LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Genomic and Transcriptomic Profiling of Coeliac Disease

Ciara Coleman

A thesis submitted to the University of Dublin for the degree of Doctor of Philosophy

Department of Genetics

University of Dublin

Trinity College

2016

TRINITY COLLEGE DUBLIN
18 APR 2016
ACADEMIC REGISTRY

PhD
Medicine.

TRINITY LIBRARY
2 7 JUL 2016
DUBLIN

Thesis 11019

Declaration

I hereby declare that this thesis has not been previously submitted for a degree at this or any other university and that the work described in this thesis is entirely my own work. I hereby give my permission to the library of Trinity College Dublin to lend or copy this thesis upon request for the purpose of study.

Summary

Coeliac Disease (CD) is a chronic immune-mediated disease triggered by the ingestion of gluten. It has an estimated prevalence of approximately 1% in European populations. Genome-wide association studies (GWAS) have identified a large number of CD risk loci to date that have added substantially to our understanding of the genetic complexity of CD. However despite these major advancements, replication of such findings is scarce. Therefore replication of disease-associated loci in independent populations is considered to be a necessary validation step.

This thesis set out to replicate findings from a recent large Immunochip analysis in an independent Irish case-control sample. We show that direction of effect of previously identified risk alleles highly correlated with those in the Irish population. A meta-analysis of both datasets, raised two new loci harbouring the *ZNF335* and *NFIA* genes beyond the genome-wide significance threshold, taking the total number of CD susceptibility loci to 43. NFIA is a transcription factor while ZNF335 forms part of a chromatin-remodelling complex, implicating chromatin remodelling/epigenetics as a process that may be altered in CD.

The ability to identify susceptibility loci is dependent of the study having insufficient power for associations to meet the current genome-wide significant threshold ($P < 5 \times 10^{-8}$). Therefore disease relevant loci that fail to meet this threshold will be missed. In order to circumvent this problem in our study, we used Polygene Risk Score (PRS) analysis to estimate the proportion of the variance that could be explained in the Irish population by previously identified loci and to determine whether a proportion of

variance in our CD cohort could be explained by the non-GWS SNPs. We found that up to 35% of the genetic variance was explained by these PRS ($P = 9x10^{-75}$). When limited to non-HLA loci, we explain a maximum of 4.5% of the genetic variance ($P = 3.6x10^{-18}$) at a significance level of $P < 0.05$. Our findings are comparable with previous studies using similar methods. We also find that there is potential for our results to be clinically beneficial as a weighted burden of risk alleles can be used to distinguish coeliac case status.

Interpreting the functional roles these CD risk loci play in disease represents a challenge. As the majority of these variants reside in non-coding regions, they are believed to exert a regulatory role by altering the expression of disease-associated genes. Recently there has been considerable emphasis on characterising cell type specific and stimulus dependent genetic variants. Therefore in this study we used RNA-seq to profile the transcriptomes of $CD4^+$ T-cells, a cell type crucial for CD pathogenesis, in both stimulated and resting samples from individuals with CD and unaffected controls. All individuals with CD recruited for this study were on a gluten-free diet (GFD) and therefore were not in active disease. We identified numerous transcriptional changes across all conditions, with >1000 genes differentially expressed in anti-CD3/CD28 stimulated samples. *IFNG* was the most strongly up regulated gene (log2 fold change 4.6; $P_{adjusted} = 2.40x10^{-11}$), validating its previously established role in CD. Co-expression network analysis highlighted both *IFNG* and the genetically associated transcription factor *BACH2* as key regulatory genes in CD. We show for the first time significant differential expression of BACH2 regulated genes in CD ($P < 2.2x10^{-16}$). *BACH2* expression is significantly down regulated in our data

(log2FC -1.75; $P_{adjusted} = 3.6x10^{-3}$), indicating that reduced expression of this master regulator of T-cell differentiation promotes a pro-inflammatory response and strongly corroborates genetic evidence that BACH2 plays an important role in CD pathogenesis. Finally efforts to extend this study into a larger sample size to connect genotype information with gene expression data for eQTL analysis could not be completed during this time.

Therefore this study has verified that the genetic architecture in the Irish population is highly similar to other European populations. We have discovered two new CD susceptibility loci and have identified key regulatory networks that may be important for the pathogenesis of CD.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Professor. Ross McManus for giving me the opportunity to pursue this research and for your continuous encouragement, support and expertise over the last five years.

I would also like to thank Dr. Anthony Ryan for all the help and advise you gave me whenever I needed it.....which seemed like most of the time!

To Dr. Emma Quinn, I cannot thank you enough for everything you have done for me. Your help, encouragement and support have been invaluable and I greatly appreciate it.

To Sinead, Cat, and Jill – we have been through so much together over the past few years and I cannot think of three more amazing people I'd rather have gone through this journey with! I would not have completed this PhD without you guys - thank you for everything! To Mikaela and Ben and everyone else in the reading room, thank you so much for making me laugh everyday and for making the basement a great place to work. I have made friends for life and I wish we could work together forever!

Outside of work I would like to say a huge thank you to all my friends especially Julie, Nicola, Sharifa, Naomi and Ciara – thank you for your encouragement, support and for putting up with the friend who just wouldn't leave college! Thank you for the endless wine nights; although you's may be responsible for extending my PhD by a few months

;) To Sean, thank you so much for your constant motivation and always believing in me even when I didn't believe in myself! I promise I will learn to drive now!!

Thank you to all my grandparents, aunts and uncles for your endless support.

I would like to say a massive thank you to my amazing family – to my brother Mark, thank you for being there to listen to me and for always offering to help in anyway you can. To my parents Noelle & Colin, who have supported me from day one and have given me everything I could ever need or want. Everything I have ever achieved would not have been possible without your constant love and support.

Lastly but by no means least, I would like to extend a special thank you to my late grandad Vincent, you always took an interest in my studies and believed in me from the very beginning. This thesis is dedicated to you.

Statement of Work

This work was undertaken by the author with the Genetics research group, Department of Clinical Medicine, Institute of Molecular Medicine, Trinity College Dublin and St James' Hospital.

Publications arising from this thesis

1. Coleman C*, Quinn EM*, Ryan AW, Conroy J, Trimble V, Mahmud N, Kennedy N, Corvin AP, Morris DW, Donohoe G, O'Morain C, MacMathuna P, Byrnes V, Kiat C, Trynka G, Wijmenga C, Kelleher D, Ennis S, Anney RJL, McManus R., 2015. Common Polygenic Variation in Coeliac Disease and Confirmation of ZNF335 and NIFA as Disease Susceptibility Loci. *European Journal of Human Genetics.*

2. Quinn EM*, Coleman C*, Molloy B, Dominguez Castro P, Cormican P, Trimble V, Mahmud N, McManus R., 2015. Transcriptome Analysis of CD4+ T-cells in Coeliac Disease Reveals Imprint of BACH2 and IFNγ Regulation. *PloS one.*

*These authors contributed equally

3. Coleman, C., Quinn, E.M. and McManus, R., 2015. Quality Control Procedures for High-Throughput Genetic Association Studies. *Celiac Disease: Methods and Protocols*, pp.203-215.

List of abbreviations

AcoRD Academic Centre on Rare Diseases

Ad12 Adenovirus type 12

AGA Anti-gliadin antibodies

AITD Autoimmune thyroid disease

APC Antigen presenting cell

AUC Area Under Curve

BACH2 BTB and CNC homology 1, basic leucine zipper transcription factor 2

CD Coeliac Disease

CEPH Centre d'Etude du Polymorphisms Humain

CF Cystic Fibrosis

CI Confidence Interval

CNV Copy number variant

CTLA4 Cytotoxic T-lymphocyte-associated protein 4

DAPPLE Disease Association Protein-Protein Link Evaluator

DE Differentially expressed

DGP Deamidated gliadin peptide

EMA Endomysium antibodies

ESPGHAN European Society for Paediatric Gastroenterology, Hepatology and Nutrition

eQTL Expression Quantitative Loci

FDR False Discovery Rate

GEO Gene Expression Omnibus

GLM Generalized linear model

GO Gene Ontology

GRAIL Gene Relationships Among Implicated Loci

GSEA Gene Set Enrichment Analysis

GWAS Genome-wide association studies

GWLS Genome-wide linkage studies

GWS Genome-wide significant

HCFC1 Host cell factor C1

HLA Human leukocyte antigen

HGP Human Genome Project

HTA Human Transcriptome Array

HWE Hardy-Weinberg equilibrium

IBD Inflammatory bowel disease

IBD Identity by Descent

IELs Intraepithelial lymphocytes

IFNα Interferon alpha

IFNγ Interferon gamma

Ig Immunoglobulin

IL Interleukin

IL18RAP Interleukin 18 receptor activating protein

INRICH Interval-based Enrichment Analysis Tool for Genome Wide Association Studies

JAK Janus kinase

KEGG Kyoto Encyclopedia of Genes and Genomes

LD Linkage disequilibrium

Lnc-RNAs Long non-coding RNAs

LPLs Lamina propria lymphocytes

NCBI National Centre for Biotechnology Information

NF1A Nuclear Factor 1A

NKG2 Natural Killer Group 2

MAF Minor allele frequency

MDS Multi-dimensional scaling

MICA MHC class I polypeptide-related sequence A

MM Module membership

MMPs Matrix metalloproteinases

MS Multiple sclerosis

NK cell Natural killer cell

OR Odds ratio

PCA Principal Component Analysis

PCIT Partial Correlation and Information Theory

PMA Phorbal 12-myristate 13-acetate

PRS Polygenic risk score

PTPN2 Protein tyrosine phosphatase non-receptor type 2

QC Quality Control

QQ Quantile-Quantile

RA Rheumatoid arthritis

RNA Ribonucleic acid

ROC Receiver Operating Curve

SLE Systemic lupus erythematosus

SNP Single nucleotide polymorphism

SOCS1 Suppressor of cytokine signalling 1

STAT Signal transducer and activator of transcription

T1D Type 1 diabetes

TOM Topological Overlap Measure

TSC The SNP Consortium

tTG Tissue transglutaminase

UTR Untranslated regions

VP7 Virus Protein 7

WGCNA Weighted Gene Co-expression Network Analysis

WTCCC Wellcome Trust Case Control Consortium

ZNF335 Zinc Finger Protein 335

Table of Contents

Declaration	i
Summary	ii
Acknowledgements	v
Statement of work	vii
Publications arising from this thesis	viii
List of Abbreviations	ix
Table of Contents	xii
Index of Figures	xvii
Index of Tables	xxii
Chapter I: General Introduction	1
1.1 Presentation of Coeliac Disease	2
1.2 Diagnosis	4
1.2.1 Serology	4
1.2.2 Endoscopy	5
1.2.3 HLA Typing	7
1.3 Epidemiology	9
1.4 Environmental Factors	13
1.4.1 Gluten	13
1.4.2 Infant feeding	14
1.4.3 Infections	14
1.5 Role of Human Leukocyte Antigen (HLA) locus in CD pathogenesis	16
1.5.1 HLA molecules	16
1.5.2 The Immunological role of HLA risk variants	19
1.6 Immunological model of CD pathogenesis	20
1.7 Role of non-HLA in CD pathogenesis	23
1.7.1 Genome-wide linkage studies	23
1.7.2 Genome-wide linkage studies in CD	24
1.7.3 Genetic association studies	26
1.7.4 Genome-wide association studies	27
1.7.5 Genome-wide association studies in CD	30

1.8 Immunochip high-density array 32
1.9 Polygene Risk Score (PRS) analysis 37
1.10 Functional assessment of genetic associations 39
1.10.1 Microarray technology 40
1.10.2 RNA-sequencing 41
1.10.3 Affymetrix GeneChip Human Transcriptome (HTA) 2.0 41
1.10.4 Expression Quantitative trait loci (eQTL) analysis 42
1.10.5 Pathway and network analysis 44
1.11 Conclusion 46
1.12 Aims and Objectives 48
Chapter II: Immunochip & Polygene analysis 49
2.1 Introduction 50
2.2 Aims and Objectives 56
2.3 Materials and Methods 57
2.3.1 Subjects 57
2.3.2 DNA sample preparation 57
2.3.3 Preparing the plates for Immunochip analysis 57
2.3.4 Workflow for Immunochip assay for whole-genome genotyping 58
2.3.5 Genotyping 59
2.3.6 Data Quality Control (QC) 60
2.3.6.1 Per-Individual QC 60
2.3.6.1.1 Missing genotypes 60
2.3.6.1.2 Heterozygosity 61
2.3.6.1.3 Identity by Descent (IBD) 62
2.3.6.1.4 Principle Component Analysis (PCA) 62
2.3.6.2 Per-Marker QC 65
2.3.6.2.1 Missing genotypes/differential missingness 65
2.3.6.2.2 Hardy-Weinberg Equilibrium (HWE) 65
2.3.6.2.3 Minor Allele Frequency (MAF) 66
2.3.7 Genetic Association 66
2.3.8 Meta-analysis 69
2.3.9 Sign test 69

2.3.10 Polygenic Risk Score (PRS) analysis 70
2.3.11 Receiver Operating Characteristic (ROC) curve 71
2.4 Results 73
2.4.1 Genetic Association 73
2.4.2 Quantile-Quantile (QQ) plot 77
2.4.3 Replication 81
2.4.4 Meta-analysis 81
2.4.5 Polygenic Risk Score (PRS) analysis 85
2.5 Discussion 91
2.5.1 *KIAA1109/ADAD1/IL21/IL2* 92
2.5.2 *PTPN2* 93
2.5.3 *IL12A* 94
2.5.4 Nuclear Factor 1A *(NFIA)* 97
2.5.5 *ZNF335* 98
2.6 Conclusion 102
Chapter III: Transcriptome analysis of $CD4^+$ T-cells in CD 104
3.1 Introduction 105
3.1.1 Transcription analysis of CD to date 106
3.1.2 cDNA microarray and RNA-sequencing 107
3.1.3 Pathway analysis 108
3.1.4 Weighted Genome Co-expression Network Analysis (WGCNA) 110
3.2 Aims and Objectives 111
3.3 Materials and Methods 112
3.3.1 Collection and processing of blood samples 112
3.3.2 RNA extraction 114
3.3.3 mRNA-sequencing 114
3.3.4 Alignment and differential expression analysis 114
3.3.5 Power calculation for RNA-seq 115
3.3.6 Pathway analysis of differentially expressed genes 116
3.3.7 Analysis of differential exon usage using DEXSeq 116
3.3.8 Weighted Genome Co-expression Network Analysis (WGCNA) 117
3.3.9 Testing for enrichment for genetic associations 118

3.3.10 Gene Set Enrichment Analysis (GSEA) 120
3.4 Results 121
3.4.1 mRNA-sequencing 121
3.4.1.1 Power analysis of transcriptome data 121
3.4.1.2 PCA analysis 124
3.4.2 Differential expression analysis 127
3.4.3 T-cell subpopulations 132
3.4.4 Pathway analysis of differentially expressed genes 134
3.4.4.1 anti-CD3/CD28 stimulation 138
3.4.4.2 PMA stimulation 138
3.4.4.3 Unstimulated 138
3.4.5 Differential exon analysis 140
3.4.6 Relating differentially expressed genes to disease association 141
3.4.7 Differentially expressed genes are enriched for coeliac association signals 144
3.4.8 Weighted Genome Co-expression Network Analysis (WGCNA) 146
3.4.9 Gene Set Enrichment Analysis (GSEA) 153
3.5 Discussion 157
3.6 Conclusion 171
Chapter IV: Gene expression analysis in an extended case-control Population 172
4.1 Introduction 173
4.2 Aims and Objectives 180
4.3 Materials and Methods 181
4.3.1 Collection of case blood samples 181
4.3.2 Collection of control blood samples 181
4.3.3 DNA sample preparation 182
4.3.4 Preparing the 96 well plates for Immunochip analysis 182
4.3.5 Genotyping 182
4.3.6 Stimulation of $CD4^+$ T-cells 182
4.3.7 RNA extraction 183
4.3.8 Checking RNA quality and concentration 183

4.3.9 Microarray analysis 183
4.3.9.1 Probe/Gene array 183
4.3.9.2 Fluidics Station 187
4.3.9.3 GeneChip Scanner 187
4.3.9.4 Computer workstation 188
4.3.10 Preprocessing 188
4.3.10.1 Affymetrix Expression Console (EC) Software 189
4.3.10.2 Robust Multi-array Average (RMA) 189
4.3.10.2.1 Background correction 189
4.3.10.2.2 Normalisation of the data 190
4.3.10.2.3 Summarisation 191
4.3.11 Quality Control 192
4.3.11.1 Report Controls 192
4.3.11.1.1 Hybridisation Controls 192
4.3.11.1.2 Labelling Controls 195
4.3.11.2 Box Plot 197
4.3.11.2.1 Log Probe Cell Intensity box plot 197
4.3.11.2.2 Relative Log Expression Signal box plot 199
4.3.11.3 Signal Histogram 201
4.3.11.4 Pearson's Correlation (Signal) 203
4.3.11.5 Principle Component Analysis 205
4.3.11.6 Removal of samples following QC 205
4.3.12 Transcriptome Analysis Console (TAC) 2.0 Software 210
4.3.13 qRT-PCR 210
4.3.13.1 Calculating gene expression using threshold amplification cycle data (Ct) 211
4.4 Results 212
4.4.1 Differential gene expression 212
4.4.2 Comparison of microarray results with differentially expressed RNA-seq genes 216
4.4.3 Comparison of microarray differentially expressed genes with CD associated genes 220

4.4.4 Differential exon analysis 222
4.4.5 Comparison of microarray differentially expressed exons with CD associated genes 222
4.4.6 qRT-PCR 223
4.4.7 Genotyping 226
4.5 Discussion 227
4.6 Conclusion 236
Chapter V: General Discussion 238
Appendices 256
Bibliography 266

Index of Figures

Figure 1.1 Presentation of CD symptoms in children and adults 3

Figure 1.2 Duodenal biopsy samples 6

Figure 1.3 Worldwide prevalence of CD 12

Figure 1.4 HLA associations in CD 18

Figure 1.5 Enzymatic action of tTG 20

Figure 1.6 Immunological model of CD pathogenesis 22

Figure 1.7 GWS publications made between 2005 – 2013 30

Figure 1.8 Chronological line highlighting the number of SNP discoveries that have been discovered since the first GWAS in 2007 until 2014 35

Figure 2.1 Workflow of Immunochip assay for whole-genome genotyping 59

Figure 2.2 Genotype failure rate vs heterzygosity rate across all Individuals in the study pre-QC 61

Figure 2.3 Multi dimensional scaling (MDS) plot demonstrating ancestry clustering based on genome-wide association data pre-QC 64

Figure 2.4 Multi dimensional scaling (MDS) plot demonstrating ancestry clustering based on genome-wide association data post-QC 64

Figure 2.5 Unsuccessful, low quality clustered SNPs 67

Figure 2.6 Successful, low quality clustered SNPs 68

Figure 2.7 Manhattan plot of GWA results using the Immunochip and encompassing the HLA locus 74

Figure 2.8 Manhattan plot of GWA results excluding the HLA locus 74

Figure 2.9 Locus association plot for SNP rs13132308, which has the strongest association to CD in our dataset 75

Figure 2.10 Locus association plot for SNP rs11875687, which has the second strongest association to CD in our dataset 76

Figure 2.11 QQ plot of the data representing both HLA and non-HLA SNPs 79

Figure 2.12 QQ plot of the data representing non-HLA SNPs 80

Figure 2.13 Manhattan plot from the meta-analysis of the Irish and

Trynka *et al* datasets excluding the HLA locus 82
Figure 2.14 The proportion of variance explained by PRS at each of the different thresholds 87
Figure 2.15 ROC curve showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for all markers in our study 89
Figure 2.16 ROC curve showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for non-HLA only markers in our study 89
Figure 2.17 ROC curve showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for HLA markers only in our study 90
Figure 3.1 Purity of (A) $CD4^+$, (B) $CD8^+$ and (C) $CD14^+$ T-cells from human peripheral blood leukocytes 113
Figure 3.2 Percentage of detected genes at different levels of sequencing depth at different fold changes across all annotated genes 124
Figure 3.3 PCA plots based on normalised read counts from the RNA-seq datasets 126
Figure 3.4 Scatter plot of fold-change in expression in cases versus controls (y-axis) against expression level (x-axis) 128
Figure 3.5 Venn diagram indicating the overlap in significantly differentially expressed genes (up/down) in each group 131
Figure 3.6 The JAK/STAT signalling pathway 136
Figure 3.7 The cytokine-cytokine receptor pathway 137
Figure 3.8 QQ plot of Immunochip association signals within genes that were differentially expressed between our anti-CD3/CD28 stimulated coeliac and control samples 144
Figure 3.9 Protein interaction networks for modules 3 and 6 149
Figure 3.10 Enrichment plot of the distribution of BACH2 regulated Genes from our dataset when compared to the Vahedi *et al* dataset 154
Figure 3.11 A heat map showing the expression values of genes in the dataset sorted by correlation with phenotype 155

Figure 3.12 Enrichment plot of the distribution of BACH2 regulated genes from our dataset when compared to the Roychoudhuri *et al* dataset 156
Figure 4.1 Line graph showing spike controls for each sample on the array 194
Figure 4.2 Line graph showing poly-A controls for each sample on the array 196
Figure 4.3 Box plots of the probe cell intensity values 198
Figure 4.4 Box plot generated from the probe set signal values following normalisation and summarisation 200
Figure 4.5 A density plot of the log-intensity distribution of each array 202
Figure 4.6 A heat map displaying the pairwise Pearson's Correlation (r^2) of the signal values from each array in our study 204
Figure 4.7 PCA plot of our coeliac samples versus controls 207
Figure 4.8 PCA plot of protein coding genes only, from coeliac patients separated by age at the time of blood collection 208
Figure 4.9 PCA plot of protein coding genes only, from coeliac patients separated based on age of diagnosis 208
Figure 4.10 A Scatter plot of controls versus coeliac cases 214
Figure 4.11 A Volcano plot of controls versus coeliac cases 215
Figure 4.12 Venn diagram showing comparison of differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data 217
Figure 4.13 Venn diagram showing comparison of anti-CD3/CD28 significantly differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data 218
Figure 4.14 Venn diagram showing comparison of anti-CD3/CD28 significantly differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data at an FDR of 5% 220
Figure 4.15 Venn diagram showing the overlap of genes differentially expressed in our microarray data (no filters applied) with previously identified CD susceptibility genes 221
Figure 4.16 Venn diagram showing the overlap of genes differentially

expressed in our microarray data (linear fold change of < -2 and > 2 and FDR P-value <0.05) with previously identified CD susceptibility genes 221

Figure 4.17 Venn diagram showing overlap of genes showing evidence of alternative splicing in our microarray data with previously identified CD susceptibility genes 223

Figure 4.18 Bar chart showing levels of gene expression, represented by Log2FC across all three platforms for all five chosen genes 225

Index of Tables

Table 1.1 Comparison of Marsh/Oberhuber and Corazza classification systems 6

Table 1.2 Algorithm proposed by the ESPGHAN to diagnose CD in symptomatic individuals 8

Table 1.3 Algorithm proposed by the ESPGHAN to diagnose CD in asymptomatic individuals 8

Table 2.1 11 of the 57 previously independent CD signals were nominally significant in our Irish sample cohort 83

Table 2.2 Meta-analysis results of the Trynka *et al* dataset combined with our data 84

Table 2.3 Variance explained (R2) for the three sets of SNPs tested (All, HLA only and non-HLA) at the different thresholds tested 86

Table 2.4 The area under the receiver operating characteristic curve (AUC) for each threshold 88

Table 3.1 Sample information for Individuals with CD and controls sequenced in each group (CD3/CD28, PMA and UNS) 122

Table 3.2 Sample information for Individuals with CD and controls sequenced in this study 123

Table 3.3 The top 25 most significant differentially regulated genes in the anti-CD3/CD28 group 129

Table 3.4 Levels of transcript expression between our CD3 stimulated coeliac cohort and controls for a subset of $CD4^+$ Th1 *(TBX21)*, Th2 *(GATA3),* Th17 *(RORC)* and T-reg *(Foxp3)* gene markers 133

Table 3.5 Pathways identified by GOSeq as significantly enriched for genes differentially expressed in anti-CD3/CD28 stimulated samples 135

Table 3.6 Pathways identified by GOSeq as significantly enriched for genes differentially expressed in PMA stimulated samples 139

Table 3.7 Differential exon usage in our anti-CD3/CD28 stimulated samples at 4 genes that have been genetically associated with CD 140

Table 3.8 DESeq2 results across all groups for genes that have been

genetically associated with CD and are also differentially expressed in our data 143

Table 3.9 WGCNA identified 15 modules in our anti-CD3/CD28 stimulated dataset 148

Table 4.1 Variables examined using PCA analysis 206

Table 4.2 The top 25 most significantly up regulated and down regulated genes 213

Table 4.3 Log2FC values of five genes analysed via qRT-PCR across all three platforms 225

Chapter I

General Introduction

1.1 Presentation of Coeliac Disease

Coeliac Disease (CD) is defined as a common, chronic inflammatory disease of the small intestine that occurs in genetically predisposed individuals and is triggered by exposure to the storage protein of wheat – gluten – and similar proteins in related grains (Schuppan et al., 2009). The word coeliac, meaning "hollow" in Greek, was first described by the Greek physician Aretaeus in the first century AD (Thomas, 1945), and although the first real description of CD came in 1888 by Samuel Gee (Gee, 1888), the link between CD and diet wasn't established until the 1950s when Willem Dicke, a Dutch paediatrician, highlighted the role wheat played in the onset of CD and how its exclusion from the diet led to dramatic improvements in patients symptoms (Dicke, 1950). The link between gluten and CD was made in 1952 by Anderson *et al* (1952). However the histological classification of CD was not formalised until 1992 when Marsh (Marsh, 1992) established a grading system for the intestinal damage seen in the coeliac lesion. It consisted of 3 stages ranging from lymphocyte infiltration and crypt hyperplasia to total villous atrophy. Thought of as primarily a paediatric disease for a very long time, it is now clear that the onset of CD can occur at any age, from early infancy to late adulthood. CD symptoms vary between infants, adolescents and adults. Young children generally present with "classical" CD symptoms such as diarrhoea, failure to thrive and abdominal distension (Vivas et al., 2008), older children and adolescents often show atypical gastrointestinal symptoms such as pain, vomiting or constipation. They also present with extraintestinal symptoms such as arthritis, anaemia and neurological symptoms, whilst adults normally present with diarrhoea, anaemia, osteoporosis, IBS, bloating, chronic fatigue and signs of villous atrophy among others (Tengah et al., 2002) (Rampertab et al., 2006) (Fig 1.1). Currently the

only effective therapy for the treatment of CD is a gluten free diet (GFD), and if left untreated, CD has been associated with increased morbidity and mortality (Corrao et al., 2001).

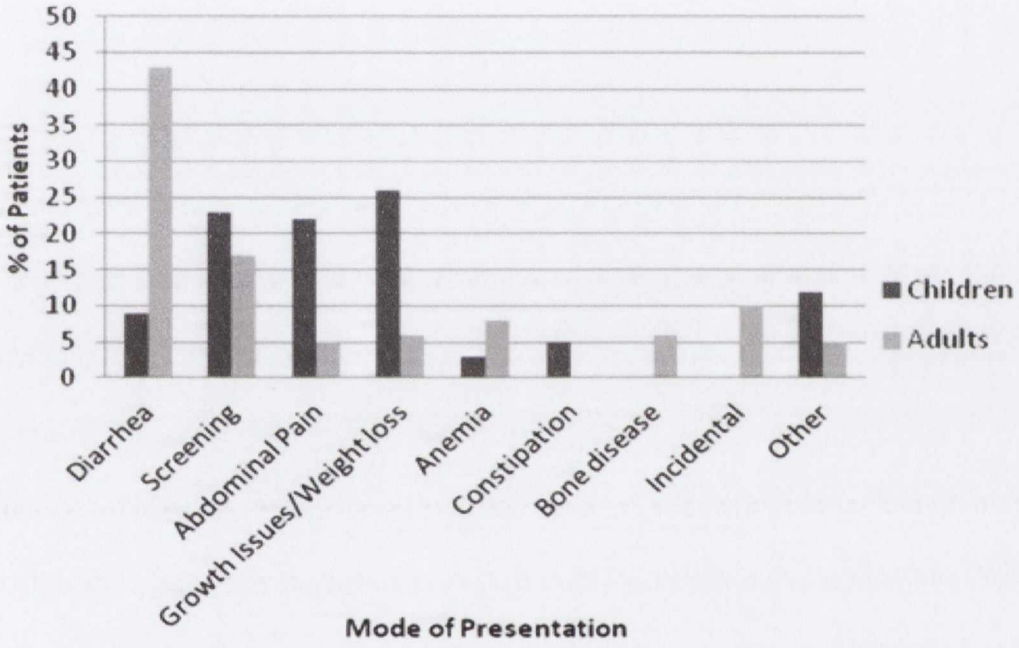


Figure 1.1: Presentation of CD symptoms in children and adults. Adults *n*=1,499, children *n*=318.

1.2 Diagnosis

The European Society for Paediatric Gastroenterology, Hepatology and Nutrition (ESPGHAN) guidelines for the diagnosis of CD was first published in 1990, requiring both accurate histological and serological testing for CD diagnosis (Walker-Smith et al., 1990). Thus, the requirement of both histological and serological testing has been the gold standard for diagnosing CD for over 20 years.

1.2.1 Serology

As previously mentioned, serological testing is used in the diagnosis of CD. The most common serological tests clinically available include the presence of IgA or IgG anti-gliadin antibodies (AGA), IgA endomysium antibodies (EMA) or IgA tissue transglutaminase antibodies (tTG). tTG is a calcium dependent enzyme that has many functions; in the context of CD, this includes modification of gluten as an immunoreactive substrate. Measurement of IgA anti-tTG antibodies is recommended for initial testing due to its high sensitivity (>93%) and high specificity (96%) (Lewis and Scott, 2010). Measurement of IgA anti-EMA antibodies is also highly sensitive and specific for the diagnosis of CD (94% and 98% respectively) (Lewis and Scott, 2006). However it is not as readily used due to its limitations, for example, it is expensive, qualitative and subjective, as the test to detect anti-EMAs requires fluorescent microscopy, which is subject to flexibility of interpretation. Poor sensitivity and specificity of AGA (>52% and >71% for IgA, >57% and >47% for IgG) limits its use as a diagnostic tool (Rashtak et al., 2008). 2-3% of coeliac individuals are IgA deficient (McGowan et al., 2008); therefore, incorporating IgG based testing into serology testing is an option whereby IgG anti-tTG antibodies or the IgG class of deamidated

gliadin peptide (DGP) antibodies are measured which can be used to diagnose CD in these individuals (Lewis and Scott, 2010).

1.2.2 Endoscopy

As mentioned, histological confirmation by biopsy of the small intestine has been used as the gold standard for decades in order to confirm CD diagnosis. Alterations seen in the coeliac lesion include a change in crypt pattern and villous atrophy (Fig 1.2). These histological changes were first classified according to the Marsh classification system (Marsh, 1992). This system was subsequently modified by Oberhuber *et al* (1999) and more recently by Corazza and colleagues (2005) who combined and simplified both previous classification systems (Table 1.1). The finding of increased intraepithelial lymphocytes (IELs) (>25 IELs/100 enterocytes) in the small intestine is another important diagnostic tool in CD (Walker et al., 2010). IELs are cytotoxic T-cells responsible for the mucosal damage characteristic of CD. Their function is mediated by IL15, which leads to increased expression of NK receptors (NKG2D and NKG2C) on their surface, and in turn bind to their respective ligands (MICA and HLA-E) on the surface of enterocytes inducing mucosal epitheial cell damage (Meresse et al., 2004).

However, the status of the duodenal biopsy being used as a reference standard for the diagnosis of CD has been called into question. Histologic changes in the small intestine may be patchy meaning that biopsy samples taken from different areas of the duodenum may not be uniform, thus leading to false-positive and false-negative diagnosis. Therefore, there is now some concern that over reliance on observing

abnormal histology to definitively diagnose CD may actually result in a failure to diagnose CD accurately.

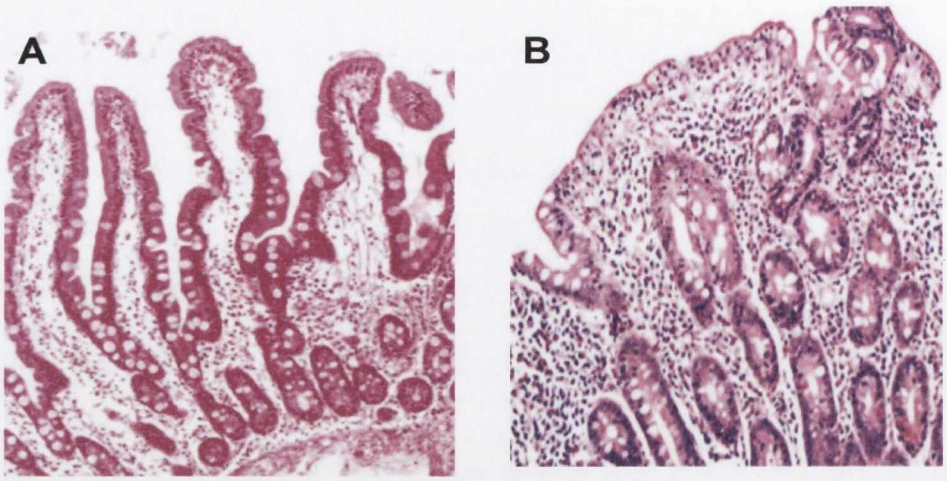


Figure 1.2: A) Normal duodenal mucosa. B) Duodenal mucosa with total villous atrophy (Walker and Murray, 2011).

Table 1.1: Comparison of Marsh/Oberhuber and Corazza classification systems (Walker and Murray, 2011).

	Morphology		IELs/100 enterocytes
Marsh			
Type 0	No detectable changes in inflammation or crypt/villous height ratio (preinfiltrative)		>40
Type 1	Increase in the number of IELs (infiltrative)		>40
Type 2	Inflammation, villous blunting and increased crypt:villous height ratio (hyperplastic)		>40
Type 3	Severe inflammation, flat villi; hyperplastic crypts (destructive)		>40
Oberhuber			
Type 0	Normal mucosa		>40
Type 1	Increased IELs; normal villous architecture and normal height of crypts		>40
Type 2	Normal villous architecture, increased IEL, crypt hyperplasia		>40
Type 3	Destructive type with varying degrees of villous atrophy; in all subtypes, there is increased crypt height and influx of inflammatory cells		>40
Type 3a	Partial villous atrophy; villi blunt and shortened with a villous:crypt ratio, 1:1		>40
Type 3b	Subtotal villous atrophy; villi atrophic but still separate and recognizable		>40
Type 3c	Total villous atrophy; villi rudimentary or absent; mucosa resembles colonic mucosa		>40
Type 4	Atrophic hypoplastic lesion; flat mucosa with normal crypt height and no significant inflammation with normal IEL counts		>40
Corazza	Morphology	Corresponds to Oberhuber/ Marsh classification	IELs/100 enterocytes
Grade A	Non-atrophic architecture normal	Type 0 deleted Type 1 and type 2	>25
Grade B1	Atrophic, villous to crypt ratio is less than 3:1	Type 3a and Type 3b	>25
Grade B2	Atrophic, villi are no longer detectable	Type 3 Type 4 deleted	>25

1.2.3 HLA Typing

In recent years, HLA genotyping has been proposed as a powerful screening tool to diagnose CD alongside serological and histological testing. HLA genotyping has a poor positive predictive value as up to 40% of the population may carry HLA-DQ2 and/or HLA-DQ8 heterodimers but fail to develop CD; however, it has a very strong negative predictive value (close to 100%), as the disease is not likely to develop in individuals lacking both HLA-DQ2 and HLA-DQ8 (Karell et al., 2003).

The increase in our knowledge of CD and the availability of accurate antibody tests over the past two decades prompted ESPGHAN to revise their existing guidelines and set out new criteria for CD diagnosis (Husby et al., 2012). The new guidelines focused on the crucial role of serological tests in symptomatic individuals and HLA typing in asymptomatic individuals belonging to groups at risk of CD. They also stated that under specific conditions, exclusion of a duodenal biopsy is acceptable and is now within the new recommended guidelines for the diagnosis of CD in children and adolescents (Table 1.2 and Table 1.3).

Table 1.2: Algorithm proposed by the ESPGHAN to diagnose CD in symptomatic individuals (Tonutti and Bizzaro, 2014).

Child/adolescent with symptoms suggestive of CD
↓
Anti-tTG IgA and total IgA → anti-tTG negative and normal total IgA → no CD
↓
Anti-tTG positive → <10x cutoff → intestinal biopsy
↓
Anti-tTG >10x cutoff

EMA and HLA DQ2/DQ8
↓
EMA pos and HLA pos
↓
Diagnosis of CD
↓
Gluten-free diet and follow up

(CD, celiac disease; EMA, endomysial antibodies; tTG, tissue transglutaminase).

Table 1.3: Algorithm proposed by the ESPGHAN to diagnose CD in asymptomatic individuals at risk for CD (Tonutti and Bizzaro, 2014).

At risk asymptomatic individual
↓
HLA DQ2/DQ8 → negative → no CD and no risk for CD
↓
Positive
↓
Anti-tTG → positive < 3x cutoff → EMA → EMA positive → biopsy
↓
Anti-tTG pos > 3x normal
↓
Biopsy
↓
Marsh 2 or 3
↓
Diagnosis of CD and gluten-free diet

(CD, celiac disease; EMA, endomysial antibodies; tTG, tissue transglutaminase).

1.3 Epidemiology

Until the late 1970s, diagnosis of CD was made based mainly on its clinical symptoms such as diarrhoea and weight loss and was considered to be a rare disease with a prevalence of 0.03% worldwide (Lehtonen and Kuusela, 1986; Aromaa, 1989). However, more recent population screenings have shown CD to be one of the more frequent inherited diseases with susceptible individuals estimated at approximately 1% of the European population (Mustalahti et al., 2010). CD not only affects individuals of European ancestry but is also common in individuals of Asian, North African and Middle Eastern (Rawashdeh et al., 1996) origin. Some African populations, specifically of the Maghreb region in North Africa - Morocco, Algeria, Egypt (Medhat et al., 2011), Tunisia (Mankaï et al., 2006) and Libya have very high prevalence's of CD. In fact the Saharawi of Algeria reportedly have the highest incidence of CD worldwide (Catassi et al., 1999). Until recently, CD was thought to be a rare disease in the USA with studies showing a low prevalence of the disease (Talley et al., 1994;Rossi et al., 1993). However these studies were local surveys, not large epidemiological studies, and significantly underestimated the occurrence of CD. Fasano *et al* (2003) carried out an extensive epidemiological study incorporating 32 states and found that CD had a similar incidence in the USA as in European populations, estimated at approximately 1:133. According to Ludvigsson *et al* (2013), the incidence of CD in the North-American population has actually been increasing over the past decade. They estimated there to be a 1.5 fold increase in the prevalence of CD between 2000 and 2010.

As to how common CD is in the Asia-Pacific region remains largely unknown, but there are a few countries for which the prevalence has been published. One retrospective Australian study (of a European derived population) showed at least 1:178 confirmed by biopsy, although 1:64 individuals were positive for anti-tTG antibodies (Chin et al., 2009) and when combined with positivity for HLA-DQ2 or DQ8, approximately 1% were positive for both.

Due to the continuous improvement of diagnosis, the documented incidence of CD has also increased in India, for example, 130 cases were reported from 1966-2000 whilst 517 cases were reported from 2001-2005. Hospital based studies in Northern India examining a paediatric patient population estimated prevalence at approximately 1% (Bhattacharya et al., 2009) and the first community based study of both adults and children in Northern India gave a similar result of 1.04% (Makharia et al., 2011). CD appears to be more common in Northern India where wheat consumption is part of the staple diet and has been detected more frequently in children rather than adults (Yachha and Poddar, 2006). There is a paucity of reports of CD in China and Japan, which suggests that CD is either uncommon or under diagnosed in these countries. One study published in 2009 diagnosed 4/62 patients from China with chronic diarrhoea as having CD (Jiang and Wong, 2009).

The prevalence of CD in the Middle East is similar to that found in Europe with rates of approximately 1:166 in Iran (Shahbazkhani et al., 2003), 1:157 in Israel (Shamir et al., 2002), 1:62 in Syria (Challar et al., 2004) and 1:87 in Turkey (Tatar et al., 2004). In the

majority of studies females are diagnosed two to three times more frequently than males (Bai et al., 2005).

CD is now recognised as a worldwide public health problem (Fig 1.3), with the need for better awareness, screening and treatment especially for those in developing countries with a substantial dietary component of wheat, where a gluten free diet (GFD) is almost impossible to sustain.

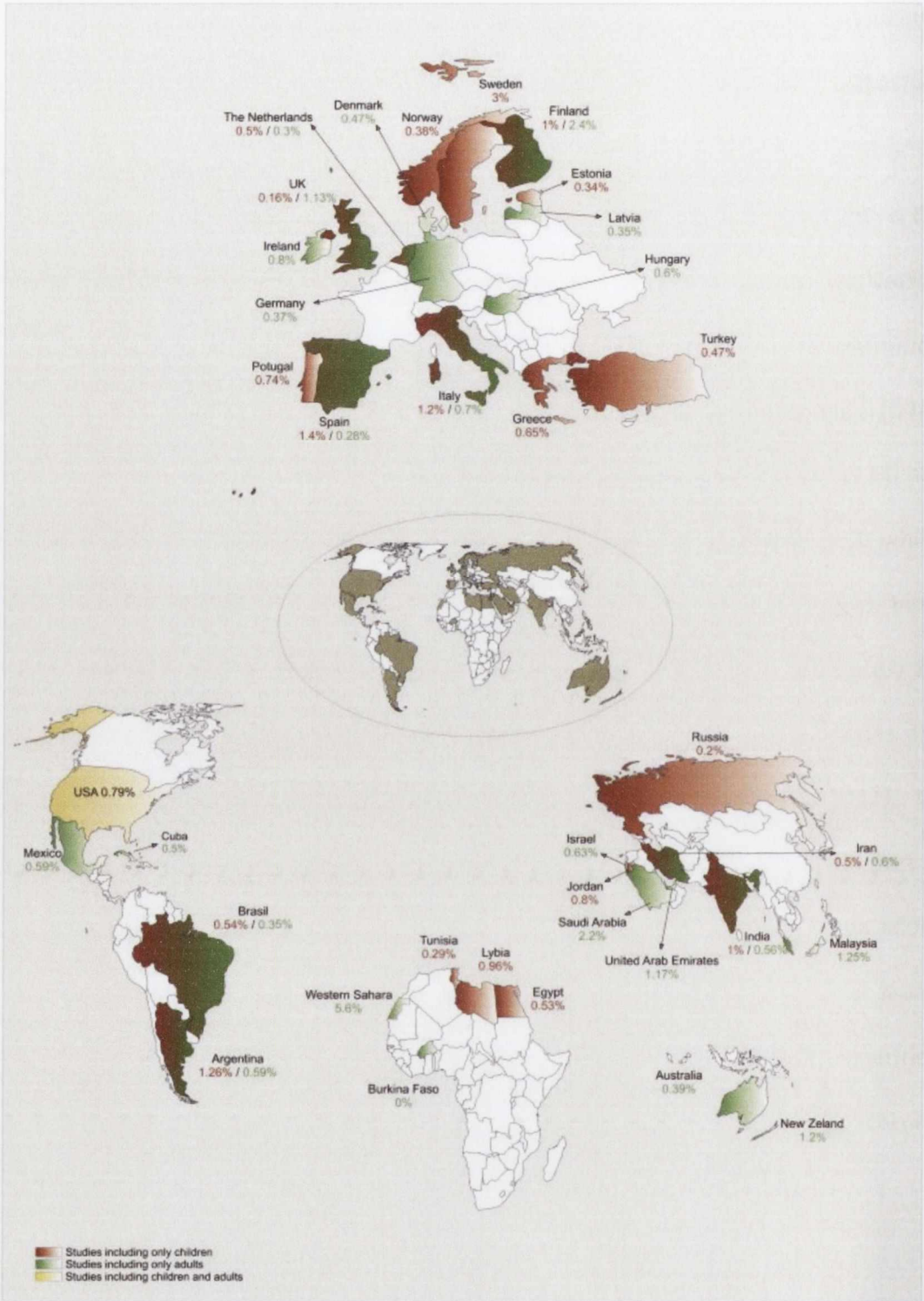


Figure 1.3: Worldwide prevalence of coeliac disease. Data was taken from published studies. Data from children is shown in red, adults in green or both in yellow (Dieli-Crimi et al., 2015).

