Cystitis	N30.gwas.imputed_v3.both_sexes	6,273	Neale lab[166]
Other/unspecified rheumatoid arthritis	RHEUMA_NOS.gwas.imputed_v3.both_sexes	5,106	Neale lab[166]
Ulcerative colitis, NAS	ULCERNAS.gwas.imputed_v3.both_sexes	7,572	Neale lab[166]
Endometriosis (self-reported)	20002_1402.gwas.imputed_v3.female	11,811	Neale lab[166]
Endometriosis	N80.gwas.imputed_v3.female	5,938	Neale lab[166]

2.1.3 Additional diseases

I additionally obtained a set of 173 general disease GWAS. In cases where it is not specifically stated that the analysis focused on immune system diseases, the immune system disease dataset was combined with the following additional UK Biobank disease GWAS, obtained from the Neale lab website[166]. The same thresholds and criteria as for the immune dataset were applied for the final disease selection.

Table 2: Additional disease GWAS.

Trait	Key	Neff
Acute appendicitis	K35.gwas.imputed_v3.both_sexes	9,552
Acute myocardial infarction	I21.gwas.imputed_v3.both_sexes	23,400
Allergy/hypersensitivity/anaphylaxis (self-reported)	20002_1374.gwas.imputed_v3.both_sexes	10,412
Angina (diagnosed by doctor)	6150_2.gwas.imputed_v3.both_sexes	44,053
Angina (self-reported)	20002_1074.gwas.imputed_v3.both_sexes	44,048
Angina pectoris	I20.gwas.imputed_v3.both_sexes	24,552
Appendicitis (self-reported)	20002_1502.gwas.imputed_v3.both_sexes	12,128
Arthritis (self-reported)	20002_1538.gwas.imputed_v3.both_sexes	11,062
Arthrosis	M13_ARTHROSIS.gwas.imputed_v3.both_sexes	92,999
Asthma	J45.gwas.imputed_v3.both_sexes	6,740
Asthma (hospital admissions)	ASTHMA_HOSPITAL1.gwas.imputed_v3.both_sexes	7,900
Atrial fibrillation (self-reported)	20002_1471.gwas.imputed_v3.both_sexes	11,223
Atrial fibrillation and flutter	I48.gwas.imputed_v3.both_sexes	24,977
Atrophic disorders of skin	L12_ATROPHICSKIN.gwas.imputed_v3.both_sexes	6,994
Barret oesophagus	K11_BARRET.gwas.imputed_v3.both_sexes	7,128
Bladder problem (not cancer) (self-reported)	20002_1201.gwas.imputed_v3.both_sexes	8,509

Trait	Key	Neff
Blood clot in the leg/DVT (diagnosed by doctor)	6152_5.gwas.imputed_v3.both_sexes	28,939
Blood clot in the lung (diagnosed by doctor)	6152_7.gwas.imputed_v3.both_sexes	11,837
Breast cysts (self-reported)	20002_1367.gwas.imputed_v3.female	5,536
Cardiac arrhytmias, COPD co-morbidities	CARDIAC_ARRHYTM.gwas.imputed_v3.both_sexes	34,346
Cataract (self-reported)	20002_1278.gwas.imputed_v3.both_sexes	19,898
Cellulitis	L03.gwas.imputed_v3.both_sexes	16,788
Cerebral infarction	I63.gwas.imputed_v3.both_sexes	9,351
Certain infectious and parasitic diseases	AB1_INFECTIONS.gwas.imputed_v3.both_sexes	29,492
Certain infectious and parasitic diseases	I_INFECT_PARASIT.gwas.imputed_v3.both_sexes	36,231
Cervical spondylosis (self-reported)	20002_1478.gwas.imputed_v3.both_sexes	9,970
Cholecystitis	K81.gwas.imputed_v3.both_sexes	7,679
Cholelithiasis	K80.gwas.imputed_v3.both_sexes	40,854
Cholelithiasis/gall stones (self-reported)	20002_1162.gwas.imputed_v3.both_sexes	23,373
Chronic fatigue syndrome (self-reported)	20002_1482.gwas.imputed_v3.both_sexes	6,606
Chronic ischaemic heart disease	I25.gwas.imputed_v3.both_sexes	49,270
Chronic obstructive airways disease/COPD (self-reported)	20002_1112.gwas.imputed_v3.both_sexes	5,122

Trait	Key	Neff
Chronic sinusitis	J32.gwas.imputed_v3.both_sexes	4,701
Chronic sinusitis (self-reported)	20002_1416.gwas.imputed_v3.both_sexes	8,972
COPD differential diagnosis	COPD_EXCL.gwas.imputed_v3.both_sexes	98,939
COPD, early/later onset	COPD_EARLYANDLATER.gwas.imputed_v3.both_sexes	7,548
Coronary atherosclerosis	I9_CORATHER.gwas.imputed_v3.both_sexes	55,061
Coxarthrosi/arthrosis of hip	COX_ARTHROSIS.gwas.imputed_v3.both_sexes	36,659
Cutaneous abscess, furuncle and carbuncle	L02.gwas.imputed_v3.both_sexes	6,756
Deep venous thrombosis/DVT (self-reported)	20002_1094.gwas.imputed_v3.both_sexes	28,368
Diabetes (self-reported)	20002_1220.gwas.imputed_v3.both_sexes	54,250
Diabetes diagnosed by doctor	2443.gwas.imputed_v3.both_sexes	65,786
Diaphragmatic hernia	K44.gwas.imputed_v3.both_sexes	31,452
Diarrhoea and gastro-enteritis of presumed infectious origin	A09.gwas.imputed_v3.both_sexes	8,592
Diseases of appendix	K11_APPENDIX.gwas.imputed_v3.both_sexes	11,715
Diseases of pulp and periapical tissues	K04.gwas.imputed_v3.both_sexes	6,407
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	III_BLOOD_IMMUN.gwas.imputed_v3.both_sexes	39,251
Diseases of the circulatory system	IX_CIRCULATORY.gwas.imputed_v3.both_sexes	201,476

Trait	Key	Neff
Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	I9_DISVEINLYMPH.gwas.imputed_v3.both_sexes	45,908
Diseases of vocal cords and larynx, not elsewhere classified	J38.gwas.imputed_v3.both_sexes	4,657
Disorders of gallbladder, biliary tract and pancreas	K11_GALLBILPANC.gwas.imputed_v3.both_sexes	53,542
Disorders of lachrymal system	H04.gwas.imputed_v3.both_sexes	5,880
Diverticular disease of intestine	K57.gwas.imputed_v3.both_sexes	48,872
Doctor diagnosed COPD (chronic obstructive pulmonary disease)	22130.gwas.imputed_v3.both_sexes	5,348
Duodenal ulcer	K26.gwas.imputed_v3.both_sexes	5,146
Duodenal ulcer (self-reported)	20002_1457.gwas.imputed_v3.both_sexes	5,971
Dyspepsia	K30.gwas.imputed_v3.both_sexes	29,707
Ear/vestibular disorder (self-reported)	20002_1415.gwas.imputed_v3.both_sexes	12,305
Endocrine, nutritional and metabolic diseases	IV_ENDOCRIN_NUTRIT.gwas.imputed_v3.both_sexes	28,295
Endometriosis, IBD co-morbidity	IBD_ENDOMETRIOSIS.gwas.imputed_v3.female	6,017
Epilepsy (self-reported)	20002_1264.gwas.imputed_v3.both_sexes	11,959
Essential hypertension (self-reported)	20002_1072.gwas.imputed_v3.both_sexes	6,602
Eye/eyelid problem (self-reported)	20002_1242.gwas.imputed_v3.both_sexes	11,271
Fibroblastic disorders	M13_FIBROBLASTIC.gwas.imputed_v3.both_sexes	12,647

Trait	Key	Neff
Fibromyalgia related co-morbidities	FIBRO_COMORB.gwas.imputed_v3.both_sexes	9,161
Fissure and fistula of anal and rectal regions	K60.gwas.imputed_v3.both_sexes	8,387
Ganglion	M13_GANGLION.gwas.imputed_v3.both_sexes	8,900
Gastric ulcer	K25.gwas.imputed_v3.both_sexes	7,299
Gastric/stomach ulcers (self-reported)	20002_1142.gwas.imputed_v3.both_sexes	10,175
Gastro-oesophageal reflux / gastric reflux (self-reported)	20002_1138.gwas.imputed_v3.both_sexes	58,278
Gastro-oesophageal reflux disease	K21.gwas.imputed_v3.both_sexes	41,694
Gastroduodenal ulcer	K11_GASTRODUOULC.gwas.imputed_v3.both_sexes	13,735
GI-bleeding	K11_GIBLEEDING.gwas.imputed_v3.both_sexes	20,636
Glaucoma	H40.gwas.imputed_v3.both_sexes	6,827
Glaucoma (self-reported)	20002_1277.gwas.imputed_v3.both_sexes	14,985
Gout (self-reported)	20002_1466.gwas.imputed_v3.both_sexes	20,399
Haemorrhoids	I84.gwas.imputed_v3.both_sexes	46,786
Heart arrhythmia (self-reported)	20002_1077.gwas.imputed_v3.both_sexes	8,007
Heart attack (diagnosed by doctor)	6150_1.gwas.imputed_v3.both_sexes	32,390
Heart attack/myocardial infarction (self-reported)	20002_1075.gwas.imputed_v3.both_sexes	32,204

Trait	Key	Neff
Heart failure	HEARTFAIL.gwas.imputed_v3.both_sexes	5,598
Heart valve problem/heart murmur (self-reported)	20002_1078.gwas.imputed_v3.both_sexes	9,745
Heart/cardiac problem (self-reported)	20002_1066.gwas.imputed_v3.both_sexes	4,689
Hernia	K11_HERNIA.gwas.imputed_v3.both_sexes	97,188
Hiatus hernia (self-reported)	20002_1474.gwas.imputed_v3.both_sexes	32,590
High blood pressure (diagnosed by doctor)	6150_4.gwas.imputed_v3.both_sexes	283,834
Hordeolum and chalazion	H00.gwas.imputed_v3.both_sexes	5,844
Hypertension	I9_HYPTENS.gwas.imputed_v3.both_sexes	4,931
Hypertension (self-reported)	20002_1065.gwas.imputed_v3.both_sexes	277,286
Hypertensive diseases	I9_HYPERTENSION.gwas.imputed_v3.both_sexes	5,233
Hyperthyroidism/thyrotoxicosis (self-reported)	20002_1225.gwas.imputed_v3.both_sexes	10,837
Hypothyroidism/myxoedema (self-reported)	20002_1226.gwas.imputed_v3.both_sexes	66,875
Inguinal hernia	K40.gwas.imputed_v3.both_sexes	50,674
Inguinal hernia (self-reported)	20002_1513.gwas.imputed_v3.both_sexes	5,963
Iron deficiency anaemia	D3_ANAEMIA_IRONDEF.gwas.imputed_v3.both_sexes	14,985
Iron deficiency anaemia (self-reported)	20002_1330.gwas.imputed_v3.both_sexes	7,908

Trait	Key	Neff
Irritable bowel syndrome (self-reported)	20002_1154.gwas.imputed_v3.both_sexes	33,341
Ischaemic heart disease, wide definition	I9_IHD.gwas.imputed_v3.both_sexes	78,610
Ischaemic Stroke, excluding all haemorrhages	I9_STR_EXH.gwas.imputed_v3.both_sexes	13,134
Joint disorder (self-reported)	20002_1295.gwas.imputed_v3.both_sexes	6,938
Kidney stone/ureter stone/bladder stone (self-reported)	20002_1197.gwas.imputed_v3.both_sexes	11,361
Major coronary heart disease event	I9_CHD.gwas.imputed_v3.both_sexes	39,486
Migraine (self-reported)	20002_1265.gwas.imputed_v3.both_sexes	41,332
Mouth ulcers	6149_1.gwas.imputed_v3.both_sexes	132,245
Muscle/soft tissue problem (self-reported)	20002_1297.gwas.imputed_v3.both_sexes	11,357
Myocardial infarction	I9_MI.gwas.imputed_v3.both_sexes	27,527
Nasal polyps (self-reported)	20002_1417.gwas.imputed_v3.both_sexes	6,320
Nasal/sinus disorder (self-reported)	20002_1413.gwas.imputed_v3.both_sexes	5,562
Non-rheumatic valve diseases	I9_NONRHEVALV.gwas.imputed_v3.both_sexes	6,395
Nontraumatic intracranial haemmorrhage	I9_INTRACRA.gwas.imputed_v3.both_sexes	4,995
Oesophagitis/barretts oesophagus (self-reported)	20002_1139.gwas.imputed_v3.both_sexes	4,562
Osteoarthritis (self-reported)	20002_1465.gwas.imputed_v3.both_sexes	110,185
Osteoporosis (self-reported)	20002_1309.gwas.imputed_v3.both_sexes	22,580

Trait	Key	Neff
Other anaemias	D3_OTHERANAEMIA.gwas.imputed_v3.both_sexes	16,972
Other arthritis	RHEU_ARTHRITIS_OTH.gwas.imputed_v3.both_sexes	4,832
Other arthrosis	M13_ARTHROSIS_OTH.gwas.imputed_v3.both_sexes	20,376
Other bacterial diseases	AB1_OTHER_BACTERIAL.gwas.imputed_v3.both_sexes	7,805
Other cataract	H26.gwas.imputed_v3.both_sexes	43,808
Other chronic obstructive pulmonary disease	J44.gwas.imputed_v3.both_sexes	6,098
Other diseases of anus and rectum	K62.gwas.imputed_v3.both_sexes	53,394
Other diseases of intestine	K63.gwas.imputed_v3.both_sexes	31,448
Other diseases of oesophagus	K22.gwas.imputed_v3.both_sexes	21,642
Other diseases of stomach and duodenum	K31.gwas.imputed_v3.both_sexes	9,434
Other diseases of the digestive system	K11_OTHDIG.gwas.imputed_v3.both_sexes	27,700
Other disorders of bladder	N32.gwas.imputed_v3.both_sexes	16,753
Other disorders of eyelid	H02.gwas.imputed_v3.both_sexes	16,972
Other disorders of urinary system	N39.gwas.imputed_v3.both_sexes	40,971
Other functional intestinal disorders	K59.gwas.imputed_v3.both_sexes	14,695

Trait	Key	Neff
Other ILD-related CVD-co-morbidities	OTHER_ILD_CVD_COMORB.gwas.imputed_v3.both_sexes	9,958
Other joint disorder (self-reported)	20002_1467.gwas.imputed_v3.both_sexes	9,047
Other neurological problem (self-reported)	20002_1434.gwas.imputed_v3.both_sexes	6,007
Other or unspecified ileus, impaction or obstruction	K11_OTHILEUS.gwas.imputed_v3.both_sexes	5,380
Other pulmonary diagnosis	PULMONARYDG.gwas.imputed_v3.both_sexes	94,390
Other renal/kidney problem (self-reported)	20002_1405.gwas.imputed_v3.both_sexes	5,991
Other serious eye condition	6148_6.gwas.imputed_v3.both_sexes	24,876
Other viral diseases	AB1_OTHER_VIRAL.gwas.imputed_v3.both_sexes	4,705
Ovarian cyst/s (self-reported)	20002_1349.gwas.imputed_v3.female	11,691
Palmar fascial fibromatosis/Dupuytren	M13_DUPUTRYEN.gwas.imputed_v3.both_sexes	11,696
Paralytic ileus and intestinal obstruction	K11_ILEUS.gwas.imputed_v3.both_sexes	8,359
Paralytic ileus and intestinal obstruction without hernia	K56.gwas.imputed_v3.both_sexes	7,366
Paroxysmal tachycardia	I47.gwas.imputed_v3.both_sexes	6,709
Peripheral artery disease	I9_PAD.gwas.imputed_v3.both_sexes	4,903
Peritonitis (self-reported)	20002_1190.gwas.imputed_v3.both_sexes	5,026
Phlebitis and thrombophlebitis	I80.gwas.imputed_v3.both_sexes	9,098

Trait	Key	Neff
Pleurisy (self-reported)	20002_1125.gwas.imputed_v3.both_sexes	4,864
Polyarthropathies	M13_POLYARTHROPATHIES.gwas.imputed_v3.both_sexes	12,981
Pulmonary embolism	I26.gwas.imputed_v3.both_sexes	8,422
Pulmonary embolism (self-reported)	20002_1093.gwas.imputed_v3.both_sexes	11,896
Senile cataract	H25.gwas.imputed_v3.both_sexes	24,884
Sleep apnoea	G6_SLEEPAPNO.gwas.imputed_v3.both_sexes	8,940
Sleep apnoea (self-reported)	20002_1123.gwas.imputed_v3.both_sexes	4,637
Spinal stenosis	M13_SPINSTENOSIS.gwas.imputed_v3.both_sexes	7,600
Spine arthritis/spondylitis (self-reported)	20002_1311.gwas.imputed_v3.both_sexes	12,506
Spondylopathies	M13_SPONDYLOPATHY.gwas.imputed_v3.both_sexes	17,358
Stroke	C_STROKE.gwas.imputed_v3.both_sexes	24,166
Stroke (diagnosed by doctor)	6150_3.gwas.imputed_v3.both_sexes	22,002
Stroke (self-reported)	20002_1081.gwas.imputed_v3.both_sexes	19,085
Stroke, excluding SAH (spontaneous subarachnoid hemorrhage)	I9_STR.gwas.imputed_v3.both_sexes	15,165
Stroke, including SAH (spontaneous subarachnoid hemorrhage)	I9_STR_SAH.gwas.imputed_v3.both_sexes	17,713
Synovitis and tenosynovitis	M65.gwas.imputed_v3.both_sexes	11,160

Trait	Key	Neff
Tonsillitis (self-reported)	20002_1598.gwas.imputed_v3.both_sexes	10,203
Transient cerebral ischaemic attacks and related syndromes	G45.gwas.imputed_v3.both_sexes	6,942
Transient ischaemic attack (self-reported)	20002_1082.gwas.imputed_v3.both_sexes	5,455
Type 2 diabetes (self-reported)	20002_1223.gwas.imputed_v3.both_sexes	9,110
Ulcer of oesophagus	K11_OESULC.gwas.imputed_v3.both_sexes	12,286
Unstable angina pectoris	I9_UAP.gwas.imputed_v3.both_sexes	13,625
Urethral stricture	N35.gwas.imputed_v3.both_sexes	8,102
Urinary tract infection/kidney infection (self-reported)	20002_1196.gwas.imputed_v3.both_sexes	6,760
Uterine fibroids (self-reported)	20002_1351.gwas.imputed_v3.female	21,430
Varicose veins (self-reported)	20002_1494.gwas.imputed_v3.both_sexes	5,177
Varicose veins of lower extremities	I83.gwas.imputed_v3.both_sexes	34,202
Venous thromboembolism	I9_VTE.gwas.imputed_v3.both_sexes	18,244
Ventral hernia	K43.gwas.imputed_v3.both_sexes	8,940

Expression datasets

In order to explore the effects of identified pleiotropic variants on gene expression I used tissue or cell specific expression quantitative trait locus (eQTL) data from several sources. In all cases, cis-eQTL data was used.

2.1.4 GTEx

Genotype-Tissue Expression (GTEx) v.8 summary statistics for lung tissue, sigmoid and transverse colon were obtained from the GTEx Portal on 09.05.2022[59]. The summary statistics were calculated from RNA sequencing data of 578 lung tissue samples, 373 sigmoid colon tissue samples and 406 transverse colon samples.

2.1.5 eQTLGen

Whole blood eQTL summary statistics (cis-eQTL) were downloaded from the eQTLGen Consortium website on 10.05.2022[85]. The expression data was calculated using samples from 31,684 individuals.

2.1.6 OneK1K

Single-cell RNA sequencing summary statistics data for peripheral blood mononuclear cells was obtained from OneK1K[81]. Data was collected from 982 individuals and split into 14 different cell types: CD4+ naïve and ventral memory T cells, CD4+ effector memory and central memory T cells, CD4+ SOX4 T cells, CD8+ naïve and central memory T cells, CD8+ effector memory T cells, CD8+ S100B T cells, natural killer cells, natural killer recruiting cells, immature and naïve B cells, memory B cells, plasma cells, classical monocytes, non-classical monocytes and dendritic cells. On average 1291 cells were sequenced per individual, with 12 out of the 14 distinct cell type populations consisting of an analysis sample size > 930 after quality control.

2.1.7 Drug data

Information on which drugs are approved or in trial to treat the analysed diseases, as well as data on drug target molecules, action types and completed drug trial phases was downloaded from Open Targets between 24.04.2023 and 08.05.2023[176]. Open Targets aggregates this information by mining drug labels through the ChEMBL bioactivity database[177]. GWAS phenotypes were matched with Open Targets phenotypes (Supplementary Table 1).
Neale lab GWAS phenotypes were identified by matching them to either ICD-10 codes (International statistical classification of diseases and related health problems, version : 2016)[178], UK Biobank definitions[179], or FinnGen (release 8) definitions[180] as applicable.

The drug information was filtered to only include the highest phase trial per drug for a disease to avoid duplicated findings.

Reference panels

2.1.8 1000 Genomes

Where not otherwise specified, European data from the 1000 Genomes phase 3 reference panel was used. The reference panel was obtained from the PLINK 2.0 website at https://www.cog-genomics.org/plink/2.0/resources, version: 2016-05-05 primary release (build 37). The file was transformed to PLINK 1 binary format (.bed) and filtered for data of European ancestry using PLINK 2.0.

2.1.9 UK Biobank

For the SuSiE analysis of GWAS created from UK Biobank data, the local linkage disequilibrium matrix was calculated by Dr. Konrad Rawlik using a UK Biobank reference panel. The reference panel was computed using imputed UK Biobank genotypes v3, drawing on data from 380,605 unrelated individuals of European ancestry (as part of UKB project 788).

2.2 Analyses

2.2.1 General Software Information

The analyses were performed in R version 4.0.5, Python version 3.7.9, Jupyter Notebook and JavaScript D3 (via Observable[181]). Affinity Designer was used for illustration and graphic design.

2.2.2 HDL

Genome-wide genetic correlations between GWAS were calculated using high-definition likelihood (HDL) inference[182]. HDL is an extension to the conventional LDSC (linkage disequilibrium score regression) method[183,184]. LDSC relies on the polygenicity-based principle that if a SNP is in LD with a higher number of other SNPs, its likelihood to be correlated with a trait-associated variant is higher, and therefore its association test

statistic will be higher. Using this relationship, the genetic correlation between traits is estimated using the slope from the regression of the product of the GWAS test statistics of the SNPs on their LD score. HDL further improves the precision of the estimated genetic correlations by integrating more information on the LD structure, fitting an additional variance-covariance LD matrix, which improves power by reducing the estimate variance.

The HDL European ancestry reference panel of UK Biobank imputed HapMap3 SNPs was used for these calculations. GWAS had a SNP overlap above 99% with the reference panel as recommended, with the exception of 87.1% overlap for critical Covid-19 and 87.3% overlap for Churg-Strauss syndrome. Significant correlations were defined as $p < 0.0401$ (corrected for a false discovery rate of 5%). For visual display, the genetic correlations were squared (to include negative correlations) and weighted for certainty by multiplication with the inverse of their standard error (correlations with lower standard errors have greater weight, and correlations with higher standard errors are down weighted).

2.2.3 HyPrColoc

The Bayesian algorithm HyPrColoc (Hypothesis Prioritisation for multi-trait Colocalization)[24] uses GWAS summary statistics to detect genetic colocalization across large numbers of traits simultaneously. It can identify subsets of traits among the data which colocalize at distinct associated variants. I performed HyPrColoc analyses to identify genetic colocalization across the genome in the collected GWAS dataset. These analyses were carried out using the R package hyprcoloc version 1.0.

HyPrColoc assesses a local genomic region and calculates the probability of colocalization using two criteria: The regional association probability, which describes the likelihood of all traits having an associated variant within the region and the alignment probability, which describes the likelihood that all trait association signals are due to the exact same genetic variant (Figure 3). If these criteria are not satisfied for all GWAS, HyPrColoc then tests if there are subgroups of traits that share an associated variant.

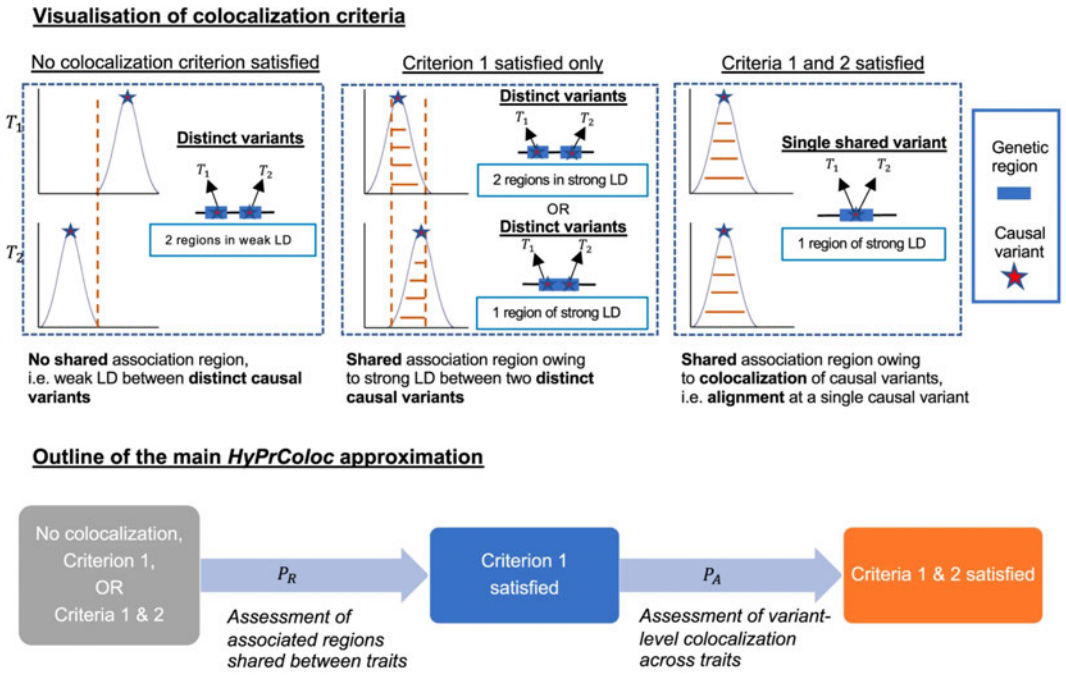


Figure 3: HyPrColoc's colocalization criteria. The top half of the figure shows the basic principle of HyPrColoc, using an example with two traits. In the illustration on the left, the traits do not share an association region, so there is no colocalization. In the scenario depicted in the middle, even though the traits share an association region, their associated variants do not align. Colocalization is only achieved in example three to the right, where both criteria are satisfied. The probability estimations of each criterion are combined to achieve the resulting colocalization probability. This figure was created and published by Foley et al.[24].

Additionally, HyPrColoc makes three assumptions: 1. The local LD structure is crucial to the analysis, therefore all GWAS should be from the same population. I limited my dataset to European ancestry to fulfil this criterion. 2. There can only be one lead variant in a local region. If this assumption is violated, the probability HyPrColoc assigns to a colocalization will be low. Follow up tests, such as fine-mapping to detect multiple signals in the same locus (SuSiE, see below), can be performed to detect or rule out this scenario. 3. There must be high quality information about the associated variant in the dataset, either directly tested in the GWAS or through imputation.

As discussed in the HyPrColoc publication as well as in supplementary information on the method, wrongly assuming that there is no sample overlap and treating the traits as independent gives a much faster yet comparable result to accounting for sample overlap. I therefore did not correct for sample overlap in the input datasets.

As HyPrColoc is a regional method, I split the GWAS into pre-defined LD blocks prior to the analysis, mapping them to the LD blocks identified by Berisa et al. for European

data[185]. These are regions where genetic variants are in high LD with each other, so trait association signals within them can likely be linked back to one or few lead variants. Additionally, variants were filtered to those with minor allele frequencies > 0.01.

HyPrColoc analysis strategy for identifying variants shared between critical Covid-19 and other diseases

For chapter 4, where I take a detailed look at colocalizations between critical Covid-19 and the other diseases, I performed multiple rounds of HyPrColoc, as well as follow up analyses to identify the largest possible number of true positives (see Figure 6 in chapter 4).

For the first HyPrcoloc analysis screen, I set the value of the conditional colocalization prior (which describes the likelihood that a variant that is associated with one trait is associated with a second trait) to the recommended 0.02. This translates to the assumed probability for a variant that is already associated with one trait to be associated with a second trait being 1 in 50, that it is associated with a third trait 1 in 25, and so forth. I then at this stage further inspected all reported colocalizations with critical Covid-19, regardless of the assigned posterior colocalization probability. There are three potential reasons why posterior probability could be low: 1. There is no colocalization in this area. 2. There are multiple traits in a colocalization group and one or more of them have less evidence supporting colocalization, reducing the probability for the whole group even though some links may be strong. 3. One or more of the colocalizing traits have multiple associations in the area, violating HyPrColoc's single associated variant assumption. To distinguish scenarios 2 and 3 from the first and rescue any additional true positives hampered by these circumstances, I performed multiple follow-up analysis.

First, I performed HyPrColoc's sensitivity analysis for all colocalizations with critical Covid-19 to test for cases that fall into scenario 2. This test looks for weak links in a group of colocalizations by investigating the sensitivity to changes in the prior and threshold parameters. The HyPrColoc analysis is repeated multiple times for the region of interest, each time with increasingly strict values for the conditional colocalization prior (0.05, 0.02, 0.01 and 0.005) as well as the two threshold parameters that in combination result in the colocalization probability – the regional association probability (that the traits have at least one associated variant) and the alignment probability (the associated variant is shared between the traits)(PR=PA=0.5,0.6 and 0.7). The result is can be depicted as a heat map where the darker the cell, the more often a trait pair is found to colocalize (see chapter 4, Figure 7 and Figure 8). This helps to easily

identify weak links in a colocalization group, which are then excluded, so only colocalizations meeting a posterior colocalization threshold of > 0.7 remain. This threshold was found to be sufficient in balancing detection of true positive results and exclusion of true negative results by Foley et al[24].

Next, I tested whether the remaining colocalizations would meet the probability threshold using the more stringent conditional colocalization prior of 0.01, a failure of which could suggest a trait with multiple associations in the region. I performed an additional HyPrColoc analysis for the regions of interest to check for traits that would not colocalize with a probability of > 0.7 under the stricter prior. Any traits that failed to meet this threshold were then subjected to a SuSiE analysis (see below) to test whether they had multiple associated variants in the region, and if so, if either of those variants were shared between the traits.

I performed the entire analysis pipeline twice, the first time focusing only on the connection between critical Covid-19 and the 54 traits in the immune system GWAS dataset, the second time analysing its colocalizations with all 227 disease GWAS. The reasoning behind this was twofold: From a biological point of view, immune system diseases were a particularly relevant choice in exploring connections with critical Covid-19, as the inflammatory response to SARS-CoV-2 infection has a major impact on COVID-19 pathogenesis[137,138]. Regarding the methodology, it was shown in the publication by Foley et al. that HyPrColoc's analysis power slightly decreases with a large number of traits. Additionally, since HyPrColoc operates on the assumption that there is a single associated variant for a trait in a given region, for traits with multiple associated regional variants this can lead to some of their colocalizations being missed, as once a trait is found to colocalize it is taken out of subsequent analyses rounds testing that region.

All HyPrColoc analyses checking the entire genome were performed in two iterations. This was done to achieve the best chance of detecting the true associated variant as well as finding as many of the colocalizing traits as possible. HyPrColoc requires a complete overlap in variants between the assessed GWAS, so only variants present in every single GWAS can be analysed. However, the more variants can be included in the analysis, the higher the chance of identifying the true associated variant. The first iteration enabled me to probe a larger number of variants for those GWAS, while the second iteration added more traits. In cases where different lead variants were found for colocalization groups between iteration one and two, the variant in iteration one was considered the

true lead variant. For the analysis exclusive to immune system disease, the first iteration was restricted to critical Covid-19, the Neale lab UK Biobank GWAS, as well as ieu-a-996 (Eczema), ebi-a-GCST004133 (Ulcerative colitis (de Lange)), ieu-a-30 (Crohn's disease), ieu-a-31 (Inflammatory bowel disease) and ieu-a-32 (Ulcerative colitis (Liu)) resulting in 49 included traits and 5713907 included variants. The second iteration included all 54 immune system disease traits and 4359796 variants. For the analysis looking at all disease GWAS, the first iteration included critical Covid-19 and the Neale lab UK Biobank GWAS (218 traits, 6217133 variants) and the second iteration included all GWAS (228 traits, 4359796 variants).

HyPrColoc analysis strategy for identifying pleiotropic variants between 228 diseases

For chapter 6, where the large volume of investigated colocalizations (analysing all 228 diseases) was prohibitive to an in depth investigation of potential multi-associated scenarios, I used the all-disease HyPrColoc analysis performed for chapter 4 (with conditional colocalization prior 0.02) and reported results for traits above a posterior colocalization probability of 0.7, which was found to generally work well to minimise false positives and identify true positives in the HyPrColoc publication. I additionally performed the HyPrColoc sensitivity analysis for all detected colocalizations, to identify cases where although the colocalization group as a whole did not reach the probability threshold, a subset of traits colocalized with a probability of > 0.7. Finally, I corrected the colocalizations involving critical Covid-19 to account for which colocalizations persisted and which were flagged as likely false positives after more in depth follow up investigations in the more detailed chapter 4 analysis.

2.2.4 Coloc

In order to pinpoint genes with expression that may be influenced by the identified pleiotropic variants, I used the coloc R package[44] to establish colocalization between GWAS and gene expression data. Coloc uses a Bayesian algorithm to calculate the likelihood of five different outcomes: That there are no associated variants with either trait, that there is only association with trait 1, that there is only association with trait 2, that both traits are associated but have different lead variants or that both traits are associated with the same variant.

The GWAS signals I sought to colocalize with expression data were identified in prior HyPrColoc analysis, which detected colocalization between multiple GWAS. For

colocalization with expression data, I chose to use the GWAS with the lowest p-value in the HyPrColoc colocalization group. The analysis region was chosen based on the LD blocks utilised in the prior HyPrColoc analysis.

As my aim was a hypothesis-generating study, all nominally supported colocalizations were potentially of interest for follow up investigation. While this leads to a higher risk of false positives it reduces the risk of missing true positives, which is of particular interest in a crisis situation such as the Covid-19 pandemic. For this purpose, I classified colocalizations in three different categories according to the level of evidence supporting them. GWAS and expression quantitative trait locus (eQTL) signals were considered as colocalizing with strong support if they had a colocalization probability of > 0.9 at a default prior probability of 10^{-5}. Colocalizations with a probability > 0.7 were classified as having medium support. Two were used to establish colocalizations with low support: 1. The probability of colocalization was > 0.5 at a prior colocalization probability of 5×10^{-5} 2. The probability that both traits were associated with different variants was not the leading hypothesis at a prior colocalization probability of 10^{-5} (default threshold). For regions with multiple associations I used coloc.susie instead (see below).

2.2.5 SuSiE

SuSiE (Sum of Single Effects)[45] is a genetic fine-mapping method that tests for multiple independent signals in a local region. The method calculates the sum of individual regressions each representing one unidentified associated variant. The main strength of this analysis is that it separates the statistical support for each variant conditional on the signal being considered. In this way, SuSiE can establish credible sets of variants based on the strength of evidence by which they are responsible for a local signal. This can be used to improve the accuracy of colocalization analysis by first employing SuSiE to distinguish between multiple signals and then Coloc to test for colocalization between all pairs of signals between two traits. This combinatory approach enables for multiple signals to be evaluated simultaneously rather than stepwise, thereby improving accuracy, while also allowing priors that account for the likelihood of shared variants to be considered, as for example variants associated with one disease are more likely to also be associated with another[21]. After SuSiE establishes the signals for each trait, Coloc uses a Bayesian algorithm to calculate the likelihood of five different outcomes: That there are no associated variants with either trait, that there is only association with trait 1, that there is only association with trait 2, that both traits are associated but have different lead variants or that both traits are associated with the same variant.

I performed SuSiE using the R package coloc version 5.1.0.1. I used SuSiE for two purposes: 1. In chapter 4 to detect whether traits that were flagged as such in the HyPrColoc analysis (see above) actually were likely to have multiple associations, and how that affected the regional colocalization between traits. 2. In chapter 5 to identify genes in connection with these GWAS variants.

GWAS-GWAS colocalization

SuSie was used to establish whether two GWAS colocalized in a region for cases where the HyPrcoloc analysis flagged them as potentially having multiple associated variants (see above). Beta, varbeta (squared standard error) and location data from the GWAS summary statistics for the LD block used in the HyPrColoc analysis were evaluated to test the entire region in question. Variants were filtered to those with minor allele frequencies >0.01. The corresponding correlation matrices for the included variants were computed with PLINK v.1.9, using data from the 1000 Genomes reference panel for GWAS that did not originate from the UK Biobank. For UK Biobank GWAS, Dr. Konrad Rawlik kindly provided me with correlation matrices calculated using a UK Biobank reference panel. To allow for better evaluation of the necessary effect sizes, GWAS sample sizes were provided to the runsusie function, which identifies local lead variants per GWAS. GWAS were considered to colocalize at a variant if the probability for a shared variant as calculated by the subsequent coloc.susie function was > 0.9.

GWAS-eQTL colocalization

For the colocalization between GWAS and expression signals in regions where I had identified the GWAS to be associated with multiple separate variants, I used SuSiE and its subsequent function coloc.susie instead of Coloc. The data was processed as outlined above. The 1000 Genomes reference panel was used for calculating the correlation matrices for the GTEx expression data. Both 1000 Genomes and UK Biobank reference panels were tested for calculating the correlation matrices for the eQTLGen expression data. The OneK1K expression data could not be used for the multi-associated regions as the necessary information (the standard error to the beta) was not included in the available dataset. The same criteria for colocalization as in the Coloc analysis were used to evaluate the results.

2.2.6 Pathway enrichment analysis

I used pathway enrichment analysis to further explore the genes which had been identified as differently expressed due to variants associated with critical Covid-19 and its colocalizing diseases. Multiple web services were used for comparison: g:Profiler g:GOSt[186] (version e110_eg57_p18_4b54a898, organism hsapiens), Enrichr[187,188] and GeneTrail[189] (version 3.2). The Benjamini–Hochberg procedure, which uses a modified sequential Bonferroni correction for multiple comparisons[190], was used to control the false discovery rate for Enrichr and GeneTrail and the g:SCS algorithm, which accounts for hierarchical terms in the probed data in addition to correcting for multiple testing, was used for g:Profiler. Pathways were considered enriched for the gene list when the adjusted P-value was <0.05. The analysis was performed on 26.09.2023 with the following input gene list: AP4M1, APBA3, ATP11A, ATP5MF, BANK1, BDH2, BUD31, CNPY4, CPSF4, DAPK3, DCUN1D2, DPP9, EBI3, ENSG00000242687, ENSG00000254531, ENSG00000268536, EPO, FEM1A, FIS1, GNB2, GPC2, KDM4B, LAMP1, MANBA, MAP2K2, MCF2L, MEPCE, MICOS13, MOSPD3, MUC12-AS1, NFKB1, PCOLCE, PIAS4, PPP1R35, PPP3CA, PTPRS, SAFB2, SH3GL1, SIRT6, SLC39A8, SLC9B1, SRRT, STAP2, TICAM1, TJP3, TMCO3, TMEM225B, TNFAIP8L1, TRIM4, TRIM56, TRIP6, TUBGCP3, ZCWPW1, ZNF789, ZSCAN21.

2.2.7 Evaluation of variant effects

To ascertain intron variant effects, I used the RegSNPs-intron web application, which evaluates the disease-causing probability of intron variants based on their impact on splicing regulation and the resulting effect on protein-structure features[191]. To test for variant effects on micro RNA target sites I queried the miRNASNP-v3 database[192]. All tests were performed on 04.10.2023.

2.2.8 Identification of drug repurposing candidates

I used the results of my HyPrColoc analysis across all diseases to prioritise drugs that could be potential candidates for repurposing. Where two diseases were colocalizing, I identified drugs that were only currently used or in trial for one of them (see drug dataset above). The next step was to then identify whether the drug's target molecule was associated with the disease colocalization. I ascertained whether the protein products of

colocalization-associated genes were targeted by the drug. Additionally, the products of genes in the same functional pathway might be druggable to achieve a similar effect, so I also obtained a wider functional genetic network and tested if the drug target was among them.

Genes associated with the colocalization SNPs were identified using the NCBI dbSNP database on 27.05.2023[193]. Of the 397 unique SNPs involved in the colocalizations, 122 did not have associations in the dbSNP database and were excluded from this analysis. For those variants where the gene was encoding an antisense RNA, the RNA's corresponding mRNA gene was used for the analysis. 29 colocalization variants were associated with genes for non-coding RNAs; these were kept in the analysis but not included in the subsequent STRING search.

The wider functional network for the 244 identified colocalization-associated protein genes was obtained from the STRING database[194], downloading full networks (functional and physical protein associations) with a maximum of 50 nodes via the API on 29.05.2023 (STRING version 11.5). Where a different version of the gene name was used by dbSNP and STRING, both name versions were included in the drug repurposing candidate analysis. The data download was restricted to human origin, which resulted in a lack of network information for the *MUC19* gene, which was only available on the STRING database for other organisms.

To ascertain that gene matches between the drug target data downloaded from the Open Targets database and the functional network data downloaded from the STRING database could be correctly identified, the drug target genes were tested against preferred gene nomenclature in STRING via the STRING API. Where human data was available for the gene in the STRING database, only two genes had a different preferred name: *GBA1* was called *GBA* in the STRING data and *OPRL1* was called *PNOC* in the STRING data. Out of these two genes, only *GBA* was present in the tested functional networks and was renamed to *GBA1* prior to the drug repurposing candidate identification analysis.

Drugs with a trial status of withdrawn or terminated were included in the drug repurposing analysis, so as to not identify drugs for repurposing that in actuality had already been tested and found not to have the desired effect. After the repurposing drug candidates were identified, drug matches based on these trials were excluded from the results and from further evaluation.

2.2.9 Bootstrapping

I used bootstrapping to answer two questions in chapter 6: Do diseases with colocalizations have GWAS with a higher effective sample size than those without? Do disease colocalizations predict the likelihood of shared drugs?

For the first question I compared the two distributions of effective sample sizes for the GWAS data (excluding GWAS without available effective sample size – Daytime dozing / sleeping (narcolepsy), Churg-Strauss syndrome (ANCA-positive) and (ANCA-negative)). In order to determine whether a colocalization between diseases affected their probability of sharing at least one drug, I calculated the probability of a disease to share at least one drug with another disease given that they share a genetic link, and the probability of a disease to share at least one drug with another disease given that they do not share a genetic link. The probability of sharing at least one drug was chosen over shared number of drugs to normalise for the large variation of known drugs per disease. I then tested the two resulting probability distributions. One factor that could bias this analysis is broadly used drugs with low disease-specificity; I tested the distribution of such drugs and found no enrichment of them in colocalizing versus non-colocalizing disease pairs (see supplementary figure 1 in the appendix).

To test for significant difference, I bootstrapped the median difference between the distributions[195,196]. The bootstrap method was chosen over parametric statistical tests as the underlying effective sample size and probability distributions were not normally distributed and exhibited a floor and ceiling effect respectively (see Figure 19 in chapter 6 and supplementary figure 2 in the appendix), and over non-parametric tests as these tend to be less reliable when floor or ceiling effects are present[197].

In each bootstrap replicate, the two distributions were resampled 50 times with replacement and subtracted and the median difference constituted the bootstrap sample. This procedure was replicated 10000 times to yield a distribution of median differences. The distribution of differences was visually inspected to determine whether the bootstrap had converged (i.e. differences were normally distributed as per the central limit theorem[198]). A confidence interval (CI) for the median difference was drawn using the percentile method, by which the 2.5th and 97.5th percentiles of the distribution comprised the 95% CI[199]. Significant differences were determined by the 95% CI[196]. Median difference distributions in which the 95% CI overlapped zero (i.e. the 2.5th and 97.5th percentiles had different arithmetic signs) indicated no significant difference, and

vice versa. The median of the bootstrapped distribution was reported as the effect size. P-values were derived from the 95% CI using the formula described by Altman and Bland (2011)[200].

2.3 Visualisation

2.3.1 Network displays

The disease networks were created using JavaScript D3 via Observable[181].

2.3.2 Protein structures

Protein structures were based on predictions by AlphaFold[201,202] and displayed using PyMOL 2.5. PyMOL was also used to interrogate amino acid changes and local interaction between amino acid residues.

Chapter 3
Covid-19 and the genetic disease network

3.1 Introduction

The Covid-19 pandemic created an urgent need for a swift understanding of the underlying disease mechanisms in order to find effective treatments. Genome-wide association studies contributed to these efforts by determining many genetic variants that confer susceptibility to the disease[135], leading to the trial prioritisation of the drug baricitinib that subsequently showed therapeutic benefits to hospitalised Covid-19 patients[5,136]. A promising follow-up avenue of investigation to learn more about disease biology and potentially identify additional associated genetic variants is to explore pleiotropy between Covid-19 and other diseases.

Pleiotropy occurs when a gene or a genetic variant affects multiple traits and is widespread throughout the genome[21,183]. A common means of exploring pleiotropy between traits is through genetic correlation analysis[182,183,203]. Genetic correlation quantitatively describes the relationship between two traits based on common genetic architecture, which are averaged across the genome. Although correlation alone cannot identify causal mechanisms linking traits[204], it provides evidence for potentially causal relationships, which can be further explored to expand our understanding of disease biology. Follow up analyses can help prioritise associated variants[205] or improve power to identify new associated variants through meta analysis of the correlated traits[23]. Further analysis can also provide additional evidence of causal relationships[206]. Variants contributing to the correlation can be identified[207], which is useful in improving interpretation of shared disease biology underlying the correlation.

In this chapter, I use high-definition likelihood (HDL) inference[182] to calculate genetic correlations between disease GWAS, with a focus on genetic correlations with critical Covid-19. HDL was created as an extension to the conventionally used linkage disequilibrium score regression (LDSC) analysis[183,184]. LDSC estimates genetic correlation between traits based on the fact that for polygenic traits, if a SNP is in linkage disequilibrium (LD) with a higher number of other SNPs, its likelihood to be

correlated with an associated variant is higher and therefore its association test statistic will be higher. HDL further improves upon this calculation by integrating additional LD structure information.

The Covid-19 pandemic created rapidly evolving circumstances, and much of the scientific effort worldwide focused on expanding our understanding of the disease in search of medical intervention strategies. While genetic correlations with critical Covid-19 had yet to be studied when I embarked on this analysis, they have now been widely reported in the literature[160]. This chapter will provide a brief summary of my results, which following chapters will then build on with novel investigations exploring pleiotropy between critical Covid-19 and other diseases in local genetic areas.

3.2 Results

In this chapter, I used genetic correlation analysis to explore pleiotropy between Covid-19 and other diseases. A good overview of the identified disease relationships can be gained from a network graph, which allows for visualisation of the strength and density of the connections. Figure 4 shows the weighted genetic correlation network, which was highly connected. Correcting for a false discovery rate of 5% ($p < 0.0401$), there were 9018 correlations linking all 228 disease GWAS. Of these, 1883 correlations between 180 diseases were identified with enough certainty to meet the much more stringent Bonferroni correction threshold adjustment for multiple testing ($p < 0.00000096$). An interactive correlation network (Supplementary Interactive Figure 1) can be accessed through the digital appendix at https://marie-zz.github.io/digital_appendix/.

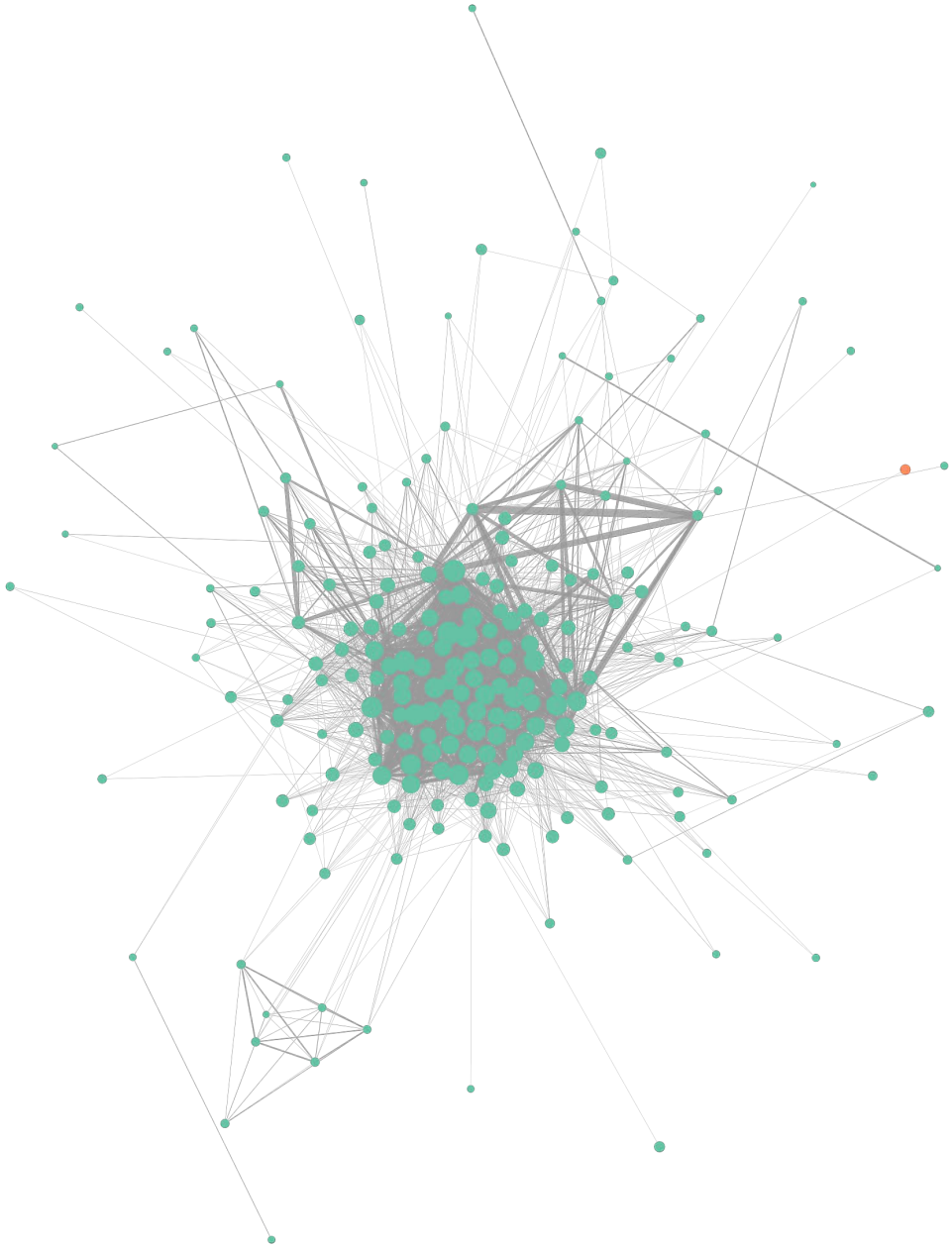


Figure 4: The genetic correlation disease network. Nodes represent the disease GWAS, edges the genetic correlations between them. Correlations were squared to include negative correlations. To account for uncertainty, the correlations were weighted by multiplying them with the inverse of their standard error (correlations with lower standard errors have greater weight, and correlations with higher standard errors are down weighted)(range after weighting was 3.48-810). For display purposes, a threshold of weighted genetic correlations > 15 is used, which excludes 22 disconnected nodes. Critical Covid-19 is highlighted in orange, with only one of its connections strong enough to pass the display threshold. Edge thickness is based on edge weight, edge distance is based on edge weight and repulsive force in the network for readability. Node size is based on number of connection partners.

Critical Covid-19 showed significant correlations with 56 traits, including diseases of the respiratory system, diseases of the circulatory system, diseases of the digestive system and diseases of the musculoskeletal system and connective tissue (Figure 5, Table 3). Standard errors for these connections were high, and none of the correlations met the more stringent Bonferroni correction threshold.

