# *Genomic approaches to pre-clinical drug target prioritisation in immune-related traits*

Bogdan Knezevic




Trinity College
University of Oxford

A thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Michaelmas Term, 2019

*Genomic approaches to pre-clinical drug target prioritisation in immune-related traits*

Bogdan Knezevic, Trinity College, Michaelmas Term 2019
*Submitted as part of the requirements for the degree of Doctor of Philosophy*

Drug discovery is time-consuming and expensive; bringing a novel compound to regulatory approval can take several years and billions of dollars. A significant portion of costs stem from compounds that have failed in clinical development, for reasons including a lack of clinical efficacy. Simultaneously, the emergence of genomic approaches like genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) has enabled better functional understanding of human genetics and robust mapping of effects of gene perturbations on phenotype. This thesis focuses on the aggregation of genomic approaches for the purposes of drug target prioritisation in pre-clinical stages of development, in immune-related traits.

The Priority Index (Pi) approach combines GWAS knowledge with cell-specific eQTLs, promoter-capture Hi-C, ontologies, and network information to prioritise genes in a disease-specific manner, using solely genetics ('discovery mode') or combining genetic data, clinical trial information, and machine learning ('supervised mode'). The outputs of Pi are computationally validated and benchmarked using multiple sclerosis as a case example, and the performance of Pi as a drug discovery prioritisation pipeline is quantified.

Pi is further implemented in the context of myasthenia gravis, and analysis is conducted at the gene, pathway, and network-level to develop an understanding of drivers of pathogenesis, functionally-relevant cell types, and genetic differences between early and late onset forms of disease. Moreover, the outputs are refined through statistical fine-mapping, eQTL, and colocalisation analysis, in addition to comparison with externally-generated datasets. Lastly, experimental validation is conducted to test hypotheses generated by Pi analysis. Specifically, prioritised targets for Dupuytren's disease and systemic lupus erythematosus are examined via epigenetic probe inhibition and siRNA knockdown, within the context of patient-derived tissue. Effects are interrogated from the perspective of both gene expression and cellular phenotype/function.

This thesis demonstrates the utility of integrative genomics for target prioritisation, highlighting the potential for informing and enabling pre-clinical drug discovery.

# Acknowledgments

First and foremost, I want to thank my supervisors, Julian Knight, Catherine Green, and Lahiru Handunnetthi. As group lead, Julian directed the project my thesis was part of, including guiding my work and helping me navigate the ULTRA-DD consortium, as well as providing me numerous opportunities for personal development. Additionally, Julian was incredibly understanding and supportive of my evolving research interests, going so far as to help facilitate my transition to a completely new project after the first year of my degree. Similarly, Cath was understanding of changing personal interests; while she was initially directly involved in guiding the experimental work in my first year, Cath remained very supportive of my DPhil process after I switched projects, and was a fantastic guide in structuring my thesis strategy, including synthesis and composition. Lahiru joined the group halfway into my project and was brought on as my supervisor when it became apparent that we had similar and synergistic research interests. He was absolutely instrumental in enabling me to feel confident in taking leadership and initiative within the ULTRA-DD project, and was a fantastic weekly touch-point for guiding experimental work and analyses.

All members of the Julian Knight Group deserve my sincere gratitude for their assistance to my work at various points in time. In particular, I would like to thank Hai Fang who patiently guided me in my transition from someone with a neuroscience background (rooted in behavioural research) to someone who could read, understand, implement, and write code for bioinformatic analysis. Additionally, I want to thank Katie Burnham for her patience and understanding when answering my endless questions around use of numerous computational tools/approaches, as well as fellow DPhil candidates Julie Osgood and Yuxin Mi for being good friends and even better vent-partners-in-crime.

My project would also not have been possible without the ULTRA-DD consortium and the numerous people/groups involved in coordinating collaborative experiments. In particular, I want to thank Michael Sundström, who led ULTRA-DD from the European side and who was a staunch believer and supporter of mine from my very first year within the project; Louise Berg, who generously guided me in my collaborative wet lab work with her, along with Mingmei Shang; and Fiona McCann and Lynn Williams, who openly collaborated with us and provided me with opportunities to contribute

directly to their work while also informing my own thesis aims.

Lastly, none of this would have been possible without the continuous support of my family and friends. My parents, Nataša and Jovan, who have unconditionally loved me, guided me, and inspired me throughout all my personal and professional endeavours; my grandma, Savka, who planted and nurtured my love for mathematics and problem solving from a young age; my uncle and aunt, Ivan and Maria, who listened to me complain about thesis hardships and made me feel normal; my three cousins, Anastasia, Katarina, and Sava, for unknowingly being constant reminders of the important things in life; my partner, Deanna, for riding all the highs and lows of the last four years, on top of working hard with me to ensure our long-distance relationship not only survived but also thrived; my RI crew – Jess (and George!), Tin, Jonas, and Jade – for all the South African and New Zealand love anyone could ask for; my housemate, Jason, for 8-track guitar jam sessions, commitment to translating caloric intake to gym output (broritos!), and legendary cocktail-making skill development; and last but definitely not least, the entire Union Street Astro Turf football crew, for providing me with an outlet for pent up energy, stress, and competitiveness, for being the light at the end of the tunnel no matter how difficult of a week I was having, and for generally keeping me grounded amidst the bubble that is Oxford academia.

It may seem like a peculiar thing to do, but I feel compelled to dedicate this work to J.R.R. Tolkien, the creator of a childhood obsession of mine, whose incredible weaving of words enchanted me for so many years. It was my fascination with Tolkien's work that first introduced me to the Dreaming Spires, and that planted the early seedling of a desire to study at Oxford. So, by extension, this work is also for all the dreamers out there who may have been inspired by a book or a poem or a movie like I was – words are immensely powerful and can carry you far on their large wings, if you allow them to.

# Declarations

I declare that, unless otherwise stated, all work presented in this thesis is my own. Several aspects of the study relied upon collaboration where part of the work was conducted with or by others. These were done across groups in the Unrestricted Leveraging of Targets for Research Advancement and Drug Discovery (ULTRA-DD) consortium, which my project has been a part of since 2016.

The Priority Index (Pi) tool was developed in-house; I was part of this team which was led by Dr. Hai Fang, who was responsible for the development of Pi, including implementation and associated software. My primary role in Pi development related to curating Golden Standard drug-targets, conducting performance evaluation, and supporting development of Pi's supervised mode, along with aspects of Pi debugging and visualisation features. All application, evaluation, and interpretation of the Pi pipeline in the context of the diseases presented in this thesis (multiple sclerosis, myasthenia gravis, Dupuytren's disease, systemic lupus erythematosus, among others) was done by me.

The myasthenia gravis work was entirely my work, except for the CD4+ T-cell eQTL data which was generated by Dr. Silva Kasela and her colleagues. All analyses and interpretations (from curated GWAS SNPs through Priority Index processing and downstream work) were my own.

The Dupuytren's disease work was done in collaboration with Drs. Fiona McCann and Lynn M. Williams, from the Kennedy Institute of Rheumatology, University of Oxford. Patient sample collection and processing (including RNA extraction and sequencing) was done by Drs. Fiona McCann and Lynn M. Williams. I processed the raw sequencing data, including quality control, mapping, counting, and differential gene expression analysis, for the purposes of this thesis. Interpretation of data, Pi processing, and downstream analysis was led by me.

The systemic lupus erythematosus work was done in collaboration with Drs. Louise Berg and Mingmei Shang from the Karolinska Institute. The initial probe screen was done by Dr. Louise Berg; bioinformatic analysis to identify potential targets was done by me; experimental design and planning of targets for siRNA validation was done together; patient sample collection and siRNA transfection was led by Dr. Mingmei Shang; and overall interpretation of data was led by me.

# Abbreviations

**AS**: ankylosing spondylitis
**AUC**: area under the curve
**BCR**: B-cell receptor
**ChIP**: chromatin immunoprecipitation
**CRO**: Crohn's disease
**DD**: Dupuytren's disease
**EBV**: Epstein-Barr Virus
**EO**: early-onset
**eQTL**: expression quantitative trait loci
**FACS**: fluorescence-activated cell sorting
**FC**: fold-change
**FDR**: false discovery rate
**GSN**: Gold Standard Negative
**GSP**: Gold Standard Positive
**GWAS**: genome-wide association studies
**IQR**: interquartile range
**LD** : linkage disequilibrium
**LO**: late-onset
**MG**: myasthenia gravis
**MGI**: Mouse Gene Informatics
**MS**: multiple sclerosis
**OT**: OpenTargets
**PBMC**: peripheral blood mononuclear cell
**Pi**: Priority Index
**PR**: precision-recall
**RA**: rheumatoid arthritis
**RES**: running enrichment score
**ROC**: receiver operating characteristic
**SGC**: Structural Genomics Consortium
**siRNA**: small interfering RNA
**SLE**: systemic lupus erythematosus
**SNP**: single nucleotide polymorphism
**SVM**: support vector machine
**TCR**: T-cell receptor
**TLR**: Toll-like receptor
**TNF**: tumor necrosis factor
**UC**: ulcerative colitis
**ULTRA-DD**: Unrestricted Leveraging of Targets for Research Advancement and Drug Discovery

# Contents


1 Introduction **1**

1.1 Drug discovery: roadblocks along a complex pipeline . . . . . 2
1.2 Translation of genetic information to informativeness for drug development . . . . . 8
1.3 Existing tools leveraging genomics for drug discovery . . . . . 13
1.4 Hit validation for establishing the clinical efficacy of targets . . . . . 19
1.5 Aims of this work . . . . . 21

2 Materials and methods **23**

2.1 Disease-specific input SNPs . . . . . 23
2.2 Priority Index . . . . . 25
2.3 Annotator datasets . . . . . 27
2.4 Databases . . . . . 28
2.5 Precision-recall and receiver operating characteristic curves . . . . . 30
2.6 The hypergeometric test and gene set enrichment analysis . . . . . 30
2.7 Tissue-level variant enrichment analysis . . . . . 33
2.8 Linkage disequilibrium . . . . . 35
2.9 Statistical fine-mapping . . . . . 35
2.10 CD4+ cell eQTL analysis . . . . . 36
2.11 Colocalisation analysis . . . . . 36
2.12 Leading edge analysis . . . . . 37
2.13 Structural Genomics Consortium probes . . . . . 38
2.14 RNA-seq analysis for Dupuytren's disease . . . . . 38
2.15 Proteomic analysis for Dupuytren's disease . . . . . 39
2.16 B-cell maturation assay . . . . . 40
2.17 Visualisations . . . . . 42

3 Computational prioritisation of genes **43**

3.1 Introduction . . . . . 43
3.2 Results . . . . . 47

3.2.1 Priority Index discovery mode applied to MS . . . 47
3.2.2 Golden standards for performance assessment . . . 51
3.2.3 Performance evaluation of discovery mode . . . 54
3.2.4 Benchmarking performance against existing tools . . . 63
3.2.5 Priority Index performance with restricted set of clinical golden standards . . . 64
3.2.6 Confirmation of performance beyond MS . . . 66
3.2.7 Priority Index supervised mode . . . 68
3.2.8 Performance assessment of supervised mode . . . 74
3.2.9 Supervised mode captures differentially expressed genes based on MS subgroup transcriptional profiling . . . 78
3.2.10 Use of supervised mode to inform re-purposing potential in MS . . . 78
3.3 Discussion . . . 79
3.4 Conclusion . . . 86

4 Application of the prioritisation pipeline to myasthenia gravis **87**
4.1 Introduction . . . 87
4.2 Results . . . 90
4.2.1 SNP selection and integration with functional annotations . . . 90
4.2.2 Functional annotation of MG SNPs . . . 91
4.2.3 The disease-associated SNPs influence gene expression in T-cells . . . 92
4.2.4 Autoimmune disease-associated genes were prioritised in MG . . . 95
4.2.5 Comparison of prioritisation output to active versus remission-state MG . . . 101
4.2.6 Contrasting prioritised genes for early onset and late onset MG . . . 101
4.2.7 Pathway analyses highlight further differences between EO and LO disease . . . 103
4.2.8 Interrogation of EO vs LO differences using independently-generated data-sets 106
4.2.9 Statistical fine-mapping of the CTLA4 locus . . . 110
4.3 Discussion . . . 116
4.4 Conclusion . . . 124

5 Genomic validation of prioritisation outputs in the context of Dupuytren's disease **125**
5.1 Introduction . . . 125
5.2 Results . . . 128
5.2.1 Pi prioritisation of disease-relevant genes . . . 128
5.2.2 RNA-seq analysis of CREBBP/EP300 inhibitor-treated cells . . . 131
5.2.3 RNA-seq analysis of healthy versus disease cells . . . 133
5.2.4 Proteomic analysis of probe-treated cells . . . 136
5.3 Discussion . . . 136
5.4 Conclusion . . . 140

6 Gene-level experimental validation of prioritisation outputs in the context of systemic lupus erythematosus **142**
6.1 Introduction . . . 142
6.2 Results . . . 145
6.2.1 SLE Priority Index output . . . 145
6.2.2 Epigenetic target screen in SLE . . . 148
6.2.3 SLE target nomination . . . 151
6.2.4 siRNA knockdown . . . 153
6.3 Discussion . . . 159
6.4 Conclusion . . . 168

7 Discussion **170**
7.1 Computational prioritisation of genes . . . 171
7.2 Application of the prioritisation pipeline to myasthenia gravis . . . 174
7.3 Experimental validation of computationally prioritised targets . . . 176
7.4 Conclusion . . . 179

References **199**

8 Appendices **200**

*She ran across the field after it, and was just in time to see it pop down a large rabbit-hole under the hedge... In another moment down went Alice after it, never once considering how in the world she was to get out again.*

Lewis Carroll

# 1

# Introduction

To state that drug discovery is difficult would be a gross oversimplification of an extremely complicated and nuanced process. Unpacking and truly understanding these difficulties remains a challenge in the field that has attracted the attention of numerous researchers looking to overcome such obstacles. Moreover, the emergence of the field now colloquially referred to as 'big data' has pushed the scientific community to completely rethink the way that we conceptualise both the problem and the solution spaces. The field of *genomics* is a manifestation of 'big data' within the realm of human health and disease, and provides us with a tool with which to tackle the emerging problems within drug discovery. It is my hope that, with thoughtful and diligent research, analysis, and execution, genomic techniques can pave the way for the better understanding of complex conditions like immune-related diseases, ultimately leading to improved patient outcomes. Perhaps the easiest way to get an understanding of the overarching challenges is to jump into the numbers themselves, and become familiar with current benchmarks as well as recent trends. Once thus oriented, the opportunity for advancing the field becomes a little more apparent, and the rationale for the

motivation behind this thesis more than a little clearer.

## 1.1 Drug discovery: roadblocks along a complex pipeline

Put simply, drug discovery has become increasingly more expensive and less efficient over time. The average cost to bring a novel drug entity to market has been steadily growing; it is estimated to be at around 1.8 billion US dollars, while the estimated time to achieve this goal sits at around 10 years of discovery and development [1]; in fact, some more recent studies propose that the cost per novel drug is actually in the 3-10 billion USD range, depending on the pharmaceutical company in question [2]. Moreover, the average number of novel drugs being approved has been steadily declining over the last few decades [3], meaning that there are fewer new drugs emerging each year, and the ones that do reach the market have cost significantly more resources to get there. The pharmaceutical industry has responded by relying on a volume-based approach, by doubling or tripling early-stage candidates to try and keep the number of approved drugs (and, thus, revenue-generation) steady; effectively, this shifts the focus from truly understanding disease biology to simply aiming for arbitrary goals/numbers to hit, often involving numerous back-up or 'me too' drugs [4]. Essentially, this means that the sustainability of drug discovery as an entire industry is largely being driven by a 'more shots on goal' type of approach [5], a method that is unsustainable in the long run.

The reasons behind this decline in research and development progress are complex and multifaceted. They involve a combination of factors, including a general increased industry aversion to risk, a complicated reimbursement and payer landscape, an upward trend of attempts to treat increasingly complex diseases (i.e. the 'low hanging fruit' has already largely been picked), and a complicated emerging regulatory environment [3, 4, 6]. Further, large pharmaceutical companies have shifted focus to later stages of drug development over the last couple of decades, while the early-mid stages have largely been outsourced to academia and biotech [5]. Therefore, lower success rates in recent years may be due to generally larger numbers of small biotech companies who, in turn, take on riskier projects or less validated drug classes, and who are resource-limited when it comes to hiring people and developing or executing complex projects or trials [3]. This mutually-connected combination of the aforementioned factors makes it extremely difficult to isolate individual drivers of the troubling

trends we are seeing in the drug development industry at a macro-scale. Instead, a productive question we can ask is whether there are statistics that can help us to identify pieces at the micro-scale; in other words, can we better understand where in the scientific pipeline things seem to be going wrong, in order to identify opportunities for improvement moving forward. The benefit of taking such a data-centric approach that focuses on understanding existing hurdles in the scientific pipeline is that it should, in theory, identify options for improvement that are independent of the 'messier' questions around regulatory landscape or biotech company structure.

A useful place to start is by analysing the full drug development pipeline, from discovery through to development and finally to market. Thankfully, some studies have already done this to a detailed and comprehensive degree, including an analysis of 835 drug developers examined between 2003 and 2011 (consisting of more than 4000 drugs across more than 7300 independent drug development paths, including both lead and non-lead indications), as well as a detailed portfolio review of AstraZeneca's pipeline, including 142 drug discovery projects across all therapeutic areas that were active between 2005 and 2010 [3, 4]. Additionally, there have been studies that considered a narrower focus of larger pharmaceutical pipelines [6], or that instead leveraged pipeline insights to build models in order to capture the effects that influencing concrete factors would have on overall pharmaceutical output [1]. Many of the findings from these various studies provided the impetus for the work done throughout this thesis; in particular, they helped inform and shape the focus of both the work contained within these pages, as well as the wider project that this thesis is part of.

The first drug pipeline statistic worth sharing (and perhaps the one that best summarizes the problem) is this: one in ten development paths that start in Phase 1 are ultimately approved by the Food and Drug Administration [3, 4]. This statistic can be broken down further into its constituent parts by considering factors such as whether the compound development pipeline was for a lead indication or a secondary one, as well as what disease category it fell under. However, even when adjusting for these things, the numbers do not look overly promising; for instance, lead indications were found to have a likelihood of approval of about 15.3% from Phase 1 [3], which is better than the 10% number usually reported, but not by much (Fig. 1.1.1). Similarly, the likelihood of approval for the subset of autoimmune-immunological groups (the focus of this thesis) is 12.7% [3], which takes us back to the

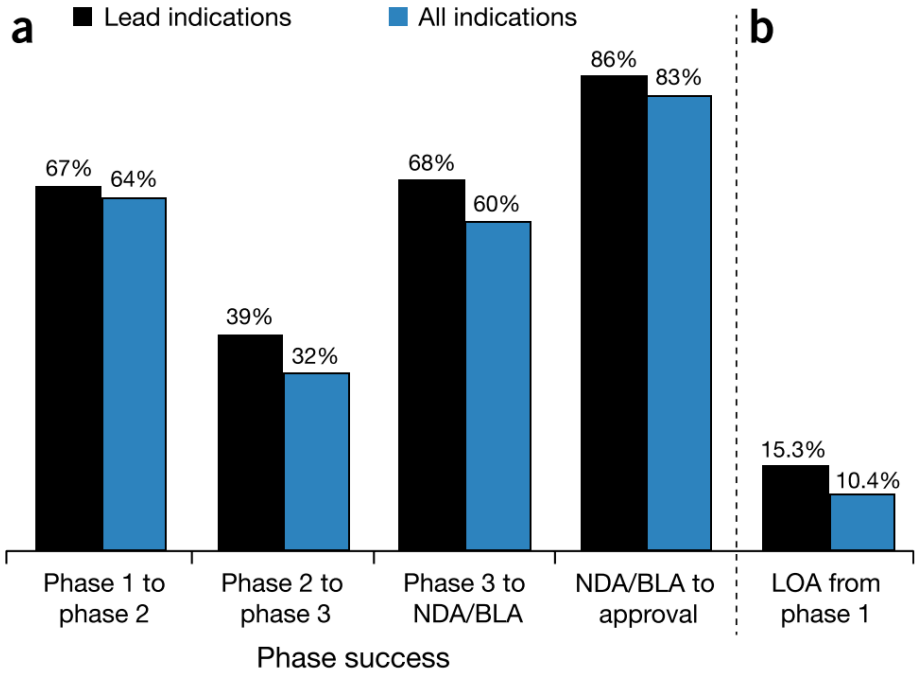



**Figure 1.1.1: Clinical trial success by development phase.** Phase success rates and likelihood of approval (LOA) of 5820 independent drug development paths, reproduced from the study by Hay and colleagues [3]. **(a)** Probability that a drug will successfully advance to the next phase of development. **(b)** LOA from Phase 1 through to approval.

question of why? More specifically, what is driving these failures, and at which points in the development pipeline are they most pronounced?

A breakdown of the different Phase transitions starts to answer the above questions; the success rate for Phase 1 to 2 is approximately 64.5%, Phase 2 to 3 is 32.4%, Phase 3 to 4 is 60.1%, and new drug application or biologic license application is 83.2% [3]. Therefore, it is apparent that the greatest rate of failure is in the transition from Phase 2 to Phase 3, suggesting that this is an area worth focusing on with regards to improvement. To put this into context, 63% of costs of bringing a new drug to market are associated with clinical development between Phase 1-3 [1] meaning that the average cost of clinical development in this zone is roughly 1.13 billion USD; having a compound fail in the later stages of this range is thus not only time-consuming but also extremely financially expensive. Further, and perhaps concerning, more than 80% of the costs of taking a compound through to market are due to clinical attrition, so most of that 1.13 billion mentioned above is due to failure along the development pipeline; safety and efficacy account for the majority of these failures (as high as 70% between the two), with lack of efficacy as the major contributor to attrition in the Phase 2 and 3 stages

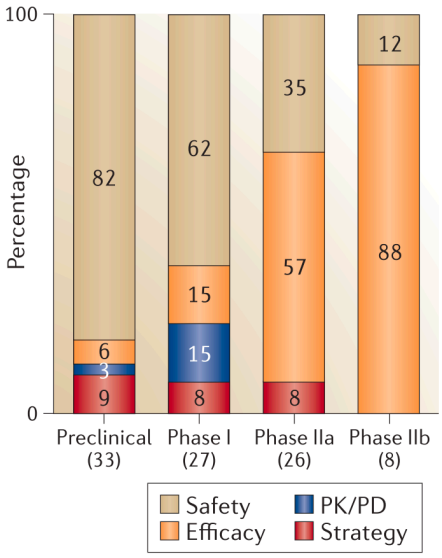



**Figure 1.1.2: Reasons for project closure by phase of development.** Stacked bar plots of the primary reasons behind drug project closures, based on a study of 142 drug discovery and development projects at AstraZeneca, covering projects from all therapeutic areas that had been active in pre-clinical through to and including Phase 2 stages, between 2005–2010, reproduced from the study by Cook and colleagues [4]. The percentage of project closures at each of the indicated phases of development is shown in the bars and the number of projects assessed is shown in brackets below the bars. Reasons for project closure include failure due to safety (toxicology or clinical safety), efficacy (failure to achieve sufficient efficacy), pharmacokinetics/pharmacodynamics (PK/PD) or the strategy.

of development, when committed resource have become quite high [3, 5]. To get a better sense of these numbers, we can refer to the analysis of AstraZeneca's internal pipeline which revealed that 57% and 88% of project closures were due to lack of efficacy in Phase 2a and Phase 2b stages, respectively (Fig. 1.1.2); overall, 65% of failure in Phase 2 was due to lack of efficacy [4].

In other words, the majority of costs of developing a drug are due to failures, and the majority of failures in the stage with the highest rates of attrition are due to the compound having no clinical effect in the patient group. Moreover, while the transition success rate from Phase 2 to Phase 3 was calculated to be 32.4% between 2003 and 2011 [3], the success rate for the same Phase 2 to Phase 3 transition was estimated to be 45% between the years of 1993-2004 [6]. Even taking into account that the studies may have used slightly different definitions or categorizations of compounds, it is nevertheless clear that we as a drug development community have not improved in our collective ability to mitigate Phase 2 to Phase 3 attrition. In fact, of the subset of drugs failing in Phase 2 due to lack of efficacy, around 40% are projects that had no initial link between target and disease to begin

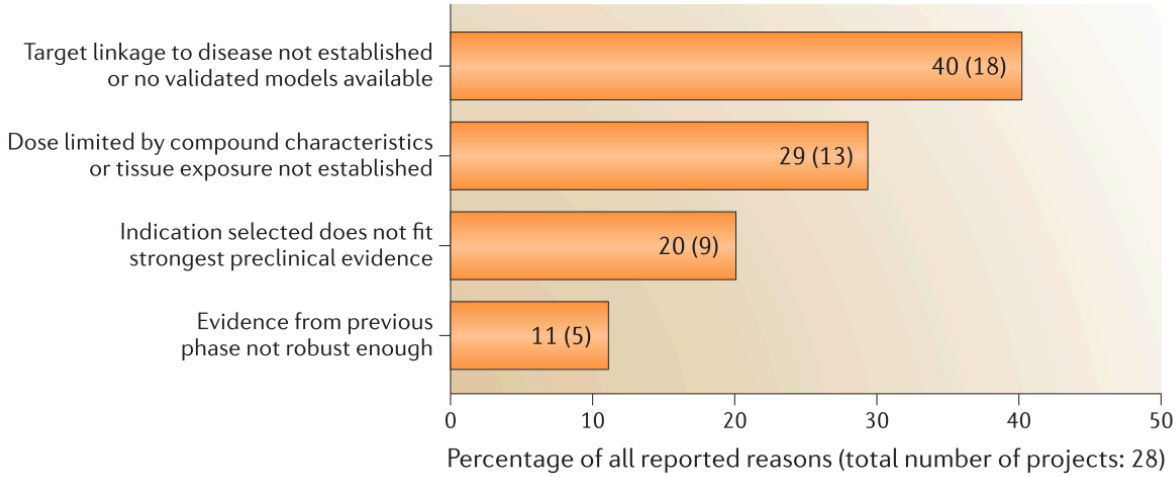



**Figure 1.1.3: Reasons for lack of clinical efficacy.** Reasons why drug projects failed, in instances when failure was categorised as a lack of clinical efficacy, based on a study of 142 drug discovery and development projects at AstraZeneca, covering projects from all therapeutic areas between 2005–2010, reproduced from the study by Cook and colleagues [4]. Projects could have more than one contributing factor for the lack of clinical efficacy observed. Percentages of project responses for each category are given in bars. The numbers in brackets represent the number of projects that reported each category as a contributing factor for the failure of the project.

with [4], which should be very troubling to hear (Fig. 1.1.3).

It is clear that drug development is a time and resource-intensive process, that has only gotten more complicated and less efficient over the last few decades; all this has happened in spite of massive leaps in our collective knowledge and technical capabilities over the last several decades. But how much of an impact on drug development costs can improving the Phase 2 to Phase 3 transition success really have? The short answer: a very substantial one. When drug development costs are analysed in a model taking into account the various levers involved (time, probability of success, productivity, etc.), it becomes possible to toggle certain numbers to gain an appreciation for the downstream effects of making improvements in specific areas within this process. For example, using the model proposed by Paul and colleagues [1] reveals that if Phase 2 attrition can be reduced from 66% to 50%, the cost per novel drug entity can reduce by as much as 25%, which amounts to a saving of approximately *450 million USD per drug* (using the 'lower' price tag of 1.8 billion USD; this value is considerably higher if we instead use the 3-10 billion USD range proposed by Schuhmacher and colleagues [2]). If similar improvements can be made to Phase 3 attrition rates as well, then the results look even more exciting; combining improvements across Phase 2 and Phase 3 (and all else being equal) can mean that the

same level of research and development investment currently being deployed can yield as much as 2-3x more novel compounds in a given time frame [1].

Clinical efficacy remains one of the central, addressable points in this equation; indeed, lowering attrition in Phase 2-3 and Phase 3-4 as mentioned above largely hinges on reducing the number if inefficacious targets in this stages. How this can be achieved remains an open-ended question. However, one logical way forward would be to ask whether existing or emerging tools could be used earlier on in the drug development process (e.g. in pre-clinical stages) in order to better funnel the number of targets proceeding formally through clinical trials. In other words, even if one argues that novel or game-changing drug discovery is like looking for a needle in a haystack, it would nonetheless be extremely favourable to reduce the size of that haystack to begin with. One such way of doing so involves leveraging human genetics for target selection, which this thesis will focus on (human genetics is also being used for determining the right patient population for drug testing based on who is likely to respond to a drug intervention [4], among other things, but this is beyond the scope of the work outlined here). Additionally, this work has taken advantage of emerging scientific translational partnerships; specifically, these open-sourced, public-private-partnerships allow for a marked improvement in drug discovery as a field, by minimizing the chance of needlessly duplicated effort and achieving the right balance between basic science, translational research, and drug development as a process [5]. Examples of successful public-private-partnership stories include the single nucleotide polymorphism consortium (between ten large pharmaceutical companies and IBM, Motorola, and the Wellcome Trust) and the Structural Genomics Consortium (between the University of Toronto, the University of Oxford, the Karolinska Institute, Merck, Novartis, GSK, the Wellcome Trust, the governments of Canada and Sweden, and others) [5]; the latter of the two is one of the driving forces of the work contained within these pages.

## 1.2 Translation of genetic information to informativeness for drug development

How do we know that human genetics will prove helpful in drug discovery, and how can this information be practically useful in a way that allows systematic adoption in the research pipeline? To answer the first of the two questions, we can look retrospectively at drug pipelines in order to understand how human genetics may have been helpful then, particularly in establishing clinical efficacy earlier on.

Promisingly, targets with genetic support for clinical efficacy display a marked increase in probability of successful transition from Phase 2 to Phase 3; one analysis showed that success rates at this transition point was 73% for targets with genetic support, compared to 43% for targets without this data (Fig. 1.2.1) [4]. In fact, the relationship often cited is that, overall, drugs with genetic support for mechanism are twice as likely to succeed from Phase 1 to approval compared to those without [7]. Quite recently, this value was interrogated in greater detail by King and colleagues (2019), who used statistical models to account for various target and indication properties and a collection of data sourced from Informa Pharmaprojects, GWAS catalogue, and OMIM in order to try and confirm this finding [8]. They concluded that targets with genetic support from OMIM, for instance, were as likely and, in some instances, more likely to be successful than previously reported by Nelson and colleagues.

Over the decades, the forms of genetic support for any given target and its role in disease have adapted in tandem with the technologies at our disposal. Tracking the discovery of statins provides a good case example here. The first genetic link between cholesterol and heart attacks was made in the 1930s through the study of co-inheritance of these two phenotypes in large families, and was extensively researched over the next three decades [9]. Following these initial observations, further epidemiological studies were carried out in order to more reliably correlate cholesterol levels to risk of heart attack; importantly, these studies also uncovered the specific role of low density lipoprotein (LDL) in disease risk. These discoveries paved the way for more in-depth molecular and mechanistic

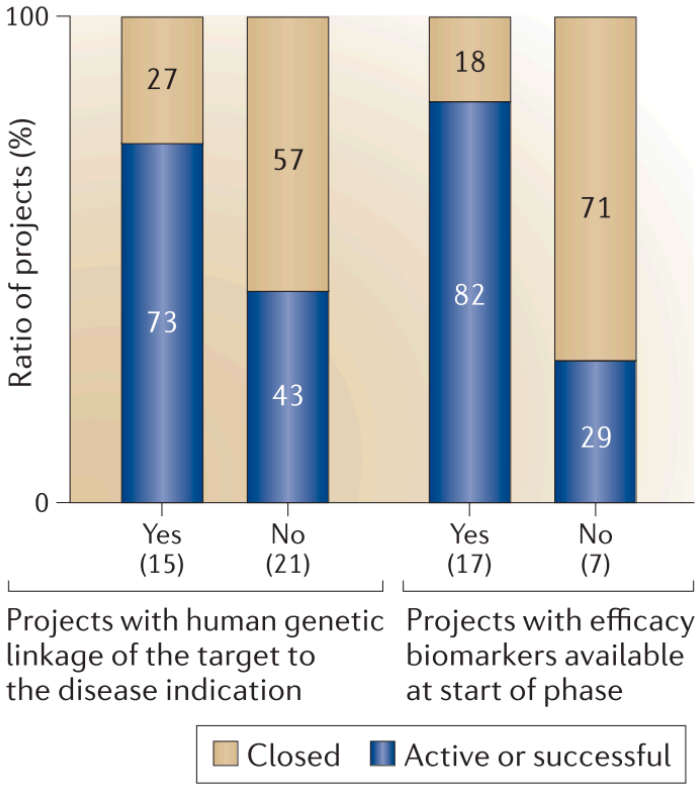



**Figure 1.2.1: Evidence of genetic link between target and disease. Left:** An analysis of all AstraZeneca projects in Phase II (between 2005-2010) for evidence of genetic linkage of biological target to disease indication and whether the project was active/successful or closed, reproduced from the study by Cook and colleagues [4]. **Right:** An analysis of projects in Phase IIa to determine whether efficacy biomarkers were available before these projects entered Phase IIa development. Percentages for each category are shown within the bars and the number of projects are shown in brackets along the x axis.

studies, ultimately pointing to the role of HMG-CoA in LDL and underscoring an opportunity for lowering disease risk via its inhibition; in 1987, thanks to Endo and colleagues, statins were approved by the FDA for clinical use [9].

As advancements were made in genetic sequencing, the links between cardiovascular disease, LDL cholesterol, and HMG-CoA were solidified. In particular, the emergence of genome sequencing following the Human Genome Project meant that the fields of genetics and epidemiology could converge in revolutionary ways via genome-wide association studies (GWASs) that allowed for genetic links to be made at an unprecedented scale. For instance, a common non-coding variant in *HMGCR* was found to be associated with risk of cardiovascular disease in the general population [10]; statins, successfully used since 1987 for lowering LDL cholesterol and reducing cardiovascular disease illness, target *HMGCR*, showcasing how GWAS can in principal uncover knowledge that previously might have taken decades of study to initially link together [7, 11]. Further, although statins emerged in the late 1980s, limited progress was made in the advancement of the cardiovascular disease therapeutics space for many years following [12, 13]. However, the evolving field of genetics led to the implication of *PCSK9* as a target for lowering LDL-C, based on a number of studies that associated variants in the gene with high cholesterol levels [14–17]. More specifically, the initial reported mechanism was gain-of-function whereby mutations decreased the quantity of cell surface LDL receptors resulting in hypercholesterolemia; shortly after, loss-of-function mutations were also reported, this time associated with lower cholesterol. Ultimately, this helped inform the drug pipeline that developed Alirocumab as a PCSK9 inhibitor for lowering cholesterol, including in patients who were not responding well to statins [18–20].

The term *experiments of nature* is often used to describe these approaches. In essence, experiments of nature are allelic associations to phenotypes, resembling a dose-response curve and thus hinting at the potential for therapeutic development [13]. Autosomal-dominant Huntington's Disease and the 'CAG' trinucleotide repeat in the *HTT* gene is a classic example of a genetic dose response, where the number of CAG repeats correlates to likelihood of clinical severity and age-of-onset of disease, in a dose-dependant manner [21]. In other words, understanding the link between a naturally-occurring genetic factor and corresponding presence or severity of disease can be a powerful way to establish

target validity for therapeutic purposes. Further, understanding these effects from the context of loss-of-function or gain-of-function mutations, as in the case of hypercholesterolemia, can be useful in determining whether inhibition or activation should be used in a therapeutic approach, based on the associations between genetics and disease [13]. Another good example of translating human genetic insight into therapeutics is in the context of osteoporosis. Initially, early studies examining the underpinnings of sclerosteosis, a disease characterised by bone overgrowth, discovered loss-of-function mutations in *SOST* which resulted in disease [22–24]. Further, genome-wide studies examining associations with bone mineral density identified a number of variants near the *SOST* gene which were associated with decreased bone mineral density, with evidence of allele dose effect [25, 26]. This paved the way for a new class of therapeutics for osteoporosis, including romosozumab, which was approved for the treatment of osteoporosis [27–30].

So how can we continue to inform this drug discovery process in fruitful ways that ensure further advancement of novel therapeutics, and what is the size of the opportunity? Early estimates predicted that about 10-15% of the human genome was druggable by small molecules, based on current techniques, known protein family characteristics, and binding site information, with varying distributions across protein classes [31, 32]. Follow-on studies that factored in advancements in genome sequencing reached more ambitious conclusions; namely, current estimates sit around 22% of protein-coding genes being druggable [33]. However, this still leaves thousands of candidates across numerous disease indications to untangle. Further, just because a target is druggable does not necessarily mean that there will be an adequate clinical effect, as there needs to be an additional functional link between target and disease. Therefore, even when considering known 'druggable targets', one needs a reliable way to functionally link those targets to disease, at scale.

Thankfully, the growing body of GWASs has provided a good stepping stone for approaching this problem productively. GWAS-identified genes are significantly enriched for druggable targets compared to what is expected by random chance [34]; more specifically, the authors found that 21% of GWAS-identified genes were druggable by small molecules as compared to 17% of genes when considering the entire genome, representing a statistically significant difference between the two groups. This is a vital first consideration when it comes to connecting druggability to functional

impact of a target and ensuring that there will likely be tractability for a target within a clinical pipeline. Importantly, the utility of GWAS-implicated targets extends beyond technical druggability to also include efficacy, which we have already established as a fundamental hurdle within drug discovery. For example, a study of the rheumatoid arthritis (RA) therapeutic landscape found that there was a significant overlap between approved drug targets for the treatment of RA and RA risk genes, and this overlap remained significant even when adjusted for (i) a background overlap with all drug target genes, as well as for (ii) any potential immune-mediated disease bias [35]. In fact, this finding extends beyond RA to most well-studied indications, as approved drug target genes are significantly enriched among genes with identified variants, and the enrichment of genetic support for a target increases further along the drug development pipeline (i.e. drugs in later stages of clinical trials have greater enrichment for genetic support) [7]. In addition to helping with novel target identification, the growing body of GWAS information presents an opportunity for drug repositioning by taking advantage of genetic pleiotropy, or the influence of one gene on multiple phenotypes which can initially seem unrelated. For instance, *TNFSF11* is the target of denosumab for the treatment of fracture in post-menopausal women, but has also been associated with Crohn's disease; thus, denosumab may also be a viable option for the treatment of Crohn's disease [34].

Despite this growing foundation of genetic association knowledge, deconvoluting signals using GWAS - particularly when it comes to the mechanism of complex disease- remains a challenge. This is especially true given what we know now about variants, such as their tendency to fall outside of protein coding regions more often than not [13]; a recent review estimates that nearly 90% of variants lie in non-coding regions [36]. Continuing on the theme of RA treatment, CCR5 antagonism is a good example of instances where human genetic signals may not be sufficient [4]; the target had association to rheumatoid arthritis [37] and even pre-clinical model support [38, 39] but failed to show clinical benefit in patient populations [40–42]. Importantly, the meta-analysis which supported the linkage of CCR5 to rheumatoid arthritis was limited in that it simply took into consideration five combined studies, each of which investigated the association in a fairly low number of patients versus controls in the $n_{cases+controls}$ = 100-700 range [37]. This challenge is not unique to CCR5 and RA; when broadly considering GWAS-identified targets and probability of drug approval, the picture is often hazy, largely due to a poor signal:noise ratio. For instance, while association specifically between

GWAS evidence and likelihood of drug approval shows general positive correlation, the magnitude of association is generally very low and not always significant in independent replication sets [8]. Although this may seem troubling, it is important to remember that broad stroke, simple analysis of GWAS data for drug discovery is only a naive baseline starting point. When the same relationship between GWAS-specific evidence and likelihood of drug approval is analysed with a bit more granularity, for instance, by accounting for variant and linkage disequilibrium properties, the correlation becomes stronger and statistically significant [8].

The potential to resolve GWAS signals from a more robust statistical perspective is one important step in better translating this research to clinic. Building on that, there is general consensus that in order to make the biggest gains in drug discovery in the current 'omics era', we need better overall integrative approaches that combine multiple layers of data points by taking advantage of essential, open-sourced databases and complementary data-sets [5, 43]. Even a basic convergence of GWAS and gene expression can be powerful, for example – an analysis found that the ability of GWAS to recover known disease targets was markedly improved when the data was combined with gene expression information compared to either method on its own [43, 44]. Further, despite our recent advances in data generation and analysis, roughly 40% of the druggable human genome remains therapeutically unexplored [45]. Thus, the key to better identifying therapeutic opportunities within complex disease, and unlocking untapped or under-explored targets, lies in a more systematic and integrative approach to genomic information downstream of GWAS.

## 1.3 Existing tools leveraging genomics for drug discovery

There is no doubt that human genetics has played and will continue to play a pivotal role in the drug discovery industry. Since the milestone of the human genome project, the number of compounds in clinical trials have increased 1.62x while research and development expenditure has gone up 2x [3]. Despite the limitations discussed thus far, progress has been made in better systematic use of GWAS data for drug discovery. For example, Finan and colleagues (2017) leveraged experimental ontologies in order to harmonize clinical nomenclature and more systematically map GWAS variants to target genes, identifying numerous novel links or combinations between disease and putative targets [33].

Further, they used this information to design the Illumina DrugDev array for specifically testing for variation in genes that are known or predicted to be druggable, allowing for future systematic yet focused sequencing of patient populations for the purposes of novel therapeutic discovery [33]. Thus, it has now become possible to conduct drug discovery-focused GWAS as a first-line method, which should yield interesting results in the coming several years as more researchers adopt this type of protocol.

### Open-sourced databases

Thankfully, there have been large strides in open-sourced data for drug discovery with numerous tools and databases emerging, which can be used to complement GWAS studies and inform downstream work. For example, ChEMBL is one of the dominant resources in the field, providing detailed information on target and drug data, including elements like bioactivity, structure, binding sites, strength of interaction, supporting evidence, and others [46]. Similarly, DrugBank has emerged as a useful open-sourced project with similar aims to provide information on chemical and molecular features of various compounds, and a particular focus on drug - target interactions [47]. Both of these resources have also focused significant energy into making sure that the databases are searchable at scale and updated in regular intervals in addition to comprehensive in nature, making them invaluable tools in the drug discovery pipeline [46, 47]. Further, efforts have been made to nuance our understanding of existing drug development pipelines by a more manual canvassing of drugs in development as well as assigning links to disease-relevant pathways on the basis of these drug targets [48]. Moreover, work has also been done to tie in indirect but equally important pieces of information, such as adverse drug reactions and side effects, which are significant factors when conducting novel drug discovery research; databases like SIDER have focused efforts on mining data to isolate adverse reaction events across a wide array of drug compounds and medical terms in a pairwise manner [49].

### Integrative tools

Recognising the limited ability of GWAS data alone in moving the needle on novel drug discovery, and that there is a marked increase in recovering known targets when combining GWAS with gene expression information [43, 44], methodologies to integrate expression quantitative trait locus (eQTL) studies with GWAS have also emerged. For example, Zhu and colleagues (2016) developed a method called SMR that bridges GWAS and eQTL in order to associate changes in gene expression with changes in phenotypic traits, akin to a Mendelian randomization analysis [50]. A major advantage of this approach is that it allows for systematic integration of GWAS and eQTL and separation of linkage, pleiotropy, and causality, all of which are key features in a drug target discovery pipeline. However, the focus means that insight is inherently limited to a convergence of eQTL and GWAS, the former of which has not yet been achieved at large depth and scale across enough cell types, presenting a challenge for translating this insight into broad clinical utility.

Some methods have attempted to overcome the limitations of eQTL studies (as well as existing public dataset annotation bias) by casting a wider net when nominating gene targets. In particular, network-based approaches have increasingly emerged over the last several years, as they allow users to leverage known global interactions in order to propagate direct target evidence to identify potential downstream or upstream targets indirectly implicated by these networks. For instance, Emig and colleagues (2013) implemented a network approach that combined gene expression microarray data with global molecular networks and known drug targets in order to identify opportunities for drug target repositioning [51]. By annotating genes based on relevant changes in gene expression and propagating that signal to neighbours based on algorithmic approaches like random walk (which uses stochastic processes to inform network connectivity and which has been the basis of many widespread tools including the PageRank algorithm behind Google's search engine [52, 53] as well as gene interaction analyses [54]), they were able to implicate indirect genes that may be contributing to disease. These annotated networks can then be overlayed with a catalogue of known drug targets, in order to identify where repositioning potentials may lie [51]. While these approaches have clear advantages, they remain limited by the input data; namely, the method uses gene expression data, in order to allow relatively straightforward systematic collection and normalization using public

repositories and bioinformatic scripts. While this is undoubtedly useful for assessing indirect implications of gene expression changes, the outputs are limited in that they are a reflection of one dimension or layer of support (gene expression) rather than a more holistic collection of factors. Attempts have been made to overcome this limitation by considering more layers of evidence; DEPICT, for instance, defined gene co-regulation datasets that assign probabilistic membership to genes for various biological and molecular pathways as well as phenotypic gene sets, and then look for enrichment of these genes in GWAS data to identify the most likely causal pathways and tissues [55]. However, the use of such *guilt by association* techniques to define gene co-expression relationships has been disputed as a reliable way of determining functionality, as they are sensitive to network outliers that cannot reliably be generalized to the rest of the gene network [56].

Recognising that there has been general consensus that the future of drug discovery revolves around better integration of multiple data sources [35, 43, 57, 58], a number of more complex approaches have emerged. These methods have aimed to combine some of the approaches outlined in the previous few paragraphs, in order to build more robust systems for drug target discovery. For example, a well known study by Okada and colleagues (2014) conducted a gene prioritisation analysis for RA based on several lines of information. Specifically, they prioritised any genes with RA-associated missense variants; any genes that were identified in eQTL analyses of PBMCs, CD4+ T-cells, and CD14+ monocytes; any genes that were identified through PubMed text-mining; any genes that were associated with haematological cancer mutations; any genes with significant protein-protein interactions; any genes with immune-related phenotypic labels in mouse studies; and any genes that were identified in significantly enriched, RA-associated molecular pathways [35]. This approach was an exciting way to systematically prioritise close to 100 biological candidate genes for drug discovery, while collectively considering a wide array of supporting pieces of evidence. However, it was not without its challenges; notably, gene scores were calculated on the basis of the number of supporting pieces of evidence, meaning that each data source was essentially treated as an equally important element in the equation. Further, a number of the predictors were biased to favour genes that have been extensively studied, like the text-mining and phenotypic annotation predictors, or the existing drug protein-protein interactions which, when taken in tandem with the issue of uniform predictor importance, is prone to heavily boosting scores of well-studied genes. Lastly, scoring for each gene

across each predictor was binary – either the gene overlapped with the predictor or it did not, which sacrifices a significant amount of nuance that is normally afforded by the data sources used to construct the predictors.

This challenge of providing nuance to predictors for better drug target prediction was perhaps best addressed by the Open Targets initiative [58], which remains one of the most exciting standards in open-sourced drug discovery. Similarly to Okada and colleagues, Open Targets aimed to integrate a wide array of biological and genomic data in order to reliably prioritise targets with therapeutic relevance. Open Targets brings together disease genetics, cancer mutation information, transcriptomics, clinical trial and therapeutic information, animal models, literature text mining, and pathway information in an organised and easy-to-use tool for pre-clinical discovery. Moreover (and improved from the method proposed by Okada and colleagues), Open Targets takes into consideration frequency, strength, and confidence of evidence for each individual dataset, in order to build a model that weighs predictors on the basis of these three pieces, before outputting a final/overall per-target priority score [58]. Thus, this powerful method allows for a granular yet holistic per-gene score on the basis of uniform handling of individual datasets, combined with the flexibility to augment scores as new data emerges.

### Machine learning approaches

Additional methods have also attempted to overcome some of the issues of early 'holistic approaches' by introducing an element of machine learning into the prioritisation pipeline. For instance, techniques such as support vector machines (SVMs) have been used previously in drug design and prediction of structure - activity relationships [59], and have since been adapted for other forms of target prioritisation. For example, Jeon and colleagues (2014) investigated therapeutic prediction opportunities in the context of cancer. Specifically, they combined gene essentiality, multi-array average expression information, and DNA copy numbers across dozens of breast, pancreatic, and ovarian cancer cell lines, quantified mutation rates compiled from the COSMIC database, superimposed this onto a protein-protein interaction network, and built an SVM classification algorithm [57]. The algorithm was subsequently trained using a set of positives (sourced from

DrugBank) and negatives (meeting any number of a variety of criteria, including lack of annotation in the drug target databases), and putative genes were then systematically categorised using this machine learning model. A major advantage to such an approach is that it allowed the authors to integrate across an extensive variety of information sources, while taking into account and controlling for redundant features that pose a risk of overfitting the model [57].

Similarly, algorithmic methods have attempted to use elements of the data contained within Open Targets in order to train models on known drug information. Ferrero and colleagues (2017) extracted the pathway, animal model, genetic association, RNA expression, and somatic mutation columns, for instance, and annotated each entry based on whether the target was contained within the PharmaProjects database as a drug target; this was then used for training several different types of classification algorithms, including an SVM, a random forest, a neural network, and a gradient boosting machine [60]. The idea is that this final model, which recovers known drug targets fairly well with 71% accuracy, can subsequently be used to inform the likelihood of a novel target succeeding in clinical development. Naturally, the advantages of this approach are in the ability to apply the techniques to large sets of genes in order to form a baseline prediction of success, as well as in the flexibility of the model to be altered as new datasets emerge.

However, the disadvantages are equally important to consider. First, many machine learning approaches (especially neural networks, which Ferrero and colleagues decided to use [60]) suffer from the *black box* problem – the path to reaching conclusions or predictions is often very muddled and hard to decipher from a human-intuition standpoint. This makes it difficult to know what to do with candidate targets from a development point of view. Further, training models on known drug information or including known drug information in the scoring, as done by both Ferrero and colleagues as well as by Open Targets [58, 60], introduces an inherent bias into the output; namely, one that optimizes for 'more of the same' rather than attempting truly bias-free, novel discovery. The authors themselves, for instance, acknowledge that while animal models seem to be the most important predictor of clinical development in their algorithm, this is likely at least in part due to the fact that better animal models are biased towards well studied genes [60]. Indeed, there is much debate in the drug community around the merits of the use of animal models as a central pillar in drug

discovery [3, 13]. Lastly, machine learning approaches increase in performance as they are fed greater amounts of data, incentivizing the pooling of both target sets and predictor sets; however, not all data is likely to be equally important across every disease category, so more granularity is needed in the building of these models.

## 1.4 Hit validation for establishing the clinical efficacy of targets

Optimistically, it is apparent that the last several years have brought a wave of new methodologies to leverage genomics and other datasets for the purposes of improved drug discovery. From smarter, more robust analyses of GWAS, to multi-faceted tools that aggregate diverse data such as gene expression and text mining, and all the way through to novel machine learning algorithmic approaches that learn from prior knowledge, the broad landscape of genomics and drug discovery has shifted significantly in recent years. The prioritisation of novel compounds, however, brings with it the need for validation of these hits, in order to reliably demonstrate a functional effect of inhibiting or agonizing **target X** on **phenotype Y**, in a way that is clinically relevant. Traditionally, the use of animal models has been a substantial element in the drug discovery pipeline; various animal models of complex disease exist, and can serve as proxies for quantifying the likelihood of a successful candidate drug target [61]. However, animal models can be poor predictors of outcomes in human clinical groups – there are a high proportion of targets that show extremely promising results in animal models but translate poorly to patient groups, as they often lack relevance to human disease at the target level [3, 13].

Therefore, there has been a challenge of developing novel approaches to target validation, in a way that maximizes the chance that a positive hit is actually successful in mid or late stage human clinical trials. While GWAS have enabled the study of allelic effects at scale, GWAS by themselves can be limited in validating targets in complex disease due to a number of factors, including but not limited to the relatively small effect size of most SNPs or the inability to detect rare variants using typical arrays [21]. Additionally, almost 90% of alleles associated with complex diseases are not protein-coding and are thus harder to contextualise functionally; this, combined with the fact that complex disease variants rarely work in isolation to affect phenotype, makes truly unraveling the link between allele and

phenotype to establish causality (and thus therapeutic tractability of the corresponding target), a major challenge [13, 36].

Thankfully, there has been a strong emergence of efforts focused on better understanding of complex diseases and improved therapeutic development in these areas, as demonstrated by the establishment of the Structural Genomics Consortium (SGC) or the Unrestricted Leveraging of Targets for Research Advancement and Drug Discovery (ULTRA-DD) consortium [62]. The aim of these initiatives is to build on genomic techniques in order to better and more reliably predict the therapeutic effects of target modulation in disease, in a validated way that is significantly more likely to clear clinical trial development standards. The development of reliable phenotypic screens for drug development has been one major area of focus for the field; the aim of these systems is to understand the effects of target perturbations within a complex assay that reliably recapitulates important aspects of the disease in question [21]. In particular, a shift toward patient-derived cellular assays has been embraced by many scientific groups interested in complementing animal model data with more human-centric approaches. In fact, patient-derived cells and associated assays have previously proven successful in drug target validation, as demonstrated by anti-TNF approaches to treating rheumatoid arthritis, which relied on culturing cells originating from disease tissue in order to establish key findings [62, 63]. More recently, Jeon and colleagues (2014) tested their computational classification algorithm by generating inhibitors against prioritised targets, which were then delivered to specified human cells via lentiviral methods (in the case of peptides) and via chemical compound libraries (in the case of small molecules), with cell viability as the chosen output measure given the cancer context [57]. In addition to compound screening within disease-specific assays, other approaches have also been increasingly utilized. For example, RNA approaches offer the significant advantage of specificity compared to compound screen – siRNA studies with well-crafted controls therefore provide an attractive opportunity for target validation studies [61], as they allow for specific modulation of a candidate target, in a disease-relevant context, which can subsequently be used to establish a firmer link between inhibition and phenotypic effect.

## 1.5 Aims of this work

It is clear that we need to reverse the trends of the drug discovery industry if we want to continue developing novel therapeutics at a sustainable rate. Fortunately, the meteoric explosion of genomic technologies and techniques in the last decade or so has provided access to an unprecedented scale of data at our disposal. While early strides have been made in translating this to clinical utility, obstacles remain around how to best leverage these genomic insights into therapeutic opportunity. In particular, the branch of complex immune-mediated disease presents an interesting challenge in identifying tangible signal from the swath of noise that accompanies it. Further, compelling target identification or prioritisation necessitates some form of experimental validation, in order to de-risk novel clinical development tracks and avoid sunk costs later down the pipeline.

Thus, the aims of this thesis are to bring these elements together using a combination of approaches, showcasing how they can be implemented and assessed at each point in the pipeline. More specifically, I will focus on four integral component aims in my four main results chapters:

I. Performance assessment of a novel prioritisation pipeline, Priority Index [64], in order to provide computational support and validation for a new way of analysing genomic data. This uses a combination of GWAS, genomic annotators and, given the compelling evidence for the propensity of drug targets to be located in central network positions [57], protein network information;

II. Application of the genetic-led target prioritisation algorithm to a novel disease case, myasthenia gravis, in order to demonstrate the utility of using such an approach in early, pre-clinical stages of drug development;

III. Experimental validation of highly networked, genetic-led prioritised epigenetic targets in Dupuytren's disease, focusing on broad gene and protein-level changes, as they pertain to disease progression;

IV. Experimental validation of prioritised targets in systemic lupus erythematosus, a disease with more established genomic underpinnings, in order to provide evidence in support of the

overall computational approach of using genetic evidence in de-risking single-target genes for therapeutic consideration.

The thesis was thus structured in order to capture the longitudinal utility of this computational approach to leverage human genetics in the context of drug discovery, a task which would have essentially been impossible to do otherwise (i.e. if the focus was on one case, start to finish) given the very long term nature of drug discovery as a whole. Naturally, it is my hope that the work and narrative presented here will be compelling enough to advance the field in the direction of better understanding and systematic use of genomic information for novel therapeutic discovery and prioritisation for development.

*Science is a way of equipping yourself with the tools to interpret what happens in front of you.*

Neil deGrasse Tyson

# 2

# Materials and methods

In this chapter, I outline key methods used throughout the thesis; specifically, I explain methods relevant across multiple chapters. Any chapter-specific adaptations made to any of the methods outlined here are explained within the relevant results chapter – for instance, the Priority Index pipeline is explained below, while any customised alterations for the purposes of a specific disease use-case would be mentioned in the corresponding results chapter.

## 2.1 Disease-specific input SNPs

Many of the analyses done throughout the work rely on a collection of disease-specific SNPs, either for direct analysis as is the case with eQTL overlap, for instance, or for indirect analysis by first using these SNPs to inform the Priority Index pipeline [64], and then utilising this output for further work. Therefore, SNP collection needed to be done carefully since it formed the foundation of most analyses. Primarily, SNPs were sourced from the GWAS Catalog [65, 66], which contains a manually

curated list of publicly available SNP-trait associations. This approach allowed systematic and scalable collection of disease-relevant SNPs and p-values, which could then be used as input in various ways. In general, SNPs for computational prioritisation using Priority Index were restricted to those associations with statistical significance of $p < 5 \times 10^{-8}$ unless otherwise stated. In instances where analysis was done using SNPs with reported p values greater than the cutoff used by GWAS Catalog (i.e. $p < 5 \times 10^{-5}$), or with SNPs from very recent publications not yet represented in GWAS Catalog, variants and p-values were sourced directly from the studies in question (both main figure tables or supplementary material). More details on disease-specific GWA studies used are below.

### Multiple sclerosis

The primary study used for sourcing multiple sclerosis SNP associations was the recently published paper by the International Multiple Sclerosis Genetics Consortium [67]. This consisted of 47000 cases and 68000 controls of European descent, which identified 233 statistically independent associations.

### Myasthenia gravis

Three main studies were used for sourcing myasthenia gravis-associated variants. This included the 2012 study by Gregersen and colleagues [68] which recruited 649 patients of North European ancestry and 2596 controls of European ancestry; the 2015 study by Renton and colleagues [69] which consisted of 1032 patients and 1998 controls as well as an independent cohort of 423 patients and 467 controls, of European ancestry; and the 2015 study by Seldin and colleagues [70], which was conducted with 532 cases and 2128 controls, all of European ancestry.

### Dupuytren's disease

The primary source of input SNPs for the analysis of Dupuytren's disease was the 2017 genome-wide study across three populations; 5408 British cases, 2195 Danish cases, and 768 German cases [71]. This is the largest Dupuytren's GWAS to date, which validated all 9 previous disease-associated

variants along with 17 additional ones for a total of 26 SNPs meeting a statistical significance of association of $p < 5 \times 10^{-8}$.

#### Systemic lupus erythematosus

A number of studies were used for the sourcing of systemic lupus erythematosus-associated variants, primarily based on information available in GWAS catalog [66] and restricted to input SNPs with association significance of $p < 5 \times 10^{-8}$. This included the 2008 study by Kozyrev and colleagues, with 794 European cases [72]; the 2008 study by Harley and colleagues, with 3037 European cases [73]; the 2008 study by Hom and colleagues, with 4651 European cases [74]; the 2008 study by Graham and colleagues, with 2586 European cases [75]; the 2009 study by Gateva and colleagues, with 9169 European cases [76]; the 2012 study by Lee and colleagues, with 4948 European cases [77]; the 2013 study by Martin and colleagues, with 10911 European cases [78]; the 2014 Armstrong study, with 2000 European cases [79]; the 2015 study by Demirci and colleagues, with 1148 European cases [80]; the 2015 study by Bentham and colleagues, with 14267 European cases [81]; the 2016 study by Márquez and colleagues, with 16546 European cases [82]; the 2016 study by Morris and colleagues, which included 10995 European cases [83]; the 2017 study by Langefeld and colleagues, which included 18264 European cases [84]; and the 2018 study by Julià and colleagues, with 16966 European cases [85].

### 2.2 Priority Index

Gene prioritisation was done using Priority Index [64], or Pi, implemented in R. Pi takes as input a set of disease-specific SNPs and p-values from corresponding GWA studies; these studies were sourced from the GWAS catalog [66], and the default input was SNP – p-value pairs for any variants meeting genome-wide significance (i.e. $p < 5 \times 10^{-8}$). Additionally, any SNPs in linkage disequilibrium ($r^2 > 0.8$) with input variants were included in prioritisation, based on 1000 Genomes Project data for the corresponding population in question.

Once input SNPs were specified, mapping to relevant genes was done to define genomic seed genes,

across three categories: nGenes, cGenes, and eGenes. nGenes were defined as genes near the SNPs in question, taking into consideration topologically associated domains generated for GM12878 based on immune relevance [86], and scoring on the basis of distance from the SNPs. Given that proximity to genes is not the only factor relevant to functional effects of SNPs, cGenes were next defined on the basis of physical interaction. Specifically, cGenes were identified by mapping input SNPs to gene promoters that they physically interact with, based on promoter-capture Hi-C studies in various immune-relevant cell types [87]. The last class of genomic seed genes, eGenes, was determined by mapping input SNPs to genes where there was a significant association with gene expression, sourced from across a number of eQTL studies in relevant immune cell types [88–90].

Next, any seed genes identified with the above genomic predictors were further scored through the use of annotation predictors. These predictor classes included fGenes, dGenes, and pGenes; fGenes (immune function genes) consisted of genes annotated to an immune response based on experimental data, as defined using Gene Ontology [91]; dGenes (disease genes) were defined as genes related to rare genetic disease as defined by Online Mendelian Inheritance in Man [92] as well as annotated to immune system disease more broadly, as defined by Disease Ontology [93]; and pGenes (immune phenotype genes) included genes annotated to immune system abnormalities via Human Phenotype Ontology [94], and to immune/hematopoietic system phenotypes using Mammalian Phenotype Ontology [95].

Given the indirect functional effects that seed genes could have based on other gene proxies, non-seed genes were identified in order to propagate the seed gene scores across the entire gene network. This was separately done for each type of seed gene, and involved superimposing seed gene scores onto a global gene interaction network, defined by the STRING database [96] and using only high-confidence STRING interactions (which corresponded to approximately 15,000 gene nodes), and then applying a Random Walk with Restart algorithm [52, 54]. This way, each seed predictor type (nGene, cGene, eGene, pGene, dGene, fGene) would capture both direct and indirect functional effects, and all scores could be converted to a value that is comparable across predictor types (i.e. the network seed gene affinity score). Lastly, these affinity scores were combined across all predictors (similar to a meta-analysis approach) to create a final gene priority list for a specific disease with the

given GWAS input data.

## 2.3 Annotator datasets

A number of genomic annotator datasets were used for the scoring of genes in the Priority Index pipeline, based on the functional effect that disease-associated variants may be exerting. More specifically, cGenes and eGenes were identified based on mapping of SNPs to genes on the basis of physical promoter interaction and association with gene expression, respectively.

For cGenes, data was sourced from the 2016 study by Javierre and colleagues [87]. In brief, the study conducted promoter capture Hi-C analysis on 17 primary blood cell types, including megakaryocytes, erythroblasts, neutrophils, monocytes, Mo-M2 macrophages, endothelial precursors, naive B-cells, total B-cells, fetal thymus, naive CD4+ T-cells, total CD4+ T-cells, non-activated total CD4+ T-cells, activated total CD4+ T-cells, naive CD8+ T-cells, and total CD8+ T-cells; each promoter capture HiC experiment had three or more biological replicates. In total, the authors identified 31253 promoter interacting regions – on average, 175000 high-confidence interactions were detected per cell type. Further, principle component analysis of the interaction scores revealed high clustering between biological replicates and high dispersion of individual cell types, supporting the high degree of cell-type specificity of the interactomes.

For eGenes, data was sourced from a number of published papers reporting on eQTLs and focusing on a range of different immune cell types. These included the 2012 study by Fairfax and colleagues done in monocytes and B-cells, with a total of 288 samples [88], the 2015 study by Naranbhai and colleagues done in neutrophils, with a total of 101 samples [89], the 2017 study by Kasela and colleagues done in T-cells with a total of 313 individuals [90], and unpublished natural killer cell data from 228 healthy individuals of European ancestry. In brief, each study collected blood from their respective donors and separated/sorted cells in order to isolate their corresponding cell types in question. RNA was extracted from these cells and gene expression array was used to quantify the level of gene expression for each sample. Further, genomic DNA was extracted from the whole blood and genotyped. eQTL mapping was then done in order to examine the association between alleles and

gene expression, and the statistical significance was reported as a false discovery rate, to control for multiple testing.

## 2.4 Databases

### Golden standard target sourcing

The primary source of data for drug information was ChEMBL [46, 97], which collectively contains drug names, phases of development, indications, and target information; the various databases were downloaded and organised locally. For a given disease analysis, the indication was first queried within the ChEMBL drug-indication tables using SQL via Sequel Pro (version 1.1.2). Any resulting drugs were then queried against the data frame containing target information, via their drug identifier number, which could be used to map drugs to corresponding targets across all possible combinations. Next, the corresponding UniProt identifier was identified for each target in question; these values were then mapped to downloaded UniProt data [98], in order to associate the protein targets to gene ID numbers. Lastly, GeneID's could be mapped to Entrez Gene symbols [99], resulting in a final master table containing drugs, phase of development, protein targets, and gene targets, for any given indication.

### REACTOME

The REACTOME database [100] was the primary source of pathway information; more specifically, REACTOME was used to identify pathways and define the genes that comprise those pathways. REACTOME is an open-sourced tool of manually-curated human pathways and processes; pathway annotations are done by expert biologists, and capture over 12000 human reactions. In total, this comprises over 2200 pathways of 11000 proteins and 12000 complexes, supported by 30000 literature references. Pathway information and visualisation options are available on https://reactome.org.

### STRING

The STRING database [96] was the primary source of network information. More specifically, STRING was used for the sourcing of gene network information critical to the Priority Index pipeline. STRING contains information for over 2000 organisms in total; for homo sapiens, there are 19257 proteins captured, along with the corresponding network connections. These protein-protein interactions include both direct, or physical, interactions, as well as indirect, or functional, interactions. Information on proteins interactions and visualisation options are available on https://string-db.org.

### Pathway Commons

In addition to the STRING database, Pathway Commons [101] was also used for capturing global gene interaction for the purposes of Priority Index network affinity calculation. This database integrates information from a number of sources; in total, it captures 4900 pathways and 2.3 million interactions, covering 18400 genes. Information on interactions is available at https://www.pathwaycommons.org.

### Mouse Gene Informatics

The Mouse Gene Informatics (MGI) Consortium includes a collection of resources relevant to mouse genetics, such as the Mouse Genome Database and MouseMine [102, 103]. These tools contain data pertaining to mouse genes, gene function, phenotype, and mouse models, and can be accessed via http://www.mousemine.org/mousemine/begin.do. For the purposes of this thesis, MGI data was used to map mouse genes to human orthologues, in the *Application of the prioritisation pipeline to myasthenia gravis* chapter.

## 2.5 Precision-recall and receiver operating characteristic curves

Both precision-recall (PR) and receiver operating characteristic (ROC) curves were generated using the R package 'ROCR' (version 1.0.7). For PR curves, this involved using the *prediction* function which takes the following form:

*PR-curve-setup <- prediction(predictions, labels)*

where **predictions** is a vector of the Priority Index ranking of the targets in question, and **labels** is a vector with the tags to denote whether the targets are golden standard positives. The output object, **PR-curve-setup**, is then passed through the **performance** function, where the user denotes that the subsequent output should be a precision versus recall calculation, which can then be plotted to visualize the results.