1.4 Environmental Factors

CD, like type 1 diabetes, rheumatoid arthritis and multiple sclerosis is a multifactorial disease where both genetics and environmental factors are responsible for disease development. Unlike these other autoimmune diseases however, the principal environmental factor that gives rise to CD – gluten – is known. Besides gluten, other environmental factors have also been shown to be associated with the onset of CD.

1.4.1 Gluten

Gluten is the protein fraction of wheat, barley and rye, which is responsible for the induction of CD. The gluten protein itself can be divided into ethanol-soluble prolamins (gliadin) and ethanol-insoluble glutenins. The storage proteins (prolamins) of wheat are enriched in glutamine (>30%) and proline (>15%) (Schuppan, 2000). Similar storage proteins are present in barley (hordeins) and rye (secalines). Several of the gluten epitopes are immune-stimulatory, however some are more vigorous than others. An immune-dominant peptide of α-gliadin has been identified that is composed of 33 amino acids (residues 57-89) that are resistant to degradation by human intestinal, gastric and pancreatic proteases (Shan et al., 2002; Di Sabatino and Corazza., 2009). An inflammatory response is generated against derivatives of this peptide primarily in the upper small intestine in patients with CD. This reaction is mediated by the innate and adaptive immune systems and is characterized by the infiltration of intraepithelial lymphocytes (IELs) and lamina propria lymphocytes (LPLs) to the small intestine alongside villous atrophy. This process will be discussed in more detail in section 1.5.2.

1.4.2 Infant feeding

The subject of the introduction of gluten into the infant diet as a cause for CD has been the source of much debate since the 1980s. In 2008 ESPGHAN stated that gluten should be introduced to the diet before 4 months and no later then 7 months of age and that this should be done gradually whilst the infant is still being breast-fed as this may reduce the risk of CD and other autoimmune diseases (Agostoni et al., 2008). This decision was based on studies by Akobeng *et al* (2006) who found an association between increasing the duration of infant breastfeeding and a decreased risk of developing CD. Furthermore, lessons learned from the Swedish CD epidemic (1984-1996) supported this decision (Myléus et al., 2009). The gluten content of infant food was increased which led to a threefold increase in CD prevalence in children born during this time when compared to children born after the epidemic where gluten was introduced gradually while continuing breastfeeding (Myléus et al., 2009). However, the results of these studies are in contrast to more recent studies which show that neither breastfeeding nor delayed introduction of gluten to the infant diet is associated with the onset of CD (Lionetti et al., 2014). Furthermore, the gradual introduction of small quantities of gluten to the infant diet also failed to reduce the risk of CD (Vriezinga et al., 2014).

1.4.3 Infections

The occurrence of autoimmune diseases following infection has been well documented (Amital et al., 2008; Maya et al., 2008; Cadwell et al., 2010) and it has been suggested that the mechanisms whereby infection can contribute to CD include molecular mimicry or immune modulation. The development of CD following an

encounter with an infectious agent causing intestinal damage has been reported by Shoenfeld *et al* (2008). Adenovirus type 12 (Ad12) is a virus usually isolated from the intestinal tract and has been implicated in the development of CD in certain studies (Kagnoff et al., 1984;Lahdeaho et al., 1993). This link has been proposed due to the discovery that the 54kb E1b protein of Ad12 shares amino acid sequence homology with A-gliadin, a major α-gliadin component known to cause CD as previously mentioned. Therefore, it has been suggested that this viral protein might play a role in the pathogenesis of CD by immunological cross reactivity between surface features shared by Ad12 and α-gliadins. However, the role Ad12 E1B plays in the onset of CD remains controversial as it is also found in the duodenal mucosa of unaffected individuals (Lawler et al., 1994).

Rotavirus infection is a common cause of gastrointestinal infection and inflammation, the frequency of which has been associated with the development of CD in genetically predisposed children (Stene et al., 2006). Serum immuno-globulins from patients with active CD were shown to recognize a peptide that shares homology with the rotavirus serotype 1 major neutralizing protein VP7 and tTG. These antibodies were shown to increase epithelial cell permeability thus enhancing infiltration of gluten peptides into the small intestine (Zanoni et al., 2006).

The treatment for other diseases such as Hepatitis C (Ruggeri et al., 2008) have also been linked to the development of CD. Studies by Cammarota and colleagues (2000) found that using interferon alpha (IFNα) as a drug treatment for hepatitis C infection could trigger CD in certain individuals. IFNα aids the differentiation of Th2 cells to Th1

cells and enhances T-cell cytotoxicity. Such imbalance of cytotoxic/suppressor T-cells may increase immune activity in certain tissues contributing to the onset of autoimmune disorders in certain individuals treated with IFNα.

1.5 Role of the Human Leukocyte Antigen (HLA) locus in CD pathogenesis

1.5.1 HLA molecules

A large body of evidence supports a genetic contribution to the pathogenesis of CD. There is a high degree of familial aggregation with about 5-15% of first-degree relatives being affected and a much larger concordance rate, 83-86% among monozygotic twins compared with 11% in dizygotic twins (Greco et al., 2002). Genetic linkage studies have shown that the disease is most strongly associated with HLA-DQ genes. 90% of CD patients carry a variant of the DQ2 (DQ2.5 - composed of DQA1*0501-DQB1*0201 chains or DQ2.2 - DQA1*0201-DQB1*0202) heterodimer whilst the majority of the remaining patients are carriers of the DQ8 (DQA1*0301-DQB1*0302) heterodimer (Gutierrez-Achury et al., 2011). The HLA-DQ2.5 variant can be encoded on several haplotypes. Both DQA1*0501 and DQB1*0201 are located on the same DR3-DQ2 haplotype and thus the HLA-DQ2.5 heterodimer is encoded in *cis* in this instance. Alternatively, the DQ2.5 variant is encoded in *trans* when these two genes (DQA and DQB) are located on different haplotypes, for example, DR5-DQ7 (DQA1*0505) and DR7-DQ2 (DQB1*0202). Both of these haplotypes produce a similar disease risk (Abadie et al., 2011). However, individuals that carry two copies of the DQB1*02 allele i.e. homozygous for the DR3-DQ2 haplotype or heterozygous for the

DR3-DQ2/DR7-DQ2 haplotype have a particularly increased risk for CD (Sollid and Lie, 2005;Abadie et al., 2011) (Fig 1.4). It is reported that approximately 41% of the genetic susceptibility of CD can be explained by the HLA molecules (Gutierrez-Achury et al., 2015). However, the HLA molecules alone are not sufficient to cause CD as approximately 40% of the general population carry either HLA-DQ2/DQ8, but only 1% actually develops CD. This shows that although these alleles are necessary, they are not sufficient to cause CD, thus supporting the role of environmental susceptibility factors as well as further genetic risk factors in CD pathogenesis.

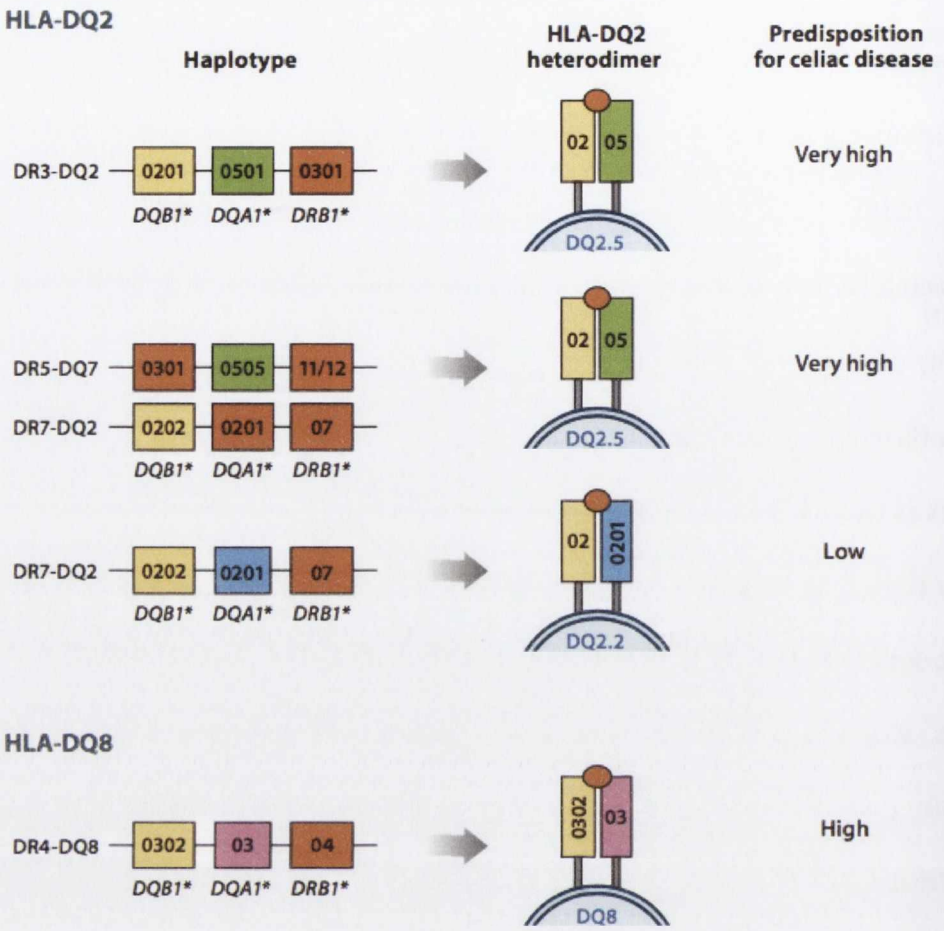


Figure 1.4: HLA associations in CD. The majority of coeliac patients carry the HLA-DQ2 heterodimer. This heterodimer can be encoded by either DQ2.5 (DQA1*0501 – DQB1*0201 chains) or DQ2.2 (DQA1*0201 – DQB1*0202). The HLA DQ2.5 variant can be encoded on several haplotypes, in cis on the DR3-DQ2 haplotype and in trans in individuals who are DR5-DQ7 and DR7-DQ2 heterozygous. The majority of the remaining coeliac patients carry the HLA-DQ8 heterodimer (DQA1*0301 – DQB1*0302) (Abadie et al., 2011).

1.5.2 The Immunological role of the HLA risk variants

The ingestion of gluten in genetically susceptible individuals triggers a cascade of immunological processes that lead to the development of CD. In CD, the HLA class II DQ2 and DQ8 molecules expressed on the surface of antigen-presenting cells bind and present gluten peptides to $CD4^+$ T-cells in the coeliac intestinal mucosa (Lundin et al., 1993). Peptides with negatively charged residues at key anchor positions (P4, P6, and P7) are favoured by the positively charged peptide-binding groove of HLA-DQ2 and HLA-DQ8 molecules (Kim et al., 2004). However, in their native state, gluten peptides lack the negatively charged amino acids preferred by the HLA-DQ2 and HLA-DQ8 heterodimers unless acted upon by tissue transglutaminase (tTG) (Dieterich et al., 1997). tTG is a multifunctional enzyme with numerous physiological roles *in vivo*. During tissue injury it is released into the intestinal mucosa were it plays a role in matrix assembly and tissue repair (Upchurch et al., 1991). However, tTG also has high avidity for gluten peptides in patients with CD and can deamidate glutamine, which has a neutral charge, into the negatively charged glutamic acid (Fig 1.5). tTG therefore allows gluten peptides to bind to the DQ2 and DQ8 molecules with greater affinity to be later recognized by $CD4^+$ T-cells in individuals with CD (Kagnoff, 2005).

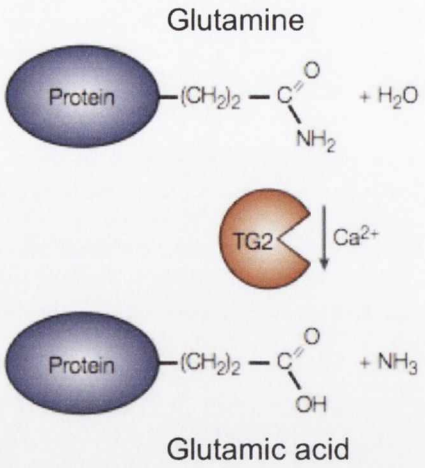


Figure 1.5: Enzymatic action of tTG. tTG is a Ca2+- dependent enzyme that deamidates specific polypeptide bound glutamines into the negatively charged glutamic acid in a reaction involving water (Sollid, 2002).

1.6 Immunological model of CD pathogenesis

Several types of immune cells are required to create the lesions associated with CD. One of these cell types is the gluten reactive $CD4^+$ T-cell, which are found in the peripheral blood and gut mucosa of coeliac patients (Christophersen et al., 2014). They recognise deamidated gluten and initiate a cascade of immune processes that are required for CD pathogenesis. Therefore these cells are sometimes referred to as the gatekeepers/master regulators of the immune response, which lead to CD. These cells infiltrate the lamina propria (LP) of the active coeliac lesion and initiate a Th1 type immunological response with the production of IFNγ and IL21 (Bodd et al., 2010). IFNγ is considered to play a key role in the mucosal damage associated with CD. It can adjust the intestinal permeability and lead to the release and activation of matrix metalloproteinases (MMPs) that can damage the mucosa. Activated gluten reactive

$CD4^+$ T-cells also up regulate the expression of IL15 produced by enterocytes, LP dendritic cells and macrophages, which plays a central role in influencing the activity of the cytotoxic $CD8^+$ IELs. IL15 synergizes with IL21 produced by $CD4^+$ T-cells leading to activation of IELs. IL15 also up regulates the NKG2D and the NKG2C receptors on the surface of IELs and induces epithelial lesions via the interactions of these receptors with their respective ligands MICA and HLA-E expressed on enterocytes (Sollid and Jabri, 2013) (Fig 1.6).

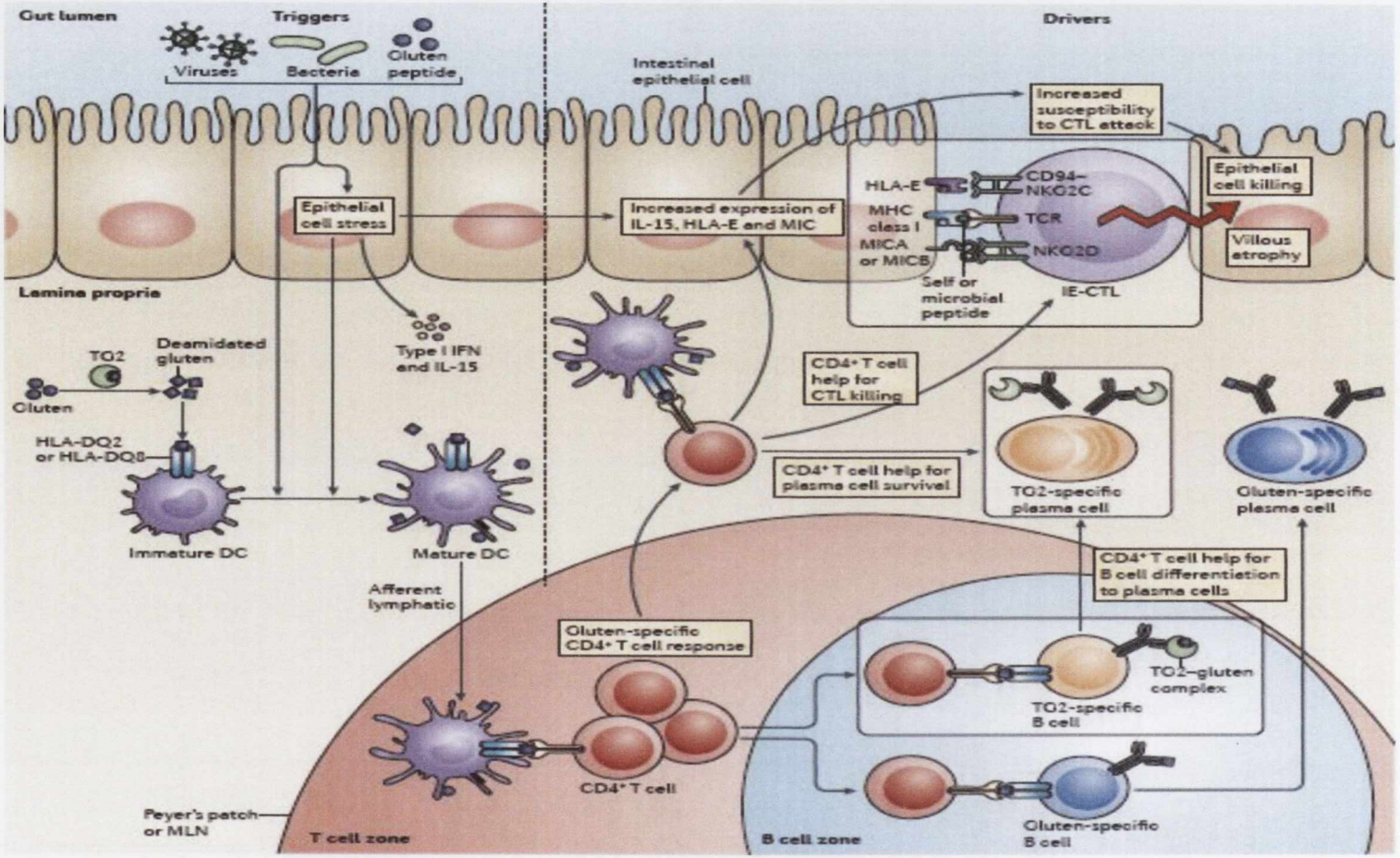


Figure 1.6: Immunological model of CD pathogenesis (Sollid and Jabri, 2013).

1.7 Role of non-HLA loci in CD pathogenesis

As previously mentioned, CD is a multifactorial disease with both environmental (e.g. gluten) and genetic factors required for disease pathogenesis. The HLA-DQ2/DQ8 molecules are necessary but not sufficient to cause the disease as only a small percentage of carriers develop CD; the contribution of HLA locus to the genetic heritability of CD is estimated at ~41%, with the clear implication that the majority of heritable risk factors lie outside the HLA.

1.7.1 Genome-wide linkage studies

Genome-wide linkage analysis has been the traditional means for identifying genes that cause genetic diseases. This involves analysing a group of DNA polymorphisms across the genome in families where the disease phenotype of interest segregates. A number of small families or if sufficiently informative, one large multi-generational family may be used to detect broad genomic regions encompassing disease susceptibility loci (Botstein and Risch, 2003). Traditionally genome-wide linkage analysis was carried out by examining low-density maps of microsatellite markers spaced at regular intervals (5-10cM) across the genome in such related individuals. The earliest linkage studies led to the localization of the Huntingtons disease locus in 1983 (Gusella et al., 1983) and the Cystic Fibrosis (CF) locus on chromosome 7q31 in 1989 (Kerem et al., 1989).

Linkage studies have successfully uncovered genetic variants for hundreds of other Mendelian disorders, which are generally caused by one or, a small number of genes,

which strongly predispose to disease and thus the presence or absence of disease variants strongly corresponds to the presence or absence of the disease. These risk variants are usually rare but have a high-risk effect. In contrast, in complex diseases it is believed interacting genetic and environmental factors play a role in disease pathogenesis, with disease heritability due to a whole array of variants of various frequencies and risk effects. Therefore by utilising the linkage approach identifying rare, high-risk effect variants or private mutations in complex diseases is possible and has been successfully used to locate disease variants in complex diseases, for example the discovery of ADAM33 for asthma (Van Eerdewegh et al., 2002) and NOD2 for Crohn's disease (Hugot et al., 2001).

1.7.2 Genome-wide linkage studies in CD

To date, several genome wide linkage studies have been performed in CD implicating a number of loci that may predispose to CD. Strong linkage to the HLA locus has been consistently identified due to its large effect size (van Heel et al., 2005). A study of Italian families performed by Greco *et al* (1998) identified a locus on chromosome 5q31-33 which was also found to be significantly associated with CD in the Scandinavian (Naluai et al., 2001), Finnish (Liu et al., 2002 ;Koskinen et al., 2009), Hungarian (Koskinen et al., 2009) and Irish populations (Zhong et al., 1996). Holopainen *et al* (1999) discovered the link between CD and the chromosomal region 2q33 in the Finnish population in 1999 and identified *CTLA4* as a strong candidate susceptibility gene at this locus. The *CTLA4* gene is an excellent candidate gene for CD as it is involved in the suppression of immune responses by acting as a negative regulator of T-cell activation and has a strong link with type 1 diabetes (Nistico et al.,

2006). This region was subsequently linked to CD in several other populations including the Irish (Brophy et al., 2005) and UK populations (van Heel et al., 2005) and has ultimately been confirmed in genome wide association studies which will be discussed in further detail in chapter 3. In 2003, Van Belzen *et al* (2003) performed one of the largest studies comprising 82 families from the Netherlands. They reported significant linkage to the chromosomal region 19p13.1. They also identified suggested linkage to the 6q21-22 region, ~70cM downstream of the HLA region. A follow-up study by Einarsdottir *et al* (2011), performed fine mapping of this region in a Dutch population and a replication study in Finnish and Hungarian families showing linkage to the 6q16-22 region. Collectively these results indicate that the 6q21-22 region may be involved in susceptibility to CD. Another study carried out by Einarsdottir and colleagues (2009) was the first to report the linkage of the *IL23R* region to CD. Thus genome-wide linkage studies have been successful in identifying promising CD loci that may contribute to CD pathogenesis. However there is considerable disparity in results from linkage studies and genome wide association studies (GWAS) which are currently unexplained: for example, while the CTLA4 association has been found in both, neither 5q31-33 or 19p13.1 have been found in CD GWAS (www.immunobase.org).

1.7.3 Genetic association Studies

As previously mentioned, linkage studies have successfully identified loci and in some cases genes that exert a high-risk effect in both mendelian and complex diseases. However, since the variants responsible for complex diseases fall within a wide spectrum from rare risk variants with large disease effects to common risk variants with low to moderate disease effects which are less tractable to linkage analysis, it was evident that other study designs would be required to identify lower risk variants in these disorders. This led to the era of genetic association studies where, rather than analysing familial inheritance patterns, the frequencies of disease susceptibility variants are compared between unrelated affected (cases) and unaffected (control) individuals (Altshuler et al., 2008). Association studies have been widely employed in identifying and eliminating suggested susceptibility loci for a range of disorders i.e. blood group O and duodenal ulcers in the 1950's (Clarke et al., 1956) and the HLA locus in several autoimmune diseases such as ankylosing spondylitis (Ryder et al., 1979), type 1 diabetes (Nisticò et al., 1996), and multiple sclerosis (Burwick et al., 2006). However, these studies were based on the candidate gene approach whereby specific 'candidate' genes are investigated based on previous reports of disease association, potential function or simply because they were known and were thus very selective and not at all comprehensive. Furthermore the amount of genetic variation interrogated at these sites was usually very limited meaning that even if there was an association it could be missed. This was inevitable, since very little genomic variation had been catalogued at this time; nonetheless it was apparent that more comprehensive approaches were needed in order to understand the genetic architecture of complex diseases.

1.7.4 Genome-wide association studies (GWAS)

The idea of systematic genome-wide association studies (GWAS), that is to develop a catalogue of common human genetic variants and to test them for association to disease risk became popular in the mid 1990's to try to tackle some of these problems. GWAS provided an alternative and unbiased approach to candidate gene studies for identifying disease-associated variants, as prior knowledge of function was not required. GWAS utilise a set of common sequence variants across the genome that could serve as genetic markers to detect association between a particular genomic region and disease. Common variants were primarily focused on due to the limited genetic variation seen in human populations, with the small number of common variants sufficient to explain the majority of the heterozygosity. Approximately 90% of heterozygous sites in each individual are common variants and characteristically shared among continental populations. The idea that common genomic variants (Minor Allele Frequency ≥5%) may play a role in the onset of common diseases, the "common disease-common variant" hypothesis meant that mapping of these common variants across the genome could be used to identify loci contributing to common disease.

However, it took over ten years to develop this idea into reality, since this required the cataloguing of millions of common variants in the human population. Pivotal steps including the deciphering of the human genome sequence in 2001 (Lander et al., 2001) and the simultaneous development of techniques and analysis methods to affordably genotype these variants in large study cohorts. Initially The SNP Consortium (TSC) and the public Human Genome Project (HGP) built a catalogue of 1.42 million SNPs

(Sachidanandam et al., 2001). As of 2015, this number has grown to 112 million (www.ncbi.nlm.nih.gov/SNP).

However as the SNP catalogue grew it became apparent that testing all of the variants in the genome for disease association would be impossible. Fortuitously however SNPs are commonly found within genomic 'blocks' where there is little evidence of historical recombination and there is limited diversity, known as haplotype blocks. Haplotypes can be described as 'the particular combinations of variant alleles physically arrayed along a chromosome, observed in a population' (Gabriel et al., 2002). The recombination rate within these haplotype blocks is low, so consequently disease alleles within them typically show association with nearby marker alleles for many generations (Altshuler et al., 2008), a process known as linkage disequilibrium (LD). Since much of the genetic variation is resident within common haplotypes, the information content of common SNPs can be derived using a much smaller number of haplotype tag SNPs (Gabriel et al., 2002) meaning that GWAS do not require the testing of all SNPs in the genome known for association with a disease phenotype.

The LD structure was investigated in 2002 by the International HapMap project (International HapMap Consortium, 2005;Gibbs et al., 2003) with the aim of characterising SNP frequencies and LD patterns across the genome of European, Asian and West African population samples. This consortium set up a project designed to build a database of common sequence variation (MAF ≥5%), to determine allele frequencies and to identify LD connections across the genome. The data collected determined that 500,000 SNPs provided substantial power to test >90% of common

SNP variation in out of Africa populations, with 1 million SNPs required in African populations (The International HapMap Consortium, 2005). In addition to cataloging common variation, a database for rare variants was set up by the 1000 Genomes Project in 2008 to identify rare variants (MAF <1%). This included many more populations and sub-populations thereby providing a much more extensive and diverse map of the human genome (1000 Genomes Consortium, 2010). As of 2011, the 1000 Genomes Project had catalogued information on SNP calls, exome alignments and genotypes for 1185 individuals (Buchanan et al., 2012).

The development of such sequence-based reference datasets has since paved the way for the development of a set of genotyping microarrays that could affordably be used to efficiently profile genetic variation in subjects, utilizing between 500,000 and 10^6 SNPs (e.g. Illumina Quad660 and Affy 6.0 genechips).

GWAS have accordingly led to revolutionary advances in the study of complex disease with the first studies published from 2005 (e.g. Klein et al., 2005; DeWan et al., 2006). However the Wellcome Trust Case Control Consortium (WTCCC) published a study in 2007 (Burton et al., 2007), which is largely considered to be the starting point for GWAS of complex diseases (Fig 1.7). It was the first large, well-designed GWAS that took advantage of a SNP chip that had good coverage of the genome. To date, thousands of loci have been identified that contribute to the susceptibility of diseases such as cardiovascular disease (Schunkert et al., 2011), various cancers (Thomas et al., 2008; Pharoah et al., 2013) and several autoimmune diseases, including rheumatoid

arthritis (Eyre et al., 2012), inflammatory bowel disease (IBD) (Rivas et al., 2011), and systemic lupus erythematosus (SLE) (Armstrong et al., 2014) to name a few.

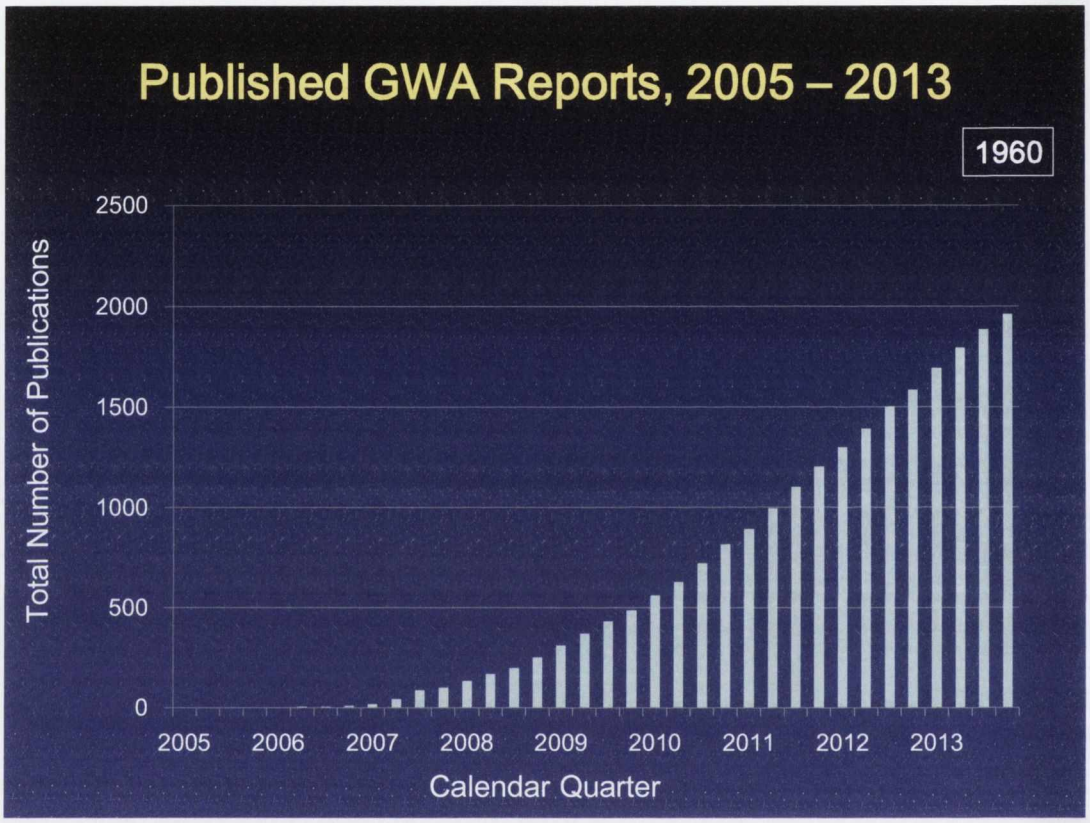


Figure 1.7: GWA publications made between 2005-2013. Data obtained from the Published GWAS Catalog (Hindorff et al., 2015).

1.7.5 GWAS in CD

The first GWA study to be performed in CD was carried out by Van Heel *et al* in 2007. They tested 310,605 SNPs for association in 778 patients and 1,422 controls with follow up of selected variants in the Irish and Dutch populations. Outside the HLA region, they identified risk variants in the 4q27 region containing the *IL2* and *IL21* genes. Hunt *et al* (2008) carried out a follow up study in 2008 were they genotyped

1,020 of the most strongly associated non-HLA markers in an additional 1,643 cases and 3,406 controls from the UK, Ireland and the Netherlands. They identified 7 previously unknown risk variants, *IL12A, IL18RAP, RGS1, SH2B3, TAGAP, CCR3* and *LPP*. A second, larger GWA study carried out by Dubois and colleagues using multiple European populations (Dubois et al., 2010) revealed an additional 13 new CD risk loci, *TNFRSF14, RUNX3, PLEK, CCR4, CD86, BACH2, PTPRK, ZMIZ1, ETS1, SOCS1, ICOSLG* and 2 genes of unknown function bringing the total number of non-HLA CD loci to 26 as of 2010. These loci however contributed only a small increase in the known heritability, thus it was assumed many more non-HLA variants remained to be discovered. Several of the loci identified contained more than one gene therefore finer localization of the risk signal was also required.

Another important factor emerging from GWA studies is that there is a strong overlap between the CD associated genes identified to date and other autoimmune and inflammatory diseases (David A van Heel et al., 2007). CD shares a distinct overlap of genes with rheumatoid arthritis (RA), with *REL, TNFRSF14, ICOS-CTLA4, IL2-IL-21, TNFAIP3* and *TAGAP* common to both diseases. Zhernakova *et al* (2011) also identified a further eight new loci contributing to both the risk of CD and RA in a study of more than 50,000 combined case/control samples, four of which had not been significantly associated with either disease previously (*CD247, UBE2L3, DDX6* and *UBASH3A*) and four genes that had been linked to one of the two disorders in previous studies (*SH2B3, STAT4, 8q24.2* and *TRAF1-C5).* Smyth *et al* (2008). identified seven loci that were present in both CD and type 1 diabetes (T1D), *RGS1, IL18RAP, TAGAP, CCR5, PTPN2, CTLA4* and *SH2B3.* Out of the 64 identified CD GWS associated genes Abadie *et*

al (2011) found using both GO and KEGG pathway analysis that the strongest overlap was between CD and T1D, where 35% of the CD associated genes were also found in T1D pathways (FDR = 1.13×10^{-7}). *CTLA4, ICOS, BACH2, PTPN2, SOCS1* and the *KIAA1109/ADAD1/IL2/IL21* haplotype block are just some of the genes overlapping between these two autoimmune diseases. They also found a 2% overlap between CD genes and inflammatory disease pathways such as Crohn's disease and Ulcerative Colitis. As many as 14 loci are shared between CD and inflammatory bowel disease (IBD), including the *IL2/IL21* gene region on ch4q21 (van Heel et al., 2007; Festen et al., 2009). Four genes, *IL18RAP, PTPN2, TAGAP,* and *PUS10* have also been shown to overlap between CD and Crohn's disease (Festen et al., 2011).

1.8 Immunochip high-density array

In 2011, the Immunochip Consortium was developed to explore some of the unresolved issues regarding the genetic disposition to CD. These issues included trying to explain the 'missing' heritability, to identify additional common risk variants, to attempt to identify rare variants and to try to pinpoint causal variants, or at least finely localise the risk signal.

Early GWAS chips particularly, consisted of a low density of SNPs to interrogate the genome resulting in associations of low resolution, making it difficult to pin point the exact causal SN responsible for disease pathogenesis. The human genome consists of more than 3 billion basepairs whilst early genotyping chips only contained in the region of 300,000 and more latterly, more than 500,000 unambiguous SNPs. In 2011

the Wellcome Trust Case-Control Consortium (WTCCC) promoted the development of the Immunochip, a custom Illumina Infinium high-density array, to allow the use of a common high density set of probes targeted at known susceptibility loci identified from various datasets (e.g. 1000 Genomes project and disease specific resequencing data) containing common, low frequency and rare variants (Cortes and Brown, 2011). It contained 196,524 polymorphisms from 186 loci reaching genome-wide significance in a range of autoimmune diseases including rheumatoid arthritis, systemic lupus erythematosus, type 1 diabetes, ankylosing spondylitis, multiple sclerosis, autoimmune thyroid disease, Crohn's disease, psoriasis, ulcerative colitis and coeliac disease. Approximately 3000 GWA SNPs from each disease are present on the chip facilitating deep replication and fine mapping of disease associated loci for both common and rare variants. Importantly it allows cross comparison of diseases using a uniform marker set which will allow the question of whether the same or different variants affect genes linked to multiple diseases to be addressed.

The Immunochip is much more cost effective than other GWAS chips, with the positive effect of enabling genotyping of larger cohorts than could previously have been achieved. In 2013, the International Genetics of Ankylosing Spondylitis Consortium used the Immunochip to genotype 10,619 affected individuals and 15,145 controls identifying 13 new risk loci and 12 additional ankylosing spondylitis-associated haplotypes (International Genetics of Ankylosing Spondylitis Consortium, 2013). In CD, the Immunochip was used by Trynka *et al* (2011) in which 12,041 affected individuals and 12,228 controls were genotyped identifying 13 new CD risk loci bringing the total to 40 including the HLA locus. This comprised 57 independent non-HLA signals, the

majority of which are common with frequencies above 5% and moderate effect sizes with odds ratios between 1.124 and 1.360. Due to the higher density of SNPs present on the Immunochip, Trynka *et al* were also able to refine the association signal to a single gene for 29 of the 39 non-HLA loci identified at that time. However the Immunochip also has its limitations, in that it does not cover the entire genome and relies on previous GWAS studies for its markers.

As mentioned following the Trynka *et al* (2011) study a total of 39 non-HLA loci had been identified (Fig 1.8). Subsequently, a recent study by Garner and colleagues (2014) identified *FRMD4B* as a coeliac susceptibility locus and findings from this study (Coleman et al., 2015) have identified a further two coeliac susceptibility loci, *NIFA*, and *ZNF335* bringing the total number of known coeliac loci, including the HLA locus to 43. The 43 non-HLA genome-wide significant loci encompass in excess of 70 candidate genes.

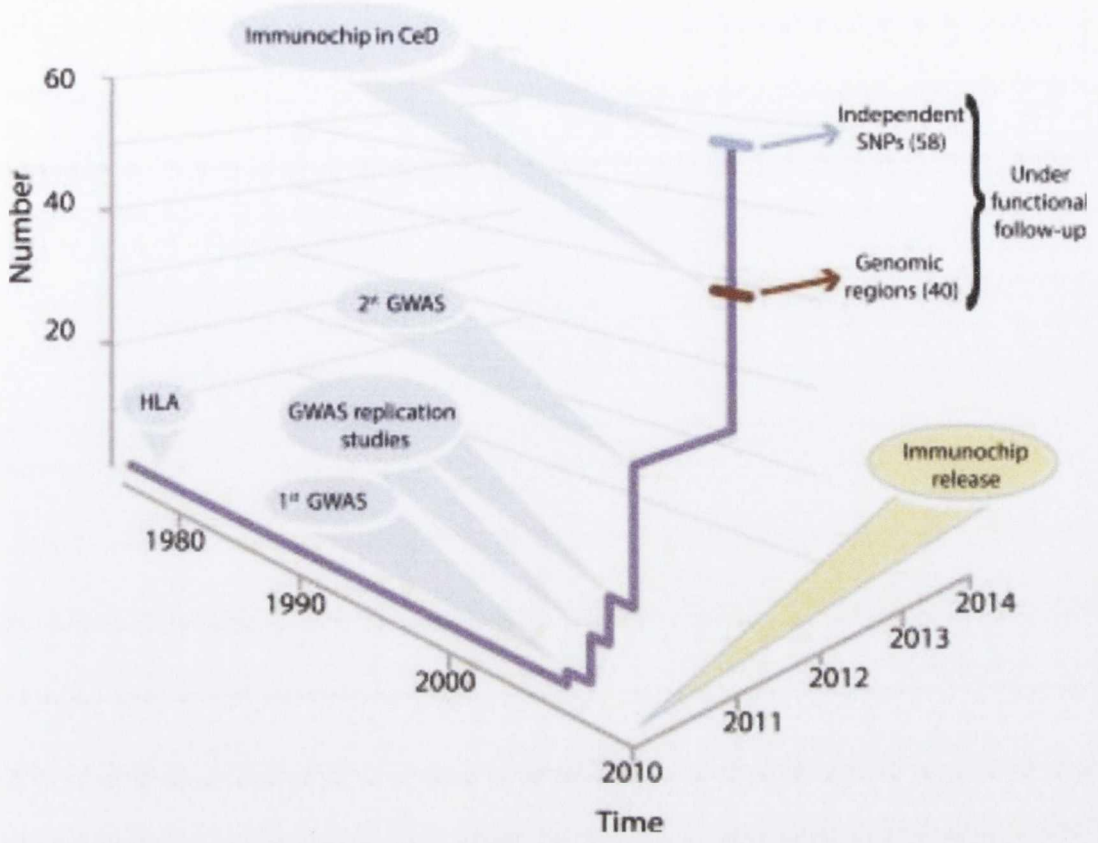


Figure 1.8: Chronological line highlighting the number of SNP discoveries that have been discovered since the first GWAS in 2007 until 2014 (Ricaño-Ponce et al., 2015).

It has commonly been reported that approximately 30-50% of the heritability of CD can be explained by the HLA locus and the 39 non-HLA loci (Kumar et al., 2012; Lundin and Sollid, 2014; Gutierrez-Achury et al., 2015). Therefore all of the CD associated SNPs discovered by GWAS to date cannot account for the total genetic variation of the disease, which begs the question, where is the remaining missing heritability?

It was previously thought that rare variants of large effect were contributing to the missing heritability not covered by common GWAS SNP arrays (Gibson, 2012). However a recent study by Hunt *et al* (2013) found no evidence that the missing heritability for common autoimmune diseases can be accounted for by rare variants but perhaps is the result of many common-variant loci of weak effect. Currently a SNP must reach a stringent threshold of $P< 5x10^{-8}$ to be considered genome-wide significant (GWS) and with many studies underpowered to achieve this threshold for SNPs with modest effect sizes on disease risk, many true positive associations remain unidentified. Yang *et al* (2010) showed that 40% of the phenotypic variance for height could be explained by common SNPs not reaching GWS and subsequently Stahl *et al* has indicated that thousands of non GWS SNPs from coeliac GWAS could explain an additional 43% of the disease risk (Stahl et al., 2012). Therefore, it has been debated that most of the heritability may not in fact be missing; it may just have failed to be detected due to the individual effect sizes of SNPs being too small to reach genome-wide significant thresholds given current study sizes.

It has been proposed that the total heritability of diseases has been overestimated generating what has been termed "phantom heritability". This overestimation is

derived from the fact that estimates of total heritability assume that there are no genetic interactions (epistasis) present among loci (Zuk et al., 2012). In fact, several recent studies have found evidence that epistasis does occur in complex diseases and contributes to disease pathogenesis and has been demonstrated in both psoriasis (Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, 2010) and multiple sclerosis (Lincoln et al., 2009). Therefore the total heritability may be considerably less when genetic interactions are accounted for, and thus the proportion of heritability explained much larger. A very recent study by Yang *et al* (2015) examining the missing heritability in human height and BMI has also further validated this idea of 'phantom heritability'. They show that previous estimates of heritability in these traits were overestimated in family studies and that their current results represent the maximum amount of heritability obtainable for these traits leaving little scope for other sources of missing heritability.