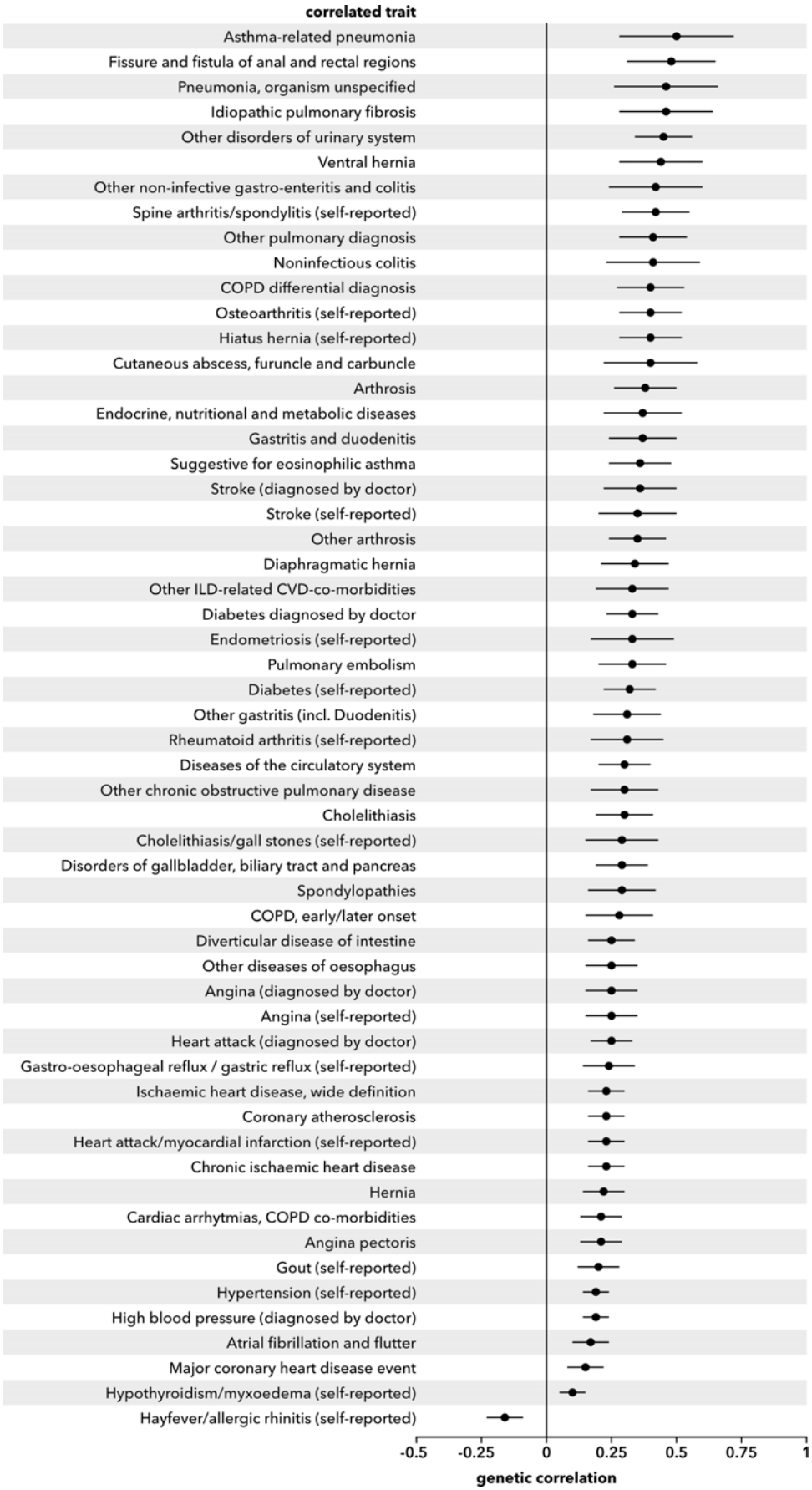

correlated trait
Asthma-related pneumonia
Fissure and fistula of anal and rectal regions
Pneumonia, organism unspecified
Idiopathic pulmonary fibrosis
Other disorders of urinary system
Ventral hernia
Other non-infective gastro-enteritis and colitis
Spine arthritis/spondylitis (self-reported)
Other pulmonary diagnosis
Noninfectious colitis
COPD differential diagnosis
Osteoarthritis (self-reported)
Hiatus hernia (self-reported)
Cutaneous abscess, furuncle and carbuncle
Arthrosis
Endocrine, nutritional and metabolic diseases
Gastritis and duodenitis
Suggestive for eosinophilic asthma
Stroke (diagnosed by doctor)
Stroke (self-reported)
Other arthrosis
Diaphragmatic hernia
Other ILD-related CVD-co-morbidities
Diabetes diagnosed by doctor
Endometriosis (self-reported)
Pulmonary embolism
Diabetes (self-reported)
Other gastritis (incl. Duodenitis)
Rheumatoid arthritis (self-reported)
Diseases of the circulatory system
Other chronic obstructive pulmonary disease
Cholelithiasis
Cholelithiasis/gall stones (self-reported)
Disorders of gallbladder, biliary tract and pancreas
Spondylopathies
COPD, early/later onset
Diverticular disease of intestine
Other diseases of oesophagus
Angina (diagnosed by doctor)
Angina (self-reported)
Heart attack (diagnosed by doctor)
Gastro-oesophageal reflux / gastric reflux (self-reported)
Ischaemic heart disease, wide definition
Coronary atherosclerosis
Heart attack/myocardial infarction (self-reported)
Chronic ischaemic heart disease
Hernia
Cardiac arrhytmias, COPD co-morbidities
Angina pectoris
Gout (self-reported)
Hypertension (self-reported)
High blood pressure (diagnosed by doctor)
Atrial fibrillation and flutter
Major coronary heart disease event
Hypothyroidism/myxoedema (self-reported)
Hayfever/allergic rhinitis (self-reported)
-0.5
-0.25
0
0.25
0.5
0.75
1
genetic correlation

Figure 5: Genetic correlations with critical Covid-19. Strength of the estimated genetic correlation of traits with critical Covid-19 and the standard error of the estimation. Correlations shown achieved a significance of $p < 0.0401$ (5% FDR corrected).

Table 3: Genetic correlations with critical Covid-19.

correlated trait	genetic correlation	standard error	P value
Asthma-related pneumonia	0.50	0.22	0.021
Fissure and fistula of anal and rectal regions	0.48	0.17	0.005
Pneumonia, organism unspecified	0.46	0.20	0.023
Idiopathic pulmonary fibrosis	0.46	0.18	0.009
Other disorders of urinary system	0.45	0.11	0.00004
Ventral hernia	0.44	0.16	0.006
Other non-infective gastro-enteritis and colitis	0.42	0.18	0.016
Spine arthritis/spondylitis (self-reported)	0.42	0.13	0.002
Other pulmonary diagnosis	0.41	0.13	0.001
Noninfectious colitis	0.41	0.18	0.022
COPD differential diagnosis	0.40	0.13	0.002
Osteoarthritis (self-reported)	0.40	0.12	0.001
Hiatus hernia (self-reported)	0.40	0.12	0.001
Cutaneous abscess, furuncle and carbuncle	0.40	0.18	0.032
Arthrosis	0.38	0.12	0.002
Endocrine, nutritional and metabolic diseases	0.37	0.15	0.013
Gastritis and duodenitis	0.37	0.13	0.005
Suggestive for eosinophilic asthma	0.36	0.12	0.003
Stroke (diagnosed by doctor)	0.36	0.14	0.012
Stroke (self-reported)	0.35	0.15	0.017
Other arthrosis	0.35	0.11	0.002
Diaphragmatic hernia	0.34	0.13	0.009
Other ILD-related CVD-co-morbidities	0.33	0.14	0.017

correlated trait	genetic correlation	standard error	P value
Diabetes diagnosed by doctor	0.33	0.10	0.001
Endometriosis (self-reported)	0.33	0.16	0.040
Pulmonary embolism	0.33	0.13	0.013
Diabetes (self-reported)	0.32	0.10	0.001
Other gastritis (incl. Duodenitis)	0.31	0.13	0.019
Rheumatoid arthritis (self-reported)	0.31	0.14	0.029
Diseases of the circulatory system	0.30	0.10	0.003
Other chronic obstructive pulmonary disease	0.30	0.13	0.019
Cholelithiasis	0.30	0.11	0.005
Cholelithiasis/gall stones (self-reported)	0.29	0.14	0.033
Disorders of gallbladder, biliary tract and pancreas	0.29	0.10	0.005
Spondylopathies	0.29	0.13	0.024
COPD, early/later onset	0.28	0.13	0.030
Diverticular disease of intestine	0.25	0.09	0.007
Other diseases of oesophagus	0.25	0.10	0.014
Angina (diagnosed by doctor)	0.25	0.10	0.013
Angina (self-reported)	0.25	0.10	0.014
Heart attack (diagnosed by doctor)	0.25	0.08	0.001
Gastro-oesophageal reflux / gastric reflux (self-reported)	0.24	0.10	0.019
Ischaemic heart disease, wide definition	0.23	0.07	0.002
Coronary atherosclerosis	0.23	0.07	0.001
Heart attack/myocardial infarction (self-reported)	0.23	0.07	0.002
Chronic ischaemic heart disease	0.23	0.07	0.001
Hernia	0.22	0.08	0.005
Cardiac arrhythmias, COPD co-morbidities	0.21	0.08	0.006
Angina pectoris	0.21	0.08	0.016
Gout (self-reported)	0.20	0.08	0.011

correlated trait	genetic correlation	standard error	P value
Hypertension (self-reported)	0.19	0.05	0.0003
High blood pressure (diagnosed by doctor)	0.19	0.05	0.0004
Atrial fibrillation and flutter	0.17	0.07	0.015
Major coronary heart disease event	0.15	0.07	0.029
Hypothyroidism/myxoedema (self-reported)	0.10	0.05	0.032
Hayfever/allergic rhinitis (self-reported)	-0.16	0.07	0.027

3.3 Discussion

What we are aiming to find is evidence of shared disease biology between traits. This is referred to as biological or horizontal pleiotropy, and occurs if one gene or genetic variant independently affects two traits[20]. However, multiple underlying factors can lead to the detection of genetic correlations. Using genetic correlation analysis, horizontal pleiotropy is impossible to differentiate from mediated or vertical pleiotropy, where a gene affects one trait, and this trait then influences another. It has been shown that genetic correlations tend to be made up of a combination of both horizontal and vertical pleiotropy[206,208].

Additionally, a variety of factors can lead to the detection of spurious pleiotropy. LDSC – and therefore its extension HDL – accounts for population stratification (difference in allele frequencies between subpopulations) due to genetic drift or overlapping subjects[209], but some bias may remain[210]. Misdiagnosis, where other phenotypes are mixed into the cases used to calculate the GWAS, can be another source of bias[211]. Much of the data used in this analysis was from the UK Biobank, which itself is not representative of the general population[37]. This selection bias can lead to collider bias in the estimated correlations, where two traits both influence a third variable, which their association is conditioned on[212].

I detected genetic correlations between critical Covid-19 and 56 other diseases. Standard errors for these connections were large, introducing a big amount of uncertainty. Due to this, I could not identify the strength of these correlations, or if critical Covid-19 belongs to any particular network cluster in the disease network. A reason for this could be that critical Covid-19 had lower overlap than recommended with the reference panel

I used to establish genetic correlations, which weakens analysis power. It could be that not all connections were detected, or that not all detected connections are accurate.

However, multiple published studies have found similar results using different datasets, further supporting those connections. A study looking at connections between Covid-19 and cardiovascular diseases using HDL analysis found positive genetic correlations with coronary artery disease, hypertension and type 2 diabetes[157]. While this investigation used the same GWAS for Covid-19 as I did, alternative non-UK Biobank data was used for the other traits. Another study found evidence of genetic correlation between critical Covid-19 and ischaemic stroke, using different data for both traits[158]. Another paper reported all UK Biobank correlations I found with either severe or hospitalised Covid-19, except for three GWAS – atrial fibrillation, cardiac arrhythmia and hypothyroidism[159]. Though the original discovery was performed on the same UK Biobank data I used, they replicated their findings in different data where available, including for hypertension, coronary artery disease, heart failure and type 2 diabetes. Finally, a paper studying overlap between Covid-19 and idiopathic pulmonary fibrosis found correlations between the two using an extended idiopathic pulmonary fibrosis meta analysis GWAS compared to the one in my dataset[160].

The association between more severe forms of Covid-19 and cardiovascular disease appears particularly strong[213]. Prior cardiovascular conditions such as hypertension and diabetes have been shown to be a risk factor for hospitalised Covid-19 and lead to worse outcomes[214,215], and Covid-19 has been noted to exacerbate existing cardiac conditions[216] and may directly injure the cardiovascular system[217]. Although it is undetermined if this relationship is due to shared etiological factors, one mechanistic way by which critical Covid-19 and cardiovascular disease may be connected could be through the enzyme ACE2 (angiotensin-converting enzyme 2), which SARS-CoV-2 uses as an entry point to invade cells[218]. ACE2 is highly expressed in the heart[219] and has been implicated in heart function, hypertension and diabetes[220]. It has been shown that SARS-CoV-2 can infect the heart, vascular tissues, and circulating cells[221]. In turn it has been reported that plasma ACE2 levels are genetically correlated with vascular diseases as well as severe Covid-19[222].

Genetic correlation is an average for pleiotropy across the whole genome, but pleiotropy can vary strongly among local regions. Opposing directions of effect at local associations can result in a genetic correlation of zero, masking shared loci that could hold vital information on shared disease biology and potential for drug repurposing. In the next

chapter, I will explore which such local pleiotropic effects critical Covid-19 shares with other diseases.

Chapter 4
Genetic colocalization between critical Covid-19 and a large range of diseases

4.1 Introduction

As we have established thus far, there is widespread pleiotropy throughout the genome – genetic variants often influence multiple seemingly unrelated traits[21]. Colocalization is a method that can be used to identify such points of pleiotropy between diseases, by testing whether both traits are the consequence of the same genetic variant[223]. In contrast to the whole-genome genetic correlation analysis in the previous chapter, colocalization is established locally for a specific region. The analysis is based on a Bayesian algorithm, which uses probability prior settings of the assumed likelihood for a shared variant and calculates the probability of a shared lead variants between traits. This avenue of investigation allows for the identification of genetic variants that have already been discovered in association with other disease phenotypes, but may impact the outcome of critical Covid-19 as well. By exploring multiple GWAS in combination, we use the genetics of similar traits to add power to the original GWAS, with the potential to detect additional associated variants that are not identifiable from the Covid-19 GWAS alone.

In this chapter, I utilise the 2021 extension to the original GWAS summary statistic colocalization method Coloc, HyPrColoc (Hypothesis Prioritisation for multi-trait Colocalization)[24]. This new version adds the ability to test for colocalizations across a wide range of traits with speed. Rather than reporting only pairwise colocalizations, HyPrColoc can identify whether any number of included traits share putative causal variants. Additionally, this expansion to more traits improved the performance of the analysis compared to Coloc: More correct colocalizations were identified, while false positive rates were shown to remain similarly low.

I chose to iteratively analyse the dataset, concentrating on a subset of diseases at first and subsequently widening the analysis to include a broader range of diseases. In the first

instance, I focused on diseases related to the immune system, such as autoimmune disease and infectious disease. It has been shown that the inflammatory response to SARS-coronavirus-2 (SARS-CoV-2) infection has a major impact on Covid-19 pathogenesis[137,138]. Additionally, there is pervasive pleiotropy among genetic associations with autoimmune disease[139], suggesting that there may be widespread pleiotropy in the immune system.

While Covid-19 has been closely studied through many approaches, only a small number of studies focused on colocalization with other GWAS. These investigations have either been limited in their scope, by exploring shared genetics between Covid-19 and one other specific trait[160], or utilised colocalization as a follow up investigation of individual pleiotropic variants identified through different means[224]. Others have explored different comparison datasets, such as studying colocalizations between host proteins involved in SARS-CoV-2 infection[225]. This leaves an untapped opportunity to discover yet unrecognised connections between the genetics of critical Covid-19 and other diseases.

4.2 Results

To establish if critical Covid-19 shares localised genetic connections with other diseases I performed a genome-wide multi-trait colocalization analysis using HyPrColoc[24].

In contrast to the genetic correlation methods in the last chapter that probed similarity over the whole genome, colocalization is a much more localised semblance test. I split the GWAS data into blocks based on linkage disequilibrium between variants to distinguish between association signals[185]. HyPrColoc tests each region to establish whether a local associated variant is present and, if so, whether it is common to different traits.

My first approach was to limit the analysed dataset to immune system related diseases. This is in part due to practical considerations regarding the method. HyPrColoc operates on the assumption that there is a single associated variant in a given local region, but this is often violated by actual genetic data. Once a trait is found to colocalize, it is taken out of subsequent analysis rounds of that local area. For a trait with multiple associated local SNPs this can lead to only one or even none of its colocalizations being found, with Coloc potentially unable to identify associated variants. A larger number of traits in the analysis increases the likelihood of this occurring. Additionally, analysis power slightly

decreases with a large number of traits, as the number of hypotheses for possible colocalizations rises[24].

Comparing our GWAS of critical Covid-19 with 56 GWAS of infectious and autoimmune diseases, I found 15 potential colocalizations (Table 4). Additional follow up analyses were performed to identify which were true positives and to exclude true negatives (Figure 6). Colocalizations were considered true positives when a) they reached a posterior probability of > 0.7 under the discovery colocalization prior 0.02 and b) they reached a posterior probability of > 0.7 under the more stringent colocalization prior 0.01, unless follow up analysis showed the involved traits were multi-associated in the local area.

Table 4: All potential HyPrColoc colocalizations involving critical Covid-19. Colocalizations are presented according to the LD block they were detected in. The posterior probability describes the likelihood that the colocalization is real, while "posterior explained by variant" indicates the probability that the colocalization variant (lead SNP ID) is chosen correctly. Colocalization groups are reported for the first HyPrColoc analysis run with fewer traits but more variants (initial groups) and second analysis run with more traits but fewer variants (additional traits for existing groups or newly detected groups). Where a new lead variant is chosen for a colocalization group in the second run, I confirmed this is due to removal of the original lead from the analysis and report probabilities for both variants. Where the second run results in a new group of traits rather than adding traits to an existing group, it is reported in a separate row and was treated as a separate colocalization in downstream analyses.

LD block	lead SNP ID	posterior probability	posterior explained by variant	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)
chr1 7247335-9365199	rs7523335	0.2943	0.2502	Critical Covid-19, Ulcerative colitis (de Lange), Inflammatory bowel disease, Ulcerative colitis (Liu)	

LD block	lead SNP ID	posterior probability	posterior explained by variant	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)
chr2 110572432-113921856	rs72837826	0.4765	0.9813		Critical Covid-19, Asthma (diagnosed by doctor), Asthma (self-reported), Ulcerative colitis (de Lange), Inflammatory bowel disease, Ulcerative colitis (Liu), Churg-Strauss syndrome (ANCA-negative), Churg-Strauss syndrome (ANCA-positive)
chr4 100678360-103221356	rs13107325 / rs13135092	0.8467 / 0.7253	0.8647 / 1	Critical Covid-19, Hayfever/allergic rhinitis or eczema (diagnosed by doctor), Hayfever/allergic rhinitis (self-reported), Inflammatory bowel disease, Crohn's disease	Daytime dozing / sleeping (narcolepsy), Allergic disease (asthma, hay fever or eczema)
chr6 31571218-32682664	rs687308	0.4178	1	Critical Covid-19, Emphysema/chronic bronchitis (diagnosed by doctor), Childhood asthma (age<16), Suggestive for eosinophilic asthma, Rheumatoid arthritis, Rheumatoid arthritis (self-reported), Other rheumatoid arthritis, Other/unspecified rheumatoid arthritis, Inflammatory bowel disease	
chr7 98715474-100196651	rs2897075	0.962	0.5485		Critical Covid-19, Idiopathic pulmonary fibrosis

LD block	lead SNP ID	posterior probability	posterior explained by variant	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)
chr11 58780549-62223771	rs174535	0.3986	0.7894	Critical Covid-19, Eczema, Asthma (diagnosed by doctor), Asthma (self-reported), Suggestive for eosinophilic asthma, Bronchitis, Unspecified acute lower respiratory infection, Inflammatory bowel disease, Crohn's disease	
chr12 132807034-133841511	rs11614702	0.9836	1	Critical Covid-19, Asthma (childhood onset), Asthma (adult onset)	
chr13 112247592-113572488	rs3742238	0.9969	0.9078		Critical Covid-19, Idiopathic pulmonary fibrosis
chr19 4348967-5811852	rs12610495	0.999	1		Critical Covid-19, Idiopathic pulmonary fibrosis
chr19 46102697-47150082	rs16980051	0.5444	1	Critical Covid-19, Asthma (childhood onset), Asthma (adult onset), Inflammatory bowel disease	
chr19 46102697-47150082	rs7250497	0.6968	0.2514		Critical Covid-19, Asthma (diagnosed by doctor), Asthma (self-reported)
chr19 8347513-9238393	rs11673136	0.7431	1	Critical Covid-19, Asthma (adult onset)	

LD block	lead SNP ID	posterior probability	posterior explained by variant	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)
chr19 9238393-11284028	rs144309607	0.9928	0.969	Critical Covid-19, Psoriasis (self-reported)	
chr19 9238393-11284028	rs34725611	0.866	0.9999		Critical Covid-19, Systemic lupus erythematosus

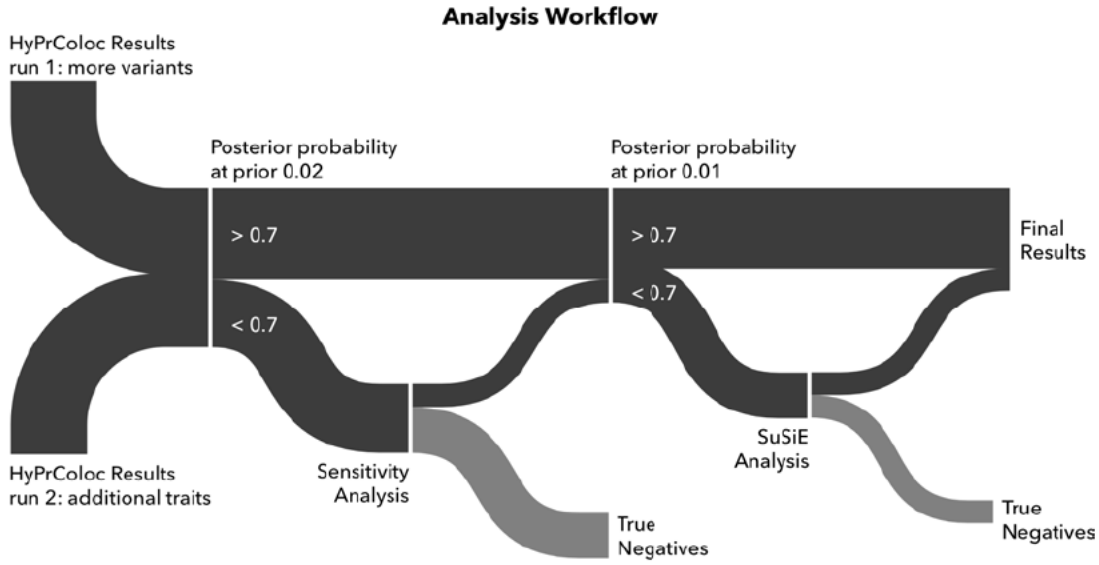


Figure 6: Analysis workflow. HyPrColoc results from two analyses runs were merged: one included more genetic variants while the other included more traits. Colocalizations that did not meet posterior probability > 0.7 when using the conditional colocalization prior 0.02 underwent a sensitivity analysis to test if only some traits in a colocalization group were unsupported. Traits or groups that failed this test were excluded from the results, while those that passed underwent further testing. Colocalizations that did not meet posterior probability > 0.7 when using the conditional colocalization prior 0.01 were tested using a second follow up analysis, SuSiE, to establish whether the local area of a GWAS contained multiple associated variants that impacted prior sensitivity in the HyPrColoc analysis. Traits that failed this test were excluded, while those with multiple associated variants were reported in the final results if one or more of those variants colocalized with other traits.

4.2.1 Quality control step 1: Prior sensitivity analysis

The first follow up step was an investigation of colocalizations that had a posterior probability < 0.7. Weak links within a group of colocalizing traits can lower the probability for the entire group. Therefore, it is prudent to investigate if any of the colocalizations within the group reach the necessary probability to suggest a true positive result. I performed the sensitivity analysis included in the HyPrColoc package by Foley et al., which tests the sensitivity of the colocalization to the conditional colocalization prior and threshold parameters, as this can indicate a false positive result. The analysis works by repeatedly calculating the colocalization probability for the region, each time tightening these settings so they are more stringent for each round. This makes it easy to distinguish colocalizations that are persistently found across priors and thresholds from those that are true negatives and need to be excluded from the results.

The sensitivity analysis revealed that some results could be rescued while others needed to be excluded (Figure 7,Figure 8, Table 5). Two of the colocalizations, located on chromosome 6 and chromosome 19, reached a posterior probability > 0.7 upon removal of weakly colocalized traits within the group. For four further tested groups, critical Covid-19 itself was among the weak traits and had to be removed, excluding those groups from further analysis in this chapter.

Table 5: HyPrColoc sensitivity analysis results. Colocalizations are presented according to the LD block they were detected in. The posterior probability describes the likelihood that the colocalization as originally detected is real, while the new posterior probability describes the likelihood after the removal of weakly colocalized traits. Where critical Covid-19 was removed I discontinued further investigation in this chapter.

LD block	lead SNP ID	posterior probability	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)	excluded in sensitivity analysis	new posterior probability
chr1 724733 5-9365 199	rs7523335	0.2943	Critical Covid-19, Ulcerative colitis (de Lange), Inflammatory bowel disease, Ulcerative colitis (Liu)		Critical Covid-19, Ulcerative colitis (de Lange), Inflammatory bowel disease, Ulcerative colitis (Liu)	NA

LD block	lead SNP ID	posterior probability	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)	excluded in sensitivity analysis	new posterior probability
chr2 110572432-113921856	rs72837826	0.4765		Critical Covid-19, Asthma (diagnosed by doctor), Asthma (self-reported), Ulcerative colitis (de Lange), Inflammatory bowel disease, Ulcerative colitis (Liu), Churg-Strauss syndrome (ANCA-negative), Churg-Strauss syndrome (ANCA-positive)	Critical Covid-19, Asthma (self-reported), Churg-Strauss syndrome (ANCA-positive)	NA
chr6 31571218-32682664	rs687308	0.4178	Critical Covid-19, Emphysema/chronic bronchitis (diagnosed by doctor), Childhood asthma (age<16), Suggestive for eosinophilic asthma, Rheumatoid arthritis, Rheumatoid arthritis (self-reported), Other rheumatoid arthritis, Other/unspecified rheumatoid arthritis, Inflammatory bowel disease		Emphysema/ chronic bronchitis (diagnosed by doctor), Inflammatory bowel disease	0.9214

LD block	lead SNP ID	posterior probability	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)	excluded in sensitivity analysis	new posterior probability
chr11 58780549-62223771	rs174535	0.3986	Critical Covid-19, Eczema, Asthma (diagnosed by doctor), Asthma (self-reported), Suggestive for eosinophilic asthma, Bronchitis, Unspecified acute lower respiratory infection, Inflammatory bowel disease, Crohn's disease		Critical Covid-19, Eczema, Suggestive for eosinophilic asthma, Bronchitis, Unspecified acute lower respiratory infection,	NA
chr19 46102697-47150082	rs16980051	0.5444	Critical Covid-19, Asthma (childhood onset), Asthma (adult onset), Inflammatory bowel disease		Inflammatory bowel disease	0.9278
chr19 46102697-47150082	rs7250497	0.6968		Critical Covid-19, Asthma (diagnosed by doctor), Asthma (self-reported)	Critical Covid-19	NA

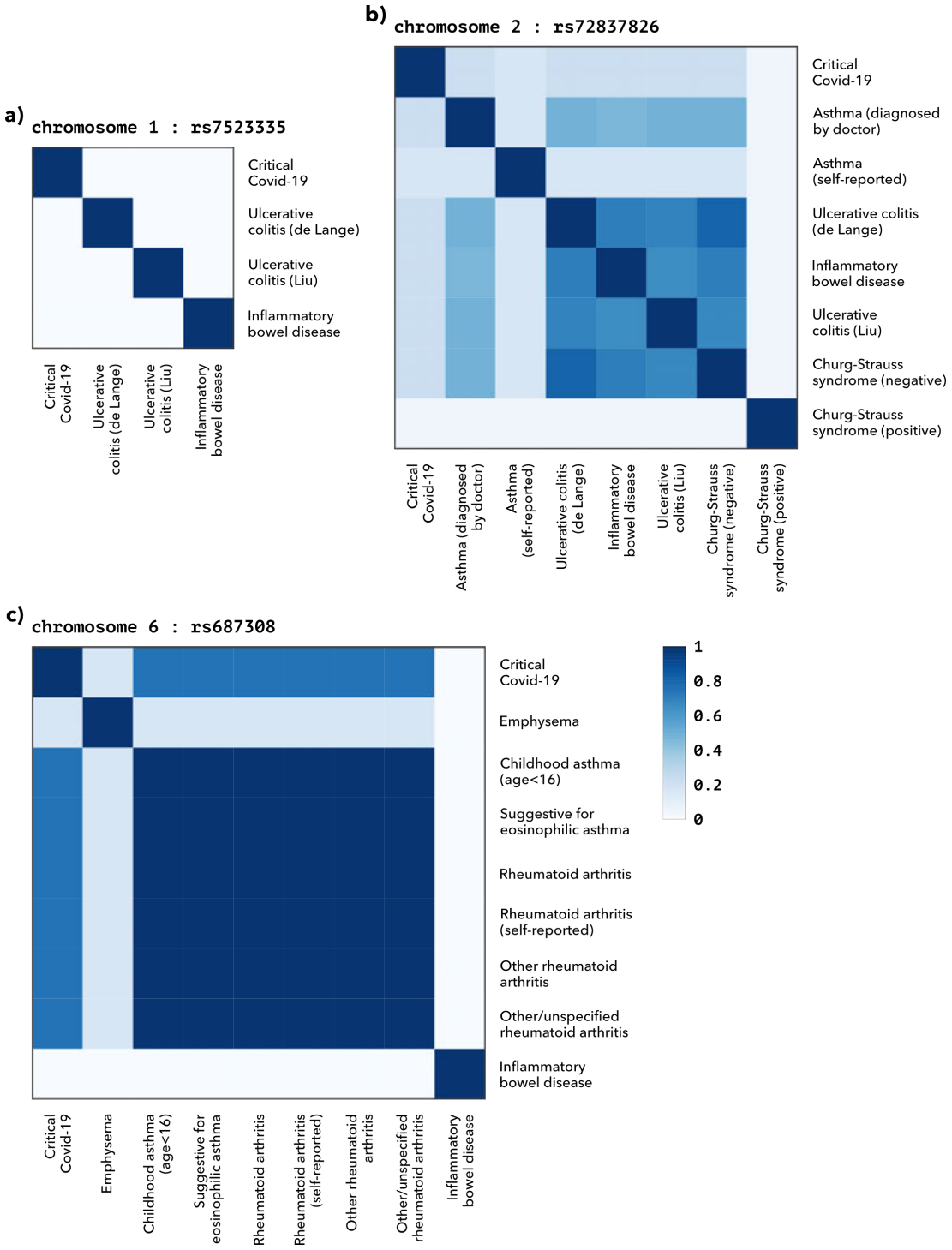


Figure 7: HyPrColoc sensitivity analysis results (part 1). This test looks for weak links in a group of colocalizations by determining the sensitivity to changes in the prior and threshold parameters. The HyPrColoc analysis for the region is repeated multiple times, each time with increasingly strict values for the conditional colocalization prior (0.05, 0.02, 0.01 and 0.005) and regional association probability as well as alignment probability threshold parameters (0.5, 0.6 and 0.7). The darker a cell in the heat map, the more often the trait pair is found to colocalize.

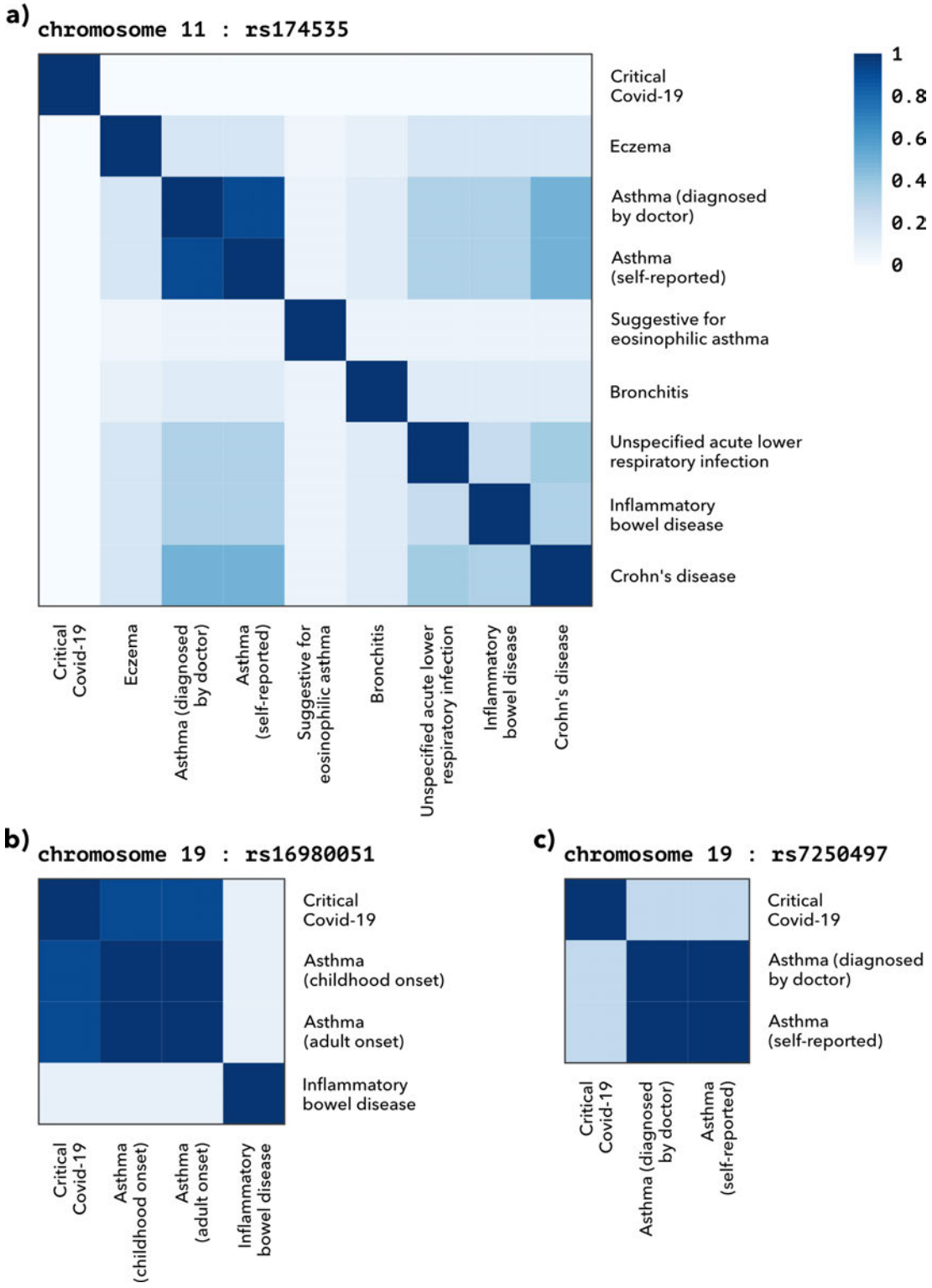


Figure 8: HyPrColoc sensitivity analysis results (part 2).

With all results remaining in the analysis having reached a posterior probability of > 0.7 under the conditional colocalization prior 0.02, the next step was to test which colocalizations could hold up to this threshold under the more stringent colocalization

prior 0.01. Colocalizations that fulfil this criterion were considered to be true positive results with strong evidence. Colocalizations that achieved the set threshold only under the more permissive prior needed to undergo a further test to decide whether they were supported by strong enough evidence. As shown in Table 6, six results were robust, while three locations needed a final verification step to establish their validity.

Table 6: Hyprcoloc results with the more stringent conditional colocalization prior 0.01. Colocalizations that did not meet posterior probability > 0.7 with this setting are marked for further follow up analysis.

LD block	lead SNP ID	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)	posterior probability (prior 0.01)	multi-association test?	traits that need investigation
chr4 100678360-103221356	rs13107325	Critical Covid-19, Hayfever/allergic rhinitis or eczema (diagnosed by doctor), Hayfever/ allergic rhinitis (self-reported), Inflammatory bowel disease, Crohn's disease	Daytime dozing / sleeping (narcolepsy), Allergic disease (asthma, hay fever or eczema)	0.5617	yes	Daytime dozing / sleeping (narcolepsy)
chr6 31571218-32682664	rs687308	Critical Covid-19, Childhood asthma (age<16), Suggestive for eosinophilic asthma, Rheumatoid arthritis, Rheumatoid arthritis (self-reported), Other rheumatoid arthritis, Other/unspecified rheumatoid arthritis		0.8572	no	
chr7 98715474-100196651	rs2897075		Critical Covid-19, Idiopathic pulmonary fibrosis	0.9263	no	
chr12 132807034-133841511	rs11614702	Critical Covid-19, Asthma (childhood onset), Asthma (adult onset)		0.9679	no	

LD block	lead SNP ID	traits colocalizing (run 1: more SNPs)	additional traits (run 2: more traits)	posterior probability (prior 0.01)	multi-association test?	traits that need investigation
chr13 112247592-113572488	rs3742238		Critical Covid-19, Idiopathic pulmonary fibrosis	0.9938	no	
chr19 4348967-5811852	rs12610495		Critical Covid-19, Idiopathic pulmonary fibrosis	0.9979	no	
chr19 4610269 7-47150082	rs16980051	Critical Covid-19, Asthma (childhood onset), Asthma (adult onset)		0.8641	no	
chr19 8347513-9238393	rs11673136	Critical Covid-19, Asthma (adult onset)		0.5884	yes	Critical Covid-19, Asthma (adult onset)
chr19 9238393-11284028	rs144309607	Critical Covid-19, Psoriasis (self-reported)		0.9857	yes	Critical Covid-19, Psoriasis (self-reported)
chr19 9238393-11284028	rs34725611		Critical Covid-19, Systemic lupus erythematosus	0.7637	yes	Critical Covid-19, Systemic lupus erythematosus

4.2.2 Quality control step 2: Multiple associated variant analysis

Sensitivity to changes in the colocalization prior can also indicate a violation of the single associated variant assumption. To establish if this was the case for the identified unstable colocalizations, I performed a SuSiE (Sum of Single Effects) fine-mapping followed by colocalization analysis[45]. This extension to Coloc can detect multiple

associated variants and use Coloc to establish colocalizations between them within one region.

The first region I identified as requiring a multi-association test was the LD block on chromosome 4 (basepairs 100678360-103221356), where the trait daytime dozing/ sleeping (narcolepsy) was weak to changes in the colocalization prior. This potential colocalization was shared with 5 other traits, but as coloc.susie can only evaluate two traits at a time, I chose to assess colocalization between narcolepsy and critical Covid-19. Due to the low P-values in the area, the sensitivity of the SuSiE analysis had to be increased by reducing coverage to 0.01. One lead SNP was detected for critical Covid-19 (rs35225200) and one credible variable set containing two SNPs was found for narcolepsy (rs111501786 and rs111229888). This indicated that neither trait had multiple associated SNPs in the area, as variants in the same set are attributed to the same signal. Furthermore, the likelihood for that variant to be shared was 2.19%. Therefore narcolepsy was excluded from this colocalization group.

The next potential colocalization in need of further validation was between critical Covid-19 and adult onset asthma at rs11673136 on chromosome 19. The P-values for asthma were too low to establish putative lead associated variants in this region and no credible variant sets could be established with SuSiE. The colocalization was therefore excluded from the results.

The final area in need of evaluation was the LD block between the basepairs 9238393-11284028 on chromosome 19. Two colocalizations with critical Covid-19 were identified in this region: with self-reported psoriasis at rs144309607 during the first analysis run (analysing more variants) and with systemic lupus erythematosus at rs34725611 in the second run (adding additional traits). Though these colocalizations were both strong and not sensitive to changes in the prior, the single associated variant assumption means the second analysis round should include all three traits if this colocalization stems from a single signal. I therefore performed SuSiE to fine-map the region and establish if the traits had multiple associated variants in this region. I identified the lead variants (credible sets) for each traits: For psoriasis I found two credible variant sets, one containing one SNP (rs539820608), the other four (rs35251378, rs11085725, rs11085727, rs34725611). For critical Covid-19 there were two credible variant sets, one with one SNP (rs34536443), the other with two SNPs (rs73510898 and rs142770866). Systemic lupus erythematosus had two credible variant sets as well, one consisting of two SNPs (rs74956615 and rs34536443), the other of

eleven (rs192560669, rs140735577, rs73510898, rs75231016, rs74908652, rs78295726, rs12720356, rs2304257, rs118115488, rs78064630 and rs7247198). Following on from this, I found two strong colocalizations for critical Covid-19 and lupus, with 100% on rs34536443 and 99.6% on rs73510898 (Figure 9). Covid-19 and psoriasis shared one colocalization with 92.1% posterior probability on rs34536443, and lupus and psoriasis with 99.9% on rs34536443. I therefore concluded that there are two colocalizations with critical Covid-19 in this region: one with lupus at rs73510898, and one with lupus and psoriasis at rs34536443.

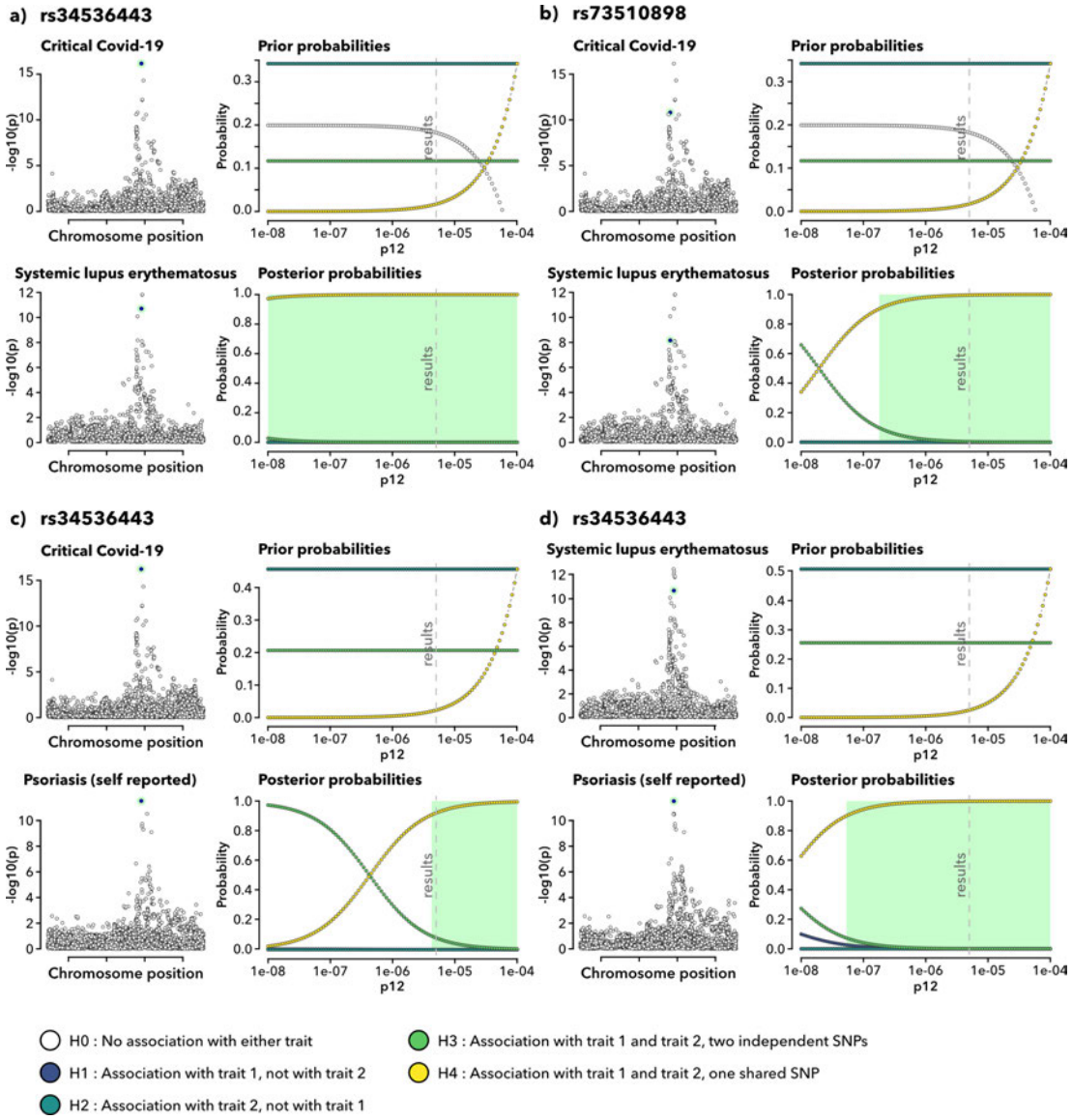


Figure 9: Sensitivity plots for colocalizations passing the SuSiE multi-association analysis test. Each subfigure (a-d) describes the colocalization at one SNP. On the left are the local manhattan plots of the colocalizing traits with the colocalization variant marked. Hypothesis outcome probabilities for a random SNP in the region are displayed on the top right, while hypothesis outcomes for the

colocalization SNP are on the bottom right. The x-axis describes changes in p12, the assumed prior probability that a random SNP in the region is jointly associated with both traits. The green shaded region marks the needed H4 > 0.9 probability threshold over the range of prior probabilities for which it is supported.

4.2.3 Widening the disease dataset

Having established the final immune disease results, I then performed an additional bigger HyPrColoc analysis which included 173 additional disease GWAS. Most colocalization loci that were detected in the smaller analysis were also identified in this large scale investigation, with the exception of the colocalizations between asthma and critical Covid-19 on chromosome 12 and 19. I found no new locations for genetic colocalizations with critical Covid-19; however, several traits were added to existing colocalization groups (Table 7). For the chromosome 4 colocalization at rs13107325 the additional traits were hypertension (self-reported), high blood pressure (diagnosed by doctor), osteoarthritis (self-reported) and diabetes diagnosed by doctor. The rs34536443 locus on chromosome 19 newly included the traits hypothyroidism/myxoedema (self-reported), rheumatoid arthritis, rheumatoid arthritis (self-reported) and other rheumatoid arthritis in the colocalization. Both expanded colocalization groups had a large probability of colocalization and were not sensitive to changes in the prior.

Table 7: HyPrColoc analysis with a larger dataset adds additional traits for existing colocalization groups. Only traits that met all follow up analyses requirements are listed.

LD block	lead SNP ID	posterior probability (prior 0.02)	posterior explained by variant	posterior probability (prior 0.01)	traits colocalizing
chr4 100678360-103221356	rs13107325	0.8628	1	0.7599	Critical Covid-19, Hypertension (self-reported), High blood pressure (diagnosed by doctor), Osteoarthritis (self-reported), Diabetes (diagnosed by doctor), Hayfever/ allergic rhinitis or eczema (diagnosed by doctor), Hayfever/allergic rhinitis (self-reported)

LD block	lead SNP ID	posterior probability (prior 0.02)	posterior explained by variant	posterior probability (prior 0.01)	traits colocalizing
chr19 9238393-11 284028	rs34536443	0.8374	1	0.7229	Critical Covid-19, Psoriasis (self-reported), Hypothyroidism/ myxoedema (self-reported), Rheumatoid arthritis, Rheumatoid arthritis (self-reported), Other rheumatoid arthritis

4.2.4 The fault in two asthma GWAS

At this point it seemed as though my analysis strategy of performing a separate, smaller scale immune system focused HyPrColoc analysis was successful, as I had identified two additional Covid-19 colocalizations compared to the large scale analysis. These shared associated variants were found between critical Covid-19, childhood onset asthma and adult onset asthma at rs11614702 on chromosome 12 as well as at rs16980051 on chromosome 19. However, as I was preparing the figures for this thesis I noticed a seemingly contradictory detail: although the reported effect sizes for these SNPs in the asthma GWAS were large (~1) and their standard errors small (0.01-0.02), the listed corresponding P-values were so large as to suggest the signal was merely background noise (0.04-0.001). Further investigation revealed that the column labelled effect size in the harmonised MRC IEU GWAS database version of the GWAS I downloaded was in actuality the unconverted odds ratio originally reported by Ferreira et al.[170] in the original publication describing both asthma GWAS. This made small changes in odds ratio appear as large effect sizes, leading to HyPrColoc identifying incorrect colocalizations with both asthma GWAS. I therefore removed those colocalizations from the final results. Additionally, I repeated the full scale HyprColoc analysis without these GWAS to ensure that this would not cause a change to the results other than the removal of the GWAS in question. There were no other changes to the results presented in this chapter.

4.2.5 Final colocalization results for critical Covid-19

Having pruned the results to only those that were strongly supported, the final results are depicted in Figure 10 and 11.

On chromosome 4, in the LD block 100678360-103221356 I found colocalizations with critical Covid-19 at rs13107325 (effect allele T / other allele C, minor allele T frequency 0.08) in analysis run one (analysing more variants) and rs13135092 in analysis run two (adding additional traits). As LD between the two variants is high (0.93 as calculated using the Ensembl Linkage Disequilibrium Calculator[226] with British 1000 Genomes data) I assumed that both represent the same signal. Since rs13107325 is chosen when both SNPs are in the analysis, I considered it the colocalization SNP. The traits sharing this putative causal variant were critical Covid-19, Crohn's disease, inflammatory bowel disease, allergic disease (asthma, hay fever or eczema), hayfever, allergic rhinitis or eczema (diagnosed by doctor), hayfever, allergic rhinitis (self-reported) hypertension (self-reported), high blood pressure (diagnosed by doctor), osteoarthritis (self-reported) and diabetes (diagnosed by doctor). Effect sizes for this variant in the GWAS I utilised suggest a largely shared direction of effect, with the minor allele being harmful for all traits except hypertension – although effect sizes were small for all traits except Covid-19 and the inflammatory bowel diseases (Figure 11).

The second colocalization I identified was at variant rs687308 (effect allele T / other allele C, minor allele T frequency 0.12) in LD block 31571218-32682664 on chromosome 6. The traits colocalizing at this position were critical Covid-19, childhood asthma (age < 16), suggestive for eosinophilic asthma, rheumatoid arthritis, rheumatoid arthritis (self-reported), other rheumatoid arthritis and other/unspecified rheumatoid arthritis. Effect size data from the GWAS shows that the minor allele confers a protective effect against critical Covid-19 as well as having a small detrimental effect to the other traits.

There were three colocalizations between critical Covid-19 and idiopathic pulmonary fibrosis. The first was at variant rs2897075 (effect allele T / other allele C, minor allele T frequency 0.38) in LD block 98715474-100196651 on chromosome 7, the second at variant rs3742238 (effect allele T / other allele C, minor allele T frequency 0.21) in LD block 112247592-113572488 on chromosome 13 and the third was at rs12610495 (effect allele G / other allele A, minor allele G frequency 0.31) in LD block 4348967-5811852 on chromosome 19. GWAS effect sizes indicate that the chromosome 7 and 19 minor alleles are detrimental for both traits. The allele change at rs3742238 shows opposing effects, suggesting it is detrimental for critical Covid-19 and protective against idiopathic pulmonary fibrosis.

At SNP rs34536443 (effect allele C / other allele G, minor allele C frequency 0.05) in LD block 9238393-11284028 on chromosome 19, I found a colocalization between critical Covid-19, systemic lupus erythematosus, psoriasis (self-reported), hypothyroidism/myxoedema (self-reported), rheumatoid arthritis, rheumatoid arthritis (self-reported) and other rheumatoid arthritis. Effect sizes in the GWAS show a detrimental effect for critical Covid-19, a strongly protective effect against systemic lupus erythematosus and a mildly protective effect for the other traits.

The final colocalization I found was between critical Covid-19, systemic lupus erythematosus at SNP rs73510898 (effect allele A / other allele G, minor allele A frequency 0.09) in LD block 9238393-11284028 on chromosome 19. Again, the GWAS effect sizes at this position suggest that the minor allele is detrimental in critical Covid-19, but protective against systemic lupus erythematosus.

