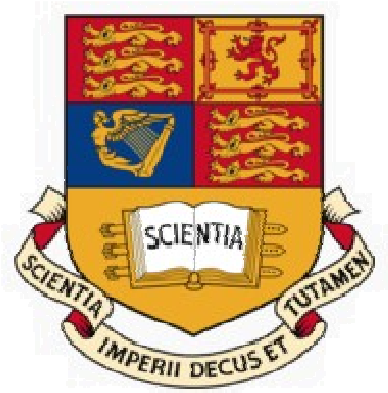
Pathway and gene-based analysis of genome wide association studies (GWAS)

Charikleia Eleftherochorinou

This thesis is presented for the

Degree of Doctor of Philosophy



Department of Paediatrics, Division of Medicine

Centre of Epidemiology and Biostatistics, School of Public Health

Imperial College

London, 2011

Abstract

My PhD thesis comprises the development and application of novel strategies to analyse genome-wide association studies (GWAS) in the context of functional pathways. I propose pathway and gene-centric methodologies as complementary tools to the conventional single-marker analyses to mine further the GWAS hidden information. I developed the cumulative trend (*CT*) test statistic that assesses the cumulative genetic variation of single nucleotide polymorphisms (SNPs) of genes that interact in the same biological pathway and tests the association between a disease and the pathway as an entity. I applied this methodology to the genotypic data of the Wellcome Trust Case Control Consortium (WTCCC) study on Crohn's disease (CD), type I diabetes (T1D), rheumatoid arthritis (RA), bipolar disorder, hypertension, type II diabetes, coronary artery disease; I identified highly significant associations between the autoimmune diseases (CD, T1D, RA) and inflammatory pathways; almost no association was identified between the same pathways and the non-inflammatory conditions. I extended my approach to a pathway-based gene stability selection methodology, which selects associated genes in the context of associated pathways. This methodology can be used to prioritise genes for follow up studies. I applied it on two GWAS of RA with different ethnic background and typed on different platforms and I demonstrated replication at the pathway, gene and *in-silico* functional levels. I finally extended my approach on family trios designed GWAS. I applied it on two case-control and family trio datasets of Kawasaki disease (KD). I explored the association between the TGF-β pathway and KD susceptibility. The involvement of this pathway in KD was further validated at the gene expression and protein levels. My proposed methodologies were tested on real datasets and provided reproducible results, which indicates rigor and robustness. I would therefore suggest their application to single or multiple GWAS as a complement to conventional single-SNP analysis.

Table of contents

Abstract **2**

Table of contents **3**

List of figures **6**

List of tables **8**

Abbreviations **10**

Declaration of originality **11**

Publications **12**

Acknowledgements **13**

Preface **15**

Chapter 1 A background review **18**

1.1 Brief overview of GWAS 19

1.2 The motivation for the pathway-based approach 21

1.3 Pathway analysis scenarios 21

1.4 Pathway-centric statistical approaches 22
1.4.1 Generic framework 22
1.4.2 Broad categories of testing procedures 24

1.5 Technical challenges 28
1.5.1 Pathway collections 30
1.5.2 SNP-to gene mapping 32
1.5.3 LD pattern 32
1.5.4 Size matters 33
1.5.5 Significant hits 33
1.5.6 Overlapping pathways 34
1.5.7 *P*-value threshold bias 37
1.5.8 Genotyping platform and genome coverage 37

1.6 Limitations 39
1.6.1 Associations that will be missed 39
1.6.2 Discordant findings 40
1.6.3 Lack of replication 40
1.6.4 Follow up on the results 41

1.7 Advantages and advances 42
1.7.1 Overcoming the limitations in single-SNP analysis 42
1.7.2 Robustness in genetic heterogeneity 43
1.7.3 Insight into disease 43

1.8 Summary 45

Chapter 2 Pathway-driven analysis methodology 46

2.1 Hypothesis 47

2.2 Aims 48

2.3 Data 49

2.3.1 GWAS datasets 49

2.3.2 Pathway selection 51

2.4 Test for pathway association - cumulative trend test statistic 52

2.5 Variable selection and logistic regression 57

2.6 Model evaluation 58

2.7 Adverse and protective genotype analysis 59

2.8 Summary 60

Chapter 3 Application to three auto-immune diseases 61

3.1 Shared and unique pathways 62

3.2 Probing the signal of association 64

3.3 Identifying key genes in associated pathways 70

3.4 Genetic prediction of disease risk 73

3.5 Validation Study 73

3.6 Visualisation of genomic risk 75

3.7 Discussion 79

3.8 Summary 81

Chapter 4 Pathway-driven gene stability selection 82

4.1 Introduction 83

4.2 Pathway-based analysis 86

4.3 Gene- based stability selection 89

4.4 Estimation of the significance of the observed overlap between studies 95

4.5 Gene Ontology functional enrichment analysis 97

4.6 Summary 97

Chapter 5 Application on two GWAS of rheumatoid arthritis 98

5.1 Pathway associations 99

5.2 Gene selection on associated pathways 101

5.3 Functional enrichment of selected genes 111

5.4 Discussion 117

5.5 Summary 120

Chapter 6 Application to Kawasaki disease 121

6.1 Introduction 122
6.2 Data and Methods 123
6.2.1 Clinical perspective 123
6.2.2 Subjects 124
6.2.3 Materials 125
6.2.4 Methods 128
6.3 Results 130
6.3.1 TGF-β signaling pathway and KD susceptibility 130
6.3.2 TGF-β signaling pathway and coronary artery outcome 133
6.3.3 TGF-β signaling pathway and aortic root dimension 135
6.3.4 TGF-β signaling pathway and IVIG treatment response in the US subjects 135
6.3.5 TGF-β pathway transcript abundance and plasma levels in acute and convalescent KD 137
6.4 Discussion 140
6.5 Summary 142
Chapter 7 Conclusions 143
7.1 What has this thesis achieved 144
7.2 Future directions 146
7.3 Closing remark 147
Bibliography 149

List of figures

Figure 1. 200 pathways as downloaded from KEGG and their gene sharing as a fraction of the smallest pathway under comparison....36
Figure 2. Inflammatory response to a pathogen....48
Figure 3. Pipeline of the pathway-based strategy for the analysis of GWAS data....51
Figure 4. Histogram, density function and QQ plots of various distribution fits to permutation data of the cumulative trend test statistic....55
Figure 5. Skew normal distribution fit for 1,000 and 10,000 permutations for the IL1 pathway with CD....57
Figure 6. Fraction of the associated pathways selected on average by variable selection....65
Figure 7. Genes identified by variable selection in all 10 folds of cross-validation for T1D, RA and CD....71
Figure 8. ROC curves showing the average predictive performance of the models fitted for each disease during 10-fold CV....74
Figure 9. The allelic architecture of the logistic regression models of disease risk....76
Figure 10. Ratio of adverse/protective genetic variants carried by cases and controls and predicted log odds of disease....77
Figure 11. Variant SNPs carried by cases with T1D and controls....78
Figure 12. Schematic diagram of the proposed methodology....85
Figure 13 χ_1^2 transformation of the CT statistics in NARAC and WTCCC....87
Figure 14. Overlap of NARAC and WTCCC pathway rank statistics....88
Figure 15. Log-transformed average model size as a function of the HyperLasso penalty f'....91
Figure 16. Gamma prior probability distributions over the penalty *f'*....92
Figure 17. Number of selected gene-blocks as a function of the selection probability cut-off π_{thr}. that defines stable selection....93
Figure 18. The trade-off between *r* common genes and FDR for different π_{thr} and E[size]....94
Figure 19. Fraction of selected genes in WTCCC and NARAC which are common for E[size]=400 at different values of π_{thr}....94
Figure 20. Comparison of typed SNP content per gene block between selected and unselected gene blocks....96
Figure 21. Methodological flowchart of my approach as implemented on two GWAS of RA....99
Figure 22. Similarity matrix plot in terms of gene composition for the 42 identified pathways..102

Figure 23. Stability paths of the 60 and 64 gene-blocks selected in WTCCC (a) and NARAC (b). 103
Figure 24. Stability paths of known RA associations (a,b) and not selected genes (c,d). 104
Figure 25. Cellular location of products of the selected genes. 112
Figure 26. Cellular and functional map of non HLA genes identified by stability selection. 117
Figure 27. Schematic representation of how genetic variation in genes in the second messenger signalling systems cause altered intensity and duration of cellular responses to multiple receptor mediated stimuli. 120
Figure 28. TGF-β signaling pathway 123
Figure 29. The cohorts and the phenotypes that were examined in different stages of analysis. ..124
Figure 30. Subjects analyzed in expression studies 125
Figure 31. Pedigree of a family trio. 128
Figure 32. Significant SNP locations for *TGFB2* (A), *TGFBR2* (B) and *SMAD3* (C) 132
Figure 33. Heatmaps for Lymphchip array and Agilent array results for 5 genes of interest 137
Figure 34. RT-PCR analysis of transcript abundance in whole blood from KD patients during the acute and convalescent phase and association between genotype and gene expression 139
Figure 35. (A) Plasma levels of TGF-β2 in KD patients during the acute and convalescent phase 139
Figure 36. Pathway-based analyses of GWAS in the perspective of an integrative genomics approach 147

List of tables

Table 1. Publicly available servers that will perform pathway analysis on the user's data.....28
Table 2. Generic characteristics of pathway-based methods.....29
Table 3. Publicly available pathway sources from which pathway gene sets can be down loaded as flat text files on the user's desktop.....30
Table 4. Summary of technical challenges in pathway-based analyses strategies of GWAS.....38
Table 5. GWAS genotype data.....52
Table 6. Summary statistics of the overall study and the pathway trend test.....62
Table 7. *P*-values of the pathway statistic over all the examined pathways and the seven diseases. 63
Table 8. Summary statistics of variable selection and logistic regression step.....64
Table 9. Correlation coefficient of –log10 (*P*-value) with pathway size and number of significant hits.....66
Table 10. List of examined pathways with the corresponding numbers of genes and SNPs. 68
Table 11. *P*-values of the pathway statistic over the examined pathways without any SNPs with single-marker $P<5x10^{-7}$.....69
Table 12.Top 42 pathways identified through pathway-based analysis in WTCCC (W) and NARAC (N) studies.....100
Table 13. Pathway collection that the 42 associated pathways were taken from.....101
Table 14. The 60 gene-blocks selected in WTCCC.....105
Table 15. The 64 gene-blocks selected in NARAC.....107
Table 16. Mapping of the union of WTCCC and NARAC selected genes back to the 42 associated pathways. 109
Table 17. The 21 identified genes that replicated in both WTCCC and NARAC and their most selected SNP. 110
Table 18. Gene ontology (GO) process enriched categories as derived by Gorilla for $\pi t_{hr}=0.4$..115
Table 19. Gene ontology (GO) functional enriched categories as derived by Gorilla for $\pi t_{hr}=0.4$ 116
Table 20. Characteristics of subjects in their respective cohorts 125
Table 21. Characteristics of subjects in microarrays, RT-PCR and ELISA.....126
Table 22. Contigency table of transmitted and not-transmitted alleles in family trios.....129
Table 23. Versions of the TGF-β pathway and *CT* test statistic results.....131

Table 24. TDT analysis of genetic variants and KD susceptibility in US/UK/Australian trios (n=451). 133
Table 25. Association of genetic variants and KD susceptibility in Dutch KD subjects and controls 133
Table 26. Case-control analysis of genetic variants and coronary artery aneurysms (CAA) in US subjects 134
Table 27. Analysis of genetic variants and CA Z-max in US KD subjects (n=176) 135
Table 28. Association of genetic variants with Ao Z-max in US KD subjects (n=98) 136
Table 29. Association of genetic variants with IVIG treatment response in US KD subjects 136
Table 30. Transcript abundance patterns for differentially expressed genes in the TGF-β pathway. 140

Abbreviations

Abbreviation	Definition
ANOVA	Analysis Of VAriance
AoR	Aortic root
AUC	Area Under the Curve
BD	Bipolar Disorder
CA	Coronary Artery
CAA	Coronary Artery Aneurysm
CAD	Coronary Artery Disease
cAMP	Cyclic adenosine monophosphate
CD	Crohn's Disease
CT	Cumulative Trend
CV	Cross-Validation
DAG	diacyglycerol
df	Degrees of freedom
eQTL	expression Quantitative Trait locus
FDR	False Discovery Rate
GO	Gene Ontology
HT	Hypertension
IHP	International HapMap Project
IKDGC	International Kawasaki Disease Genetics Consortium
ITP	Inositol triphosphate
ITPR	Inositol triphosphate receptor
kb	kilobase pairs
KD	Kawasaki Disease
LD	Linkage Disequilibrium: the non-random association of alleles at two or more genomic loci
MHC	Major Histocompatibility Complex
NARAC	North American Rheumatoid Arthritis Consortium
PCA	Principal Component Analysis
PCR	Polumerase Chain Reaction
QC	Quality Control
RA	Rheumatoid Arhtiris
ROC	Receiver Operating characteristic Curve
RT-PCR	Real Time PCR
SNP	Single nucleotide polymorphism
T1D	Type I Diabetes
T2D	Type II diabetes
tag-SNPs	A representative single nucleotide polymorphism (SNP) in a region of the genome with high linkage disequilibrium (LD).
TDT	Transmission Disequilibrium Test
TGF-β	Transforming growth factor β
TNF	Tumor Necrosis Factor
WTCCC	Wellcome Trust Case Control Consortium

Declaration of originality

This thesis describes my work undertaken jointly in the group of Prof. Michael Levin and the group of Dr. Lachlan Coin at the Department of Paediatrics and the Department of Epidemiology and Biostatistics at Imperial College, London. It is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy. The work described here has not been submitted for any degree, diploma, or any other qualification. This thesis does not exceed 300, single-sided pages of double spaced text, not including the bibliography and appendices.

This dissertation is the result of my own work and the close supervision and guidance of my supervisors and includes nothing that it is the outcome done in collaboration except what is detailed in the text below.

The pathway collection that is described in Chapter 3 was assembled by Dr. Victoria Wright, who is a Research Associate at the group of Prof. Michael Levin. Ingenuity Pathway Analysis, as described in Chapter 5, was also carried out by Dr. Victoria Wright with the supervision of Prof. Michael Levin. The cellular map of the genes identified in Chapter 5 was put together by Prof. Michael Levin, as well as the plausible scenario that suggests their involvement in rheumatoid arthritis susceptibility.

Microarray data analysis, RT-PCR and TGF-β2 measurements in plasma, as they are described in Chapter 6 were performed by Dr. Chisato Shimizu and Prof. Jane Burns' group, Department of Paediatrics, University of California, San Diego and the International Kawasaki Disease Genetics Consortium. The interpretation of the results, other than the TGF-β pathway association with KD susceptibility was also done by Prof. Jane Burns and Dr. Chisato Shimizu.

Charikleia Eleftherochorinou

London, 29 June 2011

Publications

Publication list arising from the work described in this thesis at the time of submission.

1. **H Eleftherohorinou**, CJ Hoggart, VJ Wright, M Levin, LJM. Coin, **2011**. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways, *Hum Mol Genet.* doi: 10.1093/hmg/ddr248

2. C Shimizu, S Jain, KO Lin, D Molkara, JR Frazer, S Sun, AL Baker, JW Newburger, AH Rowley, ST Shulman, S. Davila, ML Hibberd, D Burgner, WB Breunis, TW Kuijpers, VJ Wright, M Levin, **H Eleftherohorinou**, L Coin, SJ Popper, DA Relman, W Fury, C Lin, S Mellis, AH Tremoulet, JC Burns. **2010.** "Transforming growth factor-beta signaling pathway in patients with Kawasaki disease", *Circulation: Cardiovascular Genetics*, 4(1)16-25

3. **H Eleftherohorinou**, V Wright, CJ Hoggart, AL Hartikainen, MR Jarvelin, LJ Coin, M. Levin, **2009.** Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. *PLoS ONE* 4(11): e8068. doi:10.1371/journal.pone.0008068

Acknowledgements

This is the end of an era! And maybe my one and only opportunity to thank all the people who were by my side on this quite tough intellectual journey called a "PhD".

I feel the greatest need to first thank my supervisor Prof. Michael Levin, who gave me the opportunity to work with him that day of October 2007, when I knocked on his door. I deeply thank him for believing in me from the first instance, for always being supportive, encouraging and there to listen. I want to thank him for being so inspiring, for having big ideas, for giving me perspective, for keeping me enthusiastic and for all the trips that he has given me. I feel exactly the same need to thank my other inspiring supervisor, Dr. Lachlan Coin, who I look up to greatly for his super-bright mind and his admirably deep knowledge of the field. He has been like an elder brother to me at my workplace, very supportive and I am ever so grateful to him! I thank them both for their generosity, for letting me learn from their expertise and for letting me be myself at work.

I really thank the Division of Medicine scholarship committee for funding my research plan for 3 years and especially Prof. Nikolas Mazarakis for all his support and for having faith in me.

I would also like to thank Dr. Clive Hoggart for being instrumental in my knowledge of statistics and although he used to make fun of me in the beginning, he is one of the people that I really enjoy working with.

A very special thank you to Prof. Jane Burns and Dr. Chisato Shimizu for trusting me with their data, for believing in me, for the exciting work we carried out together, for always being willing to collaborate and for inviting me to San Diego.

I would like to say a very big thank you to Dr. Jethro Herberg, for being a supportive third supervisor and for always expressing interest in my well-being. Thank you to Dr. Victoria Wright for always helping me to understand the biology aspects of the work.

A very big thank you to Prof. Marjo-Riitta Jarvelin for trusting me with the NFBC data, always showing interest in my work and for offering me opportunities to present it in meetings and conferences.

I would also like to thank Ernest Turro for our conversations, the laughs and the food (!) that we shared the past two years sitting next to each other.

I want to thank Koulis Papastathis, who was my supervisor in Greece and literally pushed me to come to the UK to follow my dream.

A special thank you from the bottom of my heart to my friends in London, first Agapi Papadamou, who could potentially defend this thesis in my place, after having listened to it a million times; Reem Saleh, for always correcting my use of English and proof-reading my reports even when she left London; Dr. Georgios Tzimiropoulos and Dr. Stefanos Zafeiriou for the constructive feedback and the stimulating conversations on machine learning and statistical approaches and the invaluable advices; Kostas Chalikias, Aliki Stroumpouli, Ed Zukowski, Valentina Cipriani, Jennifer Daothong, Argiris Tzakas, Margarita Stapountzi, Lia Chatzinasiou, who have been by my side throughout this 3-year journey and all of them for having listened to every single presentation of mine, have given me feedback on figure drawings, concept formation, decision making and for always encouraging me to keep on going.

My best friends in Greece, with whom I used to be inseparable, before I left and I miss them so much (!!): Eva, the one and only, for the unforgettable times, every single summer, and for bringing me back to reality every time I was drifting away, Kleopatra for constantly caring and transmitting her "positive" energy, Fotini, Natalia, Vaso and Anna for being by my side, via skype, mobile, emails, pushing me further and for making me forget about the stresses of work!

Most of all, I need to thank my family, my shelter! My brother George, who will always find the way to make me laugh, my father Ilias, my role-model, who inspires me to be a good scientist and my mother Chaido, who inspires me to be a better person! (not that my father doesn't ☺). Thank you for always being there and for being open-minded to letting me go.

Preface

Genome-wide association studies (GWAS) have revolutionized the field of human genetics in the post human genome era and have evolved to powerful new tools currently employed to explore human disease. Typical GWAS are designed to investigate associations between genetic markers, predominantly single-nucleotide polymorphisms (SNPs) and a trait of interest, predominantly a binary disease status. Over 900 GWAS have been completed to date and they have led to the discovery of >1,300 of genomic loci, which are associated with >220 human traits. At the same time, GWAS, albeit their relative success, they have also caused a considerable amount of disappointment in that the delivered findings are disproportional to the cost and effort behind their design; the effect size of the identified loci so far is usually small with odds ratios (OR) varying from 1.1 to 3 and the studies to date have managed to explain ~5-10% of heritability of the common complex diseases under examination.

There is much debate about the source of this 'missing heritability', but one possibility is that any common complex disease, arises on the basis of the cumulative effect of many variants yet to be discovered, because they are either rare (with potentially large effect) or have a small effect (but are potentially common). Such variants are challenging to identify when performing standard analyses of GWAS data that are ultimately limited by the number of genotyped cases in each study.

The conventional statistical analyses of GWAS have largely focused on single-marker analysis, where each typed SNP is tested, independently to other typed SNPs, for its association to the disease. This strategy has been efficient to deliver the most prominent SNP-disease associations; nevertheless it has failed to expose the biological aspects of the polygenic basis of common disease susceptibility, as it is restricted by the multiple hypothesis problem. As hundreds of thousands or even >1M SNP statistical tests of association are carried out simultaneously, a proportion of them is expected to show evidence of association just out of pure chance; this is the multiple hypothesis problem. Stringent corrections need to be applied to the single SNP test statistics in order to adjust their *P*-values accordingly. Such multiple hypothesis adjustments procedures have been proven to allow only SNPs with great main effects to be discovered (e.g. OR>~3) leaving the majority of markers with modest effects not identified.

The conceptual starting point of my work was that the analysis of GWAS data in the context of biological pathways could be a powerful way to overcome the limitations imposed by the stringent significance thresholds in the multiple hypothesis framework. It would also be a logical and biologically-driven basis for analysis compared to single-marker approaches that exploit no prior biological knowledge; I anticipated that such an approach could identify the cumulative signal of multiple moderately associated markers of interacting genes in functionally related pathways. Furthermore, a pathway-based strategy could provide insights into the combined effects of low penetrance variations underlying disease susceptibility and may indicate biomarkers for disease diagnosis.

This thesis describes a novel approach to the analysis of GWAS in the context of biological pathways that could be used as a complement to conventional single SNP analysis strategies. The conceptual path goes from a cumulative trend statistic, that tests the association between the disease and a pathway as an entity, to a gene-based methodology that identifies the key player genes within associated pathways; it then continues to the use of those pathways and genes to predict disease outcome and to then gaining functional insight into disease pathogenesis mechanisms. It is an efficient approach for identifying the broader pathway functional setting and then prioritizing the candidate genes for follow up studies.

In Chapter 1, I present a summary of pathway-driven and gene-based methodologies, where I have tried to provide a comprehensive overview of the field rather than being exhaustive. In the next chapters I describe the methodologies that I developed along with their application on complex diseases in the same order that my intellectual journey evolved.

In Chapter 2, I describe the materials and methods of the pathway-based approach that I developed during the first year of my PhD.

In Chapter 3, I show the application of my methodology to the WTCCC study of seven common diseases, where I examined the association of inflammatory pathways with three autoimmune diseases and I present the original results that provided clear evidence to support my hypothesis.

In Chapter 4, I describe my refined pathway-driven gene stability selection methodology that goes beyond simple pathway association testing to actually identifying the genes within associated pathways that drive predominantly the signal of pathway association.

In Chapter 5, I present the application of the refined methodology on two GWAS of rheumatoid arthritis and I show how the approach provides reproducible results at the three levels of pathway, gene and *in-silico* cellular function.

In Chapter 6, I describe the extension of my methodology to make it applicable on family-trio-based GWAS data, further to the typical case-control design, and I show the work I carried out on exploring the association between the TGF-β pathway and the Kawasaki disease.

In Chapter 7, I summarize the milestones of this thesis, I discuss the conclusions of my pathway and gene-based analyses and I suggest potential future extensions of such analysis strategies.

Chapter 1

A background review

GWAS have emerged as a powerful new approach to explore the genetic basis of complex disease susceptibility. The approach has been made possible by 1) the revolution in high throughput sequencing and genotyping technologies, which have facilitated the way to study large cohorts of cases and controls, 2) the identification of common sequence variation in a population, 3) the databases where genetic information is being deposited, which were made publicly accessible and 4) the HapMap project that revealed ~3,1 million of the ~11 millions of common SNPs within the four major ethnic groups in Europe, Asia and Africa, that is the Yoruba population in Africa, Asian individuals from both Japan and China and samples of northern and western European ancestry.

In this chapter, I briefly touch on the basic design of a GWAS and I provide a short outline of the early analysis strategies and the motivation that led to the development of pathway-based methodologies. I present the underlying hypothesis and the rationale behind pathway-based analysis of GWAS and I describe the generic characteristics of relevant methodologies. I highlight the main technical challenges and I summarize the ways to overcome them as described in the published relevant literature. The main focus on this review is upon pathway-driven approaches, as their complexity is somewhat higher than that of gene-centric methods. However, the latter will also be covered.

1.1 Brief overview of GWAS

A typical GWAS is designed to identify associations between a phenotype of interest on one hand, which is usually a binary disease status, and one or more genetic markers on the other. The typed genetic markers are predominantly SNPs and investigators look for allele-frequency differences between cases (individuals with a disease/condition) and controls (healthy individuals). They systematically investigate tag-SNPs selected to cover the entire human genome without an *a priori* specified hypothesis on the location. The approach that is followed is primarily indirect as these studies do not necessarily assume that the typed SNPs are causal but rather that they may be in close vicinity to functional mutations and therefore associated with the disease due to linkage disequilibrium (LD) - the non-random association of alleles at two or more genomic loci.

Early association studies focused only on SNPs in candidate genes or regions that seemed likely to be associated with the disease because of biological function or previous study results (termed candidate genes studies). The major impediment to the application of these studies at a genome-wide level had been the high cost of genotyping along with the need for a large number of markers to provide human genome coverage. The large number of cases and controls to design a well-powered association study has also been a major hindrance. Early association studies were also limited by the modest number of polymorphisms available and it was not until the completion of phase I of the International HapMap Project (IHP)[1] in 2005 that GWAS flourished. The progress in the high throughput genotyping technology was predominantly the reason why GWAS flourished, as cheap and fast genotyping platforms enabled the genotyping of millions of SNPs on tens of thousands of individuals.

For the systematic search of the entire human genome with a length of approximately 3.3 $x10^9$ base-pairs, in subjects of European ancestry only approximately 300k tag-SNPs are theoretically required to capture most of the genomic variation. The average number of typed SNPs per individual varies from 350k to 1M is some large studies and the number of individuals in a study varies from a couple of hundred to a few thousands. The genotype platforms also very between studies and it is important to note that chips are designed in different ways. For example, the 1M BeadChip Illumina specifically contains 950k tag-SNPs from the IHP project and 100k non-HapMap SNPs. In comparison, Affymetrix SNP Array 6.0 used an "unbiased selection of 482k SNPs from the SNP Array 5.0" and 424k SNPs from various sources (recombination hot-spots, mitochondria, etc) thus not following the tag-SNP approach based on the HapMap project[2]. Both platforms contain SNPs whose minor allele frequency is >5% in the

population. The newly introduced Omni 2.5M and 5M Illumina chips contain SNPs from the HapMap and the 1,000 Genomes Project (http://www.1000genomes.org/); they allow the genotyping of markers with minor allele population frequency as low as 1% providing better genome wide coverage in both common and rare variants .

Modern GWAS, in contrast to several years of relatively unsuccessful attempts, have identified and validated genetic determinants for several common diseases; to name a few, type II diabetes[3], schizophrenia[4], breast cancer[5]; prostate cancer[6], multiple sclerosis[7], rheumatoid arthritis[8] and other common diseases[9]. A rough estimate brings the number of successfully identified susceptibility genomic loci to >1,300 associated with >220 traits (first quarter of 2011) as reported in the catalog of published GWAS[10]. The recent explosion of results from GWAS might seem like a sudden development. Furthermore, the mass of generated data has also posed new challenges for statisticians and analysts who need to address the computational problems by developing novel robust and efficient analysis strategies.[11]

A number of reviews on the contribution of GWAS to our understanding of disease, have highlighted the imbalance between the international effort and cost that has been devoted to undertake them, (often involving many thousands of patients and controls), and the value of delivered results so far.[12-14] There is a general view that the GWAS to date have identified only the most significant associations, often explaining only a small fraction of the genetic contribution to disease. They have so far provided little biological insight into the pathophysiological disease mechanisms leading to each disease and the reason for that is mostly attributed to the standard analysis strategies.[15-17]

The dominant approach to analyzing GWAS data is the single marker analysis, where each SNP is tested independently for its association to the trait. Naturally, when one is conducting simultaneously >500k SNP-disease association statistical tests (as many as the typed SNPs in the study), a fraction of those could show association due to chance, hence stringent "multiple testing" adjustments are required to control the number of false positive conclusions.[18] For many GWAS analysis strategies, the single-SNP association test statistic P-value$<10^{-7}$ is taken as the significance level required to account for multiple testing. For this reason, conventional analysis strategies of GWAS have allowed only the most prominent SNP-disease associations to be discovered. Many of the true associations may be hidden within the expected false positive associations with P-values $10^{-7} < P < 0.05$. There has thus been increasing recognition, that the single SNP analysis needs to be supplemented by novel approaches to fully utilize the huge amount of genotypic information available from GWAS.[17,19]

Several approaches have been suggested to enable multiple SNPs-disease association testing; amongst them are multivariate logistic regression with SNPs as covariates and the binary trait as the response variable and test for the significance of the main effects[20-23]; haplotype methods that test the association between multiple SNPs in genomic proximity and a disease[24-28]; epistatic effect modelling that tests interactions between variants.[16,29-31] The results of multivariate logistic regression and epistasis modelling, albeit promising, have been limited by the huge space of possible marker combinations and haplotype strategies have been restricted by the computational burden of a genome wide scale scanning. Furthermore, these approaches have failed to expose the biological aspects of the polygenic basis of common disease susceptibility, and as a result myself and others have suggested that the joint variants in multiple genes interacting in functional pathways may provide a better understanding of disease pathogenesis.

1.2 The motivation for the pathway-based approach

Biological pathway-driven approaches may be a powerful way to overcome the limitations imposed by the stringent significance thresholds in the multiple hypothesis framework and may serve as a way to restrict the modelling of statistical interactions in the space of biological interactions. Such approaches could adopt and combine components of the aforementioned multiple-SNPs strategies under the umbrella of a "biological pathway". They benefit from the prior biological knowledge on gene function and interaction within pathways to leverage the deeper mining of genomic data. Pathway-centric approaches hypothesize that complex disease pathogenesis is likely to depend on the genetic variation in multiple genes controlling each biological pathway, more or less as a domino effect. The predisposing or protective variants might differ at an individual's level, but the cumulative effect across the genes in the pathway is expected to associate with a disease as a whole.

Pathway-centric methodologies are appealing in that they bridge the biological knowledge with the statistical analysis of GWAS. Compared to single-SNP analysis that might come up with a hit in a gene dessert area and whose function is unknown, gene and pathway associations place the results into a functional perspective, are easily interpretable and can elucidate potential disease mechanisms.

1.3 Pathway analysis scenarios

There are three main scenarios under which a pathway-driven methodology seems the most suitable strategy to pursue. The first is to have already formulated a hypothesis that a pathway is likely to be linked to a condition. In such a scenario, the pathway(s) under examination are usually assembled based on the presumed biological processes involved in the

pathogenesis of a specific disease. The disease linked "pathway" is assembled from known biological interactions between genes in each functional group and then the prior hypothesis is tested via means of a pathway-statistic.[32-35] A typical manifestation of this case is testing inflammatory pathways for their association to autoimmune diseases[33] or the nicotine receptors pathway with smoking behaviour.[34] That is a "hypothesis-driven" pathway-based strategy.

The second scenario involves focusing on a restricted number of SNPs/genes, usually the top hits of a GWAS, and then examining the "enrichment" of those results in functional categories. Investigators that follow this strategy, usually rank the single-SNP results by decreasing significance and they choose an arbitrary *P*-value cut-off (e.g. $P<10^{-4}$ or $P<10^{-3}$) to define the SNPs to be taken forward to enrichment analysis; a list of molecular pathways or a dictionary of cellular processes is also necessary to provide the functional bins that the SNPs will be tested against.[36-42] This strategy is quite popular as it does not require the formulation of a prior hypothesis and in addition it is an effective way to expose the aggregative trends of single-SNP results into molecular pathways. That is the "hypothesis-generating" strategy.

The third scenario is about the testing of several pathways for their link to a trait, with no prior hypotheses or specific focus on a set of SNPs.[36,43,44] This is the "hypothesis-free" strategy as it blindly looks for pathway associations, in the same manner that one would perform single-SNP analysis to look for hits in the genome without any *a priori* speculations. The list of pathways is usually assembled by public databases, as they will be detailed in the following section. This strategy encompasses all available pathway knowledge and exploits a greater fraction of the typed genetic variation as opposed to the "hypothesis-generating" methodology that disregards markers that are not included in the final set. In this respect, this scenario takes into account SNPs that by themselves reach only a nominal association signal (e.g. *P*>10-4 or *P*<10-3) and can provide added information when regarded in the presence of other biologically interacting markers.

1.4 Pathway-centric statistical approaches

The statistical approaches to identifying genetic variation within groups of genes in pathways requires a pathway association test statistic, just as single-SNP approaches require an association test statistic.

1.4.1 Generic framework

The pathway statistic builds upon 3 generic steps: the null hypothesis of no association, the summary of the genetic variation along a pathway and the pathway statistic *per se*.

Null hypothesis. They broadly fall in two categories: the "self-contained" and the "competitive" methods, as nicely outlined in two recent reviews.[45,46] Self-contained methods examine the *absolute* association between a pathway and a trait; that is they test the null hypothesis of no association between the joint effect coming from the set of pathway SNPs and the disease. The competitive methods, on the other hand, examine the *relative* association between the pathway and the trait; they test the null hypothesis of no association between the pathway SNPs and the disease conditioned upon the association of all the other SNPs/pathways in the study. Competitive methods are almost all "enrichment" procedures as described in the second scenario above.

Summary of genetic variation. The genetic variation can either be summarized at one level where all pathway SNPs are considered jointly or at two levels, where the variation per gene is summarized at the first step and then all pathway genes' variation is combined at the second step. This can be implemented either by combining the single-SNP test statistic *P*-values into a single testing framework or by using the raw genotypes.

Pathway test statistic. To assess the significance of a pathway, one needs to use a pathway statistic to compare the observed genetic variation within each pathway against the expected variation under the null hypothesis of no association. The null distribution of no pathway-trait association is usually derived in 3 ways;

(1) It is approximated analytically by a known distribution such as the χ^2, the F or the hypergeometric; hence the pathway test statistic is one of the widely used χ^2, F or the hypergeometric test statistics.

(2) It is approximated empirically by a permutation-based procedure where sample phenotypes are permuted against the genotypes, the genetic variation is re-calculated under each permutation and the pathway *P*-value is defined as the fraction of permuted pathway statistics that are equal or bigger than the observed one.

(3) Genotypes are randomized against phenotypes, in that random SNPs are selected across the genome to create "random" pathways of the same size as the one under examination; the pathway *P*-value is calculated as in (2). This procedure is trying to reflect the "random" genetic variation across the genome, nevertheless it is confounded by the differences in the gene size and LD patterns between the real pathway and the "random" versions of it.[47]

1.4.2 Broad categories of testing procedures

Depending on the combination of null hypothesis, genetic variation summary technique and pathway statistic, I have broadly observed five trends of pathway methodologies. Those categories have evolved based on the published literature so far and they are not exhaustive or restrictive.

***P*-value combination.** These methods are self-contained, they require just the single SNP test statistic *P*-values as input and the pathway statistic is approximated analytically in the majority of the cases by a well-known test-statistic. A general framework for combining single SNP *P*-values assumes that *K P*-values are independent random variables and could be combined into a single statistic using the Fisher's combination test[48], which follows a χ^2 distribution with 2*K* degrees of freedom (*df*), Sidak or the Simes' combination test[43]. An alternative way would be to transform them into χ^2 statistics with 1 *df* and sum them up to a single statistic whose null distribution is a χ^2 distribution with *K* df. Another route is to transform them into z-statistics[49] or to use directly the single-SNP test statistic, which could either be the Cochran-Armitage test for trend[50] or the MAX-statistic[51] and sum them or multiply them. The null distribution can again be approximated by a χ^2 distribution with *K* df, by the multivariate Hotelling's T^2-statistic[52] or empirically by permutations of the phenotype labels. Building upon *P*-value combining methods, the adaptive rank truncated product (ARTP)[34] calculates the product of *K* smallest SNP *P*-values of a gene and then combines the products of the genes of a pathway into a test-statistic for pathway association. This category of methods benefits from the computational speed but assumes independence between the SNPs, which is strongly violated by the underlying LD patterns. Under the violation of this assumption, Fisher's combination method should not be applied, Sidak's and Simes' methods underperform and the sum of χ^2 dependent distributed variables does not follow the null χ^2 distribution with *K* df.

Multiple regression. A straightforward solution is multiple regression using the SNPs of a pathway as covariates and testing the goodness of fit via a test-statistic or via a permutation-based approach on randomized SNP sets. Nevertheless it suffers from the large number of df in the model. Principal component analysis (PCA) is used to summarize the variation across SNPs of a pathway and to reduce the *df* of the multiple logistic regression model. Multiple regression is then performed on a few principal components instead of using all the SNPs. PCA is a useful way to summarize the information across multiple SNPs, but it does not identify the most informative SNPs. The weights of the SNPs on each the first principal components are indicative of a SNP's effect, nevertheless PCA is not a variable selection procedure. Chen *et al*[86] used PCA

to summarize the pathway gene structure and then fitted the eigen-vectors in to a regularized regression model; on a similar basis Gauderman *et al*[53] used the first *K* PCs that explain 80% of total variance as covariates on a multiple logistic regression model. Li *et al*[54] used PCA on weighted scored SNPs before conducting regression analysis with a set of principal components as covariates; the weights were used to reflect the level of LD within loci. These methods are self-contained, use raw genotypes and statistical significance is either assessed by a test statistic that tests for the null hypothesis of all β coefficients being zero or by permutations of phenotypes, where the β coefficients of the fitted model are compared to the null distributions of the β coefficients of the permuted models.

Genetic similarity across individuals. The foundation of these approaches is based upon the genetic similarity across subjects and their test statistic resembles somewhat an analysis of variance (ANOVA) setting. These methods are self-contained, they use raw genotype data and the pathway statistic can be approximated analytically or empirically. In this category of methods, first the genetic similarity of *K* pathway typed markers is measured among all pairs of individuals and a test statistic compares the means between cases and controls. The genetic similarity amongst subjects is scored by "kernel" functions that measure the allele sharing between two individuals. Schaid *et al*[55] used the Zglobal statistic, when Wang *et al*[56] developed a modification of Zglobal by applying weights on SNPs based on a Fourier transformation. These statistics are greatly affected by the scoring of the kernel function; hence they are very powerful in the case of multiple associated SNPs with risk effects in the same direction, but quite weak in an opposite scenario. Mukhopadhyay *et al*[57] developed the kernel-based association test (KBAT) that makes no assumptions about the directions of the SNP effects and does not use the correlation between the SNPs. Wessel *et al*[58] on a similar note used the genetic similarity matrix and an *F*-test statistic. Wu *et al*[59] used logistic kernel-machine regression model to model the genetic distances amongst subjects, incorporated them into a logistic model and tested for association using the Q statistic, which they compared to a χ^2 scaled distribution.

Enrichment. These methodologies are competitive, they use mostly single-SNP *P*-values and statistical significance is assessed either analytically or by permutation of genotypes/phenotypes. As already mentioned, this category of methods focuses only at the top of ranked single SNP *P*-values and tests for the enrichment of those in certain pathways. They require an arbitrary threshold to be set that will define which single-SNP results will be kept for subsequent analysis. The GSEA-SNP[60], modified GSEA[61] and SNP ratio test[44] calculate the enrichment statistic as a proportion of the SNPs in a pathway that are significant compared to

fraction of significant hits in the rest of typed SNPs. Statistical significance is assessed via permutation of phenotypes. ALIGATOR[37] uses a preselected *P*-value threshold to define a set of significantly associated SNPs and counts the number of genes that contain these SNPs, with each gene counted once and instead of permuting phenotypes to derive the null distribution it uses resampling of SNPs.

Web-based enrichment of gene lists. These tools are similar to the previous category in that they also test for "enrichment" in biological pathways, but they differ in that they accept gene-lists as input and as they are web-based they used their already predefined pathway collections and gene annotations. Some widely used tools are summarized in

Table 1. One could use them as in Elbers *et al*[52], where they first selected all SNPs with single-SNP $P<0.003$; they then assigned them to functionally relevant genes using "Prioritizer"[63], which is a Bayesian approach that prioritizes genes by reconstructing functional gene networks based upon interactions from the Human Protein Reference Database (HPRD) , Biomolecular Interaction network Database (BIND) and KEGG. They then collected all those genes and explored 5 different pathway tools "Webgestalt KEGG"[64], GATHER[65], DAVID[66], PANTHER[67] and "Webgestalt BioCarta"[64], to investigate whether they would provide consistent results. DAVID uses functional similarity scores to allocate gene lists into functionally related groups based on the functional similarity scores. GATHER assigns gene lists to functional groups/pathways via means of a Bayes factor[68]; this assesses the hypothesis that the distribution of functional groups/pathways vary across the genes under examination against the hypothesis that the distribution was identical. PANTHER maps the genes to pathways without conducting a statistical test, instead provides a list of all pathways those genes belong and then provides summary statistics on the mapped pathways. Webgestalt[64] uses Fisher's exact test to test for enrichment. A large repertoire of tools for enrichment analysis in Gene Ontology (GO) terms has also been developed in recent years, including GoMiner[69], FatiGO[70], BiNGO[71], GO-stat[72] and GeneTrail[73].

Table 1. Publicly available servers that will perform pathway analysis on the user's data

Name	Enrichement method	Description	URL	Ref
PANTHER	None (by gene id)	Maps genes to ~150 in-house pathways by protein sequence evolution	http://www.pantherdb .org/pathway/	67
DAVID	functional similarity scores	Groups genes upon functional similarity, then provides enrichment in KEGG, GO pathways, etc	http://david.abcc. ncifcrf.gov/home.jsp	74
GATHER	Bayes factor	Mpas gene lists to a variety of pathways, GO terms, protein binding	http://gather.genome .duke.edu/	65
WebGestalt	Fisher's exact	KEGG, BioCarta	http://bioinfo.vanderbilt .edu/webgestalt/	64
Reactome	hypergeometric	Manually curated and peer-reviewed pathway gene sets	http://www.reactome.org/ ReactomeGWT/ entrypoint.html#FrontPage.8093	75
GOrilla	hypergeometric	GO terms	http://cbl-gorilla.cs.technion.ac.il/	76

1.5 Technical challenges

Genes are variable in size, as well as LD pattern, they are variable in number of typed SNPs within and between genotyping platforms and pathways vary in the number of genes they contain. Pathways are differently defined between pathway collections; pathway gene lists are either incomplete or over-estimated. There is no golden standard as how to map SNPs to genes and genes to pathways. In addition to these technical challenges it is often necessary to distinguish between settings where a single SNP in the pathway is the causal or in high LD with the causal variant which may be driving the pathway association versus settings where additional independent susceptibility loci of modest effect are present.[59] These differences reflect upon the various pathway-centric analyses and results could be biased for one or the other reason. All pathway-driven analyses regardless of the test of association *per se*, should adjust for these cofounders in order to ensure rigor and robustness. In the following sections I present those challenges and the way that some authors deal with them. Table 2 summarizes various pathway-based approaches with an outline of their technical characteristics.

.

Table 2. Generic characteristics of pathway-based methods

The first column shows the name, columns 2-4 list the input data, the null hypothesis tested and the pathway collection used, columns 6-9 show the technical steps each method followed, when columns 10-12 show whether the results were replicated and the most influential genes in the associated pathways were identified. Y=yes, N=no, NA=not applicable

					Adjustment for				Findings		
Name	**Input**	**Hypothesis**	**Pathway collection**	**SNP-to-gene mapping**	**LD**	**Gene size**	**Pathway size**	**Single hits**	**Replication**	**Follow up on genes**	**Ref**
Multiple tools	*P*-values	competitive	KEGG, BioCarta, DAVID	SNPs with P<0.003, map SNPs to LD blocks	Y	N	N	Exclusion of MHC only	Limited	Y	62
ALIGATOR	genotypes	competitive	GO	2 approaches:within exons, 20kb, N:1	Y	Y	Y	N	N	N	37
U-statistics	genotypes	self-contained	2 pathways	Not-specified	Y	N	N	Y	N	Y	55
Kernel-based genetic similarity	genotypes	self-contained	Simulated data	simulation	Y	Y	Y	NA	N	N	57
Conditional logistic regression	genotypes	self-contained	Axon guidance pathway	Not-specified	N	N	N	N	Y	Y	32
Regression-based distance-matrix analysis	genotypes	self-contained	Not specified	Not specified	Y	NA	NA	N	N	N	58
ARTP	genotypes	self-contained	Not specified	Not specified	Y	Y	N	N	N	N	34
GRASS**	genotypes	self-contained	KEGG	20kb, N:1	Y	Y	N	N	N	Y	36
GSEA-SNP	genotypes	competitive	In-house assembled gene-sets	User-defined	Y	N	Y	N	N	N	60
Modified GSEA	genotypes	competitive	BioCarta, KEGG,GO	500kb, 1:1 most significant,	Y	Y	N	N	N	N	61
SNP ratio test	*P*-values	competitive	KEGG	Not specified	Y	N	N	N	N	N	44
EVA	*P*-values	competitive	GO	1:1	N	N	N	N	Limited	Y	41
MetaCore	*P*-values	competitive	GO	5kb, 1:1 weighted	N	Y	N	N	N	N	40
ATOM**	genotypes	self-contained	Genes	Not-specified	Y	N	NA	N	N	NA	54
Multiple regression on PCs	genotypes	self-contained	Genes	Not-specified	Y	Y	NA	N	N	NA	53
VEGAS**	*P*-values	self-contained	Genes	50kb, N:1	Y		NA	N	N	NA	77
Logistic kernel machine model**	genotypes	self-contained	KEGG	Annotation files	Y	Y	NA	N	N	NA	59
Combination of *P*-values*	*P*-values	competitive	KEGG, BioCarta	Not-specified	N	N	N	N	Y	Y	43
PRP*	*P*-values	competitive	KEGG	500kb, 1:1 highest risk statistic	N	N	N	NA	N	N	42
PINBPA*	*P*-values	competitive	KEGG	1:1	N	N	N	N	N	Y	49

* Pathway-based method with some sort-of gene-based analysis

** Gene-based method

1.5.1 Pathway collections

The validity of pathway association studies is highly dependent upon the quality of pathway collection used. Incomplete, not-curated or potentially erroneous pathway gene collections are likely to lead to false associations. It is of surprise that the most widely used pathway collection is KEGG[78] –also obvious from Table 2, albeit the fact that it is one of the least curated. BioCarta, Pathway Commons, MetaCyc, the Molecular Signatures Database (MSigDB) are a few more widely used (Table 3). Gene Ontology[79] is a vocabulary of cellular processes and function that also contains pathways, but in the majority of the cases it is utilized as a functional category enrichment dictionary. In Pathguide investigators can find links and information on 325 biological pathway related and molecular interaction related resources to choose from for their analysis.

Table 3. Publicly available pathway sources from which pathway gene sets can be down loaded as flat text files on the user's desktop

Name	Description	URL	Ref
KEGG	Pathway maps and gene lists on metabolism, cellular processes, human diseases, drug development and more	http://www.genome.jp/kegg/pathway.html	78
Gene Ontology	A controlled vocabulary of terms for describing gene product characteristics and gene product annotation data	http://www.geneontology.org/	79
BioCarta	Pathway maps and gene lists on metabolism, immunity, cell signalling, cell cycle regulation and more.	http://www.biocarta.com/genes/index.asp	NA
Pathguide	Lists all biological pathway related resources	http://www.pathguide.org/	NA
Pathway Commons	A collection of~ 1,500 pathways as assembled by pathway and interaction information from nine other sources	http://www.pathwaycommons.org/pc/	80
HumanCyc	List of human classic metabolic pathways	http://humancyc.org/	81
Reactome	Manually curated and peer-reviewed pathway gene sets	http://www.reactome.org/ReactomeGWT/entrypoint.html#FrontPage.8093	75

KEGG and MetaCyc have collected mostly metabolic pathways, while Signal Transduction Knowledge Environment (STKE) and Reactome contain well curated, publicly available data on signaling pathways. STKE provides probably the most comprehensive description of signaling pathways, but these pathways are not connected to computationally-accessible data of the participating molecules, such as sequence accession numbers.[67]. Webgestalt[64] uses its own pathway database that is assembled by various gene and gene product sources, as well as pathway collections, such as KEGG and BioCarta.. Of major concern is the fact that the definition and classification of pathways is different between those public databases. For instance, "MAPK signalling" in KEGG is represented by 210 genes when it has only 87 in BioCarta. As another example, the "IL18 pathway" of KEGG, when examined for enrichment

in GO terms using GOrilla[76] it comes up as "immune system response" and "interleukin 1 receptor binding" with respect to most enriched process and function. Furthermore, updating and curation of databases and relevant tools is not synchronized and does not always reflect the biological knowledge to date[67]. At the same time, this variety of pathway sources, without any standards that they all comply to not only creates a problem of choice but also drives inconsistency of results[62].