1.9 Polygene Risk Score (PRS) Analysis

As previously mentioned, there is a gap between what is believed to be the actual genetic component of complex diseases and what is explainable given traditional GWAS and our current assumptions regarding heritability. GWAS examine common (MAF ≤5%) rather than rare variation (MAF ≤1%) (Manolio et al., 2009) and their capacity to detect associated polymorphisms is determined by the risk effect of the SNP and the size of the study populations. Thus polymorphisms with low effect sizes have not been detected due to the currently unattainable and perhaps unrealistically large sample sizes required to achieve genome wide significance. In the largest CD

study to date (with over 12,000 patients and a similar amount of controls) the lower limit of effect size passing GWS is of the order of 1.12 (Trynka et al., 2011). Although low, it is possible that SNPs with even lower odds ratios can contribute some of the remaining heritability and thus play a significant role in disease pathogenesis.

In 1910, East proposed that commonly varying traits may be polygenic in nature and by 1918, Fisher proposed a polygenic model of disease whereby the genetic architecture underlying disease susceptibility involves a wide range of allele frequencies and effect sizes (Fisher, 1919; Kempthorne, 1968) and developed a pipeline for linking Mendelian factors with quantitative traits. With this polygenic model of disease in mind, studies have shown that by combining loci of small effect, including those that do not achieve genome wide significance, it is possible to explain a greater proportion of the variance in a range of complex traits such as schizophrenia (Purcell et al., 2009) and multiple sclerosis (The International Multiple Sclerosis Genetics Consortium (IMSGC), 2010). This approach has greatly improved the ability to determine the overall contribution of common variants to disease risk. Two statistical approaches that have been developed to determine this genetic liability are the *polygenic risk score* (PRS) described by Purcell and colleagues (2009) and the *mixed linear modeling* method (Yang et al., 2010). We have used the PRS in this study. The PRS method combines the effects of all common variants which are found at different frequencies in the cases and controls to give a score which indicates the likelihood of the individual being either case or control and can also determine the total amount of heritability accounted for by the SNPs analysed. The major benefit of this method is that any SNPs can be included, regardless of whether they surpass GWS or not. It has

been successfully used to examine the influence of risk alleles *en masse* to disease susceptibility in a range of complex disorders. Romanos *et al* (2014) has developed a similar approach to PRS in CD developing a coeliac genetic risk model by combining HLA plus the 57 non-HLA SNP associated loci and Abraham and colleagues (2014) who developed genomic risk scores for CD based on approximately 250 SNPs. Our study differs from these in that we have used the Immunochip as the basis for our study; this is an important and widely used genotyping platform and thus our findings could be relevant in other studies where this platform is used. Polygene risk scores and their relationship with CD will be discussed further in chapter 2.

1.10 Functional Assessment of Genetic Associations

GWAS have allowed for the identification of multiple disease-associated loci involved in many complex diseases. The challenge now facing researchers is to interpret and understand the biological impact of these variants in complex diseases. Genetic variation combined with environmental and epigenetic factors may contribute to alterations in gene expression levels, including genes whose perturbation may undoubtedly lead to different phenotypic traits, including susceptibility to disease. It has now been shown that ~90% of genetic variants are found in the non-coding regions of the genome and are thus considered likely to exert a regulatory effect (Freedman et al., 2011). Genetic variants with regulatory effects may thus play a very large role in influencing disease susceptibility through their impact on gene expression patterns. Therefore gene expression profiling of complex diseases is necessary to further elucidate the genetic architecture of such diseases.

1.10.1 Microarray technology

The high-throughput and rapid quantification of the transcriptome became possible with the development of gene expression microarrays (Pickrell et al., 2010). They quantify gene expression levels using one or a few probes targeting specific sections of each gene. By 2010 the Gene Expression Omnibus (GEO) contained more than 520,000 individual experiments, most of which were produced using microarrays (Malone and Oliver, 2011). Microarrays have been used to determine gene expression profiles of patient samples for disease predisposition, diagnostics and personalized treatments in a range of diseases such as cancer (Alizadeh et al., 2000) (Colombo et al., 2011), Alzheimers disease (Pasinetti, 2001), schizophrenia (Miklos and Maleszka, 2004) and inflammatory bowel disease (IBD)(Costello et al., 2005) to name a few. There have been a small number of gene expression studies carried out in CD using different microarray platforms (Diosdado et al., 2004; Bracken et al., 2008). Despite there being many successful applications of microarray technology, they are not without their limitations. These include high background noise, cross-hybridisation between genes, low dynamic range making detection and quantification of low-abundance transcripts difficult and their inability to detect novel transcripts due to their reliance on previously characterised genes for the array development (Shendure, 2008).

1.10.2 RNA-sequencing

The subsequent development of high-throughput sequencing technologies such as RNA-seq has surpassed microarray technology as the gold standard for transcriptome

profiling in recent years. It has several advantages over the traditionally used microarray technology. Firstly, it can detect novel transcripts not only from existing genomic sequences, it can identify exons and exon boundaries to a single base resolution, thus identifying alternative splicing events (Jiang and Wong, 2009). RNA-seq also has low background signal as the vast majority of DNA sequences can be mapped specifically to individual regions of the genome and has a large dynamic range over which transcripts can be detected, well beyond the range achieved by standard microarrays (Mortazavi et al., 2008). RNA-sequencing has been used to examine the gene expression profiles of many complex diseases including Alzheimers disease (Twine et al., 2011), psoriasis (Li et al., 2014) and many cancers including breast (Edgren et al., 2011) and prostate (Nacu et al., 2011).

1.10.3 Affymetix GeneChip Human Transcriptome Array (HTA) 2.0

In 2013, Affymetrix developed the GeneChip® Human Transcriptome Array (HTA) 2.0 which they claim represented a big improvement on previous microarrays. It purportedly is the highest resolution microarray currently available for gene expression profiling, containing >6 million probes, covering all known coding and non-coding transcript isoforms. Affymetrix claim it has several advantages over RNA-seq, such as generating accurate gene expression data regardless of transcript abundance whilst RNA-seq depends on both the depth of sequencing as well as expression level to generate precise gene expression data. This array is also highly cost efficient compared to RNA-seq. The HTA 2.0 array will be discussed in more detail in chapter 4.

1.10.4 Expression Quantitative Trait Loci (eQTL) analysis

As mentioned, GWA studies have identified multiple risk loci associated with CD but the identification of the exact causal gene and SNP(s) within these loci remains elusive. More than 90% of SNPs associated with complex diseases (certainly inflammatory diseases) map to intergenic regions, introns and 5' or 3' untranslated regions (UTRs) as discussed earlier. Specifically, for immune-mediated diseases only 2.6% of SNPs analysed map to exons, whilst 2.1% map to UTRs, 42.3% to introns and 53% to intergenic regions (Ricaño-Ponce and Wijmenga, 2013). They are therefore assumed to play a role in gene regulation, most obviously by altering gene regulatory elements resulting in the abundance of the gene transcript being modified. Indeed, approximately 50% of the genetic variants associated with immune mediated diseases identified thus far have actually been demonstrated to be associated with the expression of nearby genes (Dubois et al., 2010). These regulatory polymorphisms capable of modifying gene expression are known as expression quantitative trait loci (eQTL). If gene expression is treated as a quantitative trait, it is possible to correlate genotypes of disease associated SNPs with transcription levels of the genes in a process known as eQTL mapping (Gilad et al., 2008). eQTLs can exist as either *cis*-eQTLs or *trans*-eQTLs depending on their physical location with respect to the regulated gene. In general, if the genetic variability at a particular locus affects proximal genes they are known as *cis*-eQTLs, whereas if the genetic variability affects distal genes, they are said to be *trans*-eQTLs. eQTL analysis can therefore help to prioritise specific causal genes from regions of the genome containing more than one gene. A small number of eQTL studies have been carried out in CD to date, with Hunt *et al* (2008) identifying the first eQTL gene related to CD, *IL18RAP*. Dubois *et al* (2010)

also investigated the effects of common variants on the expression of nearby genes in CD. They assessed 38 genome-wide significant and suggestive non-HLA CD risk loci for *cis* expression genotype correlations from 1,469 unselected human whole blood samples. Their results showed that 38% of CD SNPs were *cis*-eQTLs, a result significantly higher than would be expected by chance. The strongest eQTL was found for SNP rs917997 linked to the *IL18RAP* gene ($P = 7.4 \times 10^{-87}$) validating the previous eQTL identified by Hunt *et al*. There was a 9-fold increase in IL18RAP expression in carriers of two wild type alleles. eQTL mapping in this instance was successfully able to isolate the causal gene as *IL18RAP* is a gene present in a locus containing multiple genes including *IL18R1, IL1RL1* and *IL1RL2,* none of which displayed a *cis*-eQTL effect. eQTLs have also been demonstrated in Crohn's disease (Fransen et al., 2010), asthma (Moffatt et al., 2007) and many more complex diseases. Whilst eQTLs do not always prove causality of a specific variant, they help to identify plausible candidate genes that can be taken for further analysis.

However, the eQTL analysis carried out in these studies used the expression data derived from whole blood samples, which are composed of heterogenous cell populations creating high levels of transcriptome background noise making the detection of the true signal more complex. Consequently there have been several studies carried out highlighting the importance of cell and tissue type specificity for the detection of *cis*-eQTLs (Fu et al., 2012; Amundsen et al., 2014). Adding another layer of complexity to the detection of causal risk variants is the discovery that many eQTLs are also stimulous dependent (Fairfax et al., 2014; Plaza-Izurieta et al., 2014). These observations emphasise the importance of carrying out studies in the cellular and

environmental context of interest. eQTL analysis and the findings of these studies will be discussed further in chapter 4.

1.10.5 Pathway and network analysis

GWAS and high throughput sequencing technologies such as RNA-seq have revolutionised our understanding of the genetic architecture of many complex diseases however this has generated enormous amounts of data much of which remains to be processed to give insights into underlying biology of the disease. Thus a wide variety of data analysis tools have been developed to derive biologically meaningful information from associated genetic variants and lists of differentially expressed genes. As mentioned previously, most risk loci identified to date contain multiple variants and to add further complexity, these SNPs are very often in high LD with other SNPs that may be involved in disease pathogenesis. Therefore these methods can be extremely useful for filtering large numbers of variants into smaller, more meaningful pools of variants and/or identifying the causal variant involved in disease. Many of such analytical tools allowing the linking of genes and markers to functional information are publically available and include Gene Relationships Among Implicated Loci (GRAIL), pathway analysis tools such as Kyoto Encylopedia of Genes and Genomes (KEGG), the Gene Ontology (GO) database and the Biomolecular Interaction Network Database to name a few. More elaborate pathway and network analysis methods have been developed that incorporate the enrichment of gene sets using RNA or protein-interaction networks and sub networks (modules) enriched with associated genes. Such methods include Weighted Gene Co-expression Network Analysis (WGCNA), Partial Correlation and Information Theory (PCIT) and Gene Set

Enrichment Analysis (GSEA). Other tools used in this thesis include Interval-based Enrichment Analysis Tool for Genome Wide Association Studies (INRICH) which we used to evaluate the enrichment of CD associated regions for differentially expressed genes and dapple which looks for protein interactions between specified gene sets. These methods will be discussed in more detail in chapters 2 and 3.

To date, pathway analysis in CD has implicated multiple immunological pathways, as one would expect given the nature of the coeliac lesion. Some of these pathways include chemokine receptor activity, cytokine binding, T-cell and NK cell activation (Abadie et al., 2011). A study by Kumar and colleagues (2012) in CD confirmed that CD genes were enriched for a number of cellular pathways including T-cell, NK cell and B-cell pathways and as described earlier, there is significant overlap between associated regions in a variety of inflammatory diseases and other regions that are unique to specific diseases.

1.11 Conclusion

The genetic landscape of CD has changed radically in the last decade since the discovery of SNPs and development of array-based genotyping technology. GWAS have provided a hypothesis free method of associating genetic variants and CD and as we have seen the hunt for genes contributing to this disease has been particularly successful with 43 genome-wide significantly associated loci identified to date. However despite these major advancements a large proportion of the genetic variation of the disease remains to be identified. What has become clearer is that CD is a polygenic disorder in which many common alleles with small to moderate effect sizes play a role in disease phenotype, with the vast majority of variants found in non-coding regions of the genome. Therefore future post-GWAS studies of CD need to focus on the development of risk prediction models and functional analyses of the genes and biological pathways regulated by these genetic variants to begin identify the causal mechanisms behind the pathogenesis of CD. In this study we wished to investigate the genetic architecture in CD in the Irish population and we show a high degree of correlation with other European populations. We also sought to investigate if inherited differences in gene expression were evident between individuals with coeliac disease and unaffected individuals. We show substantial differences between these groups and use some of the methods outlined above to interpret this data. We identify distinctive biological pathways, which seem particularly prominent in the gene expression data of the cases giving clues to mechanism. We amalgamated these data sets and show significant enrichment of genes from CD associated regions of the genome among our differentially expressed genes, providing functional corroboration of the genetic findings. We also wished to extend our data sets to allow for the

analysis of eQTLs in our study samples however this part of the project was not completed successfully.

Here follows a specific statement of the aims and objectives of this study.

1.12 Aims and Objectives

In this study we wished to

1. Investigate the genetic architecture of coeliac disease in the Irish population using the Immunochip SNP array.

2. Investigate whether gene expression in peripherally derived $CD4^+$ T-cells from individuals with CD differs from that of individuals without CD.

3. To combine gene expression data with genotype data to carry out eQTL analysis.

4. To utilise pathway and network analysis tools to examine gene expression data to identify biological processes involved in CD.

Chapter II

Immunochip & Polygene analysis

2.1 Introduction

The HLA locus is the main genetic factor contributing to disease pathogenesis. 99.6% of all coeliac patients carry either one of two types of the HLA-DQ heterodimer, HLA-DQ2 or HLA-DQ8. However, up to 40% of the general population may also express these HLA heterodimers and never develop CD, therefore we can say that the HLA is necessary but not sufficient to cause the disease. The HLA locus has variously been estimated to account for between 20-41% of the genetic variance of the disease (Gutierrez-Achury et al., 2015), with the remaining genetic variance assumed to be resident in non-HLA genomic locations.

To date, eleven genome-wide linkage studies have been carried out in CD alongside numerous region specific linkage studies in the attempt to identify non-HLA CD genetic risk factors. These studies provided evidence of linkage to several regions of the genome, such as chromosome 5q31-33 and chromosome 19p13.1 with the most promising discovery being made on chromosome 2q33, which encompasses the CTLA4/CD28 gene region (Holopainen et al., 1999). However with the exception of the CTLA4/CD28, many of these findings have not been replicated in subsequent genetic association studies and thus whether they represent true coeliac susceptibility loci has not been established.

In recent years there have been multiple genome-wide association studies (GWAS) carried out to identify these additional genetic risk factors, which are believed to be distributed among an unknown number of non-HLA loci, each of which is thought to contribute only a small effect. As previously described (section 1.7.5), the first GWAS

for CD was performed in a UK sample of 778 patients and 1,422 controls and led to the first unequivocally identified non-HLA risk locus, the IL2/IL21 region of chromosome 4. Van Heel *et al* (2007) identified this region as approaching GWS ($P = 2 \times 10^{-7}$) and with subsequent follow up studies in separate Dutch and Irish cohorts they brought the signal up to genome wide significance levels specifically for this region ($P = 4.8 \times 10^{-11}$, OR = 0.66, CI (0.58-0.74)). A subsequent replication study by Hunt *et al* discovered a further 13 CD susceptibility regions outside of the HLA (Hunt et al., 2008). This was followed up in 2010 by Dubois *et al* were they performed a much larger CD GWAS, identifying another 13 previously unknown CD regions, bringing the total number of established non-HLA CD susceptibility loci identified to 26 (Dubois et al., 2010).

However, many variants might have remained undetected in these previous GWAS due to their low marker density and thus early GWAS SNP arrays failed to adequately capture much of the genetic variance of the CD associated loci. Consequently, associated loci were generally poorly defined and spanned large regions, making it difficult to identify the location of the true risk signal. The Immunochip Consortium (Cortes and Brown, 2011) was set up to rectify this problem by designing a new custom array, using emergent datasets of common, low frequency and rare variants to interrogate immune-related disease susceptibility loci. Initiated by the Wellcome Trust Case-Control Consortium (WTCCC), the Immunochip is a custom Illumina Infinium high-density array that contains 196,524 common and rare polymorphisms (718 small insertion deletions and 195,806 SNPs) designed to perform dense genotyping of previously established GWAS significant loci from the major autoimmune and inflammatory diseases. SNPs identified by the 1000 genomes project, and other

sequencing data available to the consortium (Trynka et al., 2011), are included on this array. 186 loci, comprising loci reaching genome-wide significance ($P < 5x10^{-8}$) from 12 diseases were selected by the Consortium. These diseases are: Crohn's disease, ulcerative colitis, type-1 diabetes, rheumatoid arthritis, psoriasis, systematic lupus erythematosus, primary biliary cirrhosis, ankylosing spondylitis, auto-immune thyroid disease, IgA deficiency, multiple sclerosis and coeliac disease. Wildcard SNPs from the 12 disease groups and SNPs from non-immunological diseases identified by the WTCCC2 were also included on the chip (Cortes and Brown, 2011).

As well as allowing fine mapping of established CD susceptibility loci, Immunochip analysis has been successful in identifying novel risk loci for a number of complex diseases, including primary billary cirrhosis (PBC)(Juran et al., 2012), autoimmune thyroid disease (AITD)(Cooper et al., 2012), bipolar disorder (Green et al., 2012). Trynka and colleagues (2011), using the Immunochip in a study of 12,014 CD patients and 12,228 controls from seven geographic regions, reported 57 independent celiac disease association signals across 39 non-HLA loci. These consisted of loci meeting genome-wide significance ($P < 5x10^{-8}$) in their data and/or in previous CD GWAS and replication data sets, and included 13 loci not previously associated with CD. This brought the total number of CD risk loci to 40 including the HLA locus. Similar to other complex diseases, the individual effect size for most of these observed variants is small (OR 1.12-1.36) and together they only help explain a small proportion of disease risk. Therefore, despite major advances with the development of the Immunochip, for most if not all-complex traits studied to date, only a modest proportion of the genetic heritability is accounted for, leaving us still with the enigma of the 'missing heritability'.

Heritability is defined as 'the portion of phenotypic variation in a population attributable to additive genetic factors'.

Many rare variants that may contribute to disease pathogenesis may be absent from the Immunochip because they failed to reach genome wide significance in the early discovery studies. It also fails to type rare copy number variants (CNVs) and perhaps most importantly, the Immunochip does not cover the whole genome so therefore relies exclusively on the power of previous GWA studies for its marker selection and thus could miss other potentially associated CD loci. The requirement that a SNP must achieve a strict threshold of $P<5x10^{-8}$ to be considered GWS is also likely to be part of the problem as even with current sizes, GWAS are without doubt underpowered to reach GWS for SNPs with small effect sizes on disease risk as can be seen from studies of schizophrenia and anthropomorphic characteristics where study sizes are considerably larger.

One of the ways of approaching these issues and trying to explain the remaining missing heritability of CD is to combine multiple genetic markers not achieving genome wide significance to generate a risk score associated with disease status. This method, known as polygenic risk score (PRS) analysis, can help to explain a substantial proportion of heritable phenotypic variation in complex disease (Dudbridge,2013).

Multiple statistical models have been developed to determine the effects non-genome-wide significant SNPs have on complex diseases; these include machine learning approaches (Abraham et al., 2014), mixed linear modelling (Yang et al., 2010)

and polygenic analysis (Purcell et al., 2009). We used the latter method in our study, which is discussed in more detail here.

It has been shown that by combining loci of small effect, including those that do not achieve genome wide statistical significance, it is possible to explain a greater proportion of the variance in complex traits than by excluding non-significant loci. The polygenic risk score (PRS) described by Purcell and colleagues (2009) is one such approach. The PRS combines the effects of multiple common variants to derive a risk score for any individual in the study and has been successfully used to examine the influence of risk alleles *en masse* to disease susceptibility in a range of complex disorders such as psychiatric illness (Purcell et al., 2009), multiple sclerosis (The International Multiple Sclerosis Genetics Consortium (IMSGC), 2010), and cardiovascular disease (Dudbridge, 2013). In essence, the PRS is a risk score for an individual calculated from the cumulative burden of risk alleles carried by that individual; these risk alleles are then weighted according to their observed effect size on disease (odds ratio) to come up with the final PRS. In CD, similar approaches to PRS have been applied by Romanos and colleagues (2014), who developed a CD genetic risk model by combining the HLA plus the 57 non-HLA disease associated loci and Abraham and colleagues (2014), who developed genomic risk scores for CD based on approximately 250 SNPs.

Therefore we wanted to use the PRS approach to examine whether adding non-GWS; non-HLA susceptibility variants to the known HLA variants could improve CD risk prediction in our Irish population cohort. Considering our Irish study lacked sufficient

power, we also wanted to compare our Irish population cohort to a much larger dataset generated by Trynka *et al* (2011) to determine if risk variants identified in our study shared the same direction of effect.

2.2 Aims and objectives

1. The aim of our study was to examine genetic associations in CD in the Irish population; to replicate the findings from the large Immunochip analyses carried out by Trynka and colleagues; and to examine whether the risk alleles identified by Trynka *et al*, were similar in the direction and size of effect.

2. We also wanted to test, by performing a meta-analysis, if our data could strengthen the association signals to surpass the level of genome wide significance of any loci nominally associated in the larger study.

3. To examine the polygenic contribution to CD, by deriving polygene risk scores for HLA and non-HLA CD loci identified and weighted according to the findings from the Trynka *et al* association study (discovery set) and calculated the proportion of variance explained for our Irish study and to investigate the discriminatory capacity of the PRS derived from our study to distinguish case-status in this sample.

2.3 Materials and Methods

2.3.1 Subjects

All study individuals are Caucasian and have self-reported Irish ancestry as defined by four grandparents of exclusively Irish origin. Genomic DNA was extracted from peripheral blood samples. CD patients were assigned on the basis of positive serology to antibody-based tests (anti-tTG and anti-EMA) followed by examination of the histological appearance of small intestinal biopsy samples and noting clinical response to a gluten-free diet. Controls were an anonymous unselected (i.e. not screened for CD) sample from the general population provided by the TCD/IMM population DNA biobank. Written and informed consent was obtained from all subjects, and the study was approved by the Research Ethics Committees of St James's Hospital Dublin and all other contributing centres.

2.3.2 DNA sample preparation

Genomic DNA for all cases and controls was isolated by the TCD biobank, using the Gentra nucleic acid extraction system (QIAGEN, Dusseldorf, Germany). All DNA stock samples were standardised to a concentration of 100ng/ul.

2.3.3 Preparing the plates for Immunochip anaysis

5ul of DNA was added to 5ul deionized H_2O in each well of a 96 well plate giving a DNA concentration of 50ng/ul. In total 959 individuals, 480 cases and 479 control samples were genotyped using the Immunochip platform.

2.3.4 Workflow of Immunochip assay for whole-genome genotyping

1. A whole genome amplification step increases the amount of DNA up to 1000 fold and the DNA is then fragmented.

2. The DNA is captured on the beadchip. The beadchip contains many copies of an oligonucleotide probe that targets a specific locus in the genome. As genomic DNA passes over the beadchip, each probe will bind to a complementary sequence stopping one base before the locus of interest.

3. Allele specificity is conferred by a single base extension that incorporates one of four labeled nucleotides.

4. Once laser excited, the nucleotide label emits a signal that's detected by a scanner. Intensity values given for each colour convey information about the allelic ratio of a given locus. When the intensity levels for both alleles are the same, this is indicative of a heterozygote genotype and when the intensity values for one allele are stronger then the other allele this is indicative of a homozygote genotype (Fig 2.1).

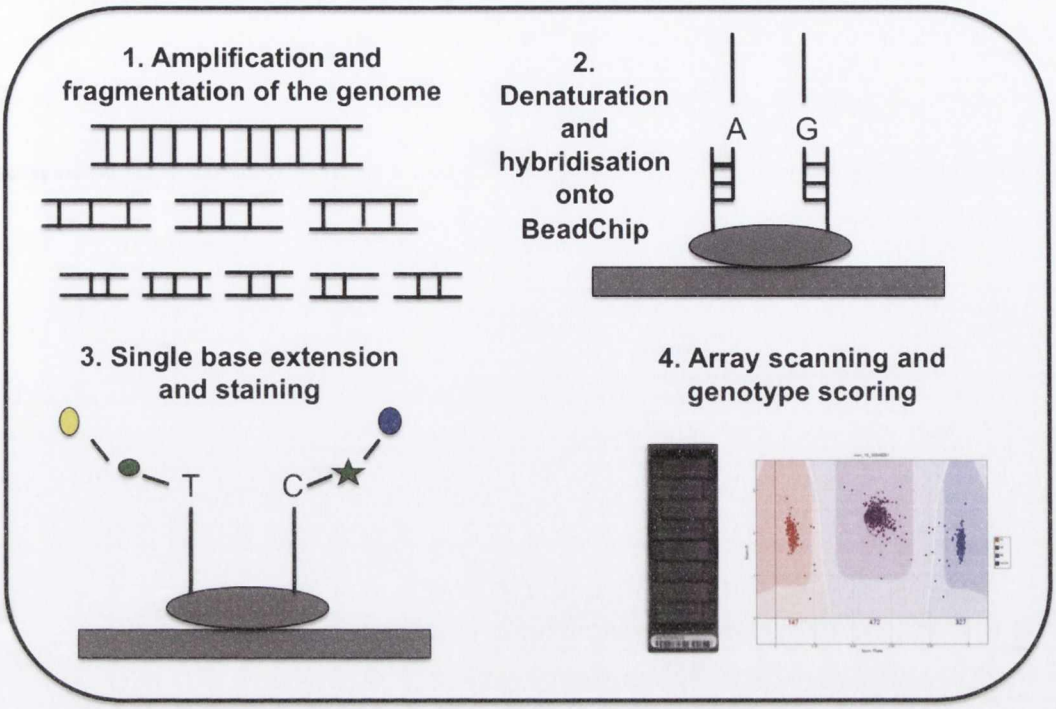


Figure 2.1: Workflow of Immunochip assay for whole-genome genotyping.

2.3.5 Genotyping

Samples were genotyped on the Immunochip according to illumina's protocols at the UCD Academic Centre on Rare Diseases (ACoRD) by Dr. Judith Conroy. Bead intensity data was processed, normalized and genotypes were called using the supplied Illumina GenomeStudio GenTrain 2.0 software using the cluster set of 172,242 autosome/x-chromosome variants applied to all of the samples in the Trynka *et al* dataset (2011). NCBI build 36 (hg18) was used.

2.3.6 Data Quality Control (QC)

Due to the large quantity of marker loci tested in GWA studies, even small sources of random and/or systematic error or bias can give rise to erroneous results. Therefore careful attention to data quality control (QC) is an essential component of all GWA studies. To maximise the numbers of markers and potential disease associations in the study, we began the QC by removing individuals failing our QC steps prior to conducting QC on a 'per-marker' basis. All quality control of the genotype data was performed using PLINK v1.07 (Purcell et al., 2007).

2.3.6.1 Per-Individual Quality Control

We excluded Individuals based on their per-sample call rate, excessive heterozygosity, evidence of non-European ancestry, duplication and being closely related to another individual in the study, as described below.

2.3.6.1.1 Missing genotypes

Empirically it can be seen that the quality of DNA samples vary, such that significant effects on genotype efficiency (call rate) and genotype accuracy occur; therefore samples with systematically low call rates (i.e. the rate at which genotypes at a given locus are successfully identified) were first excluded. The sample call rate is described as the fraction of called SNPs per sample over the total number of SNPs assayed. 25 Individuals with >5% missing genotypes were removed from our data set.

2.3.6.1.2 Heterozygosity

The proportion of loci expected to be heterozygous in an individual (ranging from 0 to 1.0) is known as the heterozygosity rate. Any individuals showing high levels of heterozygosity were removed. Individuals with an excessive proportion of heterozygote genotypes may be an indication of DNA sample contamination, individuals with a reduced proportion of heterozygote genotypes may be indicative of inbreeding or an inadequate population size. We excluded all individuals with a heterozygosity rate ± 3 SD from the mean. 13 individuals were removed for excessive heterozygosity (Fig 2.2).

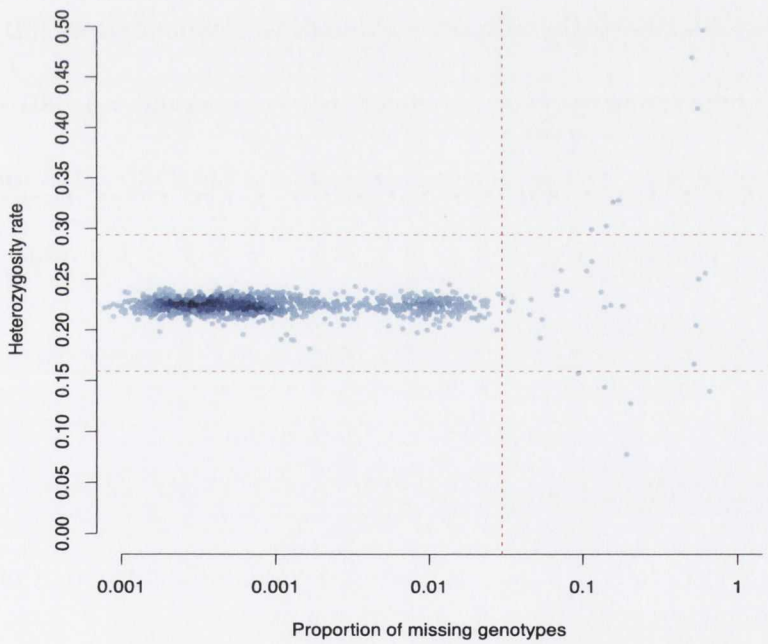


Figure 2.2: Genotype failure rate vs. heterozygosity rate across all individuals in the study pre QC. Shading indicates individual sample density and dashed lines represents QC thresholds.

2.3.6.1.3 Identity by Descent (IBD)

It is important in a standard population based case-control association study that all the samples are unrelated. If samples are related or there are duplicate samples included in the study, a bias may be introduced whereby the genotypes within families will be over represented, and therefore the sample may no longer be a fair reflection of the allele frequencies in the entire population. To identify duplicate and related individuals, a metric (identity by decent, IBD) is calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding sex chromosomes). Only independent SNPs are included in this analysis; regions of extended linkage disequilibrium (LD) were entirely removed from the dataset and the remaining regions were pruned so that no pair of SNPs within 50kb has an r2 greater then 0.2. Relatedness between Individuals was estimated by IBD. Duplicate Individuals or monozygotic twins were defined as having an IBD = 1, IBD = 0.5 for first-degree relatives, IBD = 0.25 for second-degree relatives and IBD 0.125 for third degree relatives (first cousins) (Anderson et al., 2010). 33 related individuals were identified in this study, which were removed at this stage.

2.3.6.1.4 Principal Component Analysis (PCA)

PCA is a multivariate statistical method, which can be used to identify individuals of divergent ancestry and to define an homogeneous set of samples. Our population was restricted to a proxy-North Western European ancestry subset as defined by similarity to the HapMap Centre d'Etude du Polymorphisms Humain (CEPH) reference dataset. When detecting ancestry, a principal component model is built using genome-wide genotype data from populations of known ancestry. In this study we used HapMap

genotype data from Europe (CEU), Asia (CHB [Chinese] and JPT [Japanese]) and Africa (YRI [Yoruban from Nigeria]). Population outliers were identified by multi-dimensional scaling and individuals >3 SD from the CEPH mean component score for the first 2 principal components were removed. Based on this analysis, we removed 10 control individuals that appeared to be outliers with respect to the rest of the Irish population (Fig 2.3 and Fig 2.4). We received a number of control samples from Dr. Gary Donohoe from the University of Ireland Galway and unfortunately 10 individuals were not of Irish ancestry.

Following per-individual QC the dataset consisted of 878 individuals, 425 cases and 453 controls.

Figure 2.3: Multi Dimensional Scaling (MDS) plot demonstrating ancestry clustering based on genome-wide association data pre QC. Our Irish population can be seen in blue. HapMap3 reference samples: CEU (orange), YRI (turquois), JPT (pink) and CHB (green).

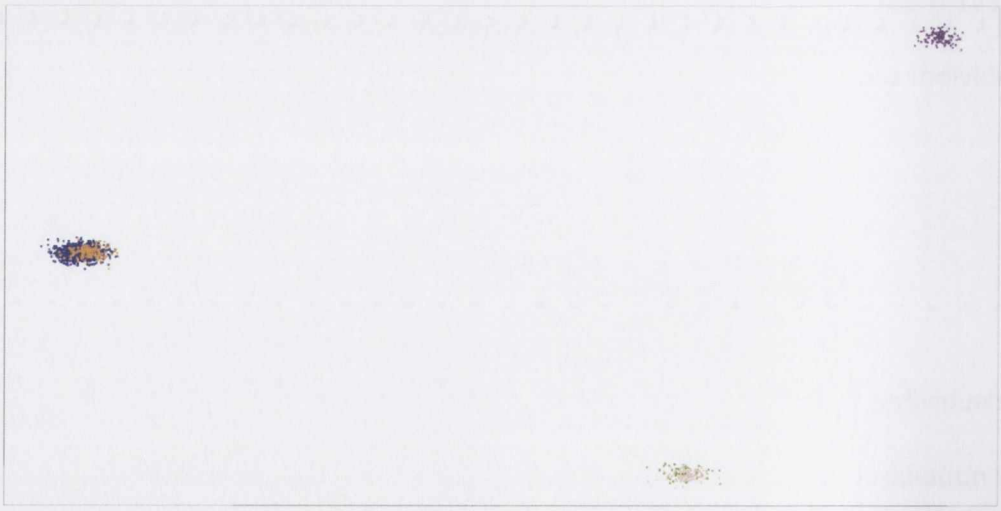


Figure 2.4: MDS plot demonstrating ancestry clustering based on genome-wide association data post QC. Well-defined groups are represented. Our Irish population can be seen in blue. HapMap3 reference samples: CEU (orange), YRI (blue), YRI (turquois), JPT (pink) and CHB (green).

2.3.6.2 Per-Marker QC

We began our per-marker QC with 172,242 SNPs (Trynka *et al* dataset).

2.3.6.2.1 Missing genotypes/differential missingness

Substandard markers were removed from the data as they may result in false positives and reduce the ability to identify true associations correlated with disease risk. We used a call rate cut off of 95%, meaning that SNPs that could not be assigned a genotype in >95% of samples were excluded from further analysis and SNPs with differential missingness between cases and controls ($P < 10^{-6}$) were also removed from the dataset. These are SNPs that have differences in call rate between cases and controls. 1758 SNPs were removed for missing genotypes and 2241 SNPs were removed for differential missingness between cases and controls.

2.3.6.2.2 Hardy-Weinberg equilibrium (HWE)

Hardy-Weinberg equilibrium (HWE) is one of the most important principles in population genetics. Under Hardy-Weinberg assumptions, both allele and genotype frequencies remain constant from generation to generation. Therefore, population allele frequencies at given loci can be used to calculate the expected genotypic proportions. Departure from this equilibrium may represent inbreeding, population stratification or, of particular relevance in these types of studies, genotyping errors. Only control samples are assessed for deviations from HWE, as it is assumed there is potentially a biological explanation for deviations from HWE at particular loci in patient samples - if for example the variant contributes to disease. We found that 34 marker

loci showed deviations from HWE in the control samples ($P < 10^{-5}$) and thus these markers were removed from the data set.

2.3.6.2.3 Minor Allele Frequency (MAF)

Some SNPs will contain one allele that is shown at a very low frequency across individuals in a study, making the power to detect the association between a phenotype and a rare allele in very difficult. Therefore to avoid this problem, filtering SNPs based on MAF is often used to exclude low MAF SNPs. SNPs with a MAF <0.05 were removed from the dataset.

Following quality control, 878 individuals (425 CD cases and 453 controls) genotyped across 143,074 markers were available for analysis.

2.3.7 Genetic association

Association analysis was performed in PLINK v1.07 (Purcell et al., 2007) using logistic regression adjusting for gender and 10 population covariates; the default setting of principal components generated from the multi-dimensional scaling (MDS) routine in PLINK to adjust for population stratification. Raw genotype intensity plots for all markers were visually inspected to confirm cluster separation and genotyping quality (Fig 2.5 and Fig 2.6).

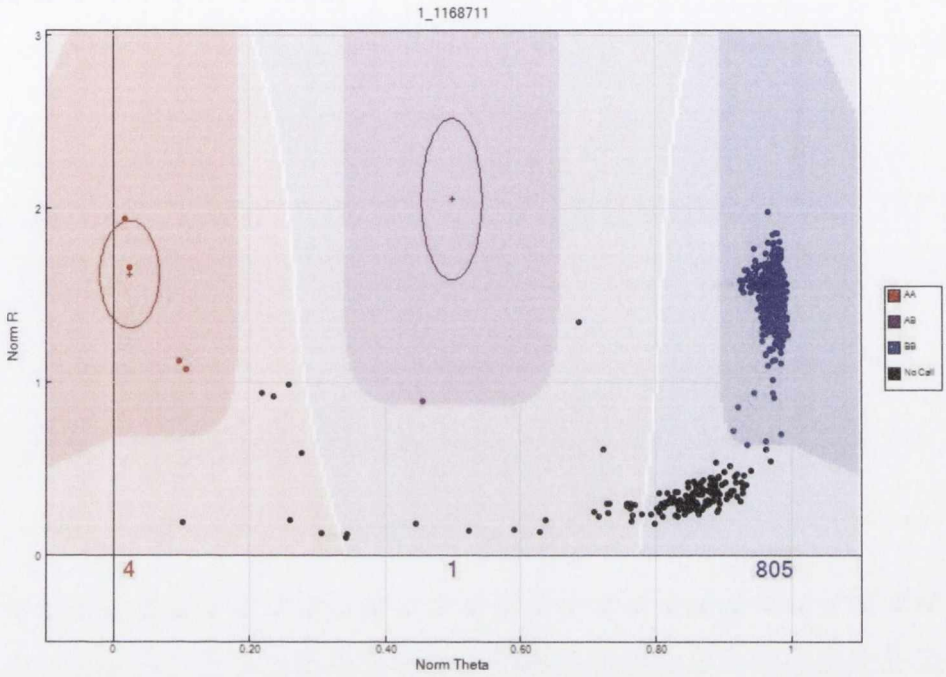


Figure 2.5: Unsuccessful, low-quality clustered SNPs. The x-axis represents the normalized θ value, the angle of deviation from the pure A signal. 0 signifies a pure A signal and 1 denotes a pure B signal. The y-axis represents normalized R indicating the intensity of the B allele. The colours for the clusters are defined as follows: red=AA, purple=AB, blue=BB and black=no call. The number shown above the x-axis shows the number of individuals included in this cluster.

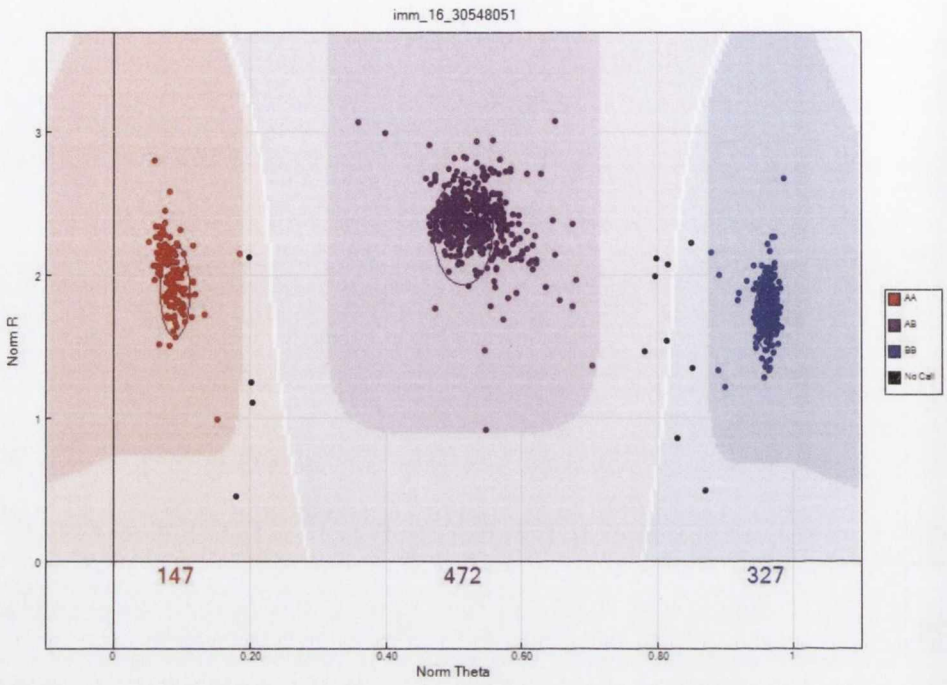


Figure 2.6: Successful, high-quality clustered SNPs. The x-axis represents the normalized θ value, the angle of deviation from the pure A signal. 0 signifies a pure A signal and 1 denotes a pure B signal. The y-axis represents normalized R indicating the intensity of the B allele. The colours for the clusters are defined as follows: red=AA, purple=AB, blue=BB and black=no call. The number shown above the x-axis shows the number of individuals included in this cluster.

2.3.8 Meta-analysis

Meta-analyses combining our findings with those of Trynka *et al* (2011) were performed using the software package METAL (Willer et al., 2010). METAL can use one of two approaches to achieve this, i). The direction of effect (OR) and P-value for each allele is converted into a *Z*-score. These scores can then be combined across all samples using a weighted approach to gain an overall *Z*-score. The weighted score is proportional to the square root of the sample size in each study and ii). The effect size estimates are weighted by their estimated standard errors.

2.3.9 Sign test

A key measure of whether our findings are concordant with those of the earlier study is to determine whether the disease associated alleles have the same direction of effect in the two studies. Thus we performed a binomial sign-test with markers previously identified as genome-wide significant by Trynka *et al* (2011). It is necessary to limit the effect of linkage disequilibrium on these data, thus linkage-independent SNPs were identified in the discovery (Trynka *et al* dataset) set by applying the "–clump" flag in PLINK. Clumps of SNPs are formed around an index SNP chosen due to having the lowest P-value in the dataset. SNPs which are within a 1Mb window from the index variant, that are correlated with it taking a cut-off of r^2>0.2 and have a P-value smaller than P <$5x10^{-8}$ are assigned to that index variants clump unless they have previously been assigned to another clump. SNPs that do not meet these thresholds are excluded. 280 LD independent GWS SNPs remained after this linkage disequilibrium-pruning step.