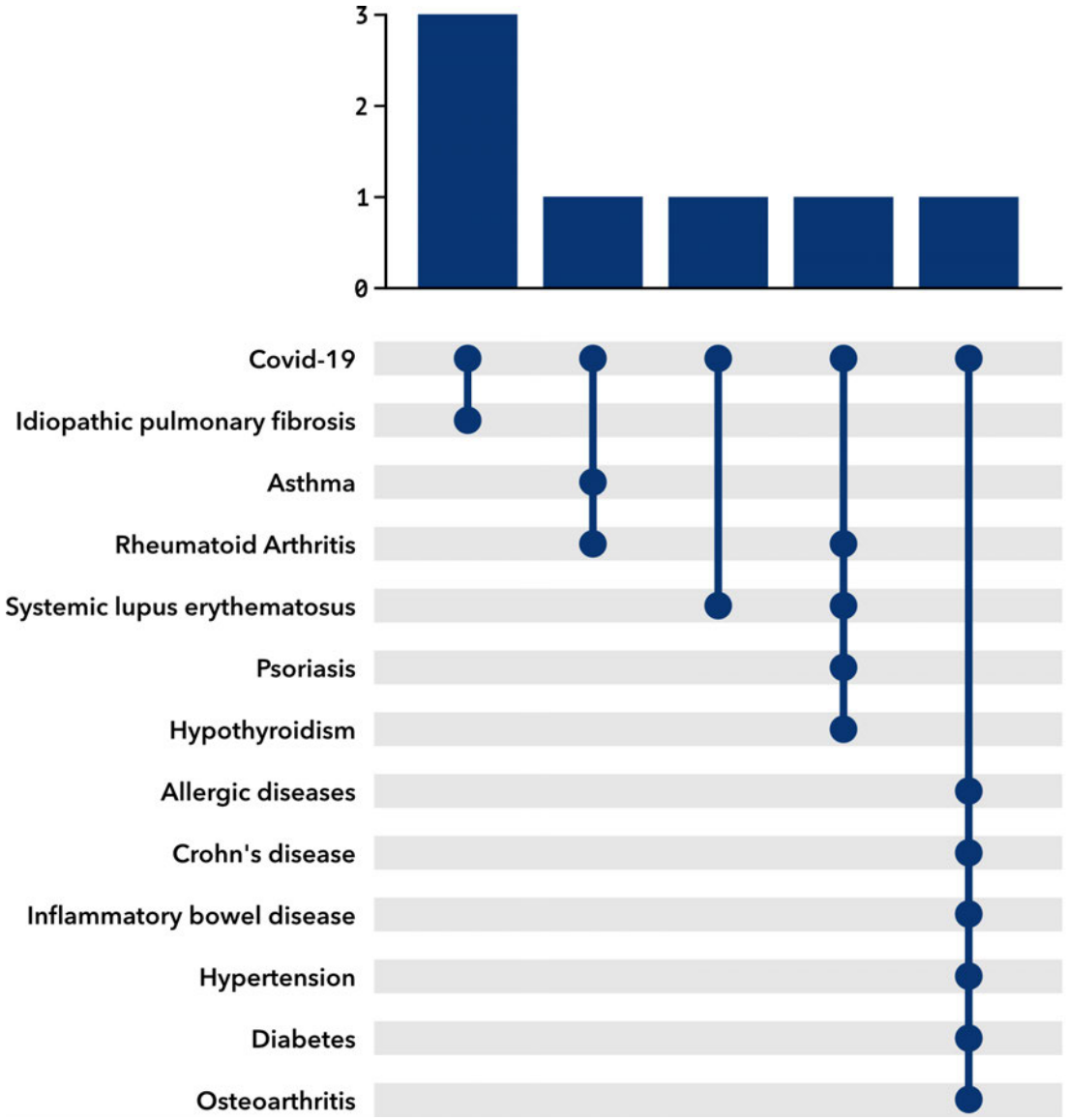


Figure 10: How often does each trait colocalize with critical-Covid-19? Idiopathic pulmonary fibrosis is the most frequent colocalization partner, with 3 instances. Both rheumatoid arthritis and systemic lupus erythematosus colocalize with Covid-19 twice, although each time grouped with different additional traits.

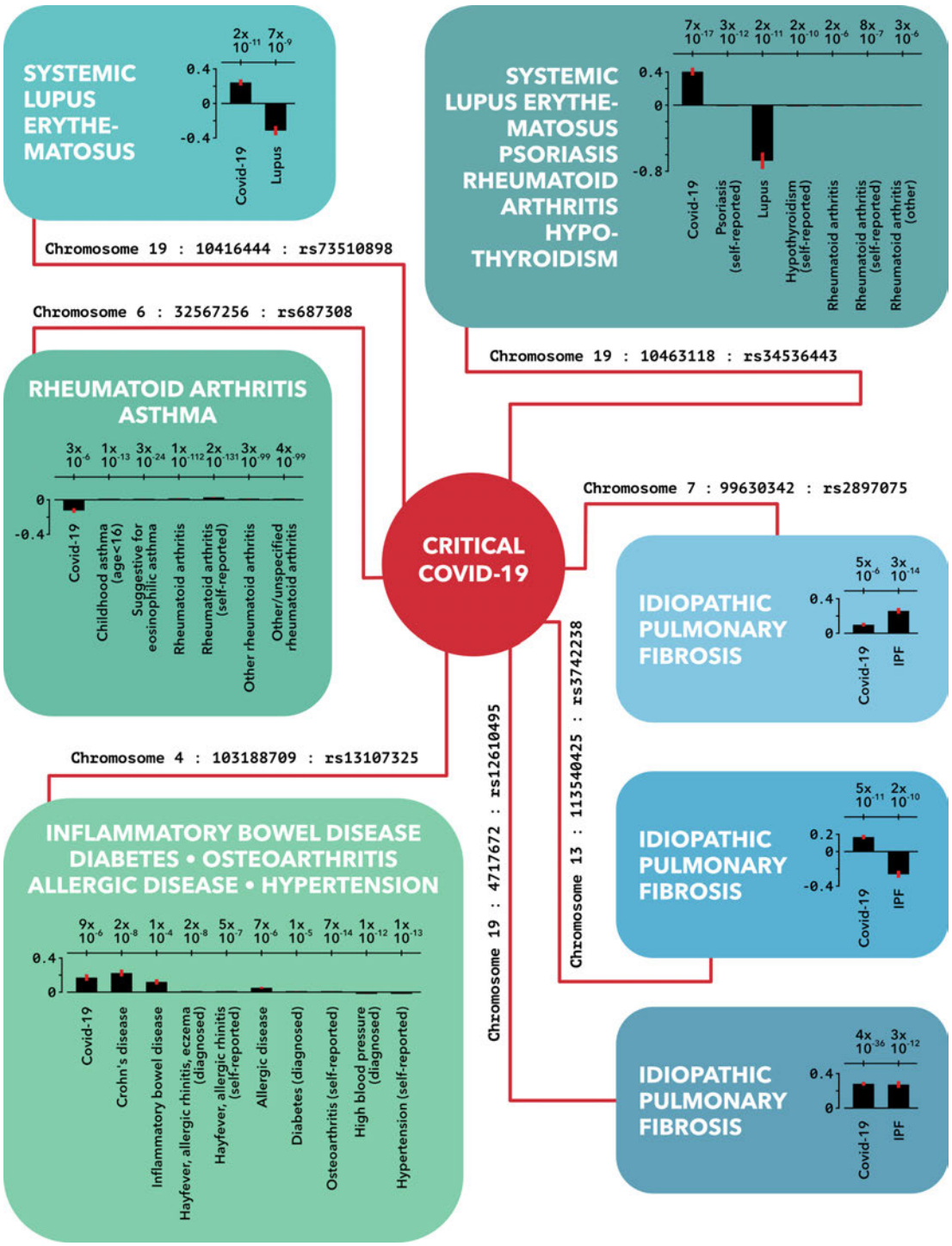


Figure 11: Final results for colocalizations with critical Covid-19. For each colocalization group, a bar chart compares effect sizes for the lead variant in the original GWAS, with the accompanying P-value displayed above each bar. Effect sizes are depicted with regards to the minor allele.

4.3 Discussion

The HyPrColoc analysis yielded seven colocalizations with critical Covid-19.

On chromosome 4, I found a colocalization at rs13107325 between critical Covid-19, Crohn's disease, inflammatory bowel disease, allergic disease (asthma, hay fever or eczema), hayfever, allergic rhinitis or eczema (diagnosed by doctor), hayfever, allergic rhinitis (self-reported), hypertension (self-reported), high blood pressure (diagnosed by doctor), osteoarthritis (self-reported) and diabetes (diagnosed by doctor). The variant is considered highly pleiotropic, with 425 trait associations listed in the GWAS Catalogue[168] as of 16.03.2023. Amongst its known associations are the colocalizing traits Crohn's disease[227,228], eczema[229], blood pressure[230], osteoarthritis[231] and diabetes[232]. Neither rs13107325 nor any other chromosome 4 SNPs reached genome-wide significance in our paper describing the critical Covid-19 GWAS that I examined in this analysis. Recently, the COVID-19 Host Genetics Initiative has reported an association between this variant and general susceptibility to Covid-19[233].

The SNP rs13107325 is located in the gene *SLC39A8*, which encodes the divalent metal ion transporter ZIP8. Although it was initially described as a zinc transporter[234], its highest affinity in mammalian cells is with manganese[235] and it additionally can transport iron[236] and cadmium ions[235]. Transporter proteins work by forming a tunnel made of multiple transmembrane helices traversing the cell membrane, through which ions are brought into the cell. The allele change in rs13107325 constitutes a missense variant, which leads to an amino acid change in the translated protein. This results in a switch from alanine to threonine at position 391 (called A391T). Based on homology to X-ray crystallography of bacterial ZIP4[237], as well as AlphaFold's[201] structural prediction for ZIP8, alanine 391 is located at the extracellular loop connecting transmembrane helices 6 and 7 (Figure 12). Alanine is a small and hydrophobic amino acid, while threonine is medium sized and hydrophilic. One way this change could affect transporter activity is by changing the shape of the protein. Investigating local molecular interactions that could affect such changes based on the newly introduced polarity using PyMol 2.5 turned up no new polar interactions within 4 ångström. However, being located so close to the boundary between the part of the protein that is within the membrane and the extracellular space, another way the change could take effect is by influencing how the protein sits in the cell membrane. The membrane is made up of a lipid bilayer, while the extracellular space is primarily water-based. The new hydrophilic

amino acid could shift this part of the protein away from the hydrophobic lipid layer and shift the transporter tunnel entrance in a way that might impact how well ions are able to enter the cell. It has also been speculated that the rs13107325 variant could affect the protein formation of ZIP8[238].

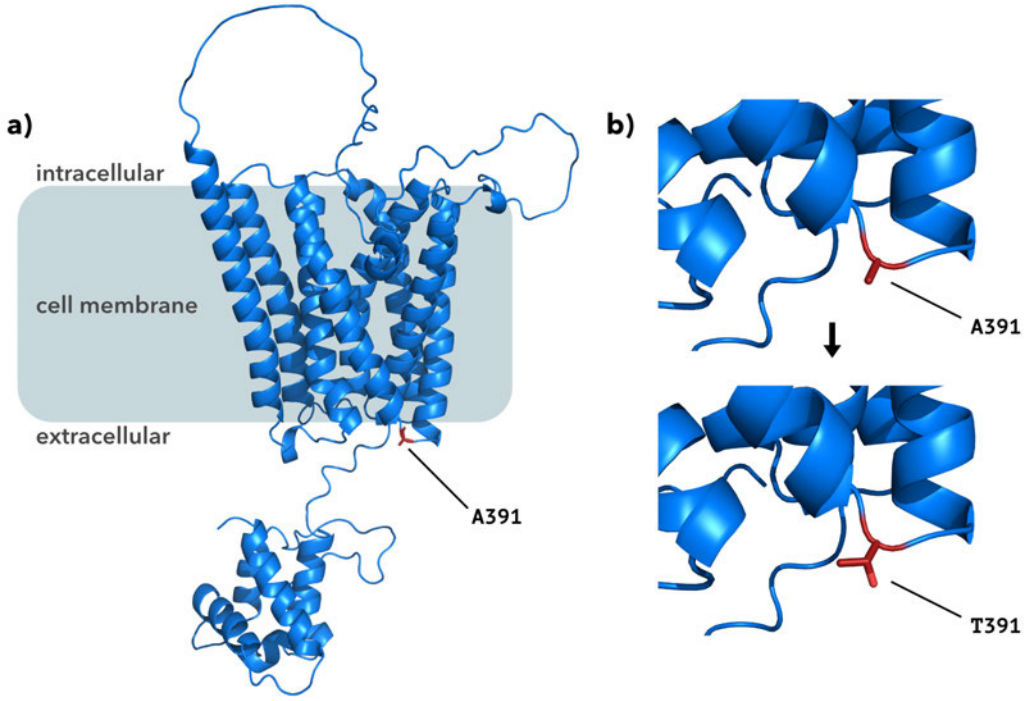


Figure 12: The missense mutation in *SLC39A8* leads to a switch from alanine to threonine at position 391 in the protein ZIP8. These models are based on the AlphaFold structural prediction for Solute carrier family 39 member 8 (*SLC39A8*/ZIP8; UniProtKB entry Q9C0K1). a) ZIP8 is an ion transporter sitting in the cellular membrane. Alanine 391 (A391) is located at the extracellular loop connecting transmembrane domain helices 6 and 7. b) Allele change at SNP rs13107325 leads to an amino acid switch.

It has been shown that the A391T variant has an impact on ZIP8's transporter functionality. The variant was found to lead to impaired manganese homeostasis in mice[239] and decreased manganese and cadmium uptake in chicken cells[240]. Verouti et al. reported an A391T-mediated effect on ZIP8 transporter activity through reduced expression of the protein in vitro, as well as reduced expression in lung and kidney membrane in a mouse model[241]. A disruption of manganese homeostasis has been documented in people who are carriers of A391T as well, with reduced manganese levels in the serum of patients with schizophrenia[242] and lowered plasma manganese levels in idiopathic scoliosis patients[243]. One way by which this is thought to lead to disease is due to its impact on the many cellular glycosyltransferases that require manganese as an

obligatory co-factor to function[239,244]. Notably, a large number of the proteins that make up SARS-CoV-2 need to be glycosylated to be correctly folded and functional[245]. This could result in the virus competing with cellular protein production for a vital and already depleted resource and so further exacerbating viral cell stress, especially as the Human Protein Atlas (Version 22.0) marks ZIP8/*SLC39A8* as enriched in lung alveolar cells[246,247], which are a infected by SARS-CoV-2[248] and damaged in critical Covid-19[249].

Another way by which impaired ZIP8 could be contributing to critical Covid-19 is through its function as a zinc transporter. It has been shown that *SLC39A8* transcription is controlled by NF-κB[250], which is also responsible for driving proinflammatory cytokine expression[251]. Liu et al. describe that in turn ZIP8 then downregulates NF-κB as a negative feedback regulator, via zinc-mediated suppression of the IκB kinase complex[250]. Reduction in ZIP8 transporter activity could lead to prolonged NF-κB activation, which may contribute to the dysregulated and aggravated host inflammatory response thought to be the cause of critical Covid-19[252]. Additionally, it has been shown that zinc deficiency exacerbates mechanical ventilator–induced lung injury in mice[253] and ZIP8 has been reported to be induced by proinflammatory cytokine TNF-α in human lung epithelia, where it is essential to zinc-mediated cytoprotection at the onset of inflammation[254].

There are no current drugs targeting ZIP8 specifically, and as the malfunction appears to be related to impaired transporter activity – rather than for example upregulation, which could be inhibited – it may be particularly difficult to target and influence directly. One feasible method to repair a broken transporter could be through future gene therapy, by correcting the mutation or supplying healthy *SLC39A8* mRNA – this may be particularly helpful to patients with severe ZIP8 deficiencies.

ZIP8's ion transporter function could suggest zinc or manganese homeostasis as targets of intervention in critical Covid-19. Zinc supplements have been suggested as a potential avenue of treatment in Covid-19 before, but trial outcomes are yet inconclusive[255,256]. Previously tested manganese therapy of two patients with severe ZIP8 deficiency showed a promising clinical improvement[257]. However, perhaps the most promising approach in determining if the *SLC39A8* association could point us towards treatment options is to focus on the pathways downstream of ZIP8. One drug that is already used to treat critical Covid-19 that takes effect this way is dexamethasone, which upregulates the inhibitor IκBα and thereby suppresses NF-kB signalling[258,259]. A downstream effector of NF-κB, nuclear factor kappa B kinase subunit beta (IKKβ),

has been suggested as a potential target for treatment as well[259]. However, while dexamethasone has been shown to reduce mortality in Covid-19 patients receiving mechanical ventilation or supplemental oxygen[258], it is important to note that the drug impacts inflammation through multiple pathways[260], the synergy of which may be crucial to its beneficial effects.

At SNP rs687308 on chromosome 6, I found a colocalization between critical Covid-19, childhood asthma (age < 16), suggestive for eosinophilic asthma, rheumatoid arthritis, rheumatoid arthritis (self-reported), other rheumatoid arthritis and other/unspecified rheumatoid arthritis. There are no reported associations with these traits for this SNP or variants in high LD (>0.8) within the GWAS catalog[168]. However, nearby genes *HLA-DRB1* and *HLA-DQA1*, located downstream of the variant, have been associated with these traits before. We reported both of these genes in association with critical Covid-19 in the paper describing the GWAS I used in this analysis[134]. Both genes have also been associated with eosinophil counts[261], childhood asthma[262] and rheumatoid arthritis[263]. They are part of the human leukocyte antigen (HLA) system, which is central to the immune system through its role in antigen presentation and self-recognition by immune cells[264]. This region is gene-dense and highly polymorphic, with a complex linkage disequilibrium structure, which renders it difficult to analyse using standard methods[265]. Imputation of the region is fraught with difficulties as well, requiring specific HLA imputation methods[266], which further complicates the identification of the correct associated variant and affected genes. It is not unlikely that the signal I detected might in actuality stem from a different genetic variant, that has already been described in association with these traits. Due to the high local complexity I am excluding this colocalization from further analysis in the next chapter, as identification of the correct associated gene(s) would be unreliable.

There were three colocalizations between critical Covid-19 and idiopathic pulmonary fibrosis. The first was at rs2897075 on chromosome 7. Allen et al. recently published a paper on genetic overlap between idiopathic pulmonary fibrosis and Covid-19[160], in which they reported colocalizations between the two traits at this variant as well. No lead variant was found on chromosome 7 in the whole-genome sequencing GWAS of critical Covid-19 I used for the analysis, although our meta-analysis of critical Covid-19 GWAS has since identified rs2897075 as a lead variant[131]. The SNPs rs2897075 is an intron variant in the gene *ZKSCAN1*, which encodes a transcription factor. One way by which variants in intron regions can take effect is by impacting splicing (during which introns are removed from the mRNA before it is translated into a protein)[267]. Additionally, they

can potentially influence gene expression[268], which I will focus on in more detail in the next chapter.

The second colocalization between critical Covid-19 and idiopathic pulmonary fibrosis was at rs3742238 on chromosome 13. On chromosome 13, Allen et al. found a colocalization at rs9577395, which is in high LD with the variant I found (0.96 as calculated using the Ensembl Linkage Disequilibrium Calculator[226] with British 1000 Genomes data). In our critical Covid-19 GWAS paper, we reported yet another lead variant in this region – rs9577175, which is in high LD with both rs3742238 and rs9577395 as well (0.89 and 0.86 respectively). Neither of these variants were included in my analysis, though based on their high LD it is likely that all three represent the same signal. If rs9577395 and rs9577175 were present in my analysis, it is possible that either one would have been detected as the colocalization variant. All three SNPs are based close to the gene *ATP11A*, with my colocalization variant rs3742238 situated in its 3 prime UTR region (the untranslated area just after the gene that is part of the mRNA but not the encoded protein), the Allen et al. variant rs9577395 in an intron within *ATP11A* and the critical Covid-19 lead variant rs9577175 located downstream of the gene. These variants are likely to have an impact on gene expression and will therefore be further explored in the next chapter.

The final colocalization between critical Covid-19 and idiopathic pulmonary fibrosis I found was at rs12610495 on chromosome 19. The aforementioned paper by Allen et al. described this colocalization as well[160], and the variant was also a lead variant in our critical Covid-19 GWAS. SNP rs12610495 is an intron variant in the gene *DPP9*, which encodes a serine protease with roles in antiviral signalling[269], antigen presentation[270] and inflammasome activation[271]. I further investigate the variant's potential influence on gene expression in the next chapter.

At variant rs34536443 on chromosome 19, I found a colocalization between critical Covid-19, systemic lupus erythematosus, psoriasis (self-reported), hypothyroidism/ myxoedema (self-reported), rheumatoid arthritis, rheumatoid arthritis (self-reported) and other rheumatoid arthritis. We reported rs34536443 as a lead variant in our paper describing the critical Covid-19 GWAS. Associations with systemic lupus erythematosus[272], psoriasis[273], hypothyroidism[229] and rheumatoid arthritis[274] have been previously described as well, and Wang et al. reported a colocalization between severe Covid-19 and systemic lupus erythematosus at rs34536443[275]. The allele change in rs34536443 leads to a missense variant in the gene *TYK2*. The resulting amino acid

switch from proline to alanine (P1104A) is located in the enzymatic kinase domain of the TYK2 (tyrosine kinase 2) protein and reduces its activity[276]. TYK2 is involved in mediating the immune response through its role in cytokine signalling[277]. We first reported the connection between TYK2 and critical Covid-19 in our 2020 publication[5], although contrary to what is suggested by the rs34536443 risk variant, we found that increased TYK2 expression is detrimental to patients. In the same publication, we suggested the use of the JAK inhibitor drug baricitinib as a potential treatment for severe cases of the disease, as it targets TYK2's signalling pathway[278]. This led to the drug's inclusion in a large clinical trial, which showed a clear therapeutic benefit for hospitalised patients[136]. Baricitinib is also used to treat rheumatoid arthritis[279], and has shown promise in treatment of psoriasis[280]. Clinical trials for its treatment of systemic lupus erythematosus were discontinued after Phase 3 due to a lack of efficacy[281].

Lastly, I found another colocalization between critical Covid-19 and systemic lupus erythematosus at rs73510898 on chromosome 19. We reported rs73510898 as a lead variant in our paper describing the critical Covid-19 GWAS, however the variant does not seem to have been reported in connection with systemic lupus erythematosus before. SNP rs73510898 is an intron variant in the gene *ZGLP1*, which encodes a transcription regulator[282]. The variant is likely to influence gene expression, which is more closely described in the next chapter.

Overall I found seven colocalizations, three of which were not detected as lead variants (or in high LD with lead variants) when looking solely at the critical Covid-19 GWAS. Two variants were missense mutations, leading to amino acid changes in the proteins ZIP8 and TYK2. The rest were intronic or intergenic variants, which likely take effect through changes in gene expression, as is explored in the next chapter.

It is of note that while the pleiotropic variants I investigated in this chapter were the lead colocalization variants between critical Covid-19 and other diseases, they are not necessarily associated variants. Their signal could be conflated with that of the actual associated variants if the two SNPs are in high linkage disequilibrium with each other[283]. Not all possible variants are present in the available data, and the number of SNPs had to be further reduced prior to GWAS-GWAS colocalization analysis to fulfil HyPrColoc requirements, so the actual associated variant may not have been part of the analysis. Additionally, as the Covid-19 GWAS I used in this analysis was measuring severity of the phenotype rather than contraction of the disease, the lead variants are likely not directly causal of the disease, but instead may impact its clinical manifestation.

Of the 56 Covid-19 links with other diseases detected via genetic correlation in chapter 3, only 9 overlap with the 19 colocalizing diseases identified in this chapter. These are diabetes (diagnosed by doctor), hayfever/allergic rhinitis (self-reported), suggestive for eosinophilic asthma, high blood pressure (diagnosed by doctor), hypertension (self-reported), hypothyroidism/myxoedema (self-reported), idiopathic pulmonary fibrosis, osteoarthritis (self-reported) and rheumatoid arthritis (self-reported). This is likely because diseases sharing associated variants affecting a certain gene do not necessarily share common genetic architecture across the whole genome and vice versa.

As described previously, there are several limitations to HyPrColoc as a technique. One issue is that while large numbers of traits can be analysed, SNPs must have been directly typed in every single GWAS. This forced me to exclude many older and smaller GWAS when selecting the dataset, as their inclusion would have removed a large number of variants for all traits. With more time available I could have performed imputation to make it possible to include more GWAS. Additionally, imputing the GWAS that did have enough coverage to be included in the analysis but required exclusion of variants that were present in many other traits could have rendered the iterative analysis process I used to be able to include more variants where possible unnecessary.

Another limitation of the HyPrColoc analysis is its forced assumption of a single associated variant per trait in a local region. Traits are taken out of the regional analysis loop as soon as they are found to colocalize with another trait. For GWAS with multiple regional associated variants this means that potential colocalizations at those additional associated variants can be missed.

However, HyPrColoc remains a well-suited choice for this line of investigation due to its ease of use and scalability. What makes Coloc-based analyses particularly accessible is that they require only summary statistic data, which is the most widely – and often publicly – available GWAS format, while other colocalization methods use harder to acquire individual participant data[284,285]. The main feature introduced by HyPrColoc that made this analysis feasible was its fast performance in large scale analyses, which enabled me to establish colocalizations between many traits simultaneously.

Ultimately, the goal of finding disease relevant genetic variants is to gain further insight into disease pathophysiology and pinpoint potential avenues of treatment. In the next chapter, I will use these results to further investigate the molecular disease architecture of

critical Covid-19 and the diseases sharing putative causal genetic variants by exploring effects on gene expression.

Chapter 5
Investigation of shared molecular disease architecture

5.1 Introduction

In the previous chapter I identified genetic variants that have shared association between critical Covid-19 and several other diseases. In this chapter, I explore their effects on gene expression in order to establish their potential biological consequence.

The majority of genetic variants associated with disease fall within regions that are not protein coding[13]. Five out of seven pleiotropic variants identified in the previous chapter were outwith protein-coding regions. Three of these variants were in intronic regions (genes *ZKSCAN1*, *DPP9* and *ZGLP1*), one in the untranslated 3 prime UTR region (gene *ATP11A*) and one in the intergenic region near genes *HLA-DRB1* and *HLA-DQA1*. Non-coding variants are predicted to impact cell function by affecting gene expression levels, for example by causing changes in transcription factor binding sites or by altering allelic chromatin states[14]. They can also influence regulatory non-coding RNAs such as lncRNAs (long non-coding RNAs) or miRNAs (micro RNAs), as well as their binding sites[286]. Variant changes in transcribed but untranslated regions can regulate gene expression on the RNA level, potentially impacting mRNA stability, localisation and translation[15,16], while intronic variants can dysregulate RNA splicing[17, 287].

Identifying which genes are affected by non-coding variants is challenging, as the influence is not necessarily on the nearest gene[288,289]. Expression quantitative trait locus (eQTL) studies provide a measure of how genetic variation influences RNA expression levels in different tissues[59,85]. Colocalization analysis offers a way to test whether the eQTL and GWAS association signals within a region occur due to the same variant[44,72], providing an opportunity to identify candidate genes affected by the GWAS variants.

Once a list of associated genes is established, the next challenge is to explore how changes in their expression might affect disease biology. This is especially the case if the list is sizeable, rendering manual consultation and interpretation of the literature impractical. An efficient way to pinpoint trends amongst the gene results is via pathway enrichment analysis (PEA), a computational method that identifies biological pathways that are overrepresented amongst the genes compared to what would be expected by chance[290]. A variety of different PEA tools are available to query a broad range of databases, such as pathway databases WikiPathways[291] and KEGG (Kyoto Encyclopedia of Genes and Genomes)[292], or the biomolecular annotations database Gene Ontology[293], which structures genes into biological processes, cellular components and molecular functions.

In this chapter I used colocalization analysis to detect genes that have their expression modulated by genetic variants that I had previously identified to have a shared association between critical Covid-19 and other diseases. I then performed pathway enrichment analysis to inspect trends among the identified genes and pinpoint pathways that may be implicated in critical Covid-19 and the other diseases associated with respective the variants. Finally, I investigated whether the non-coding variants identified in the previous chapter could impact RNA splicing or miRNA regulation.

5.2 Results

5.2.1 Colocalization between GWAS and expression signals

I first used colocalization analysis to identify overlaps between GWAS and expression signals. The analysis region was chosen to match the regions in which the respective GWAS colocalizations were detected in the previous chapter. For each region, I analysed the GWAS with the lowest P value out of the GWAS colocalization group. Expression datasets were chosen to match tissues affected in the diseases in the GWAS colocalization group. Whole blood, lung and peripheral blood mononuclear cells expression data was analysed for all regions. Sigmoid and transverse colon expression data was additionally tested for colocalization at the chromosome 4 locus, due to the presence of Crohn's disease in the GWAS colocalization group.

As my aim was a hypothesis-generating approach, I chose to further investigate all colocalizations with nominal evidence. For this purpose, I classified colocalizations in three different categories according to the level of evidence supporting them.

Colocalizations with a probability of > 0.9 at the default prior colocalization probability 10^{-5} were classified as strongly supported, colocalizations with a probability > 0.7 were classified as having medium support. For low support colocalizations, two rules were used to determine colocalization: The probability of the GWAS and the eQTL datasets to have different association variants was not the leading hypothesis at the default prior colocalization probability 10^{-5} and the probability of GWAS and eQTL colocalization was > 0.5 at a slightly higher prior colocalization probability of 5×10^{-5}. Figure 13 depicts the sensitivity analysis results for the genes associated with the highest probability per analysed region; the remaining sensitivity plots are available in the appendix (Supplementary figures 3-18). Table 8 contains a summary of the identified genes as well as the tissues or cell types in which the GWAS signal-related expression changes were detected. Out of 55 identified genes, 3 were strongly supported (*ATP11A* in blood at the chromosome 13 locus, *PTPRS* in lung and *SAFB2* in CD4 naïve/central memory T cells at the chromosome 19 locus), a further 4 had medium support (*TRIM4* and *MUC12-AS1* in blood and *CPSF4* in classic monocytes at the chromosome 7 locus, and *DAPK3* in CD4 effector memory/ TEMRA cells at the chromosome 19 locus), and 48 had weak support. 40 of the result genes have previously been associated with Covid-19 in the literature: 15 genes were associated with expression changes, 11 genes were linked to Covid-19 by genetic association, 6 genes were associated via computational prediction, 3 genes were associated via theoretical prediction, 3 genes were identified as potential drug targets, 2 were interacting with virus proteins and 1 gene each was associated with methylation changes, ubiquitination changes, potential drug target interaction and virus enhancing effects (Table 8).

Table 8: Genes with expression affected by the GWAS signals shared between critical Covid-19 and other diseases. The LD block is the analysed region, chosen to match the regions used for HyPrColoc. The GWAS colocalization SNP is the lead variant detected by HyPrColoc for colocalizations between critical Covid-19 and other diseases in chapter 4. Genes and their respective tissues are in bold if they have been found with medium (colocalization probability >0.7) or high support (colocalization probability >0.9).

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
AP001816.1	CD8 effector memory T cell	rs13107325	chr4 100678360-103221356	

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
BANK1	naïve/immature B cell	rs13107325	chr4 100678360-103221356	expression changes[294]
BDH2	natural killer recruiting cell	rs13107325	chr4 100678360-103221356	
MANBA	sigmoid colon, CD8 effector memory T cell	rs13107325	chr4 100678360-103221356	
NFKB1	blood	rs13107325	chr4 100678360-103221356	genetic association[295]
PPP3CA	transverse colon	rs13107325	chr4 100678360-103221356	expression changes[296]
SLC39A8	CD8 S100B T cell	rs13107325	chr4 100678360-103221356	expression changes[297]
SLC9B1	natural killer cell	rs13107325	chr4 100678360-103221356	computational prediction[298,299]
AC004893.11	CD4 naïve/central memory T cell	rs2897075	chr7 98715474-100196651	
AP4M1	CD4 naïve/central memory T cell	rs2897075	chr7 98715474-100196651	
ATP5J2	memory B cell, CD8 effector memory T cell	rs2897075	chr7 98715474-100196651	
BUD31	CD8 effector memory T cell	rs2897075	chr7 98715474-100196651	ubiquitination changes[300]

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
CNPY4	CD8 naïve/central memory T cell, CD4 naïve/central memory T cell	rs2897075	chr7 98715474-100196651	genetic association[131,233]
CPSF4	**classic monocyte**	rs2897075	chr7 98715474-100196651	computational prediction[301]
EPO	lung	rs2897075	chr7 98715474-100196651	potential drug target[302],theoretical prediction[303]
FIS1	blood	rs2897075	chr7 98715474-100196651	potential drug target[304]
GNB2	lung	rs2897075	chr7 98715474-100196651	expression changes[305,306]
GPC2	CD8 S100B T cell	rs2897075	chr7 98715474-100196651	
MEPCE	plasma cell	rs2897075	chr7 98715474-100196651	genetic association[307]
MOSPD3	natural killer recruiting cell	rs2897075	chr7 98715474-100196651	genetic association[131]
MUC12-AS1	**blood**	rs2897075	chr7 98715474-100196651	
PCOLCE	lung	rs2897075	chr7 98715474-100196651	
PPP1R35	naïve/immature B cell	rs2897075	chr7 98715474-100196651	

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
SRRT	CD8 effector memory T cell	rs2897075	chr7 98715474-100196651	virus protein interaction[308]
TMEM225B	CD8 naïve/central memory T cell	rs2897075	chr7 98715474-100196651	
TRIM4	**blood**, natural killer cell	rs2897075	chr7 98715474-100196651	computational prediction[309]
TRIM56	lung	rs2897075	chr7 98715474-100196651	expression changes[310],theoretical prediction[311]
TRIP6	non-classic monocyte	rs2897075	chr7 98715474-100196651	expression changes[312]
ZCWPW1	classic monocyte	rs2897075	chr7 98715474-100196651	
ZNF789	classic monocyte	rs2897075	chr7 98715474-100196651	methylation changes[313]
ZSCAN21	CD4 effector memory T/ TEMRA cell	rs2897075	chr7 98715474-100196651	genetic association[131]
ATP11A	**blood**	rs3742238	chr13 112247592-113572488	genetic association[134]
DCUN1D2	CD8 naïve/central memory T cell	rs3742238	chr13 112247592-113572488	expression changes[314]
LAMP1	CD8 effector memory T cell	rs3742238	chr13 112247592-113572488	enhances virus[315]

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
MCF2L	CD4 effector memory T/ TEMRA cell	rs3742238	chr13 112247592-113572488	expression changes[316]
TMCO3	CD8 S100B T cell	rs3742238	chr13 112247592-113572488	expression changes[314]
TUBGCP3	classic monocyte	rs3742238	chr13 112247592-113572488	computational prediction[317]
AC005523.3	natural killer cell	rs12610495	chr19 4348967-5811852	
APBA3	CD8 naïve/central memory T cell, natural killer recruiting cell	rs12610495	chr19 4348967-5811852	genetic association[318]
C19orf70	CD8 effector memory T cell, CD4 naïve/central memory T cell, memory B cell, classic monocyte, naïve/immature B cell	rs12610495	chr19 4348967-5811852	virus protein interaction[308]
DAPK3	**CD4 effector memory/ TEMRA cell**	rs12610495	chr19 4348967-5811852	expression changes[319]
DPP9	natural killer recruiting cell	rs12610495	chr19 4348967-5811852	genetic association[5]
EBI3	naïve/immature B cell	rs12610495	chr19 4348967-5811852	expression changes[320,321]
FEM1A	plasma cell	rs12610495	chr19 4348967-5811852	expression changes[322]
KDM4B	plasma cell	rs12610495	chr19 4348967-5811852	theoretical prediction (preprint)[323]

gene	tissue or cell type(s)	GWAS colocalization lead SNP	LD block	previous Covid-19 association
MAP2K2	blood	rs12610495	chr19 4348967-5811852	expression changes[324,325],potential drug target[326,327]
PIAS4	natural killer recruiting cell	rs12610495	chr19 4348967-5811852	genetic association[131],potential drug target interaction[328]
PTPRS	**lung**	rs12610495	chr19 4348967-5811852	genetic association[329]
SAFB2	**CD4 naïve/central memory T cell**, natural killer cell	rs12610495	chr19 4348967-5811852	computational prediction[330]
SH3GL1	CD4 effector memory/ TEMRA cell	rs12610495	chr19 4348967-5811852	computational prediction[331]
SIRT6	CD8 naïve/central memory T cell	rs12610495	chr19 4348967-5811852	
STAP2	CD8 effector memory T cell	rs12610495	chr19 4348967-5811852	
TICAM1	natural killer recruiting cell	rs12610495	chr19 4348967-5811852	genetic association[120]
TJP3	CD8 effector memory T cell	rs12610495	chr19 4348967-5811852	expression changes[332]
TNFAIP8L1	CD4 SOX4 T cell	rs12610495	chr19 4348967-5811852	expression changes[322,333]

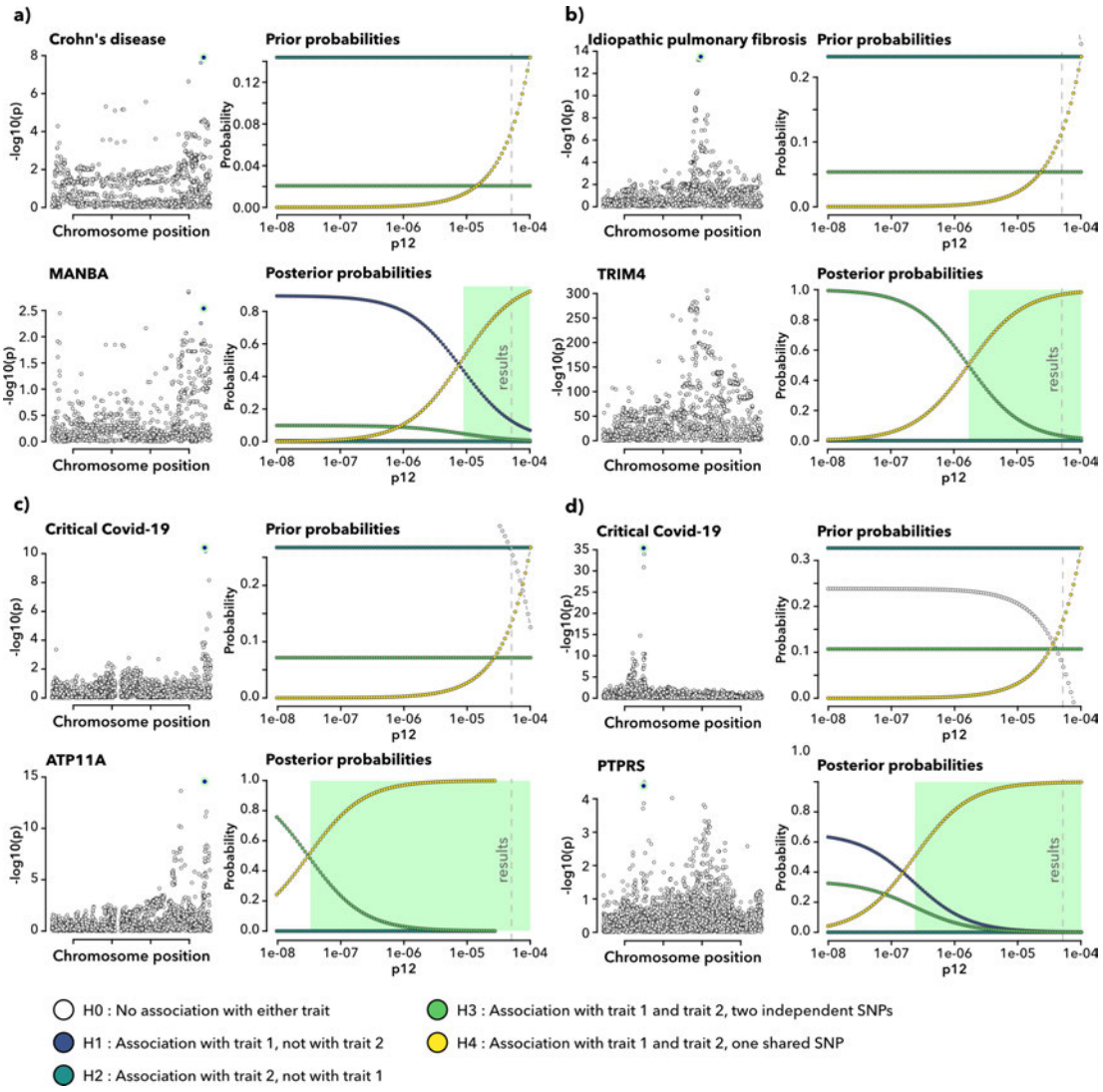


Figure 13: Sensitivity plots for GWAS-eQTL colocalizations with the highest probability per region. Each subfigure (a-d) describes the colocalization between a GWAS and an eQTL signal: a) Crohn's disease and *MANBA* expression in sigmoid colon on chromosome 4 region 100678360-103221356, b) idiopathic pulmonary fibrosis and *TRIM4* expression in blood on chromosome 7 region 98715474-100196651, c) critical Covid-19 and *ATP11A* expression in blood on chromosome 13 region 112247592-113572488 and d) critical Covid-19 and *PTPRS* expression in lung on chromosome 19 region 4348967-5811852. On the left are the local manhattan plots of the colocalizing traits with the lead colocalization variant marked. Hypothesis outcome probabilities for a random SNP in the region are displayed on the top right, while hypothesis outcomes for the lead colocalization SNP are on the bottom right. The x-axis describes changes in p12, the assumed prior probability that a random SNP in the region is jointly associated with both traits. The green shaded region marks the needed H4 > 0.5 probability threshold over the range of prior probabilities for which it is supported.

In the previous chapter, we saw that region 9238393-11284028 on chromosome 19 had multiple variants associated with critical Covid-19, with a colocalization between

Covid-19 and lupus at rs73510898, and one with Covid-19, lupus and psoriasis at rs34536443. As these lead variants were in close proximity to each other (10463118 and 10416444), I used the genetic fine-mapping method SuSiE, which is able to establish multiple associated variants, to identify lead variants for GWAS and expression data in the area prior to calculating colocalization with coloc.susie[45]. No GWAS-eQTL colocalizations were detected for this area.

5.2.2 Pathway enrichment analysis

I performed pathway enrichment analysis to investigate if any of the identified genes were members of shared pathways, which could pinpoint pathways of particular importance to the associated diseases. To identify replicable results I used three web-based tools: g:Profiler g:GOSt[186], Enrichr[187,188] and GeneTrail[189]. Three KEGG database[292] pathways were significantly enriched for the expression-affected genes: Kaposi sarcoma-associated herpesvirus infection, PD-L1 expression and PD-1 checkpoint pathway in cancer and B cell receptor signaling pathway (Figure 14 a). The WikiPathways database[291] entry Cardiac hypertrophic response was also found to be significantly enriched.

Pathway enrichment analysis for subsets of a gene list can reveal more precise results that may be obscured when studying the entire list[334]. I therefore performed additional enrichment analyses for the subsets of genes detected for each region, a sub-selection of only the genes identified with strong to medium confidence, as well as the functional subnetworks of genes detected by STRING web application[194] (Figure 14 b). No results of interest were detected for the subset analyses, with result pathways reported as enriched for only one or two genes.

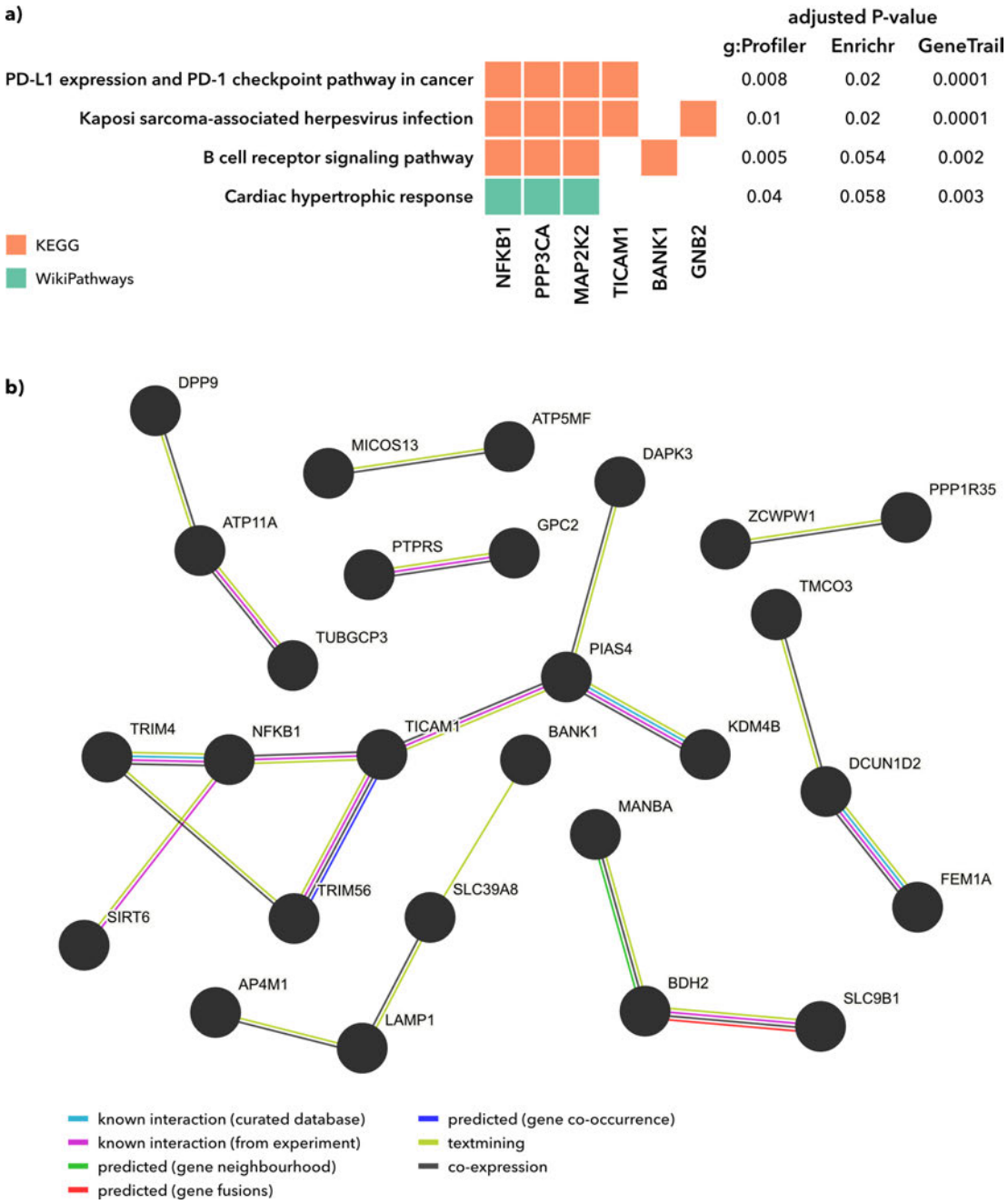


Figure 14: Trends among the genes with expression affected by the GWAS signals shared between critical Covid-19 and other diseases. a) Pathway enrichment analysis results for g:Profiler g:GOSt, Enrichr and GeneTrail. b) Functional subnetworks among the expression-modulated genes identified by the STRING web application.

5.2.3 Variant effects on splicing and RNA regulation

In the previous chapter I hypothesised that the intron variants rs2897075 in *ZKSCAN1* and rs73510898 in *ZGLP1* may affect those genes by influencing their expression. However, neither gene was detected to have an overlapping eQTL in the colocalization

analysis. To further ascertain whether they could still be affected by the variants, I used the RegSNPs-intron web application, which evaluates the disease-causing probability of intron variants based on their impact on splicing regulation and the resulting effect on protein-structure features[191]. Both variants were classified as benign by the algorithm, with variant rs2897075 having an estimated disease-causing probability of 0.2 and a false-positive rate of 0.77 and variant rs73510898 having an estimated disease-causing probability of 0.29 and a false-positive rate of 0.51.

As 3 prime UTR regions have micro RNA binding sites, I searched the miRNASNP database of SNPs and disease-related variations in miRNAs and miRNA targets for the *ATP11A* variant rs3742238 (C/T)[192]. The variant was found to be associated with a gain of 11 and a loss of 3 miRNA target sites when compared to the wild type.

5.3 Discussion

In this chapter I explored the molecular consequences of the pleiotropic variants shared between critical Covid-19 and other diseases established in the previous chapter. I identified genes that may have their expression modulated by these variants and investigated trends among these genes using pathway enrichment analysis. Additionally, I explored whether the non-coding pleiotropic variants could have an impact on mRNA splicing and microRNA regulation.

I identified 55 genes with potential expression changes related to the pleiotropic disease variants, 40 of which have previously been connected to Covid-19 in the literature. In this chapter, I established colocalization between GWAS and eQTL signals using Coloc. Although using HyPrColoc – as in the previous chapter – would have allowed testing of all disease and expression data at the same time, the need for complete overlap of available SNPs in every dataset combined with the lack of coverage in much of the expression data would have resulted in a considerable loss of information. As my aim was a hypothesis-generating study, I analysed all nominally supported colocalizations rather than only focusing on those with strong supporting evidence. This likely resulted in some false positive results. Follow up analyses and in vitro confirmation will be necessary to determine which expression changes are truly related to the GWAS signal.

One potential choice of follow up analysis would be Mendelian randomization which can be valuable as a complementary method to colocalization[67]. Mendelian randomization makes use of the random allocation of alleles in the population to select

groups in which to study the effect of an exposure on an outcome, akin to a randomized controlled trial[335]. It assumes that the studied genetic variants affect the exposure trait – and therefore the outcome – in the absence of additional confounders linking the genetic variants to the outcome. While colocalization as a technique is agnostic to causality between the studied traits, Mendelian randomization seeks to estimate causality of the exposure on the outcome, allowing for the prediction of a putative causal relationships between potential drug targets and diseases. Although this type of analysis is constrained by strong assumptions, such as the associated genetic variants having been correctly identified, it offers an additional source of information on potential links between gene expression and GWAS phenotypes.

One of my regions of interest, 9238393-11284028 on chromosome 19, contained two different GWAS colocalizations. As the pleiotropic lead variants were in close proximity to each other, I attempted to use SuSiE and coloc.susie to identify potentially multiple signals in the area and calculate their colocalization. I detected no GWAS-eQTL colocalizations, which could mean that these variants simply have no effect on gene expression. However, the SuSiE algorithm needed many iterations to detect lead variants in the expression datasets, which could also mean that the tested linkage disequilibrium reference panels were not a good match for the expression data. The accuracy of the results using this method are greatly dependent on the accuracy of the disequilibrium reference[336,337], so it is possible that a better suited reference panel would enable the detection of GWAS-eQTL colocalizations.

PEA results found an enrichment in four pathways: PD-L1 expression and PD-1 checkpoint pathway in cancer, Kaposi sarcoma-associated herpesvirus infection, B cell receptor signaling pathway and cardiac hypertrophic response. Programmed cell death 1 (PD-1) and its ligand (PD-L1) play a key role in cancer by inhibiting immune response and promoting self-tolerance through their regulation of T cells[338]. The pathway has also been implicated in several autoimmune diseases, including type 1 diabetes, multiple sclerosis, inflammatory bowel disease and rheumatoid arthritis[339]. PD-L1 levels have been proposed as a potential biomarker for predicting Covid-19 severity[340 342]. Additionally PD-1/PD-L1 and their downstream pathways have also been proposed as potential therapeutic targets in Covid-19[340,342].

Kaposi sarcoma is a viral-induced cancer, caused by the human herpesvirus 8 and primarily affecting immunodeficient individuals[343]. B cells are an important part of the humoral adaptive immune response, binding foreign antigens and producing an

antibody response[344]. Antigens binding the B cell receptor, a membrane protein complex, lead to the induction of downstream signalling pathways activating the expression of genes involved in processes such as B cell proliferation, differentiation and immunoglobulin production[345]. As a viral infection and a fundamental function of the immune system important to antiviral defense, it stands to reason that both pathways would have overlapping biology with critical Covid-19.

Cardiac hypertrophy is a growth of the heart in response to hemodynamic stress, for example as an escalation of hypertension and valvular disease. While presumed to compensate for the stressors in the short term[346] it leads to an increased risk of further cardiovascular disease and mortality[347]. On a molecular level, multiple signalling pathways are involved in the changes in gene expression that lead to hypertrophy, including Ca^{2+}, mitogen-activated protein kinases (MAPK) and NF-κB signalling cascades[348]. As highlighted in chapter 3, there is a strong link between severe forms of Covid-19 and cardiovascular disease[214 217] that may be partially due to a shared underlying disease biology[157].

NFKB1, *PPP3CA* and *MAP2K2* were result genes associated with all of these pathways. *NFKB1* encodes nuclear factor kappa B subunit 1, the DNA binding subunit of the transcription factor NF-κB[349]. Activated by a variety of stimuli, such as cytokines, oxidative stress, bacterial or viral infection, NF-κB can stimulate the expression of a large number of target genes depending on cell type and state[350]. NF-κB regulates cell differentiation, proliferation and survival[351,352] and plays a vital role in immune system responses[353]. *PPP3CA* encodes the calmodulin-binding catalytic subunit of calcineurin, a serine/threonine protein phosphatase involved in several Ca^{2+}-dependent cellular signalling pathways, including T cell regulation[354,355]. *MAP2K2* encodes Mitogen-Activated Protein Kinase Kinase 2 (MEK2)[356]. MEK2 is a dual-specificity kinase that phosphorylates and thereby actives ERK1 and ERK2[357]. ERK1 and ERK2 have hundreds of identified substrate and interaction partners, with their signalling pathways regulating many core cellular processes such as proliferation, differentiation, survival and cell motility[358,359].