Similarly, ROC curves were generated using the same initial step as above, using the **prediction** function; the notable difference with ROC curves is that the **labels** argument denotes whether a target is a gold standard positive or a gold standard negative (i.e. two classes rather than one). The output of this test is then passed through the **performance** function, this time specifying true positive rate and false positive rate as the parameters for plotting.

## 2.6 The hypergeometric test and gene set enrichment analysis

Perhaps the most frequently used analysis in this body of work (in addition to Pi prioritisation) is the enrichment test. Unless otherwise specified, this utilized the hypergeometric test to answer critical questions about what information a set of output objects is capturing. At its most basic, an enrichment test is asking the following question:

**Does this specific set of objects contain any particular set of features more often than expected by random chance?**

The objects in question can be any number of things – in this thesis, these are primarily genes and

proteins, but this can be defined in any number of ways depending on the question being asked. Likewise, the features in question can be defined according to what the question in mind is; for example, it may be a set of disease biomarkers or drug targets. Therefore, enrichment tests were used to answer the question of whether a set of objects being analysed significantly overlapped with a set of objects belonging to a particular feature family. Each feature of an enrichment test is defined below, along with a conceptual example of these elements (written in red). Accordingly, the key elements of an enrichment test can be broken down as follows:

**Defining the question:** This is a critical first step to inform the downstream parameter choices of an enrichment test. More specifically, this involves clarifying what is being tested for as well as why, in order to robustly define the parameters that ultimately impact the result of the test. Every Friday, Dumbledore invites one Hogwarts student to dine privately with him. He claims that the decision on which students to invite for the year is random, selected by a draw of 52 names from the Sorting Hat. You suspect that Dumbledore is secretly picking favourites so you want to analyse his 'random' choices using an enrichment test.

**Defining the positive objects:** Effectively, this specifies which objects, from a full set of possible objects, are considered 'positive hits'. In other words, which subset of objects would be 'interesting' to find in the set of outputs being analyzed, based on some common feature(s)? Based on a number of observations over the course of Dumbledore's tenure as Headmaster (such as him finding every excuse to award bonus House Points to Harry Potter and friends), you suspect that Dumbledore is strongly biased towards Gryffindor student. In total, there are 50 Gryffindor students this year at Hogwarts, which will constitute your positive set.

**Defining the background:** Next, and equally important to defining the positive set, is defining the background set (full set of possible objects). This includes both objects that are in the family of features of interest, as well as all other objects not in that family. In other words, what are all the objects (positive or negative) that are possible to find in this experiment? Often, this total pool is a subset of all objects in existence rather than a complete list. In addition to the 50 Gryffindor students, there are 50 from each other house (Slytherin, Hufflepuff, and Ravenclaw), along with 25 teaching staff, 20 mythical beasts, and 5 goblins residing in Hogwarts, for a total of 250 magical beings.

However, Dumbledore only invites students to his weekly dinner, so the background in your enrichment test is defined as the 200 students from the four houses.

**Defining the overlap:** This involves quantifying how many objects from the testing set (i.e. the experimental group) contain objects from the positive set defined above. In other words, from all objects belonging to the group being analysed, how many are considered positive hits based on the features identified in the *Defining the positive objects* step? After Dumbledore released his yearly list of dinner invitees, you took a look through and noted that, of the 52 students selected (i.e. the testing set), 34 are Gryffindor students. This means that 34/52 dinner guests overlap with your positive set.

Once these key elements have been defined and quantified, they can be used to run an enrichment analysis using a hypergeometric distribution, which is a discreet probability distribution to describe the likelihood of successfully 'drawing' a certain number of objects with specified features from a pool of finite size, without replacement. Put differently, the hypergeometric distribution can be used to calculate the probability of a feature being over or underrepresented amongst a subset of objects, based on the prevalence of that feature in the total population being drawn from. In the case of Dumbledore's Weekly dinner, mentioned above, a hypergeometric test would be run (in R) as follows:

*>phyper(q = 34-1, m = 50, n = 150, k = 52, lower.tail = FALSE)*

where **q** refers to the overlap between the experimental set and positive set (34 Gryffindor students were on the final dinner list), **m** refers to the total size of the positive set (50 Gryffindor students are in Hogwarts in total), **n** refers to the background, excluding the positive hits (there were 150 non-Gryffindor dinner candidates), **k** refers to the draw size, or the total size of the experimental set (52 students in total were selected for the yearly dinner schedule), and **lower.tail = FALSE** refers to the fact that we are interested in understanding the probability of selecting at least as many positives as were selected in the experimental set (i.e. what are the chances that at least 34 Gryffindor students were actually selected by chance), which is why **q** actually represents the overlap - 1, so that the upper tail of the calculated distribution will include the observed (34) positives.

Computing the above function gives a probability of $p = 1.094 \times 10^{-13}$ of actually observing 34 Gryffindor students on a list of 52 randomly selected Hogwarts students from a total pool of 200

students containing 50 Gryffindors in total. In other words, the dinner guest list is significantly enriched for Gryffindor students, and very statistically unlikely to have occurred by random chance!

Fig. 2.6.1 illustrates the above concepts; in this example, **Fig. 2.6.1A** shows a scenario where there is no statistically significant enrichment while **Fig. 2.6.1B** shows a scenario of a statistically significantly enriched set. In the case of gene set enrichment tests (used in different forms throughout this thesis), the question usually being asked is something along the lines of: are the genes captured by this analysis or affected by this probe enriched for any particular features (these features could be, for example, a specific pathway membership, which would indicate that the output in question is enriched for particular biological pathways)? Elaborating on this, the **positive objects** would be all gene members of any pathways of interest; the **background** would be all genes relevant to the context of the experiment (for e.g. if differential gene expression in Th17-cell treatment is being measured as the output, then an adequate background might be all genes that could be expressed in Th17 cells, and not simply all possible genes); and the **overlap** would be all significantly differentially expressed genes following treatment, which are also members of the pathways of interest.

The **xPierGSEA** function in the *XGR* package [104] (various versions, spanning from 2015-2019) was used for computing gene set enrichment analyses, the **xEnricher** function for computing pathway enrichment analyses, and the **xEnrichCompare** function for comparing two or more enrichment analyses; all were implemented in R.

## 2.7 Tissue-level variant enrichment analysis

In order to understand the functional implication of genomic regions harbouring disease-associated SNPs, tissue-level enrichment analysis was conducted in the *Application of the prioritisation pipeline to myasthenia gravis* chapter. Specifically, GARFIELD [105] was used to analyse disease-associated SNPs and the regulatory regions they fall in, from a tissue-specific point of view. GARFIELD takes as input a set of disease-associated SNPs and their association p-values, the genomic coordinates of regulatory regions of interest (sourced from the Encyclopedia of DNA Elements [106] and the Roadmap Epigenomics [107] projects), and linkage disequilibrium tags for the reference population

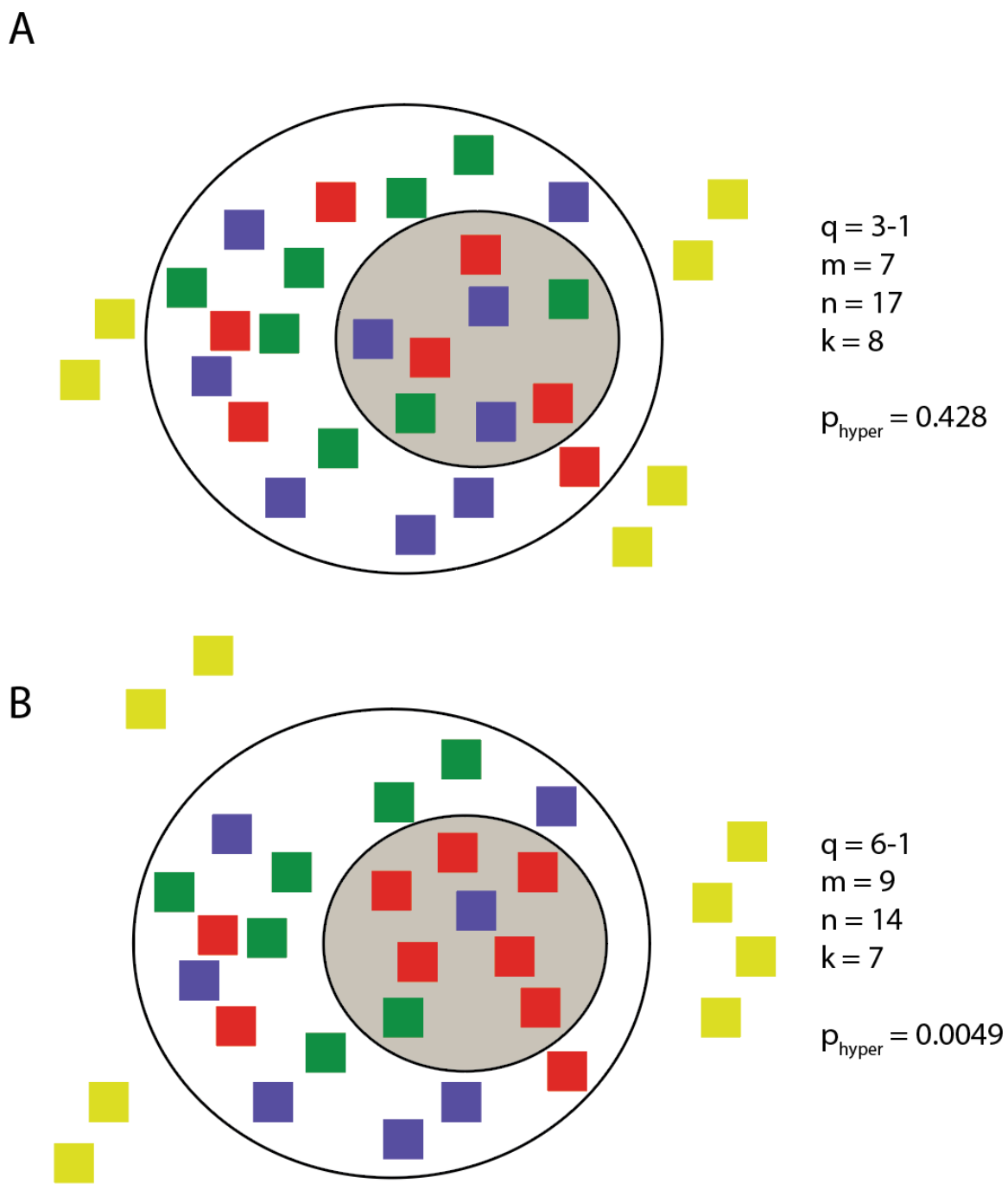



**Figure 2.6.1: A conceptual overview of an enrichment test** A simplified schematic of the premises behind an enrichment analysis, in this instance used to determine whether the experimental group (grey circle) is enriched with red squares. The white circle represents the full, background set – in both panels, yellow squares fall outside of the experimental context, and are therefore not counted in the background. **(A)** The overlap between the experimental subset and objects with the desired feature (i.e. red squares) is three; based on the hypergeometric distribution, there is a 42.8% chance that the three red squares were drawn by random chance. **(B)** The overlap between the experimental subset and objects with the desired feature (i.e. red squares) is six; based on the hypergeometric distribution, there is only a 0.49% chance that this result would be observed by random chance.

of interest. GARFIELD then 'prunes' this input set of SNPs using 'greedy pruning' where any variants in LD of $r^2 > 0.01$ are removed from the input list. Next, the algorithm computes a value called $M_{eff}$ which represents total enrichment of input signal (regardless of tissue source) as compared to a random set of SNPs, to calculate how significantly enriched with regulatory sites the input variant list is overall. Further, enrichment for regulatory regions is calculated per tissue; the $M_{eff}$ value is used as a threshold for adjusting individual p-values for each cell type, with the criteria being that individual enrichment peaks are called as tissue-wide significant if their calculated p-values are more significant than the significance of the total overlap of all tissue regulatory sites and all input SNPs (i.e. if tissue-level enrichment is $> M_{eff}$). Results of this analysis can then be visualised for each type of regulatory element and across all tissues; plots can be overlaid to account for multiple GWAS p-value cutoffs (i.e. simultaneous plotting of enrichment signals considering any number of p-value cutoffs for GWAS variant associations).

## 2.8 Linkage disequilibrium

In order to identify variants in linkage disequilibrium (LD) with other variants (e.g. SNPs in LD with input SNPs for Priority Index processing, such as in the *Application of the prioritisation pipeline to myasthenia gravis* chapter), the SNAP [108] tool was used. Effectively, this tool takes as input a list of variants, a population in question (e.g. European ancestry), and an $r^2$ threshold, and returns a set of variants that are in LD with the input SNP (meeting the specified $r^2$ threshold), based on information from the HapMap project [109].

## 2.9 Statistical fine-mapping

In order to conduct statistical fine-mapping of variants of interest in the *Application of the prioritisation pipeline to myasthenia gravis* results chapter, the PAINTOR tool [110] was used. PAINTOR is a Bayesian framework that combines various functional annotations (such as whether variants fall within enhancers, areas of open chromatin likely to be involved in gene expression regulation, etc.) with genetic association data; this added step allows for tuning of the fine-mapping algorithm on the

basis that SNPs falling in cell-specific regulatory regions are more likely to be influencing mechanism compared to those that are not. PAINTOR implements a variant of the Expectation Maximization algorithm to perform maximum likelihood estimation and to estimate the probability of each variant being causal, while allowing for the possibility of multiple causal SNPs at each locus. Functional annotations are used to adjust the weighing of the scores for SNPs that fall within regulatory regions. The input into PAINTOR are the GWAS summary statistics for the disease in question along with the list of relevant regulatory regions to be considered for functional annotation. For fine-mapping of the CTLA4 locus, these included RoadMap assayed narrow peaks [107] in H3K27ac, H3K27me3, H3K4me1, and H3K4me3, as well as DNase I hypersensitivity sites (hotspots) [111]. The output of PAINTOR is a table with the input SNPs and their posterior probabilities of being causal.

## 2.10 CD4+ cell eQTL analysis

Following statistical fine-mapping of the CTLA4 region, the two most probable causal SNPs (rs231770 and rs231735) were analysed for association with *CTLA4* gene expression. Specifically, an eQTL analysis was conducted for the two SNPs, in CD4+ T-cells based on data generated by Kasela and colleagues [90]. Briefly, CD4+ cells were isolated and purified using magnetic-activated cell sorting and microbeads, from peripheral blood mononuclear cells of 313 healthy European individuals. RNA was extracted for genome-wide gene expression analysis and genotyping was done for each individual donor; imputation via the 1000 Genomes reference panel was also conducted. In total, 6 million SNPs and expression from 38,839 probes in 23,704 genes were included in the analysis, following quality control and filtering. Cis-eQTL calculation was done by testing for association between SNPs and gene expression within 1Mb.

## 2.11 Colocalisation analysis

Recognising the value of colocalisation analysis when integrating GWAS and eQTL data, colocalisation analysis was conducted for the two SNPs (rs231770 and rs231735) being studied in the context of myasthenia gravis GWAS and CTLA4 eQTL signal in CD4+ T-cells. The most

widely-adopted method, coloc [112], was used for this analysis; for allele-matched SNPs within a region (gene), coloc implements a Bayesian framework to estimate the posterior probability that a SNP is causal in both GWAS and eQTL studies/traits (referred to as hypothesis 4, or H4). The default priors in coloc were used ($1 \times 10^4$ for association with either trait; $1 \times 10^5$ for association with both traits). The output of coloc is the posterior probability that the alleles in question are (i) not associated with neither the GWAS nor the eQTL; (ii) associated with one trait but not the other; (iii) associated with both traits via two independent SNPs; (iv) associated with both traits via a shared SNP.

## 2.12 Leading edge analysis

In addition to performing discrete enrichment analyses (e.g. focused on top 1% Pi output genes), certain results were followed up by a leading edge analysis [113]. This approach is useful for dissecting which particular genes from an enriched gene set are driving the enrichment signal most strongly. Effectively, this works as follows: a ranked list of genes (i.e. the ranked Pi MG output) is compared to a gene set of interest (e.g. differentially expressed genes in active vs remission-state MG) by taking steps down the ranked list. After each step is taken, a running-sum statistic is calculated based on whether or not the gene encountered in that step belongs to the gene set of interest, effectively asking 'how enriched for targets of interest is the ranked list thus far?'. This is repeated until the steps have gone all the way down to the end of the ranked list, and the running-sum statistic (also known as the running enrichment score, RES) can be plotted against the list ranks. Normally, the RES gradually increases with each step to a certain point (assuming there is enrichment), after which is begins to gradually decrease. The term 'leading edge' refers to the set of genes up until the peak of this plot (i.e before the RES starts decreasing), and can be interpreted as the core body of genes of interest that are driving the enrichment signal. Significance is calculated by using a permutation test that scrambles gene labels and recalculates the enrichment score, thus generating a null distribution against which the observed enrichment score can be compared. Leading edge analysis was conducted using the *xPierGSEA* function in Priority Index [64], implemented in R.

## 2.13 Structural Genomics Consortium probes

The Structural Genomics Consortium (SGC), a public–private partnership between the Karolinska Institute, the University of Oxford, the University of Toronto, and a number of pharmaceutical partners, focuses on uncovering the structure and understanding the function of various proteins relevant to human health [5]. One important aspect of work within the SGC is the development of probes against a number of epigenetic targets, referred to as the Epigenetic Probes Project. The goal of this project is to understand the biology of epigenetic proteins by developing small-molecule inhibitors and antagonists that meet a number of criteria. Specifically, probes must have an $IC_{50}$ or $K_d$ < 100nM, selectivity > 30-fold over proteins in the same family, and significant cellular activity at 1$\mu$M. The full set of epigenetic probes was screened in the *Gene-level experimental validation of prioritisation outputs in the context of systemic lupus erythematosus* chapter, and the CBP30 probe specifically was used in the *Genomic validation of prioritisation outputs in the context of Dupuytren's disease* chapter.

## 2.14 RNA-seq analysis for Dupuytren's disease

Nodular tissue was obtained from patients with Dupuytren's disease undergoing dermofasciectomy, in line with previous experiments [114]. Informed written consent was obtained through approval of the regional ethics committee (REC 07/H0706/81), in accordance with the Declaration of Helsinki.

Tissue samples were dissected into small pieces and digested in DMEM (Gibco,UK) with type I collagenase 4mg/ml (Worthington Biochemical Corporation) and DNase I (Roche Diagnostics) for up to 3 hours at 37°C. Digested tissue fragments were filtered through a 70-µm cell strainer, and the cell suspension was centrifuged at 1500 rpm for 10 minutes. Isolated cells were cultured in DMEM with 5% FBS and 1% penicillin–streptomycin and used until passage 2. Myofibroblasts were then treated with either the SGC-CBP30 probe (2.5µM) or with DMSO for three days.

RNA was isolated from cells using the Quick-RNA MiniPrep kit (Zymo) according to the manufacturer's protocol, and sample quality was verified by electrophoresis on Tapestation (Agilent). RNA-seq libraries were prepared using the NEBNext UltraTM RNA library prep kit for Illumina

according to the manufacturer's protocol. The concentration of each library was determined using Agilent Tapestation. The resulting libraries were sequenced on a NextSeq 500 platform (Illumina) using a paired-end run 2 x 80 bp, to an average depth of 20 x 106 paired-end reads/sample. The concentration of each library was determined using Agilent 2200 Tapestation.

For RNA-seq experiments, a computational pipeline was modelled off of one from the CGAT toolkit [115] to analyze the next generation sequencing data. Specifically, FastQC v0.11.8 [116] was used to verify the quality raw reads. Sequencing reads were then mapped to the reference human genome sequence (GRCh37 (hg19) assembly) using HISAT2 v2.0.5 [117]. To count the reads uniquely mapped to individual genes, featureCounts v1.4.6 was used [118]. The resulting table with counts was used for differential expression analysis, which was done with DESeq2 v1.12.3 [119]. Differential expression was calculated between the CBP30 probe versus DMSO groups, as well between myofibroblasts and donor-matched palmar fibroblasts groups.

## 2.15 Proteomic analysis for Dupuytren's disease

In addition to RNA-seq experiments, a proteomic analysis was conducted to compare CBP30-treated and DMSO-treated myofibroblasts. Briefly, methanol/choloroform extraction was done for protein precipitation, followed by overnight digestion in 0.2 µg/ml Trypsin (Promega) at 37 °C and a SPEC-PAK C18 based purification protocol. Peptides were analyzed by nanoscale ultraperformance liquid chromatography coupled to tandem mass spectrometry using a Dionex Ultimate 3000 coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). Label-free quantification of proteins was performed using Progenesis QI for Proteomics (version 4.1, Waters) and proteins were identified using MASCOT (Matrix Science) by searching against the Uniprot reference human proteome.

## 2.16 B-cell maturation assay

### Sample collection and siRNA transfection

Blood was collected from healthy donors and came from the blood bank and the Karolinska Institute University Hospital; Buffycoat of each donor was ordered one day prior to the experimental protocol. Buffycoat was carefully added into Falcon tubes with Ficoll-Paque Plus (GE Health Care Life Sciences, Cat#17-1440-03), which were then centrifuged for 25 minutes at 500G. The mononuclear layer between the Ficoll and plasma was transferred to new Falcon tubes with PBS (Sigma Aldrich, Cat#D8537, Lot#RNBG6035/8372) and spun down at 400 G for 8 minutes. Samples were then washed and spun down twice more. Platelets were depleted with two 10 minutes spins at 200 G, pellets transferred to another tube and resuspended in PBS.

siRNA transfection was done using a Neon Transfection System (ThermoFisher). Specifically, a combination of siRNA (Dharmcon, On-Target Smart Pool), siGLO, and buffer T (ThermoFisher, Neon transfection kit) was added to each sterile electroporation tube such that the final concentration of siRNA was 50nM. Additionally, a mock transfection and a non-targeting (NT) siRNA transfection were conducted as controls. Following transfection, cells were transferred to 1mL of warm medium containing RPMI 1640 Complete medium (+10% FBS + glutamax + sodium pryvate; RPMI1640 (Sigma Aldrich, Cat# R0883), L-glutamine (Sigma Aldrich, Cat#G7513), Sodium Pryvate (Sigma Aldrich, Cat#S8636)), and the stimulation cocktail, comprised of IL4 (R&D, Cat#204-IL, 0.2 ug/ml), IL21 (R&D, Cat#8879-IL, 50 ng/ml), IL10 (R&D, Cat#1064-IL, 0.2 ug/ml), sCD40L (R&D, Cat#6245-CL, 500 ng/ml), and ODN2006 (Miltenyi, Cat#130-100-105, 1 ug/ml). One unstimulated and one mock stimulation condition were done as controls. Cells were then plated on a 96 well U-bottom plate (Sarstedt, 83.3925, Lot#6022611), and transferred to the Incucyte. In total, there were around 5 million cells per condition.

### Cell sorting, RNA extraction, and supernatant analysis

FACS sorting was conducted on day 1 and 6 of the protocol. Cells were placed in a V-bottom plate (Corning Costar, Cat#3363) for FACS surface staining, with around 1 million cells per condition. Cells were washed with cold PBS and treated with viability near infrared (IR) dye for 30 minutes in the fridge, before being washed again and spun down twice at 500 G for 3 minutes. Antibodies for cell surface markers were added to each well, including IgD (BD, Cat#555778), CD38 (BD, Cat#561106), CD19 (Biolegend, Cat#302216), IgG (Miltenyi, Cat#130A093A194), ViabIR (Invitrogen, Cat#L10119), CD27 (Biolegend, Cat#302824), and CD45 (BD, Cat#563204); cells were then resuspended and stored in the fridge for 10 minutes. Lastly, cells were washed again, treated with staining buffer (PBS + 0.5% FCS), and spun down at 500 G for 3 minutes, before being resuspended in BD Fixation buffer (BD Biosciences, Cat#339860), transferred to FACS insert tubes, and stored in the fridge until analysis using a Gallios FACS machine.

RNA extraction was conducted on day 3 of the protocol post-transfection. Specifically, sample plates containing 1.5 million cells/condition were removed from the Incucyte and centrifuged at 500 G for 3 minutes. The pellet was washed with 200 uL/well of cold PBS, and the wells were pooled into a sample collection tube that was spun down at 500 G for 3 minutes. PBS was carefully removed, and the pellets were snap frozen at -80 degrees for storage. RNA isolation was done using the Qiagen RNeasy Mini Kit, following the kit protocol.

Supernatant was collected on day 3 and day 6 of the protocol post-transfection. Plates were removed from the Incucyte and centrifuged at 500 G for 3 minutes. 140 uL/well of supernatant was removed for analysis. The Human IgG ELISA development kit (Mabtech, Cat#3850-1AD-6) was used for IgG secretion measurement, following the manufacturer protocol. Additionally, one plate was reserved for cell viability analysis on day 6, containing 0.5 million cells/condition; 50 uL of CellTiterGlo reagent (Promega, Cat#G755A, Lot#0000297658) was added per well and mixed gently. The whole cell lysate was then transferred to an opaque plate, and luminescence was measured using a Multiple Detection System from Promega.

## 2.17 Visualisations

All plots were generated using *ggplot2* implemented in R, unless otherwise specified. In some instances, base R was used for less complicated visuals such as Venn diagrams.

*In the beginning there was nothing, which exploded.*

Terry Pratchett

# 3

# Computational prioritisation of genes

## 3.1 Introduction

Validating the results of a computational pipeline for the purposes of drug discovery is a complex task that can be approached from a number of different angles. Perhaps the clearest or most compelling form of validation is to show how a novel, computationally-priortised target (or, ideally, set of targets) can successfully be taken through clinical trials, ultimately resulting in an approved therapeutic for the treatment of a particular disease. However, such an approach would take several years and considerable time and financial investment to complete, during which the uncertainty as to whether the underlying computational pipeline is good, remains. Thus, there is a compelling need to demonstrate some level of validation early on, in order to be able to more confidently proceed with pipeline implementation in the drug discovery and design process itself. Before outlining how such a method of validation can be conducted, however, it is important to briefly outline the prioritisation approach itself, especially since it will be implemented in subsequent chapters in this thesis.

### Priority Index pipeline

The pipeline in question is Priority Index, or Pi, which was developed as part of the **U**nrestricted **L**everaging of **T**argets for **R**esearch **A**dvancement and **D**rug **D**iscovery (ULTRA-DD) consortium [64]. The rationale for this project was fairly straightforward, even if the underlying problem is not: how can we better leverage the growing body of genomic (e.g. gene expression, promotor-capture, etc.) and network (e.g. protein-protein interaction) data, in order to gain more insight into the information captured by genome-wide association studies (GWAS)? Specifically, Pi was developed as a tool to integrate a number of complementary datasets in order to link GWAS-identified SNPs to target genes, which could then be scored based on the available evidence to quantify the likely functional importance of each gene for each disease.

Therefore, the input for a Pi analysis is the GWAS summary statistics (SNPs and p-values) for a particular disease of interest. Next, the disease associated SNPs are attributed to the genes whose function they are most likely to modulate based on a set of genomic predictors including:

I. proximity to input GWAS SNPs, taking into consideration linkage disequilibrium (LD) and genomic structure, specifically location within the same topologically associated domain **(nGenes)**.

II. modulation of gene expression, based on expression quantitative trait loci (eQTL) in a range of immune cells **(eGenes)**, including neutrophils [89], T-cells [90], NK-cells, B-cells and monocytes [88].

III. evidence of physical interaction with gene promoters, based on chromosomal conformation capture in immune cells **(cGenes)** [87].

The genes identified by the above are referred to as **genomic seed genes**. These seed genes are then further scored based on intersection with a set of annotation predictors; namely, immune function (fGene), immune phenotype (pGene) and rare genetic diseases related to immunity (dGene). Next, seed genes are superimposed onto a gene interaction network (one predictor at a time) based on the STRING database [96], and the strength of connectivity between seed genes and all other genes

(non-seeds) is determined using a Random Walk with Restart algorithm [54] to capture how connected (related) gene nodes are on the basis of the interacting networks they are part of [52, 54], allowing for scoring of all interacting genes, regardless of whether they are direct seeds or not. Finally, each of the sets of gene scores are combined across each predictor type in a meta-analysis approach, and a final per-gene score is determined. Thus, Pi allows for genetics-led target prioritisation that takes into account GWAS, functional genomics, annotations, and network effects, and which can be relatively easily scaled across numerous immune-related diseases. Importantly, Pi does not take into account known drug target information, allowing for an unbiased, genetics-first approach.

### Defining drug targets and golden standards

It addition to prioritising genes on the basis of disease relevance and genetic underpinnings, it is critical to be able to compare putative targets to known targets of effective therapeutics, in order to validate novel algorithms, understand similarities or dissimilarities to existing targets, or predict which patient subgroups are most likely to respond, on the basis of what is already known to be effective or ineffective within a particular therapeutic class for a specific disease. Defining drug targets, broadly speaking, is itself an important step in improving therapeutic advances, and has been the sole topic of recent reviews aiming to aggregate and classify existing data [120]. A number of databases have also emerged in order to address this important questions; for instance, ChEMBL is one of the dominant resources in the field, providing detailed information on target and drug data, including elements like bioactivity, structure, binding sites, strength of interaction, supporting evidence, and others [46, 97]. Similarly, DrugBank is another similar open-sourced resource with aims to provide information on chemical and molecular features of various compounds, and a particular focus on drug - target interactions [47]. The advantage of such approaches is that they are structured so as to be organised and searchable at scale and updated in regular intervals, as well as being comprehensive in nature, making them invaluable tools in the drug discovery pipeline.

Indeed, a number of previous studies have leveraged such tools when building predictive models based on known drug target information. For example, work has been done to build a support vector machine classification algorithm for cancer drug prediction; the classifier sourced a number of genetic

information across various cancer cell lines, and then trained this algorithm using a set of positives sourced from DrugBank [57]. Similarly, a study mapping genetic associations, pathway information, and RNA expression sourced positive targets from the PharmaProjects database in order to uncover underlying patterns [60]. While many of the resources for defining true positives, or Gold Standard Positives, for drug targets have been made readily available for open-access use, understanding Gold Standard Negatives remains a hurdle in the field. In some instances, negatives can be defined as targets that lack drug or therapeutic annotations in the above-mentioned databases [57]; however, this is limited by how comprehensive the initial annotations are, as well as prone to false negatives (i.e. defining a target as a gold standard negative when it may actually be the case that the target has simply never been investigated in a clinical setting). This problem is compounded when considering that supervised predictive algorithms are trained based on the labels provided; in other words, specifying that a target is a gold standard negative means that the algorithm will be heavily biased to classify any similar target as a negative as well, with high confidence. Additionally, sourcing failed clinical trial information (containing enough granularity to understand true reason for failure) remains a challenge due to the competitive incentive for pharmaceutical companies to not make reasons for failures public knowledge; it is therefore important to continue promoting public-private partnerships so that such information can become more accessible to researchers looking to advance the therapeutic field of drug discovery [5]. In lieu of such 'true' gold standard negative information, alternative techniques can be implemented for these purposes, discussed in greater detail in later sections of this chapter.

### Aims

Thus, following gene prioritisation, one critical question remains: how reliable is the output from a Pi analysis in its ability to highly prioritise therapeutically-relevant targets? Without some sort of quantification of performance, it is very difficult to justify using the Pi pipeline to inform drug discovery. In this chapter, my goal is to explain how Pi's performance can be evaluated through the use of 'Golden Standard' targets. Specifically, the aims are as follows:

I. Demonstrate how Priority Index can be used to process multiple sclerosis (MS) GWAS data, and how these outputs can be assessed in a way that can be systematically applied to any trait in

question;

II. Validate the results of the above analyses, using a computational approach that defines and utilises Golden Standard Positives (GSPs) and Negatives (GSNs) in a way that can easily scale beyond a single disease;

III. Implement a complementary form of Pi analysis, 'supervised mode', which leverages known drug target information (Golden Standards) to prioritise targets on the basis of genomic similarity to existing therapeutics, and contrast this output to that of the one in step I above.

## 3.2 Results

The results outlined here describe the application of Pi to MS, including both discovery mode and supervised mode, with a particular focus on the computational assessment of these outputs. Multiple sclerosis was selected as a trait due to the relatively large availability of drug - target combinations (approved drugs in clinical use with evidence of efficacy) used in the treatment of the disease, allowing for credible performance assessment based on a high number of GSPs (and, correspondingly, more reliable simulated GSNs).

### 3.2.1 Priority Index discovery mode applied to MS

The first output generated and interrogated in this chapter was done using discovery mode, which is a purely genetics-led approach to gene prioritisation, leveraging GWAS summary statistics, genomic predictors, annotation predictors, and network connectivity [64]. Specifically, GWAS data for MS was sourced from the GWAS Catalog [65], and included a collection of MS associations, captured in the recently-published paper by Patsopoulos and colleagues [67]. The genomic predictors were comprised of proximal genes, nGenes, located close to the input SNPs, taking into account topologically associated domains; cGenes, or genes with evidence of physical interaction between an input SNP and the gene promoter [87]; and eGenes, or genes with association between expression levels and SNPs in question (eQTLs), in various immune cell types including monocytes and B-cells

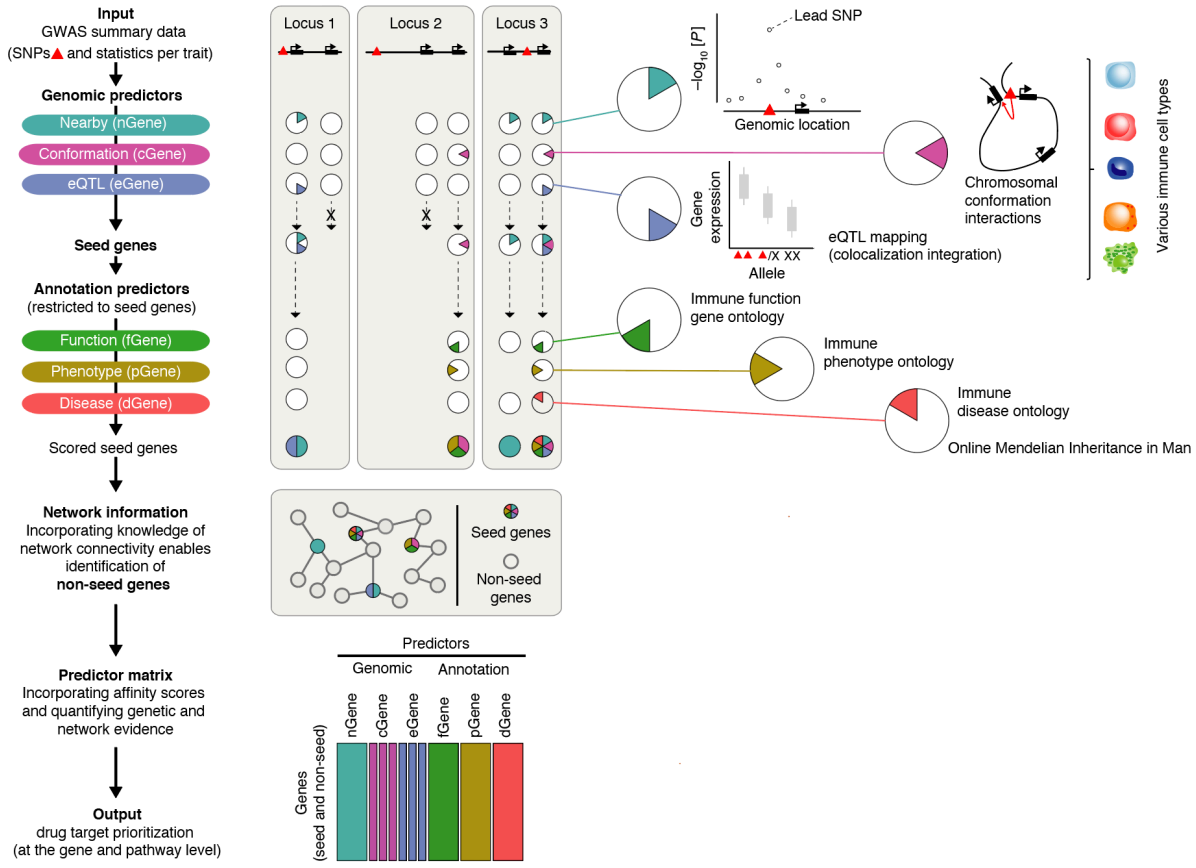



**Figure 3.2.1: Overview of the Pi pipeline**. A schematic outlining the Pi pipeline workflow, beginning with GWAS summary statistics as input and ending with a prioritised gene list as the output, reproduced from the 2019 Fang *et al.* study [64]. The per-trait input SNPs are matched to functionally relevant genes through a number of steps, beginning with genomic predictor mapping (on the basis of proximity (nGene), physical interaction with gene promoters (cGene), and association with gene expression (eGene)) to define seed genes. Any such defined genes are then further scored on the premise of annotation predictors (immune function ontology, immune phenotype ontology, and immune disease ontology), and scored genes are mapped onto a gene interaction network. Scoring is then propagated across all genes in the interaction network via Random Walk with Restart, for each unique predictor class. The final list of prioritised genes is calculated by incorporating all predictor scores per gene, and ranking genes based on this final score.

[88], neutrophils [89], T-cells [90], and unpublished NK cell data from 228 healthy individuals of European ancestry. Annotation predictors (which were only applied to genes that had also been scored by the genomic predictors) consisted of immune function genes (fGene), or genes annotated to an immune response based on experimental data, as defined using Gene Ontology [91]; disease genes (dGenes), defined as genes related to rare genetic disease as defined by Online Mendelian Inheritance in Man [92] as well as annotated to immune system disease more broadly, as defined by Disease Ontology [93]; and immune phenotype genes (pGenes), capturing genes annotated to immune system abnormalities via Human Phenotype Ontology [94], and to immune/hematopoietic system phenotypes using Mammalian Phenotype Ontology [95]. Additionally, out-of-box parameters were used for Pi analysis, including an $r^2 = 0.8$ for LD calculation, all available eQTL and promoter-interaction datasets, along with all annotation predictors. Implementation of Pi (version 1.10.0) and subsequent gene and pathway-level analyses were conducted in R (version 3.5.1). An overview of the pipeline can be found in Fig. 3.2.1.

### Gene-level prioritisation

Fig. 3.2.2 shows the top 30 genes from the discovery mode, which are useful in giving initial insight into the comparison between Pi's output and existing/published MS data and testing the hypothesis that Pi outputs reliably capture known information about disease, particularly at the top end of the prioritisation output.

For example, *CD40* was one of the highest priority genes in the discovery mode output, and was prioritised on the basis of being simultaneously classified as an nGene, eGene, and cGene; its role in T-cell activation (together with CD40L) and MS progression has been well documented previously and extensively reviewed since [121–125]. Moreover, while previous Phase 2 studies of CD40 as a therapeutic target were halted due to severe side effects, more recent work examining the CD40 signalling pathway and small molecule inhibition has been more promising, and thus may revitalize the underlying therapeutic potential [125, 126]. Further, *IFNAR2* was highly prioritised in the output, due to multiple cGene mappings between MS SNPs and the gene promoter, in immune cell types [87]. Previous analysis of MS patients revealed a significant decrease in *IFNAR2* mRNA

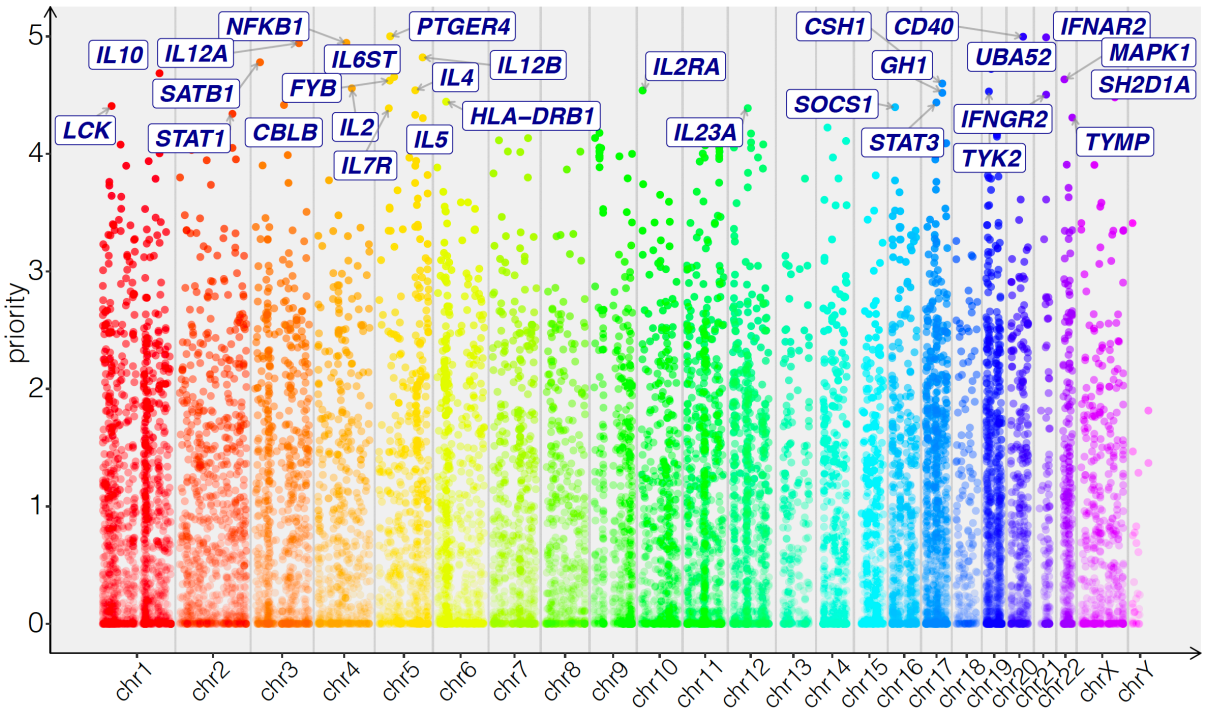



**Figure 3.2.2: Top prioritised genes in Pi discovery mode analysis of MS.** A Manhattan plot of the top 30 genes from the Pi discovery mode output, using MS GWAS summary data as input.

expression in patients compared to controls and, further, there was a significant difference in expression amongst interferon beta responders and non-responders [127]. Additionally, *IL2* and *IL2RA* were both highly prioritised, which is in line with known associations between polymorphisms in these genes and MS risk [128]; while *IL2RA* was identified as an nGene, *IL2* was prioritised on the basis of its high network connectivity to seeds. Similarly, a meta-analysis of MS GWAS, which took into account 2624 cases and 7220 controls, identified *IL12A* and *PTGER4* as loci with suggestive evidence of link to MS [129]; a more recent study also found differences in gene expression of *PTGER4* between MS patient blood as compared to healthy controls [130]; this gene was highly ranked in the Pi analysis on the premise of multiple SNP-gene promoter interactions [87].

### Pathway analysis

Moving beyond gene-level outputs and into a pathway-level view confirmed that Pi discovery mode captured highly disease-relevant targets in the top 1% of the prioritisation list. Top enriched pathways (Fig. 3.2.3) included TCR signaling in CD8+ T-cells (FDR = $2.7 \times 10^{-33}$), which has been implicated




**Figure 3.2.3: Top prioritised genes in Pi discovery mode analysis of MS.** A Manhattan plot of the top 30 genes from the Pi discovery mode output, using MS GWAS summary data as input.

in MS via postmortem analysis of MS patient brain lesions [131], as well as TCR signaling in CD4+ T-cells (FDR = 9.9 x $10^{-33}$), which has been implicated as a key factor in rapid patient transition to progressive MS [132]; IL12-mediated signaling (FDR = 6.1 x $10^{-23}$), previously implicated in MS in a number of studies as well as linked via GWAS evidence for the *IL12A* gene and MS susceptibility [129, 133–135]; and IL2-mediated signaling (FDR = 3.5 x $10^{-14}$), which is in line with literature implicating mutations in corresponding genes IL2 and IL2RA and MS risk [128]. Taken together, these results confirm that Pi analysis of MS genomic data outputs a list of prioritised genes that are broadly aligned with current knowledge and understanding of MS disease pathogenesis, specifically when it comes to the highest priority genes and pathways involved.

### 3.2.2 Golden standards for performance assessment

In order to quantify how well Pi was performing, a set of genes to benchmark against was needed. Firstly, a set of 'golden standard positives' (GSPs) would serve as a type of positive control; in other words, these genes would, ideally, be captured in the outputs of Pi, without explicitly being

programmed into the calculation of gene scoring. Since the ultimate aim of Pi is to predict which targets would be most favourable to take forward from pre-clinic to clinical trials, a set of 'successful' drugs (by disease) could serve as a credible GSP set. More specifically, a GSP list could contain the targets of known, effective therapeutics, which could then serve as a benchmark for Pi's ability to prioritise these targets in the disease-specific outputs.

In order to generate such a list, ChEMBL [97] was used as the source material, due to the comprehensive nature of the database, along with the multiple layers of supporting evidence for each drug. Each disease of interest was systematically queried in the ChEMBL database, and all targets of all drugs (supported by >2 sources linking target to drug) were collected and categorized by clinical phase cutoff (completed Phase 2, completed Phase 3, completed Phase 4). The information was organised locally as a collection of SQL-searchable databases, such that it was possible to search for various combinations of drugs, diseases, targets, etc., map them to corresponding human genes, and sort by stage of development as needed. Fig. 3.2.4 shows an outline of how the various databases were interlinked in order to preform these searches, and Fig. 3.2.5 shows an example of the kind of SQL query that was run in order to conduct a specific search to bring all the information together.

Collectively, the targets gathered in this manner represented the GSPs against which various evaluation tests could be run, with the added flexibility of being able to define 'positives' in a few different ways (i.e. by phase cutoffs). Since some drugs were outliers in that they were highly promiscuous (i.e. had many targets, making it difficult to determine which targets had a significant clinical effect), a filtering step was applied to remove such outliers. All drugs with > 5 targets were thus removed, since these could be classified as extreme outliers on the basis of being more than 3 x IQR (interquartile range) above quartile 3, based on the distribution of number of targets per drug across all drugs.

In addition to compiling GSPs, a set of negatives (GSNs) was needed in order to run certain performance tests. Negatives were more complicated to define, for a number of reasons. Firstly, substantially more evidence is needed to determine that something is a true negative compared to evidence needed to determine that something is an effective (positive) therapeutic. Second, negative

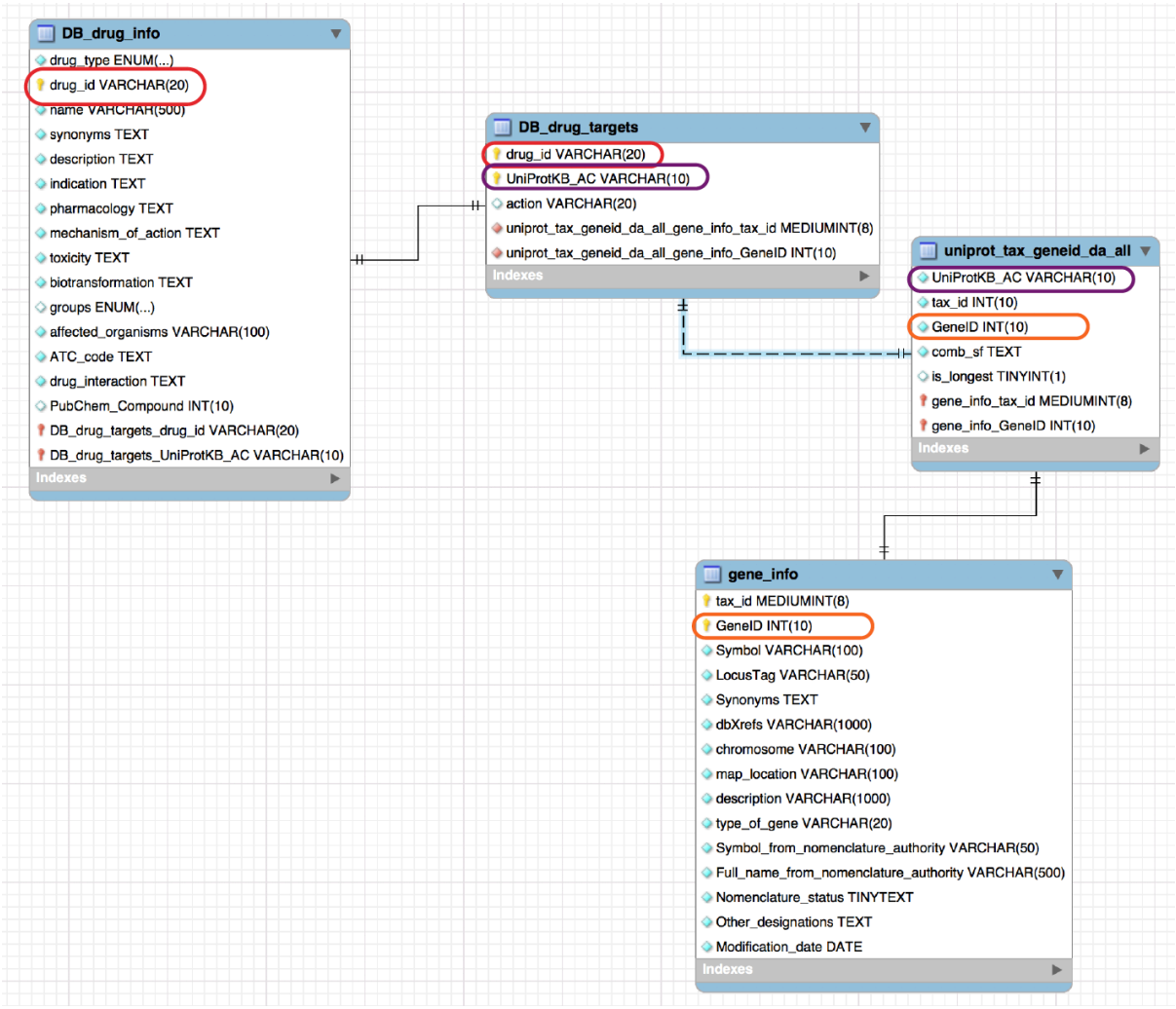



**Figure 3.2.4: Structure of databases relating to drug targets.** A schematic showing how various downloaded databases were linked, allowing for queries to tie together drug, protein, gene, and phase information. Linker terms used for queries are contained in the coloured boxes, and box colour reflects where matching terms were intersecting.

```
select di.drug_id, dt.UniProtKB_AC as 'drug targets', dt.action, di.name, gene.GeneID, gi.Symbol, di.synonyms, di.description, di.indication, di.groups

from DB_drug_info as di
join DB_drug_targets as dt
on di.drug_id = dt.drug_id

join uniprot_tax_geneid_da_all as gene
on gene.UniProtKB_AC = dt.UniProtKB_AC

join gene_info as gi
on gi.GeneID = gene.GeneID

where di.affected_organisms like '%Human%'
and di.indication like '%%' #INSERT DISEASE IN THIS ROW
and di.groups in ('approved','investigational','experimental')
and gene.tax_id = 9606;
```

**Figure 3.2.5: SQL query structure.** An example of the SQL code used to bring together information on drugs, drug targets at the protein level, drug targets at the gene level, phase cutoff, gene description, and other information. This approach was used in order to efficiently compile the necessary GSP information for each disease of interest.

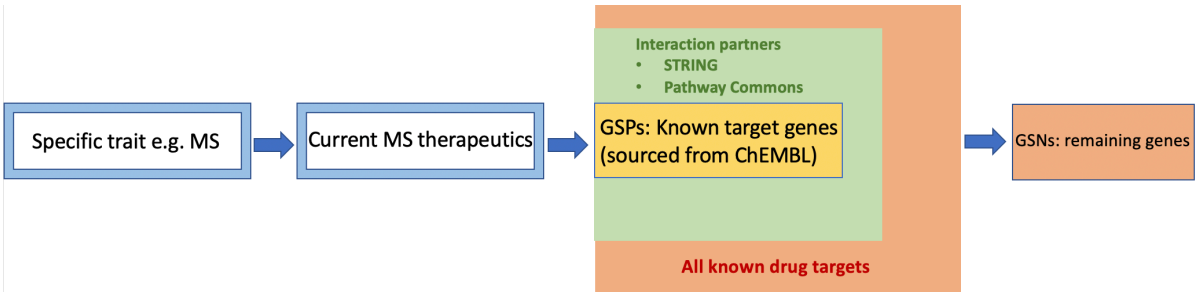



**Figure 3.2.6: Workflow for simulating GSNs.** A schematic outlining how Golden Standard Negatives (GSNs) were simulated, on a per-trait basis.

results are, on the whole, poorly recorded, especially in the clinical trial space; there are numerous targets that remain in 'clinical limbo'. Third, the reason for failure of a particular target is often opaque, and not always due to a target's failure to treat disease; targets may be labelled as failures due to severe side effects, for example, or even due to the particular pharmaceutical company's de-prioritisation of a compound in favour of another drug/target. Therefore, defining 'true negatives' remains a challenge in the field and presented an interesting obstacle. The approach taken here was slightly different: instead of looking for failed targets (per disease) and attempting to discern whether failure was 'true failure', GSNs were instead simulated. This involved a process with a few steps (Fig. 3.2.6): first, a trait was selected (e.g. multiple sclerosis). Second, the known 'effective' (GSP) drug space was identified for the trait (defined as Phase 2 or above). Third, all direct gene network interaction partners of these GSPs were identified, based on merged STRING [96] and Pathway Commons [101] databases. Fourth, a credible 'drug landscape' was defined as all targets of all drugs, regardless of disease indication. Lastly, the list of GSPs and all interacting partners was removed from the drug landscape list, and the remaining genes were labelled as GSNs. This method was useful since it allowed for simulation of negatives at scale, it accounted for network effects of known GSPs (by not including network neighbours), and it minimized type I error by removing therapeutics in the early clinical trial pipeline (i.e. treated them as potential GSPs).

### 3.2.3 Performance evaluation of discovery mode

Two approaches were used for performance assessment: precision-recall (PR) curves and receiver operating characteristic (ROC) curves, with an emphasis on the latter. PR curves were generated by

labelling each GSP gene in the Pi output as a positive, and then plotting precision (True Positives/(True Positives + False Positives)) on the y-axis against recall (True Positives/(True Positives + False Negatives)) on the x-axis, moving step-wise down the priority list. In contrast, ROC curves were generated by plotting sensitivity (True Positive Rate, or (True Positives/(True Positives + False Negatives)) on the y-axis against specificity (False Positive Rate, or (False Positives/(False Positives + True Negatives)) on the x-axis, moving step-wise down the priority list. Additionally, Area Under the Curve (AUC) was computed in order to quantify the 'performance' of each ROC curve, with values ranging between 0 - 1 (1 being a perfect score).

Initially, precision-recall was selected due to the high class imbalance between GSPs and non-GSPs (i.e. of the 15,000 genes per disease-output, only a relatively small number are GSPs). Thus, the aim of constructing a PR curve was to ascertain whether the Pi output would prioritise GSPs most highly, over other genes. Fig. 3.2.7 shows the PR curve for MS across three definitions of GSPs: (i) completed Phase 2 and above, (ii) completed Phase 3 and above, and (iii) completed Phase 4 and above.

As evident in Fig. 3.2.7, the Pi output for MS did not perform very well at prioritising a substantial proportion of GSPs at the very top of the output (peak precision was 0.2, with a recall < 0.1 at this threshold). However, the primary purpose of Pi is to nominate novel, putative targets, and discovery mode does not utilize known knowledge of GSPs in its algorithm. Since putative targets technically count as non-GSPs, and significantly outnumber GSPs, it was not surprising that the PR curves looked the way they did; novel, viable targets are treated as false hits in a PR analysis which, naturally, does not reflect their potentially biological disease significance.

Therefore, an alternate approach of performance evaluation was considered: ROC AUC values. This approach asks a different question compared to PR; instead of looking at how highly prioritised GSPs were compared to all other genes, ROC AUC would determine how well Pi's output was separating GSPs from GSNs. In other words, does the genetics-led prioritisation approach successfully prioritise GSPs over GSNs, without being fed knowledge of what these GSPs and GSNs are beforehand.

As apparent in Fig. 3.2.8, AUC values for the discovery mode MS output were quite high; all three GSP cutoff thresholds had AUC > 0.80, with AUC increasing as the GSP threshold approached 'truer

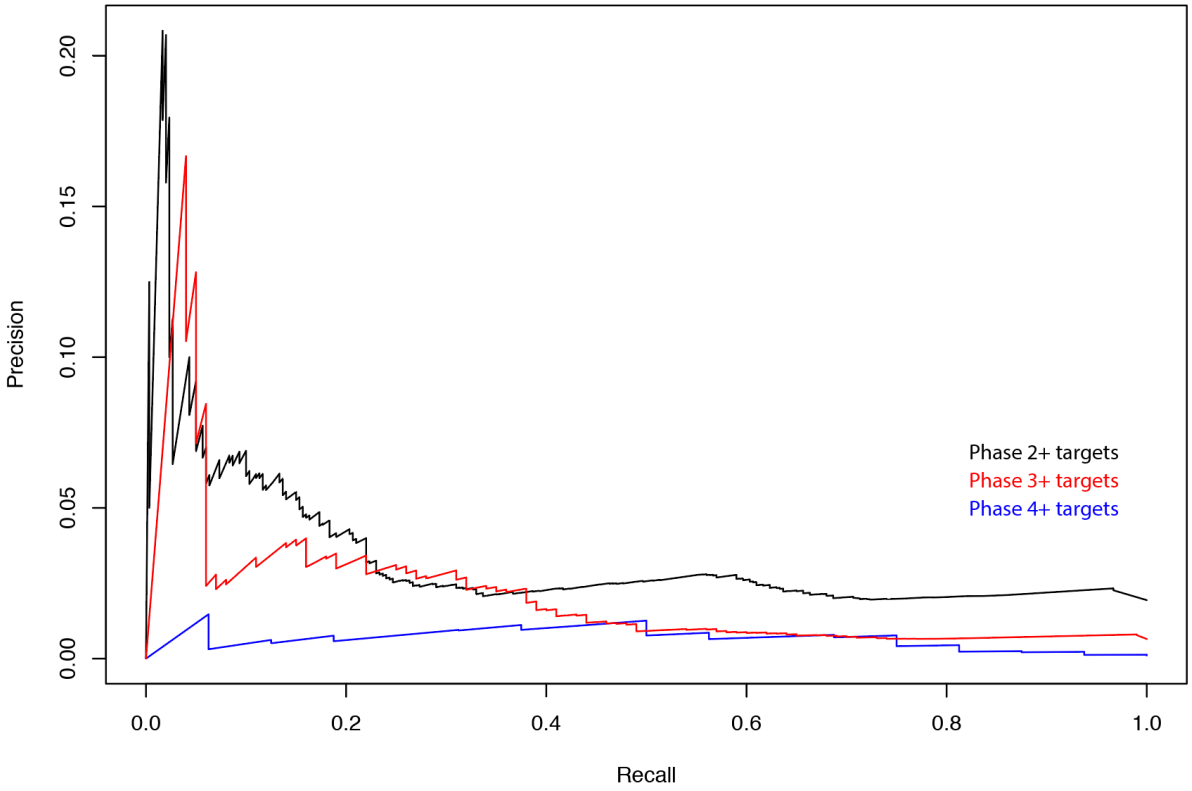



**Figure 3.2.7: A PR-curve of Pi's discovery mode.** A plot of precision versus recall of the Pi discovery mode MS output, across a range of different GSP definitions (depicted by the different coloured lines).

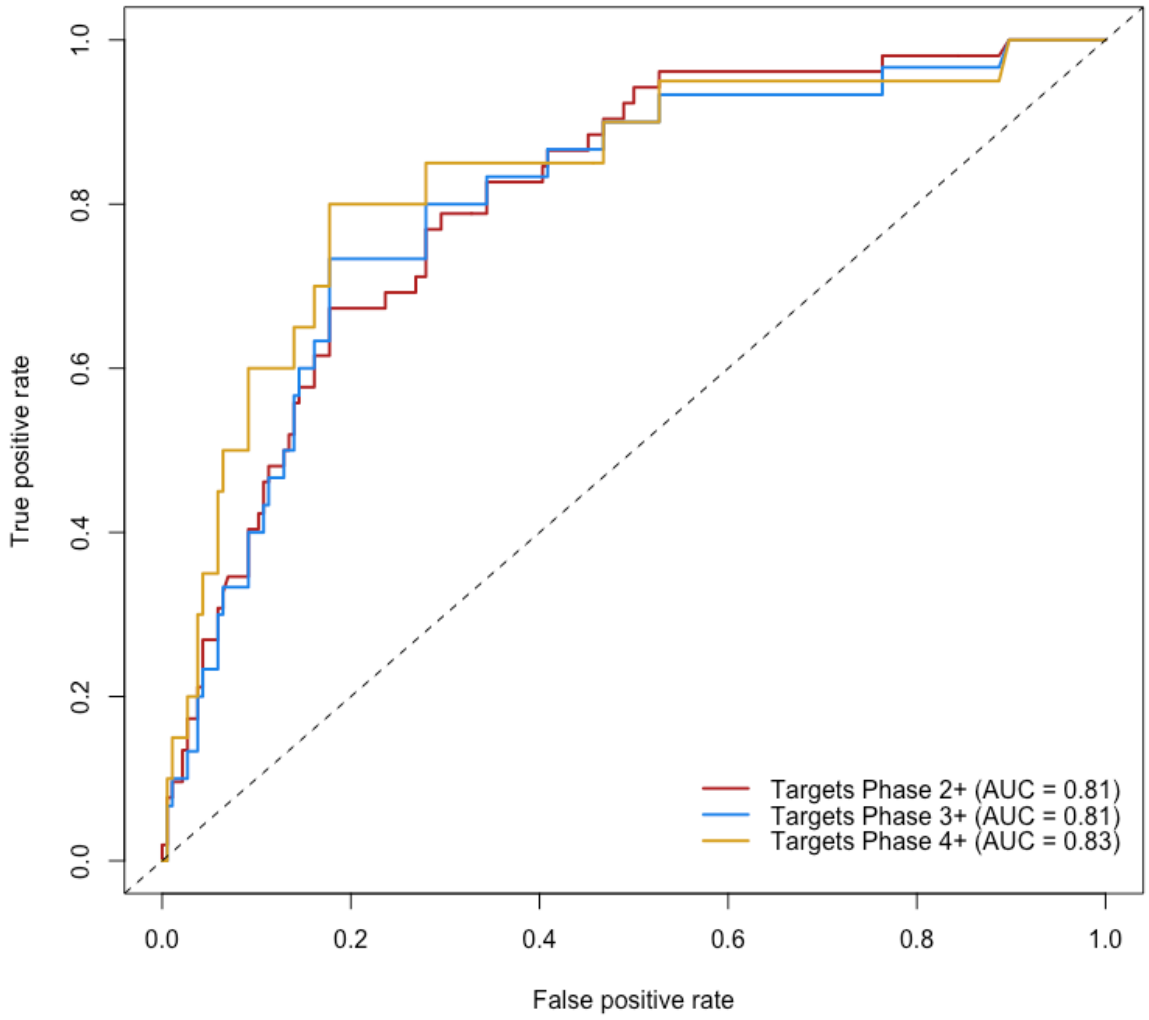



**Figure 3.2.8: ROC curves of Pi's discovery mode.** A plot of sensitivity versus specificity of the Pi discovery mode MS output, across a range of different GSP definitions (depicted by the different coloured lines). Area Under the Curve (AUC) for each GSP threshold is noted in brackets.

positives' (i.e. transitioning from targets in the mid-stage of clinical trials to those in the latter stages). As noted earlier, a trade-off between ROC vs PR curves is that PR curves give context to how highly prioritised GSPs are compared to all other genes, while ROC curves only consider the prioritisation of GSPs over GSNs. Therefore, it was important to confirm that, in addition to the high AUC shown in Fig. 3.2.8, GSPs were ranking relatively highly compared to GSNs.

Plotting GSP vs GSN density of Phase 3+ targets (i.e. targets that have substantial evidence of clinical efficacy; Fig. 3.2.9) confirmed this: Phase 3+ GSPs tended to fall in the top third of the Pi output, while simulated GSNs were more uniformly distributed throughout the output. In fact, a density plot broken down by individual Phases of targets (Phase 2, Phase 3, Phase 4) complemented the AUC finding in that 'truer GSPs' (i.e. those already completed Phase 4) were more concentrated at the top of the Pi output, compared to Phase 2 and Phase 3 targets (Fig. 3.2.10). Additionally, the top 1% of prioritised genes was significantly enriched for GSPs (9.4-fold enrichment, $p = 3.99 \times 10^{-3}$), as was the top 5% of genes (7.5-fold enrichment, $p = 1.99 \times 10^{-8}$), and these cutoffs were used as general guidelines to define genes of interest throughout the thesis.

The next phase of analyses aimed to test a series of controls, in order to further interrogate the reliability of the Pi output and confirm that the high AUC values summarized in Fig. 3.2.8 were reflective of true performance. Each control is described below, and the corresponding ROC curves are shown in Fig. 3.2.11.

### True baseline control

Normally, an AUC value of 0.5 is used as the baseline against which AUC values are compared; AUC = 0.5 is the expected value if positives and negatives were randomly distributed. However, this may not be the best way to define baseline in the context of drug discovery, since targets are not usually chosen at complete random. Instead, one method of investigating a potential target is to ask whether the target of interest has been investigated before in another clinical setting/for another indication. Therefore, a more conservative baseline was implemented as a control for the Pi MS output: instead of 0.5, all genes were ranked based on how frequently they were targeted by any drug, and this ROC

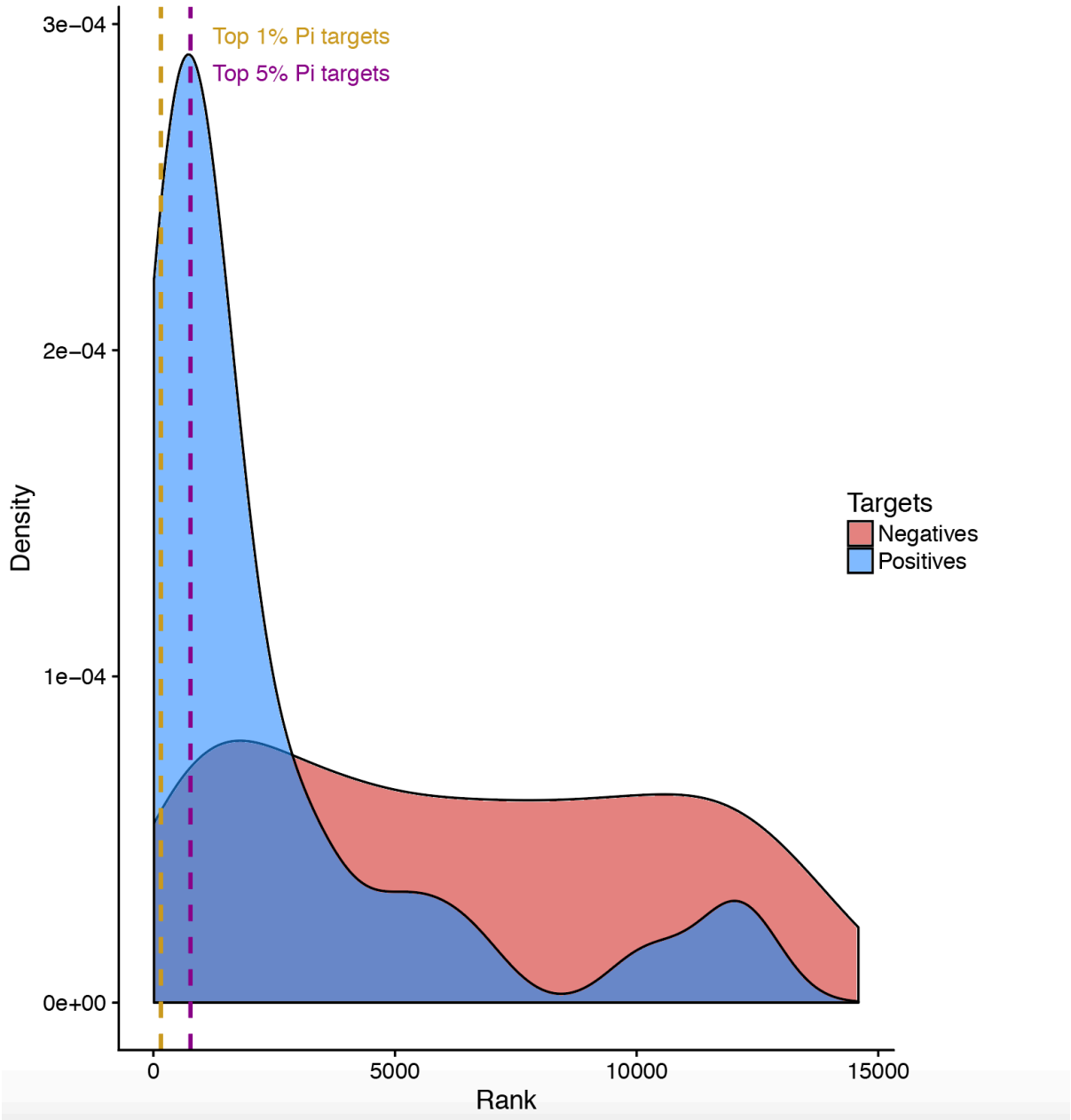



**Figure 3.2.9: Phase 3+ MS target density.** A density plot of all Phase 3+ MS targets in the discovery mode Pi output (GSPs) and corresponding GSNs. Dotted lines represent 1% and 5% rank cutoffs.