2.3.10 Polygene risk score analyses

All polygene risk scores (PRS) were calculated for the Irish study individuals using the --score flag in PLINK v1.07. All SNPs were corrected for strand inconsistencies; any SNPs where strand ambiguity was possible (A/T or C/G) were excluded from analysis. In total, we calculated PRS for eight groups of SNPs each defined by surpassing a specific association threshold in the discovery analyses. These thresholds were $P < 5x10^{-8}$, $P < 0.01$, $P < 0.05$, $P < 0.1$, $P < 0.2$, $P < 0.3$, $P < 0.4$ and $P < 0.5$. All analyses were calculated on the putative risk allele and a relative weight was applied to each allele as the log of the odds ratio of the association signal observed in the discovery analyses. Individual PRS were then calculated as the sum of all risk variants carried by that individual weighted for the risk effect size. To limit the effect of linkage disequilibrium on these data, linkage independent SNPs were identified in the discovery set as described above.

Put more simply, a PRS is created by adding the genotypic values for all loci that surpass a user-defined P value threshold for association in the sample of genotyped individuals under study. For example, if we consider locus A (SNP) that has two alleles A1 and A2, with the A2 allele more associated with disease phenotype, we can then assign each genotype a value. Thus Individuals with an A1A1 genotype at locus A are given a 0 value, A1A2=1 and A2A2=2. The values created at each locus are then weighted by the log of the risk effect size (odds ratio) of the association signal in the discovery analysis to give a score. This score can then be added to the score of locus B

and so on for all the loci of interest to give the total polygenic score at a given threshold in each individual.

To examine the strength of the derived PRS to predict disease status (i.e. case or control) in our sample, a nested logistic regression was performed. All scores were converted from a scale of 0-1 and individual SNP missingness was corrected by mean imputation from available SNPs genotyped in the individual. Two nested logistic regression models were applied; (i) case-status predicted by score, gender, and population covariates and (ii) case-status predicted by gender, and population covariates.

The genetic variance in case/control status explained by the score statistic (Vm) was calculated as the difference in variance (Nagelkerke's pseudo $R^{2)}$ between the two models outlined above - i.e. the model including the score and the covariates (full model) versus the covariates alone (reduced model). In addition to performing score analyses for all associated markers as described above, analyses of SNPs within the HLA locus only and SNPs not within the HLA locus (non-HLA) were also carried out. In these analyses, we examined (i) case-status predicted by score_HLA, score_non-HLA, gender, and population covariates and (ii) case-status predicted by score_non-HLA, gender, and population covariates.

2.3.11 Receiver Operator Characteristic (ROC) curves

The Student's t-test was used to test for significance of the differences in risk score between groups. We also calculated Receiver Operator Characteristic curves for the

scores in the Irish study to assess diagnostic efficiency. ROC curves plot the true positive rate (sensitivity) against the false positive rate (specificity) and calculates the area under the curve (AUC). The AUC calculates the probability that the predictor is greater for cases than controls. It can range from 0.5 to 1 with a score of 1 representing perfect separation between cases and controls.

Dr. Richard Anney performed the polygene risk score analysis and additional calculations and visualizations in STATA v13 using custom scripts.

2.4 Results

2.4.1 Genetic association

As expected, a highly significant association was observed within the HLA region in our study of 425 Irish coeliac patients and 453 population controls (Fig 2.7). The strongest association was observed for rs3891175, which maps to *HLA-DQB1* ($P = 1.9 \times 10^{-32}$, OR = 6.22 (95%CI 4.6-8.4)). No SNPs outside of the HLA region achieved GWS (Fig 2.8). The 'top' 2 genes with the largest *P*-values in our dataset following replication was the *KIAA1109/ADAD1/IL21/IL2* gene block (Fig 2.9) and *PTPN2* (Fig 2.10).

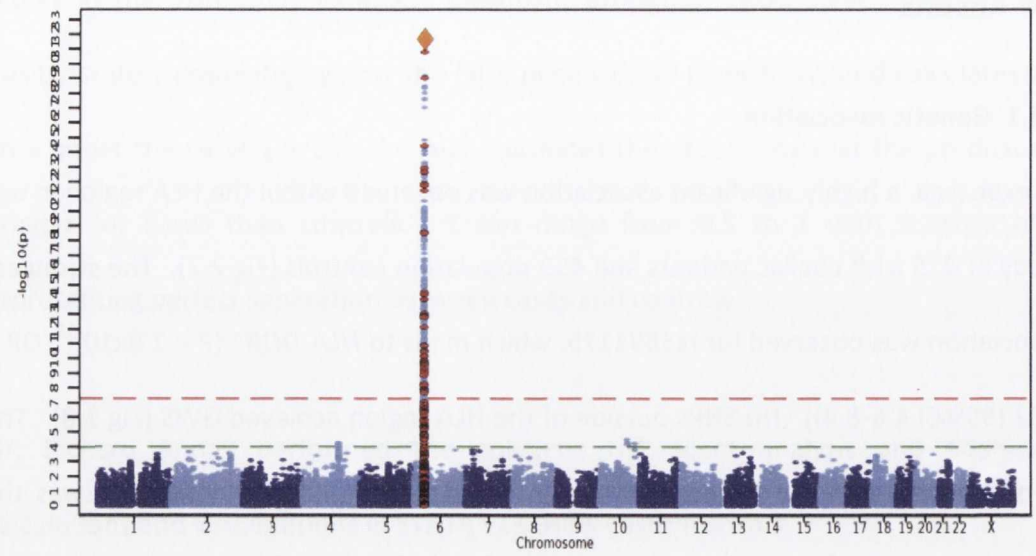


Figure 2.7: Manhattan plot of genome-wide association results using the Immunochip and encompassing the HLA locus. The X-axis represents the genome in physical order and the Y-axis shows log10 P-values for all SNPs. The blue line represents $P < 5x10^{-5}$ and the red line indicates the $P < 5x10^{-8}$ significance threshold.

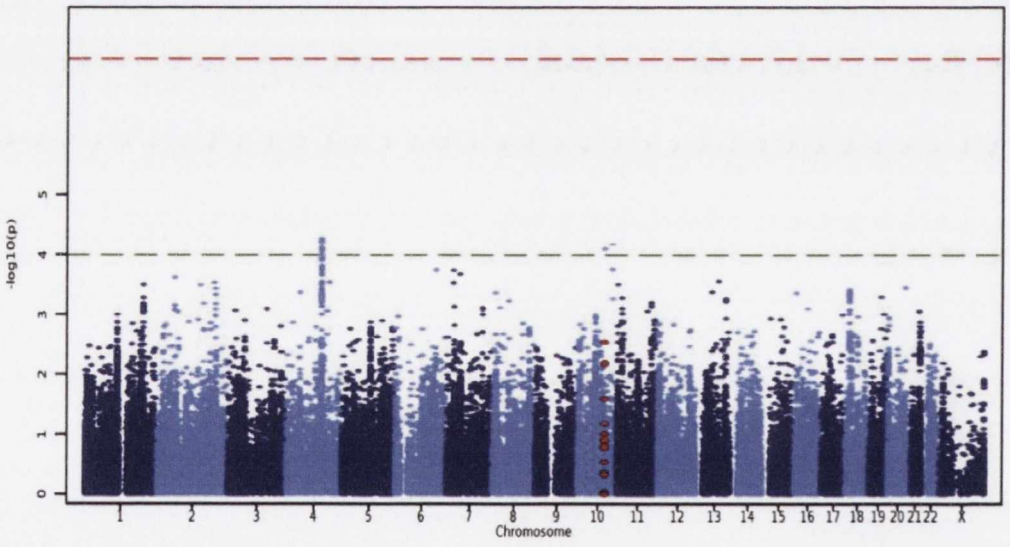


Figure 2.8: Manhattan plot of genome-wide association results excluding the HLA locus. The X-axis represents the genome in physical order and the Y-axis shows log10 P-values for all SNPs. Removal of the HLA locus of chromosome 6 shows several loci reaching P-values of approximately 10^{-4}. The red line indicates the $P < 5x10^{-8}$ significance threshold.

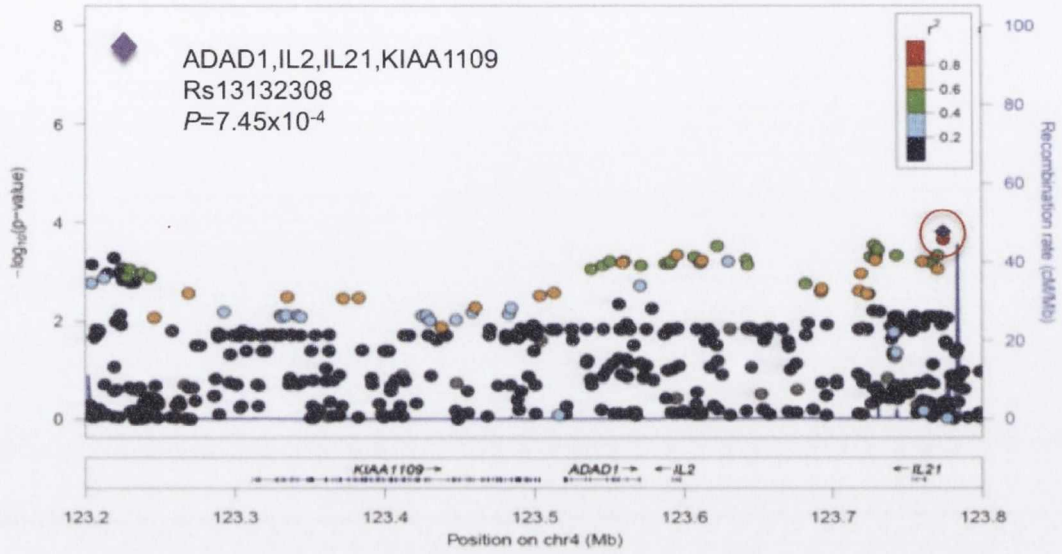


Figure 2.9: Locus association plot for SNP rs13132308, which has the strongest association in our dataset to CD. SNPs are plotted according to their chromosomal positions with –log10 P-values from the Immunochip analysis in this region on chromosome 4. The lead SNP is shown in purple. Strength of LD (r^2 values) between the lead SNP and other SNPs is indicated by colour. Genes within the region of interest are annotated at the bottom with arrows indicating the direction of transcription. Plots were generated using LocusZoom.

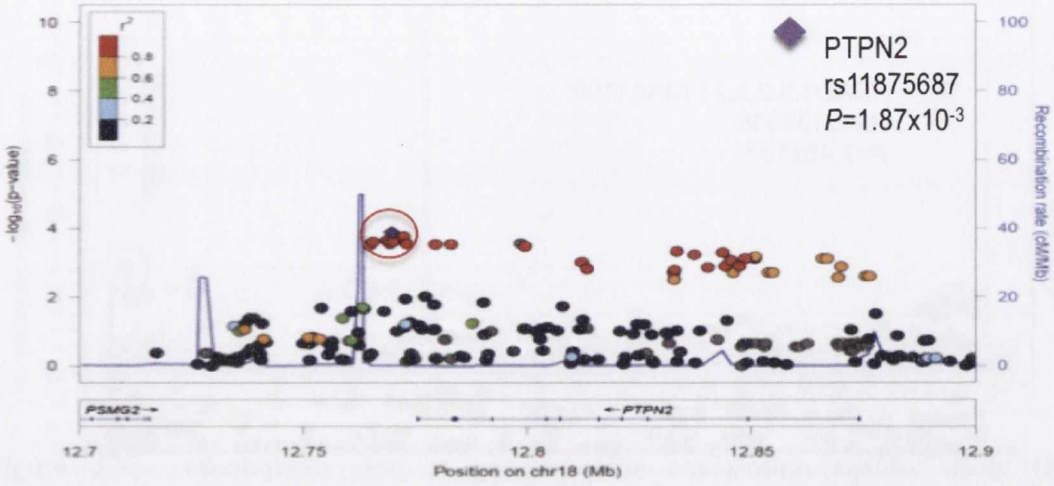


Figure 2.10: Locus association plot for SNP rs11875687, which has the second strongest association in our dataset to CD. SNPs are plotted according to their chromosomal positions with –log10 P-values from the Immunochip analysis in this region on chromosome 18. The lead SNP is shown in purple. Strength of LD (r^2 values) between the lead SNP and other SNPs is indicated by colour. Genes within the region of interest are annotated at the bottom with arrows indicating the direction of transcription. Plots were generated using LocusZoom.

2.4.2 Quantile-Quantile (QQ) plot

Care must be taken when comparing characteristics between two different groups, such as those with a disease and those without, since deviations can be caused by multiple and complex factors. The observed differences may not only be due to genetic differences but there can be differences arising from multiple factors, such as whether cases and controls were genotyped in the same laboratory or if the DNA was collected using the same method as well as the most common confounder, population stratification. Allele frequency can vary between different population subgroups such as those divided due to ethnicity or geographic origin which might in turn differ in their risk for the disease. Therefore GWA studies may identify these subgroups as related to the disease if they are unequally represented in the case and control populations. Excess relatedness amongst cases, a not-unlikely scenario for an inherited disease, can also lead to different allele frequencies between the groups which may be unrelated to susceptibility and must be controlled for as described above. Thus it is necessary to assess population structure. One way to do this is by generating a quantile-quantile (QQ) plot of the data using the R software program. The *P*-values for each SNP are ranked from the smallest to the largest, these are known as the observed association statistics and plotted against their expected values had they been taken from a distribution of known theoretical form ie: the distribution of our SNPs under the null hypothesis where none of our SNPs are associated with disease state. Deviation along the length of the X=Y line suggests a constant difference between cases and controls across the entire genome, suggesting bias. Whilst, a 'clean' QQ plot should show a line matching the X=Y line with a sharp curve at the end thus representing the number of true associations among thousands of unassociated SNPs.

The QQ plot representing our data with both the HLA and non-HLA SNPs included shows a sharp deviation from the X=Y line in Fig 2.11 indicating an excess of significantly associated SNPs in the case population and that the null hypothesis of no association being present should be rejected. The genomic inflation factor lambda (λ) was also calculated as λ=1.12. This factor indicates measurable population stratification/bias across the samples. If there is no bias, λ=1 however, if λ is greater than 1, it may be indicative of bias with the data. The removal of the HLA locus in Fig 2.12 left residual deviation from the line also identifying the presence of associated SNPs outside of the HLA region. The genomic inflation statistic was calculated as λ=1.02 indicating no measurable population stratification/bias across the samples after the removal of the HLA locus.

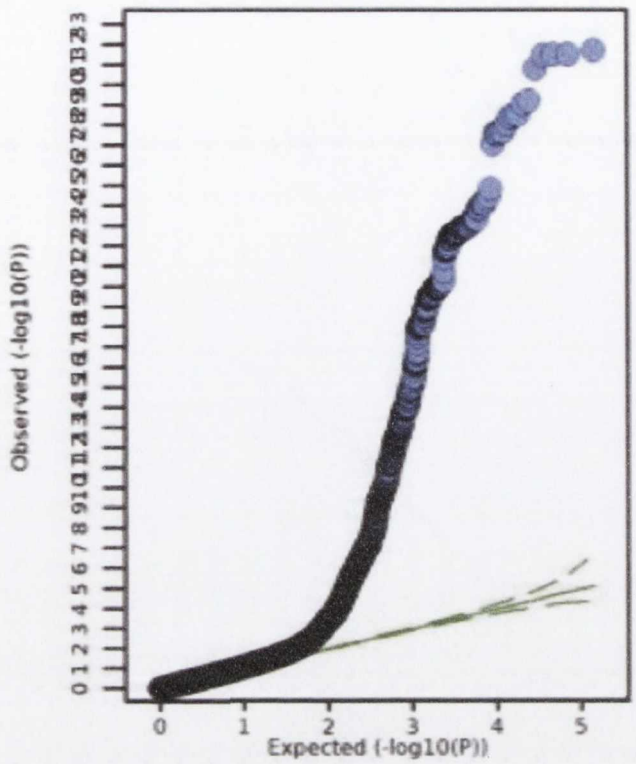


Figure 2.11: QQ plot of the data representing both HLA and non-HLA SNPs. The observed association statistics of all polymorphic SNPs (black/blue) vs the expected null distribution (solid green line) incorporating the HLA locus. The dashed green lines represent the upper and lower 95% confidence intervals. The observed association statistics (-log10 P-values) for each SNP were calculated and ranked in order from smallest to largest on the y-axis and plotted against the values expected under the null hypothesis of no association on the x-axis. The genomic inflation factor lambda (λ) was calculated as 1.12.

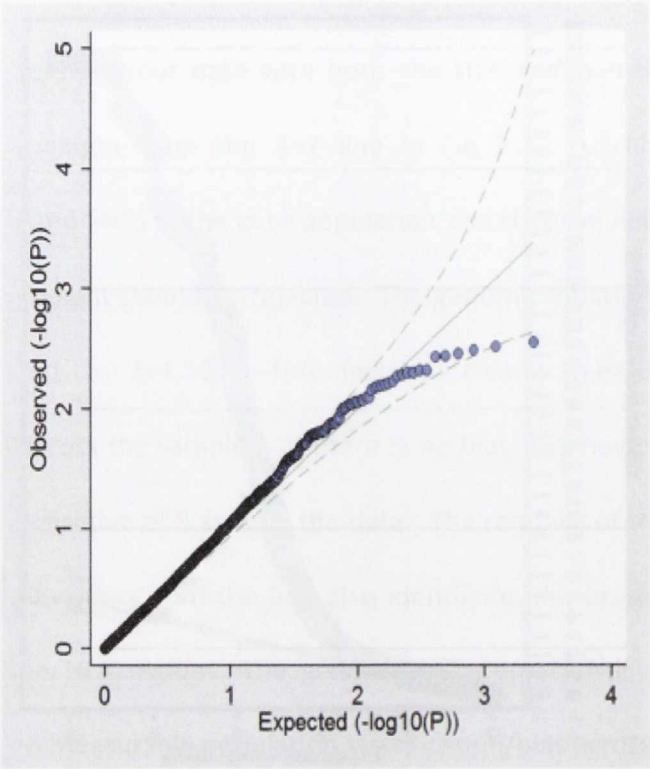


Figure 2.12: QQ plot of the data representing non-HLA SNPs only. The observed association statistics of all polymorphic SNPs (black/blue) vs the expected null distribution (solid green line) after the removal of the HLA locus. The dashed green lines represent the upper and lower 95% confidence intervals. The observed association statistics (-log10 P-values) for each SNP were calculated and ranked in order from smallest to largest on the y-axis and plotted against the values expected under the null hypothesis of no association on the x-axis. The inflation statistic for all markers is $\lambda = 1.02$ after the removal of the HLA locus.

2.4.3 Replication

Trynka *et al* (2011) reported 57 GWS, LD-independent, non-HLA, CD association signals that reached levels of genome wide significance or had been identified as such in earlier CD GWAS and replication data sets (Hunt et al., 2008; Dubois et al., 2010). 56 of these SNPs passed QC in our dataset (imm_21_44453549 was omitted), and of these, we observed 5 SNPs at P <0.005 and 6 SNPs below or at the nominal association level (P <0.05) (Table 2.1). Based on the null hypothesis, we would expect to observe 0 at P <0.005 and less than 3 at P <0.05. This represents strong evidence for replication and provides additional support for a role of at *IL2/IL21* (P_{irish} <$1x10^{-4}$), *SCHIP1*, *IL12A*, *PTPN2*, *SH2B3*, *ETS1*, *TAGAP* (P_{irish} <$1x10^{-3}$), *PVT1*, *DDX6* (P_{irish} <$1x10^{-2}$) and *RUNX3*, *RGS1* (P_{irish} <0.05) as risk genes for CD.

2.4.4 Meta-analysis

We performed a meta-analysis by combining the data of Trynka *et al* (2011) and this study to examine if combining our results added sufficient power to identify new loci exceeding the level of genome wide significance (GWS; P <$5x10^{-8}$). Indeed, the association signal was strengthened at several loci (Figure 2.13) and two additional loci were identified which surpassed genome wide significance levels in the combined analysis; *NFIA* (rs6691768, $P_{combined}$ $2.2x10^{-8}$) and *ZNF335* (rs6032606 ,$P_{combined}$ $4.79x10^{-8}$) (Table 2.2). These are thus confirmed as CD susceptibility loci for the first time. To test the consistency of the direction of associations with those previously reported, a binomial sign test was used. There was highly significant evidence of concordance in direction of effect between studies. 218 of 280 genome-wide

significant (GWS) SNPs (78%) show effect in the same direction (Pr (K>=18)=2.2x10^{-16}). Notably, this trend was maintained when analyses were restricted to the 57 GWS, LD-independent, non-HLA, CD association signals, where 43 of 56 (77%) show the same direction of effect (Pr(K>=43)=5.6x10^{-05}).

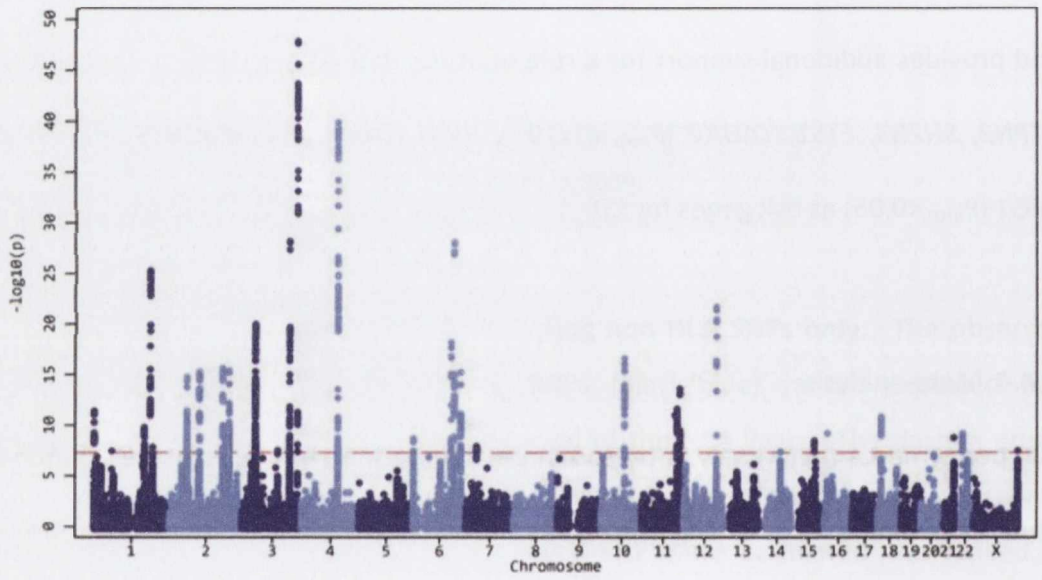


Figure 2.13: Manhattan plot from the meta-analysis of the Irish and Trynka *et al* datasets excluding the HLA locus.

Table 2.1: 11 of the 57 previously reported independent coeliac disease signals were nominally significant in our Irish sample cohort.

Chromosome	Position	Immunochip ID	RSID	OR_{Irish}	P_{Irish}	OR_{Trynka}	P_{Trynka}	Nearest Gene
4	123770564	imm_4_123770564	rs13132308	0.614	7.45E-04	0.7054	1.87E-38	ADAD1, IL2, IL21, KIAA1109
18	12833137	imm_18_12833137	rs11875687	1.527	1.87E-03	1.173	1.92E-10	PTPN2
3	161120372	imm_3_161120372	rs76830965	1.547	3.06E-03	1.36	2.56E-27	IL12A
12	110368991	imm_12_110368991	rs3184504	1.391	3.41E-03	1.192	5.42E-21	ATXN2, SH2B3
11	127897147	imm_11_127897147	rs61907765	1.459	3.69E-03	1.175	3.43E-13	ETS1
6	159418255	imm_6_159418255	rs1107943	1.819	6.31E-03	1.221	7.95E-09	TAGAP
8	129333242	imm_8_129333242	rs10808568	0.7258	1.03E-02	0.9132	2.20E-05	PVT1
11	118085075	rs10892258	rs10892258	0.7165	1.11E-02	0.8611	1.73E-11	DDX6, TREH
1	25162321	imm_1_25162321	rs72657048	0.777	2.26E-02	0.918	3.79E-06	RUNX3
4	123257745	imm_4_123257745	rs62323881	1.568	2.98E-02	1.204	6.87E-08	ADAD1, IL2, IL21, KIAA1109
1	190808095	imm_1_190808095	rs1359062	0.7605	5.00E-02	0.7691	2.55E-25	RGS1

Table 2.2: Meta-analysis results of the Trynka *et al* dataset combined with our data. The association signal was strengthened at several previously identified loci and two loci previously unassociated with CD were found to exceed the level of genome-wide significance, *NFIA* and *ZNF335*.

Chromosome	Immunochip ID	Variant ID*	Reference allele	Variant allele	Test Allele	LD range	$OR_{trv/e}$	$P_{trv/e}$	OR_{Ilm}	P_{Ilm}	P_{meRnd}	Direction	Effect*	Hetisq*	Hetpval*	Closest gene
1	**rs6691768**	**rs6691768:g.61564451G>A**	G	**A**	**G**	**chr1:61564451..61564451**	**0.9001**	**5.32E-08**	**0.8645**	**0.1811**	**2.28E-08**	**++**	**0.1065**	**0**	**0.7152**	**[NFIA]**
1	imm_1_171026745	rs859715:g.171026745A>T	A	T	T	chr1:171024719..171058589	1.112	5.30E-08	1.16	0.2071	2.77E-08	--	-0.1073	0	0.7232	[5' of FASLG]
1	imm_1_199142520	rs59655222:g.199142520T>C	T	C	C	chr1:199141865..199145350	0.8932	6.17E-08	0.8723	0.2444	3.12E-08	++	0.1137	0	0.8426	[C1orf106]
2	imm_2_61018233	rs6726160:g.61018233T>G	T	G	T	chr2:61018233..61329962	0.9027	5.92E-08	0.8418	0.1042	1.91E-08	--	-0.1045	0	0.5165	[PUS10]
2	imm_2_204416891	rs231804:g.204416891C>T	C	T	C	chr2:204399783..204453248	0.9024	5.03E-08	0.8759	0.2287	2.42E-08	++	0.1035	0	0.7896	[CTLA4]
2	imm_2_191678575	rs13401064:g.191678575C>G	C	G	G	chr2:191678575..191678575	0.8342	6.45E-08	0.7453	0.1384	2.46E-08	++	0.1844	0	0.5755	[STAT4]
2	imm_2_204392042	rs62184863:g.204392042G>C	G	C	C	chr2:204392042..204392042	1.222	6.19E-08	1.403	0.1387	2.47E-08	++	0.204	0	0.5511	[CD28]
2	imm_2_102458945	rs1468788:g.102458945C>T	C	T	C	chr2:102458945..102458945	1.106	5.06E-08	1.114	0.3207	3.32E-08	--	-0.101	0	0.9479	[SLC9A4]
2	imm_2_204308702	rs45620941:g.204308702G>T	G	T	T	chr2:204308702..204308702	1.137	5.56E-08	1.139	0.3697	4.04E-08	++	0.1284	0	0.9905	[CD28]
3	imm_3_161112061	rs9873580:g.161112061C>T	C	T	T	chr3:161108239..161145664	0.8516	8.28E-08	0.7935	0.1429	3.04E-08	--	-0.1631	0	0.6602	[Intergenic region between CCR3 and CCR2]
3	imm_3_46341958	rs2187670:g.46341958T>C	T	C	T	chr3:46328478..46341958	1.121	1.01E-07	1.192	0.1609	3.98E-08	++	0.116	0	0.6296	[Intergenic region between SCHIP1 & IL12A]
4	imm_4_123592583	rs2069772:g.123592583T>C	T	C	C	chr4:123360504..123592583	1.118	7.24E-08	1.318	0.02957	1.34E-08	--	-0.1158	38.9	0.2009	[ADAD1,IL2,KIAA1109]
4	imm_4_123480980	rs62321692:g.123480980A>C	A	C	C	chr4:123257527..123722990	1.203	7.38E-08	1.6	0.02397	1.39E-08	--	-0.1924	45.2	0.1767	[KIAA1109]
6	imm_6_128198581	rs4897233:g.128198581A>G	A	G	G	chr6:128197641..128198581	1.109	5.41E-08	1.155	0.1775	2.01E-08	--	-0.1047	0	0.7084	[THEMIS]
6	imm_6_159434848	rs3120630:g.159434848G>A	G	A	A	chr6:159434848..159434848	0.8703	5.22E-08	0.7984	0.1578	2.16E-08	--	-0.1411	0	0.5932	[RSPH3,TAGAP]
6	imm_6_159254113	rs9457515:g.159254113T>C	T	C	C	chr6:159254113..159388565	1.203	1.34E-07	1.544	0.04487	3.25E-08	--	-0.1912	22.8	0.2552	[RSPH3,TAGAP]
7	1kg_7_37371503	rs77778931:g.37371503C>A	C	A	A	chr7:37366994..37404142	1.172	5.78E-08	1.296	0.1263	2.22E-08	++	0.1616	0	0.5592	[ELMO1]
10	imm_10_80735780	rs1250557:g.80735780G>T	G	T	G	chr10:80693358..80735780	1.108	7.19E-08	1.324	0.01513	1.17E-08	--	-0.1073	56.7	0.1288	[ZMIZ1]
11	imm_11_127920339	rs4128561:g.127920339T>C	T	C	C	chr11:127911985..127920339	0.9044	8.39E-08	0.7723	0.01438	1.15E-08	++	0.1053	53.9	0.1406	[ETS1]
11	imm_11_118001329	rs45565037:g.118001329G>A	G	A	A	chr11:117993992..118041050	0.886	5.41E-08	0.795	0.0823	1.60E-08	--	-0.124	0	0.4182	[PHLDB1,TREH]
14	imm_14_68341722	rs971290:g.68341722C>A	C	A	A	chr14:68341722..68341722	1.128	5.64E-08	1.181	0.222	2.79E-08	++	0.1216	0	0.7388	[5' of ZFP36L1]
15	rs936229	rs936229:g.72919372A>G	A	G	A	chr15:72919372..72919372	1.113	1.13E-07	1.276	0.04088	2.56E-08	++	0.1109	21.6	0.2587	[ULK3]
20	**imm_20_44029614**	**rs6032606:g.44029614C>G**	C	**G**	**C**	**chr20:44029614..44029614**	**0.7972**	**8.25E-08**	**0.7753**	**0.305**	**4.81E-08**	**--**	**-0.2274**	**0**	**0.9119**	**[ZNF335]**

2.4.5 Polygenic risk score analyses

We performed score analyses on three sets of SNPs; all markers, HLA only and non-HLA. We calculated PRS for all markers at or below 8 association P-value thresholds in our target sample ($P < 5x10^{-8}$, $P < 0.01$, $P < 0.05$, $P < 0.1$, $P < 0.2$, $P < 0.3$, $P < 0.4$ and $P < 0.5$) each weighted by the log odds ratio from the discovery sample (described above in methods section 2.3.10). To assess whether the combined scores reflect CD risk, we tested for a higher mean score in the target cases compared to controls. When all markers are included, higher PRS were found to be significantly associated with CD case status at each of the 8 thresholds (Table 2.3). This was also true when we examined HLA SNPs only. For the non-HLA SNP set, higher PRS were significantly associated with coeliac case status at the thresholds $P < 5x10^{-8}$, $P < 0.01$, $P < 0.05$ and $P < 0.1$.

We then carried out logistic regression analysis to test the relationship between the PRS and CD status in our target population. Figure 2.14 shows the proportion of variance explained by PRS at each of the different thresholds. Up to 35% of the variance was explained by the PRS when all markers were considered (SNPs threshold $P < 0.05$; Student's T test, $P = 9.23x10^{-75}$). Furthermore, up to 4.5% of the variance could be explained by the non-HLA SNPs alone (SNPs threshold $P < 0.01$; $P = 3.6x10^{-18}$). Up to 31% of the variance was explained by the HLA SNPs alone (SNPs threshold $P < 0.1$; $P = 2x10^{-70}$).

For each threshold we also estimated the area under the receiver operating characteristic curve (Table 2.4). For all markers the highest AUC was 0.83 (SNPs threshold P <0.05) (Fig 2.15); for non-HLA markers the AUC was 0.67 (SNPs threshold P <0.01) (Fig 2.16); and for HLA only, the highest AUC was 0.82 (SNPs threshold P <0.5) (Fig 2.17).

Table 2.3: Variance explained (R^2) for the three sets of SNPs tested (All, HLA only and non-HLA) at the different thresholds tested. P-values were calculated for the difference in score between cases and controls using Student's t test.

Threshold	Variance explained (delta R^2)			P values		
(-log10(P))	All	Non-HLA	HLA only	All	Non-HLA	HLA only
$5x10^{-8}$	0.327	0.041	0.259	$1.1x10^{-69}$	$1.3x10^{-17}$	$5.6x10^{-66}$
0.01	0.343	0.045	0.250	$2x10^{-73}$	$3.6x10^{-18}$	$5.7x10^{-66}$
0.05	0.349	0.017	0.294	$9.2x10^{-75}$	$8.3x10^{-10}$	$3.1x10^{-71}$
0.1	0.321	0.005	0.310	$1.3x10^{-58}$	0.006	$2x10^{-70}$
0.2	0.281	$8.9x10^{-5}$	0.323	$3.8x10^{-47}$	0.081	$1.5x10^{-70}$
0.3	0.266	$7.5x10^{-5}$	0.328	$5.6x10^{-43}$	0.128	$2.3x10^{-71}$
0.4	0.196	$2.8x10^{-5}$	0.291	$6.4x10^{-26}$	0.213	$2.2x10^{-62}$
0.5	0.225	$4x10^{-4}$	0.334	$3.7x10^{-26}$	0.318	$5.3x10^{-72}$

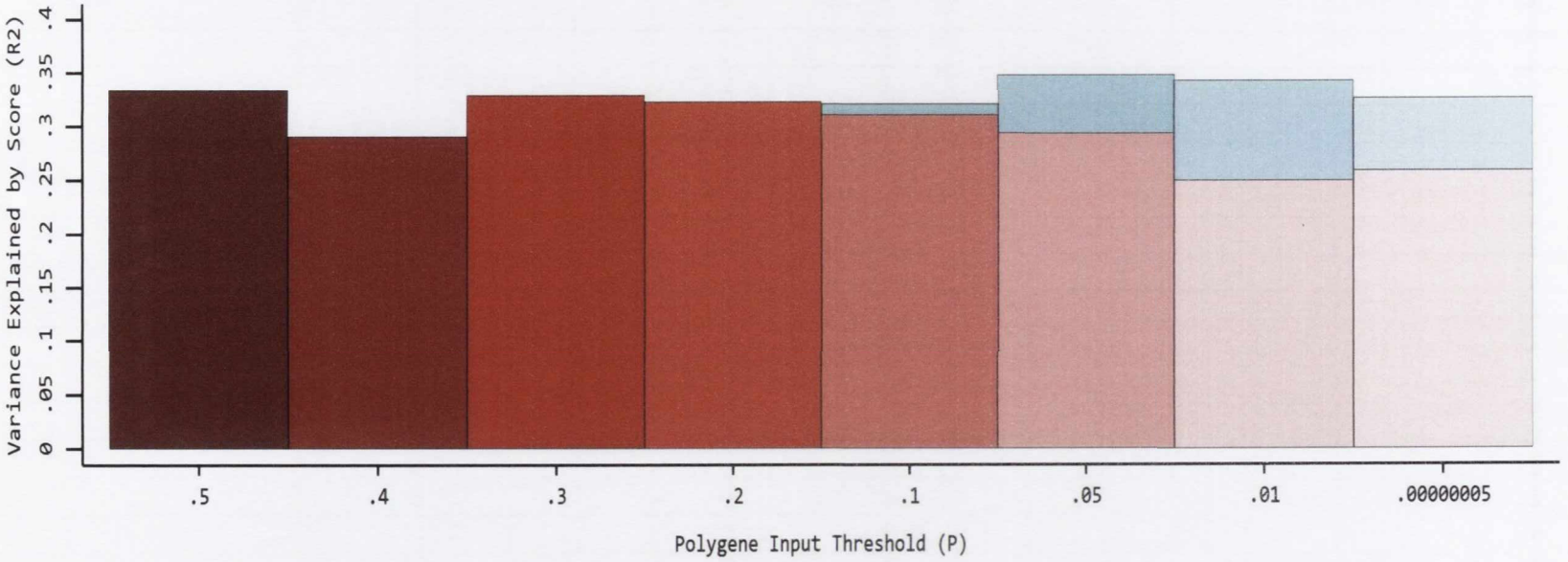


Figure 2.14: The proportion of variance explained by PRS at each of the different thresholds. Red = all SNPs included in the analysis. Blue = non-HLA SNPs only.

Table 2.4: The area under the receiver operating characteristic curve (AUC) for each threshold.

SNP set	THRESHOLD	COUNT	DELTA R2	N CASE	MEAN CASE	SD CASE	N CONTROL	MEAN CONTROL	SD CONTROL	P value	AUC	AUC SE
ALL SNPs	0.00000005	560	0.32658756	425	172.31	9.04	453	160.62	8.9089	1.09E-69	0.82	0.01
ALL SNPs	0.01	3088	0.34314925	425	417.38	9.64	453	404.76	9.1351	1.98E-73	0.83	0.01
ALL SNPs	0.05	7892	0.3486049	425	929.34	11.31	453	914.35	10.739	9.23E-75	0.83	0.01
ALL SNPs	0.1	12572	0.3205447	425	1502.54	12.19	453	1488.77	11.2968	1.32E-58	0.81	0.01
ALL SNPs	0.2	21044	0.28065362	425	2414.66	14.28	453	2400.79	12.6295	3.83E-47	0.78	0.02
ALL SNPs	0.3	28682	0.26646788	425	3135.82	15.78	453	3121.41	13.7124	5.60E-43	0.78	0.02
ALL SNPs	0.4	35750	0.19561343	425	3749.62	16.51	453	3738.43	14.1767	6.38E-26	0.72	0.02
ALL SNPs	0.5	42684	0.22548913	425	4331.99	20.01	453	4318.07	17.957	3.69E-26	0.71	0.02
NON HLA	0.00000005	112	0.04117338	425	9.69	0.85	453	9.19	0.8735	1.28E-17	0.66	0.02
NON HLA	0.01	2480	0.04462574	425	208.18	2.13	453	206.93	2.0605	3.59E-18	0.67	0.02
NON HLA	0.05	7152	0.01668193	425	688.68	3.35	453	687.41	2.8416	8.28E-10	0.64	0.02
NON HLA	0.1	11790	0.00471613	425	1258.97	6.86	453	1257.86	6.1669	0.006	0.57	0.02
NON HLA	0.2	20170	0.0000889	425	2157.81	9.39	453	2156.98	8.0801	0.081	0.55	0.02
NON HLA	0.3	27644	7.53E-06	425	2852.07	11.11	453	2851.28	9.3553	0.128	0.55	0.02
NON HLA	0.4	34646	0.00002745	425	3463.72	13.50	453	3463.05	11.4024	0.213	0.55	0.02
NON HLA	0.5	41452	0.0003539	425	4021.32	16.85	453	4020.82	14.9565	0.318	0.55	0.02
HLA ONLY	0.00000005	436	0.25940777	425	162.62	8.97	453	151.43	8.7895	5.58E-66	0.81	0.01
HLA ONLY	0.01	608	0.2497022	425	209.21	9.24	453	197.83	8.8314	5.73E-66	0.81	0.01
HLA ONLY	0.05	734	0.29354292	425	240.66	10.44	453	226.95	10.3322	3.12E-71	0.82	0.01
HLA ONLY	0.1	780	0.31028779	425	243.57	9.76	453	230.92	9.5492	2.03E-70	0.82	0.01
HLA ONLY	0.2	874	0.32273192	425	256.85	10.13	453	243.81	9.7544	1.52E-70	0.82	0.01
HLA ONLY	0.3	1038	0.32847623	425	283.75	10.52	453	270.13	10.0972	2.27E-71	0.82	0.01
HLA ONLY	0.4	1100	0.29050774	425	285.90	8.78	453	275.37	8.5117	2.15E-62	0.81	0.01
HLA ONLY	0.5	1220	0.33387895	425	310.66	10.12	453	297.25	10.06	5.25E-72	0.82	0.01

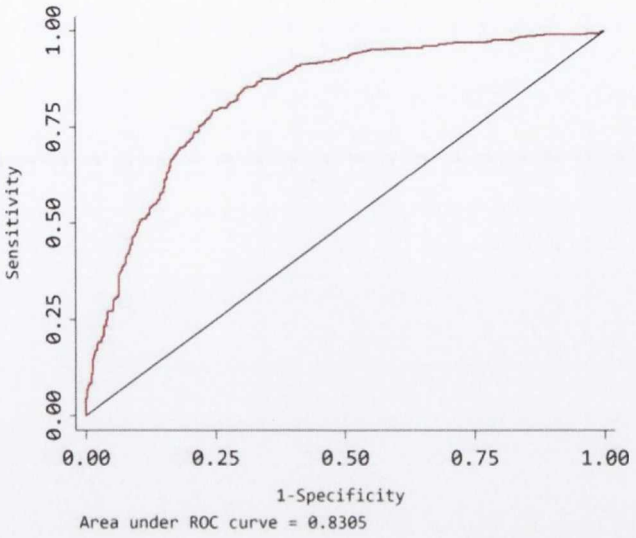


Figure 2.15: ROC showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for all markers in our study. The highest AUC was 0.83 at a SNP threshold of P <0.05.

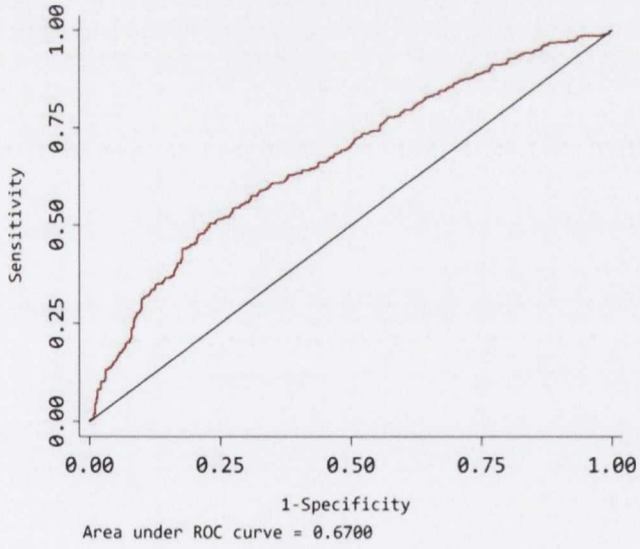


Figure 2.16: ROC showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for non-HLA markers only in our study. The highest AUC was 0.67 at a SNP threshold of P <0.01.