Three of the PEA result pathways were also enriched for additional genes among my results. Both PD-L1 expression and PD-1 checkpoint pathway in cancer and Kaposi sarcoma-associated herpesvirus infection were also enriched for *TICAM1*, which encodes the protein TIR Domain Containing Adaptor Molecule 1. TICAM1 mediates interaction between Toll-like receptors and downstream effector molecules and is

involved in activation of NF-κB[360]. Kaposi sarcoma-associated herpesvirus infection was additionally also enriched for *GNB2*, which encodes for a G protein (Guanine nucleotide-binding protein) beta subunit, signalling factors involved in many cellular processes[361], including the ERK1/2 pathway[362]. The B cell receptor signalling pathway was additionally enriched for *BANK1*, which encodes for a B-cell-specific scaffold protein that is involved in B-cell antigen receptor induced Ca^{2+} mobilisation[363].

None of the PEA result pathways are linear cascades of interactions; rather, they are a network of multiple different signalling pathways, triggered by several often not interconnected signals and resulting in multiple expression programme changes. The enriched genes are part of different pathways within these networks in the KEGG[292] and Wikipathways[291] databases, which does not provide a strong indication for any of them as a potential drug target. *NFKB1* is part of tumor necrosis factor (TNF) signalling in the Kaposi sarcoma and cardiac hypertrophic response pathways, marked for Toll-like receptor (TLR) signalling and T cell receptor signalling pathway in the PD-L1/PD-1 axis and simply as NF-κB signalling pathway in the B cell receptor signalling network. *PPP3CA* is involved in the Ca^{2+} signalling cascades in all of the identified enriched pathways, as is *BANK1* in B cell receptor signalling. *MAP2K2* is generally marked as part of the MAPK/ERK signalling pathway, as well as chemokine signalling in Kaposi sarcoma-associated herpesvirus pathway together with *GNB2*. The role of these proteins and the changes in expression provoked by these pathways can strongly depend on cell type and state, making it difficult to predict whether they may play a similar role in critical Covid-19 as in the identified enriched pathways. Furthermore the expression changes in relation to the disease-associated variants I identified were predicted for varying tissues and cell types, rendering assignment to a single enrichment pathway tenuous. Future analysis of further cell type specific gene expression data could be useful in answering these questions.

Despite these complications, the identified pathways may still be valid targets for treatment of Covid-19. Among the PEA results, PD-L1/PD-1 presents perhaps the strongest opportunity for a potential drug target, as it is a single node impacting multiple of the enrichment gene pathways, affecting NF-κB, MAPK/ERK and Ca^{2+} signalling downstream. As noted above, it has previously been suggested as a potential therapeutic target in Covid-19[340,342]. Calcium channels have been suggested as a treatment strategy for Covid-19 due to the importance of Ca^{2+} in viral infection cascades[364]. MAPK signalling related biomarkers may be informative of different clinical features in Covid-19, potentially identifying patients at higher risk of severe

complications[365]. MEK1/2 inhibitor ATR-002, which is currently undergoing clinical trials as an anti-influenza drug, has been proposed as a potential therapeutic agent in Covid-19 as well, showing promise by blocking SARS-CoV-2 propagation and decreasing virus-induced expression of pro-inflammatory cytokines[326]. Dexamethasone, which upregulates an NF-κB inhibitor, is already successfully used in the treatment of hospitalized Covid-19[258] and further NF-κB inhibiting drugs have been proposed as well[366].

Another potential consequence of non-coding genetic variants I explored in this chapter was their impact on RNA splicing or microRNA regulation. As intron variants are known to potentially dysregulate RNA splicing[17,287], I investigated whether the three pleiotropic intron variants identified in the last chapter may have such an impact by consulting the regSNP-intron web application. Variants rs2897075 in *ZKSCAN1* and rs73510898 in *ZGLP1* were both classified as benign and likely do not have a relevant impact on mRNA splicing. A future follow up analysis could confirm this by integrating splicing quantitative trait loci data[367]. There were no regSNP-intron results for the rs12610495 variant in *DPP9*, possibly because it has been classified as an intron or non coding transcript exon variant depending on the transcript. *DPP9* has multiple isoforms[368]. The variant associated with critical Covid-19 and idiopathic pulmonary fibrosis, rs12610495 (G>A), has been shown to affect *DPP9* splicing, increasing the excision of an intron in lung[369]. This may contribute to deciding which of the two major isoforms of the protein is produced, a cytosolic or nucleus-localized version[369,370].

Micro RNAs regulate biological pathways by binding target sites in the 3 prime UTR region of mRNAs, sequestering or degrading the mRNA and thereby silencing gene expression[371]. Single nucleotide changes in miRNA target sequences can disrupt miRNA binding or create new target sequences for different miRNAs[372]. The miRNASNP database marks the pleiotropic variant I identified in the 3 prime UTR region of *ATP11A* as leading to 11 gained and 3 lost miRNA targets. This suggests that the variant change might affect *ATP11A* expression by changing miRNA regulation.

As discussed in the previous chapter, the putative pleiotropic variants I explored in this chapter were the lead colocalization variants between critical Covid-19 and other diseases, but are not guaranteed to be the actual associated variants. Their signal could be conflated with that of the true associated variants if the two SNPs are in high linkage disequilibrium with each other[283]. Not all possible variants are present in the available data, and the number of SNPs had to be further reduced prior to GWAS-GWAS

colocalization analysis, so the true associated variant may not have been part of the analysis. This is an important caveat when considering these results.

Disease associated genetic variants can offer insight into their molecular pathophysiology and point to potential drug targets. In this chapter, I identified genes affected by the pleiotropic variants critical Covid-19 shares with other diseases and explored their signalling pathways. In future, follow up Mendelian randomization analysis and in vitro testing will help provide further evidence on whether the uncovered genes are truly associated with the genetic variants of interest.

Chapter 6
Genetic colocalization across disease identifies drug repurposing candidates

6.1 Introduction

In the previous two chapters, I demonstrated the potential of colocalizations between diseases for drug target discovery. In this chapter I widen the lens to look at local connections across a wide range of diseases using HyPrColoc and investigate how these indications of shared biology can lead towards possible candidates for drug repurposing.

As seen in chapter 4, HyPrColoc is a method for exploring localised connections between diseases, using a Bayesian algorithm to calculate if traits share associated variants in a genetic region. This analysis can be performed across the whole genome by splitting it into subsections based on local linkage disequilibrium. Due to the large number of connections detected within my dataset of 228 disease GWAS, I performed a simplified version of the post-analysis quality control, accepting all colocalizations meeting the standard probability threshold without further investigation into potentially multi-associated regions. This approach may lead to more false positive results, but limits the number of false negatives.

In order to prioritise which disease connections might be the most conducive to clinical applications based on the shared biology, I screened my HyPrColoc results for drug repurposing candidates. Finding new therapeutic applications for existing drugs is an attractive opportunity, as the discovery of new therapies is expensive, time intensive and fraught with failure at all stages[140 142]. The increase in available "omics" and other biomedical data, such as medical records, has made computational repurposing approaches a promising avenue of exploration. As genetically supported targets increase the likelihood of successful drug application[4,146], genetic data is a valuable source of information to integrate in these efforts[147]. Drug repurposing analyses using phenotype-based disease similarity[373,374] and analyses using GWAS candidate genes[19], often in combination with other biological data[148,152,153], have led to new drug recommendations.

Integrating large-scale colocalization data could add to existing approaches by opening up the possibility to detect new connections, as the method improves power compared to investigating only those variants that reach genome wide significance within each GWAS separately. Additionally, studying genetic links between diseases at the level of the variant can reveal connections that otherwise might stay hidden due to erroneous assumptions about the affected genes in the disease GWAS, as matching GWAS signals to target genes is a difficult process.

In this chapter I performed genome-wide colocalization analysis of 228 disease GWAS to identify points of potential overlap in their underlying biological mechanisms. In order to prioritise amongst the colocalizations, I integrated drug data to reveal connections that may present candidates for drug repurposing.

6.2 Results

6.2.1 Colocalizations between 228 disease GWAS

I used HyPrColoc to detect localised genetic pleiotropy between a wide range of diseases. Briefly, this method allows for the detection of SNPs that are associated with multiple phenotypes and may or may not be causal for them. An in depth examination like in chapter 4, where I investigated colocalizations specific to critical Covid-19 with several follow up analyses to account for multiple associated alleles in the area, would be beyond the scope of this thesis. I employed a simplified version of the analysis, performing only one follow up test to identify cases where although the colocalization group as a whole did not reach the recommended probability threshold, a subset of traits did. Out of 945 results, 628 reached the > 0.7 probability threshold. An additional 149 colocalization subgroups reached the threshold after exclusion of traits with weaker evidence. As I uncovered more exact information about the colocalizations involving critical Covid-19 through more in-depth testing in chapter 4, I corrected those colocalizations to account for the identified multi-association loci. This gives us a final colocalization count of 779.

Splitting groups of colocalizing diseases into pairwise colocalizations adds up to 4001 colocalizations between 123 diseases. These pairwise connections are depicted in Figure 15. As many of these diseases shared more than one colocalization SNP, there are a total of 1045 connections between the 123 diseases. An interactive version of this figure can be found in the digital appendix at https://marie-zz.github.io/digital_appendix/ (Supplementary Interactive Figure 2).

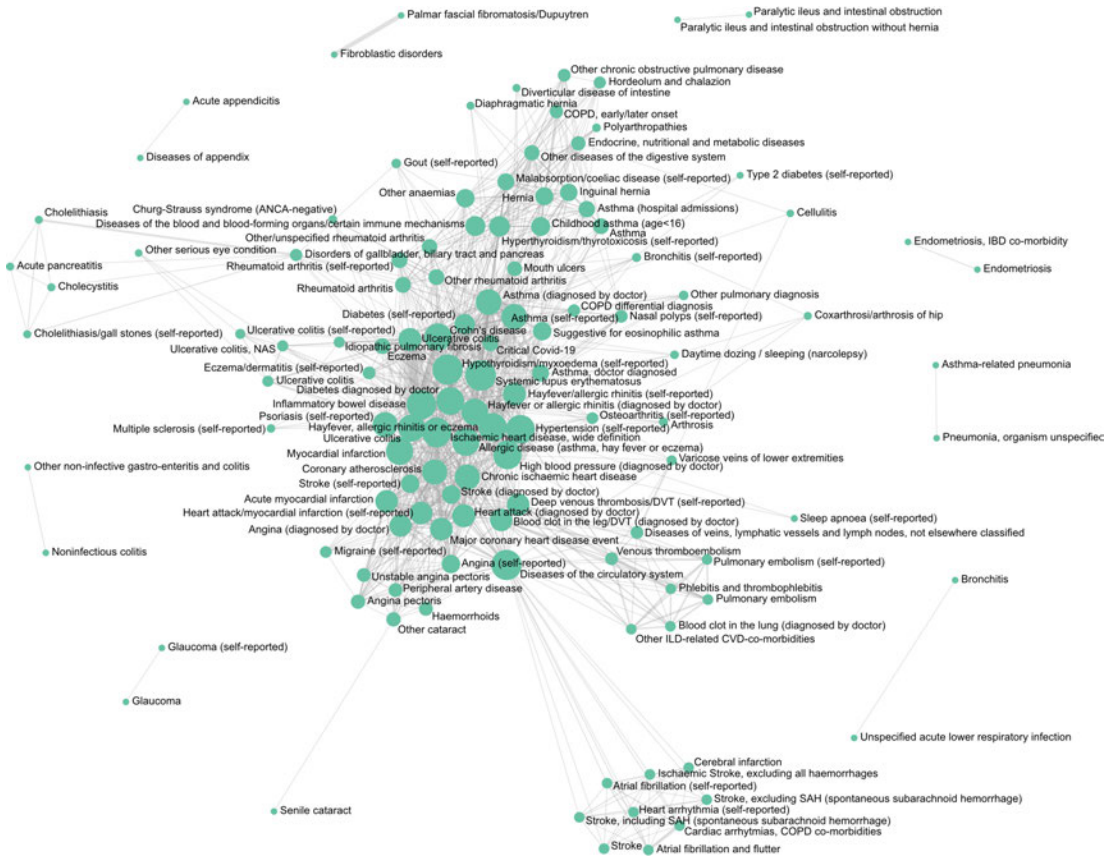


Figure 15: The HyPrColoc disease network. The network encompasses 4001 links between 123 disease nodes. Diseases are connected if they share colocalizations with each other. The edges are weighted by the number of colocalizations between the traits. Disease nodes are bigger when the disease shares links with more nodes.

The colocalizations differed in how many diseases were involved. 533 colocalizations were between two traits, 246 between groups of 3-25 (Figure 16 a). Of the 228 diseases I analysed, 105 showed no colocalizations with any other disease. To test if this could be related to the power of the involved GWAS, I explored whether there was a difference in effective sample size between GWAS with colocalizations and those without (Figure 16 b). Bootstrapping the median differences of the two distributions revealed that on average effective sample sizes for diseases with colocalizations was significantly greater by 9479.5 (95% CI [3174.8 19356.6], p=0.021). This suggests that the GWAS without any colocalization may in fact have shared associations with other GWAS in the dataset, but do not have a large enough sample size for these associations to be detectable.

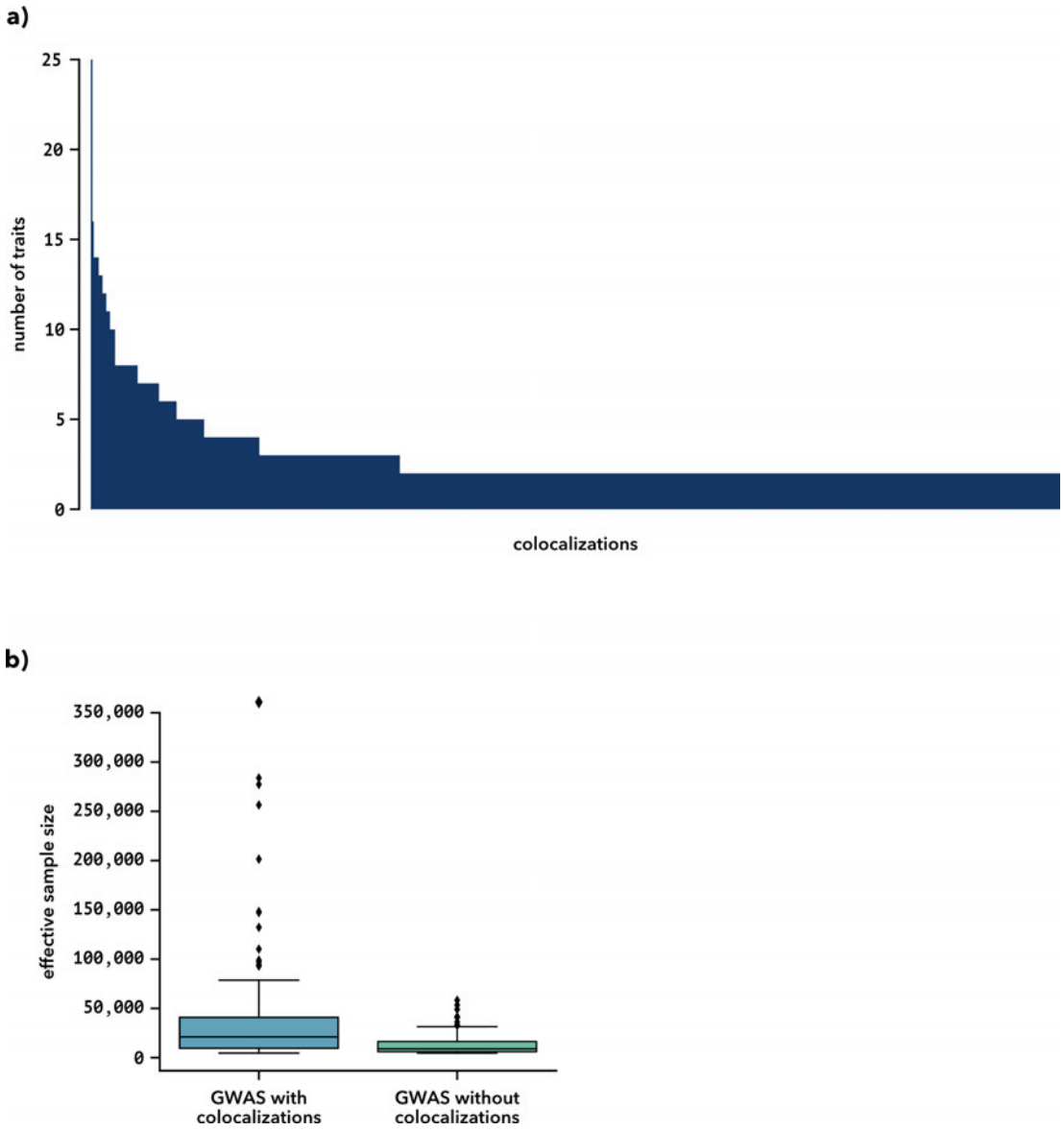


Figure 16: Colocalizations details a) Distribution of the number of traits in a colocalization group. Most colocalizations occurred between two diseases, but colocalization group sizes of up to 25 diseases were detected as well. b) Comparison of the effective sample sizes of GWAS that showed colocalizations and those were no colocalizations were found. Diseases with colocalizations had a statistically higher likelihood of having a greater effective sample size (On average greater by 9479.5, 95% CI [3174.8 19356.6], p=0.021, bootstrapping the median differences).

The disease network of HyPrColoc connections is densely packed in the centre, with many different types of diseases connected to each other (Figure 17). Disease clusters on the outer regions of the network tend to be within disease types, as for example with certain diseases of the digestive system involving the gallbladder and pancreas, as well as a circulatory system disease cluster with several stroke and heart arrhythmia GWAS. Immune system diseases, which have been reported to be highly pleiotropic[139], largely

find themselves at the centre of the network (Figure 18). While they are tightly connected with each other, they are also colocalizing with many non-immune system diseases. The hyperconnected and complex nature of the network renders it difficult to identify which connections have the potential to be the most informative. Moreover, just because two diseases are more tightly connected and have many colocalizations between them, those colocalizations are not necessarily more likely to add to our biological understanding compared to two diseases connected through a single colocalization. This poses the problem of how to best gain useful information from this rich dataset.

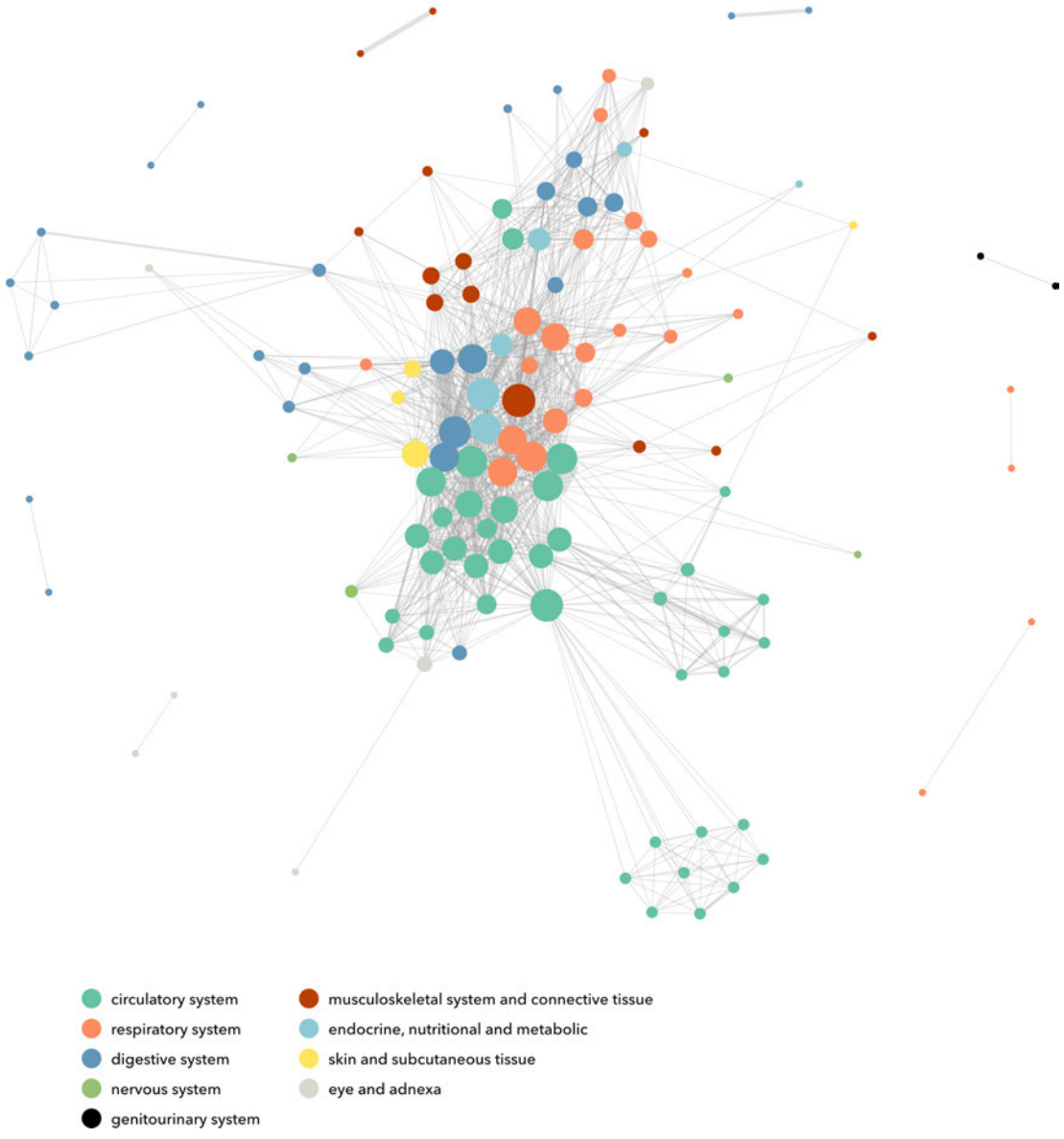


Figure 17: Disease groupings in the HyPrColoc network. Disease nodes are coloured based on which part of the body they affect.

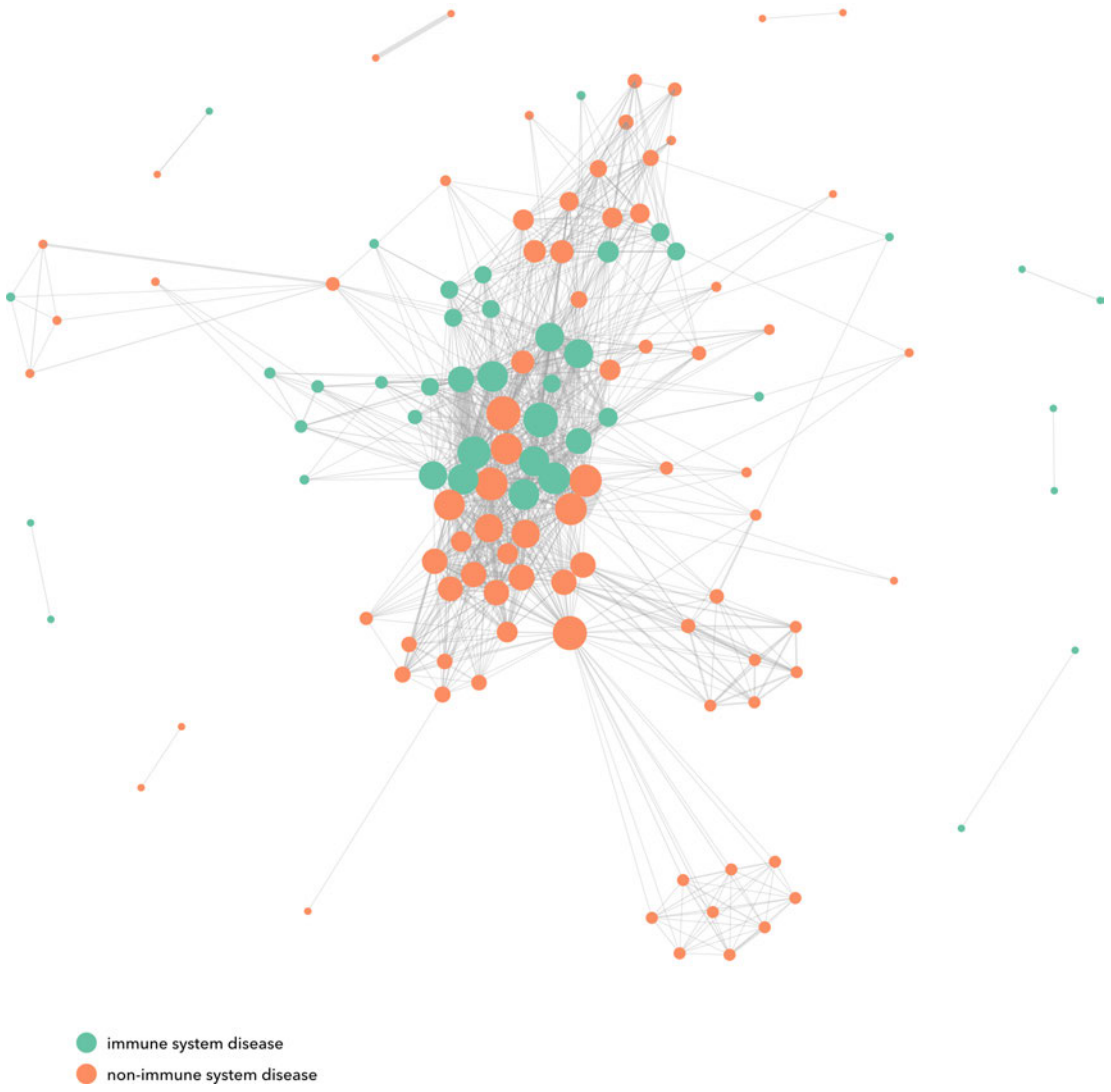


Figure 18: Immune system related diseases in the HyPrColoc network.

6.2.2 Colocalization as a source for drug repurposing

Although delving deeper into the connection SNPs and related genes can reveal additional insight into the linked diseases (as shown in the previous two chapters), exploring all 779 connections identified in this chapter in such detail would be a considerable undertaking. To prioritise colocalizations with the potential to be particularly insightful, I sought to identify potential candidates for drug repurposing. Where two diseases are connected, could their colocalization variant reveal drug repurposing opportunities?

Colocalization between diseases is suggestive of a potential shared biological pathway that may be targetable by the same drug. I first set out to test if disease pairs that had

colocalizations with each other were overall more likely to share drugs than pairs without using drug information from the ChEMBL bioactivity database obtained via Open Targets[176,177]. I calculated the probability of a disease to share at least one drug with another disease given that they share a genetic link, and the probability of a disease to share at least one drug with another disease given that they do not share a genetic link. Bootstrapping the median difference between the two probability distributions (Figure 19) revealed that a shared colocalization increases the probability of sharing a drug by 9.13% (95% CI [6.59, 15.88], p=0.00013).

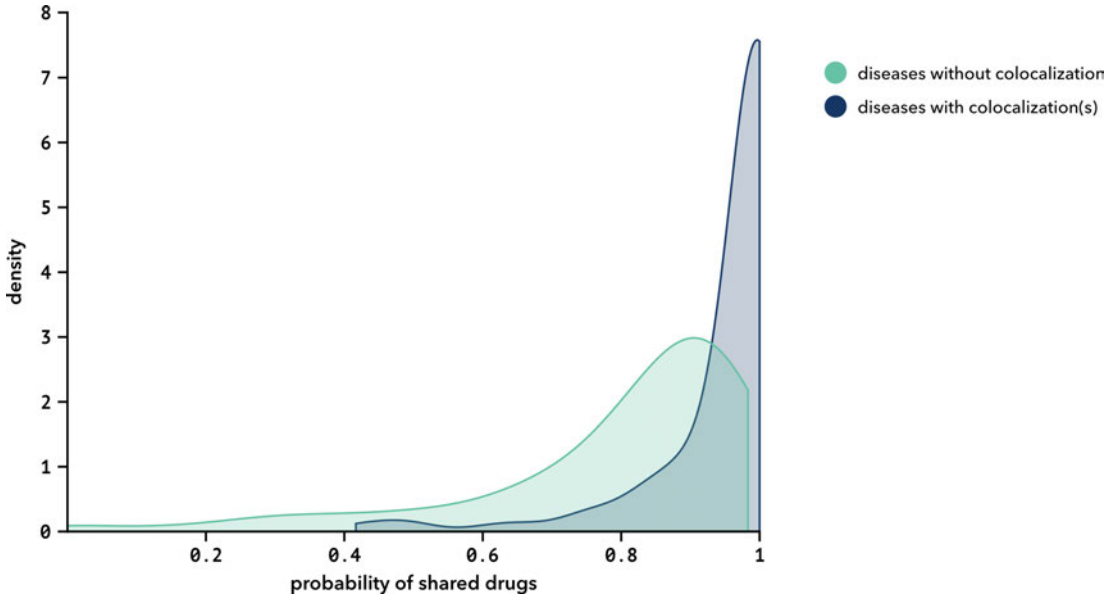


Figure 19: Are diseases that colocalize more likely to share drugs? Probability density distributions for diseases that share at least one colocalization to share at least one drug, compared to diseases without colocalizations. Kernel density estimate plot was chosen for a better comparison between the two distributions as the two populations have a big difference in sample size (1045 disease pairs with colocalizations, 6458 disease pairs without).

To see if colocalization SNPs might reveal drug repurposing opportunities (Figure 20) I first asked: Are there drugs that are only used or currently in trial for one of the connected diseases? To answer this question I obtained the drug information from the ChEMBL bioactivity database via Open Targets[176,177]. Having identified potential candidates I then narrowed down this list by exploring which of them were supported by the respective HyPrColoc result. Is the drug target biologically relevant to the potentially shared molecular pathway revealed by the colocalization? I used the NCBI dbSNP database[193] to gain information on genes associated with the 397 unique SNPs involved in the colocalizations, and found that 275 of the SNPs had listed gene associations. As genes in the same functional pathway might be druggable to achieve a similar effect, I

also obtained a wider functional genetic network for these association genes from the STRING database where available[194].

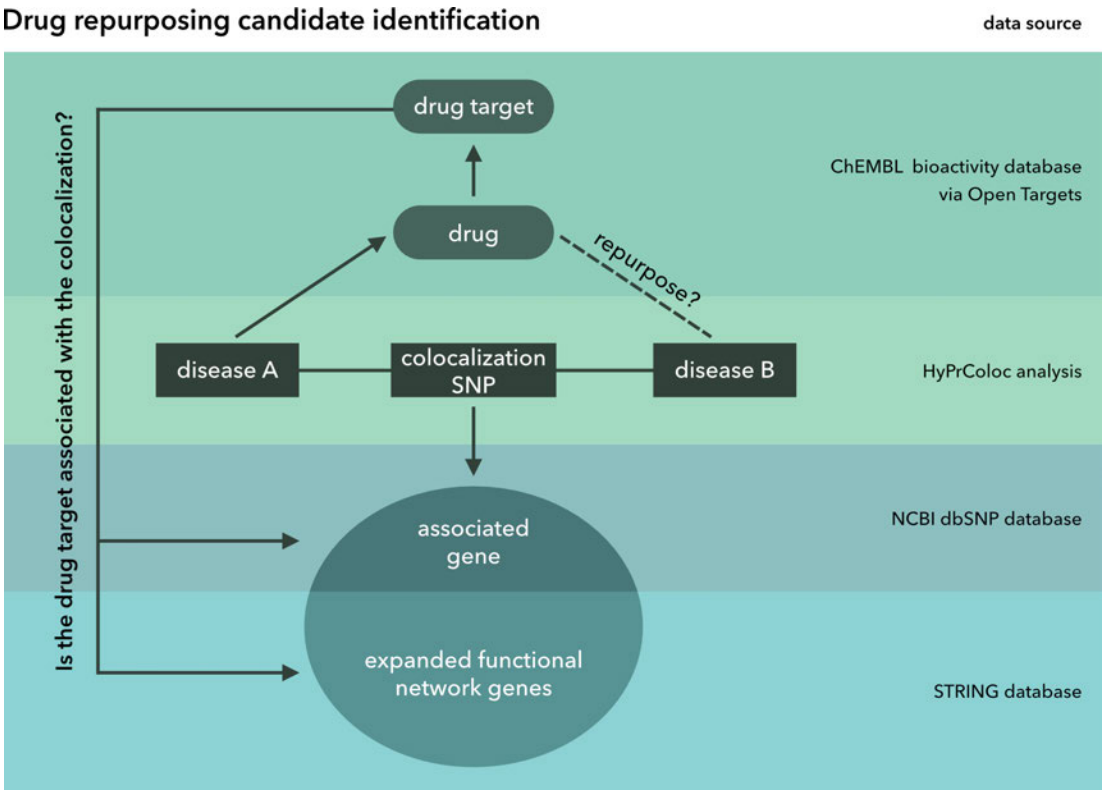


Figure 20: Drug repurposing candidate identification. Using the results of my HyPrColoc analysis, I sought to prioritise candidates for drug repurposing. Where two diseases are colocalizing I identified drugs that are only currently used or in trial for one of them. I then tested whether the drug target was associated with the colocalization. To do so, I identified genes that had an association with the colocalization SNP and checked whether their protein products were targeted by the drug. Additionally, genes in the same functional pathway might be druggable to achieve a similar effect, so I also obtained a wider functional genetic network and checked if the drug target was among them. Drug information was obtained from the ChEMBL bioactivity database via Open Targets, genes associated with the colocalization SNPs were identified using the NCBI dbSNP database and the wider functional network for those association genes were obtained from the STRING database.

Using this method, I identified 3625 drug repurposing candidates with drug targets falling in their respective functional colocalization gene networks. The number of suggested repurposing drugs varied greatly depending on the colocalizing disease, with the highest reported numbers for psoriasis, allergic disease, systemic lupus erythematosus, diabetes and a variety of cardiovascular diseases, including ischaemic heart disease and myocardial infarction (Figure 21). Similarly, the number of repurposing candidates varied strongly by colocalization SNP, with the top 5 accounting for 67% of drug repurposing suggestions (Figure 22).

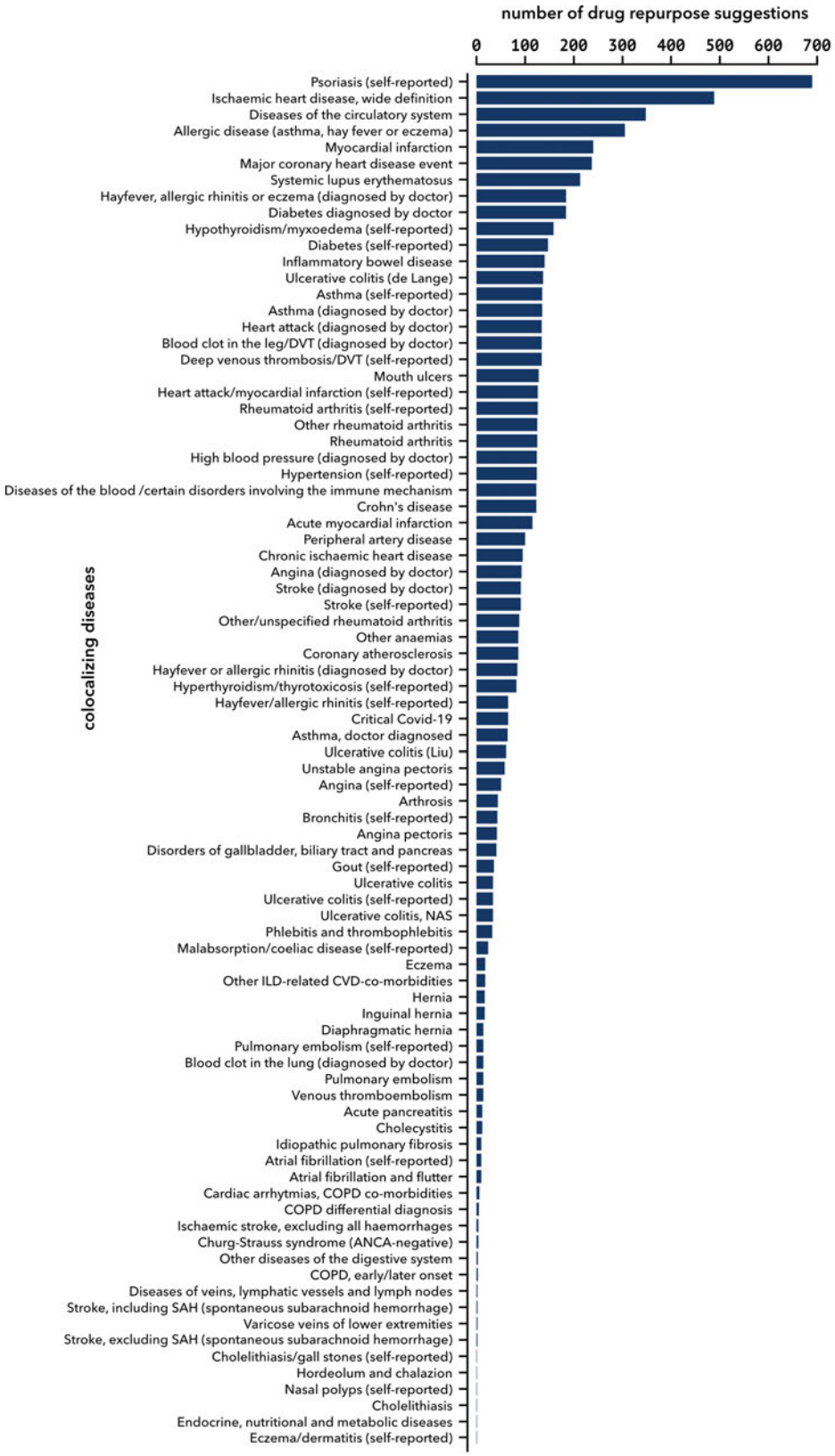


Figure 21: Number of drug repurposing suggestions per colocalizing diseases.

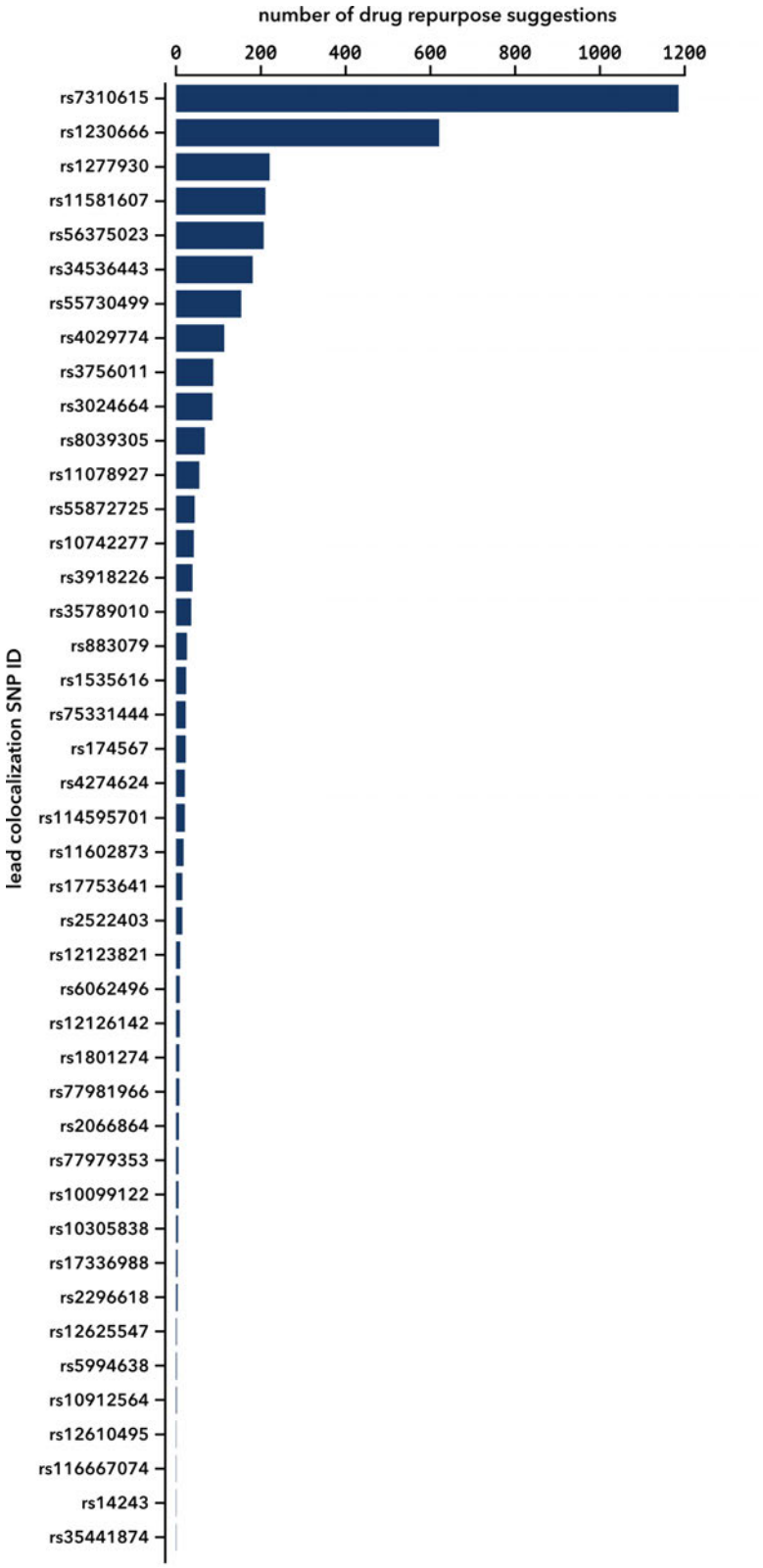


Figure 22: Number of drug repurposing suggestions per lead colocalization SNP.

In order to focus on the drug repurposing candidates which were most strongly supported by the data, I applied the following selection criteria for closer inspection.

Firstly, I mainly focused on drugs that were currently in or have passed phase 4 clinical trials, and I disregarded drug suggestions where another drug with the same target was already in use or under investigation for repurposing. I disregarded instances where a drug was suggested to be repurposed between two closely related diseases - such as from one cardiovascular disease to another - as such repurposing candidates were more likely to already have been considered. Finally, I disregarded repurposing suggestions between wide-reaching phenotypes and specific diseases included within them, such as allergic disease and eczema.

The colocalization with lead SNP rs7310615 resulted in the highest number of potential drug repurposing suggestions. The associated gene for this colocalization is *SH2B3*, a regulator of cytokine signalling and hematopoiesis[375], though none of the suggested drugs target it directly. In its functional network, targeting the erythropoietin receptor (EpoR), its agonist epoetin beta was suggested to be repurposed from myocardial infarction (through seven GWAS) to psoriasis, systemic lupus erythematosus and hypothyroidism/myxoedema, from myocardial infarction (through six GWAS) to allergic disease (through three GWAS), inflammatory bowel disease and ulcerative colitis, from myocardial infarction (through five GWAS) to diabetes and from chronic ischaemic heart disease (ischaemic cardiomyopathy) to psoriasis, systemic lupus erythematosus, diabetes and hypothyroidism/myxoedema. These diseases all share the same direction of effect at rs7310615, except for the three allergic disease GWAS (Figure 23 a).

The intracellular signalling pathway kinases JAK1 and JAK2 are among the drug targets with the highest number of potential repurposing candidates, though none of the suggested drugs have passed phase 4 trials for the original target disease. Multiple of these suggestions are based on the colocalization with lead SNP rs7310615 (associated gene *SH2B3*) as well, but in this case the suggested repurposing direction is from immune system related diseases to cardiovascular diseases. Drugs inhibiting one or both Janus kinases were suggested to be repurposed from allergic disease (atopic eczema – ruxolitinib, abrocitinib, brepocitinib, tofacitinib, baricitinib), hayfever, allergic rhinitis or eczema (eczema – ruxolitinib, delgocitinib), psoriasis (itacitinib, solcitinib, baricitinib, ruxolitinib, ruxolitinib phosphate), systemic lupus erythematosus (baricitinib, brepocitinib, filgotinib), ulcerative colitis (peficitinib, filgotinib, Crohn's disease (filgotinib) and diabetes (baricitinib) to diseases of the circulatory system, myocardial infarction (through four GWAS), angina, deep venous thrombosis/DVT (through two GWAS), major coronary heart disease event, stroke (through two GWAS) and

hypothyroidism/myxoedema. The drugs for psoriasis and systemic lupus erythematosus were additionally also suggested as repurposing candidates for coronary atherosclerosis, hypertension (through two GWAS), chronic ischaemic heart disease and ischaemic heart disease, wide definition. These diseases all share the same direction of effect at rs7310615, with the exception of the two eczema (allergic disease) GWAS (Figure 23 a).

Several drugs targeting the tyrosine-protein kinase TYK2 were suggested as repurposing candidates. Most of the suggested target diseases were either already treated with or currently in trials for other drugs targeting TYK2, with the exception of hypothyroidism/myxoedema. Janus kinase inhibitor tofacitinib was suggested for repurposing from rheumatoid arthritis (through three GWAS) to hypothyroidism. Additionally multiple drugs currently in development were proposed for the treatment of hypothyroidism, including upadacitinib from systemic lupus erythematosus, deucravacitinib from systemic lupus erythematosus, psoriasis and rheumatoid arthritis, brepocitinib from psoriasis and rheumatoid arthritis and tofacitinib additionally also from Covid-19 and psoriasis. Notably, repurposing of upadacitinib and deucravacitinib from systemic lupus erythematosus were supported through two different colocalizations: with lead SNP rs34536443, associated directly with *TYK2* and with lead SNP rs4274624, associated with *STAT4*. Hypothyroidism, systemic lupus erythematosus, rheumatoid arthritis and psoriasis have the same direction of effect at rs34536443, while Covid-19 has an opposing effect direction (chapter 4 Figure 11). Hypothyroidism and systemic lupus erythematosus also have the same direction of effect at rs4274624 (Figure 23 d).

Several vitamin D receptor agonists were suggested repurposing candidates based on colocalizations with lead SNP rs56375023, associated with the gene *SMAD3*, which encodes a transcription factor regulating cell proliferation[376]. Alfacalcidol, calcifediol, calcitriol and ergocalciferol were noted as candidates for repurposing from asthma (through three GWAS) to coronary atherosclerosis and ischaemic heart disease (through two GWAS). Coronary atherosclerosis and ischaemic heart disease were also suggested as repurposing targets for ergocalciferol and cholecalciferol from Crohn's disease (through two GWAS) and seasonal allergic rhinitis (through four GWAS) respectively. Asthma, Crohn's disease and seasonal allergic rhinitis have opposing effect size directions compared to the cardiovascular diseases at SNP rs56375023 (Figure 23 b).

Multiple drug repurposing candidates were suggested based on colocalizations with lead SNP rs1230666 (associated with the gene *MAGI3*). Among them were drugs that target

adrenergic receptors beta 1 and beta 2 (encoded by the genes *ADRB1* and *ADRB2*), which mediate the physiological effects of the hormone epinephrine and the neurotransmitter norepinephrine[377,378]. Repurposing to rheumatoid arthritis (through four GWAS) was suggested for a variety of drugs treating ischaemic heart disease, including the beta-1 adrenergic receptor antagonists esmolol and metoprolol and adrenergic receptor agonist epinephrine (also targeting ADRB2) used for acute coronary syndrome, beta-1 adrenergic receptor antagonists metoprolol tartrate, bisoprolol, metoprolol succinate, nadolol and atenolol as well as the beta-2 adrenergic receptor antagonists carvedilol and propranolol hydrochloride used for angina pectoris and the beta-1 adrenergic receptor antagonists atenolol, propranolol and propranolol hydrochloride (also targeting ADRB2) used in treatment of coronary atherosclerosis. Beta blockers metoprolol and propranolol (also targeting ADRB2), used to treat Graves ophthalmopathy (hyperthyroidism) and hyperthyroidism/thyrotoxicos respectively, as well as the myocardial infarction drugs blockers nadolol and metoprolol tartrate and the beta-1 adrenergic receptor agonist dopamine hydrochlorid were suggested for repurposing to rheumatoid arthritis (through four GWAS) and Crohn's disease. Beta blocker atenolol was additional suggested to be repurposed from diabetes (through 2 GWAS) to rheumatoid arthritis (through four GWAS), Crohn's disease and systemic lupus erythematosus. These diseases have the same direction of effect at SNP rs1230666, with the exception of Crohn's diseases (Figure 23 c).

Several drugs targeting ATP-Sensitive inward rectifier potassium channel 11 (gene *KCNJ11*) were suggested as potential repurposing candidates. Diazoxide, a potassium channel opener, was suggested for repurposing from hypertension (through two GWAS) to arthrosis based on colocalization with lead SNP rs55872725, associated with the gene *FTO*, which encodes for an RNA demethylase regulating energy homeostasis[379]. Based on colocalization with same lead SNP, potassium channel blockers glipizide, glyburide, nateglinide and tolazamide were suggested for repurposing from diabetes (through two GWAS) to arthrosis. These diseases have the same direction of effect at rs55872725 (Figure 24 b). Glipizide, glyburide, nateglinide and tolazamide were also sugested to be repurposed from diabetes to heart attack (through two GWAS) based on colocalization with lead SNP rs77981966 (associated gene *THADA*). Diabetes and heart attack have opposite directions of effect at rs77981966 (Figure 24 d).

There were multiple drug suggestions due to colocalization with lead SNP rs35789010, associated with the gene *CARMIL1*, which encodes for a cytoskeleton formation protein[380]. Probenecid – an inhibitor of solute carrier family 22 member 11 (SLC22A11,

a transporter of conjugated steroids[381]) – and benzbromarone, lesinurad, sulfinpyrazone, which inhibit solute carrier family 22 member 12 (SLC22A12, a urate transporter[382]), were suggested for repurposing from gout to asthma (through two GWAS), systemic lupus erythematosus, malabsorption/coeliac disease, hyperthyroidism/thyrotoxicosis and hypothyroidism/myxoedema. All of these diseases have the same direction of effect at SNP rs35789010 (Figure 24 a).

Several repurposing suggestions were based on colocalizations with lead SNP rs11581607 (the associated gene *IL23R* encodes the interleukin-23 cytokine receptor). The drugs are inhibitors for the proinflammatory cytokine interleukin-17A (encoded by gene *IL17A*)[383]. The repurposing candidates are ixekizumab and secukinumab from psoriasis (psoriasis vulgaris) to ulcerative colitis (through five GWAS), with ixekizumab additionally also being suggested for repurposing to Crohn's disease and inflammatory bowel disease. All of these diseases have the same direction of effect at SNP rs11581607 (Figure 24 c).