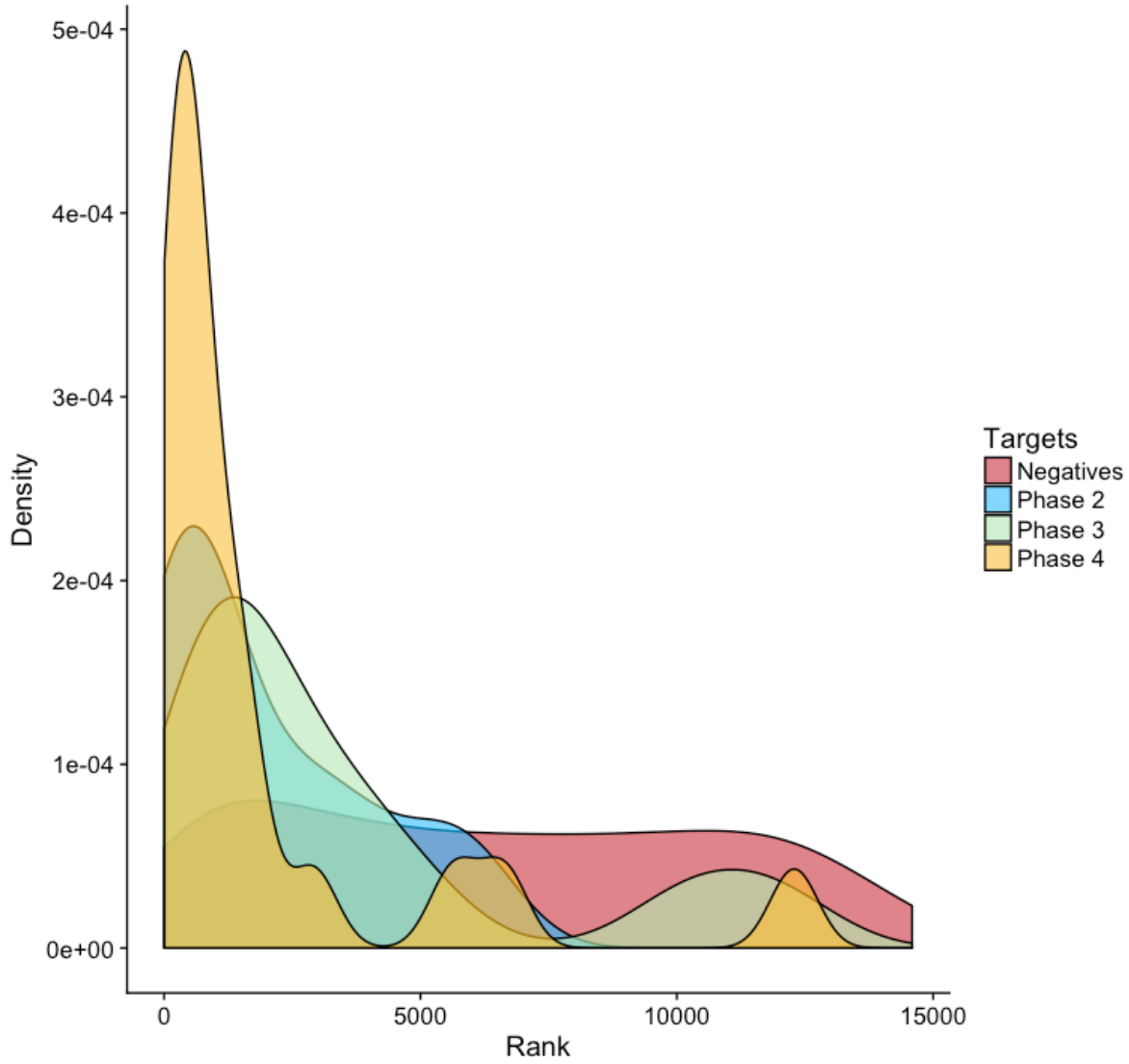



**Figure 3.2.10: MS target density by clinical Phase.** A density plot of all MS targets (GSPs) in the discovery mode Pi output, broken down by Phase of clinical trial, and corresponding GSNs.

curve was plotted as a baseline against which each test ROC was compared. Effectively, this answered the question: "did the Pi output do a better job of prioritising potential drug target genes than simply ranking genes by how often they are targeted by existing drugs?". This new baseline is represented by the golden line in Fig. 3.2.11, and has an AUC of 0.64.

### Positive control: alternate GSN simulation

The method of GSN simulation implemented in this chapter had a number of advantages/strengths, such as being easily scaled across numerous diseases. However, there are other possible ways to define negatives for a given disease. Therefore, an alternate method of defining GSNs was implemented as a type of positive control, to simultaneously check the robustness of Pi performance. This method involved defining GSNs as the targets of approved drugs in an unrelated indication from the one in question, resulting in a smaller number (but higher quality) set of negatives. For the MS context explored here, approved targets for treatment of heart failure were used as GSNs; this control ROC curve had an AUC of 0.83 (represented by the lighter red line in Fig. 3.2.11), which is comparable to the AUC of the simulated GSNs (0.81). Thus, the Pi MS output performed equally well in both cases of GSN definition.

### Negative controls: specificity of MS output and MS drug targets

In addition to the positive control discussed above, two negative controls were implemented. The first control involved testing the AUC of the MS output using ankylosing spondylitis (AS) targets as GSPs. Effectively, this control served to check whether the MS output performance was specific to recovering MS targets, and not simply recovering more general autoimmune/inflammatory targets. AS was chosen since it is an immune-related disease that differs from MS but that simultaneously has well-established genetic/GWAS links. This control ROC curve (represented by the dark blue line in Fig. 3.2.11) had an AUC of 0.61, which fell below the baseline value of 0.64 defined above, supporting the fact that the MS output was indeed MS-specific. The second negative control served to check whether MS targets were generally MS-specific, as opposed to generalizing well across

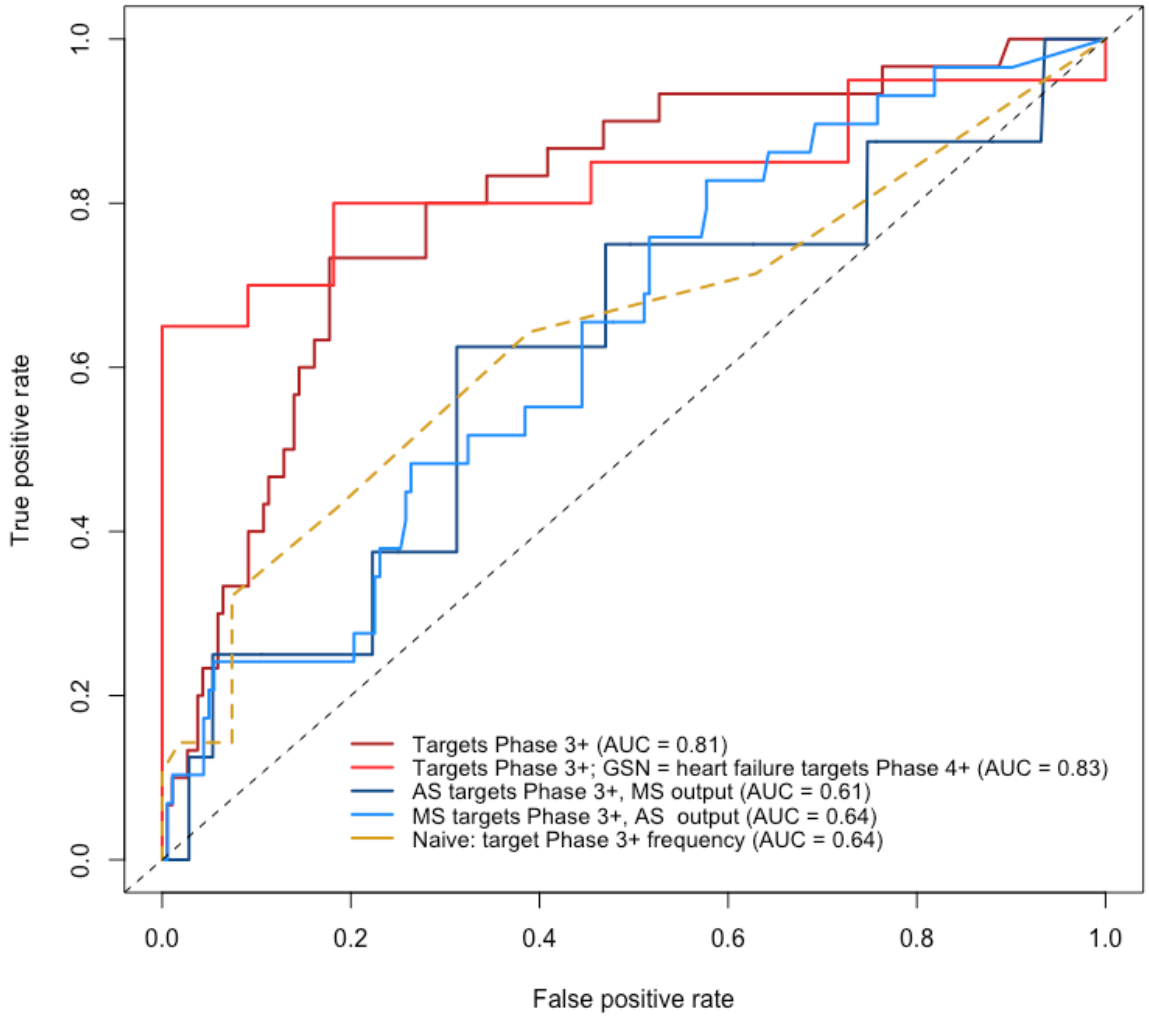



**Figure 3.2.11: Control ROC curves for MS in discovery mode.** A series of controls for the Pi MS discovery mode output, plotted against the test ROC curve, where GSPs are defined as all targets having completed Phase 3 clinical trials and above. The gold line represents a more conservative and biologically relevant definition of baseline; the light red line tests an alternate form of GSN definition; and the two blue lines test for the specificity of the MS output and of MS drug targets.

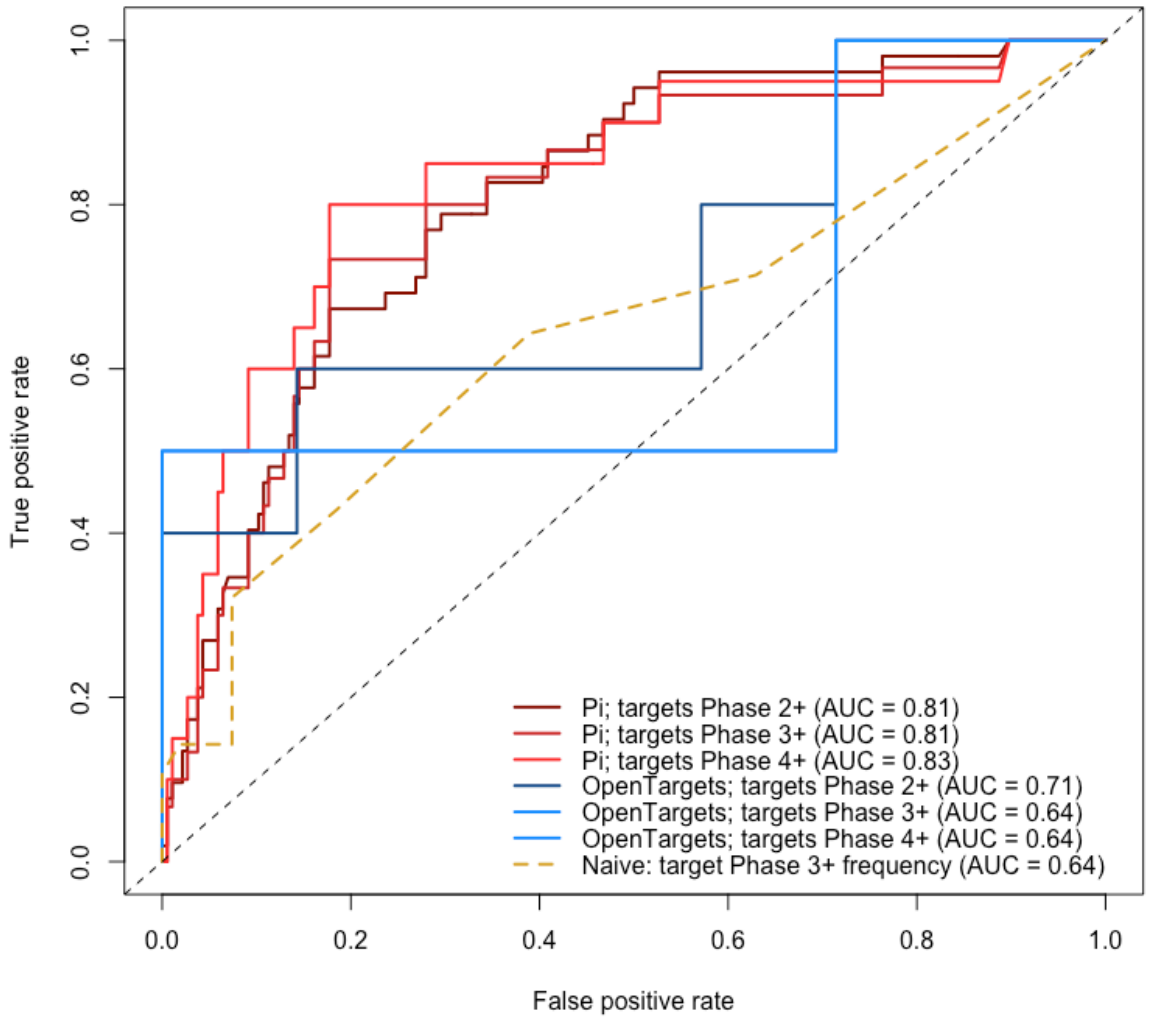



**Figure 3.2.12: Pi discovery mode versus OpenTargets genetic predictor.** A comparison between the Pi MS discovery mode output and OpenTargets' MS output, restricted to the genetic predictor. ROC curves are shown across a range of GSP definitions, and contrasted to the target frequency baseline (in gold).

autoimmune/inflammatory conditions. Thus, a ROC curve was constructed by testing MS target recovery in a Pi discovery mode AS output (represented by the lighter blue line in Fig. 3.2.11). This curve had an AUC of 0.64, which matched that of the baseline ROC curve, supporting the specificity of MS targets to the MS output.

### 3.2.4 Benchmarking performance against existing tools

In addition to benchmarking AUC values against the baseline defined above (i.e. against how often genes were targeted by existing drugs in any indication), the Pi output was benchmarked against the

closest existing tool: OpenTargets, an initiative aimed at leveraging genomic information for the purposes of drug discovery [58]. OpenTargets utilizes a number of different elements in its prediction pipeline, including human genetics, mouse studies, and known drug information. Therefore, in order to compare like to like, Pi MS discovery mode output was benchmarked against the OpenTargets MS output, restricted to the 'genetics' predictor column, which scores genes based on a collection of GWAS catalog [65], PheWAS catalog [136], UniProt [98], and European Variation Archive (https://www.ebi.ac.uk/eva/) data. Results are shown in Fig. 3.2.12. Across every GSP definition, Pi discovery mode outperformed OpenTargets genetic predictor, which narrowly outperformed the baseline defined by drug target frequency.

### 3.2.5 Priority Index performance with restricted set of clinical golden standards

Thus far, GSPs were defined as targets of therapeutics having completed Phase 2, 3, or 4, regardless of other factors such as how often the drugs were prescribed in clinic. Therefore, an interesting hypothesis to test was whether Pi performed better or worse when additional requirements were set out for GSPs to be considered truly 'golden standard'. Thus, the existing GSP list was sent to collaborating consultant neurologists, who were asked to identify which drugs were considered 'true golden standards' in the treatment of MS. The consensus list (10 drugs comprising 19 targets in total) was then used to define a restricted set of GSPs. In order to minimize false positives, the GSN list remained unchanged (i.e. was simulated using all MS therapeutics, not just the restricted GSP list). ROC curves were then generated to compare the performance of Pi using this new list, and results are shown in Fig. 3.2.13. Pi's performance at prioritising positives over negatives increased substantially when positives were defined as clinical golden standards (AUC of 0.89 with clinical golden standards, compared to 0.81 using all Phase 3+ targets), supporting the hypothesis that Pi prioritisation performs well at recovering known targets of effective therapeutics.

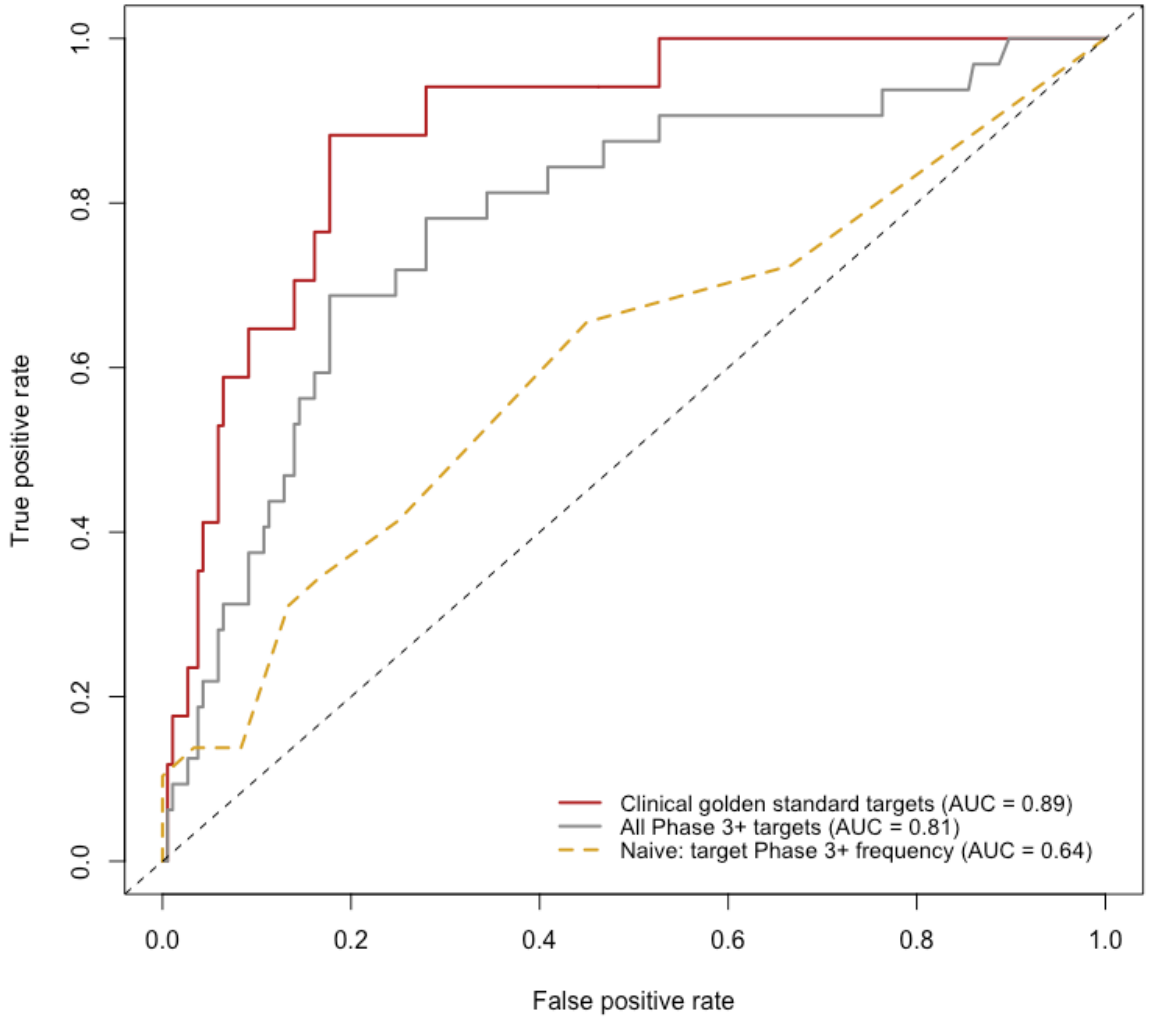



**Figure 3.2.13: Pi discovery mode using clinical GSPs versus all GSPs.** A comparison between the Pi MS discovery mode output performance when GSPs are defined as any target of a drug having completed Phase 3 and above (grey line) versus GSPs defined as clinical golden standards (n = 19,sourced from collaborating neurologists; red line).

### 3.2.6 Confirmation of performance beyond MS

In addition to performance evaluation in an MS context, a number of other diseases were also investigated using Pi outputs, including ulcerative colitis (UC), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and Crohn's disease (CRO). Similarly to MS, these diseases have a large number of GWAS-identified variants that served as inputs into the Pi pipeline, as well as a number of drug-target combinations that are useful for performance evaluation.

#### Benchmarked ROC curves across Phase cutoffs

As with MS, the main measure for performance was ROC curve AUC, to interrogate how well the Pi output separated GSPs from GSNs in these novel disease contexts. Thus, after running Pi for each disease individually, defining GSPs based on curated drug-target combinations, and correspondingly simulating GSNs, ROC curves were plotted for each disease individually. Both Phase 2+ and Phase 3+ GSP definitions were included, along with the corresponding OpenTargets genetics-only disease outputs. Results are summarized in Fig. 3.2.14. Pi's performance in each disease was comparable to the earlier MS plots, with AUC ranging from 0.83 - 0.87 for Phase 3+ GSPs (compared to AUC = 0.81 in MS). Additionally, Pi performed as well or better than OpenTargets in all definitions of GSPs, with RA being the closest AUC between the two tools (Phase 3+ Pi AUC = 0.84, Phase 3+ OpenTargets AUC = 0.83).

#### GSP target density distribution

In addition to plotting ROC curves for GSP versus GSN separation, it was important to once again confirm that GSPs were, in general, being prioritised near the top end of the Pi output, and not simply concentrated with GSNs near the middle or bottom of the list. Fig. 3.2.15 shows density plots for RA, SLE, UC, and CRO, including both distribution of GSPs (Phase 3+) as well as GSNs. Similarly to the MS targets, distribution plots for RA, SLE, UC, and CRO revealed that the greatest density of GSPs was near the top of the prioritisation output, whereas GSNs were more uniformly spread out across

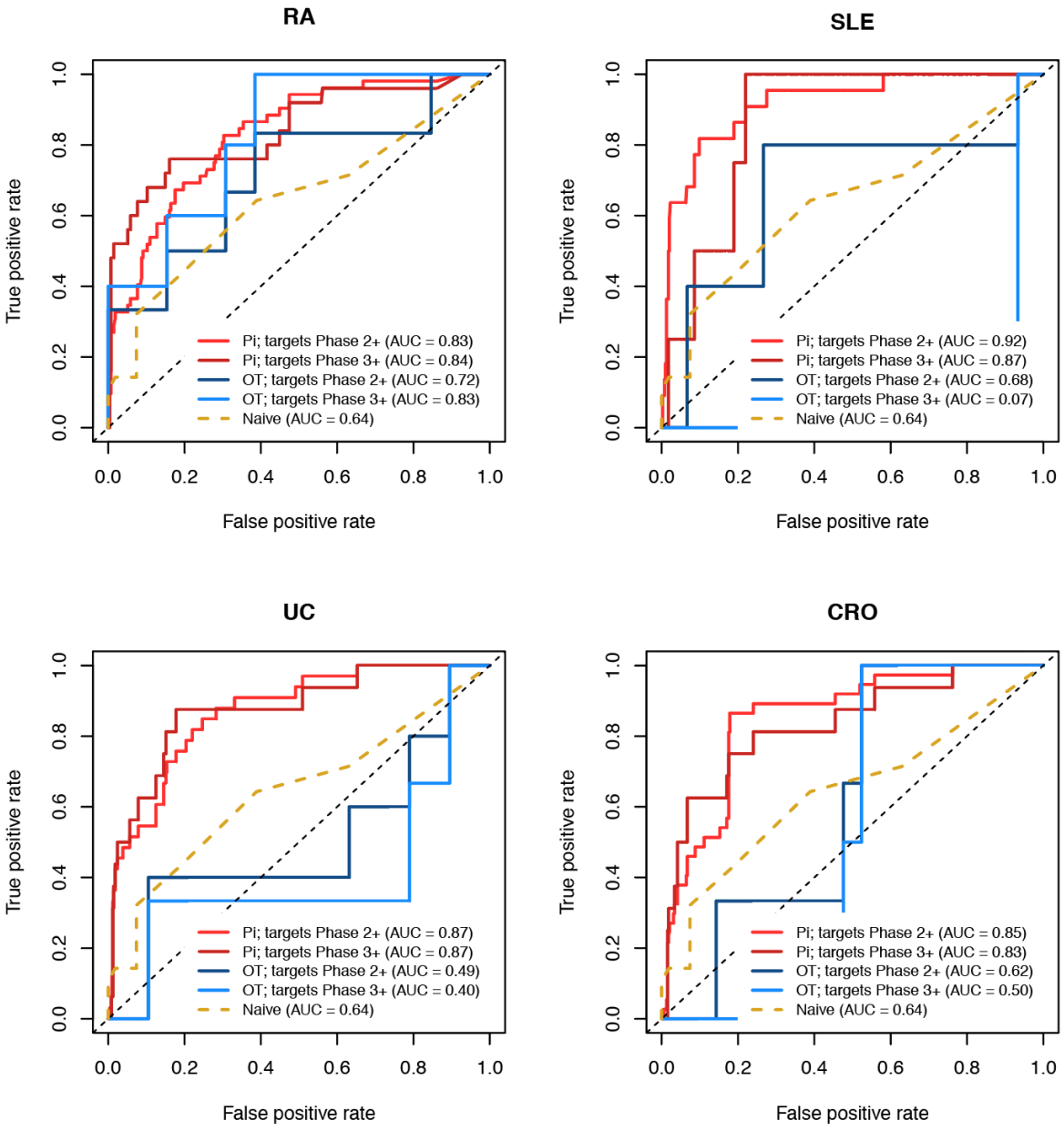



**Figure 3.2.14: Pi discovery mode performance across multiple diseases.** Performance of Pi discovery mode across four diseases (RA, SLE, UC, and CRO), as measured by the abllity of the output to prioritise GSPs over GSNs. ROC curves are shown across two GSP definitions (red lines), contrasted to the target frequency baseline (in gold), and benchmarked against the corresponding genetics-only OpenTargets outputs (blue lines).

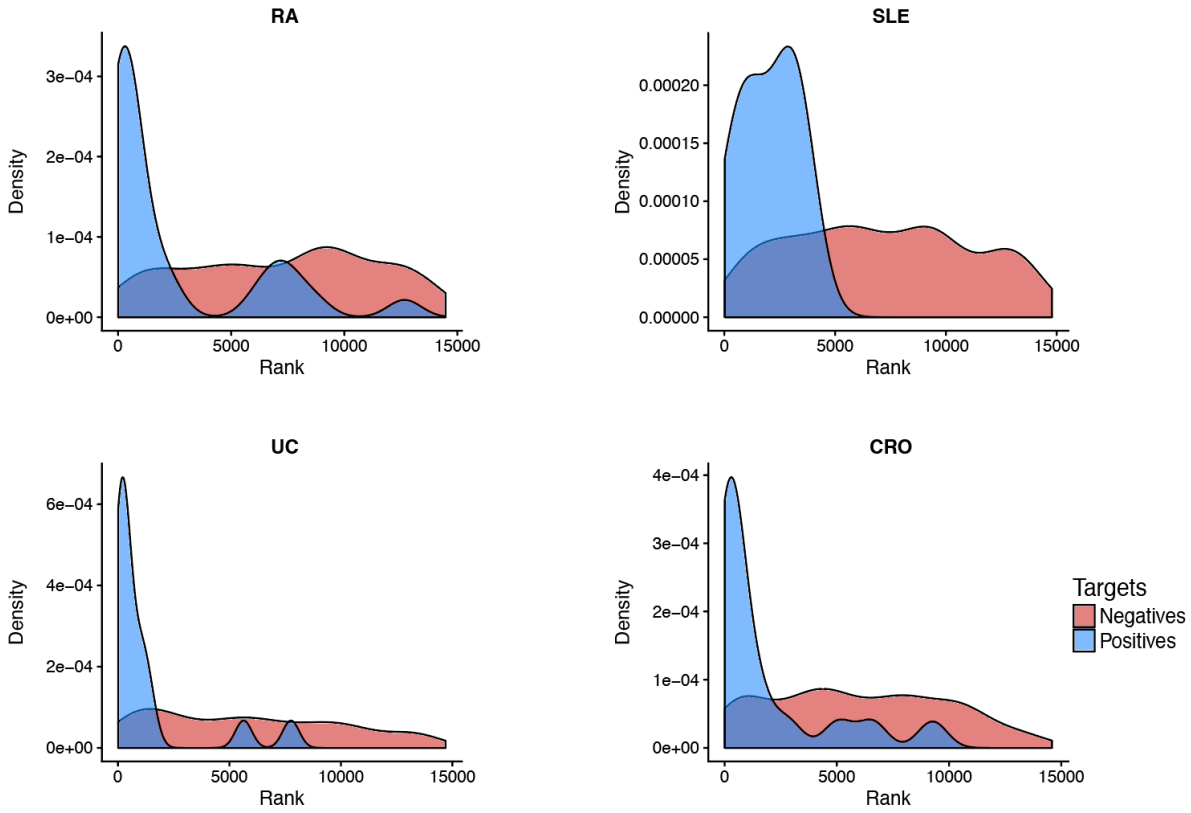



**Figure 3.2.15: Distribution of Phase 3+ GSPs and GSNs across multiple diseases.** Density plots for RA, SLE, UC, and CRO, with blue representing Phase 3+ GSPs and red representing simulated GSNs.

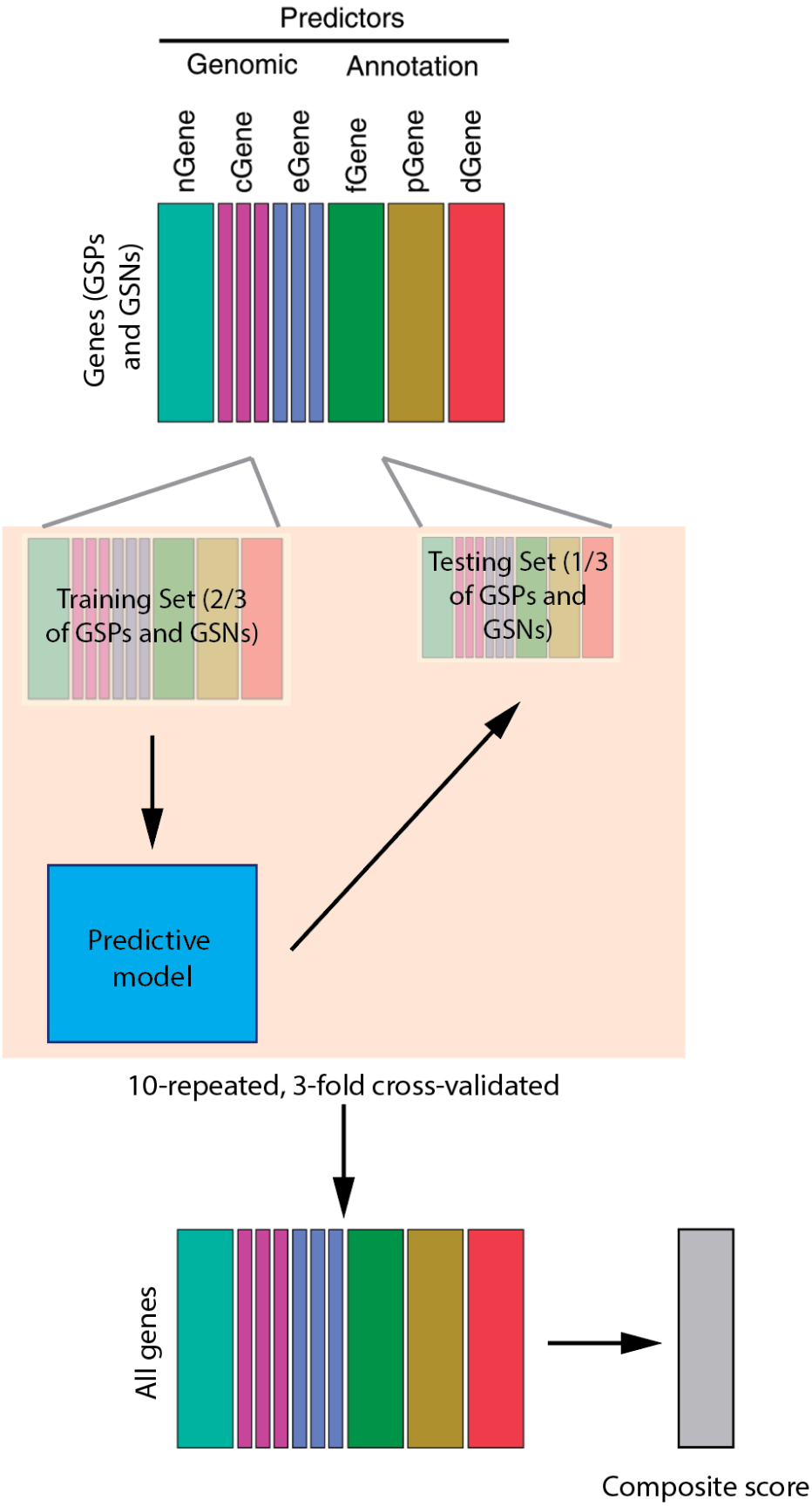
the full priority list.

### 3.2.7 Priority Index supervised mode

An alternate mode of Pi - 'supervised mode' - was developed in addition to the core 'discovery mode'. Supervised mode allowed for the use of GSP and GSN information earlier in the prioritisation pipeline; namely, GSPs and GSNs were used to inform the prioritisation algorithm itself, in order to predict genes similar to targets of approved drugs.

Fig. 3.2.16 depicts the steps involved in supervised mode. First, the gene-predictor matrix was generated for GSPs and GSNs, using genomic and annotation predictors (as done for discovery mode). Next, the GSPs and GSNs were divided into a training (2/3 of genes) and a testing (1/3 of genes) set, ensuring that the ratio of GSPs:GSNs remained representative of the whole. A classifier

**Figure 3.2.16** ***(following page)*****: Construction of Pi supervised mode.** A schematic outlining how Pi's supervised mode was constructed. The initial predictor gene matrix (top) was generated the same was as in discovery mode. From there, GSP and GSN targets were broken down into training and testing sets, and a random forest predictive model was built and tested using these groups. Each model was 3-fold cross-validated, and the entire process was repeated 10 times. The resulting 30 'mini models' were applied to the full gene list, and a final composite score was determined for each gene, representing the likelihood that a given target was a positive (0-1 scale).


Predictors
Genomic
Annotation
nGene
cGene
eGene
fGene
pGene
dGene
Genes (GSPs and GSNs)
Training Set (2/3 of GSPs and GSNs)
Testing Set (1/3 of GSPs and GSNs)
Predictive model
10-repeated, 3-fold cross-validated
All genes
Composite score

was then built using randomForest (version 4.6.14) in R. This is an ensemble learning classification method that consists of an aggregation of many decision trees, built on the training data and applied to the testing data. randomForest utilizes a bagging, or bootstrap aggregating, technique, where each decision tree is constructed on a random sample of the training data and 'sees' a random sample of the predictors. Thus, each tree is safeguarded from over-fitting to the data but, when taken collectively, the combined 'forest' of trees can classify very accurately. For the two main parameters used by randomForest (tree depth, or how many decision levels each tree has, and tree number, or the total size of the forest), values were optimised by running all possible combinations of tree depth and number, and selecting the values that had the best performance for input into the model. The 2/3 training, 1/3 testing subsets of GSPs/GSNs were then swapped (so that one of the training thirds became a testing set, and the testing set became part of the 2/3 training set), and model construction was repeated for each combination, resulting in a 3-fold cross-validated model. The entire process of splitting data, training, testing, and cross-validating was then repeated 9 additional times, resulting in an overall 10-repeated, 3-fold cross-validated model, or 30 randomForest models in total. The model building and optimisation steps were done using the Caret package (version 6.0-84) in R, in order to streamline the workflow. Finally, each of the 30 models was applied to the full gene-predictor matrix (all GSPs, GSNs, and putative targets), and an aggregate score was determined for each gene; this aggregate score represented the likelihood of a gene being classified as a GSP, ranging from 0-1, based on the genomic, annotation, and network feature vector; essentially, this quantified the similarity between GSPs and each non-gold-standard target. The output was then analysed in similar ways to the discovery mode output discussed previously.

A contrast of the top 30 prioritised genes in discovery versus supervised mode can be found in Table 3.2.1. As expected, there are a number of differences between the two outputs. Many of the genes in the top 30 of the supervised mode output, for example, are ranked so highly due to being targets of various existing MS drugs. The sphingosine-1-phosphate receptors (S1PRs), for example, are the targets of fingolimod, a well established, approved therapeutic for the treatment of MS [137]. Similarly, *IL2RB* is among the set of top rated genes (in both supervised and discovery mode), and is the target of daclizumab, which serves as an alternative option to other disease-modifying therapies

**Table 3.2.1: Top MS prioritised genes in Pi discovery versus supervised mode.** Top 30 genes based on discovery mode as well as supervised mode. Any genes that fall within the top 1% of both prioritisation modes are shaded in green.

| Gene Discovery Mode | Rank | Gene Supervised Mode | Rank |
|---|---|---|---|
| CD40 | 1 | S1PR4 | 1 |
| NFKB1 | 2 | S1PR3 | 2 |
| SATB1 | 3 | S1PR2 | 3 |
| IL12A | 4 | CCR2 | 4 |
| MAPK1 | 5 | S1PR1 | 5 |
| IL12B | 6 | CXCL10 | 6 |
| IL6ST | 7 | CCR1 | 7 |
| UBA52 | 8 | SST | 8 |
| IL10 | 9 | OPRM1 | 9 |
| IL2 | 10 | CXCL11 | 10 |
| IL4 | 11 | OPRD1 | 11 |
| IFNAR2 | 12 | IL2RB | 12 |
| IL2RA | 13 | IL2RG | 13 |
| ADRA1B | 14 | CCR9 | 14 |
| TYK2 | 15 | CXCR4 | 15 |
| STAT3 | 16 | CCL5 | 16 |
| IL7R | 17 | CCR5 | 17 |
| TYMP | 18 | PENK | 18 |
| CD27 | 19 | CCR3 | 19 |
| NFKBIA | 20 | CCL21 | 20 |
| TNFRSF1A | 21 | PYY | 21 |
| CBLB | 22 | PDYN | 22 |
| LCK | 23 | DRD2 | 23 |
| STAT1 | 24 | GRM6 | 24 |
| SOCS1 | 25 | ADRA2A | 25 |
| CD4 | 26 | SSTR2 | 26 |
| TRAF3 | 27 | CXCL1 | 27 |
| CD40LG | 28 | KNG1 | 28 |
| CSH1 | 29 | CXCL9 | 29 |
| CD8A | 30 | SAA1 | 30 |

for the treatment of relapsing MS [138]. Additionally, the C-C motif chemokine receptor 2 (*CCR2*) is the target of MLN-1202, which completed Phase 2 trials for the treatment of MS (results pending).

Unsurprisingly, most top-rated discovery mode genes were not at the top of the supervised mode output. The few that were, like *STAT3*, *IL10*, and *IL2*, were logical candidates. For example, *STAT3* has been GWAS-identified in the context of MS and has been replicated in a separate study [139], meaning that it was categorised as a genomic seed gene in both modes and received high seed score. *IL10* has a long-standing tie to MS; it was initially implicated in MS in the 1990s due to it's link to interferon beta, and has since been supported by other studies that analysed its role with respect to the clinical advantages of interferon beta in MS treatment [140, 141]. *IL2*, on the other hand, has been implicated through GWAS studies (along with *IL2RA* and *IL2RB*) and, more recently, was taken into clinical trials [128]; the drug MS-IL2 is recruiting for Phase 2 testing for its potential role as an MS therapeutic (Clinical Trials identifier: NCT02424396).

Similarly, from a pathway perspective the supervised mode output had both points of overlap as well as differentiation from the discovery mode. For instance, IL2 signaling was significantly enriched among the top genes of the supervised mode output (FDR = $7.5 \times 10^{-11}$) as well as in discovery mode (FDR = $1.7 \times 10^{-9}$), which is unsurprising given the above-mentioned link between *IL2* and MS. Likewise, IL23-mediated signaling was enriched in both supervised (FDR = $2.3 \times 10^{-14}$) as well as in discovery mode (FDR = $6.5 \times 10^{-15}$), which is supported by the established role of IL23 signalling in autoimmune inflammation of the brain, and its role in the development and maintenance of Th17 cells [142, 143]. On the other hand, CXCR3-mediated signalling events was uniquely enriched in supervised mode (FDR = $1.9 \times 10^{-8}$), and is supported by literature that has connected CXCR3 to the accumulation of T-cells in the central nervous system of MS patients; in fact, a number of related genes (CXCL10, CXCL12, CXCR4, CXCR7) have been previously implicated in regulation of remyelination and reviewed for their potential therapeutic roles in MS [144–146]. A comprehensive plot of similar versus unique pathways (between the two modes) can be found in Supplementary Figure 8.0.1.

### 3.2.8 Performance assessment of supervised mode

As with discovery mode, it was necessary to assess the performance of supervised mode in its ability to prioritise GSPs over GSNs. Additionally, since supervised mode utilized knowledge of GSPs and GSNs within the algorithm that did the prioritisation, higher performance metrics were expected relative to discovery mode. Once again, the focus within this section is on MS, due to the high number of drug targets that were available for machine learning training in the prioritisation pipeline.

#### ROC curve assessment and discovery mode comparison

Similar to the analysis done within the discovery mode context, ROC curves were plotted for different GSP definitions. Specifically, the Pi supervised mode output was assessed for its ability to prioritise GSPs over GSNs, with GSPs defined as either Phase 3+ targets or Phase 4+ targets. Fig. 3.2.17 shows the supervised mode ROC curves for these two GSP cutoffs, alongside the corresponding ROC curves for discovery mode. In both definitions of GSPs, Pi supervised mode greatly outperformed discovery mode (AUC of 0.92 vs 0.81 in Phase 3+; 0.94 vs 0.83 in Phase 4+), which was the expected result.

#### Target distribution in supervised mode

Within the discovery mode output, testing for target distribution served one primary purpose: to ensure that GSPs had a tendency to fall near the top of the priority list. GSNs, on the other hand, were more evenly distributed throughout the list. In contrast, target distribution in supervised mode tested for two elements: in addition to GSP clustering at the top of the priority list, GSNs should be clustering near the bottom of the list, due to the fact that an algorithm was trained with explicit GSP vs GSN labels. Fig. 3.2.18 shows this GSP/GSN target distribution based on the supervised mode output, with a Phase 3+ GSP definition. In contrast to the discovery mode plot, GSPs in supervised mode were clustered more tightly at the top of the priority list, while GSNs were heavily skewed toward the bottom of the list. Putative targets (those not defined as GSPs or GSNs) were, as expected,

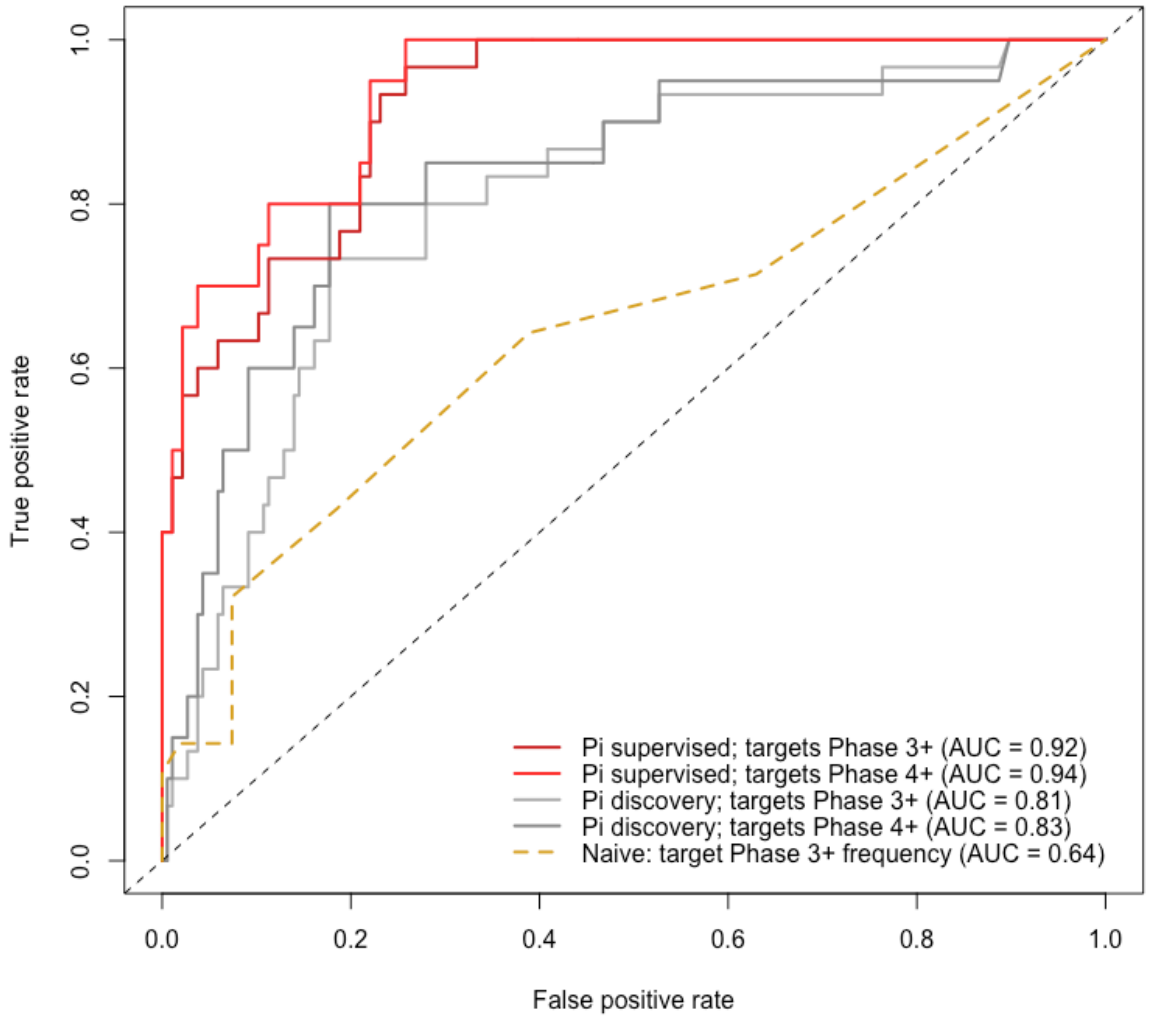



**Figure 3.2.17: Performance assessment of Pi supervised mode in MS.** ROC curves for Phase 3+ and Phase 4+ GSPs, as captured by the supervised mode MS output of Pi (red lines), compared to the same GSPs captured by the discovery mode (grey lines).

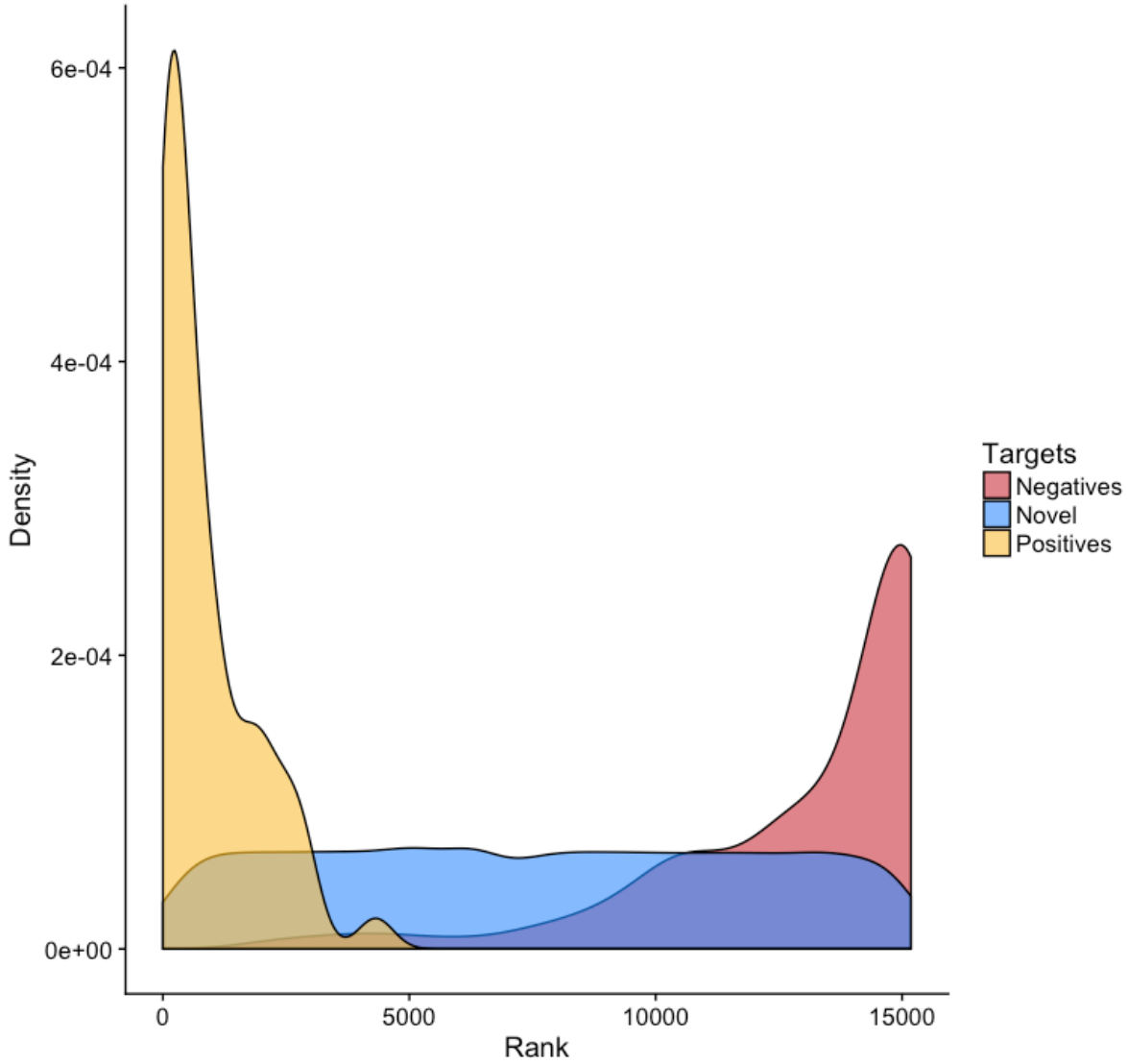



**Figure 3.2.18: Target distribution of Pi supervised mode in MS.** A density plot of Phase 3+ GSPs in the supervised mode MS output, along with corresponding GSNs, and putative target genes.

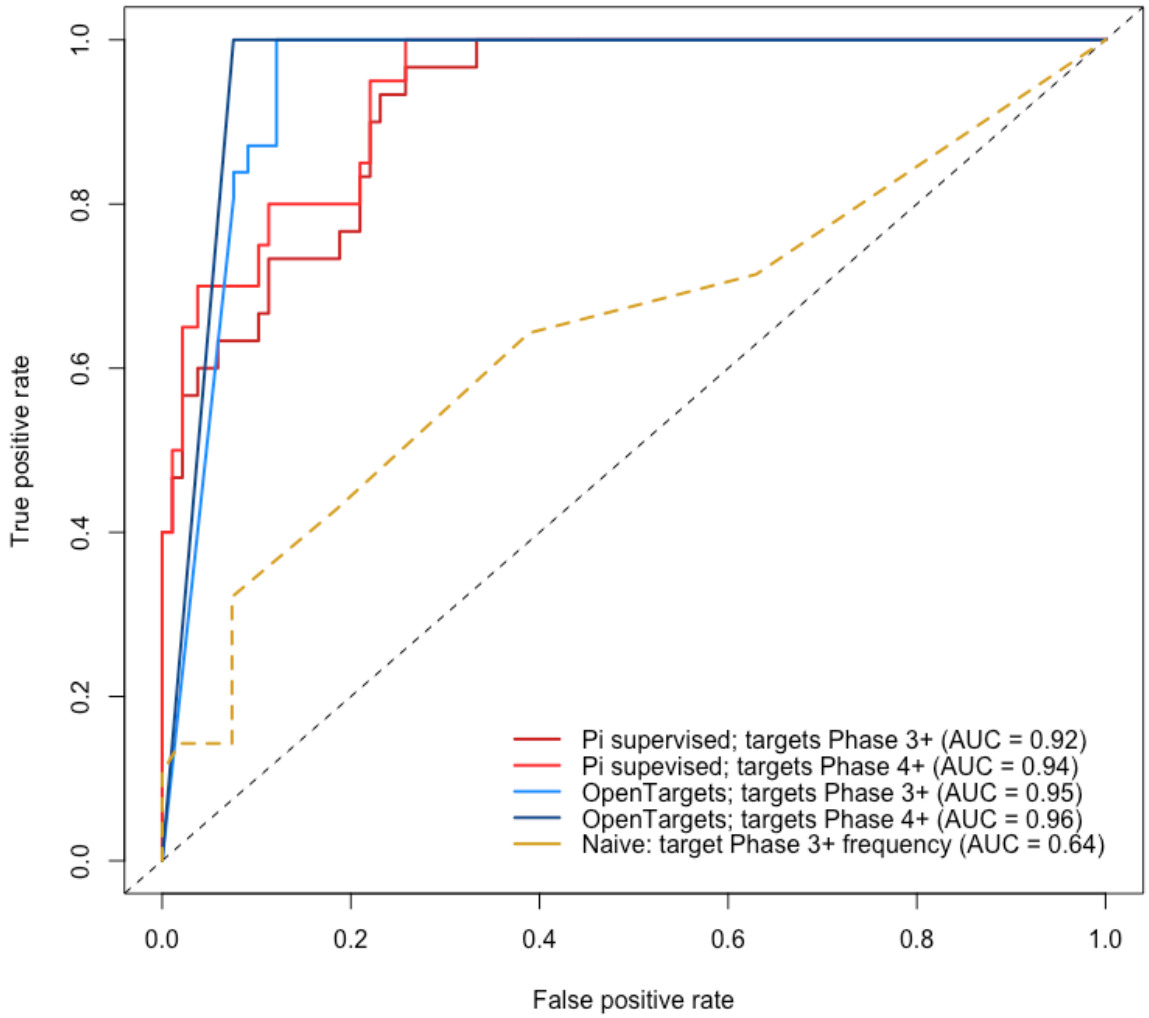



**Figure 3.2.19: Supervised mode benchmarked against OpenTargets.** ROC curves for the MS supervised mode output (red) and the corresponding OpenTargets MS output (blue), using the full, unrestricted set of OpenTargets predictors, across Phase 3+ and Phase 4+ GSP definitions.

evenly distributed throughout, with those near the top of the list (e.g. ranked 1-5000) sharing more genomic feature similarity with GSPs than with GSNs.

### Comparison of supervised mode to OpenTargets

Since supervised mode utilised knowledge of known GSPs and GSNs, performance in this section was benchmarked to the unrestricted outputs of OpenTargets. Fig. 3.2.19 shows the AUC values for the Pi vs OpenTargets outputs, across both a Phase 3+ and a Phase 4+ definition of GSPs. In both instances, OpenTargets slightly outperformed Pi (e.g. Pi AUC Phase 3+ = 0.92, OpenTargets AUC Phase 3+ =

0.95), which was not surprising given the more comprehensive nature of the data-sets used to build the OpenTargets predictors. For instance, OpenTargets incorporates human genetics, known drug information, RNA expression, text mining, and animal models into its combined output, whereas Pi supervised mode does not use RNA-seq, text mining, or animal models in building the predictor.

### 3.2.9 Supervised mode captures differentially expressed genes based on MS subgroup transcriptional profiling

Given the ability of discriminating GSPs from GSNs with high AUC using supervised mode, a related hypothesis to test was whether supervised mode could also capture differences between MS patients groups differing in disease activity, following treatment. In order to do this, data from the Ottoboni *et al.* (2012) study was used [147]; specifically, the genes which the authors identified as differentially expressed between patients with lower versus higher disease activity (in PBMCs) following treatment with either of two first-line MS disease-modifying treatments (glatiramer acetate and interferon beta). These genes (meeting FDR < 0.05, fold-change > 2) were contrasted with the Pi supervised mode output; among top 1% Pi MS genes, there was a 4.69-fold enrichment of the differential disease activity genes ($p_{hyper} = 0.026$). The top Pi-ranked genes from the transcriptional profiling study were *CCR2*, *PIK3R1*, and *PIK3CA*.

### 3.2.10 Use of supervised mode to inform re-purposing potential in MS

In addition to prioritising genes that are similar to known targets of effective therapeutics, there was the potential to test for overlap between top prioritised MS targets and targets of non-MS immune therapeutics, in order to identify re-purposing potential (i.e. for the use of non-MS drugs that target high priority genes in an MS-context). First, approved, non-MS, immune-related drugs were mapped to their targets using ChEMBL [46]. Next, targets of approved MS therapeutics and their first degree protein interaction partners, mapped using the HuRI database [148], were combined with top 1% MS supervised mode genes. Lastly, these two lists (approved, non-MS drug targets and top priority Pi MS genes) were contrasted; results are summarised in Table 3.2.2, and overlap between the two groups

**Table 3.2.2: Overlap between non-MS immune drugs and top priority MS genes.** The intersect between targets of non-MS immune-related drugs and top priority genes in the Pi supervised mode.

| Drug | Targets | Pi Supervised Rank |
|---|---|---|
| USTEKINUMAB | IL12B | 38 |
| CANAKINUMAB | IL1B | 114 |
| BASILIXIMAB | IL2RA, IL2RB, IL2RG | 105, 12, 13 |
| ADALIMUMAB | TNF | 65 |
| CERTOLIZUMAB | TNF | 65 |
| GOLIMUMAB | TNF | 65 |
| INFLIXIMAB | TNF | 65 |

included canakinumab (targets IL1B) and basiliximab (targets IL2RA/B/G).

## 3.3 Discussion

The results presented in this chapter outlined how a computational tool (Pi) could be used to prioritise disease-relevant genes using GWAS data and downstream processing which combines functional genomics, annotations, and gene network interactions. More importantly, this chapter outlined a series of performance evaluation approaches that interrogated the predictive value of Pi outputs, in order to encourage the use of such an approach in a pre-clinical drug discovery pipeline. Overall, the performance evaluation results strongly supported the use of Pi, as evaluation metrics were relatively high across a range of different tests. Further, an alternate version of Pi was developed (supervised mode), which leverages known drug information via machine learning in order to conduct prioritisation.

### Application of discovery mode to MS

Before conducting a purely quantitative, systematic performance evaluation, a more MS-specific analysis was done using the discovery mode output. Namely, top prioritised genes and corresponding enriched pathways were analysed in the context of what is known about MS. Importantly, a number of previously identified candidates were captured at the gene-level, including *CD40* [125, 126], *IFNAR2*

[127], *PTGER4* [129, 130], and *STAT1* and *STAT3* [149, 150]. Likewise, the pathway-level data also captured key information that has been previously reported with regard to MS development and progression; for instance, TCR signaling in both CD4+ and CD8+ T-cells is important in MS disease context [131, 132], as is IL12-mediated signalling [129, 133–135]. Establishing these connections between Pi outputs and existing literature was a key step in validating the framework as a whole. In particular, the fact that discovery mode – a genetics-led approach that consolidates GWAS data with various additional genomics predictors – was able to capture highly relevant targets and pathways at the top end of the prioritisation output was reassuring and lends a broad level of credibility to other, more novel results at the top end of the output. While the majority of this chapter focused on more systematic, gold-standard-based interrogation of outputs, the following results chapter (*Application of the prioritisation pipeline to myasthenia gravis*), explores how such Pi outputs can be further investigated and supplemented with various complementary pieces of data, as well as taken further to a more specific understanding of disease mechanism.

### Performance evaluation using Golden Standards

As mentioned above, the main approaches taken for systematic performance evaluation relied on the leveraging of gold standards; initially, precision-recall was selected in order to determine how highly Pi was prioritising GSPs. However, as discussed earlier, this resulted in poor PR values, which was unsurprising given that Pi (a) does not take into account known GSPs when conducting prioritisation and (b) the primary aim of the algorithm is to uncover novel targets, none of which would be classified as positives in a PR analysis. Therefore, the bulk of performance evaluation utilised ROC curves, which were quantified with AUC values. The advantage of this approach is that it tests for the ability to discriminate between positives versus negatives, without taking putative targets into consideration. If high AUC is achieved, then one can more reliably say that putative genes that rank near known positives are likely candidate targets themselves.

While AUC values are a useful quantification of performance, they will depend entirely on the question being asked (i.e. how positives and negatives are being defined). Therefore, it was critical to consider multiple angles from which to approach ROC curve analysis, in order to ensure that high

AUC was not due to an unintentional confounding factor(s). First, AUC was calculated for a sliding window of GSP definition, starting from 'completed Phase 2' to 'completed Phase 4'. The fact that AUC increased as GSP definition encompassed an increasing proportion of later-stage targets was highly promising, since this suggests that Pi is better at separating positives from negatives when a higher bar is set for something to count as a positive. In fact, AUC was highest when 'clinical golden standards' were used to define GSPs (Fig. 3.2.13), which affirms this conclusion. Importantly, the GSN pool remained unchanged throughout these analyses; since GSNs were simulated, it was crucial to ensure that a very stringent mechanism was applied in this simulation, so that targets labeled as GSNs would have a very low probability of actually being undiscovered positives.

One key caveat to the use of AUC in the context of Pi evaluation is that it does not inherently take into consideration where positives and negatives are ranked. In other words, it is theoretically possible to have a very high AUC value, with all GSPs and GSNs ranked in the bottom portion of the full priority list (since putative/unknown targets do not affect AUC). The target density analyses were therefore key in showing that this was not the case; GSPs actually were clustering near the top of the ranked output, while GSNs were more evenly distributed throughout. This density distribution was further interrogated using enrichment tests, which showed that there was a highly statistically significant enrichment of GSPs in the top of the output. Given that known drug information was not used to construct these lists, this result was reassuring, since it implies that putative targets at the top of the ranked output are clustering with the majority of GSPs. A breakdown of GSPs by phase showed that Phase 4 targets were most densely contained in the top of the output relative to other phases, which again suggested that Pi was more highly rating targets that had strong supporting evidence of efficacy in humans.

As discussed, a consistent set of parameters was used to define simulated GSNs. However, the results presented still hinge on the certainty that the simulation approach used is adequate in reflecting 'true negatives'. Therefore, the series of control analyses in Fig. 3.2.11 were instrumental in testing the robustness of Pi's performance output, including both an alternate form of GSN definition, as well as disease and drug specificity to MS. This finding was very important because it supported the use of GSN simulation as an alternate to manually curated negatives (i.e. since the AUC values remained

unchanged), which allows for widely scalable and adaptable use across other diseases, and it confirmed that high AUC value was disease-specific and unlikely to be occurring by chance or due to confounding variables. The further performance tests across four other diseases (RA, SLE, UC, and CRO) had comparable AUC values, confirming that good performance was not unique to MS as a disease. Moreover, benchmarking against Open Targets was useful in demonstrating the advantages of using Pi to integrate a number of varied data-sets, when the goal is to have a purely human, genetics-led approach focusing on GWAS interpretation.

### Supervised mode analysis

In addition to the discovery mode outputs, a supervised mode output was similarly analysed. There was some overlap between the top-rated genes of these two prioritisation lists, such as in the case of *STAT3*, *IL10*, and *IL2*, all of which were unsurprising given previously established genetic and therapeutic links between these genes and MS [128, 139–141]. Likewise, the areas of differentiation between the supervised mode and discovery mode were expected; for instance, prioritisation of S1PRs made sense given the therapeutic benefits of fingolimod in MS [137]. Perhaps the more interesting targets identified by this approach were less obvious potential candidates like *CCR2* and the other CC-motif chemokine receptors prioritised alongside this gene; results for a Phase 2 trial of MLN-1202, which targets *CCL2*, in MS patients are pending. It is possible that this family of targets is viable for therapeutic advancement of MS treatment – *CCR5*, for example, contains druggable pockets based on an *fpocket* analysis of the target [151], and therefore may be worth exploring further. In fact, CCR5 has previously been proposed and reviewed as a relevant target in the context of MS [152], which Pi supervised mode analysis supports. While *CCR5* is not simultaneously in the top 1% of the discovery mode output, it is still relatively highly ranked (top 3%) supporting the genetics linking this target to disease relevance.

Interestingly, *CCR2* was also identified in the analysis which contrasted Pi supervised mode to the post-treatment, differential transcriptional profile of two MS subgroups, categorised on the basis of disease activity [147]. The study found that *CCR2* was a significantly differentially expressed gene between MS patients with high versus low disease activity post-treatment; taken with the

above-mentioned exploration of MLN-1202 for MS treatment, as well as the high priority score of this gene in Pi supervised more, it is possible that this family of targets may have a potential therapeutic role in the treatment of MS. The other promising aspect of this analysis was the fact that the Pi supervised mode output top 1% targets were significantly enriched for the genes identified in the Ottoboni *et al.* study [147]. Effectively, this served to extend the validation of the Pi supervised mode output beyond simply higher GSP-GSN AUC, to also include post-treatment gene expression differences between MS patients with high versus low disease activity. While this finding does not provide a definitive set of novel targets for MS treatment, it does support the hypothesis that Pi supervised mode captures disease and therapeutic-relevant targets, and extends the utility of that mode beyond simple prioritisation of known targets of existing therapeutics to also include those that are newly emerging in the clinical development space or those that have complementary pieces of data supporting their role in or relevance to disease.