1.5.2 SNP-to gene mapping

Another technical complexity is the mapping of SNPs to genes. To assign SNPs to genes, one can use absolute genetic distance within a window of each side of the gene or use relative distance that allocates each SNP to the nearest gene. Several windows from each side of a gene coordinates' have been used, varying from 5kb[40] to 500kb[42]. There is no gold standard. There is naturally a trade-off between capturing real genetic variation linked to a specific gene, and bringing in random noise due to variation unrelated to any gene. A choice of -+5kb reduces the likelihood of including random SNPs, nevertheless it is likely to miss potential true association that might lie further away from the gene boundaries; +-50kb is likely to include some *cis* acting promoters, +-500kb will include more regulatory elements, but this may be at the expense of a great amount of noise. It has been suggested[36,37] that choosing 20kB is not entirely arbitrary since a recent study of gene–expression showed that the majority of expression Quantitative Trait Loci (eQTLs) lie within 20kb of either side of genes.[82] Other approaches consider only SNPs in exons[37] or limit SNPs to those which are non-synonymous.[55] Several methods even assign only one SNP to a gene, usually the most significant.[41,42,61] This strategy is not advised as it results in loss of power; it does not use the information from the LD pattern of a gene nor does it summarize the gene's genetic variation effectively and it cannot identify multiple associated variants acting cumulatively within a gene. An alternative approach is that by Torkamani *et al*[40], where they assigned one SNP to a gene but with a weighting that approximates the association-statistics of the gene.

Inconsistency on the SNP-to-gene mapping between tools/studies could also be attributed to the human genome reference set used, the gene coordinates, the SNP reference number versions. Some tools use public reference sets, such as NCBI and EBI, when others use in-house composed references (Webgestalt, PANTHER, DAVID, GATHER). for all these reasons, researchers are advised to always report the version of the databases and the annotation files used, for their results to be reproducible. As a general comment, the lack of standardisation may explain differences in results between studies.

1.5.3 LD pattern

Gene and pathway-centric approaches are prone to being biased by the inflation of certain genomic areas due to differing LD patterns across the genome. This issue has been recognised by researchers and therefore several strategies have been employed to deal with it. One way is to permute the phenotype labels and keep the underlying LD undisrupted.[36,61] This

approach preserves the underlying LD patterns of the pathway SNPs but in the presence of population stratification or genotyping errors it may yield biased results.[47] Other ways are PCA simulations based on LD structure as derived from HapMap.[36,77] On a similar basis, another group proposed an optimally weighted markers association test, where weights are assigned to each marker in a manner that is proportional to the amount of information the marker captures for the trait locus, based on LD patterns as derived from HapMap.[54] Holmans *et al*[37] corrected for the LD between significant SNPs by normalizing the LD of a region to the average LD across the genome. In another study they first allocated SNPs to genomic regions based on LD structure and then mapped those blocks to genes.[62] In Schaid *et al*[45] they modelled the genetic similarity between cases and that of controls using U vectors, where the LD pattern was accounted for by the covariance matrices.

1.5.4 Size matters

Elbers *et al*[42] assessed five different pathway tools on the same dataset and demonstrated explicitly that large genes, genes that reside in large LD blocks and large pathways are more likely to be denoted as associated, regardless of the pathway-approach used. The same has been argued by several other authors.[36,37,45,83,84] With regard to pathway size, some investigators prefer to restrict their analysis to pathways of a certain size; for instance between 20 and 200 genes.[84] Holmans *et al*[37] adjusted for GO category size via sampling random genes from the genome and simulating random gene lists with as many genes as the category contained in order to approximate the true null distribution of no association for every category.

Others deal with gene size by giving smaller weights to the larger genes in their ridge regression model[36]. ALIGATOR[37] allows for the varying sizes of genes by creating "random" genes with as many SNPs as the original. Thus, the probability that a gene is added to the list is proportional to the number of SNPs that it contains. GRASS[36] uses permutation of phenotypes and standardization of the statistic contributed by each gene, but even so the authors found a modest correlation between gene/pathway size and the estimated *P*-values.

1.5.5 Significant hits

Another potential limitation is that pathway statistics can be driven by very significant hits in one pathway. It is possible that a single gene if not a single SNP along a pathway could be enough to drive the entire association signal of the pathway.[84] For example, the MHC genomic region is strongly associated with most inflammatory and autoimmune diseases[85] and therefore when testing inflammatory pathways that include this region, such as "Antigen processing and

presentation" pathways or the "Cell adhesion molecules", the results might be biased. One strategy would be remove those genes and repeat the analysis: Elbers *et al*[52] showed that the *P*-values were affected when significant hits were removed, but the significance of several pathways remained. Another study[84] on a pathway analysis of two GWAS on type II diabetes (T2D) picked-up an association with the "Wnt-signaling pathway", which contains the *TCFL2* gene, one of the strongest T2D associations. By excluding this gene and repeating the analysis, "Wnt-signaling pathway" only reached a nominal *P*-value=0.002, which disappeared after Bonferroni-correction at the pathway level. In an alternative strategy to measure the effect of single hits, others used a stepwise procedure where at each step the most significant SNP was removed and the statistic of association was re-computed.[55] In another study, in their weighted genetic similarity matrices of individuals they weighed more the functionally relevant SNPs and therefore did not adjust for bias from single hits.[58] Nevertheless, their proposed methodology could be reversed by giving less weight to the significant SNPs in order to measure the remaining joint effect of the not-significant SNPs. In one of the early studies that showed association between the axon-guidance pathway and Parkinson's disease (PD)[32] the researchers used 1,412 SNPs of the pathway to fit a logistic model when 142 SNPs, that is 10.1% of the total were individually associated with PD. Nevertheless, the authors did not demonstrate which fraction of disease prediction those significant hits accounted for and what was the added value of the remaining pathway SNPs.

1.5.6 Overlapping pathways

The way that the gene products interact with each other at a molecular level is highly complicated. The same genes are involved in multiple different biological pathways and as a result pathways overlap in terms of gene composition. This is explained by many cellular processes using the same genes for signalling, cell activation, cell turnover and repair. Figure 1 shows the gene similarity matrix of 200 random pathways as downloaded from KEGG. It is obvious that the majority of them have at least half of their genes in common with other pathways. In a scenario, where a core of truly associated genes is involved in several pathways, any pathway approach would identify multiple pathway associations, when in reality only a small number of genes could account for the signal. Pathway-based approaches suffer from this limitation, as the implication resides in the complex biology and not the robustness of the method *per se*. Even multiple hypothesis adjustment at the level of pathways would not be adequate to solve this problem, as the interconnected associated pathways would still lie at the top of associations even though with adjusted *P*-values. Some investigators adopted a combined pathway and gene-driven methodology to look directly for the genes within associated pathways

that drive the signal for association. For example, Chen *et al*[56] applied a dual regularization function that selects SNP effects within a gene by lasso penalty while simultaneously shrinking the estimates of the genes by ridge penalty. Schaid *et al*[55] used a stepwise variable selection procedure to identify the most influential markers of the associated pathway(s). As I suggest in Table 4 my proposed strategy would be first to identify the associated pathways and then look for the genes that are most instrumental within the associated pathways.

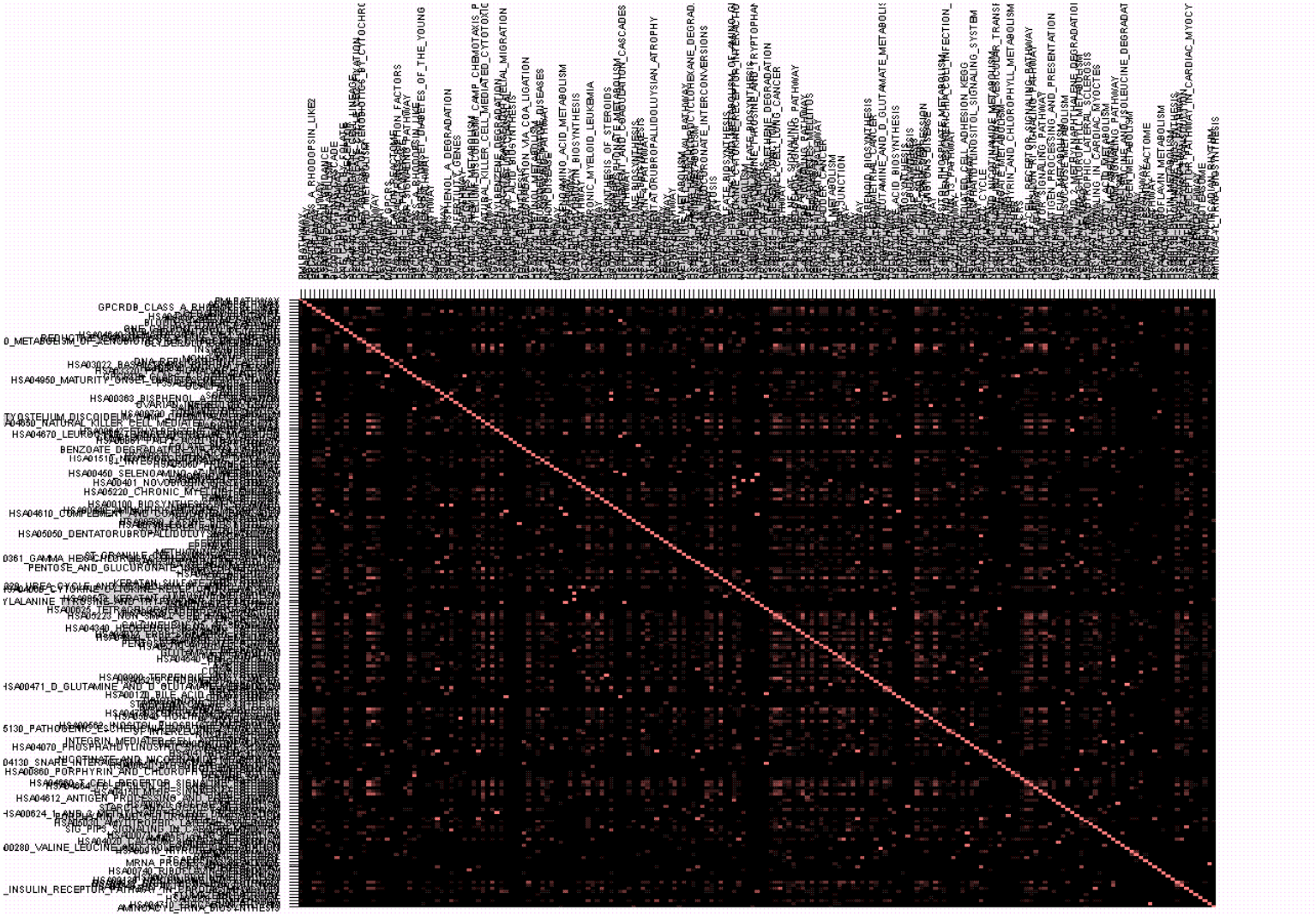


Figure 1. 200 pathways as downloaded from KEGG and their gene sharing as a fraction of the smallest pathway under comparison
Bright orange dots denote 100% genes in common and black stands for no genes in common. The gradient indicates the degree of overlap.

1.5.7 *P*-value threshold bias

Methods that use a *P*-value cut-off to define a list of significant SNPs to be tested for enrichment in pathways will be biased by the inherent arbitrariness of the cut-off selection. O'Dushlaine *et al*[44] argued that different cut-offs resulted in different associated pathways. Application of ALIGATOR[37] showed that by varying the cut-off from P=0.0001 to P=0.05 different enriched categories arose. They also demonstrated that in the application of ALIGATOR on a GWAS of bipolar disorder (BD) the significance of over-represented GO categories increased as the cut-off was loosened, whereas the opposite was observed in Crohn's disease (CD). It is likely that the cut-off should be chosen in relation to the genetic architecture of the disease and the number of significant hits in the GWAS. In the field of CD, which is a very diverse disease with more than 70 associated loci identified to date[86] the choice of a very a stringent threshold is likely to create a SNP list with several true associations. Conversely in BD, whose genetic architecture is almost entirely unknown and very few hits have been discovered so far and with no outstanding *P*-values[87], a more generous threshold should be used to allow for SNPs with modest effects to define enriched categories. Elbers *et al*[62] demonstrated that with *P*-value threshold<0.003 no GO category would show enrichment due to the small number of significant genes of T2D.

1.5.8 Genotyping platform and genome coverage

Elbers *et al*[62] showed that the Affymetrix GeneChip Human Mapping 500k array covers the whole genome but not all areas equally. For instance, they showed that the average number of SNP per pathway is 7.84 for the KEGG "fatty acid biosynthesis" pathway and 441.5 for the "antigen processing and presentation". This affects the analysis as some pathways are better covered therefore more likely to come up in the analysis. Naturally one could argue that this also depends on the number of genes per pathway and the LD pattern across the interacting genes. O'Dushlaine *et al*[44] also observed that for a pathway to be denoted as significant using their method, it would require at least 20 SNPs, therefore adequate SNP coverage of a pathway is essential for effective testing.

Table 4. Summary of technical challenges in pathway-based analyses strategies of GWAS

Technical challenge	Ways to tackle it	Pros/Cons	My suggestion
Pathway collection	KEGG, BioCarta, HumanCyc	+: Pathway gene lists can be downloaded and customized -: Pathway versions might be incomplete, contain errors	Try several versions of the same pathway; very discordant results should be suspicious; concordance of results on different pathway versions should be indicative of a real signal
	Gene Ontology (GO)	+: Well curated vocabulary of cellular functions -: Not all GO terms are supported with experimental evidence, caution required for misuse of terms	
	In-house pathway databases of online tools	+: Quick computation of pathway enrichment statistics -: Rely on the quality of the provided pathways, no means of modifying them or testing users' custom pathway	
SNP-to-gene mapping	1:1, the most significant SNP of the gene	+: Not biased by LD patterns, gene size -: Miss multiple associated hits in a gene, loses power in the presence of modest associations in moderate LD	+-20kb is a good trade-off between random noise and informative SNPs, lots of *cis*-regulatory elements fall within these boundaries
	N:1, using a window along the sides of a gene or focus on SNPs in exons only, etc	+: multiple SNPs that act additively on the same gene -:Need to summarize LD, adjust for gene size, find a good trade-off between bringing in informative SNPs and random noise	
LD	Phenotype randomization	+: preserves underlying LD -: in presence of population stratification yields biased results, might be computationally slow depending on sample size	Phenotype randomization Assign SNPs to LD blocks and then LD blocks to genes; use either a scoring function or PCA on blocks
	Normalizing to average LD across the genome	-: Does not favour pathways enriched in genes with LD patterns higher that those of the average LD across the genome;	
	PCA	+: reduces the df in a test statistic or the number of covariates in a regression model -:does not take into account the phenotype, thus PCs might not be the most discriminating between case/control	
Gene/ Pathway size	Test statistics that are summarized by one weighted SNP	+: all gene sizes become 1 -: Loss of power if few hits on the gene / inflated result if several associated SNPs in a large gene	
	PCA	As PCA above	
Single hits	Remove then and repeat analysis	+: downstream results are unbiased by single hits -: interactions of modest risk variants with the single hits might result in loss of power	Follow up on pathway associations to find the key associated genes
	Stepwise variable selection	-: model selection is cumbersome	
Overlapping pathways	Stepwise variable selection	-: biased by LD, model selection is cumbersome	Adopt a gene-selection methodology on associated pathways
	Dual penalized regression on pathway and gene	+: identifies pathways and genes simultaneously	

Table 4.(continued)

Technical challenge	Ways to tackle it	Pros/Cons	My suggestion
***P*-value threshold**	Rank SNPs in order of significance; compare the distribution of ranks between genes in a pathway to the rest	+: based on the distribution of ranks; cut-off -free -: favours pathways in which a few genes have very large test statistics in comparison to pathways with several genes of approximately similar significance.	Try multiple thresholds/inspect results, decide based on the study sample size, the number of single hits in the study and the complexity of the genetic architecture of the disease
	Stringent vs Generous cut-off	+-: Generous: favours pathways/diseases with several variants of modest association; in diseases with several variants of strong associations it is likely to bring in noise +-:Stringent: Does not favour pathways with few associations and a lot of non-associated genes or large genes with several non-associated SNPs	
Genome coverage on chip	Imputation	+: more coverage -: quality of imputed genotypes	

1.6 Limitations

1.6.1 Associations that will be missed

As already discussed, the success of pathway analyses depends upon the accurate annotation of the genes within the pathway as well as the richness of the pathway collections. What complicates matters is the lack of a consensus on pathway definition and the lack of a universal standard on what constitutes a pathway. Furthermore, current pathway gene lists represent the knowledge to date and their quality also depends on the electronic database maintenance and update.[33] Pathway versions can also be quite different between public databases, as discussed above. As a consequence, genes that are not included in the list of pathways under examination for any of the aforementioned reasons, genes whose function is unknown and therefore are not involved in a molecular pathways, as well as SNPs that lie in gene desserts are bound to be missed. For example, when using a window of 20kb on each side of the ~18k human genes as downloaded from NCBI and mapping them to 1034 pathway gene lists as downloaded from KEGG and BioCarta, ~91K SNPs of the Affymetrix 550k chip and ~107k of the Illumina 550k BeadArray were mapped to pathways, which in practice means that 4/5 of the SNPs on the chip will not be analyzed by pathway approaches, unless the window around the genes is widened.

Defining whether a particular SNP is part of a gene is also problematic, and can result in trans-regulatory SNPs being missed. Even *cis*-regulatory elements will be missed

unless the window of SNP-to-gene assignment is wide enough. It is up to the researcher to decide whether promoters, enhancers, suppressors and the overall transcription regulation machinery should be considered as part of a gene function within a pathway. It should be noted that in the majority of public pathway gene lists, genes coding for regulators are not included as components of the pathways.

1.6.2 Discordant findings

GWAS have been plagued with false positive results at the single-SNP level and there are lessons to be learnt to avoid similar false positive associations using pathway approaches. Pathway-based results to date have a high degree of discordance between them. Pathway methods applied on the same datasets by different groups did reveal the same associated pathways. Elbers *et al*[42] identified inflammatory pathways as being associated with T2D by analyzing two published GWAS, but did not validate the well-established involvement of β-cell development and function in T2D pathogenesis. Other groups identified an association between the axon guidance pathway and Parkinson's disease,[32,88]. Another study identified no signals of association between the axon guidance pathway and Parkinson's disease by applying both single and multiple marker analysis on two PD sample sets. [89] As pointed out in a recent review[45], the WTCCC BD dataset was analyzed by 4 different groups and the associated pathways did not match.[37,40,41,90] It is likely that the results of pathway-based analyses will be difficult to compare between groups unless the same settings are chosen. All the technical issues discussed above hinder replication.

1.6.3 Lack of replication

The majority of pathway-oriented approaches published to date, were based on simulated datasets and a single GWAS.[34,36,42,44,53-55,57,59,77] Therefore the results are questionable until they have been replicated on at least another independent dataset or they are further examined on gene-expression or functional studies.

Very few groups have shown validation of their findings on real data. Wang *et al*[53] showed an association of the IL12/IL23 pathway with Crohn's disease in 4 different datasets. Nevertheless, as 5 of the 20 pathway genes were confirmed CD susceptibility loci (*IL18R1*, *IL12RB2* in moderate LD with major hit *IL23R*, *JAK2*, *IL12B*, *TYC2*)[91], this replicated association should only be regarded as proof of principle. The study by Peng *et al*[43] on RA demonstrated some level of replication by applying 4 different

enrichment methodologies on two independent datasets. However, the different methodologies yielded very discordant results when compared to each other. Moreover, the association signals were not corrected for multiple testing of pathways (they tested 465 pathways). Their application of Fisher's exact test identified 10 pathways that replicated, all of which contained the major single hits of RA and especially the MHC HLA-genes ("Cell adhesion molecules", "Antigen processing and presentation pathway", "Type I diabetes mellitus pathway", "Bystander β-cell activation pathway", "Th1/Th2 differentiation pathway"). No adjustment was performed to remove the bias arising from the single SNP hits. They also reported that their pathway replication did not conform to their gene association results, as the associated genes were not components of the associated pathways. Lesnick *et al*[92] validated their multiple logistic regression models on two GWAS of Parkinson's disease (PD), but the prediction of disease status was done independently on the two datasets with refitting of the models and it was assessed by a likelihood ratio test. This is not an out-of-sample validation; on the contrary this approach allows for over-fitting on both GWAS. As it was later shown by another group that attempted to replicate the results with a proper out-of-sample cross-validation method, the obtained predictive power (area under the receiver operating characteristic curve (AUC)=56%) did not live up to the originally reported (AUC=90%) by Lesnick *et al*[92]. The same was observed by another group that did not manage to replicate the original published results.[92]

In general, published attempts to replicate pathway approaches have yielded discordant results and over-fitting due to lack of secondary independent cohorts for testing has made interpretation of results difficult. As replication is now essential for single SNP analyses of GWAS to be considered true findings, we should expect to see replication in an independent datasets as a requirement for accepting pathway associations from GWAS.

1.6.4 Follow up on the results

In addition to the need for replication of the pathway findings on independent datasets, further investigation of associated pathways at the gene level is needed for clinical and biological use of the results. The majority of pathways are interconnected sharing several genes as I have shown before. As a result they tend to show up as associations in clusters. For instance, Peng *et al*[43] identified a cluster of associated metabolic pathways ("Lysine degradation", "Glycerolipid metabolism",

"Glycerophospholipid metabolism", "Ether lipid metabolism") with RA. Taking into account that the KEGG "Ether lipid metabolism pathway" (18 genes) is contained entirely in the "lysine degradation pathway" (31 genes) which in its turn shares 7 genes with "glycerolipid metabolism" (42 genes), which also shares 12 genes with "Glycerophospholipid metabolism" (43 genes), then it would be reasonable to speculate that a core set of genes might be driving the multiple pathway associations. More importantly, without further dissection of the reasons for the association of multiple pathways, false inferences may be drawn from the results. Very few groups have tried to follow up on their results at the gene level. A very efficient approach was GRASS[36], a regularized logistic regression algorithm that selects SNPs within gene by lasso penalty while simultaneously shrinking the estimates of the genes by ridge penalty. In essence, this is a way of testing for association of a pathway with a disease and simultaneously identifying the key genes within a pathway. Other groups[37] prioritized the genes within associated GO categories based on their single-SNP *P*-value, which in essence contradicts the pathway concept that by definition supports the idea that SNPs with low *P*-values on their own may contribute altogether to the overall association of a pathway. Schaid *et al*[35] used stepwise variable selection and re-calculation of the pathway test statistic to denote the most influential genes. Peng *et al*[43] replicated five pathways across two studies for T2D, but their gene based methodology identified genes that were not parts of those pathways and this discordance was because the two methodologies-pathway and gene-based were applied separately. A safe strategy would be to apply a pathway-based approach followed by a gene-centred methodology on the associated pathways, so that the broader functional setting of a disease is identified initially and then the individual key genes can be exposed.

1.7 Advantages and advances

Pathway-driven approaches might have several technical challenges to face but they are powerful methodologies and if applied as a complement to conventional single-SNP analysis strategies, they have the potential to both drive the field of GWAS forward and shed light into complex disease genetic mechanisms.

1.7.1 Overcoming the limitations in single-SNP analysis

Despite the technical challenges of pathway-based approaches, as discussed above, there is growing evidence that pathway-driven methodologies are powerful new

tools that can overcome the limitations imposed by the stringent multiple hypothesis adjustments in the single-SNP analysis field.[45]

Pathway-based methods would be expected to have increased power compared to single-SNP analysis when applied to small studies. Combining GWAS of the same disease to perform meta-analysis has been an effective way of increasing the sample size and the power to detect additional variants[86,93] which has been used to increase the associations detected by single SNP analysis; however, this strategy is ultimately limited by the number of genotyped cases in each study. Furthermore, meta-analysis cannot always be applied, particularly for relatively rare diseases where all available cases have been used in the primary analysis.[94,95] Approaches that allow genes and pathways to be identified using smaller cohorts are therefore urgently needed, and pathway-oriented methods are promising ways of increasing the value of GWAS data.

1.7.2 Robustness in genetic heterogeneity

Pathway approaches are expected to be more robust than single-SNP approaches to the effect of genetic heterogeneity between individuals, ethnic backgrounds, genotyping platforms and studies. Different individuals might possess different disease variants of the same gene or even in different genes that are bound to be missed by single-SNP analyses, nevertheless they will be captured by a powerful pathway-centric approach, as long as these genes are members of the same functional pathway.[37] For that reason, true associations should be easier to replicate across different studies without the need to increase drastically the study size as would be required in meta-analysis. Multiple GWAS could be analyzed under a pathway-based concept in a way that is expected to be more powerful than simple meta-analysis based on *P*-values across studies.[83] The genes might differ to an extent but the pathways should be resistant to genetic heterogeneity. As an example, Elbers *et al*[62] replicated pathway results across two T2D datasets but reported that they were picking different genes.

1.7.3 Insight into disease

A pathway approach is more likely to yield insight into disease mechanisms than single-SNP analysis, in that associated biological pathways by definition implicate molecular function and interaction.[37,45,83,96]

Several studies to date have documented the importance of biological pathways in disease pathogenesis For instance, accumulated evidence implicates the complement

pathway with age related macular degeneration,[97] the important role of "IL23 pathway" in CD,[33,98,99] the "Wnt-signalling pathway" association with T2D,[100,101] the apolipoprotein E receptor mediated signalling pathways with Alzheimer's disease,[102] several interconnected metabolic pathways with obesity.[103]

When findings from GWAS are viewed from the perspective of biological pathways, remarkable pictures begin to emerge. The genetic variants associated with diseases have been observed to cluster in related biological processes. Substantial evidence has linked immune response pathways to autoimmune diseases,[104-108] metabolic pathways to metabolic traits,[103] intracellular signaling pathways with neuropsychiatric disorders including bipolar BD,[109-111] whose biological foundation does not seem to resemble that of classic neurodegenerative disorders such as Parkinson disease and Alzheimer disease, where recent cumulative evidence implicates the mitochondrial pathway as pathogenic lever.[111-113]

On a different note, exciting unexpected pathway associations expose common genetic architecture between diseases of differing pathophysiology. For instance, the "Wnt signalling pathway" has been associated with T2D,[101,114] as it has been involved in cancers of the blood, thyroid, breast, lung, prostate, and colon.[115,116] Increasing evidence indicates that individuals with T2D are at elevated risk for several cancers, mostly of the colon, pancreas, and liver, the risk being more than twice as high for these T2D patients than for the general population.[117-120] A hypothetical functional disease model upon the link between the "Wnt-signaling pathway", diabetes and colorectal cancer was suggested in a recent review[121].

This example serves as a proof of concept that pathway-driven methodologies can pave the way to discovering new biology. The shared pathways between different yet co-occurring diseases may not explain entirely susceptibility to each disease but they might point to the mechanisms of elevated risks of one disease when already suffering from the other. On this basis, GWAS of different diseases, but yet with some common aetiology or co-occurrence, could be pulled together to increase the discovery power. This approach has already been followed for autoimmune diseases to expose common underlying autoimmune mechanisms[43].

1.8 Summary

In this review of existing pathway approaches I have discussed the range of different approaches used to date, and some of the advantages and disadvantages of methods proposed for pathway analysis.

As already demonstrated, pathway-based approaches differ to a great extend with each other, all described methodologies have their relative strengths and drawbacks, they have several conceptual and technical challenges to address and as this pathway-centric concept is still in its infancy, no best pathway-based practice exists as yet.

The published literature to date suggests that pathway approaches may be a powerful strategy to complement single-SNP analyses of GWAS; they offer advantages in enabling significant findings on smaller data sets and they allow biological interpretation of findings to be more readily established. Nevertheless they also present with several technical challenges that remain to be addressed. All these technical limitations have been presented and discussed above.

As the pathway based methodologies evolve to account for these drawbacks, and a more standardized approach to the application of pathway-driven analyses is adopted, it is likely that the pathway-centric methods will become an important step in the analysis of GWAS.

In the next chapters of this thesis I describe my own contribution to the development of pathway-driven methods, and my efforts to overcome the methodological difficulties and develop an improved analysis strategy.

Chapter 2

Pathway-driven analysis methodology

In Chapter 1, I reviewed the background in the conventional single-SNP analysis of GWAS, the need and rationale for pathway-based approaches, as well as the technical challenges of such methods. In this chapter I describe my work to develop a pathway-centric method to GWAS analysis. Although, at this time (2011), the relevant literature includes many publications that report pathway-based methods, at the time I commenced my PhD (2008), there were very few published algorithms that approached the problem of GWAS analysis in a biological pathway-driven way. The landmark Wellcome trust Case Control Consortium (WTCCC) study of seven common diseases had recently been published, only with the single-SNP analysis results, whose numbers of identified loci were disproportional to the size and the cost of the study. The genotypic data from WTCCC were made available to my supervisors, and they were used as the initial dataset on which to develop a pathway based approach, as a complementary strategy to GWAS analysis.

2.1 Hypothesis

The principle hypothesis that drove my research is that evolutionary selection is likely to act on multiple genes that interact through functional pathways that are responsible for a phenotype, rather than only on single genes.

If we consider the genetic regulation of the immune response, multiple genes contribute to the nature and intensity of the response to any pathogen - some acting as positive and others as negative regulators.[122] The pattern of gene variants within inflammatory pathways retained in the population during evolution would be likely to depend on their overall effect on the immune phenotype that ensures survival against the major infectious diseases that cause death in early life[123,124]. At the same time, those gene variants, which result in rapid activation of a vigorous inflammatory response to infection may have the disadvantage of increasing the risk of auto-immune and inflammatory diseases later in life,[125,126] as also demonstrated in Figure 2. Pro-inflammatory mediators such as TNF, IL-12, IL-1, IL-6 and IFN-γ, which are essential for containment of microbial pathogens,[127,128] are also associated with the inflammatory processes seen in common auto-immune diseases such as RA,[129] CD,[130] or T1D[131]. Conversely, treatments that reduce inflammation, such as anti-TNF therapy for CD or RA[132] are associated with increased risk of opportunistic infection,[133] suggesting that pathways which are involved in inflammatory diseases are also involved in resistance to infectious diseases.

I, therefore, postulated that the genetic contribution to common autoimmune diseases would be determined by multiple gene variants in the same inflammatory pathways involved in host response to infectious diseases. I selected CD, T1D and RA, diseases with a substantial body of evidence suggesting involvement of inflammatory processes,[129-131] as model diseases to test this hypothesis.

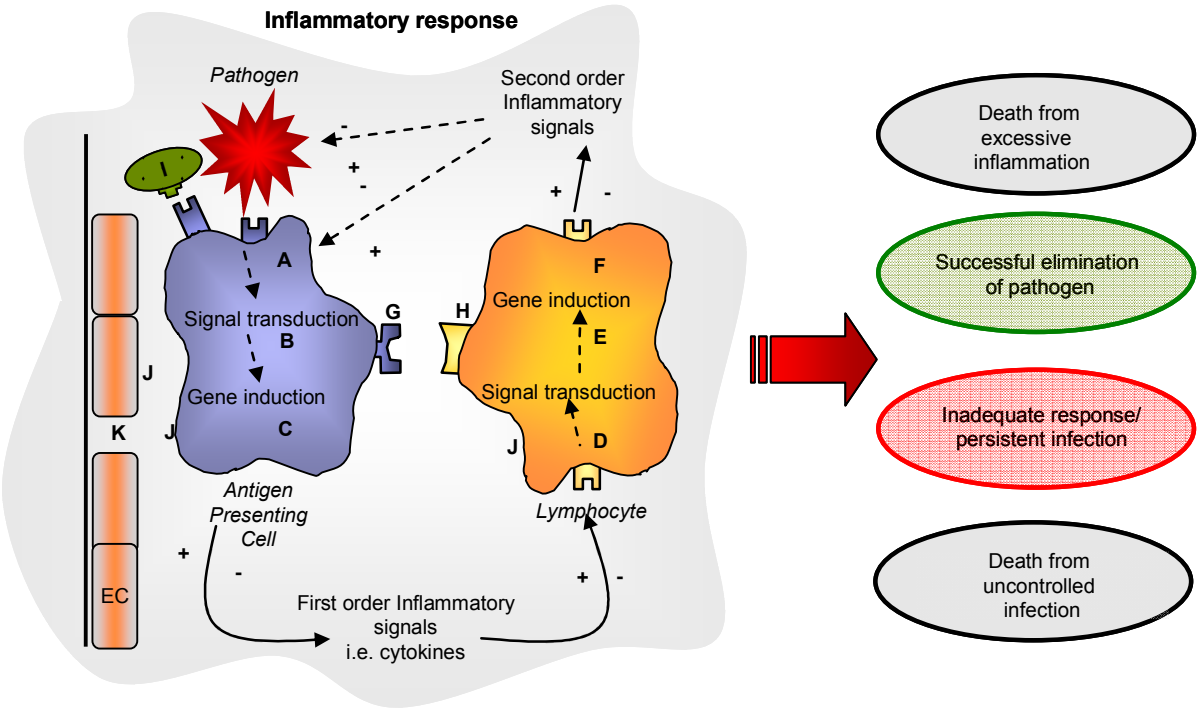


Figure 2. Inflammatory response to a pathogen.

Pathogen recognised by pattern recognition receptors on phagocytic cell (A) or plasma opsonins (I). Signal induction (B) and first order inflammatory genes (C) are induced leading to release of inflammatory signals. These bind to receptors (D), leading to activation of signal transduction pathways and gene induction of second order inflammatory mediators (E, F). These act as effectors of the inflammatory response (Red Arrow) or as positive or negative regulators. Inflammation upregulates cell adhesion molecules (J) and those involved in transendothelial migration (K). Genetic variants (A-J) will interact to alter the intensity and nature of the response. EC= endothelial cell

2.2 Aims

The original aims of my PhD were:

1. To develop a statistical method that captures the combined effect of multiple genetic variants with biological pathways and assesses their association to the disease as an entity.
2. To validate the results of the pathway statistic by comparing the same pathways in different diseases. I postulated that immune-mediated pathways would be involved in autoimmune and inflammatory diseases, but they would be unlikely to show evidence of association with non-inflammatory diseases.
3. To develop a methodology for identifying the key susceptibility loci within associated pathways.
4. To utilize the pathway information in building predictive models of disease and assess the added value to the findings of single-SNP analysis.
5. To validate the predictive models in a second independent dataset.

Over the first year of my PhD, I developed a test statistic that captures the cumulative trend of multiple SNPs assigned to genes in a pathway and tests for association of the pathway with a disease of interest. I then incorporated this test statistic in a methodology for pathway-based analysis of GWAS data.

In the proposed strategy, I initially construct the gene lists of the pathways to be examined and map SNPs to genes within 10kb of either side of the gene (this can vary, depending on user's criteria). I apply the cumulative trend (CT) test statistic on the genotype data of the case-control study to obtain the list of pathways that show evidence of association. I then employ a variable selection procedure only on the SNPs within associated pathways to identify the individual markers primarily responsible for the signal of association. Variable selection and logistic model fitting are carried out under the framework of 10-fold cross validation to adjust for over-fitting. I then evaluate the predictive power of the models with Receiver Operating Characteristic (ROC) curves and explore the allelic architecture of the disease models by various methods. I finally identify the covariates that are persistently selected in all 10 logistic models and I performed functional annotation and extra pathway exploration, as those genes which show constant association with the disease therefore are likely to represent true findings. The flowchart of the proposed methodology is depicted in Figure 3, while each step is described in detail in the following sections.

2.3 Data

For the first phase of my work I used available GWAS datasets from published studies and pathways from public databases.

2.3.1 GWAS datasets

The WTCCC data. This data is publicly available (on application) genotypic data from the WTCCC study on 14,000 Caucasian UK patients and 3,000 controls genotyped on the Affymetrix 500K Mapping Array GWA GeneChip. The cohort included 7 common diseases; Crohn's disease (CD), rheumatoid arthritis (RA), type I diabetes (T1D), hypertension, type II diabetes (T2D), bipolar disorder (BD) and coronary artery disease (CAD). The full clinical details, genotype data and initial analyses are described in The WTCCC[134]. In the original work, 16,179 individuals remained in the study after filtering for contamination, false identity, non-Caucasian ancestry and relatedness, while 468,919 SNPs passed all the quality control (QC) filters and the visual

inspection of the genotype call cluster plots. The QC filters excluded SNPs with MAF<1%, SNPs with a study-wide missing data rate >5%, or.1% with a study-wide MAF<5 and SNPs with Hardy–Weinberg exact *P*-value<$5x10^{-7}$. I performed no extra QC procedures and I used the exact same filtered dataset as in the original publication[134]. In all analyses the X chromosome was excluded because I did not have access to the X chromosome of the study described below, therefore I would not be able to validate any models at the time.

The 1966 Northern Finland Birth Cohort (NFBC). The cohort followed 12,058 infants, born in 1966 for over 30 years, with documentation of all health events, and illnesses and genotyped using the Illumina Infinium 370cnvDuo array[135]. The same QC filters were used in this study too. Over the 30 years of follow up, 30 of the 4,763 genotyped individuals (i.e. 0.6%) were diagnosed as having T1D. For that reason, this cohort was used to validate my findings from the WTCCC data analysis.

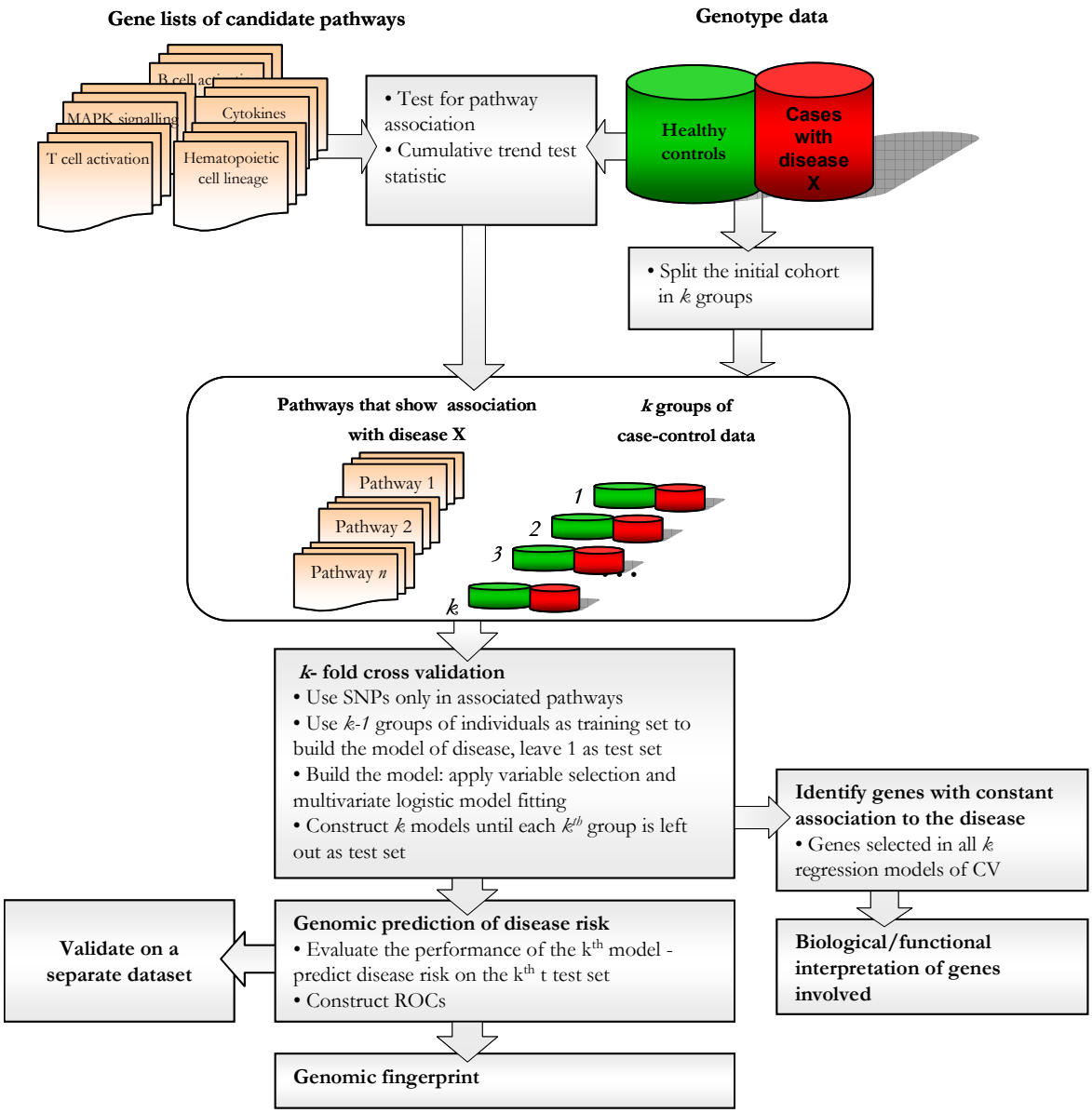


Figure 3. Pipeline of the pathway-based strategy for the analysis of GWAS data.

2.3.2 Pathway selection

The pathways under investigation were defined from published annotation, and using previously defined "canonical pathways" from the KEGG and Ingenuity Pathway Analysis 6 databases. The majority of these pathways were collated by Dr. Victoria Wright, at the Department of Paediatrics, Imperial College, and the pathways downloaded from KEGG were also examined by her for their biological validity. The full details of the pathways used in my analysis, the genes comprising each pathway, and the overlap in genes between pathways are provided on the webpage http://www1.imperial.ac.uk/medicine/people/l.coin/.

I examined key canonical pathways associated with innate and acquired immune response to pathogens, such as those involving pattern recognition receptors (Figure 2A); signal transduction triggered through these receptors (B); the inflammatory genes induced (C); second order cytokine receptors (D) and their signal transduction pathways (E); leukocyte adhesion and transendothelial migration pathways (F), antigen processing and presentation (G), the complement pathway (H), and T and B cell activation pathways (I). As a negative control, I examined metabolic pathways such as the urea cycle, amino group metabolism and glycolysis, which biologically were not expected to impact directly on inflammatory disease susceptibility. The physical positions of all human genes (N=18,272) were downloaded from the ftp service of NCBI. SNP positions were taken from the Affymetrix annotation files. Gene to SNP mapping was performed by a routine that I developed. I used all SNPs within 10kb of either end of the gene. I examined 84 pathways and sub-pathway combinations in total, where 1415 genes are involved and 20,309 SNPs are assigned to them. In this thesis, all genes are referred to by their HGNC ID and SNPs are referred to by their dbSNP id.

2.4 Test for pathway association - cumulative trend test statistic

Having formulated the hypothesis, I had to develop the pathway statistic to assess the genetic variation within all genes of a biological pathway and its relation to the disease of interest. I opted for a permutation-based pathway statistic that would account for the underlying LD structure without being too computationally intensive.

The data available from a case-control, study are represented in Table 5, where A is the high risk allele and a is the other.

Table 5. GWAS genotype data

	aa	***aA***	***AA***	**Total**
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Assume that (r_0, r_1, r_2) follows a trinomial distribution with probabilities for genotypes aa, aA, AA equal to p_0, p_1 and p_2 respectively and that (s_0, s_1, s_2) follows a

trinomial distribution with probabilities q_0, q_1, and q_2. The null hypothesis of no association can be written as:

$$H_0 : p_i = q_i, i = 0,1,2 \tag{2.1}$$

To test H_0 against any alternative using the Cochran – Armitage trend test, a set of scores x=(x_0, x_1, x_2) must be assigned to genotypes (*aa*, *aA*, *AA*). From Sasieni[50], x = (0, 1, 2) is assigned for the additive model, x = (0, 1, 1) for the dominant and x = (0, 0, 1) for the recessive. I also assigned x = (0, 1, 0) for the overdominance heterozygous advantage/disadvantage model. Given the score x, the Cochran-Armitage test can be written as:

$$Z_T^* = \frac{U}{[\mathrm{var}_{H_0}(U)]^{1/2}} \tag{2.2}$$

where

$$U = \frac{1}{N}\sum_{i=0}^{2} x_i (Sr_i - Rs_i) \tag{2.3}$$

and

$$\mathrm{var}_{H_0}(U) = N\sigma_0^2 = \frac{RS}{N}[\sum_{i=0}^{2} x_i^2 q_i - (\sum_{i=0}^{2} x_i q_i)^2] \tag{2.4}$$

In practice, the q_i in (2.5) might not be known, therefore σ_0 is estimated from the data by representing q_i as n_i / N. Thus, I have

$$\widehat{\mathrm{var}}_{H_0}(U) = N\hat{\sigma}_0^2 = \frac{RS}{N^3}[\sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2] \tag{2.5}$$

The CA_{MAX} statistic is the maximum of the 4 Cochran-Armitage test statistics under the 4 genetic models.

$$CA_{MAX} = \max\left(CA|_{x=(0,1,2)}, CA|_{x=(0,1,1)}, CA|_{x=(0,0,1)}, CA|_{x=(0,1,0)}\right) \tag{2.6}$$

To obtain the cumulative trend test statistic for pathway association, I sum the CA_{MAX} Armitage trend test statistic over all of the *n* SNPs in a given pathway. If we

denote as $CA_{MAX,i}$ the MAX Armitage trend test statistic for the i[th] SNP, then the cumulative trend test statistic for the pathway is

$$CT_{pathway} = \sum_{i=1}^{n} CA_{MAX,i}\ , \tag{2.7}$$

where *n* is the number of SNPs in the pathway.

If the null hypothesis of no association was true, we would expect the sum of independent χ^2 test statistics to follow a χ^2 distribution with *df* as many as the number of test statistics in the sum. The ease of the expected null distribution was the reason that I developed the pathway statistic around the sum of single-SNP statistics. In this case though, as the test statistics are not independent, I had to approximate the null distribution parametrically. To estimate a parametric approximation of the null distribution of the statistic I fitted several continuous theoretical distributions. Goodness of fit was assessed by QQ-plots of the observed and theoretically distributed data and subsequently by the Kolmorogov-Smirnof goodness of fit test statistic that assesses the null hypothesis H_0: The distribution fits the data. The skew-normal[136] approximation was finally preferred as it exhibited better fit than the Normal, Weibull, Gamma and χ^2 distributions (Figure 4) in the vast majority of pathways (80%). In the rest of the pathways the skew normal and the gamma distributions fitted almost equally well, suggesting a *P*-value of the same magnitude, therefore I opted for the skew normal fit on all pathways.

Given N ordered data points Y_1, Y_2,..., Y_N the Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1<i<N}(F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i)) \tag{2.8}$$

where F is the theoretical cumulative distribution of the distribution being tested

I show the fit of the skew normal to the empirical null distribution of cumulative χ^2 statistics in the top-left panel of Figure 4.

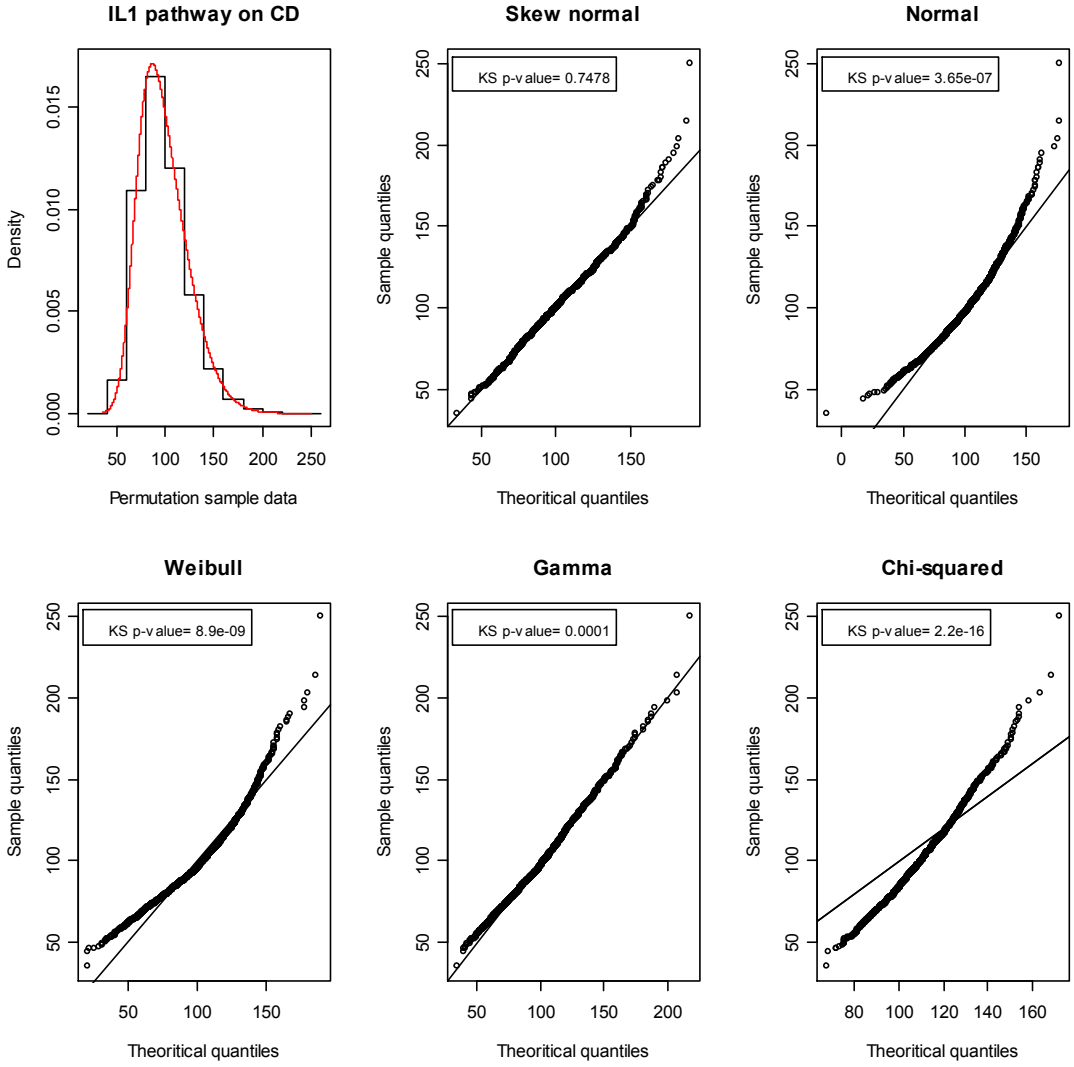


Figure 4. Histogram, density function and QQ plots of various distribution fits to permutation data of the cumulative trend test statistic

The plot on the top left corner shows the histogram and the skew-normal density function fit of the cumulative trend test statistic calculated from 10000 permutations of cases/control label for the IL-1 pathway in CD. The next five plots correspond to QQ-plots of the skew-normal, normal, weibull, gamma and chi-squared distribution fits to the same null distribution. The *P*value of Kolmogorov-Smirnof goodness of fit test statistic is depicted in the legend of each plot.