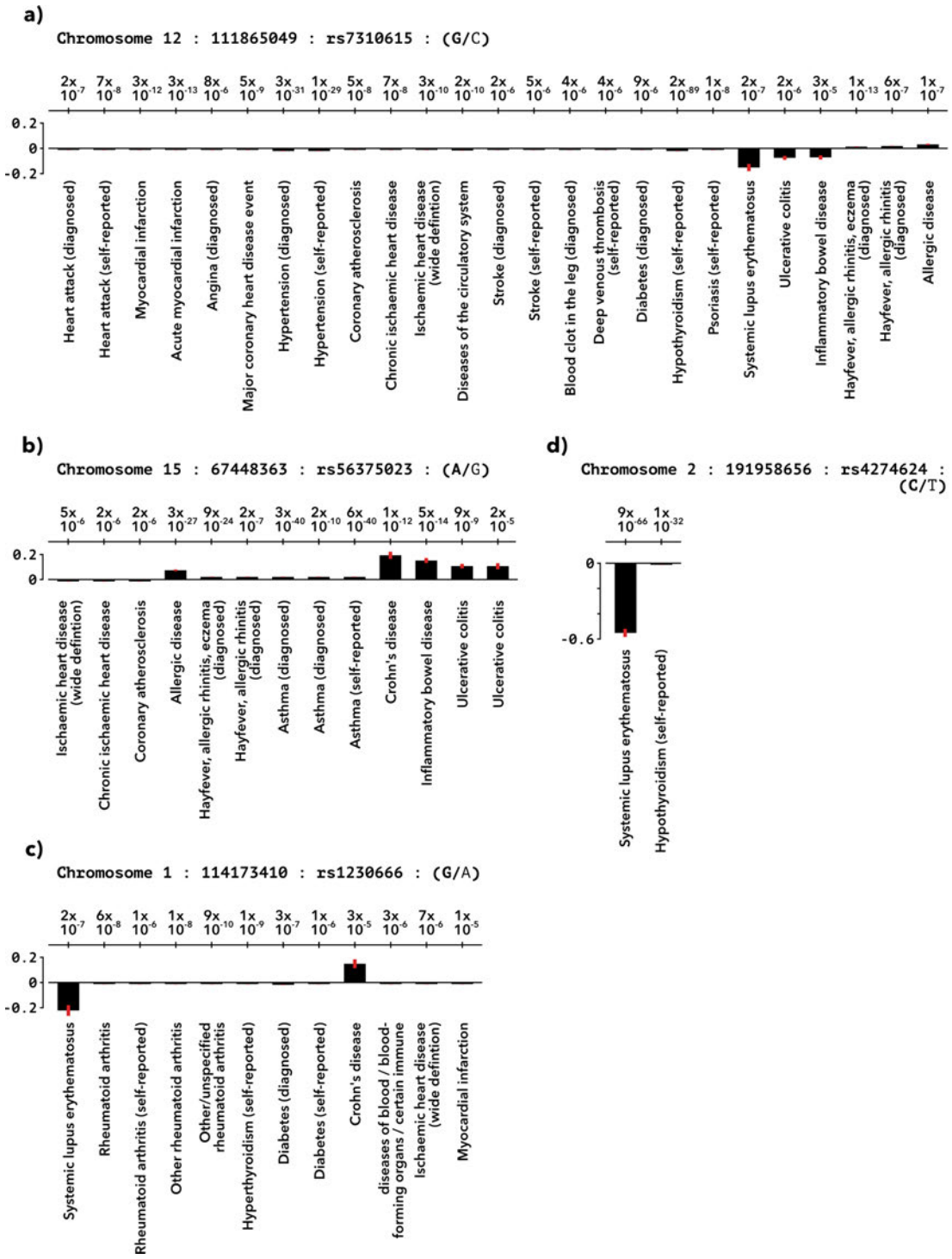


Figure 23: Effect sizes at the lead colocalization SNPs. For each colocalization group, a bar chart compares effect sizes for the variant in the original GWAS, with the accompanying P value displayed above each bar. The reference allele for the effect size is listed in bold, the other allele in regular font (reference allele/other allele).

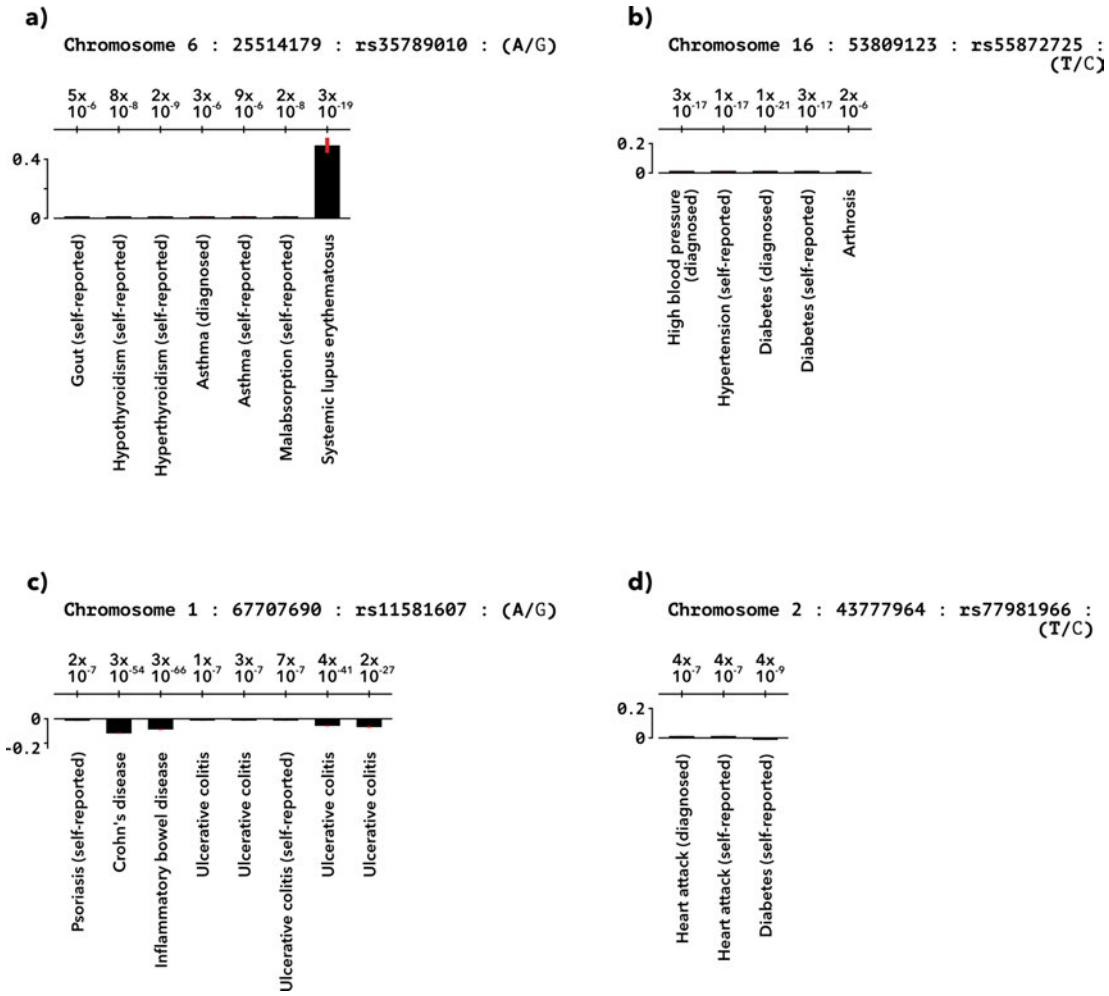


Figure 24: Effect sizes at the lead colocalization SNPs. For each colocalization group, a bar chart compares effect sizes for the variant in the original GWAS, with the accompanying P value displayed above each bar. The reference allele for the effect size is listed in bold, the other allele in regular font (reference allele/other allele).

6.3 Discussion

In this chapter I have shown that a genome-wide colocalization analysis of a large number of diseases reveals a tightly packed disease network, with many localised connections between the different phenotypes. These connections point to a potential etiological overlap in disease biology. Due to the large volume of colocalizations, I decided to prioritise colocalizations by their potential to reveal drug repurposing opportunities. This candidate screening study revealed several possible avenues for further investigation.

Colocalization analysis is a powerful technique for revealing the pleiotropic effects of one genetic variant on multiple traits. Due to the large number of investigated colocalizations, I performed a simplified version of the HyPrColoc analysis, reporting results for traits above a posterior colocalization probability of 0.7 when using conditional colocalization prior 0.02. This was found to generally work well to minimise false positives and identify true positives in the publication describing the technique[24]. By not performing an additional analysis round with conditional colocalization prior 0.01 to identify potential areas biased by multiple associated variants, this risks the inclusion of some false positive results. As the analysis was mainly aimed at exploratory discovery, this liability was chosen over the potential risk of the exclusion of true positives in a more stringent analysis setting without the resources for a quality control step to rescue such true positives within multi-association regions.

In chapter 3, I explored genetic correlations between the disease GWAS. There were 9018 genetic correlations between diseases, while in this chapter 1045 disease pairs were shown to be connected through local colocalization. Out of these, only 192 disease connections were present in both sets of results. Connections that were identified with both methods included both strong and weak correlations as well as high and low number of colocalizations between the diseases. Diseases that did not achieve overall genetic similarity were found to share associated variants that may provide insight into the pleiotropic effects connecting them on a biological level. By analysing similarity on a local level, large scale colocalization unveils connections that may be missed when employing more widely used correlation techniques.

The identified lead colocalization SNPs led to 3625 drug repurposing candidates. The number of suggested repurposing candidates varied greatly by lead colocalization SNP, with the top 5 accounting for 67%. It is likely that this in large part reflects the number of diseases that are colocalizing. The SNP rs7310615 had the largest number of drug candidate suggestions (1186) as well as the largest number of colocalizing traits (25). Given that most colocalizations were between pairs of diseases, the other top ranking lead SNPs showed a similar trend at 14 diseases and 621 drug candidates, 12 diseases and 221 drug candidates, 8 diseases and 211 drug candidates and 14 diseases and 207 drug candidates.

Multiple drug repurposing candidates were based on a colocalization with lead SNP rs7310615, associated with the gene *SH2B3*, which encodes for Lnk, a regulator of cytokine signalling and hematopoiesis[375]. Among these candidates were several drugs

targeting the erythropoietin receptor (EpoR) and the intracellular signalling kinases JAK1 and JAK2. EpoR and JAK2 are part of a common signalling pathway; EpoR activates JAK2 and thereby its downstream signalling cascades, regulating erythropoiesis and tissue protection, as well as immunomodulatory activity[384,385]. Lnk is part of the same pathway. It phosphorylates EpoR, and thereby inhibits erythropoiesis and downstream JAK2 signalling pathways[386,387]. A connection between Lnk and JAK1 is less evident; their STRING network association evidence score was much lower than the one between Lnk and JAK2 (0.523 and 0.989), and largely based on shared disease associations, which on their own are not clear evidence of a shared functional pathway. Biological evidence does not support a shared pathway between Lnk and JAK1[388]. It therefore seems likely that drugs targeting JAK1 but not JAK2 (abrocitinib, brepocitinib, itacitinib, solcitinib and filgotinib) are not truly supported by the colocalization evidence and should not be considered.

EpoR agonists were suggested as repurposing candidates from cardiovascular disease to immune system disease and diabetes, while JAK2 inhibitors were suggested as repurposing candidates from immune system disease and diabetes to cardiovascular disease. As EpoR agonists and JAK2 inhibitors have opposing effects, it may seem unlikely that this would be beneficial. However, the situation is more complex in biological practice. JAK inhibitors have been noted to reduce cardiovascular events by curbing inflammation, although they also have possible prothrombotic effects[389,390]. Further research is needed to establish if the potential therapeutic benefits could outweigh the risks. Additional well-powered randomized control trials could be used to distinguish whether there are subgroups of patients (for example depending on their age or comorbidities) that could benefit from the treatment without sustaining a substantially heightened risk. There is an indication that some JAK inhibitors may pose a higher risk of adverse cardiovascular effects than others[391], though it has also been noted that adverse effects may be dosage-dependant[389] – additional data could clarify these effects. Another avenue of interest would be longer term follow up studies focusing on the cardiovascular health in patients currently receiving JAK inhibitor treatment for other conditions.

It has been shown that erythropoietin (EPO) plays a role in tissue protection and immune regulation by binding a heterodimeric receptor with an EpoR and CD131 subunit (also named cytokine β common subunit), called the tissue-protective receptor[385]. Through this pathway, EPO protects cells surrounding an injury from apoptosis[392,393], recruits endothelial progenitor cells[394,395] and suppresses recruitment of

inflammatory cells[396]. The tissue-protective receptor has a lower affinity to EPO than EpoR does[397], therefore a much higher dose of EPO is required to activate its effects. However, high doses of EPO increase stroke risk via its erythropioetic pathways[398]. To avoid the hematopoietic and procoagulant effects of EPO while utilising its immunomodulatory function, modified versions (carbamylated EPO and ARA290) were created to selectively bind only to the tissue-protective receptor[398 400]. Erythropoietin and its modified derivatives have been suggested for treatment of systemic lupus erythematosus[401]. Animal models have shown the potential benefits of EPO treatment in autoimmune disease: EPO and ARA290 reduced inflammation in mouse models of systemic lupus erythematosus[402,403] and EPO reduced disease severity in a mouse model of autoimmune colitis[404] and reduced colonic inflammation while increasing survival rate in another[405]. ARA290 improved recovery in a rat model of autoimmune neuritis[406], while EPO reduced inflammation in a rat model of autoimmune encephalomyelitis[407]. EPO has been suggested as a treatment option for diabetes mellitus, due to its potential for treating multiple of its complications[408] and promising results from animal and cellular studies, including reduced blood glucose levels[409], normalised inflammatory response[410] and tissue protective effects[411]. Further investigation of modified EPO for the suggested target diseases systemic lupus erythematosus, inflammatory bowel disease, diabetes and potentially psoriasis could open new avenues of treatment.

Hypothyroidism was listed as a repurposing option for both activation and inhibition of the EpoR-JAK2 pathway, suggesting a link to the disease. Anaemia is a common comorbidity in thyroid disease[412]. Thyroid hormones induce erythropoietin expression[413] and hypothyroid patients have shown reduced EPO plasma levels[414]. Patients with hypothyroidism that were undergoing hemodialysis were found to require a higher dosage of erythropoietin[415]. This makes it seem likely that the impact on EpoR-JAK2 is a downstream effect of hypothyroid pathogenesis, and not a good target for treatment. However, there is additional evidence supporting JAK2 as a treatment target in hypothyroidism through its involvement in cytokine pathways.

A range of TYK2 inhibitors, as well as Janus kinase inhibitors targeting TYK2, JAK1, JAK2 and JAK3, were suggested for repurposing to hypothyroidism. Two of the candidate drugs, upadacitinib and deucravacitinib, were supported through multiple colocalizations; one in direct association with *TYK2*, and the other in association with *STAT4* through the functional gene network. STAT4 is a transcription factor, which is activated downstream of a variety of cytokine receptors including IL-12 and IL-23[416].

These receptors activate TYK2 and JAK2 which in turn leads to STAT4 phosphorylation and activation[417]. These pathways are heavily implicated in immune system diseases, many of which are currently treated with or investigated for TYK2 inhibitors[418,419]. TYK2 inhibitors have also recently been suggested as a potential treatment for hypothyroidism based on a Mendelian randomization study[420]. In geographical regions with iodine sufficiency, the most common cause of hypothyroidism is Hashimoto's thyroiditis, an autoimmune disease[421]. TYK2 inhibitors may therefore prove a promising avenue of exploration in the treatment of autoimmune hypothyroidism.

Several vitamin D receptor agonists were suggested repurposing candidates from asthma, Crohn's disease and seasonal allergic rhinitis to coronary atherosclerosis and ischaemic heart disease. The vitamin D receptor is a nuclear hormone receptor, which translocates into the nucleus upon binding vitamin D, where it functions as a transcription factor influencing many processes such as calcium homeostasis, cell proliferation, and cell differentiation[422]. The repurposing suggestions are based on a colocalization associated with SMAD3, a transcription factor regulating cell proliferation[376]. SMAD3 can bind to the *VDR* promoter and VDR can bind to the *SMAD3* promoter[423] and may be able to co-occupy regulatory segments at the same time[424]. Multiple effects of the two transcription factors on each other have been proposed – among them are reciprocal activation, where activity of one leads to the expression of the other[423], or genomic antagonism, where SMAD3 binding alters the local genomic landscape in a way that facilitates VDR binding, which in turn reduces SMAD3 recruitment[424]. From a molecular point of view, there is some evidence to support that VDR could be a drug target supported through colocalization at a SMAD3 associated variant, but the exact functional connection as well as the direction of effect between the two remains unclear. However, a pilot trial testing the effects of vitamin D repletion in patients with established coronary artery disease failed to demonstrate any benefits on inflammation or surrogate markers of cardiovascular health[425]. Additionally, a large meta analysis found that vitamin D supplementation did not reduce the risk of major adverse cardiovascular events, myocardial infarction, stroke or cardiovascular disease mortality[426]. It is likely that this repurposing suggestion was only proposed by the analysis because these failed trials were only listed for closely related phenotypes. Vitamin D supplements are therefore not supported as an effective treatment for cardiovascular disease.

Multiple drugs that target adrenergic receptors beta-1 and beta-2 were suggested for repurposing from cardiovascular disease and hyperthyroidism to rheumatoid arthritis, Crohn's disease and systemic lupus erythematosus. These G-protein-coupled receptors

bind epinephrine and norepinephrine and mediate their physiological effects; beta-1-adrenergic receptor binds both molecules with approximately the same affinity[377], while beta-2-adrenergic receptor binds epinephrine with an 30-fold greater affinity[378]. The gene associated with the colocalization these repurposing suggestions are based on is *MAGI3*, which encodes for a membrane-associated guanylate kinase that regulates multiple signalling processes including downregulation of cell proliferation and migration[427,428]. The molecular mechanisms of beta blockers are well characterised and currently not thought to involve MAGI3[429]. However, there is a possibility that MAGI3 could still be relevant as it was found to potentially physically interact with the beta-2-adrenergic receptor to mediate extracellular signal-regulated kinase (ERK 1/2) activation[430]. Similarly, beta-1-adrenergic receptor has been shown to interact with MAGI3, impacting the ERK 1/2 pathway as well[431].

While beta blockers could be potentially reasonable repurposing options based on the molecular evidence, clinical data reveals that their use in this case is inadvisable. The concomitant use of beta blockers reduced remission rates in patients with rheumatoid arthritis[432] and is associated with an increased relapse risk in patients with inflammatory bowel disease[433]. Furthermore, beta blockers have been observed to cause drug induced lupus erythematosus[434,435,435]. While beta blockers had an effect on the suggested target diseases, in clinical practice they were found to worsen those diseases and should not be considered as repurposing options.

Multiple drugs targeting potassium inwardly rectifying channel subfamily J member 11 (KCNJ11) were suggested as potential repurposing candidates from hypertension and diabetes to arthrosis, as well as from diabetes to heart attack. KCNJ11 is a subunit of an ATP-sensitive potassium channel that transports potassium into cells[436,437]. The colocalization the repurposing suggestion is based on is associated with *FTO* for arthrosis and *THADA* for heart attack. FTO is an RNA demethylase that regulates fat mass, adipogenesis and energy homeostasis[379,438]. THADA has been associated with thermogenesis and obesity in Drosophila[439] and with metabolic function through its impact on β-cells in mice[440]. STRING association between both colocalization genes and the drug target gene was based purely on shared publications, as all three are associated with diabetes – KCNJ11 through its involvement in insulin signalling and THADA and FTO through interaction with environmental factors[441]. A potential functional connection between the pairs as part of one druggable pathway is therefore tenuous.

Neither arthrosis nor heart attack are good targets for the suggested drug candidates. It has been shown that FTO affects osteoarthritis via vertical pleiotropy, mediated through its impact on body weight[442]. Drugs should therefore not be based on this association. Several drugs targeting KCNJ11 have been in trials for cardiovascular diseases beyond its current use in acute hypertension; for example, the blocker glyburide showed positive results in an early phase trial for ischaemic stroke[443]. Several other potassium channel blockers showed no clinical benefit in phase 3 or 4 trials for multiple cardiovascular diseases, including glyburide for subarachnoid hemorrhage[444], glimepiride for cardiovascular events or coronary artery disease[445] and glipizide for atherosclerosis[446]. Given these results trials for treatment of heart attack may be inadvisable.

Solute carrier family 22 member 11 (SLC22A11, also called OAT4) inhibitor probenecid and solute carrier family 22 member 12 (SLC22A12, also called URAT1) inhibitors benzbromarone, lesinurad and sulfinpyrazone were suggested for repurposing from gout to asthma, systemic lupus erythematosus, malabsorption/coeliac disease, hyperthyroidism and hypothyroidism. SLC22A11 is a transporter of conjugated steroids involved in renal reabsorption of urate and derived steroid sulfates[381,447]. SLC22A12 transports urate, helping to maintain blood levels of uric acid through its renal reabsorption[382]. Expression of both drug target genes is highly tissue specific, with SLC22A11 being enriched in the kidney, epididymis and placenta and SLC22A12 expression restricted to the kidney[246,247]. The repurposing suggestions are based on a colocalization associated with the gene *CARMIL1*, which encodes for LRRC16A, a cytoskeleton regulation protein that has been proposed to be involved in the reabsorption of urate via urate transportome formation[380,448]. The functional network association between these genes was based entirely on their shared mention in publications due to all three genes having been associated with hyperuricemia and gout[449].

Probenecid, enzbromarone, lesinurad and sulfinpyrazone are all used to treat hyperuricemia in gout. Hyperuricemia is also a factor that connects the suggested target diseases in several ways. Patients with systemic lupus erythematosus frequently develop hyperuricaemia[450], likely due to nephritis and treatment with diuretics[451], however they rarely develop gout[450,452]. Reports on whether thyroid disorders increase risk of developing gout have been conflicting, with some studies finding a link between both hypo- and hyperthyroidism and hyperuricaemia or gout[453,454], while others report no increase of risk[455]. Regardless, thyroid hormones have an impact on renal function, affecting many aspects such as the glomerular filtration rate, renal hemodynamics,

sodium and water homeostasis and nephron transport systems[456]. Patients with celiac disease have been found to have heightened uric acid levels in plasma, potentially working as an antioxidant response to enhanced oxidative stress[457]. Gout has also been reported as a risk factor for allergic asthma[458]. While the colocalization appears to be linking these diseases through kidney function, it seems unlikely that this is an actual connection of shared aetiology. The lupus-related renal damage and thyroid hormone-impacted renal function may be further exacerbated by genetic variants that are causal for gout, but the underlying pathology works through functionally distinct mechanisms. In celiac disease, decreasing uric acid levels may be detrimental to oxidative stress management. The genetic variant may be associated with asthma through vertical pleiotropy, affecting development of gout which then in turn heightens risk of allergic asthma. It therefore seems unlikely that drug repurposing would be beneficial in this case.

Proinflammatory cytokine interleukin-17A (IL-17) inhibitors ixekizumab and secukinumab were suggested for repurposing from psoriasis to inflammatory bowel disease (IBD). On paper, this seems like a great repurposing opportunity; IL-17 is produced by activated T-cells and induces neutrophil mobilization and activation[383], and was in the functional network of the actual colocalization-associated gene *IL23R*, which encodes for the interleukin-23 (IL-23) cytokine receptor. Targeting IL-17 instead of IL-23 seems promising, as IL-23 induces the differentiation of naive T cells into highly pathogenic helper T cells that produce IL-17, functionally working upstream[459]. Patients with IBD have increased IL-17 expression in their colonic mucosa[460] and genetic studies have associated the IL-23 receptor and its pathway with IBD[461,462]. Ustekinumab, an antibody that blocks IL-23, has been approved for treatment of Crohn's disease[463]. Studies in animal models reported conflicting results, with some showing that blocking either IL-23 or IL-17 both had significantly protective effects in both IBD[464] and multiple sclerosis[465], while others reported blocking IL-17 had a worsening effect on colitis[466].

Notably, in my drug repurposing results secukinumab was suggested for repurposing only to ulcerative colitis while ixekizumab was suggested for both ulcerative colitis and Crohn's disease. The wealth of supporting evidence detailed above already led to a clinical trial for secukinumab as a treatment for Crohn's disease[467]. Paradoxically, the trial had to be terminated as patients were suffering from worsening of disease symptoms. Another drug targeting IL-17 via its receptor, brodalumab, was trialed for Crohn's disease but also ended up worsening the disease[468]. Additionally, new-onset IBD has

been described in patients receiving secukinumab or ixekizumab treatment[469]. It is therefore clear that IL-17 inhibitors should not be repurposed for IBD. It has been suggested that the surprising opposite effect of IL-23 and IL-17 neutralization in colitis may be because blocking IL-23 reduces helper T cell autoimmunity, with several other cell lines still producing IL-17, while blocking IL-17 impairs intestinal homeostasis and tissue repair[470]. Overall, although this drug repurposing candidate did not succeed in clinical practice, it shows the colocalization-based repurposing approach can pinpoint opportunities in line with other biological data, and underlines the difficulties of accurately predicting clinical reality, especially for such complex processes as the human immune system.

Though many of the repurposing suggestions I have explored in greater depth are unlikely to be good candidates, these only scratch the surface of a rich dataset. The analysis resulted in many more repurposing proposals than possible to explore within the scope of this thesis. Furthermore, several lead colocalization SNPs did not have gene associations in the dbSNP database, or had associations with long non-coding RNA genes, which in turn did not have associated functional genetic networks. Alternate ways of linking the colocalization variants to genes likely would have uncovered further associated genes to explore, adding even more connections. I mapped the lead colocalization SNPs to associated genes using the dbSNP database, which lists the nearest genes for variants. Although variants do not necessarily affect the nearest gene, multiple studies have shown that in many cases the nearest gene is the most likely candidate[471 473]. The dbSNP database does not assign a gene to every SNP, thereby missing many nearest genes, which could be avoided by using the Ensembl genome database to identify nearest genes[226]. However, dbSNP only assigns genes if the variant is within their coding region or the SNP is otherwise known to affect the gene, providing an opportunity to avoid nearest genes that are less likely to be affected. I obtained the SNP-gene associations from a database because of speed and ease of use. In the future, a much more labour intensive but more in depth combination of eQTL-GWAS colocalization analysis (as performed in the previous chapter) and Mendelian randomization analysis could offer a more accurate means of identifying relevant genes, as well as pinpoint the tissues they are affected in[67].

Pathway-based drug repurposing analyses offer a wider lens to find opportunities for medical interventions. A limitation to such approaches is that they tend to prioritise known biology[147,474]. Using the functional association data provided by the STRING database may offer a more flexible approach, as many different sources of information

are aggregated to form the association scores. A downside to this less rigid approach is that it makes follow up investigation and validation of the connections of interest an even greater necessity.

GWAS effect direction at the lead colocalization SNP seemed disconnected from whether two diseases benefited from having a pathway modulated in the same direction, with some diseases already sharing a drug having opposing effect directions. A variety of factors make predicting the necessary direction of modulation challenging. Variants may influence genes through multiple means, either by directly affecting protein folding, function and truncation or through an impact on transcription of the gene, leading to more or less of the protein product than necessary for healthy function. The exact influence is difficult to anticipate without follow up *in silico* or *in vitro* studies. Diseases may be affected in different cell types, requiring different directions of modulation based on the affected tissue. Additionally, the lead colocalization SNP is likely not the only variant affecting the disease, and not necessarily the only SNP affecting the disease through the identified pathway or gene. If multiple independent SNPs are affecting a gene, they may not all have the same direction of effect. This becomes even more difficult to parse when the drug of potential interest affects the product of another gene in the functional network. In combination, these factors result in a disconnect between GWAS effect at the shared lead colocalization SNP and needed drug modulation direction. Incorporating expression data into the repurposing analysis pipeline could help shed light onto these intricacies, offering information on affected tissues and direction of effect.

Missed opportunities for drug discovery might result from false negative results in GWAS, brought on by the necessary stringent multiple testing corrections[39]. In an effort to cast a wider net, some methods try to circumvent this issue by looking at all nominally significant associations[156,475]. Although mitigated by the integration of additional biological data, this risks an increase of false positive results. Widening the number of investigated associations by improving detection power through large-scale colocalization data instead could offer similar benefits, while retaining quality control measures that weed out false positive results. Additionally, drug repurposing studies that include genetic data tend to identify disease connections based on shared affected genes[147]. Matching GWAS signals to affected genes is a difficult, error-prone process. Studying genetic links between diseases at the level of the variant can reveal connections that otherwise might stay hidden due to inaccurate assumptions about which genes are affected.

As we have seen in this chapter, colocalization analysis is a powerful tool for studying the genetic relationship between diseases. Its ability to widen the investigation to genetic variants that would normally fall below detection threshold within single GWAS enhances the ability to select possible drug repurposing candidates. Colocalization analysis could therefore be a fruitful addition to existing bioinformatic drug repurposing pipelines combining a large number of information sources.

While I have identified several drug repurposing candidates that warrant further investigation, it is important to note that although computational approaches can help with candidate prioritisation, studies in animal models and ultimately patient randomized control trials are required to confirm their efficacy in clinical practice.

Chapter 7
Concluding Remarks and Future Directions

My aim was to identify pleiotropic genetic variants shared between diseases in order to pinpoint potential therapeutic targets for critical Covid-19 and find opportunities for drug repurposing. To address this, I explored the similarity between critical Covid-19 and a wide range of diseases on a genome-wide level through genetic correlation (Chapter 3). I then identified putatively causal genetic variants shared between critical Covid-19 and other diseases (Chapter 4) and investigated the effect of those variants on gene expression (Chapter 5). Finally, I pinpointed shared genetic associations between 228 disease GWAS and subselected variants linked to drug targets that could provide repurposing opportunities between the associated diseases (Chapter 6).

7.1 Cardiovascular disease and idiopathic pulmonary fibrosis are genetically correlated with critical Covid-19

In chapter 3, I showed genetic correlations between critical Covid-19 and 56 other diseases. While genetic correlations with Covid-19 had yet to be studied when I started this analysis, they have now been widely reported in the literature[160]. Although the standard errors for my results were large, my findings matched previously published results – in particular with regard to the links between critical Covid-19 and cardiovascular disease and between critical Covid-19 and idiopathic pulmonary fibrosis – lending further credence to these correlations.

Multiple underlying factors can lead to the detection of genetic correlations. They are often caused by a combination of both horizontal pleiotropy where the traits share underlying biology, and vertical pleiotropy where a variant or gene affects one trait, and this trait then influences another[206,208]. Either could be interesting: while horizontal pleiotropy is informative in terms of understanding shared pathology between diseases, vertical pleiotropy may be useful in identifying strategies for disease prevention. As

genetic correlation analysis is unable to distinguish the source of pleiotropy, follow up tests are necessary to disentangle this signal. Several studies have already shed more light on the connection between Covid-19 and cardiovascular disease. Both influence each other; prior cardiovascular conditions such as hypertension and diabetes have been shown to be a risk factor for hospitalised Covid-19 and lead to worse outcomes[214,215], and Covid-19 has been noted to exacerbate existing cardiac conditions[216] and may directly injure the cardiovascular system[217]. However, whether this relationship is based on shared underlying disease biology needs further investigation.

I described a genetic variant with a shared association between critical Covid-19 and hypertension, as well as multiple shared variants between critical Covid-19 and idiopathic pulmonary fibrosis in chapter 4, pinpointing potential avenues for future exploration. As laid out in chapter 4 as well as in the next section, the rs13107325 variant associated with both critical Covid-19 and hypertension is located in the gene *SLC39A8* and is known to impact the corresponding protein ion transporter ZIP8. In those sections I describe how NF-kB or interleukin-10 signalling might be the pathways affected by ZIP8 function in critical Covid-19. To test this connection and whether these or other players downstream of ZIP8 affect hypertension as well could be tested in cellular or animal models of the disease. As it is known that the variant in question affects ZIP8's transporter function, this could be modelled by impairing the transporter, for example by blocking the channel or silencing its expression, and testing the effects on disease-relevant cell or organism fitness in models of disease context. The three shared variants associated with critical Covid-19 and hypertension were located in or near *ZKSCAN1*, *ATP11A* and *DPP9* and may impact the diseases through these genes. Additionally, I found evidence that these variants may affect the expression of several genes in a cell or tissue-specific manner, as shown in chapter 5. The expression of *ATP11A* in blood, *PTPRS* in the lung and *SAFB2* in CD4 naïve/central memory T cells was associated with these variants with high confidence and the expression of *TRIM4* and *MUC12-AS1* in blood, *CPSF4* in classic monocytes and *DAPK3* in CD4 effector memory/TEMRA cells was associated with medium confidence. Here as well these genes could be a starting point for exploration of a potential aetiological relation with critical Covid-19 and idiopathic pulmonary fibrosis by disrupting these genes or their associated proteins in cellular models of disease context.

7.2 A novel critical Covid-19 association in *SLC39A8*

Genetic correlation is an average for pleiotropy across the whole genome, but pleiotropy can vary strongly among local regions. Opposing directions of effect at local associations can result in a genetic correlation of zero, masking shared loci that could hold vital information on shared disease biology and potential for drug repurposing. To address this I explored local pleiotropic effects between critical Covid-19 and other diseases in chapter 4. Using genome-wide multi-trait colocalization analysis I identified seven putatively causal variants shared between critical Covid-19 and several other diseases. Among these results was a Covid-19 association with a variant in the gene *SLC39A8* that did not reach detection threshold in the critical Covid-19 GWAS itself. It suggests pathways involving the divalent metal ion transporter ZIP8 as potential therapeutic targets.

The variant in *SLC39A8* was in shared association with critical Covid-19, Crohn's disease, inflammatory bowel disease, allergic disease, medically diagnosed hayfever, allergic rhinitis or eczema, self-reported hayfever, allergic rhinitis, self-reported hypertension, high blood pressure, self-reported osteoarthritis and diabetes. The associated allele change introduces a missense mutation into the gene, which then leads to an amino acid change in the transcribed protein. This results in a switch from alanine to threonine at position 391 in the ZIP8 protein that reduces its transporter function[239,240,242,243].

ZIP8's ion transporter function could suggest zinc or manganese homeostasis as targets of intervention in critical Covid-19. Zinc supplements have been suggested as a potential avenue of treatment in Covid-19 before, but trial outcomes are yet inconclusive[255,256]. Previously tested manganese therapy of two patients with severe ZIP8 deficiency showed a promising clinical improvement[257]. There are no current drugs targeting ZIP8 specifically, and as the malfunction appears to be related to impaired transporter activity – rather than for example upregulation, which could be inhibited – it may be particularly difficult to target and influence directly. Perhaps the most promising approach in determining if the *SLC39A8* variant association could point us towards treatment options is to focus on the pathways downstream of ZIP8. One drug that is already used to treat critical Covid-19 that takes effect this way is dexamethasone, which upregulates the inhibitor I κ B α and thereby suppresses NF-kB signalling[258,259]. A downstream effector of NF- κ B, nuclear factor kappa B kinase subunit beta (IKK β),

has been suggested as a potential target for treatment as well[259]. However, while dexamethasone has been shown to reduce mortality in Covid-19 patients receiving mechanical ventilation or supplemental oxygen[258], it is important to note that the drug impacts inflammation through multiple pathways[260], the synergy of which may be crucial to its beneficial effects.

ZIP8 is the only zinc transporter that has been observed to be greatly upregulated in immune cells upon infection[250,476 478]. This suggests downstream zinc-dependent signalling could be a treatment target in those cells during the viral replication phase of Covid-19. Pyle et al. have described that ZIP8 reduces interleukin-10 (IL-10) expression in human macrophages after induced inflammation[477]. The import of zinc into the cell reduces nuclear localisation of the transcription factor C/EBP β, which in turn downregulates IL-10 expression. IL-10 functions primarily as an anti-inflammatory cytokine. It is possible that an exaggerated increase in IL-10 levels in early stages of Covid-19 impairs proper viral clearance by downregulating the immune response[479]. Though it is unclear whether ZIP8 is involved in mediating IL-10 expression in chronic inflammation, IL-10 certainly is an important factor in the hyperinflammatory phase of severe and critical Covid-19[480]. A massive rise in IL-10 has been reported in severe forms of the disease[481,482]. Higher IL-10 levels are correlated with ICU admission[483] and reduced patient survival[484,485]. Persistent deficiency of IL-10 has previously been described to result in increased and chronic inflammation and tissue damage in inflammatory and autoimmune disease[486 488]. Counterintuitively, it may be possible that prolonged and elevated IL-10 secretion in the later stages of severe or critical Covid-19 is facilitating the production of proinflammatory cytokines and chemokines and thereby worsening systemic inflammation[480,481]. This would also be congruent with the fact that baricitinib, a JAK inhibitor that has directly opposing effects downstream of the JAK1-activating IL-10, has been shown to improve the condition of hospitalised Covid-19 patients[136]. IL-10, the upstream pathways responsible for its expression or its downstream signalling cascades could therefore also make promising therapeutic targets in later phases of critical Covid-19.

The multi-trait colocalization analysis in chapter 4 was performed using the Bayesian algorithm based method HyPrColoc. This method offers multiple benefits, allowing for the study of many traits at once and improving power to detect novel associations through simultaneous analysis of traits that share local genetic similarity. However, the technique also has several limitations, such as its assumption of only one associated variant being present within a region. HyPrColoc tests whether sub-clusters of analysed

traits share an associated variant, and traits are taken out of the regional analysis loop as soon as they are found to colocalize. For GWAS with multiple regional associated variants this means that potential colocalizations at additional variants may be missed. Another major restriction is that only variants present in all datasets can be analysed. A wide coverage of genetic variants is crucial to successful fine-mapping of the associated variant behind a GWAS signal. Multiple independent GWAS of interest had to be excluded from the analysis as they would have drastically lowered the number of included variants. In the future, this could be improved by imputation of variants that were not directly genotyped from population-based reference sequencing data.

7.3 The PD-1/PD-L1 checkpoint in critical Covid-19

Five out of seven pleiotropic variants identified in chapter 4 were located outwith protein coding regions. Non-coding variants are predicted to impact cell function by affecting gene expression levels. In chapter 5, I explored the effect of the identified pleiotropic variants on gene expression through colocalization with tissue-specific expression quantitative trait locus data. This method uncovered 55 genes with altered expression in whole blood, lung, transverse colon or certain cell type subsets of peripheral blood mononuclear cells, 40 of which had previously been connected to Covid-19 in the literature. Due to a lack of statistical power in much of the expression data, I used relaxed colocalization threshold criteria which may have resulted in some false positive results. Further analyses, such as Mendelian randomization, and in vitro confirmation will be necessary to determine which expression changes are truly related to the GWAS signal.

Follow up pathway enrichment analysis of these genes pointed to the cellular checkpoint programmed cell death 1 (PD-1) and its ligand (PD-L1) as a possible drug target. While multiple studies have linked severe and critical Covid-19 with PD-1 dysregulation[340 342, 489,490] its potential etiological role in the disease is far from certain. The direction of causation is unclear – is a higher expression of PD-1/PD-L1 a maladaptive response that leads to a perpetuation of the illness, or a sign of the body trying to stop excessive damage caused by the host immune system? Though it has many downstream consequences, the PD-1-axis is generally viewed as a safety checkpoint against an excessive immune system attack[491]. Its effect could also depend on the stage of the disease; elevated levels of PD-L1 in hospitalised patients[341] could be representative of an attempted tissue-protective response, while observed elevation in the amount of PD-L1

earlier in the disease[340] could hinder the antiviral response by protecting infected cells. However, it is also possible that higher levels of PD-L1 during earlier stages of the disease are compensating for a malfunction in another pathway under a completely different causal mechanism. Sabbatino et al. have argued that treatment with PD-1/PD-L1 inhibitors in early Covid-19 progression could assist in viral clearance, as PD-L1 upregulation could serve as an escape mechanism for infected cells against both innate and adaptive immune responses[340]. The same study suggested that one of the effects of JAK1/JAK2 inhibitors on SARS-CoV-2 infected cells is reduced PD-L1 transcription levels, which could be part of the therapeutic benefit of JAK1/JAK2 inhibitor baricitinib in patients hospitalised with Covid-19[136]. However, PD-1 or PD-L1 inhibitors could also have detrimental effects that potentiate immune system damage[492]. If PD-1 proves to be causal for a more severe progression of the disease, using inhibitors to treat patients that are at high risk of developing critical Covid-19 and show elevated levels of PD-L1 in early disease stages could be clinically valuable. As a powerful modulator, PD-1 has the potential to exacerbate or prohibit severity in Covid-19, so a cautious approach in evaluating its effects is warranted.

7.4 Colocalization as a tool for drug repurposing

In chapter 6 I widened the scope of my investigation, identifying genetic variants shared between a broad range of diseases. Performing genome-wide multi-trait colocalization for a dataset of 228 GWAS I detected 779 shared associated variants, featuring 4001 pairwise colocalizations between 123 diseases. Common variant associations between diseases are suggestive of a potential shared biological pathway that may be targetable by the same drug. I prioritised among the large number of established disease links by identifying which connections might be the most conducive to clinical applications through drug repurposing. Potential repurposing candidates were selected by finding drugs currently used or in clinical trials for one of two diseases sharing a variant and testing if the drug targets matched genes functionally connected to the shared variant. This method identified 3625 drug repurposing candidates, suggesting an untapped potential in drug repurposing opportunities.

Notably, there was particularly strong support for the use of TYK2 inhibitors in the treatment of autoimmune hypothyroidism. TYK2 is involved in pathways that are heavily implicated in immune system diseases, many of which are under currently treated with or investigated for TYK2 inhibitors[418,419]. Drug repurposing was supported

through two colocalizations, one in association with TYK2 itself and the other in association with the transcription factor STAT4, which is activated by TYK2 activity[417]. A recent Mendelian randomization study has also found evidence suggesting TYK2 inhibitors as a potential treatment option for hypothyroidism[420]. Further experimental and clinical studies will be required to confirm TYK2 inhibitor efficacy and safety in autoimmune hypothyroidism in clinical practice.

There were multiple limitations to my drug repurposing analysis. Although genetic variants often affect the expression of nearby genes, linking the shared genetic variants to their nearest gene does not identify the correct associated gene in all cases. A much more labour intensive but more in depth combination of eQTL-GWAS colocalization analysis (as in chapter 5) and Mendelian randomization analysis could offer a more accurate means of identifying relevant genes, as well as pinpointing the tissues they are affected in. This information would be particularly valuable as diseases may be affected in different cell types and require different directions of modulation based on the affected tissue. The necessary direction in which a drug would need to modulate the shared disease pathway is another factor currently not identified by the analysis. A drug identified through shared associations could in be beneficial for one disease but detrimental to the other. Information on such potential side effects in patients with certain comorbidities would still hold clinical value, but would not encourage drug repurposing. Follow up Mendelian randomization analysis could be used to gain insight into the needed direction of modulation in the repurposing candidate disease. However, as cases such as the failed trials of IL-17 inhibitors to treat inflammatory bowel disease[467,468] illustrate, detrimental effects can still be observed in clinical practice after a wealth of supporting evidence from a variety of biological research, including genetic studies[460 462,464]. The efficacy and safety of any drug will ultimately have to be tested in randomized clinical trials.

Using colocalization analysis to identify shared disease biology that could lead to drug repurposing offers multiple advantages compared to the general strategy of selecting disease links through shared affected genes. Matching GWAS signals to their biological consequence is difficult; both fine-mapping of the associated variant and linking the variant to the correct target gene can be uncertain. Studying genetic links between diseases by detecting the probability of a shared variant can reveal connections that otherwise might stay hidden due to erroneous assumptions about the affected genes. Additionally, missed opportunities for drug discovery might result from false negative results in GWAS, brought on by the necessary stringent multiple testing corrections.

Some studies have tried to circumvent this issue by looking at all nominally significant associations[156,475]. Although mitigated by the integration of additional biological data, such an approach runs a much higher risk of mistakenly focusing on the effects of variants that are not truly associated with the disease. Improving detection power through large-scale colocalization analysis instead could offer similar benefits, while retaining quality control measures that weed out false positive results. By exploring multiple GWAS in combination, the genetics of similar traits can add power to the original study, with the potential to identify additional associated variants that would normally fall below detection threshold within single GWAS.

Colocalization analysis is a powerful tool for studying the genetic relationship between diseases. Its ability to detect shared associations on the level of genetic variants and to reveal new associations by increasing the power of similar disease GWAS offers opportunities to discover candidates for drug repurposing. This analysis could be improved by introducing a more reliable method of linking the shared variants to genes, such as by eQTL colocalization and Mendelian randomization. Further data such as protein quantitative trait loci, splicing quantitative trait loci and chromatin state could be integrated to more accurately select candidates. I identified multiple pleiotropic variants as potentially affecting lncRNAs, information on which is currently in progress of being studied and collated under the FANTOM6 project[493], which may provide another valuable future resource. Overall, my results demonstrate that colocalization will be a valuable metric to include in future drug repurposing analyses.

7.5 Concluding Remarks

Genome-wide association studies present a valuable opportunity for advances in precision medicine. This is true not only in the sense of personalised medicine – everyone has a unique genetic code which could inform perfectly targeted therapeutic interventions – but also in the sense of *precise* medicine. Genetic associations have the power to pinpoint the exact protein or molecular pathway that is in need of treatment in a specific disease. Over the last two decades, GWAS have uncovered much information about how genetic variation influences complex traits. At time of writing, the congregation platform GWAS catalogue hosts 559539 top associations from 6624 publications[168]. While this has led to a better understanding of several disease mechanisms and highlighted new drug targets, compared to the rate of discovery of

genetic associations progress on their functional biological interpretation has unfolded much slower.

The vast number of uncovered variants may paradoxically be partially to blame for this disparity. Few genetic associations have been thoroughly functionally characterised through *in vitro* and *in vivo* experiments, as such studies are costly and time intensive, creating a bottleneck for our understanding of disease and the identification of drug targets. The prioritisation of candidate variants is therefore crucial. As I illustrated in this thesis, post GWAS computational analyses hold the potential to select variants that are particularly promising and identify associated cellular mechanisms for further investigation. This in tandem with experimental validation will be a key strategy to deliver on the promise of genetic studies.

Appendix

Digital Appendix

This thesis is accompanied by a digital appendix containing Supplementary Interactive Figures 1 and 2 which is publicly hosted at https://marie-zz.github.io/digital_appendix/ .

Supplementary Tables and Figures

Supplementary Table 1: Matched phenotypes between the GWAS dataset and the drug dataset. The first column lists the phenotypes in my GWAS dataset. The second lists the phenotypes as they were named on the Open Targets website, and notes additional detail where only part of the drug data was used to more accurately adhere to the GWAS phenotype definition. The third column names corresponding disease ontology identifications as listed for the drug data disease phenotypes on Open Targets for further specification: the EMBL-EBI Experimental Factor Ontology (EFO)[494], the Human Phenotype Ontology (HPO)[495] or the Mondo Disease Ontology (MONDO)[496].