### Utility of supervised mode for drug repositioning

Further, the re-purposing analysis highlighted another potential role for Pi's supervised mode; namely, it showcased the potential to identify new opportunities for the use of existing immune drugs in an MS context. More specifically, the rationale was as follows: since supervised mode first processes genes on the basis of aggregated genomic and network information and then prioritises them on the basis of feature similarity to existing, disease-relevant therapeutics, top priority genes in supervised mode are by definition more similar to targets of existing drugs, and are thus, in theory, more likely to be of therapeutic relevance. Since many immune-related drugs can be used in multiple disease settings, the question at hand was whether a contrast of the targets of non-MS drugs against high-priority MS targets would yield any overlap. If so, then this overlap could be used to inform a re-purposing event, where a non-MS drug could be considered for clinical trials in an MS context, on the basis of the similarity between that drug's target(s) and MS genetic and therapeutic knowledge. Conducting this analysis in MS yielded a number of interesting results, with two groups of drugs emerging. The first group (adalimumab, certolizumab, golimumab, and infliximab) are anti-TNFs, meaning they are unlikely to be beneficial in an MS context given the reverse effect this has compared

to a disease setting like rheumatoid arthritis, for example; in fact, there is documented evidence of MS disease onset with following adalimumab treatment of rheumatoid, for example [153]. The other group of drugs/targets were more relevant to the question; while ustekinumab (targeting IL12B) has been trialed in a Phase 2, double-blinded, placebo-controlled study without success [154], canakinumab (targeting IL1B) and basiliximab (targeting IL2RA/B/G) are potentially more promising. For instance, the treatment of inflammatory disease via blockage of interluekin-1 has been previously reviewed, with some evidence suggesting a potential role in MS [155], and thus a potential for canakinumab, a human IgG1 antibody designed against IL1B [156, 157], to be therapeutically relevant; a study which investigated anakinra (targeting interleukin-1) found a decrease in demyelinating events following treatment [158], supporting this idea of IL1B target re-purposing for MS. Similarly, the potential for basiliximab in MS treatment is supported by existing evidence linking the targets to disease; *IL2* and *IL2RA* have known associations between polymorphisms in these genes and MS risk [128], while *IL2RB* is the target of daclizumab, which serves as an alternative option to other disease-modifying therapies for the treatment of relapsing MS [138]. Thus, there is a potential for the investigation of basiliximab in an MS context based on the approach outlined here. Moreover, this approach can be systematically implemented across multiple diseases and applied in a batch manner, which would allow for a wide 'scan of re-purposing potential' that could be used to inform the starting point of such future studies in a number of other diseases as well.

The implementation of supervised mode brings with it some important considerations regarding potential advantages versus disadvantages of such an approach. The underlying rationale is clear: using feature vectors (comprised of genomic, annotation, and network data) to rank genes based on similarity and dissimilarity between their feature vectors and those of GSPs versus GSNs, respectively. However, such an approach is inherently biased in that it essentially prioritises maximisation of feature matching between putative and GSP targets in order to come out with a final rank, rather than using biological/genetic data to drive ranking, as discovery mode does. Moreover, how those specific features/predictors are actually used in the prioritisation is itself not straightforward, and the 'black box' argument holds true in this instance. In other words, it becomes quite difficult to disentangle the reasoning behind how the algorithm reached a specific priority rank, especially when there are so many predictors involved. Some measures, like predictor importance as a function of change in AUC

when the predictor is removed from the model, can be useful proxies for interrogating the supervised learning model, but still do not always yield a very clear/intuitive or complete answer.

On the other hand, the clearest advantage to the supervised mode approach is that it performs much better at separating GSPs from GSNs; by extension, this implies that highly prioritised putative targets are more likely to be 'positives' in a therapeutic sense. Accordingly, AUC values were significantly higher in the supervised mode of Pi as compared to discovery mode. Further, the very clear boundary in target density distribution was encouraging, and could be informative in novel target nomination. Specifically, because there is simultaneously a very high concentration of GSPs and a very low concentration of GSNs at the top of the supervised priority output (Fig. 3.2.18), putative genes that fall into this top portion are likely to share many genomic features with GSPs (and, by extension, be quite dissimilar to GSNs). Since one trade off to using a supervised learning approach is that prioritisation is being biased by existing drug information, it might be interesting to leverage this GSP/GSN density split as a downstream step to the unbiased discovery mode. In other words, one could overlay the supervised mode output onto the discovery mode output, and see which genes fall into the top end of both prioritisation modes. This way, genes that are highly prioritised by discovery mode (i.e. unbiased, genetics-led, with intuitive reasoning) and that also cluster with GSPs in supervised mode (i.e. share very similar predictor patterns/signatures) form compelling targets for clinical follow-on.

Lastly, it is very important to note that the utility of performance evaluation based on GSP/GSN discrimination will always depend entirely on the confidence in the quality of said golden standards. The approach used in this chapter was selected for a combination of reasons:

I. There is extensive, publicly available data on disease-specific drug-target combinations, which includes information on clinical phase of development that can be leveraged for defining GSPs across diseases;

II. It is relatively straightforward to tweak the script necessary for simulating a large enough group of GSNs, meaning that it can be adapted to any diseases where GSPs have been identified;

III. A simulation approach allows one to avoid having to navigate the noisy data surrounding why

certain targets failed in clinical trials.

The results presented here showed how such an approach could be implemented for MS, and provided a series of controls to demonstrate the robustness of the high AUC values when other methods were implemented (e.g. when a more manual approach was used for GSP and GSN definition). However, there may be instances where a simulation-based approach is misleading in that it leads to a number of targets being classified as false negatives. Therefore, there would be great utility in having a list of negatives that was directly and systematically compiled from the results of failed clinical targets, *where there is high confidence in the experimental design of the trial and where the reason for failure was strictly due to lack of efficacy in treating disease*. Something like this, however, would only be possible with internal knowledge of various pharmaceutical drug development pipelines; until such a time, alternate approaches of defining golden standards like the ones described here are possible substitutes.

## 3.4 Conclusion

In conclusion, the results of this chapter demonstrated how genomic information could be used to prioritise genes through the use of the Pi pipeline, which brings together GWAS data, functional genomics, annotations, and network information in order to score genes in a disease-relevant manner. Moreover, the work highlighted a robust method for computationally evaluating the performance of Pi, a critical step for supporting the adoption of Pi in the broader drug development community, as a pre-clinical tool that can be either be used to prioritise or de-prioritise existing putative targets, or nominate entirely novel genes. Further and importantly, results showed that Pi outperforms existing genetics-led approaches for target prioritisation. Lastly, an alternate mode of Pi - supervised mode- was explained, showcasing an opportunity for the use of known drug target information during computational prioritisation, which can be a powerfully complementary approach to target ranking.

*Not all those who wander are lost.*

J.R.R. Tolkien

# 4

# Application of the prioritisation pipeline to myasthenia gravis

## 4.1 Introduction

Myasthenia gravis (MG) is an autoimmune disorder that is characterised by autoantibodies against components of the post-synaptic muscle endplate, which manifests clinically in muscle weakness and fatigue, with a prevalence of around 20-90 per 100,000 [159, 160]. Symptoms tend to worsen after prolonged activity, and can affect eye movement, facial expression, chewing, talking, and swallowing, among other things [160]. Diagnosing myasthenia gravis is not always straightforward, due to similarity of symptoms to other diseases, as well as the fluctuating nature of symptoms, and often requires lab testing or serological examination [159]. Clinical manifestations are brought about due to an autoimmune response against receptors at the neuromuscular junction (usually the alpha subunit of acetylcholine receptors, but sometimes to muscle-specific tyrosine kinase instead), causing a reduced ability of motor nerves to induce muscular contraction.

### Treatment options

Accordingly, patients with MG receive symptomatic relief through acetylcholinesterase inhibitors but often require additional treatment with immunosuppression [159, 160]. For example, drugs like pyridostigmine, neostigmine, and ambenonium chloride are preferred candidates for symptomatic relief, as they all work to increase the amount of acetylcholine at the neuromuscular endplate in order to help overcome the inability of muscle contraction; long-term use of these acetylcholinesterase inhibitors is generally deemed safe as cumulative side effects of the treatment have not been reported [160]. For patients who do not respond well to this type of disease management, drugs like prednisone, prednisolone, and azathioprine are used as first in line immunosuppressants, and work well across all myasthenia gravis subtypes; for patients who do not respond to any of these, mycophenolate mofetil, rituximab, and methotrexate are treatment options as well [160]. Additionally, a recent large randomised controlled trial has demonstrated that thymectomy is beneficial in non-thymoma-associated patients [161] providing support for a central role of the thymus in the disease pathogenesis and another option for treatment. Moreover, epidemiological studies show that there are peaks in age of onset [159]; early-onset (EO) MG is usually defined as onset of symptoms before the age of 50, while late-onset (LO) MG encompasses onset after age 50. Studies have shown that EO patients are more likely to suffer from other autoimmune disorders in their families and respond better to thymectomy compared to LO [160]. However, relatively little else is known about underlying differences between the two forms of disease, or about what specific biological mechanisms are driving the development of the autoimmune response in MG.

### Genetic associations to disease

Simultaneously, genome-wide association studies (GWAS) have uncovered a number of genetic associations in MG [68–70]. The three primary studies focused on distinct patient groups; Renton and colleagues [69] collected DNA from white Americans with mixed onset of disease (both as a pooled group as well as separately for EO and LO forms), for a total of 1032 patients and 1998 controls as well as an independent cohort of 423 patients and 467 controls; Gregersen and colleagues [68] focused on early onset disease, and recruited 649 patients of North European ancestry and 2596

controls of European ancestry; Seldin and colleagues [70] conducted their study on the late onset form of disease, with 532 cases and 2128 controls, all of European ancestry. All three studies were done with acetylcholine receptor antibody-positive patients.

In line with clinical differences, these GWAS have identified genetic differences between EO and LO forms of the disease highlighting differences in the underlying pathological processes. For instance, single nucleotide polymorphisms (SNPs) in *TNIP1*, a gene involved in the regulation of NF-kB activation, are associated with EO myasthenia gravis [68, 162]. On the other hand, SNPs in *PTPN22*, a gene involved in affecting the responsiveness of T and B-cells, show associations in both groups [68, 70]. Despite this early success, GWAS regions often contain many candidate genes and further work is needed to identify the causal genes and the cell types in which these associated genes exert functional effects. Moreover, the growing body of other available genomic data-sets (such as various cell or tissue-specific RNA-seq, genome-wide CRISPR-Cas9 screens of immune regulatory genes, etc.) provides an opportunity for leveraging multiple levers of evidence in order to uncover key genes or components of disease progression, in order to better understand development and more effectively tailor treatment.

### Aims

In chapter three, I introduced the computational pipeline Pi, which allows a user to prioritise disease-relevant genes based on downstream analysis of GWAS data, by aggregating a variety of genomic predictors and utilising knowledge of genome-wide interaction networks. The hope is that, through the use of Pi and through the growing body of publicly available genomic data-sets, it will be possible to glean new insight into a variety of diseases, and translate this insight into actionable follow-up. Therefore, the aims of this chapter are as follows:

I. Develop an understanding of disease-relevant tissues by directly analysing current myasthenia gravis GWAS data (i.e. significant, disease-associated variants).

II. Demonstrate how Priority Index can be applied to this novel disease context in order to prioritise target genes on the basis of aggregated genomic predictors and network effects, and

uncover potential drivers of pathogenesis. Myasthenia gravis was selected for a few different reasons; namely, it is (i) an immune-related disease with (ii) many unknowns about true pathogenesis but with (iii) a growing body of publicly-available genomic data-sets, making it an ideal candidate for Priority Index application;

III. Process/analyse the Pi outputs at the pathway and network levels, and contrast this to previously published and complementary data-sets;

IV. Develop a hypothesis around key disease-driving targets based on aims I-III, and test this hypothesis at the variant-level.

## 4.2 Results

### 4.2.1 SNP selection and integration with functional annotations

SNPs were collected from three previously published GWAS in MG [68–70]. In order to investigate functional signals that may be underlying the disease associations, a genome-wide enrichment test was done using the GARFIELD tool [105]. Briefly, the GWA signals were first pruned (greedy pruning, $r^2 > 0.01$) in order to interrogate the most likely signal-driving SNPs from the input summary data. Next, association signals were tagged with respective genomic annotations, taking into account population LD structure ($r^2 > 0.8$). Finally, a generalized linear model is applied to the variant-annotation matrix to test for enrichment across a range of GWAS p-values, while also correcting for features such as distance to transcription start site and number of LD proxies.

In order to then continue with downstream analysis of MG-associated SNPs, all SNPs that met a combined p-value cut-off of $5 \times 10^{-5}$ were selected, and any SNPs in the major histocompatibility complex were removed due to the high linkage disequilibrium (LD) in this region. Once pooled, each SNP was categorised into either EO or LO groups based on reported findings from the original studies. Given that GWAS reported SNPs are not always causal, LD proxies to reported SNPs were included in the subsequent analyses. For LD calculation, the Broad Institute's web-based tool SNAP [108] was used to identify all SNPs that were in LD ($r^2 = 0.8$ or higher) with the GWAS reported

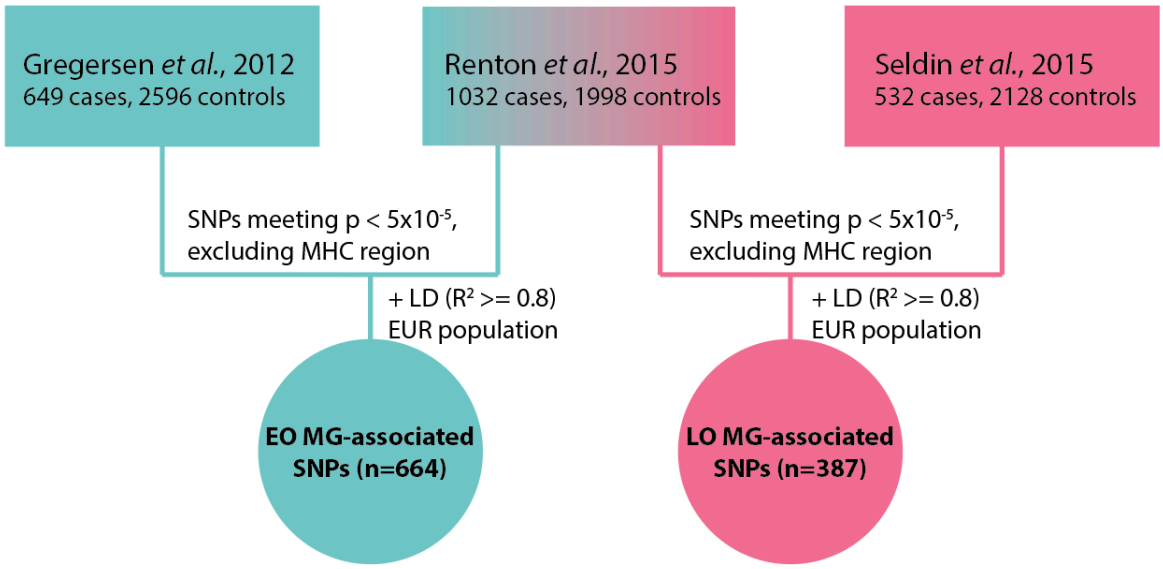



**Figure 4.2.1: Sourcing and pooling of MG SNPs.** A schematic outlining where MG-associated SNPs were sourced from and how sources were pooled for EO and LO analysis. Any SNPs that showed an association p-value of $< 5 \times 10^{-5}$ were included from each of the three studies.

SNPs in the European population. Subsequent analyses were carried out separately for EO and LO SNPs, as well as for a combined group. A schematic of SNP selection/pooling is shown in Fig. 4.2.1.

### 4.2.2 Functional annotation of MG SNPs

The first aim was to develop an understanding of disease-relevant information captured directly by existing myasthenia gravis GWAS data (i.e. significant, disease-associated variants), before any downstream analysis of GWAS was conducted. To do so, I wanted to understand whether disease-associated SNPs were directly enriched for any cell or tissue-specific annotations. In other words, were MG GWAS signals overlapping with any regulatory elements more often than expected by random chance and, if so, where were those regulatory elements located. The rationale here was that looking across the entire spectrum of tissues would help identify any highly significant signals suggesting a functional effect of MG SNPs on gene regulation, in a particular cell type. This could then be used to inform or complement later, downstream Pi analyses.

Therefore, disease-associated SNPs were analyzed for enrichment with GARFIELD, a tool that allows the user to measure enrichment across tissue; Fig. 4.2.2 shows enrichment results based on DNase

I-hypersensitivity sites (hotspots), sourced from ENCODE [106] and the NIH Roadmap Epigenomics data [107], across a range of tissue types. Through this analysis, it was immediately apparent that blood cell types were of relevance, as a substantial portion of the enriched signals were found in blood. More specifically, notable peaks include CD3 primary cells (OR = 7.24, p = 0.019), mobilized CD4 primary cells (OR = 7.05, p = 0.009), CD19 primary cells (OR = 5.30, p = 0.032), CD8 primary cells (OR = 5.18, p = 0.023), and CD56 primary cells (OR = 4.89, p = 0.029). Critically, there were smaller groups of overlap in fetal thymus and muscle (among others) which was encouraging due to the known role of those tissues in MG. It is important to note, however, that the peaks represented in Fig. 4.2.2 do not meet GARFIELD-adjusted p-value cutoffs that account for multiple testing, and so the above-mentioned peaks cannot be said to be statistically significant tissue-wide based on this analysis. This may be due to the p-adjustment method implemented by GARFIELD; effectively, the algorithm first computes a value called $M_{eff}$ which represents total enrichment of input signal (regardless of tissue source) as compared to a random set of SNPs, to calculate how significantly enriched with regulatory sites the input is overall. This $M_{eff}$ value is then used as a threshold for adjusting individual p-values for each peak, with the criteria being that individual peaks are called as tissue-wide significant if their calculated p-values are more significant than the significance of the total overlap of all tissue regulatory sites and all input SNPs. Therefore, it is likely the case that this approach, combined with the fact that there are relatively few MG-associated SNPs to begin with and that most initial input SNPs were pruned in the greedy pruning step of GARFIELD, results in a conservative final output that underestimates true significance of peaks. Thus, this result was used as an initial first-pass test to see which tissues MG SNP-associated regulatory elements were most significantly enriched in, and whether this was later confirmed by other, independent analyses in this chapter.

### 4.2.3 The disease-associated SNPs influence gene expression in T-cells

Given the immune-related nature of MG and the indication of blood cell regulatory element importance in disease based on GARFIELD analysis, an interesting question emerged: were MG

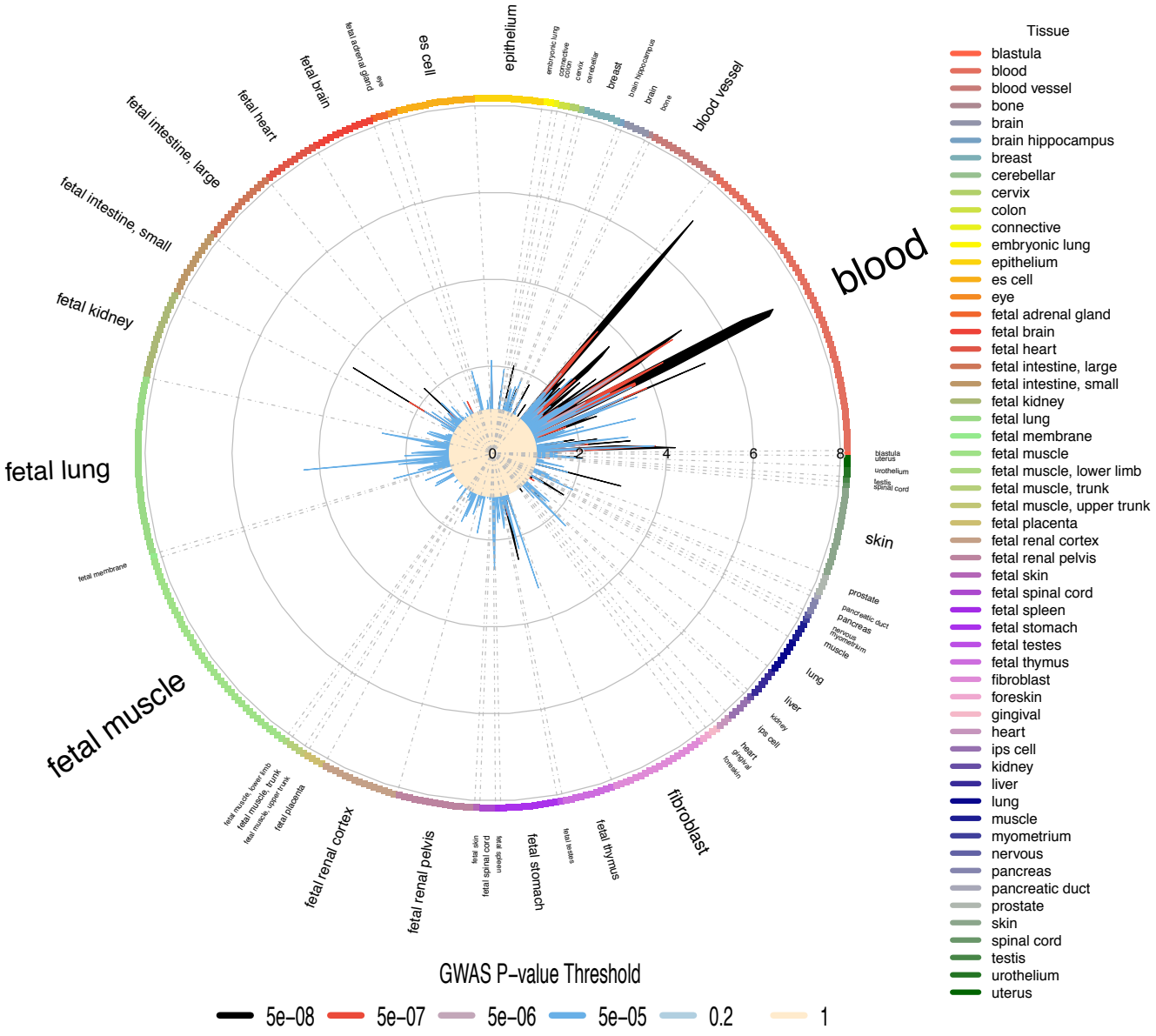



**Figure 4.2.2: Enrichment of MG GWAS signals in DHS hotspots.** Overlap between MG-associated SNPs and tissue-specific DNase I-hypersensitivity sites from ENCODE and Roadmap Epigenomics data. Radial lines denote odds ratio (OR), and tissue label sizes represent number of cell types.

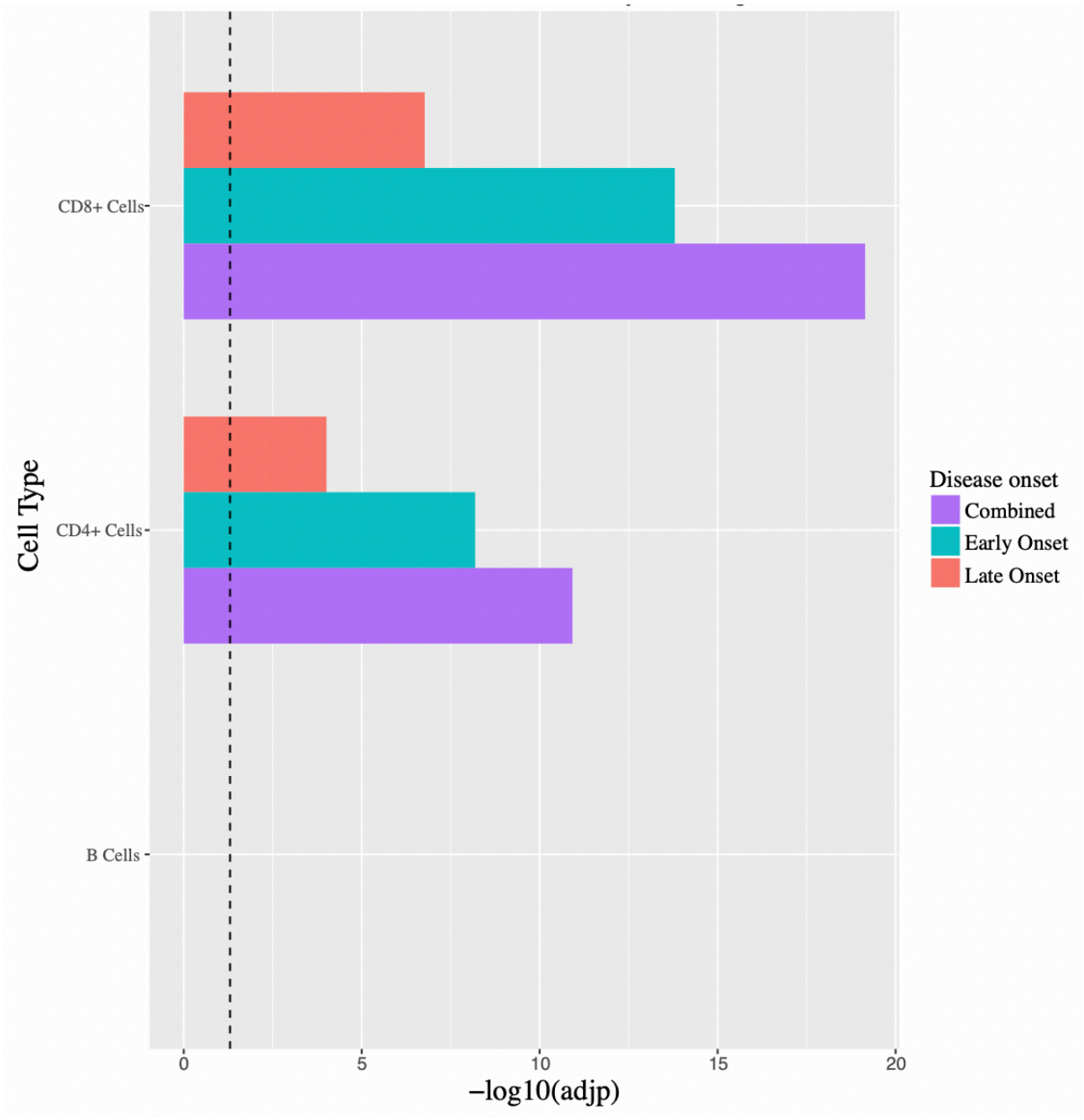



**Figure 4.2.3: eQTL enrichment in MG SNPs.** Enrichment results of the overlap between MG-associated SNPs (EO vs LO vs combined) and cell-specific eQTL (B-cells, monocytes, macrophages, neutrophils, NK cells, and $CD4^+$ and $CD8^+$ T-cells), with all GWAS SNPs meeting significance of $p < 5.0 \times 10^{-5}$ as background. Vertical dotted line represents a significance threshold of $p_{adj} = 0.05$.

disease-associated SNPs significantly overlapping with known eQTLs in blood as compared to immune disease-associated SNPs in general? Rather than focusing directly on regulatory element region overlap, this analysis instead looks at the overlap of disease association signals with measures of gene expression at the SNP level, to see whether the input pool of SNPs is associated with changes in gene expression more often than a random set of immune-related disease-associated SNPs and, if so, in which cell type. Thus, MG-associated SNPs were analysed to see whether they were significantly over-represented among immune cell eQTLs compared to all known immune GWAS identified variants meeting a p-value cut-off of $5 \times 10^{-5}$ in GWAS Catalogue [65]. A collection of previously catalogued internal and collaborator eQTLs were used, conducted in a range of immune cells including monocytes and B-cells [88], neutrophils [89], T-cells [90], and unpublished NK cell data from 228 healthy individuals of European ancestry. An enrichment analysis was carried out by comparing the observed overlaps between eQTL and MG-associated variants (or in linkage disequilibrium; $r^2 > 0.8$ in Caucasian populations) against the expected overlaps, which were estimated by random sampling (without replacement) from all known immune disease variants (i.e. a hypergeometric distribution). Analysis of MG SNPs with regard to cell-specific data uncovered that disease-associated SNPs were significantly enriched among $CD4^+$ and $CD8^+$ T-cell eQTLs in EO ($CD4^+$ $p_{hyper} = 6.5 \times 10^{-9}$, $CD8^+$ $p_{hyper} = 8.1 \times 10^{-15}$), LO ($CD4^+$ $p_{hyper} = 9.8 \times 10^{-5}$, $CD8^+$ $p_{hyper} = 1.7 \times 10^{-7}$) and combined ($CD4^+$ $p_{hyper} = 1.2 \times 10^{-11}$, $CD8^+$ $p_{hyper} = 3.6 \times 10^{-20}$) groups, as shown in Fig. 4.2.3. Interestingly, there was no significant overlap observed in other immune cell types, despite the traditional implication of B-cells in autoimmunity, suggesting that T-cells play an out-sized role in MG pathogenesis as compared to other immune-related diseases, from a genetics point of view. This finding was useful as it suggests that downstream analysis should focus on T-cell processes for better understanding of disease mechanism and potential ties to genetic polymorphisms.

### 4.2.4 Autoimmune disease-associated genes were prioritised in MG

Following direct analysis of MG GWAS-identified SNPs via overlap with regulatory elements and eQTL studies, disease SNPs were analysed using Pi [64], in order to move from a SNP-centric mode of analysis to a gene and pathways-centric approach instead; analysis was conducted with the

complete pooled set of significant SNPs as well as done separately for EO and LO variants. Settings used for Pi analysis included LD $r^2$ >0.8 for European population and all genomic predictors.

### Gene-level prioritisation

First, all reported GWAS were considered, irrespective of timing of disease onset. The most highly prioritised genes for the combined set of SNPs are shown in Fig. 4.2.4 as well as summarised in Table 4.2.1; these included *TNIP1*, a gene previously identified in myasthenia GWAS, which was highly prioritised on the basis of being an nGene (proximal to input SNP(s)) as well as being a cGene (evidence of interaction between input SNPs and the gene promoter in immune cell types [87]). Additionally, *IKZF3* (prioritised as an nGene, eGene (eQTL signal), and cGene), *MDM2* (prioritised on the basis of network connectivity to other seed genes), and *PIN1* (similarly a network-prioritised non-seed gene), were highly ranked and all three are genes in which mutations or abnormal activation have been linked to autoimmune disease development and progression such as systemic lupus erythematosus [163–165]. Taken together, these results demonstrated two important points: Pi prioritisation was able to recover known disease-associated genes with strong GWAS support of association to MG and Pi analysis recovered genes that have previously not been implicated in MG but have strong links to autoimmunity in other disease contexts. This provides an opportunity for considering the targeting of genes that may be novel to a MG context but with support in other, related processes and immune mechanisms.

### Pathway analysis

Results were also analysed from a pathway perspective. Downstream analyses were carried out using the top 1% of the prioritised list of genes. Enrichment analyses was conducted to test whether these top 1% genes were over-represented among the member genes of any known biological pathways in the REACTOME [100] database (restricted to immune pathways) more often than expected by chance, against a background of the full gene priority list. There was significant enrichment for diverse signaling pathways involving interleukins as well as various toll-like receptors (Fig. 4.2.5). In order to

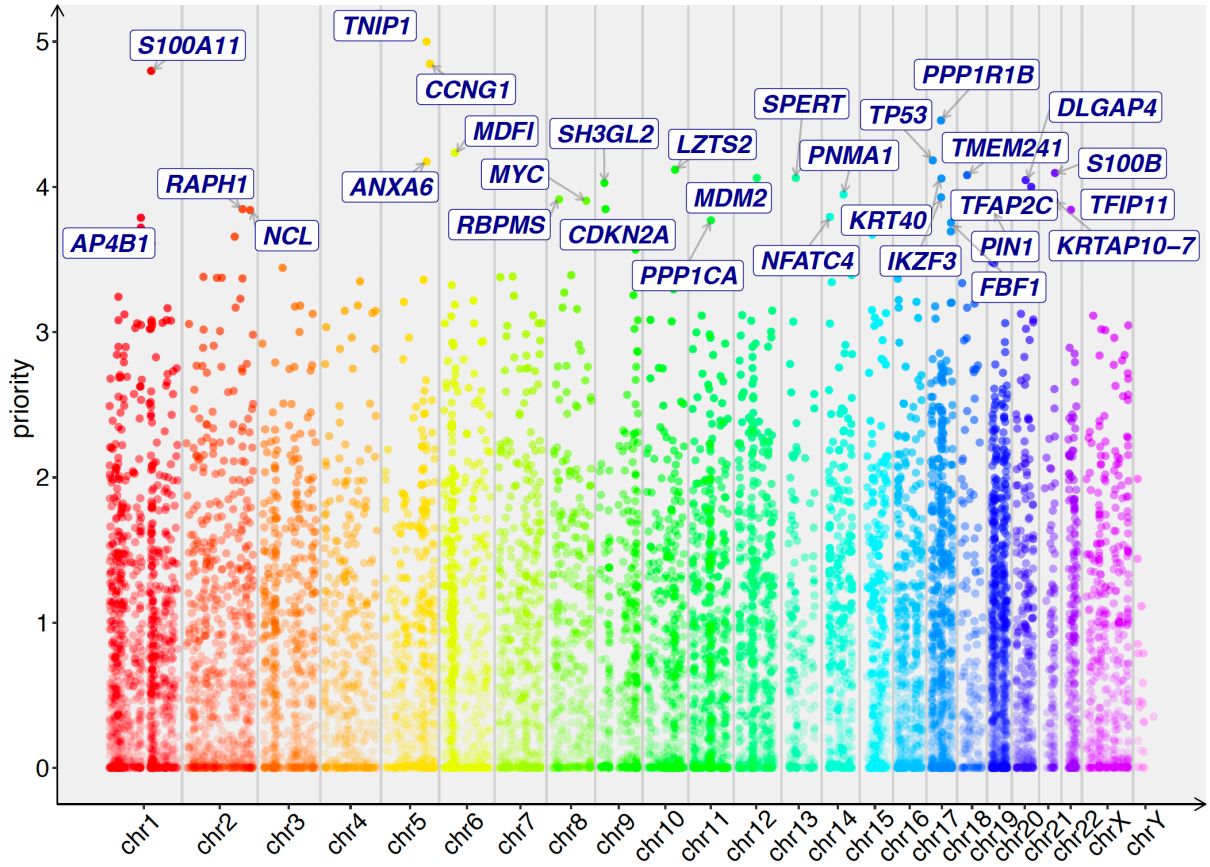



**Figure 4.2.4: Manhattan plot of MG prioritised genes.** A Manhattan plot with the 30 top prioritised genes from the Priority Index analysis labeled. Input was the combined set of myasthenia gravis GWAS SNPs, meeting a significance cutoff of $p < 5 \times 10^{-5}$. The y-axis shows the gene priority scores of 12177 genes, and the x-axis shows their chromosome position.

**Table 4.2.1: Top 30 Pi-prioritised MG genes.** A summary of the top 30 genes prioritised by Pi based on all MG GWAS SNPs meeting a disease association significance of $p < 5 \times 10^{-5}$. The Pi ranking is included for each gene, along with the number of nGene (proximity to SNP), eGene (eQTL signal), and cGene (physical SNP – gene promoter interaction) annotations captured by the input SNP list (collectively representing seed genes). Genes without any nGene, eGene, or cGene annotations (i.e. non-seed genes) were prioritised on the basis of strong network connectivity to seed genes, based on the STRING gene interaction map (limited to high confidence interactions).

| GENE | RANK | NGENE ANNOTATION | EGENE ANNOTATION | CGENE ANNOTATION |
|---|---|---|---|---|
| TNIP1 | 1 | 1 | 0 | 5 |
| CCNG1 | 2 | 0 | 0 | 10 |
| S100A11 | 3 | 0 | 0 | 11 |
| PPP1R1B | 4 | 0 | 0 | 11 |
| MDFI | 5 | 0 | 0 | 0 |
| TP53 | 6 | 0 | 0 | 0 |
| ANXA6 | 7 | 1 | 0 | 7 |
| LZTS2 | 8 | 0 | 0 | 0 |
| S100B | 9 | 0 | 0 | 0 |
| TMEM241 | 10 | 0 | 0 | 10 |
| SPERT | 11 | 0 | 0 | 0 |
| MDM2 | 12 | 0 | 0 | 0 |
| KRT40 | 13 | 0 | 0 | 0 |
| DLGAP4 | 14 | 0 | 0 | 0 |
| SH3GL2 | 15 | 0 | 0 | 0 |
| TFAP2C | 16 | 0 | 0 | 0 |
| PNMA1 | 17 | 0 | 0 | 0 |
| IKZF3 | 18 | 1 | 1 | 7 |
| RBPMS | 19 | 0 | 0 | 0 |
| MYC | 20 | 0 | 0 | 0 |
| KRTAP10-7 | 21 | 0 | 0 | 0 |
| CDKN2A | 22 | 0 | 0 | 0 |
| RAPH1 | 23 | 0 | 0 | 10 |
| TFIP11 | 24 | 0 | 0 | 0 |
| NCL | 25 | 0 | 0 | 0 |
| PIN1 | 26 | 0 | 0 | 0 |
| NFATC4 | 27 | 0 | 0 | 0 |
| AP4B1 | 28 | 0 | 0 | 10 |
| PPP1CA | 29 | 0 | 0 | 0 |
| FBF1 | 30 | 0 | 0 | 0 |

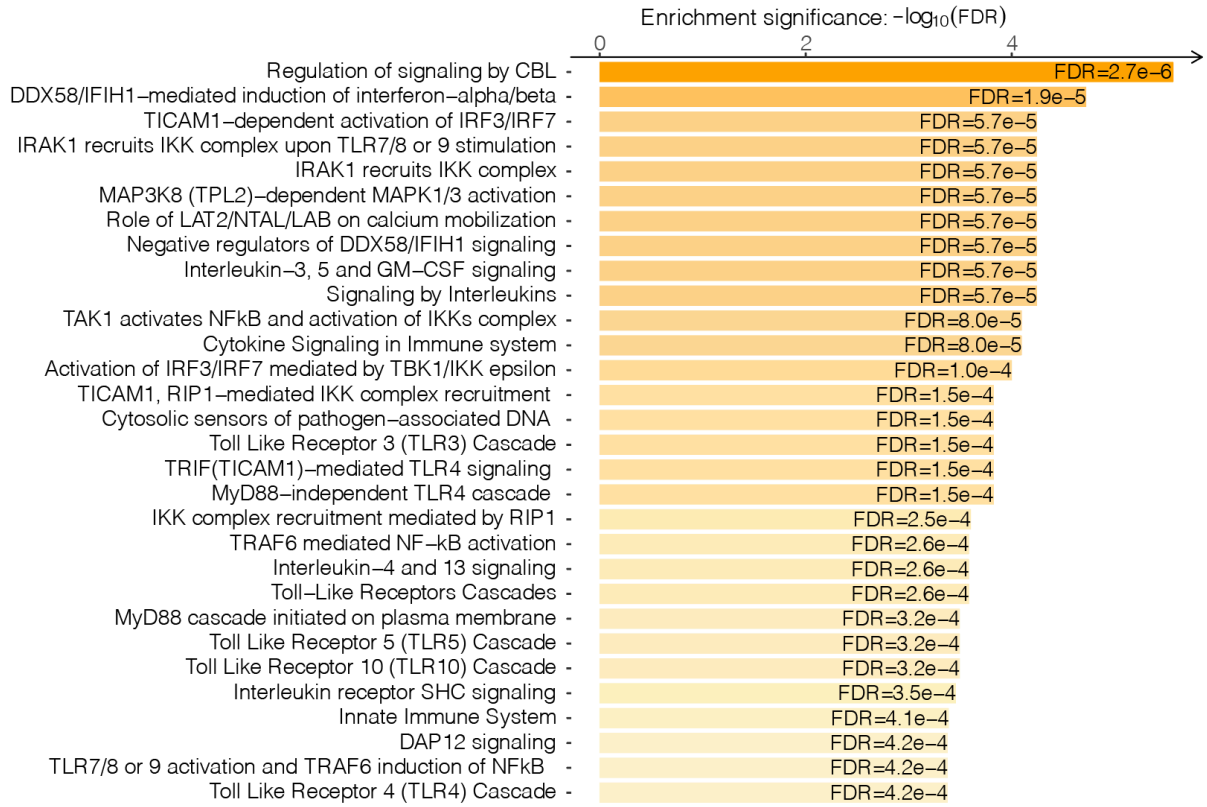



**Figure 4.2.5: Enriched pathways captured by Pi MG analysis.** The top 30 significantly enriched REACTOME pathways (FDR $< 0.05$) captured by top 1% genes in the Pi output of all myasthenia gravis variants (i.e. both early and late onset) meeting disease association significance $p < 5 \times 10^{-5}$.

interrogate this finding further and identify significantly enriched, disease-associated pathways that contained the maximal number of highly prioritised genes, a gene subnetwork cross-talk analysis was conducted [166] from a pathway-merged KEGG interaction map [167]. Essentially, this analysis heuristically solves a prize-collecting Steiner tree problem to answer the question: what gene interaction subnetwork (of specified size) can be identified based on the input data (i.e. containing as many highly prioritised genes as possible), allowing for some lower-rated 'linker genes' when required. Once identified, an enrichment test was conducted to identify which pathways were significantly over-represented in this gene sub-network, in order to uncover key disease-relevant pathways that are directly interacting with one another and which may serve as starting points for developing hypotheses around disease mechanism. This identified several genes of interest (Fig. 4.2.6), including *MAPK4* (p38a), a target investigated in the context of autoimmunity due to links between a loss of p38 signaling and enhanced T-regulatory cell induction [168].

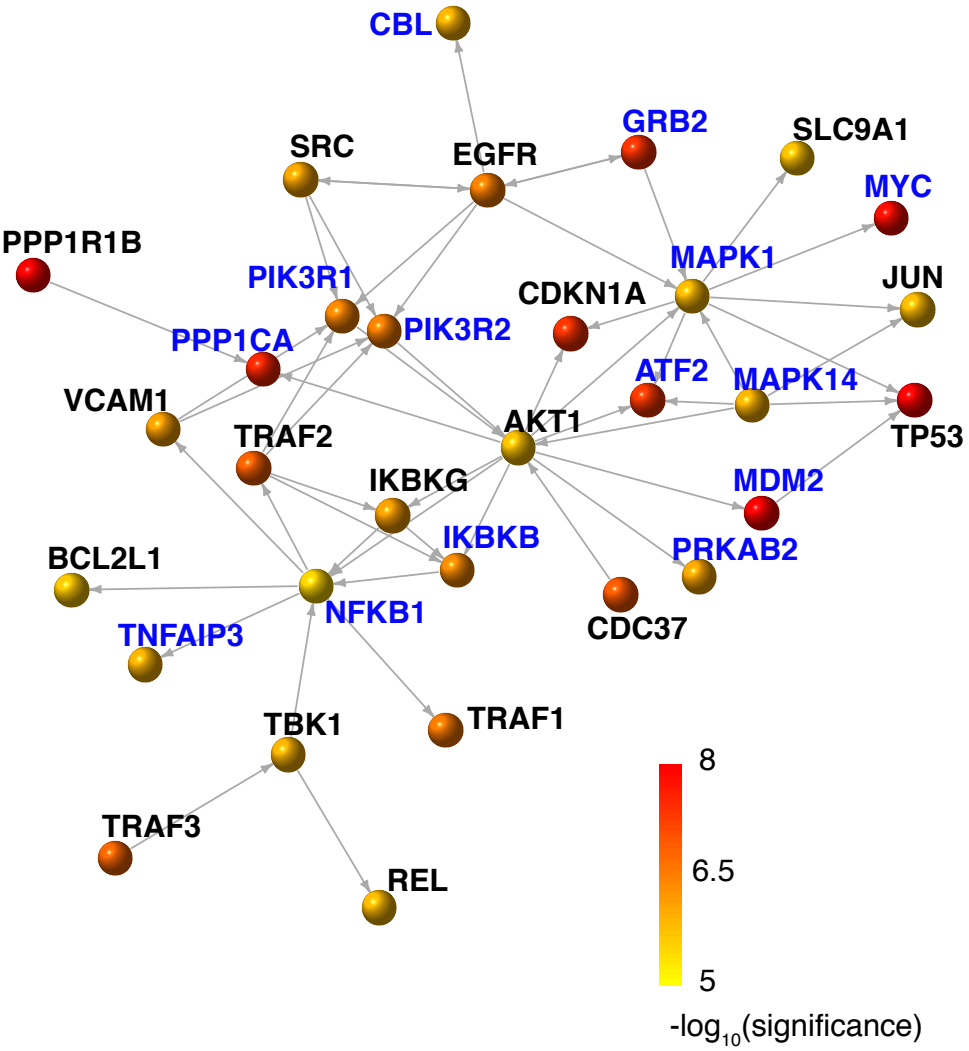



**Figure 4.2.6: Subnetwork pathway cross-talk analysis.** A subnetwork cross-talk analysis of KEGG-merged pathways, constructed using the combined myasthenia gravis prioritised gene list, and maximized for high-priority genes. The darker the gene node, the more highly prioritised the gene. Genes that are FOXP3-bound in activated T-regulatory cells are labelled in blue.

### 4.2.5 Comparison of prioritisation output to active versus remission-state MG

We hypothesised that disease relevant genes identified through our genomic approach would overlap with gene expression signatures of myasthenia gravis patients. Park *et al.* (2016) published a pilot blood transcriptomic analysis of patients with active disease (MG) versus those in remission, using a mixed group of early and late onset patients [169]; specifically, RNA was purified from isolated peripheral blood mononuclear cells, from five patients in remission and five with active disease (either *de novo* disease or refractory MG despite long-term immunosuppressive treatment), as defined by the Myasthenia Gravis Foundation of America Clinical Classification system [170]. This list of differentially expressed genes was contrasted with the combined Pi output. An enrichment analysis was done, which demonstrated a significant enrichment of differentially expressed MG disease genes in the top 1% of the prioritisation output ($p_{hyper} = 0.018$). Additionally, leading edge analysis (Fig. 4.2.7) recovered 7 out of 21 differentially expressed genes, at a significance of $p = 0.0095$. Moreover, 5 of these 7 genes in the leading edge (i.e. the majority of genes driving the enrichment signal near the top of the Pi output) were identified by the authors of the blood profiling study as likely candidates for molecular profiling of myasthenia disease activity; these included *S100P*, a protein mediating MAPK signaling with implications in cancer [171]; *GAB2*, which has been shown to be important in mediation of T-cell activation in knockout mice models [172]; *NFKBIA*, which contains polymorphisms that have been associated with autoimmunity and inflammation [173]; *TNFAIP3*, a gene implicated in lupus via a likely causal SNP within its enhancer element in a genome-edited mouse model [174] that was also found in the subnetwork cross-talk analysis; and *PPP1R15A*, an inhibitor of which has been shown to be of potential therapeutic benefit for the treatment of multiple sclerosis [175].

### 4.2.6 Contrasting prioritised genes for early onset and late onset MG

In addition to analysing the pooled set of MG GWAS data, separate Pi analyses were implemented for the subset of significant variants associated with early onset MG versus those associated with the late onset form of the disease. Gene prioritisation outputs were visualised with Manhattan plots for EO

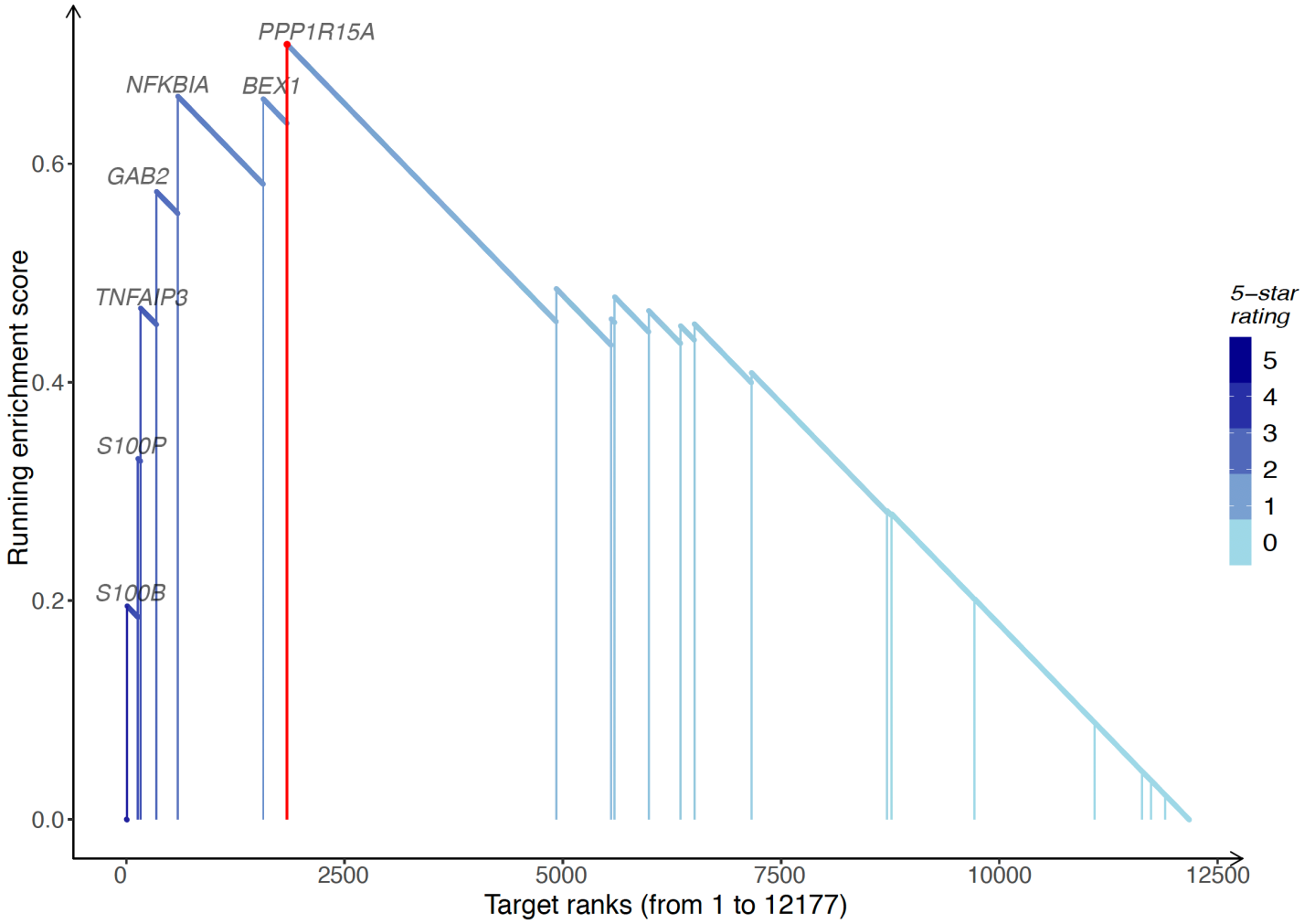



**Figure 4.2.7: Enrichment of differentially expressed (active vs remission-state) MG genes in the leading edge of the Pi output**. A leading edge analysis of the combined EO and LO MG SNPs Pi output, with respect to genes identified in a blood transcriptomic analysis of active vs remission-state MG. 7/21 of differentially expressed genes (fold-change >2, $p < 0.05$) were recovered in the leading edge, at a significance of $p = 0.0095$.

versus LO (Fig. 4.2.8). Presented is a summary of prioritised genes in Table 4.2.2 and a complete list of top 1% of genes in EO and LO can be found in Supplementary Tables 8.0.1 and 8.0.2, respectively. This analysis prioritised some previously identified MG GWAS genes like *TNIP1* [68]; additionally, *TRAF3* was identified, a gene previously associated with multiple sclerosis and allergic disease [176, 177] that was captured in the analysis via physical interactions between EO MG SNPs and the gene promoter (cGene), and that thus may serve as a putative target to investigate in a MG context. Similarly, *AGER* was indirectly prioritised in the EO MG output due to strength of gene network connectivity with genomic seed genes, and thus may be worth investigating with respect to a gene pathway-level understanding of disease mechanism. This gene was also identified in previous GWAS for inflammatory bowel disease [178]. On the other hand, LO prioritisation identified genes like *RAPH1*, which has been implicated in rheumatoid arthritis and type 1 diabetes on the basis of functional interaction with casual variants [87, 179], as well as *PRKDC*, a gene with mutations associated with impaired T-cell tolerance and autoimmunity [180]. Both of these targets thus form likely candidates for investigation of functional relevance in MG.

### 4.2.7 Pathway analyses highlight further differences between EO and LO disease

Recognising that there is significant heterogeneity when considering all myasthenia gravis patients, differences between early and late onset disease were further interrogated through analysis of underlying pathways. To uncover any differences, a pathway enrichment test was conducted using the top 1% of prioritised genes, with all myasthenia prioritised genes used as the testing background (Fig. 4.2.9). In line with what is known about the disease, B-cell receptor signaling was highly enriched in both forms (EO odds ratio = 3.01, FDR = $2.4 \times 10^{-2}$; LO odds ratio = 5.19, FDR = $1.6 \times 10^{-3}$) together with immunoglobulin Fc epsilon receptor signaling important in B-cell activation and proliferation (EO odds ratio = 2.57, FDR = $3.7 \times 10^{-2}$; LO odds ratio = 5.40, FDR = $8.40 \times 10^{-4}$), and interleukin signaling (notably Il-1, IL-4 and Il-13 signaling, some of which have previously been implicated in myasthenia pathogenesis on the basis of differential cytokine profiles between patients and controls [181]). This uncovered that exclusively EO genes were enriched in innate recognition

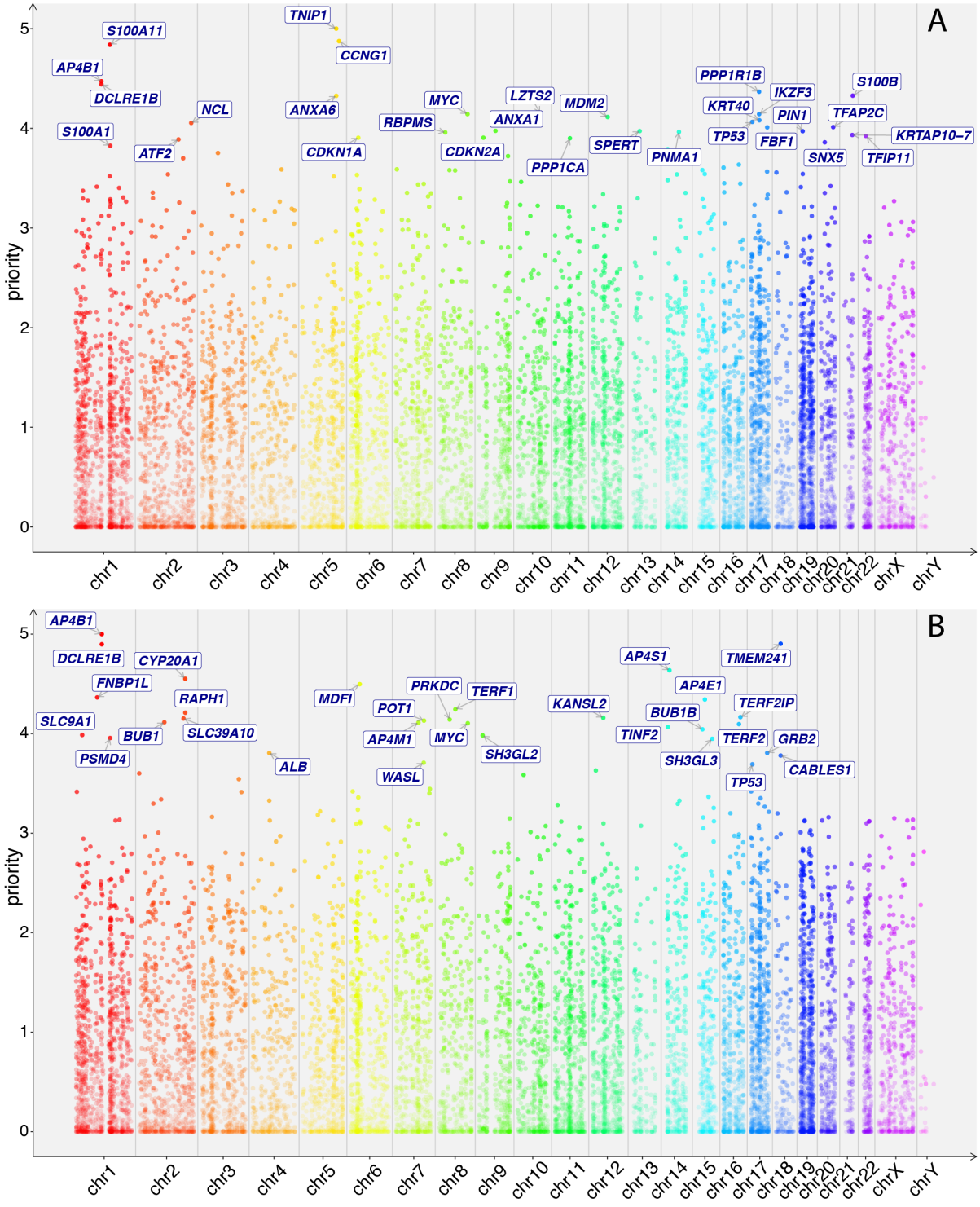



**Figure 4.2.8: Distribution of prioritised genes.** Manhattan plots of prioritised genes using **(A)** EO MG SNPs as input and **(B)** LO myasthenia gravis SNPs as input. The x-axis shows genomic location and the y-axis shows Pi priority score. The top 30 genes in each prioritisation output are labelled.

**Table 4.2.2: Gene prioritisation results summary.** A summary of selected genes from the top 1% of either the EO or LO prioritised gene list, describing how they were prioritised (nGene, eGene, cGene, or network evidence), whether they were previously identified in MG GWAS, and any other clinical implications, defined as whether genes have been implicated in other indications or disease GWAS (derived from GWAS Catalog).

| GENE | DESCRIPTION | ONSET | REPORTED IN MG GWAS | REASON FOR HIGH PRIORITISATION | OTHER CLINICAL IMPLICATIONS |
|---|---|---|---|---|---|
| **AGER** | Advanced Glycosylation End-Product Specific Receptor | EO | No | Network gene | inflammatory bowel disease (IBD) and ulcerative colitis, GWAS |
| **AKT2** | AKT Serine/Threonine Kinase 2 | LO | No | Network gene | Diastolic blood pressure GWAS |
| **ANXA1** | Annexin A1 | EO | No | Network gene | Parkinson's disease and schizophrenia GWAS; rheumatoid arthritis and leukemia |
| **GRB2** | Growth Factor Receptor Bound Protein 2 | EO | No | Network gene | Systemic lupus erythematosus (SLE) GWAS |
| **PIK3R1** | Phosphoinositide-3-Kinase Regulatory Subunit 1 | LO | No | Network gene | Schizophrenia and breast cancer GWAS; insulin resistance |
| **PRKDC** | Protein Kinase, DNA-Activated, Catalytic Polypeptide | Both | No | Network gene | Immunodeficiency 26 |
| **PSMB1** | Proteasome Subunit Beta 1 | LO | No | Network gene | Type 1 diabetes GWAS |
| **TNIP1** | TNFAIP3 Interacting Protein 1 | EO | Yes | nGene, cGene | SLE, psoriasis, systemic sclerosis, IBD, psoriatic arthritis GWAS |
| **TRAF2** | TNF Receptor Associated Factor 2 | EO | No | Network gene | Lymphoma |
| **TRAF3** | TNF Receptor Associated Factor 3 | EO | No | nGene, cGene | Multiple sclerosis and allergic disease GWAS |

pathways for viral and bacterial pathogens such as TRAF6-mediated NFkB activation (odds ratio = 14.6, FDR = 1.3 x $10^{-3}$), viral sensing (DDX58/IFIH1; odds ratio = 7.59, FDR = 1.4 x $^{-3}$), and TLR4 signaling (odds ratio = 4.59, FDR = 8.3 x $10^{-3}$), which is supported by earlier studies that have, for instance, suggested a pathogenic role of EBV in myasthenia gravis [182, 183]. By contrast, for late-onset disease we found that signaling pathways related to DAP12 signaling (odds ratio = 11.7, FDR = 8.4 x $10^{-4}$), T-cell (CD28 co-stimulation; odds ratio = 13.2, FDR = 2.9 x $10^{-4}$) and B-cell activation by antigens (odds ratio = 12.7, FDR = 8.4 x $10^{-4}$) as well as ubiquitination (CBL signal regulation; odds ratio = 24.1, FDR = 4.0 x $10^{-5}$) were enriched, which was interesting given previous ties between Cbl-b and CD40 in B-cell proliferation and evidence showing suppression of experimental myasthenia gravis in mice with CD40 ligand blockade [184, 185].

In order to isolate the genes of the highest significance enriched pathways, a subnetwork analysis was conducted separately for EO and LO outputs using the 30 top ranked genes in each instance (Fig. 4.2.10). For EO genes, this revealed a series of cross-talking KEGG pathways, including those relevant in EBV infection, TNF signalling, IL-17 signalling, and NF-kB signalling, corroborating the findings mentioned above. These four pathways intersect via *TRAF2, TRAF3, TNFAIP3, NFKB1, IKBKB,* and *IKBKG* genes. In contrast, subnetwork analysis of the LO output elucidated an overlap between T-cell receptor signalling, chronic myeloid leukemia, and MAPK signalling pathways via *GRB2, AKT2,* and *MAPK3*.

### 4.2.8 Interrogation of EO vs LO differences using independently-generated data-sets

In order to further investigate some of the onset-specific differences, additional enrichment analyses were performed using independently generated data sets. First, previous subnetwork cross-talk analysis (*Pathway analyses highlight further differences between EO and LO disease* section) uncovered a specific TLR4 innate – adaptive differentiation between EO and LO outputs along with a significant enrichment of NFkB pathways in EO disease. In order to further corroborate this finding, the prioritised gene lists were compared to regulators of TLR4 signaling, a key pathway involved in the activation of NFkB signalling and the innate immune response.

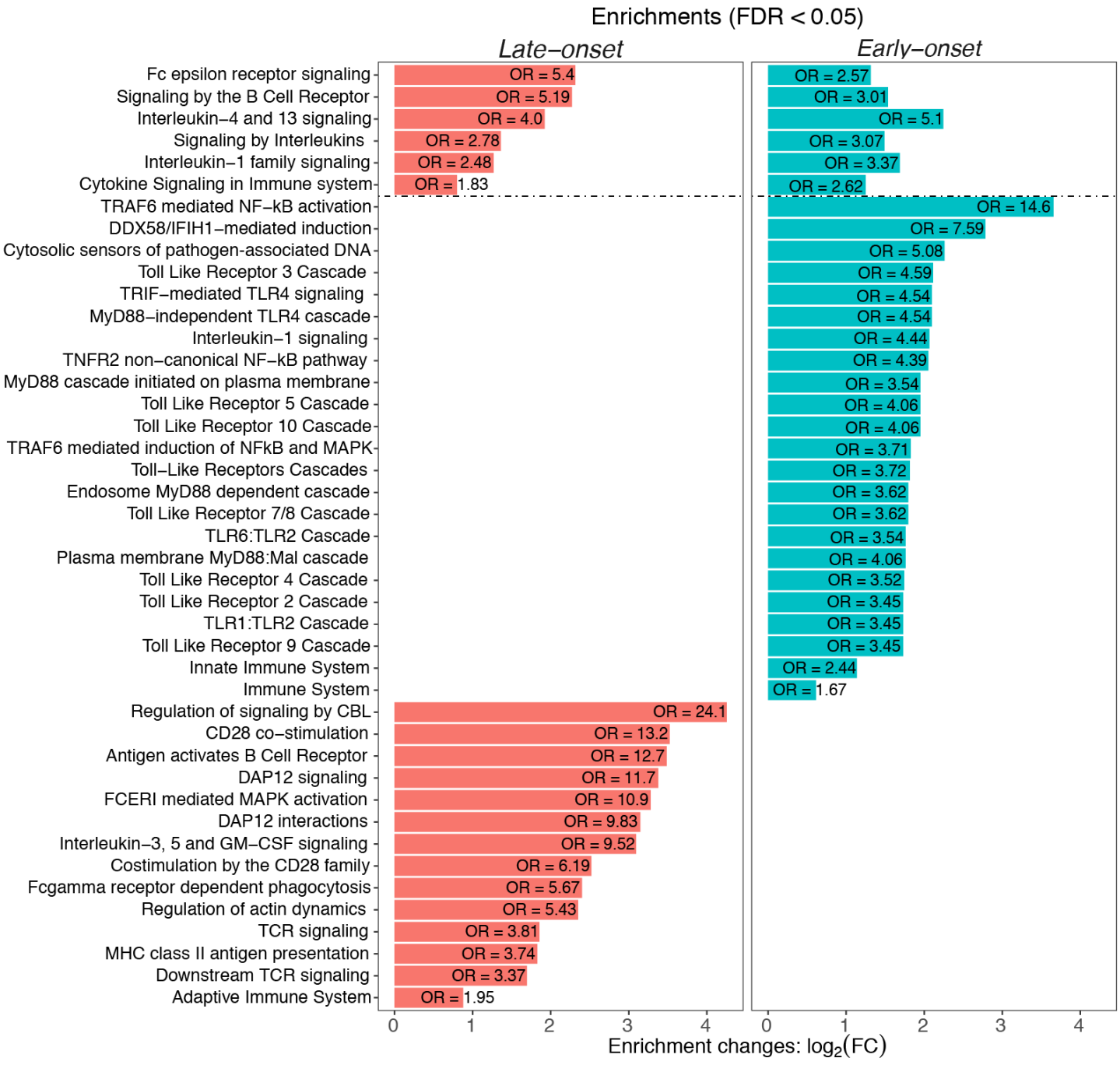



**Figure 4.2.9: Comparison of EO vs LO prioritisation.** A downstream analysis of the top 1% of prioritised genes from the EO and LO outputs; shared and unique EO vs LO REACTOME pathways enriched for in top genes; display is limited to the top 45 results that met a significance threshold of FDR $< 0.05$. Odds ratio (OR) listed for each pathway.