I estimated the parameters of a parametric approximation of the null distribution of the statistic by fitting a skew normal distribution to results obtained from 1000 random permutations of case/control label. For each disease and pathway I chose location (μ), scale (σ) and shape (λ) parameters which maximise the likelihood

$$P(CT_j \mid \hat{\mu}, \hat{\sigma}, \hat{\lambda}), j \in [1,1000] \tag{2.9}$$

where each CT_j is calculated as the cumulative statistic under a random permutation of the case control labels and *j* is the index of a random permutation. The values of each parameter vary greatly with pathway size and single SNP significance. To

give an example, the mean of location (μ) over 800 canonical KEGG pathways is 923 and the standard deviation is σ=1367.

I then estimated the significance as

$$P = \int_{CT_{pathway}}^{\infty} P(x \mid \hat{\mu}, \hat{\sigma}, \hat{\lambda}) dx \tag{2.10}$$

where $CT_{pathway}$ is the pathway statistic under no permutation and $P(x \mid \hat{\mu}, \hat{\sigma}, \hat{\lambda})$ is the fitted skew normal distribution. This procedure was carried out separately for each disease and given pathway using the "sn" package in R[137].

I performed only 1000 permutations to save computational time. However, I examined the goodness of fit for a few pathways under 1000 and 10000 permutations. As it is shown in Figure 5 for the IL1 pathway with CD, the approximate distributions fit equally well the data and the derived *P*-value for IL1 was ~0.0026 for 1000 permutations and ~0.0028 for 10000.

The statistical significance level was adjusted using the Bonferroni correction

$$\alpha_{adj} = \frac{\alpha}{n} \tag{2.11}$$

where α_{adj} is the adjusted significance level for n null hypotheses that are tested simultaneously with a significance level α per hypothesis.

In terms of computational time performance, permutations for a pathway of <100 SNPs run in a couple of seconds and they scale up to 5 minutes for pathways of ~5,000 SNPs. The algorithm is very fast and in the case that the pathways can be run in parallel on a server, the entire pathway-based analysis would take up only a few minutes.

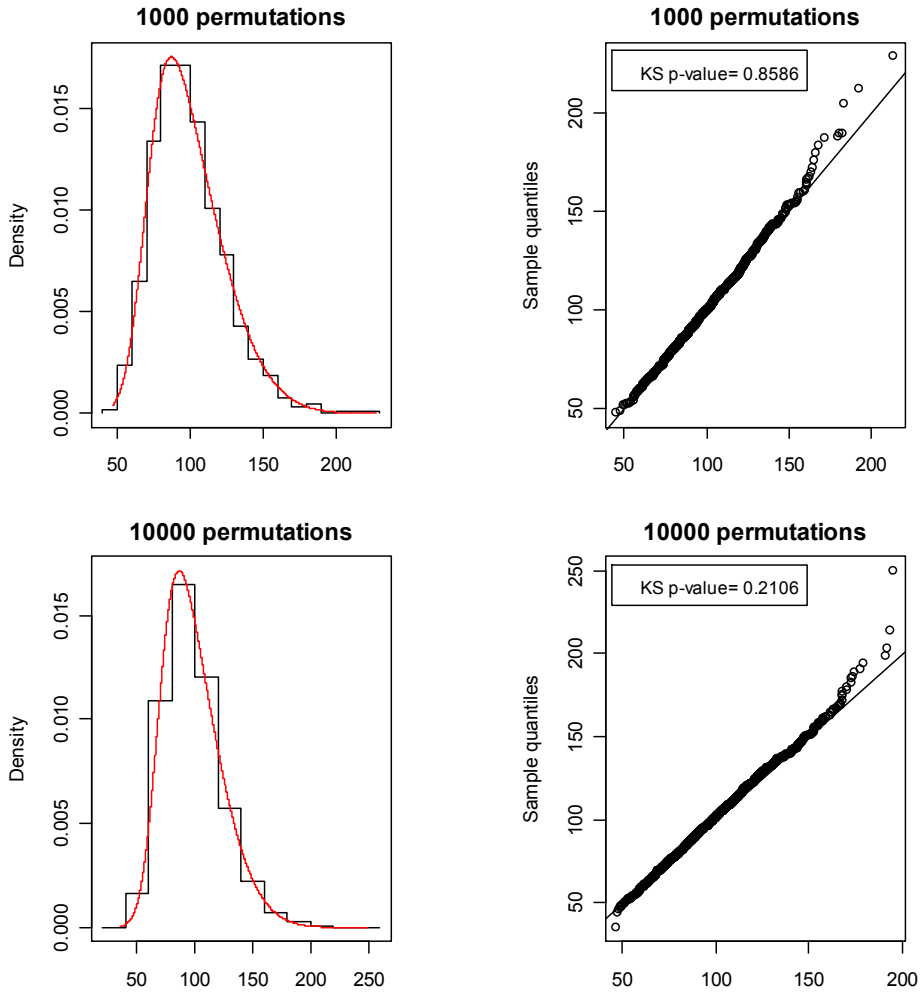


Figure 5. Skew normal distribution fit for 1,000 and 10,000 permutations for the IL1 pathway with CD

The *CT P*-value derived by the top distribution was ~0.0026 and by the second ~0.0028.

2.5 Variable selection and logistic regression

The cumulative trend test statistic captures the effect of multiple SNPs assigned to genes in a pathway that may be implicated in a disease, but does not identify the individual markers primarily responsible for the evidence of association. I therefore employed a variable selection procedure to determine a subset of genes in associated pathways that are predominantly driving the pathway association signal in each inflammatory disease.

Variable selection is a procedure of selecting a set of "most" influential variables (in this case the key responsible SNPs/genes) of an original pool of variables (in this case all the SNPs/genes within associated pathways) to build learning models (in this case predictive disease models). The procedure can be implemented by various algorithms.

Each algorithm fit models on subsets of variables and assesses the "relevance" of the subset to the outcome of interest based on a criterion (e.g. likelihood of the data given the model, FDR, a test-statistic of goodness of fit and others). The variable selection algorithm iterates until an optimal criterion is reached (e.g. maximum likelihood, minimum FDR, significant *P*-value of a test statistic and others) and the variables in the model are the "selected" ones.

Variable selection and model fitting was performed with the program HyperLasso[138]. No statistical interactions were taken into account. The algorithm is designed to fit a logistic regression model while performing variable selection to generate models with relatively few SNP covariates. The regression coefficients are given by maximum posterior estimates and sparse solutions are achieved by using a prior with a sharp peak at zero. The prior used was the normal–exponential–gamma (NEG) distribution and was specified on effect size at all SNPs resulting in an algorithm that chooses those SNPs which explain the greatest variation in the phenotype.

HyperLasso was applied as to consider four types of SNP effects - additive, recessive, dominant, heterozygous advantage/disadvantage(overdominance) – and encoded genotypes accordingly – [0,1,2], [0,0,x], [0,y,y], [0,z,0], where the original genotypes are ordered as [AA, Aa, aa], where *a* is the minor allele. The values of *x*, *y* and *z* are arbitrary and affect the probability of each effect type being included in the model; I chose the values of *x*, *y* and *z* for each SNP separately such that the standard deviation over all SNPs of all four effect types was the same. This resulted in choosing the effect type with the largest gradient of the log-likelihood at each SNP, which is equivalent to choosing the effect type with the minimum *P*-value. At each position the encoded genotypes were centred to have mean zero. The algorithm runs with two parameters, shape λ and penalty *f'* of the prior distribution. In all analyses shape was set equal to 3.5 and penalty equal to 40. These values were selected after sensitivity analysis on the trade-off between known hits and novel SNPs to be allowed in the models.

2.6 Model evaluation

To evaluate the performance of the predictive logistic models, I performed 10-fold cross-validation (CV). As in the general case of *k*- fold CV, the dataset was divided into *k* subsets (in this study *k*=10) and each time, one of the *k* subsets was used as the test set and the other *k-1* subsets were considered together to form a training set.

Variable selection and model fitting was performed on each k_{th} training dataset separately and the predictive power of the k_{th} fitted logistic model was evaluated on the corresponding k_{th} training set. Then the average sensitivity/specificity across all k trials was computed to produce the average ROC plot. The area under the ROC curve (AUC), shows the predictive power of the model; an AUC of 100% represents a perfectly fitted model and an AUC of 50% represents a random model. The AUC of each k_{th} fitted logistic model as well as the average AUC of all 10 models for each disease were calculated following the process outlined in Mason and Graham[139].

The procedure was repeated for each disease. Specifically, the initial cohorts of T1D, RA and CD (1963, 1860 and 1748 cases), were split into 10 datasets each of 196, 186 and 174 individuals respectively. The control group of 2938 individuals was split into subsets of 293 individuals. During each trial, a control subset with each disease specific subset formed the *test* dataset for the particular disease, while the 9 remaining disease and controls subsets were put together to form the *training* set.

2.7 Adverse and protective genotype analysis

To better understand the genomic architecture underlying disease susceptibility at the allelic level, I probed the structure of my predictive models of disease risk. Considering that within each pathway some components act as positive and others as negative regulators of the overall pathway output, I reasoned that genotypes for each SNP may have different effects, either increasing or decreasing the risk of disease depending on the functional effect on the pathway. The contribution at each SNP was assigned as additive, recessive, dominant or heterozygous. Genotypes were encoded numerically on the basis of the genetic model at each SNP with respect to the minor allele (refer to Chapter 3, section §3.5). For each disease, I picked the logistic model trained on the first fold of the 10 fold CV to demonstrate the *adverse* against *protective* genotype effect. I repeat the same procedure for all models and the plots remain the same. The effect type at each SNP was also categorised as *adverse* (increasing disease risk) or *protective* on the basis of the sign of the coefficient (+ve or –ve) relative to this numerical encoding. I summed encoded genotypes at *adverse* and *protective* SNPs separately for each individual and calculated their ratio.

I displayed all cases and controls in relation to the variants carried by each individual. Genotypes were coded by the variable selection algorithm coding scheme.

Cells were calculated as the product of the coefficients of the logistic model with each individual's genotypes. Thus, for cell $c(i,j)$ it would be

$$c(i,j) = \beta(i) * gen(i,j) \tag{2.12}$$

where $\beta(i)$ is the logistic model coefficient at SNP *i* and $gen(i,j)$ is the genotype of individual *j* at SNP *i*. Cells were coded red at adverse SNPS, and green at protective SNPS. The intensity of the colour in each cell was calculated using the cumulative distribution function (cdf) of the normal distribution as

$$intensity = 2 * (cdf(abs(c(i,j)) \mid \mu = 0, \sigma) - 0.5)\ ,\ \sigma = \sqrt{\sum_j c(i,j)^2} \tag{2.13}$$

SNPs were ordered via a neighbour-joining based tree-clustering algorithm designed to place rows with similar patterns next to each other. To do this I first calculated a distance between each SNP i and k as

$$d(i,k) = \sum_j (c(i,j) - c(k,j))^2 \tag{2.14}$$

and then calculated a neighbour joining tree from this distance matrix in a manner similar to Howe *et al*[140].

2.8 Summary

In this chapter I described the driving hypothesis of my research and my original methodology to detect pathway-disease associations from GWAS data. I presented the datasets that I used and the statistical analysis that I applied. In the next chapter I showcase the results of this approach to the WTCCC study of seven common diseases.

Chapter 3

Application to three auto-immune diseases

As described in the previous chapter I aimed to test the hypothesis that the genetic contribution to common autoimmune diseases would operate through biological pathways involved in the immune response. I developed a pathway-centric analysis methodology of GWAS and I applied on the WTCCC GWAS dataset of 7 common diseases; my analysis tested the association between 84 inflammatory pathways and three autoimmune diseases, RA, T1D and CD. The four non-autoimmune diseases of the WTCCC study (i.e. BD, CAD, T2D, HT) were used as negative controls. In this chapter, I present my results.

3.1 Shared and unique pathways

The number of pathways, genes within them and SNPs used in the analysis are shown in Table 6. When the cumulative trend test statistic was applied to the seven diseases of the WTCCC study, highly significant associations were observed between several of the inflammatory pathways and the three autoimmune diseases (Table 7). The "Jak-STAT signalling", "antigen processing and presentation", "T cell activation pathway", "cell adhesion molecules", "hematopoietic cell lineage" and "NK cell mediated cytotoxicity" showed high levels of significance in all three diseases ($P<10^{-4}$ to $P<10^{-20}$). However certain pathways showed evidence for association with one disease; the pathways of "NOD2" ($P<10^{-4}$to $P<10^{-15}$), "IL-23" and "TNFα" ($P<10^{-20}$), "IL-12" and "TGF-β " ($P<10^{-06}$), "TLR2 signalling" ($P<10^{-5}$ to 10^{-9}), "classical MAPK" ($P<10^{-10}$) and "B cell activation" ($P<10^{-8}$) with CD; "TLR3" ($P<10^{-4}$), "T-cell activation via PLC" and "via NFAT" ($P<10^{-4}$) and "ABC transporters" ($P<10^{-16}$) with T1D; "TLR9 signalling via IRF5" ($P<10^{-6}$) and "purine metabolism" ($P<10^{-5}$) with RA (Table 7).

Almost all of the pathways under investigation showed no evidence of association with the non-autoimmune diseases (Table 7). However, some signal of association was detected in pathways with a previously identified or biologically plausible link to a non-inflammatory disease. For instance, the association between the B cell signalling via protein kinase C pathway in B cells and bipolar disorder is consistent with reports that PKC (a constituent of this pathway) activity has a role in the pathophysiology of bipolar disease.[141] No association was seen for metabolic pathways that are not expected to have a biological link to the autoimmune diseases (Table 7).

Table 6. Summary statistics of the overall study and the pathway trend test

Overall study			
Inflammatory pathways		84	
Genes within inflammatory pathways		1415	
SNPs mapped to genes		20309	
Pathway statistic analysis results	**T1D**	**RA**	**CD**
Associated pathways	24	15	38
Genes in associated pathways	989	878	1124
SNPs in associated pathways	12663	12215	16284

Table 7. *P*-values of the pathway statistic over all the examined pathways and the seven diseases.

P-values in bold and scientific format were denoted significant after Bonferroni correction for the number of pathways (N=84) at the 5% significance level. *P*-values equal to 0 represent $P < 10^{-20}$. CD= Crohn's disease, T1D=type 1 diabetes, RA=rheumatoid arthritis, HT=hypertension, BD=bipolar disorder, CAD=coronary artery disease, T2D, type 2 diabetes.

PATHWAY	CD	T1D	RA	HT	BD	CAD	T2D
Pattern recognition receptors							
TLR2[1]	**3.5E-05**	0.029	0.041	0.257	0.187	0.386	0.357
TLR2/1[1]	**9.1E-06**	0.030	0.026	0.197	0.185	0.318	0.367
TLR2/6[1]	**2.4E-05**	0.020	0.065	0.129	0.129	0.179	0.316
TLR3[1]	0.001	**3.3E-04**	0.037	0.319	0.198	0.246	0.145
TLR4 (MyD88 dependent)[1]	0.020	0.010	0.040	0.199	0.102	0.469	0.335
TLR4 (MyD88 independent)[1]	0.018	0.018	0.029	0.152	0.124	0.343	0.262
TLR5/TLR7/TLR8/TLR9[1 2]	0.127	0.062	0.397	0.380	0.154	0.342	0.119
TLR2-IRF5	**2.7E-06**	0.225	0.001	0.116	0.085	0.389	0.477
TLR2/1-IRF5	**3.1E-09**	0.228	0.001	0.129	0.066	0.453	0.442
TLR2/6-IRF5	**8.1E-05**	0.171	0.001	0.179	0.066	0.102	0.330
TLR4 (MyD88 dependent)-IRF5	0.022	0.076	0.017	0.173	0.078	0.555	0.482
TLR5/TLR7/TLR8-IRF5[2]	0.441	0.804	0.032	0.315	0.013	0.394	0.004
TLR9-IRF5	0.002	0.058	**1.6E-06**	0.220	0.112	0.638	0.361
Signal transduction							
Jak-STAT signalling	**2.5E-07**	**1.9E-12**	**4.4E-09**	0.005	0.011	0.166	0.045
MAPK: All	**7.7E-09**	**1.5E-06**	0.119	0.005	0.054	0.020	0.003
MAPK: Classical	**1.8E-10**	0.001	0.164	0.009	0.121	0.033	0.001
MAPK: JNK & p38	**2.2E-06**	2.72E-02	0.139	0.053	0.158	0.046	0.030
NFKB	0.030	0.016	0.010	0.842	0.533	0.333	0.294
NOD1[1]	0.097	0.324	0.633	0.003	0.141	0.001	0.489
NOD2 (via GRIM19)[1]	**4.4E-04**	0.055	0.237	0.184	0.102	0.205	0.210
NOD2 (via RICK)[1]	**1.3E-15**	0.587	0.528	0.005	0.151	0.045	0.287
Second order cytokines							
IL-1[1]	0.003	0.217	0.127	0.101	0.157	0.516	0.496
IL-6	**3.5E-04**	**4.2E-04**	0.064	0.140	0.826	0.221	0.426
IL-10	**8.4E-06**	**2.1E-04**	0.083	0.378	0.683	0.116	0.836
IL-12	**4.9E-06**	4.32E-03	0.003	0.324	0.949	0.523	0.747
IL-18	0.017	0.062	0.032	0.220	0.391	0.287	0.287
IL-23	**0.0E+00**	0.017	0.063	0.012	0.709	0.637	0.301
TNF[1]	**0.0E+00**	0.017	0.026	0.045	0.244	0.635	0.283
TGF-β	**2.6E-05**	0.004	0.035	0.102	0.502	0.155	0.218
Antigen processing and presentation							
All[1]	**1.4E-04**	**0.0E+00**	**0.0E+00**	0.012	0.016	0.256	0.577
MHC I	0.073	**0.0E+00**	**0.0E+00**	0.291	0.266	0.420	0.521
MHC II	**1.9E-05**	**0.0E+00**	**0.0E+00**	0.021	0.043	0.810	0.545
B – cell activation							
All[1]	**5.2E-08**	0.005	0.656	0.005	0.026	0.196	0.006
AKT[1]	0.039	0.550	0.223	0.015	0.021	0.059	0.004
AP1	**7.6E-11**	0.037	0.578	0.146	0.136	0.403	0.004
NFAT	**1.7E-04**	1.72E-03	0.702	0.010	0.009	0.221	0.067
PKC[1]	0.037	0.050	0.553	0.245	**1.1E-04**	0.323	0.135
T – cell activation							
All[1]	**2.5E-11**	**1.3E-07**	**2.4E-05**	**4.1E-05**	0.004	0.069	0.002
AP1	**5.3E-07**	0.002	0.112	0.124	0.293	0.516	0.003
NFAT	0.004	**1.0E-04**	0.025	0.021	0.004	0.452	0.047
PLC[1]	0.002	**1.3E-04**	0.009	0.028	0.153	0.480	0.007
ICOS/CD28[1]	**4.1E-05**	0.227	0.029	0.001	0.060	0.024	0.352
Cytokines/receptors	**4.3E-04**	**6.0E-06**	**2.5E-06**	0.156	0.005	0.055	0.123
Cytokines/receptors/Jak-STAT/suppressors	**2.4E-04**	**1.2E-09**	**1.1E-04**	0.007	0.011	0.023	0.014
IFN-γ	**1.7E-05**	**8.6E-06**	0.018	0.065	0.089	0.186	0.630
Signalling molecules and interaction							
Cell adhesion molecules: All	**0.0E+00**	**0.0E+00**	**0.0E+00**	0.081	**3.1E-04**	**4.6E-04**	**2.0E-04**
APC: T cell	**5.0E-06**	**0.0E+00**	**0.0E+00**	0.211	0.001	0.618	0.184
Tc cell: target cell	0.010	**0.0E+00**	**0.0E+00**	0.577	0.284	0.561	0.438
Th cell: B cell	**2.5E-04**	**0.0E+00**	**0.0E+00**	0.068	0.035	0.363	0.554
Leukocyte : platelet	0.208	0.868	0.631	0.208	**3.5E-05**	0.680	0.139
Leukocyte : endothelial cell	0.029	0.095	0.257	0.194	**2.2E-04**	0.190	0.452
Neural cells	**7.6E-05**	0.123	0.125	0.191	0.045	0.001	0.001
Cytokine-cytokine receptor interactions	**4.2E-15**	**2.7E-12**	0.002	0.042	0.026	0.194	0.104
Others							
ABC transporters	0.021	**2.2E-16**	0.004	0.350	0.003	0.276	0.166
Cell communication	0.014	**8.3E-05**	0.007	0.064	**1.6E-04**	0.051	0.009
Complement: All	0.080	0.323	0.051	0.245	0.551	0.484	0.206
Haematopoietic cell lineage: All	**6.3E-10**	**0.0E+00**	**0.0E+00**	0.006	**2.8E-04**	0.019	0.070
Leucocyte transendothelial migration	0.013	0.575	0.102	0.418	0.003	0.059	0.203
Natural killer cell mediated cytotoxicity	**3.3E-10**	**0.0E+00**	**1.4E-09**	0.003	0.001	0.100	0.431
Neutrophil activation	**4.5E-05**	0.002	0.043	0.061	0.290	0.043	0.605
Purine metabolism	1.04E-02	0.026	**3.2E-05**	0.174	0.474	0.010	0.367
Pyrimidine metabolism	7.02E-02	0.160	0.039	0.389	0.377	0.590	0.280
Type 1 diabetes	**4.8E-07**	**0.0E+00**	**0.0E+00**	0.231	0.045	0.672	0.940
Non-inflammatory pathways							
Urea cycle	0.928	0.081	0.771	0.318	0.153	0.257	0.069
Citrate Cycle (TCA cycle)	0.829	0.935	0.331	0.197	0.235	0.761	0.327
Arachidonic Acid metabolism	0.411	0.844	0.289	0.157	0.806	0.338	0.192

[1] Addition of NFKB did not change the result

[2] Although shown together, these were considered as separate pathways

3.2 Probing the signal of association

In order to identify the genes within associated pathways that are primarily responsible for the evidence of each pathway association, I collected all the SNPs within genes of pathways with $P<5x10^{-4}$ (significance level $\alpha = 0.05$, Bonferroni corrected for approximately 100 pathways) and I applied HyperLasso[138], a variable selection procedure to determine a subset of genes in associated pathways that predominantly account for the genetic effect. The procedure, under the framework of 10-fold CV to adjust for over-fitting, was applied separately for CD, T1D and RA as different pathways (thus different sets of genes and SNPs) were found to be significant in each disease (Table 7 and Table 8). As shown in Table 8, the models developed during CV consisted on average of 205 SNPs (~149 genes) in T1D, 350 SNPs (~189 genes) in RA and 493 SNPs (~277 genes) in CD.

Table 8. Summary statistics of variable selection and logistic regression step[1]

Variable selection and model fitting	T1D	RA	CD
SNPs (in at least 1 fold)	665	1032	1589
Genes (in at least 1 fold)	299	316	478
SNPs in every fold	32	57	62
Genes in every fold (any SNP of the gene)	52	88	118
Genes in every fold (the same SNP selected)	29	48	55
Average SNPs (per model, over all folds)	205	350	493
Average genes (per model, over all folds)	149	189	277
Average dominant SNPs (per model, over all folds)	54	77	100
Average recessive SNPs (per model, over all folds)	48	104	145
Average additive SNPs (per model, over all folds)	48	73	112
Average heterozygous SNPs (per model, over all folds)	53	97	134

The pathway signal might arise from a powerful effect of a small number of variants within each pathway or from multiple variants each contributing a small effect. To explore both scenarios, I displayed for each associated pathway, the fraction of the pathway genes selected on average by variable selection (Figure 6).

For some large pathways such as T cell activation (composed of 273 genes), 30-45% of the genes contributed to the effect in all three diseases, while some small pathways (such as NOD2 or IL-23, composed of 4 and 7 genes respectively) were selected almost in their entirety.

[1] Numbers were rounded to the closest integer.

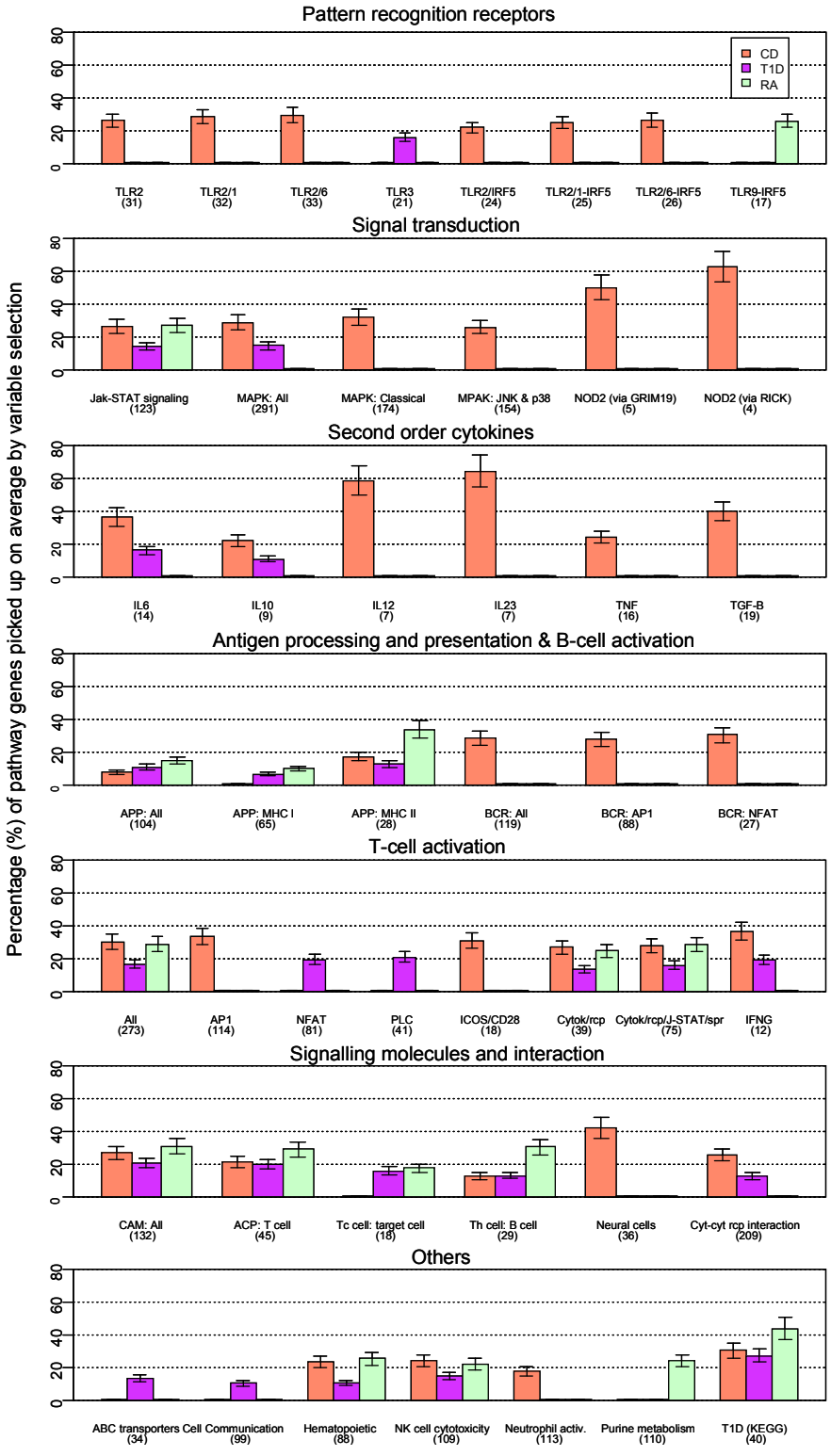


Figure 6. Fraction of the associated pathways selected on average by variable selection.
The bar charts show for each disease the average number of genes in a pathway that were picked out by the variable selection algorithm, divided by the total number of genes in that pathway and expressed as a percentage. Only associated pathways are shown. The total number of genes in a pathway is shown in parenthesis below the pathway name. Bar-charts are colour coded with respect to the disease of association. Pathways are shown grouped in functional categories. The error bars show the standard deviation as calculated by the 10 models.

The percentage of genes in a pathway contributing to the association differed in each disease, however not substantially. For example, in the leukocyte migration pathway, 20% of the pathway was selected in CD, 17% in RA and 12% in T1D.

To examine potential bias that might be introduced by the significant hits in each pathway I calculated Pearson's correlation coefficient (r^2) between the $-\log_{10}(P)$ and the number of significant hits per pathway, where P is the P-value of the pathway statistic. To assess the bias from the pathway size on the cumulative trend test statistic, I calculated the r^2 between the $-\log_{10}(P^*)$ and the pathway-size in number of SNPs, where P^* is the P-value of the pathway statistic after removal of significant hits. The correlation coefficient with respect to the pathway size was not alarming, while the strong correlation with the number of significant hits was somewhat expected. It would be surprising to see pathways with significant hits within them not to show association with the disease. Nevertheless, I was mostly keen to examine the extent to which SNPs with great marginal effects drove the association of a pathway as a whole.

Table 9. Correlation coefficient of –log10 (*P*-value) with pathway size and number of significant hits

	Correlation coefficient with		
	T1D	**RA**	**CD**
Pathway size (no of genes)	0.04	-0.017	0.490
Significant hits	0.75	0.799	0.351

In Table 10 I show the number of significant hits in each pathway. It is clear that several pathways were associated with a disease, although they had no mapped SNPs with individual single SNP trend test $P<5\text{x}10^{-7}$ (WTCCC threshold). To name a few, "Jak-STAT signalling" with all three autoimmune diseases, the "B-cell activation" sub-pathways and "Cell adhesion molecules" with CD, "IL6" and "IL10" with CD and T1D and "TLR2 signalling via IRF5" with RA. Interestingly, "T-cell activation via AP1", although it contains 3 significant SNPs it was not found associated with either T1D or with RA, and "TNF" pathway showed no association with T1D despite its 1 hit. Of note is also the "Cell communication" pathway, which has 3 significant SNPs associated with T1D and 2 with RA, and was found to be only associated with T1D.

These findings encouraged me to believe that my pathway-based approach could be used as a complementary strategy to single-SNP analysis, as my results did not seem to be substantially biased by the single-SNP analysis results. Of course in the case of a pathway with numerous significant hits, such as the "Cell adhesion molecules" with several significant SNPs from the MHC genes, the pathway statistic is driven by the single-marker test results. However, variable selection on such a pathway is likely to identify additional components that may be associated, other than the obvious HLA genes and as it is shown in Figure 6, the variable selection algorithm picked up the ~40% of the Cell adhesion molecules pathway (~53 genes).

To further investigate this matter, I repeated the cumulative trend test statistic calculation for each pathway with the significant hits excluded. As shown in Table 9 the majority of pathways retained the statistical significance even after the removal of highly associated markers. This provided evidence that the association was accumulated from several moderately significant markers and not only from a few hits. Not surprisingly very few pathways, such as the "TLRs" with CD, the "ABC transporters" with T1D and the "Haematopoietic cell lineage" with RA did not retain their significance when the highly significant hits were removed.

Table 10. List of examined pathways with the corresponding numbers of genes and SNPs.

The column of significant hits shows the number of SNPs with individual trend test $P < 5x10^{-7}$ within each pathway. The asterisk denotes association of a pathway with the disease. Pathways are color-coded based on their broader functional setting as in Table 7. The cut-off $P < 5x10^{-7}$ is used as rule of thumb in GWAS single hit definition[134].

PATHWAY	Genes	SNPs	CD	T1D	RA
Pattern recognition receptors					
TLR2	31	219	**6**	0	0
TLR2/1	32	221	**6**	0	0
TLR2/6	33	244	**6**	0	0
TLR3	21	167	0	**0**	0
TLR4 (MyD88 dependent)	38	315	0	0	0
TLR4 (MyD88 independent)	39	503	0	0	0
TLR9	21	179	0	0	0
TLR2-IRF5	24	142	**6**	0	0
TLR2/1-IRF5	25	144	**6**	0	0
TLR2/6-IRF5	26	167	**6**	0	0
TLR4 (MyD88 dependent)-IRF5	30	218	0	0	0
TLR5-IRF5	11	61	0	0	0
TLR9-IRF5	17	119	0	0	**0**
Signal transduction					
Jak-STAT signalling	123	1094	**0**	**0**	**0**
MAPK: All	291	4757	**0**	**1**	**0**
MAPK: Classical MAPK	174	3502	**0**	0	0
MAPK: JNK & p38 MAPK	154	1623	**0**	1	0
NFKB	20	210	0	0	0
NOD1	5	32	0	0	0
NOD2 (via GRIM19)	5	69	**5**	0	0
NOD2 (via RICK)	4	53	**5**	0	0
Second order cytokines					
IL-1	19	230	0	0	0
IL-6	14	143	**0**	**0**	0
IL-10	9	114	**0**	**0**	0
IL-12	7	79	**0**	0	0
IL-18	16	165	0	0	0
IL-23	7	70	**5**	0	0
TNF	16	106	**6**	1	0
TGF-β	19	298	**0**	0	0
Antigen processing and presentation					
all	104	436	**0**	**52**	**23**
MHC I	65	269	0	**16**	**5**
MHC II	28	194	**0**	**31**	**16**
B – cell activation					
All	119	1660	**0**	0	0
AKT	18	159	0	0	0
AP1	88	1272	**0**	0	0
NFAT	27	546	**0**	0	0
PKC	21	499	0	0	0
T – cell activation					
All	273	3883	**0**	**2**	**1**
AP1	114	1713	**0**	1	1
NFAT	81	1728	0	**1**	1
PLC	41	755	0	**1**	1
ICOS/CD28	18	153	**0**	0	0
Cytokines/receptors	39	311	**0**	**1**	**0**
Cytokines/receptors/Jak-STAT/suppressors	75	644	**0**	**1**	**0**
IFN-γ	12	117	**0**	**0**	0
Signalling molecules and interaction					
Cell adhesion molecules: All	132	3330	**0**	**48**	**24**
APC: T cell	45	336	**0**	**36**	**17**
Tc cell: target cell	18	93	0	**5**	**1**
Th cell: B cell	29	187	**0**	**31**	**16**
Leukocyte : platelet	9	82	0	0	0
Leukocyte : endothelial cell	28	381	0	0	0
Neural cells	36	2352	**0**	0	0
Cytokine-cytokine receptor interactions	209	1880	**0**	**1**	0
Others					
ABC transporters	34	616	0	**9**	3
Cell communication	99	1500	0	**3**	2
Complement: All	28	320	0	0	1
Haematopoietic cell lineage: All	88	798	**0**	**9**	**4**
Leucocyte transendothelial migration	89	1401	0	0	0
Natural killer cell mediated cytotoxicity	109	1192	**0**	**12**	**2**
Neutrophil	113	736	**0**	1	0
Purine metabolism	110	2226	0	0	**0**
Pyrimidine metabolism	59	556	0	0	0
Type 1 Diabetes	40	484	**0**	**38**	**18**
Non-inflammatory pathways					
Urea cycle	28	164	0	1	0
Citrate Cycle (TCA cycle)	24	243	0	0	0
Arachidonic Acid metabolism	40	343	0	0	0

Table 11. *P*-values of the pathway statistic over the examined pathways without any SNPs with single-marker *P*<5x10^{-7}.

P-values in bold and scientific format correspond to pathways that preserved their statistical significance even after the removal of the significant hits. *P*-values in italics stand for pathways that were statistically significant when the significant hits were included but not when they were excluded from analysis. Pathways are color-coded based on their broader functional setting as in Table 7.

PATHWAY	CD	T1D	RA
Pattern recognition receptors			
TLR2	*0.199*	0.029	0.041
TLR2/1	*0.157*	0.030	0.026
TLR2/6	*0.156*	0.020	0.065
TLR3	0.001	**3.3E-04**	0.037
TLR4 (MyD88 dependent)	0.020	0.010	0.040
TLR4 (MyD88 independent)	0.018	0.018	0.029
TLR9	0.127	0.062	0.397
TLR2-IRF5	*0.662*	0.225	0.001
TLR2/1-IRF5	*0.540*	0.228	0.001
TLR2/6-IRF5	*0.502*	0.171	0.001
TLR4 (MyD88 dependent)-IRF5	0.022	0.076	0.017
TLR5-IRF5	0.441	0.804	0.032
TLR9-IRF5	0.002	0.058	**1.6E-06**
Signal transduction			
Jak-STAT signalling	**7.4E-09**	**1.9E-12**	**4.4E-09**
MAPK: All	**6.4E-05**	*0.001*	0.119
MAPK: Classical MAPK	**2.2E-05**	0.001	0.164
MAPK: JNK & p38 MAPK	*0.025*	0.055	0.139
NFKB	0.825	0.016	0.010
NOD1	0.833	0.324	0.633
NOD2 (via GRIM19)	*0.085*	0.055	0.237
NOD2 (via RICK)	*0.160*	0.587	0.528
Second order cytokines			
IL-1	0.003	0.217	0.127
IL-6	**3.5E-04**	**4.2E-04**	0.064
IL-10	**8.4E-06**	**2.1E-04**	0.083
IL-12	**4.9E-06**	0.004	0.003
IL-18	0.017	0.062	0.032
IL-23	**4.5E-11**	0.017	0.063
TNF	*0.091*	0.017	0.026
TGF-β	**2.6E-05**	0.004	0.035
Antigen processing and presentation			
all	**1.4E-04**	**3.1E-14**	**5.2E-11**
MHC I	0.073	**4.1E-05**	0.004
MHC II	**1.9E-05**	**1.4E-05**	**1.5E-10**
B – cell activation			
All	**5.2E-08**	0.005	0.656
AKT	0.039	0.550	0.223
AP1	**7.6E-11**	0.037	0.578
NFAT	**1.7E-04**	0.002	0.702
PKC	0.037	0.050	0.553
T – cell activation			
All	**2.5E-11**	**1.6E-06**	**1.4E-04**
AP1	**5.3E-07**	0.002	0.189
NFAT	0.004	*0.082*	0.158
PLC	0.002	*0.002*	0.182
ICOS/CD28	**4.1E-05**	0.227	0.029
Cytokines/receptors	**4.3E-04**	**9.6E-05**	**2.5E-06**
Cytokines/receptors/Jak-STAT/suppressors	**2.4E-04**	**7.1E-08**	**1.1E-04**
IFN-γ	**1.7E-05**	**8.6E-06**	0.018
Signalling molecules and interaction			
Cell adhesion molecules: All	**0.0E+00**	*0.008*	*0.005*
APC: T cell	**5.0E-06**	**1.6E-10**	**0.0E+00**
Tc cell: target cell	0.010	**5.5E-07**	**8.3E-10**
Th cell: B cell	**2.5E-04**	**4.2E-08**	**0.0E+00**
Leukocyte : platelet	0.208	0.868	0.631
Leukocyte : endothelial cell	0.029	0.095	0.257
Neural cells	**7.6E-05**	0.123	0.125
Cytokine-cytokine receptor interactions	**4.2E-15**	**6.3E-05**	0.002
Others			
ABC transporters	0.021	*0.004*	0.312
Cell communication	0.014	*0.345*	0.265
Complement: All	0.080	0.323	0.528
Haematopoietic cell lineage: All	**6.3E-10**	**5.5E-05**	*0.016*
Leucocyte transendothelial migration	0.013	0.575	0.102
Natural killer cell mediated cytotoxicity	**3.3E-10**	*0.046*	**5.2E-05**
Neutrophil	**4.5E-05**	0.003	0.043
Purine metabolism	0.010	0.026	**3.2E-05**
Pyrimidine metabolism	0.070	0.160	0.039
Type 1 Diabetes	**4.8E-07**	**8.0E-08**	**6.8E-10**
Non-inflammatory pathways			
Urea cycle	0.928	0.284	0.771
Citrate Cycle (TCA cycle)	0.829	0.935	0.331
Arachidonic Acid metabolism	0.41	0.84	0.29

3.3 Identifying key genes in associated pathways

I reasoned that the genes, which were identified in all ten CV models for each of the three diseases, would represent the "core" genes showing consistent association with each disease. This core comprised a total of 52 genes in T1D, 88 genes in RA and 118 genes in CD, and their overlap is shown in the Venn diagram (Figure 7). Only 12 genes were common to all three diseases, 11 were shared by CD and T1D, 5 shared by T1D and RA, and 26 were shared by RA and CD. The majority of identified genes were unique to each disease (Figure 7). Common to all three diseases are genes of the major histocompatibility complex (MHC): the class II molecule *HLA-DQB1* (the selected SNP in the models was rs9273363 in T1D; rs9275418 in RA; rs9469220 in CD), the class I molecule *HLA-C* (rs9366778 in T1D; rs2524067 in RA; rs7452890 in CD) and the foetally-expressed non-classical class I paralogue*, HLA-G* (rs1632882 in T1D; rs1063320 in RA; rs2253981 in CD); the protein phosphatase regulatory subunit *PPP3R2* (rs10512298 in T1D; rs10990238 in RA; rs10123826 in CD), the phospholipase *PLA2G4A* (rs12088162 in T1D, rs10912209 in RA; rs10494612 in CD) and genes associated with cell signalling: inositol 1,4,5-triphosphate receptor type 1 *ITPR1* (rs9850086 in T1D; rs2291861 in RA; rs2053502 in CD), the guanine nucleotide exchange factor *VAV3* (rs4914946 in T1D; rs12752530 in RA; rs1418992 in CD), and the kinase *PAK7* (rs6118724 in T1D, RA, CD) from the T cell activation pathways; and cell adhesion molecules: activated leukocyte cell adhesion molecule *ALCAM* (rs995470 in T1D; rs3772556 in RA; rs7643160 in CD), a neuronal cell surface protein *NLGN1* (rs7430111 in T1D; rs9838752 in RA and CD), an integrin receptor subunit *ITGA1* (rs7731949 in T1D; rs11745801 in RA; rs2448420 in CD) and the cadherin *CDH2* (rs1404432 in T1D; rs1496479 in RA and rs2591108 in CD).

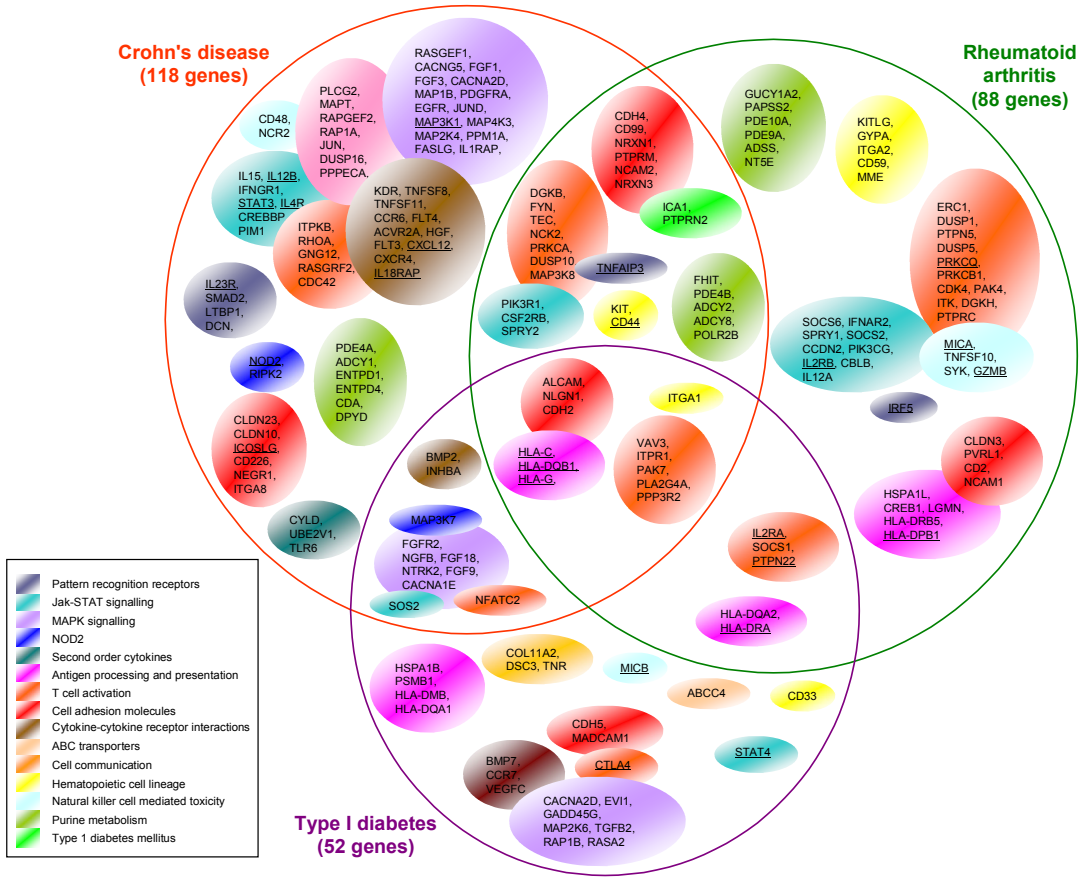


Figure 7. Genes identified by variable selection in all 10 folds of cross-validation for T1D, RA and CD.

Key genes that showed consistent association with each disease in all 10 logistic models of cross validation. Genes are grouped in their pathways, which are shown as bubbles. Pathways are colour-coded in agreement with the colors in Table 7, which in their turn represent broader cellular processes where the pathways are involved. Overlapping bubbles represent pathways that share key genes. Genes that were previously associated in GWAS are underlined. They "Type 1 diabetes mellitus pathway" highlighted in green contains genes that are exclusively found in this and no other of the rest of the pathways.

My pathway-based approach identified several genes that have been associated with autoimmune diseases in previous GWAS studies[104,134,142-144] (Figure 7). These include for T1D, genes in the HLA region, the negative regulator of T cell activation *CTLA4* (rs11571304, rs3087243), the suppressor of T cell activation *PTPN22* (rs6679677) and the interleukin 2 receptor alpha *IL-2RA* (rs2104286, rs6602360). For RA, *PRKCQ* (rs7096622), *IL2RA* (rs10795791) and *IL2RB* (rs3218253). Similarly for CD, the well established associations with *NOD2* (rs2076756), *IL23R* (rs11805303, rs17375018) and the recently implicated *IL12B* (rs4921227) and *STAT3* (rs744166) were all confirmed and these genes were selected in all models of CV. However, in addition to these, I identified other components of key pathways contributing to disease, such as *RIPK2* (rs160448, rs160410) and *MAP3K7* (rs791059, rs2069314) from the *NOD2* pathway in CD. I also identified novel associated pathways for further exploration. For example in T1D, a number of genes controlling T cell activation were selected such as the calcium channel *ITPR1* (rs9850086); the calcium dependent phospholipase *PLA2G4A* (rs12088162), the regulatory subunit of calcineurin *PPP3R2* (rs10512298) and the calcineurin dependent transcription factor *NFATC2* (rs6067742) suggesting a role for calcineurin /nuclear factor of activated T-cells (*NFAT*) signalling in susceptibility of T1D.

TNF plays a critical role in inflammation in RA and CD[129,130] and has been a major target for therapeutic antibody treatments. The TNF pathway was significant in CD only, and key components selected include two enzymes that are regulators of NF-κB signalling – a negative regulator of NF-κB, the deubiquitinating enzyme *CYLD* (rs7342715) [145] and *TNFAIP3* (rs7753394) both a deubiquitinating enzyme and a ubiquitin ligase.[146] Both of these enzymes deubiquitinate NEMO and, when knocked out in vivo, lead to inflammatory bowel disease.[147] Genetic variants near *TNFAIP3* have recently been associated with RA (see below) as well as Ulcerative colitis and CD[93,134] and was associated in my study for both CD (rs7753394) and RA (rs6920220) (see below).

Considering the possible involvement of TLR signalling in RA, I found that only TLR9 signalling via IRF5 was significant among TLR pathways. *TLR9* is constitutively expressed on ß-cells that are critical in the pathogenesis of rheumatoid arthritis.[148] The plausibility of this pathway is strengthened by the finding that several genes downstream of *TLR9* were also selected; *IRF5* (rs3807306) and the negative regulators *SOCS1* (rs11074956, rs243325) and *TNFAIP3* (rs6920220). It is of interest that recent studies

show an association of the *IRF5* gene with RA[149], and 2 studies link region 6q23, flanked by *TNFAIP3* and *OLIG3* with RA susceptibility.[150,151]

3.4 Genetic prediction of disease risk

I next investigated how well the pathway-centric combination of gene variants selected by variable selection could predict the disease status in individuals. The sensitivity and specificity of all 10 models for each disease are shown in ROC curves (Figure 8A-C). The area under the average ROC is 91%, 85% and 60% for T1D, RA and CD respectively, which correlates inversely to the number of SNPs fitted in the models for each disease.