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Acute appendicitis	appendicitis	EFO_0007149
Acute myocardial infarction	acute myocardial infarction	EFO_0008583
Acute pancreatitis	acute pancreatitis	EFO_1000652
Allergic disease (asthma, hay fever or eczema)	allergic disease	MONDO_0005271
Angina (diagnosed by doctor)	angina pectoris	EFO_0003913
Angina (self-reported)	angina pectoris	EFO_0003913
Angina pectoris	angina pectoris	EFO_0003913
Arthrosis	osteoarthritis	MONDO_0005178
Asthma	asthma	MONDO_0004979
Asthma (diagnosed by doctor)	asthma	MONDO_0004979

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Asthma (hospital admissions)	asthma	MONDO_0004979
Asthma (self-reported)	asthma	MONDO_0004979
Asthma-related pneumonia	viral pneumonia, bacterial pneumonia	EFO_0007541, EFO_1001272
Asthma, doctor diagnosed	asthma	MONDO_0004979
Atrial fibrillation (self-reported)	atrial fibrillation	EFO_0000275
Atrial fibrillation and flutter	atrial fibrillation	EFO_0000275
Blood clot in the leg/DVT (diagnosed by doctor)	deep vein thrombosis	EFO_0003907
Blood clot in the lung (diagnosed by doctor)	pulmonary embolism	EFO_0003827
Bronchitis	bronchitis	EFO_0009661
Bronchitis (self-reported)	bronchitis	EFO_0009661
Cardiac arrhytmias, COPD co-morbidities	cardiac arrhythmia (excluding: cardiac arrest)	EFO_0004269
Cellulitis	cellulitis	EFO_0003035
Cerebral infarction	cerebral infarction	MONDO_0002679
Childhood asthma (age<16)	asthma	MONDO_0004979
Cholecystitis	Cholecystitis	HP_0001082
Cholelithiasis	cholelithiasis	EFO_0004799
Cholelithiasis/gall stones (self-reported)	cholelithiasis, gallstones	EFO_0004799, EFO_0004210
Chronic ischaemic heart disease	atherosclerosis, coronary aneurysm, Aortic dissection, ischemic cardiomyopathy	EFO_0003914, EFO_1000881, HP_0002647, EFO_0001425
Churg-Strauss syndrome (ANCA-negative)	Churg-Strauss syndrome	EFO_0007208
COPD differential diagnosis	chronic obstructive pulmonary disease	EFO_0000341
COPD, early/later onset		EFO_0000341

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	chronic obstructive pulmonary disease	
Coronary atherosclerosis	coronary atherosclerosis	MONDO_0021661
Coxarthrosi/arthrosis of hip	osteoarthritis, hip	EFO_1000786
Critical Covid-19	COVID-19	MONDO_0100096
Crohn's disease	Crohn's disease	EFO_0000384
Daytime dozing / sleeping (narcolepsy)	narcolepsy	MONDO_0021107
Deep venous thrombosis/DVT (self-reported)	deep vein thrombosis	EFO_0003907
Diabetes (self-reported)	diabetes mellitus	EFO_0000400
Diabetes diagnosed by doctor	diabetes mellitus	EFO_0000400
Diaphragmatic hernia	hernia (excluding: Inguinal hernia, Umbilical hernia, ventral hernia)	HP_0100790
Diseases of appendix	appendicitis	EFO_0007149
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	anemia (excluding: Lesch-Nyhan syndrome, neonatal anemia), Abnormality of blood and blood-forming tissues (excluding: Arterial thrombosis, deep vein thrombosis, Ecchymosis, epistaxis, Gastrointestinal hemorrhage, hematoma, hemorrhage, Menorrhagia, Portal vein thrombosis, postpartum hemorrhage, Recurrent thrombophlebitis, Thromboembolism, Thrombophlebitis, Venous thrombosis), blood coagulation disease (excluding: acquired thrombotic thrombocytopenic purpura, immunoglobulin-mediated membranoproliferative glomerulonephritis, systemic lupus erythematosus, thrombotic	MONDO_0002280, HP_0001871, EFO_0009314, MONDO_0002243, MONDO_0010518

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	microangiopathy, thrombotic thrombocytopenic purpura), hemorrhagic disease (excluding: acquired thrombotic thrombocytopenic purpura, thrombotic thrombocytopenic purpura, thrombocytopenic purpura), Wiskott-Aldrich syndrome	
Diseases of the circulatory system	heart disease (excluding: 22q11.2 deletion syndrome, Alagille syndrome, Aortic Coarctation, atrial septal defect 1, ATTRV122I amyloidosis, Becker muscular dystrophy, carcinoid heart disease, Churg-Strauss syndrome, congenital heart disease, diffuse scleroderma, Duchenne muscular dystrophy, Fabry disease, Friedreich ataxia, Hypereosinophilic syndrome, hypoplastic left heart syndrome, Leber hereditary optic neuropathy, limited scleroderma, long QT syndrome 3, Noonan syndrome, post-operative atrial fibrillation, primary systemic amyloidosis, Pseudoxanthoma elasticum, Steinert myotonic dystrophy, systemic scleroderma, transposition of the great arteries, ventricular septal defect), vascular disease (excluding: adrenal cortex carcinoma, angiosarcoma, Aortic Coarctation, ataxia telangiectasia, Behcet's syndrome, blue rubber bleb nevus, choroideremia, choroiditis, Churg-Strauss syndrome, compartment syndrome, diabetic foot, diabetic macular edema, diabetic retinopathy, Ehlers-Danlos syndrome, vascular type, epistaxis, epithelioid	EFO_0003777, EFO_0004264, HP_0002619, MONDO_0002052

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	hemangioendothelioma, Fabry disease, familial multiple nevi flammei, famililal cerebral cavernous malformations, Granulomatosis with Polyangiitis, HELLP syndrome, hemangioblastoma, hemangioma, hemangioendothelioma, hemangiopericytoma, hemorrhoid, hepatic veno-occlusive disease, Hepatopulmonary Syndrome, hereditary angioedema, hereditary hemorrhagic telangiectasia, hypertension, pregnancy-induced, Kaposi's sarcoma, kaposiform hemangioendothelioma, lymphangioma, microscopic polyangiitis, migraine disorder, migraine with aura, migraine without aura, mucocutaneous lymph node syndrome, neurofibromatosis type 1, non-infectious posterior uveitis, non-proliferative diabetic retinopathy, Noonan syndrome, ocular hypertension, ocular vascular disease, persistent fetal circulation syndrome, Polyarteritis Nodosa, portal hypertension, preeclampsia, proliferative diabetic retinopathy, Proteus syndrome, pseudotumor cerebri, pulmonary venoocclusive disease, renal artery obstruction, retinal artery occlusion, retinal vasculitis, retinal vein occlusion, severe pre-eclampsia, temporal arteritis, transient ischemic attack, tufted angioma, vascular anomaly, vascular dementia, vascular malformation), Varicose veins, lymphadenitis	

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified	Phlebitis, Thrombophlebitis, Venous thrombosis, Varicose veins, lymphadenitis, lymphedema	EFO_1001395, HP_0004418, HP_0004936, HP_0002619, MONDO_0002052, MONDO_0019297
Disorders of gallbladder, biliary tract and pancreas	gallbladder disease (excluding: gallbladder cancer, gallbladder carcinoma, gallbladder neoplasm), biliary tract disease (excluding: Alagille syndrome, ampulla of Vater adenocarcinoma, bile duct cancer, bile duct carcinoma, biliary tract cancer, biliary tract neoplasm, cholangiocarcinoma, combined hepatocellular carcinoma and cholangiocarcinoma, familial intrahepatic cholestasis, gallbladder cancer, gallbladder carcinoma, gallbladder neoplasm, hilar cholangiocarcinoma, intrahepatic cholangiocarcinoma, intrahepatic cholestasis, malignant tumor of extrahepatic bile duct, progressive familial intrahepatic cholestasis), pancreas disease (excluding: diabetes mellitus, diabetic ketoacidosis, familial hyperinsulinism, functional pancreatic neuroendocrine tumor, gestational diabetes, hyperinsulinism, insulinoma, islet cell tumor, latent autoimmune diabetes in adults, MODY, monogenic diabetes, pancreatic adenocarcinoma, pancreatic adenosquamous carcinoma, pancreatic carcinoma, pancreatic ductal adenocarcinoma, pancreatic endocrine carcinoma, Pancreatic Gastrinoma, pancreatic insulinoma, pancreatic neoplasm,	EFO_0003832, EFO_0009534, EFO_0009605

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	pancreatic neuroendocrine tumor, prediabetes syndrome, somatostatinoma, Stiff-Person syndrome, type 1 diabetes mellitus, type 2 diabetes mellitus, Undifferentiated Pancreatic Carcinoma, Zollinger-Ellison Syndrome)	
Diverticular disease of intestine	diverticular disease	EFO_0009959
Eczema	Eczema	HP_0000964
Eczema/dermatitis (self-reported)	Eczema, dermatitis	HP_0000964, MONDO_0002406
Endocrine, nutritional and metabolic diseases	endocrine system disease (including: acromegaly, ACTH Syndrome Ectopic, acute intermittent porphyria, Adenomyosis, adrenal gland disease, adrenal gland hyperfunction, adrenocortical insufficiency, adrenogenital syndrome, adrenoleukodystrophy, adrenomyeloneuropathy, Albright hereditary osteodystrophy, amenorrhea, Central precocious puberty, chronic primary adrenal insufficiency, Combined hyperlipidemia, congenital adrenal hyperplasia, congenital hypothyroidism, Cushing syndrome, diabetes mellitus, diabetic ketoacidosis, endocrine system disease, erythropoietic protoporphyria, euthyroid sick syndrome, familial hyperinsulinism, genetic endocrine growth disease, goiter, Graves disease, Graves ophthalmopathy, growth hormone insensitivity syndrome, Hashimoto's thyroiditis, hyperaldosteronism, hyperandrogenism, hyperinsulinemic	EFO_0001379, EFO_0000589

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	hypoglycemia, hyperinsulinism, Hyperlipoproteinemia type 1, Hyperlipoproteinemia type 4, hyperparathyroidism, hyperprolactinemia, hyperthyroidism, hypogonadism, hypogonadotropic hypogonadism, hypoparathyroidism, hypopituitarism, Kallmann syndrome, Laron syndrome, myxedema, nodular goiter, parathyroid disease, pituitary dwarfism, pituitary gland disease, pituitary-dependent Cushing's disease, polycystic ovary syndrome, porphyria cutanea tarda, precocious puberty, primary adrenal insufficiency, primary aldosteronism, primary hyperparathyroidism, primary ovarian insufficiency, pseudohypoparathyroidism, pseudohypoparathyroidism type 1A, secondary hyperparathyroidism, subacute thyroiditis, type 1 diabetes mellitus, type 2 diabetes mellitus, Zollinger-Ellison Syndrome), metabolic disease (excluding: anorexia nervosa, binge eating, bulimia nervosa, cardiac amyloidosis, chondrocalcinosis, eating disorder, familial intrahepatic cholestasis, Fanconi anemia, gestational diabetes, gout, hyperkalemic periodic paralysis, hypokalemic periodic paralysis, inborn errors of metabolism, Leber hereditary optic neuropathy, Leigh syndrome, Lennox-Gastaut syndrome, Lewy body dementia, macroglobulinemia, multiple system atrophy, Prader-Willi syndrome, prediabetes syndrome, progressive familial intrahepatic cholestasis,	

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	retinitis pigmentosa, Rubinstein-Taybi syndrome, Smith-Lemli-Opitz syndrome, Smith-Magenis syndrome, Stiff-Person syndrome, Waldenstrom macroglobulinemia, xeroderma pigmentosum, Zellweger syndrome)	
Endometriosis	endometriosis	EFO_0001065
Endometriosis, IBD co-morbidity	endometriosis	EFO_0001065
Fibroblastic disorders	Plantar Fasciitis	EFO_1001909
Glaucoma	glaucoma	MONDO_0005041
Glaucoma (self-reported)	glaucoma	MONDO_0005041
Gout (self-reported)	gout	EFO_0004274
Haemorrhoids	hemorrhoid	EFO_0009552
Hayfever or allergic rhinitis (diagnosed by doctor)	allergic rhinitis	EFO_0005854
Hayfever, allergic rhinitis or eczema (diagnosed by doctor)	allergic rhinitis, Eczema	EFO_0005854, HP_0000964
Hayfever/allergic rhinitis (self-reported)	allergic rhinitis	EFO_0005854
Heart arrhythmia (self-reported)	cardiac arrhythmia	EFO_0004269
Heart attack (diagnosed by doctor)	myocardial infarction	EFO_0000612
Heart attack/myocardial infarction (self-reported)	myocardial infarction	EFO_0000612
Hernia	hernia	HP_0100790
High blood pressure (diagnosed by doctor)	hypertension	EFO_0000537
Hordeolum and chalazion	hordeolum	EFO_0007315
Hypertension (self-reported)	hypertension	EFO_0000537
Hyperthyroidism/thyrotoxicosis (self-reported)	hyperthyroidism, Thyrotoxicosis	EFO_0009189, EFO_0009190
	hypothyroidism, myxedema	

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Hypothyroidism/myxoedema (self-reported)		EFO_0004705, EFO_1001055
Idiopathic pulmonary fibrosis	idiopathic pulmonary fibrosis	EFO_0000768
Inflammatory bowel disease	inflammatory bowel disease	EFO_0003767
Inguinal hernia	Inguinal hernia	HP_0000023
Ischaemic heart disease, wide definition	coronary artery disease, angina pectoris, acute myocardial infarction, ST Elevation Myocardial Infarction, anterolateral myocardial infarction, Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction	EFO_0001645, EFO_0003913, EFO_1000013, EFO_1001375, EFO_0008583, EFO_0008585, EFO_1000812, EFO_0008584
Ischaemic Stroke, excluding all haemorrhages	Ischemic stroke, cerebral infarction, stroke	HP_0002140, MONDO_0002679, EFO_0000712
Major coronary heart disease event	coronary artery disease, angina pectoris, acute myocardial infarction, ST Elevation Myocardial Infarction, anterolateral myocardial infarction, Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction, cardiac arrest	EFO_0001645, EFO_0003913, EFO_1000013, EFO_1001375, EFO_0008583, EFO_0008585, EFO_1000812, EFO_0008584, EFO_0009492
Malabsorption/coeliac disease (self-reported)	Malabsorption, celiac disease	HP_0002024, EFO_0001060
Migraine (self-reported)	migraine disorder	MONDO_0005277
Mouth ulcers	Oral ulcer	HP_0000155
Multiple sclerosis (self-reported)	multiple sclerosis	MONDO_0005301
Myocardial infarction	myocardial infarction	EFO_0000612
Nasal polyps (self-reported)	Nasal polyposis	HP_0100582
Noninfectious colitis	colitis	EFO_0003872
Osteoarthritis (self-reported)	osteoarthritis	MONDO_0005178

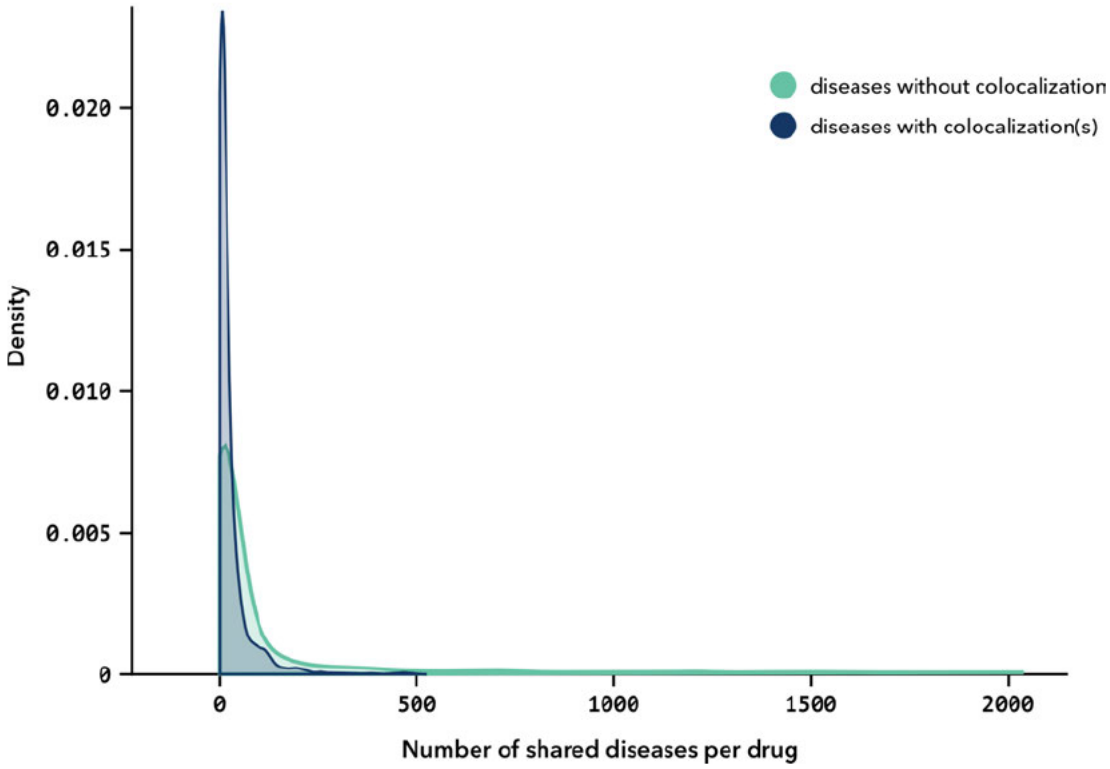
GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Other anaemias	anemia	MONDO_0002280
Other cataract	cataract	MONDO_0005129
Other chronic obstructive pulmonary disease	chronic obstructive pulmonary disease	EFO_0000341
Other diseases of the digestive system	Malabsorption, celiac disease, blind loop syndrome, postgastrectomy syndrome, Gastrointestinal hemorrhage	HP_0002024, EFO_0001060, EFO_0007175, MONDO_0004566, HP_0002239
Other ILD-related CVD-co-morbidities	pulmonary embolism, pulmonary hypertension, chronic pulmonary heart disease	EFO_0003827, MONDO_0005149, MONDO_0001493
Other non-infective gastro-enteritis and colitis	colitis, gastroenteritis	EFO_0003872, EFO_1001463
Other pulmonary diagnosis	respiratory system disease (excluding: acute lung injury, adenosquamous lung carcinoma, Alpha-1-antitrypsin deficiency, altitude sickness, asphyxia neonatorum, Biphasic Mesothelioma, bronchial neoplasm, bronchoalveolar adenocarcinoma, bronchogenic carcinoma, bronchopulmonary dysplasia, Churg-Strauss syndrome, congenital diaphragmatic hernia, cystic fibrosis, diffuse scleroderma, Eisenmenger syndrome, high altitude pulmonary edema, hypopharyngeal carcinoma, hypopharyngeal squamous cell carcinoma, idiopathic and/or familial pulmonary arterial hypertension, idiopathic pulmonary arterial hypertension, juvenile dermatomyositis, large cell lung carcinoma, laryngeal carcinoma, laryngeal neoplasm, laryngeal squamous cell carcinoma, limited scleroderma, lung adenocarcinoma, lung cancer, lung carcinoma, lung	EFO_0000684

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	neoplasm, Lung Sarcomatoid Carcinoma, malignant peritoneal mesothelioma, malignant pleural mesothelioma, Malignant Pleural Neoplasm, Middle East respiratory syndrome, nasal cavity cancer, nasal cavity carcinoma, nasopharyngeal neoplasm, neoplasm of hypopharynx, neoplasm of oropharynx, newborn respiratory distress syndrome, non-small cell lung carcinoma, non-small cell squamous lung carcinoma, obstructive sleep apnea, Olfactory Neuroblastoma, oropharyngeal carcinoma, oropharynx cancer, oropharynx squamous cell carcinoma, paranasal sinus cancer, paranasal sinus neoplasm, persistent fetal circulation syndrome, pharynx cancer, Pleuropulmonary blastoma, primary ciliary dyskinesia, pulmonary arterial hypertension, Pulmonary arterial hypertension associated with portal hypertension, pulmonary neuroendocrine tumor, pulmonary sarcoidosis, pulmonary tuberculosis, pulmonary venoocclusive disease, respiratory distress syndrome in premature infants, respiratory system neoplasm, rhinoscleroma, severe acute respiratory syndrome, sleep apnea, small cell lung carcinoma, spasmodic dystonia, squamous cell lung carcinoma, systemic scleroderma, tonsil cancer, Tracheal Squamous Cell Carcinoma, voice disorders)	
Other rheumatoid arthritis	rheumatoid arthritis (excluding: ankylosing spondylitis, psoriatic arthritis)	EFO_0000685
Other serious eye condition		EFO_0003966

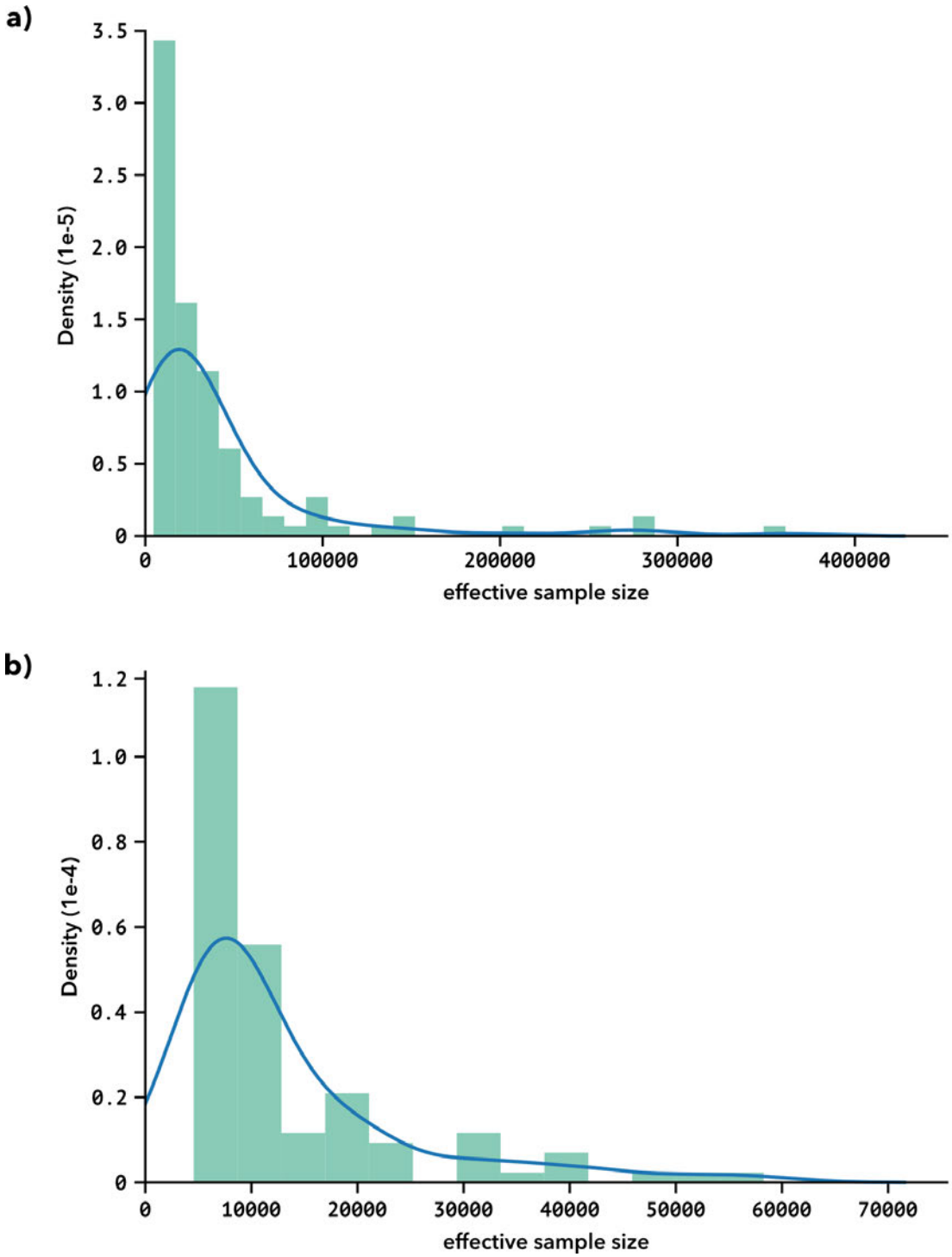
GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
	eye disease (excluding: age-related macular degeneration, angle-closure glaucoma, atrophic macular degeneration, cataract, choroidal melanoma, conjunctival tumor, cystoid macular edema, diabetic macular edema, diabetic maculopathy, diabetic retinopathy, glaucoma, low tension glaucoma, macular degeneration, macular holes, macular retinal edema, macular telangiectasia type 2, neovascular glaucoma, non-proliferative diabetic retinopathy, Ocular Melanoma, open-angle glaucoma, proliferative diabetic retinopathy, retinoblastoma, unilateral retinoblastoma, Uveal Melanoma, wet macular degeneration)	
Other/unspecified rheumatoid arthritis	rheumatoid arthritis (excluding: ankylosing spondylitis, psoriatic arthritis)	EFO_0000685
Palmar fascial fibromatosis/ Dupuytren	Dupuytren Contracture	EFO_0004229
Paralytic ileus and intestinal obstruction	Paralytic ileus	HP_0002590
Paralytic ileus and intestinal obstruction without hernia	Paralytic ileus	HP_0002590
Peripheral artery disease	peripheral vascular disease (excluding: erythromelalgia, hereditary hemorrhagic telangiectasia, intermittent vascular claudication, primary erythermalgia, telangiectasis)	EFO_0003875
Phlebitis and thrombophlebitis	Phlebitis, Thrombophlebitis	EFO_1001395, HP_0004418
Pneumonia, organism unspecified	pneumonia	EFO_0003106
Polyarthropathies	polyarticular arthritis	MONDO_0024280

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Psoriasis (self-reported)	psoriasis	EFO_0000676
Pulmonary embolism	pulmonary embolism	EFO_0003827
Pulmonary embolism (self-reported)	pulmonary embolism	EFO_0003827
Rheumatoid arthritis	rheumatoid arthritis (excluding: ankylosing spondylitis, psoriatic arthritis)	EFO_0000685
Rheumatoid arthritis (self-reported)	rheumatoid arthritis	EFO_0000685
Senile cataract	cataract	MONDO_0005129
Sleep apnoea (self-reported)	Apnea	HP_0002104
Stroke	stroke, intracerebral hemorrhage, cerebral infarction, subarachnoid hemorrhage, transient ischemic attack	EFO_0000712, EFO_0005669, MONDO_0002679, EFO_0000713, EFO_0003764
Stroke (diagnosed by doctor)	stroke	EFO_0000712
Stroke (self-reported)	stroke	EFO_0000712
Stroke, excluding SAH (spontaneous subarachnoid hemorrhage)	stroke, intracerebral hemorrhage, cerebral infarction	EFO_0000712, EFO_0005669, MONDO_0002679
Stroke, including SAH (spontaneous subarachnoid hemorrhage)	stroke, intracerebral hemorrhage, cerebral infarction, subarachnoid hemorrhage	EFO_0000712, EFO_0005669, MONDO_0002679, EFO_0000713
Suggestive for eosinophilic asthma	Nasal Cavity Polyp, Churg-Strauss syndrome, eosinophilic pneumonia	EFO_1000391, EFO_0007208, EFO_0007257
Systemic lupus erythematosus	systemic lupus erythematosus	MONDO_0007915
Type 2 diabetes (self-reported)	type 2 diabetes mellitus	MONDO_0005148
Ulcerative colitis	ulcerative colitis	EFO_0000729
Ulcerative colitis (de Lange)	ulcerative colitis	EFO_0000729
Ulcerative colitis (Liu)	ulcerative colitis	EFO_0000729
Ulcerative colitis (self-reported)	ulcerative colitis	EFO_0000729

GWAS phenotype	Open Targets phenotype(s)	disease ontology ID(s)
Ulcerative colitis, NAS	ulcerative colitis	EFO_0000729
Unspecified acute lower respiratory infection	lower respiratory tract disease (including: lung disease)	EFO_0009433
Unstable angina pectoris	angina pectoris	EFO_0003913
Varicose veins of lower extremities	Varicose veins	HP_0002619
Venous thromboembolism	venous thromboembolism	EFO_0004286



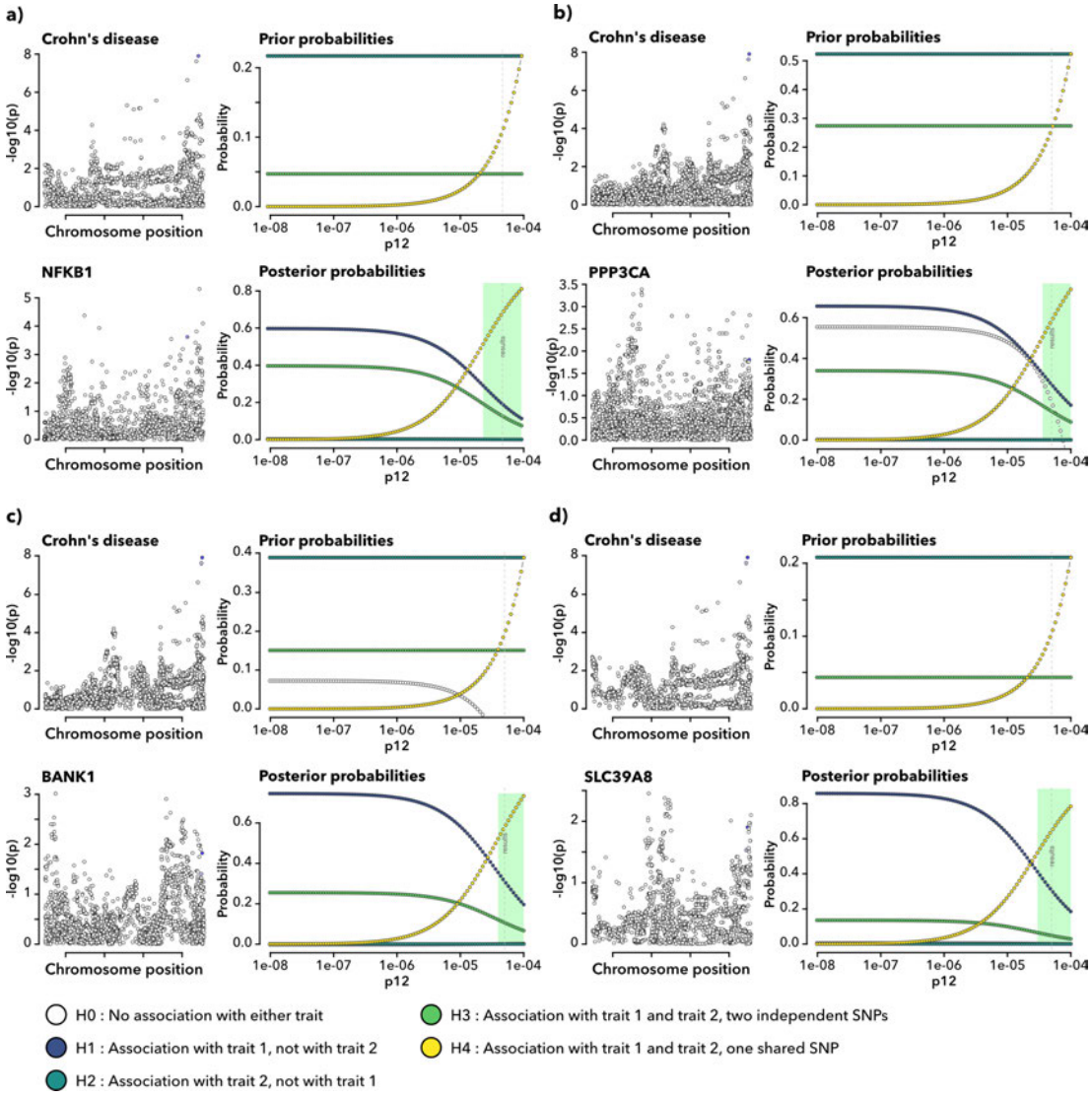
Supplementary Figure 1: Distribution of the number of disease pairs sharing a drug. Each entry on the x axis is for one drug and its y value is the density of disease pairs that share it.



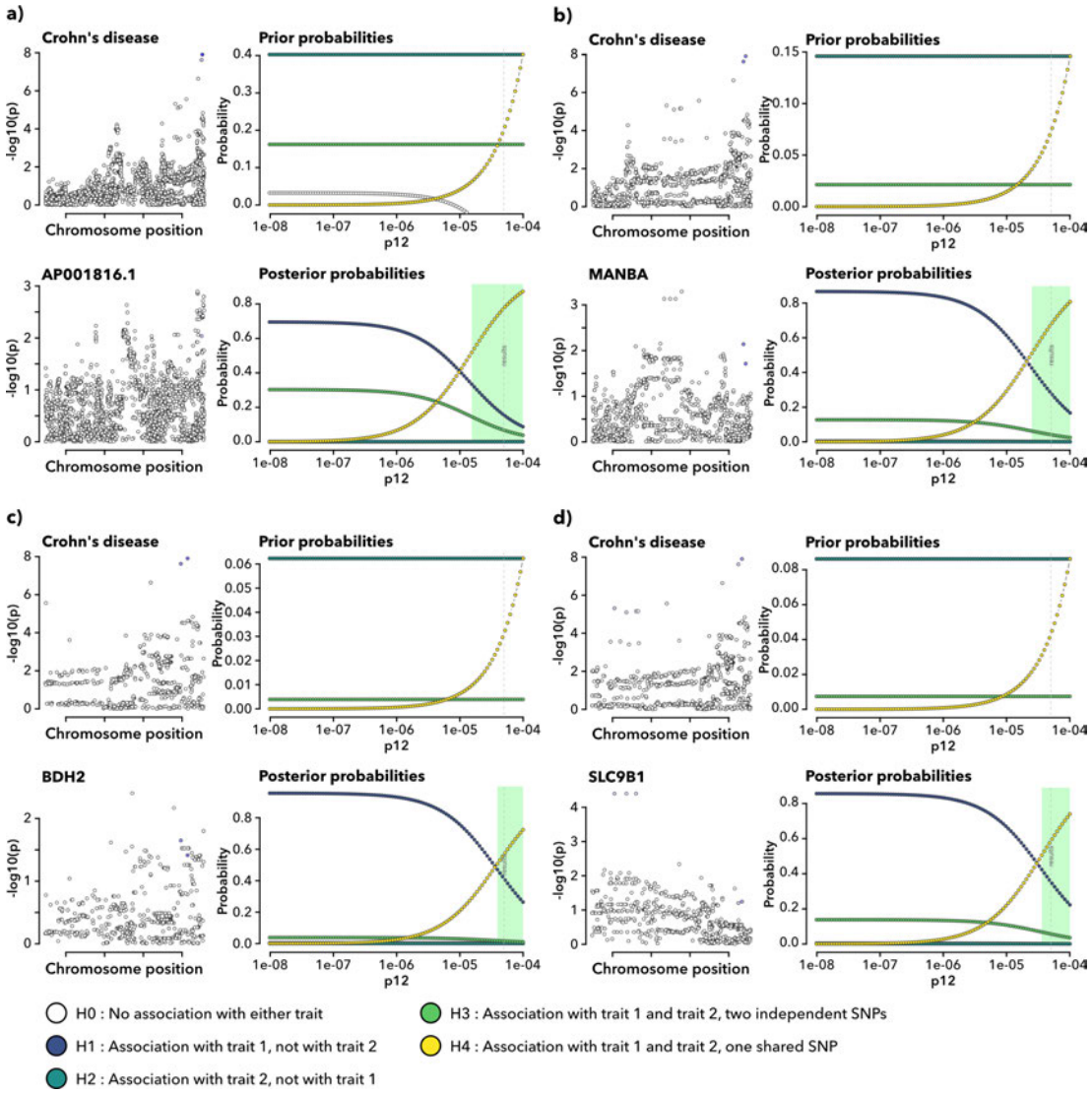
Supplementary Figure 2: Distribution of effective GWAS sample sizes for a) GWAS with colocalizations and b) GWAS without colocalizations.

GWAS and expression data colocalization sensitivity plots

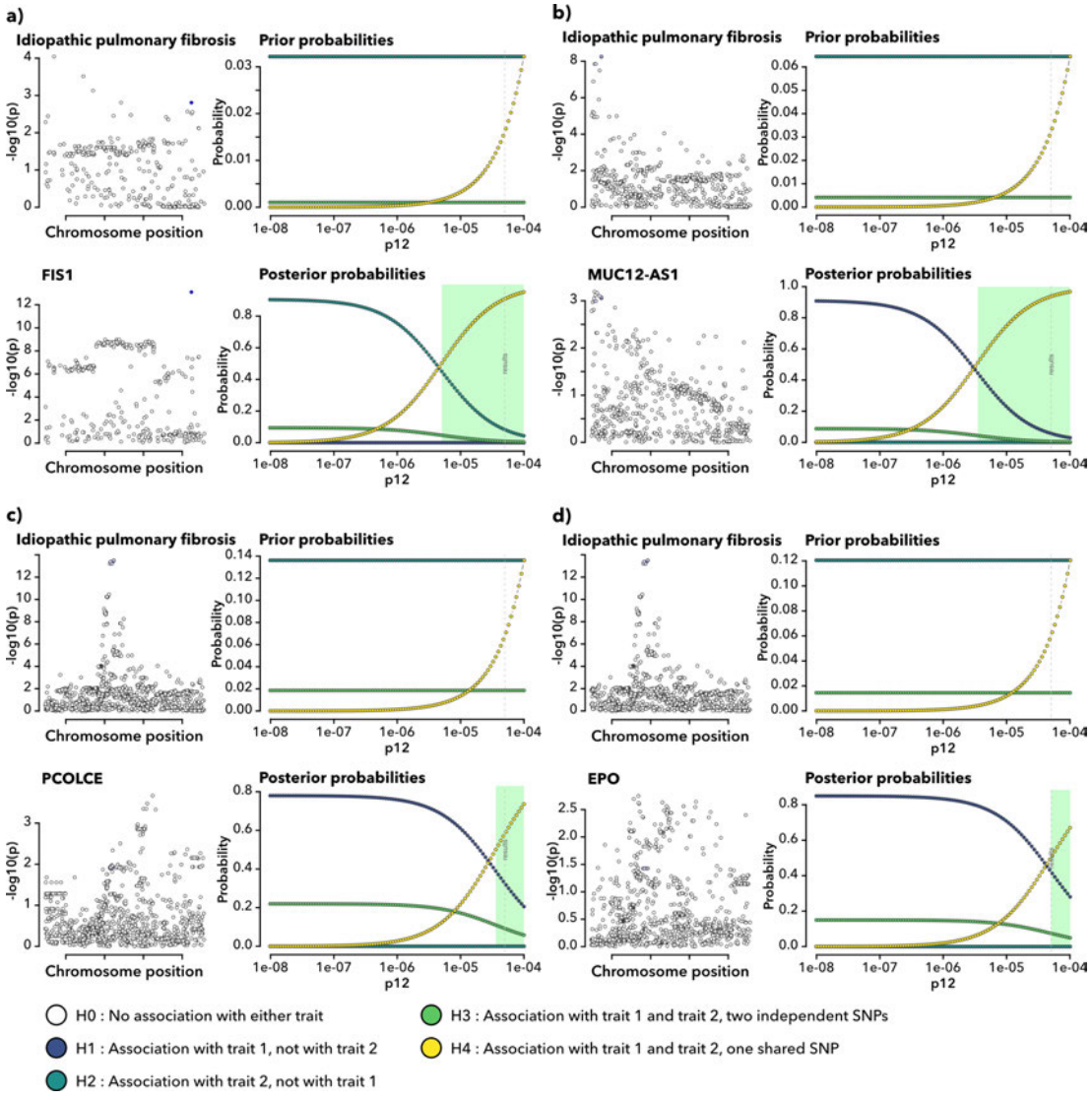
Supplementary figures 3-18 contain the sensitivity plots for colocalizations between GWAS and eQTL signals as discussed in chapter 5. Each subfigure contains the same elements: On the left are the local manhattan plots of the colocalizing traits. Hypothesis outcome probabilities for a random SNP in the region are displayed on the top right, while hypothesis outcomes for the lead colocalization SNP are on the bottom right. The x-axis describes changes in p12, the assumed prior probability that a random SNP in the region is jointly associated for both traits. The green shaded region marks the needed H4 > 0.5 probability threshold over the range of prior probabilities for which it is supported.



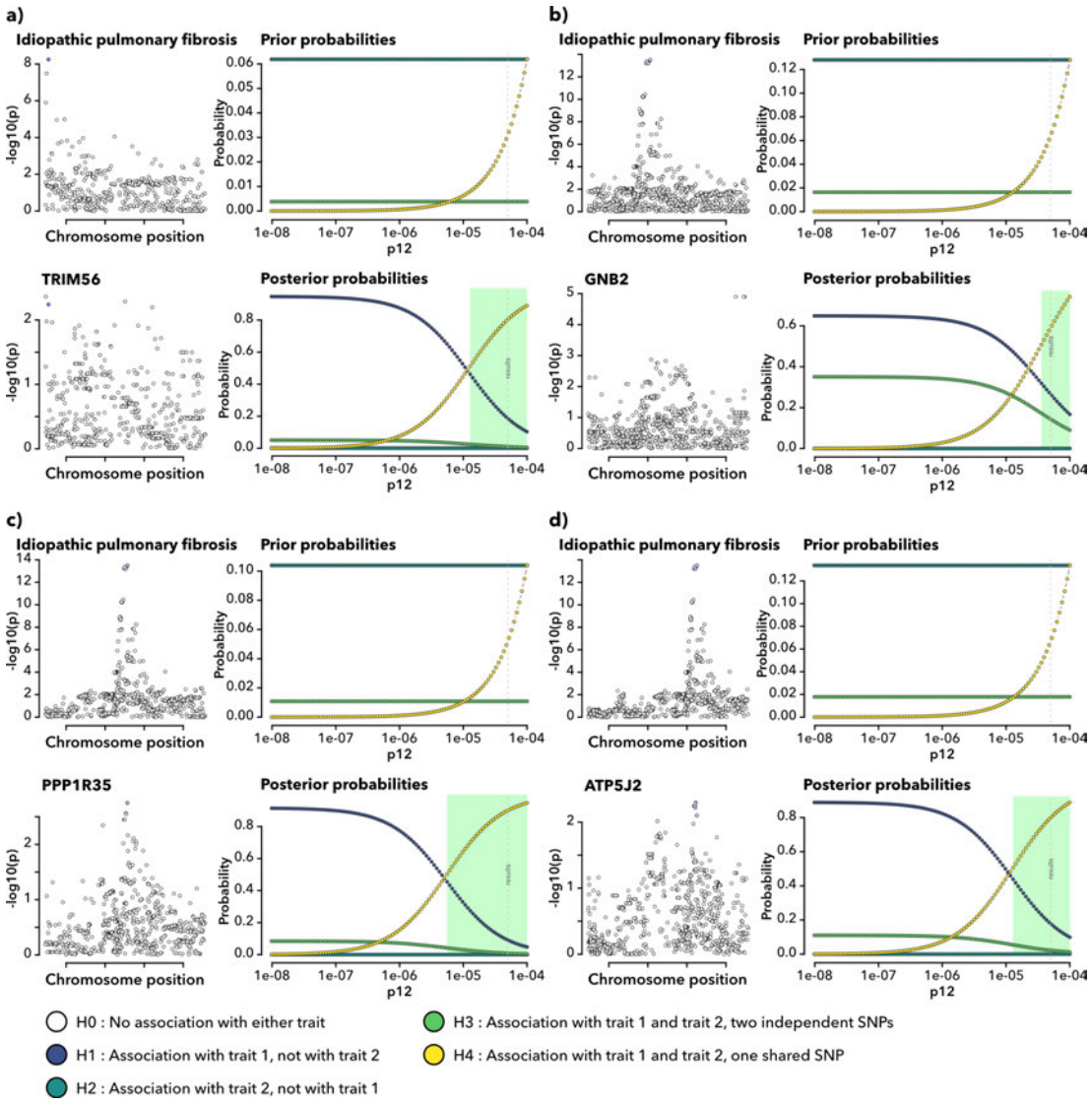
Supplementary Figure 3: Sensitivity plots for GWAS and eQTL signal colocalizations. a) Crohn's disease and NFKB1 expression in blood on chromosome 4 region 100678360-103221356, b) Crohn's disease and PPP3CA expression in transverse colon on chromosome 4 region 100678360-103221356, c) Crohn's disease and BANK1 expression in naïve/immature B cells on chromosome 4 region 100678360-103221356 and d) Crohn's disease and SLC39A8 expression in CD8 S100B T cells on chromosome 4 region 100678360-103221356.



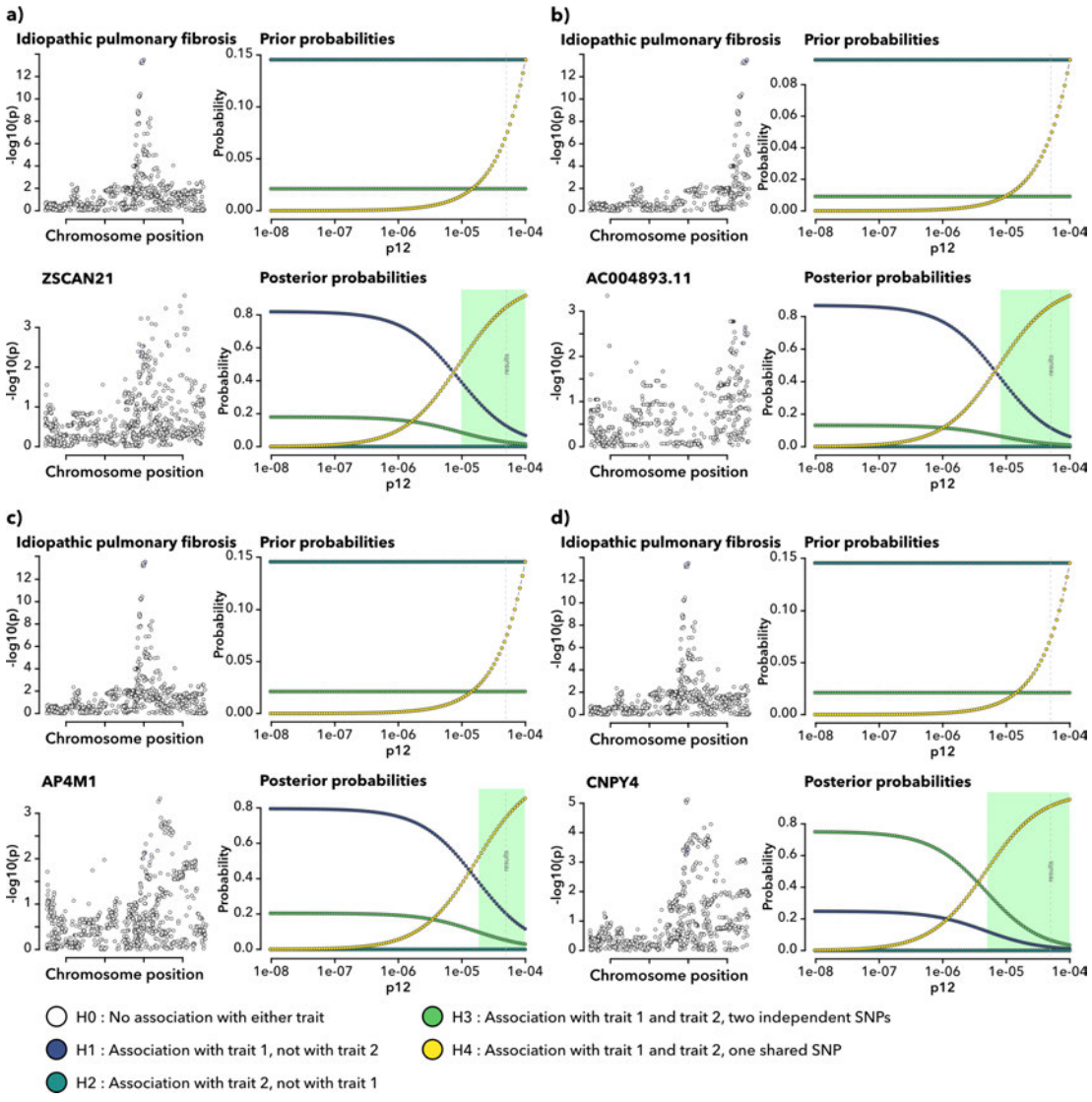
Supplementary Figure 4: Sensitivity plots for GWAS and eQTL signal colocalizations. a) Crohn's disease and AP001816.1 expression in CD8 effector memory T cells on chromosome 4 region 100678360-103221356, b) Crohn's disease and MANBA expression in CD8 effector memory T cells on chromosome 4 region 100678360-103221356, c) Crohn's disease and BDH2 expression in natural killer recruiting cells on chromosome 4 region 100678360-103221356 and d) Crohn's disease and SLC9B1 expression in natural killer cells on chromosome 4 region 100678360-103221356.



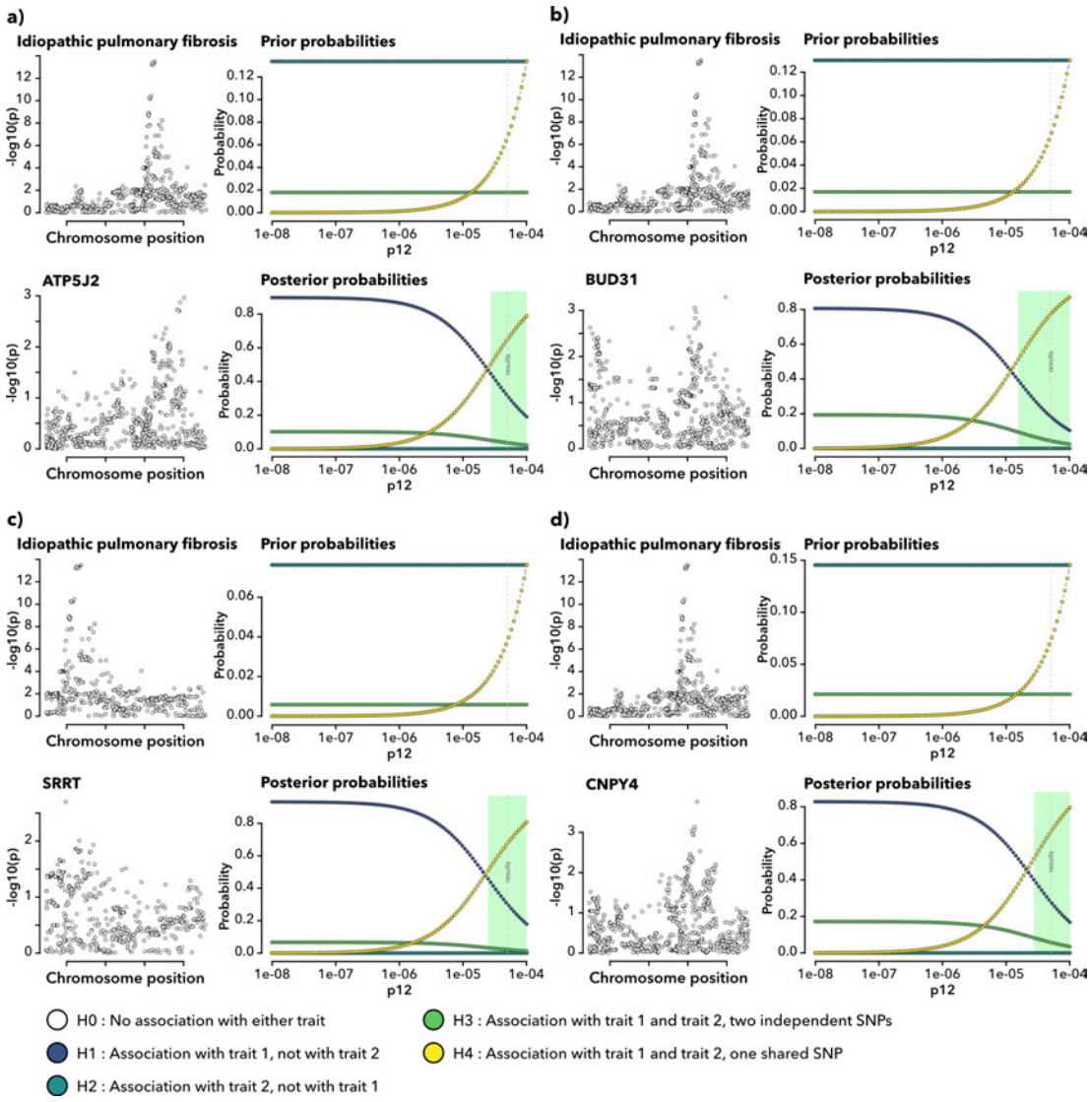
Supplementary Figure 5: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and FIS1 expression in blood on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and MUC12-AS1 expression in blood on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and PCOLCE expression in lung on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and EPO expression in lung on chromosome 7 region 98715474-100196651.



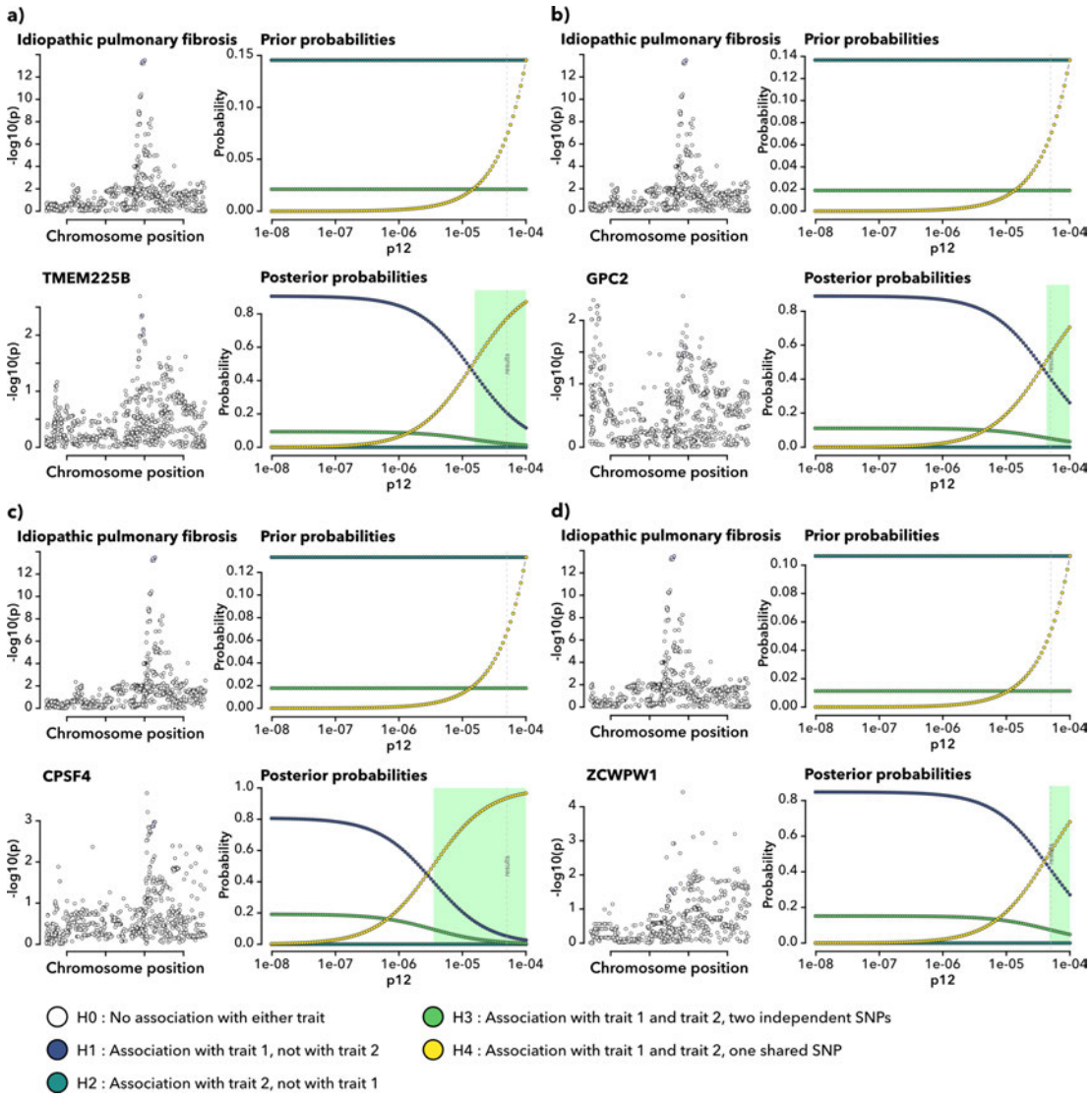
Supplementary Figure 6: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and TRIM56 expression in lung on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and GNB2 expression in lung on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and PPP1R35 expression in naïve/immature B cells on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and ATP5J2 expression in memory B cells on chromosome 7 region 98715474-100196651.



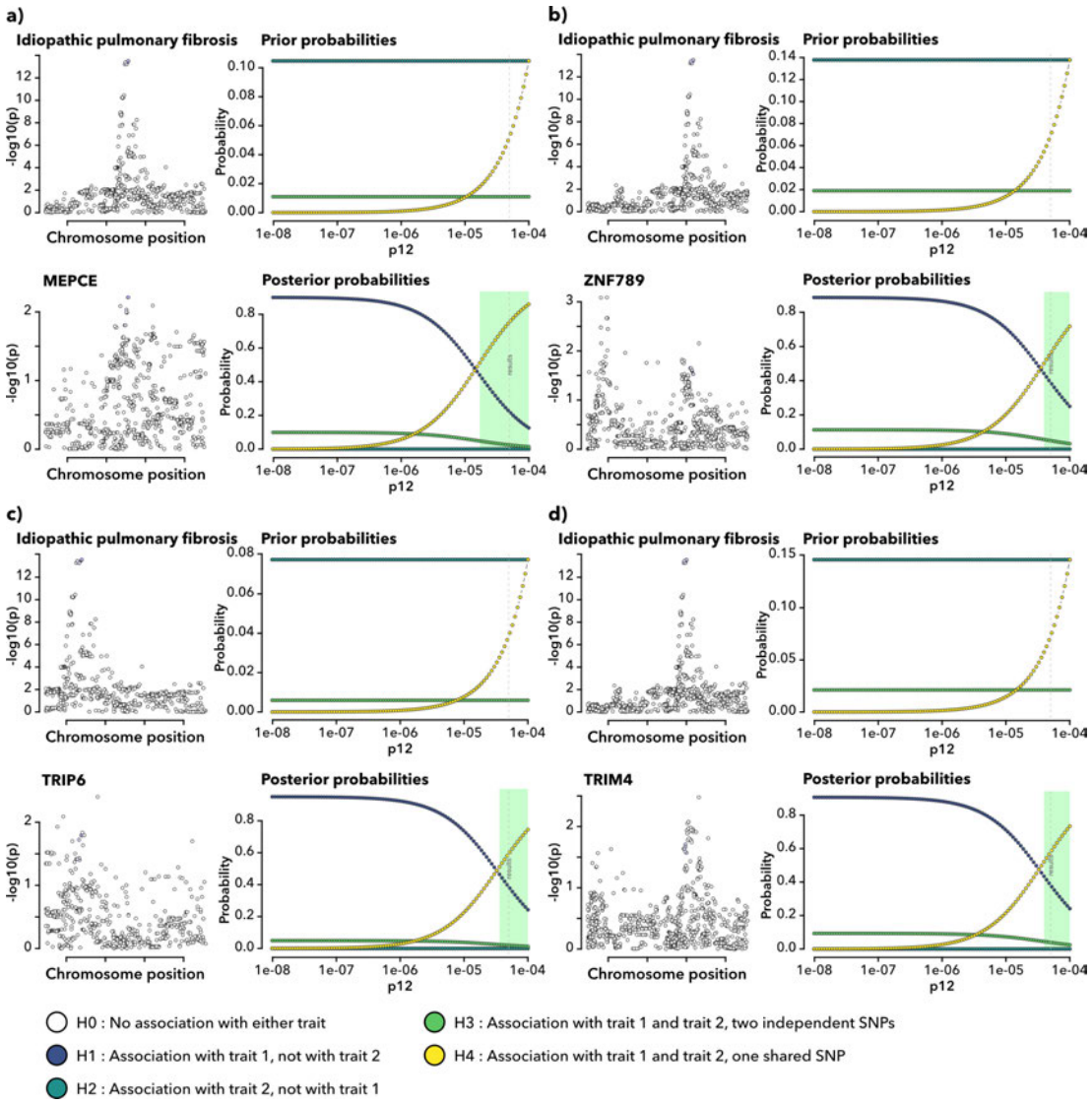
Supplementary Figure 7: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and ZSCAN21 expression in CD4 effector memory T/TEMRA cells on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and AC004893.11 expression in CD4 naïve/central memory T cells on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and AP4M1 expression in CD4 naïve/central memory T cells on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and CNPY4 expression in CD4 naïve/central memory T cells on chromosome 7 region 98715474-100196651.