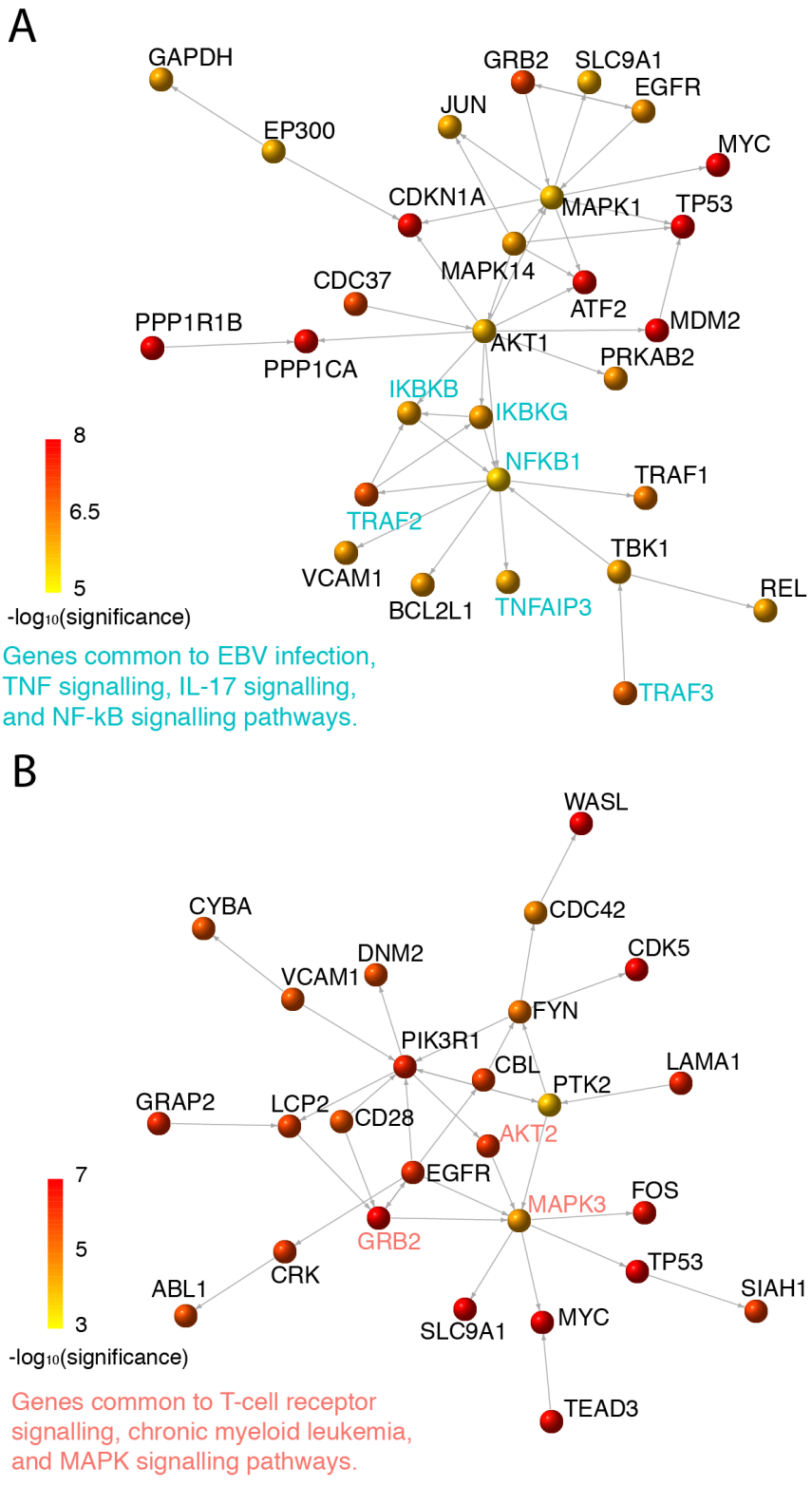



**Figure 4.2.10: Prioritised gene pathway cross-talk.** A pathway cross-talk subnetwork containing a maximal number of high priority genes from **(A)** the EO output and **(B)** the LO output. Gene nodes were limited to $< 30$ total, while significance refers to Priority Index gene significance score. Blue coloured genes show the nodes common to the most highly enriched pathways in the EO output, while red coloured genes show the nodes common to the most highly enriched pathways in the LO output.

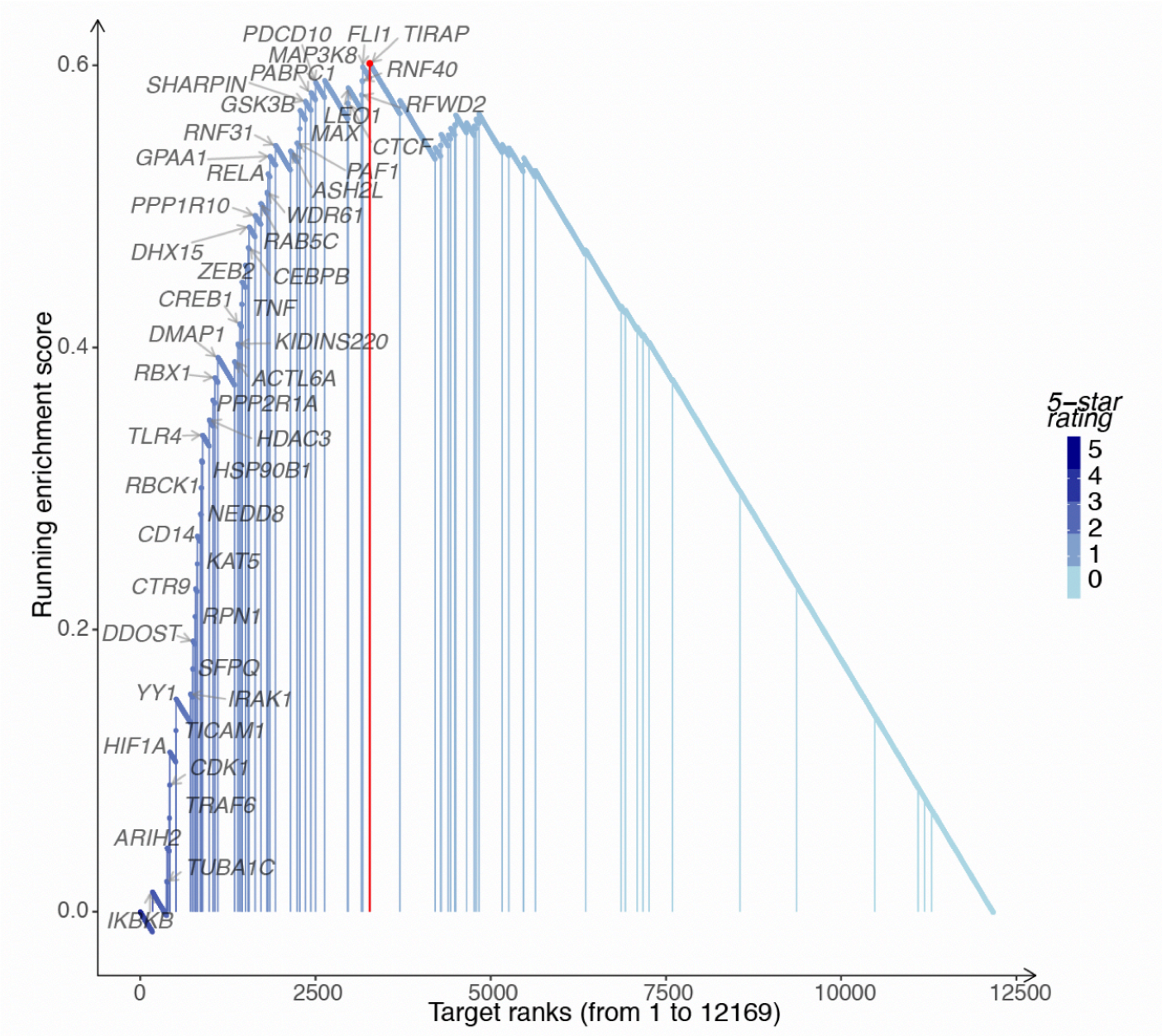



**Figure 4.2.11: Enrichment of key Tlr4 regulators in the leading edge of the EO MG output**. A leading edge analysis of the EO MG Pi output, with respect to genes identified in a genome-wide CRISPR screen of Tlr4 regulators during tumor necrosis factor activation. 51/81 of the Tlr4-regulatory genes were recovered in the leading edge, at a significance of $p = 0.002$.

Specifically, I investigated how the genes identified in EO vs LO disease captured known regulators of Tlr4-mediated TNF induction, identified in a mouse genome-wide CRISPR screen of Tlr4 immune regulators [186]; more specifically, this study used the screen in order to identify key genes in the induction of tumor necrosis factor by lipopolysaccharide, mediated by the Tlr4 pathway. The hypothesis was that the genome-wide analysis of EO SNPs would more comprehensively capture regulators of TLR4. I found that the top 10% of EO prioritised genes were indeed significantly enriched for these regulators (3.85-fold enrichment, $p = 6.4 \times 10^{-3}$), while LO genes were not ($p = 0.746$; Fig.4.2.11). Moreover, a leading edge analysis of Tlr4 regulators in the EO output recovered 51/81 targets ($p = 2.0 \times 10^{-3}$). This confirmed the onset-specific TLR4 signaling finding, and demonstrated that the original enrichment reported in the pathway-level analysis above was not simply driven by a few high-priority targets but rather by a large group of genes.

Second and similar to the above, I used an infection-related gene signature to test whether the EO output-specific EBV enrichment signal (and, more generally, infection-related signals) were supported by data from an existing, independent study of the Epstein-Barr Virus (EBV) infection transcription profile [187], which looked at EBV-induced gene transcription changes in healthy individual; EO genes were found to be significantly enriched for both infection-associated genes generally ($p = 0.019$) as well as EBV-unique genes more specifically ($p = 0.041$) while LO genes were not ($p = 0.117$ and $p = 0.274$, respectively; Fig.4.2.12). Additionally, there was a weak, positive correlation between EBV infection-induced expression fold-change and Pi ranking (Spearman correlation = 0.24, $p = 1.08 \times 10^{-5}$). Both of these findings reaffirmed the EO vs LO differences highlighted earlier, suggesting a difference in underlying disease mechanism and pathogenesis.

### 4.2.9 Statistical fine-mapping of the CTLA4 locus

Thus far, disease analysis was conducted genome-wide and in an open-ended fashion; downstream gene and pathway-level results were informed by the initial, unbiased GWAS input data, which included all disease-associated SNPs. In order to complement some of the insight drawn from these approaches, I next considered specific genetic associations in the context of gene expression, and how fine mapping and use of genomic annotators may enable mechanistic insights. Thus, emphasis was

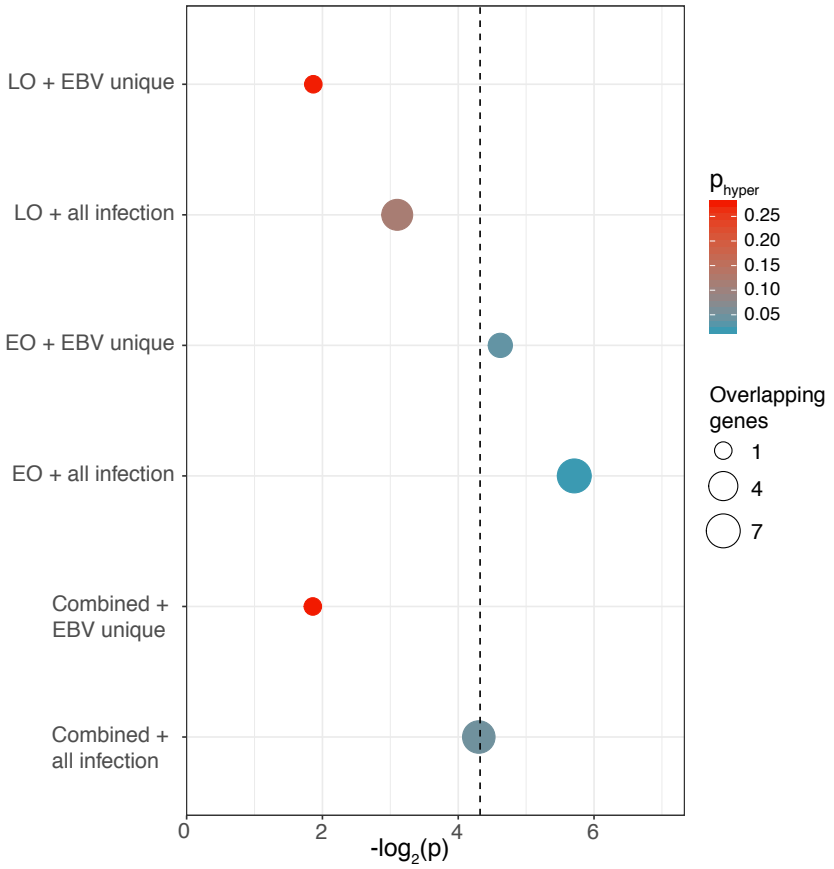



**Figure 4.2.12: EBV infection-related gene signature enrichment.** Enrichment analysis of infection-associated genes and Pi myasthenia gravis outputs (combined, early, and late onset). Enrichment was calculated using both the full infection gene list ('all infection') as well as the EBV-unique gene subset ('EBV unique'). Infection genes ($p < 0.05$) were sourced from the Dunmire *et al.* study (2014) [187]. The dotted vertical line represents a $p_{hyper}$ cutoff of 0.05.

placed on *CTLA4*, for three main reasons:

I. An early reported genetic association linking *CTLA4* to myasthenia gravis [69].

II. Implications of *CTLA4* methylation in myasthenia disease pathogenesis [188].

III. Evidence from recent use of check point inhibitors for cancer on myasthenia gravis treatment [189].

In order to begin dissecting the *CTLA4* signal, fine-mapping of the CTLA-4 locus was done using PAINTOR [110], a Bayesian framework that combines various functional annotations (such as whether variants fall within enhancers) with genetic association data. Given the crucial role of CTLA-4 in CD4+ T-regulatory cells,functional annotation scoring was limited to include specific CD4+ T-regulatory datasets. These included RoadMap assayed narrow peaks [107] in H3K27ac, H3K27me3, H3K4me1, and H3K4me3, as well as DNase I hypersensitivity sites (hotspots) in CD4+ cells [111]. SNPs for fine-mapping were sourced from the combined MG study [69] which looked at combined EO and LO myasthenia gravis cases. Any SNP listed in the supplementary materials that was in the CTLA-4 locus was included in the fine-mapping analysis, which took summary statistics as input. Two SNPs, rs231770 and rs231735, were identified with high posterior probability of causality (0.98 and 0.91, respectively) in the CTLA-4 locus (Fig.4.2.13; Supplementary Table 8.0.3).

Given this finding, the two SNPs were investigated for their role in influencing the expression of *CTLA4* in CD4+ T-cells using an eQTL analysis. Briefly, CD4+ cells were purified from peripheral blood mononuclear cells of 313 healthy European individuals, and genome-wide gene expression analysis, genotyping, and imputation (via the 1000 Genomes reference panel) was conducted. 6 million SNPs and expression from 38,839 probes in 23,704 genes were included in the analysis, following quality control and filtering. Cis-eQTL calculation was done by testing for association between SNPs and gene expression within 1Mb intervals [90]. For *CTLA4*, association between allele and gene expression was calculated using three probes; both rs231770 and rs231735 were indeed eQTLs influencing the expression of the *CTLA4* gene in CD4+ cells and the effect alleles matched that of the GWAS-identified variants and negatively correlated with gene expression. Colocalisation analysis [112] confirmed that there was a high statistical likelihood of shared signal

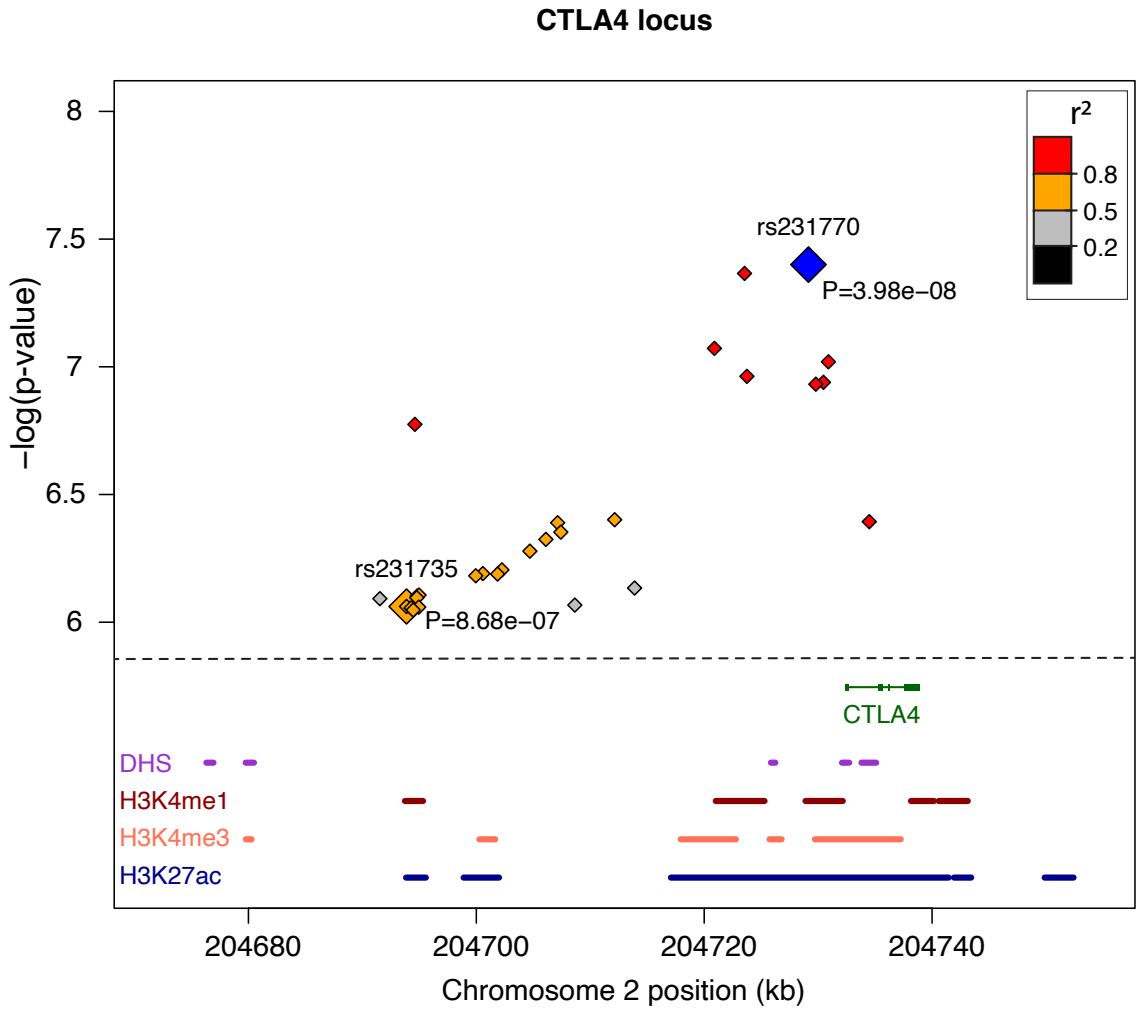



**Figure 4.2.13: CTLA-4 statistical fine-mapping.** A local association plot of significant myasthenia gravis GWAS-identified SNPs (combined onset), with chromosome position on the x-axis and GWAS significance on the y-axis. The two PAINTOR fine-mapped SNPs are labelled with their GWAS p-values and the lead SNP in blue. Coloured lines below the dotted line correspond to regions of overlap with epigenetic marks based on studies in CD4+ T-regulatory cells that were used during fine-mapping (the top 5 peaks in the region from each study are displayed).

between the eQTL and GWAS datasets (posterior probability = 0.925). In both cases, probe 4010767 was the most significant, with a rho of -0.66 ($p = 1.28 \times 10^{-38}$) and of -0.52 ($p = 7.01 \times 10^{-22}$) for rs231735 and rs231770, respectively, both meeting an FDR < 0.05 cutoff (Fig.4.2.14).

### Pi output is enriched for FOXP3-regulated genes in activated T-reg cells

The results of the colocalisation and fine-mapping analysis implicated two MG SNPs with regard to *CTLA4* expression; this functional link was supported by the fact that both SNPs, rs231770 and rs231735, were associated with decreased *CTLA4* expression in the CD4+ eQTL study. Given the well-established, multi-functional role of CTLA-4 in regulatory T-cells to help prevent autoimmunity [190] this seemed promising, as it suggested that MG-associated SNPs could be contributing to disease pathogenesis via CTLA-4. Therefore, an interesting hypothesis to test was whether there was evidence to support the role of known regulators of CTLA-4 in disease, based on the MG-specific prioritisation outputs outlined in earlier sections of this chapter. Specifically, interrogating Pi outputs with respect to FOXP3 seemed interesting, for a number of reasons: FOXP3 has a critical role in programming the development of regulatory T-cells, as has been demonstrated in Foxp3-mutant mice [191]; FOXP3 plays a role in up-regulating the expression of *CTLA4* and structure-guided mutations interfere with this [192]; and there has been previous evidence of FOXP3-related functional defects in myasthenia gravis patient T-regulatory cells [193].

One way to test whether a role for FOXP3 activity in MG is reflected in the Pi outputs is to check for the presence of related genes; in other words, were FOXP3-related genes prioritised highly in the Pi disease analysis? To test this hypothesis, data was sourced from a published ChIP-seq study that identified FOXP3-regulated genes in activated versus resting state regulatory T-cells [194]. This gene set was then contrasted to both the EO and the LO outputs, and the significance of overlap between the data-sets was measured (Fig. 4.2.15). Compellingly, there was significant overlap between top 1% Pi prioritised genes and differentially expressed genes, both in EO ($p_{hyper} = 0.032$) and LO ($p_{hyper} = 0.022$) outputs. Next, a more restrictive set of genes was used in the analysis: rather than testing for naive overlap between the data-sets, analysis was focused on genes that were driving the most highly enriched pathways in the combined Pi output, as captured by the subnetwork cross-talk analysis. This

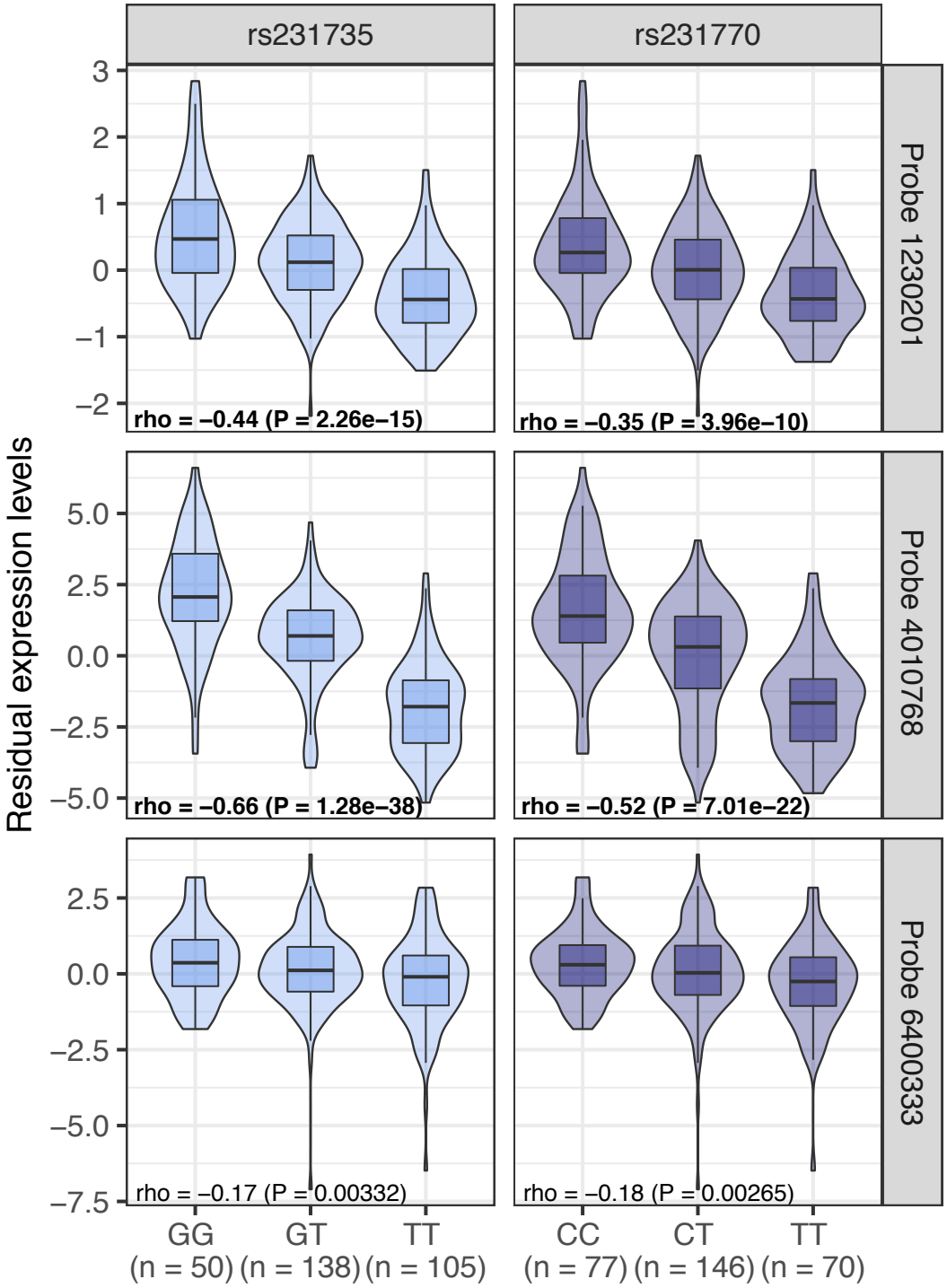



**Figure 4.2.14: *CTLA4* eQTL in CD4+ cells.** The allelic effect of the SNP (x-axis) on the expression levels (y-axis, residual gene expression levels) of different probes in the *CTLA4* gene in CD4+ cells are shown by boxplots within violin plots. Violin plot shows the density plot of the data on each side, the lower and upper border of the box correspond to the first and third quartiles, respectively, the central line depicts the median, and whiskers extends from the borders to ±1.5xIQR, where IQR stands for inter-quartile range, the distance between the first and third quartiles. Effect sizes and corresponding P-values are reported for the tissue-by-tissue analysis. Significant eQTLs in the tissue-by-tissue analysis at FDR 5% are highlighted in bold.

output was indeed also enriched for highly-prioritised, FOXP3-regulated genes (5.46-fold enrichment, $p = 4.771 \times 10^{-5}$; Fig. 4.2.6, genes labelled in blue), including *TNFAIP3*, one of the genes identified in the myasthenia blood transcriptomic study [169], and *IKBKB*, a key regulator of TNF induction identified in the CRISPR Tlr4 screen [186].

In addition to the link between FOXP3, which is a key component in peripheral immunity, and MG that the Pi analysis highlighted, it was interesting to consider potential evidence supporting central immunity. Autoimmune Regulator (AIRE) is one such candidate, as its role in central tolerance via regulation of expression and negative selection in the thymus has been extensively studied and well established [195]. Therefore, I investigated whether AIRE-regulated or associated genes were over-represented in the prioritisation outputs for MG. Specifically, an Aire-knockdown mouse model was used in conjunction with differential RNA-seq analysis to identify changes in gene expression associated with AIRE, in thymic epithelial cells [196]. These significantly differentially expressed genes, defined as fold-change > 2, FDR < 0.05 following AIRE knockdown, were mapped to human orthologs using data pulled from the Mouse Gene Informatics database [103] and then contrasted to the Pi outputs, in order to test for statistically significant overlap between groups. In contrast to overlap with FOXP3-regulated genes, there was no significant overlap between top 1% Pi genes and AIRE-associated genes ($p = 0.998$); this remained true even when onset-specific Pi outputs were considered ($p = 0.999$ and $p = 0.987$, EO and LO respectively). Further, there was no overlap between subnetwork cross-talk genes and AIRE-associated genes.

## 4.3 Discussion

The analyses conducted in this chapter demonstrated the utility of leveraging a computational tool like Pi to add significant nuance to the information contained within GWAS data. Myasthenia gravis provided a great case example for the methodology – it falls into the category of immune-related diseases for which Pi as a tool is optimised; it has poorly defined pathogenesis with no readily available cure; and it has preliminary GWAS data that has uncovered some early signals and generated initial hypotheses. Therefore, the use of Pi for MG GWAS analysis was an opportunity to integrate

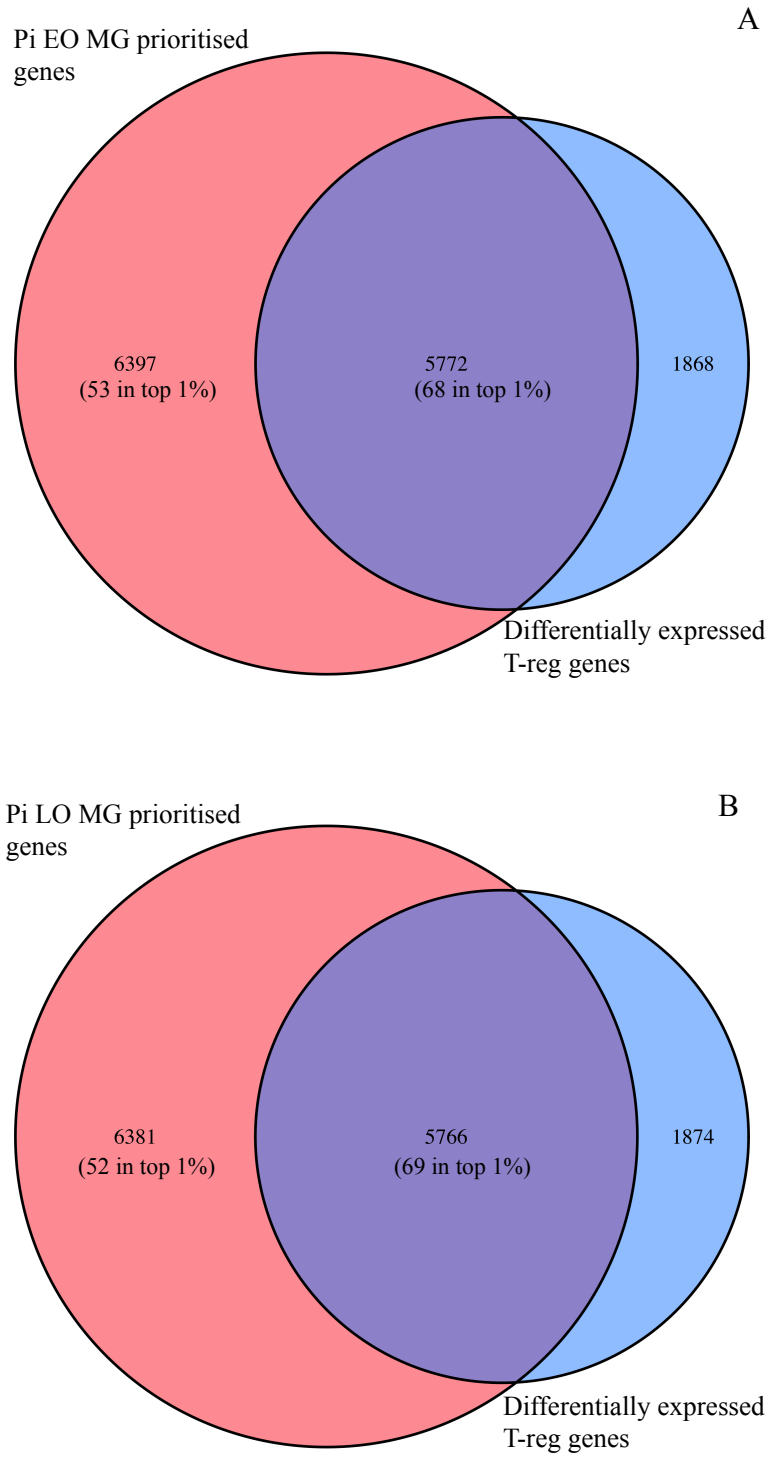



**Figure 4.2.15: Overlap with T-reg cell RNA-seq data.** A comparison of top 1 % targets and differentially expressed genes in activated $T_{reg}$ cells. Both **(A)** EO genes ($p_{hyper} = 0.03$) and **(B)** LO genes ($p_{hyper} = 0.02$) were significantly enriched for activated $T_{reg}$ cell genes.

various genomic predictors and deconvolute the noise surrounding this data in order to uncover true signal. Further, the outputs generated from this approach could be contrasted with existing, independently-generated, complementary data to investigate areas of overlap that may indicate where future work should be focused. Lastly, there was an opportunity to use statistical tools to focus in on particular genomic loci of interest, in order to create support for an opportunity for tangible, experimental follow-on to understand disease mechanism more closely.

### Analysis of MG-associated SNPs in the context of tissue relevance

Initially, MG-associated SNPs were analysed directly, before application of any genomic algorithms. The question behind this was simple: was there any information captured directly by the disease variants that could be used to set a baseline of expectations or at least generate certain hypotheses, and how does this match or not match what is known about the disease? The GARFIELD analysis was one way to set a baseline that matched prior knowledge: it identified blood cell types as disease relevant (based on a propensity for MG-associated SNPs to fall in the corresponding regulatory regions), as well as confirmed peaks in muscle and thymus related tissue, in line with what is generally known about the disease. It is important to once again note that these GARFIELD peaks dropped below significance threshold when adjusted for multiple-testing, but this very likely due to the conservative approach used by the algorithm in pruning input SNP lists and defining significance, as well as to the limited nature of MG GWAS data to begin with. Conversely, when MG-associated SNPs were analysed with respect to overlap with immune-related eQTLs, significant overlaps were robust to multiple-testing adjustment. This analysis highlighted an important piece of information: compared to other immune-related SNPs, MG-associated variants were enriched for T-cell eQTLs but not other immune cell-type eQTLs. This finding suggests an important functional role of T-cells in disease, as indicated by the underlying genetic basis of MG.

### Pi prioritisation and gene-level analysis

Following direct analysis of MG-associated SNPs, GWAS information was processed using the Pi pipeline. Two main modifications were applied to the standard Pi pipeline: the threshold for SNP

significance was lowered to $5 \times 10^{-5}$ in order to account for the relatively low number of SNPs identified by initial MG GWA studies, and only genomic predictors were considered in the gene prioritisation step (i.e. annotation predictors were omitted) in order to generate a naive, entirely genomics-based output. The result of the Pi prioritisation of all significant (pooled) MG SNPs was convincing for a few reasons. First, top prioritised genes included a mix of both known candidates (such as *TNFIP1*, with established ties to risk for MG [68]) and novel targets, including those implicated by gene expression, physical promoter interactions, and network effects. This simultaneously achieves two things: it serves as pseudo-validation since known knowledge about disease is captured, and it allows for reliable novel hypothesis generation based on the putative targets. In fact, when the priority output was compared to a blood transcriptomic study of MG patients [169], there was a statistically significant amount of overlap between the genes identified by those authors and the genes in the Pi output leading edge. Moreover, of the full set of genes profiled by Park and colleagues, a subset of 7 was proposed by the authors as likely candidates for a disease activity molecular signature, on the basis of differences between patients with highly active forms of disease and those in remission. Of those 7 genes, 5 were in the Pi leading edge, indicating a potential for the use of Pi in identifying likely candidates as disease markers, particularly when there is a need to filter a wide panel of putative targets to a more manageable sized list.

### Pathway and network-level analysis of Pi outputs

Further analysis of prioritised genes via network cross-talk presented another interesting opportunity for the utility of the computational approach outlined in this chapter: the re-purposing of existing drugs for a MG disease context, by taking advantage of genetic pleiotropy. For instance, *MAPK14* (p38a) was a central node in the pathway cross-talk analysis of MG; a previous study has shown that loss of p38 signaling results in enhanced $T_{reg}$ cell induction [168]. Therefore, inhibition of this target could be a viable way forward to achieve improved self-tolerance in a MG disease context. Further, p38 is a downstream target of a drug for the treatment of rheumatoid arthritis, with evidence linking the effect of p38 inhibition to reduced immune response specific to CD4+ cells [197], highlighting the potential for investigating this compound in a myasthenia setting. Thus, there is an opportunity to

systematically leverage Pi-generated disease outputs in order to highlight novel options for drug re-purposing. Specifically, this might be achieved through the following steps:

I. Creation of a directional drug – target data frame; this would include information linking all drugs to their targets (as presented in this thesis), as well as information on whether the interaction is stimulatory or inhibitory.

II. Creation of a drug – indication data frame; this would include information linking drugs to indications for which they have been approved in or in which there has been clinical traction (a more nuanced version of the approach described in the first results chapter of this thesis).

III. Generation of Pi outputs for a desired set of diseases.

IV. Systematic screening of every top priority set of genes from each of the outputs in the previous step, against the drug – target – indication database generated in steps i and ii.

This way, one could draw hypotheses such as: "**target X** is very highly prioritised in **disease Y**, and is already successfully being modulated by **drug Z** in a separate indication, with the desired direction of effect. Therefore, perhaps **drug Z** could be used to treat **disease Y**". Importantly, this type of systematic screen could be restricted to only the Pi outputs of those diseases for which the underlying internal genomic data (i.e. mainly immune-associated, blood cell-driven) is compelling, and can then be expanded to include additional indications as novel data-sets emerge and can be integrated into the Pi pipeline.

### Uncovering genetic differences between early and late onset forms of MG

In addition to investigating the pooled SNP Pi output, and recognising that there is significant heterogeneity when considering all myasthenia gravis patients, separate analyses were also conducted based on whether SNPs were identified in early onset or late onset patients. This approach was able to highlight logical similarities between both forms of disease (such as activity of B-cell receptor pathways, as expected in an autoimmune context, or enrichment of T-cell processes, confirming the importance of this cell type [160]), as well as some interesting differences. For instance, there was a distinct contrast with regard to toll-like receptor cascades, with EO prioritised genes enriched for a

wide array of these cascades while LO genes were not. This finding was additionally supported by a leading edge analysis using CRISPR-screened Tlr4 regulators [186], which showed that the majority of TLR4 regulators were captured in the EO leading edge to a high degree of statistical significance, but not in the LO one. Accordingly, modulation of a subset of toll-like receptors may be one way to focus therapeutic development in EO MG cases. On the other hand, LO prioritised genes were significantly enriched for CBL-driven signal regulation while EO genes were not. The role of CBL in tuning the autoimmune response as a negative regulator has been well documented and it has been investigated in a cancer immunomodulatory context [198, 199]; the results presented here indicate a potential for better understanding LO MG mechanism via a more in-depth analysis of CBL-signaling in patients.

The pathway cross-talk analysis identified additional differences between the two forms of disease, with one that particularly stood out: EO subnetworks were enriched for EBV-associated genes while LO subnetworks were not. Similarly, when the full set of top prioritised Pi genes for EO versus LO MG was compared to EBV-associated gene expression changes [187], there was an EO-specific enrichment of these EBV-affected genes in the Pi output. Interestingly, previous studies have observed a link between MG and viral infections and, specifically, Epstein-Barr virus. For example, EBV shows persistence and reactivation in MG patient thymuses, and may be a contributing factor to immune changes that drive or even initiate the disease [182, 183, 200]. Further, of the three MG-EBV studies conducted, two were done entirely with EO patients, while the third was done with a mixture of EO and LO. Therefore, it may be that this previously established link between MG and EBV is being driven by the EO form of disease, and it would be worth examining this further and perhaps focusing on the genes driving our EBV enrichment signal in greater detail.

### Uncovering a functional effect of MG-associated SNPs via the *CTLA4* gene

Another aim this chapter set out to achieve was to provide insight into potential disease mechanism, which could be used to inform future experimental design. In order to do so, a Bayesian statistical fine-mapping framework was implemented to interrogate a locus, CTLA-4, with prior implication in a MG context. PAINTOR [110] was carefully selected as the tool for this portion of the chapter, due to

its added functionality of integrating a standard Bayesian fine-mapping approach with functional genomic annotation data, akin to how Pi leverages multiple levels of genomic evidence in gene prioritisation. Given the highly context-dependant nature of CLTA-4 to regulatory T-cells, the PAINTOR approach allowed us to inform the fine-mapping using highly relevant, $T_{reg}$ data, such as RoadMap ChIP-seq peaks [107] and DHSs [111], in the specified cell type. Thus, the SNPs implicated by the model to be causal variants in the CTLA-4 locus (rs231770 and rs231725) are supported by functional genomic evidence in the cell type where CTLA-4 is most relevant, providing a compelling justification for causality. The fact that we were able to also associate the two SNPs to CTLA-4 expression via eQTL analysis suggests that the mechanism by which they likely exert their effect (and by which MG pathogenesis might be driven) is through modulation of CTLA-4 expression, a gene which previous studies have implicated in this disease via complementary approaches [69, 188, 189]. Ideally, these specific variants would be followed-up on in greater detail in larger cohorts of MG patients and from a T-cell phenotypic point of view, as well as in other autoimmune contexts that have implicated abnormalities with regard to *CTLA4*, such as systemic lupus erythematosus [201].

### Relevance of FOXP3 to disease

The last major piece addressed in this chapter was the connection between the MG analysis and FOXP3 regulation. This was relevant for two main reasons: it was important to consider given the causal link implicating MG-associated SNPs and CTLA-4 expression, due to the intertwined nature of CLTA-4 and FOXP3 in regulatory T-cells [191, 192]; and it was one method by which to interrogate the central vs peripheral tolerance question as it relates to MG. For the former, I looked for and established significant overlap between top Pi-prioritised MG genes and FOXP3-regulated genes in activated $T_{reg}$ cells [194], indicating that perhaps this underlies MG pathogenesis; this FOXP3 link to disease is supported by previous studies [193]. As for the latter of the two reasons, establishing FOXP3 relevance to MG was interesting since, on the other hand, we were not able to show evidence to support AIRE relevance as reflected in the Pi output. It is important to note the limitations of the AIRE analysis; namely, it was based on an analysis of gene expression in an Aire mouse knockdown

model, which brings with it a number of complications (e.g. non-human model, complexity around mouse – human gene mapping, etc.). Nevertheless, it was an interesting finding given the complexity around the role of the thymus in MG; while randomised trials of thymectomy have been shown to be beneficial in improving clinical outcomes of patients, not all patients subgroups respond to this approach [160, 161]. The results from this chapter suggest that peripheral tolerance may be a key driver of disease, which may be a factor in parallel with the traditional implications of the thymus and central tolerance in disease.

### Limitations

It is important to highlight the limitations underpinning the analyses conducted throughout this chapter. As mentioned earlier, perhaps the most obvious is the limited nature of MG GWAS; with only three early studies conducted, each with only roughly 500-1000 MG patients, the ability to detect disease-association remains limited. While lowering the threshold for significance for the purposes of Pi analysis was one way to combat this, it would be ideal to conduct a larger scale MG GWA study in order to confirm the various findings and potentially uncover new hits. Similarly, the PAINTOR analysis was based on the CTLA-4 locus, which only included 28 SNPs based on the data available in the main and supplementary figures of the study from which they were sourced, since they were limited to SNPs meeting the authors' threshold [69]. Therefore, it would be advisable to conduct statistical fine-mapping using the raw, unfiltered data, which may affect the posterior probabilities presented here. Further, in the eQTL analysis of CTLA-4 (post-fine-mapping), data presented was for general (bulk) CD4+ cells, rather than specifically for regulatory CD4+ T-cells. While eQTL data exists in more specific cell types, it is still fairly limited in sample numbers and thus under-powered. For instance, the DICE database is an exciting resource compiled from an extensive eQTL study across numerous immune cell types, including specifically regulatory T-cells [202]. However, the study was conducted with <100 individuals, and so is prone to potential failure to detect associations. In fact, neither of the two SNPs discussed in this chapter were present in the DICE database, in any cell type, despite being highly significant eQTLs in T-cells based on a previous study [90], which had >300 donors involved. Therefore, one would ideally repeat the eQTL analysis using CD4+ $T_{reg}$ cells in

at least double the number of donors as done by Schmiedel and colleagues. Lastly, it would be very compelling to take some of these findings into a model system, in order to interrogate them in greater detail. Experimental autoimmune mouse models of MG exist [203], and would provide a logical starting point for testing some of the specific targets outlined in this chapter. With the recent advent of targeted CRISPR editing in mice [204, 205], it would be possible to test the effects of altering specific genes or binding sites on the disease model, particularly in instances where more direct parallels can be drawn between human and mouse.

## 4.4 Conclusion

In conclusion, this chapter highlighted a wide array of insights that can be drawn from computational, genomics-based analysis of GWAS findings in the context of MG. In addition to prioritising a number of genes and pathways based on these approaches, the work highlighted a number of differences between early and late-onset forms of MG, presenting an opportunity to better characterise patients and develop more targeted therapies. The outputs derived from Pi were supported by evidence from complementary and independent data-sets, thus building the case for a number of targets to be explored in further detail. Lastly, the findings provided support for T-cell causal events driving MG pathogenesis, and in particular highlighted the importance of T-regulatory cells and a role for FOXP3. Ideally, the work presented here could be used to inform future follow-on studies, particularly those that aim to translate early experimental work into patient-derived or mouse model systems.

*Luck is what happens when preparation meets opportunity.*

Seneca

# 5

# Genomic validation of prioritisation outputs in the context of Dupuytren's disease

## 5.1 Introduction

Thus far, the focus of the thesis has been on how one would approach computational prioritisation of genes. Specifically, the first results chapter demonstrated a computational pipeline, Priority Index, that aggregates multiple layers of genomic data-sets in order to link disease-associated SNPs to functionally-relevant genes. The idea is that, by prioritising genes based on a combination of direct genomic evidence and indirect network effects, one can better predict potential targets that would be of therapeutic benefit, narrowing down the list of pre-clinical targets that go through the clinical development pipeline, and increasing the likelihood of success for those that do. Moreover, the chapter demonstrated how such a pipeline can be assessed computationally, without the need of significant time and resources to reach disease-specific experimental validation at scale. The use of known drug golden standards and simulated negatives provided an input into a framework that

quantified the performance of Pi, showcasing how a purely genetics-led approach could successfully prioritise targets of known therapeutic benefit over targets that can be classified as negatives.

Next, the second results chapter demonstrated how such a pipeline could be applied to a novel disease context, myasthenia gravis, which has early GWAS data available but lacks nuanced knowledge of disease pathogenesis or viable therapeutic targets. Through the use of Pi, I showed how myasthenia gravis GWAS data could be harnessed to yield a gene priority list. Further, I validated the gene output by comparing it to a variety of disease-relevant experimental data sets, building a case for how the prioritised candidates shed light on disease mechanism. Lastly, I demonstrated how the convergence of functional genomics, network interactions, and statistical mapping approaches could identify potential therapeutic targets and disease-relevant cell types, paving the way for moving into pre-clinical experimental validation.

Accordingly, this results chapter aims to demonstrate how the approach outlined above can be taken a critical step further, as computationally prioritised genes are taken into an experimental setting where they can be modulated in patient cells. This chapter focuses on fibrosis which has been the focus of a number of research initiatives, largely due to the morbidity and mortality associated with fibrosis of various visceral organs [206]. Dupuytren's disease (DD) is one such fibro-proliferative disorder and is characterized by fibrotic nodule formation in the hand, which causes fingers to permanently contract. DD is common in patients of northern European descent, affecting >30% of Scandinavians over the age of 60 and >4% in the U.K. [207]. Currently, treatments of DD include surgery (for the removal of the nodule) or collagenase injection (for the breakdown of the collagen-rich nodule cords) which is of limited efficacy as approximately 80% of patients have recurring disease within a few years of receiving injection [208]. Therefore, there is a pressing need to develop more effective therapeutics that will both treat existing DD nodules as well as prevent future ones from forming.

The genetics of DD have been previously studied and a number of loci implicated in disease risk [71, 209]. A study of 30,330 Danish twin pairs established a strong case for a genetic component to the disease, given their findings of significantly higher concordance rates and odds ratios amongst monozygotic twins as compared to dizygotic twins [209]. Further, a recent genome-wide study across three populations (5408 British cases, 2195 Danish cases, and 768 German cases) uncovered 26 SNPs

significantly associated ($p < 5 \times 10^{-5}$) with Dupuytren's disease, including a number of variants near relevant WNT-signaling genes like *WNT2*, *WNT4*, and *WNT7B* ligands as well as WNT antagonist *SFRP4* [71]. Additionally, various environmental factors have been proposed, including smoking, alcohol consumption, BMI, and others [210, 211]. Moreover, epigenetic dysregulation has been shown to influence the onset of cardiac fibrosis, liver fibrosis, and systemic sclerosis, all of which are related to DD [212–215].

With regards to DD specifically, it has been established that the myofibroblast cell is responsible for the excessive production of collagens, proteoglycans, and other extracellular matrix proteins that contribute to nodule formation [216, 217]. Unlike in fibrotic disorders of the heart or kidneys, tissue from DD patients is more easily accessible, which provides a unique opportunity for the study of novel therapeutic targets on patient-derived tissue, and which has already contributed to the understanding of some of the immune cell composition of the nodules and surrounding areas [114, 218, 219]. Accordingly, methods exist to extract Dupuytren's nodular tissue from patients and then dissect the tissue so as to isolate the disease-driving myofibroblasts; thus, it becomes possible to study therapeutic interventions in a cell system that is patient-derived and highly disease-relevant [114]. Further, the Structural Genomics Consortium (Oxford) has developed a focused library of small-molecule inhibitors against numerous epigenetic targets, which presents an opportunity to leverage these tools for the validation of specific targets, within the context of patient-derived myofibroblasts and their role in disease pathogenesis.

### Aims

Given the above, DD presented an interesting opportunity for using a genetics-led approach to prioritise novel therapeutic targets, which could then be validated using patient-derived tissues and highly specific small molecule inhibitor probe sets. Specifically, the aim of this chapter are as follows:

I. Demonstrate how Priority Index can be applied to recently published Dupuytren's disease GWAS to prioritise genes;

II. Experimentally test the hypothesis that the prioritised genes are likely to be disease-relevant

based on application of probes to disease-context tissue and downstream analysis of effects of these probes on phenotype;

III. Develop a hypothesis on potential targets for further therapeutic evaluation based on aims I and II and aggregation of results contained within, as they pertain to involvement in disease progression.

## 5.2 Results

### 5.2.1 Pi prioritisation of disease-relevant genes

Similar to previous results chapters looking at myasthenia gravis and multiple sclerosis, the initial result to interrogate was the Pi prioritisation output based on disease input SNPs (in this instance, DD). The top results from the prioritisation output (Table 5.2.1) highlight a number of different genes which include *CFTR*, mutations in which have long been established to associate with cystic fibrosis [220]; *GTSE1*, which has been implicated in modulation of lung fibrosis in mouse models [221]; *TAF1*, which has been identified as a potential biomarker for the progression of liver fibrosis [222]; and *MYC*, whose activation via IL-1 beta has been implicated as a necessary switch in the progression of fibrotic kidney disease [223].

Additionally, top 1% genes were analysed from a pathway perspective, in order to have a higher-level understanding of what the functional effects or interactions of these high-priority genes may be. Specifically, overlap between top 1% genes and REACTOME immune pathways [100] was analysed for evidence of statistically-significant enrichment. A number of disease-relevant pathways were identified through this analysis. For instance, C-type lectin receptor activity was enriched for (odd ratio = 4.93, $p_{adj}$ = 0.0027), which is supported by previous findings that have implicated the role of C-type lectin producing Langerhans cells in Dupuytren's contractures [224, 225]; regulation of actin dynamics (odds ratio = 9.24, $p_{adj} = 6.2 \times 10^{-4}$), supported by previous work that has demonstrated the specific role of a-smooth muscle actin regulation in Dupuytren's cells [226]; a number of toll-like receptor signaling cascades, consistent with findings that have investigated differentiation of primary

**Table 5.2.1: Top 30 Pi-prioritised DD genes.** A summary of the top 30 genes prioritised by Pi based on all DD GWAS SNPs in GWAS catalog, meeting a disease association significance of $p < 5 \times 10^{-8}$. The Pi ranking is included for each gene, along with the number of nGene (proximity to SNP), eGene (eQTL signal), and cGene (physical SNP – gene promoter interaction) annotations captured by the input SNP list (collectively representing seed genes), as well as each of the annotation predictors (disease ontology, phenotype ontology, and immune function ontology). Genes without any predictor annotations (i.e. non-seed genes) were prioritised on the basis of strong network connectivity to seed genes, based on the STRING gene interaction map (limited to high confidence interactions).

| GENE | RANK | DGENE | PGENE | FGENE | NGENE | EGENE | CGENE |
|---|---|---|---|---|---|---|---|
| GABARAPL2 | 1 | 0 | 0 | 0 | 1 | 0 | 7 |
| GTSE1 | 2 | 0 | 0 | 0 | 1 | 0 | 8 |
| MYC | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| NEDD4 | 4 | 0 | 0 | 0 | 1 | 3 | 0 |
| CFTR | 5 | 1 | 1 | 0 | 1 | 0 | 0 |
| UBC | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| TAF1 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| NTRK1 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| POLR2A | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| JUN | 10 | 0 | 0 | 0 | 0 | 0 | 0 |
| CALU | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| APP | 12 | 0 | 0 | 0 | 0 | 0 | 0 |
| EMC2 | 13 | 0 | 0 | 0 | 1 | 0 | 7 |
| E2F4 | 14 | 0 | 0 | 0 | 0 | 0 | 0 |
| ATF2 | 15 | 0 | 0 | 0 | 0 | 0 | 0 |
| RBL2 | 16 | 0 | 0 | 0 | 0 | 0 | 0 |
| ESR1 | 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| HIST3H3 | 18 | 0 | 0 | 0 | 0 | 0 | 0 |
| SP1 | 19 | 0 | 0 | 0 | 0 | 0 | 0 |
| HNF4A | 20 | 0 | 0 | 0 | 0 | 0 | 0 |
| E2F1 | 21 | 0 | 0 | 0 | 0 | 0 | 0 |
| YBX1 | 22 | 0 | 0 | 0 | 0 | 0 | 0 |
| LNX1 | 23 | 0 | 0 | 0 | 0 | 0 | 0 |
| PRPF40A | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
| TP53 | 25 | 0 | 0 | 0 | 0 | 0 | 0 |
| XPO1 | 26 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRPK2 | 27 | 0 | 0 | 0 | 0 | 0 | 0 |
| GABARAPL1 | 28 | 0 | 0 | 0 | 0 | 0 | 0 |
| NXF1 | 29 | 0 | 0 | 0 | 0 | 0 | 0 |
| SQSTM1 | 30 | 0 | 0 | 0 | 0 | 0 | 0 |

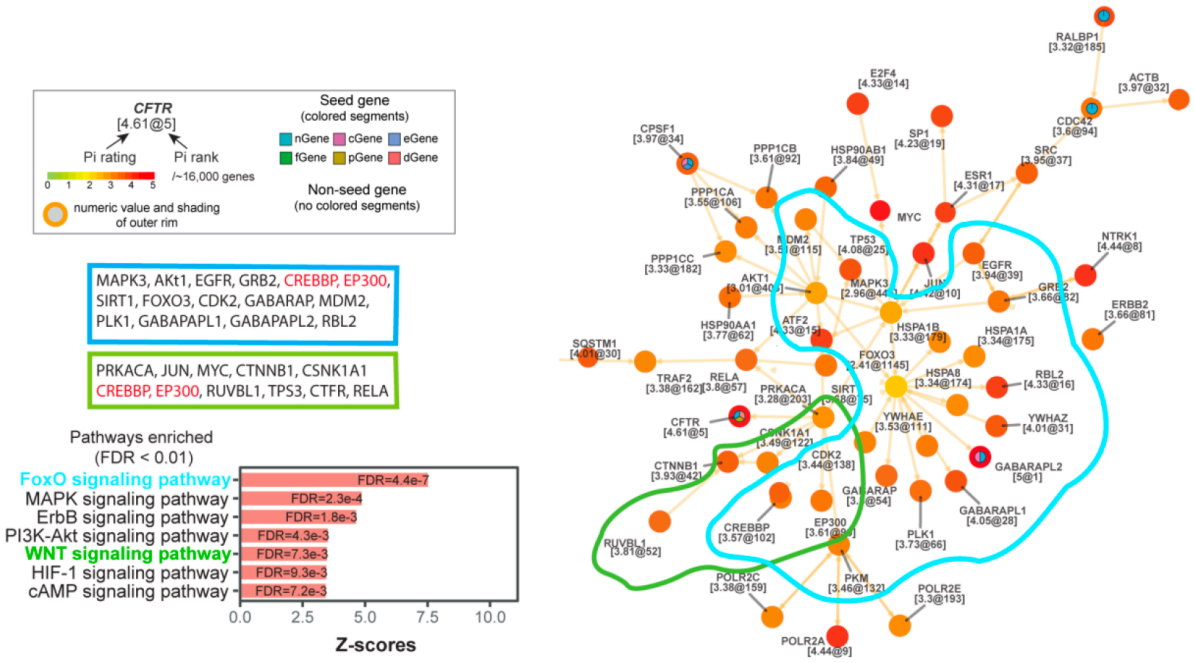



**Figure 5.2.1: Cross-talk analysis of top Pi prioritised genes.** Results of the subnetwork cross-talk analysis using Pi-prioritised genes and limited to 50 nodes. Top enriched pathways are shown in the bar plot; gene members driving enrichment for each pathways are shown in the blue and green boxes, and are circled accordingly in the network plot. The intersect of these significant pathways is in the nodes CREBBP and EP300.

cells into myofibroblasts, derived from DD patients [227]; and, more broadly, various immune signalling pathways, consistent with increasing evidence of immune signalling and system relevance in the disease [228].

Lastly, the Pi prioritisation output was analysed using the subnetwork cross-talk approach mentioned in earlier sections of this thesis. Effectively, this analysis served as a way of answering the question of which directly-interacting genes are driving the most significant enrichment of disease-relevant pathways, and at which nodes do these pathways intersect? Fig. 5.2.1 shows the results of this analysis and captures several main pathways. The most significantly enriched pathway was FoxO signaling (FDR = $4.4 \times 10^{-7}$), while WNT signaling, aberrations in which have been implicated as key factors in fibromatosis in DD [229], was also one of the top enriched pathways (FDR = $7.3 \times 10^{-3}$). These two pathways converged on two nodes, *CREBBP* and *EP300*, which are two established transcription regulators, highlighting a potential pair of targets for further analysis on the basis of the subnetwork cross-talk identified here.

### 5.2.2 RNA-seq analysis of CREBBP/EP300 inhibitor-treated cells

Given the findings around *CREBBP* and *EP300* summarised in the previous section, and the opportunity to leverage high quality SGC-developed probes, we were interested in investigating the potential disease-modifying effects of inhibiting targets, specific to DD. In particular, the SGC probe CBP30 is an inhibitor of both targets of interest, thus providing a tractable way forward. Specifically, SGC-CBP30 was applied to patient-derived myofibroblasts, and an RNA-seq transcriptomic analysis was completed in order to measure the effects of these inhibitors on gene expression, in 6 donor samples. Fig. 5.2.2 summarizes the results of this analysis at the gene-level. 660 genes were identified as differentially expressed (fold-change >1.5, FDR < 0.05) in the probe-treated myofibroblasts as compared to the DMSO-treated cells; of these, 192 were significantly up-regulated while 468 were significantly down-regulated. Differential expression identified a number of interesting genes that are supported by existing evidence. For instance, both *BMP6* and *BMPR* meet fold change and FDR criteria, and have been implicated as regulators of fibrotic processes [230, 231]. Similarly, *PRG4*, *WISP1*, and *ADAMTS8* have independently been studied for the roles in fibroblast proliferation and phenotypic association with fibrotic diseases, and also meet the significance cutoff from both a fold-change and an FDR standpoint [232–234].

Moreover, it was important to develop an understanding of pathway-level changes that were occurring due to probe treatment of diseased cells. Fig. 5.2.3 shows the pathway enrichment summary based on the significantly differentially expressed genes (fold-change > 2, FDR < 0.05). Notably, top enriched pathways were ones relevant to the extracellular matrix and its organisation, as well as collagens and proteoglycans, which is line with established literature on DD and its underlying mechanisms [206, 235–238]. Importantly, the background to testing for enrichment was the full set of all differentially expressed myofibroblast genes, thus controlling for the cellular context of the probe treatment.

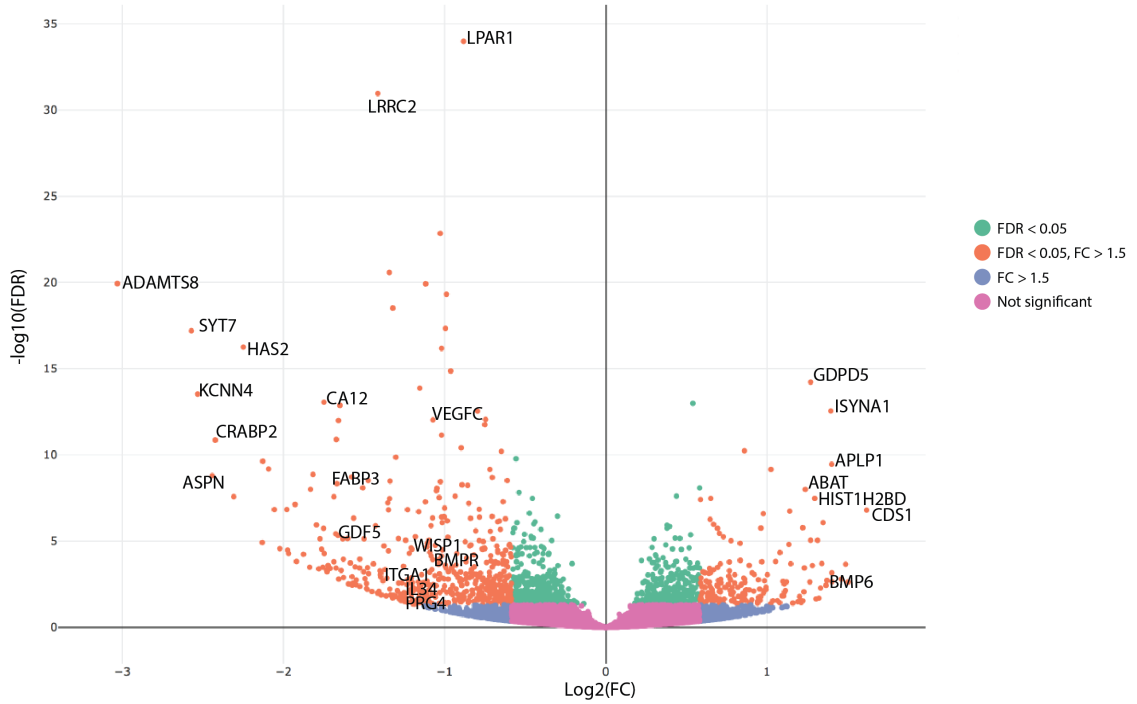



**Figure 5.2.2: RNA-seq results of probe-treated myofibroblasts vs control (DMSO).** A volcano plot of differentially expressed genes based on CBP30 versus DMSO treatment of patient-derived myofibroblasts. A fold-change cutoff $> 1.5$ was used, as well as FDR $< 0.05$. Highly significant (based on fold-change, FDR, or both) genes are labelled, along with select, fibrotic-related genes.

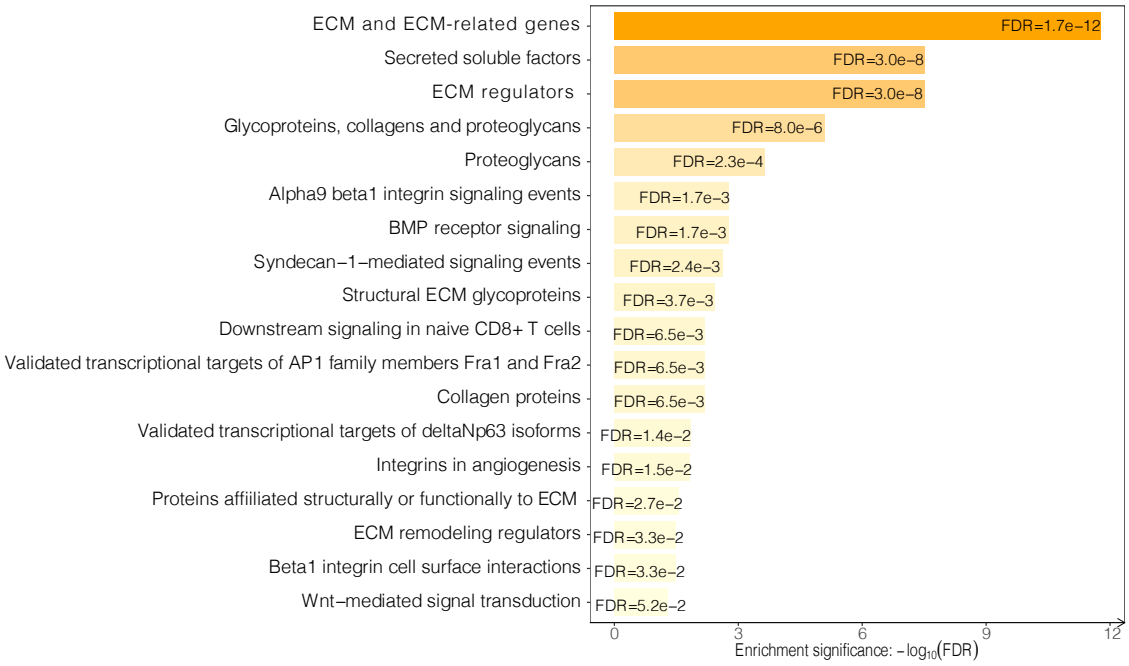



**Figure 5.2.3: Pathway enrichment of differentially expressed CBP30-treated myofibroblasts.** Top enriched pathways captured by significantly (FC $> 2$, FDR $< 0.05$) differentially expressed genes in CBP30-treated myofibroblasts.

### 5.2.3 RNA-seq analysis of healthy versus disease cells

Following RNA-seq analysis of probe-treated cells, it was clear that the CBP30 inhibitor was generally affecting fibrotic genes; however, it was not clear how relevant the affected pathways were to the specific Dupuytren's disease context in question, and whether downstream effects of CBP30 treatment was likely to be beneficial in counteracting disease-driving mechanisms. Therefore, in order to contextualise the results on gene expression of probe-treated myofibroblasts and investigate genes important in driving healthy cells into a diseased state, a separate RNA-seq analysis was conducted using donor-matched healthy (palmer fibroblasts) and disease (myofibroblasts) cells, across 5 donors. In total, 201 genes were significantly differentially expressed (fold-change >2, FDR < 0.05), with 107 genes significantly up-regulated and 94 significantly down-regulated. Fig. 5.2.4 summarizes the gene-level results of this analysis, which include a number of matrix metalloproteinases whose role in fibrotic lung micro-environments has been reviewed [239]; *CHI3L1*, a regulator of fibroproliferative processes that has been shown to serve as a bio-marker to distinguish stage of liver fibrosis [240, 241]; and *SFRP4*, which has been investigated in the progression of renal fibrosis and has previously been implicated as a potential anti-fibrotic therapeutic target [242, 243].