The fitted models, as expected, contained some of the well-established associated markers with each disease. To evaluate the added value of genes identified by my approach versus a single-SNP approach, I refitted the models excluding SNPs with single-SNP $P < 5\text{x}10^{-7}$ as well as any SNPs in LD with these ($r^2 \geq 0.3$), and (for T1D and RA) any SNPs in the HLA genes; I also refitted the models with only covariates these excluded SNPs. As seen in Figure 8A-C, for all three diseases, the complete set of SNPs (green curves) had greater predictive value than the significant hits alone (red curves). For RA, the SNPs identified by my pathway approach excluding the significant hits (blue curves), have higher predictive value than the significant hits alone (red curves). For CD and T1D a large proportion of the predictive power comes from established associations (such as the HLA SNPs), however the SNPs identified by the pathway analysis have predictive power in the absence of these hits and also increase the predictive power in their presence.

3.5 Validation Study

The 10 CV logistic regression models trained on 90% subsets of the WTCCC T1D dataset were tested on 4763 individuals from the Northern Finland Birth Cohort of 1966 (NFBC1966) of whom 30 developed T1D by age 30.[135] As shown in Figure 8D, I predicted on average over 60% of the Finnish T1D cases with a false positive rate of 10%, compared to 73%-10% in the original UK sample (Figure 8A). The AUC was 0.79. Most importantly the predictive power of the newly identified genes (blue curve) is 70%, this is exactly the same as in the WTCCC study.

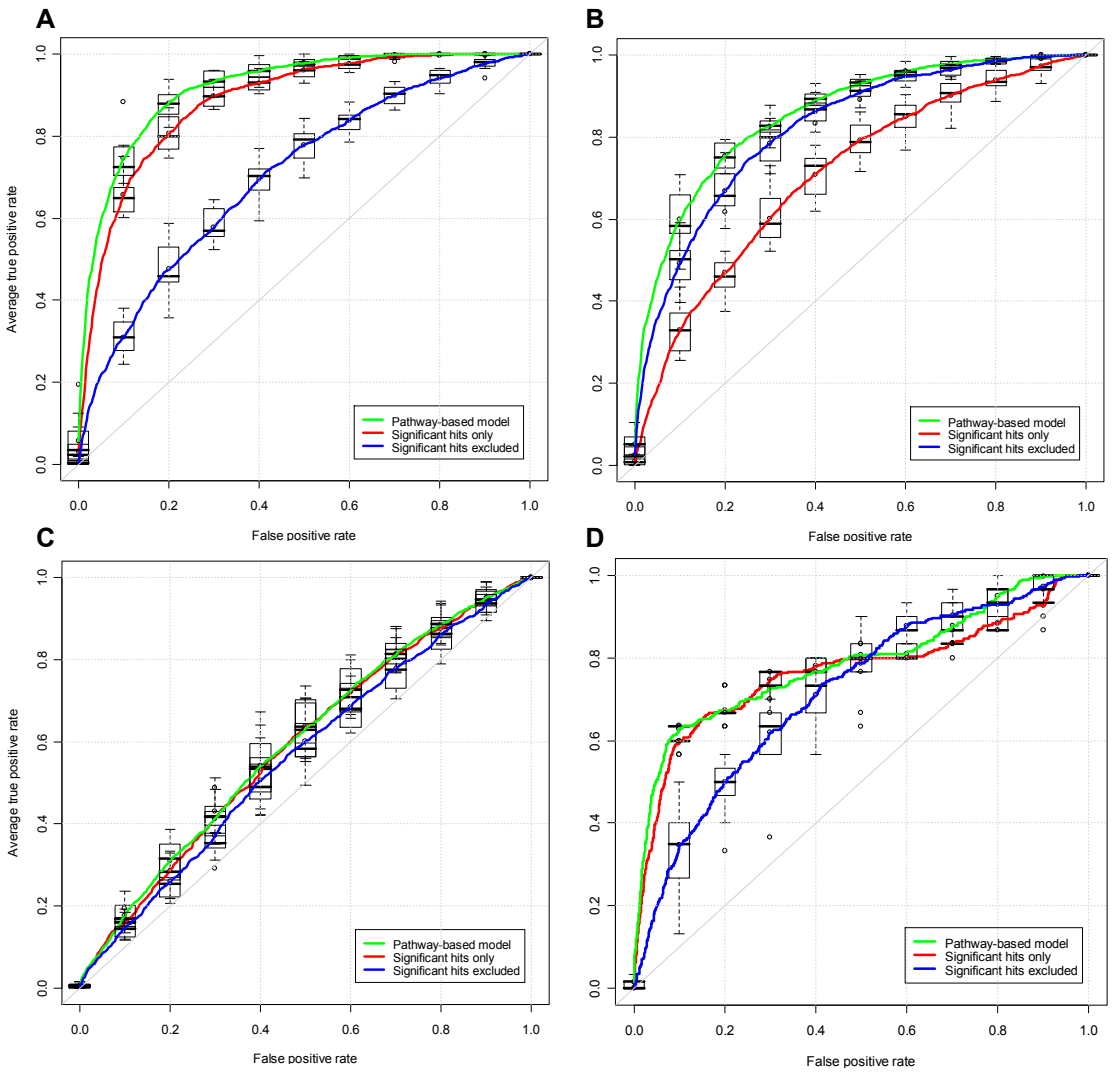


Figure 8. ROC curves showing the average predictive performance of the models fitted for each disease during 10-fold CV.

True positive rate and false positive rate for predicting case-control status for T1D (A), rheumatoid arthritis (B), and Crohn's disease (C) of the WTCCC dataset. Each green coloured curve corresponds to the average ROC of the 10 fitted logistic models of CV. Red curves stand for the average performance of the models fitted only with pathway SNPs with single-SNP $P<5x10^{-7}$ and any SNPs in LD with these ($r^2 \geq 0.3$). Blue curves show the average performance of the 10 models fitted during CV, when all SNPs with single-SNP $P<5x10^{-7}$ and any SNPs in LD with these ($r^2 = 0.3$) were excluded from analysis. The plot in panel (D) shows the performance of a single model – this justifies the jagged lines - trained on the WTCCC T1D cases and controls dataset and tested on a separate study of the Northern Finland Birth Cohort of 1966 (NFBC1966) with 4763 individuals, 30 of whom developed T1D by age 30. This model has the same AUC as the average of the 10 CV models fitted on the WTCCC and trained on the NFBC.

Although the SNPs meeting GWAS significance accounted for a considerable proportion of the predictive power, the additional SNPs identified by my approach contribute to the predictive power (Figure 8D, red curve) and have predictive value on their own (Figure 8D, blue curve).

3.6 Visualisation of genomic risk

I reasoned that each SNP has a different effect on disease predisposition either increasing or decreasing the risk due to its functional effect on the regulation of the overall pathway output. I categorized SNPs as *adverse* or *protective* on the basis of their coefficient in the CV model (relative to the minor allele for additive effects). As shown in Figure 9, each SNP exerts an *adverse* or *protective* effect through dominant, recessive, additive or heterozygous modes. In T1D for example, there are two *adverse* HLA-DQB1 variants (rs9273363/rs9275418) and two *protective* HLA-DQA1 variants (rs9272723/rs9270986), and PTPN22 (rs6679677) is *adverse* for both T1D and RA. Very few SNPs were selected in more than one diseases and I observed either same or opposite direction of the coefficients in the models of different disease. In the case of a real effect, this could be attributed to differential patterns of LD with the actual SNP or in my opinion to some residual over-fitting that I didn't account for. As a general comment, in T1D I observed a number of SNPs with large *adverse* effects acting in a dominant and additive manner. Furthermore, there are more *adverse* than *protective* SNPs with large effect size (>0.3). In RA, there are SNPs with large additive *adverse* and *protective* effects (middle panel, Figure 9). In contrast, in CD there are fewer SNPs with large effect size (>0.3) but these act in a heterozygous and recessive manner (bottom panel, Figure 9). Three SNPs of PPM1A have large heterozygous effects (rs7154773 having a risk effect, and rs8019531 and rs8012815 having *protective* effects). Smaller recessive *adverse* effects were seen in the case of IL6R (rs11265618) and RHOA (rs9814873), and *protective* effects in FHIT (rs10510813).

To explore these differences further, for each genetic mode I calculated the sum of encoded genotypes at SNPs which act according to this genetic mode, weighted by their coefficient in the logistic regression model. This is a partitioning of the estimated log-odds of disease according to genetic mode. I calculated the significance of differences in these sums between cases and controls in the 10% test data set. For T1D, cases and controls had a significantly different sum of additive (*t*-test, $P=2\text{x}10^{-6}$) and dominant (*t*-test, $P=2\text{x}10^{-6}$) effects; whereas for RA additive (*t*-test, $P=2\text{x}10^{-14}$), dominant (*t*-test, $P=6\text{x}10^{-4}$) and recessive (*t*-test, $P=0.014$) are significantly different; and in CD additive ($P=4\text{x}10^{-3}$) and heterozygous (*t*-test, $P=0.04$) are significantly different.

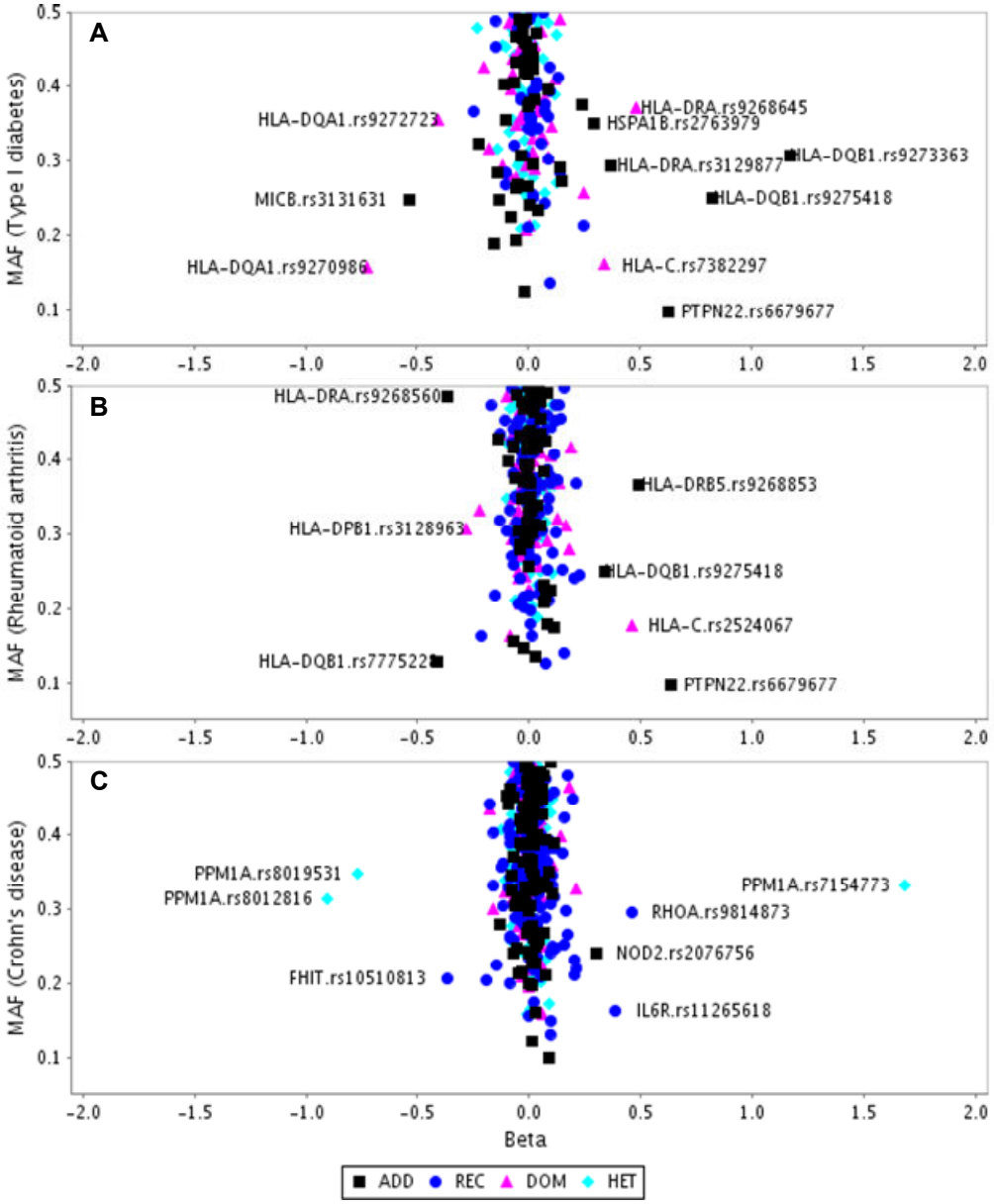


Figure 9. The allelic architecture of the logistic regression models of disease risk.
Minor allele frequency vs. beta coefficient for each SNP retained in the fitted logistic regression model from first round of cross-validation. Additive, recessive, dominant and heterozygous effects are displayed by black squares, blue dots, mauve triangles and cyan diamonds respectively. Adverse vs. protective SNPs have a positive vs. negative beta value respectively. Labels are given for all SNPs with beta greater than 0.25

The ratio of *adverse* to *protective* variants carried by each individual was positively correlated with the log odds of disease risk (Figure 10). In Figure 11, I plot for every individual (columns) their genotype at each SNP, with red indicating an *adverse* and green indicating a *protective* SNP (rows). The intensity of colour reflects the effect of the SNP in the model (see section 3.8). Each individual seems to carry a unique combination of *adverse* and *protective* variants, which may represent a personal "genomic fingerprint" of disease predisposition. Despite the vast number of different combinations, a pattern is

revealed where patients (left side) compared to healthy individuals (right side) are seen to carry a higher number of *adverse* genotypes and fewer number of *protective* genotypes and vice versa.

.

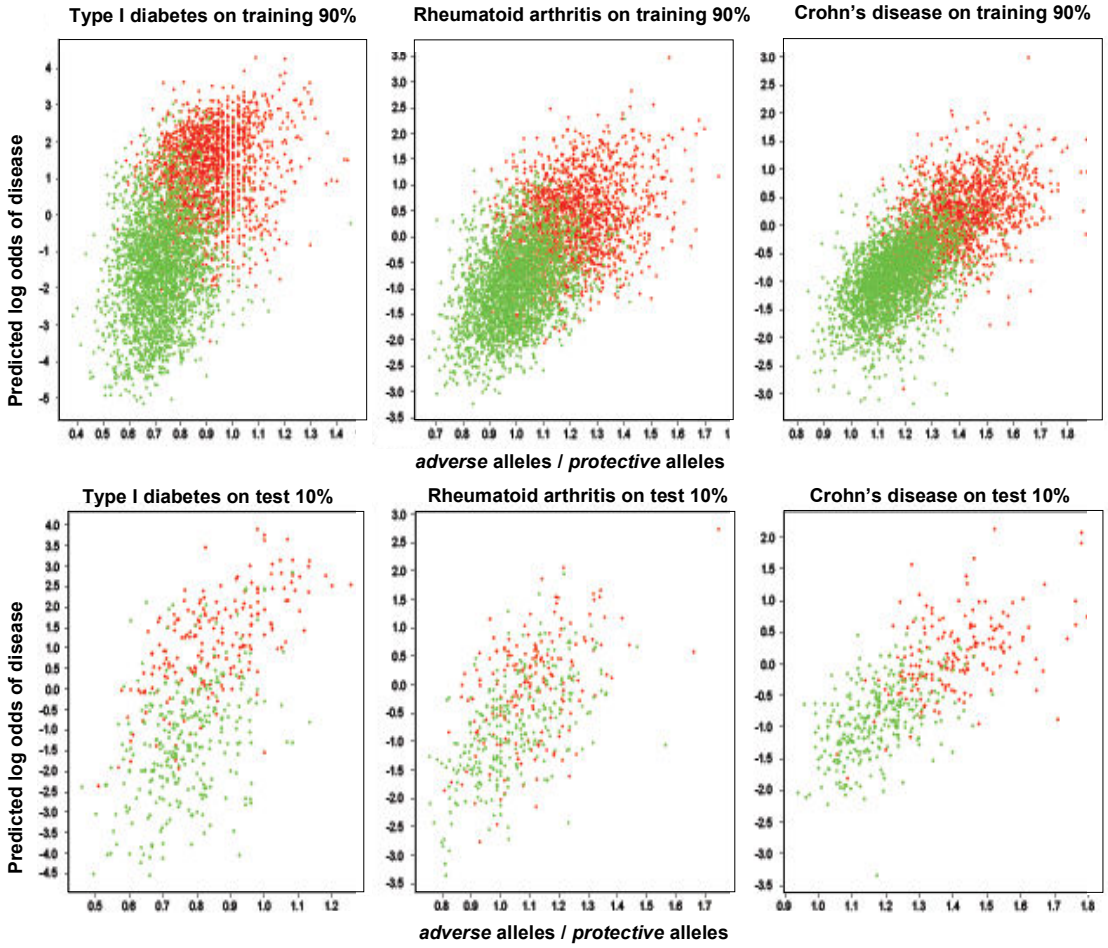


Figure 10. Ratio of adverse/protective genetic variants carried by cases and controls and predicted log odds of disease.

Predicted disease log-odds (y axis) in cases (red) and controls (green) vs. the ratio of the number of adverse to protective encoded genotypes according to genetic mode for T1D, RA and CD in the training set (top 3 panels), and the test set (bottom 3 panels). For each dataset, the log-odds of disease increases with increasing ratio. (Pearson's r^2 examined on the test set was 0.51, 0.54, 0.58 for T1D, RA, and CD respectively)

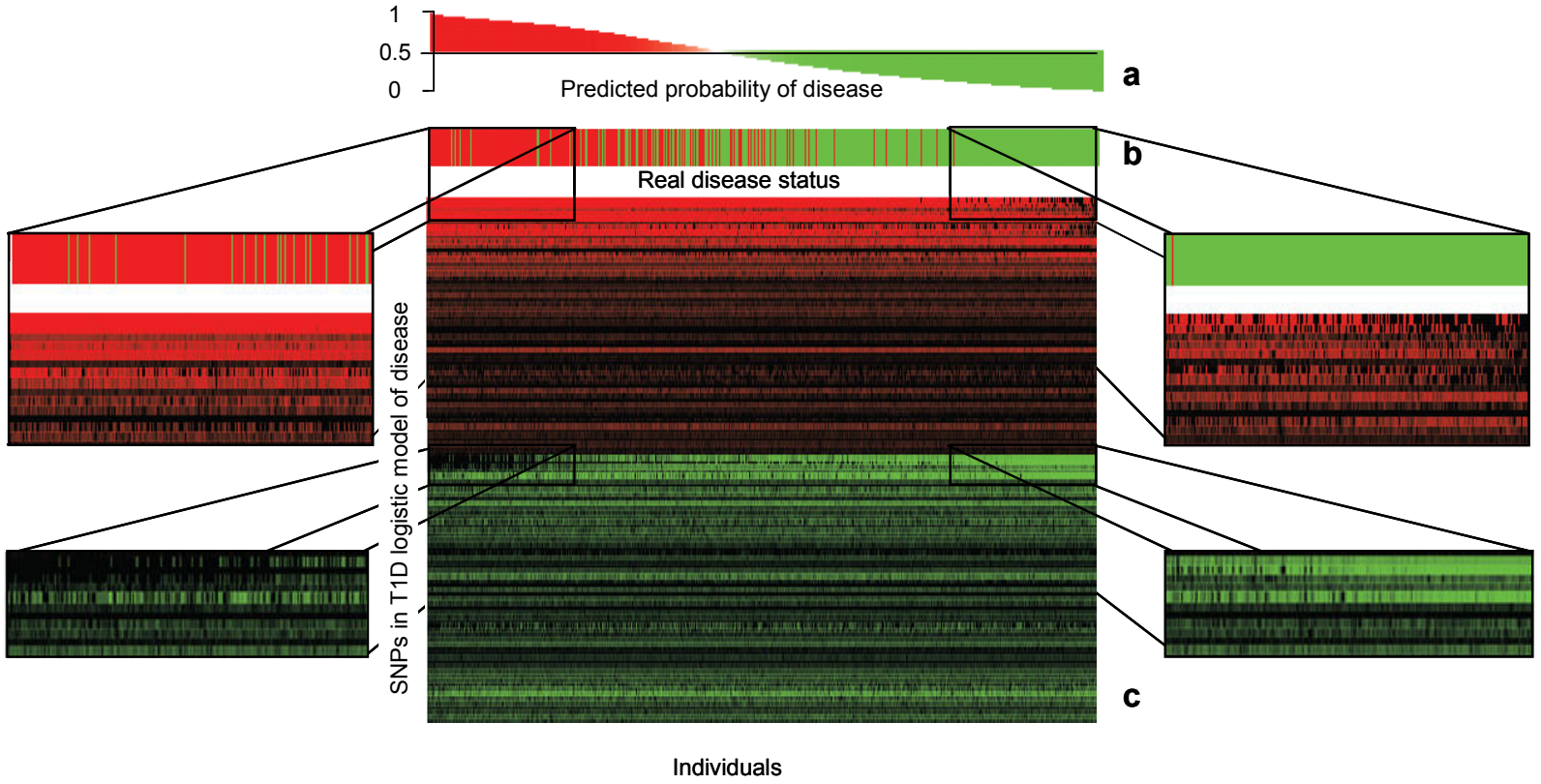


Figure 11. Variant SNPs carried by cases with T1D and controls.

(a) Predicted probability of being a case, (b) Actual case or control status. Patients shown in red and controls in green. The model correctly assigns the majority of cases at the extreme left, and controls at extreme right, with less predictive ability in the middle. (c) Individual patients or controls are displayed in columns and each row represents one SNP. Red indicates an *adverse* and green a *protective* SNP. Black indicates zero contribution of the SNP as the individual carried zero of the SNP adverse/protective allele. Intensity of colour indicates disease log-odds from the predictive model. The magnified sections show regions where very marked differences between cases and controls can be readily seen.

3.7 Discussion

My analysis places the genetic basis of common autoimmune disease in a pathway context. Although there has been increasing evidence[15,104,152,153] that multiple genes, which functionally interact in a variety of pathways, are associated with autoimmune diseases, I have shown directly that variation in conserved canonical inflammatory pathways underlies genetic susceptibility to T1D, RA, and CD. My results suggest that generic activities involved in innate and adaptive immunity, such as T-cell activation, antigen presentation and cell adhesion, constitute common underlying processes of autoimmune predisposition. In addition to these shared pathways I identified a number of disease-specific pathways, which indicates distinct modulating mechanisms for each condition.

At the gene level, my analysis identified new genes associated with each disease, as well as replicated previously identified loci, which confirm the consistency of my approach to results from previous studies. Of particular interest is the fact that my analysis identified *IL12B*, *ICOSLG* and *STAT3* for CD, were only identified in a meta-analysis of three separate GW scans[144], as well as STAT4 with T1D , which was found associated with the disease .in a follow-up candidate gene study[154] and not on the original WTCCC study. The latter findings serve as a proof of concept that genes with moderate effects, which are missed by single-SNP analysis of GWAS can be extracted via a pathway based approach.

Pathway-based approaches have been employed in studies of obesity[155], Parkinson disease[32], Breast cancer[22], hyperlipidaemia[156], age-related Macular degeneration[157] and the WTCCC dataset[40], but have focused less on the underlying mechanisms that affect disease occurrence. My approach provided a new insight in the number of pathways, the genes implicated and the underlying genetic architecture for disease risk, due to my strong unifying hypothesis that autoimmune genetic influence operates through functionally interacting genes involved in inflammatory pathways. In many diseases the relevant pathways are not yet as clearly identified. However, as the understanding of function of genes increases[158,159], the same approach may be readily applied to other diseases.

My findings provide new insight into the genomic architecture underlying common disease risk. Most studies to date have reported the odds ratio at each single locus separately as a measure of increased risk. My results suggest that disease in any individual occurs as a result of altered balance between *adverse* and *protective* variants at several loci that affect the overall function of key pathways in opposite directions such as through positive or negative regulatory effects.

Similar mechanisms have been suggested in studies that used simulated data to model the effect of multiple genetic variants.[160,161] Furthermore the majority of studies have focused on additive SNP genetic models, but I have shown, in addition, that dominant, recessive and heterozygous effects contribute risk in all three diseases. The relative contribution of each type of effect differs between the three diseases, with dominant and additive effects contributing more in T1D, dominant and recessive effects in RA and heterozygous effects in CD.

I have been able to identify gene/SNP sets within inflammatory pathways which can be used as predictive markers of disease. The high performance of the logistic models in T1D and RA raises the possibility of genomic disease prediction and diagnosis, with greater accuracy than by using combinations of single genes in previous analysis.[156,160]

The validation of my T1D models on a different population provides the most stringent test of my approach. The predictive models trained on the UK WTCCC samples predicted development of T1D in 60% of the Finnish patents who developed disease by age 30. I speculated that the level of prediction did not reach that of the WTCCC CV models, partially due to the use of SNPs imputed in the Finnish samples with a imperfectly matched reference population and partially because the Finnish population differs significantly from UK Caucasians in haplotypic structure.[162] However, the predictive power was kept very high and to my knowledge higher than any other study has shown.[163-165] Although much of the predictive power in T1D was due to HLA genes effects, I have shown that remaining SNPs in my predictive models have substantial predictive value in their own right, and in RA, the genes not meeting GWAS significance provide better predictive value than the significant hits. The lower predictive performance of the models in CD may reflect that this disease (which has a number of well recognised clinical phenotypic subgroups[160]) is more heterogeneous in its genetic origins, a possibility also suggested by the higher number of SNPs identified by variable selection in CD.

Nevertheless, environmental factors are beginning to be identified for some autoimmune disorders and integrated into models of pathogenesis. For example, smoking is an established risk factor for RA[166] and there has also been evidence of a strong interaction between HLA-genes and smoking for the development of future RA. Similarly, smoking and dietary influences have been identified as an important risk factor for CD.[167] I would expect that the inclusion of environmental factors, gene-gene and gene-environment interactions (if data is available) will improve these predictive models.

My approach suggested a new picture of how variation in multiple genes linked in functional pathways contributes to autoimmune disease susceptibility and provides a useful tool to reveal the hidden information of GWAS that would be missed in single SNP analysis.

3.8 Summary

In this chapter, I presented the results of my pathway-based analysis on the WTCCC dataset and I showed that it is a powerful approach to directly identify plausible pathway-disease associations between inflammatory pathways and autoimmune diseases. I also showed that my methodology is able to reveal additional susceptibility markers (further to the single-SNP hits) that drive predominantly the pathway associations and can also assist in disease prediction. In the next chapter, I go on to improve the pathway-driven analysis by addressing some of the limitations of my approach and by taking the gene selection part a step forward.

Chapter 4

Pathway-driven gene stability selection

The application of my proposed strategy on the WTCCC dataset yielded promising results; nevertheless I identified a number of methodological limitations that had to be addressed before applying it on multiple datasets.

To start with, the gene prioritization methodology was entirely based on the gene inclusion frequency in 10 models of a CV procedure. It is argued that 10-fold CV is a robust technique to provide a reliable estimate of the out-of-sample prediction error but this does not necessarily imply that it works as efficiently in identifying relevant predictors. It has been suggested that relevant predictors could be efficiently identified by the inclusion frequency in

resampled datasets[168,169] and on these grounds I set out to refine my gene-based selection procedure.

At the same time, my methodology had not been substantially validated at the pathway and gene level on two independent GWAS of the same disease, because I initially had no access to such datasets. During the second year of my PhD, I obtained access to a second GWAS on RA and I used it to verify my findings. In this chapter I go beyond a simple pathway-based analysis of GWAS, to identify the genes within associated pathways that are predominantly responsible for the pathway association signal. To achieve this I recast the question of how to associate genes with a disease as a pathway-driven stability selection problem.

4.1 Introduction

Stability selection is a technique that is based on subsampling (in this case of individuals) in combination with variable selection algorithms.[170] Variable selection methods attempt to identify combinations of variables (in this case SNPs) that best explain the outcome (disease status). Regularization parameters control the number of variables selected: smaller models (stronger regularization) result in lower FDR but may miss real associated variables; larger models (weaker regularization) detect more real associated variables but also admit more false positives. Stability selection[170] involves variable selection on multiple random sub-samples of the data (in this case individuals), and across multiple regularization parameter settings.[170] Variables which are consistently chosen are reported as 'stably selected'.

I employed the stability selection concept in a pathway-based analysis setting that accommodates for validation on a second GWAS. The suggested strategy is called "pathway-driven gene stability selection" and it is conducted in the following steps: Pathway-based analysis of a GWAS is carried out using the *CT* test statistic, as described in Chapter 2. Pathway association results are adjusted for multiple testing. The same analysis is carried out in parallel on a second GWAS dataset. The *CT* test statistic *P*-values are ranked in ascending order in both studies. A pathway rank percentage cut-off analysis is carried out to identify the threshold on pathway rank significance, above which, all pathways will be considered to show evidence of association with the disease in both GWAS datasets. Only the associated pathways are taken forward to the gene-selection procedure. Gene-based stability selection is applied to identify the most influential genes within associated pathways. This procedure is performed separately for each GWAS, but the settings are such to adjust for each study sample size and equalize the discovery power between studies. Equalizing the power between studies aims for the same

amount of information to be extracted by smaller studies as in the larger and then hypergeometric modelling is used to assess the overlap. The selected genes in each study are compared between studies and the significance of the observed overlap is calculated. I consider that any genes which are associated with disease in both studies and go beyond the expected overlap by chance are true positive associations. The genes that are identified in one study but not the other are deemed "suggestive associations", as GWAS undertaken on two different populations may indeed identify different genes underlying the disease in each population. The selected genes are taken forward to an *in-silico* functional enrichment analysis in GO categories. I consider as very plausible true associations the genes that are enriched in the same functional categories in both studies, on the grounds that the same biological pathway might be associated with a disease at a population level, but different genes and variants could be responsible at an individual's level. Nevertheless those genes, albeit different, might be involved in the same cellular process having a similar function along the pathways. The flowchart of my proposed methodology is given in Figure 12.

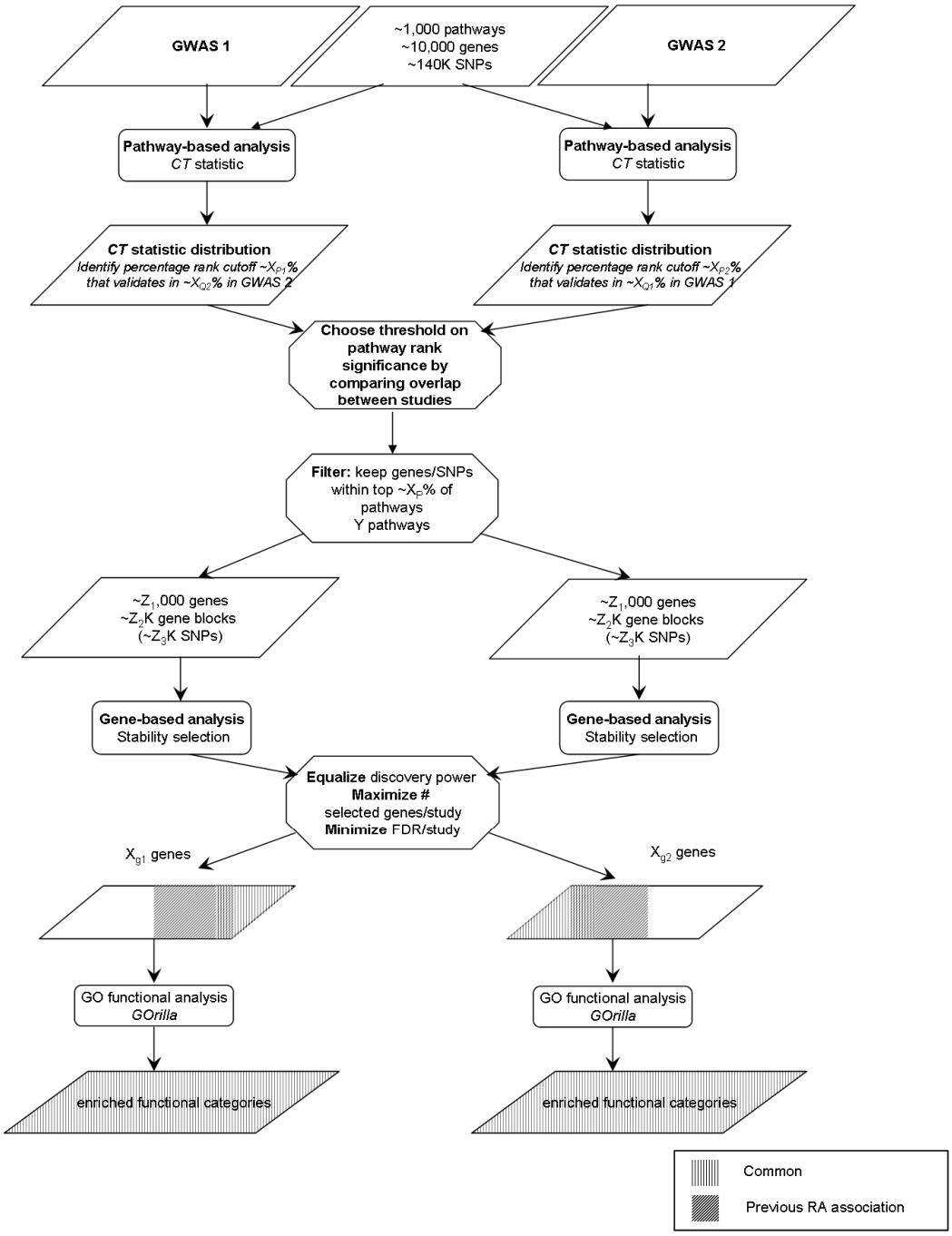


Figure 12. Schematic diagram of the proposed methodology

The technical steps of the proposed methodology are explained in detail in the following sections of this chapter. The methodology was developed and is illustrated through the application of my approach on two RA case control GWAS; the WTCCC RA study, of 1,860 cases and 2,938 controls typed on the Affymetrix 500k chip[9] and the North American Rheumatoid Arthritis Consortium (NARAC) study of 908 cases and 1260 controls genotyped on

the Illumina Humanhap550k.[8] The results of this analysis are shown separately and discussed in Chapter 5.The overlap in typed SNPs between these two platforms is ~10%.

4.2 Pathway-based analysis

Pathway-disease associations were examined by the *CT* test statistic. To recap, the *CT* test statistic[171] assesses the null hypothesis that the pathway does not comprise any genes that are associated with the disease. The null distribution is obtained by fitting a skew normal distribution to 1000 *CT* test statistic under permuted phenotype labels. In this work, I observed that the 'sn' package in R, which is the package I used to fit the skew normal distribution, does not allow the calculation of the lower tail of the cumulative distribution in log space, which enforces a lower bound on the reported significance levels. To address this problem, I calculated *P*-values by fitting a gamma distribution. I used the Kolmogorov-Smirnof test statistic, as well as visual inspection of QQ plots to establish that the gamma distribution provides as good a fit as the skew-normal. Correction for multiple testing and dependence between pathways due to shared genes was performed via the method of Benjamini-Hochberg-Yekutieli, which controls the FDR under dependency. The FDR was set at α=0.05. I tested 700 curated, biological pathways, as assembled from KEGG, BioCarta and GenMAPP. By using 3 public databases, one could argue that I am creating an additional multiple testing problem by bringing in different versions of the same pathways. In my case and as I show below, I do not employ and multiple testing correction, as I tackle the problem of overlapping pathways with a gene selection procedure and I control the FPR using the hypergeometric distribution. The procedure described above was applied independently to WTCCC and NARAC.

I calculated the χ_1^2-scores from the *CT P*-values and found that the correlation between the two studies was highly significant (Spearman's r^2=0.69, $P < 2.2\text{x}10^{-16}$). This correlation persisted even after removal of pathways that included strong and confirmed associated loci with RA, such as the HLA genes within the MHC region (Spearman's r^2=0.32, $P < 2.2\text{x}10^{-16}$). Pair-wise correlation of pathway statistics between studies (which were carried out in independent populations and on different genotyping platforms) should indicate common genetic variation associated with RA. Ranking of the pathways in both studies revealed that the top 6% of ranked pathways in NARAC laid in the top 10% of pathways in WTCCC.

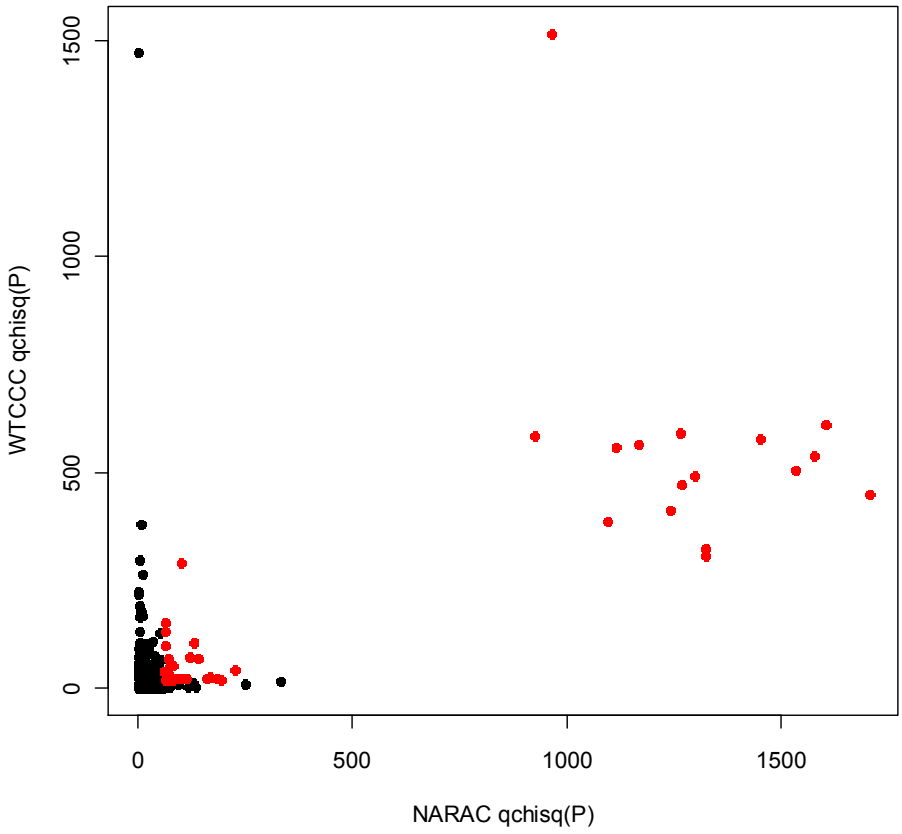


Figure 13 χ_1^2 transformation of the CT statistics in NARAC and WTCCC.

The plot shows the χ_1^2 scores of the *CT* P values in NARAC (x-axis) plotted against those in WTCCC (y-axis); the 42 pathways from the top ranks of both distributions are shown in red.

I then plotted the overlap between WTCCC and NARAC pathways as a function of pathway rank cut-off in Figure 14a. The maximum overlap was achieved for the first 2% of pathways (i.e. 14 pathways) that are mostly dominated by the known RA associations. I observed a second peak in overlap at the 6% top ranked pathways, where 70% of those were common between the studies. In the plot of Figure 14b I show (as a function of NARAC pathway rank) the maximum pathway rank in WTCCC amongst all pathways more significant than a given NARAC pathway rank, and vice-versa in Figure 14c). I observed a step-change in the plot b) at ~6% of all pathways (42 pathways), at which the maximum WTCCC rank goes from ~10% to ~45%. Conversely, in Figure 14, I observed a similar step change at 2% (14 pathways), where the rank goes from ~10% to ~70%, which indicates that WTCCC derived pathway statistics may be less reliably reproducible in NARAC. In other words, the top 6% of NARAC ranked pathways laid within the top 10% of WTCCC ranked pathways. By selecting 6% as the threshold of pathway rank in NARAC, I achieved a good trade-off between choosing a small set of highly overlapping pathways (which may miss genuine signals) and a larger set of less overlapping

pathways (which may bring in more noise). At the same time it provided a good balance between introducing noise and increasing the computational demands of the variable selection step, and potentially missing genuine association signals. I suggest using the inflection point in maximum path rank as a useful diagnostic for deciding which threshold to apply in future applications of this methodology.

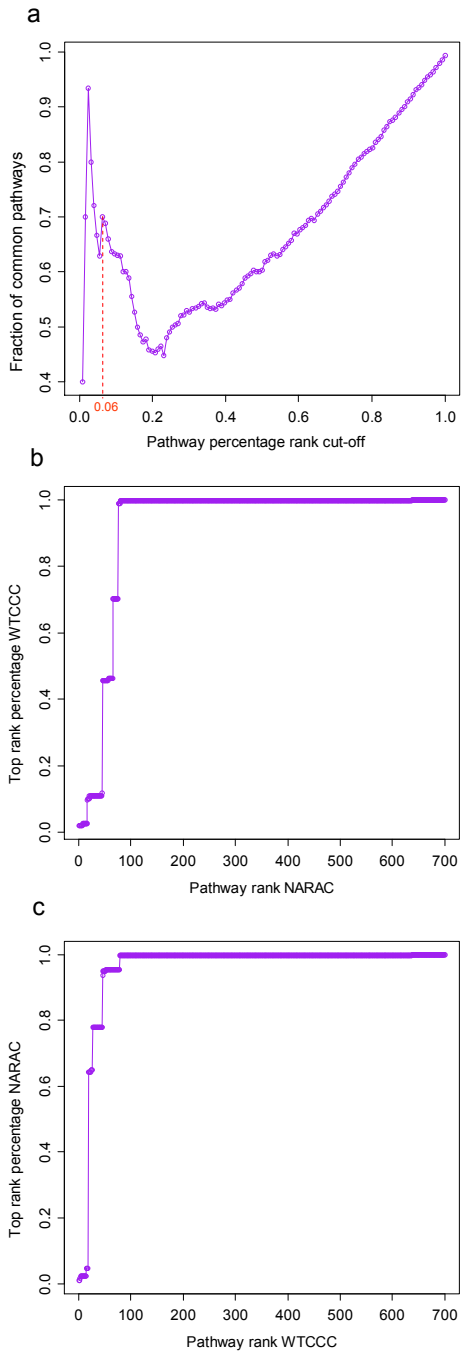


Figure 14. Overlap of NARAC and WTCCC pathway rank statistics.

a) Overlap of WTCCC and NARAC selected pathways as a function of percentage rank cut-off. The x-axis shows the percentage of ranked pathways that is used as a cut-off. The y-axis shows the fraction of pathways that are common between WTCCC and NARAC for a certain percentage rank cut-off . b) Top rank percentage in WTCCC as a function of NARAC rank threshold. c) Top rank percentage in NARAC as a function of WTCCC rank threshold.

4.3 Gene- based stability selection

The genes that comprised the 42 associated pathways were taken forward to the gene stability selection methodology.

The gene selection methodology was entirely based on the concept of "stability selection"[170] adjusted to be applicable on SNP/gene data. Stability selection, as mentioned in §4.1, requires a variable selection algorithm with a regularization parameter that controls the number of variables selected. In the original implementation of stability selection the Lasso variable selection algorithm was used and the authors then showed that a "randomised Lasso" algorithm gave better selection properties. I used the HyperLasso variable selection algorithm[138], which has similar properties with the randomised Lasso (Phil Brown and Jim Griffin contribution to the discussion of Meinshausen and Puhlman paper[170] and the author response). The HyperLasso has two parameters, shape λ and penalty f, the regularisation parameter. I fixed the shape parameter at $\lambda = 1$ as it was found to have negligible impact on the size of the selected models. I defined a grid of penalty values $f \in \Lambda \subset (0,45)$, in increments of 3. For each value of $f \in \Lambda$ I generated 1000 random sub-samples each consisting of 50% of the individuals, sampled without replacement, case control ration remained constant, and generated a set of selected SNPs for each sub-sample using HyperLasso.[138] The upper and lower bounds were chosen to fit models of plausible size (between 1 and ~2000 covariates).

I then calculated the selection probabilities for each gene. In order to avoid spuriously selecting large genes, I first split genes into LD based gene-blocks on the basis of recombination rates estimated using LDHat[172] on HapMap Phase II data. I assumed that the LD structure of the study samples was the same as that in HapMap. I assigned every position with recombination rate >= 15cM/Mb as the border of an LD region (corresponding to LD-regions with average size 35kb). The selection probability of a gene-block $\pi(g|f)$ was defined as the fraction of sub-samples in which at least one SNP in a gene-block g was selected. The stability path of a gene-block was defined as the selection probability over the range of all penalty values $f \in \Lambda$.

In order to extract the 'stably selected' variables, Meinshausen and Buhlman[170] suggest choosing a threshold π_{thr}, such that $0<\pi_{thr}<1$ and retaining everything with $\pi(g|f) > \pi_{thr}$, for some small $f \in \Lambda$. However, this focuses only on a single point of the stability path. I propose instead to weight the selection probabilities along all points of the path according to a prior belief $p(f)$ as to the reliability of inferences made at each $f \in \Lambda$ and then retain all genes with

weighted sum $\sum_{f' \in \lambda} p(f')\pi(g \mid f') > \pi_{thr}$ The WTCCC study had greater power as it was more than twice the size of NARAC, and so it generated substantially larger models for the same value of *f'*. In order to equalise the power between these two datasets, I derived study-specific prior weights so as to equalise the expected model size (E[size]), such that in the smaller study more weight is given to the models fitted under looser values of *f'* and vice-versa.

I assumed a gamma prior probability for the penalty with shape parameter k and scale parameter θ,

$$\Gamma(f' \mid \kappa, \theta) = f'^{\kappa-1} \frac{1}{\Gamma(\kappa)\theta^k} e^{-\frac{f'}{\vartheta}} \tag{4.1}$$

with support truncated to lie within the lower and upper bound of the grid Λ. Therefore the expected probability of inclusion of a gene *g*, was given by the integral

$$\pi(g \mid \kappa, \theta) = \int \pi(g \mid f')\Gamma(f' \mid \kappa, \theta) df' \tag{4.2}$$

which I approximated by

$$\pi(g \mid \kappa, \theta) \approx \frac{\sum_{f' \in \Lambda} \pi(g \mid f')\Gamma(\frac{1}{f'} \mid \kappa, \theta)\Delta_{f'}}{\sum_{f' \in \Lambda} \Gamma(\frac{1}{f'} \mid \kappa, \theta)\Delta_{f'}} \tag{4.3}$$

where Δ_f is the size of the intervals over which I approximated the continuous integrands by their values at *f'*. In order to define the parameters of the gamma prior, I noticed that the model size could be approximated by an exponential function of *f'* (Figure 15), and fit the model

$$\log(size(f')) = A + Bf' \tag{4.4}$$

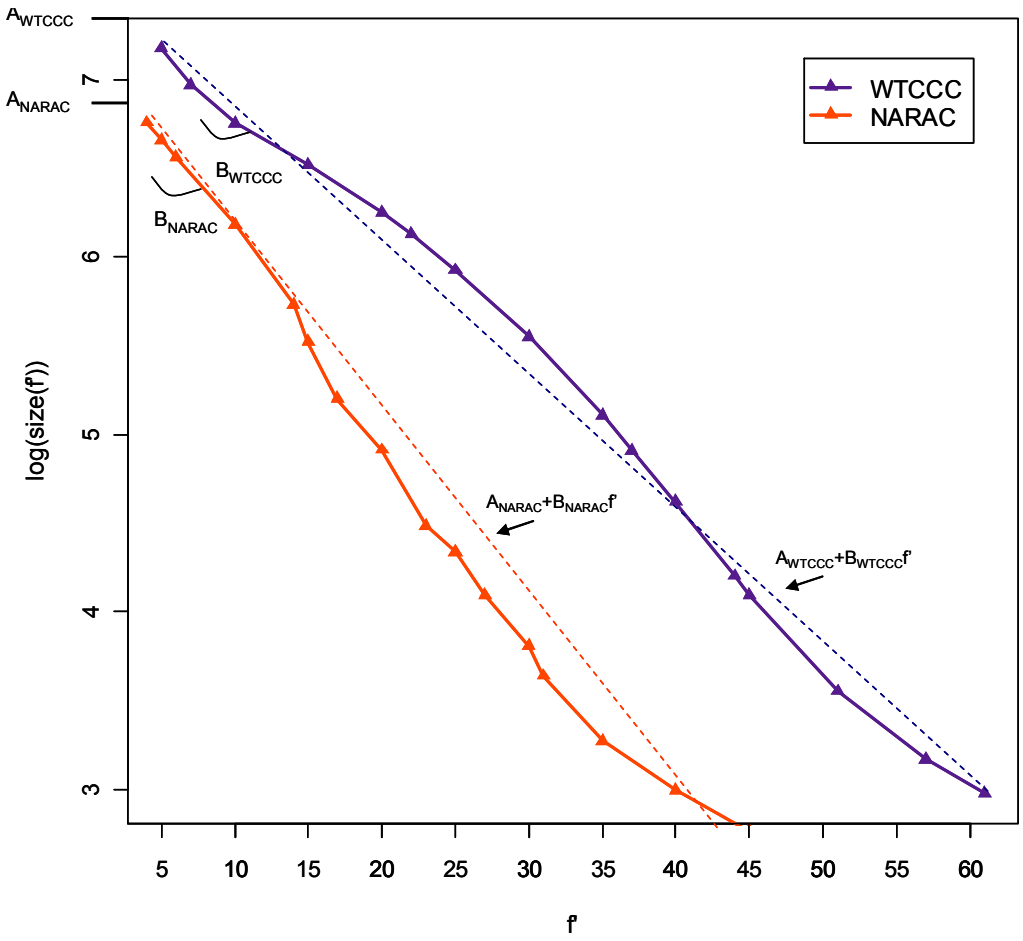


Figure 15. Log-transformed average model size as a function of the HyperLasso penalty f'.
The y-axis shows the log-transformed model sizes and the x-axis shows the penalty f'. The dashed lines represent the fitted least-squares line as described in Equation [4.4]. WTCCC models are almost twice as large as the NARAC models for the same' *f*, due to the dependency of *f* on study sample size.