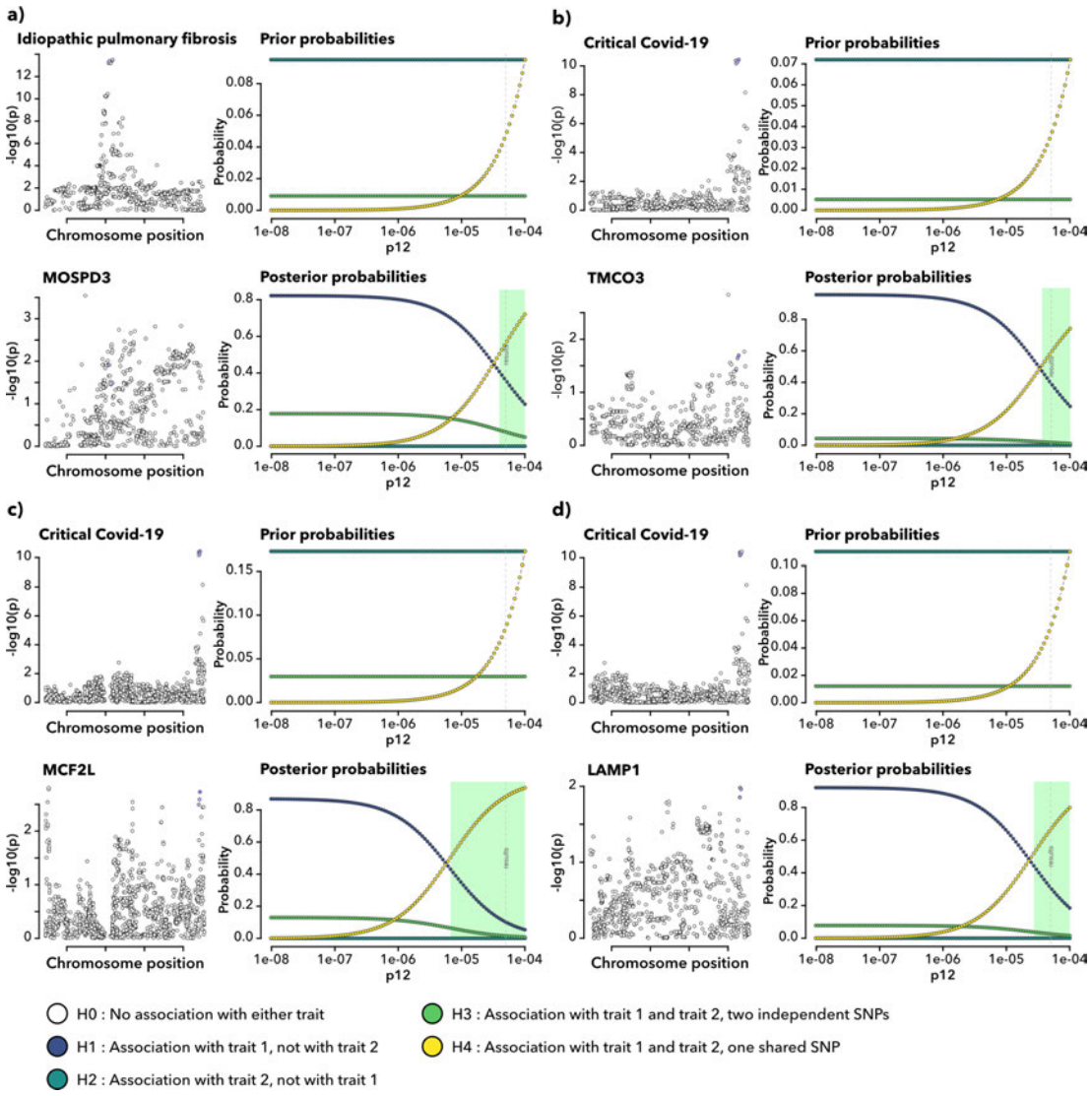
Supplementary Figure 8: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and ATP5J2 expression in CD8 effector memory T cells on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and BUD31 expression in CD8 effector memory T cells on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and SRRT expression in CD8 effector memory T cells on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and CNPY4 expression in CD8 naïve/central memory T cells on chromosome 7 region 98715474-100196651.



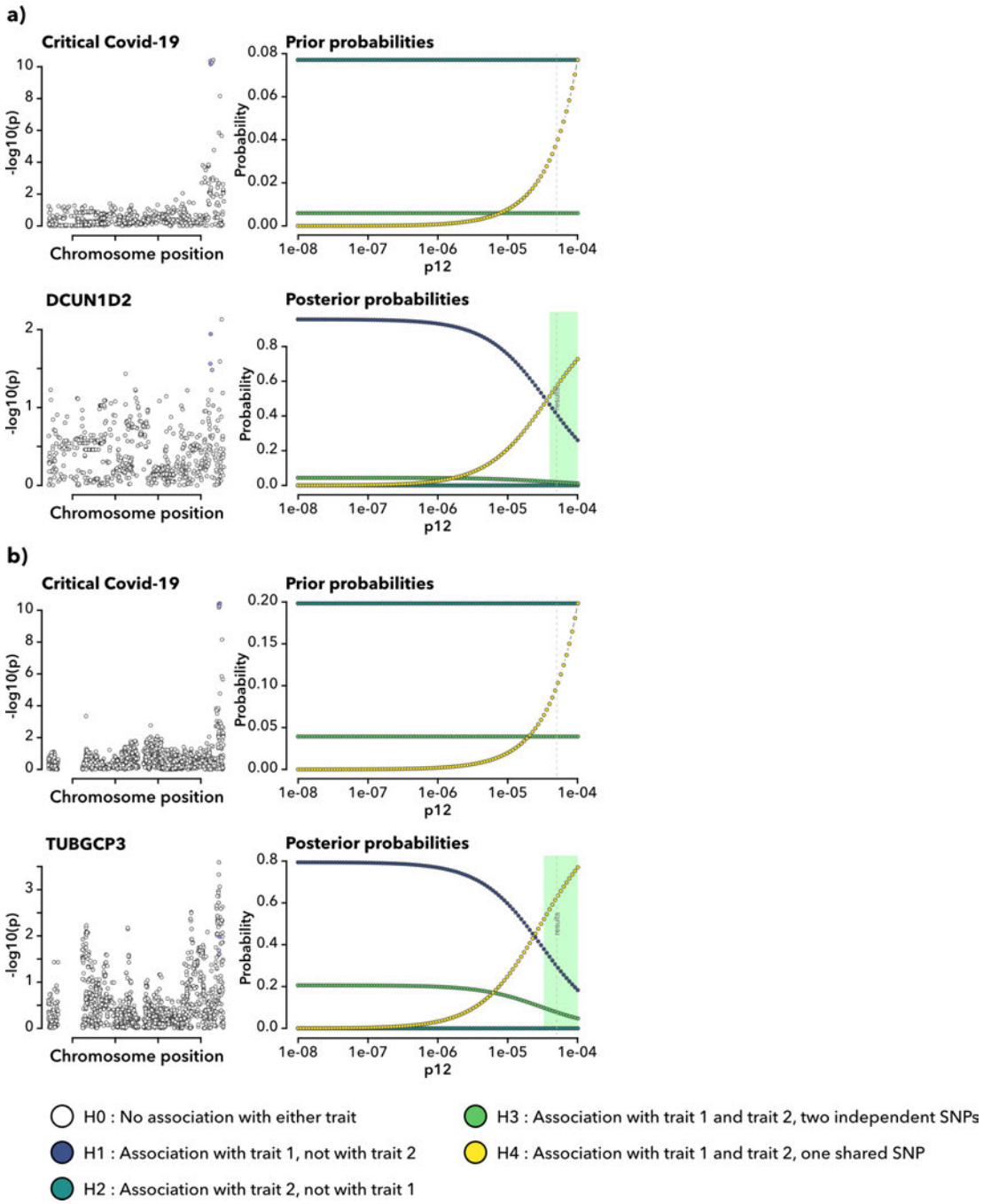
Supplementary Figure 9: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and TMEM225B expression in CD8 naïve/central memory T cells on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and GPC2 expression in CD8 S100B T cells on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and CPSF4 expression in classic monocytes on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and ZCWPW1 expression in classic monocytes on chromosome 7 region 98715474-100196651.



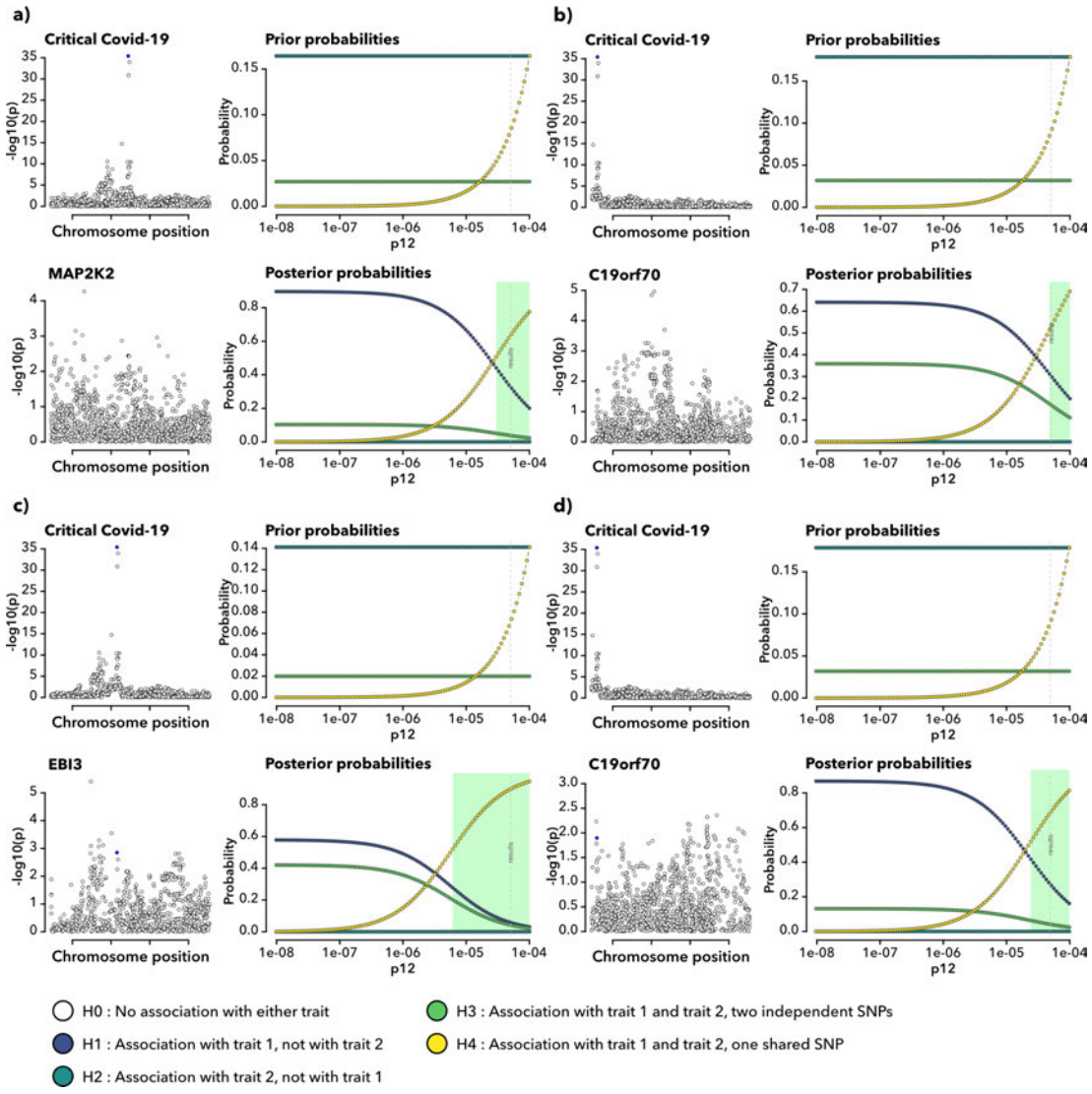
Supplementary Figure 10: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and MEPCE expression in plasma cells on chromosome 7 region 98715474-100196651, b) idiopathic pulmonary fibrosis and ZNF789 expression in classic monocytes on chromosome 7 region 98715474-100196651, c) idiopathic pulmonary fibrosis and TRIP6 expression in non-classic monocytes on chromosome 7 region 98715474-100196651 and d) idiopathic pulmonary fibrosis and TRIM4 expression in natural killer cells on chromosome 7 region 98715474-100196651.



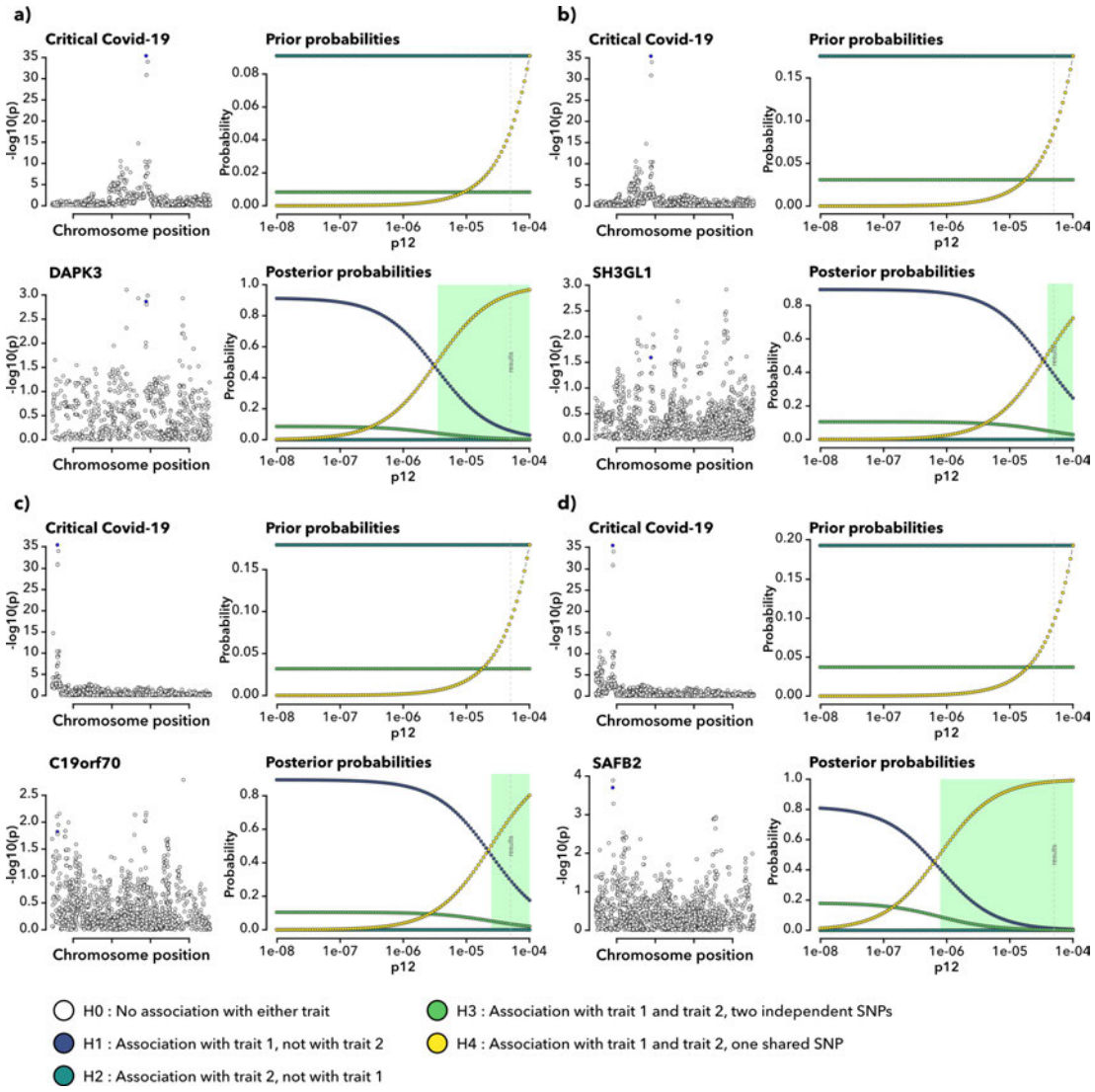
Supplementary Figure 11: Sensitivity plots for GWAS and eQTL signal colocalizations. a) idiopathic pulmonary fibrosis and MOSPD3 expression in natural killer recruiting cells on chromosome 7 region 98715474-100196651, b) critical Covid-19 and TMCO3 expression in CD8 S100B T cells on chromosome 13 region 112247592-113572488, c) critical Covid-19 and MCF2L expression in CD4 effector memory T/TEMRA cells on chromosome 13 region 112247592-113572488 and d) critical Covid-19 and LAMP1 expression in CD8 effector memory T cells on chromosome 13 region 112247592-113572488.



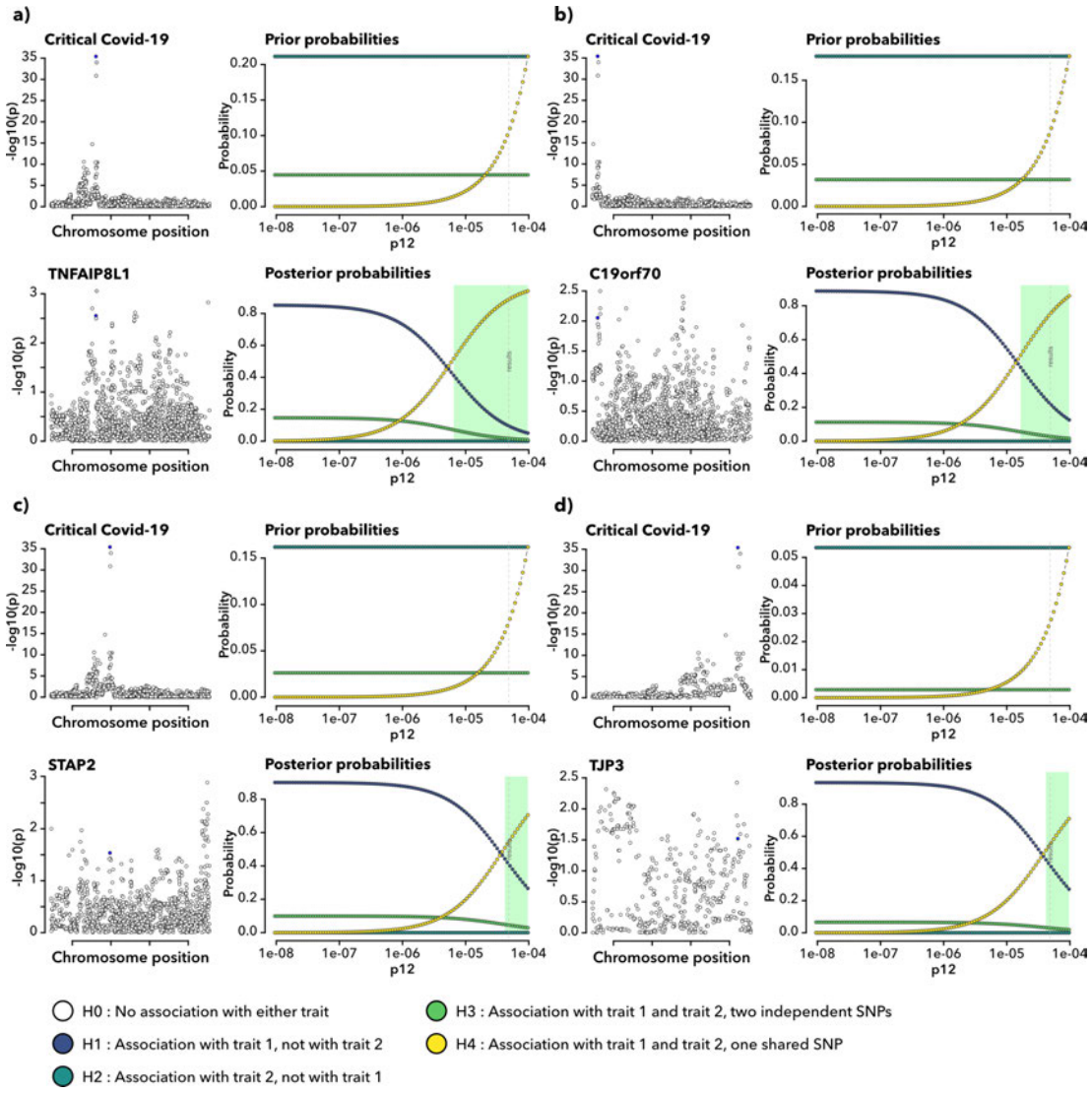
Supplementary Figure 12: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and DCUN1D2 expression in CD8 naïve/central memory T cells on chromosome 13 region 112247592-113572488 and b) critical Covid-19 and TUBGCP3 expression in classic monocytes on chromosome 13 region 112247592-113572488.



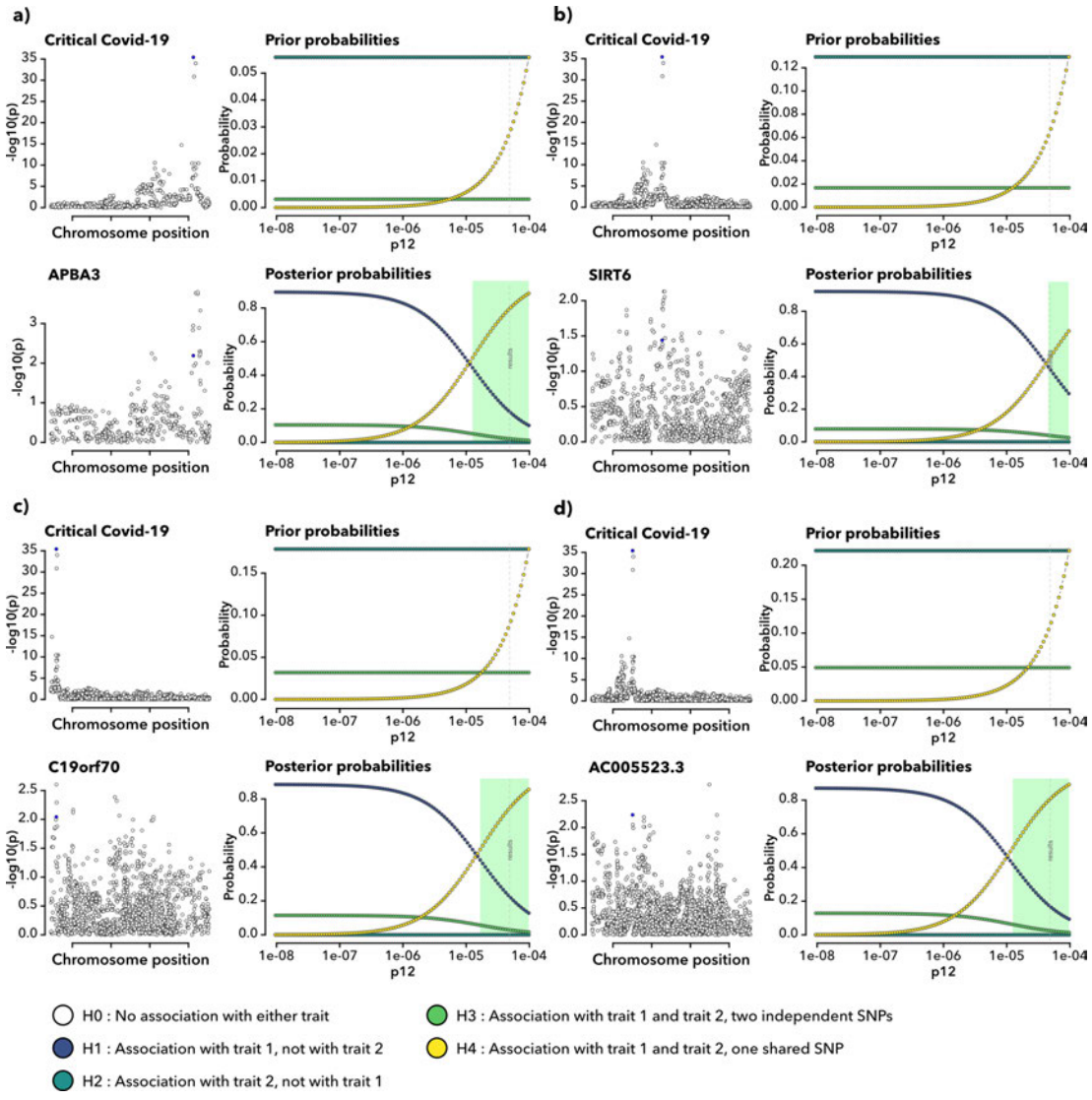
Supplementary Figure 13: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and MAP2K2 expression in blood on chromosome 19 region 4348967-5811852, b) critical Covid-19 and C19orf70 expression in naïve/immature B cells on chromosome 19 region 4348967-5811852, c) critical Covid-19 and EBI3 expression in naïve/immature B cells on chromosome 19 region 4348967-5811852 and d) critical Covid-19 and C19orf70 expression in memory B cells on chromosome 19 region 4348967-5811852.



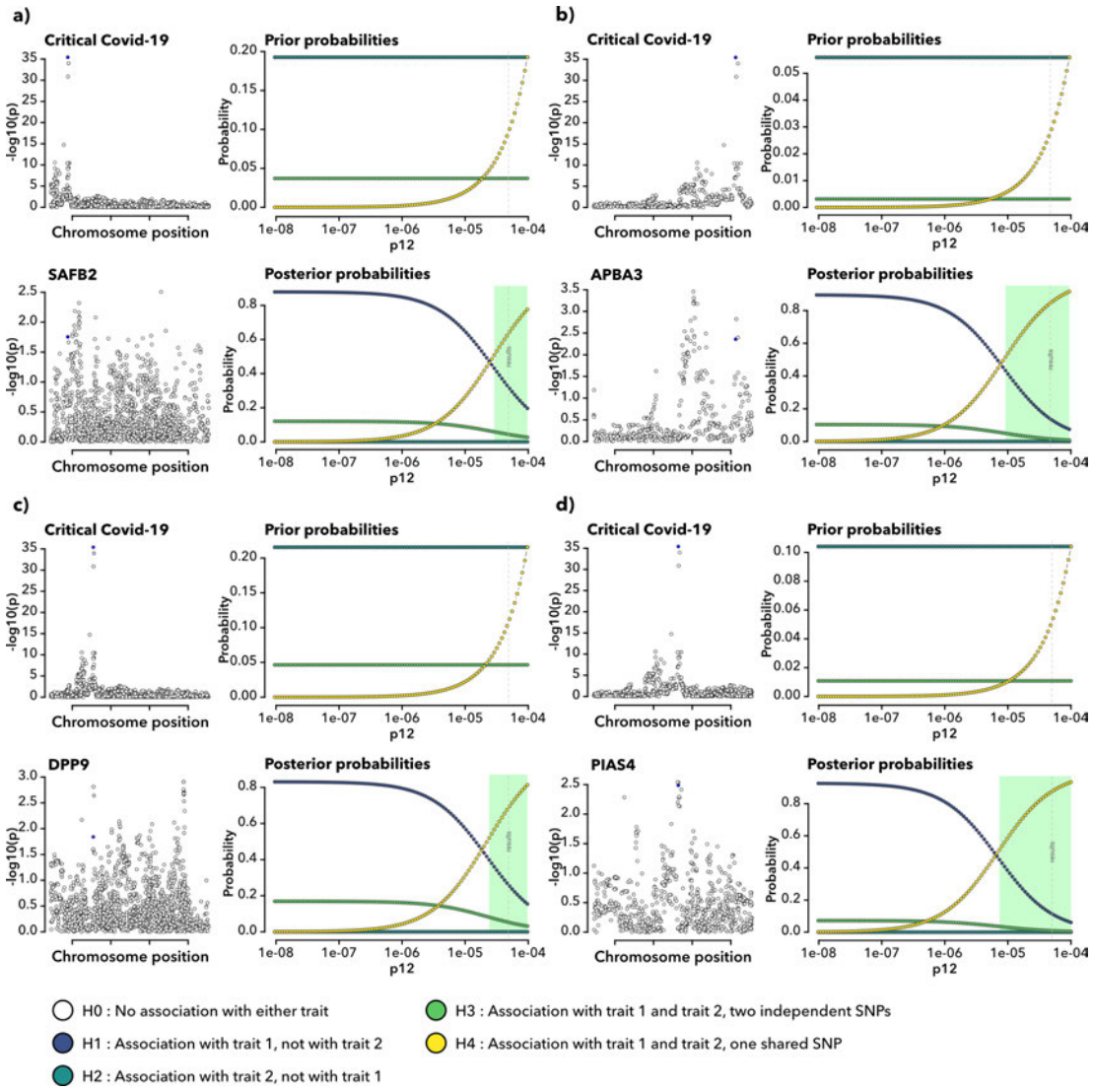
Supplementary Figure 14: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and DAPK3 expression in CD4 effector memory/TEMRA cells on chromosome 19 region 4348967-5811852, b) critical Covid-19 and SH3GL1 expression in CD4 effector memory/TEMRA cells on chromosome 19 region 4348967-5811852, c) critical Covid-19 and C19orf70 expression in CD4 naïve/central memory T cells on chromosome 19 region 4348967-5811852 and d) critical Covid-19 and SAFB2 expression in CD4 naïve/central memory T cells on chromosome 19 region 4348967-5811852.



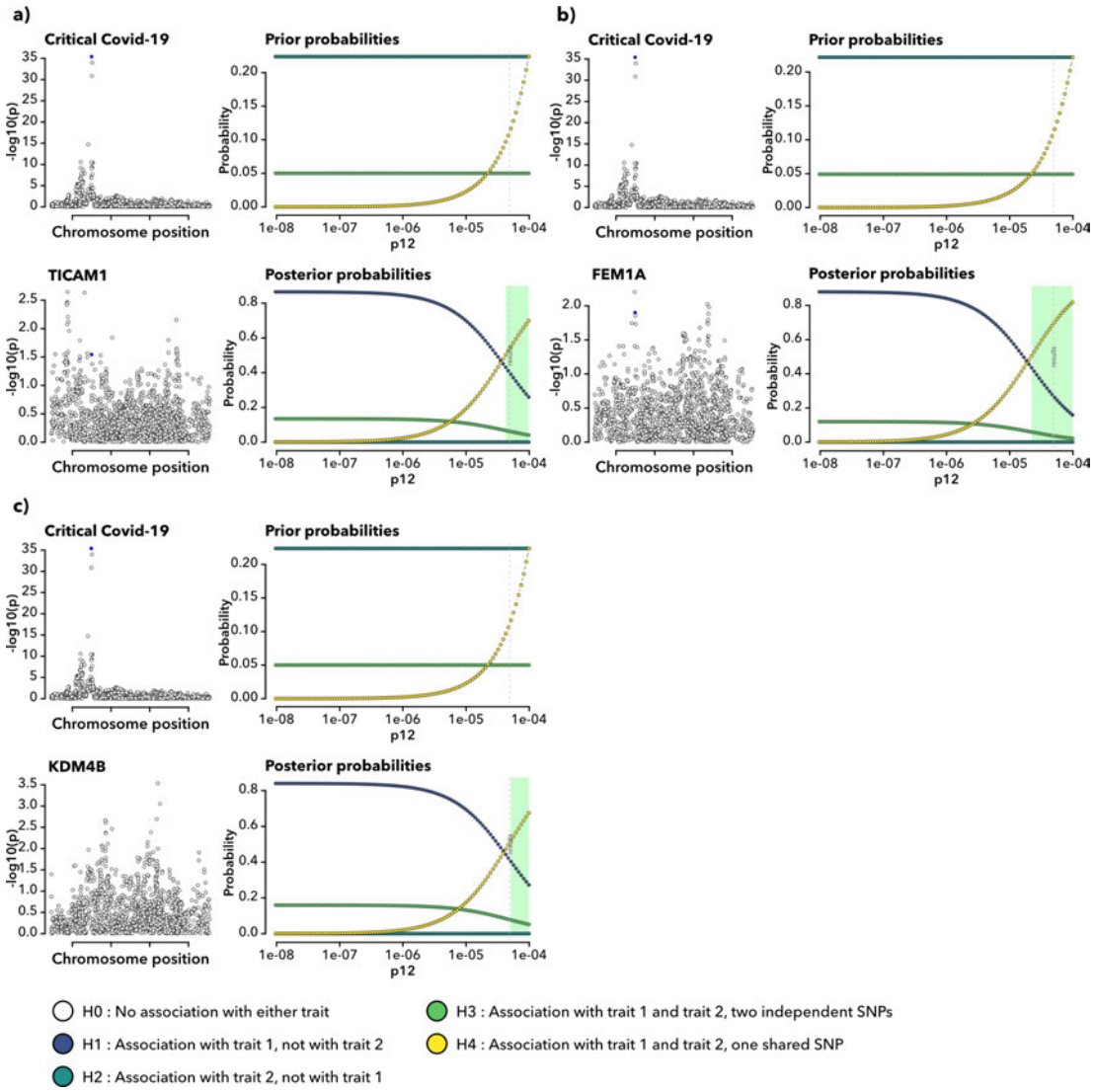
Supplementary Figure 15: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and TNFAIP8L1 expression in CD4 SOX4 T cells on chromosome 19 region 4348967-5811852, b) critical Covid-19 and C19orf70 expression in CD8 effector memory T cells on chromosome 19 region 4348967-5811852, c) critical Covid-19 and STAP2 expression in CD8 effector memory T cells on chromosome 19 region 4348967-5811852 and d) critical Covid-19 and TJP3 expression in CD8 effector memory T cells on chromosome 19 region 4348967-5811852.



Supplementary Figure 16: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and APBA3 expression in CD8 naïve/central memory T cells on chromosome 19 region 4348967-5811852, b) critical Covid-19 and SIRT6 expression in CD8 naïve/central memory T cells on chromosome 19 region 4348967-5811852, c) critical Covid-19 and C19orf70 expression in classic monocytes on chromosome 19 region 4348967-5811852 and d) critical Covid-19 and AC005523.3 expression in natural killer cells on chromosome 19 region 4348967-5811852.



Supplementary Figure 17: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and SAFB2 expression in natural killer cells on chromosome 19 region 4348967-5811852, b) critical Covid-19 and APBA3 expression in natural killer recruiting cells on chromosome 19 region 4348967-5811852, c) critical Covid-19 and DPP9 expression in natural killer recruiting cells on chromosome 19 region 4348967-5811852 and d) critical Covid-19 and PIAS4 expression in natural killer recruiting cells on chromosome 19 region 4348967-5811852.



Supplementary Figure 18: Sensitivity plots for GWAS and eQTL signal colocalizations. a) critical Covid-19 and TICAM1 expression in natural killer recruiting cells on chromosome 19 region 4348967-5811852, b) critical Covid-19 and FEM1A expression in plasma cells on chromosome 19 region 4348967-5811852 and c) critical Covid-19 and KDM4B expression in plasma cells on chromosome 19 region 4348967-5811852.

Bibliography

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

2. Holland, D. *et al.* Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model. *PLOS Genetics* **16**, e1008612 (2020).

3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal of Human Genetics* **101**, 5–22 (2017).

4. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature Genetics* **47**, 856–860 (2015).

5. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).

6. Dendrou, C. A. *et al.* Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Science translational medicine* **8**, 363ra149 (2016).

7. Voight, B. F. *et al.* Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *Lancet* **380**, 572–580 (2012).

8. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47**, D1005–D1012 (2019).

9. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* **4**, 587–597 (2003).

10. Nica, A. C. & Dermitzakis, E. T. Using gene expression to investigate the genetic basis of complex disorders. *Human Molecular Genetics* **17**, R129–R134 (2008).

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

12. McMahon, A. *et al.* Sequencing-based genome-wide association studies reporting standards. *Cell Genomics* **1**, 100005 (2021).

13. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).

14. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (New York, N.Y.)* **337**, 1190–1195 (2012).

15. Mignone, F., Gissi, C., Liuni, S. & Pesole, G. Untranslated regions of mRNAs. *Genome Biology* **3**, reviews0004.1–reviews0004.10 (2002).

16. Jansen, R. P. mRNA localization: Message on the move. *Nature Reviews. Molecular Cell Biology* **2**, 247–256 (2001).

17. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nature Reviews Genetics* **17**, 19–32 (2016).

18. Amariuta, T. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *American Journal of Human Genetics* **104**, 879–895 (2019).

19. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics* **102**, 717–730 (2018).

20. Paaby, A. B. & Rockman, M. V. The many faces of pleiotropy. *Trends in Genetics* **29**, 66–73 (2013).

21. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics* **51**, 1339–1348 (2019).

22. Visscher, P. M. & Yang, J. A plethora of pleiotropy across complex traits. *Nature Genetics* **48**, 707–708 (2016).

23. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **50**, 229–237 (2018).

24. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications* **12**, 764 (2021).

25. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics* **14**, 483–495 (2013).

26. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: From theory to practice. *Nature Reviews Genetics* **20**, 567–581 (2019).

27. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry* **72**, 291–336 (2003).

28. Gommans, W. M., Mullen, S. P. & Maas, S. RNA editing: A driving force for adaptive evolution? *BioEssays : news and reviews in molecular, cellular and developmental biology* **31**, 1137 (2009).

29. J, H. Seven types of pleiotropy. *The International journal of developmental biology* **42**, (1998).

30. Rieck, M. *et al.* Genetic Variation in PTPN22 Corresponds to Altered Function of T and B Lymphocytes1. *The Journal of Immunology* **179**, 4704–4710 (2007).

31. Menard, L. *et al.* The PTPN22 allele encoding an R620W variant interferes with the removal of developing autoreactive B cells in humans. *The Journal of Clinical Investigation* **121**, 3635–3644 (2011).

32. Major Lipids, Apolipoproteins, and Risk of Vascular Disease. *JAMA : the journal of the American Medical Association* **302**, 1993–2000 (2009).

33. Byars, S. G. & Voskarides, K. Antagonistic Pleiotropy in Human Disease. *Journal of Molecular Evolution* **88**, 12–25 (2020).

34. Elguero, E. *et al.* Malaria continues to select for sickle cell trait in Central Africa. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 7051–7054 (2015).

35. Inusa, B. P. D. *et al.* Sickle Cell Disease—Genetics, Pathophysiology, Clinical Presentation and Treatment. *International Journal of Neonatal Screening* **5**, 20 (2019).

36. Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 1–21 (2021).

37. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **186**, 1026–1034 (2017).

38. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics* **53**, 663–671 (2021).

39. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* **20**, 467–484 (2019).

40. Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11**, 724 (2010).

41. Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D. & Province, M. A. Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology* **34**, 100–105 (2010).

42. Sesia, M., Bates, S., Candès, E., Marchini, J. & Sabatti, C. False discovery rate control in genome-wide association studies with population structure. *Proceedings of the National Academy of Sciences* **118**, e2105841118 (2021).

43. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* **10**, 681–690 (2009).

44. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).

45. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics* **17**, e1009440 (2021).

46. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

47. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).

48. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nature Genetics* **46**, 220–224 (2014).

49.

Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241–251 (2009).

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

51. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–569 (2010).

52. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *The American Journal of Human Genetics* **99**, 139–153 (2016).

53. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512–517 (2004).

54. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. Conditions Under Which Genome-Wide Association Studies Will be Positively Misleading. *Genetics* **186**, 1045–1052 (2010).

55. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

56. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genetics* **6**, (2010).

57. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015).

58. Aguet, F. *et al.* Molecular quantitative trait loci. *Nature Reviews Methods Primers* **3**, 1–22 (2023).

59. THE GTEX CONSORTIUM. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

60. Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature Genetics* **53**, 994–1005 (2021).

61. Sanderson, E. *et al.* Mendelian randomization. *Nature Reviews Methods Primers* **2**, 1–21 (2022).

62. Glymour, M. M., Tchetgen Tchetgen, E. J. & Robins, J. M. Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *American Journal of Epidemiology* **175**, 332–339 (2012).

63. Keele, L., Zhao, Q., Kelz, R. R. & Small, D. Falsification Tests for Instrumental Variable Designs With an Application to Tendency to Operate. *Medical Care* **57**, 167 (2019).

64. Taylor, A. E. *et al.* Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates. *Economics and Human Biology* **13**, 99–106 (2014).

65. Hartwig, F. P., Davies, N. M. & Davey Smith, G. Bias in Mendelian randomization due to assortative mating. *Genetic Epidemiology* **42**, 608–620 (2018).

66. Brumpton, B. *et al.* Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications* **11**, 3519 (2020).

67. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *American Journal of Human Genetics* **109**, 767–782 (2022).

68. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091–1098 (2015).

69. Mai, J., Lu, M., Gao, Q., Zeng, J. & Xiao, J. Transcriptome-wide association studies: Recent advances in methods, applications and available databases. *Communications Biology* **6**, 1–10 (2023).

70. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

71. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics* **48**, 481–487 (2016).

72. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *American Journal of Human Genetics* **99**, 1245–1260 (2016).

73. Chun, S. *et al.* Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nature Genetics* **49**, 600–605 (2017).

74. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends in Genetics* **37**, 109–124 (2021).

75. Connally, N. J. *et al.* The missing link between genetic association and regulatory function. *eLife* **11**, e74970 (2022).

76. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics* **55**, 1866–1875 (2023).

77. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics* **52**, 626–633 (2020).

78. Hukku, A. *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: Promise and limitations. *The American Journal of Human Genetics* **108**, 25–35 (2021).

79. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* **49**, 139–145 (2017).

80. Kim-Hellmuth, S. *et al.* Cell type–specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).

81. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).

82. Ota, M. *et al.* Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021.e17 (2021).

83. Mu, Z. *et al.* The impact of cell type and context-dependent regulatory variants on human immune traits. *Genome Biology* **22**, 122 (2021).

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

85. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**, 1300–1310 (2021).

86. Pierce, B. L. *et al.* Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 Sou th Asians. *PLOS Genetics* **10**, e1004818 (2014).

87. Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35**, i108–i116 (2019).

88. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods* **18**, 1196–1203 (2021).

89. Morris, J. A. *et al.* Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **380**, eadh7699 (2023).

90. Van de Sande, B. *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery* **22**, 496–520 (2023).

91. Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics* **11**, (2020).

92. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. *The American Journal of Human Genetics* **89**, 496–506 (2011).

93. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *The American Journal of Human Genetics* **101**, 686–699 (2017).

94. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods* **10**, 1213–1218 (2013).

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

96. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

97.

Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

98. Yue, X. *et al.* Simultaneous profiling of histone modifications and DNA methylation via nanopore sequencing. *Nature Communications* **13**, 7939 (2022).

99. Consortium, T. E. P. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57 (2012).

100. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).

101. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics* **45**, 10.1038/ng.2504 (2013).

102. Schmidt, E. M. *et al.* GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601–2606 (2015).

103. Soskic, B. *et al.* Chromatin activity at GWAS loci identifies T cell states driving complex immune diseases. *Nature genetics* **51**, 1486–1493 (2019).

104. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *American Journal of Human Genetics* **97**, 139–152 (2015).

105. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *American Journal of Human Genetics* **94**, 559–573 (2014).

106. Iotchkova, V. *et al.* GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nature genetics* **51**, 343–353 (2019).

107. Yuan, Y., Jiao, B., Qu, L., Yang, D. & Liu, R. The development of COVID-19 treatment. *Frontiers in Immunology* **14**, 1125246 (2023).

108. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* **324**, 782–793 (2020).

109. Tsai, P.-H. *et al.* Clinical manifestation and disease progression in COVID-19 infection. *Journal of the Chinese Medical Association* **84**, 3 (2021).

110. Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: A systematic analysis. *The Lancet* **399**, 1469–1488 (2022).

111. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Prospective observational cohort study. *BMJ (Clinical research ed.)* **369**, m1985 (2020).

112. Dorward, D. A. *et al.* Tissue-Specific Immunopathology in Fatal COVID-19. *American Journal of Respiratory and Critical Care Medicine* **203**, 192–201 (2021).

113.

Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet (London, England)* **395**, 497–506 (2020).

114. Merad, M., Blish, C. A., Sallusto, F. & Iwasaki, A. The immunology and immunopathology of COVID-19. *Science* **375**, 1122–1127 (2022).

115. Hou, Y. J. *et al.* SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. *Cell* **182**, 429–446.e14 (2020).

116. Katsura, H. *et al.* Human Lung Stem Cell-Based Alveolospheres Provide Insights into SARS-CoV-2-Mediated Interferon Responses and Pneumocyte Dysfunction. *Cell Stem Cell* **27**, 890–904.e8 (2020).

117. Gallo, O., Locatello, L. G., Mazzoni, A., Novelli, L. & Annunziato, F. The central role of the nasal microenvironment in the transmission, modulation, and clinical progression of SARS-CoV-2 infection. *Mucosal Immunology* **14**, 305–316 (2021).

118. Mick, E. *et al.* Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. *Nature Communications* **11**, 5854 (2020).

119. Fajnzylber, J. *et al.* SARS-CoV-2 viral load is associated with increased disease severity and mortality. *Nature Communications* **11**, 5493 (2020).

120. Zhang, Q. *et al.* Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. *Science (New York, N.y.)* **370**, eabd4570 (2020).

121. Bastard, P. *et al.* Autoantibodies against type I IFNs in patients with life-threatening COVID-19. *Science (New York, N.Y.)* **370**, eabd4585 (2020).

122. Zhou, R. *et al.* Acute SARS-CoV-2 Infection Impairs Dendritic Cell and T Cell Responses. *Immunity* **53**, 864–877.e5 (2020).

123. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science (New York, N.Y.)* **369**, eabc8511 (2020).

124. Kaneko, N. *et al.* Loss of Bcl-6-Expressing T Follicular Helper Cells and Germinal Centers in COVID-19. *Cell* **183**, 143–157.e13 (2020).

125. Russell, C. D. *et al.* Tissue Proteomic Analysis Identifies Mechanisms and Stages of Immunopathology in Fatal COVID-19. *American Journal of Respiratory Cell and Molecular Biology* **66**, 196–205.

126. Wang, Y. & Perlman, S. COVID-19: Inflammatory Profile. *Annual Review of Medicine* **73**, 65–80 (2022).

127. Merad, M. & Martin, J. C. Pathological inflammation in patients with COVID-19: A key role for monocytes and macrophages. *Nature Reviews Immunology* **20**, 355–362 (2020).

128. Rendeiro, A. F. *et al.* The spatial landscape of lung pathology during COVID-19 progression. *Nature* **593**, 564–569 (2021).

129.

Sefik, E. *et al.* A humanized mouse model of chronic COVID-19. *Nature Biotechnology* **40**, 906–920 (2022).

130. Upadhya, S., Rehman, J., Malik, A. B. & Chen, S. Mechanisms of Lung Injury Induced by SARS-CoV-2 Infection. *Physiology* **37**, 88–100 (2022).

131. Pairo-Castineira, E. *et al.* GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* **617**, 764–768 (2023).

132. Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, multimorbidity and COVID-19. *Nature Medicine* **29**, 334–343 (2023).

133. Sørensen, T. I., Nielsen, G. G., Andersen, P. K. & Teasdale, T. W. Genetic and environmental influences on premature death in adult adoptees. *The New England Journal of Medicine* **318**, 727–732 (1988).

134. Kousathanas, A. *et al.* Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022).

135. Ferreira, L. C., Gomes, C. E. M., Rodrigues-Neto, J. F. & Jeronimo, S. M. B. Genome-wide association studies of COVID-19: Connecting the dots. *Infection, Genetics and Evolution* **106**, 105379 (2022).

136. Abani, O. *et al.* Baricitinib in patients admitted to hospital with COVID-19 (RECOVERY): A randomised, controlled, open-label, platform trial and updated meta-analysis. *The Lancet* **400**, 359–368 (2022).

137. Ong, E. Z. *et al.* A Dynamic Immune Response Shapes COVID-19 Progression. *Cell Host & Microbe* **27**, 879–882.e2 (2020).

138. Anka, A. U. *et al.* Coronavirus disease 2019 (COVID-19): An overview of the immunopathology, serological diagnosis and management. *Scandinavian Journal of Immunology* **93**, e12998 (2021).

139. Cotsapas, C. *et al.* Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLOS Genetics* **7**, e1002254 (2011).

140. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **20**, 273–286 (2019).

141. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **47**, 20–33 (2016).

142. Kinch, M. S. & Griesenauer, R. H. 2017 in review: FDA approvals of new molecular entities. *Drug Discovery Today* **23**, 1469–1473 (2018).

143. Rm, P., Em, S. & D, A. Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery* **12**, (2013).

144. Ghofrani, H. A., Osterloh, I. H. & Grimminger, F. Sildenafil: From angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery* **5**, 689–702 (2006).

145. Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery* **3**, 673–683 (2004).

146. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genetics* **15**, e1008489 (2019).

147. Lau, A. & So, H.-C. Turning genome-wide association study findings into opportunities for drug repositioning. *Computational and Structural Biotechnology Journal* **18**, 1639–1650 (2020).

148. Nabirotchkin, S. *et al.* Next-generation drug repurposing using human genetics and network biology. *Current Opinion in Pharmacology* **51**, 78–92 (2020).

149. Heinig, M. Using Gene Expression to Annotate Cardiovascular GWAS Loci. *Frontiers in Cardiovascular Medicine* **5**, 59 (2018).

150. Pan, Y., Cheng, T., Wang, Y. & Bryant, S. H. Pathway Analysis for Drug Repositioning Based on Public Database Mining. *Journal of Chemical Information and Modeling* **54**, 407–418 (2014).

151. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nature Reviews Genetics* **12**, 56–68 (2011).

152. Mullen, J., Cockell, S. J., Woollard, P. & Wipat, A. An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. *PLoS ONE* **11**, e0155811 (2016).

153. Sadegh, S. *et al.* Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nature Communications* **12**, 6848 (2021).

154. Wang, Z., Zhou, M. & Arnold, C. Toward heterogeneous information fusion: Bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics* **36**, i525–i533 (2020).

155. Li, H. *et al.* Novel disease syndromes unveiled by integrative multiscale network analysis of diseases sharing molecular effectors and comorbidities. *BMC Medical Genomics* **11**, 112 (2018).

156. International Multiple Sclerosis Genetics Consortium. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. *American Journal of Human Genetics* **92**, 854–865 (2013).

157. Guo, H., Li, T. & Wen, H. Identifying shared genetic loci between coronavirus disease 2019 and cardiovascular diseases based on cross-trait meta-analysis. *Frontiers in Microbiology* **13**, (2022).

158. Zuber, V. *et al.* Leveraging Genetic Data to Elucidate the Relationship Between COVID-19 and Ischemic Stroke. *Journal of the American Heart Association* **10**, e022433 (2021).

159. Chang, X. *et al.* Genetic correlations between COVID-19 and a variety of traits and diseases. *The Innovation* **2**, (2021).

160. Allen, R. J. *et al.* Genetic overlap between idiopathic pulmonary fibrosis and COVID-19. *European Respiratory Journal* **60**, (2022).

161. Caulfield, M. *et al.* The 100,000 Genomes Project Protocol. (2017).

162. GWAS Harmonization And Imputation. *GitHub*.

163. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Briefings in Bioinformatics* **14**, 144–161 (2013).

164. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

165. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.

166. UK Biobank. *Neale lab*.

167. Lyons, P. A. *et al.* Genome-wide association study of eosinophilic granulomatosis with polyangiitis reveals genomic loci stratified by ANCA status. *Nature Communications* **10**, 5120 (2019).

168. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research* **51**, D977–D985 (2023).

169. Allen, R. J. *et al.* Genome-Wide Association Study of Susceptibility to Idiopathic Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine* **201**, 564–574 (2020).

170. Ferreira, M. A. R. *et al.* Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. *The American Journal of Human Genetics* **104**, 665–684 (2019).

171. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics* **47**, 1449–1456 (2015).

172. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics* **49**, 256–261 (2017).

173. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics* **47**, 979–986 (2015).

174. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature genetics* **49**, 1752 (2017).

175. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics* **47**, 1457–1464 (2015).

176. Ochoa, D. *et al.* The next-generation Open Targets Platform: Reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353–D1359 (2023).

177. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2019).