Further, the two sets of differentially expressed genes (i.e. the differentially expressed genes from the probe treatment versus control group, and the differentially expressed genes from the disease cells versus healthy cells group) were contrasted at both the gene and the pathway levels, in order to determine areas of overlap that may be of therapeutic value. Notably, several disease-relevant pathways were significantly enriched and common to both groups including elastic fiber formation (myofibroblast RNA-seq FDR = $3.6\text{x}10^{-4}$, CBP30 RNA-seq FDR = $3.3\text{x}10^{-6}$); collagen degradation (myofibroblast RNA-seq FDR = $3.9\text{x}10^{-4}$, CBP30 RNA-seq FDR = $6.7\text{x}10^{-3}$); and extracellular matrix degradation (myofibroblast RNA-seq FDR = $2.4\text{x}10^{-3}$, CBP30 RNA-seq FDR = $1.9\text{x}10^{-2}$). The relevance of these pathways to disease is supported by previous studies; for instance, analysis of DD contracture cords shows a significant decrease in elastic fibers as compared to clinically unaffected regions [244]. Similarly, the role of altered collagen deposition and subtype ratios has been well established in the context of DD, and collagenase injection is one available therapeutic approach to

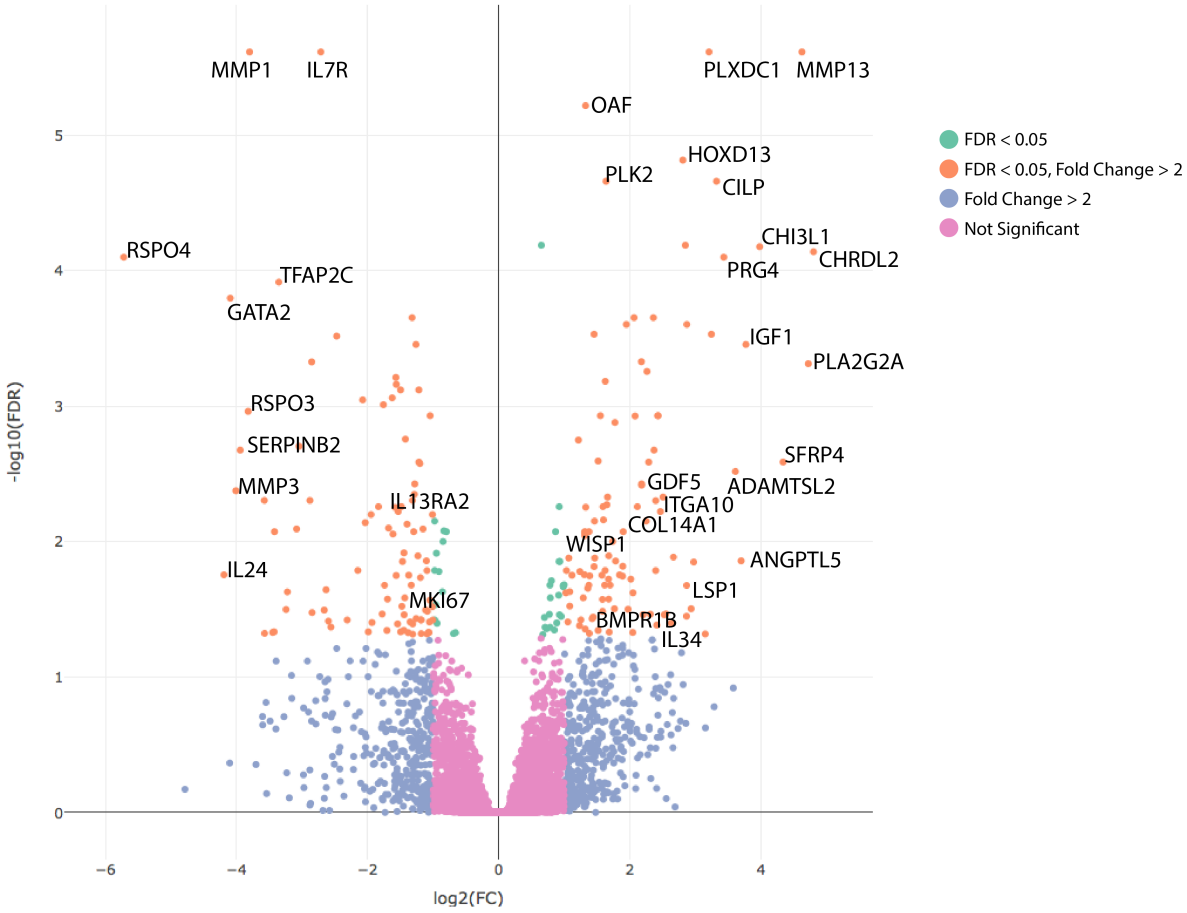



**Figure 5.2.4: RNA-seq results of myofibroblasts vs palmer fibroblasts.** A volcano plot of differentially expressed genes based on myofibroblasts versus donor-matches palmer fibroblasts from patient-derived tissue. A fold-change cutoff $> 2$ was used, as well as FDR $< 0.05$. Highly significant (based on fold-change, FDR, or both) genes are labelled, along with select, fibrotic-related genes.

**Table 5.2.2: Genes up-regulated in disease and down-regulated by SGC-CBP30.** Genes meeting a fold-change cut-off $> 2$ and adjusted p-value $< 0.05$, which were up-regulated in disease state (as measured by myofibroblast vs palmer fibroblast RNA-seq) and down-regulated by the CBP30 probe (as measured by CBP30 vs DMSO RNA-seq).

| Gene | Differential expression in CBP30-treated cells ($Log_2$(fold-change)) | Adjusted p-value | Differential expression in myofibroblasts ($Log_2$(fold-change)) | Adjusted p-value |
|---|---|---|---|---|
| *BMPR1B* | -1.084 | $7.26 \times 10^{-5}$ | 1.423 | $3.70 \times 10^{-2}$ |
| *CMKLR1* | -1.980 | $1.47 \times 10^{-7}$ | 2.868 | $2.00 \times 10^{-3}$ |
| *GDF5* | -1.675 | $3.81 \times 10^{-6}$ | 2.179 | $4.00 \times 10^{-3}$ |
| *HSPB6* | -1.303 | $1.38 \times 10^{-10}$ | 1.514 | $3.00 \times 10^{-3}$ |
| *IL34* | -1.233 | $3.36 \times 10^{-3}$ | 2.410 | $4.10 \times 10^{-2}$ |
| *ITGA10* | -1.396 | $7.03 \times 10^{-4}$ | 2.393 | $5.00 \times 10^{-3}$ |
| *LSP1* | -1.565 | $4.68 \times 10^{-7}$ | 2.865 | $2.10 \times 10^{-2}$ |
| *PLXDC1* | -1.659 | $1.01 \times 10^{-12}$ | 3.206 | $2.43 \times 10^{-6}$ |
| *PRG4* | -1.215 | $6.82 \times 10^{-3}$ | 3.431 | $7.95 \times 10^{-5}$ |
| *SYT7* | -2.571 | $6.35 \times 10^{-18}$ | 2.357 | $2.22 \times 10^{-4}$ |
| *TRIM9* | -1.341 | $3.44 \times 10^{-8}$ | 1.891 | $1.51 \times 10^{-2}$ |
| *WISP1* | -1.208 | $2.50 \times 10^{-5}$ | 1.309 | $8.54 \times 10^{-3}$ |

disease management [235, 236]. Likewise, excessive accumulation of extracellular matrix molecules is a common feature in fibrotic diseases and has been identified as an important therapeutic angle [206, 237]. The fact that these core pathways were enriched in both healthy vs disease as well as probe vs no probe groups implicates shared, functionally disease-relevant genes that are underpinning disease progression and that may be viable therapeutic targets.

In order to further interrogate the therapeutic potential of targeting differentially expressed genes, overlap between the two RNA-seq data-sets was analysed while taking into consideration the direction of change. More specifically, a subset of genes (Fig. 5.2.2) was identified, representing targets that were simultaneously up-regulated in disease cells compared to healthy cells, as well as down-regulated by the probe inhibitors in the probe vs no probe group. These genes included *ITGA10* and *WISP1*, both of which have recently been implicated in a study of differential gene expression of DD tissue [245]; *BMPR1B*, whose differential expression in diseased versus healthy cells has been previously investigated [246]; *IL34*, which has been found to contribute to pro-fibrotic processes during chronic hepatitis C virus infection as well as recently reported to enhance lung fibrosis in a mouse model [247, 248]; and *HSPB6*, whose role in the development and maintenance of fibrosis via interaction with connective tissue growth factor has been previously reviewed [249].

### 5.2.4 Proteomic analysis of probe-treated cells

To complement the transcriptomic analysis of SGC-CBP30 treatment, an additional proteomic analysis was completed of the CBP30-treated myofibroblasts collected from the 6 donors, following 3 days of treatment. This identified a total of 40 significant (fold-change > 1.5, $p < 0.05$) proteins, including 14 down-regulated and 26 up-regulated (Supplementary Table 8.0.4). Given the fact that proteomic and transcriptomic approaches provide complementary points of view, the pathway enrichment results of the proteomic screen was contrasted to those of the previously mentioned RNA-seq analyses. Fig. 5.2.5 summarizes the most significant and overlapping pathways between the three sets of data, which is largely comprised of pathways related to extracellular matrix structure, remodelling, and regulation. These findings are consistent with well-established literature around disease pathology and the role of excess extracellular matrix driven by disease-state myofibroblasts [206, 237, 238]. Additionally, a correlation analysis between the fold-change values of the proteomic-identified targets (CBP30 to DMSO proteomics) and of the corresponding RNA-seq-identified targets (CBP30 to DMSO transcriptomics) shows a significant correlation between the two groups (Pearson's correlation = 0.56, $p < 0.002$).

## 5.3 Discussion

The results presented in this chapter highlight how a computational approach can be used to further process GWAS information in a way that integrates other genomic datasets and assigns scores that can be used to inform downstream analyses. The use of the Pi tool on recent DD GWAS information implicated two epigenetic regulators, *CREBBP* and *EP300*, as central nodes at the intersect of highly enriched, disease-relevant pathways captured by prioritised genes. Further, targeting of these two nodes via the inhibitor probe SGC-CBP30 in diseased patient cells resulted in gene and pathway-level changes that were reflective of key disease processes. Moreover, the CBP30-induced gene signature had relevant overlap with a differential gene expression signature of healthy vs diseased cells, implicating the potential of CBP30-related targets as therapeutic options worth investigating. These similarities were further reflected at the protein-level, as measured by proteomic analysis, indicating

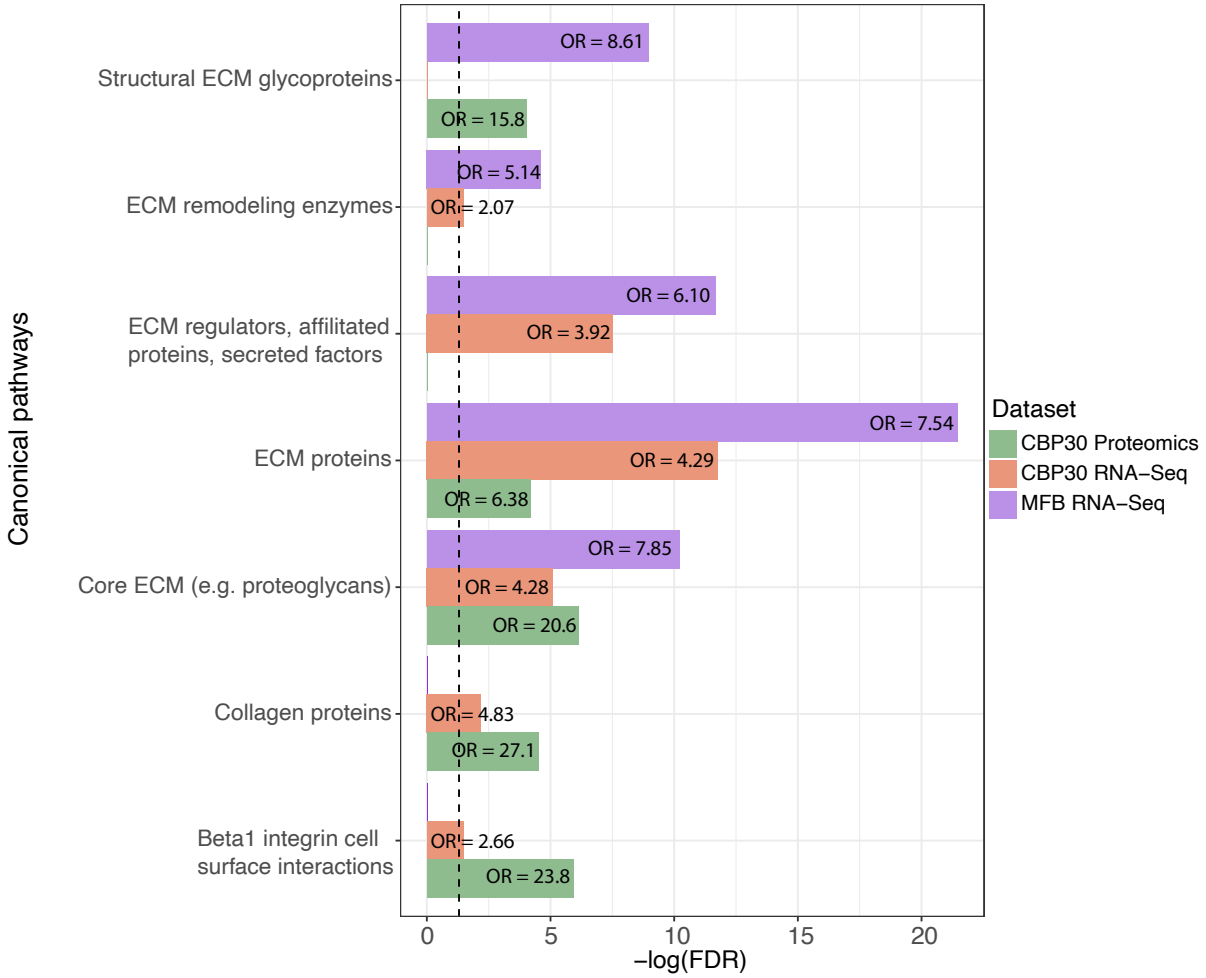



**Figure 5.2.5: Pathway enrichment overlap between groups.** Overlap of enriched canonical REACTOME pathways captured by significant genes based on (i) CBP30-treated vs DMSO proteomic analysis of myofibroblasts; (ii) CBP30-treated vs DMSO RNA-seq analysis of myofibroblasts; (iii) myofibroblasts vs palmer fibroblasts RNA-seq analysis. The dashed vertical line represents an FDR cutoff of 0.05.

that additional work to understand the therapeutic potential is warranted.

Utilising tools to measure genome-wide and proteome-wide changes between treatment or cell groups was useful in understanding broad disease relevance in different contexts. First of all, the healthy versus disease cell analysis using donor-matched palmer fibroblasts and myofibroblasts was an important step to confirm previously well-established findings regarding drivers of disease pathogenesis at both a gene and a pathway level. Second, establishing this baseline in our own hands allowed us to more directly compare the results of probe treatment to that of healthy versus diseased state. The fact that treatment of disease-driving myofibroblasts with CBP30 resulted in similar pathway changes as compared to disease versus healthy cells was reassuring. These pathways, largely related to collagen deposition and extracellular matrix organisation and degradation, are in line with previous, well-established literature around fibrotic disease development as well as specifically DD progression [206, 216, 235, 237, 238]. Therefore, this suggests both that probe treatment of disease-driving cells influences key genes underpinning the pathogenic mechanisms in question, as well as that the changes affected by the probe mirror those changes found in the progression of healthy cells (palmer fibroblasts) into a diseased state (myofibroblasts). These findings were further reflected in the proteomic data, which not only showed overlap at a broad, pathway enrichment-level, but also in terms of significant correlation of magnitude of effect (fold-change). In other words, not only were similar targets affected at an mRNA and a protein level, but they were also affected to a strongly, statistically significant, level. It is also critical to note that all enrichment tests were conducted with a background that took into account the cellular context in question. For instance, enrichment calculation using the most significantly differentially expressed genes in probe-treated cells was done with a background of all differentially expressed genes from patient myofibroblasts, thus taking into account all of the actively expressed genes in diseased cells (as compared to healthy controls) and establishing that cellular context as a baseline. This means that identification of enriched pro-fibrotic processes in CBP30-treated cells is most likely due to a direct effect of the probe on key fibrotic genes, and not simply a residual background finding due to the cells in questions (myofibroblasts) which would inherently contain a higher proportion of active fibrotic genes.

While having a broad picture view of these different treatment and disease groups was useful in

identifying disease-relevant patterns, the ultimate goal of identifying therapeutic candidates necessitates a closer, target-level understanding of changes occurring. For instance, the most abundant peptides identified in CBP30-treated cells were Collagen Type VI, including Col6A1/2/3. Past literature has examined the role of Type VI Collagen and uncovered a role in regulation of extracellular matrix assembly and deposition as demonstrated by comparing wild type to Type VI-deficient mice [250]. Additionally, evidence exists that shows an over-expression of Type VI collagen in patients with pulmonary fibrosis, with a specific tie to the IL-23 signaling axis that leads to excessive inflammation [251]. Given the epigenetic role of the direct CBP30 targets, *CREBBP* and *EP300*, and their involvement as co-factors for a number of transcription factors [252, 253], understanding indirect effects of CBP30 (such as via Collagen Type VI) may provide a more tractable way to potential therapeutics. Interestingly, some early follow-up work done with CBP30 (in extension to the results presented here) seems to suggest that the probe does not globally affect all pro-fibrotic genes. For instance, *TGFB1* expression remained unaffected following probe treatment.

In fact, work is being done along this 'target-subset' tangent to take the findings presented here further. For instance, in order to understand the regulatory effects of *CREBBP* and *EP300*, chromatin immunoprecipitation (ChIP)-Seq analysis was done on DD myofibroblasts treated by CBP30 using antibodies specific to those targets, in two independent biological replicates. This way, it becomes possible to get a better understanding of where these co-factors may physically be binding in the genome, in order to better understand what is under their direct control from a gene-expression standpoint. Association of the identified binding regions to the nearest annotated genes revealed a significant enrichment of integrin and extracellular matrix interactions, at the pathway level, confirming the broad RNA-seq findings presented in this chapter. Further analysis of these identified genes (for example, via RNAi targeting) would allow for a more nuanced understanding of disease progression, and could be investigated in the context of fibroblast to myofibroblast transition for more specific insight into therapeutic potential. Additionally, work is being done to add robustness to the above-mentioned Collagen Type VI findings, via RNA-seq analysis of DD patient myofibroblasts following application of Col6A1/A2/A3 RNAi oligos.

Furthermore, translating the largely expression-based results here into more concrete biological

understanding of disease progression would also benefit from complementary, disease-relevant cellular assays. For instance, further investigating probe effects on cell morphology is a possible next step in translating target-level insights into a more holistic understanding of the downstream manifestation of these changes as it pertains to disease progression. Moreover, cell contractility force is another useful lens through which to analyze the direct role of targets in disease, as it specifically pertains to DD, and these experiments are also underway as follow-on to the work here. Taken together, this would allow for a more end-to-end understanding of the potential tractability of novel therapeutic targets in the context of DD: from broad gene and protein-level changes, to their more concrete relevance in disease-driving pathways, to their specific role in cellular functions that pertain to DD (extracellular matrix deposition, contraction force, cell morphology, chemotaxis, and others), and ideally/ultimately to their de-risking for entry into early stage clinical trials.

Naturally, some limitations do exist with the type of work described in this chapter. For instance, deciding on cut-offs to use for calling differential RNA-seq results is fairly arbitrary. Generally, an FDR < 0.05 is considered acceptable along with a fold-change of >2. However, an argument can be made that fold-change should be more relaxed (to, for example, > 1.5, as done in one portion of the results section here, when total genes called is low), or even raised to a more stringent threshold. Further, weighing the trade-off between FDR and fold-change is not always an obvious decision, and gene-level results should therefore be interpreted cautiously (hence the pathway-level analyses prevalent in this chapter). Similarly, a number of putative targets may be missed if they fall just outside the bounds of FDR and fold-change criteria, which is another important consideration when drawing conclusions from such experiments. This chapter attempted to circumvent some of these limitations by layering multiple sources of data (RNA-seq, proteomics, etc.) on top of each other, to add greater confidence to differences identified in any one single experiment.

## 5.4 Conclusion

In conclusion, this chapter demonstrated how we can move from disease-specific GWAS data (1000s of genes) to a genomic-and-network prioritised list (10s of genes), and how select targets from this subset can be interrogated via RNA-seq and proteomic analyses. Specifically, the results showed a

relevance of the epigenetic modulators CREBBP/EP300 in Dupuytren's disease, as demonstrated by the overlapping and complementary pathways affected by inhibiting those two targets, as compared to pathways relevant to the progression of healthy cells into a diseased state. As such, these methods can help inform the disease-relevance of chosen targets at a broad, genome-wide and proteome-wide level, in order to identify any areas of potential therapeutic tractability. Moreover, the overlap of and correlation between affected genes and pathways at the RNA and protein levels uncovered a potential for more targeted investigation of genes modulated by CREBBP/EP300, such as Collagen Type VI. Accordingly, the resulting findings require closer, single-gene-level investigation in order to de-risk them as clinical targets, which is the topic of the following and last results chapter.

*It's gone! It's done... I can see the Shire. The Brandywine River. Bag End. Gandalf's fireworks, the lights, the party tree... I'm glad to be with you... here at the end of all things.*

Frodo Baggins

# 6

# Gene-level experimental validation of prioritisation outputs in the context of systemic lupus erythematosus

## 6.1 Introduction

Thus far, the work in the thesis has demonstrated how a novel computational approach, Priority Index, can be used to aggregate genomic data in order to analyse disease-specific GWAS information to prioritise candidate genes for downstream analysis, based on their functional links to disease-associated SNPs. The application of Priority Index to myasthenia gravis, for example, demonstrated how the pipeline could be applied to a novel disease context which has early GWAS data available but lacks nuanced knowledge of disease pathogenesis or viable therapeutic targets, and how GWAS data could be harnessed to yield a gene priority list. This output was combined with a variety of disease-relevant experimental data-sets, building a case for how the prioritised candidates

can help guide hypotheses on disease mechanism. Further, a combination of functional genomics, network interactions, and statistical fine-mapping approaches were used to identify potential therapeutic targets and disease-relevant cell types, paving the way for moving into pre-clinical experimental validation.

Additionally, work in the *Genomic validation of prioritisation outputs in the context of Dupuytren's disease* chapter demonstrated how computational prioritisation could be used as a precursor to experimental validation; more specifically, the work demonstrated an opportunity for using highly specific inhibitor probes to validate predicted targets, in disease-relevant contexts such as patient-derived cells. By comparing the gene expression profiles of probe-treated cells to controls, as well as contrasting this expression 'fingerprint' to that of disease versus healthy cells, I was able to demonstrate pathway-level relevance of the two prioritised targets (*CREBBP* and *EP300*) to disease, as well as nominate single gene targets (such as Col6) for further investigation. Moreover, these results were supported by proteomic analysis as well, demonstrating how a convergence of relevant datasets could be useful for informing early pre-clinical investigation. Accordingly, this final results chapter aims to demonstrate how this outlined approach can be taken a critical step further, as computationally prioritised genes are taken into an experimental setting where they can be validated in disease-relevant assays, at a single target-level.

The chapter focuses on systemic lupus erythematosus (SLE), an autoimmune disorder that affects multiple organs, including skin, joints, and kidneys, among others. Substantial gains have been made in understanding SLE through GWAS; to date, more than 50 risk loci have been associated with SLE [75, 81, 85]. For instance, one early study that examined 431 cases and 2155 controls found a number of disease-associated variants near *TNFAIP3* that met genome-wide statistical significance, along with variants near *STAT4* and *BLK* [75]. Later, a much larger GWAS analysing 7219 cases and 15991 controls confirmed these findings along with 10 others, identifying a total of 43 susceptibility loci [81]. Moreover, the authors included information on mouse model immune phenotype annotations, coding variants, and immune cell eQTL signals in addition to the GWA information in order to propose the most likely causal genes based on the confluence of these data types, which included *TYK2*, *IRF5/7*, and *IL12A*, among others. Further, studies have attempted to interrogate these

findings in greater detail, through the use of combined techniques including allele-specific gene expression analysis and machine learning-informed classification. For example, Almlof and colleagues (2017) used genotype data from 1160 SLE patients and 2711 controls to train a random forest classifier that assigned a gene importance score based on the data, in order to identify gene regions that contain alleles which increase SLE risk, which was then combined with allele specific gene expression information in B and T-cells to understand potential functional effects of these variants on development of disease [254]; among the top 40 proposed genes, the authors identified three primary novel candidates (*ZNF804A*, *CDK1*, and *MANF*) as highly likely to be SLE-relevant, along with other genes previously studied in an SLE context. However, disease complexity and heterogeneity make it difficult to therapeutically modulate disease; there currently exists no cure for the disease, treatments generally focus on symptomatic management of inflammation, and the potential therapeutic space consists of a complex web of interlinked immune cell populations, with numerous failed attempts at treatment [255].

Therefore, there is an opportunity to use the genomic prioritisation approach outlined in this thesis, for the purposes of prioritising novel candidate targets for clinical development in the context of lupus. Moreover, the ULTRA-DD consortium (which this project is a part of) is comprised of a variety of disease expert groups, each of which focus on advancing therapeutic knowledge for their corresponding disease. One such group, based out of the Karolinska Institute in Sweden, works on SLE, and has developed a disease-relevant assay for screening pre-clinical targets. Specifically, this assay uses peripheral blood mononuclear cells (PBMCs) isolated from patients or healthy controls, which are then stimulated in order to activate B-cell maturation, a process which occurs in the presence of autoimmune disease. The readout of the assay is then used to quantify the effects of an intervention on this process.

### Aims

Given the above, SLE presented an interesting opportunity for using Pi to prioritise targets, which could then individually be validated using patient-derived tissues and knockdown experiments. Specifically, the aim of this chapter are as follows:

I. Apply Priority Index to published systemic lupus erythematosus GWAS to prioritise genes by genomic and network analysis;

II. Nominate targets of interest on the basis of the Pi output and various downstream analyses such as subnetwork cross-talk, druggability, and others;

III. Test the hypothesis that the targets identified in Aim II are relevant to disease progression by validating them via siRNA knockdown in a patient-derived B-cell maturation assay.

## 6.2 Results

### 6.2.1 SLE Priority Index output

#### Gene-level analysis

In order to determine the initial shortlist of genes for experimental follow-up, the Pi tool [64] was used. Specifically, the algorithm was used to prioritise genes based on SLE GWAS summary data, sourced from GWAS catalog [66], together with all genomic and annotation predictors [87–95]. Fig. 6.2.1 shows the top 30 prioritised genes based on Pi analysis of SLE. As expected, a number of previously-implicated genes were highly prioritised in this analysis. For instance, the association between HLA genes and SLE is well documented in various populations, with both risk and protective allelic effects demonstrated [256, 257]. Further, high ranking Pi genes like *TNIP1*, *PRDM1*, and *BLK* have also been implicated within an SLE context [258]; these genes, along with high-ranking targets *PHRF1*, *STAT4*, and *PXK*, have various genomic evidence supporting their role in SLE. Table 6.2.1 summarises the supporting evidence for each of the top 30 Pi prioritised genes. Specifically, a number of them are cGenes, which is to say that SLE-associated SNPs physically interact with the gene promoters, and are thus likely functionally modulating downstream processes [87]. Lastly, some of these top genes (like *PHRF1*) also have eGene evidence, directly linking disease-related SNPs to gene expression, in disease-relevant cell types like B-cells [88]. Taken together, the Pi-prioritised gene list provided a holistic set of candidate genes that included those that have been previously linked to SLE based on numerous genomic analyses. For the purposes of

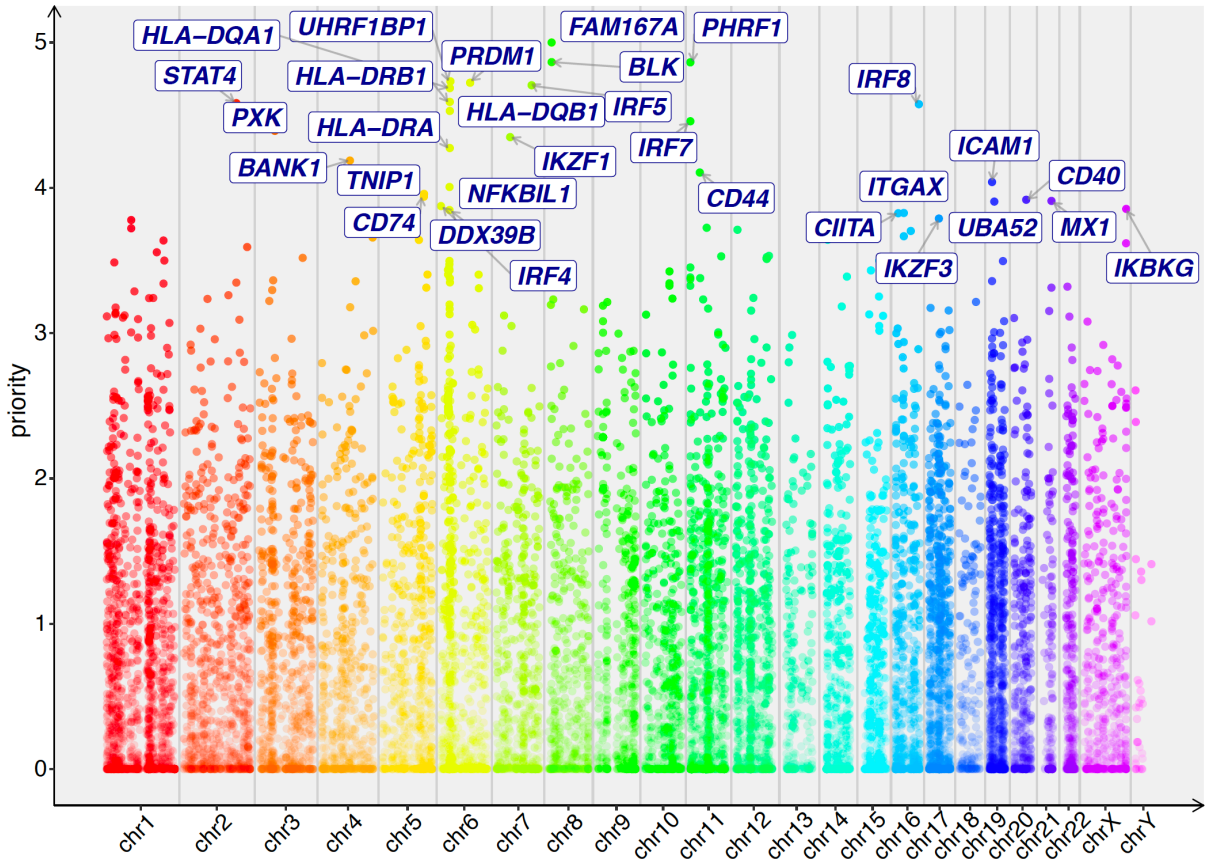



**Figure 6.2.1: Top SLE Pi-prioritised genes.** A Manhattan plot of the top 30 prioritised genes based on the Pi discovery mode output. Input was a collection of SLE GWAS summary data, and prioritisation was done using both genomic and annotation-based predictors. Genomic location is shown along the x-axis, while Pi priority score is along the y-axis.

experimental validation, the top 1% of genes ( 150) were selected as a shortlist.

### Pathway-level analysis

In addition to Pi prioritisation, potential targets were analysed from a pathway enrichment level (using the top 1% genes) as well as through a subnetwork cross-talk approach. Fig. 6.2.2 shows the results of these two analyses. As expected, pathway enrichment showed that many of the expected signals were captured in the top 1% of genes, including cytokine and interleukin signaling (odds ratio = 41.2, $p_{adjusted} = 4.30 \times 10^{-88}$). Moreover, TLR cascades were significantly enriched for as well (odds ratio = 15, $p_{adjusted} = 6.80 \times 10^{-14}$), which is consistent with past studies that have shown that TLRs likely play a role in initial B-cell activation and SLE disease progression or exacerbation, through both microbial

**Table 6.2.1: Top 30 Pi-prioritised SLE genes.** A summary of the top 30 genes prioritised by Pi based on all SLE GWAS SNPs in GWAS catalog, meeting a disease association significance of $p < 5 \times 10^{-8}$. The Pi ranking is included for each gene, along with the number of nGene (proximity to SNP), eGene (eQTL signal), and cGene (physical SNP – gene promoter interaction) annotations captured by the input SNP list (collectively representing seed genes), as well as each of the annotation predictors (disease ontology, phenotype ontology, and immune function ontology). Genes without any predictor annotations (i.e. non-seed genes) were prioritised on the basis of strong network connectivity to seed genes, based on the STRING gene interaction map (limited to high confidence interactions).

| GENE | RANK | NGENE | EGENE | CGENE | DGENE | PGENE | FGENE |
|---|---|---|---|---|---|---|---|
| FAM167A | 1 | 1 | 0 | 7 | 0 | 0 | 0 |
| BLK | 2 | 1 | 4 | 7 | 1 | 0 | 0 |
| PHRF1 | 3 | 1 | 3 | 6 | 0 | 0 | 0 |
| UHRF1BP1 | 4 | 1 | 7 | 1 | 0 | 0 | 0 |
| PRDM1 | 5 | 1 | 0 | 9 | 1 | 0 | 0 |
| IRF5 | 6 | 1 | 5 | 0 | 1 | 1 | 0 |
| HLA-DQA1 | 7 | 1 | 0 | 8 | 1 | 1 | 1 |
| HLA-DRB1 | 8 | 0 | 0 | 7 | 1 | 0 | 0 |
| STAT4 | 9 | 1 | 1 | 6 | 1 | 1 | 0 |
| IRF8 | 10 | 1 | 2 | 6 | 1 | 1 | 1 |
| HLA-DQB1 | 11 | 1 | 0 | 11 | 1 | 1 | 1 |
| IRF7 | 12 | 1 | 4 | 3 | 1 | 1 | 1 |
| PXK | 13 | 1 | 1 | 6 | 0 | 0 | 1 |
| IKZF1 | 14 | 0 | 0 | 10 | 1 | 1 | 0 |
| HLA-DRA | 15 | 1 | 1 | 10 | 1 | 0 | 1 |
| BANK1 | 16 | 1 | 1 | 0 | 1 | 0 | 0 |
| CD44 | 17 | 0 | 0 | 6 | 1 | 0 | 0 |
| ICAM1 | 18 | 0 | 0 | 7 | 1 | 0 | 1 |
| NFKBIL1 | 19 | 0 | 0 | 11 | 1 | 0 | 0 |
| TNIP1 | 20 | 1 | 0 | 2 | 1 | 0 | 0 |
| CD74 | 21 | 0 | 0 | 0 | 0 | 0 | 0 |
| CD40 | 22 | 1 | 5 | 3 | 1 | 1 | 1 |
| MX1 | 23 | 0 | 1 | 0 | 1 | 0 | 1 |
| UBA52 | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
| IRF4 | 25 | 0 | 0 | 0 | 0 | 0 | 0 |
| IKBKG | 26 | 0 | 0 | 5 | 1 | 1 | 1 |
| DDX39B | 27 | 0 | 0 | 11 | 1 | 0 | 0 |
| ITGAX | 28 | 1 | 1 | 0 | 1 | 0 | 0 |
| CIITA | 29 | 0 | 0 | 0 | 0 | 0 | 0 |
| IKZF3 | 30 | 1 | 3 | 6 | 1 | 0 | 0 |

and non-microbial nucleic acid binding [255, 259]. The subnetwork crosstalk analysis confirmed some of the findings mentioned above with, for example, a group of interacting HLA genes identified in two branches of the network. Similarly, a number of interleukin-related genes like *IL2* and *IL12A* were identified which, along with the connecting nodes like *IFNGR1*, present an opportunity to explore therapeutic intervention; in fact, the role of the IL12/23 axis in SLE has previously been explored, with some early data to suggest clinical potential of these targets [260].

### 6.2.2 Epigenetic target screen in SLE

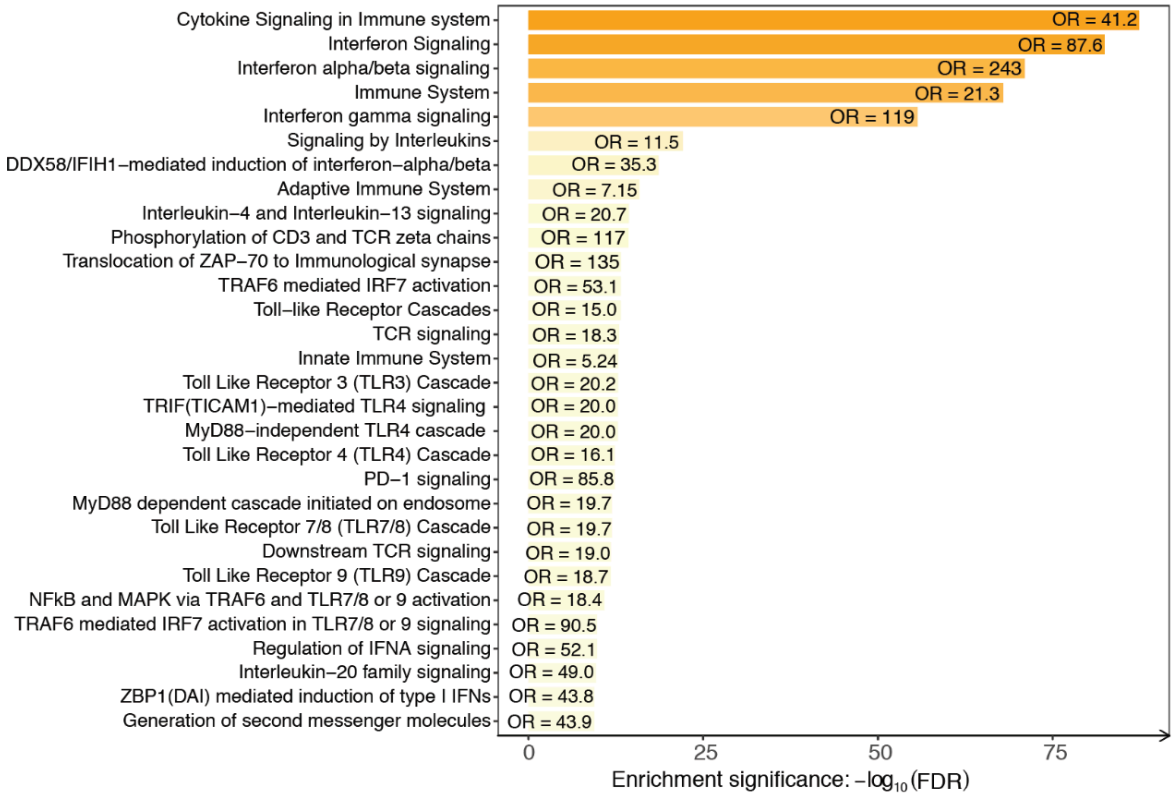
Prior to conducting a series of customised siRNA gene knockdowns, existing resources were leveraged to get a broad picture of Pi's predictive ability in the context of SLE. Specifically, a set of 24 epigenetic probes were functionally assayed in a B-cell IgG secretion protocol (refer to the *Methods* chapter for more details on the assay). This experiment was done for two main reasons:

I. There is evidence of dysregulated DNA methylation and histone acetylation in the pathogenesis of SLE, including significantly variable methylation in disease discordant monozygotic twins [261], as well as decreased CD4+ T-cell histone acetylation in SLE patients [262] along with established therapeutic benefit from histone deacetylase inhibition in a mouse model [263];

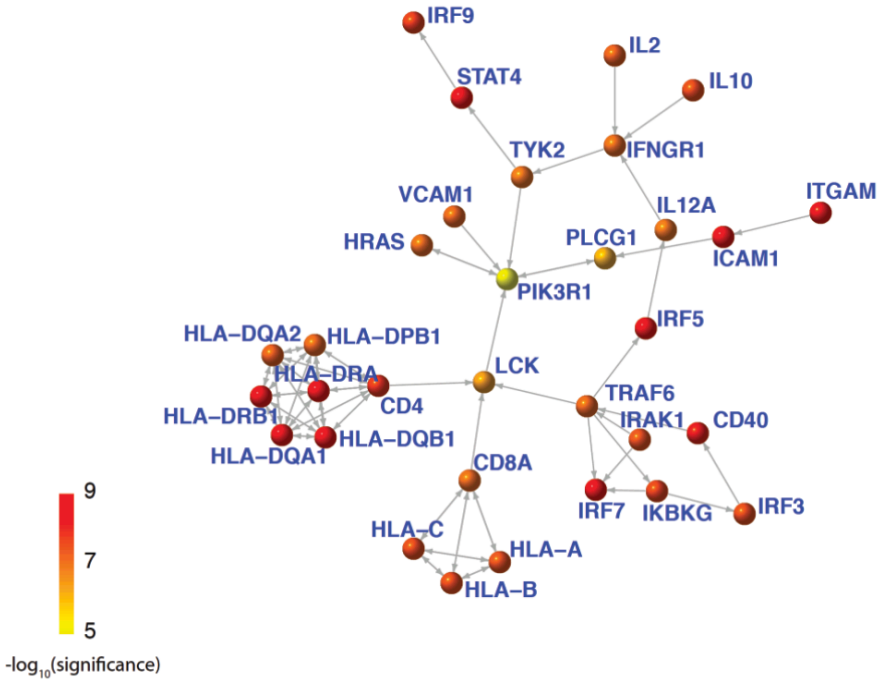
II. The Structural Genomics Consortium (SGC), which the ULTRA-DD is affiliated with, had developed a set of highly specific/selective small-molecule inhibitors against various epigenetic targets, to assist in the study of the biological roles of epigenetic proteins; there was interest from within the ULTRA-DD/SGC to test these probes in various immune-context patient-derived assays.

**Figure 6.2.2 *(following page)*: Pathway analysis of top prioritised genes.** An analysis of the top 1% Pi-prioritised genes in SLE. **(A)** The top 30 enriched pathways captured by top 1% genes, sourced from REACTOME Immune Systems, and meeting an FDR $< 0.05$ threshold; odds ratio (OR) for each pathway is shown within each bar. **(B)** A 30-node subnetwork crosstalk analysis of top 1% genes; shading represents significance of each gene, as determined by overall priority score.

A


Cytokine Signaling in Immune system
Interferon Signaling
Interferon alpha/beta signaling
Immune System
Interferon gamma signaling
Signaling by Interleukins
DDX58/IFIH1−mediated induction of interferon−alpha/beta
Adaptive Immune System
Interleukin−4 and Interleukin−13 signaling
Phosphorylation of CD3 and TCR zeta chains
Translocation of ZAP−70 to Immunological synapse
TRAF6 mediated IRF7 activation
Toll−like Receptor Cascades
TCR signaling
Innate Immune System
Toll Like Receptor 3 (TLR3) Cascade
TRIF(TICAM1)−mediated TLR4 signaling
MyD88−independent TLR4 cascade
Toll Like Receptor 4 (TLR4) Cascade
PD−1 signaling
MyD88 dependent cascade initiated on endosome
Toll Like Receptor 7/8 (TLR7/8) Cascade
Downstream TCR signaling
Toll Like Receptor 9 (TLR9) Cascade
NFkB and MAPK via TRAF6 and TLR7/8 or 9 activation
TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling
Regulation of IFNA signaling
Interleukin−20 family signaling
ZBP1(DAI) mediated induction of type I IFNs
Generation of second messenger molecules
OR = 41.2
OR = 87.6
OR = 243
OR = 21.3
OR = 119
OR = 11.5
OR = 35.3
OR = 7.15
OR = 20.7
OR = 117
OR = 135
OR = 53.1
OR = 15.0
OR = 18.3
OR = 5.24
OR = 20.2
OR = 20.0
OR = 20.0
OR = 16.1
OR = 85.8
OR = 19.7
OR = 19.7
OR = 19.0
OR = 18.7
OR = 18.4
OR = 90.5
OR = 52.1
OR = 49.0
OR = 43.8
OR = 43.9
0
25
50
75
Enrichment significance: −log10(FDR)

B


IRF9
IL2
STAT4
IL10
TYK2
IFNGR1
VCAM1
IL12A
ITGAM
HRAS
PLCG1
ICAM1
PIK3R1
IRF5
HLA−DQA2
HLA−DPB1
HLA−DRA
LCK
TRAF6
HLA−DRB1
CD4
IRAK1
CD40
HLA−DQB1
HLA−DQA1
CD8A
IRF7
IKBKG
IRF3
HLA−C
HLA−A
HLA−B
9
7
5
-log10(significance)

Therefore, there was an opportunity to apply these epigenetic probes to an SLE-relevant cellular assay and compare the relative output of each treatment to the relative rank of the corresponding probe target in the Pi analysis of SLE, acknowledging the fact that the assay captures only one aspect of disease mechanism (B cell maturation and IgG secretion).

More specifically, probes were applied to donor PBMCs in the experimental protocol, and then ranked according to their ability to inhibit IgG production. Pi's ability to predict relative functional effect was then assessed through the calculation of a Spearman rank correlation between the Pi ranking and relative assay effectiveness of each target gene. As a control, the correlation with the SLE Pi output was compared to the correlation between probe effectiveness and the Pi ranking of corresponding gene targets in 29 other immune trait outputs Results of this correlation are shown in Fig. 6.2.3; there was a high, statistically significant correlation between Pi target ranking and inhibitor assay activity (Spearman'srank correlation = 0.69, $p = 2.0 \times 10^{-4}$).

The top hit was *EHMT2*, a gene that encodes a methyltransferase and promotes nuclear stability, which is in line with evidence that implicates nuclear structure alterations in SLE autoimmunity [264]. Additionally, *EHMT2* is an interacting partner with a number of SLE genomic seed genes like *PRDM1*, and is itself a seed gene due to the physical interaction between the *EHMT2* promoter and SLE SNP rs558702. Moreover, the correlation between Pi ranking and assay effectiveness was both higher and more statistically significant in the SLE Pi output as compared to the outputs of the other 29 immune traits analysed by Pi, indicating that relative prediction was disease-specific.

This strong, statistically significant correlation primarily served as a form of broad validation of Pi's SLE output, which demonstrated an ability to predict the relative effectiveness of each probe, rather than as a precursor to specific target follow-up; given the broad, gene-regulatory role of many epigenetic proteins, the focus of the remainder of this chapter was on more specific, non-epigenetic targets. Effectively, this was a demonstration of the ability of the Pi output to predict gene target effectiveness, in a disease-relevant context, which presents an exciting opportunity for implementation of Pi in pre-clinical target screening. For example, when considering a large-scale panel of compounds for therapeutic intervention, one way to prioritise which targets to invest significant time and resources on (experimentally) could be to compare their relative disease-relevant

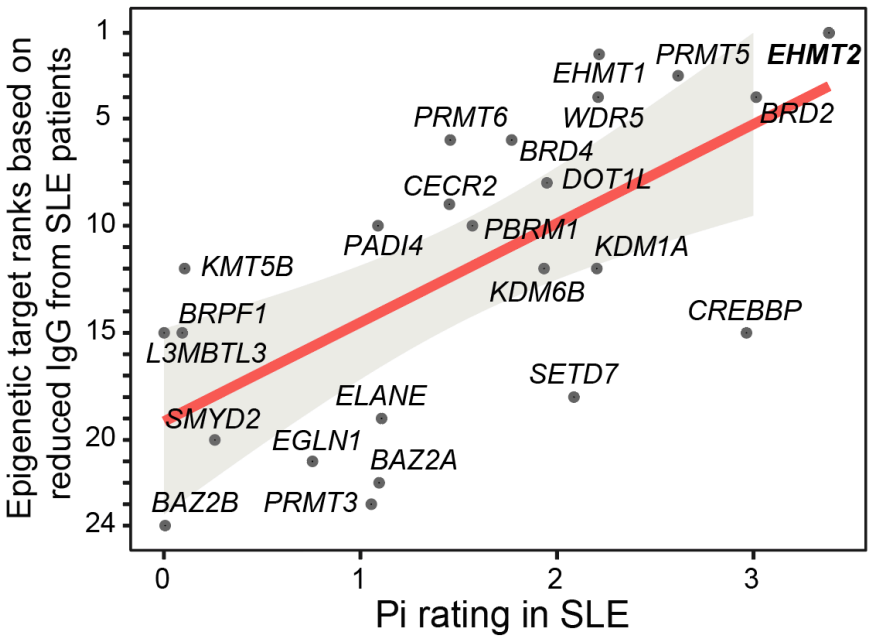



**Figure 6.2.3: Epigenetic probe activity and Pi ranking.** A Spearman rank correlation between epigenetic probe activity (measured by inhibition of IgG secretion in 5 SLE patient PBMC samples) and Pi ranking in the SLE output, reproduced from the 2019 Fang *et al.* study [64]. The rank correlation was 0.69 with a significance $p = 2.0 \times 10^{-4}$. Significance represents an empirical p-value based on a randomized test.

Pi scores, and only focus on those with high priority, relative to others.

### 6.2.3 SLE target nomination

Following Pi prioritisation of SLE genes and comparison of epigenetic target rank to the SGC probe screen, a series of steps were taken to nominate several promising targets for experimental follow-on. Specifically, genes were tagged with information on whether they were: (i) present in the subnetwork cross-talk analysis, implemented using Pi's xSubneterGenes function; (ii) highly prioritised in at least one other SLE-similar trait, including Pi outputs for RA, MS, and T1D, and cross-referenced with the ULTRA-DD consortium's list of nominated high-priority targets, which was compiled based on input from multiple disease-specialists; (iii) druggable, according to an implementation of the fpocket algorithm [151] which tests for accessible protein pockets; and (iv) compatible with the stimulation protocol used in the planned assay, ensuring that none of the potential targets of interest would interfere with the stimulation portion of the assay. The final selection of targets are summarized in

Table 6.2.2. Each target was nominated based on multiple layers of supporting data, outlined below:

**(i) Network crosstalk.** The final Pi priority list was processed using the subnetwork crosstalk approach, which identified 30 highly prioritised genes that are functionally interacting with one another, sourced from the STRING and Pathway Commons databases [96, 101]. This method identified targets that belong to enriched pathways in the SLE output, as well as any linker genes that may serve as an intermediary between two highly ranked targets, providing an opportunity to shortlist the full Pi output on the basis of disease-relevant functional interaction.

**(ii) Multi-disease relevance.** Certain SLE-prioritised targets were also highly ranked in other autoimmune and inflammatory traits that share commonalities with SLE (such as RA, MS, and T1D), presenting an opportunity to build on our understanding of broader autoimmune implications of target therapeutic intervention. Further, the ULTRA-DD Consortium has a list of high-priority targets of interest, nominated by various disease specialists and consultant physicians. Thus, genes that were highly prioritised in other Pi immune outputs and that were contained in the ULTRA-DD list were deemed favourable as targets for assaying in the SLE B-cell stimulation protocol.

**(iii) Druggable pockets.** Since the downstream goal of this analysis was to identify targets that may serve as good candidates for therapeutic intervention/novel drug design, the druggability of each target was a factor to consider. In order to reach scale (i.e. assess druggability of all 15000 SLE-prioritised genes), the fpocket algorithm [151] was implemented due to it's relative efficiency of computing druggability based on the presence or absence of certain pockets in the targets of interest. While a lack of a positive hit based on fpocket does not necessarily mean a target is not druggable, it was useful to recognise targets that do contain druggable pockets, and thus may be more readily taken into a drug development pipeline.

**(iv) Interference with experimental protocol.** Ultimately, any nominated target would be tested in the context of the PBMC experimental protocol described in the Methods section. Since this protocol relies on a cytokine cocktail to stimulate B-cell maturation (comprised of IL4, IL21, IL10, CD40, and ODN2006), it was critical to ensure that knocking down of targets of interest would not interfere with the cytokine signalling pathway, which would confound any results. Therefore, any targets that were directly involved in the signaling pathway of any of the cytokines in the cocktail were flagged for

**Table 6.2.2: Final list of targets nominated for experimental follow-on.** A summary of the final 11 targets taken into lab for siRNA knockdown. Information on network cross-talk, priority across multiple immune disease outputs (multi-trait), presence of druggable pockets (based on fpocket), and genomic seed context is listed for each gene. nGene denotes proximity of gene to a disease-associated SNP, eGene denotes implication of eQTL between gene and disease-associated SNP, and cGene denotes interaction between gene promoter and disease-associated SNP.

| TARGET | CROSSTALK | MULTI-TRAIT | DRUGGABLE POCKET | GENOMIC SEED |
|---|---|---|---|---|
| IL2 | Yes | Yes | Yes | nGene |
| IFNGR1 | Yes | Yes | Yes | cGene- monocytes, macrophages, CD8 cells, neutrophils |
| IL12B | Yes | Yes | Yes | No |
| RELA | Yes | Yes | Yes | No |
| ATF6B | Yes | Yes | No | nGene<br>cGene: B-cells, monocytes, macrophages, T-cells<br>eGene: CD4 |
| IKBKG | Yes | No | Yes | cGene: B-cells, T-cells, neutrophils |
| IRF4 | Yes | No | Yes | No |
| CHUK | Yes | No | Yes | No |
| VCAM1 | Yes | Yes | No | No |
| IL12A | Yes | No | Yes | nGene<br>cGene: B-cells, monocytes |
| MAPK14 | Yes | Yes | Yes | No |

removal from consideration.

**(v) Genomic seed gene.** An additional column was included to denote whether a nominated target was a genomic seed gene; that is, whether there was direct genomic evidence linking the target to SLE-associated SNPs, as well as in what cell-type context this evidence was derived from. Genes that were not genomic seeds were highly prioritised based on their network interaction with genomic seeds. This way, validation would include both genes with direct genetic evidence of implication in disease, as well as highly networked nodes with indirect but strong support for disease involvement.

### 6.2.4 siRNA knockdown

An overview of the siRNA knockdown protocol can be found in Fig. 6.2.4, while Fig. 6.2.5 shows the gating strategies used during analysis; detailed methods can be found in the *Methods* chapter.

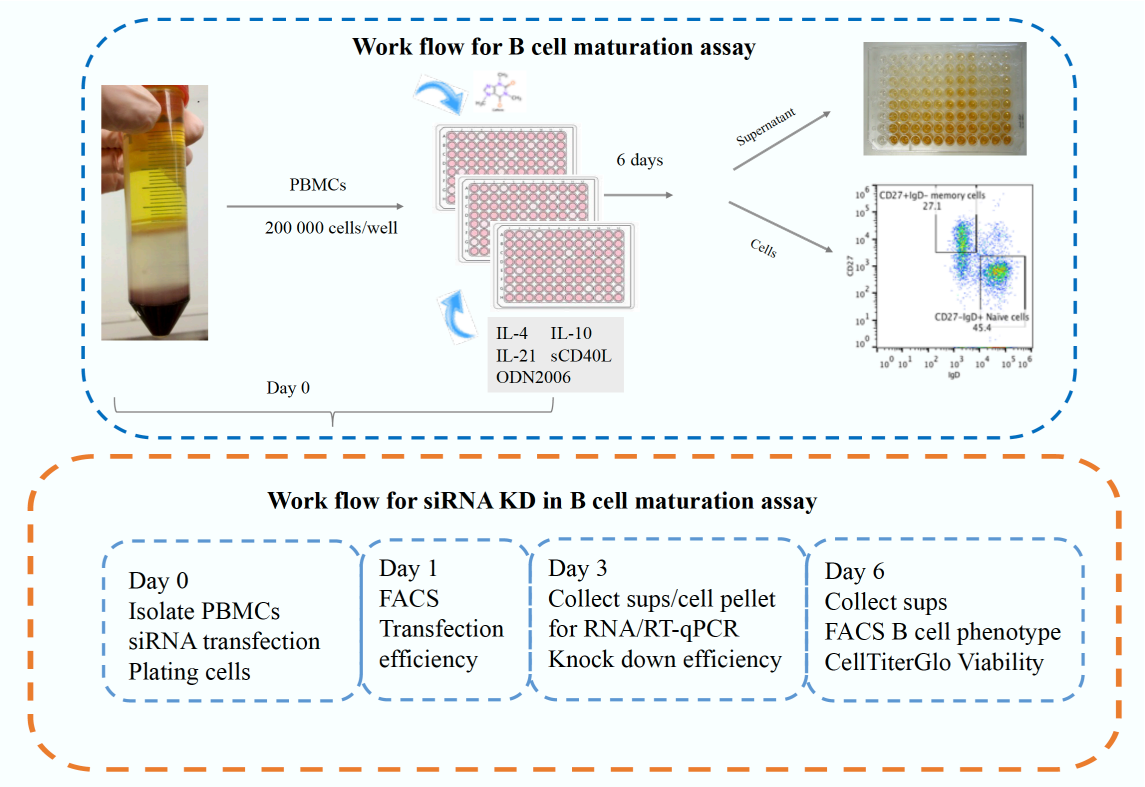
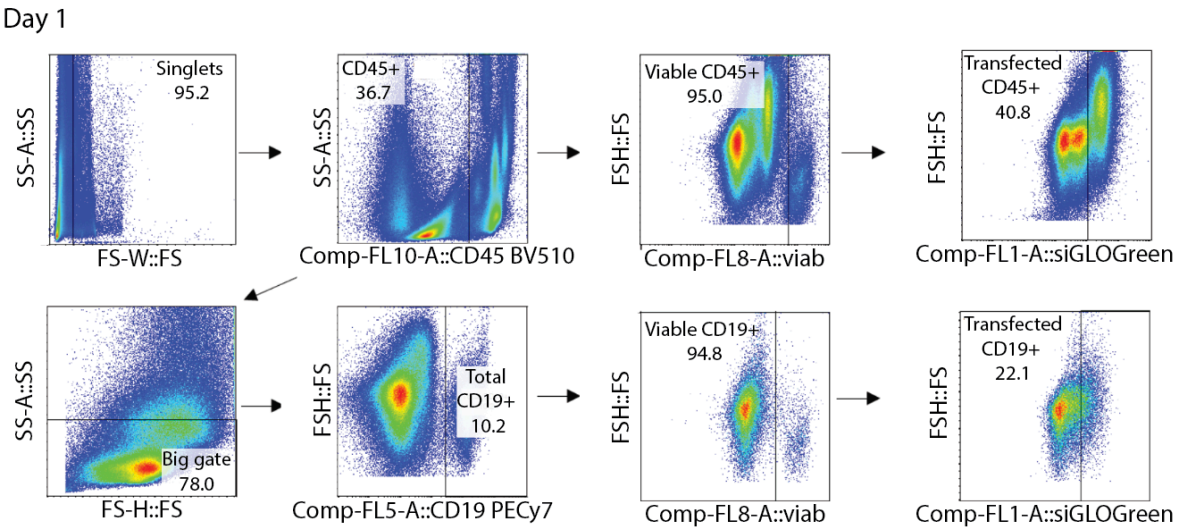


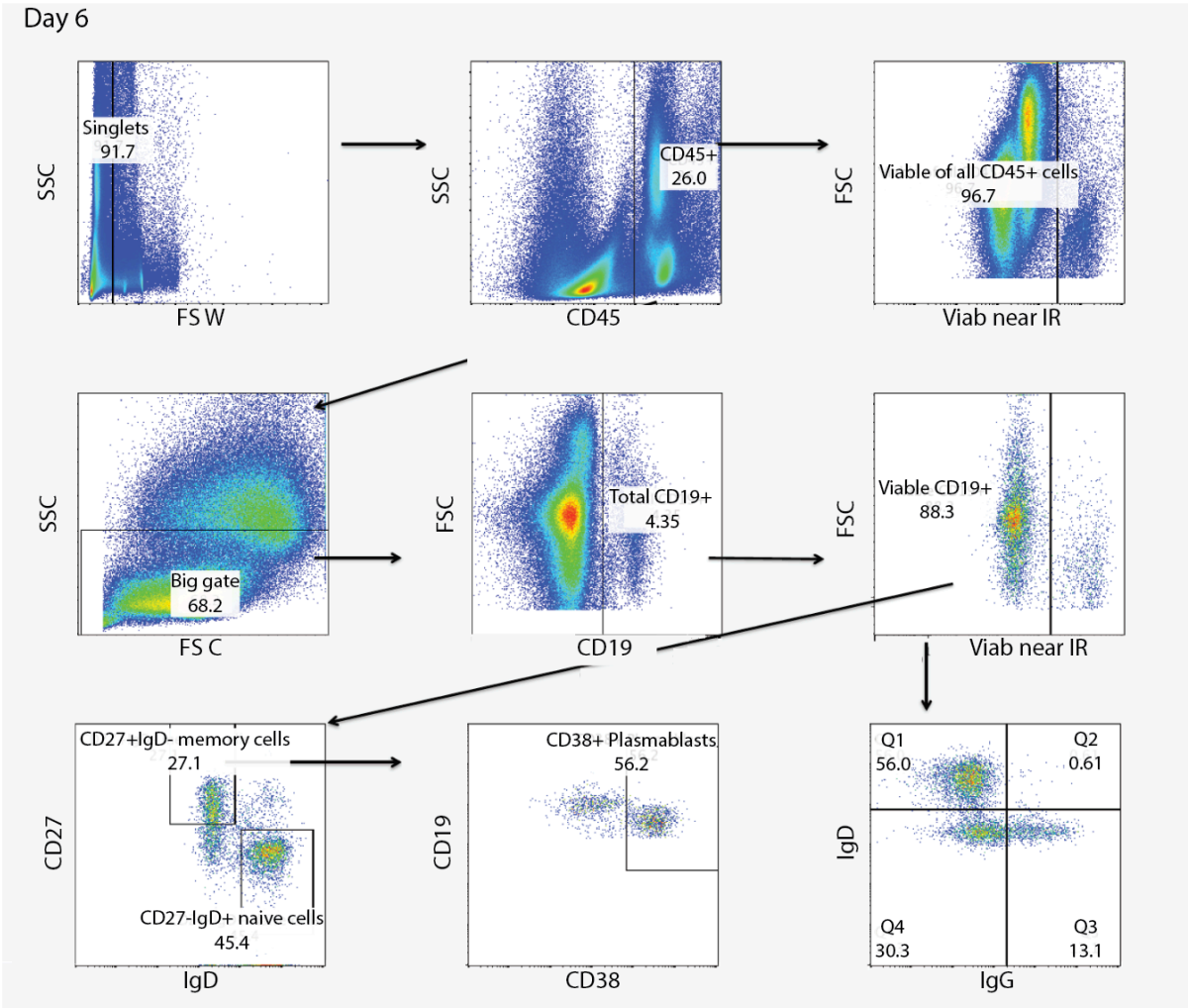

**Figure 6.2.4: siRNA knockdown workflow.** A summary of the workflow for siRNA knockdown of target genes in a B-cell assay, including measures of transfection efficiency, gene knockdown efficiency, cell viability, IgG secretion, and cell phenotype measurement.

### Transfection efficiency

In order to ensure that target genes of interest were in fact being successfully knocked down, transfection efficiency was measured for each gene, in CD19+ B-cells; a summary of the results can be found in Fig. 6.2.6. Transfection efficiency was measured using the siGLO Green indicator, which localises in the nucleus and can thus indicate whether transfection was successful. Based on this, transfection efficiency was between 20-35% across the different genes (at a concentration of 50nM),

**Figure 6.2.5 *(following page)*: Gating strategy for FACS.** The workflow for FACS gating for day 1 (measuring transfection efficiency) and day 6 (measuring cell phenotype). Day 1 gating had two main streams; the first (top row) focused on transfection efficiency in total leukocyte populations, as measured by siGLOGreen, while the second (second row) focused on transfection efficiency in B-cells specifically. SS (side scatter) and FS (forward scatter) refer to measures used for cell size gating. Day 6 gating was used to measure cell phenotype and consisted of a series of gates; these served to separate out leukocytes (top row), followed by viable B-cells (based on cell size and near IR viability dye; second row), and lastly IgG expressing mature B-cells (based on a CD38+ marker in plasmablasts; bottom row). The text within each plot describes what the gating was done on and what percentage of cells passed the criteria.


Day 1
Singlets
95.2
SS-A::SS
FS-W::FS
CD45+
36.7
Comp-FL10-A::CD45 BV510
Viable CD45+
95.0
FSH::FS
Comp-FL8-A::viab
Transfected
CD45+
40.8
Comp-FL1-A::siGLOGreen
Big gate
78.0
FS-H::FS
Total
CD19+
10.2
Comp-FL5-A::CD19 PECy7
Viable CD19+
94.8
Transfected
CD19+
22.1


Day 6
Singlets
91.7
SSC
FS W
CD45+
26.0
CD45
Viable of all CD45+ cells
96.7
FSC
Viab near IR
Big gate
68.2
FS C
Total CD19+
4.35
CD19
Viable CD19+
88.3
CD27+IgD- memory cells
27.1
CD27
IgD
CD27-IgD+ naive cells
45.4
CD38+ Plasmablasts
56.2
CD38
Q1
56.0
Q2
0.61
Q4
30.3
Q3
13.1
IgG

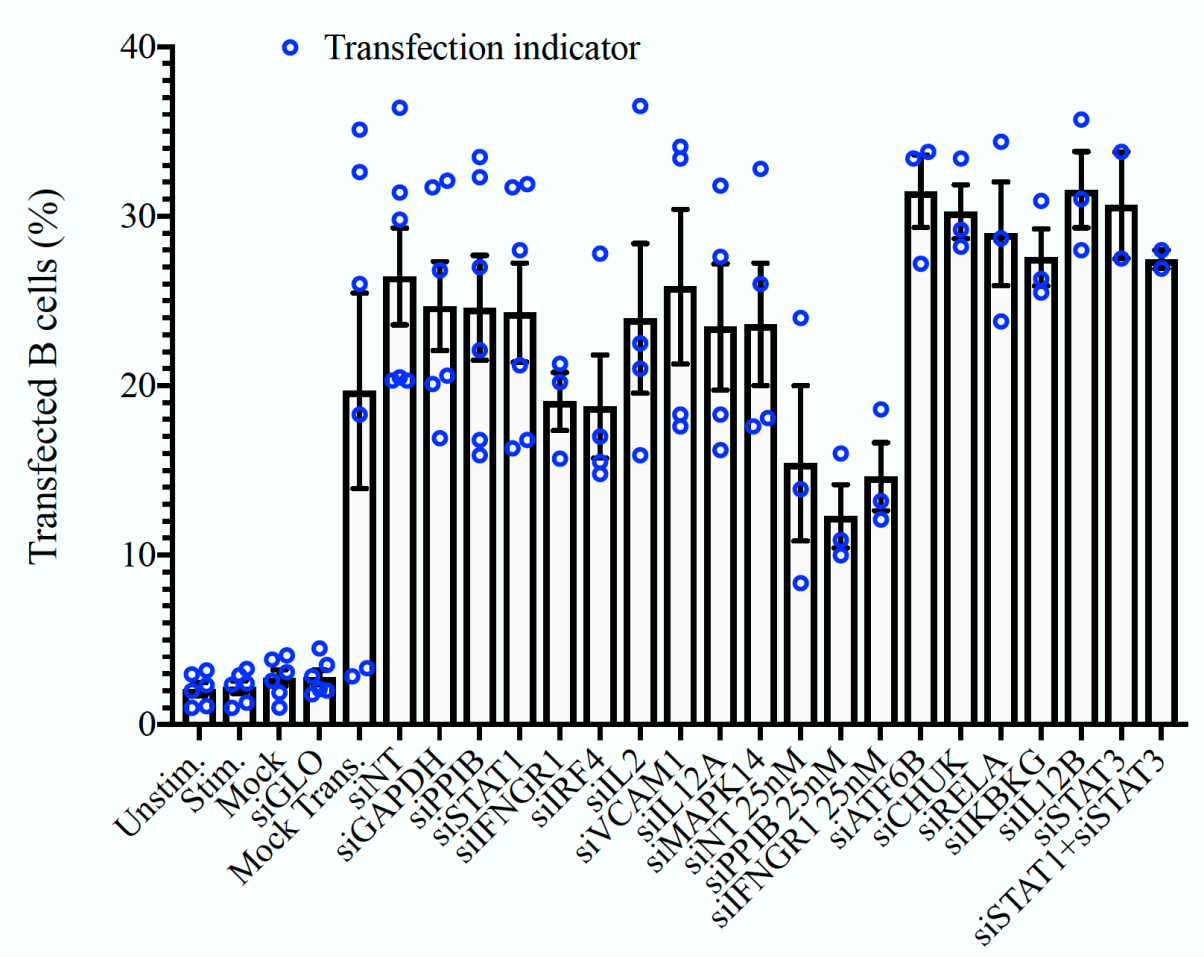



**Figure 6.2.6: siRNA transfection efficiency.** A bar plot showing the transfection efficiency of each targeting siRNA (error bars representing SEM), in CD19+ B-cells, as measured using the siGLO Green indicator. siRNA targets include *GAPDH*, *PPIB*, *STAT1*, *IFNGR1*, *IRF4*, *IL2*, *VCAM1*, *IL12A*, *MAPK14*, *ATF6B*, *CHUK*, *RELA*, *IKBKG*, *IL12B*, and *STAT3*. siRNA concentrations were all 50nM, except in the three bars labelled with 25nM. Additionally, several control groups are shown, including unstimulated, stimulated, mock (only buffer), siGLO (no transfection, no siRNA), and mock transfection conditions. Each dot represent one of 6 donors.

which was low but deemed an acceptable level. The control groups also ensured that the protocol was working as expected, and demonstrated that successful CD19+ B-cell transfection was due to the combined use of siRNA and stimulation cocktail.