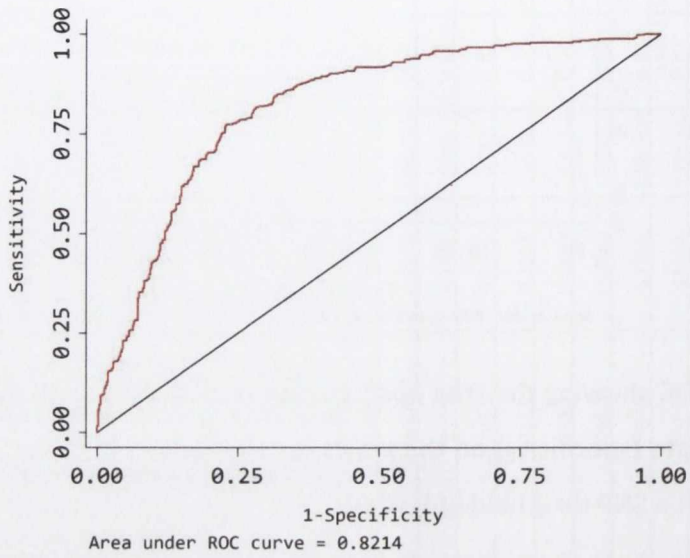


Figure 2.17: ROC showing the true positive rate (sensitivity) on the y-axis versus the false positive rate (specificity) on the x-axis for HLA markers only in our study. The highest AUC was 0.82 at a SNP threshold of P <0.05.

2.5 Discussion

Recent association studies have added substantially to our understanding of the genetic complexity of coeliac disease, making it one of the best understood complex diseases. However, there is still a low probability of replicating genome wide significant results, so it is important that disease associated loci for complex diseases should be replicated in independent populations (Liu et al., 2008).

In this study, we sought to replicate in an independent Irish case-control sample, findings from a recent large Immunochip analysis carried out by Trynka and colleagues (2011). We defined replicated SNPs as those showing nominally significant association ($P < 0.05$) and a consistent direction of effect. Trynka and colleagues reported 57 independent non-HLA association signals for CD that had reached genome wide significance, either in their dataset, or earlier genome-wide studies. We found evidence of replication at 11 of these SNPs, which was almost 4 times greater than would be expected by chance at the $P < 0.05$ level. Genes at these loci include *IL2-IL21*, *PTPN2*, *IL12A*, *SH2B3*, *ETS1*, *TAGAP*, *PVT1*, *TREH-DDX6*, *RUNX3* and *RGS1*. Furthermore, there was highly significant evidence of concordance in direction of effect for the majority (77%) of SNPs that were GWS in the Trynka *et al* study in our data (both HLA and non-HLA SNPs). This included 12 of the new loci identified by Trynka *et al* that had not been previously implicated in coeliac disease, adding further support for these loci as true coeliac susceptibility loci.

The 'top' 3 genes with the largest p values in our data following replication were the *KIAA1109/ADAD1/IL-21/IL-2* gene block, *PTPN2* and *IL12A.*

2.5.1 KIAA1109/ADAD1/IL21/IL2

rs13132308 SNP is the most significant SNP emerging from the Trynka *et al* dataset (P = 1.87 x 10^{-38}). It is located within a ~700kb LD block on chromosome 4q27 encompassing three known protein coding genes, *IL2, ADAD1, IL21* and a gene of unknown function *KIAA1109*. This SNP is also the most highly associated SNP in our Irish population dataset with a P-value of $7.45x10^{-4}$. In standard studies SNPs are only considered GWS if they reach the P-value threshold of $5x10^{-8}$, however this threshold is considered too stringent for replication studies of autoimmune diseases where a SNP with a confirmed significant association in one autoimmune disease has a higher prior probability of being associated with another autoimmune disease. Therefore P-values of 10^{-4} are accepted and SNPs are considered to be disease associated even if the SNP does not reach P = $5x10^{-8}$ in the replication study (Zhernakova et al., 2011). *ADAD1*, also known as *TENR* (Testis nuclear RNA-binding protein) is specifically expressed in the testis and *KIAA1109* transcripts are widely expressed as multiple splice variants in multiple tissues, and therefore are perhaps unlikely candidates for the causal CD susceptibility genes within this LD block. Thus, the most convincing biological candidates appear to be the cytokine genes *IL21* and *IL2*. IL2 has paradoxical functions in T-cell homeostasis. It stimulates T-cell activation and proliferation and promotes the differentiation of cytotoxic T-lymphocytes and NK cells. However, it also

plays a role in maintaining immunologic self-tolerance and terminating T-cell responses. It has the ability to eliminate self-reactive T-cells and maintain $CD4^+$ $CD25^+CD127^-Foxp3^+$ regulatory (T-reg) T-cells. Previous studies have shown that activation of $Foxp3^+$ T-reg cells have lead to the development of autoimmune or inflammatory diseases in humans and therefore could also potentially play a major role in the development of CD (Dominguez-Villar et al., 2011;McClymont et al., 2011). IL21 is secreted from active $CD4^+$ T-cells. It targets a broad range of cells; it regulates the functions of T-cells, NK cells, B-cells and myeloid cells and regulates the production of IFNγ. The interplay between IL21 and IL15 in CD has been reported. These cytokines work synergistically to stimulate the activation and expansion of $CD8^+$ T-cells leading to tissue damage characteristic of CD. Association of this region with type 1 diabetes (Burton et al., 2007), Graves disease (Burton et al., 2007) and CD (Hunt et al., 2008; van Heel et al., 2007; Dubois et al., 2010) has previously been reported.

2.5.2 *PTPN2*

SNP rs11875687 is located within the *PTPN2* gene. It the second most strongly associated SNP in our data with a nominal P-value of 1.87×10^{-3}. Protein tyrosine phosphatase non-receptor type 2 (PTPN2) encodes the enzyme tyrosine-protein phosphatase non-receptor type 2, a member of the protein tyrosine kinases (PTP) superfamily. There are two isoforms of PTPN2 identified to date, a major TC45 isoform (45kDa) and a minor TC48 isoform (48kDa). TC45 contains a nuclear localization sequence and upon activation exits the nucleus and has multiple targets, including

Janus kinases (JAKs), signal transducer and activator of transcription (STAT) 1 and 3, and the insulin receptor β (IRβ) (Glas et al., 2012). PTPN2 plays a role in maintaining the epithelial barrier and the structure of tight junctions. Studies have shown that it is activated by IFNγ and protects against IFNγ induced epithelial permeability (Scharl et al., 2009) suggesting a functional role for PTPN2 in disease pathogenesis. *PTPN2* has been identified as a candidate susceptibility gene for Crohn's disease (Franke et al., 2010), type 1 diabetes (T1D) (Moore et al., 2009), ulcerative colitis (UC) (Franke et al., 2008) and CD (Smyth et al., 2008).

2.5.3 *IL12A*

SNP rs76830965 is located within the chromosomal region 3q25.33 containing the *IL12A* gene. It is the second most significantly associated SNP in the Trynka *et al* study ($P = 2.56 \times 10^{-27}$) and a nominal association was also observed in our dataset ($P = 3.06 \times 10^{-3}$). The *IL12A* gene encodes a protein known as Interleukin 12 subunit alpha. IL-12A is a 35kDa subunit (IL12p35), which together with subunit IL12B/IL12p40 (40kDA) constitutes the heterodimeric cytokine IL12. IL12 is produced mainly by antigen-presenting cells such as dendritic cells (DCs), macrophages and B-cells in response to antigenic stimulation (Mattner et al., 1993) and plays a key role in the Th1 immune response. It aids in the differentiation of Th1 cells and inhibits Th2 cells when produced at the early stages of infection and inflammation. IL12 is also a pro-inflammatory cytokine inducing IFNγ secreting cytotoxic $CD8^+$ IELs and activating Natural Killer (NK) cells (Wang et al., 2000), one of the immunological hallmarks of CD.

In CD these NK cell receptors NKG2D and NKG2C interact with their ligands MICA and HLA-E that are upregulated on enterocytes, to induce epithelial lesions of the small intestine. *IL12A* was originally associated with CD in 2008 by Hunt *et al* (2008).

Increasing the sample size in association studies has been shown to have more impact in terms of SNP discovery than improving array coverage even with imputation (Lindquist et al., 2013). More risk loci for CD and related diseases remain to be detected. With even the current large sample sizes, many true signals of association fall short of the threshold for GWS. In the largest recent GWAS of anthropometrics (Wood et al., 2014), Crohn's disease (Jostins et al., 2012) and schizophrenia, increases in sample size have yielded a non-linear increase in the identified GWS loci. In schizophrenia, for example, an increase in sample participants from approximately 3000, to 6000, to 9000 and more recently 35000 resulted in the discovery of 1, 2, 6, and 108 GWS loci respectively (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Due to our small sample size we could not detect any SNPs reaching GWS outside of the well-established HLA locus on Ch 6p21. However examining QQ plots one might assume that the associations we did see are still in fact true associations and could potentially reach GWS if the sample size was increased. In figure 2.11 we can see a sharp deviation above the expected X^2 value of approximately 2.5 due to the HLA locus. Exclusion of the SNPs at the HLA locus figure 2.12 left a residual upward

deviation identifying further associated SNPs with a higher observed X^2 value than expected under the null hypothesis. The bulk of the distribution lies on the 'expected' line like we would anticipate, as the majority of the SNPs tested would not be associated with CD. However those that do deviate from the line could possibly represent a small number of true associations among thousands of unassociated SNPs. Since we corrected for population structure during our GWAS QC steps it wouldn't be unreasonable to suggest that these are in fact true associations between a number of SNPs and CD pathogenesis. The QQ plots were also used to generate an estimated genomic inflation factor (lambda), which is used to determine any systematic bias in our data. With the inclusion of the HLA locus, the lambda inflation factor is 1.12. This is quite high and is a sign of potential genetic structure within a population. However when the HLA locus is removed lambda falls to 1.02, suggesting no population stratification associated with case/control status.

Due to the complexity of GWAS and the knowledge that (i) previously identified susceptibility loci for common disease have a small effect and (ii) the difficulties in obtaining large population cohorts with which to perform GWAS, the power to detect true disease associations remains limited. Zondervan and Cardon (2004) showed that for an odds ratio of 1.3 sample sizes of 2000-10,000 cases and controls are needed to attain 80% power. These kind of sample numbers are out of reach for most laboratories therefore combining data from multiple GWAS in a meta-analysis seems a reasonable way around such a problem. A meta-analysis is defined as a 'quantitative

approach to combine systematically the results of previous research to arrive at a conclusion about a body of evidence' (Munafo and Flint, 2004).

Our study was underpowered to detect such GWS loci however when we combined our Irish dataset and the existing Immunochip dataset to create a modestly increased sample size, this added sufficient power to raise an additional two loci beyond the GWS threshold; rs6691768 ($P_{Trynka} = 5.3\times10^{-8}$; $P_{combined}$ $2.2x10^{-8}$) and rs6032606 ($P_{Trynka} = 8.25\times10^{-8}$; $P_{combined}$ $4.79x10^{-8}$).

2.5.4 Nuclear Factor 1A *(NFIA)*

rs6691768 is an intronic SNP located in the *NFIA* gene which is a member of the NF1 (nuclear factor 1) family of transcription factors. It had been identified by Dubois *et al* (2010) as a locus with suggestive evidence for coeliac disease in an earlier GWAS ($PDubois_{combined}$ 1.19×10^{-7}). A marked enrichment for this locus was found by Trynka *et al* (2011), increasing the P-value to $P = 5.3\ x10^{-8}$. NFIA is also associated with a malformation syndrome of the central nervous system and ureteral and renal defects (Lu et al., 2007). Perhaps the most relevant function of NFIA in the context of our study, is the role it plays in the inflammatory response. NFIA plays a role in granulopoiesis (Fazi et al., 2005), a process whereby undifferentiated blood precursors differentiate into functionally active granulocytes (neutrophils, eosinophils, basophils

and mast cells). Neutrophils play a crucial role in infection and are one hallmark of acute inflammation. The link between neutrophils and inflammatory conditions has already been made in rheumatoid arthritis (RA) (Wright et al., 2010), where inappropriately activated neutrophils cause persistent inflammation through the secretion of inflammatory mediators such as cytokines, chemokines, prostaglandins and leucotrienes and through antigen presentation stimulating T-cell activation and proliferation. Neutrophils also cause direct damage to tissue, bone and cartilage through the release of proteases and reactive oxygen species (ROS). With this link being made it serves to reason that the inappropriate regulation of NFIA in CD could act in the same way prolonging the inflammatory response and potentially causing tissue damage of the small intestine.

2.5.5 *ZNF335*

rs6032606 is a missense SNP located in the gene *ZNF335* which has not previously been specifically implicated in coeliac disease and is not strongly associated with other immune mediated diseases, although located in an extended region identified as a risk locus for Crohn's disease, multiple sclerosis, and rheumatoid arthritis (ImmunoBase; www.immunobase.org). ZNF335 (also known as NIF-1) was originally shown to modulate nuclear hormone receptor coactivator (NRC) activity, thus indirectly affecting the function of a wide variety of nuclear ligand-bound and other transcription factors including Fos, Jun and NF_kB (Garapaty et al., 2009). More recently, it has been shown to be a component of the trithorax chromatin remodelling complex (Yang et al.,

2012), with H3K4 methyltransferase activity. Deletion of *ZNF335* results in embryonic lethality. It is a binding partner of another transcriptional co-regulator with chromatin-remodelling activity HCFC1 (Mahajan et al., 2002). Interestingly, *HCFC1* is encoded in a CD susceptibility locus on Xq28 (2011). Thus, by virtue of its role as a transcriptional regulator, ZNF335 had the capacity to profoundly influence gene expression and in combination with HCFC1, implicates chromatin remodeling as a molecular mechanism in CD pathogenesis.

In a PRS analysis using the Trynka and colleagues (2011) GWAS summary statistics as a discovery dataset, we examined the collective influence of risk alleles on disease susceptibility and assessed the predictive power of these SNPs present on the Immunochip. Similar approaches have already been undertaken in CD (Romanos et al., 2014; Abraham et al., 2014) however using different sets of SNPs than those used here, i.e. the Immunochip. By simultaneously considering variants in the Immunochip data *en masse,* we observed strong evidence that regardless of P-value threshold used, alleles associated with CD from the earlier studies can explain a significant proportion of the variance in our CD cohort. Increase in the liability is not limited to markers reaching the standard benchmarks of association e.g. GWS or $P < 0.05$, but also extends to those with $P < 0.5$, indicating there are more disease associated markers to be found amongst these non-genome wide significant SNPs.

The highest proportion of variance in disease status explained by the score in our dataset was 35% when all markers are included ($P < 0.05$ threshold) and up to 4.5%

excluding the HLA locus (P <0.01). The latter finding is in accordance with previous observations in CD and other in complex conditions including schizophrenia (Purcell et al., 2009) and multiple sclerosis (The International Multiple Sclerosis Genetics Consortium (IMSGC), 2010) in which common non-HLA variants have been found to explain approximately 3% of the variance in disease risk. In contrast with other studies however, the proportion of variance explained decreased when more liberal discovery thresholds (P >0.1) were used to calculate the polygenic scores, indicating perhaps that the signal was being diluted by the addition of more biologically non-relevant SNPs.

Our findings are consistent with Abraham and colleagues (2014) who, using a machine learning approach generated a PRS based on multiple SNPs that explains 30-35% of disease variance when applied across several CD cohorts. Their analysis was performed predominately on datasets genotyped on the Ilumina Infinium array platform but also included a dataset genotyped on the Immunochip (however only SNPs in common to both were included – approximately 4% overlap). Their data showed a predictive power of 0.86-0.9 (AUC) which was moderately higher than that reported by Romanos *et al* (2014) who assessed the predictive power of just the 57 non-HLA susceptibility variants combined with HLA genotypes (AUC=0.854). The highest AUC in our dataset using SNPs present on the Immunochip was 0.83 (SNPs threshold P <0.05) including the HLA and 0.67 (SNPs threshold P <0.01) with non-HLA markers only.

Despite our small sample size being a limitation of this study, meaning that we had insufficient power to observe associations of small effect (<4% power to detect an association at P $<5x10^{-8}$ (MAF 0.5, OR <1.36)), we nevertheless observe PRS scores, which are sufficiently different between cases and controls and have the ability to predict case status (AUC 0.83) in line with previous reports.

Although assessments of the known heritability vary, it appears that a substantial proportion of 'missing heritability' remains to be uncovered and to date there is little evidence that rare variants are contributing substantially to the heritability of CD (Hunt et al., 2013). This polygene study provides further evidence that common variants not reaching GWS are contributing to the disease. Being able to predict CD risk through PRS analysis is also beneficial for accurately genetically diagnosing CD as described by Romanos *et al* (2014). Currently HLA-typing is being used alongside serological and histological testing for the diagnosis of CD. However, HLA-DQ2 and HLA-DQ8 have a poor positive predictive value and are therefore not entirely adequate at identifying individuals with CD. Thus, using genomic risk score
approaches such as PRS we have shown in our study that SNPs that fail to reach genome-wide significance when considered together can contribute modest improvements in accuracy of disease status prediction above HLA status alone. Being able to predict with accuracy those individuals at a higher risk of developing CD through genotyping and PRS methods will vastly improve the diagnosis of individuals suffering from CD and those at a higher risk of developing it.

One of the biggest advantages of the Immunochip is that it costs a fraction of standard GWAS chips and is applicable to other immune diseases. This study shows it has added value as a predictor of CD case status and it will be instructive to see if this also applies to other related diseases. Studies of inflammatory bowel disease indicates this to be likely, where AUCs of 0.86 and 0.83 were predicted for Crohn's disease and ulcerative colitis respectively (Wei et al., 2013). PRS is a flexible method that, used in conjunction with low cost genotyping platforms similar to the Immunochip, could represent a valuable and cost efficient tool with clinical application in assessing the risk of developing these diseases.

2.6 Conclusion

In conclusion, we have replicated in an independent Irish CD sample a number of the findings from a larger Immunochip study confirming associations at 11 loci. The direction of effect of GWS signals were highly concordant and our findings show that the Immunochip can be used to predict CD case status accurately using a polygene score method. Finally, by combining our data with earlier Immunochip studies, we have identified *NFIA* and *ZNF335* as GWS coeliac disease susceptibility loci.

Chapter III

Transcriptome analysis of $CD4^+$ T-cells in Coeliac Disease

3.1 Introduction

As previously mentioned Coeliac Disease (CD) is a common, chronic inflammatory disease of the small intestine that occurs in genetically predisposed individuals and is triggered by exposure to gluten and similar proteins in related grains (Schuppan et al., 2009). The HLA heterodimers, HLA-DQ2 and HLA-DQ8 are necessary but not sufficient to cause the disease, accounting for ~40% of the genetic susceptibility to CD. The remaining genetic risk is believed to be distributed among an unknown number of non-HLA genes where each gene contributes only a small effect. To date, genome wide association studies (GWAS) have identified 42 non-HLA CD risk loci harboring in excess of 70 candidate genes (Trynka et al., 2011; Garner et al., 2014; Coleman et al., 2015).

The majority of genetic variants associated with inflammatory diseases, as revealed by GWAS, are located in non-protein coding sequences and are thought to exert their affect by altering the expression of disease associated genes, many of which remain to be definitively identified (Fairfax et al., 2014; Raj et al., 2014). Genetic fine mapping allied to functional studies, including epigenomic analysis and gene expression and eQTL analysis in relevant cell types, will aid in the conclusive identification of the actual disease associated genes and causal genetic variants. Analysing gene expression profiles of patients may allow us to identify dysregulated gene expression and thus molecular pathways with altered activity in disease. Furthermore, by cross referencing with data from gene mapping studies, it may help pinpoint the most likely disease associated genes and genetic variants in the context of specific cells and activation statuses. Such information could potentially be used to refine disease prognosis and

risk prediction and may be useful in monitoring disease status (Montgomery and Dermitzakis, 2011; (Ayalew et al., 2012; Simon and Roychowdhury, 2013).

3.1.1 Transcriptome analysis of coeliac disease to date

To date genome wide expression studies carried out in CD have examined the expression patterns of genes from epithelial cells (Bracken et al., 2008), whole duodenal (Diosdado et al., 2004) and jejunal (Juuti-Uusitalo et al., 2004) biopsy samples using cDNA microarrays. Such studies have highlighted genes that may play a role in CD pathogenesis but their heterogenous cell populations, expressing distinct genetic programmes make it difficult to disentangle gene expression networks with the aim of pinpointing causal genes for further functional studies. Therefore recently there has been considerable emphasis on characterising cell type specific and stimulus dependent genetic variants (Dimas et al., 2009; Gerrits et al., 2009; Fairfax et al., 2014).

$CD4^+$ T-cells play a crucial role in the pathogenesis of CD as described in section 1.6. Antigen-presenting cells (APCs) bind and present deamidated gluten peptides to naïve $CD4^+$ T-cells in the lamina propria of individuals with CD promoting T-cell activation and upregulating a Th1 type immunological response with the production of IFNγ and IL21 (Bodd et al., 2010). This ultimately leads to the activation of cytotoxic intraepithelial lymphocytes (IELs), which cause much of the intestinal damage by directly killing mucosal epithelial cells. Expression profiling of $CD4^+$ T-cells in individuals with coeliac disease has not been undertaken thus far therefore

investigating gene expression in these cells could potentially uncover critical orchestrators of the immune response in CD. Thus focusing studies on disease relevant cell types and keeping studies as physiological as possible may be valuable in improving our understanding of disease pathogenesis.

3.1.2 cDNA microarray and RNA-sequencing

Early GWA studies were carried out using cDNA microarrays which feature several limitations, including that previous knowledge about the genome sequence is required for array design which are thus limited to detecting known transcripts. Furthermore, cross hybridization as a result of repetitive DNA sequences can generate high background levels of hybridization i.e., hybridization to a probe irrespective of the corresponding transcripts expression level (noise) which can lead to inaccurate expression measurements and the lack of reproducibility of results across studies (Okoniewski and Miller, 2006; Shendure, 2008). Therefore mRNA-sequencing (RNA-seq) has become a powerful alternative to microarrays and is considered a more robust and sensitive methodology (Wang et al., 2009; Li et al., 2014) not constrained by limitations affecting microarrays. It is the first sequencing-based method that allows the entire transcriptome to be examined in a high-throughput and quantitative manner and does not require previous knowledge of existing genomic sequences. It provides a much greater sensitivity and specificity to detect transcript expression levels compared to standard microarrays, as the dynamic range of expression levels over which transcripts can be detected spans at least five orders of magnitude, a range well beyond that achieved by standard microarrays .

3.1.3 Pathway analysis

High-throughput sequencing has revolutionized biological disease research; however identifying differentially expressed genes is only one of the steps in beginning to understand the functional aspects of the transcriptome. It is now thought to be essential to take into consideration the underlying biological structure and gene-gene interactions in order to accurately understand the mechanisms behind the disease. Functional insight into the regulation of a number of diseases can be achieved by looking for patterns of expression changes within sets of genes and identifying the biological pathways within which the products of these genes play a role. For example, if the expression of two genes appears to be highly correlated, does that mean they are co-regulated, or is one of them the positive regulator of the other; or are they even involved in the same biological process? Genes can be grouped into different categories based on some common biological properties such as sequence, function, interactions, biological pathways etc and then these categories are tested to find which ones are over represented amongst the differentially expressed genes. The genes genetically associated with CD pathogenesis have been shown to be enriched for immune function, particularly genes involved in chemokine receptor activity, cytokine binding and T-cell activation (Abadie et al., 2011). Gene Ontology (GO) categories are most commonly used (Gene Ontology Consortium, 2004) and there are many tools emerging that perform GO analysis, such as GeneGo, Ingenuity Pathway Analysis (IPA), GSEA and DAVID to name a few. However, these standard methods can

give biased results on RNA-seq data. This is because the read count for a gene is proportional to a gene's expression level multiplied by its length. Longer transcripts will have a greater number of reads mapping to it and therefore are more likely to be detected as DE even when they are not DE compared to short/lowly expressed transcripts of the same gene. Thus a more recent pathway analysis tool called Goseq was developed that corrects for gene length bias and allows GO analysis of RNA-seq data (Young et al., 2010). In our study we use Goseq to conduct pathway enrichment analysis for DE genes using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations. KEGG is a major resource for pathway analysis (Kanehisa and Goto, 2000). It exploits pathway knowledge from public databases, however it is also a standalone system that has the ability to link genomes to cellular functions as it contains numerous genes, molecules and interaction networks. Its database contains various biological processes divided into 7 main groups, Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems and Human Diseases.

3.1.4 Weighted gene correlation network analysis (WGCNA)

Alongside pathway analysis, network analysis approaches have been developed to reveal further insight into complex transcriptional regulation by trying to identify groups of highly co-regulated genes. Genes can be organized into functional groups (called modules in the case of WGCNA analysis) according to their cellular processes and analysed as a system/network rather then as individual parts. These modules therefore represent genes that are interconnected and co-regulated with one another. Several methods for identifying gene co-expression networks are available, Partial Correlation and Information Theory (PCIT) method (Kadarmideen and Watson-haigh, 2012) and the Lemon-Tree project (Joshi et al., 2009). However, one such method that has proven its superiority over both of these methods is the Weighted Gene Correlation Network Analysis (WGCNA) (Langfelder and Horvath, 2008). WGCNA is a systems biology method for describing correlation patterns among genes. It can be used to group genes together into modules of highly correlated genes that are biologically meaningful in that they may form or be enriched for biological pathways that correspond to clusters of correlated transcripts. It identifies clinically interesting modules from thousands of genes and uses gene significance (correlation between gene expression profile and a sample trait) to distinguish key genes in the disease pathway for further investigation. Thus the biological information within the network helps to reveal the function of the novel genes identified by RNA-seq.

3.2 Aims and Objectives

1. To use RNA-seq to assay the transcriptome of $CD4^+$ T-cells, a cell type crucial for disease pathogenesis, in individuals with CD and those without CD, using various stimuli to determine characteristic gene expression patterns in CD.

2. To combine this information with CD associated variants to establish which genetic associations are mirrored by altered gene expression in a physiologically relevant cell model.

3. To conduct pathway enrichment analysis for differentially expressed genes using Goseq and KEGG pathway annotations to identify key biological pathways involved in CD.

4. To use weighted gene co-expression network analysis (WGCNA) to group genes into biologically related expression modules composed of potentially functionally related genes

3.3 Materials and Methods

3.3.1 Collection and processing of blood samples

Detailed sample information is given in Table 3.1 and Table 3.2. Written and informed consent was obtained from all subjects, with approval of St. James's Hospital, and the Adelaide and Meath and National Children's Hospital (AMNCH) Research Ethics Committee. Our control samples were healthy volunteers (not screened for CD) from the general population. None of our volunteers were taking immunosuppressive drugs nor had any known autoimmune disease or family history of any autoimmune disease. All individuals are Caucasian and have Irish ancestry. All coeliac patients were on a gluten-free diet (GFD) at the time of blood collection. 30mls of blood was collected from each coeliac patient (mean age 53 +/-12) and control subjects (mean age 51 +/-9) with 24mls used for $CD4^+$ T-cell isolation and subsequent RNA extraction.

$CD4^+$ T-cells were isolated using positive selection from each blood sample using $CD4^+$ microbeads using the manufacturer specified procedures (Miltenyi Biotec GmBH). $CD4^+$ T-cells were then divided into 3 separate groups and stimulated with either 10ul of immobilized anti-CD3 antibody (1mg/ml)(HIT3a, Becton Dickinson) and 1ul anti-CD28 antibody (1mg/ml)(CD28.2, Becton Dickinson) for 24 hours at 37°C, 10ul (1ng/ml) Phorbal 12-myristate 13-acetate (PMA) for 4 hours at 37°C or left unstimulated (UNS) to represent basal conditions and incubated for 24 hours at 37°C. The purity of our $CD4^+$ T-cells was quantified at various stages throughout the isolation process with yields exceeding 99% purity post culture (Figure 3.1 and Appendix Fig A).

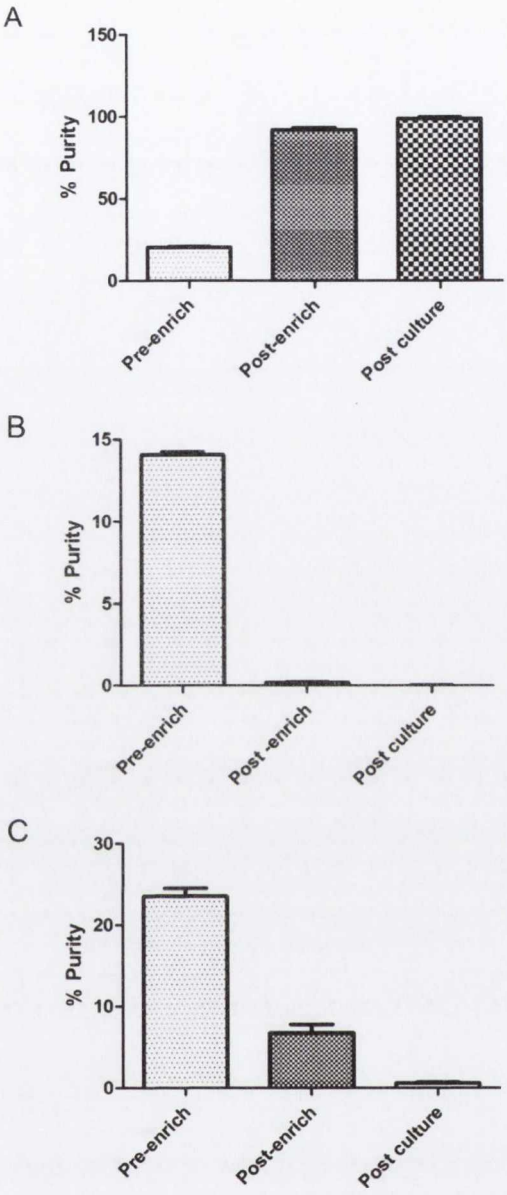


Figure 3.1: Purity of (A) $CD4^+$, (B) $CD8^+$ and (C) $CD14^+$ T-cells from human peripheral blood leucocytes. Pre-enrichment for $CD4^+$ T-cells using CD4+ microbeads, following enrichment for $CD4^+$ T-cells using CD4+ microbeads and after incubation (post culture) for 24 hours at 37° C.

3.3.2 RNA extraction

Total poly(A)$^+$ RNA was extracted from ~10^{-6} cells using the Qiagen RNeasy Micro kit according to the manufacturers protocol for all samples. The RNA quality was checked using an Agilent 2100 bioanalyser and only samples with a RIN of 8 or greater were included in the study.

3.3.3 Construction of the cDNA libraries and mRNA-sequencing

The cDNA library samples are prepared using Illumina ® TruSeq™ RNA Sample Preparation Kit. cDNA library preparation and sequencing was performed on an Illumina HiSeq™ 2000 using 50 base pair single end reads by GATC Biotech (Konstanz, Germany). Where necessary, samples were re-sequenced to ensure comparable coverage across samples and read data from these technical replicates merged prior to alignment. 4 samples failed library preparation leaving a total of 74 samples. On average across the 74 samples we generated 50 +/- 11 (mean +/-SD) million reads per sample. Sequence data quality control was evaluated using the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and any reads with a substandard quality or adapter contamination were removed. The QC steps undertaken are displayed in appendix Fig B-E.

3.3.4 Alignment and Differential expression analysis

Reads were aligned to the UCSC hg19 human reference genome using Tophat2 v2.0.3 (Kim et al., 2013). Up to two mismatches were allowed per uniquely aligned read.

Alignment metrics were then collected using the *PICARD CollectRNASeqMetrics (http://broadinstitute.github.io/picard/).* On average across the 74 samples; 93% (+/- 0.4) of reads were uniquely mapped with an average of 84% mapping to mRNA (57% and 27% mapping to coding and UTR regions respectively). After alignment, the read counts for each gene were summarized using the program *Htseq-count* (Anders et al., 2014) from the HTseq package. Normalisation and differential expression between each group of cases and controls was evaluated using DESeq2 (Love et al., 2014) version 1.6.0, implemented in R 3.01.1. DESeq2 uses a negative binomial generalised linear model (GLM) to evaluate differential gene expression. For significance testing, DESeq2 uses a Wald test statistic following normalisation based on size factors and variance. We adjusted for gender, sequencing batch and age covariates in the analysis. A gene base mean cut off of 10 and adjusted P-value (P_{adj}) ≤ 0.05, using the Benjamini-Hochberg method and a log2 fold change of above 1 or below -1 was used as the default threshold to define differential expression. Venn diagrams of overlapping genes between each group were generated using Venny *(http://bioinfogp.cnb.csic.es/tools/venny/index.html).*

3.3.5 Power calculations for RNA-sequencing

The Bioconductor tool *RNASeqPower* was used to calculate the power of our sample size to observe significant changes in expression. We estimated this based on our sample size of 11 biological replicates in the control group and 15 in the coeliac group at a range of read depths using a coefficient of variation (CV) of 0.4. The coefficient of

variation is the biological variability in gene expression intensity for each gene. A CV threshold of 0.4 was set in order to detect genes that vary above what is expected based on their mean expression across all samples.

3.3.6 Pathway analysis of differentially expressed genes

GOseq (Young et al., 2010) was used to identify gene ontology terms and KEGG biological pathways that were significantly enriched in differentially expressed genes from the RNA-seq analysis. GOseq specifically adjusts for the higher abundance of reads from long or highly expressed genes in RNA-seq experiments when assessing a gene list for over-representation. The false discovery rate from multiple-hypothesis testing was accounted for with Benjamini–Hochberg correction. GO terms and KEGG pathways with adjusted P values less than 0.05 were considered significantly enriched with differentially expressed genes. To evaluate modules identified in WGCNA we used the commercially available software MetaCore (http://www.genego.com/metacore.php) which conducts pathway analysis based on manually curated database of transcription factors, protein-protein interactions and signaling and metabolic pathways. Pathways with adjusted *P*-value <0.05 were considered significantly enriched for module genes.

3.3.7 Analysis of differential exon usage using DEXSeq

We used the biooconductor package DEXSeq (v1.12) (Anders et al., 2012) to identify differential expression of exons between coeliacs and controls. For a given exon,

DEXSeq considers the ratio of the number of reads mapping to this exon compared to the number of reads mapping to any other exon of the same gene in each sample. Normalisation and dispersion estimates for each exon are calculated similarly to DESeq2 and the ratio of counts across conditions is used to infer differential usage of exons using a generalized linear model (GLM). Data was prepared as outlined in the package vignette correcting for age, gender and sequencing batch in the GLM.

3.3.8 Weighted Gene Co-expression Network Analysis (WGCNA)

The R package WGCNA (Langfelder and Horvath, 2008) was used to identify co-expression modules in a dataset consisting of the anti-CD3/CD28 stimulated cases and controls. Counts transformed using the *rlg* transformation implemented in DESeq2 (Love et al., 2014) were used as input. To reduce background noise, genes with less than 10 read counts in each sample were removed, resulting in a dataset consisting of 10,787 genes. A co-expression network was constructed using WGCNA using standard procedures. Here we used a 'signed' network construction method in which the similarity between genes reflects the sign of the correlation of their expression profiles. It identifies modules with more specific expression patterns then an 'unsigned' network and is more biologically meaningful. The adjacency matrix was created by calculating the bi-weight mid-correlation coefficients between all genes raised to a β power of 20 (based on the scale-free topology criterion recommendations for datasets of this size). Next genes were clustered into network modules using the Topological Overlap Measure (TOM). The dissimilarity TOM (1-TOM) was then used as

input for hierarchical clustering, and modules (clusters of highly interconnected genes) were detected as branches of the dendrogram using the *DynamicTreeCut* algorithm. Highly correlated module genes are represented and summarized by their first principal component (referred to as the module eigengene, or ME). The ME is used to define measures of module membership (MM) or connectivity (kME), the correlation between a gene expression value and the module eigengene. Module hub genes (highly connected genes) were identified by ranking genes by module membership. Pearson's correlations were calculated between the module eigengene and disease status in order to find modules that were correlated with coeliac disease. We used the *in silico* tool Disease Association Protein-Protein Link Evaluator (DAPPLE) (Rossin et al., 2011) to test if there was an excess of protein-protein interactions within co expression modules. Genes from each module with a MM > 0.7 were submitted to DAPPLE as seed nodes and the resulting direct and indirect (through other proteins) interaction networks were built based on data from the InWeb database of protein interaction using 1,000 permutations and an interacting binding degree cut-off of 2.

3.3.9 Testing for enrichment for genetic associations

We extracted the Immunochip assocation P-values from the Trynka *et al* study comprising 12,014 individuals with coeliac disese and 12,228 controls for SNPs located within non-HLA genes differentially expressed in our anti-CD3/CD28 stimulated samples (excluding genes located within the MHC locus) and we also carried out this analysis having removed the GWS SNPs to determine whether our genes were

enriched for genes falling short of GWS. The Quantile-quantile (QQ) plot was generated using Stata v10. We used INRICH (interval-based enrichment analysis for genome-wide association studies) (Lee et al., 2012) to evaluate whether differentially expressed genes between coeliacs and controls were enriched for Coeliac Immunochip P-values (Trynka et al., 2011). INRICH implements a permutation-based algorithm to test for enrichment whilst correcting for possible biases induced by marker density, linkage disequilibrium (LD) structure, and gene size. We first summarised Coeliac Immunochip results using the "*--clump*" algorithm in PLINK (Purcell et al., 2007; Chang et al., 2015). P-value cutoffs of P_1 <0.01 (significance threshold for index SNPs) and P_2 <0.05 (the secondary significance threshold for clumped SNPs) and an r^2 threshold of 0.2 was used. INRICH tests whether there is enrichment within these associated regions (intervals) for particular sets of targets (in this instance; genes differentially expressed in our dataset). We applied 10,000 permutations in our first round of INRICH enrichment analysis and 5,000 in the bootstrap replication round for correction of multiple testing.

All raw sequence data and processed files from htseq have been submitted to the GEO (currently awaiting a reference ID).

3.3.10 Gene Set Enrichment Analysis (GSEA)

We used gene set enrichment analysis (GSEA) (Subramanian et al., 2005) to quantify enrichment of genes identified as regulated by BACH2 amongst genes that were

differentially regulated between cases and controls using all three conditions (anti-CD3/CD28, PMA and UNS). We used the gene sets discovered by Roychaudhuri *et al* (Roychoudhuri et al., 2013) and Vahedi *et al* (Vahedi et al., 2015). We used only genes which we could identify in the human genome from the Roychaudhuri gene set, giving a total of 123 genes and 98 in the Vahedi gene set. We ran the GSEA against normalised gene counts for the entire transcriptome for each condition with 1000 permutations and used an FDR q-value of 0.05 as significant.

3.4 Results

3.4.1 mRNA-sequencing

We sequenced the $CD4^+$ T-cell transcriptomes of 15 individuals with coeliac disease and 11 controls under three different conditions; cells were unstimulated (UNS), stimulated with anti-CD3/CD28 antibody and stimulated with phorbal 12-myristate 13-acetate (PMA), comprising a total of 74 transcriptomes. Samples are described in Table 3.1.

3.4.1.1 Power analysis of transcriptome data

It has been previously shown that with a sequencing depth of 10 million reads, approximately 90% of all genes will be covered by at least 10 reads (Wang et al., 2011). Furthermore, 30–40 million reads is sufficient to allow technically precise measurement of gene expression for most genes surveyed. As stated above, across the 74 samples we generated 50 ± 11 (mean ± SD) million reads per sample. Using an RNA-seq power calculator (Wang et al., 2011), it was estimated that, in a sample size of 11 controls versus 15 cases at average transcript coverage of 10 read counts, we had 93% power to detect a 2-fold change in expression. This increases to over 97% at 20x coverage (Fig 3.2) and approaches 100% for more highly expressed genes, indicating that the study was adequately powered to reliably detect changes in gene expression.

Table 3.1: Sample Information for individuals with coeliac disease and controls sequenced in each group (CD3/CD28, PMA and UNS).

Characteristic		Median or n	%[1]	Range	anti-CD3/CD28	PMA	Unstimulated
	Coeliac individuals						
Age, n 15		57.0		22-75	15	14	13
Females/Males, n 15		11/4	73.3/26.7				
Diagnosis age, n 15		39.0		0.5-70			
Diagnosis by age range, n 15	< 18 years	5	33.3				
	≥18-35 years	2	13.3				
	>35-55 years	5	33.3				
	>55	3	20.0				
Clinical presentation at diagnosis*, n 12	Classical CD	9	75.0				
	Non-classical CD	2	16.7				
	Subclinical CD	1	8.3				
Family history of coeliac disease, n 15	Yes	5	33.3				
	No	10	66.7				
Family history of other autoimmune disease, n 15	Yes	5	33.3				
	No	10	66.7				
Presence of other autoimmune diseases, n15	Yes	4	26.7				
	No	11	73.3				
	Control individuals						
Age, n 11		50.0		40-70	11	11	10
Females/Males, n 11		7/4	63.6/36.3				

* Clinical presentation at diagnosis was classified according to the Oslo definition of CD into; Classical, Non-classical and Subclinical

† Percentages provided are valid percentages

Table 3.2: Sample information for all individuals with coeliac disease and controls sequenced in this study.

Characteristic		Median or n				P-value
	Coeliac Patients	Coeliac/Coeliac		Control/Coeliac		
			Range		Range	
Age n 15 (years)		58.0	22-75	56.0	49-58	**0.242**
Age of diagnosis n 15 (years)		44.5	2-70	2.0	0.5-45	**0.057**
Clinical presentation at diagnosis*, n 12	Classical CD	5		4		
	Non-classical CD	1		1		
	Subclinical CD	1		0		**0.559**
Family history of coeliac disease, n 15	Yes	5		0		
	No	5		5		**0.101**
Family history of other autoimmune disease, n 15	Yes	2		3		
	No	8		2		**0.251**
Presence of other autoimmune diseases , n 15	Yes	4		0		
	No	6		5		**0.231**
Compliance with diet at sample collection, n 15	Yes	6		5		
	No	4		0		**0.231**
tTG at sample collection , n 15 (U/ml)		5.35	0.2-101	0.8	0.4-3.7	**0.098**

* Clinical presentation at diagnosis was classified according to the Oslo definition of CD into; Classical, Non-classical and Subclinical

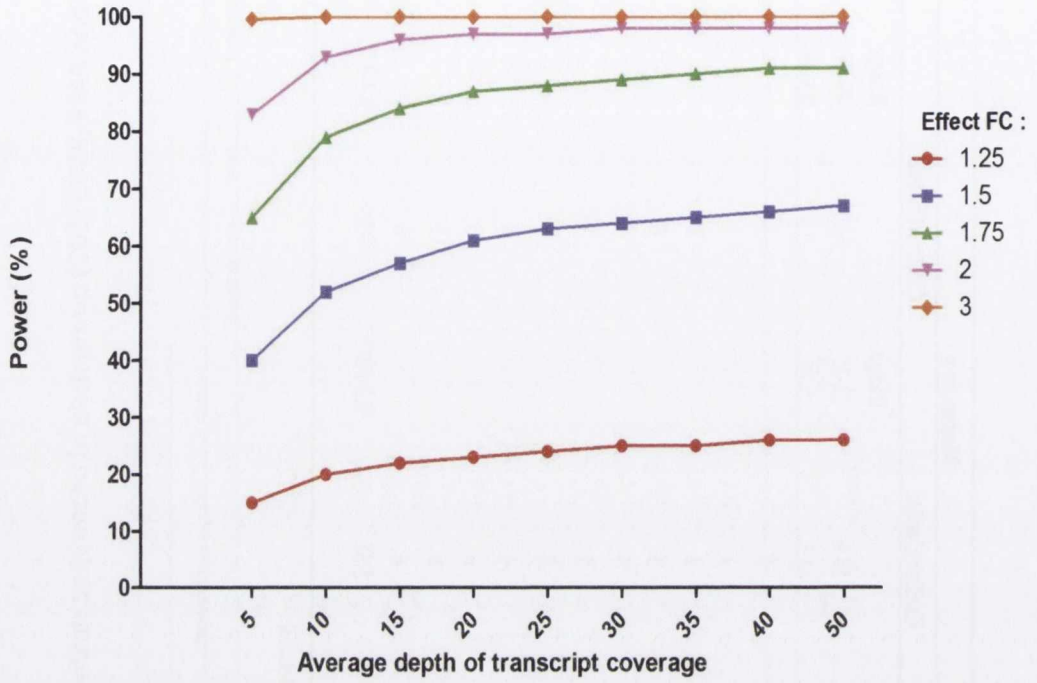


Figure 3.2: Percentage of detected genes at different levels of sequence depth at different fold changes across all annotated genes.