The expected posterior model size was given by

$$
\begin{aligned}
E[size] &= \int_0^\infty e^{A} e^{Bf'} \Gamma(f' \mid \kappa, \theta) df' \\
&= \left(\left(\frac{1}{1 - B\theta} \right)^{\kappa} e^{A} \right) * \int_0^\infty \Gamma(f' \mid \kappa, \frac{\theta}{1 - B\theta}) df' \\
&= \left(\left(\frac{1}{1 - B\theta} \right)^{\kappa} e^{A} \right)
\end{aligned}
\qquad (4.5)
$$

,

which leads to an expression for the shape parameter $\varkappa$ in terms of the scale parameter θ and the expected model size E[size]

$$\kappa = \frac{\log(E[size]/e^{A})}{\log(1/(1-B\theta))} \tag{4.6}$$

Furthermore, I sought to equalise the spread of the posterior distribution of the model size by setting the scale parameter of the integrand in Equation 4.5 to be the same in both studies, which leads to an expression of the scale parameter of one study in terms of the other.

$$\theta_{WTCCC} = \frac{\theta_{NARAC}}{1-\theta_{NARAC}*(B_{NARAC}-B_{WTCCC})} \tag{4.7}$$

Subsequently, the stably selected gene-blocks were defined as those with $\pi(g|\kappa,\theta) >= \pi_{thr}$. I calculated results for a range of E[size]$\in$[100,1000] (increment of 100) and $\pi_{thr} \in$[0.01,0.9] (increments of 0.1).

Throughout I set the mean of the NARAC gamma prior distribution to be θ_{NARAC} / κ_{NARAC} = 10, from which for each E[size] I could calculate θ_{NARAC} and κ_{NARAC} using Equation 4.6, subsequently calculate θ_{WTCCC} from Equation 4.7 and then κ_{WTCCC} again from Equation 4.6. The corresponding prior distributions for each study and for different expected sizes E[size] are shown in Figure 16.

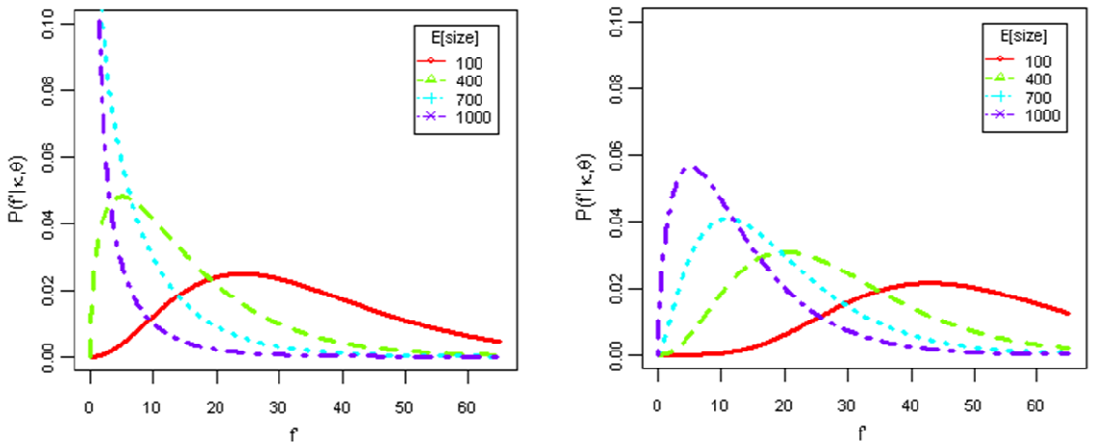


Figure 16. Gamma prior probability distributions over the penalty *f'*.
The distributions are given for different expected sizes E[size], as calculated from Equation (4.4) and Equation (4.5) for WTCCC (left panel) and NARAC (right panel). The expected model size (E[size]) controls the distribution in such a way that in the smaller study more weight is given to the models fitted under looser values of *f'* and vice-versa

The total number of selected genes per study for various values of E[size] is shown in Figure 17. I calculated the number of selected genes which overlapped between the two studies, the median FDR and its 95% confidence interval (CI) amongst these common genes. The FDR

was defined as the fraction of overlapping genes which would have been expected to be selected by chance alone and was calculated using the hypergeometric distribution. I defined the FDR in terms of genes rather than gene-blocks as I found this to be more conservative, due to the fact that the number of possible gene-blocks is much larger than the number of possible genes, whereas the two studies almost always selected the same gene-block within a common gene for π_{thr} >0.2.

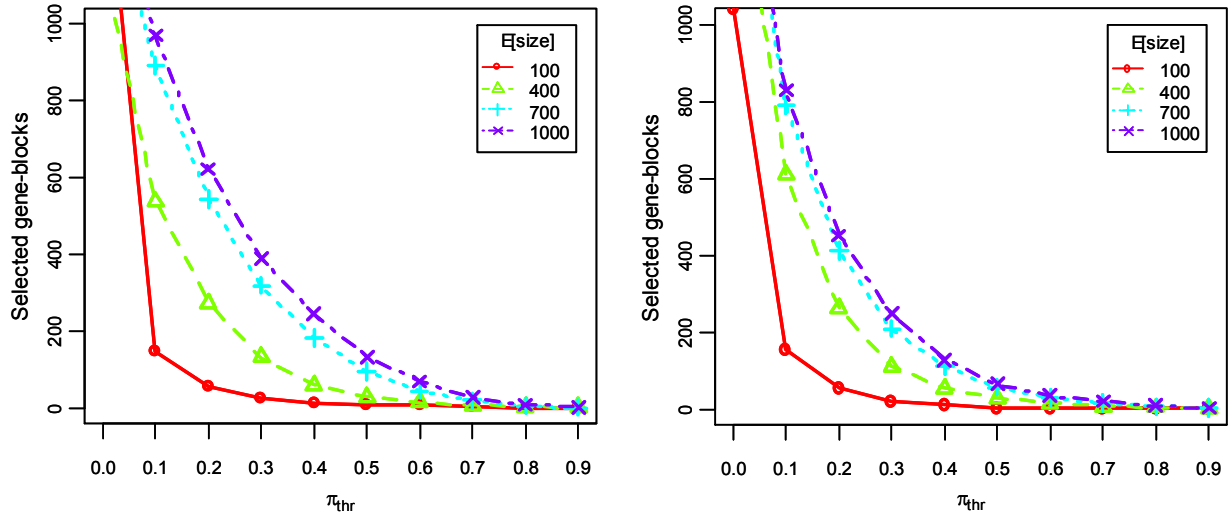


Figure 17. Number of selected gene-blocks as a function of the selection probability cut-off π_{thr}. that defines stable selection.

Results are depicted for WTCCC (left panel) and for NARAC (right panel) and they are color-coded to reflect different expected sizes.

In this study, I opted for π_{thr} =0.4 and E[size]=400, as thus I achieved a good trade-off between maximising the number of observed common gene-blocks between studies, while controlling the FDR (Figure 18). This resulted in selecting 58 genes in WTCCC and 61 genes in NARAC (that is 60 gene-blocks in WTCCC and 64 gene-blocks in NARAC), of which 21 were common (hypergeometric test $P = 10^{-21}$) and the median FDR is 5% with 95% CI of [0%-19%] as shown in Figure 19. These findings are fully discussed in Chapter 5, but they are briefly mentioned here to justify the steps and the decision-making in the analytical process.

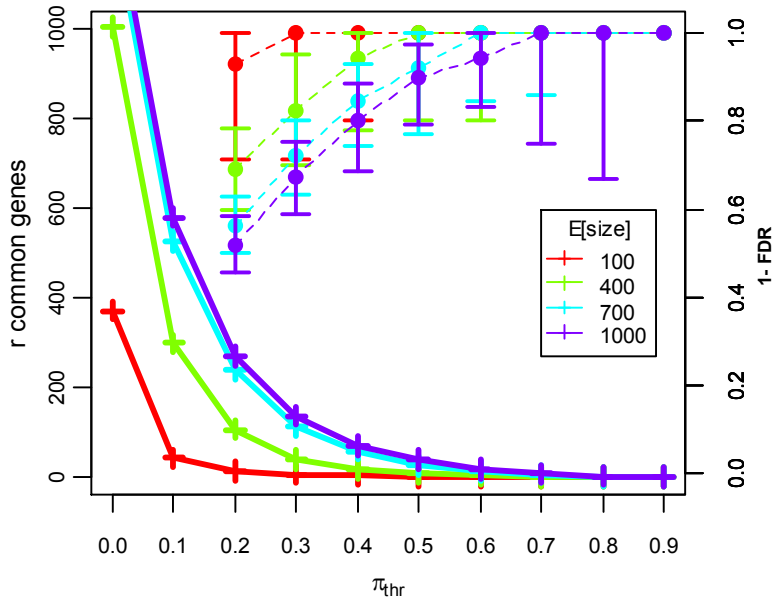


Figure 18. The trade-off between *r* common genes and FDR for different π_{thr} and E[size].

The solid lines show on the left y-axis the number of common genes selected in WTCCC and NARAC for different values of π_{thr} (x-axis) and expected model sizes E[size] (depicted in different colors). Dashed lines show (1- median FDR) on the right y-axis and the 95% CI of the median is shown as error bars.

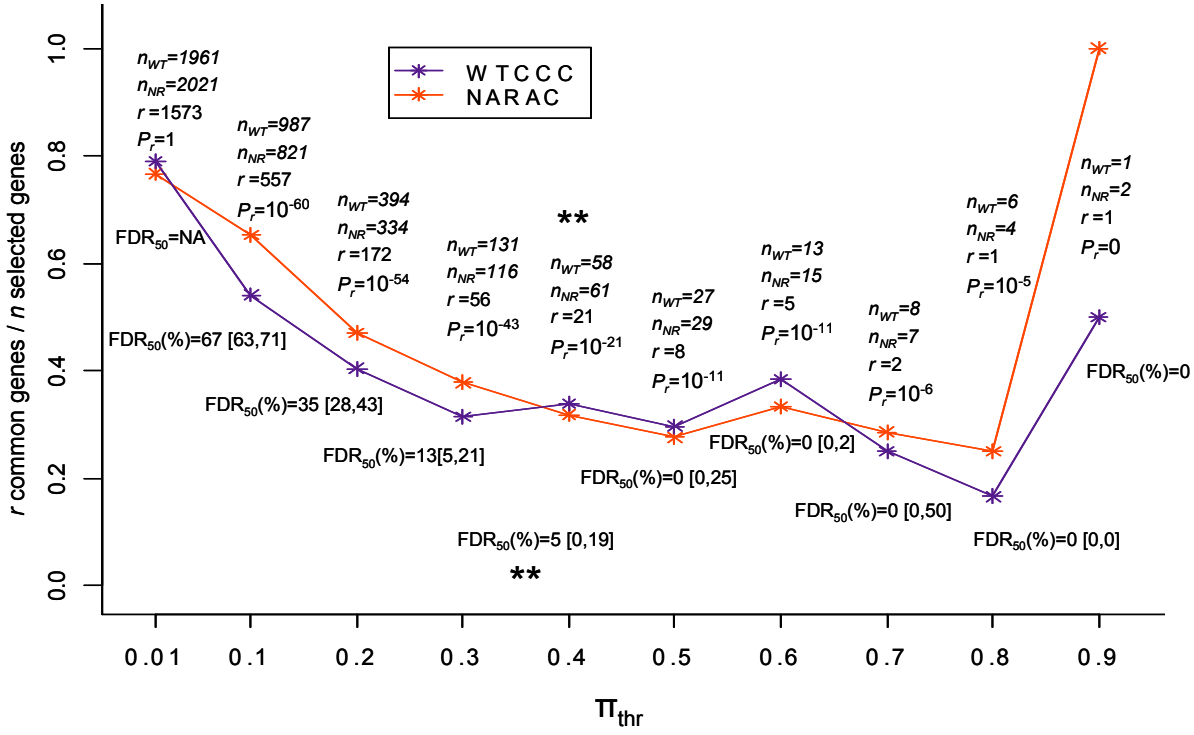


Figure 19. Fraction of selected genes in WTCCC and NARAC which are common for E[size]=400 at different values of π_{thr}.

The number of selected genes per study (n_{WTCCC}, n_{NARAC}) is shown above the curves followed by the number of genes in common (*r*) and the hypergeometric probability of the observed overlap by chance. The median FDR and the 95% CI around it are given below the curve. The asterisk shows that for π_{thr} =0.4, I achieve the maximum number of genes that overlap with a median FDR=5% [0,19%].

4.4 Estimation of the significance of the observed overlap between studies

The significance of the observed overlap in the selected genes between studies, as shown in Figure 19 was calculated using the hypergeometric distribution.

In order to investigate the possible bias in favour of selecting larger genes, I initially calculated the correlation between the number of gene-blocks and the physical length of a gene. I found that the physical length of a gene (in Mbases) correlated with the number of gene-blocks (r^2=0.59). Hence, larger genes are more likely to be selected just because they contain more gene-blocks. This is the reason I decided to calculate all of the overlap statistics at the gene-block level, rather than the gene level, and I only consider replication at the gene-block level. Thus, while there is a bias to select large genes (which is also the case in GWAS analysis, in the sense that large genes are more likely by chance to be reported as associated by virtue of containing more SNPs), I do not think that this is adversely impacting the reported overlap statistics.

On a similar note, a potential confounding effect might lie in the fact that larger gene-blocks are more likely to be selected by chance because they contain more SNPs. I therefore compared the distribution of number of SNPs per gene-block (#SNPs) between selected and non-selected gene-blocks in both NARAC and WTCCC on the same chip of the study and cross-chip (Figure 20). As it is demonstrated from all these plots, there is an over-representation of selected gene-blocks in the region ~20<#SNPS<~70, and otherwise an under-representation of selected gene-blocks.

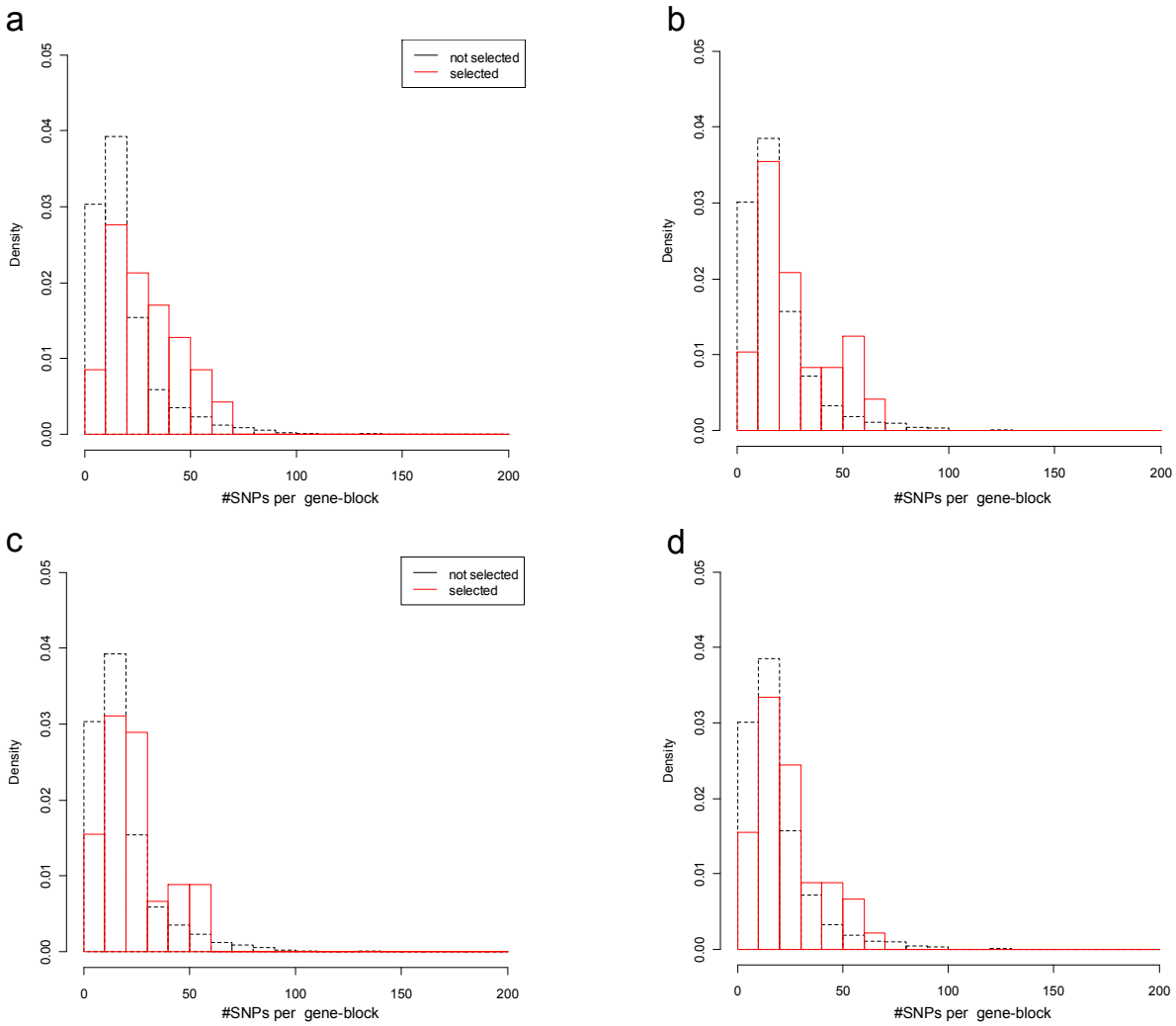


Figure 20. Comparison of typed SNP content per gene block between selected and unselected gene blocks. The x-axis shows the number of typed SNPs per gene-block in bins of size 10 and the y-axis the fraction of selected/non-selected gene-blocks in each bin. . Gene blocks selected by stability selection are depicted with a solid red line and gene blocks not selected (but still contained within one of the 42 pathways taken forward to stability selection) are depicted with a dashed black line. a,b) NARAC based stability selection; c,d) WTCCC based stability selection; a,c) # typed SNPs based on Illumina 550k chip; b,d) # typed SNPS based on Affymetrix 500k chip.

Given the observed bias in the distribution of the number of typed SNPs (#SNPs) per selected gene block between selected and non-selected gene blocks, and in order to make sure that it is not inflating the overlap statistics, I simulated the null distribution of number of gene-blocks in the overlap based on this empirical distribution of #SNPs per gene block. I recalculated the significance of the observed overlap assuming non-random gene block selection according to #SNPs. I first calculated the empirical probability of selecting a gene-block with n SNPs as $\pi(n \mid \text{selected}) \sim$ #(gene selected gene blocks with n SNPS). I then smoothed this probability distribution by setting $\pi(n \mid \text{selected})$ to be equal to its average in a moving 5-SNP window. I then calculated the probability of selecting each given gene-block g, with #SNPs =

#SNPs(g) as P(g | selected) = π (#SNPs(g) | selected) * 1/#(geneblocks with #SNPs=#SNPs(g))

I could then sample two vectors of selected gene-blocks, of length N_1 and N_2 corresponding to the number of NARAC and WTCCC selected gene-blocks, and finally calculate the number of common gene-blocks selected. While I could not simulate enough data to get the same level of significance, I did retain significance of the overlap of selected gene-blocks of with $P<10^{-4}$.

4.5 Gene Ontology functional enrichment analysis

As a final step, I conducted a gene ontology (GO) enrichment analysis using GOrilla[76] on the genes that the gene selection methodology identified. The selected genes were compared to all the typed genes on the chip of each study. I carried out this analysis on each study separately and I observed that the same functional categories were enriched. I repeated the analysis on the union of the identified genes to derive a stronger signal of enrichment. I also carried out a sensitivity analysis on the effect of different π_{thr} on the enriched GO categories and I demonstrate in Chapter 5 that the enriched functional categories are robust to the selection of π_{thr}.

In terms of performance, the gene stability selection step is computationally very intensive in both computational time and memory resources. A single model fit varies between an hour (for the stringiest penalty) to a day (for the loosest penalty) and one should consider that I am fitting 1000 models per penalty value, which is ~16,000 in total.

4.6 Summary

In this chapter I described in detail the technical steps of my pathway-driven gene stability selection. It is a procedure that goes beyond simple pathway-based analysis of GWAS to identify the key responsible genes within associated pathways that are driving the pathway signal. I provided the generic methodological framework and I illustrated the steps and the decisions along the process through the application of my methodology on two independent GWAS of RA, the WTCCC and the NARAC study. In the next chapter, I present the results in detail and I discuss future directions of the findings.

Chapter 5

Application on two GWAS of rheumatoid arthritis

In Chapter 4, I introduced my methodology on pathway-driven gene stability selection. In this chapter, I go on to present the results of the application of this approach on two RA case control GWAS; the WTCCC RA study, of 1,860 cases and 2,938 controls typed on the Affymetrix 550K chip[9] and the NARAC study of 908 cases and 1260 controls genotyped on the Illumina Humanhap550[8]. The generic flowchart of my approach was shown in Chapter 4, Figure 12; the implementation of it is given in Figure 21. In all analyses the X chromosome was omitted, as I did not have access to the X chromosome data of the NARAC study.

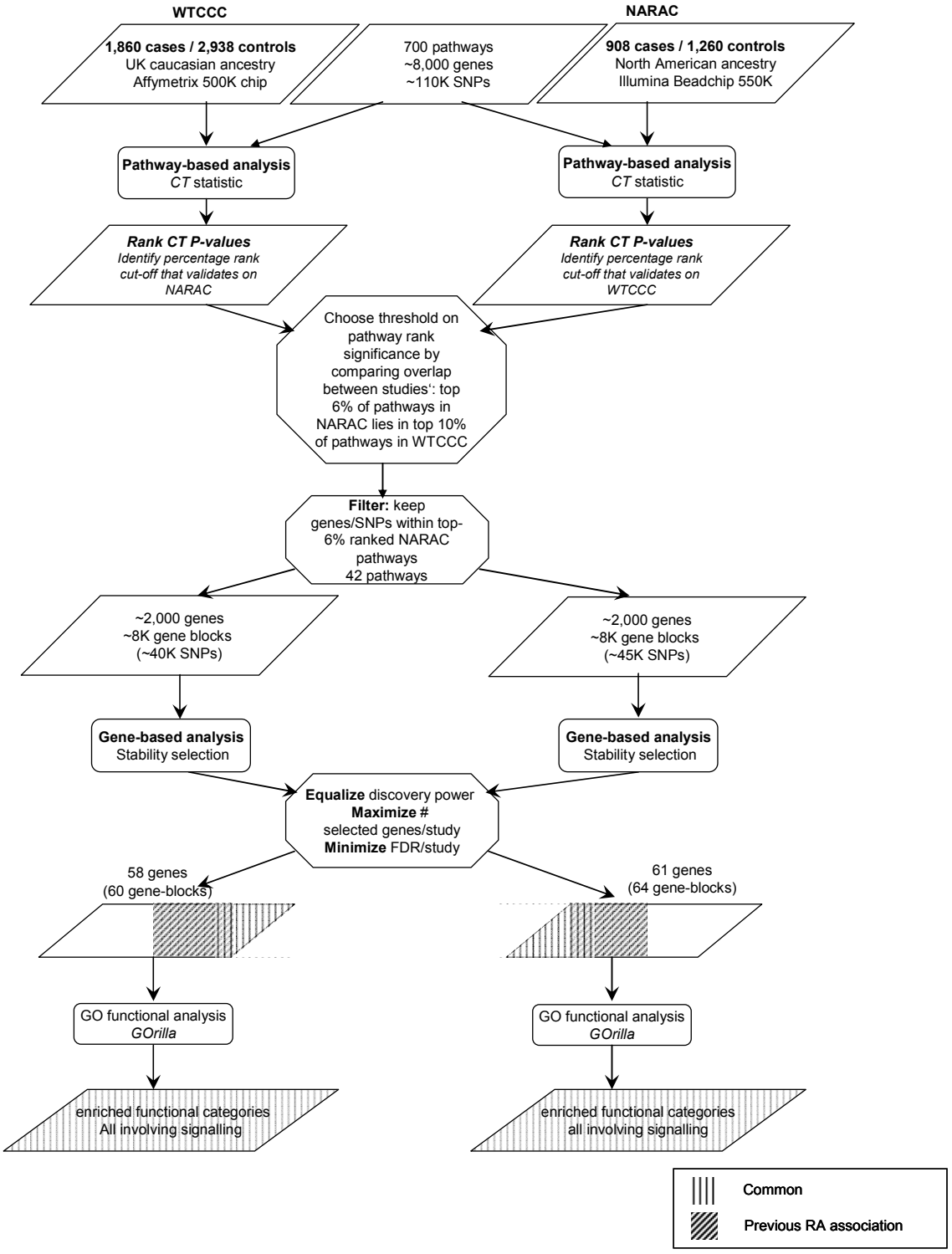


Figure 21. Methodological flowchart of my approach as implemented on two GWAS of RA.

5.1 Pathway associations

The 42 associated pathways with the *CT P*-values and their ranks in both studies are listed in Table 12. Their database source is listed in Table 13.

Table 12.Top 42 pathways identified through pathway-based analysis in WTCCC (W) and NARAC (N) studies.

The cumulative trend (CT) statistic *P*-values are adjusted at α=0.05 with the Benjamini-Hochberg method that controls the FDR. Number of SNPs per pathway might differ between studies, because they refer to the typed markers. The pathways appear in the order of ascending NARAC ranks.

Pathway definition	*P*-value		Rank		#Genes	#Typed SNPs	
	N	W	N	W		Illumina 550k	Affy 500k
Bystander B Cell Activation	1.00E-371	1.00E-98	1	11	4	66	53
Type I diabetes mellitus	1.00E-349	1.00E-133	2	2	41	846	745
IL-5 Signalling Pathway	1.00E-344	1.00E-118	3	7	10	164	134
Antigen Dependent B Cell Activation	1.00E-334	1.00E-111	4	8	8	114	92
Activation of Csk by cAMP-dependent Protein Kinase inhibited signaling through the T cell receptor	1.00E-332	1.00E-123	5	5	20	246	250
T cell receptor signaling	1.00E-331	1.00E-211	6	1	264	5215	4613
Cell adhesion molecules	1.00E-289	1.00E-68	7	15	127	4159	3811
The Role of Eosinophils in the Chemokine Network of Allergy	1.00E-288	1.00E-71	8	14	8	66	76
Cytokines and Inflammatory Response	1.00E-283	1.00E-108	9	9	28	285	258
B Lymphocyte Cell Surface Molecules	1.00E-276	1.00E-104	10	10	10	153	125
The Co-Stimulatory Signal During T-cell Activation	1.00E-275	1.00E-130	11	3	17	227	186
Th1/Th2 Differentiation	1.00E-270	1.00E-91	12	12	17	244	256
Lck and Fyn tyrosine kinases in initiation of TCR Activation	1.00E-254	1.00E-122	13	6	10	153	134
Antigen processing and presentation	1.00E-243	1.00E-85	14	13	71	615	549
Hematopoietic cell lineage	1.00E-238	1.00E-128	15	4	81	1132	1022
Notch signaling pathway	1.00E-74	7.90E-06	16	62	43	826	606
Proteasome Complex	1.00E-43	1.91E-05	17	64	17	142	119
Complement Pathway	1.00E-41	5.31E-06	18	59	16	269	242
Classical Complement Pathway	1.00E-37	4.56E-06	19	56	13	199	190
Natural killer cell mediated cytotoxicity	1.00E-35	9.28E-05	20	70	125	2122	1775
Focal adhesion	1.00E-31	1.00E-48	21	16	195	5026	4586
Cytokine-cytokine receptor interaction	1.00E-29	1.00E-42	22	17	249	2956	2854
Calcium signaling pathway	1.00E-28	5.51E-12	23	37	168	5604	5109
MAPK signaling	1.00E-27	1.00E-38	24	18	252	5335	4848
Glycerophospholipid_metabolism	1.00E-25	4.63E-06	25	57	49	988	838
Glycerolipid_metabolism	1.00E-24	7.03E-06	26	61	45	943	794
Neuroactive ligand receptor interaction	1.00E-22	1.00E-37	27	19	234	5642	5323
Toll like receptor 9 signaling via IRF5	1.00E-21	1.60E-06	28	53	17	169	184
Smooth muscle contraction	1.00E-18	2.56E-06	29	54	138	4356	3871
Phosphatidylinositol signalling system	1.00E-17	5.41E-10	30	43	85	2887	2569
Tight junction	1.00E-17	5.92E-12	31	38	133	4125	3673
Purine metabolism	1.00E-16	2.34E-15	32	29	143	3532	3165
Gnrh signaling pathway	1.00E-15	5.34E-07	33	49	96	2714	2381
Wnt signaling pathway	1.00E-14	1.00E-34	34	21	144	2830	2636
Gap junction	1.56E-14	1.00E-22	35	25	92	2823	2582
Valine leucine and isoleucine biosynthesis	1.77E-14	8.78E-05	36	69	12	171	183
Cell communication	2.81E-14	8.83E-10	37	44	132	2304	2162
Regulation of actin cytoskeleton	1.55E-13	1.00E-29	38	23	204	4206	3932
Ecm receptor interaction	2.38E-13	1.36E-12	39	36	87	2541	2386
Jak stat signaling pathway	1.30E-12	3.89E-15	40	30	149	1735	1699
Taste transduction	1.70E-10	6.87E-07	41	50	49	1033	959
Glycan structures biosynthesis 1	4.28E-08	4.81E-14	42	32	113	2985	2631

Table 13. Pathway collection that the 42 associated pathways were taken from
The pathways appear in the same order as in Table 1 of the main manuscript.

Pathway description	Collection/Source
Bystander B Cell Activation	BioCarta
Type I diabetes mellitus	KEGG
IL-5 Signaling Pathway	BioCarta
Antigen Dependent B Cell Activation	BioCarta
Activation of Csk by cAMP-dependent Protein Kinase inhibited signaling through the T-cell receptor	BioCarta
T cell receptor signaling	KEGG modified in house
Cell adhesion molecules	KEGG
The Role of Eosinophils in the Chemokine Network of Allergy	BioCarta
Cytokines and Inflammatory Response	BioCarta
B Lymphocyte Cell Surface Molecules	BioCarta
The Co-Stimulatory Signal During T-cell Activation	BioCarta
Th1/Th2 Differentiation	BioCarta
Lck and Fyn tyrosine kinases in initiation of TCR Activation	BioCarta
Antigen processing and presentation	KEGG
Hematopoietic cell lineage	KEGG
Notch signaling pathway	KEGG
Proteasome Complex	BioCarta
Complement Pathway	BioCarta
Classical Complement Pathway	BioCarta
Natural killer cell mediated cytotoxicity	KEGG
Focal adhesion	KEGG
Cytokine cytokine receptor interaction	KEGG
Calcium signaling pathway	KEGG
MAPK signaling	KEGG
Glycerophospholipid_metabolism	BioCarta
Glycerolipid_metabolism	BioCarta
Neuroactive ligand receptor interaction	KEGG
Toll like receptor 9 signaling via IRF5	KEGG modified in house
Smooth muscle contraction	BioCarta
Phosphatidylinositol signaling system	BioCarta
Tight junction	KEGG
Purine metabolism	KEGG
Gnrh signaling pathway	KEGG
Wnt signaling pathway	KEGG
Gap junction	KEGG
Valine leucine and isoleucine biosynthesis	KEGG
Cell communication	KEGG
Regulation of actin cytoskeleton	KEGG
Ecm receptor interaction	KEGG
Jak stat signaling pathway	KEGG
Taste transduction	KEGG
Glycan structures biosynthesis 1	KEGG

5.2 Gene selection on associated pathways

A gene composition similarity matrix, as shown in Figure 22 revealed that many of the significant pathways share the same genes.

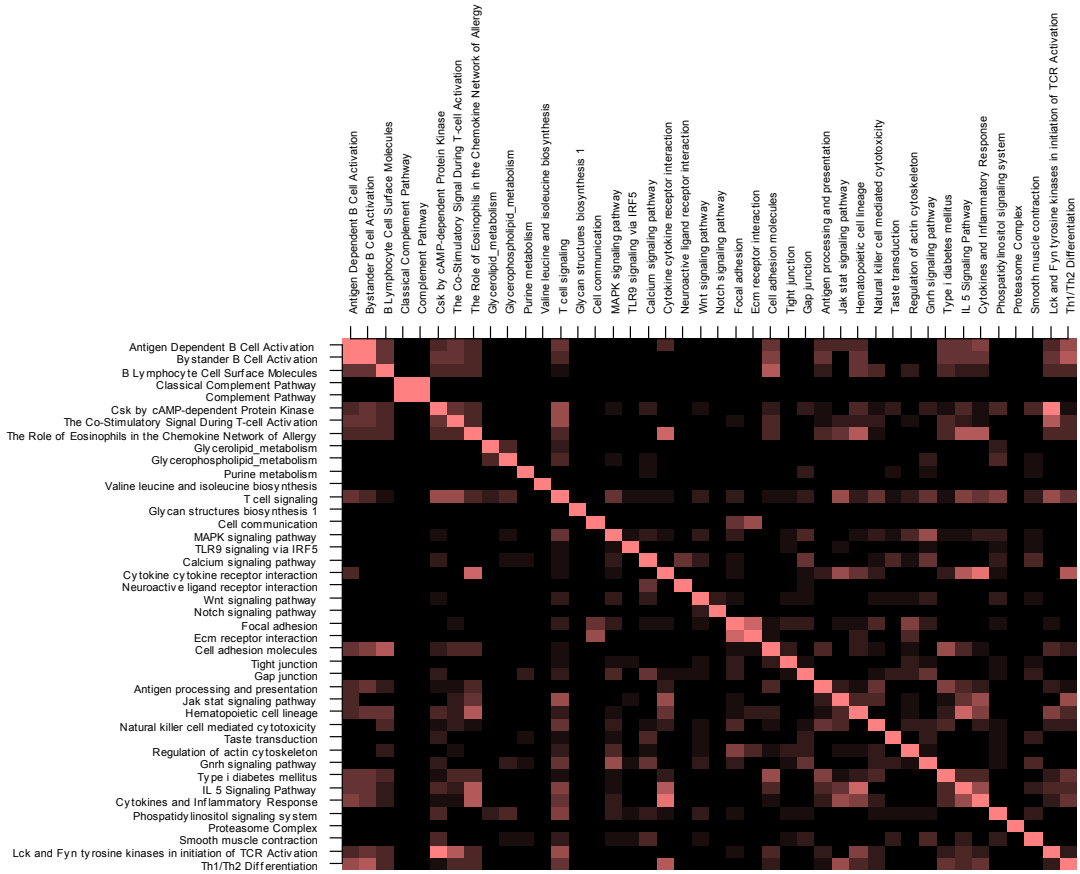


Figure 22. Similarity matrix plot in terms of gene composition for the 42 identified pathways.
Spots are color coded to represent the number of shared genes between two pathways as a fraction of the smallest pathway size under comparison. Light red stands for 100% of the genes in one pathway being involved in the other pathways and black stands for zero genes in common. Pathways are ordered by the rank of the CT *P* value in WTCCC and subsequently in NARAC. As an example to read the figure, the *ECM-receptor interaction* pathway ($P < 10^{-11}$) shares 50% of its 87 genes with *Cell Communication* ($P < 10^{-13}$) and 96% of the 45 genes involved in *Glycerolipid Metabolism* ($P_{WTCCC} = 4 \times 10^{-6}$, $P_{NARAC} < 10^{-17}$) are also present in *Glycerophospholipid Metabolism* (49 genes). *Calcium signaling* ($P_{WTCCC} = 5 \times 10^{-12}$, $P_{NARAC} < 10^{-17}$) shares 20% of its 166 genes with *GnRH signaling* of 95 genes ($P_{WTCCC} = 5 \times 10^{-7}$, $P_{NARAC} = 4 \times 10^{-13}$), which in turn is 40% similar to *Focal adhesion* with 195 genes ($P < 10^{-17}$). The 8 genes of the *Bystander B cell activation* ($P < 10^{-17}$) are entirely a subset the 12 genes of *Antigen Dependent B Cell Activation* ($P < 10^{-17}$, top left corner).

I therefore speculated that there is an underlying group of genes acting in the associated pathways, which may be driving this correlation in pathway significance. I selected the 2156 genes comprising the 42 pathways significant in both studies to carry forward to my gene stability selection procedure, as it was described in detail in Chapter 4, §4.3. To recap, this methodology retains genes that exhibit "stable selection". Stable selection is inferred by stability paths, which in essence are plots of the selection probability of a gene in fitted models on the data under random sub-sampling of the individuals and varying regularization parameters of the variable selection algorithm. The models vary in size from the smallest possible (a single covariate in the model) to the largest (several genes in the model), thus genes with a constantly high or increasing stability path are likely to be denoted as "stably selected genes".

I broadly observed that stability paths with consistently high selection probability $\pi(g|f)$ across $f' \in \Lambda$ corresponded to the major hits found in single SNP analyses in each study; for instance *TRAF1/C5*, *HLA-DRB1* and *MICA* in NARAC; *IL2RA*, *PTPN22*, *HLA-DRA* and *MICA* in WTCCC. At the other extreme, I expected that stability paths with low $\pi(g|f)$ across $f' \in \Lambda$ comprise non-associated gene-blocks. In between these two extremes laid many stability paths that have low selection probability under strong regularisation and high selection probability under weak regularization.

My pathway-based gene stability selection methodology identified as "stably selected" genes, 58 genes in WTCCC and 61 genes in NARAC (that is 60 gene-blocks in WTCCC and 64 gene-blocks in NARAC), of which 21 are common ($P = 10^{-21}$) and the median FDR is 5% with 95% CI of [0%-19%] as shown in Figure 19 (Chapter 4, §4.3). The selected gene-blocks in WTCCC and NARAC are shown in Table 14 and Table 15 respectively. For the majority of genes, only one gene block was selected, but for *ADCY2* and *FHIT* in WTCCC and *FHIT* and *PTPRN2* in NARAC, two/three LD blocks exhibited stable selection. I considered a gene found in one study to replicate in the other study, only if the same gene-block was selected in both. Known associations comprised 10 of 58 in WTCCC; 13 of 61 in NARAC and 5 of 21 common genes. Of the 58 and 61 genes selected independently in NARAC and WTCCC from an original pool of 2,156 genes, more than 50% were either common between studies or previous RA associations and hence can be considered as validated. The stability paths of the selected gene-blocks, in each study, are shown in Figure 23c/d; the subset of stability paths for previously known associations is shown in Figure 24a,b and the stability paths of gene blocks that were not selected are shown in Figure 24c,d.

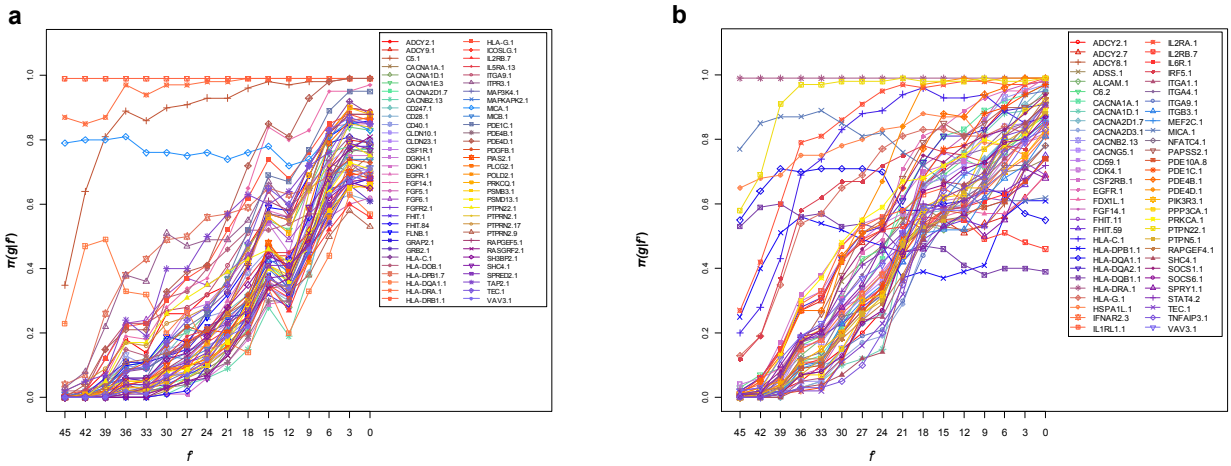


Figure 23. Stability paths of the 60 and 64 gene-blocks selected in WTCCC (a) and NARAC (b).
Gene names are shown in the legend followed by a number that indicates the selected gene-block.

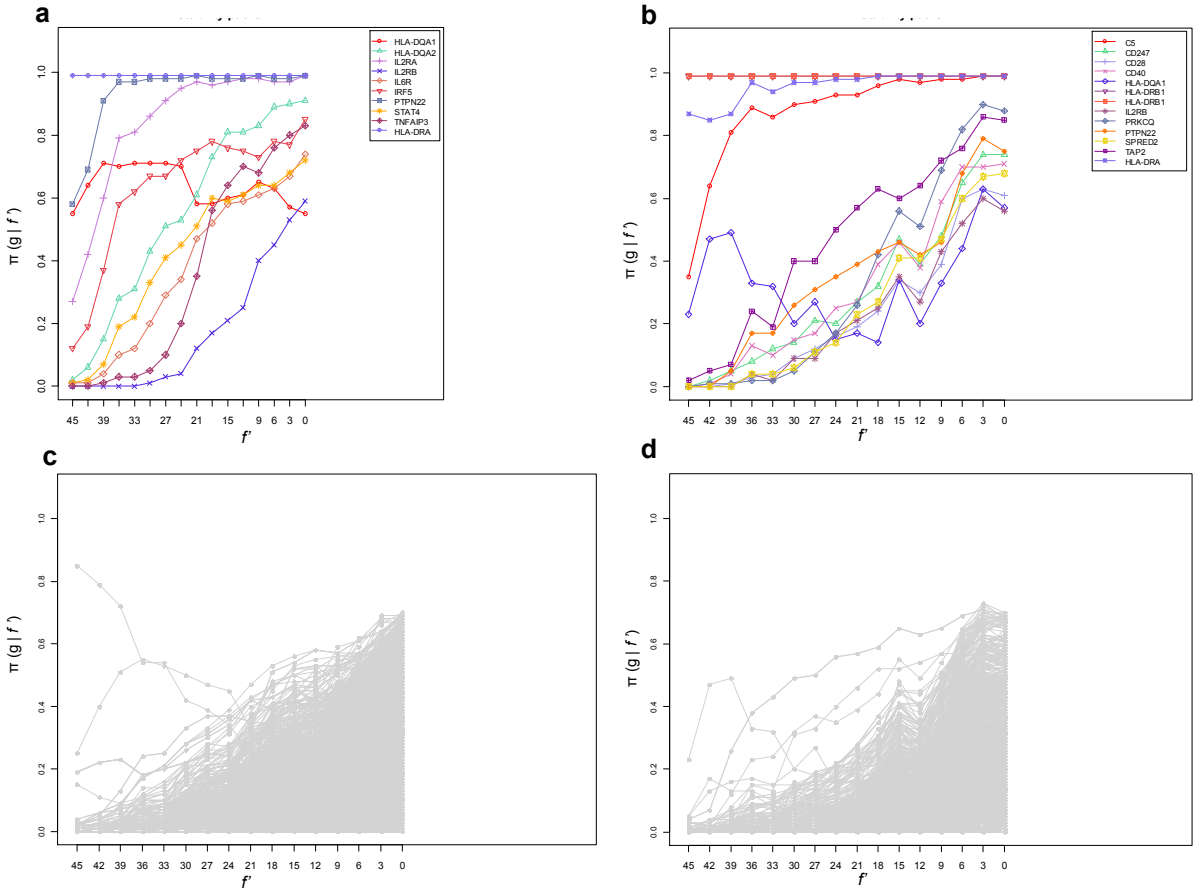


Figure 24. Stability paths of known RA associations (a,b) and not selected genes (c,d).

a, b) Stability paths of the previously reported associations with RA which were validated by the gene stability selection in a) WTCCC (10 genes) and b) NARAC (13 genes). c, d) Stability paths of all the genes within the selected pathways that were rejected as noise by my stability selection methodology in c) WTCCC and d) NARAC. The gene in c) plot that follows a decreasing trend is *HLA-DRB1*, the most well known association with RA. The paradoxical curve is explained by different *HLA* genes of the MHC being preferentially included as the models became larger.

Table 14. The 60 gene-blocks selected in WTCCC.

For the majority of genes, only one LD block was selected, but for ADCY2 and FHIT two LD blocks exhibited stable selection.

Gene symbol	LD block index	Chrom. pos. start	Chrom. pos. end
ADCY2	1	7428754	7593069
ADCY2	7	7594566	7712781
ADCY8	1	131840404	131847376
ADSS	1	242618418	242702036
ALCAM	1	106548402	106798434
C6	2	41160720	41312856
CACNA1A	1	13155623	13331522
CACNA1D	1	53474897	53575012
CACNA2D1	7	81573468	81851660
CACNA2D3	1	54105141	54141360
CACNB2	13	18578665	18827827
CACNG5	1	62283591	62287922
CD59	1	33658907	33671217
CDK4	1	56408269	56452431
CSF2RB	1	35615840	35653760
EGFR	1	55033429	55128489
FGF14	1	101150762	101335129
FHIT	11	59733714	59749954
FHIT	59	60247400	60278544
HLA-C	1	31324507	31367834
HLA-DPB1	1	33128365	33132110
HLA-DQA1	1	32693160	32739407
HLA-DQA2	1	32796968	32822121
HLA-DQB1	1	32715634	32762444
HLA-DRA	1	32495624	32540802
HLA-G	1	29882734	29926878
HSPA1A	1	31871269	31913698
ICAM1	1	10222516	1.0278291
IFNAR2	3	33510364	33570678
IL1RL1	1	102274193	102334507
IL2RA	1	6073511	6164278
IL2RB	7	35874863	35888302
IL6R	1	152624292	152726812
IRF5	1	128338976	128347997
ITGA1	1	52099877	52169023
ITGA4	1	182007979	182060813
ITGA9	1	37448558	37577446
ITGB3	1	42666206	42765076
MEF2C	1	88031921	88234780
MICA	1	31459349	31511069
NFATC4	1	23887093	23938650
PAPSS2	1	89389455	89517442
PDE10A	8	165719604	165756281
PDE1C	1	31775674	31790442
PDE4B	1	66010515	66545937
PDE4D	1	58280444	58658855
PIK3R3	1	46258398	46391295

Table 13(continued)

Gene symbol	LD block index	Chrom. pos. start	Chrom. pos. end
PPP3CA	1	102143609	102507376
PRKCA	1	61708068	61762419
PTPN22	1	114137960	114235857
PTPN5	1	18685226	18703485
RAPGEF4	1	173287843	173492013
SHC4	1	46882243	46992983
SOCS1	1	11235774	11277540
SOCS6	1	66087116	66168414
SPRY1	1	124517405	124564359
STAT4	2	191582912	191681279
TEC	1	47812556	47986571
TNFAIP3	1	138210273	138266142
VAV3	1	107894802	108046933

Table 15. The 64 gene-blocks selected in NARAC.

For the majority of genes, only one LD block was selected, but for FHIT and PTPRN2 two and three LD blocks respectively exhibited stable selection.