### Control genes: assay readouts

In total, 4 control genes (*PPIB, GAPDH, STAT1,* and *STAT3*) were tested in parallel to the main genes of interest; Fig 6.2.7 shows the results of each of these controls. *PPIB* and *GAPDH* served as negative

controls, which is to say that there was no expected effect on cell viability, phenotype, or IgG production following knockdown of these two genes. In contrast, *STAT1* and *STAT3* served as positive controls; these two genes were selected on the basis of being involved in the signaling pathway of the cytokine stimulation cocktail used in the protocol. Thus, knocking down of *STAT1* and *STAT3* was expected to result in a reduction of IgG production, since knockdown would result in interference with B-cell maturation. In general, all four control genes had the predicted effect (Fig 6.2.7); *PPIB* and *GAPDH* did not result in any changes in cell survival, B-cell phenotype, or IgG production, whereas *STAT3* knockdown dramatically reduced IgG production on both days 3 and 6. Interestingly, *STAT1* knockdown had no effect on B-cell maturation or IgG production; this is likely due to compensation by *STAT3*, which has experimentally been shown to be a critical driver of B-cell differentiation, whereas *STAT1* was shown to not have a significant effect on the maturation process [265].

### Target genes: assay readouts

In addition to the 4 control genes, 11 targets of interest were assayed in the siRNA knockdown protocol. Of these 11 genes, 8 showed promising results: *IFNGR1*, *IL2*, *IL12A*, *MAPK14*, *ATF6B*. *CHUK*, *RELA*, and *IKBKG* (Fig 6.2.8). More specifically, knockdown of these genes with targeted siRNA resulted in a significant reduction of IgG secretion on days 3 or 6 (i.e. < 80% secretion compared to non-targeting siRNA) and/or a reduction in the number of matured B-cells (plasmablasts), as indicated by pink columns labeled IgG3, IgG6, and PlasmaB in Fig 6.2.8. Moreover, the results supported an important role for *IFGNR1*, *IL2*, *IL12A*, and *MAPK14*; in each of these instances, knockdown achieved *both* a reduction in IgG secretion *and* an overall reduction in mature B-cell phenotype, suggesting a potential for mediating the effect of disease-driven B-cell development via targeted therapies. In fact, *IL2* has previously been implicated in SLE through studies looking at IL2 production in patient versus control T-cells [266]. Likewise, *IL12A* has been identified as a likely causal gene driving SLE pathogenesis, via both GWAS with integrated functional data (including mouse phenotype annotations and eQTL signals) as well as computational predictive approaches based on machine learning classifiers processing patient versus control genotype data [81, 254].

Importantly, cell viability was not significantly reduced by any of these siRNA gene knockdowns,

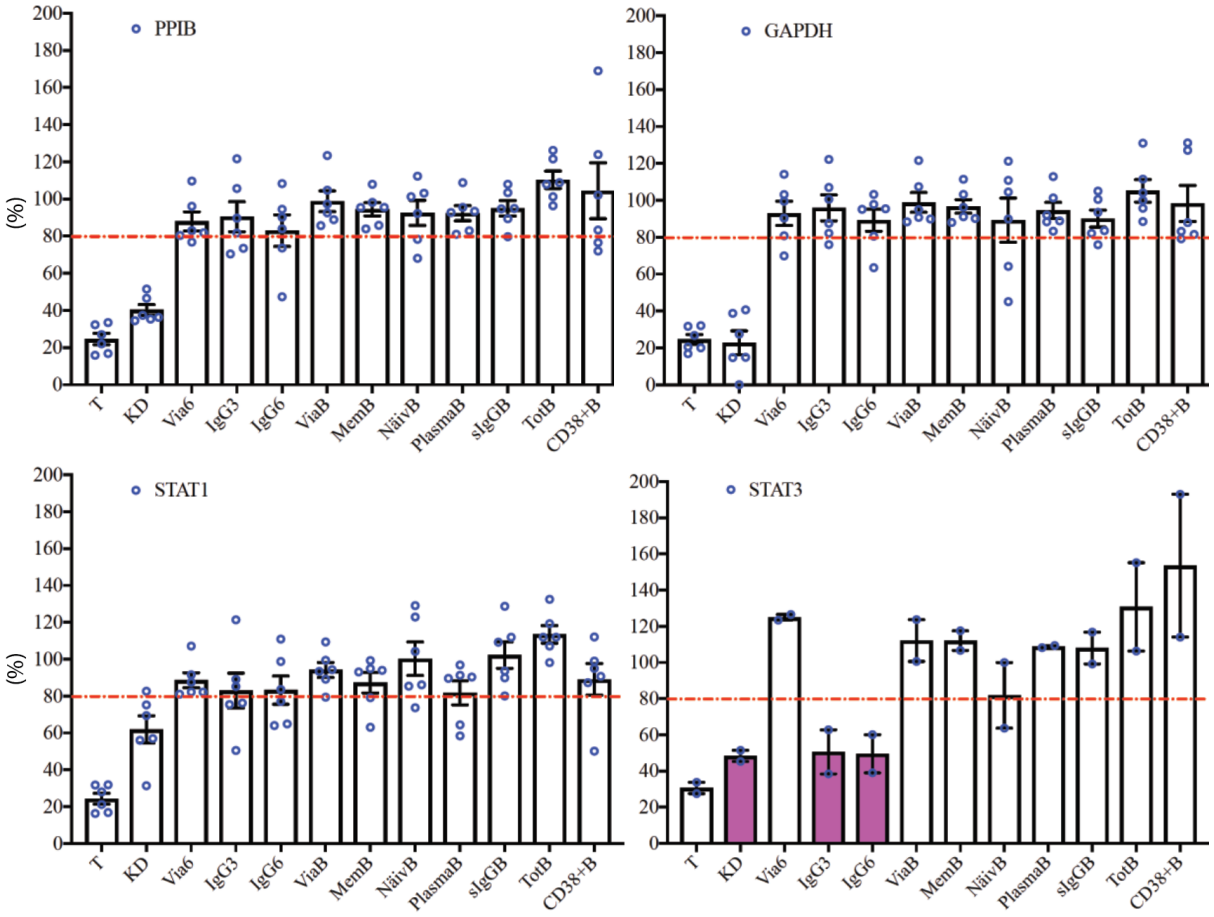


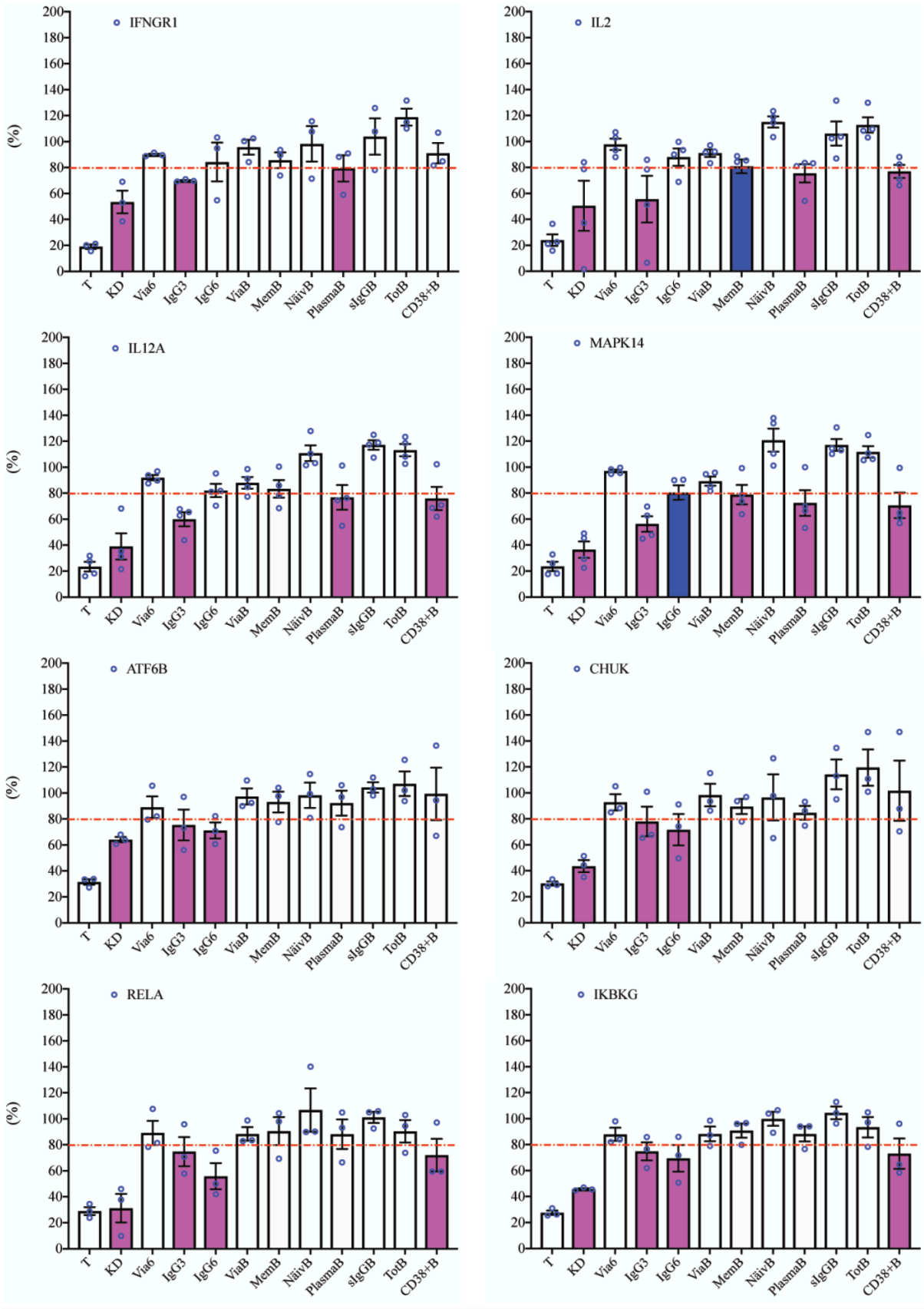

**Figure 6.2.7: Effects of siRNA knockdown on control genes.** Bar plots showing the effects of siRNA gene knockdown on four control genes, *PPIB*, *GAPDH*, *STAT1*, and *STAT3*. The red dotted line indicates threshold for significance and pink-shaded bars indicate a significant effect of the siRNA knockdown; in instances when there was a significant effect of siRNA knockdown, pink-shaded KD bars indicate adequate knockdown efficiency. Error bars represent SEM. From left to right: transfection efficiency (T), knockdown efficiency (KD), day 6 viability (Via6), day 3 IgG secretion (IgG3), day 5 IgG secretion (IgG6), viable CD19+ B-cells (ViaB), CD19+CD27+IgD- memory B-cells (MemB), CD19+CD27-IgD+ naive B-cells (NaivB), CD19+CD27+IgD-CD38+ plasmablasts (PlasmaB), CD19+IgG+ B-cells (sIgGB), total B-cells (TotB), and CD19+CD38+ B-cells (CD38+B). All values were normalised to siNT. Each dot represents one of 6 donors.

which was critical to demonstrate so as to ensure that reduction in mature and IgG-secreting B-cells was not simply due to cell death, but rather due to inhibition of the cell maturation process itself. On the other hand, while *RELA* and *IKBKG* had effects on both IgG secretion as well as B-cell phenotype, they also resulted in reduced cell viability barely above the acceptable threshold of 80% and so could be exerting part of their effects via cell death. Fig 6.2.10 shows just the cell viability plot across all tested genes, for ease of comparison. In contrast, *IRF4*, *VCAM1*, and *IL12B* did not result in any major phenotypic changes (Fig 6.2.8); the only exception is *IRF4*, which had an effect on day 3 IgG secretion, but no effect on day 6 or on mature B-cell populations.

## 6.3 Discussion

The work outlined in this chapter showcases the practical utility of using Pi prioritisation in the context of pre-clinical drug discovery and demonstrates how one could use Pi, in tandem with disease-relevant assays, to nominate promising targets for further investigation. The first section - computational prioritisation of genes using Pi - identified a number of candidate targets, including many of the well-established genes in the context of SLE. This, coupled with the pathway-level analyses that were significantly enriched for disease-relevant processes, demonstrated the utility of using such a multifaceted data set aggregation tool.

**Figure 6.2.8 *(following page)*: Results of siRNA knockdown of promising genes.** Bar plots showing the effects of siRNA gene knockdown on eight target genes, *IFNGR1*, *IL2*, *IL12A*, *MAPK14*, *ATF6B*. *CHUK*, *RELA*, and *IKBKG*. The red dotted line indicates threshold for significance, and pink-shaded bars indicate a significant effect of the siRNA knockdown, and blue-shaded bars indicate a near-significant finding (at threshold); error bars represent SEM. From left to right: transfection efficiency (T), knockdown efficiency (KD), day 6 viability (Via6), day 3 IgG secretion (IgG3), day 5 IgG secretion (IgG6), viable CD19+ B-cells (ViaB), CD19+CD27+IgD- memory B-cells (MemB), CD19+CD27-IgD+ naive B-cells (NaivB), CD19+CD27+IgD-CD38+ plasmablasts (PlasmaB), CD19+IgG+ B-cells (sIgGB), total B-cells (TotB), and CD19+CD38+ B-cells (CD38+B). All values were normalised to siNT. Each dot represents one of 6 donors.


IFNGR1
IL2
IL12A
MAPK14
ATF6B
CHUK
RELA
IKBKG
(%)
T
KD
Via6
IgG3
IgG6
ViaB
MemB
NäivB
PlasmaB
sIgGB
TotB
CD38+B

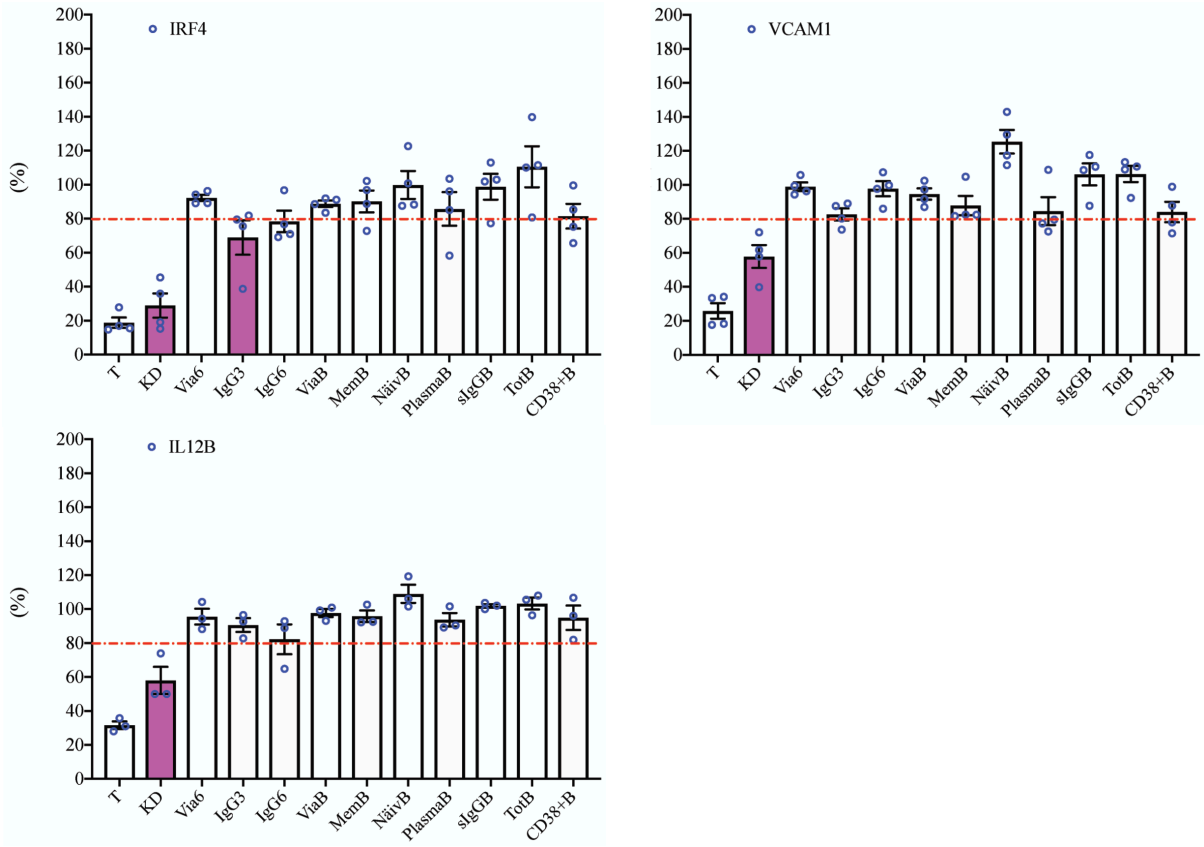



**Figure 6.2.9: Results of siRNA knockdown of unpromising genes.** Bar plots showing the effects of siRNA gene knockdown on three target genes, *IRF4*, *VCAM1*, and *IL12B*. The red dotted line indicates threshold for significance, and pink-shaded bars indicate a significant effect of the siRNA knockdown; error bars represent SEM. From left to right: transfection efficiency (T), knockdown efficiency (KD), day 6 viability (Via6), day 3 IgG secretion (IgG3), day 5 IgG secretion (IgG6), viable CD19+ B-cells (ViaB), CD19+CD27+IgD- memory B-cells (MemB), CD19+CD27-IgD+ naive B-cells (NaivB), CD19+CD27+IgD-CD38+ plasmablasts (PlasmaB), CD19+IgG+ B-cells (sIgGB), total B-cells (TotB), and CD19+CD38+ B-cells (CD38+B). All values were normalised to siNT. Each dot represents one of 6 donors.

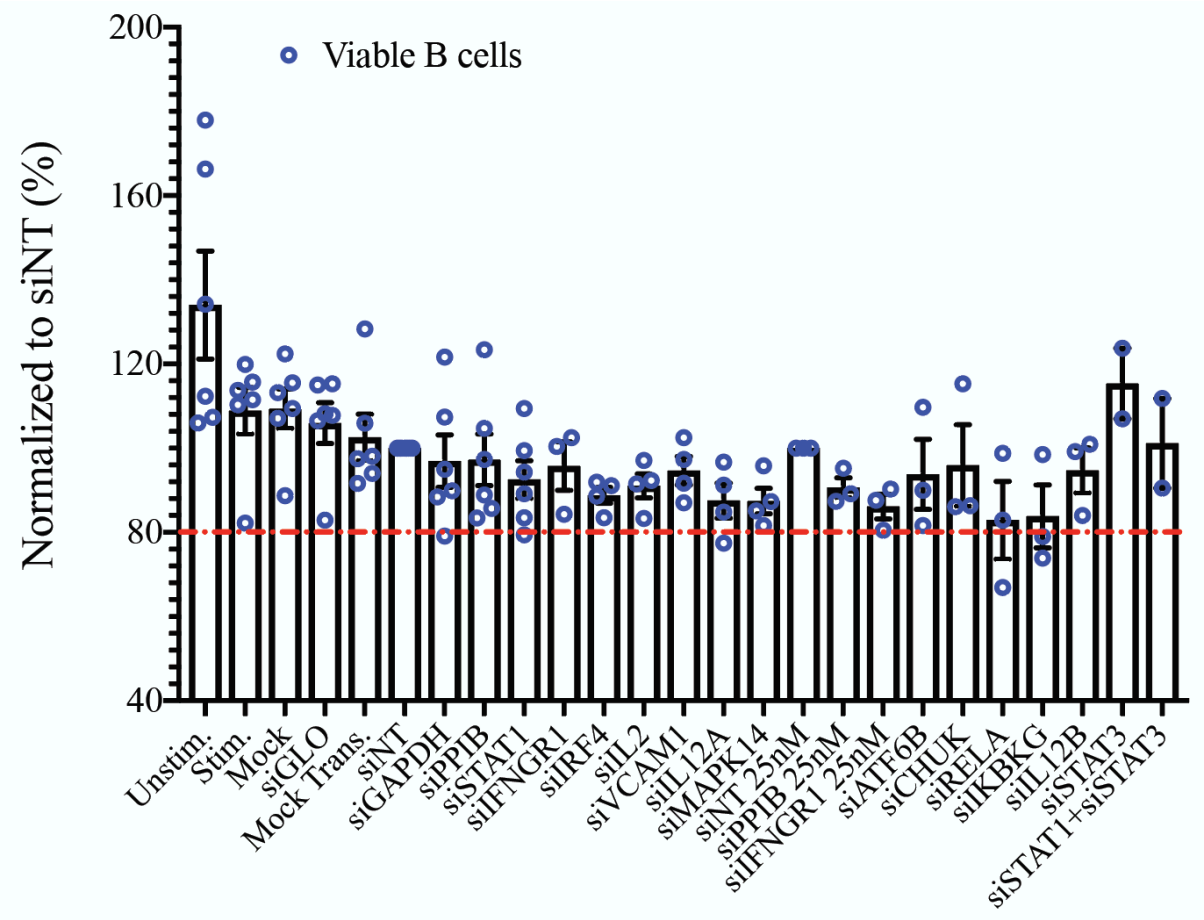



**Figure 6.2.10: Post-siRNA cell viability readout.** The cell viability readout on day 6 of the assay, as measured by CellTiter Glo. Results are normalised to the siNT, and the red dotted line indicates the threshold for a significant effect.

### Epigenetic probe screen

The epigenetic probe screen highlighted the informativeness of Pi in ranking targets of interest relative to one another. Specifically, the results of probe effectiveness at inhibiting disease phenotype correlated strongly and significantly with the relative ranking of probe targets in the Pi output. In other words, the prioritisation output did a good job at predicting which targets would perform better or worse in a disease-relevant context. The implications of this for use in drug discovery are exciting; one could imagine instances where an academic or pharmaceutical lab are faced with dozens (or, in some cases, hundreds) of potential compounds that need to be validated experimentally before they can transition to clinical trials. These experiments are often time and resource expensive, and so having a reliable, scalable, and adaptable method of prioritising or de-prioritising targets *before* they are even taken into pre-clinical testing would have positive effects on overall resource expenditure. The results presented in this chapter suggest that Pi could be used in such a manner, effectively narrowing down the candidate set or solution space to a more manageable number, which can then be validated experimentally.

### Prioritisation of genes for experimental validation

Accordingly, the second portion of this chapter focused on how one might go about exploring novel targets, from computational prioritisation to experimental validation. First, I outlined a set of possible steps one could take in order to move from a prioritised set of around 150 disease-relevant genes to a more manageable number of roughly 10 targets for wet lab validation. I focused on four major areas with this funnel approach: network crosstalk, multi-disease relevance, druggable pockets, and compatibility with experimental protocol. For instance, genes that were interaction partners across multiple, disease-relevant and cross-talking pathways received higher priority than those that were not, since they likely play an important role in driving disease. Additionally, targets that were relevant to multiple, related diseases, and ones that had known druggable pockets were favoured over others. Lastly, targets that were likely to (unintentionally) interfere with the experimental protocol itself were removed, as positive results from such targets could be due to confounding factors and not target

effectiveness itself.

### siRNA knockdown effects on B-cell maturation

Lastly, this chapter demonstrated the critical step of early validation of any computationally-prioritised targets, through the use of a disease-relevant assay. This entailed testing the effects of siRNA gene knockdown in a B-cell stimulation protocol, to see whether targeting genes of interest would have any effect on inhibiting the process of B-cell maturation, which is a key component of SLE disease progression. The protocol had to be carefully thought out and designed in such a way as to ensure that any effects/outputs were true results and not arising due to confounding variables. This meant ensuring a number of things, including that: none of the individual steps in the protocol were causing inhibition of disease phenotype (e.g. the reagents themselves, the Neon pipette, etc.); the readout was successful when known stimulation pathway targets were knocked down (i.e. positive controls) and unsuccessful when unrelated targets were knocked down (i.e. negative controls); the gene of interest could actually be knocked down (i.e. transfection and knockdown efficiencies); the results of genes of interest could be compared across experiments (i.e. were normalised to a non-targeting control); and the effects of gene knockdown on cell phenotype were not simply due to cell death but actually due to interference with the underlying mechanism (i.e. passable cell viability measures). Additionally, targets that affected both cell phenotype (B-cell maturation) as well as IgG secretion were favoured over those that only had an effect in one or the other.

Ultimately, knockdown of *IL2*, *IL12A*, *IFNGR1*, and *MAPK14* resulted in the most interesting results, based on the above-mentioned criteria, which was interesting for a number of reasons. First, *IL2*, *IL12A*, and *IFNGR1* were identified in one, directly interacting cluster in the pathway crosstalk subnetwork analysis presented in Fig. 6.2.2, which strengthens the case for a significant, interlinking role of those genes in mechanism underlying disease. For instance, while early studies have well established the role of IL12 in immune response via Th1 induction [267], a more nuanced picture implicating dynamic interactions between IL12 and IFNGamma later emerged [260, 268]. Thus, it is likely the case that complex immune diseases like SLE will require multi-target or pathway-level therapeutic modulation. In fact, the IL12 and IL23/Th17 axis has become increasingly interesting as a

therapeutic entry point into SLE treatment [260]. Ustekinumab, a monoclonal antibody targeting the IL12/IL23 axis and used in the treatment of psoriasis, has actually recently been investigated for the treatment of SLE; a double blind, Phase 2, randomised, controlled study had positive findings and established efficacy for use of ustekinumab in SLE patients, and recently transitioned from Phase 2 to Phase 3 trials [269].

It is worth noting that, unlike *IL12A*, siRNA knockdown of *IL12B* had no effect on cell phenotype (Fig 6.2.9), indicating that more work is needed in order to understand the intricacies of how IL12 response is mediating B-cell maturation. Past studies have shown that IL12B has a dual role in the IL12 pathway, serving not only as a subunit of IL12, but also as an inhibitor on its own that regulates the signalling response [270, 271]. Therefore, it could be this IL12B balancing role that complicates the output when it is selectively knocked down, resulting in a lack of a positive result.

As mentioned, the role of IL12 in autoimmunity via modulation of the IFNGamma pathway has been debated and investigated and, by extension, the role of IFNGR1. The role of the receptor in complex immune system tuning has been studied extensively in mice; deletion of IFNGR1 in lupus-prone mice, for example, results in better survival and prognosis, as well as reduced disease-related renal complications [272]. Additionally, other studies have investigated IFNGamma/IFNGR1 in human PBMCs, and found that SLE patient peripheral blood T-cells produce excess IFNGamma, which contributes to SLE parthenogenesis through induction of soluble B lymphocyte stimulator to promote B-cell activation and maturation [273]. More recently, a 2016 study used a chimeric model of B-cell-driven autoimmunity to demonstrate that B-cell IFNGR1 signaling accelerated lupus disease development [274]. Therefore, the readout of the siRNA assay coupled with results from past studies seem to support the potential for therapeutic tractability of IL12/IFNGR1 inhibition in SLE treatment.

The results from *IL2* knock down are a little more complicated to understand/interpret, given the role of IL2 in the balance between immunity and tolerance [275]. Work has been done to understand the relevance of IL2 to autoimmunity, and studies have demonstrated marked differences in IL2 between patients and controls; for instance, T-cells from patients with SLE show significant reduction in IL2 production when compared to normal, healthy T-cells, which seem to be driven by increased amount

of transcriptional regulators [266]. Moreover, there is clinical evidence to support low-dose IL2 treatment for use with SLE patients, which is currently being investigated in Phase 2 clinical trials [276].

### Indirect target identification

Another approach for informing future work is to incorporate the knockdown results into a broader-level network view. More specifically, given that three of the most significant siRNA knockdown targets (*IL12A*, *IL2*, and *IFNGR1*) were identified in the subnetwork analysis, it would be interesting to investigate downstream nodes that are likely involved in the disease-relevant signaling processes. For instance, *TYK2* and *PIK3R1* are both nodes directly downstream of the *IL12A*, *IL2*, and *IFNGR1* cluster; *TYK2*, a JAK family member that is involved in the propagation of cytokine signaling from multiple sources, has recently been implicated in SLE by a study that proposed a role for the gene in generating specialized T-cells relevant to the pathogenesis of the disease [277]. On the other hand, mutations in *PIK3R1* have been linked to human immunodeficiency and, more recently, evidence has emerged showing that a mutation in *PIK3CD* (which encodes a catalytic subunit of Class IA PI3Ks; *PIK3R1* encodes one of the regulatory subunits) caused pediatric SLE [278, 279]. Therefore, it is possible that *PIK3R1*, which is a central node in the Pi-prioritised SLE output and which serves as a signal convergence point downstream of promising siRNA knockdown targets, may of therapeutic interest in this disease context. Moreover, both *PIK3R1* and *TYK2* are near the HLA cluster of genes (Fig. 6.2.2), a region that is known to be associated with SLE, such as in the case of HLA-DRB1, which is one of the nodes within this cluster [257, 280, 281]. In fact, *TYK2* is currently being investigated in a Phase 2 clinical trial for the treatment of SLE [282]. The randomised, double blind, placebo-controlled, global study (Clinicaltrials.gov NCT03252587) is currently recruiting patients and will use a selective inhibitor, BMS-986165, targeted against *TYK2*, to block SLE-specific cytokine signaling. Thus, by leveraging the combination of genetics-driven Pi prioritisation, assay readout of selective siRNA knockdown, and knowledge of network signaling and crosstalk, it becomes possible to identify indirect targets whose role in disease is supported by multiple pieces of evidence.

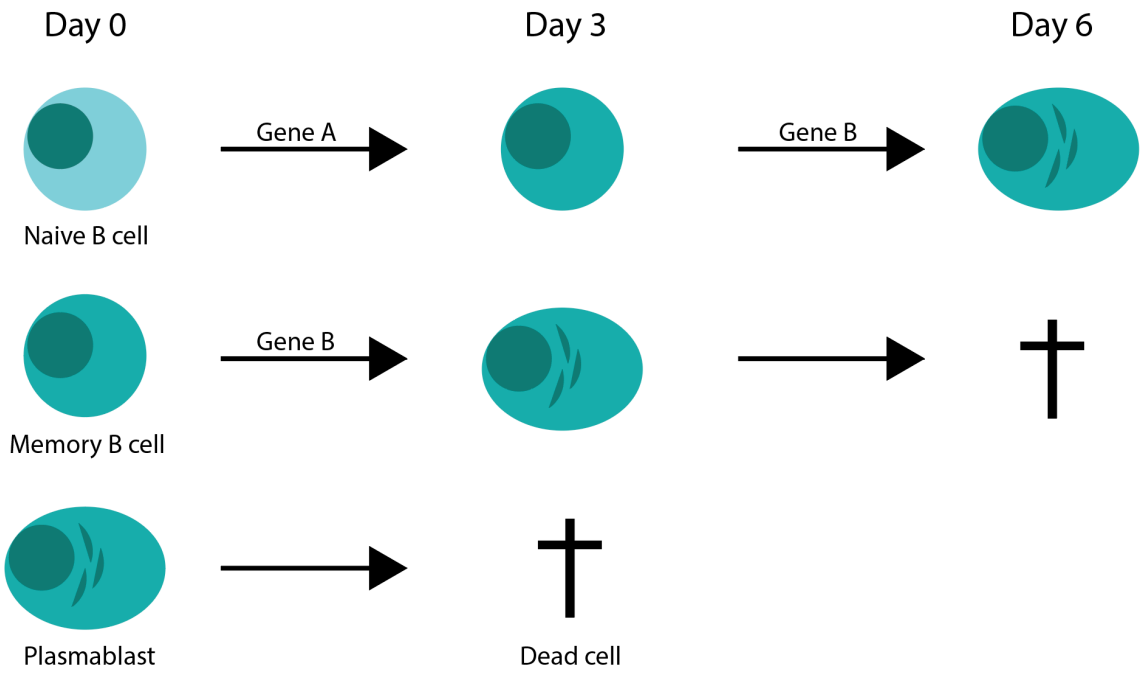



**Figure 6.3.1: Schematic of B-cell maturation.** An overview of the main stages of B-cell maturation, over the course of the assay protocol, with respect to day 0, 3, and 6 time points. *Gene A* and *Gene B* may or may not be the same gene.

### Limitations

It's critical to note some of the main caveats to the findings presented in this chapter. First, the particular assay used presents one way to investigate B-cell response to targeted siRNA, based on a cytokine stimulation protocol. There are, of course, numerous other ways to investigate B-cell signaling (e.g. through the B-cell receptor) which may give a different picture of what a particular siRNA effect may be. Therefore, it may be prudent to investigate the genes of interest in other, complementary B-cell assays in order to get a more comprehensive understanding. Moreover, the assay itself is capturing a 'snapshot' of a number of stages of B-cell maturation at any given point in time (Fig 6.3.1). Specifically, not all B-cells are necessarily at the same stage of maturation on a given day, since this is a function of what stage they were in to begin with. Therefore, more nuance may be required in order to better map sub-populations of B-cells over the time course of the protocol, particularly when it comes to interpreting whether a gene knock down resulted in a 'favourable' outcome, and where in the maturation pathway that gene may be acting.

Further, since the assay is done with PBMCs (and not on isolated B-cells), more work is needed in

order to truly understand how a certain target is exerting its effect on B-cell maturation. For example, it may be the case that knocking down a particular gene affects B-cell maturation indirectly via an effect on T-cells, for instance, something that would be important to understand from a clinical standpoint. In this particular example, something that could be done would be to isolate T-cells from donor blood, complete the transfection protocol on the isolated cells, and then add these back into the full PBMC sample. This way, any significant differences between siRNA and control groups on B-cell phenotype would likely be due to effects modulated by T-cells.

With regard to calling significance, it's important to note that cutoffs for calling significant effects on cell phenotype or IgG production were fairly arbitrary; the 80% threshold used seemed appropriate, and was a good cutoff to differentiate between the positive and negative controls (i.e. negative controls *PPIB* and *GAPDH* were > 80% while positive controls *STAT1* and *STAT3* were at or below 80%). However, this decision to use 80% can be interrogated in greater detail, which may be beneficial for future assay result stratification, particularly when greater and greater numbers of genes of interest are considered.

Lastly, although each candidate gene was knocked down in 3-4 individuals (replicates), there were cases where knock down effects were quite variable between individuals, which is to be expected when dealing with measuring complex immune response between individuals. Nevertheless, increasing the number of replicates would be favourable, particularly for the genes that seemed most promising following this first set of siRNA experiments.

## 6.4 Conclusion

In conclusion, these findings present one possible approach of pre-clinical target assessment and screening. By moving from GWAS-identified disease associations to prioritised genes and pathways via multi-faceted computational tools like Pi, it becomes possible to funnel hundreds or thousands of potential targets to an order-of-magnitude more manageable number of putative genes, backed by multiple layers of human genomic evidence and cell-type relevance to SLE. In turn, this presents an opportunity for experimental validation of targets of interest, through the use of disease-relevant

cellular assays and siRNA gene knockdown experiments. By capturing key processes like B-cell maturation in these assays, genes can be evaluated on the spectrum of promising - non-promising targets based on their effects on cell phenotype. This process identified *IL2*, *IL12A*, *IFNGR1*, and *MAPK14* as the most promising targets, based on the significant reduction in IgG secretion as well as inhibited B-cell maturation in patients cells where these targets were knocked down via siRNA. Moreover, *IL2*, *IL12A*, and *IFNGR1* were identified as a cluster of genes in an SLE-specific subnetwork crosstalk analysis of Pi-prioritised targets, further supporting the continued study of these genes and pathways for potentially advancing SLE therapeutics. Taken together, this chapter highlights an exciting approach to efficiently inform the therapeutic process for complex, immune-related diseases that require better treatment options, in a scalable way.

*In one case out of a hundred a point is excessively discussed because it is obscure; in the ninety-nine remaining it is obscure because it is excessively discussed.*

Edgar Allen Poe

# 7 Discussion

While numerous challenges do remain, the work summarized in the chapters of this thesis has demonstrated how certain approaches can be used to inform the therapeutic pipeline for immune-related diseases. Specifically, the main results sections outlined:

I. A novel prioritisation pipeline, Priority Index [64], including how it can be leveraged to prioritise drug targets at the gene and pathway level and how it can be computationally validated in a robust and scalable way;

II. The application of Priority Index to a novel disease case, myasthenia gravis, in order to demonstrate the utility of using such an approach in early, pre-clinical stages of drug development;

III. Experimental validation of highly networked, prioritised epigenetic targets in Dupuytren's disease, focusing on broad gene and protein-level changes, as they pertain to disease progression;

IV. Experimental validation of prioritised targets in systemic lupus erythematosus, a disease with more established genomic underpinnings, demonstrating the utility of the computational approach in de-risking single-target genes for therapeutic consideration.

Below, I discuss the key takeaways from each chapter, including (broadly) the major findings and limitations of each section. Additionally, I outline avenues for future work that can serve to both broaden and deepen the findings presented thus far; some of this work is already underway and will hopefully translate to publishable findings in the near future; other work is more long-term, but can nevertheless be used to inform future experiments or projects to be taken on.

## 7.1 Computational prioritisation of genes

The first results chapter focused on Priority Index [64], or Pi, and how this computational tool provides an exciting opportunity for the integration of various genetic-based predictors and network information in the context of novel drug target discovery. Namely, I highlighted how this robust pipeline can be used to prioritise genes based on a series of genomic data, incorporating GWAS, significance thresholds, physical gene promoter interaction, gene expression correlations, cell type specificity, gene annotations, and global gene protein networks [87–96]. While some tools do exist for achieving certain aspects of this such as SMR[50], which combines GWAS and eQTL for integrative analysis, or Open Targets[58] which leverages human data, animal data, text mining, and machine learning to prioritise targets, Pi has a number of advantages over these. For instance, unlike the SMR method, Pi is not limited to a convergence of eQTL and GWAS. Rather, genes are scored independently across a range of predictors, which is then extrapolated with network approaches, and finally combined for an aggregate score; this way, a lack of one type of predictor data can be compensated for by other predictors or by indirect network effects. Further, Pi's discovery mode allows for completely human genetics-led prediction without model building (unlike Open Targets), resulting in more unbiased putative target prioritisation.

However, validation remains an integral part of the development of any novel computational approach; while experimental approaches do exist for validating certain aspects of the pipeline, it

remains a challenge to do so at rapidly and at scale. Therefore, a core focus of the first chapter was centred around defining Gold Standard Positives (GSPs) and Negatives (GSNs) in a scalable way; multiple sclerosis was used as a case study throughout, in order to demonstrate how such an approach could be practically applied. The quality of the Pi algorithm was measured through a series of AUC tests in order to demonstrate how well GSPs were separated from GSNs; importantly, a number of different definitions of GSPs and GSNs were highlighted, in order to robustly demonstrate the performance of Pi across a range of scenarios and definitions. For instance, AUC values actually increased when the definition of GSPs was made stricter, to include only clinical golden standard MS treatments. On the other hand, the use of alternate forms of GSN definition (from simulated targets, as initially proposed, to approved targets in other indications) did not negatively impact AUC values, indicating that the pipeline is likely to perform well under a range of forms of scrutiny. Further, the quality of target prioritisation was benchmarked against Open Targets [58], which is perhaps the closest existing tool, and Pi was shown to outperform this tool, when it comes to a purely genetics-led approach (i.e. 'discovery mode').

Additionally, a 'supervised mode' was developed as an alternate form of analysis, with the primary differentiating point being that supervised mode leverages machine learning in order to translate knowledge of existing therapeutics to the gene-level scoring and ranking system. The supervised mode outputs performed better when assessed under the criteria used for discovery mode, which was expected given that information on GSPs and GSNs was used in the prioritisation pipeline itself. Further, the chapter outlined key similarities and differences between the two prioritisation modes in terms of the gene and pathway outputs themselves, demonstrating a particular opportunity for drug repositioning via the supervised mode.

Naturally, there were a number of limitations to the above mentioned findings. First, the Pi pipeline itself depends on available, disease-specific data, and its utility grows as more data is accessible for integration. This factors in across a number of levels: GWAS information, for instance, serves as the initial input into the Pi pipeline, so the quality of output will always be constrained by the quality of this input. Further, the utility of eQTL data-sets, for example, is in part dictated by the underpinnings of the disease in question, since these data are limited in some cell types. In an immune context, a

pipeline like Pi can discriminate well within immune cell types, since there are numerous eQTL studies focusing on these. However, certain tissues have very limited available data at scale, so the use of Pi will remain constrained in these scenarios, at least until such a time as when more data emerges in the public domain. Second, while the Pi output does provide relative target ranking, allowing for prioritisation between targets of interest, it does not necessarily capture direction of effect. For example, a target like TNF would rank highly in both rheumatoid arthritis and in multiple sclerosis outputs, but the effects of anti-TNFs are very different in these two instances [153]. This limitation is perhaps a little trickier to tackle given that direction of effect is not only a function of disease state, but also one of cell type; at the moment, the Pi output is at the disease-level, not the cell-type level, so providing directionality is not a straightforward task. Lastly, there are limitations around the generation of GSPs and GSNs; namely, while GSN simulation was shown to be a good option (particularly when it comes to deployment at scale, for the purposes of performance assessment), this does not capture 'true negatives' in the way clinical golden standards can capture true GSPs. This problem is a complex one to solve given the lack of data transparency around pharmaceutical studies and drug trials, particularly when it comes to failed compounds (either intentionally, as a way of preserving competitive advantage by not disclosing failed results to other companies or, less cynically, due to general disinterest in spending the time to publish negative results). It is the hope of various public-private partnerships that this sort of data will come to light more quickly [5, 62], which will hopefully translate to more robust definitions of GSPs and GSNs in the future.

There are a number of ways that the work outlined in this chapter can be taken forward. From a high level perspective, there needs to be continued support for and curation of high quality GWAS data – with sequencing costs drastically reducing over time, this type of data will continue to improve the outputs generated by approaches like Pi. Similarly, as the methods and techniques surrounding eQTL data improve, so too will that feature of Pi outputs; the pipeline itself is set up in a way that will allow for constant iteration and improvement, both within immune-related diseases and beyond, which is a marked strength in the Pi approach. Further, as these data-sets become more widespread and integrated into the pipeline, there is an emerging opportunity to improve the segmentation of Pi results by cell type by disease; in other words, there will be an opportunity to improve Pi's prioritisation by altering the weighting of information fed into Pi's network prioritisation on the basis

of cell type relevance which, at the moment, is binary (i.e. include cell type or do not include). Nuancing this on the basis of new information can provide a more informed gene prioritisation output in the future. From the perspective of Pi's supervised mode, continued partnerships with industry will be an important way forward; in particular, more transparent curation of failed drug trials, with clearer understanding of root cause of failure (perhaps even a database of failed trials which includes drug, target, indication, reason for failure). This way, the training of supervised mode models will depend less on simulated negatives for a particular disease and more on 'clinical negatives' which will be useful in further refining the outputs.

As discussed in the results chapter, the outputs of supervised mode can also be leveraged for the purposes of drug discovery – in particular, it would be interesting to combine gene expression analysis, drug signature data, and Priority Index for a systematized method of identifying drug repositioning opportunities. More specifically, the Pi supervised mode output could be utilised to identify highly ranked genes for a particular disease which are simultaneously targets of approved therapeutics in a separate indication; from here, the gene expression signature of the drug in question could be contrasted to the gene expression signature of patients versus controls in the novel disease context, in order to identify opportunities where therapeutic intervention could reverse changes brought about by disease state. Lastly, there is an exciting opportunity for a more systematic integration of discovery and supervised mode – in essence, a 'dual mode' to enable the aggregation of results of both runs. This way, it becomes possible to uncover the intersect of genetically novel, high confidence putative targets with those containing feature vector similarity to existing, effective therapeutics in diseases of interest.

## 7.2 Application of the prioritisation pipeline to myasthenia gravis

The second results chapter outlined how the Priority Index computational tool could be used in a novel disease context, myasthenia gravis, where early GWAS have identified potential genes of interest [68–70] but where the underlying mechanism and the corresponding therapeutic potential remains largely under-explored. The work demonstrated how Priority Index could be used to aggregate a number of existing genetic and genomic data-sets in order to prioritise a set of genes; these genes were then analysed downstream of Pi prioritisation at the gene level, pathway level, and network level, in

order to uncover possible mechanistic drivers of disease. For instance, there was a significant enrichment of T-cell relevant genes in the high priority list, consistent with past studies that have suggested a role for these in disease [160]. Further, a subnetwork cross-talk analysis identified a number of nodes that are FOXP3-bound in activated T-regulatory cells [194], giving even further granularity to the T-cell relevance in myasthenia gravis. Importantly, the overall priority list was supported by external data in that high priority genes from the Pi output captured genes that were previously reported as drivers of disease activity [169], which served as one form of validation of this approach in this disease context.

Moreover, given the differences in genetic underpinnings of early versus late onset myasthenia gravis [68–70], it was possible to segment analyses in such a way as to better understand the pathogenic implications of those genetic differences, via Pi and downstream tests. For instance, while interleukin 4 and 13 signalling was prevalent in both forms of disease (based on prioritised genes using the segmented sets of GWAS data), early onset disease genes showed a significant enrichment for EBV-associated targets [187] as well as for important TLR4 regulators during TNFa activation. These findings suggest that the underlying drivers of the two forms of disease may be markedly different, which would carry important implications around the therapeutic approaches for the two patient groups.

Lastly, given the T-cell relevance uncovered by the Priority Index analysis, there was an opportunity to dive into those genetic drivers as well. In particular, the CD4+ disease relevance finding was interrogated via statistical fine-mapping, informed by functional, cell-type specific data [110], which uncovered two candidate causal SNPs in the CTLA-4 locus. These two SNPs were analysed with respect to correlation to gene expression in a CD4+ eQTL dataset [90], and were found to be statistically significantly associated with changes in *CTLA4* gene expression, suggesting that they likely play a role in the sequence of events that may lead to disease progression.

While the results presented were undoubtedly exciting, it is important to note some of the key limitations to the work outlined. First and foremost, the amount of available GWAS data for myasthenia gravis remains limited; there were three major studies from which SNPs and p-values were sourced [68–70]. From here, only 11 SNPs reached genome-wide significance in total, making it

more difficult to infer which genes could be strongly implicated in disease on the basis of GWAS significance. It would therefore be important to expand the body of GWAS work in this context, before re-running the analyses to see whether outputs change significantly or not. Similarly, the limitation of existing eQTL data remains a limiting factor in the approaches outlined in this chapter. For example, while it was possible to investigate fine-mapped SNP gene expression correlations in CD4+ cells, it would be favourable to be able to do so in T-regulatory cells more specifically. While some databases do exist with this kind of information captured [202], many of these studies are limited in the number of samples used in the eQTL analyses, thus limiting their ability to identify significant correlations between gene expression and SNPs.

More broadly, the work in this results chapter focused on aggregating complementary but disjointed sources of data in order to provide structure to a myasthenia gravis-specific downstream analysis of GWAS information; before firm conclusions can be reached on which targets to potentially take into clinical trials, however, the results would need to be systematically tested and validated in a series of unified experiments with specific therapeutic hypotheses, taking into account targets of interest and specific cell types, which could then be tested in disease models. Expanding on this, it could be very interesting to structure a set of experiments that begin by systematically canvassing the myasthenia-relevant network cross-talk genes against targets of clinical phased drugs in other indications, as outlined in the drug repositioning section of the *Computational prioritisation of genes* chapter. Next, these targets could be investigated in the context of T-regulatory cell gene expression, given the findings presented here on the relevance of this cell type in disease. Genes thus identified, with a link between myasthenia-specific GWAS data, Pi priority, and cell-type specific relevance, could then be tested in existing models of disease [203], either via siRNA knockdown (for a broader testing of many targets) or via administration of repositioned drugs to the animal disease models.

## 7.3 Experimental validation of computationally prioritised targets

The final two results chapters focused on taking target prioritisation a step beyond computational analysis, into lab where effects of modulation could be tested. The first of the chapters demonstrated how specific targets could be validated from an -omic perspective, in the context of Dupuytren's

disease. The newly published Dupuytren's GWAS [71] presented a great opportunity for Pi analysis; once genes were prioritised and examined from a network level, two targets, *CREBBP* and *EP300*, stood out as particularly interesting, due to their positioning within cross-talking pathways and relevance to fibrotic disease [229]. Thus, targeting these two genes (via the CBP30 probe) allowed for an analysis into the therapeutic potential of inhibiting the disease-relevant epigenetic modulators. Moreover, access to patient-derived tissue meant that the disease context could be preserved throughout, and that two things could be simultaneously achieved: an understanding of gene expression changes between probe treatment versus no probe treatment, in disease tissue, could be established (via RNA-seq testing, in relevant cell types), and this expression profile could be contrasted to that of donor-matched healthy versus disease cell RNA-seq. Thus, not only did the results identify a set of genes whose probe-induced expression change signature captured key fibrotic pathways, but also that a very similar set of genes/pathways was affected when comparing untreated, healthy versus diseased tissue. Further, a proteomic analysis of probe treatment confirmed the RNA-seq findings and helped further implicate targets downstream of *CREBBP* and *EP300*, such as Collagen Type VI, which has been investigated in other fibrotic processes [251].

The final results chapter presented a complementary, target-specific approach to validation, this time in the context of systemic lupus erythematosus. Initially, a similar Priority Index analysis was done as outlined above, where genes were prioritised on the basis of GWAS summary statistics and other genetic data. Top priority genes were then further assessed across a number of other relevant metrics, such as their presence in multiple disease-relevant pathways, their association across other immune-related Pi outputs/traits, and their protein structure from the perspective of druggable pockets [151]. Thus, the gene list was reduced to a set of 11 targets, which could then be addressed individually. Doing so involved a series of siRNA experiments, conducted on patient-derived cells, in a B-cell assay meant to emulate the maturation of B-cells in response to a diseased state like systemic lupus erythematosus and investigate the effects of knocking down individual targets of interest. The results presented demonstrated how it is possible to take large, genome-scale data, process it via a computational pipeline like Pi to identify targets of interest, further filter this list down to an experimentally-manageable number, and demonstrate disease-relevant utility of inhibiting these targets in a patient-derived cellular assay. Moreover, the outputs of the knockdown experiments

(which identified *IL2*, *IL12A*, *IFNGR1*, and *MAPK14* as the genes with the greatest effect on B-cell maturation, as measured by plasmablast numbers and IgG secretion) were interesting in that they were supported by evidence from previous studies [81, 254, 266], including clinical trial data, as in the case of ustekinumab (targeting the IL12/IL23 axis) which recently progressed from Phase 2 to Phase 3 trials in lupus patients [269].

From a therapeutic perspective, a few challenges exist before the work presented in these chapters can be taken forward. One such challenge is the nature of the tested targets themselves – *CREBBP* and *EP300* are both epigenetic modulators that affect a number of downstream genes; thus, it is important to understand what the implications of inhibiting such targets may have on overall cell function and health. It may be the case that processes modulated by these two genes are central to cell function and survival, such that it is not feasible to inhibit those two targets directly, from a treatment perspective, as it may lead to inadvertent side effects. Further, while the B-cell maturation assay in the case of the lupus experiments was useful in validating the role of putative targets in B-cell development in patient-derived tissue, it is important to note that any assay captures only a part of the overall picture/biology of disease. Accordingly, any interesting results from those experiments would ideally need to be analysed in complementary assays as well as tested in animal models prior to taking them forward into clinical trials.

Building on the above, future work involves a number of additional avenues to better understand how the computationally-driven findings could be further validated from a therapeutic standpoint. Firstly, the computational pipeline itself could be refined as outlined in the *Computational prioritisation of genes* section above; this would involve informing gene prioritisation by weighting predictor data-sets according to cell-type relevance specific to the diseases in question, as well as combining discovery and supervised mode outputs to more confidently establish target potential. Next, in the case of Dupuytren's disease, more work can be done downstream of *CREBBP* and *EP300*, so as to hopefully identify more isolated genes that could serve as more adequate therapeutic targets for inhibiting drugs to be designed against. On the other hand, the lupus data would benefit from being supplemented by experiments that analyse other aspects of disease parthenogenesis, including both within B-cell biology (e.g. stimulating through the B-cell receptor) as well as across other cell types (e.g.

understanding T-cell role and effects of gene knockdown on these cells). Naturally, the most compelling form of target validation would consist of successfully navigating clinical trials, at least until the Phase 3 stages when target efficacy is de-risked. One way to fast-track such validating data would be to focus on the drug repositioning aspect; that is, taking existing approved therapies for other indications into trials for the treatment of either systemic lupus erythematosus or Dupuytren's disease.

## 7.4 Conclusion

In summary, the work outlined in this thesis demonstrated three key points: that there is a potential for the use of aggregated genetic data-sets to inform gene prioritisation in a scalable and adaptable way; that such a computational pipeline could successfully be used to identify new insight into disease pathogenesis and uncover potential for novel therapies; and that the genes prioritised through the pipeline can be experimentally validated in a number of complementary lab procedures that give both omic-level and cell phenotype-level information into target relevance to disease. While the 'ideal' outcome of this work (to have adoption of the Priority Index and downstream analyses in future drug discovery pipelines) relies on a number of future experiments to both expand the scope of potential as well as further validate the established findings, the results presented here have moved us a step closer to this. It is my hope that the body of findings discussed in these pages will meaningfully inform and guide future experiments, so as to allow for more robust, systematic, and insightful analysis of genomic data, that translates to the discovery of better therapeutics across a range of immune-related disease.

# References


[1] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9 (3):203–14, 2010.

[2] Alexander Schuhmacher, Oliver Gassmann, and Markus Hinder. Changing R&D models in research-based pharmaceutical companies. *Journal of Translational Medicine*, 14(1):1–11, 2016.

[3] Michael Hay, David W. Thomas, John L. Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1): 40–51, 2014.

[4] David Cook, Dearg Brown, Robert Alexander, et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, 13 (6):419–31, 2014.

[5] Aled Edwards. Open-source science to enable drug discovery. *Drug Discovery Today*, 13 (17-18):731–733, 2008.

[6] J. A. Dimasi, L. Feldman, A. Seckler, and A. Wilson. Trends in risks associated with new drug development: Success rates for investigational drugs. *Clinical Pharmacology and Therapeutics*, 87(3):272–277, 2010.

[7] Matthew R Nelson, Hannah Tipney, Jeffery L Painter, et al. The support of human genetic evidence for approved drug indications. *Nature Genetics*, 47(8):1–7, 2015.

[8] Emily A. King, J. Wade Davis, and Jacob F. Degner. 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. *bioRxiv*, page 513945, 2019.

[9] Akira Endo. A historical perspective on the discovery of statins. *Proceedings of the Japan Academy, Series B*, 86(5):484–493, 2010.

[10] Sekar Kathiresan, Olle Melander, Candace Guiducci, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, 40(2):189–197, 2008.

[11] Sekar Kathiresan, Cristen J. Willer, Gina M. Peloso, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics*, 41(1):56–65, 2009.

[12] Andrew S. Plump and Pek Yee Lum. Genomics and Cardiovascular Drug Development. *Journal of the American College of Cardiology*, 53(13):1089–1100, 2009.

[13] Robert M Plenge, Edward M Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*, 12(8):581–594, 2013.

[14] Annick Prat, Aurélie Derré, Bernard Chanu, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics*, 34(2):154–6, 2003.

[15] Rossi P. Naoumova, Isabella Tosi, Dilip Patel, et al. Severe hypercholesterolemia in four British families with the D374Y mutation in the PCSK9 gene: Long-term follow-up and treatment response. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 25(12):2654–2660, 2005.

[16] S. E. Humphries, R. A. Whittall, C. S. Hubbart, et al. Genetic causes of familial hypercholesterolaemia in patients in the UK: Relation to plasma lipid levels and coronary heart disease risk. *Journal of Medical Genetics*, 43(12):943–949, 2006.

[17] Marianne Abifadel, Jean Pierre Rabès, Martine Devillers, et al. Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9(PCSK9) gene in cholesterol metabolism and disease. *Human Mutation*, 30(4):520–529, 2009.

[18] Keith C. Ferdinand and Samar A. Nasser. PCSK9 Inhibition: Discovery, Current Evidence, and Potential Effects on LDL-C and Lp(a). *Cardiovascular Drugs and Therapy*, 29(3):295–308, 2015.

[19] Alirocumab (Praluent) to Lower LDL-Cholesterol. *JAMA*, 314(12):1284–1285, 2019.

[20] Johann Auer and Robert Berent. Alirocumab as add-on therapy to statins: current evidence and clinical potential. *Therapeutic Advances in Cardiovascular Disease*, 12(7):191–202, 2018.

[21] Samir Wadhawan, Heiko Runz, and Julja Burchard. The genome as pharmacopeia: association of genetic dose with phenotypic response. *Biochemical Pharmacology*, 94(4):229–240, 2015.

[22] Wendy Balemans, Martin Ebeling, Neela Patel, et al. Increased bone density in sclerosteosis is due to the deficiency of a novel secreted protein (SOST). *Human Molecular Genetics*, 10(5): 537–544, 2001.

[23] Mary E Brunkow, Jessica C Gardner, Jeff Van Ness, et al. Bone Dysplasia Sclerosteosis Results from Loss of the SOST Gene Product, a Novel Cystine Knot–Containing Protein. *American Journal of Human Genetics*, 68:577–589, 2001.

[24] Elke Piters, Cavit Culha, Martiene Moester, et al. First missense mutation in the SOST gene causing sclerosteosis by loss of sclerostin function. *Human Mutation*, 31(7):1526–1543, 2010.

[25] Andre G. Uitterlinden, Pascal P. Arp, Bryan W. Paeper, et al. Polymorphisms in the sclerosteosis/van Buchem disease gene (SOST) region are associated with bone-mineral density in elderly whites. *American Journal of Human Genetics*, 75(6):1032–1045, 2004.

[26] Unnur Styrkarsdottir, Bjarni V. Halldorsson, Solveig Gretarsdottir, et al. New sequence variants associated with bone mineral density. *Nature Genetics*, 41(1):15–17, 2009.

[27] Ken Garber. Two pioneering osteoporosis drugs finally approach approval. *Nature Reviews Drug Discovery*, 15(7):445–446, 2016.

[28] F. Cosman, D. B. Crittenden, J. D. Adachi, et al. Romosozumab treatment in postmenopausal women with osteoporosis. *New England Journal of Medicine*, 375(16):1532–1543, 2016.

[29] Kenneth G. Saag, Jeffrey Petersen, Maria Luisa Brandi, et al. Romosozumab or alendronate for fracture prevention in women with osteoporosis. *New England Journal of Medicine*, 377(15): 1417–1427, 2017.

[30] Michael R. McClung. Romosozumab for the treatment of osteoporosis. *Osteoporosis and Sarcopenia*, 4(1):11–15, 2018.

[31] Andrew L Hopkins and Colin R Groom. The druggable genome. *Nature Reviews Drug Discovery*, 1(9):727–30, 2002.

[32] Andreas P. Russ and Stefan Lampel. The druggable genome: An update. *Drug Discovery Today*, 10(23-24):1607–1610, 2005.

[33] Chris Finan, Anna Gaulton, Felix A. Kruger, et al. The druggable genome and support for target identification and validation in drug development. *Science Translational Medicine*, 9 (383):1–16, 2017.

[34] Philippe Sanseau, Pankaj Agarwal, Michael R Barnes, et al. Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317–320, 2012.

[35] Yukinori Okada, Di Wu, Gosia Trynka, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

[36] Hector Giral, Ulf Landmesser, and Adelheid Kratzer. Into the Wild: GWAS Exploration of Non-coding RNAs. *Frontiers in Cardiovascular Medicine*, 5(December), 2018.

[37] S. Prahalad. Negative association between the chemokine receptor CCR5-Δ32 polymorphism and rheumatoid arthritis: A meta-analysis. *Genes and Immunity*, 7(3):264–268, 2006.

[38] Michel P.M. Vierboom, Paul J. Zavodny, Chuan Chu Chou, et al. Inhibition of the development of collagen-induced arthritis in rhesus monkeys by a small molecular weight antagonist of CCR5. *Arthritis and Rheumatism*, 52(2):627–636, 2005.

[39] H. Okamoto and N. Kamatani. A CCR-5 antagonist inhibits the development of adjuvant arthritis in rats [3]. *Rheumatology*, 45(2):230–232, 2006.

[40] Daniëlle M. Gerlag, Sally Hollis, Mark Layton, et al. Preclinical and clinical investigation of a CCR5 antagonist, AZD5672, in patients with rheumatoid arthritis receiving methotrexate. *Arthritis and Rheumatism*, 62(11):3154–3160, 2010.

[41] Arno W.R. Van Kuijk, Clarissa E. Vergunst, Danielle M. Gerlag, et al. CCR5 blockade in rheumatoid arthritis: A randomised, double-blind, placebo-controlled clinical trial. *Annals of the Rheumatic Diseases*, 69(11):2013–2016, 2010.

[42] Dona L. Fleishaker, Xin Wang, Sujatha Menon, et al. Maraviroc, a chemokine receptor-5 antagonist, fails to demonstrate efficacy in the treatment of patients with rheumatoid arthritis in a randomized, double-blind placebo-controlled trial. *Arthritis Research and Therapy*, 14(1): 1–11, 2012.

[43] B. Chen and A. J. Butte. Leveraging big data to transform target selection and drug discovery. *Clinical Pharmacology and Therapeutics*, 99(3):285–297, 2016.

[44] Hua Fan-Minogue, Bin Chen, Weronika Sikora-Wohlfeld, et al. a Systematic Assessment of Linking Gene Expression With Genetic Variants for Prioritizing Candidate Targets. *Pac Symp Biocomput.*, 20:383–394, 2015.

[45] Tudor I. Oprea, Cristian G. Bologa, Søren Brunak, et al. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery*, 17(5):317–332, 2018.

[46] Mark Davies, Michal Nowotka, George Papadatos, et al. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43:W612–W620, 2015.

[47] David S Wishart, C Knox, An C Guo, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(90001):D668–D672, 2006.

[48] Hong Yang, Chu Qin, Ying Hong Li, et al. Therapeutic target database update 2016: Enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Research*, 44(D1):D1069–D1074, 2016.

[49] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, 2016.

[50] Zhihong Zhu, Futao Zhang, Han Hu, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, 48(5):481–7, 2016.

[51] Dorothea Emig, Alexander Ivliev, Olga Pustovalova, et al. Drug Target Prediction and Repositioning Using an Integrated Network-Based Approach. *PLoS ONE*, 8(4), 2013.

[52] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford InfoLab*, pages 111–112, 1998.

[53] Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan, and Eli Upfal. Fast distributed PageRank computation. *Theoretical Computer Science*, 561(PB):113–121, 2015.

[54] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*, 82(4):949–958, 2008.

[55] Tune H. Pers, Juha M. Karjalainen, Yingleong Chan, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, 6: 5890, 2015.

[56] Jesse Gillis and Paul Pavlidis. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology*, 8(3), 2012.

[57] Jouhyun Jeon, Satra Nim, Joan Teyra, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine*, 6(7):57, 2014.

[58] Gautier Koscielny, Peter An, Denise Carvalho-Silva, et al. Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Research*, 45:D985–D994, 2017.

[59] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26(1):5–14, 2001.

[60] Enrico Ferrero, Ian Dunham, and Philippe Sanseau. In silico prediction of novel therapeutic targets using gene-disease association data. *Journal of Translational Medicine*, 15(1):1–16, 2017.

[61] Mouldy Sioud. Main approaches to target discovery and validation. *Methods in molecular biology (Clifton, N.J.)*, 360:1–12, 2007.

[62] Aled M. Edwards, Cheryl H. Arrowsmith, Chas Bountra, et al. Preclinical target validation using patient-derived cells. *Nature Reviews Drug Discovery*, 14(3):149–150, 2015.

[63] Fionula Brennan, David Chantry, Andrew Jackson, Ravinder Maini, and Marc Feldmann. Inhibitory effect of TNF-alpha antibodies on synovial cell IL1 production in rheumatoid arthritis. *The Lancet*, 3:244–247, 1989.

[64] Hai Fang, Hans De Wolf, Bogdan Knezevic, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature Genetics*, 51(July), 2019.

[65] Danielle Welter, Jacqueline MacArthur, Joannella Morales, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1):1001–1006, 2014.

[66] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1): D896–D901, 2017.

[67] Nikolaos A. Patsopoulos, Sergio E. Baranzini, Adam Santaniello, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*, 365 (6460), 2019.

[68] Peter K. Gregersen, Roman Kosoy, Annette T. Lee, et al. Risk for myasthenia gravis maps to a 151Pro→Ala change in TNIP1 and to human leukocyte antigen-B*08. *Annals of Neurology*, 72 (6):927–935, 2012.

[69] Alan E. Renton, Hannah A. Pliner, Carlo Provenzano, et al. A Genome-Wide Association Study of Myasthenia Gravis. *JAMA Neurology*, 72(4):396, 2015.

[70] Michael F Seldin, Omar K Alkhairy, Annette T Lee, et al. Genome-Wide Association Study of Late-Onset Myasthenia Gravis: Confirmation of TNFRSF11A and Identification of and Three Distinct HLA Associations. *Molecular Medicine*, 21(1):1, 2015.

[71] Michael Ng, Dipti Thakkar, Lorraine Southam, et al. A Genome-wide Association Study of Dupuytren Disease Reveals 17 Additional Variants Implicated in Fibrosis. *American Journal of Human Genetics*, 101(3):417–427, 2017.

[72] Sergey V. Kozyrev, Anna Karin Abelson, Jerome Wojcik, et al. Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus. *Nature Genetics*, 40(2): 211–216, 2008. doi: 10.1038/ng.79.

[73] John B. Harley, Marta E. Alarcón-Riquelme, Lindsey A. Criswell, et al. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nature Genetics*, 40(2):204–210, 2008.

[74] Geoffrey Hom, Robert R. Graham, Barmak Modrek, et al. Association of Systemic Lupus Erythematosus with C8orf13–BLK and ITGAM–ITGAX. *New England Journal of Medicine*, 358(9):900–909, 2008.

[75] Robert R. Graham, Chris Cotsapas, Leela Davies, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nature Genetics*, 40(9):1059–1061, 2008.

[76] Vesela Gateva, Johanna K. Sandling, Geoff Hom, et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nature Genetics*, 41(11):1228–1233, 2009.

[77] Young Ho Lee, Sang Cheol Bae, Sung Jae Choi, Jong Dae Ji, and Gwan Gyu Song. Genome-wide pathway analysis of genome-wide association studies on systemic lupus erythematosus and rheumatoid arthritis. *Molecular Biology Reports*, 39(12):10627–10635, 2012.

[78] Jose Ezequiel Martin, Shervin Assassi, Lina Marcela Diaz-Gallo, et al. A systemic sclerosis and systemic lupus erythematosus pan-meta-GWAS reveals new shared susceptibility loci. *Human Molecular Genetics*, 22(19):4021–4029, 2013.

[79] D. L. Armstrong, R. Zidovetzki, M. E. Alarcón-Riquelme, et al. GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. *Genes and Immunity*, 15 (6):347–354, 2014.

[80] F. Yesim Demirci, Xingbin Wang, Jennifer A. Kelly, et al. Identification of a New Susceptibility Locus for Systemic Lupus Erythematosus on Chromosome 12 in Individuals of European Ancestry. *Arthritis and Rheumatology*, 68(1):174–183, 2016.

[81] James Bentham, David L. Morris, Graham Cunninghame, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics*, 47(12):1457–1464, 2015.