3.4.1.2 PCA analysis

PCA analysis of the normalised expression data showed that the unstimulated samples do not form any distinct groups indicating that overall gene expression patterns do not differ between case and control groups in the absence of stimulation (Fig 3.3c). However, when the T-cells were activated either by using anti-CD3/CD28 or PMA stimulation clustering is seen (Fig 3.3a and 3.3b). Interestingly, the patients separate into two groups with the same 5 coeliac samples clustering with controls upon stimulation with either anti-CD3/CD28 or PMA. When the two patient sample clusters

were assessed for differences in patient characteristics, including age at sampling, gender, clinical presentation and potential technical confounders such as sequencing batch there were no significant differences between the groups (Table 3.2). However a higher proportion of patients with control-like expression profiles were diagnosed during infancy (3/5 vs 2/10; median age in years: 2, 0.5-45.0; 44.5, 2-70 P = 0.057). Although all patients self-reported adherence to the gluten free diet at time of sampling, patients with a coeliac like expression profile for whom data was available had higher median tTG levels at the most recent time of sampling (within 6 months) (5.35 U/ml, 0.2-101; vs 0.8, 0.4-3.7; P = 0.098) and a majority of these patients had one or more instances of abnormal tTG (>7.0 U/ml) levels, whereas the control-like CD patients had consistently low tTG levels, over the previous 3 years on review of their hospital charts. Other risk factors appeared more prevalent in the CD-like group, including family history of CD and comorbidity with another autoimmune disease, however these were not significant. As these patients met the patient criteria they were included in the analysis under the speculation that they form part of a subgroup of the disease phenotype or their expression profiles were modified by long-term remission.

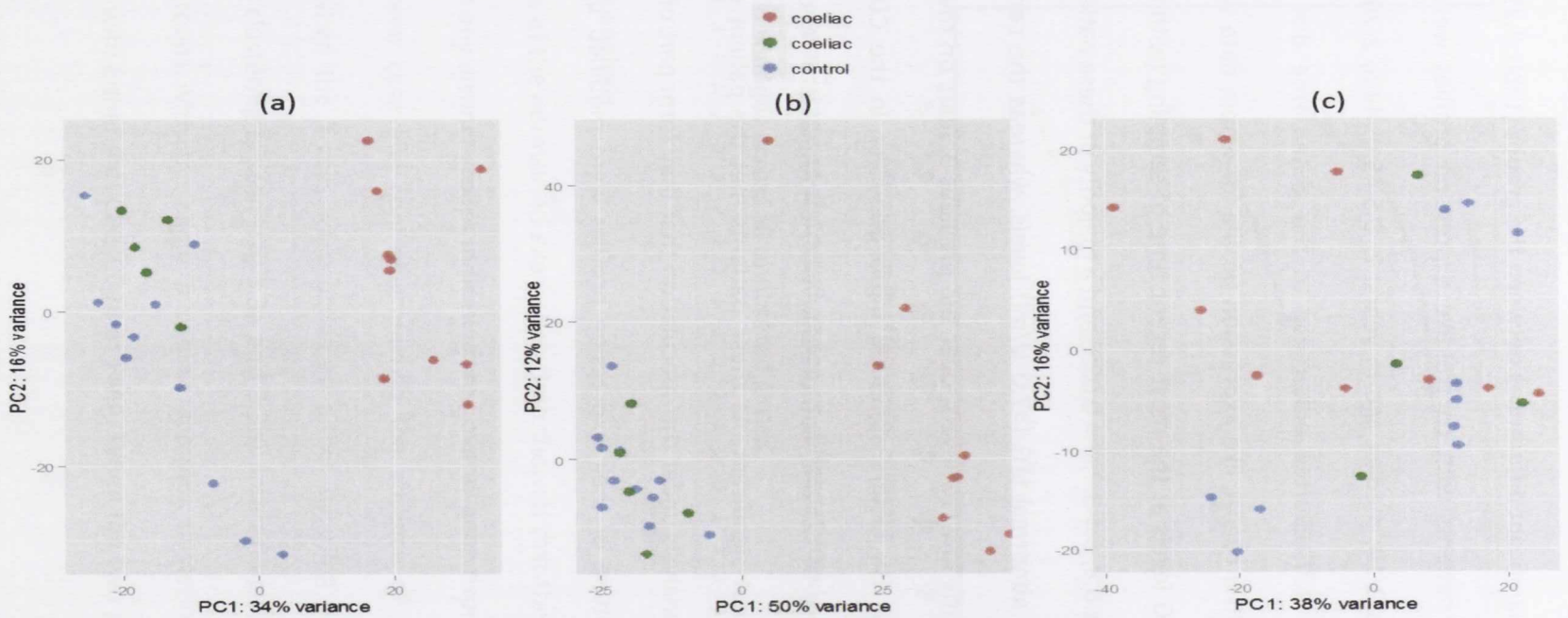


Figure 3.3: PCA plots based on normalised read counts from the RNA-seq datasets. There is distinct separation of the majority of celiac samples and controls in both anti-CD3/CD28 (a) and PMA (b) sample sets however UNS (c) do not form distinct groups.

3.4.2 Differential expression analysis

We identified 1069 differentially expressed (DE) genes between coeliacs and controls in the anti-CD3/CD28 stimulated group using DEseq2 (defined as a log two fold change of 2 [Log2FC] or greater, and a false discovery rate [FDR] of 5% or less). 657 genes were up regulated and 412 were down regulated (Fig 3.4). The most strongly up regulated gene in CD samples was *IFNG* which is known to be highly expressed in the coeliac lesion, with a log2FC of 4.6 (P_{adj} = $2.40x10^{-11}$) and the strongest gene down regulated in CD was *NOG* (Noggin) which functions in the TGFβ signaling pathway specifically inhibiting BMP, which is thought to play a role in gut epithelial cell homeostasis (Crosnier et al., 2006) (log2FC= -3.4, P_{adj} = $7.87x10^{-08}$). The top 25 most significantly differentially regulated genes in the anti-CD3/CD28 group are listed in Table 3.3.

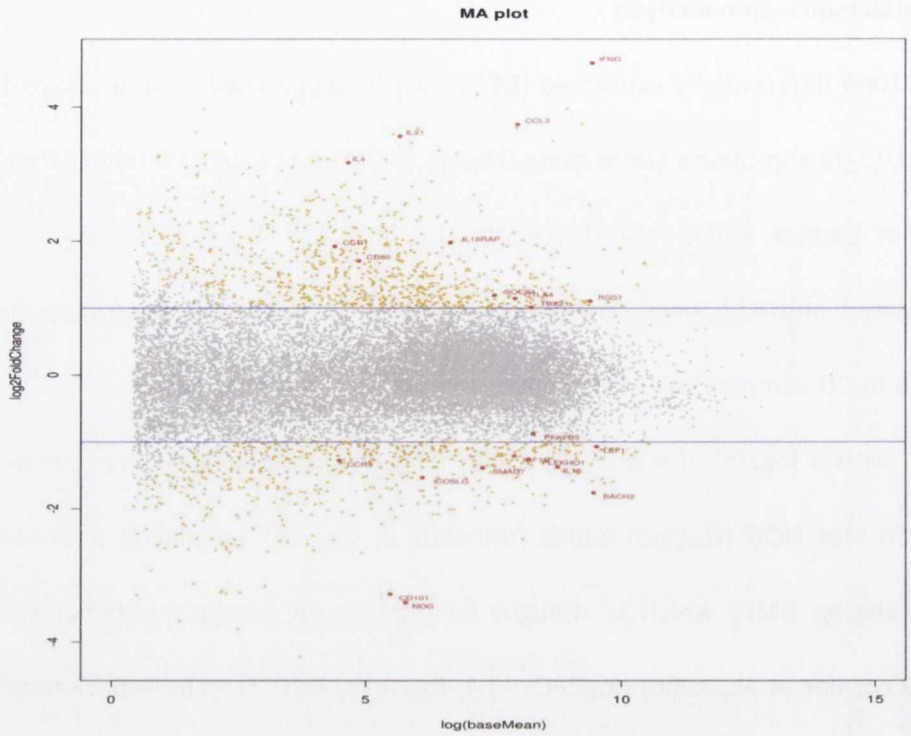


Figure 3.4: Scatter plot of fold-change in expression in cases versus controls (y-axis) against expression level (x-axis). 2 fold up or down regulated genes with adjusted P <0.05 are highlighted in yellow. Differentially expressed genes of interest are highlighted with red dots and labels. Horizontal blue lines indicate 2-fold changes in expression.

Table 3.3: The top 25 most significantly differentially regulated genes in the anti-CD3/CD28 group.

Upregulated				Downregulated			
Gene	basemean	log2foldchange	padj	Gene	basemean	log2foldchange	padj
IFNG	12425.42	4.659316	2.40E-11	NOG	324.9054	-3.407037	7.87E-08
ZBTB32	409.7372	3.67949	2.47E-11	WNT7A	137.4844	-2.662605	8.87E-08
GZMH	1311.553	4.189856	2.47E-11	PITPNC1	1717.513	-1.449887	1.05E-07
FAM40B	252.925	3.591436	7.59E-11	CD101	239.9696	-3.281365	1.16E-07
GZMB	7627.356	3.976255	8.70E-11	TSC22D3	3214.326	-2.455231	2.02E-07
ZBED2	10326.91	3.746521	5.78E-10	P2RY6	12.11319	-3.804131	3.43E-07
PRSS23	34.16801	4.069848	2.81E-09	IRF2BP2	7433.654	-1.126946	8.87E-07
IL4	105.3752	3.17195	3.56E-09	SPOCK2	26908.3	-1.194497	1.13E-06
CCL3	2906.318	3.742559	4.56E-09	TCF7	15906.77	-1.61789	1.26E-06
TNFRSF4	4150.351	1.934196	9.90E-09	CD5	12236.08	-1.332533	1.36E-06
FANCA	174.383	2.575302	1.09E-08	EPHB6	819.7302	-1.486736	2.37E-06
GLDC	60.1645	3.230646	1.09E-08	MMP28	20.62269	-3.157811	2.60E-06
IL13	620.9891	3.353638	1.09E-08	FAM102A	15210.63	-1.552252	4.28E-06
LIF	2842.316	3.686361	1.09E-08	TNFAIP8L1	1251.469	-1.343781	5.42E-06
IL21	290.0744	3.564816	1.67E-08	BNIP3L	1861.533	-1.367198	5.82E-06
OSM	879.5017	3.311707	5.10E-08	ARRDC2	3934.31	-1.776797	6.98E-06
CCL4	5816.93	3.496838	5.10E-08	IL9R	197.4241	-1.820885	1.15E-05
CSF2	1424.299	2.450557	7.45E-08	BCL9L	5004.654	-1.792264	1.15E-05
TYMP	1394.521	2.539145	1.04E-07	ICOSLG	448.6913	-1.53195	1.15E-05
IQCG	432.7306	2.364276	1.21E-07	EDAR	300.5045	-1.950019	1.20E-05
EVI5	362.2564	2.81671	1.29E-07	SIRPB1	164.5306	-1.910829	1.30E-05
IFT57	940.3505	1.09151	2.77E-07	C2orf55	34.11497	-1.767445	1.36E-05
CMAHP	947.484	1.358978	3.59E-07	TTC28	39.84447	-1.137565	1.57E-05
CD70	180.4551	2.655766	6.40E-07	ZNF853	229.6861	-1.550303	1.60E-05

In the PMA group, there were 2,007 DE genes, 1726 up regulated and 281 down regulated. The most strongly up regulated gene in CD samples was the purinergic receptor *P2RX7* (log2FC= 3.44; $P_{adj} = 8.15x10^{-10}$). P2RX7 has been shown to be heavily involved in the regulation of inflammation in the gut and increased *P2RX7* mRNA levels have been observed in biopsy tissue from individuals with IBD. The most down regulated gene in CD was the RASD family member 2 (*RASD2*) gene (log2FC=-2.8; $P_{adj} = 1.7x10^{-05}$). RASD2 is a GTP binding protein involved in the mTOR signalling pathway (Subramaniam et al., 2012). The top 25 most significantly differentially regulated genes in the PMA group are listed in appendix Table A.

In the unstimulated (UNS) group, 87 genes were differentially expressed of which 63 were up regulated and 24 down regulated. The top 25 most significantly differentially up regulated genes and the 24 down regulated genes in this group are listed in appendix Table B. In this group, the most strongly up regulated gene in coeliac samples was *GRAMD1C* with a log2FC of 2.41 ($P_{adj} = 8.45x10^{-05}$). The most strongly down regulated gene in coeliacs was the RAS-like family 10 member B (*RASL10B*) (log2FC= -2.63; $P_{adj} = 7.43x10^{-06}$). Neither of these genes have well documented functions.

Figure 3.5 displays the overlap in significantly DE genes (up or down regulated) in each group. The three groups are replicates of the same samples using different or no stimulation. There was high concordance (94% PMA, 100% UNS) in the direction of

fold change (up or down) between any genes that overlapped under the different stimulations, adding confidence that the observed changes in gene expression are real and reproducible.

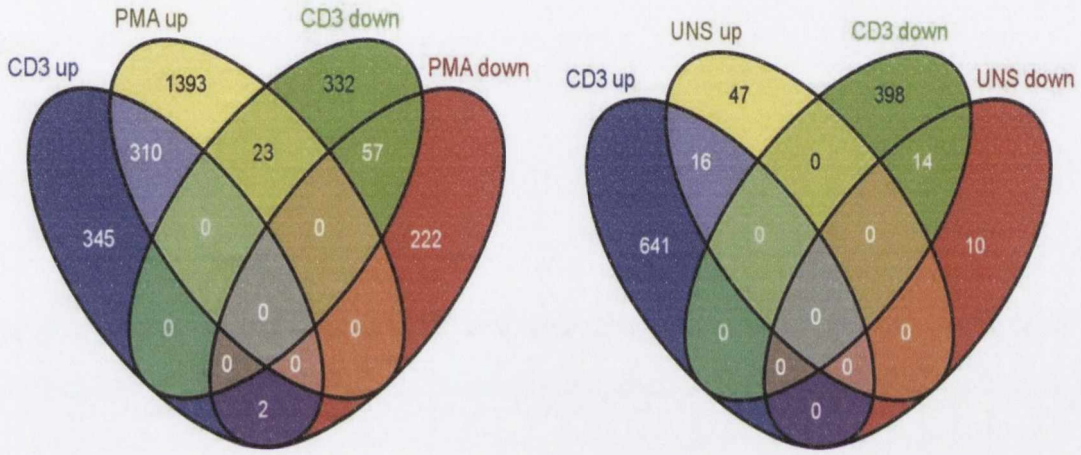


Figure 3.5: Venn diagram indicating the overlap in significantly DE genes (up/down) in each group.

3.4.3 T-cell subpopulations

To evaluate the possibility that different levels of T-cell subpopulations were influencing our gene expression results in our dataset we examined transcript expression between the anti-CD3/CD28 stimulated and unstimulated coeliac case cohort and controls for a subset of $CD4^+$ Th1 *(TBX21),* Th2 *(GATA3),* Th17 *(RORC)* and T-reg *(FOXP3)* gene markers (Table 3.4). From the four gene markers analysed, only one showed a significant divergence in transcript expression between our cases and controls at a threshold of P >0.05, the Th1 cell gene marker *TBX21* following anti-CD3/C28 stimulation (Log2FC=1.01,P_{adj} = 0.021). It should be noted that *TBX21* may be under the control of *BACH2* since it shows increased expression in Bach2 knock-out mice, and BACH2 is down regulated across all conditions (unstimulated log2FC =-0.72, P_{adj} = 0.052).

Table 3.4: Levels of transcript expression between our CD3 stimulated coeliac cohort and controls for a subset of $CD4^+$ Th1 (TBX21), Th2 (GATA3), Th17 (RORC) and T-reg (FOXP3) gene markers. Only TBX21 showed a significant divergence in transcript expression between cases and controls at a threshold of P >0.05 with a P_{adj} = 0.021.

Gene	Cell type	Base mean	Log2 foldchange	lfcSE	Stat	P-value	padj
TBX21	Th1	3743.21	1.014745	0.3374561	3.007042	0.002638	0.0216536
GATA3	Th2	2309.901	-0.4355618	0.2013452	-2.163259	0.0305213	0.1131209
RORC	Th17	1493.305	-0.5508857	0.4086495	-1.348064	0.1776378	0.3621348
FOXP3+	Treg	3411.863	0.5596505	0.3307828	1.800777	0.071738	0.2027844

3.4.4 Pathway analysis of differentially expressed genes

Pathway enrichment analyses were conducted for DE genes in each of the three groups.

3.4.4.1 anti-CD3/CD28 stimulation

Genes that play a role in the immune system were strongly represented amongst the DE genes; Gene ontology analysis of the anti-CD3/CD28 DE genes using the pathway analysis tool GOseq, showed they were highly enriched for the GO terms: "Immune system processes"; GO:0002376: (P_{adj} = $1.67x10^{-23}$) and the "Immune response"; GO:0006955 (P_{adj} = $9.28x10^{-18}$). The KEGG pathway most significantly enriched for DE genes amongst the anti-CD3/CD28 stimulated samples was 'Cytokine-cytokine receptor interaction' (P_{adj} = $2.16x10^{-17}$). Forty eight of our DE genes belong to this pathway, including 4 genes that have been genetically linked to coeliac disease: *CCR1, IL18RAP, IL21* and *IL2RA* (Dubois et al., 2010; Trynka et al., 2011). A summary of other significantly enriched interesting pathways are listed in Table 3.5. Pathways enriched for genetically associated genes that show evidence of differential expression in our dataset include: the JAK/STAT signaling pathway (Fig 3.6), type-1 diabetes, rheumatoid arthritis, chemokine signaling and T-cell receptor signaling pathways. The R tool Pathview (Luo and Brouwer, 2013) was used to generate a graphical representation of the cytokine-cytokine receptor interaction pathway as an example, indicating which genes are differentially regulated (Fig 3.7).

Table 3.5: Pathways identified by GOSeq as significantly enriched for genes differentially expressed in anti-CD3/CD28 stimulated samples.

Pathway	# DE genes	# Genes in pathway	Adjusted P value
Cytokine-cytokine receptor interaction	48	259	2.16E-17
Jak-STAT signaling pathway	27	151	1.66E-09
Osteoclast differentiation	21	128	1.36E-06
Metabolic pathways	82	1104	5.37E-05
Toll-like receptor signaling pathway	15	101	7.41E-05
Type I diabetes mellitus	9	43	1.10E-04
Allograft rejection	8	37	1.73E-04
Rheumatoid arthritis	13	89	1.89E-04
Intestinal immune network for IgA production	8	48	1.07E-03
Asthma	6	30	1.09E-03
Fc epsilon RI signaling pathway	11	79	1.24E-03
Wnt signaling pathway	17	149	1.36E-03
Galactose metabolism	6	27	1.74E-03
Graft-versus-host disease	7	41	2.18E-03
TGF-beta signaling pathway	11	84	3.15E-03
Type II diabetes mellitus	7	47	8.56E-03
Chemokine signaling pathway	16	188	1.35E-02
Folate biosynthesis	3	11	1.39E-02
Autoimmune thyroid disease	6	52	1.83E-02
MAPK signaling pathway	21	267	2.36E-02
T cell receptor signaling pathway	10	107	3.97E-02

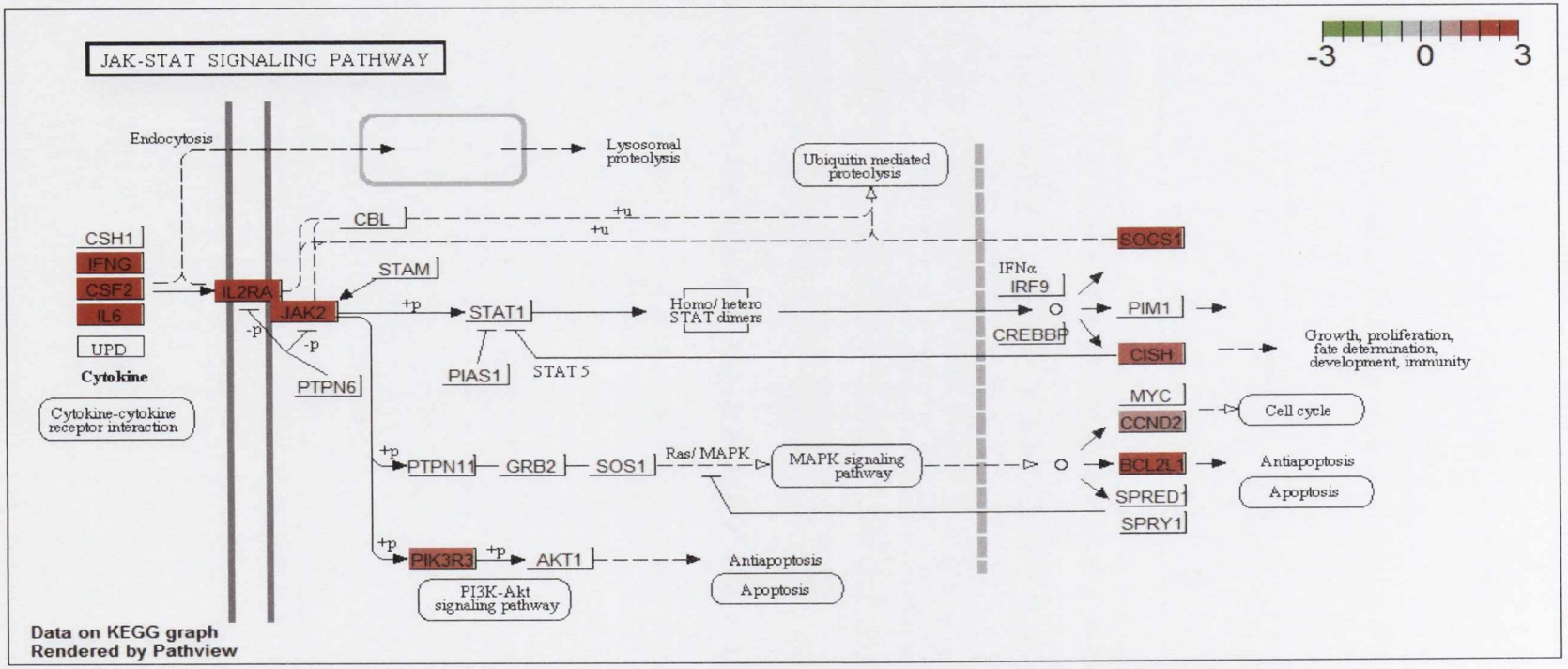


Figure 3.6: The JAK-STAT signaling pathway. Pathway analysis of genes differentially expressed in the anti-CD3/CD28 stimulated dataset found the JAK-STAT signaling pathway to be most significantly enriched. The direction of fold change in coeliac over controls for the genes involved is indicated in red (up regulated) or green (down regulated).

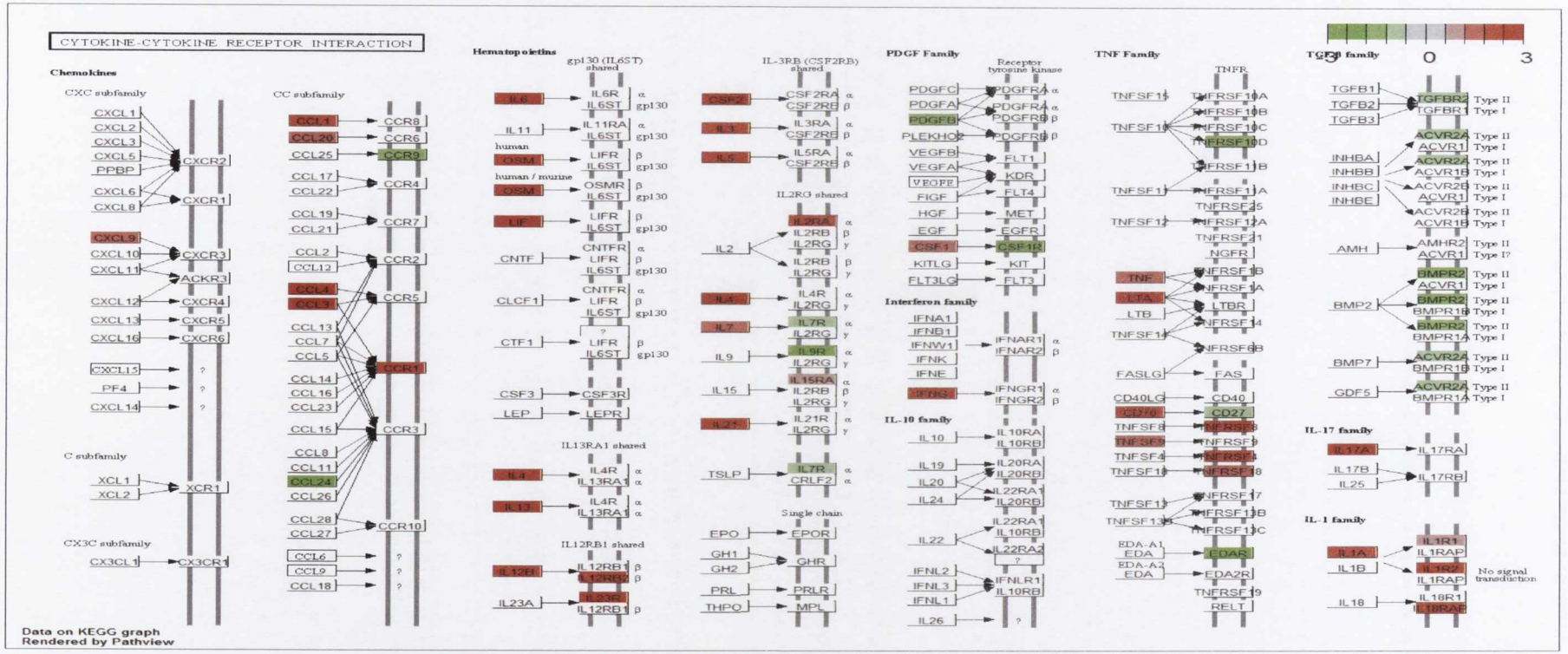


Figure 3.7: The cytokine-cytokine receptor pathway. Pathway analysis of genes differentially expressed in the anti-CD3/CD28 stimulated dataset found the Cytokine-cytokine receptor pathway to be most significantly enriched. The direction of fold change in coeliac over controls for the genes involved is indicated in red (up regulated) or green (down regulated).

3.4.4.2 PMA stimulation

Again there was significant enrichment of immune related genes amongst our PMA stimulated samples ("Immune system processes"; GO: 0002376 (P_{adj} = $2.92x10^{-16}$), "Immune response"; GO: 0006955 (P_{adj} = $1.63x10^{-15}$). There was a large overlap between pathways which were significantly enriched for DE genes in the anti-CD3/CD28 and PMA groups, which is unsurprising given the high degree of overlap in the two stimulated DE gene sets. Pathways with excess representation of DE genes after PMA stimulation are shown in Table 3.6.

3.4.4.3 Unstimulated

No Go terms or KEGG pathways showed any significant enrichment for genes differentially expressed in the unstimulated group.

Table 3.6: Pathways identified by GOSeq as significantly enriched for genes differentially expressed in PMA stimulated samples.

Pathway	#DE genes	# Genes in pathway	Adjusted P value
Metabolic pathways	178	1104	6.50E-13
Valine, leucine and isoleucine degradation	24	44	4.68E-12
Butanoate metabolism	14	30	9.95E-06
Peroxisome	23	79	1.77E-05
Cytokine-cytokine receptor interaction	46	259	3.59E-05
Propanoate metabolism	13	32	8.93E-05
Tryptophan metabolism	14	41	3.61E-04
Synthesis and degradation of ketone bodies	6	9	1.14E-03
Jak-STAT signaling pathway	29	151	1.14E-03
Fatty acid metabolism	13	43	2.64E-03
Lysine degradation	12	42	8.58E-03
Glyoxylate and dicarboxylate metabolism	7	18	1.05E-02
beta-Alanine metabolism	8	22	1.05E-02
Pyrimidine metabolism	19	97	1.05E-02
Lysosome	22	119	2.05E-02
Pyruvate metabolism	10	39	2.86E-02
Chemokine signaling pathway	28	188	4.36E-02
Rheumatoid arthritis	16	89	4.36E-02
Hepatitis C	22	133	4.76E-02
Cytosolic DNA-sensing pathway	11	56	4.87E-02

3.4.5 Differential exon analysis

We also tested for differential usage of exons between coeliac and controls using the Bioconductor tool DEXSeq. We identified 647 exons from 542 genes with differential usage between the coeliacs and controls in the anti-CD3/CD28 stimulated samples (FDR<5%). Only 26 of these genes were DE based on the gene level analysis. Interestingly, 4 coeliac associated genes; *YDJC, SH2B3, ITGA4* and *UBE2E3* were amongst those containing significantly differentially expressed exons but were not DE at the gene level (Table 3.7).

Table 3.7: Differential exon usage in our anti-CD3/CD28 stimulated samples at 4 genes that have been genetically associated with coeliac disease.

Chromosome	Exon start	Exon end	Symbol	pvalue	padj	Exon #	log2foldchange
2	182322131	182322154	ITGA4	1.91E-04	0.045	E005	-2.465
12	111843752	111844081	SH2B3	1.08E-03	0.086	E001	-2.869
12	111855923	111856681	SH2B3	8.16E-05	0.034	E003	-1.398
2	181845333	181845533	UBE2E3	2.59E-04	0.050	E008	-1.729
22	21983299	21983476	YDJC	2.19E-04	0.047	E005	-1.196
22	21983597	21983696	YDJC	2.26E-07	0.003	E007	-0.403

3.4.6 Relating differentially expressed genes to disease associations

Over 40 non-HLA genomic loci have been associated with CD in recent years (Trynka et al., 2011; Garner et al., 2014; Coleman et al., 2015). eQTL and expression studies (Hrdlickova et al., 2011) have highlighted expression differences at a number of these loci in PBMCs, biopsies (Plaza-Izurieta et al., 2014) and epithelial cells. Therefore we examined if any of these genes were amongst those DE in our anti-CD3/CD28 T-cell data. Furthermore, many of the associated loci are large, encoding more than one gene, so we examined whether the expression data could identify a particular gene at a given locus amongst other potential candidates or whether there was coordinated expression of genes in these regions indicating a potential enhancer effect.

The non-HLA loci incorporate approximately 70 protein-coding genes. Of these, 11 are significantly DE between coeliacs and controls in our anti-CD3/CD28 stimulated samples, 11 were DE between coeliacs and controls in the PMA stimulated group and 3 were DE in the UNS samples. Several show differential and concordant expression under more than one condition, for example *BACH2*, *CCR1, SOCS1* and *IL18RAP,* and in total, 21 CD candidate susceptibility genes are DE under one or more of the conditions used in this study. These are tabulated in Table 3.8, which also shows expression patterns in multi-gene susceptibility loci.

This brings to a total of 25 the genes which have been associated in genome studies and which are either differentially regulated or show alternative exon usage in $CD4^+$ T-cells under the conditions utilised in this study.

Table 3.8: DESeq2 results across all groups for genes that have been genetically associated with coeliac disease and are also differentially expressed in our data. The direction of fold change in coeliac over controls for the genes involved is indicated in red (up regulated) or green (down regulated).

CEL locus	LD Region	Size(Mb)	Gene	CD3+ Log2FC	padj CD3	PMA Log2FC	padj PMA	UNS Log2FC	padj UNS	Candidate genes
1q31.2	chr1:192469607...192548229	0.08	RGS1	1.105	2.64E-02	1.065	6.00E-02	-0.020	9.88E-01	*RGS1*
2p14	chr2:68536253...68682256	0.15	PLEK	0.898	1.50E-01	1.485	1.74E-02	1.551	1.13E-01	*PLEK*,*FBX048*
2q12.1	chr2:102855241...103268442	0.41	IL1RL1	2.396	2.06E-04	0.530	4.58E-01	1.769	NA	
2q12.1	chr2:102855241...103268442	0.41	IL18R1	0.920	8.64E-03	1.859	8.93E-05	0.575	3.26E-01	*IL1RL2*, *IL1RL1*,
2q12.1	chr2:102855241...103268442	0.41	IL18RAP	1.974	2.05E-06	1.372	1.42E-02	1.582	3.02E-02	*IL18RAP*,*IL18R1*
2q33.2	chr2:204445827...204817281	0.37	CTLA4	1.147	1.36E-03	0.201	7.19E-01	0.276	7.25E-01	*CTLA4*, *CD28*, *ICOS*
3p21.31	chr3:45908148...46626589	0.72	CCR9	-1.277	4.53E-02	-0.245	7.95E-01	0.570	5.36E-01	
3p21.31	chr3:45908148...46626589	0.72	CCR1	1.919	1.59E-03	1.775	6.55E-04	0.726	5.27E-01	*CCR3*, *CCRL2*,
3p21.31	chr3:45908148...46626589	0.72	CCR2	0.309	6.84E-01	1.697	2.12E-02	0.439	6.88E-01	*CCR2*, *CCR1*,*CCR5*, *CCR9*, *LTF*
3p14.1	chr3:69270000...69370000	0.1	FRMD4B	2.012	4.40E-05	0.784	1.95E-01	0.610	3.21E-01	*FRMD4B*
3q13.33	chr3:119102458...119300655	0.2	CD80	1.704	7.58E-05	0.931	1.01E-01	0.227	8.56E-01	*CD80*, *ARHGAP31*, *POGLUT1*
4q27	chr4:122938634...123612510	0.67	IL2	0.927	6.26E-02	1.686	4.24E-02	-1.447	1.10E-01	*ADAD1*, *IL21*,
4q27	chr4:122938634...123612510	0.67	IL21	3.565	1.67E-08	4.097	NA	1.999	7.44E-02	*KIAA1109*, *IL2*
6p25.3	chr6:380341...467414	0.09	IRF4	0.167	7.62E-01	1.672	6.84E-03	0.176	8.17E-01	*IRF4*
6q15	chr6:90806835...91026426	0.22	BACH2	-1.759	3.66E-03	-1.263	1.72E-02	-0.722	5.19E-02	*BACH2*, *MAP3K7*
6q22.33	chr6:127954614...128341901	0.39	PTPRK	0.261	6.88E-01	0.649	3.47E-01	-1.876	8.45E-05	*THEMIS*, *PTPRK*
6q25.3	chr6:159318774...159541481	0.22	TAGAP	0.237	6.56E-01	1.155	8.58E-03	0.786	1.59E-01	*TAGAP*
10p15.1	chr10:6388071...6436034	0.05	PFKFB3	-0.874	3.60E-04	-0.492	1.25E-01	-1.146	1.15E-02	*DKFZP667F0711*, *PFKFB3* , *PRKCQ*
11q24.3	chr11:128338131...128425884	0.09	ETS1	-0.483	7.60E-02	1.205	1.35E-02	0.310	5.76E-01	*ETS1*
16p13.13	chr16:11307024...11477757	0.17	SOCS1	1.191	8.29E-04	1.578	1.18E-02	0.288	7.79E-01	*SOCS1* ,*PRM1* ,*PRM2*
21q22.3	chr21:45597579...45704644	0.11	ICOSLG	-1.532	1.15E-05	-0.772	8.42E-02	-0.706	2.69E-01	*ICOSLG*

3.4.7 Differentially expressed genes are enriched for coeliac association signals

A QQ plot of SNPs located within the boundaries of these DE genes indicated that our gene list was enriched for SNPs that had been associated with coeliac disease using the Immunochip (Fig 3.8). We used the tool INRICH (Interval-based Enrichment Analysis Tool for Genome-Wide Association Studies) to quantify this finding.

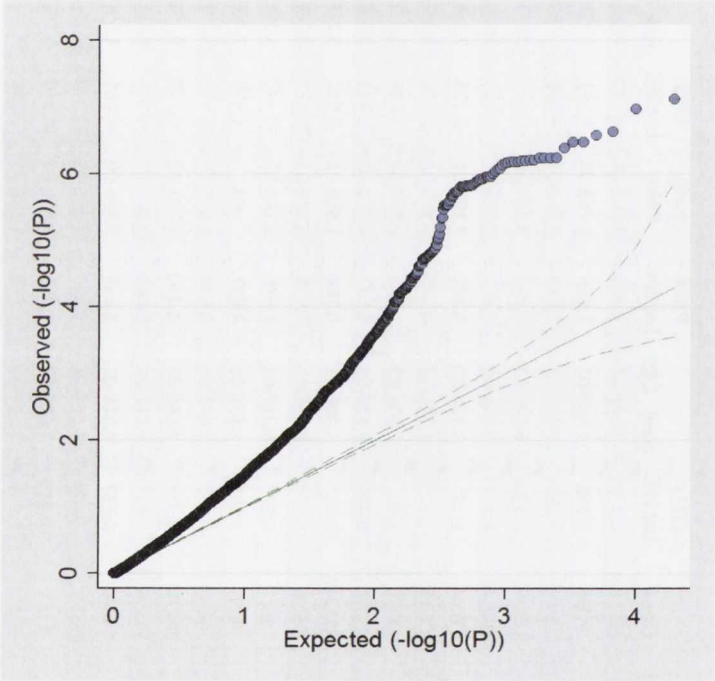


Figure 3.8: QQ plot of the Immunochip association signals within genes that were differentially expressed between our anti-CD3/CD28 stimulated coeliac and control samples. The expected –log10 P-values under the null hypothesis are represented on the x axis whilst the observed values are represented on the y axis.

After gene-set-based association analysis and multiple testing correction, there was evidence of strong enrichment for CD associated SNPs amongst our DE genes in the anti-CD3/CD28 group (P_{adj} = 0.002). Interestingly, when up or down regulated genes were tested for association separately, there was more significant enrichment for up regulated (P_{adj} up = 0.01), compared to the down regulated group (P_{adj} down = 0.07). There was no evidence for enrichment for CD associated genes in either the PMA (P_{adj} = 0.57) or UNS (P_{adj} = 0.38) DE gene sets.

We also tested whether any DE genes were found in loci falling below the threshold for GWS in CD. Using INRICH we found that the enrichment was still significant in the anti-CD3/CD28 group (P_{adj} = 0.01) after removal of GWS SNPs. This indicates that SNPs falling short of current benchmarks for GWS are nonetheless associated with differential gene expression in CD patients and could thus be contributing to disease.

3.4.8 Weighted gene co-expression network analysis

Weighted gene co-expression analysis (WGCNA) can be used to group genes into correlated biologically related expression modules comprised of potentially functionally related genes as described in section 3.1.4. Strongly correlated expression levels could imply participation in similar biological processes or common regulatory mechanisms such as regulation by a common set of transcription factors. We used WGCNA to analyse the anti-CD3/CD28 stimulated coeliac cases and controls dataset, as pathway analysis results and INRICH had indicated it to be a more biologically relevant dataset in terms of CD. WGCNA identified 15 modules (M1-M15) in the dataset (Table 3.9).

The expression levels of each module were then summarised by the first principal component (module eigengene) and correlated with disease status. Additionally, each gene in a module was ranked by its connectivity or module membership in order to identify module 'hub genes' or genes with the greatest connectivity.

Significant correlation was detected between 6 modules and CD patient status. Modules 2, 3, 5 and 8 were positively correlated with disease status, meaning that genes in these modules show greater expression in CD. Conversely, modules 6 and 11 were negatively correlated, meaning they comprise genes showing predominantly lesser expression in CD cases. R^2 and P-values for these correlations are listed in Table 3.9. Of these modules, M3 and M6 were of greatest interest as both were very significantly enriched for differentially expressed genes ($P < 2.2x10^{-16}$, Fishers exact

test) and showed the strongest correlation with disease status. In order to test whether these two modules had more protein-protein interactions than expected by chance, we used the protein interaction tool, DAPPLE as described in section 3.3.8. Both M3 (P = 0.0009) and M6 (P = 0.001) were enriched for direct protein connections (Fig 3.9a and 3.9b), which supports the interpretation that the co-expression of genes we observe in the WGCNA analysis is caused in part by direct functional interactions of these genes.

Table 3.9: WGCNA identified 15 modules in our anti-CD3/CD28 stimulated dataset.