Gene symbol	LD block index	Chrom. pos start	Chrom. pos end
ADCY2	1	7428754	7593069
ADCY9	1	3932523	3938631
C5	1	122734434	122872375
CACNA1A	1	13155623	13331522
CACNA1D	1	53474897	53575012
CACNA1E	3	179711393	179792378
CACNA2D1	7	81573468	81851660
CACNB2	13	18578665	18827827
CD247	1	165643582	165704577
CD28	1	204259442	204330802
CD40	1	44158244	44181354
CLDN10	1	94863031	94884949
CLDN23	1	8577075	8619027
CSF1R	1	149393048	149467669
DGKH	1	41500888	41721888
DGKI	1	136704876	136736520
EGFR	1	55033429	55128489
FGF14	1	101150762	101335129
FGF5	1	81383976	81426987
FGF6	1	4393568	4445041
FGFR2	1	123207590	123218650
FHIT	1	59689950	59732130
FHIT	84	60551758	60707039
FLNB	1	57949166	58153017
GRAP2	1	38607031	38717330
GRB2	1	70805355	70900365
HLA-C	1	31324507	31367834
HLA-DOB	1	32868423	32911977
HLA-DPB1	7	33160006	33163494
HLA-DQA1	1	32693160	32739407
HLA-DRA	1	32495624	32540802
HLA-DRB1	1	32634524	32685540
HLA-G	1	29882734	29926878
ICOSLG	1	44451149	44505262
IL2RB	7	35874863	35888302
IL5RA	13	3076524	3101727
ITGA9	1	37448558	37577446
ITPR3	1	33677138	33792326
MAP3K4	1	161311536	161470346
MAPKAPK2	1	204904911	204994249
MICA	1	31459349	31511069
MICB	1	31553833	31606880
PDE1C	1	31775674	31790442
PDE4B	1	66010515	66545937
PDE4D	1	58280444	58658855
PDGFB	1	37929247	37984993
PIAS2	1	42626057	42771464

Table 14 (continued)

Gene symbol	LD block index	Chrom. pos start	Chrom. pos end
PLCG2	1	80347176	80523696
POLD2	1	44099990	44101373
PRKCQ	1	6488953	6583069
PSMB3	1	34142527	34194004
PSMD13	1	206056	237102
PTPN22	1	114137960	114235857
PTPRN2	1	157003814	157259178
PTPRN2	17	157740875	157857270
PTPRN2	9	157520348	157537330
RAPGEF5	1	22103681	22118493
RASGRF2	1	80270614	80400810
SH3BP2	1	2744547	2832621
SHC4	1	46882243	46992983
SPRED2	1	65371488	65533160
TAP2	1	32877514	32911977
TEC	1	47812556	47986571
VAV3	1	107894802	108046933

I then investigated whether common genes were driving the association signals of multiple pathways by mapping the selected genes back to their original pathways (Table 16). The striking majority of selected genes, to name a few: the *MHC* genes, *ADCY2*, *ADCY8*, *CACNA1A*, *DGKH*, *FLNB*, *ITGA1*, *PRKCA*, *PRKCQ*, *SHC4*, *VAV3* were indeed involved in more than one of the 42 pathways. On the other hand, some selected genes such as *IRF5, MAPKAPK2, PDE10A, PDE1C, PSMD13* and *TEC* were only involved in one pathway (Table 16). I observed that the very top-ranked pathways were mostly driven by the MHC genes and 6 pathways were driven entirely by known associations, the *HLA-DRA*, *HLA-DRB1* and *CD28* genes in particular. The majority of pathway signals (18 pathways) were driven by a combination of previously reported associations as well as novel associations (Table 16). A further 14 pathways were driven only by the newly reported associations and surprisingly 4 pathways had no gene selected (Table 16).

Finally I calculated the inverse variance weighted meta-analysis *P*-value for the SNP selected most frequently in each study in the 21 commonly selected genes using PLINK[173].The meta-analysis *P* –value was less than 0.05 for all SNPs (Table 17).

Table 16. Mapping of the union of WTCCC and NARAC selected genes back to the 42 associated pathways.

Pathways are in the same order as they appear in Table 12. Column 2 shows the number of selected genes (SG) over the number of pathway genes (PG).Previously reported RA associations are underlined. Genes that appear in only 1 of the 42 pathways are followed by an asterisk (*). Genes without an * act in more than 1 of the 42 pathways. Pairs of genes in italics indicate genes that overlap the same selected gene-block, but are in different pathways. The first listed gene in each pair is from the pathway indicated, while the second listed has a higher selection probability and hence is reported in my final list.

Pathway	#SG\#PG	Selected genes
Bystander B Cell Activation	3\8	CD28, HLA-DRA, HLA-DRB1
Type i diabetes mellitus	12\45	CD28, HLA-C, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-G, PTPRN2
IL 5 Signaling Pathway	3\10	HLA-DRA, HLA-DRB1, IL5RA
Antigen Dependent B Cell Activation	3\12	CD28, HLA-DRA, HLA-DRB1
Activ. Csk by cAMP-dependent Protein Kinase inhibited signaling via the T cell receptor	2\24	HLA-DRA, HLA-DRB1
T cell signaling	29\273	CD28, CD247, VAV3, GRAP2*, TEC*, ITPR3, DGKH, DGKI*, PPP3CA, GRB2, RASGRF2, PRKCA, PTPN5, NFATC4, PIK3R3, TNFAIP3, PRKCQ, CDK4, IL6R, IL2RA, IL2RB, IL5RA, CSF2RB, PIAS2, SOCS1, SPRED2, SPRY1, PTPN22*, *RAP1A (VAV3)*
Cell adhesion molecules	19\134	ALCAM*, CD28, CD40, CLDN10, CLDN23, HLA-C, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-G, ICAM3*, ICAM4, ICOSLG*, ITGA4, ITGA9
Eosinophils in the Chemokine Network of Allergy	2\8	HLA-DRA, HLA-DRB1
Cytokines and Inflammatory Response	2\29	HLA-DRA, HLA-DRB1
B Lymphocyte Cell Surface Molecules	311	ICAM1, HLA-DRA, HLA-DRB1
The Co-Stimulatory Signal During T-cell Activation	4\21	CD28, GRB2, HLA-DRA, HLA-DRB1
Th1/Th2 Differentiation	5\19	CD28, HLA-DRA, HLA-DRB1, IL2RA, *IL18R1(IL1RL1)*
Lck and Fyn tyrosine kinases in initiation of TCR Activation	2\13	HLA-DRA, HLA-DRB1
Antigen processing and presentation	15\83	HLA-C, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DRA, HLA-DRB1, HLA-G, TAP1*, TAP2*, HSPA1A* (HSPA1B,HSPA1L), PSMB3, PSMD13*
Hematopoietic cell lineage	10\88	CD59*, CSF1R, HLA-DRA, HLA-DRB1, IL2RA, IL5RA, IL6R, ITGA1, IGA4, ITGB3
Notch signaling pathway	0\47	`-
Proteasome Complex	1\17	PSMB3
Complement Pathway	2\19	C5, C7*
Classical Complement Pathway	2\14	C5, C6*
Natural killer cell mediated cytotoxicity	15\132	CD247, GRB2, HLA-C, HLA-G, ICAM1, IFNAR2, MICA, MICB, NFATC4, PIK3R3, PLCG2, PPP3CA, PRKCA, SH3BP2*, SHC4, VAV3
Focal adhesion	13\200	EGFR, FLNB, GRB2, ITGA1, ITGA4, ITGA9, ITGB3, PDGFB, PIK3P3, PRKCA, RAP1A (VAV3), SHC4, VAV3
Cytokine cytokine receptor interaction	11\257	CD40, CSF1R, CSF2RB, EGFR, IFNAR2, IL2RA, IL2RB, IL5RA, IL6R, PDGFB, *IL18R1 (IL1RL1)*
Calcium signaling pathway	12\174	ADCY2, ADCY8, ADCY9, CACNA1A, CACNA1D, CACNA1E, EGFR, ITPR3, PDE1C, PLCG2, PPP3CA, PRKCA
MAPK signaling	25\257	CACNA1A, CACNA1D, CACNA1E, CACNA2D1*, CACNA2D3*, CACNB2*, CACNG5*, EGFR, FGF14, FGF5, FGF6, FGFR2, FLNB, GRB2, MAP3K4, MAPKAPK2*, MEF2C*, NFATC4, PDGFB, PPP3CA, PRKCA, PTPN5, RASGRF2, *RAP1A (VAV3)*
Glycerophospholipid_metabolism	2\51	DGKH, PLCG2
Glycerolipid_metabolism	1\46	DGKH
Neuroactive ligand receptor interaction	0\254	`-
TLR9 signaling via IRF5	4\17	IL1RL1, SOCS1, TNFAIP3, IRF5
Smooth muscle contraction	9\150	ADCY2, ADCY8, ADCY9, ITPR3, PDE4B, PDE4D, PLCG2, PRKCA, PRKCQ
Phosphatidylinositol signalling system	4\90	DGKH, PLCG2, PRKCA, PRKCQ
Tight junction	5\136	CDK4*, CLDN10*, CLDN23*, PRKCA, PRKCQ
Purine metabolism	11\145	ADCY2, ADCY8, ADCY9, ADSS*, FHIT*, PAPSS2*, PDE10A*, PDE1C, PDE4B, PDE4D, POLD2
Gnrh signaling pathway	9\97	ADCY2, ADCY8, ADCY9, CACNA1D, EGFR, GRB2, ITPR3, MAP3K4, PRKCA
Wnt signaling pathway	3\149	NFATC4, PPP3CA, PRKCA
Gap junction	8\98	ADCY2, ADCY8, ADCY9, EGFR, GRB2, ITPR3, PDGFB, PRKCA
Valine leucine and isoleucine biosynthesis	1\12	*VARS (MICB)*
Cell communication	0\138	`-
Regulation of actin cytoskeleton	12\212	EGFR, FGF14, FGF5, FGF6, FGFR2, ITGA1, ITGA4, ITGA9, ITGB3, PDGFB, PIK3R3, VAV3
Ecm receptor interaction	4\87	ITGA1, ITGA4,ITGA9,ITGB3
Jak stat signaling pathway	15\153	CSF1R*, CSF2RB*, GRB2, IFNAR2, IL2RA, IL2RB, IL5RA, IL6R, PIAS2, PIK3R3, SOCS1, SOCS6*, SPRED2, SPRY1, STAT4
Taste transduction	3\53	ADCY8, CACNA1A, ITPR3
Glycan structures biosynthesis 1	0\117	`-

Table 17. The 21 identified genes that replicated in both WTCCC and NARAC and their most selected SNP.

HLA-DRA, *HLA-DQA1*, *PTPN22*, *IL2RB*, and *MICA* have shown association in previous studies. The remaining 17 are novel associations cross validating in both cohorts The ORs were calculated with respect to the reported allele.

Gene	Chromosome	Start pos	End pos	Most selected SNP in block	Allele	P (meta-analysis)	OR (95% CI)
HLA-DRA	6	32495624	32540802	rs9268853	C	4.56E-109	2.40 (2.20 - 2.60)
HLA-DQA1	6	32693160	32739407	rs9272219	T	9.58E-46	0.52 (0.48 – 0.57)
MICA	6	31459349	31511069	rs1063635	A	1.42E-17	0.74 (0.69 – 0.79)
HLA-G	6	29882734	29926878	rs1610677	C	4.12E-15	0.76 (0.71 - 0.81)
IL2RB	22	35831016	35872708	rs743777	G	2.09E-06	1.19 (1.10 – 1.30)
HLA-C	6	31324507	31367834	rs6906846	A	2.26E-05	0.85 (0.79 – 0.92)
EGFR	7	55033429	55128489	rs17172446	A	0.0001	0.86 (0.79 – 0.93)
TEC	4	47812556	47986571	rs2457414	C	0.0016	0.89 (0.83 – 0.96)
PTPN22*	1	114137960	114235857	rs2488457	G	0.0020	1.13 (1.10 – 1.20)
PDE4D	5	58280444	58658855	rs829451	T	0.0030	1.12 (1.10 – 1.20)
SHC4	15	46882243	46992983	rs16961743	T	0.0050	1.40 (1.10 – 1.80)
CACNA1D	3	53474897	53575012	rs3774425	A	0.0078	0.90 (0.85 – 0.98)
CACNA1A	19	13155623	13331522	rs3774425	A	0.0084	0.90 (0.83 – 0.97)
VAV3	1	107894802	108046933	rs10494071	T	0.0091	1.19 (1.00 – 1.37)
CACNB2	10	18578665	18827827	rs1757217	T	0.0132	1.15 (1.00 – 1.21)
PDE1C	7	31775674	31790442	rs992185	A	0.0163	0.80 (0.67 – 0.96)
FGF14	13	101150762	101335129	rs1537930	C	0.0167	1.10 (1.00 – 1.21)
ITGA9	3	37448558	37577446	rs6788347	G	0.0196	1.08 (1.00 – 1.21)
ADCY2	5	7594566	7712781	rs7717535	T	0.0200	0.92 (0.86 – 0.99)
CACNA2D1	7	81573468	81851660	rs38564	T	0.0274	0.92 (0.85 – 0.99)
PDE4B	1	66010515	66545937	rs17128076	A	0.0376	1.23 (1.00 – 1.50)

* Original analysis on NARAC study identified rs2476601 (single SNP P=2×10-11) as the most significant SNP mapped to *PTPN22* gene. I did not replicate this result as this SNP lies ~68kb away from the *PTPN22* boundaries and I only mapped SNPs to genes within 20kb of each side of the gene.

5.3 Functional enrichment of selected genes

In order to characterise the functional role of the identified genes, I collaborated with Dr Victoria Wright and my supervisor Prof Mike Levin, who assigned the identified genes to their cellular location using Ingenuity Pathways Analysis and further explored their biological function (Figure 25). There was a significant over-representation of genes located in the cell membrane, all of which are involved in signalling from the cell exterior to the interior. I then performed a gene ontology (GO) functional enrichment analysis using GOrilla.[76] Together with my colleagues we finally created a schematic map (Figure 26) of how these genes might act together in biological processes to determine RA susceptibility.

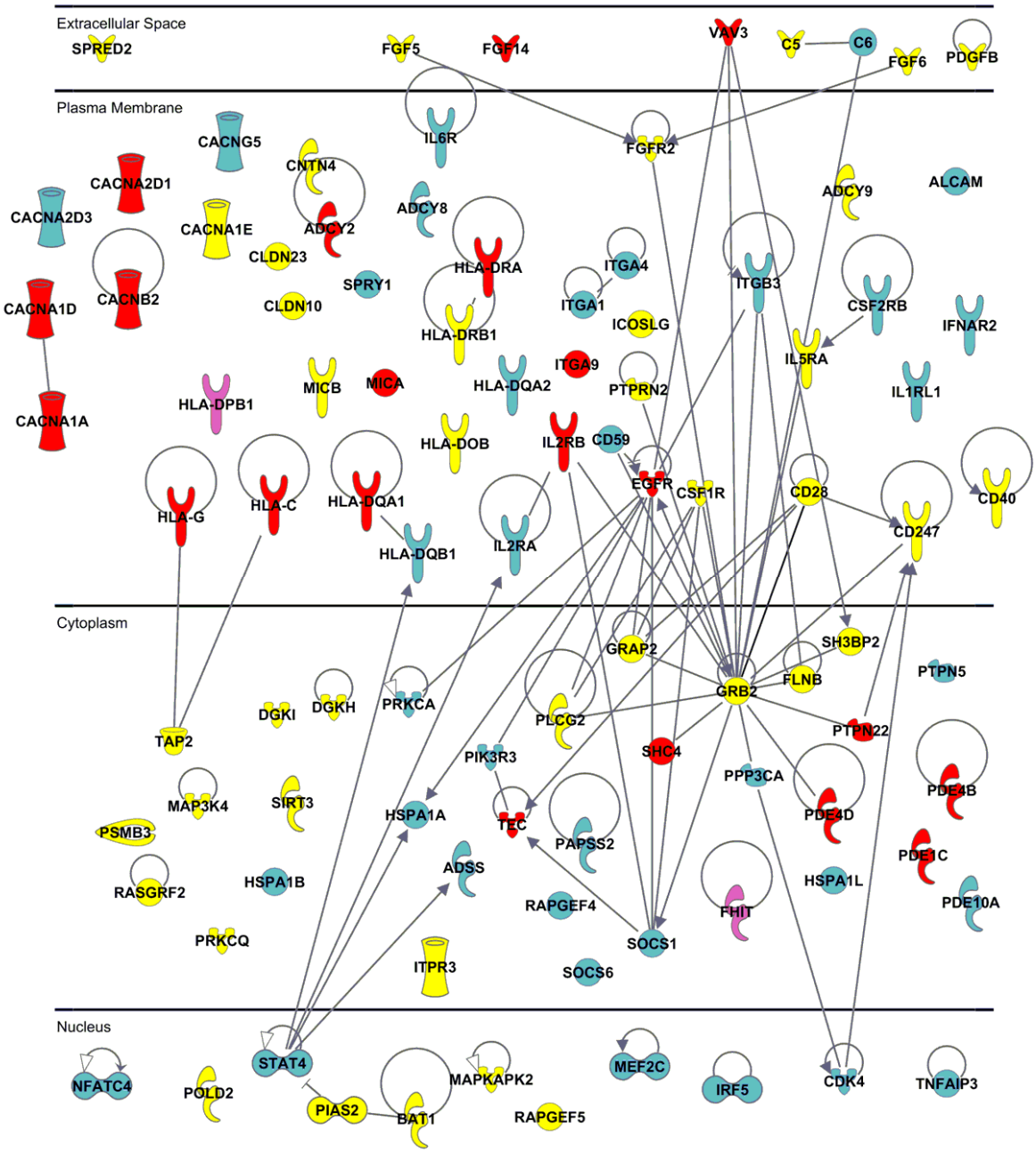


Figure 25. Cellular location of products of the selected genes.

Genes in red were selected in both WTCCC and NARAC; in yellow are genes selected only in NARAC. and in blue are genes selected only in WTCCC. Pink indicates that the same gene was identified in both studies but different gene block in each study. Genes coding for extracellular proteins are predominantly growth factors. Genes coding for intracellular proteins are predominantly involved with signal transduction. Lines connecting genes indicate direct interactions as defined from Ingenuity Pathway Systems knowledge base.

In addition to the expected involvement of HLA mediated processes (GO:0032395, P=5.1×10^{-14}), I observed significant enrichment in multiple GO functional categories, (Table 18, Table 19), including membrane receptor mediated signalling, integrin-mediated signaling (GO:0007229, P=1.53×10^{-5}), cytokine receptor (GO:0004896, P=1.72×10^{-8}) and growth factor binding (GO:0019838, P=1.09×10^{-7}), as well as adenylate cyclase activity (GO:0004016,

P=4.7×10^{-5}), 3'5'cyclic nucleotide phosphodiesterase activity (GO:0004114, P=9.05×10^{-6}), activation of protein kinase activity (GO: 0033674, P=5.3×10^{-9}) and voltage-gated channel activity (GO:0005245, P=1.11×10^{-10}), which contain genes involved in intracellular signalling through second messenger systems (SMSs)[174,175]. The SMSs are conserved, ubiquitous systems that relay signals from cell surface receptors to target molecules in the cytosol or nucleus, and greatly amplify the strength of the signal, through release of the second messengers: cyclic adenosine monophosphate (cAMP), inositol triphosphate (ITP)/diacylglycerol (DAG) and calcium ions (Ca^{2+})[174-176].

Adenylate cyclase genes (*ADCY2, ADCY8, ADCY9*) synthesise cAMP that activates multiple protein kinases (Figure 26). The activity of cAMP is then terminated by the phosphodiesterase (PDE) enzymes, of which several genes (*PDE10A*, *PDE4B*, *PDE4D*) were identified in either WTCCC or NARAC datasets suggesting control of cAMP (GO: 0004016, P=4.7×10^{-5}, GO:0004114, P=9.05×10^{-6}) as a potential key regulation point in RA pathophysiology.

The ITP/DAG system is activated by phospholipase C (PLC), which hydrolyses membrane phospholipids releasing DAG (Figure 26). DAG recruits calcium-dependent kinases that activate numerous pathways (GO: 0045859, P=6.98×10^{-9}) including RAS, ERK and NFκB signalling, interacting with the p38, NFAT and STAT signalling pathways.[174] The actions of DAG are terminated by the enzyme diacylglycerol kinase (DGK). Protein kinase C (PRKC) requires calcium ions generated from the action of PLC on the complex of inositol 1,4,5-trisphosphate (ITP) and DAG. PLC releases ITP which diffuses through the cytosol and binds to ITP receptors (ITPR) on the endoplasmic reticulum causing the release of Ca^{2+} into the cytosol (GO: 0042325, P= 2.99×10^{-10}). The identification of several genes in this system (*DGKH*, *DGKI*, *PLCG2*, *PRKCA*, *PRKCQ*, *ITPR3*) suggest the ITP/DAG signalling is another key regulatory step in RA (GO: 00045859, P=6.98×10^{-9}).

A further enriched SMS was the calcium second messenger system, which is based on cytosolic Ca^{2+} triggering numerous cellular processes.[176,177] (GO:0005262, P=1.64×10^{-8}; GO:0005245, P=1.11×10^{-11}, genes *CACNA1A*, *CACNA1D*, *CACNA1E*, *CACNA2D1*, *CACNA2D3*, *CACNB2*, *CACNG5*, *ITPR3*, *VAV3*). In addition to the release of Ca^{2+} from the endoplasmic reticulum, the entrance of Ca^{2+} into the cytosol from the exterior is also controlled by voltage gated channels, receptor-operated channels (GO: 007166, P=7.08×10^{-23}) and G-

protein coupled receptors, with an increase in cytosolic Ca^{2+} activating numerous calcium dependent proteins including PRKC (Figure 26).

Activation of protein kinase activity (GO: 0032147, $P=2.7\times10^{-7}$) and activation of phosphorylation (GO: 0042325, $P=2.99\times10^{-10}$, GO: 0001932, $P=8.47\times10^{-10,}$) were also enriched functional categories. The multiple identified genes coding for kinases and phosphatases (*PRKCA*, *PRKCQ*, *PIK3R3*, *PPP3CA*, *PTPN22*, *PTPN5*, *MAPKAPK2*, *MAP3K4*) are activated by the SMSs (Figure 26) and regulate multiple cellular functions including cell turnover, metabolism, secretion and growth [175,178,179].

Many of the remaining genes identified in either study that are not directly involved in receptor mediated signalling or SMSs are also known to regulate these systems (Figure 26). The proteins encoded by *VAV3*, *ALCAM*, *CD59*, *ITGB3* interact with those of *EGFR*, which interacts with *GRB2*, leading to *RAS* signal transduction. *EGFR* encoded protein also interacts with the *SPRY1* protein which inhibits growth factor signalling. *FGF5* and *FGF6* are ligands for *FGFR2* which interacts with *GRB2*. In terms of cytokines and their receptors, the encoded proteins of *TNFAIP3*, *IRF5*, *SOCS1*, *SOCS6*, *STAT4* and *PIAS2* are well known downstream regulators of cytokine induced inflammation, as are *RAPGEF4*, *RAPGEF5*, *RASGRF2*, *PRKCQ* and *NFATC4*. The biological interactions of all significant genes in both NARAC and WTCCC studies are shown schematically in Figure 4 as a functional cellular map, developed by my colleagues and myself in conjunction with my GO analysis.

The GO enrichment functional analysis, cell location mapping and the cellular map we created demonstrate further concordance between the findings in the two studies at the functional level. Other than the 21 genes that were identified in both WTCCC and NARAC studies, many of the genes, which were only selected in one study, had functional roles similar to genes selected in the other study.

Table 18. Gene ontology (GO) process enriched categories as derived by Gorilla for πt_{hr}=0.4

GO category	Description	WTCCC	NARAC	WTCCC ∪ NARAC
GO:0007165	signal transduction	4.89E-20	4.04E-16	2.31E-29
GO:0051716	cellular response to stimulus	1.67E-18	7.22E-14	2.50E-26
GO:0007166	cell surface receptor linked signaling pathway	1.04E-14	2.43E-13	7.08E-23
GO:0050896	response to stimulus	8.06E-14	5.98E-12	7.09E-20
GO:0050789	regulation of biological process	1.13E-13	1.32E-10	2.08E-18
GO:0050794	regulation of cellular process	1.79E-13	1.85E-10	8.15E-18
GO:0065007	biological regulation	1.10E-13	1.28E-09	1.13E-17
GO:0048583	regulation of response to stimulus	3.09E-10	8.00E-09	1.98E-17
GO:0002682	regulation of immune system process	1.22E-09	2.05E-08	1.30E-15
GO:0002376	immune system process	2.60E-08	2.71E-08	1.31E-14
GO:0050776	regulation of immune response	5.42E-09	1.52E-08	1.82E-14
GO:0023051	regulation of signaling	7.31E-09	3.75E-08	2.52E-13
GO:0002684	positive regulation of immune system process	3.44E-11	1.60E-04	2.73E-13
GO:0065009	regulation of molecular function	1.76E-11	3.20E-07	6.07E-13
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	1.80E-10	6.12E-06	2.50E-12
GO:0042325	regulation of phosphorylation	3.75E-06	1.01E-06	2.99E-10
GO:0019220	regulation of phosphate metabolic process	5.20E-06	1.37E-06	5.32E-10
GO:0050865	regulation of cell activation	6.76E-09	7.79E-04	5.43E-10
GO:0001932	regulation of protein phosphorylation	1.48E-05	4.31E-06	8.47E-10
GO:0010646	regulation of cell communication	1.37E-05	7.76E-08	9.21E-10
GO:0051249	regulation of lymphocyte activation	3.42E-08	4.15E-04	1.53E-09
GO:0033674	positive regulation of kinase activity	4.01E-08	2.08E-04	5.31E-09
GO:0045859	regulation of protein kinase activity	1.23E-06	2.09E-04	6.98E-09
GO:0051347	positive regulation of transferase activity	5.23E-08	2.39E-04	7.44E-09
GO:0032412	regulation of ion transmembrane transporter activity	2.79E-08	4.55E-07	1.72E-08
GO:0022898	regulation of transmembrane transporter activity	3.13E-08	5.06E-07	1.97E-08
GO:0034765	regulation of ion transmembrane transport	4.86E-08	6.85E-07	3.28E-08
GO:0051925	regulation of calcium ion transport via voltage-gated calcium channel activity	1.30E-07	5.57E-08	3.74E-08
GO:0046058	cAMP metabolic process	6.24E-07	4.61E-07	7.21E-08
GO:0070663	regulation of leukocyte proliferation	1.44E-06	3.08E-04	1.01E-07
GO:0007202	activation of phospholipase C activity	1.53E-05	7.97E-06	7.05E-07
GO:0010863	positive regulation of phospholipase C activity	1.61E-05	8.44E-06	7.60E-07
GO:0051924	regulation of calcium ion transport	1.44E-06	1.31E-05	1.65E-06
GO:0009187	cyclic nucleotide metabolic process	1.23E-05	9.16E-06	3.18E-06
GO:0010959	regulation of metal ion transport	3.90E-06	2.80E-05	5.18E-06
GO:0043269	regulation of ion transport	9.43E-06	6.24E-05	1.42E-05
GO:0007229	integrin-mediated signaling pathway	8.29E-07	2.05E-04	1.53E-05

Table 19. Gene ontology (GO) functional enriched categories as derived by Gorilla for πt_{hr}=0.4

GO category	Description	WTCCC	NARAC	WTCCC ∪ NARAC
GO:0032395	MHC class II receptor activity	1.18E-08	1.54E-08	5.10E-14
GO:0005515	protein binding	1.86E-11	1.55E-05	1.59E-13
GO:0005245	voltage-gated calcium channel activity	3.31E-08	1.11E-10	1.11E-10
GO:0005488	binding	9.01E-07	4.68E-04	1.42E-08
GO:0005262	calcium channel activity	3.66E-07	1.47E-07	1.64E-08
GO:0004896	cytokine receptor activity	1.77E-08	4.87E-04	1.72E-08
GO:0019838	growth factor binding	4.59E-04	2.45E-05	1.09E-07
GO:0005102	receptor binding	2.02E-06	3.89E-04	1.11E-07
GO:0004115	3',5'-cyclic-AMP phosphodiesterase activity	2.20E-05	1.75E-07	1.60E-06
GO:0004114	3',5'-cyclic-nucleotide phosphodiesterase activity	7.80E-05	1.00E-06	9.05E-06
GO:0004112	cyclic-nucleotide phosphodiesterase activity	8.84E-05	1.19E-06	1.07E-05
GO:0022843	voltage-gated cation channel activity	1.15E-04	4.73E-06	1.17E-05
GO:0004016	adenylate cyclase activity	9.16E-04	7.90E-04	4.70E-05
GO:0042578	phosphoric ester hydrolase activity	6.96E-04	5.63E-05	5.67E-05
GO:0022832	voltage-gated channel activity	3.97E-04	2.02E-05	6.55E-05
GO:0005244	voltage-gated ion channel activity	3.97E-04	2.02E-05	6.55E-05
GO:0008081	phosphoric diester hydrolase activity	2.10E-04	1.50E-04	1.08E-04
GO:0005261	cation channel activity	2.90E-04	1.69E-04	1.08E-04
GO:0022836	gated channel activity	5.26E-04	3.02E-04	2.32E-04

Many of the remaining genes identified in either study that are not directly involved in receptor mediated signalling or SMSs are also known to regulate these systems (Figure 26). The proteins encoded by *VAV3*, *ALCAM*, *CD59*, *ITGB3* interact with those of *EGFR*, which interacts with *GRB2*, leading to R*A*S signal transduction. *EGFR* encoded protein also interacts with the *SPRY1* protein which inhibits growth factor signalling. *FGF5* and *FGF6* are ligands for *FGFR2* which interacts with *GRB2*. In terms of cytokines and their receptors, the encoded proteins of *TNFAIP3*, *IRF5*, *SOCS1*, *SOCS6*, *STAT4* and *PIAS2* are well known downstream regulators of cytokine induced inflammation, as are *RAPGEF4*, *RAPGEF5*, *RASGRF2*, *PRKCQ* and *NFATC4*. The interactions of all significant genes in both NARAC and WTCCC studies are shown schematically in Figure 26 as a functional cellular map.

The GO enrichment functional analysis, cell location mapping and the cellular map I created demonstrate further concordance between the findings in the two studies at the functional level. Other than the 21 genes that were identified in both WTCCC and NARAC studies, many of the genes, which were only selected in one study, had functional roles similar to genes selected in the other study.

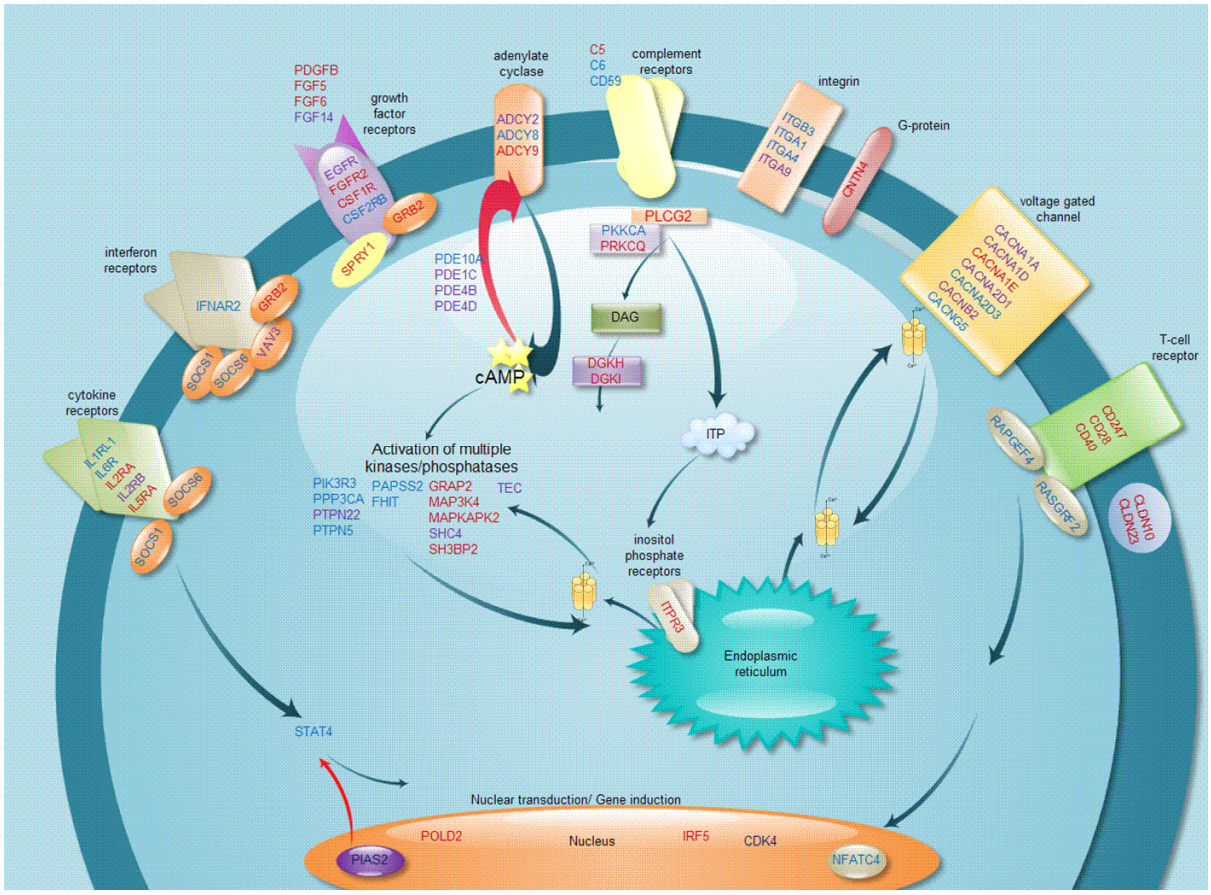


Figure 26. Cellular and functional map of non HLA genes identified by stability selection.
Purple - selected in both WTCCC and NARAC; blue- selected in WTCCC only; red – selected in NARAC only. Focus is given on the second messenger signalling systems (SMSs). In the cyclic AMP system (cAMP), adenylate cyclase (ADCY) synthesises cAMP from ATP. ADCY is activated by a range of receptors including G-protein coupled receptors. cAMP activates protein kinases (PKA) that phosphorylate target proteins. The enzymes phosphodiesterase (PDE) terminate the activity of cAMP switching off the process. The inositol triphosphate/diacylglycerol (ITP/DAG) system is activated by phospholipase C (PLC), which hydrolyses membrane phospholipids releasing DAG and ITP. DAG recruits calcium dependent protein kinase C and D (PKC, PKD). These activate numerous pathways including RAS and ERK leading to cell turnover, IKK and activation of NFKB signalling and interaction with p38 MAP kinase, NFAT and STAT signaling pathways. The actions of DAG are terminated by the enzyme diacylglycerol kinase, which converts DAG to phosphatidic acid. PKC requires calcium ions (Ca^{2+}) and these are made available by the action of ITP in the calcium signaling pathway. ITP, released from the complex with DAG by PLC, diffuses through the cytosol and binds to inositol triphosphate receptors (ITPR) on the endoplasmic reticulum causing the release of Ca^{2+} into the cytosol. The calcium second messenger system is activated by release of ITP, which binds to the ITPR on the reticuloendothelial system triggering release of Ca^{2+} from the ER into the cytosol. Entrance of Ca^{2+} into the cytosol is also controlled by voltage gated channels, such as CACNAC, ITPR or receptor operated channels and G protein coupled receptors. The rise in Ca^{2+} concentrations in the cytosol, as a result of influx from the endoplasmic reticulum or extracellular fluids through calcium channels, activates numerous calcium dependent proteins including PRKC. The actions of Ca^{2+} are terminated by active transport of Ca^{2+} out of the cytosol using calcium ATPases or sodium calcium exchanges.

5.4 Discussion

I have shown that using both biologically driven pathway information and a robust gene stability selection methodology yields improved power and consistent results across two independent RA cohorts. I have shown that this approach can identify novel, reproducible

associations from GWAS as well as replicate known associations that were not validated using single-SNP approaches using the same GWAS data. For example, the original NARAC report[8], presented the *TRAF1-C5* association and confirmed only the MHC region and the *PTPN22* gene, whereas my approach on NARAC apart from the aforementioned, also confirmed the *CD28*, *PRKCQ*, *CD40*, *IL2RB*, *SPRED2*, *CD247* and *TAP2* genes, which have been previously reported as modest associations with disease[180]. Similarly, in comparison to the original WTCCC report[9] that only confirmed the association of the MHC region and the *PTPN22* gene, I also confirmed the associations of *STAT4*, *TNFAIP3*, *IL2RA*, *IL2RB*, *IRF5* and *IL6R*[180]. Furthermore, of the 7 novel loci discovered in the recent meta-analysis of RA GWAS on a total of 41,282 samples[180], I discovered 3 (*IL2RA, IRF5, SPRED2*) and of the 10 suggestive loci I discovered 2 (*CD247*, *IL6R*) using only two studies with 4798 (WTCCC) and 2168 (NARAC) samples each.

My approach would also be applicable if only one GWAS cohort was available, although the calibration of the selection probability π_{thr} and expected model size to maximise the number of discoveries and minimise the FDR would not be possible, and I would suggest using the values derived in this work. If more than two GWAS cohorts were available, it would be straightforward to extend my approach by analyzing all studies at the pathway level separately and then taking the top ~10% of pathways to the gene stability selection methodology; genes would then be selected per study by following the exact procedure that was described in Chapter 4, but using a multivariate hypergeometric distribution to control the false discovery rate and set the threshold above which genes would be denoted as "stably selected".

My results suggest that my pathway-based and gene stability-selection analysis is robust to genetic heterogeneity between different samples and to the technical heterogeneity of different genotyping platforms. I have replicated my findings at the pathway level by identifying the same pathways in both studies lying in the top 6% of the rankings; at the gene level by identifying 21 genes in common of the 60 and 64 selected in WTCCC and NARAC studies respectively and by validating 10 (WTCCC) and 12 (NARAC) known associations that were not validated using single-SNP approaches on the same GWAS data; at the functional level through demonstrating that many genes selected in only one study have functionally similar genes selected in the other study. A recent review[45] commented that at present, there is no clear way to compare various pathway analysis methods against each other, but there is a need to replicate association results. To my knowledge, this is the first combined pathway and gene-based methodology that has replicated association findings in two real GWAS datasets.

Finally, the functional analysis of the identified genes suggests that the underlying concept of RA as a largely autoimmune and inflammatory disease should be broadened. My findings suggest that defective second messenger signalling through cAMP, ITP, DAG and Ca^{2+} may be at the heart of RA predisposition. The SMSs are used by all cells to switch on and off multiple processes including cell cycle, differentiation, growth, secretion and immune response [175,176]. SMSs regulate both the intensity and duration of cellular responses and the identification of multiple genetic variants in the SMSs suggests that RA may arise from a defect in the on /off switch controlling both the amplitude and duration of key cellular functions. The SMSs control not only immune response cells but chondrocytes, osteocytes, connective tissue and blood vessel cells, as well as processes such as osteoclast resorption, differentiation of chondrocyte and osteoblasts ,and repair and removal of senescent cells [181,182]. Thus a defect in on/off signaling may explain the widespread cellular and connective tissue disorder in RA in addition to abnormal function of inflammatory cells (Figure 27).

It is of interest that numerous cellular studies have documented alterations in cAMP, inositol phosphate and calcium signalling in RA, including abnormalities in these pathways in inflammatory cells in the joints, in *in-vivo* studies of osteocytes and chondrocytes [182-185] supporting my hypothesis. Future studies should explore the relationship between genes identified and functional cellular responses to test this hypothesis, particularly as all steps in the processes I have implicated may be amenable to pharmacological manipulation. My identification of the SMSs as playing a key role in RA susceptibility suggests a novel concept of the genetic mechanisms of the disease which potentially opens new avenues for pharmacological intervention to prevent and treat the disorder.

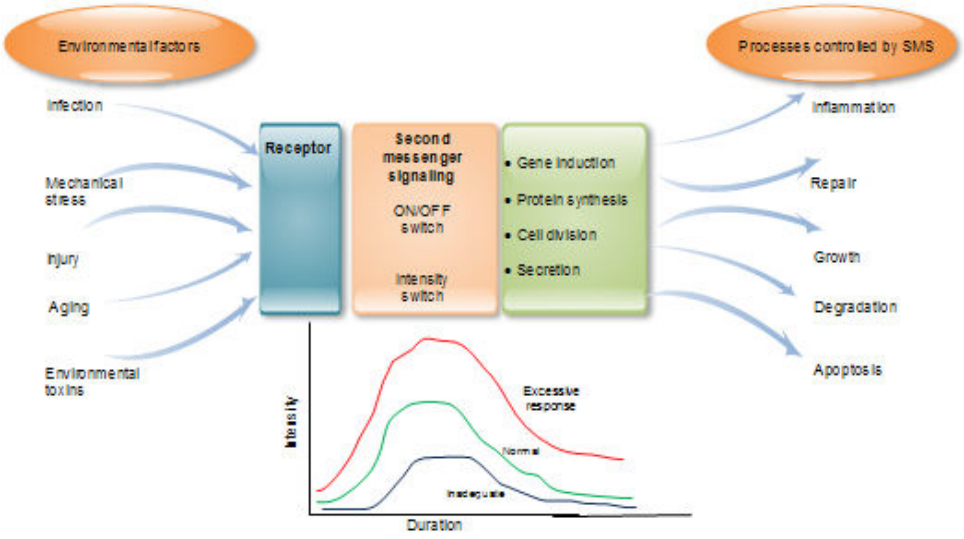


Figure 27. Schematic representation of how genetic variation in genes in the second messenger signalling systems cause altered intensity and duration of cellular responses to multiple receptor mediated stimuli.

5.5 Summary

In this chapter, I presented the results of the first application of my pathway-driven gene-stability selection methodology. I analyzed the WTCCC and the NARAC studies of ~5,000 and ~2,000 subjects respectively. I examined 700 pathways comprising ~8,000 genes. Ranking pathways by significance revealed that the NARAC top-ranked ~6% laid within the top 10% of WTCCC. Gene selection on those pathways identified 58 genes in WTCCC and 61 in NARAC; 21 of those were common ($P_{overlap}<10^{-21}$), of which 16 were novel discoveries. Amongst the identified genes, I confirmed 10 known RA associations in WTCCC and 13 in NARAC, which were not discovered using single-SNP approaches on the same data. Gene ontology functional enrichment analysis on the identified genes showed significant over-representation of signalling activity ($P<10^{-29}$) in both studies. My findings suggest a novel model of RA genetic predisposition, which involves cell membrane receptors and genes in second messenger signalling systems, in addition to genes that regulate immune responses, which have been the focus of interest previously.

Chapter 6

Application to Kawasaki disease

As another application of the pathway-based approach to GWAS analysis, I worked on a hypothesis-driven investigation of the association between the TGF-β pathway and the Kawasaki disease. I initially analyzed a case-control GWAS and I then examined a genome-wide association study of family trios. For the latter, I also had to modify my *CT* test statistic to apply to genotype data from family trios. The results were further explored by RNA expression studies of the same pathway in patients with acute disease, using microarray and reverse transcription polymerase chain

reaction (RT-PCR), as well as protein analysis. All gene-expression and protein work was carried out by my colleagues and our collaborators in the University of San Diego and the International Kawasaki Disease Genetics Consortium (IKDGC), as I have stated it in the Declaration of Authentication. In this chapter I will describe the extension of my approach to family-trio GWAS and the investigation of the association between the TGF-β pathway and the Kawasaki disease susceptibility, severity and treatment response.

6.1 Introduction

Kawasaki disease (KD) is an acute, systemic vasculitis associated with cardiovascular sequelae that include coronary artery aneurysms (CAA), aortic root (AoR) dilatation, and myocardial fibrosis[186,187]. CAA develop in 25% of untreated patients and 3-5% of patients treated with intravenous immunoglobulin (IVIG) within the first 10 days after fever onset.[188] AoR dilatation occurs in approximately 15% of treated patients.[189,190] Histological examination of tissues from KD autopsies has revealed myocardial fibrosis, dysregulation of angiogenesis and coronary artery thrombosis, stenosis, and calcification[191,192]. The aetiology of KD remains unknown, although an infectious trigger affecting genetically susceptible hosts is suspected.[193-197] The marked differential susceptibility to KD in North Asians compared to Europeans, maintained following migration to low incidence countries, together with a 10-fold increased relative risk in siblings strongly suggest a genetic predisposition to KD susceptibility. [198,199] Genetic and immunologic studies suggest that T-cell activation has an important role in the pathogenesis of KD.[192,193,197,200,201] Transforming growth factor (TGF)-β modulates the balance of pro-inflammatory and anti-inflammatory T-cells through a complex set of interactions. The importance of cardiovascular remodelling and immune activation in KD suggests that TGF-β pathway may play an important role in KD pathogenesis[192,193,197,200,201]. In collaboration with the IKDGC and my supervisor Prof. Michael Levin, we investigated this hypothesis by analyzing the genetic variation and transcript abundance patterns for genes in the TGF-β pathway. Prof. Jane Burns' group also measured levels of TGF-β2 peptide in acute and convalescent KD plasma samples. These studies revealed that genetic variation in the TGF-β pathway influences KD susceptibility, severity, and treatment response.

6.2 Data and Methods

6.2.1 Clinical perspective

TGF-β is a multifunctional peptide that regulates proliferation, differentiation, apoptosis, and migration in many cell types. In the cardiovascular system, TGF-β can induce neoangiogenesis, cardiomyocyte hypertrophy, calcification, and fibrosis[202,203]. Alteration of TGF-β signaling (Figure 28) has been implicated in the pathophysiology of several vascular disorders including Marfan and Loeys-Dietz syndromes, which are associated with CAA and AoR dilatation and vascular complications of sickle cell anemia[204-207]. In the immune system, TGF-β modulates the balance of pro-inflammatory and anti-inflammatory T-cells through a complex set of interactions.

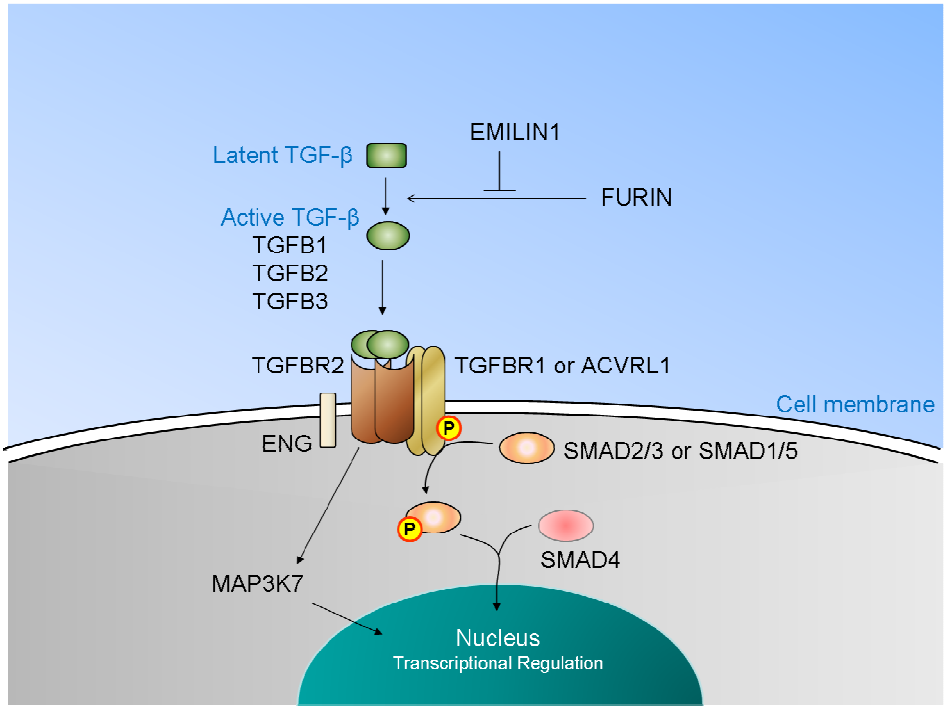


Figure 28. TGF-β signaling pathway

TGF-β is secreted in a latent form and activation is mediated by several molecules including furin and emilin. Active TGF-β peptides (TGFB1, 2, 3) bind to the Type II receptor (TGFBR2), which recruits and activates Type I receptors such as TGFBR1 or ACVRL1. The activated Type I receptor phosphorylates receptor-specific SMAD molecules: TGFBR1 phosphorylates SMAD2 and 3, and ACVRL1 phosphorylates SMAD1, 5 and 8. These activated SMADs form a larger complex with Smad4 and translocate to the nucleus where they regulate gene transcription. Accessory receptors such as ENG have a role in the balance of ACVRL1 and TGFBR1 signaling to regulate endothelial cell proliferation. In addition to the classical SMAD signaling pathway, a SMAD-independent pathway using alternative molecules such as MAP3K7 can also mediate TGF-β signaling. TGFB: Transforming growth factor-β, TGFBR: Transforming growth factor-β receptor, EMILIN1: Elastin microfibril interfacer 1, SMAD: SMAD family member, MAP3K7: Mitogen-activated protein kinase kinase kinase 7, ENG: Endoglin, ACVRL1: Activin A receptor type II-like 1, FURIN: Furin

6.2.2 Subjects

The recruitment of KD patients and the details of their clinical presentation and diagnosis are described in a previous work of IKDGC[208]. IVIG resistance was defined as previously described[209]. A total of 771 KD subjects and 159 controls were genotyped and divided into different cohorts for analysis. Cohort 1 was comprised of 128 Dutch European descent KD subjects (Table 20, Figure 29) and 159 Dutch European descent healthy adult control subjects. Cohort 2 was comprised of 451 biologic trios (mother, father, affected child, European descent 81% and mixed European descent 19%) from the US, UK and Australia (US/UK/Aus). Cohort 3 (n=435) was comprised of Dutch Cohort 1 KD subjects (n=125) plus subjects from the European descent Cohort 2 (UK n=48, Australian=186) plus additional European descent KD subjects from the UK (n=29) and Australia (n=47). Cohort 4 (n=237) was comprised of European descent US KD subjects from Cohort 2 (n=121) plus additional European descent KD subjects from the US (n=116) (Table 20, Figure 29).