[82] Ana Márquez, Laura Vidal-Bralo, Luis Rodríguez-Rodríguez, et al. A combined large-scale meta-Analysis identifies COG6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus. *Annals of the Rheumatic Diseases*, 76(1):286–294, 2017.

[83] David L. Morris, Yujun Sheng, Yan Zhang, et al. Genome-wide association meta-analysis in Chinese and European individuals identifies ten new loci associated with systemic lupus erythematosus. *Nature Genetics*, 48(8):940–946, 2016.

[84] Carl D. Langefeld, Hannah C. Ainsworth, Deborah S.Cunninghame Graham, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. *Nature Communications*, 8(May), 2017.

[85] Antonio Julià, Francisco Javier López-Longo, José J. Pérez Venegas, et al. Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus. *Arthritis Research and Therapy*, 20(1):1–10, 2018.

[86] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, 2014.

[87] Biola M. Javierre, Oliver S. Burren, Steven P. Wilder, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, 2016.

[88] Benjamin P. Fairfax, Seiko Makino, Jayachandran Radhakrishnan, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics*, 44(5):502–510, 2012.

[89] Vivek Naranbhai, Benjamin P Fairfax, Seiko Makino, et al. Genomic modulators of gene expression in human neutrophils. *Nature Communications*, 6(May):7545, 2015.

[90] Silva Kasela, Kai Kisand, Liina Tserel, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+versus CD8+T cells. *PLoS Genetics*, 13(3), 2017.

[91] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(May):25–29, 2000.

[92] David Valle, Ada Hamosh, Victor A. McKusick, Joanna Amberger, and Alan F. Scott. Online Mendelian Inheritance In Man (OMIM). *Human Mutation*, 15(1):57–61, 2002.

[93] Warren A. Kibbe, Cesar Arze, Victor Felix, et al. Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research*, 43(D1):D1071–D1078, 2015.

[94] Sebastian Köhler, Nicole A. Vasilevsky, Mark Engelstad, et al. The human phenotype ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876, 2017.

[95] Cynthia L. Smith and Janan T. Eppig. The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):390–399, 2009.

[96] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43:D447–D452, 2015.

[97] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40:1100–1107, 2012.

[98] The UniProt Consortium. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019.

[99] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research*, 33(DATABASE ISS.):54–58, 2005.

[100] David Croft, Gavin O'Kelly, Guanming Wu, et al. Reactome: A database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(SUPPL. 1):691–697, 2011.

[101] Ethan G. Cerami, Benjamin E. Gross, Emek Demir, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39:685–690, 2011.

[102] H. Motenko, S. B. Neuhauser, M. O'Keefe, and J. E. Richardson. MouseMine: a new data warehouse for MGI. *Mammalian Genome*, 26(7-8):325–330, 2015.

[103] Carol J. Bult, Judith A. Blake, Cynthia L. Smith, et al. Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806, 2019.

[104] Hai Fang, Bogdan Knezevic, Katie L. Burnham, and Julian C. Knight. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. *Genome Medicine*, 8(1):129, 2016.

[105] Valentina Iotchkova, Graham R S Ritchie, Matthias Geihs, et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals Valentina. *Nature Genetics*, 51(February):343–353, 2019.

[106] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[107] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

[108] Andrew D. Johnson, Robert E. Handsaker, Sara L. Pulit, et al. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24): 2938–2939, 2008.

[109] The International HapMap Consortium. The International HapMap Project. *Nature*, 426 (6968):789–796, 2003.

[110] Gleb Kichaev, Wen-yun Yang, Sara Lindstrom, Farhad Hormozdiari, and Eleazar Eskin. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, 10(10), 2014.

[111] Matt Maurano, Richard Humbert, Eric Rynes, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[112] Claudia Giambartolomei, Damjan Vukcevic, Eric E. Schadt, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genetics*, 10(5), 2014.

[113] Aravind Subramaniana, Pablo Tamayoa, Vamsi K. Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102 (43):15545–15550, 2005.

[114] Liaquat S. Verjee, Jennifer S.N. Verhoekx, James K.K. Chan, et al. Unraveling the signaling pathways promoting fibrosis in Dupuytren's disease reveals TNF as a therapeutic target. *Proceedings of the National Academy of Sciences of the United States of America*, 110(10): 928–937, 2013.

[115] Adam P. Cribbs, Sebastian Luna-Valero, Charlotte George, et al. CGAT-core: a python framework for building scalable, reproducible computational biology workflows. *F1000Research*, 8:377, 2019.

[116] Simon Andrews. Fastqc: a quality control tool for high throughput sequence data, 1999. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[117] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, 2015.

[118] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[119] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, 2014.

[120] Rita Santos, Oleg Ursu, Anna Gaulton, et al. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017.

[121] Koen Gerritse, Jon D. Laman, Randolph J. Noelle, et al. CD40-CD40 ligand interactions in experimental allergic encephalomyelitis and multiple sclerosis. *Proceedings of the National Academy of Sciences of the United States of America*, 93(6):2499–2504, 1996.

[122] Iqbal S. Grewal, Harald G. Foellmer, Kate D. Grewal, et al. Requirement for CD40 ligand in costimulation induction, T cell activation, and experimental allergic encephalomyelitis. *Science*, 273(5283):1864–1867, 1996.

[123] J. D. Laman, C. B.M. Maassen, M. M. Schellekens, et al. Therapy with antibodies against CD40L (CD154) and CD44-variant isoforms reduces experimental autoimmune encephalomyelitis induced by a proteolipid protein peptide. *Multiple Sclerosis*, 4(3):147–153, 1998.

[124] Louis Boon, Herbert P. M. Brok, Jan Bauer, et al. Prevention of Experimental Autoimmune Encephalomyelitis in the Common Marmoset ( Callithrix jacchus ) Using a Chimeric Antagonist Monoclonal Antibody Against Human CD40 Is Associated with Altered B Cell Responses . *The Journal of Immunology*, 167(5):2942–2949, 2001.

[125] Suzanne A.B.M. Aarts, Tom T.P. Seijkens, Pascal J.H. Kusters, et al. Inhibition of CD40-TRAF6 interactions by the small molecule inhibitor 6877002 reduces neuroinflammation. *Journal of Neuroinflammation*, 14(1):1–14, 2017.

[126] Suzanne A.B.M. Aarts, Tom T.P. Seijkens, Koos J.F. van Dorst, Christine D. Dijkstra, Gijs Kooij, and Esther Lutgens. The CD40-CD40L dyad in experimental autoimmune encephalomyelitis and multiple sclerosis. *Frontiers in Immunology*, 8, 2017.

[127] Rehab A. Karam, Noha A. Rezk, Mona M. Amer, and Hala A. Fathy. Immune response genes receptors expression and polymorphisms in relation to multiple sclerosis susceptibility and response to INF-$\beta$ therapy. *IUBMB Life*, pages 727–734, 2016.

[128] María L. Cavanillas, Antonio Alcina, Concepción Ñez, et al. Polymorphisms in the IL2, IL2RA and IL2RB genes in multiple sclerosis risk. *European Journal of Human Genetics*, 18(7): 794–799, 2010.

[129] Philip L. De Jager, Xiaoming Jia, Joanne Wang, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nature Genetics*, 41(7):776–782, 2009.

[130] S. Perga, F. Montarolo, S. Martire, P. Berchialla, S. Malucchi, and A. Bertolotto. Anti-inflammatory genes associated with multiple sclerosis: A gene expression study. *Journal of Neuroimmunology*, 279(C):75–78, 2015.

[131] Andreas Junker, Jana Ivanidze, Joachim Malotka, et al. Multiple sclerosis: T-cell receptor expression in distinct brain regions. *Brain*, 130(11):2789–2799, 2007.

[132] Evelyn Zastepa, Leslie Fitz-Gerald, Michael Hallett, et al. Naive CD4 T-cell activation identifies MS patients having rapid transition to progressive MS. *Neurology*, 82(8):681–690, 2014.

[133] L. Adorini. Interleukin-12, a key cytokine in Th1-mediated autoimmune diseases. *Cellular and Molecular Life Sciences*, 55(12):1610–1625, 1999.

[134] Christopher L. Karp, Christine A. Biron, and David N. Irani. Interferon $\beta$ in multiple sclerosis: Is IL-12 suppression the key? *Immunology Today*, 21(1):24–28, 2000.

[135] Pushpalatha Palle, Kelly L. Monaghan, Sarah M. Milne, and Edwin C.K. Wan. Cytokine Signaling in Multiple Sclerosis and Its Therapeutic Applications. *Medical Sciences*, 5(4):23, 2017.

[136] Joshua C. Denny, Lisa Bastarache, Marylyn D. Ritchie, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, 31(12):1102–1110, 2013.

[137] Ilya Ayzenberg, Robert Hoepner, and Ingo Kleiter. Fingolimod for multiple sclerosis and emerging indications: Appropriate patient selection, safety precautions, and special considerations. *Therapeutics and Clinical Risk Management*, 12:261–272, 2016.

[138] Matt Shirley. Daclizumab: A Review in Relapsing Multiple Sclerosis. *Drugs*, 77(4):447–458, 2017.

[139] Christina M. Lill, Brit Maren M. Schjeide, Denis A. Akkad, et al. Independent replication of STAT3 association with multiple sclerosis risk in a large German case-control sample. *Neurogenetics*, 13(1):83–86, 2012.

[140] V. Özenci, M. Kouwenhoven, Y. M. Huang, B. G. Xiao, P. Kivisäkk, S. Fredrikson, and H. Link. Multiple sclerosis: Levels of interleukin-10-secreting blood mononuclear cells are low in untreated patients but augmented during interferon-$\beta$-1b treatment. *Scandinavian Journal of Immunology*, 49(5):554–561, 1999.

[141] E. Ersoy, C. N.S. Kuş, U. Şener, I. Çoker, and Yaşar Zorlu. The effects of interferon-$\beta$ on interleukin-10 in multiple sclerosis patients. *European Journal of Neurology*, 12(3):208–211, 2005.

[142] Daniel J. Cua, Jonathan Sherlock, Yi Chen, et al. Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. *Nature*, 421(6924):744–748, 2003.

[143] Kerstin Göbel, Tobias Ruck, and Sven G. Meuth. Cytokine signaling in multiple sclerosis: Lost in translation. *Multiple Sclerosis Journal*, 24(4):432–439, 2018.

[144] Torben L. Sorensen, Corinna Trebst, Pia Kivisäkk, et al. Multiple sclerosis: A study of CXCL10 and CXCR3 co-localization in the inflamed central nervous system. *Journal of Neuroimmunology*, 127(1-2):59–68, 2002.

[145] Kevin S. Carbajal, Juan L. Miranda, Michelle R. Tsukamoto, and Thomas E. Lane. CXCR4 signaling regulates remyelination by endogenous oligodendrocyte progenitor cells in a viral model of demyelination. *Glia*, 59(12):1813–1821, 2011.

[146] Tianci Chu, Lisa B.E. Shields, Yi Ping Zhang, Shi Qing Feng, Christopher B. Shields, and Jun Cai. CXCL12/CXCR4/CXCR7 Chemokine Axis in the Central Nervous System: Therapeutic Targets for Remyelination in Demyelinating Diseases. *Neuroscientist*, 23(6):627–648, 2017.

[147] Linda Ottoboni, Brendan T. Keenan, Pablo Tamayo, et al. An RNA profile identifies two subsets of multiple sclerosis patients differing in disease activity. *Science Translational Medicine*, 4(153), 2012.

[148] Thomas Rolland, Murat Taşan, Benoit Charloteaux, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, 2014.

[149] Isaac Hurtado-Guerrero, Maria Jesús Pinto-Medel, Patricia Urbaneja, et al. Activation of the JAK-STAT signaling pathway after In Vitro stimulation with IFN$\beta$ in multiple sclerosis patients according to the therapeutic response to IFN$\beta$. *PLoS ONE*, 12(1):1–19, 2017.

[150] Ester Canto, Noriko Isobe, Alessandro Didonna, et al. Aberrant STAT phosphorylation signaling in peripheral blood mononuclear cells from multiple sclerosis patients. *Journal of Neuroinflammation*, 15(1):1–11, 2018.

[151] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:1–11, 2009.

[152] Luca Vangelista and Sandro Vento. The expanding therapeutic perspective of CCR5 blockade. *Frontiers in Immunology*, 8(JAN):1–7, 2018.

[153] Takumi Matsumoto, Ichiro Nakamura, Ayumi Miura, Gen Momoyama, and Katsumi Ito. New-onset multiple sclerosis associated with adalimumab treatment in rheumatoid arthritis: A case report and literature review. *Clinical Rheumatology*, 32(2):271–275, 2013.

[154] Benjamin M. Segal, Cris S. Constantinescu, Aparna Raychaudhuri, Lilianne Kim, Rosanne Fidelus-Gort, and Lloyd H. Kasper. Repeated subcutaneous injections of IL12/23 p40 neutralising antibody, ustekinumab, in patients with relapsing-remitting multiple sclerosis: a phase II, double-blind, placebo-controlled, randomised, dose-ranging study. *The Lancet Neurology*, 7(9):796–804, 2008.

[155] Charles A. Dinarello, Anna Simon, and Jos W.M. Van Der Meer. Treating inflammation by blocking interleukin-1 in a broad spectrum of diseases. *Nature Reviews Drug Discovery*, 11(8): 633–652, 2012.

[156] Eugen Dhimolea. Canakinumab. *mAbs*, (February):3–13, 2010.

[157] Eline A. Dubois, Robert Rissmann, and Adam F. Cohen. Rilonacept and canakinumab. *British Journal of Clinical Pharmacology*, 71(5):639–641, 2011.

[158] Sasha Bernatsky, Christel Renoux, and Samy Suissa. Demyelinating events in rheumatoid arthritis after drug exposures. *Annals of the Rheumatic Diseases*, 69(9):1691–1693, 2010.

[159] Vern C. Juel and Janice M. Massey. Myasthenia Gravis. *Orphanet Journal of Rare Diseases Review*, 2(44), 2007.

[160] Nils Erik Gilhus and Jan J. Verschuuren. Myasthenia gravis: Subgroup classification and therapeutic strategies. *The Lancet Neurology*, 14(10):1023–1036, 2015.

[161] Gil I. Wolfe, Henry J. Kaminski, Inmaculada B. Aban, et al. Randomized Trial of Thymectomy in Myasthenia Gravis. *New England Journal of Medicine*, 375(6):511–522, 2016.

[162] Karen Heyninck, Dirk De Valck, Wim Vanden Berghe, et al. The zinc finger protein A20 inhibits TNF-induced NF-κB-dependent gene expression by interfering with an RIP- or TRAF2-mediated transactivation signal and directly binds to a novel NF-κB-inhibiting protein ABIN. *Journal of Cell Biology*, 145(7):1471–1482, 1999.

[163] Xinze Cai, Ying Qiao, Cheng Diao, et al. Association between polymorphisms of the IKZF3 gene and systemic lupus erythematosus in a Chinese Han Population. *PLoS ONE*, 9(10):8–11, 2014.

[164] Ramanjaneyulu Allam, Sufyan G. Sayyed, Onkar P. Kulkarni, et al. Mdm2 Promotes Systemic Lupus Erythematosus and Lupus Nephritis. *Journal of the American Society of Nephrology*, 22 (11):2016–2027, 2011.

[165] Shuo Wei, Nobuya Yoshida, Greg Finn, et al. Pin1-Targeted Therapy for Systemic Lupus Erythematosus. *Arthritis and Rheumatology*, 68(10):2503–2513, 2016.

[166] Hai Fang and Julian Gough. The 'dnet' approach promotes emerging research on cancer patient survival. *Genome Medicine*, 6(8):1–15, 2014.

[167] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

[168] Morisada Hayakawa, Hiroko Hayakawa, Tsvetana Petrova, et al. Loss of Functionally Redundant p38 Isoforms in T Cells Enhances Regulatory T Cell Induction. *Journal of Biological Chemistry*, 292(5):1762–1772, 2017.

[169] K.H. Park, J. Jung, J.-H. Lee, and Y.-H. Hong. Blood transcriptome profiling in myasthenia gravis patients to assess disease activity: A pilot RNA-seq study. *Experimental Neurobiology*, 25 (1):40–47, 2016.

[170] A. Jaretzki, R. J. Barohn, R. M. Ernstoff, H. J. Kaminski, J. C. Keesey, A. S. Penn, and D. B. Sanders. Myasthenia gravis: Recommendations for clinical research standards. *Neurology*, 55 (1):16–23, 2000.

[171] Chang Xia, Zachary Braunstein, Amelia C. Toomey, Jixin Zhong, and Xiaoquan Rao. S100 proteins as an important regulator of macrophage inflammation. *Frontiers in Immunology*, 8 (JAN):1–11, 2018.

[172] Zhengqi Wang, Tamisha Y. Vaughan, Wandi Zhu, et al. Gab2 and Gab3 redundantly suppress colitis by modulating macrophage and CD8+ T-cell activation. *Frontiers in Immunology*, 10 (MAR):1–18, 2019.

[173] Guo Long Zhang, Yan Feng Zou, Xiao Liang Feng, et al. Association of the NFKBIA gene polymorphisms with susceptibility to autoimmune and inflammatory diseases: A meta-analysis. *Inflammation Research*, 60(1):11–18, 2011.

[174] Upneet K. Sokhi, Mark P. Liber, Laura Frye, et al. Dissection and function of autoimmunity-associated TNFAIP3 (A20) gene enhancers in humanized mouse models. *Nature Communications*, 9(1), 2018.

[175] Yanan Chen, Joseph R. Podojil, Rejani B. Kunjamma, et al. Sephin1, which prolongs the integrated stress response, is a promising therapeutic for multiple sclerosis. *Brain*, 142(2): 344–361, 2019.

[176] Ashley H. Beecham, Nikolaos A. Patsopoulos, Dionysia K. Xifara, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11):1353–1362, 2013.

[177] Manuel A. Ferreira, Judith M. Vonk, Hansjörg Baurecht, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature Genetics*, 49(12):1752–1757, 2017.

[178] Katrina M. De Lange, Loukas Moutsianas, James C. Lee, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, 49(2):256–261, 2017.

[179] Harm Jan Westra, Marta Martínez-Bonet, Suna Onengut-Gumuscu, et al. Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nature Genetics*, 50(10):1366–1374, 2018.

[180] Anne Laure Mathieu, Estelle Verronese, Gillian I. Rice, et al. PRKDC mutations associated with immunodeficiency, granuloma, and autoimmune regulator-dependent autoimmunity. *Journal of Allergy and Clinical Immunology*, 135(6):1578–1588.e5, 2015.

[181] A. Uzawa, N. Kawaguchi, K. Himuro, T. Kanai, and S. Kuwabara. Serum cytokine and chemokine profiles in patients with myasthenia gravis. *Clinical and Experimental Immunology*, 176(2):232–237, 2014.

[182] Paola Cavalcante, Barbara Serafini, Barbara Rosicarelli, et al. Epstein-Barr virus persistence and reactivation in myasthenia gravis thymus. *Annals of Neurology*, 67(6):726–738, 2010.

[183] Paola Cavalcante, Stefania Marcuzzo, Sara Franzi, et al. Epstein-Barr virus in tumor-infiltrating B cells of myasthenia gravis thymoma: an innocent bystander or an autoimmunity mediator? *Oncotarget*, 8(56):95432–95449, 2017.

[184] Guilin Qiao, Minxiang Lei, Zhenping Li, et al. Negative Regulation of CD40-Mediated B Cell Responses by E3 Ubiquitin Ligase Casitas-B-Lineage Lymphoma Protein-B. *The Journal of Immunology*, 179(7):4473–4479, 2007.

[185] Sin-Hyeog Im, Dora Barchan, Prasanta Kumar Maiti, Sara Fuchs, and Miriam C. Souroujon. Blockade of CD40 Ligand Suppresses Chronic Experimental Myasthenia Gravis by Down-Regulation of Th1 Differentiation and Up-Regulation of CTLA-4. *The Journal of Immunology*, 166(11):6893–6898, 2001.

[186] Oren Parnas, Marko Jovanovic, Thomas M. Eisenhaure, et al. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell*, 162(3):675–686, 2015.

[187] Samantha K. Dunmire, Oludare A. Odumade, Jean L. Porter, et al. Primary EBV infection induces an expression profile distinct from other viruses but similar to hemophagocytic syndromes. *PLoS ONE*, 9(1):1–10, 2014.

[188] Ti-kun Fang, Cheng-jun Yan, and Juan Du. CTLA-4 methylation regulates the pathogenesis of myasthenia gravis and the expression of related cytokines. *Medicine*, 97, 2018.

[189] Jan-Michael Werner, Viola Schweinsberg, Michael Schroeter, et al. Successful Treatment of Myasthenia Gravis Following PD-1/CTLA-4 Combination Checkpoint Blockade in a Patient With Metastatic Melanoma. *Frontiers in Oncology*, 9(February):1–3, 2019.

[190] N. Jain, H. Nguyen, C. Chambers, and J. Kang. Dual function of CTLA-4 in regulatory T cells and conventional T cells to prevent multiorgan autoimmunity. *Proceedings of the National Academy of Sciences*, 107(4):1524–1528, 2010.

[191] Jason D. Fontenot, Marc A. Gavin, and Alexander Y. Rudensky. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells. *Journal of Immunology*, 198(3): 986–992, 2017.

[192] Yongqing Wu, Madhuri Borde, Vigo Heissmeyer, et al. FOXP3 Controls Regulatory T Cell Function through Cooperation with NFAT. *Cell*, 126(2):375–387, 2006.

[193] Muthusamy Thiruppathi, Julie Rowin, Qin Li Jiang, et al. Functional defect in regulatory T cells in myasthenia gravis. *Annals of the New York Academy of Sciences*, 1274(1):68–76, 2012.

[194] Fabian Birzele, Tanja Fauti, Heiko Stahl, et al. Next-generation insights into regulatory T cells: Expression profiling and FoxP3 occupancy in Human. *Nucleic Acids Research*, 39(18): 7946–7960, 2011.

[195] Martti Laan and Pärt Peterson. The Many Faces of Aire in Central Tolerance. *Frontiers in Immunology*, 4(October):1–6, 2013.

[196] Stephen N. Sansom, Noriko Shikama-Dorn, Saule Zhanybekova, et al. Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Research*, 24(12): 1918–1931, 2014.

[197] Bruce E. Garfield, Troy Krahl, Stacia Appel, Sheldon M. Cooper, and Mercedes Rincón. Regulation of p38 MAP kinase in CD4+ lymphocytes by infliximab therapy in patients with rheumatoid arthritis. *Clinical Immunology*, 116(2):101–107, 2005.

[198] Christopher E Rudd and Helga Schneider. Lymphocyte signaling: Cbl sets the threshold for autoimmunity. *Current Biology*, 381:344–347, 2000.

[199] Christina Lutz-Nicoladoni, Dominik Wolf, and Sieghart Sopper. Modulation of Immune Cell Functions by the E3 Ligase Cbl-b. *Frontiers in Oncology*, 5(March):1–14, 2015.

[200] Paola Cavalcante, Lorenzo Maggi, Lara Colleoni, et al. Inflammation and Epstein-Barr virus infection are common features of myasthenia gravis thymus: Possible roles in pathogenesis. *Autoimmune Diseases*, 1(1), 2011.

[201] Elizabeth C. Jury, Fabian Flores-Borja, Hardeep S. Kalsi, et al. Abnormal CTLA-4 function in T cells from patients with systemic lupus erythematosus. *European Journal of Immunology*, 40 (2):569–578, 2010.

[202] Benjamin J. Schmiedel, Divya Singh, Ariel Madrigal, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell*, 175(6):1701–1715.e16, 2018.

[203] Bo Wu, Elzbieta Goluszko, Ruksana Huda, Erdem Tuzun, and Premkumar Christadoss. Experimental autoimmune myasthenia gravis in the mouse. *Current Protocols in Immunology*, 23(1):1–7, 2001.

[204] Rolen M. Quadros, Hiromi Miura, Donald W. Harms, et al. Easi-CRISPR: A robust method for one-step generation of mice carrying conditional and insertion alleles using long ssDNA donors and CRISPR ribonucleoproteins. *Genome Biology*, 18(1):1–15, 2017.

[205] Takayuki Koyano, Masato Ohtsuka, Shuji Takabayashi, et al. i-GONAD: a robust method for in situ germline genome engineering using CRISPR nucleases. *Genome Biology*, 19(1):1–15, 2018.

[206] Scott L. Friedman, Dean Sheppard, Jeremy S. Duffield, and Shelia Violette. Therapy for fibrotic diseases: Nearing the starting line. *Science Translational Medicine*, 5(167), 2013.

[207] Sandip Hindocha, Duncan Angus McGrouther, and Ardeshir Bayat. Epidemiological evaluation of dupuytren's disease incidence and prevalence rates in relation to etiology. *Hand*, 4(3):256–269, 2009.

[208] T. R.C. Davis. Surgical treatment of primary Dupuytren's contractures of the fingers in the UK: Surgeons' preferences and research priorities. *Journal of Hand Surgery: European Volume*, 38 (1):83–85, 2013.

[209] S. Larsen, D. G. Krogsgaard, L. Aagaard Larsen, M. Iachina, A. Skytthe, and H. Frederiksen. Genetic and environmental influences in Dupuytren's disease: A study of 30,330 Danish twin pairs. *Journal of Hand Surgery: European Volume*, 40(2):171–176, 2015.

[210] Nina S. Godtfredsen, Hasse Lucht, Eva Prescott, Thorkild I.A. Sørensen, and Morten Grønbæk. A prospective study linked both alcohol and tobacco to Dupuytren's disease. *Journal of Clinical Epidemiology*, 57(8):858–863, 2004.

[211] Alexis Descatha, Pénélope Jauffret, Jean François Chastang, Yves Roquelaure, and Annette Leclerc. Should we consider Dupuytren's contracture as work-related? A review and meta-analysis of an old debate. *BMC Musculoskeletal Disorders*, 12, 2011.

[212] Asish K. Ghosh, Rahul Rai, Panagiotis Flevaris, and Douglas E. Vaughan. Epigenetics in Reactive and Reparative Cardiac Fibrogenesis: The Promise of Epigenetic Therapy. *Journal of Cellular Physiology*, 232(8):1941–1956, 2017.

[213] Saeed Aslani, Soheila Sobhani, Farhad Gharibdoost, Ahmadreza Jamshidi, and Mahdi Mahmoudi. Epigenetics and pathogenesis of systemic sclerosis; the ins and outs. *Human Immunology*, 79(3):178–187, 2018.

[214] Eva Moran-Salvador and Jelena Mann. Epigenetics and Liver Fibrosis. *Cmgh*, 4(1):125–134, 2017.

[215] Steven O'Reilly. Epigenetics in fibrosis. *Molecular Aspects of Medicine*, 54:89–102, 2017.

[216] James J. Tomasek, Giulio Gabbiani, Boris Hinz, Christine Chaponnier, and Robert A. Brown. Myofibroblasts and mechano: Regulation of connective tissue remodelling. *Nature Reviews Molecular Cell Biology*, 3(5):349–363, 2002.

[217] Boris Hinz, Sem H. Phan, Victor J. Thannickal, Andrea Galli, Marie Luce Bochaton-Piallat, and Giulio Gabbiani. The myofibroblast: One function, multiple origins. *American Journal of Pathology*, 170(6):1807–1816, 2007.

[218] Jagdeep Nanchahal and Boris Hinz. Strategies to overcome the hurdles to treat fibrosis, a major unmet clinical need. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7291–7293, 2016.

[219] Liaquat Suleman Verjee, Kim Midwood, Dominique Davidson, David Essex, Ann Sandison, and Jagdeep Nanchahal. Myofibroblast Distribution in Dupuytren's Cords: Correlation With Digital Contracture. *Journal of Hand Surgery*, 34(10):1785–1794, 2009.

[220] Rebecca K. Rowntree and Ann Harris. The phenotypic consequences of CFTR mutations. *Annals of Human Genetics*, 67(5):471–485, 2003.

[221] Hee Jin, Ga Young Kang, Seulgi Jeon, et al. Identification of molecular signatures involved in radiation-induced lung fibrosis. *Journal of Molecular Medicine*, 97(1):37–47, 2019.

[222] Dong Ji, Guo Feng Chen, Jin Cheng Wang, et al. Identification of TAF1, HNF4A, and CALM2 as potential therapeutic target genes for liver fibrosis. *Journal of Cellular Physiology*, 234(6): 9045–9051, 2018.

[223] Dario R. Lemos, Michael McMurdo, Gamze Karaca, et al. Interleukin-1b activates a MYC-dependent metabolic switch in kidney stromal cells necessary for progressive tubulointerstitial fibrosis. *Journal of the American Society of Nephrology*, 29(6):1690–1705, 2018.

[224] Yayoi Tada, Elisabeth Riedl, Mark S Lowenthal, et al. Identification and Characterization of Endogenous Langerin Ligands in Murine Extracellular Matrix. *Journal of Investigative Dermatology*, 126(7):1549–1558, 2006.

[225] F I Qureshi, R Hornigold, J D Spencer, and S M Hall. Langerhans cells in Dupuytren's Contracture. *Journal of Hand Surgery*, 26B(4):0–5, 2001.

[226] Liaquat Suleman Verjee, Kim Midwood, Dominique Davidson, Mark Eastwood, and Jagdeep Nanchahal. Post-Transcriptional Regulation of a -Smooth Muscle Actin Determines the Contractile Phenotype of Dupuytren ' s Nodular Cells. *Journal of Cellular Physiology*, (November 2009):681–690, 2010.

[227] Maro Bujak, Ivana Ratkaj, Elitza Markova-Car, et al. Inflammatory gene expression upon TGF-$\beta$1-Induced p38 activation in Primary Dupuytren's Disease fibroblasts. *Frontiers in Molecular Biosciences*, 2(DEC):1–9, 2015.

[228] Sara Mccarty, Farhatullah Syed, and Ardeshir Bayat. Role of the HLA System in the Pathogenesis of Dupuytren ' s Disease. *Hand*, 5:241–250, 2010.

[229] Guido H Dolmans, Paul M Werker, Hans C Hennies, et al. Wnt Signaling and Dupuytren's Disease. *New England Journal of Medicine*, 365(4):307–317, 2011.

[230] Robert H. Jenkins and Donald J. Fraser. BMP-6 emerges as a potential major regulator of fibrosis in the kidney. *American Journal of Pathology*, 178(3):964–965, 2011.

[231] Blanca Herrera, Annalisa Addante, and Aránzazu Sánchez. BMP signalling at the crossroad of liver fibrosis and regeneration. *International Journal of Molecular Sciences*, 19(1), 2018.

[232] S. Klee, M. Lehmann, D. E. Wagner, H. A. Baarsma, and M. Konigshoff. WISP1 mediates IL-6-dependent proliferation in primary human lung fibroblasts. *Scientific Reports*, 6 (February):1–11, 2016.

[233] Cynthia Yu-Wai-Man, Aristides D. Tagalakis, Jinhong Meng, et al. Genotype-phenotype associations of IL6 and PRG4 with conjunctival fibrosis after glaucoma surgery. *JAMA Ophthalmology*, 135(11):1147–1155, 2017.

[234] I. I. Badshah, S. Brown, L. Weibel, et al. Differential expression of secreted factors SOSTDC1 and ADAMTS8 cause profibrotic changes in linear morphoea fibroblasts. *British Journal of Dermatology*, 180(5):1135–1149, 2019.

[235] G A C Murrell, M. J. O. Francis, and Linda Bromley. The collagen changes of Dupuytren's Contracture. *The Journal of Hand Surgery*, 16B:263–266, 1991.

[236] Marie A. Badalamente, Lawrence C. Hurst, Stony Brook, Vincent R. Hentz, and Palo Alto. Collagen as a clinical target: Nonoperative treatment of Dupuytren's disease. *Journal of Hand Surgery*, 27(5):788–798, 2002.

[237] Marike M. van Beuge, Evert Jan P.M. ten Dam, Paul M.N. Werker, and Ruud A. Bank. Matrix and cell phenotype differences in Dupuytren's disease. *Fibrogenesis and Tissue Repair*, 9(1): 1–11, 2016.

[238] A. Berndt, H. Kosmehl, D. Katenkamp, and V. Tauchmann. Appearance of the myofibroblastic phenotype in dupuytren's disease is associated with a fibronectin, laminin, collagen type IV and tenascin extracellular matrix. *Pathobiology*, 62(2):55–58, 1994.

[239] Annie Pardo, Sandra Cabrera, Mariel Maldonado, and Moisés Selman. Role of matrix metalloproteinases in the pathogenesis of idiopathic pulmonary fibrosis. *Respiratory Research*, 17(1), 2016.

[240] Yang Zhou, Hong Peng, Huanxing Sun, et al. Chitinase 3-like 1 suppresses injury and promotes fibroproliferative responses in mammalian lung fibrosis. *Science Translational Medicine*, 6(240), 2014.

[241] Haijun Huang, Tiangang Wu, Jian Mao, et al. CHI3L1 is a liver-enriched, noninvasive biomarker that can be used to stage and diagnose substantial hepatic fibrosis. *OMICS A Journal of Integrative Biology*, 19(6):339–345, 2015.

[242] Kameswaran Surendran, Susan Schiavi, and Keith A. Hruska. Wnt-dependent $\beta$-catenin signaling is activated after unilateral ureteral obstruction, and recombinant secreted frizzled-related protein 4 alters the progression of renal fibrosis. *Journal of the American Society of Nephrology*, 16(8):2373–2384, 2005.

[243] Michalis Mastri, Zaeem Shah, Karin Hsieh, et al. Secreted Frizzled-related protein 2 as a target in antifibrotic therapeutic intervention. *American Journal of Physiology - Cell Physiology*, 306 (6):531–539, 2014.

[244] Camilo Andrés Alfonso-Rodríguez, Ingrid Garzón, Juan Garrido-Gómez, et al. Identification of histological patterns in clinically affected and unaffected palm regions in dupuytren's disease. *PLoS ONE*, 9(11), 2014.

[245] Kerstin Becker, Sabine Siegert, Mohammad Reza Toliat, et al. Meta-analysis of genome-wide association studies and network analysis-based integration with gene expression data identify new suggestive loci and unravel a Wnt-centric network associated with Dupuytren's disease. *PLoS ONE*, 11(7):1–18, 2016.

[246] Steven S. Shin, Chuanju Liu, Eric Y. Chang, Cathy S. Carlson, and Paul E. Di Cesare. Expression of bone morphogenetic proteins by Dupuytren's fibroblasts. *Journal of Hand Surgery*, 29(5):809–814, 2004.

[247] Laurence Preisser, Charline Miot, Hélène Le Guillou-Guillemette, et al. IL-34 and macrophage colony-stimulating factor are overexpressed in hepatitis C virus fibrosis and induce profibrotic macrophages that promote collagen synthesis by hepatic stellate cells. *Hepatology*, 60(6):1879–1890, 2014.

[248] Munehiro Yamaguchi. IL-34 induces lung fibrosis via M2 macrophages differentiation in mouse lung injured model. *Journal of Allergy and Clinical Immunology*, 141(2):AB174, 2018.

[249] Catherine M. Dreiza, Padmini Komalavilas, Elizabeth J. Furnish, et al. The small heat shock protein, HSPB6, in muscle function and disease. *Cell Stress and Chaperones*, 15(1):1–11, 2010.

[250] Georgios Theocharidis, Zoe Drymoussi, Alexander P. Kao, et al. Type VI collagen regulates dermal matrix assembly and fibroblast motility. *Journal of Investigative Dermatology*, 136(1): 74–83, 2016.

[251] Larry Voiles, David E. Lewis, Ling Han, et al. Overexpression of type VI collagen in neoplastic lung tissues. *Oncology Reports*, 32(5):1897–1904, 2014.

[252] Qihuang Jin, Li Rong Yu, Lifeng Wang, et al. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *EMBO Journal*, 30(2):249–262, 2011.

[253] E. T. Manning, T. Ikehara, T. Ito, J. T. Kadonaga, and W. L. Kraus. p300 Forms a Stable, Template-Committed Complex with Chromatin: Role for the Bromodomain. *Molecular and Cellular Biology*, 21(12):3876–3887, 2001.

[254] Jonas Carlsson Almlöf, Andrei Alexsson, Juliana Imgenberg-Kreuz, et al. Novel risk genes for systemic lupus erythematosus predicted by random forest classification. *Scientific Reports*, 7 (1):1–11, 2017.

[255] Shinsuke Yasuda. Emerging targets for the treatment of lupus erythematosus: There is no royal road to treating lupus. *Modern Rheumatology*, 29(1):60–69, 2019.

[256] Zhili Niu, Pingan Zhang, and Yongqing Tong. Value of HLA-DR genotype in systemic lupus erythematosus and lupus nephritis: A meta-analysis. *International Journal of Rheumatic Diseases*, 18(1):17–28, 2015.

[257] Hiroshi Furukawa, Aya Kawasaki, Shomi Oka, et al. Human leukocyte antigens and systemic lupus erythematosus: A protective role for the HLA-DR6 alleles DRB1*13:02 and *14:03. *PLoS ONE*, 9(2):1–7, 2014.

[258] Minori Koga, Aya Kawasaki, Ikue Ito, et al. Cumulative association of eight susceptibility genes with systemic lupus erythematosus in a Japanese female population. *Journal of Human Genetics*, 56(7):503–507, 2011.

[259] Adeeb H. Rahman and Robert A. Eisenberg. The role of toll-like receptors in systemic lupus erythematosus. *Springer Seminars in Immunopathology*, 28(2):131–143, 2006.

[260] Maddalena Larosa, Margherita Zen, Mariele Gatto, Diogo Jesus, Elisabetta Zanatta, Luca Iaccarino, Luis Inês, and Andrea Doria. IL-12 and IL-23/Th17 axis in systemic lupus erythematosus. *Experimental Biology and Medicine*, 244(1):42–51, 2019.

[261] Biola M. Javierre, Agustin F. Fernandez, Julia Richter, et al. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Research*, 20(2):170–179, 2010.

[262] N. Hu, X. Qiu, Y. Luo, et al. Abnormal histone modification patterns in lupus CD4+ T cells. *The Journal of Rheumatology*, 5(35):804–810, 2008.

[263] Christian Michael Hedrich. Epigenetics in SLE. *Current Rheumatology Reports*, 19(9), 2017.

[264] Namrata Singh, Duncan B. Johnstone, Kayla A. Martin, Italo Tempera, Mariana J. Kaplan, and Michael F. Denny. Alterations in nuclear structure promote lupus autoimmunity in a mouse model. *Disease Models & Mechanisms*, 9(8):885–897, 2016.

[265] Elissa K. Deenick, Danielle T. Avery, Anna Chan, et al. Naive and memory human B cells have distinct requirements for STAT3 activation to differentiate into antibody-secreting plasma cells. *The Journal of Experimental Medicine*, 210(12):2739–2753, 2013.

[266] E. E. Solomou, Y.-T. Juang, M. F. Gourley, G. M. Kammer, and G. C. Tsokos. Molecular Basis of Deficient IL-2 Production in T Cells from Patients with Systemic Lupus Erythematosus. *The Journal of Immunology*, 166(6):4216–4222, 2001.

[267] Roberto Manetti, Paola Parronchi, Maria Grazia Giudizi, Marie-pierre Piccinni, Enrico Maggi, Giorgio Trinchieri, and Sergio Romagnani. Natural Killer Cell Stimulatory Factory (Interleucin 12 [IL-12]) Induces Responses and Inhibits the Dedevelopment of IL-4-producing Th Cells. *Journal of Experimental Medicine*, 177(4):1199–1204, 1993.

[268] Chiara Nembrini, Brian Abel, Manfred Kopf, and Benjamin J Marsland. Strong TCR signaling, TLR ligands, and cytokine redundancies ensure robust development of type 1 effector T cells. *Journal of Immunology*, 176(12):7180–8, 2006.

[269] Ronald F. van Vollenhoven, Bevra H. Hahn, George C. Tsokos, et al. Efficacy and safety of ustekinumab, an IL-12 and IL-23 inhibitor, in patients with active systemic lupus erythematosus: results of a multicentre, double-blind, phase 2, randomised, controlled study. *The Lancet*, 392(10155):1330–1339, 2018.

[270] Silke Gillessen, Daisy Carvajal, Frank J Podlaski, et al. Mouse interleukin-12 (IL-12) p40 homodimer: potent IL-12 antagonist. *European Journal of Immunology*, 25:200–206, 1995.

[271] F P Heinzel, A M Hujer, F N Ahmed, and R M Rerko. In vivo production and function of IL-12 p40 homodimers. *Journal of Immunology*, 158(9):4381–8, 1997.

[272] Cordula Haas, Bernhard Ryffel, and Michel Le Hir. IFN- $\gamma$ Receptor Deletion Prevents Autoantibody Production and Glomerulonephritis in Lupus-Prone (NZB x NZW)F1 Mice. *The Journal of Immunology*, 160:3713–3718, 1998.

[273] M. Harigai, M. Kawamoto, M. Hara, T. Kubota, N. Kamatani, and N. Miyasaka. Excessive Production of IFN- in Patients with Systemic Lupus Erythematosus and Its Contribution to Induction of B Lymphocyte Stimulator/B Cell-Activating Factor/TNF Ligand Superfamily-13B. *The Journal of Immunology*, 181(3):2211–2219, 2014.

[274] Shaun W. Jackson, Holly M. Jacobs, Tanvi Arkatkar, et al. B cell IFN-$\gamma$ receptor signaling promotes autoimmune germinal centers via cell-intrinsic induction of BCL-6. *The Journal of Experimental Medicine*, 213(5):733–750, 2016.

[275] Diana Gómez-Martín, Mariana Díaz-Zamudio, José Carlos Crispín, and Jorge Alcocer-Varela. Interleukin 2 and systemic lupus erythematosus. Beyond the transcriptional regulatory net abnormalities. *Autoimmunity Reviews*, 9(1):34–39, 2009.

[276] Masayuki Mizui and George C. Tsokos. Low-Dose IL-2 in the Treatment of Lupus. *Current Rheumatology Reports*, 18(11), 2016.

[277] Jacquelyn A. Gorman, Christian Hundhausen, Mackenzie Kinsman, et al. The Tyk2-p1104A autoimmune protective variant limits coordinate signals required to generate specialized t cell subsets. *Frontiers in Immunology*, 10(JAN):1–16, 2019.

[278] Marie-céline Deau, Alain Fischer, Sven Kracker, et al. A human immunodeficiency caused by mutations in the PIK3R1 gene. *The Journal of Clinical Investigation*, 124(April):3923–3928, 2015.

[279] Guo Min Li, Hai Mei Liu, Wan Zhen Guan, et al. A mutation in PIK3CD gene causing pediatric systemic lupus erythematosus: A case report. *Medicine*, 98(18):e15329, 2019.

[280] C. M. Black, K. I. Welsh, A. Fielder, G. R.V. Hughfs, and J. R. Batchelor. HLA antigens and Bf allotypes in SLE: evidence for the association being with specific haplotypes. *Tissue Antigens*, 19(2):115–120, 1982.

[281] N. Tsuchiya, A. Kawasaki, B. P. Tsao, T. Komata, J. M. Grossman, and K. Tokunaga. Analysis of the association of HLA-DRB1, TNF*a* promoter and TNFR2 (TNFRSF1B) polymorphisms with SLE using transmission disequilibrium test. *Genes and Immunity*, 2(6):317–322, 2001.

[282] Shalabh Singhal, Vaishali Shah, Chris Crater, et al. Oral selective tyrosine kinase 2 (TYK2) inhibition with BMS-986165 in patients with systemic lupus erythematosus: a Phase 2, randomized, double-blind, placebo-controlled study. *Lupus Science and Medicine*, 6(1):1–13, 2019.

*All we have to decide is what to do with the time that is given us.*

Gandalf

# 8 Appendices

This section of the thesis contains a set of figures that are referenced in the main text, but were not included within the results chapters. Broadly speaking, the figures contained here are meant to provide more detail to the primary results discussed in the main thesis body. For ease of referencing the appropriate results chapter, the figures included in this section are as follows:

I. Two tables (8.0.1 and 8.0.2) summarising the top 100 genes prioritised by Pi analysis of EO and LO myasthenia gravis-associated SNPs, respectively, as discussed in the *Application of the prioritisation pipeline to myasthenia gravis chapter*;

II. A table (8.0.3) summarising the statistical fine-mapping results of the CTLA4 locus, also in the myasthenia gravis chapter;

III. A table (8.0.4) summarising the proteomic results discussed in the *Genomic validation of prioritisation outputs in the context of Dupuytren's disease* chapter;

IV. A bar plot (Figure 8.0.1) showing the discovery versus supervised mode comparison of Priority Index outputs, as discussed in the *Computational prioritisation of genes* chapter.

**Table 8.0.1: Top Pi-prioritised genes based on analysis of EO MG-associated SNPs.** Top 100 genes prioritised by Pi analysis of MG-associated SNPs, limited to only those SNPs associated with the early-onset form of the disease and meeting GWAS p-value of $p < 5.0 \times 10^{-5}$. SNPs were sourced from the 2015 study by Renton and colleagues[69] as well as the 2012 study by Gregersen and colleagues[68]; an LD cutoff of $r^2 > 0.8$ was used, in European populations. Pi prioritisation was limited to only the use of genomic predictors.

| Gene | Rank | nGene | eGene | cGene | Gene | Rank | nGene | eGene | cGene |
|---|---|---|---|---|---|---|---|---|---|
| TNIP1 | 1 | 1 | 0 | 5 | PFDN1 | 54 | 0 | 0 | 0 |
| CCNG1 | 2 | 0 | 0 | 9 | TINF2 | 55 | 0 | 0 | 0 |
| S100A11 | 3 | 0 | 0 | 11 | TRAF2 | 56 | 0 | 0 | 0 |
| AP4B1 | 4 | 0 | 0 | 10 | VIM | 57 | 0 | 0 | 0 |
| DCLRE1B | 5 | 0 | 0 | 10 | HMOX2 | 58 | 0 | 0 | 0 |
| PPP1R1B | 6 | 0 | 0 | 11 | TERF2IP | 59 | 0 | 0 | 0 |
| S100B | 7 | 0 | 0 | 0 | POT1 | 60 | 0 | 0 | 0 |
| ANXA6 | 8 | 1 | 0 | 7 | IQCB1 | 61 | 0 | 0 | 0 |
| LZTS2 | 9 | 0 | 0 | 0 | AP4M1 | 62 | 0 | 0 | 0 |
| MYC | 10 | 0 | 0 | 0 | STK4 | 63 | 0 | 0 | 0 |
| IKZF3 | 11 | 1 | 1 | 7 | FNBP1L | 64 | 0 | 0 | 0 |
| MDM2 | 12 | 0 | 0 | 0 | ABHD17A | 65 | 0 | 0 | 0 |
| KRT40 | 13 | 0 | 0 | 0 | CDC73 | 66 | 0 | 0 | 0 |
| TP53 | 14 | 0 | 0 | 0 | MDFI | 67 | 0 | 0 | 0 |
| NCL | 15 | 0 | 0 | 0 | TXLNA | 68 | 0 | 0 | 0 |
| TFAP2C | 16 | 0 | 0 | 0 | DNAJB11 | 69 | 0 | 0 | 0 |
| FBF1 | 17 | 0 | 0 | 0 | CFAP53 | 70 | 0 | 0 | 0 |
| ANXA1 | 18 | 0 | 0 | 0 | CDIP1 | 71 | 0 | 0 | 0 |
| SPERT | 19 | 0 | 0 | 0 | FAM161A | 72 | 0 | 0 | 0 |
| PIN1 | 20 | 0 | 0 | 0 | IKZF1 | 73 | 0 | 0 | 0 |
| PNMA1 | 21 | 0 | 0 | 0 | TRAF3 | 74 | 1 | 0 | 6 |
| RBPMS | 22 | 0 | 0 | 0 | TRIM42 | 75 | 0 | 0 | 0 |
| KRTAP10-7 | 23 | 0 | 0 | 0 | SNORD17 | 76 | 0 | 0 | 9 |
| TFIP11 | 24 | 0 | 0 | 0 | UBE3C | 77 | 0 | 0 | 8 |
| CDKN2A | 25 | 0 | 0 | 0 | KANSL2 | 78 | 0 | 0 | 0 |
| CDKN1A | 26 | 0 | 0 | 0 | TK1 | 79 | 0 | 0 | 0 |
| PPP1CA | 27 | 0 | 0 | 0 | TCAF1 | 80 | 0 | 0 | 0 |
| ATF2 | 28 | 0 | 0 | 0 | RCOR3 | 81 | 0 | 0 | 0 |
| SNX5 | 29 | 0 | 0 | 9 | GEM | 82 | 0 | 0 | 0 |
| S100A1 | 30 | 0 | 0 | 0 | TMEM17 | 83 | 0 | 0 | 0 |
| NFATC4 | 31 | 0 | 0 | 0 | SMAD9 | 84 | 0 | 0 | 0 |
| DLGAP4 | 32 | 0 | 0 | 0 | TSSK3 | 85 | 0 | 0 | 0 |
| AP4S1 | 33 | 0 | 0 | 0 | ARMC7 | 86 | 0 | 0 | 0 |
| ROBO2 | 34 | 0 | 0 | 0 | TYRO3 | 87 | 0 | 0 | 0 |
| BUB1 | 35 | 0 | 0 | 0 | S100A2 | 88 | 0 | 0 | 0 |
| GOLGA2 | 36 | 0 | 0 | 0 | CEP57L1 | 89 | 0 | 0 | 0 |
| SLC39A10 | 37 | 0 | 0 | 0 | TCEANC2 | 90 | 0 | 0 | 0 |
| TCTN3 | 38 | 0 | 0 | 0 | S100A6 | 91 | 0 | 0 | 0 |
| TERF2 | 39 | 0 | 0 | 0 | TEX11 | 92 | 0 | 0 | 0 |
| UBE2I | 40 | 0 | 0 | 0 | BYSL | 93 | 0 | 0 | 0 |
| ACTB | 41 | 0 | 0 | 0 | TP53BP2 | 94 | 0 | 0 | 0 |
| SCLT1 | 42 | 0 | 0 | 0 | TNIP3 | 95 | 0 | 0 | 0 |
| BUB1B | 43 | 0 | 0 | 0 | PRKRA | 96 | 0 | 0 | 0 |
| PRKDC | 44 | 0 | 0 | 0 | NEK6 | 97 | 0 | 0 | 0 |
| TERF1 | 45 | 0 | 0 | 0 | UBQLN4 | 98 | 0 | 0 | 0 |
| GRB2 | 46 | 0 | 0 | 0 | ATPAF2 | 99 | 0 | 0 | 0 |
| NDC80 | 47 | 0 | 0 | 0 | MCM10 | 100 | 0 | 0 | 0 |
| CDC37 | 48 | 0 | 0 | 0 | | | | | |
| AP4E1 | 49 | 0 | 0 | 0 | | | | | |
| HS6ST1 | 50 | 0 | 0 | 9 | | | | | |
| ZBTB25 | 51 | 0 | 0 | 0 | | | | | |
| TUBB | 52 | 0 | 0 | 0 | | | | | |
| PSMD4 | 53 | 0 | 0 | 0 | | | | | |

**Table 8.0.2: Top Pi-prioritised genes based on analysis of LO MG-associated SNPs.** Top 100 genes prioritised by Pi analysis of MG-associated SNPs, limited to only those SNPs associated with the late-onset form of the disease and meeting GWAS p-value of $p < 5.0 \times 10^{-5}$. SNPs were sourced from the 2015 study by Renton and colleagues[69] as well as the 2015 study by Seldin and colleagues[70]; an LD cutoff of $r^2 > 0.8$ was used, in European populations. Pi prioritisation was limited to only the use of genomic predictors.

| Gene | Rank | nGene | cGene | Gene | Rank | nGene | cGene |
|---|---|---|---|---|---|---|---|
| AP4B1 | 1 | 0 | 10 | ZBTB22 | 54 | 0 | 0 |
| TMEM241 | 2 | 0 | 10 | AP3B1 | 55 | 0 | 0 |
| DCLRE1B | 3 | 0 | 10 | FOXK2 | 56 | 0 | 0 |
| AP4S1 | 4 | 0 | 0 | KAT8 | 57 | 0 | 0 |
| CYP20A1 | 5 | 0 | 10 | LAMA1 | 58 | 0 | 0 |
| MDFI | 6 | 0 | 0 | PIK3R1 | 59 | 0 | 0 |
| FNBP1L | 7 | 0 | 0 | IMPDH2 | 60 | 0 | 0 |
| AP4E1 | 8 | 0 | 0 | PHF20 | 61 | 0 | 0 |
| TERF1 | 9 | 0 | 0 | LCMT2 | 62 | 0 | 0 |
| RAPH1 | 10 | 0 | 8 | OGT | 63 | 0 | 0 |
| TERF2IP | 11 | 0 | 0 | WDR5 | 64 | 0 | 0 |
| KANSL2 | 12 | 0 | 0 | HCFC1 | 65 | 0 | 0 |
| SLC39A10 | 13 | 0 | 0 | CDC73 | 66 | 0 | 0 |
| PRKDC | 14 | 0 | 0 | APLN | 67 | 0 | 0 |
| POT1 | 15 | 0 | 0 | C19orf60 | 67 | 0 | 0 |
| BUB1 | 16 | 0 | 0 | C4orf26 | 67 | 0 | 0 |
| AP4M1 | 17 | 0 | 0 | DCANP1 | 67 | 0 | 0 |
| MYC | 18 | 0 | 0 | KIAA0040 | 67 | 0 | 0 |
| TERF2 | 19 | 0 | 0 | LRRN4 | 67 | 0 | 0 |
| TINF2 | 20 | 0 | 0 | ZNF679 | 67 | 0 | 0 |
| BUB1B | 21 | 0 | 0 | NXNL1 | 74 | 0 | 0 |
| SLC9A1 | 22 | 0 | 0 | TCF20 | 75 | 0 | 0 |
| SH3GL2 | 23 | 0 | 0 | FES | 76 | 0 | 0 |
| PSMD4 | 24 | 0 | 0 | KIF18A | 77 | 0 | 0 |
| SH3GL3 | 25 | 0 | 0 | GRAP2 | 78 | 0 | 0 |
| GRB2 | 26 | 0 | 0 | GNMT | 79 | 0 | 0 |
| ALB | 27 | 0 | 0 | ZNF318 | 80 | 0 | 0 |
| CABLES1 | 28 | 0 | 7 | AP1B1 | 81 | 0 | 0 |
| WASL | 29 | 0 | 0 | MCRS1 | 82 | 0 | 0 |
| TP53 | 30 | 0 | 0 | HMGN4 | 83 | 0 | 0 |
| STRAP | 31 | 0 | 0 | OCM2 | 83 | 0 | 0 |
| PGLS | 32 | 0 | 0 | DDX52 | 85 | 0 | 0 |
| ACP1 | 33 | 0 | 0 | RNASEH2B | 86 | 0 | 0 |
| YME1L1 | 34 | 0 | 0 | BUB3 | 87 | 0 | 0 |
| GOLIM4 | 35 | 0 | 0 | EGFR | 88 | 0 | 0 |
| CDK5 | 36 | 0 | 0 | TFB1M | 89 | 0 | 0 |
| TFAP2A | 37 | 0 | 0 | NPAT | 90 | 0 | 0 |
| SGSM2 | 38 | 0 | 0 | WDR76 | 91 | 0 | 0 |
| CEP104 | 39 | 0 | 0 | LASP1 | 92 | 0 | 0 |
| USP13 | 40 | 0 | 0 | RPL10 | 93 | 0 | 0 |
| EZH2 | 41 | 0 | 0 | PSMB1 | 94 | 0 | 0 |
| SMAD3 | 42 | 0 | 0 | XRCC1 | 95 | 0 | 0 |
| TEAD3 | 43 | 0 | 0 | ZWINT | 96 | 0 | 0 |
| KANSL1 | 44 | 0 | 0 | THEG | 97 | 0 | 0 |
| KANSL3 | 45 | 0 | 0 | TMSB10 | 98 | 0 | 0 |
| FOS | 46 | 0 | 0 | AKT2 | 99 | 0 | 0 |
| AFP | 47 | 0 | 0 | RIOK3 | 100 | 0 | 5 |
| TMEM17 | 48 | 0 | 0 | | | | |
| AP2B1 | 49 | 0 | 0 | | | | |
| RDH12 | 50 | 0 | 0 | | | | |
| USP47 | 51 | 0 | 0 | | | | |
| POLG2 | 52 | 0 | 0 | | | | |
| ARPIN | 53 | 0 | 0 | | | | |

**Table 8.0.3: Statistical fine-mapping of the CTLA4 locus.** A table summarising the findings (posterior probabilities) of the fine-mapping analysis of the CTLA4 locus. SNPs were sourced from the study by Renton and colleagues[69], and the PAINTOR tool[110] was used to conduct fine-mapping. Functional annotations for PAINTOR analysis were limited to DNase I hypersensitivity sites (hotspots) and RoadMap assayed narrow peaks in H3K27ac, H3K27me3, H3K4me1, and H3K4me3, given the T-regulatory cell context.

| Chromosome | Position | A0 | A1 | Zscore | Posterior_Prob | rsID |
|---|---|---|---|---|---|---|
| chr2 | 204729153 | C | T | 5.4917369 | 9.81E-01 | rs231770 |
| chr2 | 204693876 | G | T | 4.91942234 | 9.09E-01 | rs231735 |
| chr2 | 204720895 | T | G | 5.35681877 | 2.63E-01 | rs231763 |
| chr2 | 204713891 | G | T | 4.9518798 | 7.42E-03 | rs231811 |
| chr2 | 204691538 | T | C | -4.9331837 | 7.26E-03 | rs6745050 |
| chr2 | 204708646 | C | T | 4.92191828 | 6.97E-03 | rs231804 |
| chr2 | 204712143 | A | G | 5.07039063 | 6.07E-03 | rs231810 |
| chr2 | 204707138 | A | G | -5.0651868 | 6.03E-03 | rs11571303 |
| chr2 | 204707417 | C | T | 5.04905539 | 5.84E-03 | rs231799 |
| chr2 | 204706107 | A | G | 5.03654892 | 5.71E-03 | rs231797 |
| chr2 | 204704702 | G | A | 5.01621264 | 5.51E-03 | rs231796 |
| chr2 | 204723541 | C | T | 5.47765476 | 1.78E-03 | rs231764 |
| chr2 | 204701865 | C | A | 4.97633339 | 1.42E-03 | rs231791 |
| chr2 | 204699952 | C | T | 4.97336677 | 1.42E-03 | rs231789 |
| chr2 | 204694805 | T | C | 4.93782175 | 1.33E-03 | rs3115969 |
| chr2 | 204694784 | G | A | 4.93488013 | 1.33E-03 | rs3116514 |
| chr2 | 204723750 | A | G | -5.3110423 | 1.32E-03 | rs1427679 |
| chr2 | 204694465 | A | G | 4.91320392 | 1.31E-03 | rs231733 |
| chr2 | 204694611 | G | A | 5.23165399 | 1.15E-03 | rs3116513 |
| chr2 | 204700574 | G | T | 4.97723199 | 5.13E-05 | rs231790 |
| chr2 | 204730901 | C | T | -5.3348947 | 4.90E-05 | rs231790 |
| chr2 | 204694974 | C | T | 4.93955763 | 4.82E-05 | rs863603 |
| chr2 | 204694952 | A | G | -4.9182979 | 4.70E-05 | rs12986861 |
| chr2 | 204730465 | T | C | -5.3012698 | 4.69E-05 | rs736611 |
| chr2 | 204729795 | A | G | -5.2981217 | 4.67E-05 | rs1427680 |
| chr2 | 204734487 | C | T | 5.06706328 | 6.29E-06 | rs231779 |
| chr2 | 204702249 | A | G | 4.9836373 | 9.56E-07 | rs231792 |
| chr2 | 204694280 | G | A | 4.91695667 | 8.74E-07 | rs231734 |

**Table 8.0.4: Proteomic analysis of CBP30-treated myofibroblasts.** A table summarising the results of a proteomic analysis of CBP30-treated myofibroblasts collected from the 6 Dupuytren's disease donors, following 3 days of probe treatment (versus DMSO). In total, 40 proteins were deemed as significant (fold-change $> 1.5$, $p < 0.05$), including 14 down-regulated and 26 up-regulated.

| Accession | gene | p-value | FC | Highest mean condition | Lowest mean condition |
|---|---|---|---|---|---|
| O95793;Q9NUL3 | STAU1 | 0.03233768 | 4.52054888 | CBP | DMSO |
| Q6ZVM7 | TOM1L2 | 0.00361044 | 3.90497575 | CBP | DMSO |
| Q9UK45 | LSM7 | 0.03853575 | 3.14894104 | CBP | DMSO |
| Q15397 | PUM3 | 0.03028362 | 2.66235812 | CBP | DMSO |
| Q9NUQ9;Q9H0Q0 | FAM49B | 0.00387934 | 2.33318362 | CBP | DMSO |
| O43148 | RNMT | 0.02167447 | 2.32691256 | CBP | DMSO |
| Q8TBF2 | FAM213B | 0.00665922 | 2.24123865 | CBP | DMSO |
| P62140 | PPP1CB | 0.01781555 | 2.02546554 | CBP | DMSO |
| O43567 | RNF13 | 0.03304005 | 1.87859951 | CBP | DMSO |
| Q8WW22 | DNAJA4 | 0.02939821 | 1.86112869 | CBP | DMSO |
| O60488 | ACSL4 | 0.03014296 | 1.86089854 | CBP | DMSO |
| Q0JRZ9 | FCHO2 | 0.0019433 | 1.85095082 | CBP | DMSO |
| Q14980 | NUMA1 | 0.02209068 | 1.74752711 | CBP | DMSO |
| P61018 | RAB4B | 0.0417547 | 1.7152906 | CBP | DMSO |
| Q9Y6Y8 | SEC23IP | 0.00045237 | 1.68284815 | CBP | DMSO |
| Q10713 | PMPCA | 0.02974603 | 1.65043309 | CBP | DMSO |
| Q01105;P0DME0 | SET | 0.01927123 | 1.63655476 | CBP | DMSO |
| Q9UBG0 | MRC2 | 0.00362623 | 1.61062994 | CBP | DMSO |
| O95479 | H6PD | 0.03025296 | 1.60636407 | CBP | DMSO |
| Q9BWS9 | CHID1 | 0.01410519 | 1.57458707 | CBP | DMSO |
| P35222 | CTNNB1 | 0.03965344 | 1.54700382 | CBP | DMSO |
| O60716 | CTNND1 | 0.00111641 | 1.53851245 | CBP | DMSO |
| Q9BRF8 | CPPED1 | 0.00032223 | 1.53612182 | CBP | DMSO |
| Q7L311 | ARMCX2 | 0.00323743 | 1.52687707 | CBP | DMSO |
| A5YKK6 | CNOT1 | 0.03710004 | 1.5144819 | CBP | DMSO |
| P61769 | B2M | 0.01703301 | 1.50300152 | CBP | DMSO |
| P42226 | STAT6 | 0.01584367 | -1.5021581 | DMSO | CBP |
| P12109 | COL6A1 | 0.00346553 | -1.5179259 | DMSO | CBP |
| Q15582 | TGFBI | 0.04838408 | -1.5876565 | DMSO | CBP |
| Q9NQX4 | MYO5C | 0.00381154 | -1.615794 | DMSO | CBP |
| Q7KZ85 | SUPT6H | 0.03713178 | -1.6169386 | DMSO | CBP |
| P12111 | COL6A3 | 0.0158424 | -1.6247965 | DMSO | CBP |
| O14495 | PLPP3 | 0.00364053 | -1.6550422 | DMSO | CBP |
| P17096 | HMGA1 | 0.03744463 | -1.7048581 | DMSO | CBP |
| P12110 | COL6A2 | 0.00169552 | -1.7516379 | DMSO | CBP |
| P02751 | FN1 | 0.00954124 | -1.7663536 | DMSO | CBP |
| P28300 | LOX | 3.69E-05 | -1.8799575 | DMSO | CBP |
| P32322 | PYCR1 | 0.01767419 | -1.9516708 | DMSO | CBP |
| Q96CG8 | CTHRC1 | 0.01995653 | -2.5316061 | DMSO | CBP |
| P29373;P29762 | CRABP2 | 0.00041649 | -2.9466584 | DMSO | CBP |

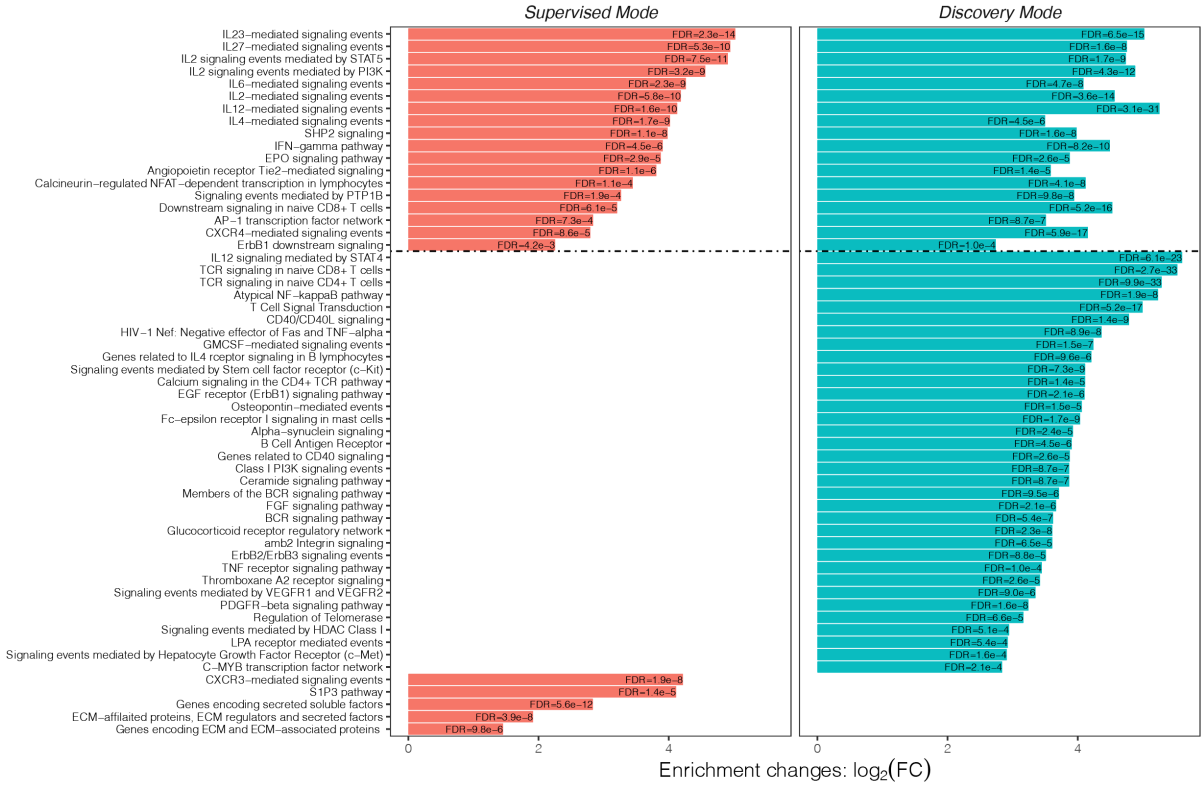



**Figure 8.0.1: Discovery versus supervised mode pathway enrichment.** A plot comparing discovery versus supervised mode enriched REACTOME pathways of top 1% genes of the Priority Index output for multiple sclerosis, meeting an FDR $< 0.005$ cutoff.