Module	# genes in module	ME trait correlation (r2)	Adjusted p-value	# DE genes	Hub genes
M1	190	-0.013	1	10	IFIH1, SP110, MYD88
M2	284	0.66	0.016	21	CNPY2, MYL6B, PSMD3
M3	675	0.93	3.2x10-8	219	IFNG, TFRC,IL2RA, EVI5
M4	416	-0.42	0.96	1	FUNDC2, RPL41, RPL39
M5	283	0.68	0.011	34	JAK2, KDELC2, BLM
M6	1,163	-0.96	4.8x10-11	203	BACH2, ZFP36L1, FBXO34, TGFBR2
M7	867	-0.4	1	24	IL11RA, AHSA2,ZNF862
M8	271	0.79	3.2x10-4	68	FAM54A, TCF19, PSMC3IP
M9	879	-0.48	0.48	10	ETS1, AKAP11, PAPOLG
M10	712	-0.55	0.144	5	KIAA1109, USP34,ZNF445
M11	204	-0.67	0.014	0	SLC30A7, ZNF845, LZTFL1
M12	147	-0.13	1	0	ADAT2,PRPF39, C1orf27
M13	869	0.55	0.144	20	UBE2L3, C17orf37,TSFM
M14	1,632	0.46	0.64	23	ACTR2, API5,USP14
M15	2,189	0.13	1	50	YDJC, CSK,S1PR4

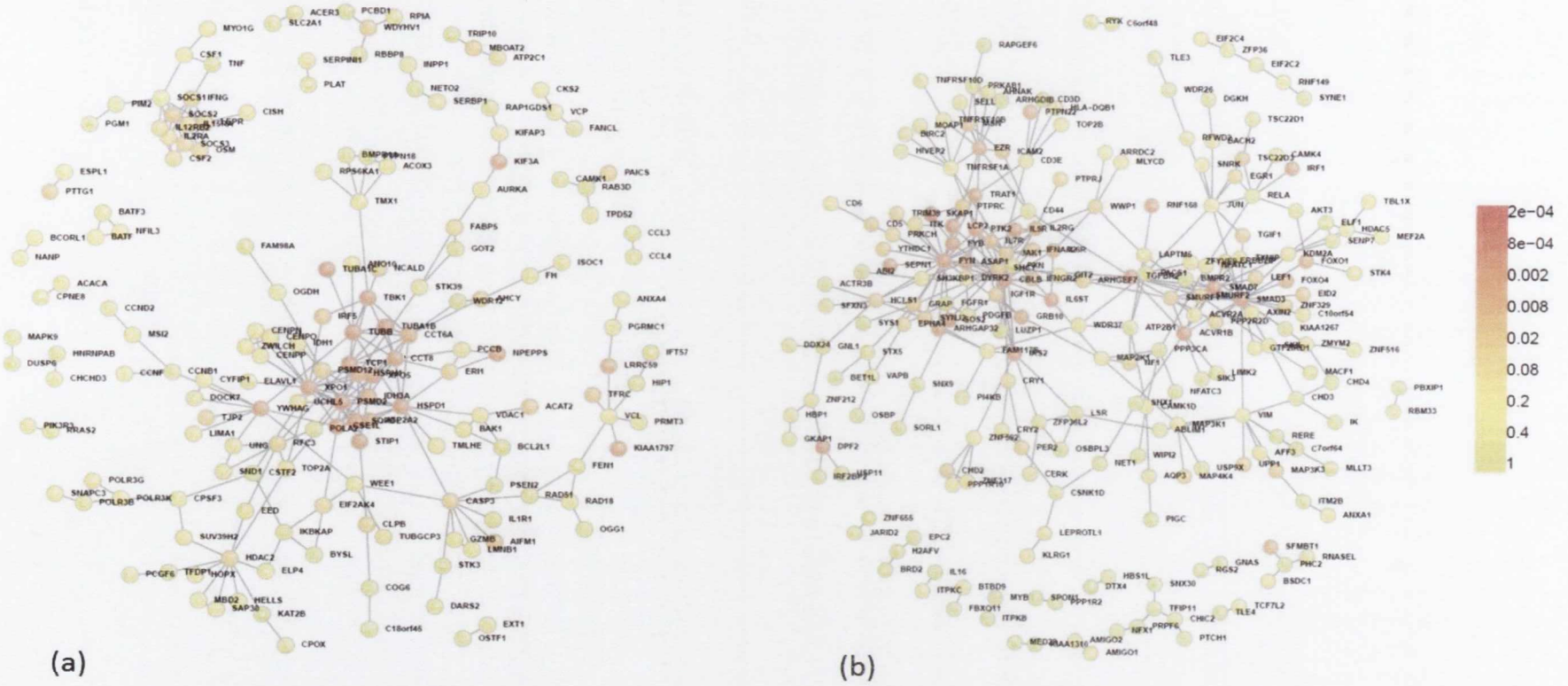


Figure 3.9: Protein interaction networks for modules 3 and 6. Genes in M3 (a) and M6 (b) showed enrichment for protein-protein interactions. Colour code represents the probability that a protein would be as connected to other proteins by chance based on 1000 permutations.

The M3 module was strongly associated with coeliac disease status (r^2=0.93, P_{adj} = $3.2x10^{-8}$) and contained over 200 DE genes. Pathway analysis of this module using MetaCore showed enrichment for the Immune response: T-cell subsets (P_{adj} = $1.8x10^{-4}$), the JAK/STAT Pathway (P_{adj} = $6.3x10^{-3}$). M3 contains a number of genes genetically associated with CD, including *IL18R1, CCR5, PTPRK, CTLA4, RGS1, IL18RAP* and *SOCS1*, which we show to be differentially expressed. DAPPLE reported a list of 56 genes to prioritise in M3 that have more direct protein connections in the module than would be expected by chance. Interestingly one of these genes is *IFNG*, which along with being the most significantly DE gene in M3 is also a hub gene in this module (MM= 0.91, P = $4.15x10^{-9}$). The immunologically related genes *IL12RA* (MM=0.95, P = $4.3x10^{-11}$) and *IL32* (MM=0.95, P = $4.6x10^{-11}$) are also hub genes in M3.

The M6 module showed the greatest association to coeliac disease status (r^2=-0.96, P_{adj} = $4.8x10^{-11}$), was significantly enriched for DE genes (P < $2.2x10^{-16}$, Fishers exact) and also showed enrichment for coeliac associated loci when tested using INRICH (P_{adj} = 0.004). Pathway analysis of this module, using MetaCore showed enrichment for genes involved in TGFβ receptor signaling (P_{adj} = $6.9x10^{-3}$) and T-cell receptor signaling (P_{adj} = $2.2x10^{-2}$). Known or putative coeliac susceptibility genes in M6 include *BACH2, ZFP36L1, SH2B3, PFKFB3, TNFAIP3, ICOSLG* and *FASLG*. Interestingly, the TGFβ receptor 2 *(TGFβR2)* is one of the top hub genes in this module (MM=0.95, P = $6.38x10^{-12}$), as is the coeliac associated gene *BACH2* (MM=0.96, P = $2.8x10^{-12}$) which is required, amongst other functions discussed below, for TGFβ induced FOXP3 expression and thus normal T-reg function (Kim et al., 2014). BACH2 was also found to

participate more often with other proteins in the interaction network built using DAPPLE than expected by chance.

BACH2 is a very interesting candidate gene for involvement in coeliac disease. It is a transcription factor, which has an important role in balancing tolerance and immunity, and appears to exert a suppressive effect on inflammation (Roychoudhuri et al., 2013; Kim et al., 2014; Tsukumo et al., 2013) by controlling T-cell differentiation. We see a significant decrease in *BACH2* expression in both our anti-CD3/CD28 (Log2FC -1.75; P_{adj} = 0.0037), PMA stimulated (Log2FC -1.26; P_{adj} = 0.017), and unstimulated (Log2FC -0.72; P_{adj} = 0.052) coeliac samples, which is concordant with the findings that suggest risk alleles for the disease are associated with *BACH2* down regulation (Dubois et al., 2010).

We investigated if genes previously identified as subject to BACH2 regulation were differentially regulated in our dataset. A recent study in mice by Vahedi *et al* (2015) looked at regions of the genome that act as super enhancers (SE) in $CD4^+$ T-cells. They identified a network of BACH2 regulated genes composed primarily of cytokines and cytokine receptors, many of which are critical for T-cell function. On examining our RNA-seq data, we found our differentially expressed genes were significantly enriched for BACH2 targets ($P < 2.2x10^{-16}$, Fishers exact test). Ninety-eight genes regulated by BACH2 were differentially expressed between coeliacs and controls in our dataset (These genes are listed in appendix Table C). Furthermore, expression patterns of these genes were correlated with a decrease in BACH2 expression in that 23/32 (72%)

genes that are apparently induced by BACH2 (WT v KO transcriptome data) are down regulated in our coeliacs and 60/66 (91%) genes that are apparently repressed by BACH2 are up regulated.

3.4.9 Gene Set Enrichment Analysis

To further assess this finding, we used Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) to examine if BACH2 regulated genes in the Vahedi *et al* (2015) data set were enriched among our DE genes across all three conditions (anti-CD3/CD28, PMA and UNS) (Appendix Table D). We also analysed if the changes in gene expression observed by Roychoudhuri *et al* (2013) in mouse CD4SP thymocytes following BACH2 knockout were significantly enriched in these data. In the anti-CD3/CD28 expression data, we observed a very extreme distribution of BACH2 regulated genes as identified by Vahedi *et al*, with all of these genes found in the upper and lower ends of the spectrum of DE genes (Fig 3.10 and Fig 3.11) (normalised enrichment score [NES] = 1.716; FDR q-value = 0.00127). The DE geneset of Roychoudhuri *et al* was not significant enriched among our DE genes (NES = -1.226; FDR q-value = 0.224). There was no evidence of enrichment of either geneset among our UNS DE genes (Vahedi FDR q-value = 0.317, NES = 1.249; Roychoudhuri FDR q-value = 0.292, NES = 1.098) or our PMA DE data set (Vahedi FDR q-value = 0.145, NES = 1.305; Roychoudhuri FDR q-value = 0.275, NES = 1.316) (Fig 3.12).

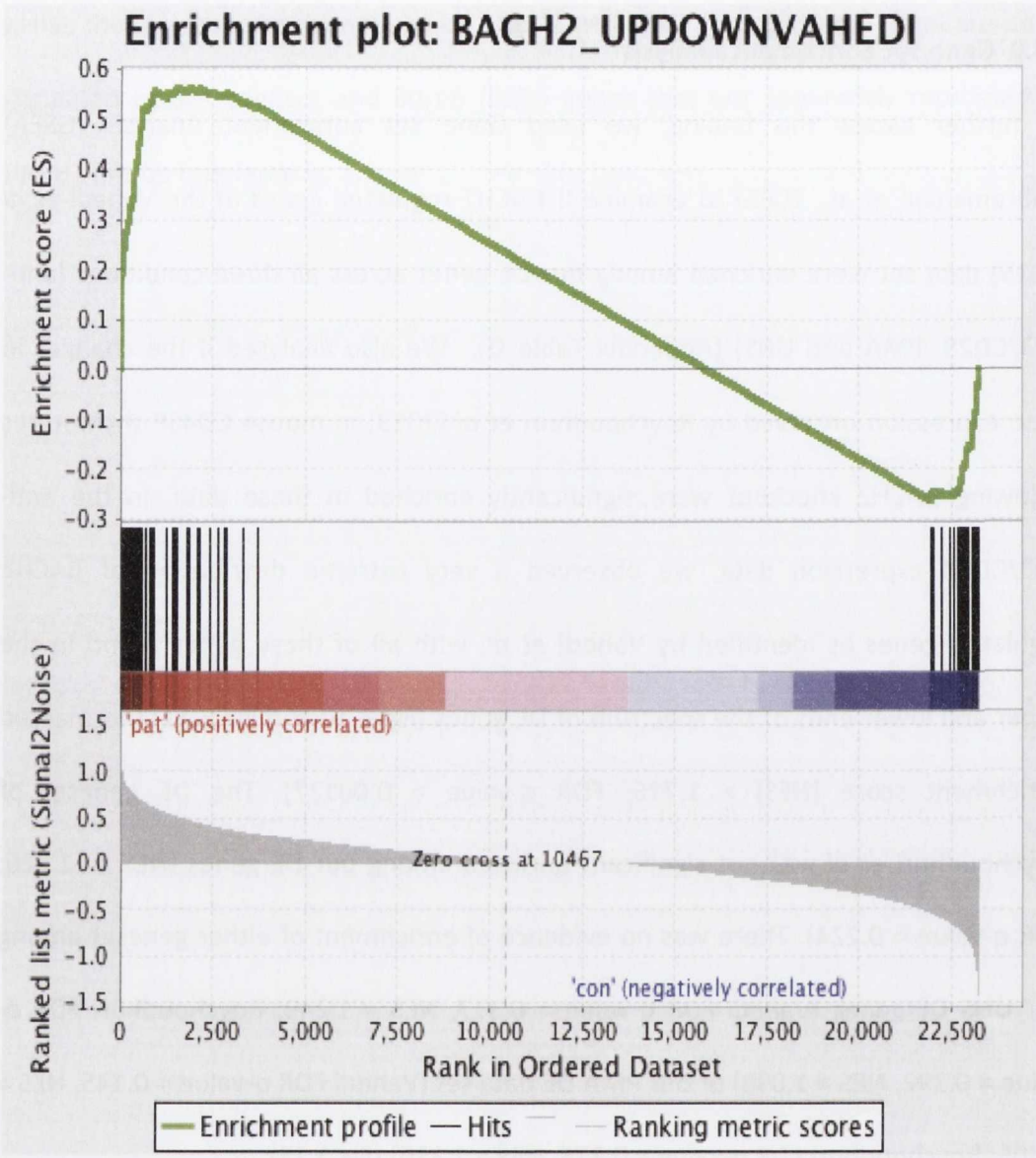


Figure 3.10: Enrichment plot of the distribution of BACH2 regulated genes from our dataset when compared to the Vahedi *et al* dataset.

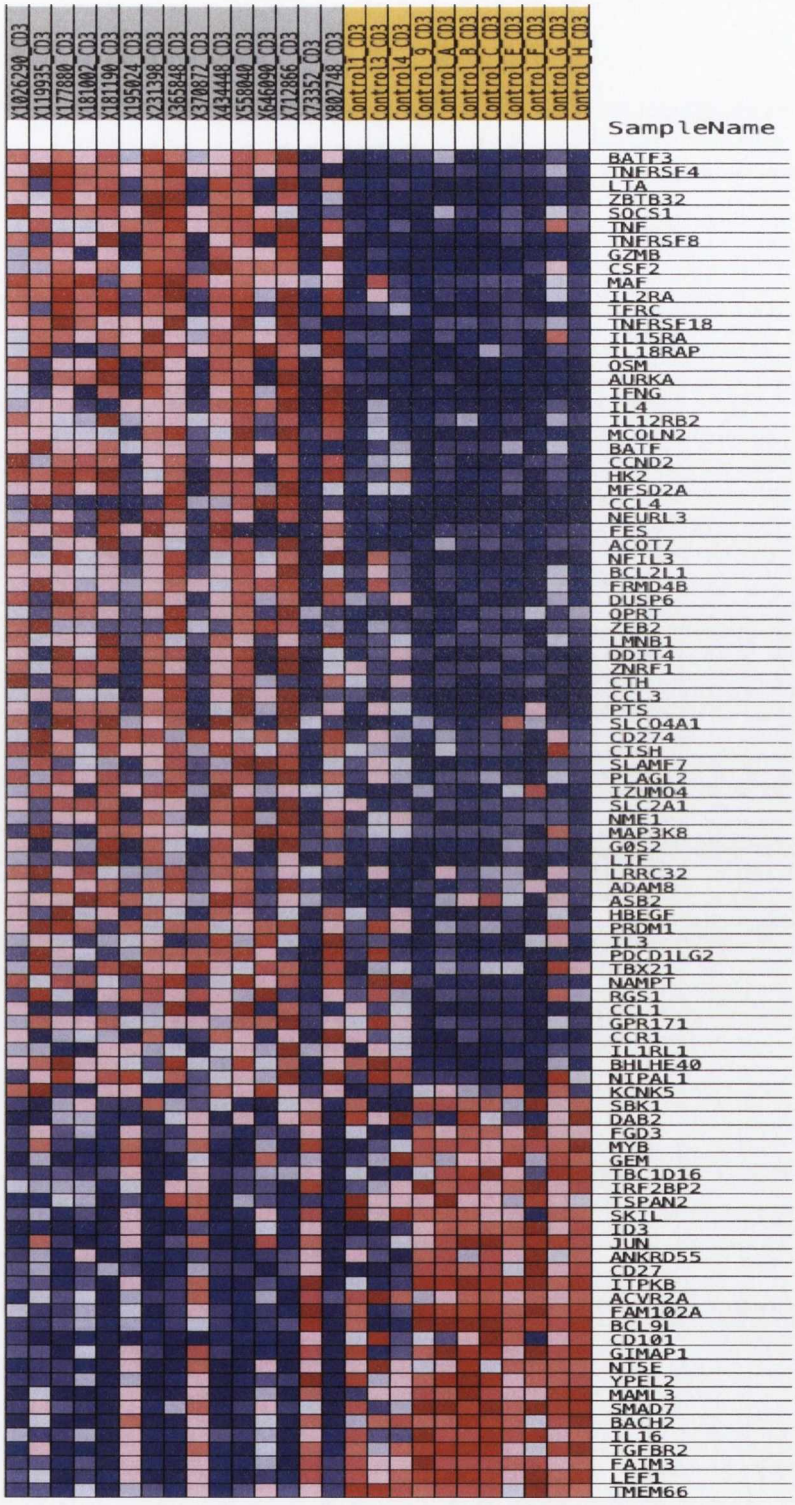


Figure 3.11: A heat map showing the expression values of genes in the dataset sorted by correlation with phenotype. The expression values are represented by colours, red=highly expressed, blue=lowely expressed.

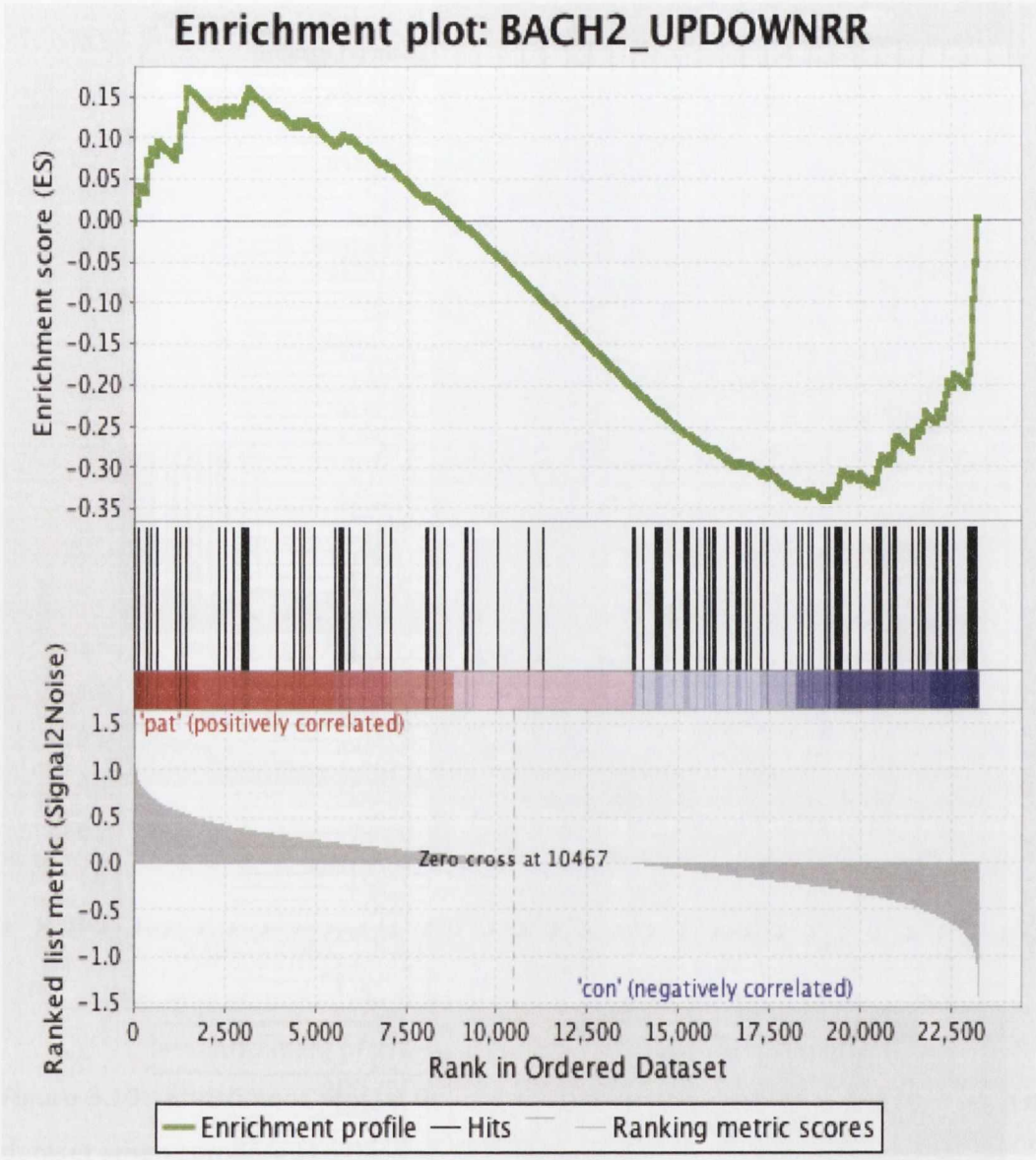


Figure 3.12: Enrichment plot of the distribution of BACH2 regulated genes from our dataset when compared to the Roychoudhuri *et al* dataset.

3.5 Discussion

With a known environmental trigger, gluten and the HLA-DQ genes well characterised as genetic risk factors, coeliac disease is one of the best understood of all complex diseases. While the general mechanism, of gluten provoking innate and adaptive immune responses orchestrated by T-cells, is well established, much of the complex network of processes that ensues remains to be elucidated - for example, how differences in the regulation of disease associated genes affects molecular circuitry in individual cells.

Recent GWAS and fine mapping studies have identified over 40 CD susceptibility loci. However, it is not always clear which is the true causal gene or genetic variant at a given locus. The possibility also exists that more than one gene at a locus may be causally associated. Associated variants are predominantly found in non-coding DNA and are believed to contribute to disease by modifying the expression of nearby or even distant genes often in a cell specific manner; altered gene expression affecting particular molecular pathways is therefore likely to be an important mechanism of disease. In our data, pathways significantly enriched with differentially expressed genes include T-cell development, T & B-cell co-stimulation, cytokines, chemokines, NK cell activation and IFNγ production among others.

Clearly, CD is caused by the interaction of several different cell types, including antigen presenting cells, B-cells, $CD4^+$ and $CD8^+$ T-cells, NK, and epithelial cells. Genomic approaches do not discriminate between cell types and it is unlikely that all genes or

SNPs identified in these studies are functional in every cell type. To date, transcriptomic studies in CD have generally used mixed populations of cells (for example whole biopsy material) even though it is certain that each has different transcriptional profiles, potentially relevant to disease activity. Thus, we aimed to profile the transcriptome specifically of $CD4^+$ T-cells, which are known to be centrally important in CD pathogenesis (Meresse et al., 2012; Sollid et al., 2012). We therefore characterised genome-wide differences in transcription between individuals with coeliac disease and unaffected controls in resting cells and cells activated using two different stimulations (anti-CD3/CD28 and PMA). This is, to our knowledge, the first case control study to apply mRNA-sequencing specifically to T-cells from coeliac disease patients. We compared the transcriptomes of unstimulated and stimulated T-cells, since the latter should more closely model the scenario in the coeliac lesion, which is the site of vigorous immune activity. Also, the use of three independent treatments on each sample allows for comparisons of the effect of each treatment and also represents an internal replication under different conditions with the results proving highly concordant.

The data gave us the opportunity to:

(i) Compare gene expression between cases and controls.

(ii) Specifically assess whether genes that have been genetically associated with the disease were being differentially expressed.

(iii) Look for known and novel aspects of pathogenesis in the transcriptome of this specific cellular compartment.

Whether or not the T-cells were stimulated had a major impact on our findings. Stimulation using anti-CD3/CD28 and PMA resulted in clustering of our coeliacs and controls, which was not obvious in unstimulated cells. Clustering patterns for PMA and anti-CD3/CD28 were highly similar. Thus, the transcriptome of peripheral T-cells from the majority of patients is considerably different from that of controls under the conditions employed here, implying that these differences are ultimately genetically determined and relevant in disease pathogenesis. However we also found evidence of a potential subgroup of coeliacs whose mRNA expression (following stimulation) was similar to controls. As discussed in the results section, this could represent a subgroup, which differs to the other CD patients in the study in terms of disease processes, genetic susceptibility or possibly adherence to the gluten free diet. The fact that the same individuals reacted similarly to two different types of stimuli argues that this is not an artefactual finding. Interestingly, this kind of transcriptional signature has been observed in a previous CD study by Anderson and colleagues were they found that approximately one third of their coeliac patient were unresponsive to gliadin stimulation (Anderson et al., 2006).

Anti-CD3/CD28 stimulation is considered more physiologically relevant as it mimics stimulation by antigen-presenting cells; however we observe that it has less extensive effects than PMA, which stimulates cells by activating the PKC and MAP kinase signaling pathways. Using both agents allowed us compare the transcriptome-wide effects of these different stimulations. Genes DE following anti-CD3/CD28 stimulation

were enriched for pathways such as dysregulated cytokine and T-cell receptor signaling, in keeping with what is known of disease mechanisms. Furthermore, this list of genes was shown to be enriched for coeliac associated loci. The PMA and anti-CD3/CD28 data were highly concordant in terms of direction of fold change (r^2=0.71) regardless of significance which acts to substantiate the findings since the samples were replicas treated independently and differently once split. However almost double the number of genes were DE after treatment with PMA with pronounced enrichment of general cell metabolism pathways (Table 3.6). Hence we concluded that the anti-CD3/CD28 data was more representative of the *in vivo* state of immune activation and so focused on this for network analysis.

Although genetic studies of CD have only identified ~70 genes in the disease, we observe large differences in mRNA expression affecting >1000 genes in the anti-CD3/CD28 (and more than 2000 in the PMA) samples. Many genetically linked genes are represented in this group. It is likely that many of the other DE genes are downstream amplification effects caused by the dysregulation of key disease associated genes such as transcription factors or signaling molecules, any perturbation of which would be predicted to have widespread effects on the transcriptome. Most of the latter molecules show only small changes in their levels of expression; however the potential importance of such changes to pathogenesis is demonstrated by IFNγ, which displayed a ~25 fold increase in expression in cases compared to controls. This shows that the documented high expression of IFNγ in the coeliac lesion is probably indirectly genetically determined even though *IFNG* itself is unlinked to CD. More broadly, subtle

differences in expression in other genes may combine to have a more cumulative downstream impact. This can be seen in the enrichment of differentially expressed genes in a number of self-evidently important immune related pathways in our data e.g. the JAK/STAT and TGFβ pathways.

We show that 21 of 70 putative CD susceptibility gene candidates are DE under one or more of the conditions tested here. A further 4 genes show differential exon usage bringing the total to 25. In regions where there is more than one candidate gene, our results distinguish several as the more likely candidate gene in $CD4^+$ T-cells. In region 2q12.1 however, three genes, both of the IL18 receptor genes (*IL18R1* and *IL18RAP*) and the IL1 receptor gene *IL1RL1*, all show greater expression. Since several studies show eQTLs in this region acting on both IL18 receptor genes, it may be that they both represent susceptibility genes for CD. This makes sense from a functional point of view, as IL18 signaling through its receptor (a heteromeric complex of α- and β-chains encoded by the *IL18R1* and *IL18RAP* genes respectively) and subsequent IFNγ activation requires the co-expression of both genes. eQTLs with overlapping effects on *IL18R1* and *IL1RL1* have also been reported (Akhabir et al., 2014) indicating complex regulatory associations with the causal variant affecting the expression of multiple genes, perhaps through an enhancer like mechanism. It is also worth considering that variants in high LD with the most associated SNP in a region such as this may also be functional, ie there may be more than one causal variant. In the chemokine receptor cluster, *CCR1* and *CCR9* are differentially regulated in this study, the latter being of particular interest since it is involved in gut homing.

Our findings of DE are in good agreement with a recent study by Plaza-Izurieta *et al* (2014), who looked specifically at the expression differences of 45 CD associated genes in whole intestinal biopsy tissue from 15 coeliac cases and controls. They showed 15 of the 45 genes to be differentially expressed. Five of these, *SOCS1, CTLA4, ZFP36L1, ICOS*, and *ICOSLG* were also significantly DE in the same direction in our anti-CD3/CD28 stimulated samples (although *ZFP36L1* and *ICOS* did not pass our FC cutoff i.e. >2) and *PLEK* in our PMA stimulated samples, while the remaining 9 genes were not significantly DE under any of our conditions. Other notable DE genes which are either genetically linked or thought to be functionally involved in the lesion and are DE in our study include *IL21* (Log2FC 3.6, P = $1.7x10^{-8}$), *IL23R* (2.5, P = $7.5x10^{-7}$), *IL15R* (1.01, P = $6.2x10^{-6}$).

Common co-morbidities of coeliac disease include type 1 diabetes, rheumatoid arthritis, ulcerative colitis and Crohn's disease. There is considerable overlap between loci for coeliac disease and other immune-related diseases and this is reflected in our data, since our DE genes are enriched for pathways relating to several of these diseases (Table 3.5). Given this sharing of molecular mechanisms, we investigated whether DE genes failing to achieve genome wide significance (GWS) for CD (and thus not recognised as CD susceptibility genes) had been genetically associated with another autoimmune disease at GWS level. This analysis found that ~14% of our anti-CD3/CD28 DE genes had been linked to at least 1 or more other autoimmune disease. Some genes of interest among these include CD226, a cell surface receptor involved in

T-cell adhesion and activation (associated with Crohn's disease, RA, T1D and UC) which is upregulated in our coeliacs and was previously strongly associated with CD, falling however short of GWS ($P = 10^{-7}$) (Trynka et al., 2011). CTSH (cathepsin H) is also upregulated in our coeliacs and is genetically associated with MS and T1D where it has been linked to β-cell function during disease progression (Fløyel et al., 2014). The CTSH SNP rs34843303 is strongly associated with CD ($P = 2.5x10^{-6}$) and is an eQTL, which affects CTSH levels in B-cells and monocytes (Guo et al., 2015). LON Peptidase N-Terminal Domain and Ring Finger 2 (LONRF2) is down regulated in coeliac patients and was previously identified as being associated with CD (Trynka et al., 2009), T1D (Todd et al., 2007) and RA (Stahl et al., 2010). Of novel interest is *ZPBP2* (log2FC 1.06, P = 0.025) which maps to a locus associated with several other immune mediated diseases (T1D, Primary Biliary Cirrhosis, Crohn's Disease, UC, RA and asthma) also encoding candidate genes *ORMDL3*, *GSDMB* and *IKZF3*, none of which are differentially expressed in our data; and *CSF2* (encoding GM-CSF; log2FC 2.45 $P = 7.5x10^{-08}$) which has been associated with CD, UC and RA. CSF2 is similarly overexpressed in a recent RNA-seq analysis of psoriatic skin lesions (Li et al., 2014).

Network analysis using WGCNA identified several co-expression modules associated with coeliac disease. Two of these modules (M3 and M6) were highly enriched ($P < 2.2x10^{-16}$) for DE genes. The most highly correlated module is M6 ($r^2=-0.96$, $P = 4.8x10^{-11}$), highlighting the CD susceptibility gene *BACH2* as a key hub gene. Polymorphisms at *BACH2* have been associated with numerous inflammatory diseases including Crohn's disease (Franke et al., 2010) and T1D (Cooper et al., 2008). We find that *BACH2* mRNA

levels are substantially decreased in coeliacs compared with controls, which is in agreement with eQTL studies in peripheral blood showing CD risk alleles at *BACH2* are associated with decreased expression (Dubois et al., 2010).

BACH2 is emerging as a crucial regulator of immune cell differentiation. It is a transcription factor, which has been shown to bind super enhancers that control immune cell identity (Vahedi et al., 2015). In $CD4^+$ T-cells, it inhibits the expression of genes needed for the development of effector cells (Th1, -2, -17), and promotes the formation of T-reg cells. This immunoregulatory role is essential, and when under expressed, the formation of T-regs is reduced, leading to fatal levels of inflammation in the case of knockout animals (Roychoudhuri et al., 2013). Notably, these animals displayed significant inflammation and immune cell infiltration of the gut. As stated, lesser *BACH2* expression is seen in coeliac patient samples under all conditions (anti-CD3/CD28 Log2FC -1.76, P_{adj} = 0.0037). Our differentially expressed genes were significantly enriched for BACH2 targets ($P < 2.2x10^{-16}$) and genes previously shown to be repressed by BACH2 were up regulated in our individuals with CD. BACH2 deficient cells produce greater amounts of effector cell related transcripts (Roychoudhuri et al., 2013; Kim et al., 2014; Tsukumo et al., 2013) and cytokines representative of Th1, Th2 and Th17 cells. We examined if these cell types show increased expression in our data, as might be expected if BACH2 represses T-effector cell activity or differentiation. However at the level of master cell fate regulators, only TBX21 is DE using our criteria, potentially showing an expansion of Th1 cells in the CD sample. This may be directly attributable to BACH2 since TBX21 shows increased expression in Bach2 knockout mice

(Vahedi et al., 2015). If replicated in the coeliac lesion, this could plausibly result in an abnormal or exaggerated local immune response. While there is no evidence of reduced T-reg populations in CD patients, several reports suggest that they are unable to suppress T-effector cells effectively (Granzotto et al., 2008; Zanzi et al., 2011; Hmida et al., 2012) a finding also seen in T-regs from *Bach2* KO mice (Roychoudhuri et al., 2013). Interestingly, TGFβR2 which is upregulated in *Bach2* KO mice, is also a hub gene in the M6 module and TGFβ signaling has been shown to require BACH2 to maintain T-reg function (Kim et al., 2014). Our pathway analysis (see Figure 4) shows receptors of the extended TGFβ/BMP family to be comprehensively down regulated although not all meet our cutoff for significance. Taken together these observations supports decreased immune suppression via T-reg / T-effector cell homeostasis as an important mechanism in CD pathogenesis.

As mentioned, a large number of our DE genes were found to be regulated by BACH2. To further evaluate this finding, We used GSEA to examine whether BACH2 regulated genes identified by Vahedi *et al* (2015) and Roychoudhuri and colleagues (2013) were enriched among our DE gene datasets. We detected a very extreme distribution of BACH2 regulated genes identified by Vahedi et al in our anti-CD3/CD28 expression data with a non-random distribution of all of these genes found in the upper and lower ends of the spectrum. The DE genes from the Roychoudhuri *et al* dataset were not significantly enriched among our genes. The different associations seen in these two genesets may relate to the fact that the Vahedi data set consists of super enhancer genetic architecture in $CD4^+$ T-cells and thus at the top of the hierarchy of BACH2

regulated genes, while the Roychoudhuri gene set is composed of BACH2 regulated DE genes which are likely amplified by various downstream regulatory circuits in this cell type (mouse CD4SP thymocytes).

M3 is the next most highly correlated module with coeliac disease (r^2=0.93; P = $3.2x10^{-8}$). *IFNG* is a hub gene in M3 and our network analysis showed that a substantial proportion of our DE genes were co-expressed with *IFNG* in $CD4^+$ T-cells, indicating that these changes are diagnostic of CD status and may be causatively associated. Our findings are interesting in the context that Kumar *et al* (2014) recently noted that 30% of plausible genetically associated coeliac genes (15/49) showed evidence of co-expression with *IFNG* and of these, 7/11 genes for which data is available in Fairfax *et al* (2014) were significantly induced by IFNγ in monocytes.

This is significant because we show that *IFNG* is the most significantly upregulated gene in peripheral $CD4^+$ T-cells. IFNγ is very highly expressed in the coeliac lesion where it plays a critical role in disease pathogenesis (Nilsen et al., 1998; Olaussen et al., 2002). It is produced by T-cells and powerfully promotes the Th1 response characteristic of CD. IFNγ signaling is via the JAK/STAT pathway, which is linked to a variety of inflammatory diseases and is strongly represented amongst DE genes enriched in our data (P = $1.66x10^{-6}$). This all goes to illustrate the centrality of IFNγ expression to the network of DE genes despite there being no genetic association of the *IFNG* locus itself and suggests that this is a downstream effect of inherited susceptibility variants which ultimately regulate IFNγ expression.

A plausible cause of this high IFNγ expression is IL18RAP, which is co-expressed with IFNγ in M3, and is DE in our data (log2FC 2, P = $2.05x10^{-6}$). *IL18RAP* (encoding IL18Rβ) has been linked to CD in several genomic studies (Dubois et al., 2010) and combines with IL18Rα (encoded by *IL18R1*) to form the IL18 receptor, which is expressed predominantly on T and NK cells. IL18 is a key IFNγ stimulating cytokine expressed by monocytes, dendritic cells and epithelial cells of the gastrointestinal tract (Dinarello et al., 2013). Thus up regulation of the IL18 receptor could be expected to lead to the increased IFNγ expression typically seen in the CD lesion. IL18R1 (IL18Rα) is also overexpressed, but to a lesser extent (anti-CD3/CD28 log2FC=0.92, P_{adj} = 0.086; PMA log2FC= 1.86, $P_{adj} = 9x10^{-5}$).

In accordance with our study, Plaza-Izurieta *et al* (2011) found that IL18RAP was significantly overexpressed in intestinal biopsies of active CD samples. A similar study by Pascual *et al* also reported increased expression of IL18RAP in duodenal tissue (Pascual et al., 2015). Koskinen *et al* (2009) also showed overexpression of IL18RAP in CD, which they suggest may be in part explained by higher expression levels of a newly identified 35kDa alternative IL18RAP isoform. However we see no evidence of differential exon usage within the IL18RAP transcript itself or in exon usage between cases and controls.

However these observations conflict with those of Dubois *et al* (2010), who reported that the IL18RAP proximal coeliac disease associated eQTL (rs917997), correlates

impressively with IL18RAP downregulation ($P = 7.4x10^{-87}$). This correlation was shown in un-stimulated peripheral blood leukocytes and the possibility exists that it behaves differently in resting compared to stimulated cells. Indeed, Fairfax *et al* (2014) showed that the majority of eQTLs were active only upon cell stimulation while many constitutive eQTLs in resting cells were not functional post stimulation and a similar observation on SNPs associated with autoimmune diseases was made by Lee *et al* in dendritic cells (Lee et al., 2014). Finally, another difference is that we are testing CD derived cells compared to controls, rather than looking for an eQTL effect which this study is insufficiently powered to do.

Interestingly another immune mediated regulatory gene *CTLA4* co-expressed with IFNγ in M3 is also significantly upregulated in our coeliac patients. IFNγ has been shown to induce CTLA4 expression in human PBMCs (Wang et al., 2001). CTLA4 is a critical negative regulatory molecule expressed on T-cells that competes with CD28 to inhibit T-cell activation and maintain central tolerance. CTLA4 polymorphisms have been associated with numerous autoimmune diseases, including systemic lupus erythematosus (SLE) (Liu and Zhang, 2013), rheumatoid arthritis (RA) (Lee et al., 2011) and celiac disease (Torinsson Naluai et al., 2000; Dubois et al., 2010; Song et al., 2013). Many CTLA4 polymorphisms identified in autoimmune disease have demonstrated a lower level of gene expression thus seemingly contributing to prolonged T-cell proliferation and pathogenesis of disease as would be expected. On the other hand, overexpression of the *CTLA4* gene has been shown in $CD4^+$ T-cells from patients with SLE. This overexpression did not subdue the T-cell response in SLE; on the contrary, it

was unable to inhibit T-cell activation pathways or proliferation in these patients thus highlighting the complex effects different polymorphisms of the *CTLA4* gene are reported to have on gene expression and gene function in different diseases . This highlights a potential role for CTLA4 over-expression and dysfunction that could potentially play a role in CD pathogenesis.

Another hub gene in the M3 module is *SOCS1*, which is upregulated in patient samples and has been previously associated with the condition. SOCS1 inhibits IFNγ signalling via the JAK/STAT pathway and is induced by IFNγ in a typical feedback inhibition loop. SOCS1 has been studied in diverse autoimmune diseases with apparently conflicting observations. On the one hand, SOCS1 deficiency was shown to result in continuous production of pro-inflammatory cytokines, mediating an inflammatory immune response (Fujimoto et al., 2004). Conversely however, over-expression of SOCS1 has been implicated in autoimmune diseases since it inhibits anti-inflammatory signalling as well. For example in RA, SOCS1 reduced IL-10 mediated STAT3 activation in $CD4^+$ T-cells causing increased IFNγ expression and a prolonged inflammatory response (Yamana et al., 2004). SOCS1 activity broadly impacts T-cells homeostasis and both pro- and anti-apoptotic functions have been observed. In psoriatic keratinocytes it inhibits cell death by sustaining the anti-apoptotic P13k/AKT pathway in response to IFNγ (Madonna et al., 2012). This type of response may contribute to the prolonged inflammatory response seen in $CD4^+$ T-cells in CD. This suggests the paradoxical possibility that rather than dampening the immune response, SOCS1 may promote pro-inflammatory autoimmune disease by reducing anti-inflammatory cytokine

signaling or preventing apoptosis. These observations indicate that caution should be exercised in attempting to interpret the implications of altered gene expression, which in some scenarios at least may have complex effects. Furthermore, while we highlight a number of interlinked hub genes of the M3 and M6 modules, many genes of these modules are dysregulated and may have an impact on disease.

3.6 Conclusion

In conclusion, using an unbiased approach, we characterised over 70 transcriptomes from coeliac patients and controls in both resting and stimulated cells. We identified large transcriptional changes that significantly mirror what has emerged from genetic studies of the disease. Differential expression is strongly associated with genes of immunological function particularly in the anti-CD3/CD28 stimulated group, with PMA stimulated cells showing a larger signature representing metabolic and proliferative changes. Genes which have been associated with CD in genome scans are enriched in our DE genes and our data provides corroborating evidence for several genes which fail to meet genome wide significance levels. Genes shown to be DE in this study are strongly associated with several co-expression networks, the most strongly of which feature *IFNG* and *BACH2*. *BACH2* is genetically associated with CD and down regulated under all conditions tested. We identify a strong association of CD with a network of BACH2 regulated genes, supporting emerging evidence of an important role of BACH2 in the regulation of T-cell differentiation and prevention of autoimmune disease.

Chapter IV

Gene expression analysis in an extended case-control population

4.1 Introduction

The discovery of novel genetic variants together with newly identified disease genes has increased our understanding of CD pathogenesis enormously in the past decade. As previously discussed, these variants appear to exert regulatory effects on disease susceptibility genes. However the precise manner by which these associated variants regulate gene expression and thus influence complex traits is still not well understood. One problem that makes this difficult is that many loci responsible for complex diseases harbour multiple variants, thus the identification of the exact causal SNP remains elusive. Many genotyping arrays used for previous GWA studies were low-density SNP arrays, therefore making the identification of the causal SNP responsible for disease pathogenesis very difficult. This complexity is added to by the fact that an estimated 58% of the loci linked to complex diseases contain more than one gene and a further 16% contain no gene (Hindorff et al., 2015). Refinement of these susceptibility loci is complex; therefore much work and a variety of methodological approaches are required to identify the causal variants that lead to disease pathogenesis.

To date, 43 loci comprising over 70 genes have been significantly associated with CD; however, the majority of these GWA SNPs do not fall within coding regions of the genome. In fact, more than 90% of these SNPs are located outside the protein-coding sections of genes; approximately 43% are located in intergenic regions and 45% are located in introns, suggesting they may have a regulatory role (Freedman et al., 2011). Only rarely does a SNP have a clear functional effect in complex disease pathogenesis;

for example, where it is present in the coding sequence of a gene and modifies protein function directly such as with *NOD2* mutations in Inflammatory Bowel Disease (IBD) (Hugot et al., 2001). Variants lying outside protein coding sequences are believed to exert their affect on gene regulation by altering gene regulatory elements such as the promoter/enhancer sequences, exon splicing junctions or other regulatory sequences like micro RNA binding sites. Such alterations may change the transcriptional or translational efficiency of the gene thereby affecting gene expression. Any variants capable of altering gene expression are known as expression quantitative trait loci (eQTL). They provide insight into the possible biological basis for a disease phenotype and can help to identify gene networks that play a substantial role in disease pathogenesis.