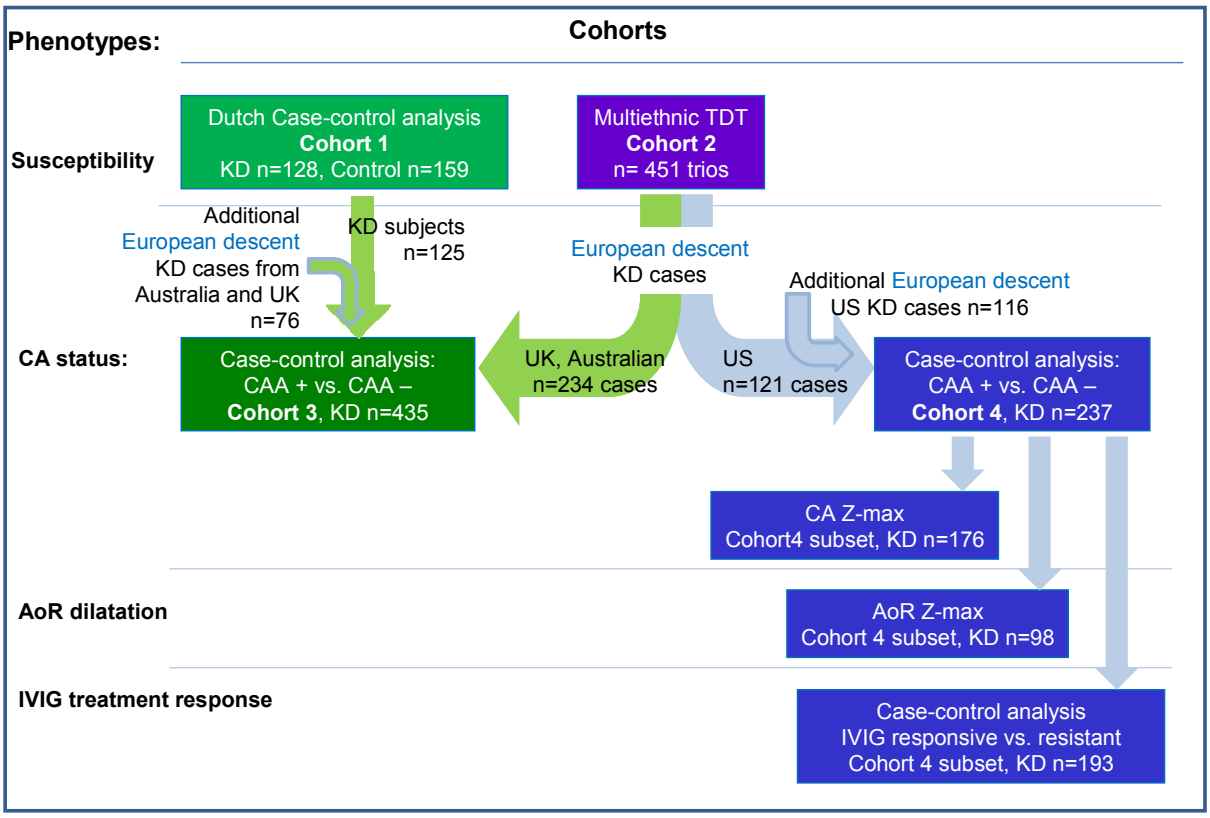


Figure 29. The cohorts and the phenotypes that were examined in different stages of analysis

Table 20. Characteristics of subjects in their respective cohorts

		Cohort 1 NL Case-control study (KD n=128, Controls n=159)				Cohort 2 US/UK/Aus TDT (KD n=451)		Cohort 3 UK/NL/Aus (KD n=435)		Cohort 4 US only (KD n=237)	
		n	**%**	**n**	**%**	**N**	**%**	**n**	**%**	**n**	**%**
Gender:	Male	77	60.1	74	46.5	270	59.9	261	60.0	153	64.6
	Female	50	39.1	57	35.9	181	40.1	174	40.0	84	35.4
	Unknown	1	0.8	28	17.6	0	0	0		0	0
CA status:	CAA+	28	21.9	NA	NA	71	15.7	73	16.8	51	21.5
	CAA-	97	75.8	NA	NA	349	77.4	362	83.2	186	78.5
	Unknown	3	2.3	NA	NA	31	6.9	0		0	0
Z worst:											
	CA	NA	NA	NA	NA	NA	NA	NA	NA	176	74.3
	Aortic root	NA	NA	NA	NA	NA	NA	NA	NA	98	41.4
IVIG treatment response*	Resistant	NA	NA	NA	NA	NA	NA	NA	NA	46	19.4
	Responsive	NA	NA	NA	NA	NA	NA	NA	NA	147	62.0

*44 US KD subjects could not be classified because they had either no treatment, other treatment or no information was available

NA: not available; NL: Netherlands, US: United States; UK: United Kingdom; Aus: Australia; TDT: Transmission disequilibrium test; CA: coronary artery; CAA: coronary artery aneurysm; IVIG: intravenous immunoglobulin

6.2.3 Materials

Blood samples. RNA was isolated from whole blood collected in PAXgene tubes from US KD subjects (Table 21,Figure 30) as described in Popper *et al.*[209] Plasma (sodium citrate) for ELISA was collected from KD patients during acute (pre-IVIG) and convalescent (≥1 month after fever onset and normalization of erythrocyte sedimentation rate and platelet count) phases. Blood samples were processed within 48 hours of collection and plasma was stored at -70°C.

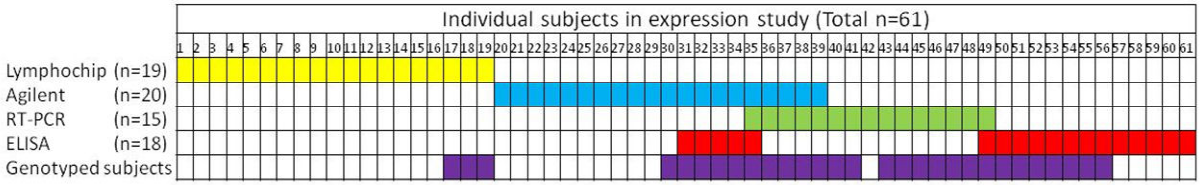


Figure 30. Subjects analyzed in expression studies

Yellow cells: subjects used for Lymphochip microarray, Blue cells: subjects used for Agilent microarray, 4 of 20 subjects were also used for RT-PCR, 4 of 20 subjects were also used for ELISA assay and 1 of 20 was used for both RT-PCR and ELISA. Green cells: subjects used for RT-PCR, Red cells: subjects used for ELISA assay, Purple cells: subjects used for genotype assays

Table 21. Characteristics of subjects in microarrays, RT-PCR and ELISA

		Microarrays				RT-PCR		ELISA	
		Lymphochip		Agilent array					
		n	%	n	%	n	%	n	%
Total subjects	n	19		20		15		18	
Gender	Male	12	63	9	45	11	73	8	44
CA status	CAA+	2	11	1	5	2	13	3	17
	CAA-	17	89	19	95	13	87	15	83
IVIG treatment	Resistant	2	11	8	40	5	33	5	28
	Responder	17	89	12	60	10	67	4	22
	Others*	0		0		0		9	50
Ethnicity	Asian	7	37	1	5	0	0	1	6
	Caucasian	4	21	6	30	8	53	13	71
	Hispanic	4	21	7	35	4	27	2	11
	≥2 races	3	16	6	30	3	20	1	6
	African American	1	5	0	0	0	0	1	6

Echocardiography. All KD subjects had echocardiographic assessment during the acute and convalescent stages of their illnesses. For all cohorts, coronary artery (CA) lesions (aneurysm or ectasia) were defined according to the Japanese Ministry of Health criteria[210]. The coronary artery status of the US subjects (Cohort 4) was assessed by echocardiography during the acute and sub-acute phase and serially thereafter as dictated by patient status. Measurements of the internal diameters of the proximal right (RCA) and left anterior descending (LAD) coronary arteries were normalized for body surface area and expressed as standard deviation units from the mean (Z scores)[211]. The worst-ever Z-score for either the RCA or LAD at any time point or the Z score of the largest aneurysm was used for the continuous variable analysis (CA Z-max).

AoR dimensions (annulus and sinus) were measured at only two US sites (Rady Children's Hospital San Diego and Boston Children's Hospital). Measurements performed within 1 year from the onset of KD were normalized for body surface area and expressed as Z-scores.[212] The variable, (AoR Z-max), which was the largest dimension of the AoR at any timepoint during the first year after disease onset, was used for the continuous variable analysis. For subjects who were not seen within the first year after disease onset, we used measurement of the AoR from the first available echocardiogram.

Genotyping. Genomic DNA from whole blood or mouth wash samples was extracted as described in a previous study of the same group[208]. SNPs (n=164) that were selected from HapMap Caucasian populations with minor allele frequencies ≥ 0.1, from 15 genes in the TGF-β signaling

pathway (Figure 28) were genotyped using a custom Illumina™ Oligo Pool Assay.[208] All SNPs met quality control criteria with genotype call rate ≥ 93% and no deviation from Hardy-Weinberg equilibrium in controls and parents of KD subjects.

Microarray analysis. Two different types of microarray platforms (Stanford University Lymphochip, 19 KD subjects, and Agilent G4112F Human Whole Genome Oligo Arrays (design #14850), 20 KD subjects) were used to study two independent cohorts of KD subjects[209,213] (Table 21, Figure 30). For the Lymphochip array, 55 clones from 15 genes in the TGF-β pathway were filtered to meet quality standards as described in Popper *et al*[209]. Data from individual cDNA clones were included if paired acute and convalescent measurements were available for at least 10 of the 19 KD subjects, and data from clones associated with a single Unigene cluster were averaged. Differences in transcript abundance in acute and convalescent samples were assessed using a paired *t*-test. For the Agilent arrays, 22 clones from 15 genes in the TGF-β pathway were examined. They were subject to the calculation of average difference, *P*-value of Student *t*-test, and q-value of FDR, based on the 22 tests, between acute and convalescent stages. Regarding the quality control (QC) of the arrays, the signal intensity of at least 80% of the samples tested was 2.5-fold above background in either the Cy5 (sample) or Cy3 (reference) channel, and a regression correlation for the two channels was at least 0.6 across each measured element.

RT-PCR. To validate the microarray results, transcript abundance levels were measured by RT-PCR for *SMAD1*, *SMAD3*, *ACVRL1, FURIN*, *TGFBR3* and *TGFBR2* for 14 KD subjects with acute and convalescent paired whole blood RNA samples (Table 21, Figure 30). Relative abundance of the target transcripts was normalized to the expression level of the house keeping gene, TATA box binding protein-associated factor, RNA polymerase I, B (*TAF1B*), as described in Popper *et al*[209] *SMAD1* (ABI Assay ID Hs00195432_m1), *SMAD3* (ABI Assay ID Hs00969208_m1), *ACVRL1* (ABI Assay ID Hs00953798_m1), *FURIN* (ABI Assay ID Hs00159829_m1), *TGFBR3* (ABI Assay ID Hs00234257_m1) and *TGFBR2* (ABI Assay ID Hs00234253_m1) Of the 14 subjects, none overlapped with the 19 subjects used for the Lymphochip analysis and five of 20 were also studied on the Agilent microarrays.

ELISA. The concentrations of TGFB2 in paired acute and convalescent plasma samples were compared in 18 KD subjects (Quantikine, R&D system) (Table 21, Figure 30).

6.2.4 Methods

6.2.4.1 Pathway analysis

Pathway analysis was applied using the *CT* test statistic as described in Chapter 2. I extended the statistic to be applicable to family trios by replacing the Cochran-Armitage test statistic with the transmission disequilibrium test (TDT)[214], as they are both approximated by a χ^2 distribution with 1 *df*.

The TDT is a family-based test of linkage and association. The TDT requires two unaffected parents and an affected offspring, which is usually called a family-trio as depicted in Figure 31. In the lack of association between the disease and the marker in the population, the parental genotypes are random. In the lack of linkage between the marker and the disease locus, the transmission of the marker allele is independent of transmission of the disease. In this respect, TDT tests for both population association and linkage.

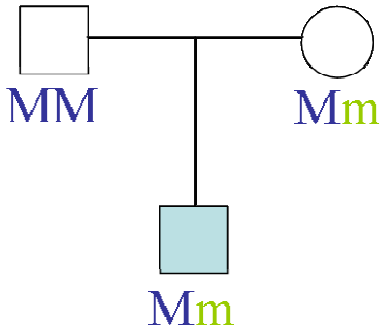


Figure 31. Pedigree of a family trio.
Square stands for male and round for female individual. Color filled rectangle stands for affected and no fill stands for unaffected. *M* and *m* are the two alleles of a bi-allelic marker. In this trio, one *M* from the father and the *m* from the mother were transmitted to the offspring.

Let us assume a marker locus with alleles *M* and *m* as depicted in Figure 31. Imagine that we have *n* family trios of *n* affected children, whose *2n* unaffected parents and their alleles can be summarized in a 2x2 table:

Table 22. Contigency table of transmitted and not-transmitted alleles in family trios.

	Not-transmitted allele		
Transmitted allele	**M**	**m**	**Total**
M	a	b	a+b
m	c	d	c+d
Total	a+c	b+d	2n

The counts *a* and *d* are uninformative, as they correspond to homozygous parents and so transmitted and untrasmitted alleles are the same. For that reason, **only heterozygous parents** are informative and thus taken into account. If there is no association or no linkage, the numbers of transmitted *M* and not-trasmitted *M* should be the same. The same applies to allele *m*. The TDT assesses whether the proportions $b/(b+c)$ and $c/(b+c)$ have 0.5 and 0.5 probabilities respectively. TDT tests this hypothesis using a χ^2 test statistic with one df:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{6.1}$$

To obtain the cumulative trend test statistic for pathway association, I sum the TDT over all of the *n* SNPs in a given pathway. If I denote as TDT_i the TDT statistic for the i^{th} SNP, then the cumulative trend test statistic for the pathway is

$$CT_{pathway} = \sum_{i=1}^{n} TDT_i \,, \tag{6.2}$$

where *n* is the number of SNPs in the pathway.

Permutations of the trio data were performed by randomizing the transmitted allele. I applied this analysis to Cohort 2 (451 biological trios). I also applied the *CT* test statistic, as described in Chapter 2 for a case-control design to Cohort 1 (128 KD cases-159 controls). Finally, I applied a unified pathway analysis to the merged dataset of Cohort 1 and the affected children of Cohort 2 (579 KD cases -159 controls).

6.2.4.2 Single-SNP analyses

The transmission disequilibrium test (TDT) was applied to test for the single-SNP association between the disease and the single markers on the family-trio data[215].

Case-control association analysis between the single SNPs and the disease, as well as the CA status was conducted using the general linear model: $Y_i = a + \beta X_i + \epsilon_i$. The genetic models that

were considered were the dominant model (genotype model:11 vs.12+22) and the additive model (allele model: 1 vs. 2). Under the dominant model, X_i=1 if the i^{th} subject has at least one minor allele or X_i=0 otherwise. Under the additive model, X_i indicates the i^{th} subject's number of minor alleles. This is equivalent to the test based on the allele frequency. All *P*-values were adjusted via means of the Bonferroni procedure to correct for multiple testing.

Association analyses between genetic variants and the Z-max for the coronary arteries were performed using non-parametric tests due to the non-normality of the Z-score. Associations between the genetic variants and AoR Z-max were analyzed using a general linear model. The genetic models that were considered were the additive, dominant and recessive models. Analysis was performed using the SNPassoc and DGCgenetics packages in the R software (http://www.r-project.org/).

Paired *t*-tests were conducted to compare transcript abundance levels, as well as to compare transcript abundance levels and plasma TGFB2 levels (PRISM v5.01). Association between genetic variants and gene expression or TGFB2 protein levels were analyzed with either the Kruskal-Wallis test or the Mann-Whitney test (PRISM v5.01).

6.3 Results

6.3.1 TGF-β signaling pathway and KD susceptibility

Analysis of the cumulative variation of 15 genes in the TGF-β pathway in the merged dataset of Cohorts 1 and 2 showed evidence of association between the pathway and KD susceptibility. In particular, I analyzed 6 different versions of the TGF-β pathway, as they were perceived and assembled by Dr. Chisato Shimizu (Table 23). I observed moderate association signals in all the versions of the pathway and the merged dataset of Cohort 1 and affected children of Cohort 2; the most significant association signal was observed for the 15 genes altogether with *P*=0.00065 (Table 23).

Table 23. Versions of the TGF-β pathway and *CT* test statistic results.
P-values are given for the analysis of Cohort 1 [*P*(1)], Cohort 2 [*P*(2)] and the merged dataset [*P*(merged)]

Pathway	#Genes	Genes	*P* (1)	*P* (2)	*P* (merged)
TGF-β	15	TGFB1, TGFB2, TGFB3, TGFBR1, TGFBR2, ACVRL1, SMAD1, SMAD2, SMAD3, SMAD4, SMAD5, ENG, MAP3K7, FURIN, ENDOGLIN	0.280	0.289	0.00065
TGF-β #1	6	TGFB2, TGFBR2, ACVRL1, SMA2, SMAD3, ENDOGLIN	0.219	0.062	0.0041
TGF-β #2	8	TGFB3, TGFBR1, TGFB2, TGFBR2, ACVRL1, SMAD2, SMAD3, ENDOGLIN	0.249	0.085	0.0025
TGF-β #3	7	TGFB1, TGFB2, TGFBR2, ACVRL1, SMAD2, SMAD3, ENDOGLIN	0.227	0.082	0.0018
TGF-β #4	6	TGFB1, TGFB2, ACVRL1, SMAD2, SMAD3, ENDOGLIN	0.099	0.004	0.0017
TGF-β #5	7	TGFB3, TGFBR1, TGFB2, ACVRL1, SMA2, SMAD3, ENDOGLIN	0.115	0.007	0.0011

Gene-based analysis on the merged dataset identified *TGFB2* (*P*=0.006), *TGFBR2* (*P*=0.08), *ACVRL1* (*P*=0.04), *SMAD3* (*P*=0.06) and *FURIN* (*P*=0.01) as the ones driving the pathway signal. Gene-based analyzed was performed exactly as pathway-based analysis, but only on the SNPs of each gene.

Single-SNP analysis of the case-control Cohort 1 (128 KD vs 159 control subjects) identified SNPs in 6 genes in the pathway with a modest influence on KD susceptibility ($P \in [0.0031\text{-}0.047]$ (Table 25). The significance of genetic variation in 3 of these 6 genes (*SMAD3*, *TGFB2*, and *TGFBR2*) was replicated in the TDT analysis of the independent US/UK/Aus subjects (Cohort 2: 451 trios) that identified 7 different SNPs associated with KD susceptibility (Table 24, Figure 32). Five of the 7 SNPs were in *SMAD3*, with the most significant result for an intronic SNP (A/G, rs4776338) for which 272 of the 453 heterozygous parents transmitted the G allele to their affected offspring (*P*=0.003, OR 1.50, 95%CI 1.25-1.81, calculated using the DGCgenetics package in R).

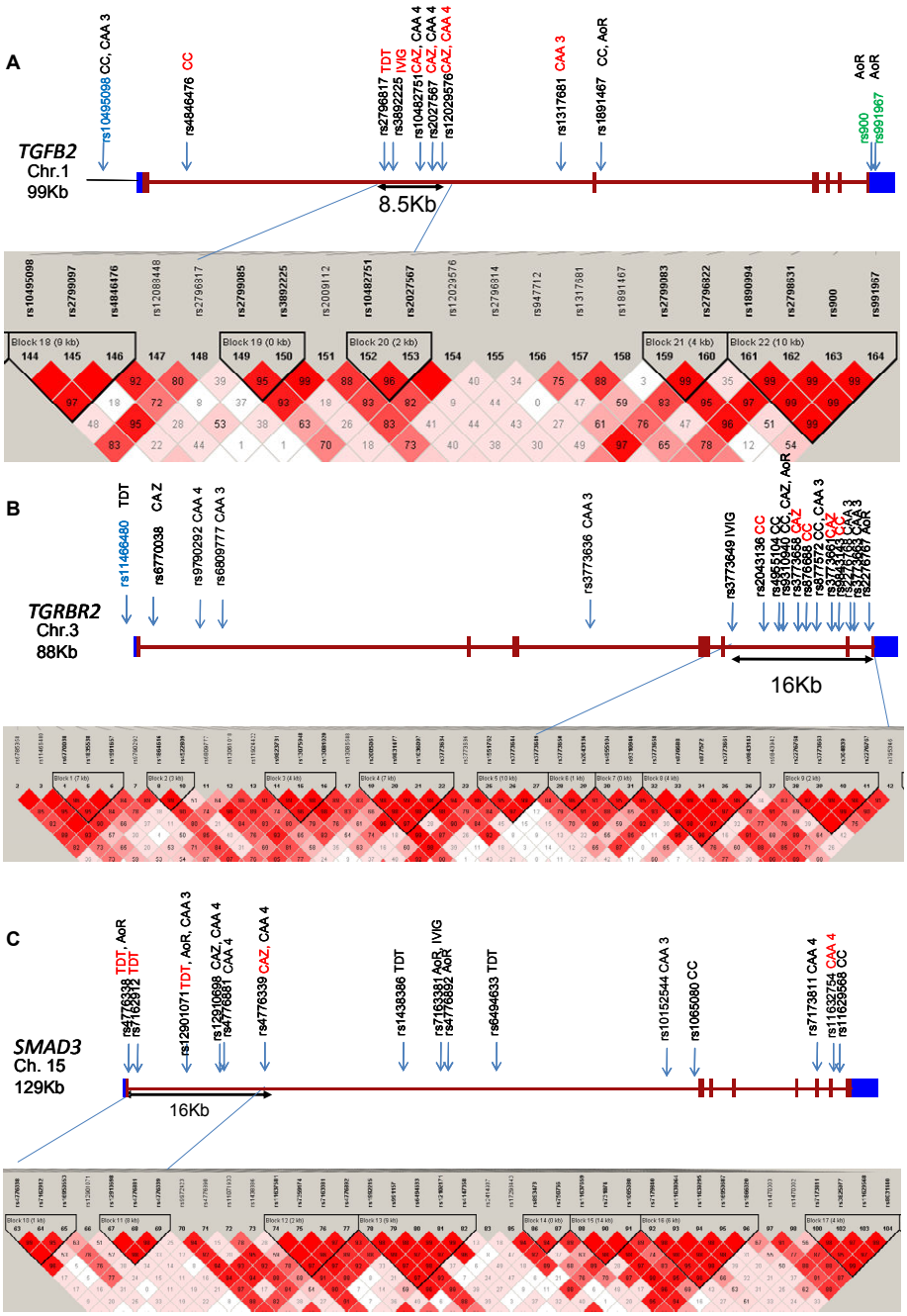


Figure 32. Significant SNP locations for *TGFB2* (A), *TGFBR2* (B) and *SMAD3* (C)

Arrows show the location of significant SNPs genotyped in this study. Gene structure and the location of SNPs are shown: boxes= exons and 3' and 5' untranslated regions;. Underlined text highlights SNPs with $P<0.01$. Letter code above SNP rs# refers to type of analysis and cohort: CC: susceptibility in Cohort 1 (Table 25), TDT: susceptibility in Cohort 2 (Table 24), CAA3 and CAA4: association with CAA in Cohort 3 and Cohort 4, respectively (Table 26), CAZ: association with CA Z-worst (Table 27), AoR; association with AoR Z-worst (Table 28), IVIG; association with IVIG treatment response (Table 29). LD maps are shown for SNPs of interest.

Table 24. TDT analysis of genetic variants and KD susceptibility in US/UK/Australian trios (n=451).
* Results are shown only for SNPs with a *P*-value <0.05. OR: odds ratio, CI: confidence interval

Gene	rs number	Allele 1/2	Frequency	Allele 2 Transmitted	Not transmitted	OR	95%CI		*P*-value
TGFB2	rs2796817	A/C	0.13	139	98	1.42	1.10	1.84	0.0092
TGFBR2	rs11466480	A/C	0.03	12	27	0.44	0.23	0.88	0.024
SMAD3	rs4776338	A/G	0.42	272	181	1.50	1.25	1.81	0.000022
	rs7162912	C/A	0.31	229	174	1.32	1.08	1.60	0.0071
	rs12901071	A/G	0.29	230	169	1.36	1.12	1.66	0.0026
	rs1438386	A/G	0.41	190	237	0.80	0.66	0.97	0.026
	rs6494633	G/A	0.43	174	219	0.80	0.65	0.97	0.026

Table 25. Association of genetic variants and KD susceptibility in Dutch KD subjects and controls
Results are shown only for SNPs with a nominal *P*-value <0.05

Chr.	Gene	SNP	Allele	Allele 2 frequencies KD n=128	Control	Allele 2 vs. Allele 1 OR	95% CI		*P*-value
1	TGFB2	rs10495098	C/A	0.35	0.44	0.68	0.48	0.96	0.028
		rs4846476	G/C	0.17	0.27	0.54	0.36	0.82	0.0032
3	TGFBR2	rs2043136	A/G	0.2	0.29	0.59	0.39	0.89	0.0095
		rs4955104	T/A	0.17	0.24	0.66	0.43	1.01	0.052
		rs9310940	C/A	0.44	0.54	0.65	0.46	0.91	0.01
		rs876688	G/A	0.42	0.3	1.66	1.18	2.34	0.0031
		rs877572	C/G	0.4	0.49	0.72	0.52	1.00	0.047
		rs9843143	C/G	0.54	0.42	1.57	1.13	2.18	0.006
4	SMAD1	rs6537355	A/G	0.12	0.11	1.08	0.65	1.78	0.041
		rs2118438	G/A	0.17	0.22	0.71	0.47	1.08	0.11
9	ENG	rs4451422	A/C	0.48	0.38	1.46	1.05	2.03	0.023
		rs10760503	A/G	0.5	0.4	1.47	1.06	2.04	0.02
		rs1330684	G/A	0.45	0.36	1.41	1.01	1.97	0.045
12	ACVRL1	rs706819	G/A	0.24	0.35	0.62	0.44	0.88	0.007
15	SMAD3	rs1065080	G/A	0.14	0.08	1.7	1.01	2.89	0.045
		rs11629568	A/C	0.39	0.47	0.72	0.51	1.00	0.052

6.3.2 TGF-β signaling pathway and coronary artery outcome

Genetic variation in *TGFB2*, *TGFBR2, and SMAD3* consistently influenced coronary artery outcome in 2 independent, non-overlapping cohorts: Cohort 3 from the UK, Australia, and the Netherlands (CAA-: n=362, CAA+: n=73) and Cohort 4 from the US (CAA-: n=186, CAA+: n=51) (Table 26). Although the associated SNPs in Cohort 3 and 4 were different, many of the SNPs co-localized to the first intron of each of the 3 genes (*TGFB2*, *TGFBR2* and *SMAD3*) (Figure 32).

The collaborators in this study analyzed the association between genetic variants and CA Z-max from a subset of Cohort 4 (US, n=176). Significant differences were revealed in the distribution of CA Z-max for 11 SNPs in 4 genes in the pathway (Table 27). Whether the analysis was performed using Z-max (n=176) or using the presence or absence of aneurysms in the complete

Cohort 4 (176 + 61= 237), 3 SNPs in *TGFB2* (rs10482751, rs2027567, rs12029576) and 2 SNPs in *SMAD3* (rs12910698, rs4776339) were consistently associated. (Table 26, Table 27, Figure 32).

Table 26. Case-control analysis of genetic variants and coronary artery aneurysms (CAA) in US subjects (CAA-: n=186, CAA+: n=51) and UK/Aus/NL subjects (CAA-: n=362, CAA+: n=73). Results are shown only for SNPs with a nominal *P*-value <0.05.

			Cohort 3			**Cohort 4**		
			Allele2 frequency			**Allele2 frequency**		
Chr.	**Gene**	**SNP**	**CAA-**	**CAA+**	***P***	**CAA-**	**CAA+**	***P***
1	TGFB2	rs10495098	0.37	0.46	0.016			
		rs10482751				0.30	0.41	0.019
		rs2027567				0.21	0.31	0.028
		rs12029576				0.27	0.42	0.0011
		rs1317681	0.15	0.24	0.0090			
3	TGFBR2	rs9790292				0.46	0.39	0.029
		rs6809777	0.27	0.18	0.034			
		rs3773636	0.25	0.32	0.035			
		rs877572	0.47	0.34	0.027			
		rs2276768	0.10	0.18	0.039			
		rs3773663	0.41	0.53	0.025			
4	SMAD1	rs6537355	0.14	0.08	0.018			
5	SMAD5	rs3764942	0.08	0.03	0.008			
12	ACVRL1	rs11169953	0.33	0.26	0.014			
15	SMAD3	rs12901071	0.39	0.32	0.030			
		rs12910698				0.44	0.54	0.014
		rs4776881				0.44	0.54	0.014
		rs4776339				0.47	0.58	0.014
		rs10152544	0.45	0.50	0.015			
		rs7173811				0.48	0.40	0.038
		rs11632754				0.23	0.12	0.0032
19	*TGFB1*	rs8179181	0.24	0.14	0.015			

Table 27. Analysis of genetic variants and CA Z-max in US KD subjects (n=176)

CA: coronary artery. SNPs with an asterisk * were found also significantly associated with CA Z-max in Cohort 4. Results are shown only for SNPs with a nominal *P*-value <0.05.

Chr.	Gene	SNP	Genotype	N	CA Z-max	Interquartile range	Mann-Whitney *P*
1	TGFB2	rs10482751*	GG	78	1.50	(0.95- 2.60)	0.0008
			AG+AA	98	1.97	(1.32- 4.44)	
		rs2027567*	AA	99	1.70	(0.99- 2.65)	0.0032
			AG+GG	77	1.94	(1.30- 4.60)	
		rs12029576*	AA	83	1.52	(0.99- 2.60)	0.0009
			AC+CC	93	2.00	(1.30- 4.10)	
3	TGFBR2	rs6770038	GG	122	1.91	(1.20- 3.70)	0.03
			AG+AA	54	1.43	(1.00- 2.75)	
		rs9310940	CC	54	1.85	(1.60- 4.02)	0.049
			AC+AA	121	1.64	(1.00- 3.10)	
		rs3773658	AA	136	1.90	(1.19- 3.73)	0.0052
			AG+GG	40	1.29	(0.89- 2.15)	
		rs3773661	GG	137	1.90	(1.20- 3.70)	0.0033
			CC+CG	39	1.28	(0.89- 2.10)	
15	SMAD3	rs12910698*	CC	52	1.43	(0.99- 2.25)	0.023
			AC+AA	123	1.94	(1.20- 3.40)	
		rs4776881	AA	52	1.43	(0.99- 2.25)	0.022
			AG+GG	124	1.94	(1.20- 3.40)	
		rs4776339*	AA	42	1.40	(0.84- 1.99)	0.0058
			AG+GG	134	1.93	(1.20- 3.63)	
15	FURIN	rs2071410	GG	80	2.05	(1.30- 3.92)	0.014
			CC+CG	92	1.60	(1.00- 2.45)	

6.3.3 TGF-β signaling pathway and aortic root dimension

The maximal internal diameter for the aortic root normalized for body surface area (AoR Z-max) was available for a subset of the US subjects (n=98) (Table 20). Twenty SNPs in 8 genes in the pathway, including *TGFB2*, *TGFBR2* and *SMAD3,* were significantly associated with AoR Z-max (Table 28). One SNP (rs9310940) within *TGFBR2* and one SNP (rs12901071) within *SMAD3* were significant in both the analysis of CA outcome and the analysis of AoR dilatation (Figure 32).

6.3.4 TGF-β signaling pathway and IVIG treatment response in the US subjects

Case-control analysis of treatment response as a function of genotype was performed for the US subjects (IVIG-resistant n=46, IVIG-responsive n=147) (Table 29). The same 3 genes (*TGFB2*, *TGFBR2* and *SMAD3*) that were associated with KD susceptibility, coronary artery aneurysm formation, and aortic root dilatation were again associated with IVIG treatment response.

Table 28. Association of genetic variants with Ao Z-max in US KD subjects (n=98)

Results are shown only for SNPs with a nominal *P*-value <0.05. SD=standard deviation. The SNP followed by an asterisk * was also significantly associated in both the analysis of CA outcome and the analysis AoR dilatation.

Chr.	Gene symbol	SNP	Genotype	n	AoR Z-max (mean)	SD	*P*
1	TGFB2	rs900	AA	46	1.60	0.10	0.022
			AT+TT	52	1.27	0.10	
		rs991967	AA	42	1.60	0.11	0.028
			AC+CC	55	1.29	0.10	
3	TGFBR2	rs9310940*	CC	34	1.63	0.12	0.038
			CA+AA	63	1.32	0.09	
		rs2276767	CC	49	1.27	0.10	0.029
			CA+AA	48	1.59	0.10	
4	SMAD1	rs714195	GG	33	1.63	0.13	0.042
			GA+AA	65	1.32	0.08	
		rs11100886	AA	32	1.63	0.14	0.043
			AG+GG	66	1.32	0.08	
6	MAP3K7	rs205347	AA	40	1.23	0.11	0.021
			AG+GG	55	1.56	0.09	
		rs3799897	AA	59	1.54	0.10	0.04
			AG+GG	39	1.24	0.10	
9	TGFBR1	rs6478974	AA	29	1.67	0.15	0.023
			AT+TT	69	1.32	0.08	
9	ENG	rs1330684	GG	38	1.24	0.11	0.039
			GA+AA	60	1.54	0.09	
15	SMAD3	rs4776338	GG	27	1.68	0.17	0.027
			GA+AA	70	1.32	0.07	
		rs12901071	AA	40	1.24	0.10	0.036
			AG+GG	58	1.55	0.10	
		rs7163381	GG	46	1.57	0.11	0.049
			GA+AA	52	1.29	0.09	
		rs4776892	AA	55	1.56	0.10	0.032
			AT+TT	42	1.25	0.10	
19	TGFB1	rs10417924	GG	62	1.30	0.09	0.021
			GA+AA	35	1.64	0.12	

Table 29. Association of genetic variants with IVIG treatment response in US KD subjects

Results are shown only for SNPs with a nominal *P*-value <0.05.

				Allele 2 frequencies		Allele2 vs. Allele 1				Genotype 11
Chr.	Gene	SNP	Allele	Resistant	Responsive	OR	95%CI		*P*	*P*
1	TGFB2	rs3892225	A/G	0.34	0.14	0.41	0.21	0.77	0.0035	0.029
3	TGFBR2	rs3773649	G/A	0.8	0.47	0.59	0.36	0.96	0.034	0.043
5	SMAD5	rs746994	G/A	0.62	0.3	0.53	0.32	0.89	0.016	0.014
		rs10056474	C/G	0.82	0.44	0.57	0.34	0.94	0.029	0.026
15	SMAD3	rs7163381	G/A	0.29	0.39	1.97	1.07	3.61	0.046	0.079

6.3.5 TGF-β pathway transcript abundance and plasma levels in acute and convalescent KD

To discover differences in transcript abundance levels of genes in the TGF-β pathway between acute and convalescent KD samples, in collaboration with Prof. Jane Burn's group we analyzed two independent microarray experiments each with 19 paired samples (Lymphochip array: 2 subjects had CAA and 2 had IVIG-resistance; Agilent array: 1 subject had CAA and 8 had IVIG-resistance). We found differential transcript abundance for all 15 genes of TGF-β pathway in both platforms. Transcript abundance levels for *SMAD3* and *TGFBR3* were significantly lower during the acute compared to convalescent stage on both platforms; transcript abundance for *ACVRL1*, *SMAD1,* and *FURIN* were higher during the acute phase on both array platforms (Table 30, Figure 33).

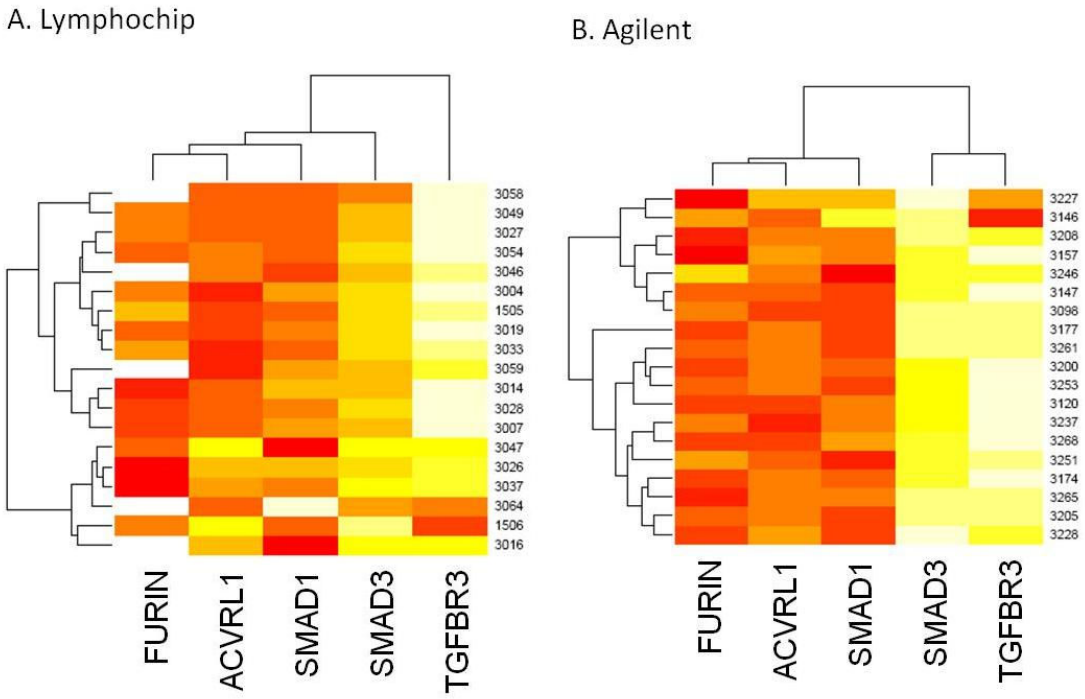


Figure 33. Heatmaps for Lymphchip array and Agilent array results for 5 genes of interest
Yellow: Convalescent levels higher than Acute levels, Red: Acute levels higher than Convalescent levels.

After Bonferroni correction, all transcripts on the Lymphochip array except for *SMAD3* remained significant. All transcripts on the Agilent array were under the q value of FDR. To validate the microarray results, they performed RT-PCR on cDNA for transcripts of *ACVRL1, SMAD1, FURIN, SMAD3,* and *TGFBR3*. Increased transcript levels for *ACVRL1* (p=0.02), *SMAD1* (p=0.001) and decreased *TGFBR3* (p=0.02) were documented during the acute phase (Figure 34A,

B, D). *FURIN* was detected in only three of 14 acute whole blood RNA samples and in none of the convalescent samples. To test whether the change in expression levels was simply a function of neutrophil number, they examined the correlation between *ACVRL1*, *SMAD1*, *SMAD3*, *TGFBR3* and *FURIN* gene expression levels and found only weak associations: $r^2 \leq 0.36$ for absolute neutrophil count (Agilent array) and $r^2 \leq 0.58$ for percent neutrophil change (Lymphochip array). Therefore, the dynamic changes in gene expression levels could not be explained simply by differences in cell populations in the peripheral blood. Genotype data were available for 14 of 15 subjects whose gene expression levels were measured by RT-PCR, (Figure 30). We analyzed the association between changes in transcript levels and genotype for the most significant SNPs in *ACVRL1* rs11169953, *SMAD1* rs6537355 and *SMAD3* rs4776338 (Figure 34E-G). No significant association was detected in this small number of subjects. Plasma concentrations of TGF-β2 during the acute phase were lower compared to the convalescent phase (P=0.002) (Figure 35) (n=18, 3 had CAA and 5 were IVIG-resistant). For the 13 subjects whose genotypes were available, we analyzed the association between changes of TGF-β2 plasma levels and genetic variants of rs12029576, the most significantly associated SNP in *TGFB2* (P=0.0011, Table 26). No significant association was detected in this small number of subjects.

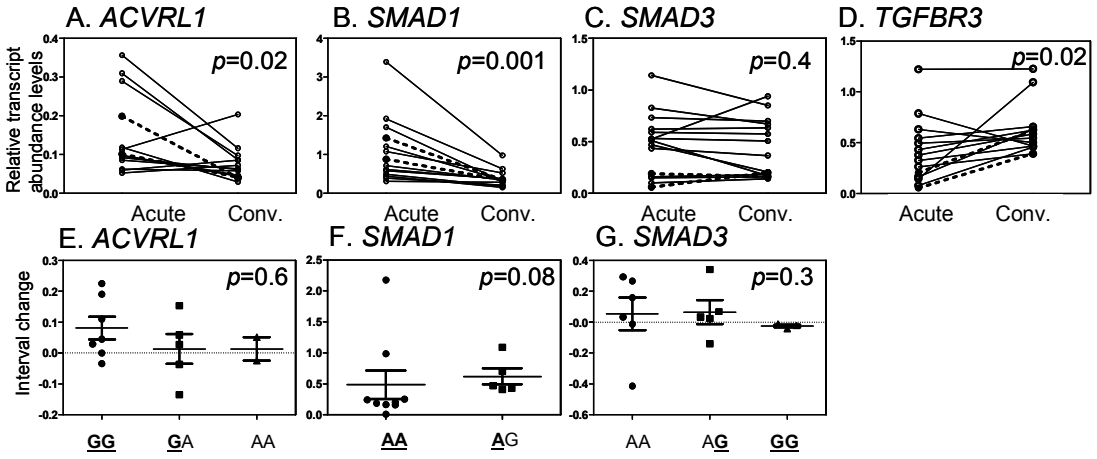


Figure 34. RT-PCR analysis of transcript abundance in whole blood from KD patients during the acute and convalescent phase and association between genotype and gene expression

Relative transcript abundance levels were normalized by *TAF1B* for *ACVRL1* (A), *SMAD1* (B), *SMAD3* (C) and *TGFBR3* (D). Solid line: subjects with normal coronary arteries, Dashed line: subjects with coronary artery aneurysms. *p* value by paired t-test. (E-G) Association between acute and convalescent gene expression levels and *ACVRL1* rs11169953 (risk allele= **G, GG**: n=7, **G**A: n=5, AA: n=2) (E), *SMAD1* rs6537355 (risk allele=**A**, **AA**: n=9, **A**G: n=5, no subject had GG genotype). (F), *SMAD3* rs4776338 (risk allele= **G** , AA: n=6, A**G**: n=5, **GG**: n=3) (G). *p* value by Kruskal-Wallis test (E, G) and Mann Whitney test (F).

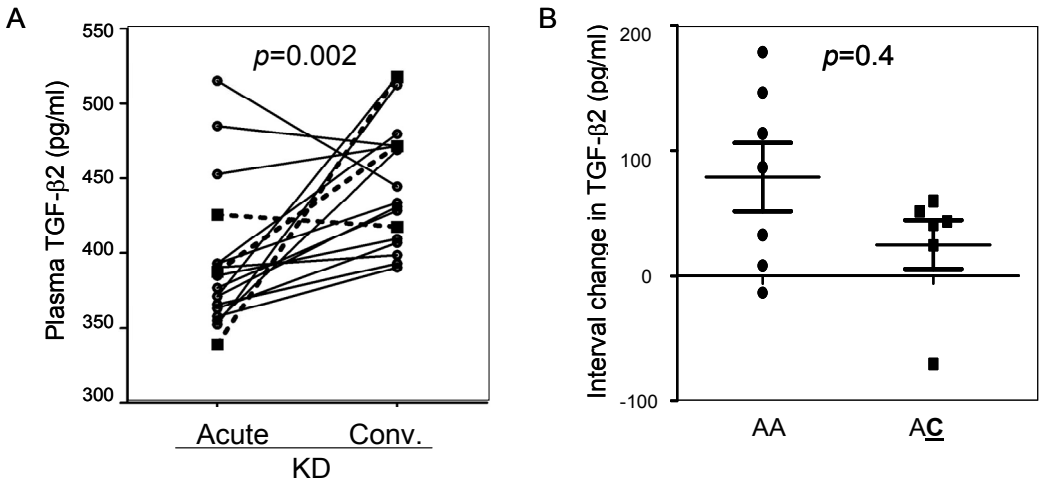


Figure 35. (A) Plasma levels of TGF-β2 in KD patients during the acute and convalescent phase

Solid line: subjects with normal coronary arteries, Dashed line: subjects with coronary artery aneurysms. *p* value by paired t-test. (B) Association between acute and convalescent plasma levels of TGF-β2 and rs12029576 in TGF-β2 (risk allele= C, AA: n=7, AC: n=6, no subject had CC genotype). *P*-value by Mann Whitney test.

Table 30. Transcript abundance patterns for differentially expressed genes in the TGF-β pathway.
Results are shown only for genes with SNPs with a *P*-value <0.05

		Lymphochip array		Agilent array	
Gene	**Transcript abundance**	***P***	**Fold**	***P***	**Fold change**
ACVRL1	**↑Acute, ↓Conv.**	0.0024	1.34	4.29E-08	2.03
SMAD1		0.00027	1.25	3.84E-06	2.37
FURIN		0.0012	1.37	6.27E-07	2.38
SMAD3	**↓Acute, ↑Conv.**	0.015	0.91	9.59E-07	0.6
TGFBR3		0.001	0.58	0.0007	0.47

6.4 Discussion

Several lines of evidence presented in this study suggest that the TGF-β signaling pathway plays an important role in KD pathogenesis and that genetic variation in three genes in the pathway (*TGFB2*, *TGFBR2* and *SMAD3*) influence KD susceptibility, CAA formation, AoR dilatation and IVIG treatment response (Figure 32, Table 24, Table 26, Table 27 ,Table 28, Table 29). This is the first genetic analysis to find an association between CA and aortic root Z-scores and genotype. Genetic variants associated with KD phenotypes clustered in discrete regions based on single-SNP analysis. On the basis of the Venter genome analysis, individuals of European descent are likely to carry 200 to 500 non-synonymous, rare (minor allele frequency <5%) and/or novel variants that affect protein function.[6,216] Sequencing the implicated regions of the three genes, *TGFB2, TGFBR2* and *SMAD3*, may identify novel or rare genetic variants associated with KD pathogenesis. The significance of genetic variation in this pathway was further highlighted by the finding of dynamic changes in transcript abundance and plasma protein levels during the course of the illness. An analysis of the dynamic change in transcript or protein levels as a function of genotype failed to show an association, potentially because the sample size was very small. Future studies with a bigger sample size will be needed to more definitively address the effect of these polymorphisms on gene expression.

Each of the three genes identified as important by this study has a plausible role in KD pathogenesis. The first gene, *TGFB2,* encodes one of the three isoforms of TGF-β and is expressed in both a tissue-specific and a developmentally regulated fashion. Only TGFB2 is expressed and required for endocardial cushion cell transformation in the mouse[217] and *Tgfb2* knockout mice suffer from a variety of cardiovascular anomalies.[218-220] Plasma levels of TGF-β2 in the KD subjects were low in the acute compared to the convalescent phase. Similar findings have been reported for TGF-β1 plasma levels and transcript abundance in acute KD but were not measured in this study.[221-223]

The second gene highlighted by this study was *TGFBR2*, which was associated with KD susceptibility, CAA formation, AoR dilatation and response to IVIG treatment. Genetic variation in this gene is linked to several aneurysm syndromes including Loey-Dietz and abdominal aortic aneurysm.[224,225] Immunohistochemistry of histologically normal and aneurysmal coronary arteries 3-12 years after disease onset demonstrated TGF-β1, TGFBR1 and TGFBR2 in intimal or medial smooth muscle cells[226,227]. This was not seen in age-matched autopsy tissues from children dying of other conditions, suggesting the importance of this pathway in arterial remodeling late after KD.

The third gene, *SMAD3*, has an essential role in down-regulating T-cells and increasing expression of FoxP3, an essential step in the differentiation of regulatory T cells[228]. Imbalance of pro-inflammatory Th17 and regulatory T-cells has been reported in acute coronary syndrome[229]. Recent data from Prof. Jane Burns' laboratory documented the emergence of peripherally induced regulatory T-cells in acute KD subjects, suggesting that this pathway may be important in disease recovery.[230] SMAD3 also has an important role in cardiovascular remodeling and fibrosis after injury, both active processes in the recovery phase from KD.[231]

Naturally, there are several limitations to this study. First, only the tag-SNPs with a minor allele frequency of >=10% were genotyped. Thus, other rare, but important, genetic variants would not have been detected by this analysis. Secondly, although I and my collaborators, we have identified genetic variants that influence disease susceptibility and CAA outcome, none of these intronic variants have a known function. The clustering of significant SNPs in introns suggests the possibility of regulatory elements that might be encoded for by these regions. Although this is the first analysis to find an association between CA and aortic root Z-scores and SNPs, the number of subjects with CAA and AoR dilatation was limited and the findings will need to be validated in independent cohorts. The cohorts that were used may have been biased toward cases with CAA because of sample availability (CAA 73/362 (20%), and 51/186 (27%) in cohort 3 and 4). This may have affected the analysis of KD susceptibility. This issue can only be clarified by testing the findings in larger cohorts of subjects without CAA. KD is highly prevalent in Asian populations and this study only addressed genetic variation in subjects of European descent. Future studies must also address the mechanism by which these variants exert their effect. Studies of T-cell activation and differentiation of regulatory T-cells in cultured peripheral blood mononuclear cells from individuals

of known genotype will contribute to our understanding of how modulation of the TGF-β pathway impacts KD pathogenesis.