178. ICD-10 Version:2016.

179. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* **12**, e1001779 (2015).

180. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

181. Collaborative data platform and canvas Observable.

182. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across human complex traits. *Nature Genetics* **52**, 859–864 (2020).

183. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47**, 291–295 (2015).

184. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236–1241 (2015).

185. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).

186. Reimand, J. *et al.* G:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research* **44**, W83–W89 (2016).

187. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

188. Kuleshov, M. V. *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90–W97 (2016).

189. Gerstner, N. *et al.* GeneTrail 3: Advanced high-throughput enrichment analysis. *Nucleic Acids Research* **48**, W515–W520 (2020).

190. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

191. Lin, H. *et al.* RegSNPs-intron: A computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biology* **20**, 254 (2019).

192. Liu, C.-J. *et al.* miRNASNP-v3: A comprehensive database for SNPs and disease-related variations in miRNAs and miRNA targets. *Nucleic Acids Research* **49**, D1276–D1281 (2021).

193. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Research* **9**, 677–679 (1999).

194. Szklarczyk, D. *et al.* The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research* **51**, D638–D646 (2023).

195. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1–26 (1979).

196. Mooney, C. Z., Duval, R. D. & Duvall, R. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. (SAGE, 1993).

197. Šimkovic, M. & Träuble, B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLoS ONE* **14**, e0220889 (2019).

198. Laplace, P.-S. Sur les approximations des formules qui sont fonctions de tres grands nombres et sur leur application aux probabilites. *Œuvres complètes* **12**, 301–345 (1810).

199. STINE, R. An Introduction to Bootstrap Methods: Examples and Ideas. *Sociological Methods & Research* **18**, 243–291 (1989).

200. Altman, D. G. & Bland, J. M. How to obtain the confidence interval from a P value. *BMJ* **343**, d2090 (2011).

201. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

202. Varadi, M. *et al.* AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).

203. Zheng, J. *et al.* LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

204. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nature Genetics* **50**, 1728–1734 (2018).

205. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).

206. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018).

207. Lu, Q. *et al.* A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. *American Journal of Human Genetics* **101**, 939–964 (2017).

208. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* **50**, 693–698 (2018).

209. Lin, D.-Y. & Sullivan, P. F. Meta-Analysis of Genome-wide Association Studies with Overlapping Subjects. *The American Journal of Human Genetics* **85**, 862–872 (2009).

210. Yengo, L., Yang, J. & Visscher, P. M. Expectation of the intercept from bivariate LD score regression in the presence of population stratification. 310565 (2018) doi:10.1101/310565.

211.

Wray, N. R., Lee, S. H. & Kendler, K. S. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *European Journal of Human Genetics* **20**, 668–674 (2012).

212. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: When selection bias can substantially influence observed associations. *International Journal of Epidemiology* **47**, 226–235 (2018).

213. Vidal-Perez, R. *et al.* Cardiovascular disease and COVID-19, a deadly combination: A review about direct and indirect impact of a pandemic. *World Journal of Clinical Cases* **10**, 9556–9572 (2022).

214. Richardson, S. *et al.* Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *JAMA* **323**, 2052–2059 (2020).

215. Li, B. *et al.* Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. *Clinical Research in Cardiology* **109**, 531–538 (2020).

216. Burger, A. L. *et al.* Direct cardiovascular complications and indirect collateral damage during the COVID-19 pandemic. *Wiener Klinische Wochenschrift* **133**, 1289–1297 (2021).

217. Rusu, I., Turlacu, M. & Micheu, M. M. Acute myocardial injury in patients with COVID-19: Possible mechanisms and clinical implications. *World Journal of Clinical Cases* **10**, 762–776 (2022).

218. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020).

219. Donoghue, M. *et al.* A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. *Circulation Research* **87**, E1–9 (2000).

220. Turner, A. J., Hiscox, J. A. & Hooper, N. M. ACE2: From vasopeptidase to SARS virus receptor. *Trends in Pharmacological Sciences* **25**, 291–294 (2004).

221. Chung, M. K. *et al.* COVID-19 and Cardiovascular Disease. *Circulation Research* **128**, 1214–1236 (2021).

222. Yang, Z. *et al.* Genetic Landscape of the ACE2 Coronavirus Receptor. *Circulation* **145**, 1398 (2022).

223. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).

224. Wang, L. *et al.* An atlas connecting shared genetic architecture of human diseases and molecular phenotypes provides insight into COVID-19 susceptibility. *Genome Medicine* **13**, 83 (2021).

225. Pietzner, M. *et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nature Communications* **11**, 6397 (2020).

226.

Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Research* **50**, D988–D995 (2022).

227. Li, D. *et al.* A Pleiotropic Missense Variant in SLC39A8 Is Associated With Crohn's Disease and Human Gut Microbiome Composition. *Gastroenterology* **151**, 724–732 (2016).

228. Collij, V. *et al.* SLC39A8 missense variant is associated with Crohn's disease but does not have a major impact on gut microbiome composition in healthy subjects. *PLOS ONE* **14**, e0211328 (2019).

229. Kichaev, G. *et al.* Leveraging Polygenic Functional Enrichment to Improve GWAS Power. *The American Journal of Human Genetics* **104**, 65–75 (2019).

230. Wain, L. V. *et al.* Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature Genetics* **43**, 1005–1011 (2011).

231. Tachmazidou, I. *et al.* Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data. *Nature Genetics* **51**, 230–236 (2019).

232. Vujkovic, M. *et al.* 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**, 680–691 (2020).

233. The COVID-19 Host Genetics Initiative *et al.* A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7–E26 (2023).

234. Jeong, J. & Eide, D. J. The SLC39 family of zinc transporters. *Molecular Aspects of Medicine* **34**, 612–619 (2013).

235. He, L. *et al.* ZIP8, Member of the Solute-Carrier-39 (SLC39) Metal-Transporter Family: Characterization of Transporter Properties. *Molecular Pharmacology* **70**, 171–180 (2006).

236. Wang, C.-Y. *et al.* ZIP8 Is an Iron and Zinc Transporter Whose Cell-surface Expression Is Up-regulated by Cellular Iron Loading*. *Journal of Biological Chemistry* **287**, 34032–34043 (2012).

237. Zhang, T., Sui, D. & Hu, J. Structural insights of ZIP4 extracellular domain critical for optimal zinc transport. *Nature Communications* **7**, 11979 (2016).

238. Ng, E. *et al.* Genome-wide association study of toxic metals and trace elements reveals novel associations. *Human Molecular Genetics* **24**, 4739–4745 (2015).

239. Sunuwar, L. *et al.* Pleiotropic ZIP8 A391T implicates abnormal manganese homeostasis in complex human disease. *JCI Insight* **5**, e140978.

240. Fujishiro, H., Miyamoto, S., Sumi, D., Kambe, T. & Himeno, S. Effects of individual amino acid mutations of zinc transporter ZIP8 on manganese- and cadmium-transporting activity. *Biochemical and Biophysical Research Communications* **616**, 26–32 (2022).

241.

Verouti, S. N. *et al.* The Allelic Variant A391T of Metal Ion Transporter ZIP8 (SLC39A8) Leads to Hypotension and Enhanced Insulin Resistance. *Frontiers in Physiology* **13**, (2022).

242. Mealer, R. G. *et al.* The schizophrenia risk locus in SLC39A8 alters brain metal transport and plasma glycosylation. *Scientific Reports* **10**, 13162 (2020).

243. Haller, G. *et al.* A missense variant in SLC39A8 is associated with severe idiopathic scoliosis. *Nature Communications* **9**, 4171 (2018).

244. Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16**, 29R–37R (2006).

245. Gong, Y., Qin, S., Dai, L. & Tian, Z. The glycosylation in SARS-CoV-2 and its receptor ACE2. *Signal Transduction and Targeted Therapy* **6**, 1–24 (2021).

246. The Human Protein Atlas.

247. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Science Advances* **7**, eabh2169 (2021).

248. Mulay, A. *et al.* SARS-CoV-2 infection of primary human lung epithelium for COVID-19 modeling and drug discovery. *Cell Reports* **35**, (2021).

249. Carsana, L. *et al.* Pulmonary post-mortem findings in a series of COVID-19 cases from northern Italy: A two-centre descriptive study. *The Lancet Infectious Diseases* **20**, 1135–1140 (2020).

250. Liu, M.-J. *et al.* ZIP8 Regulates Host Defense through Zinc-Mediated Inhibition of NF-κB. *Cell Reports* **3**, 386–400 (2013).

251. Chen, L.-F. & Greene, W. C. Shaping the nuclear action of NF-κB. *Nature Reviews Molecular Cell Biology* **5**, 392–401 (2004).

252. Montazersaheb, S. *et al.* COVID-19 infection: An overview on cytokine storm and related interventions. *Virology Journal* **19**, 92 (2022).

253. Boudreault, F. *et al.* Zinc deficiency primes the lung for ventilator-induced injury. *JCI Insight* **2**, (2017).

254. Besecker, B. *et al.* The human zinc transporter SLC39A8 (Zip8) is critical in zinc-mediated cytoprotection in lung epithelia. *American Journal of Physiology-Lung Cellular and Molecular Physiology* **294**, L1127–L1136 (2008).

255. Pal, A. *et al.* Zinc and COVID-19: Basis of Current Clinical Trials. *Biological Trace Element Research* **199**, 2882–2892 (2021).

256. Firouzi, S. *et al.* The effect of Vitamin C and Zn supplementation on the immune system and clinical outcomes in COVID-19 patients. *Clinical Nutrition Open Science* **44**, 144–154 (2022).

257. Park, J. H. *et al.* SLC39A8 deficiency: Biochemical correction and major clinical improvement by manganese therapy. *Genetics in Medicine* **20**, 259–268 (2018).

258. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384**, 693–704 (2021).

259. Kandasamy, M. NF-κB signalling as a pharmacological target in COVID-19: Potential roles for IKKβ inhibitors. *Naunyn-Schmiedeberg's Archives of Pharmacology* **394**, 561–567 (2021).

260. Sharma, A. Inferring molecular mechanisms of dexamethasone therapy in severe COVID-19 from existing transcriptomic data. *Gene* **788**, 145665 (2021).

261. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214–1231.e11 (2020).

262. Zhu, Z. *et al.* Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. *Journal of Allergy and Clinical Immunology* **145**, 537–549 (2020).

263. Burton, P. R. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

264. Thorsby, E. A short history of HLA. *Tissue Antigens* **74**, 101–116 (2009).

265. Kennedy, A. E., Ozbek, U. & Dorak, M. T. What has GWAS done for HLA and disease associations? *International Journal of Immunogenetics* **44**, 195–211 (2017).

266. Naito, T. & Okada, Y. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Seminars in Immunopathology* **44**, 15–28 (2022).

267. Douglas, A. G. L. & Wood, M. J. A. RNA splicing: Disease and therapy. *Briefings in Functional Genomics* **10**, 151–164 (2011).

268. Rose, A. B. Introns as Gene Regulators: A Brick on the Accelerator. *Frontiers in Genetics* **9**, (2019).

269. Zhang, H. *et al.* Identification of novel dipeptidyl peptidase 9 substrates by two-dimensional differential in-gel electrophoresis. *The FEBS Journal* **282**, 3737–3757 (2015).

270. Geiss-Friedlander, R. *et al.* The Cytoplasmic Peptidase DPP9 Is Rate-limiting for Degradation of Proline-containing Peptides *. *Journal of Biological Chemistry* **284**, 27211–27219 (2009).

271. Griswold, A. R. *et al.* DPP9's Enzymatic Activity and Not Its Binding to CARD8 Inhibits Inflammasome Activation. *ACS Chemical Biology* **14**, 2424–2429 (2019).

272. Langefeld, C. D. *et al.* Transancestral mapping and genetic load in systemic lupus erythematosus. *Nature Communications* **8**, 16021 (2017).

273. Tsoi, L. C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics* **44**, 1341–1348 (2012).

274. Ha, E., Bae, S.-C. & Kim, K. Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Annals of the Rheumatic Diseases* **80**, 558–565 (2021).

275. Wang, Y. *et al.* COVID-19 and systemic lupus erythematosus genetics: A balance between autoimmune disease risk and protection against infection. *PLOS Genetics* **18**, e1010253 (2022).

276. Couturier, N. *et al.* Tyrosine kinase 2 variant influences T lymphocyte polarization and multiple sclerosis susceptibility. *Brain* **134**, 693–703 (2011).

277. Rani, M. R. *et al.* Catalytically active TYK2 is essential for interferon-beta-mediated phosphorylation of STAT3 and interferon-alpha receptor-1 (IFNAR-1) but not for activation of phosphoinositol 3-kinase. *The Journal of biological chemistry* **274**, 32507–32511 (1999).

278. Nguyen, D.-T. *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research* **45**, D995–D1002 (2017).

279. Al-Salama, Z. T. & Scott, L. J. Baricitinib: A Review in Rheumatoid Arthritis. *Drugs* **78**, 761–772 (2018).

280. Papp, K. A. *et al.* A randomized phase 2b trial of baricitinib, an oral Janus kinase (JAK) 1/JAK2 inhibitor, in patients with moderate-to-severe psoriasis. *British Journal of Dermatology* **174**, 1266–1276 (2016).

281. Petri, M. *et al.* Baricitinib for systemic lupus erythematosus: A double-blind, randomised, placebo-controlled, phase 3 trial (SLE-BRAVE-II). *The Lancet* **401**, 1011–1019 (2023).

282. Nagaoka, S. I. *et al.* ZGLP1 is a determinant for the oogenic fate in mice. *Science* **367**, eaaw4115 (2020).

283. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature reviews. Genetics* **19**, 491–504 (2018).

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

285. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21**, 2815–2824 (2012).

286. Hrdlickova, B., de Almeida, R. C., Borek, Z. & Withoff, S. Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RN As in disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1842**, 1910–1922 (2014).

287. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: Identifying the splicing spoilers. *Nature Reviews Genetics* **5**, 389–396 (2004).

288. Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: Functional insights and role in human disease. *Current Opinion in Genetics & Development* **33**, 71–76 (2015).

289. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

290. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols* **14**, 482–517 (2019).

291. Martens, M. *et al.* WikiPathways: Connecting communities. *Nucleic Acids Research* **49**, D613–D621 (2021).

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

293. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research* **47**, D330–D338 (2019).

294. Mastropasqua, L. *et al.* Transcriptomic analysis revealed increased expression of genes involved in keratinization in the tears of COVID-19 patients. *Scientific Reports* **11**, 19817 (2021).

295. Camblor, D. G. *et al.* Genetic variants in the NF-κB signaling pathway (NFKB1, NFKBIA, NFKBIZ) and risk of critical outcome among COVID-19 patients. *Human Immunology* **83**, 613–617 (2022).

296. Leng, L. *et al.* Spatial region-resolved proteome map reveals mechanism of COVID-19-associated heart injury. *Cell Reports* **39**, 110955 (2022).

297. Iwabuchi, S. *et al.* Immune Cells Profiles in the Different Sites of COVID-19-Affected Lung Lobes in a Single Patient. *Frontiers in Medicine* **9**, 841170 (2022).

298. Corpas, M. *et al.* Genetic signature detected in T cell receptors from patients with severe COVID-19. *iScience* **26**, 107735 (2023).

299. Tanaka, Y. *et al.* Dynamic changes in gene-to-gene regulatory networks in response to SARS-CoV-2 infection. *Scientific Reports* **11**, 11241 (2021).

300. Xu, G. *et al.* Multiomics approach reveals the ubiquitination-specific processes hijacked by SARS-CoV-2. *Signal Transduction and Targeted Therapy* **7**, 312 (2022).

301. Gerassimovich, Y. A. *et al.* Proximity-dependent biotinylation detects associations between SARS coronavirus nonstructural protein 1 and stress granule–associated proteins. *The Journal of Biological Chemistry* **297**, 101399 (2021).

302. Ehrenreich, H. *et al.* Erythropoietin as candidate for supportive treatment of severe COVID-19. *Molecular Medicine* **26**, 58 (2020).

303. Papadopoulos, K. I., Sutheesophon, W., Manipalviratn, S. & Aw, T.-C. Age and genotype dependent erythropoietin protection in COVID-19. *World Journal of Stem Cells* **13**, 1513–1529 (2021).

304. Singh, S. P. *et al.* Mitochondrial Modulations, Autophagy Pathways Shifts in Viral Infections: Consequences of COVID-19. *International Journal of Molecular Sciences* **22**, 8180 (2021).

305. Barrett, T. J. *et al.* Platelets amplify endotheliopathy in COVID-19. *Science Advances* **7**, eabh2434 (2021).

306. Vastrad, B., Vastrad, C. & Tengli, A. Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. *Gene Reports* **21**, 100956 (2020).

307. Kasela, S. *et al.* Genetic and non-genetic factors affecting the expression of COVID-19-relevant genes in the large airway epithelium. *Genome Medicine* **13**, 66 (2021).

308. Zhang, N., Wang, S. & Wong, C. C. L. Proteomics research of SARS-CoV-2 and COVID-19 disease. *Medical Review* **2**, 427–445.

309. Meyers, J. M. *et al.* The proximal proteome of 17 SARS-CoV-2 proteins links to disrupted antiviral signaling and host translation. *PLOS Pathogens* **17**, e1009412 (2021).

310. Tavakoli, R. *et al.* Expression of TRIM56 gene in SARS-CoV-2 variants and its relationship with progression of COVID-19. *Future Virology* 10.2217/fvl-2022-0210 doi:10.2217/fvl-2022-0210.

311. Heidary, F. & Gharebaghi, R. Systematic review of the antiviral properties of TRIM56: A potential therapeutic intervention for COVID-19. *Expert Review of Clinical Immunology* **16**, 973–984 (2020).

312. Lee, A. C. *et al.* COVID-19 Severity Potentially Modulated by Cardiovascular-Disease-Associated Immune Dysregulation. *Viruses* **13**, (2021).

313. Bradic, M. *et al.* DNA methylation predicts the outcome of COVID-19 patients with acute respiratory distress syndrome. *Journal of Translational Medicine* **20**, 526 (2022).

314. Fang, K.-Y. *et al.* Screening the hub genes and analyzing the mechanisms in discharged COVID-19 patients retesting positive through bioinformatics analysis. *Journal of clinical laboratory analysis* **36**, e24495 (2022).

315. Dolskiy, A. A. *et al.* Increased LAMP1 Expression Enhances SARS-CoV-1 and SARS-CoV-2 Production in Vero-Derived Transgenic Cell Lines. *Molecular Biology* **56**, 463–468 (2022).

316. Crudele, A. *et al.* Hydroxytyrosol Recovers SARS-CoV-2-PLpro-Dependent Impairment of Interferon Related Genes in Polarized Human Airway, Intestinal and Liver Epithelial Cells. *Antioxidants* **11**, 1466 (2022).

317. Barman, R. K., Mukhopadhyay, A., Maulik, U. & Das, S. A network biology approach to identify crucial host targets for COVID-19. *Methods (San Diego, Calif.)* **203**, 108–115 (2022).

318. Picchiotti, N. *et al.* Post-Mendelian Genetic Model in COVID-19. *Cardiology and Cardiovascular Medicine* **5**, 673–694 (2021).

319. Liu, J.-F. *et al.* Proteomic and phosphoproteomic characteristics of the cortex, hippocampus, thalamus, lung, and kidney in COVID-19-infected female K18-hACE2 mice. *eBioMedicine* **90**, (2023).

320. Valdés-López, J. F. & Urcuqui-Inchima, S. Antiviral response and immunopathogenesis of interleukin 27 in COVID-19. *Archives of Virology* **168**, 178 (2023).

321. Klingler, J. *et al.* Immune profiles to distinguish hospitalized versus ambulatory COVID-19 cases in older patients. *iScience* **25**, 105608 (2022).

322. Pahl, M. C. *et al.* Implicating effector genes at COVID-19 GWAS loci using promoter-focused Capture-C in disease-relevant immune cell types. *Genome Biology* **23**, 125 (2022).

323. Zhao, W. *et al.* Histone demethylase KDM4B epigenetically controls NLRP3 expression to enhance inflammatory response. (2023) doi:10.21203/rs.3.rs-3138058/v1.

324. Yang, B. *et al.* Clinical and molecular characteristics of COVID-19 patients with persistent SARS-CoV-2 infection. *Nature Communications* **12**, 3501 (2021).

325. Ackermann, M. *et al.* Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. *New England Journal of Medicine* **383**, 120–128 (2020).

326. Schreiber, A. *et al.* The MEK1/2-inhibitor ATR-002 efficiently blocks SARS-CoV-2 propagation and alleviates pro-inflammatory cytokine/chemokine responses. *Cellular and Molecular Life Sciences* **79**, 65 (2022).

327. Chatterjee, B. & Thakur, S. S. SARS-CoV-2 Infection Triggers Phosphorylation: Potential Target for Anti-COVID-19 Therapeutics. *Frontiers in Immunology* **13**, (2022).

328. Jin, S. *et al.* Suppression of ACE2 SUMOylation protects against SARS-CoV-2 infection through TOLLIP-mediated selective autophagy. *Nature Communications* **13**, 5204 (2022).

329. D'Antonio, M. *et al.* SARS-CoV-2 susceptibility and COVID-19 disease severity are associated with genetic variants affecting gene expression in a variety of tissues. *Cell Reports* **37**, 110020 (2021).

330. Thiecke, M. J., Yang, E. J., Burren, O. S., Ray-Jones, H. & Spivakov, M. Prioritisation of Candidate Genes Underpinning COVID-19 Host Genetic Traits Based on High-Resolution 3D Chromosomal Topology. *Frontiers in Genetics* **12**, (2021).

331. Ghosh, N., Saha, I., Sharma, N. & Sarkar, J. P. Human miRNAs to Identify Potential Regions of SARS-CoV-2. *ACS Omega* **7**, 21086–21101 (2022).

332. Temena, M. A. & Acar, A. Increased TRIM31 gene expression is positively correlated with SARS-CoV-2 associated genes TMPRSS2 and TMPRSS4 in gastrointestinal cancers. *Scientific Reports* **12**, 11763 (2022).

333. Torinsson Naluai, Å. *et al.* Transcriptomics unravels molecular changes associated with cilia and COVID-19 in chronic rhinosinusitis with nasal polyps. *Scientific Reports* **13**, 6592 (2023).

334. Chicco, D. & Agapito, G. Nine quick tips for pathway enrichment analysis. *PLoS Computational Biology* **18**, e1010348 (2022).

335. Boehm, F. J. & Zhou, X. Statistical methods for Mendelian randomization in genome-wide association studies: A review. *Computational and Structural Biotechnology Journal* **20**, 2338–2351 (2022).

336. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the 'Sum of Single Effects' model. *PLOS Genetics* **18**, e1010299 (2022).

337. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *American Journal of Human Genetics* **101**, 539–551 (2017).

338. Han, Y., Liu, D. & Li, L. PD-1/PD-L1 pathway: Current researches in cancer. *American Journal of Cancer Research* **10**, 727–742 (2020).

339. Francisco, L. M., Sage, P. T. & Sharpe, A. H. The PD-1 Pathway in Tolerance and Autoimmunity. *Immunological reviews* **236**, 219–242 (2010).

340. Sabbatino, F. *et al.* PD-L1 Dysregulation in COVID-19 Patients. *Frontiers in Immunology* **12**, 695242 (2021).

341. Chavez-Galan, L. *et al.* Circulating Levels of PD-L1, TIM-3 and MMP-7 Are Promising Biomarkers to Differentiate COVID-19 Patients That Require Invasive Mechanical Ventilation. *Biomolecules* **12**, 445 (2022).

342. R Bonam, S., Hu, H. & Bayry, J. Role of the PD-1 and PD-L1 axis in COVID-19. *Future Microbiology* 10.2217/fmb-2022-0103 doi:10.2217/fmb-2022-0103.

343. Mesri, E. A., Cesarman, E. & Boshoff, C. Kaposi's sarcoma herpesvirus/ Human herpesvirus-8 (KSHV/HHV8), and the oncogenesis of Kaposi's sarcoma. *Nature reviews. Cancer* **10**, 707–719 (2010).

344. Alberts, B. *et al.* B Cells and Antibodies. in *Molecular Biology of the Cell. 4th edition* (Garland Science, 2002).

345. Tanaka, S. & Baba, Y. B Cell Receptor Signaling. *Advances in Experimental Medicine and Biology* **1254**, 23–36 (2020).

346. Berenji, K., Drazner, M. H., Rothermel, B. A. & Hill, J. A. Does load-induced ventricular hypertrophy progress to systolic heart failure? *American Journal of Physiology. Heart and Circulatory Physiology* **289**, H8–H16 (2005).

347. Artham, S. M. *et al.* Clinical impact of left ventricular hypertrophy and implications for regression. *Progress in Cardiovascular Diseases* **52**, 153–167 (2009).

348. Samak, M. *et al.* Cardiac Hypertrophy: An Introduction to Molecular and Cellular Basis. *Medical Science Monitor Basic Research* **22**, 75–79 (2016).

349. Meyer, R. *et al.* Cloning of the DNA-binding subunit of human nuclear factor kappa B: The level of its mRNA is strongly regulated by phorbol ester or tumor necrosis factor alpha. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 966–970 (1991).

350. Perkins, N. D. Integrating cell-signalling pathways with NF-κB and IKK function. *Nature Reviews Molecular Cell Biology* **8**, 49–62 (2007).

351.

Pasparakis, M., Luedde, T. & Schmidt-Supprian, M. Dissection of the NF-κB signalling cascade in transgenic and knockout mice. *Cell Death & Differentiation* **13**, 861–872 (2006).

352. Gerondakis, S., Grossmann, M., Nakamura, Y., Pohl, T. & Grumont, R. Genetic approaches in mice to understand Rel/NF-kappaB and IkappaB function: Transgenics and knockouts. *Oncogene* **18**, 6888–6895 (1999).

353. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduction and Targeted Therapy* **2**, 1–9 (2017).

354. Rusnak, F. & Mertz, P. Calcineurin: Form and Function. *Physiological Reviews* **80**, 1483–1521 (2000).

355. Regulation of T Cell Activation by Calcineurin Inhibition. *Science's STKE* **2003**, tw351–TW351 (2003).

356. Zheng, C. F. & Guan, K. L. Cloning and characterization of two distinct human extracellular signal-regulated kinase activator kinases, MEK1 and MEK2. *Journal of Biological Chemistry* **268**, 11435–11439 (1993).

357. Leicht, D. T. *et al.* Raf kinases: Function, regulation and role in human cancer. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1773**, 1196–1212 (2007).

358. Yoon, S. & Seger, R. The extracellular signal-regulated kinase: Multiple substrates regulate diverse cellular functions. *Growth Factors* **24**, 21–44 (2006).

359. von Kriegsheim, A. *et al.* Cell fate decisions are specified by the dynamic ERK interactome. *Nature Cell Biology* **11**, 1458–1464 (2009).

360. Han, K.-J. *et al.* Mechanisms of the TRIF-induced interferon-stimulated response element and NF-kappaB activation and apoptosis pathways. *The Journal of Biological Chemistry* **279**, 15652–15661 (2004).

361. Malerba, N., De Nittis, P. & Merla, G. The Emerging Role of Gβ Subunits in Human Genetic Diseases. *Cells* **8**, 1567 (2019).

362. Guo, Y. *et al.* Gβ2 Regulates the Multipolar-Bipolar Transition of Newborn Neurons in the Developing Neocortex. *Cerebral Cortex* **27**, 3414–3426 (2017).

363. Yokoyama, K. *et al.* BANK regulates BCR-induced calcium mobilization by promoting tyrosine phosphorylation of IP(3) receptor. *The EMBO journal* **21**, 83–92 (2002).

364. Berlansky, S. *et al.* Calcium Signals during SARS-CoV-2 Infection: Assessing the Potential of Emerging Therapies. *Cells* **11**, 253 (2022).

365. Cusato, J. *et al.* COVID-19: A Possible Contribution of the MAPK Pathway. *Biomedicines* **11**, 1459 (2023).

366. Gudowska-Sawczuk, M. & Mroczko, B. The Role of Nuclear Factor Kappa B (NF-κB) in Development and Treatment of COVID-19: Review. *International Journal of Molecular Sciences* **23**, 5283 (2022).

367.

Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nature Communications* **12**, 727 (2021).

368. Chen, T. *et al.* Molecular characterization of a novel dipeptidyl peptidase like 2-short form (DPL2-s) that is highly expressed in the brain and lacks dipeptidyl peptidase activity. *Biochimica et biophysica acta* **1764**, 33–43 (2006).

369. Nakanishi, T. *et al.* Alternative splicing in the lung influences COVID-19 severity and respiratory disease. (2022) doi:10.1101/2022.10.18.22281202.

370. Justa-Schuch, D., Möller, U. & Geiss-Friedlander, R. The amino terminus extension in the long dipeptidyl peptidase 9 isoform contains a nuclear localization signal targeting the active peptidase to the nucleus. *Cellular and Molecular Life Sciences* **71**, 3611–3626 (2014).

371. Zhang, R. & Su, B. Small but influential: The role of microRNAs on gene regulatory network and 3'UTR evolution. *Journal of Genetics and Genomics = Yi Chuan Xue Bao* **36**, 1–6 (2009).

372. Saunders, M. A., Liang, H. & Li, W.-H. Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 3300–3305 (2007).

373. Gottlieb, A., Stein, G. Y., Ruppin, E. & Sharan, R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology* **7**, 496 (2011).

374. Wang, H., Gu, Q., Wei, J., Cao, Z. & Liu, Q. Mining drug-disease relationships as a complement to medical genetics-based drug repositioning: Where a recommendation system meets genome-wide association studies. *Clinical Pharmacology and Therapeutics* **97**, 451–454 (2015).

375. Rudd, C. E. Lnk Adaptor: Novel Negative Regulator of B Cell Lymphopoiesis. *Science's STKE* **2001**, pe1–pe1 (2001).

376. Zhang, Y., Alexander, P. B. & Wang, X.-F. TGF-β Family Signaling in the Control of Cell Proliferation and Survival. *Cold Spring Harbor Perspectives in Biology* **9**, a022145 (2017).

377. Pak, Y., Pham, N. & Rotin, D. Direct binding of the Beta1 adrenergic receptor to the cyclic AMP-dependent guanine nucleotide exchange factor CNrasGEF leads to Ras activation. *Molecular and Cellular Biology* **22**, 7942–7952 (2002).

378. Chung, F. Z., Wang, C. D., Potter, P. C., Venter, J. C. & Fraser, C. M. Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. *The Journal of Biological Chemistry* **263**, 4052–4055 (1988).

379. Jia, G. *et al.* N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology* **7**, 885–887 (2011).

380. Yang, C. *et al.* Mammalian CARMIL inhibits actin filament capping by capping protein. *Developmental Cell* **9**, 209–221 (2005).

381. Ekaratanawong, S. *et al.* Human organic anion transporter 4 is a renal apical organic anion/dicarboxylate exchanger in the proximal tubules. *Journal of Pharmacological Sciences* **94**, 297–304 (2004).

382. Enomoto, A. *et al.* Molecular identification of a renal urate–anion exchanger that regulates blood urate levels. *Nature* **417**, 447–452 (2002).

383. Zenobia, C. & Hajishengallis, G. Basic biology and role of interleukin-17 in immunity and inflammation. *Periodontology 2000* **69**, 142–159 (2015).

384. Tóthová, Z., Tomc, J., Debeljak, N. & Solár, P. STAT5 as a Key Protein of Erythropoietin Signalization. *International Journal of Molecular Sciences* **22**, 7109 (2021).

385. Peng, B., Kong, G., Yang, C. & Ming, Y. Erythropoietin and its derivatives: From tissue protection to immune regulation. *Cell Death & Disease* **11**, 79 (2020).

386. Tong, W., Zhang, J. & Lodish, H. F. Lnk inhibits erythropoiesis and Epo-dependent JAK2 activation and downstream signaling pathways. *Blood* **105**, 4604–4612 (2005).

387. Perez-Garcia, A. *et al.* Genetic loss of SH2B3 in acute lymphoblastic leukemia. *Blood* **122**, 2425–2432 (2013).

388. Hammarén, H. M., Virtanen, A. T., Raivola, J. & Silvennoinen, O. The regulation of JAKs in cytokine signaling and its breakdown in disease. *Cytokine* **118**, 48–63 (2019).

389. Baldini, C., Moriconi, F. R., Galimberti, S., Libby, P. & De Caterina, R. The JAK–STAT pathway: An emerging target for cardiovascular disease in rheumatoid arthritis and myeloproliferative neoplasms. *European Heart Journal* **42**, 4389–4400 (2021).

390. Mehta, N. N. Potential cardiovascular implications of Janus kinase inhibitors in immune mediated diseases. *Cardiovascular Research* **114**, e81–e83 (2018).

391. Wei, Q. *et al.* Cardiovascular safety of Janus kinase inhibitors in patients with rheumatoid arthritis: Systematic review and network meta-analysis. *Frontiers in Pharmacology* **14**, 1237234 (2023).

392. Moon, C. *et al.* Erythropoietin, modified to not stimulate red blood cell production, retains its cardioprotective properties. *The Journal of Pharmacology and Experimental Therapeutics* **316**, 999–1005 (2006).

393. Brines, M. & Cerami, A. The Receptor That Tames the Innate Immune Response. *Molecular Medicine* **18**, 486–496 (2011).

394. Anagnostou, A., Lee, E. S., Kessimian, N., Levinson, R. & Steiner, M. Erythropoietin has a mitogenic and positive chemotactic effect on endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 5978–5982 (1990).

395. Satoh, K. *et al.* Important role of endogenous erythropoietin system in recruitment of endothelial progenitor cells in hypoxia-induced pulmonary hypertension in mice. *Circulation* **113**, 1442–1450 (2006).

396. Brines, M. & Cerami, A. Erythropoietin-mediated tissue protection: Reducing collateral damage from the primary injury response. *Journal of Internal Medicine* **264**, 405–432 (2008).

397. Collino, M., Thiemermann, C., Cerami, A. & Brines, M. Flipping the molecular switch for innate protection and repair of tissues: Long-lasting effects of a non-erythropoietic small peptide engineered from erythropoietin. *Pharmacology & Therapeutics* **151**, 32–40 (2015).

398. Hand, C. C. & Brines, M. Promises and pitfalls in erythopoietin-mediated tissue protection: Are nonerythropoietic derivatives a way forward? *Journal of Investigative Medicine: The Official Publication of the American Federation for Clinical Research* **59**, 1073–1082 (2011).

399. Leist, M. *et al.* Derivatives of erythropoietin that are tissue protective but not erythropoietic. *Science (New York, N.Y.)* **305**, 239–242 (2004).

400. Fiordaliso, F. *et al.* A nonerythropoietic derivative of erythropoietin protects the myocardium from ischemia–reperfusion injury. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2046–2051 (2005).

401. Eswarappa, M., Cantarelli, C. & Cravedi, P. Erythropoietin in Lupus: Unanticipated Immune Modulating Effects of a Kidney Hormone. *Frontiers in Immunology* **12**, (2021).

402. Zhang, Z., Liu, D., Zhang, X. & Wang, X. Erythropoietin Treatment Ameliorates Lupus Nephritis of MRL/lpr Mice. *Inflammation* **41**, 1888–1899 (2018).

403. Huang, B. *et al.* Non-erythropoietic erythropoietin-derived peptide protects mice from systemic lupus erythematosus. *Journal of Cellular and Molecular Medicine* **22**, 3330–3339 (2018).

404. Nairz, M. *et al.* Erythropoietin Contrastingly Affects Bacterial Infection and Experimental Colitis by Inhibiting Nuclear Factor-κB-Inducible Immune Pathways. *Immunity* **34**, 61 (2011).

405. Nakamura, S. *et al.* Erythropoietin attenuates intestinal inflammation and promotes tissue regeneration. *Scandinavian Journal of Gastroenterology* **50**, 1094–1102 (2015).

406. Liu, Y. *et al.* Erythropoietin-Derived Nonerythropoietic Peptide Ameliorates Experimental Autoimmune Neuritis by Inflammation Suppression and Tissue Protection. *PLoS ONE* **9**, e90942 (2014).

407. Agnello, D. *et al.* Erythropoietin exerts an anti-inflammatory effect on the CNS in a model of experimental autoimmune encephalomyelitis. *Brain Research* **952**, 128–134 (2002).

408. Maiese, K. Erythropoietin and diabetes mellitus. *World Journal of Diabetes* **6**, 1259–1273 (2015).

409. Katz, O. *et al.* Erythropoietin treatment leads to reduced blood glucose levels and body mass: Insights from murine models. *The Journal of Endocrinology* **205**, 87–95 (2010).

410. Chu, Q. *et al.* Differential gene expression pattern of diabetic rat retinas after intravitreal injection of erythropoietin. *Clinical & Experimental Ophthalmology* **39**, 142–151 (2011).

411. Chattopadhyay, M., Walter, C., Mata, M. & Fink, D. J. Neuroprotective effect of herpes simplex virus-mediated gene transfer of erythropoietin in hyperglycemic dorsal root ganglion neurons. *Brain* **132**, 879–888 (2009).

412. Szczepanek-Parulska, E., Hernik, A. & Ruchała, M. Anemia in thyroid diseases. *Polish Archives of Internal Medicine* **127**, 352–360 (2017).

413. Ma, Y. *et al.* Thyroid hormone induces erythropoietin gene expression through augmented accumulation of hypoxia-inducible factor-1. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* **287**, R600–607 (2004).

414. Das, K. C., Mukherjee, M., Sarkar, T. K., Dash, R. J. & Rastogi, G. K. Erythropoiesis and erythropoietin in hypo- and hyperthyroidism. *The Journal of Clinical Endocrinology and Metabolism* **40**, 211–220 (1975).

415. Ng, Y. Y. *et al.* Impact of Thyroid Dysfunction on Erythropoietin Dosage in Hemodialysis Patients. *Thyroid* **23**, 552–561 (2013).

416. Song, L. & Schindler, C. Chapter 249 - JAK-STAT Signaling. in *Handbook of Cell Signaling (Second Edition)* (eds. Bradshaw, R. A. & Dennis, E. A.) 2041–2048 (Academic Press, 2010). doi:10.1016/B978-0-12-374145-5.00249-7.

417. Watford, W. T. *et al.* Signaling by IL-12 and IL-23 and the immunoregulatory roles of STAT4. *Immunological Reviews* **202**, 139–156 (2004).

418. Korman, B. D., Kastner, D. L., Gregersen, P. K. & Remmers, E. F. STAT4: Genetics, Mechanisms, and Implications for Autoimmunity Review for Current Allergy and Asthma Reports. *Current allergy and asthma reports* **8**, 398–403 (2008).

419. Croxford, A. L., Kulig, P. & Becher, B. IL-12-and IL-23 in health and disease. *Cytokine & Growth Factor Reviews* **25**, 415–421 (2014).

420. Yuan, S. *et al.* Mendelian randomization and clinical trial evidence supports TYK2 inhibition as a therapeutic target for autoimmune diseases. *eBioMedicine* **89**, (2023).

421. Vanderpump, M. P. J. The epidemiology of thyroid disease. *British Medical Bulletin* **99**, 39–51 (2011).

422. Kato, S. The function of vitamin D receptor in vitamin D action. *Journal of Biochemistry* **127**, 717–722 (2000).

423. Yu, H., Xie, Y., Dai, M., Pan, Y. & Xie, C. SMAD3 interacts with vitamin D receptor and affects vitamin D-mediated oxidative stress to ameliorate cerebral ischaemia-reperfusion injury. *The European Journal of Neuroscience* **56**, 6055–6068 (2022).

424. Ding, N. *et al.* A Vitamin D Receptor/SMAD Genomic Circuit Gates Hepatic Fibrotic Response. *Cell* **153**, 601–613 (2013).

425. Sokol, S. I. *et al.* The effects of vitamin D repletion on endothelial function and inflammation in patients with coronary artery disease. *Vascular Medicine* **17**, 394–404 (2012).

426. Barbarawi, M. *et al.* Vitamin D Supplementation and Cardiovascular Disease Risks in More Than 83 000 Individuals in 21 Randomized Clinical Trials: A Meta-analysis. *JAMA Cardiology* **4**, 765 (2019).

427. Qian, M. A. *et al.* MAGI3 Suppresses Glioma Cell Proliferation via Upregulation of PTEN Expression. *Biomedical and Environmental Sciences* **28**, 502–509 (2015).

428. Kotelevets, L. & Chastre, E. A New Story of the Three Magi: Scaffolding Proteins and lncRNA Suppressors of Cancer. *Cancers* **13**, 4264 (2021).

429. Oliver, E., Mayor Jr, F. & D'Ocon, P. Beta-blockers: Historical Perspective and Mechanisms of Action. *Revista Española de Cardiología (English Edition)* **72**, 853–862 (2019).

430. Yang, X. *et al.* Beta-2 adrenergic receptor mediated ERK activation is regulated by interaction with MAGI-3. *FEBS Letters* **584**, 2207–2212 (2010).

431. He, J. *et al.* Proteomic analysis of Beta1-adrenergic receptor interactions with PDZ scaffold proteins. *The Journal of Biological Chemistry* **281**, 2820–2827 (2006).

432. Abuhelwa, A. Y. *et al.* Concomitant beta-blocker use is associated with a reduced rate of remission in patients with rheumatoid arthritis treated with disease-modifying anti-rheumatic drugs: A post hoc multicohort analysis. *Therapeutic Advances in Musculoskeletal Disease* **13**, 1759720X211009020 (2021).

433. Willemze, R. A., Bakker, T., Pippias, M., Ponsioen, C. Y. & de Jonge, W. J. β-Blocker use is associated with a higher relapse risk of inflammatory bowel disease: A Dutch retrospective case-control study. *European Journal of Gastroenterology & Hepatology* **30**, 161–166 (2018).

434. Hughes, G. R. Hypotensive agents, beta-blockers, and drug-induced lupus. *British Medical Journal (Clinical research ed.)* **284**, 1358–1359 (1982).

435. Bilewicz-Stebel, M., Miziołek, B., Bergler-Czop, B. & Stańkowska, A. Drug-induced Subacute Cutaneous Lupus Erythematosus Caused by a Topical Beta Blocker - Timolol. *Acta dermatovenerologica Croatica: ADC* **26**, 44–47 (2018).

436. Kane, G. C. *et al.* KCNJ11 gene knockout of the Kir6.2 K ATP channel causes maladaptive remodeling and heart failure in hypertension. *Human Molecular Genetics* **15**, 2285–2297 (2006).

437. Tinker, A., Aziz, Q. & Thomas, A. The role of ATP-sensitive potassium channels in cellular function and protection in the cardiovascular system. *British Journal of Pharmacology* **171**, 12–23 (2014).

438. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England Journal of Medicine* **373**, 895–907 (2015).

439. Moraru, A. *et al.* THADA Regulates the Organismal Balance between Energy Storage and Heat Production. *Developmental Cell* **41**, 72–81.e6 (2017).

440. Zhang, Y. *et al.* THADA inhibition in mice protects against type 2 diabetes mellitus by improving pancreatic β-cell function and preserving β-cell mass. *Nature Communications* **14**, 1020 (2023).

441. Mambiya, M. *et al.* The Play of Genes and Non-genetic Factors on Type 2 Diabetes. *Frontiers in Public Health* **7**, 349 (2019).

442. Panoutsopoulou, K. *et al.* The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: A Mendelian randomisation study. *Annals of the Rheumatic Diseases* **73**, 2082–2086 (2014).

443. Sheth, K. N. *et al.* Long-Term Outcomes in Patients Aged $\leq$70 Years With Intravenous Glyburide From the Phase II GAMES-RP Study of Large Hemispheric Infarction: An Exploratory Analysis. *Stroke* **49**, 1457–1463 (2018).

444. Costa, B. B. S. da *et al.* Glibenclamide in aneurysmal subarachnoid hemorrhage: A randomized controlled clinical trial. *Journal of Neurosurgery* 1–8 (2021) doi: 10.3171/2021.7.JNS21846.

445. Action to Control Cardiovascular Risk in Diabetes Study Group *et al.* Effects of intensive glucose lowering in type 2 diabetes. *The New England Journal of Medicine* **358**, 2545–2559 (2008).

446. García-García, H. M. *et al.* Evaluation of in-stent restenosis in the APPROACH trial (Assessment on the Prevention of Progression by Rosiglitazone On Atherosclerosis in diabetes patients with Cardiovascular History). *The International Journal of Cardiovascular Imaging* **28**, 455–465 (2012).

447. Hagos, Y., Stein, D., Ugele, B., Burckhardt, G. & Bahn, A. Human renal organic anion transporter 4 operates as an asymmetric urate transporter. *Journal of the American Society of Nephrology: JASN* **18**, 430–439 (2007).

448. Vadakedath, S. & Kandi, V. Probable Potential Role of Urate Transporter Genes in the Development of Metabolic Disorders. *Cureus* **10**, e2382.

449. Zhu, W., Deng, Y. & Zhou, X. Multiple Membrane Transporters and Some Immune Regulatory Genes are Major Genetic Factors to Gout. *The Open Rheumatology Journal* **12**, 94–113 (2018).

450. Ho, H. H. *et al.* Gout in systemic lupus erythematosus and overlap syndrome – a hospital-based study. *Clinical Rheumatology* **22**, 295–298 (2003).

451. Quilis, N. & Andrés, M. AB0890 Systemic lupus erythematosus and gout: Really an unusual association? *Annals of the Rheumatic Diseases* **76**, 1367–1367 (2017).

452. Aguiar, F., Brito, I. & Sibilia, J. Revisiting the association between systemic lupus erythematosus and gout. *Reumatología Clínica (English Edition)* **17**, 55–56 (2021).

453. Giordano, N. *et al.* Hyperuricemia and gout in thyroid endocrine disorders. *Clinical and Experimental Rheumatology* **19**, 661–665 (2001).

454. See, L.-C. *et al.* Hyperthyroid and Hypothyroid Status Was Strongly Associated with Gout and Weakly Associated with Hyperuricaemia. *PLoS ONE* **9**, e114579 (2014).

455. Bruderer, S. G., Meier, C. R., Jick, S. S. & Bodmer, M. The association between thyroid disorders and incident gout: Population-based case–control study. *Clinical Epidemiology* **9**, 205–215 (2017).

456. Mariani, L. H. & Berns, J. S. The Renal Manifestations of Thyroid Disease. *Journal of the American Society of Nephrology* **23**, 22 (2012).

457. Piatek-Guziewicz, A. *et al.* Intestinal parameters of oxidative imbalance in celiac adults with extraintestinal manifestations. *World Journal of Gastroenterology* **23**, 7849–7862 (2017).

458. Relationship between Gout and Asthma: A National Database Analysis. *ACR Meeting Abstracts*.

459. Iwakura, Y. & Ishigame, H. The IL-23/IL-17 axis in inflammation. *Journal of Clinical Investigation* **116**, 1218–1222 (2006).

460. Fujino, S. *et al.* Increased expression of interleukin 17 in inflammatory bowel disease. *Gut* **52**, 65–70 (2003).

461. Duerr, R. H. *et al.* A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science (New York, N.Y.)* **314**, 1461–1463 (2006).

462. Thompson, A. I. & Lees, C. W. Genetics of ulcerative colitis. *Inflammatory Bowel Diseases* **17**, 831–848 (2011).

463. Torres, J. *et al.* ECCO Guidelines on Therapeutics in Crohn's Disease: Medical Treatment. *Journal of Crohn's and Colitis* **14**, 4–22 (2020).

464. Yen, D. *et al.* IL-23 is essential for T cell–mediated colitis and promotes inflammation via IL-17 and IL-6. *Journal of Clinical Investigation* **116**, 1310–1316 (2006).

465. Chen, Y. *et al.* Anti–IL-23 therapy inhibits multiple inflammatory pathways and ameliorates autoimmune encephalomyelitis. *Journal of Clinical Investigation* **116**, 1317–1326 (2006).

466. Ogawa, A., Andoh, A., Araki, Y., Bamba, T. & Fujiyama, Y. Neutralization of interleukin-17 aggravates dextran sulfate sodium-induced colitis in mice. *Clinical Immunology (Orlando, Fla.)* **110**, 55–62 (2004).

467. Hueber, W. *et al.* Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: Unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* **61**, 1693–1700 (2012).

468. Targan, S. R. *et al.* A Randomized, Double-Blind, Placebo-Controlled Phase 2 Study of Brodalumab in Patients With Moderate-to-Severe Crohn's Disease. *The American Journal of Gastroenterology* **111**, 1599–1607 (2016).

469. Fauny, M. *et al.* Paradoxical gastrointestinal effects of interleukin-17 blockers. *Annals of the Rheumatic Diseases* **79**, 1132–1138 (2020).

470. Whibley, N. & Gaffen, S. L. Gut-busters – IL-17 Ain't Afraid Of No IL-23. *Immunity* **43**, 620–622 (2015).

471. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

472. Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Research* **44**, 6046–6054 (2016).

473. Stacey, D. *et al.* ProGeM: A framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Research* **47**, e3 (2019).

474. Jhamb, D., Magid-Slav, M., Hurle, M. R. & Agarwal, P. Pathway analysis of GWAS loci identifies novel drug targets and repurposing opportunities. *Drug Discovery Today* **24**, 1232–1236 (2019).

475. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics* **47**, 569–576 (2015).

476. Pyle, C. J. *et al.* Elemental Ingredients in the Macrophage Cocktail: Role of ZIP8 in Host Response to Mycobacterium tuberculosis. *International Journal of Molecular Sciences* **18**, 2375 (2017).

477. Pyle, C. J. *et al.* Zinc Modulates Endotoxin-Induced Human Macrophage Inflammation through ZIP8 Induction and C/EBPβ Inhibition. *PloS One* **12**, e0169531 (2017).

478. Aydemir, T. B., Liuzzi, J. P., McClellan, S. & Cousins, R. J. Zinc transporter ZIP8 (SLC39A8) and zinc influence IFN-gamma expression in activated human T cells. *Journal of Leukocyte Biology* **86**, 337–348 (2009).

479. Redpath, S. A., Fonseca, N. M. & Perona-Wright, G. Protection and pathology during parasite infection: IL-10 strikes the balance. *Parasite Immunology* **36**, 233–252 (2014).

480. Carlini, V. *et al.* The multifaceted nature of IL-10: Regulation, role in immunological homeostasis and its relevance to cancer, COVID-19 and post-COVID conditions. *Frontiers in Immunology* **14**, (2023).

481. Lu, L., Zhang, H., Dauphars, D. J. & He, Y.-W. A Potential Role of Interleukin 10 in COVID-19 Pathogenesis. *Trends in Immunology* **42**, 3–5 (2021).

482. Luporini, R. L. *et al.* IL-6 and IL-10 are associated with disease severity and higher comorbidity in adults with COVID-19. *Cytokine* **143**, 155507 (2021).

483. Diao, B. *et al.* Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). *Frontiers in Immunology* **11**, 827 (2020).

484. Li, J. *et al.* Dynamic changes in serum IL-6, IL-8, and IL-10 predict the outcome of ICU patients with severe COVID-19. *Annals of Palliative Medicine* **10**, 3706–3714 (2021).

485. Mehta, P. *et al.* COVID-19: Consider cytokine storm syndromes and immunosuppression. *Lancet (London, England)* **395**, 1033–1034 (2020).

486. Iyer, S. S. & Cheng, G. Role of interleukin 10 transcriptional regulation in inflammation and autoimmune disease. *Critical Reviews in Immunology* **32**, 23–63 (2012).

487. Barry, J. C. *et al.* Hyporesponsiveness to the anti-inflammatory action of interleukin-10 in type 2 diabetes. *Scientific Reports* **6**, 21244 (2016).

488. Ng, T. H. S. *et al.* Regulation of adaptive immunity; the role of interleukin-10. *Frontiers in Immunology* **4**, 129 (2013).

489. Aghbash, P. S., Eslami, N., Shamekh, A., Entezari-Maleki, T. & Baghi, H. B. SARS-CoV-2 infection: The role of PD-1/PD-L1 and CTLA-4 axis. *Life Sciences* **270**, 119124 (2021).

490. Kong, Y. *et al.* Storm of soluble immune checkpoints associated with disease severity of COVID-19. *Signal Transduction and Targeted Therapy* **5**, 192 (2020).

491. Tang, Q. *et al.* The role of PD-1/PD-L1 and application of immune-checkpoint inhibitors in human cancers. *Frontiers in Immunology* **13**, (2022).

492. Pickles, O. J. *et al.* Immune checkpoint blockade: Releasing the breaks or a protective barrier to COVID-19 severe acute respiratory syndrome? *British Journal of Cancer* **123**, 691–693 (2020).

493. Ramilowski, J. A. *et al.* Functional annotation of human long noncoding RNAs via molecular phenotyping. *Genome Research* **30**, 1060–1072 (2020).

494. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).

495. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Research* **49**, D1207–D1217 (2021).

496. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. 2022.04.13.22273750 (2022) doi:10.1101/2022.04.13.22273750.