This study supports the importance of the TGF-β signaling pathway in KD pathogenesis. Recently, Onouchi *et al*[232] discovered a functional polymorphism in inositol 1,4,5-triphosphate 3-kinase C that increases calcineurin/NFAT signaling and leads to increased levels of IL2 and T-cell activation in KD subjects. Genetic variation in the TGF-β pathway may lead to an imbalance of pro-inflammatory and regulatory T-cells by affecting Foxp3 expression mediated through *SMAD3* and *NFAT*.[228] The TGF-β pathway is also important in inflammation and tissue remodeling mediated by endothelial cells, fibroblasts, and smooth muscle cells. Investigation of the association between the calcineurin/NFAT and TGF-β pathways in T-cells and immunohistochemical study of tissue will further our understanding these pathways in patients with KD. Both of these pathways have important potential for pharmacologic intervention with the availability of calcineurin inhibitors (eg. cyclosporine and tacrolimus) and modulators of the TGF-β signaling pathway (eg. angiotensin receptor blockers). Therefore, further study of these pathways may lead to important new therapeutic intervention strategies for KD patients.

6.5 Summary

In this chapter I described the extension of my methodology on family-trio GWAS, other than the typical case-control design. I presented the application of my approach on two GWAS of Kawasaki disease, (one case-control, one family-trio), where I examined the association between the TGF-β pathway and the disease. The evidence of association was further examined by microarray, RT-PCR and protein-level experiments and the results were validated. This chapter was a stand-alone application of my methodology. In the next chapter, I go on to present the conclusions I arrived to after 3-years of my PhD and close by my general view of the field and its future.

Chapter 7

Conclusions

At the time I commenced my PhD project, there were very few published studies that utilised pathway-driven methodologies in the analysis of GWAS. In fact, my work evolved concurrently with a publication bloom on studies introducing various pathway-based approaches. For that reason, the real challenge has been the need to address the several conceptual and technical difficulties in a field that was and possibly still is at the very early stages of development.

7.1 What has this thesis achieved

All in all, in the last three years, with the guidance of my supervisors and the contribution of the scientists in our group and our collaborators, I reviewed the relevant literature and gained insight on the technical challenges, the relative advantages and disadvantages of pathway and gene-based analyses of GWAS (Chapter 1). I developed a novel test statistic that summarizes the cumulative genetic variation along a biological pathway and then tests the association between the disease and the pathway as a single entity (Chapter 2). I applied this approach to the three autoimmune GWAS of the WTCCC dataset (CD, T1D, RA) and I identified several associations with inflammatory pathways. My methodology validated almost all previously reported associated genes; it identified associations that were missed at the time of the original publication and were only found subsequently in large meta-analyses; it revealed new plausible gene associations with the diseases in question (Chapter 3). I then went on to utilize the pathway-driven methodology to build predictive models of disease and assessed them robustly via cross-validation and out-of-sample replication. Specifically, to my knowledge this was the first study at the time to have presented such a highly predictive genetic model of disease in T1D and RA. I validated the findings derived from the WTCCC dataset analysis on a population dataset of NFBC1966 (Chapter 3). I showed that pathway-driven markers can assist disease prediction and that pathway-based model building outperforms in predictive power non-pathway-based model building (Chapter 3).

I refined my original methodology to a pathway-driven gene stability selection approach that identifies pathways associated with a disease and it then narrows down the signal of association to a kernel of genes that predominantly account for the effect. On the grounds that biological pathways usually share multiple gene constituents, it identifies the genes within those pathways that drive jointly the signal of these multiple pathway associations (Chapter 4). I showcased that my methodology provides reproducible results at the pathway, gene and *in-silico* function level on two independent GWAS datasets of RA (Chapter 5). My findings on RA suggest a new model of predisposition to RA, where cellular signalling might be at the heart of RA susceptibility, compared to the MHC scenario that has been the dominant hypothesis until today (Chapter 5). Lastly I extended my approach to be applicable on GWAS of family-trios and analyzed one case-control and one family-trio dataset on KD. I demonstrated evidence for the association between the TGF-β pathway and the disease that was then validated by gene-expression and protein studies on this pathway and on the same individuals.

I have completed a thesis that has brought to light new interesting directions on the analysis of GWAS and has equipped me with a strong background on the field. At the same time, I have addressed only in a suboptimal way the several implications that are inherent in the pathway-driven analysis of GWAS. Although I suggest potential solutions and provide results to support them, I do not imply that the challenges have been resolved. As I discuss in the overview of gene-based and pathway-driven methodologies (Chapter 1), the challenges that will need to be further explored in the future include the as yet limited understanding of what constitutes a pathway; the accuracy of information that is deposited on public pathway databases and how to manage the overlapping pathways; the question of how to assign SNPs to genes; how to account for the confounding effects of LD patterns and gene size; how to address the single SNP hits or the large number of genes in a pathway that might elevate the pathway statistics; and how to identify the genes within associated pathways predominantly responsible for the pathway effect, and finally how to validate the findings on independent datasets given the genetic heterogeneity between studies.

At the same time, the evolution of technology on genotyping platforms that allows >1M SNPs to be typed per study (e.g. Omni 2.5M and Omni 5M chips are already available) and the 1000 Genomes project (http://www.1000genomes.org/) that will make available ~5 million SNPs to be used for imputation will vastly increase the computational burden, which is an issue not to be disregarded. My pathway-based methodology will not be able to scale up to millions of typed SNPs in a straightforward way, as I have already demonstrated that it is positively biased towards very large number of SNPs. In such a case, I would have to apply PCA separately on cases and controls and test the association of the first principal components of a gene rather than each SNP; I would then sum the statistics in the same manner I have described in this thesis. One way that my approach could accommodate for CNV pathway-based analysis, would be to test for the association between each CNV with the disease by fitting a logistic regression model. Then I could derive the z-statistic of the coefficient it, square it and it would then follow a $\chi 2$ distribution with 1 *df* under the null hypothesis of no association. I could then derived a similar pathway-based test statistic by summing the squared z-statistic in the same manner that I have described the sum of Cochran-Armitage test statistics.

7.2 Future directions

I anticipate that in the *omics*-era, where the advance of technology and the public awareness of genetic studies is expected to result in an overwhelming amount of generated data as well as typed individuals world-wide, pathway-driven analysis methodologies of GWAS will play a pivotal role.

As depicted in Figure 36, pathway and gene-driven analysis of GWAS could act as a guide to define the broad pathway functional setting and to then prioritize genes for further functional studies. By prioritizing the genes at the genetic level and then validating the findings at the gene-expression level, researchers would maximize the chances of pinning down functionally relevant genes to be knocked-down on animal models. Genes can be prioritized based on their strength of pathway association, statistical evidence of individual gene expression, and statistical evidence of coordinated altered gene expression of the pathways that they are involved in. An open-source solution to studies that do not have SNPs and gene-expression data on the same individuals would be the ArrayExpress Archive (AEA)[44] of the European Bioinformatics Institute (EBI). AEA is a database of 19,923 gene-expression experiments comprising 554,961 assays on 18419 conditions that can be downloaded and analyzed with any of the pathway-based gene expression approaches that are numerous and they are very well presented in a recent review.[46] On the linking of genetic and gene-expression data, a pathway approach would most probably take the form of a "network-based" approach. Researchers that have tried to combine information across these two genomic levels have so far attempted to identify the SNPs that affect gene expression and then build *de novo* gene-networks based on the correlation of differentially expressed genes.[233] These networks are subsequently tested for "enrichment" in biological processes and functions, using any of the discussed methods.

Metabonomic or proteomic analysis on the end product or intermediate products of a pathway could verify the pathway effect on disease susceptibility. The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. Many data fields are hyperlinked to other databases (KEGG, PubChem, MetaCyc, ChEBI, PDB, Swiss-Prot, and GenBank), so the end-user can actually see which metabolites act in which pathways.

Another direction could be the prediction of novel, not-discovered pathways, through capitalizing the cumulative evidence of genes that show association at multiple *omics* data.

Rare variants from sequencing studies would be expected to be grouped together in order to reveal their joint effect at a population rather than an individual's level. Therefore a pathway-driven approach could also be a natural way for researchers to look at rare variants, as rare variants could be fitted in the same pathway-based model and each of them explain different portion of the genetic variation – with an effect on disease - in the population.

On a similar note, the progress in next generation sequencing, epigenetic and proteomics studies open new exciting routes but yet poses new challenges on how pathway-based analyses will be able to extract and aggregate the knowledge across the different levels of genetic information.

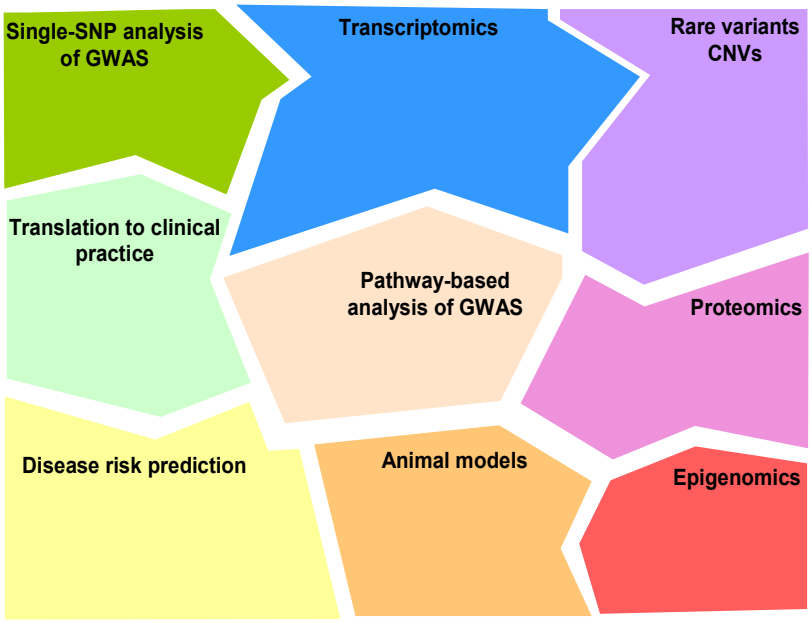


Figure 36. Pathway-based analyses of GWAS in the perspective of an integrative genomics approach

7.3 Closing remark

Pathway and gene-based analyses of GWAS differ to a great extend from each other, they have several technical challenges yet to tackle and as this concept is still in its infancy, no best pathway-based practice exists as yet. The same applies to the work presented in this thesis. Nevertheless, if I were to place my proposed methodology in the setting of relevant approaches, I would say that it has succeeded in three respects: 1) it addresses the majority of technical challenges and 2) it goes beyond other pathway-driven methodologies in that it identifies pathway-disease

associations and then identifies the genes within the top associated pathways that are driving the signal of multiple pathway associations. The latter is of great importance, as it has not been given enough attention to date and to my knowledge this methodology is the first to present such an integrated pathway and gene-based approach. 3) It has been applied and has produced replicated results in several GWAS datasets, at a time when the majority of similar methods have presented results on a single study only or simulated data.

As a general comment, pathway-based analysis strategies of GWAS are powerful new tools to mine deeper the plethora of generated data and if their main limitations are addressed, they are expected to play an instrumental role in the forthcoming years. It is crucial for researchers to realize that some common standards should be adopted to allow for comparison of results between studies. Sceptical and careful interpretation of pathway association results should be adopted. As our knowledge of the gene complex interplay is incomplete, any pathway-disease association scenario could be plausible. Validation, replication and follow up of results should be considered essential. As nowadays validation on an independent dataset is essential for a single SNP to be considered truly associated, validation should be a natural step for a pathway to be denoted associated. Researchers are also advised to cite database versions, chip annotations, human genome reference sets and any other types of annotations used to ensure that results are fully reproducible. Pathway-based approaches have a long way to go. Their initial application has delivered promising results and they are expected to play central role in the integrative genomics era. I hope that the work presented in this thesis will contribute to the evolution of this approach and I aspire to contribute more in my future work to the application of pathway-driven methods to understand human disease.

Bibliography

1. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299-1320

2. Ziegler A, Konig IR, Thompson JR (2008) Biostatistical aspects of genome-wide association studies. Biom J 50:8-28

3. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341-1345

4. O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, et al. (2008) Identification of loci associated with schizophrenia by genome-wide association and follow-up. Nature Genet 40:1053-1055

5. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-U1087

6. Eeles RA, Kote-Jarai Z, Giles GG, Al Olama AA, Guy M, Jugurnauth SK, Mulholland S, et al. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. Nature Genet 40:316-321

7. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18:767-778

8. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis - A genomewide study. N Engl J Med 357:1199-1209

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

10. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362-9367

11. Kruglyak L (2008) The road to genome-wide association studies. Nat Rev Genet 9:314-318

12. Ioannidis JPA, Thomas G, Daly MJ (2009) GENOME-WIDE ASSOCIATION STUDIES Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10:318-329

13. Lango H, Weedon MN (2008) What will whole genome searches for susceptibility genes for common complex disease offer to clinical practice? J Intern Med 263:16-27

14. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369

15. Xavier RJ, Rioux JD (2008) Genome-wide association studies: a new window into immune-mediated diseases. Nat Rev Immunol 8:631-643

16. Hoh J, Ott J (2003) Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet 4:701-709

17. McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet 17:R156-165

18. Hunter DJ, Kraft P (2007) Drinking from the fire hose--statistical issues in genomewide association studies. N Engl J Med 357:436-439

19. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832-838

20. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781-791

21. Zheng SL, Sun JL, Wiklund F, Smith S, Stattin P, Li G, Adami HO, Hsu FC, Zhu Y, Balter K, Kader AK, Turner AR, Liu WN, Bleecker ER, Meyers DA, Duggan D, Carpten JD, Chang BL, Isaacs WB, Xu JF, Gronberg H (2008) Cumulative association of five genetic variants with prostate cancer. N Engl J Med 358:910-919

22. Pharoah PD, Tyrer J, Dunning AM, Easton DF, Ponder BA (2007) Association between common variation in 120 candidate genes and breast cancer risk. PLoS Genet 3:e42

23. Clayton D, Chapman J, Cooper J (2004) Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415-428

24. Servin B, Stephens M (2007) Imputation-based analysis of association studies: Candidate regions and quantitative traits. PLoS Genet 3:1296-1308

25. Schaid DJ (2004) Evaluating associations of haplotypes with traits. Genet Epidemiol 27:348-364

26. Zhao LP, Li SYS, Khalid N (2003) A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231-1250

27. Lake SL, Lyon H, Tantisira K, Silverman EK, Weiss ST, Laird NM, Schaid DJ (2003) Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Human Heredity 55:56-65

28. Tanck MWT, Klerkx A, Jukema JW, De Knijff P, Kastelein JJP, Zwinderman AH (2003) Estimation of multilocus haplotype effects using weighted penalised log-likelihood: Analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. Ann Hum Genet 67:175-184

29. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463-2468

30. Marchini J, Donnelly P, Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genet 37:413-417

31. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95-108

32. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, Henley JR, Rocca WA, Ahlskog JE, Maraganore DM (2007) A genomic pathway approach to a complex disease: Axon guidance and parkinson disease. PLoS Genet 3:984-995

33. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JR, Russell RK, Sleiman PMA, et al. (2009) Diverse Genome-wide Association Studies Associate the IL12/IL23 Pathway with Crohn Disease. Am J Hum Genet 84:399-405

34. Yu K, Li QZ, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N (2009) Pathway Analysis by Adaptive Combination of P-Values. Genetic Epidemiology 33:700-709

35. Low YL, Li Y, Humphreys K, Thalamuthu A, Li Y, Darabi H, WedrÃ©n S, Bonnard C, Czene K, Iles MM, Heikkinen T, AittomÃ¤ki K, Blomqvist C, Nevanlinna H, Hall P, Liu ET, Liu J (2010) Multi-Variant Pathway Association Analysis Reveals the Importance of Genetic Determinants of Estrogen Metabolism in Breast and Endometrial Cancer Susceptibility. PLoS Genet 6:e1001012

36. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L (2010) Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. The American Journal of Human Genetics 86:860-871

37. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, Sklar P, Owen MJ, O'Donovan MC, Craddock N (2009) Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. The American Journal of Human Genetics 85:13-24

38. Elbers C (2009) Using Genome-Wide Pathway Analysis to Unravel the Etiology of Complex Diseases. Genetic Epidemiology 33:419-431

39. Baranzini S (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Human Molecular Genetics 18:2078-2090

40. Torkamani A, Topol EJ, Schork NJ (2008) Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 92:265-272

41. Askland K, Read C, Moore J (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet 125:63-79

42. Chen L, Zhang L, Zhao Y, Xu L, Shang Y, Wang Q, Li W, Wang H, Li X (2009) Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. Bioinformatics 25:237-242

43. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M (2009) Gene and pathway-based second-wave analysis of genome-wide association studies. Eur J Hum Genet 18:111-117

44. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, Corvin A (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics 25:2762-2763

45. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11:843-854

46. Nam D, Kim SY (2008) Gene-set approach for expression pattern analysis. Brief Bioinform 9:189-197

47. Kraft P, Raychaudhuri S (2009) Complex Diseases, Complex Genes Keeping Pathways on the Right Track. Epidemiology 20:508-511

48. Fisher R (1932) Statistical Methods for Research Workers. Oliver and Boyd, London

49. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BMJ, Kappos L, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet 18:2078-2090

50. Sasieni PD (1997) From genotypes to genes: Doubling the sample size. Biometrics 53:1253-1261

51. Freidlin B, Zheng G, Li ZH, Gastwirth JL (2002) Trend tests for case-control studies of genetic markers: Power, sample size and robustness. Human Heredity 53:146-152

52. Fan RZ, Knapp M (2003) Genome association studies of complex diseases by case-control designs. Am J Hum Genet 72:850-868

53. Gauderman WJ, Cassandra M, Frank G, David VC (2007) Testing association between disease and multiple SNPs in a candidate gene. Genetic Epidemiology 31:383-395

54. Li M, Wang K, Grant SFA, Hakonarson H, Li C (2009) ATOM: a powerful gene-based association test by combining optimally weighted markers. Bioinformatics 25:497-503

55. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN (2005) Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 76:780-793

56. Wang T, Elston RC (2007) Improved power by use of a weighted score test for linkage disequilibrium mapping. Am J Hum Genet 80:353-360

57. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A (2009) Association Tests Using Kernel-Based Measures of Multi-Locus Genotype Similarity Between Individuals. Genetic Epidemiology 34:213-221

58. Wessel J, Schork NJ (2006) Generalized Genomic Distance Based Regression Methodology for Multilocus Association Analysis. 79:792-806

59. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X (2010) Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. The American Journal of Human Genetics 86:929-942

60. Holden M, Deng SW, Wojnowski L, Kulle B (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics 24:2784-2785

61. Wang K, Li MY, Bucan M (2007) Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 81:1278-1283

62. Elbers CC, van Eijk KR, Franke L, Mulder F, van der Schouw YT, Wijmenga C, Onland-Moret NC (2009) Using Genome-Wide Pathway Analysis to Unravel the Etiology of Complex Diseases. Genet Epidemiol 33:419-431

63. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78:1011-1025

64. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Research 33:W741-W748

65. Chang JT, Nevins JR (2006) GATHER: a systems approach to interpreting genomic signatures. Bioinformatics 22:2926-2933

66. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:16

67. Mi HY, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Research 35:D247-D252

68. Gelman A (2003) Bayesian Data Analysis. CRC Press LLC, Boca Raton, FL, USA

69. Zeeberg BR, Feng WM, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4:8

70. Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. pp 578 - 580

71. Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21:3448-3449

72. Beissbarth T, Speed TP (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 20:1464-1465

73. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, Muller R, Meese E (2007) GeneTrail - advanced gene set enrichment analysis. pp 186 - 192

74. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57

75. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu GM, Birney E, Stein L (2007) Reactome: a knowledge base of biologic pathways and processes. Genome Biol 8:13

76. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:48

77. Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S (2010) A Versatile Gene-Based Test for Genome-wide Association Studies. The American Journal of Human Genetics 87:139-145

78. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28:27 - 30

79. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G (2000) Gene Ontology: tool for the unification of biology. 25:25 - 29

80. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ãzn, Anwar N, Schultz N, Bader GD, Sander C (2010) Pathway Commons, a web resource for biological pathway data.

81. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD (2005) Computational prediction of human metabolic pathways from the complete human genome. Genome Biol 6:17

82. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK (2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet 4:15

83. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86:6-22

84. Perry JRB, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, Frayling TM (2009) Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach. Diabetes 58:1463-1467

85. Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, Vyse TJ, Rioux JD (2008) Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. PLoS Genet 4:e1000024

86. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature Genet 42:1118-U1121

87. Craddock N, O'Donovan MC, Owen MJ (2008) Genome-wide association studies in psychiatry: lessons from early studies of non-psychiatric and psychiatric phenotypes. Mol Psychiatr 13:649-653

88. Newport M, Sirugo G, Lyons E, Vannberg F, Hill AVS, Bradbury LA, Farrar C, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nature Genet 39:1329-1337

89. Li YH, Rowland C, Xiromerisiou G, Lagier RJ, Schrodi SJ, Dradiotis E, Ross D, Bui N, Catanese J, Aggelakis K, Grupe A, Hadjigeorgiou G (2008) Neither Replication nor Simulation Supports a Role for the Axon Guidance Pathway in the Genetics of Parkinson's Disease. PLoS One 3:7

90. Peng Q, Zhao J, Xue F (2010) A gene-based method for detecting gene-gene co-association in a case-control association study. Eur J Hum Genet 18:582-587

91. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nature Genet 40:955-962

92. Li Y, Rowland C, Xiromerisiou G, Lagier RJ, Schrodi SJ, Dradiotis E, Ross D, Bui N, Catanese J, Aggelakis K, Grupe A, Hadjigeorgiou G (2008) Neither Replication nor Simulation Supports a Role for the Axon Guidance Pathway in the Genetics of Parkinson's Disease. PLoS ONE 3:e2707

93. Fisher SA, Tremelling M, Anderson CA, Gwilliam R, Bumpstead S, Prescott NJ, Nimmo ER, et al. (2008) Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. Nat Genet 40:710-712

94. Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, Gyapong J, Sirugo G, et al. (2010) Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nature Genet 42:739-741

95. Davila S, Wright VJ, Khor CC, Sim KS, Binder A, Breunis WB, Inwald D, Nadel S, Betts H, Carrol ED, de Groot R, Hermans PWM, Hazelzet J, Emonts M, Lim CC, Kuijpers TW, Martinon-Torres F, Salas A, Zenz W, Levin M, Hibberd ML (2010) Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nature Genet 42:772-U763

96. Hirschhorn JN (2009) Genomewide Association Studies - Illuminating Biologic Pathways. N Engl J Med 360:1699-1701

97. Dinu V, Miller PL, Zhao HY (2007) Evidence for association between multiple complement pathway genes and AMD. Genet Epidemiol 31:224-237

98. McGovern DPB, Rotter JI, Mei L, Haritunians T, Landers C, Derkowski C, Dutridge D, Dubinsky M, Ippoliti A, Vasiliauskas E, Mengesha E, King L, Pressman S, Targan SR, Taylor KD (2009) Genetic Epistasis of IL23/IL17 Pathway Genes in Crohn's Disease. Inflamm Bowel Dis 15:883-889

99. Abraham C, Cho J (2009) Interleukin-23/Th17 Pathways and Inflammatory Bowel Disease. Inflamm Bowel Dis 15:1090-1100

100. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881-885

101. Jin T (2008) The WNT signalling pathway and diabetes mellitus. Diabetologia 51:1771-1780

102. Bu GJ (2009) Apolipoprotein E and its receptors in Alzheimer's disease: pathways, pathogenesis and therapy. Nat Rev Neurosci 10:333-344

103. Yang X, Deignan JL, Qi H, Zhu J, Qian S, Zhong J, Torosyan G, et al. (2009) Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. Nat Genet 41:415-423

104. Lettre G, Rioux JD (2008) Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 17:R116-121

105. Gregersen PK (2005) Pathways to gene identification in rheumatoid arthritis: PTPN22 and beyond. Immunol Rev 204:74-86

106. Oksenberg JR, Baranzini SE, Sawcer S, Hauser SL (2008) The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. Nat Rev Genet 9:516-526

107. Ounissi-Benkalha H, Polychronakos C (2008) The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. Trends Mol Med 14:268-275

108. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, Romanos J, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nature Genet 40:395-402

109. Wang K, Zhang HT, Ma DQ, Bucan M, Glessner JT, Abrahams BS, Salyakina D, et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459:528-533

110. Law AJ (2008) Genes associated with schizophrenia and bipolar disorder converge on specific intracellular signaling pathways: The NRG1/phosphatidylinositol pathway as an emerging interface. Biol Psychiatry 63:486

111. Martinowich K, Schloesser RJ, Manji HK (2009) Bipolar disorder: from genes to behavior pathways. J Clin Invest 119:726-736

112. Hardy J, Lewis P, Revesz T, Lees A, Paisan-Ruiz C (2009) The genetics of Parkinson's syndromes: a critical review. Current Opinion in Genetics & Development 19:254-265

113. Du H, Guo L, Fang F, Chen D, Sosunov AA, McKhann GM, Yan YL, Wang CY, Zhang H, Molkentin JD, Gunn-Moore FJ, Vonsattel JP, Arancio O, Chen JX, Du Yan S (2008) Cyclophilin D deficiency attenuates mitochondrial and neuronal perturbation and ameliorates learning and memory in Alzheimer's disease. Nat Med 14:1097-1105

114. Smith U (2007) TCF7L2 and type 2 diabetes—we WNT to know. Diabetologia 50:5-7

115. Teicher BA, Bagley RG, Yagui-Beltrán A, He B, Jablons DM (2009) Wnt Signaling in Cancer: From Embryogenesis to Stem Cell Self-Renewal. Stem Cells and Cancer. Humana Press, pp 37-55

116. Polakis P (2000) Wnt signaling and cancer. Genes Dev 14:1837-1851

117. Hemminki K, Li X, Sundquist J, Sundquist K (2010) Risk of Cancer Following Hospitalization for Type 2 Diabetes. Oncologist 15:548-555

118. Hjartaker A, Langseth H, Weiderpass E (2008) Obesity and diabetes epidemics: cancer repercussions. Innovative Endocrinology of Cancer. Springer-Verlag Berlin, Berlin, pp 72-93

119. Ogunleye AA, Ogston SA, Morris AD, Evans JMM (2009) A cohort study of the risk of cancer associated with type 2 diabetes. Br J Cancer 101:1199-1201

120. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, Barzi F, Woodward M (2005) Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. Br J Cancer 92:2076-2083

121. Giouleme O, Diamantidis MD, Katsaros MG (2011) Is diabetes a causal agent for colorectal cancer? Pathophysiological and molecular mechanisms. World Journal of Gastroenterology 17:444-448

122. van der Poll T, Opal SM (2008) Host-pathogen interactions in sepsis. Lancet Infect Dis 8:32-43

123. Kwiatkowski DP, Luoni G (2006) Host genetic factors in resistance and susceptibility to malaria. Parassitologia 48:450-467

124. Hill AV (2006) Aspects of genetic susceptibility to human infectious diseases. Annu Rev Genet 40:469-486

125. Blackwell JM, Searle S, Mohamed H, White JK (2003) Divalent cation transport and susceptibility to infectious and autoimmune disease: continuation of the Ity/Lsh/Bcg/Nramp1/Slc11a1 gene story. Immunol Lett 85:197-203

126. Bach JF (2002) The effect of infections on susceptibility to autoimmune and allergic diseases. N Engl J Med 347:911-920

127. Brandtzaeg P, van Deuren M (2002) Current concepts in the role of the host response in Neisseria meningitidis septic shock. Curr Opin Infect Dis 15:247-252

128. Casanova JL, Abel L (2002) Genetic dissection of immunity to mycobacteria: the human model. Annu Rev Immunol 20:581-620

129. McInnes IB, Schett G (2007) Cytokines in the pathogenesis of rheumatoid arthritis. Nat Rev Immunol 7:429-442

130. Xavier RJ, Podolsky DK (2007) Unravelling the pathogenesis of inflammatory bowel disease. Nature 448:427-434

131. Ounissi-Benkalha H, Polychronakos C (2008) The molecular genetics of type 1 diabetes: new genes and emerging mechanisms. Trends Mol Med 14:268-275

132. Atzeni F, Doria A, Carrabba M, Turiel M, Sarzi-Puttini P (2007) Potential target of infliximab in autoimmune and inflammatory diseases. Autoimmun Rev 6:529-536

133. Lin J, Ziring D, Desai S, Kim S, Wong M, Korin Y, Braun J, Reed E, Gjertson D, Singh RR (2008) TNFalpha blockade in human diseases: an overview of efficacy and safety. Clin Immunol 126:13-30

134. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678

135. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarvelin MR, Freimer NB, Peltonen L (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41:35-46

136. Ohagan A, Leonard T (1976) Bayes Estimation Subject to Uncertainty About Parameter Constraints. Biometrika 63:201-203

137. Azzalini A, Capitanio A (1999) Statistical applications of the multivariate skew-normal distribution. Journal of the Royal Statistical Society 61:579-602

138. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4:e1000130

139. Mason SJ, Graham NE (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J R Meteorol Soc 128:2145–2166

140. Howe K, Bateman A, Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18:1546-1547

141. Hahn CG, Friedman E (1999) Abnormalities in protein kinase C signaling and the pathophysiology of bipolar disorder. Bipolar Disord 1:81-86

142. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, et al. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 40:1216-1223

143. Barton A, Thomson W, Ke X, Eyre S, Hinks A, Bowes J, Plant D, Gibbons LJ, Wilson AG, Bax DE, Morgan AW, Emery P, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Worthington J (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. Nat Genet 40:1156-1159

144. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955-962

145. Jono H, Lim JH, Chen LF, Xu H, Trompouki E, Pan ZK, Mosialos G, Li JD (2004) NF-kappaB is essential for induction of CYLD, the negative regulator of NF-kappaB: evidence for a novel inducible autoregulatory feedback pathway. J Biol Chem 279:36171-36174

146. Lin SC, Chung JY, Lamothe B, Rajashankar K, Lu M, Lo YC, Lam AY, Darnay BG, Wu H (2008) Molecular basis for the unique deubiquitinating activity of the NF-kappaB inhibitor A20. J Mol Biol 376:526-540

147. Zhang J, Stirling B, Temmerman ST, Ma CA, Fuss IJ, Derry JM, Jain A (2006) Impaired regulation of NF-kappaB and increased susceptibility to colitis-associated tumorigenesis in CYLD-deficient mice. J Clin Invest 116:3042-3049

148. Silverman GJ, Carson DA (2003) Roles of B cells in rheumatoid arthritis. Arthritis Res Ther 5 Suppl 4:S1-6

149. Sigurdsson S, Padyukov L, Kurreeman FA, Liljedahl U, Wiman AC, Alfredsson L, Toes R, Ronnelid J, Klareskog L, Huizinga TW, Alm G, Syvanen AC, Ronnblom L (2007) Association of a haplotype in the promoter region of the interferon regulatory factor 5 gene with rheumatoid arthritis. Arthritis Rheum 56:2202-2210

150. Thomson W, Barton A, Ke X, Eyre S, Hinks A, Bowes J, Donn R, Symmons D, Hider S, Bruce IN, Wilson AG, Marinou I, Morgan A, Emery P, Carter A, Steer S, Hocking L, Reid DM, Wordsworth P, Harrison P, Strachan D, Worthington J (2007) Rheumatoid arthritis association at 6q23. Nat Genet 39:1431-1433

151. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, Pe'er I, et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 39:1477-1482

152. Burgner D, Jamieson SE, Blackwell JM (2006) Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? Lancet Infect Dis 6:653-663

153. Gregersen PK, Behrens TW (2006) Genetics of autoimmune diseases - disorders of immune homeostasis. Nat Rev Genet 7:917-928

154. Martinez A, Varade J, Marquez A, Cenit MC, Espino L, Perdigones N, Santiago JL, Fernandez-Arquero M, de la Calle H, Arroyo R, Mendoza JL, Fernandez-Gutierrez B, de la Concha EG, Urcelay E (2008) Association of the STAT4 gene with increased susceptibility for some immune-mediated diseases. Arthritis Rheum 58:2598-2602

155. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, Macneil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE (2008) Variations in DNA elucidate molecular networks that cause disease. Nature

156. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, Roos C, Hirschhorn JN, Berglund G, Hedblad B, Groop L, Altshuler DM, Newton-Cheh C, Orho-Melander M (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 358:1240-1249

157. Wang K, Li M, Bucan M (2007) Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 81

158. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423-428

159. Benfey PN, Mitchell-Olds T (2008) From genotype to phenotype: systems biology meets natural variation. Science 320:495-497

160. Janssens AC, van Duijn CM (2008) Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet 17:R166-173

161. Wray NR, Goddard ME, Visscher PM (2007) Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res 17:1520-1528

162. Jakkula E, Rehnstrom K, Varilo T, Pietilainen OPH, Paunio T, Pedersen NL, Defaire U, Jarvelin MR, Saharinen J, Freimer N, Ripatti S, Purcell S, Collins A, Daly MJ, Palotie A, Peltonen L (2008) The Genome-wide Patterns of Variation Expose Significant Substructure in a Founder Population. Am J Hum Genet 83:787-794

163. Noble JA, Martin A, Valdes AM, Lane JA, Galgani A, Petrone A, Lorini R, Pozzilli P, Buzzetti R, Erlich HA (2008) Type 1 diabetes risk for human leukocyte antigen (HLA)-DR3 haplotypes depends on genotypic context: Association of DPB1 and HLA class I loci among DR3-and DR4-matched Italian patients and controls. Hum Immunol 69:291-300

164. Buzzetti R, Galgani A, Petrone A, Buono MLD, Erlich HA, Bugawan TL, Lorini R, Meschi F, Multari G, Pozzilli P, Locatelli M, Bottazzo G, Di Mario U (2004) Genetic prediction of type 1 diabetes in a poputation with tow frequency of HLA risk genotypes and low incidence of the disease (the DIABFIN study). Diabetes-Metab Res Rev 20:137-143

165. Skyler JS (2007) Prediction and prevention of type 1 diabetes: Progress, problems, and prospects. Clin Pharmacol Ther 81:768-771

166. Costenbader KH, Feskanich D, Mandl LA, Karlson EW (2006) Smoking intensity, duration, and cessation, and the risk of rheumatoid arthritis in women. Am J Med 119:503-511

167. Katschinski B, Logan RFA, Edmond M, Langman MJS (1988) Smoking and Sugar Intake Are Separate but Interactive Risk-Factors in Crohns-Disease. Gut 29:1202-1206

168. Breitling LP, Steyerberg EW, Brenner H (2009) The Novel "Genomic Pathway Approach" to Complex Diseases A Reason for (Over-)Optimism? Epidemiology 20:500-507

169. Sauerbrei W (1999) The use of resampling methods to simplify regression models in medical statistics. J R Stat Soc Ser C-Appl Stat 48:313-329

170. Meinshausen N, Buhlman P (2010) Stability Selection. Journal of the Royal Statistical Society 74:417–473

171. Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A-L, Jarvelin M-R, Balding D, Coin L, Levin M (2009) Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases. PLoS ONE 4:e8068

172. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584

173. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

174. Merida I, Avila-Flores A, Merino E (2008) Diacylglycerol kinases: at the hub of cell signalling. Biochem J 409:1-18

175. Serezani CH, Ballinger MN, Aronoff DM, Peters-Golden M (2008) Cyclic AMP - Master regulator of innate immune cell function. Am J Respir Cell Mol Biol 39:127-132

176. Berridge MJ, Lipp P, Bootman MD (2000) The versatility and universality of calcium signalling. Nat Rev Mol Cell Biol 1:11-21

177. Clapham DE (2007) Calcium signaling. Cell 131:1047-1058

178. Kholodenko BN (2006) Cell-signalling dynamics in time and space. Nat Rev Mol Cell Biol 7:165-176

179. Rozengurt E (2007) Mitogenic signaling pathways induced by G protein-coupled receptors. J Cell Physiol 213:589-602

180. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li YH, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nature Genet 42:508-U556

181. de Crombrugghe B, Lefebvre V, Behringer RR, Bi WM, Murakami S, Huang WD (2000) Transcriptional mechanisms of chondrocyte differentiation. Matrix Biol 19:389-394

182. Deshmukh K, Sawyer BD (1977) Synthesis of Collagen by Chondrocytes in Suspension Culture - Modulation by Calcium, 3'-5'-Cyclic Amp, and Prostaglandins - (Arthritis Protein Synthesis Invitro). Proc Natl Acad Sci U S A 74:3864-3868

183. Allen ME, Young SP, Michell RH, Bacon PA (1995) Altered T-Lymphocyte Signaling in Rheumatoid-Arthritis. Eur J Immunol 25:1547-1554

184. Souness JE, Aldous D, Sargent C (2000) Immunosuppressive and anti-inflammatory effects of cyclic AMP phosphodiesterase (PDE) type 4 inhibitors. Immunopharmacology 47:127-162

185. Zha YY, Marks R, Ho AW, Peterson AC, Janardhan S, Brown I, Praveen K, Stang S, Stone JC, Gajewski TF (2006) T cell anergy is reversed by active Ras and is regulated by diacylglycerol kinase-alpha. Nat Immunol 7:1166-1173

186. Gordon JB, Kahn AM, Burns JC (2009) When Children With Kawasaki Disease Grow Up Myocardial and Vascular Complications in Adulthood. J Am Coll Cardiol 54:1911-1920

187. Ravekes WJ, Colan SD, Gauvreau K, Baker AL, Sundel RP, van der Velde ME, Fulton DR, Newburger JW (2001) Aortic root dilation in Kawasaki disease. Am J Cardiol 87:919-+

188. Newburger JW, Takahashi M, Burns JC, Beiser AS, Chung KJ, Duffy CE, Glode MP, Mason WH, Reddy V, Sanders SP, Shulman ST, Wiggins JW, Hicks RV, Fulton DR, Lewis AB, Leung DYM, Colton T, Rosen FS, Melish ME (1986) The Treatment of Kawasaki Syndrome with Intravenous Gamma-Globulin. N Engl J Med 315:341-347

189. de Zorzi A, Colan SD, Gauvreau K, Baker AL, Sundel RP, Newburger JW (1998) Coronary artery dimensions may be misclassified as normal in Kawasaki disease. J Pediatr 133:254-258

190. McCrindle BW, Li JS, Minich LL, Colan SD, Atz AM, Takahashi M, Vetter VL, Gersony WM, Mitchell PD, Newburger JW (2007) Coronary artery involvement in children with Kawasaki disease - Risk factors from analysis of serial normalized measurements. Circulation 116:174-179

191. Amano S, Hazama F, Hamashima Y (1979) Pathology of Kawasaki Disease .2. Distribution and Incidence of the Vascular-Lesions. Jpn Circ J-Engl Ed 43:741-748

192. Amano S, Hazama F, Hamashima Y (1979) Pathology of Kawasaki Disease .1. Pathology and Morphogenesis of the Vascular Changes. Jpn Circ J-Engl Ed 43:633-643

193. Rowley AH, Baker SC, Orenstein JM, Shulman ST (2008) Opinion - Searching for the cause of Kawasaki disease - cytoplasmic inclusion bodies provide new insight. Nat Rev Microbiol 6:394-401

194. Burgner D, Davila S, Breunis WB, Ng SB, Li Y, Bonnard C, Ling L, Wright VJ, Thalamuthu A, Odam M, Shimizu C, Burns JC, Levin M, Kuijpers TW, Hibberd ML (2009) A Genome-Wide Association Study Identifies Novel and Functionally Related Susceptibility Loci for Kawasaki Disease. PLoS Genet 5:15

195. Burns JC, Shimizu C, Gonzalez E, Kulkarni H, Patel S, Shike H, Sundel RS, Newburger JW, Ahuja SK (2005) Genetic variations in the receptor-ligand pair CCR5 and CCL3L1 are important determinants of susceptibility to Kawasaki disease. J Infect Dis 192:344-349

196. Burns JC, Shimizu C, Shike H, Newburger JW, Sundel RP, Baker AL, Matsubara T, Ishikawa Y, Brophy VA, Cheng S, Grow MA, Steiner LL, Kono N, Cantor R (2005) Family-based association analysis implicates IL-4 in susceptibility to Kawasaki disease. Genes Immun 6:438-444

197. Onouchi Y, Gunji T, Burns JC, Shimizu C, Newburger JW, Yashiro M, Nakamura Y, et al. (2008) ITPKC functional polymorphism associated with Kawasaki disease susceptibility and formation of coronary artery aneurysms. Nature Genet 40:35-42

198. Burns JC, Glode MP (2004) Kawasaki syndrome. Lancet 364:533-544

199. Onouchi Y (2009) Molecular genetics of Kawasaki disease. Pediatr Res 65:46R-54R

200. Brown TJ, Crawford SE, Cornwall ML, Garcia F, Shulman ST, Rowley AH (2001) CD8 T lymphocytes and macrophages infiltrate coronary artery aneurysms in acute Kawasaki disease. J Infect Dis 184:940-943

201. Franco A, Shimizu C, Tremoulet AH, Burns JC (2010) Memory T-cells and characterization of peripheral T-cell clones in acute Kawasaki disease. Autoimmunity 43:317-324

202. Ruiz-Ortega M, Rodriguez-Vita J, Sanchez-Lopez E, Carvajal G, Egido J (2007) TGF-beta signaling in vascular fibrosis. Cardiovasc Res 74:196-206

203. Clark-Greuel JN, Connolly JM, Sorichillo E, Narula NR, Rapoport HS, Mohler ER, Gorman JH, Gorman RC, Levy RJ (2007) Transforming growth factor-beta 1 mechanisms in aortic valve calcification: Increased alkaline phosphatase and related events. Ann Thorac Surg 83:946-953

204. Loeys BL, Chen JJ, Neptune ER, Judge DP, Podowski M, Holm T, Meyers J, Leitch CC, Katsanis N, Sharifi N, Xu FL, Myers LA, Spevak PJ, Cameron DE, De Backer J, Hellemans J, Chen Y, Davis EC, Webb CL, Kress W, Coucke P, Rifkin DB, De Paepe AM, Dietz HC (2005) A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. Nature Genet 37:275-281

205. Neptune ER, Frischmeyer PA, Arking DE, Myers L, Bunton TE, Gayraud B, Ramirez F, Sakai LY, Dietz HC (2003) Dysregulation of TGF-beta activation contributes to pathogenesis in Marfan syndrome. Nature Genet 33:407-411

206. Steinberg MH (2009) Genetic Etiologies for Phenotypic Diversity in Sickle Cell Anemia. TheScientificWorldJOURNAL 9:46-67

207. Sebastiani P, Ramoni MF, Nolan V, Baldwin CT, Steinberg MH (2005) Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nature Genet 37:435-440

208. Burgner D, Davila S, Breunis WB, Ng SB, Li Y, Bonnard C, Ling L, Wright VJ, Thalamuthu A, Odam M, Shimizu C, Burns JC, Levin M, Kuijpers TW, Hibberd ML (2009) A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. PLoS Genet 5:e1000319

209. Popper SJ, Shimizu C, Shike H, Kanegaye JT, Newburger JW, Sundel RP, Brown PO, Burns JC, Relman DA (2007) Gene-expression patterns reveal underlying biological processes in Kawasaki disease. Genome Biol 8:R261

210. Disease RCoK (1984) Report of subcommittee on standardization of diagnostic criteria and reporting of coronary artery lesions in Kawasaki disease. Tokyo, Japan: Ministry of Health and Welfare

211. McCrindle BW, Li JS, Minich LL, Colan SD, Atz AM, Takahashi M, Vetter VL, Gersony WM, Mitchell PD, Newburger JW (2007) Coronary artery involvement in children with Kawasaki disease: risk factors from analysis of serial normalized measurements. Circulation 116:174-179

212. Ravekes WJ, Colan SD, Gauvreau K, Baker AL, Sundel RP, van der Velde ME, Fulton DR, Newburger JW (2001) Aortic root dilation in Kawasaki disease. The American journal of cardiology 87:919-922

213. Fury W, Tremoulet AH, Watson VE, Best BM, Shimizu C, Hamilton J, Kanegaye JT, Wei Y, Kao C, Mellis S, Lin C, Burns JC (2010) Transcript abundance patterns in Kawasaki disease patients with intravenous immunoglobulin resistance. Hum Immunol 71:865-873

214. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission Test for Linkage Disequilibrium - the Insulin Gene Region and Insulin-Dependent Diabetes-Mellitus (Iddm). Am J Hum Genet 52:506-516

215. Burns JC, Shimizu C, Gonzalez E, Kulkarni H, Patel S, Shike H, Sundel RS, Newburger JW, Ahuja SK (2005) Genetic variations in the receptor-ligand pair CCR5 and CCL3L1 are important determinants of susceptibility to Kawasaki disease. J Infect Dis 192:344-349

216. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241-251

217. Camenisch TD, Molin DG, Person A, Runyan RB, Gittenberger-de Groot AC, McDonald JA, Klewer SE (2002) Temporal and distinct TGFbeta ligand requirements during mouse and avian endocardial cushion morphogenesis. Developmental biology 248:170-181

218. Bartram U, Molin DG, Wisse LJ, Mohamad A, Sanford LP, Doetschman T, Speer CP, Poelmann RE, Gittenberger-de Groot AC (2001) Double-outlet right ventricle and overriding tricuspid valve reflect disturbances of looping, myocardialization, endocardial cushion differentiation, and apoptosis in TGF-beta(2)-knockout mice. Circulation 103:2745-2752

219. Sanford LP, Ormsby I, Gittenberger-de Groot AC, Sariola H, Friedman R, Boivin GP, Cardell EL, Doetschman T (1997) TGFbeta2 knockout mice have multiple developmental defects that are non-overlapping with other TGFbeta knockout phenotypes. Development 124:2659-2670

220. Azhar M, Schultz Jel J, Grupp I, Dorn GW, 2nd, Meneton P, Molin DG, Gittenberger-de Groot AC, Doetschman T (2003) Transforming growth factor beta in cardiovascular development and function. Cytokine Growth Factor Rev 14:391-407

221. Huang HP, Lai YC, Tsai IJ, Chen SY, Cheng CH, Tsau YK (2008) Nephromegaly in children with Kawasaki disease: new supporting evidence for diagnosis and its possible mechanism. Pediatric research 63:207-210

222. Kimura J, Takada H, Nomura A, Ohno T, Mizuno Y, Saito M, Kusuhara K, Hara T (2004) Th1 and Th2 cytokine production is suppressed at the level of transcriptional regulation in Kawasaki disease. Clinical and experimental immunology 137:444-449

223. Matsubara T, Umezawa Y, Tsuru S, Motohashi T, Yabuta K, Furukawa S (1997) Decrease in the concentrations of transforming growth factor-beta 1 in the sera of patients with Kawasaki disease. Scandinavian journal of rheumatology 26:314-317

224. Loeys BL, Schwarze U, Holm T, Callewaert BL, Thomas GH, Pannu H, De Backer JF, Oswald GL, Symoens S, Manouvrier S, Roberts AE, Faravelli F, Greco MA, Pyeritz RE, Milewicz DM, Coucke PJ, Cameron DE, Braverman AC, Byers PH, De Paepe AM, Dietz HC (2006) Aneurysm syndromes caused by mutations in the TGF-beta receptor. N Engl J Med 355:788-798

225. Baas AF, Medic J, van 't Slot R, de Kovel CG, Zhernakova A, Geelkerken RH, Kranendonk SE, van Sterkenburg SM, Grobbee DE, Boll AP, Wijmenga C, Blankensteijn JD, Ruigrok YM (2010) Association of the TGF-beta receptor genes with abdominal aortic aneurysm. Eur J Hum Genet 18:240-244

226. Suzuki A, Miyagawa-Tomita S, Komatsu K, Nakazawa M, Fukaya T, Baba K, Yutani C (2004) Immunohistochemical study of apparently intact coronary artery in a child after Kawasaki disease. Pediatr Int 46:590-596

227. Suzuki A, Miyagawa-Tomita S, Komatsu K, Nishikawa T, Sakomura Y, Horie T, Nakazawa M (2000) Active remodeling of the coronary arterial lesions in the late phase of Kawasaki disease: immunohistochemical study. Circulation 101:2935-2941

228. Tone Y, Furuuchi K, Kojima Y, Tykocinski ML, Greene MI, Tone M (2008) Smad3 and NFAT cooperate to induce Foxp3 expression through its enhancer. Nat Immunol 9:194-202

229. Cheng X, Yu X, Ding YJ, Fu QQ, Xie JJ, Tang TT, Yao R, Chen Y, Liao YH (2008) The Th17/Treg imbalance in patients with acute coronary syndrome. Clinical immunology (Orlando, Fla 127:89-97

230. Franco A, Shimizu C, Tremoulet AH, Burns JC (2010) Memory T-cells and characterization of peripheral T-cell clones in acute Kawasaki disease. Autoimmunity 43:317-324

231. Bujak M, Ren G, Kweon HJ, Dobaczewski M, Reddy A, Taffet G, Wang XF, Frangogiannis NG (2007) Essential role of Smad3 in infarct healing and in the pathogenesis of cardiac remodeling. Circulation 116:2127-2138

232. Onouchi Y, Gunji T, Burns JC, Shimizu C, Newburger JW, Yashiro M, Nakamura Y, et al. (2008) ITPKC functional polymorphism associated with Kawasaki disease susceptibility and formation of coronary artery aneurysms. Nature genetics 40:35-42

233. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, et al. (2008) Genetics of gene expression and its effect on disease. Nature 452:423-428