One way to determine if disease associated variants are regulatory is to investigate if they correlate with altered expression of nearby protein coding genes (so-called *cis*-eQTLs). eQTL analysis may also facilitate the prioritisation of genes from regions of the genome containing more than one gene. A small number of eQTL studies have been carried out in CD to date. The first eQTL analysis in CD was carried out by Dubois and colleagues (2010) in 1,469 whole blood samples using a microarray platform. Their results showed that 38% of CD SNPs were *cis*-eQTLs, a result significantly higher than would be expected by chance. This observation of enrichment of *cis*-eQTLs suggests that these variants may in fact be influencing CD susceptibility through a mechanism of altered gene expression.

However, while this large study improved our understanding of the functional role of CD associated SNPs it was performed in whole peripheral blood samples, which are composed of heterogenous cell populations. The composition of the aggregate population may thus vary from one sample to the next and furthermore the activation status of cells within these samples is unknown but they would be presumed to be quiescent in the majority of cases. Since it is assumed that the transcriptome of each of these cell types are different and therefore which SNPs function as eQTLs in various cells also differs, data from aggregate peripheral blood cell populations is complex and the higher levels of transcriptome background noise could be predicted to make the detection of true signals more challenging.

Therefore, not surprisingly multiple studies have since been carried out focussing on cell and tissue type specificity for the detection of *cis*-eQTLs. A study performed by Fu *et al* (2012) highlights the importance of this approach, as approximately one-third of the *cis*-eQTLs they identified acted in a tissue specific manner. Amundsen *et al* (2014) performed an eQTL analysis in thymic tissue exploring the regulatory potential of previously identified CD-associated SNPs. Their results found that 54% of the SNPs analysed showed a *cis*-eQTL effect and 11 SNPs were potentially thymus-specific *cis*-eQTLs.

cis-eQTLs are not only tissue/cell type dependent but also stimulus dependent. Fairfax *et al* (2014) found that more than 50% of the *cis*-eQTLs they identified could only be observed after IFNγ stimulation, with a high proportion of constitutively active eQTLs

being shut down upon stimulation. Plaza-Izurieta *et al* (2014) performed an eQTL analysis using intestinal biopsies from individuals with CD at both time of diagnosis and after 2 years on a GFD and compared to non-coeliac controls. They aimed to determine whether disease associated variants had any influence on 45 selected candidate coeliac susceptibility genes. They identified both *cis*- and *trans*-eQTLs activated upon gliadin stimulation and generalised inflammation present in the coeliac biopsies at the time of diagnosis. These observations further highlight the importance of performing studies in the cellular and environmental context of interest.

Therefore inspired by the results of our RNA-seq study and the identification of multiple associated genetic variants from our Immunochip study we aimed to perform eQTL analysis to correlate differentially expressed genes with associated SNPs. As our initial study of 15 cases and 11 controls was insufficient to carry out this analysis, we wanted to expand our study using a larger case-control $CD4^+$ T-cell sample, again stimulated with anti-CD3/CD28, as we have previously found this to be the most disease relevant condition.

We initially aimed to choose a select number of candidate SNPs and genes on which to carry out this eQTL study. We wanted to begin with SNPs previously associated with CD, especially if there was evidence of a nearby gene that was showing differential expression in our RNA-seq data. We identified 11 out of the 70 previously established non-HLA protein coding genes that were significantly differentially expressed between coeliacs and controls in the anti-CD3/CD28 stimulated samples. We also wanted to

focus on high priority genes e.g. *IL18RAP* ($P = 7.35 \times 10^{-87}$) and *BACH2* ($P = 1.92 \times 10^{-3}$) with previously demonstrated eQTLs in CD (Dubois et al., 2010) and were also differentially expressed in our RNA-seq study.

As mentioned, our initial plan was to select a number of genes to examine via qPCR for the replication phase of our study. However, whilst considering technologies to carry out this part of the study, we included microarray technologies, which have been the standard for whole genome expression studies in the past (Diosdado et al., 2004; Bracken et al., 2008). In 2013, Affymetrix developed a new version of their whole genome array, called the Affymetrix GeneChip® Human Transcriptome Array (HTA) 2.0. This HTA 2.0 array represented an advance in design over previous microarrays in that it contained a much-increased density of probes at over 6 million, covering both coding and non-coding transcripts. Affymetrix also claimed it generated accurate gene expression data regardless of transcript abundance. It is the highest resolution microarray currently available for gene expression profiling of all transcript isoforms in the genome. As it represented an unbiased approach, it eliminated the need for candidate gene selection, which was considered an advantage over qPCR. It was also highly cost efficient compared to all other methods assessed and technically achievable given in house experience with this platform. Thus using the Affymetrix GeneChip® Human Transcriptome Array (HTA) 2.0 seemed to be an obvious choice of methodology to use for our study.

Using such an array allowed us to modify our objectives at this point to follow on from our RNA-seq study and profile the entire transcriptome in a significantly larger sample

size of 60 cases and 20 controls. Additionally, carrying out a global gene profile on a larger group of subjects could potentially allow us to confirm and further characterise the distinct subgroup of coeliac patients which we identified in our RNA-seq study (section 3.4.3) and potentially identify any other coeliac subgroups present in a much larger sample cohort. Having the opportunity to gather gene expression information on a large number of coeliac patients using microarray technology would also give us the opportunity to carry out the eQTL study on the entire genome as opposed to a number of candidate genes as we had initially planned.

Furthermore, this array provided us the opportunity to examine microRNAs and long non-coding RNAs (lnc-RNAs) as both have been implicated in the pathogenesis of various diseases, including CD (Vaira et al., 2014; Hrdlickova et al., 2014), which was not possible to carry out extensively using our RNA-seq experimental set up.

It is estimated that >92% of human multi-exon genes undergo alternative splicing (Wang et al., 2008) which unsurprisingly due to its critical role in determining eukaryotic gene expression, has been associated with several disease phenotypes including many autoimmune diseases e.g. multiple sclerosis (MS) (Gregory et al., 2007). Therefore it was desirable to study alternative splicing in greater detail and the HTA 2.0 platform was claimed to detect all transcript isoforms by extensively measuring expression changes at the exon, sub-exon and junctional level. We had found some evidence of alternative splicing in the RNA-seq study, thus having access

to the entire genome in a larger population would give us the opportunity to more comprehensively characterise splicing such events.

Finally, since the samples were being newly collected for this phase of the gene expression study in order to perform the eQTL study as intended, we would need to genotype previously untested individuals using the Immunochip.

4.2 Aims and Objectives

1. Inspired by our RNA-seq results we wanted to expand this study into a larger sample size of 60 coeliac cases and 20 controls using the Affymetrix GeneChip® Human Transcriptome Array (HTA) 2.0.

2. To identify differentially expressed exons using the HTA 2.0 array and to correlate these results with differentially expressed exons identified using DEXSeq from our RNA study.

3. To identify differentially expressed non-coding RNAs (micro-RNAs and lnc-RNAs) between our coeliac cases and control samples.

4. To genotype newly acquired samples using the Immunochip required for eQTL analysis.

5: To combine both our gene expression data with genotypic data to carry out and eQTL study in our population cohort.

4.3 Materials and Methods

4.3.1 Collection of case blood samples

Written and informed consent was obtained from 60 coeliac patients attending routine coeliac follow up clinics in St. James's hospital, with SJH/AMNCH Research Ethics Committee approval. Twenty-five of these individuals had previously been genotyped on the Immunochip as described in section 2.3.3. All individuals are Caucasian and have Irish ancestry. None of these individuals were taking immunosuppressive drugs to the best of our knowledge; patients charts were checked for this information and patients were also asked directly before collection of the blood sample. 30mls of blood was collected from each coeliac patient. 24mls was used for $CD4^+$ T-cell isolation and subsequent RNA extraction. 6mls was used for DNA isolation and genotyping of any new coeliac patient using the Immunochip.

4.3.2 Collection of control blood samples

30mls of blood was taken from 20 healthy volunteers, with SJH/AMNCH Research Ethics Committee approval. 23mls was used for $CD4^+$ T-cell isolation and subsequent RNA extraction. 6mls was used for a tTG blood test to rule out coeliac disease in our control cohort. 1ml was used for DNA isolation and genotyping using the Immunochip.

4.3.3 DNA sample preparation

Genomic DNA for all cases and controls was isolated by the TCD biobank, using the Gentra nucleic acid extraction system (QIAGEN, Dusseldorf, Germany). All DNA stock samples were standardised to a concentration of 100ng/ul.

4.3.4 Preparing the 96 well plates for Immunochip analysis

In order to analyse eQTLs, we needed to fully genotype our cases and controls. We prioritised collection of patients previously genotyped using the Immunochip, of which 25 cases were available. Thus the remaining cases and controls needed to be genotyped using the Immunochip. Prior to loading each Immunochip, each DNA sample was added to an individual well of a 96 well plate and diluted to a concentration of 50ng/ul. In total 55 individuals, 35 cases and 20 control samples were genotyped using the Immunochip platform.

4.3.5 Genotyping

The samples were genotyped using the Immunochip according to illumina's protocols at the UCD Academic Centre on Rare Diseases (ACoRD) by Dr. Judith Conroy. NCBI build 36 (hg18) was used.

4.3.6 Stimulation of $CD4^+$ T-cells

$CD4^+$ T-cells were isolated and stimulated with 10ul anti-CD3 antibody (1mg/ml) and 1ul anti-CD28 antibody (1mg/ml) for 24 hours at 37°C as described in section 3.3.1.

4.3.7 RNA extraction

RNA was extracted as previously described in section 3.3.2.

4.3.8 Checking RNA quality and concentration

The RNA quality was checked using an Agilent 2100 bioanalyser. All samples used for gene expression studies had a quantity of >200ng and an RNA Integrity Number (RIN) of 8.

4.3.9 Microarray analysis

4.3.9.1 Probe/Gene Array

200ng of total RNA from 60 coeliac cases and 20 controls was processed for each array using the GeneChip® WT PLUS Reagent Kit (Affymetrix, Santa Clara, CA) following the manufacturers instructions. This is a 3-day process involving cDNA synthesis, cRNA synthesis, fragmenting, labelling, hybridisation, washing, staining and scanning of the arrays.

Here follows an abbreviated description of the protocol:

1: First-strand cDNA synthesis

- The First-Strand master mix (4ul First-Strand buffer + 1ul First-Strand enzyme) was prepared on ice for all RNA samples and 5ul was transferred to individual tubes. 5ul of total RNA was added to each tube which were incubated for 1hr at 25°C, then for 1hr at 42°C, then for at least 2min at 4°C.

2: Second-Strand cDNA synthesis

- 20ul of the Second-Strand master mix (18ul Second-Strand buffer + 2ul Second-Strand enzyme prepared on ice for all RNA samples) was transferred to each first-strand cDNA sample. These were incubated for 1hr at 16°C, then for 10min at 65°C, then for at least 2min at 4°C.

3: cRNA synthesis by *in vitro* Transcription

In this step the complimentary RNA was synthesized and amplified by *in vitro* transcription (IVT) of the second-stranded cDNA template using T7 RNA polymerase.

- 30ul of the IVT master mix (24ul IVT buffer + 6ul IVT enzyme) was transferred to each second-strand cDNA sample. Samples were incubated for 16hrs at 40°C, then at 4°C for at least 2min.

4: cRNA purification

In this step, enzymes, salts, inorganic phosphates, and unincorporated nucleotides are removed to prepare the cRNA for 2nd-cycle single-stranded cDNA synthesis.

- 100ul of purification beads were added to each cRNA sample, mixed and transferred to the bottom of a U-bottom plate. Incubation for 10min allowed the purification beads to bind to the cRNA samples.
- The beads were then captured by incubating the plate on a magnetic stand for ~5min during which time they form pellets against the magnets in the stand. The supernatant was then removed and the beads washed.
- The purification beads were then washed in 200ul 80% ethanol for 30 seconds. The ethanol was discarded. This step was repeated twice after which the plate was air-dried on the stand for 5min.

- To elute the cRNA, the plate was removed from the magnetic stand and 27ul nuclease-free water (65°C) was added to each well (1min). The beads were recaptured following 5min incubation on the magnetic stand.
- The eluted cRNA (supernatant) was transferred to a nuclease-free tube.

5: 2nd-cycle single-stranded cDNA synthesis

- 15ug of cRNA was prepared and added to an eppendorf tube alongside 4ul of 2^{nd}-cycle primers. The tube was incubated for 5min at 70°C, then 5min at 25°C and 2min at 4°C.
- The 2^{nd}-cycle-ss-cDNA master mix (12ul)(8ul 2^{nd}-cycle buffer + 4ul 2^{nd}-cycle enzyme) was prepared on ice for all RNA samples and transferred to each cRNA/2^{nd}-cycle primers sample. Incubation was for 10min at 25°C, then 90min at 42°C, then 10min at 70°C, then for at least 2min at 40°C.

6: RNA hydrolyses using RNase H

In this step, RNase H hydrolyses the cRNA template leaving ss-cDNA

- 4ul of RNase H is added to each 2^{nd}-cycle-ss-cDNA sample following which the samples were incubated for 45min at 37°C, then 5min at 95°C, then for at least 2min at 40°C, following which nuclease-free water was added to each hydrolysed 2^{nd}-cycle-ss-cDNA sample.

7: 2^{nd}-cycle-ss-cDNA purification

- 100ul of purification beads were added to each 2^{nd}-cycle-ss-cDNA sample, mixed and transferred to the bottom of a U-bottom plate.

- 150ul of 100% ethanol was added to each ss-cDNA/beads sample and the plate incubated for 20min at room temperature to allow the ss-cDNA in the sample to bind the purification beads.
- The plate was moved to the magnetic stand for ~5min to capture the purification beads. The supernatant was discarded.
- The beads were washed (200ul of 80% ethanol added to each well, incubated for 30sec and discarded). This wash step was repeated for a total of 3 washes. The plate was air dried for ~5min on the magnetic stand.
- The plate was removed from the magnetic stand and 30ul of preheated (65°C) nuclease-free water was added to each well and incubated for 1min.

  The plate was moved to the magnetic stand for ~5min to capture the purification beads and the supernatant containing the eluted ss-cDNA was transferred to individual nuclease-free tubes.

8: Fragmentation and labelling of ss-cDNA

5.5ug of ss-cDNA was prepared for each sample, to which 16.8ul of the fragmentation master mix was added (10ul nuclease-free water + 4.8ul 10X cDNA fragmentation buffer + 1ul 10U/ul UDG + 1ul 1,000U/ul APE for each sample). These were then incubated for 1hr at 37°C, then for 2min at 93°C and then for at least 2min at 4°C.

- 45ul of the fragmented ss-cDNA sample was then added to individual tubes.
- 15ul of the labelling master mix (12ul 5X TdT buffer + 1ul 5mM DNA labelling reagent + 2ul 30U/ul TdT) was added to each fragmented ss-cDNA sample. The

samples were incubated for 1hr at 37°C, then for 10min at 70°C and then for at least 2min at 4°C.

9: Array hybridisation

A "rotisserie" oven is used to perform the hybridisation of 16 arrays at one time at 45°C and 60RPM

- 160ul hybridisation master mix (3.7ul (3nM) control oligo B2 + 11ul 20X hybridisation controls + 110ul 2X hybridisation mix + 15.4ul DMSO +19.9ul nuclease-free water for all RNA samples at room temperature) was added to 5.2ug (~60ul) of fragmented and labeled ss-cDNA for a total of 220ul. The samples were incubated for 5min at 99°C, then 5min at 45°C.
- The samples were then injected onto the arrays and hybridized

4.3.9.2 Fluidics Station

Washing and staining of the arrays was carried out on the GeneChip® Fluidics Station 450. It is automated and contains four modules, each holding one array.

4.3.9.3 GeneChip Scanner

The arrays were scanned using the GeneChip® Scanner 3000 7G.

4.3.9.4 Computer Workstation

The GeneChip Command Console software controls the fluidics station and the scanner using pre-programmed hybridisation, wash, and stain protocols for the probe assay.

Washing, staining and scanning of the arrays was carried out by Dr. Cathy Spillane at The Coombe Women and Infants University Hospital, Dublin.

4.3.10 Preprocessing

Following completion of the laboratory part of the microarray experiment, the next step in the analysis is called preprocessing. Each gene on the HTA 2.0 array has on average 109 probes per gene and therefore 109 probe intensities assigned to each gene. Therefore each gene is represented by multiple raw intensity values and the goal is to obtain a single gene expression value (intensity value) for every gene under investigation. Preprocessing is divided into 3 steps, background correction, normalisation and summarisation.

There are two software packages supplied by Affymetrix that are used to do the analysis of the HTA 2.0 array:

1. The Expression Console (EC) Software, which generates the probe set signals and allows quality control analysis of the array data.
2. The Transcriptome Analysis Console (TAC) Software, which allows interrogation of the array data for evidence of differential expression and alternative splicing of RNA-sequences.

4.3.10.1 Affymetrix Expression Console Software (EC)

We used EC software to create summarised expression values i.e. the normalised values known as CHP files for each RAW data feature intensity CEL file and to provide the quality control metrics for evaluating the success of the hybridisations. EC utilizes several methods in order to create these normalised CHP files, MAS 5.0, RMA, or PLIER algorithms. For the purpose of our study we used the robust multi-array average (RMA) method implemented by EC.

4.3.10.2 Robust multi-array average (RMA)

RMA is an algorithm designed to measure true mRNA expression by summarising probe intensities for each probe set. Briefly, the raw intensity values are background corrected, log2 transformed and then quantile normalised. Next a linear model is fit to the normalised data to obtain an expression measure for each probe set on each array in a process known as summarisation.

4.3.10.2.1 Background correction

Every gene on an array is represented by a series of oligonucleotide probes making up what is known as a probe set. Each probe set is made up of probe pairs consisting of a perfect match (PM) and mismatch (MM) oligonucleotide probes. The PM probe sequence is 100% complimentary to the specific gene and measures the expression of the gene. The MM probe sequence differs from the PM probe by a single base substitution at the centre base position and measures the amount of non-specific

binding. The signals from each of the PM probes measures the abundance of specific RNA-sequences but can also be affected by non-specific binding and auto fluorescence of the array surface known as "background". Background correction methods estimate the background levels (noise) and remove them from the data.

Unlike MAS 5.0, RMA's background correction method only considers the PM probes and ignores the MM probes. RMA assumes that the observed PM probe signal (O) is comprised of the true signal (S) and the background/noise component (N), O = N + S and the background corrected values (E) are E = (S/O). This calculation is performed for all arrays seperately.

4.3.10.2.2 Normalisation of the data

Normalisation of the data is essential because when dealing with multiple arrays, variation can be caused, for example, due to the sample prep, the manufacturing of the arrays, the hybridisation efficiency of the probes for their mRNA targets, giving rise to variable and some times low quality, measurements. This variation will lead to scaling differences between the overall fluorescence intensity levels on each array and can affect our interpretation of true biological variation between the samples. Therefore when dealing with multiple arrays it is paramount that the data from each array be comparable before it can be analysed. In order to establish one true RNA expression value for each gene on the array, it is necessary to adjust the data for this variation in a process known as normalisation. Affymetrix have tried to insure that the

intensities are scaled so that each array has the same average value and to achieve this they use RMA. The RMA procedure uses the *quantile normalisation* algorithm, which ensures that the distribution of probe intensites for each array in a set of arrays is the same. In the *quantile normalisation* method, the highest background corrected and log transformed PM intensity on each GeneChip is determined and these values are averaged. The highest individual values on each chip are then replaced by this average value. This process is repeated with the second highest background corrected and log transformed PM intensities on each GeneChip and so on.

4.3.10.2.3 Summarisation

The intensities for each probe set must be summarised in order to establish a single expression value for every gene on the chip. Tukeys median polish is a standard algorithm used in RMA analysis and is used to obtain median values for normalised probe set data for each GeneChip. These median values serve as the log-scale expression measures associated with the particular probe set and are reported as the RMA measures of expression for each probe set.

4.3.11 Quality Control

Once the CHP files have been created they are ready for QC analysis within the Expression Console software. Quality assessment of the data is important for all types of microarrays as it ensures the results at the end of analysis are meaningful.

4.3.11.1 Report Controls

Affymetrix have established their own quality control processes for gene expression analysis to determine the quality of the assay. These include hybridisation controls and labelling controls.

4.3.11.1.1 Hybridisation controls

Spike controls, *BioB, BioC, BioD* and *Cre* are eukaryotic hybridisation controls, which are included in the hybridisation cocktail added to the samples during the experimental stage. *BioB, BioC* and *BioD* are genes from the biotin synthesis pathway of E.coli, and *Cre* is a recombinase gene from P1 bacteriophage. These genes are not expected to cross-hybrid with non-bacterial and non-viral samples. They are used to evaluate the hybridisation efficiency of each sample in the array. *BioB, BioC, BioD* and *Cre* should display increasing signal values correlating with their increasing relative concentrations.

The results we obtained are characteristic of what one would expect from the spike-in controls (Fig 4.1). Despite what looks like a sharp decline in the control intensity levels

on certain arrays, each array actually shows a high level of uniformity as the range of intensity values on the y-axis are essentially very narrow across the range. For example, the intensity values for *Bio B* (red) only span a narrow range from a minimum value of 7.07 to a maximum value of 10.40, which is considered typical of the spike-in controls.

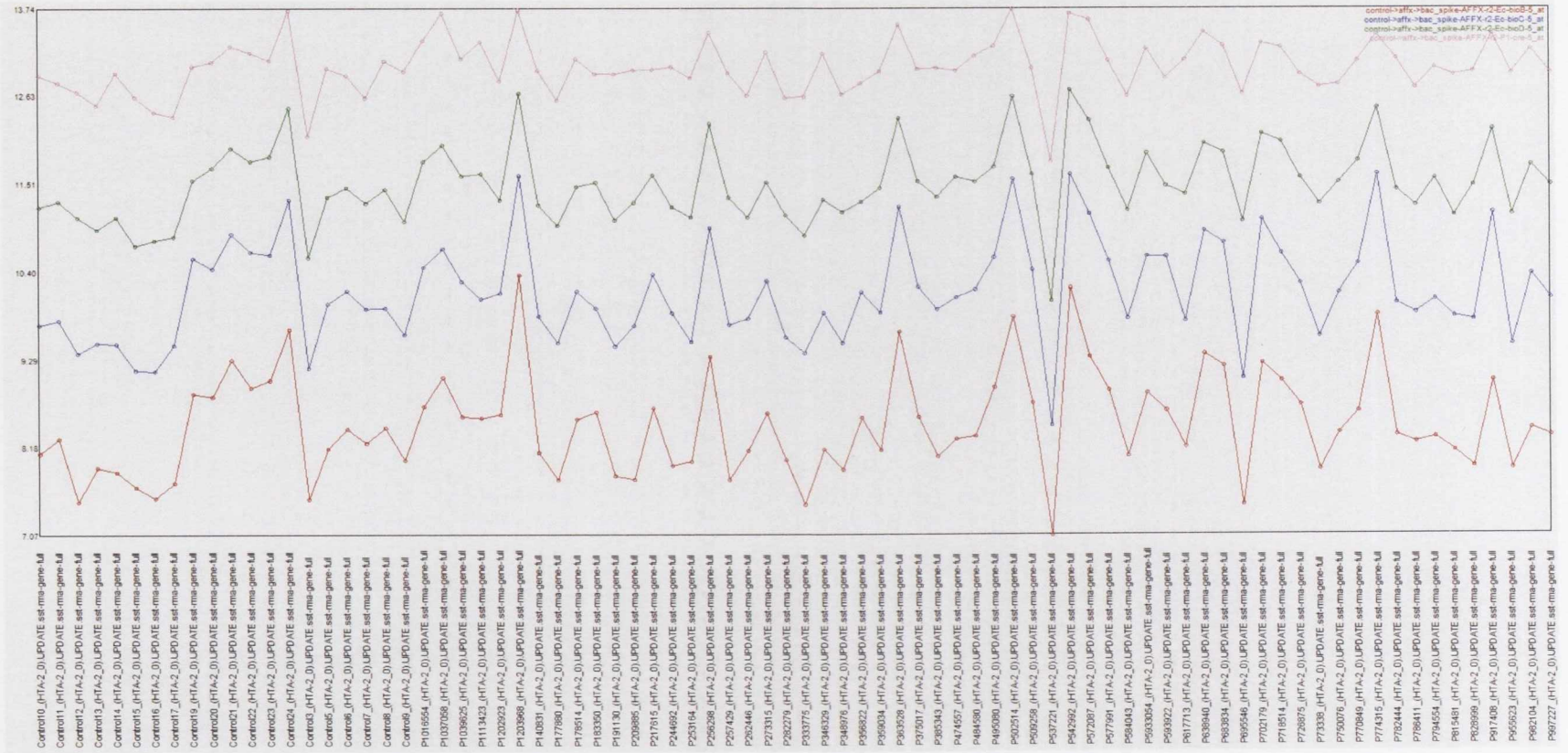


Figure 4.1: Line graph showing spike controls for each sample on the array. The x-axis represents each sample in the study and the y-axis represents the signal values for each spike control. BioB is shown in red, BioC in blue, BioD in green and Cre in pink.

4.3.11.1.2 Labelling controls

Poly-A controls are used to monitor the entire labelling process. Each array contains probe sets from a handful of *B. Subtilis* genes that are not present in eukaryotic genomic DNA (*lys, phe, thr*, and *dap*). The polyadenylated transcripts for these genes are synthesised *in vitro* and premixed at specific concentrations. These are known as Poly-A controls and are added to each RNA sample, carried through the sample preparation process and assessed like internal control genes. *Lys, phe, thr* and *dap* should display increasing signal values reflecting their increasing concentrations.

The results we obtained are characteristic of what one would expect from the poly-A controls (Fig 4.2). Despite what looks like a sharp decline in the control intensity levels on certain arrays, each array actually shows a high level of uniformity as the range of intensity values on the y-axis are essentially very tight to one another. For example, the intensity values for *Lys* (Blue) only span a narrow range from a minimum value of 3.47 to a maximum value of ~7, which is considered typical of the poly-A controls.

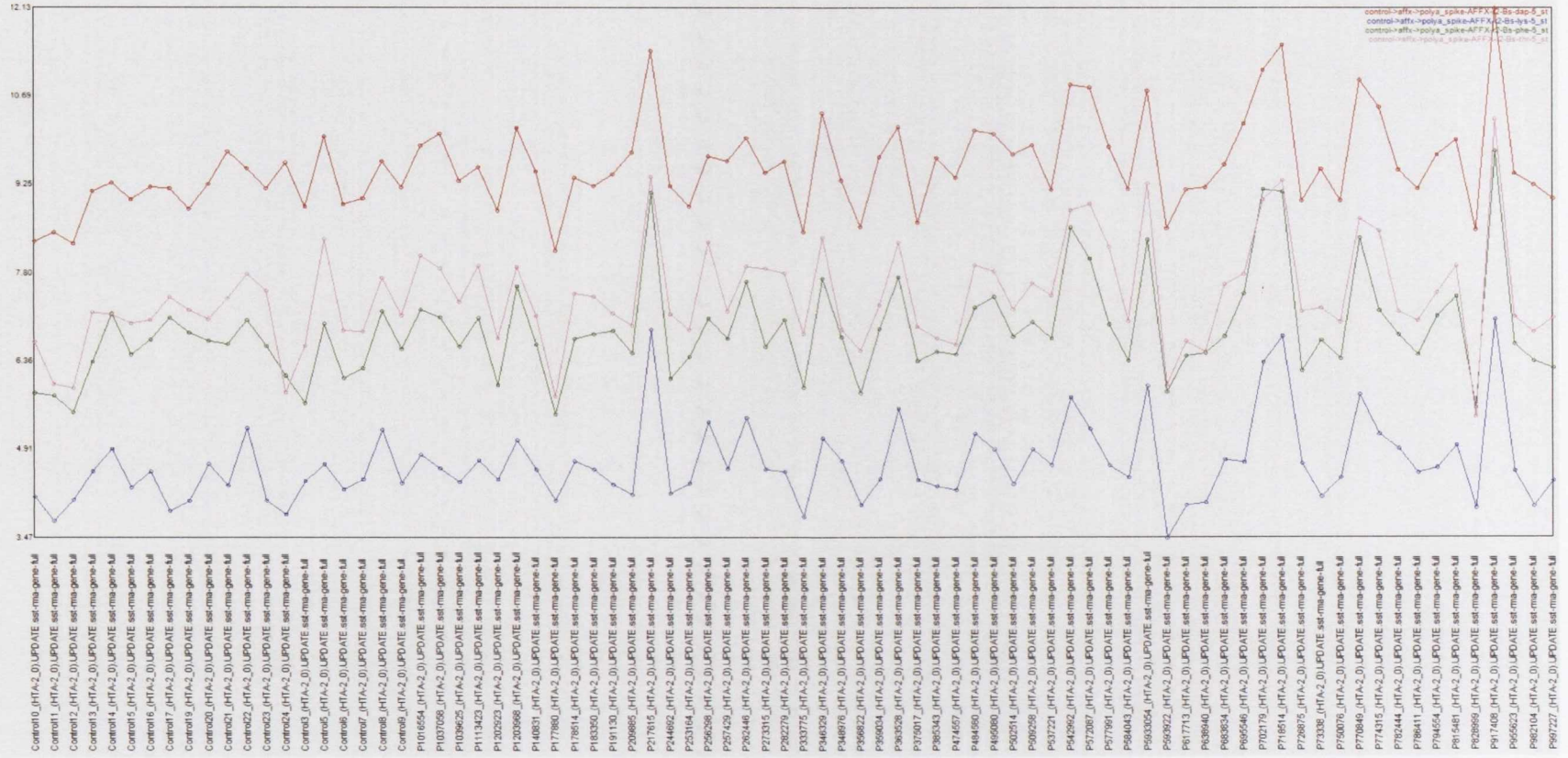


Figure 4.2: Line graph showing poly-A controls for each sample on the array. The x-axis represents each sample in the study and the y-axis represents the signal values for each poly-A control. *Lys* is shown in blue, *phe* in green, *thr* in pink and *dap* in red.

4.3.11.2 Box Plot

The Box Plot display can also be used to compare the probe intensity levels between arrays. In Expression Console (EC) there are two types of box plot available; box plots of the intensity values before normalisation or the signal box plots after normalisation.

4.3.11.2.1 Log Probe Cell Intensity box plot

The probe cell intensity creates a box plot of the log transformed probe intensity values for each array prior to normalisation (Fig 4.3). This box plot can convey some useful information. We can see that some of the arrays, highlighted by red diamonds have noticeably lower and higher overall probe intensity readings respectively than the other arrays. This could be an early indication of a problem with these arrays, however normalisation may correct these discrepancies.

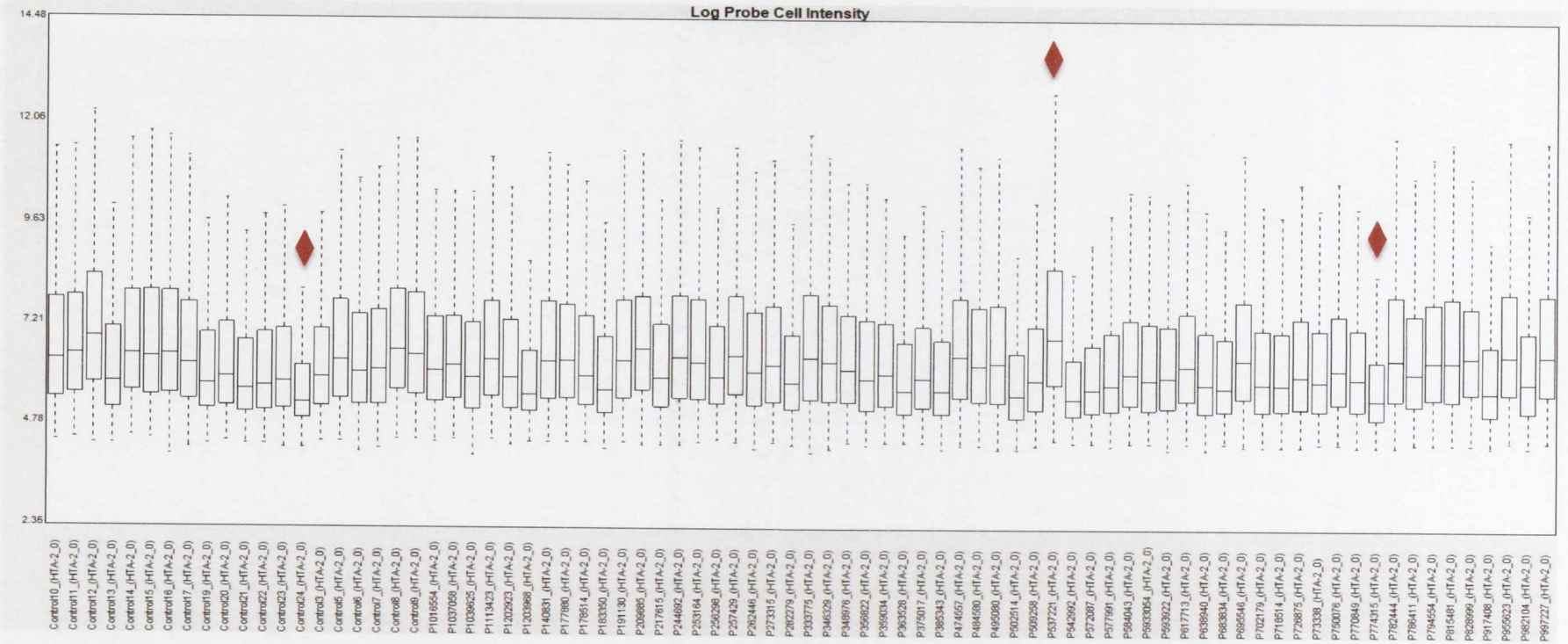


Figure 4.3: Box plots of probe cell intensity values. Each array is shown on the x-axis and the log2 intensity values are displayed on the y-axis. The line in the middle of the box represents the median and either end of the box represents the upper and lower quartile. The "whiskers" represent the smallest and largest values not considered outliers that lie within 1.5 times the interquartile range from the first of third quartile (the edges of the box).

4.3.11.2.2 Relative Log Expression Signal box plot

The relative log expression signal box plot compares the distribution of the ratio of signal for each probe set to the median probe set signal across all the selected arrays following the normalisation and summarisation steps. Fig 4.4 shows that the normalisation step has rectified the array discrepancy problem for most of the arrays, although one array in particular (marked with a red diamond) still looks noticeably different to the others. More information (i.e. further quality control steps) are required in order to make an informed decision whether we should include or exclude this array from the analysis. These QC steps are discussed below.

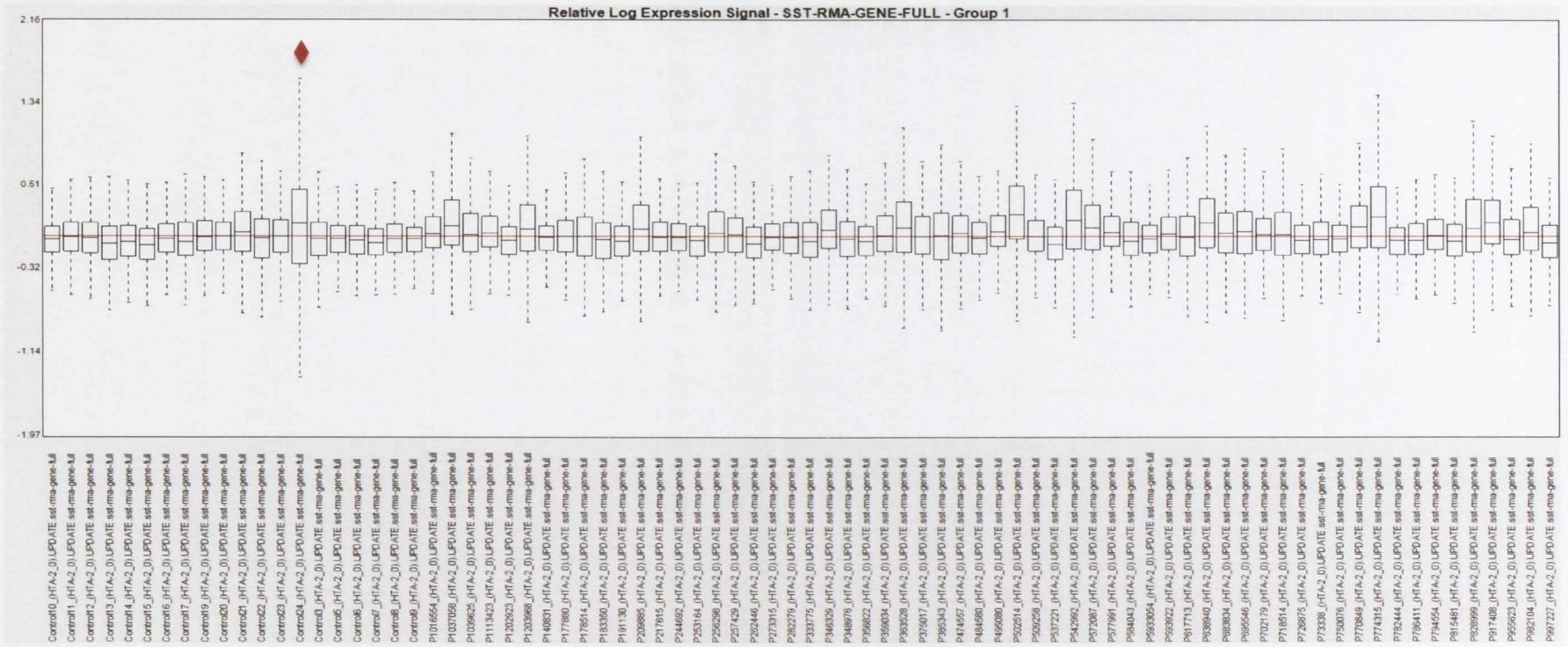


Figure 4.4: Box plot generated from the probe set signal values following normalisation and summarisation. Each array is shown on the x-axis and the log2 intensity values are displayed on the y-axis. The line in the middle of the box represents the median and either end of the box represents the upper and lower quartile. The "whiskers" represent the smallest and largest values not considered outliers that lie within 1.5 times the interquartile range from the first of third quartile (the edges of the box).

4.3.11.3 Signal Histogram

The signal histogram tells a similar story to that of the box plot and is also used to compare probe intensity between the arrays. In Fig 4.5 we can see that all of the arrays follow the same pattern and have a homogeneous signal distribution, no array appears to have a different expression pattern from any other array in the study. Any array that displays a very different signal distribution than the other arrays should be noted, and may need to be removed from the subsequent differential expression analysis.

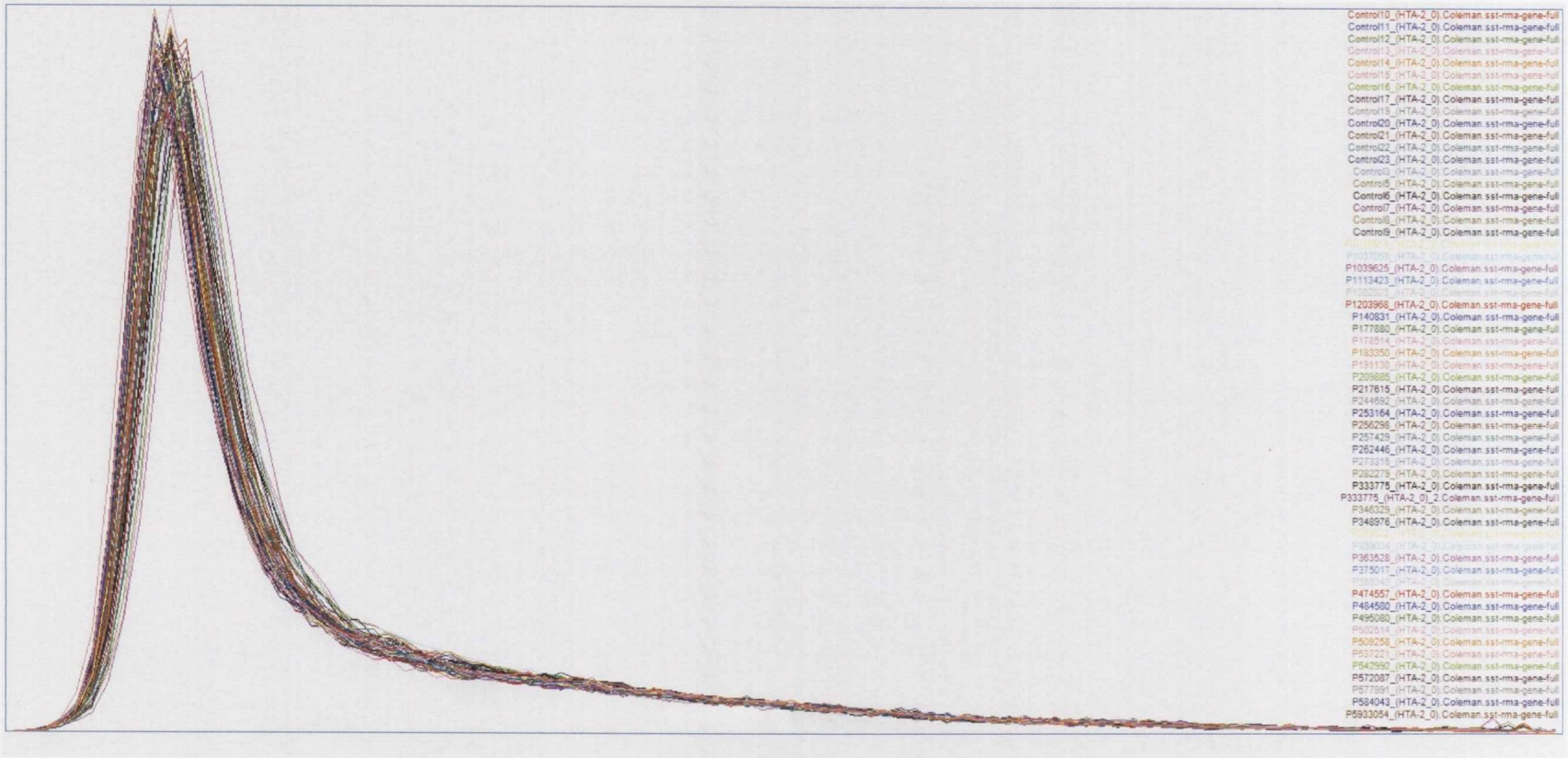


Figure 4.5: A density plot of the log-intensity distribution of each array. The probe log-intensity is represented on the x-axis and the probe density level is characterised on the y-axis. Each array is colour coded on the right hand side.

4.3.11.4 Pearson's Correlation (Signal)

The relationship between two variables is described by their correlation. The Pearson's correlation coefficient (r^2) value is used to compare the linear relationship between signal values between two individuals represented by a heat map. The closer the r^2 value is to 1 the greater the correlation between two samples. An r^2 of 1 is shown on the graph in dark red fading to dark blue for low r^2 values representing little or no correlation between samples. A green box highlights the control samples. We can see that out of our 20 control samples, 19 samples cluster very tightly with one another, which is evident from the high Pearson's correlation statistics which range from 0.97 - 1. However one control sample, control 24 (highlighted in yellow) did not appear to correlate with the rest of the control samples. The coeliac case samples represent a significantly more heterogeneous population, highlighted by an orange box, however, the majority of samples have a high Pearson's correlation statistic ranging from 0.97 – 1 (Fig 4.6).

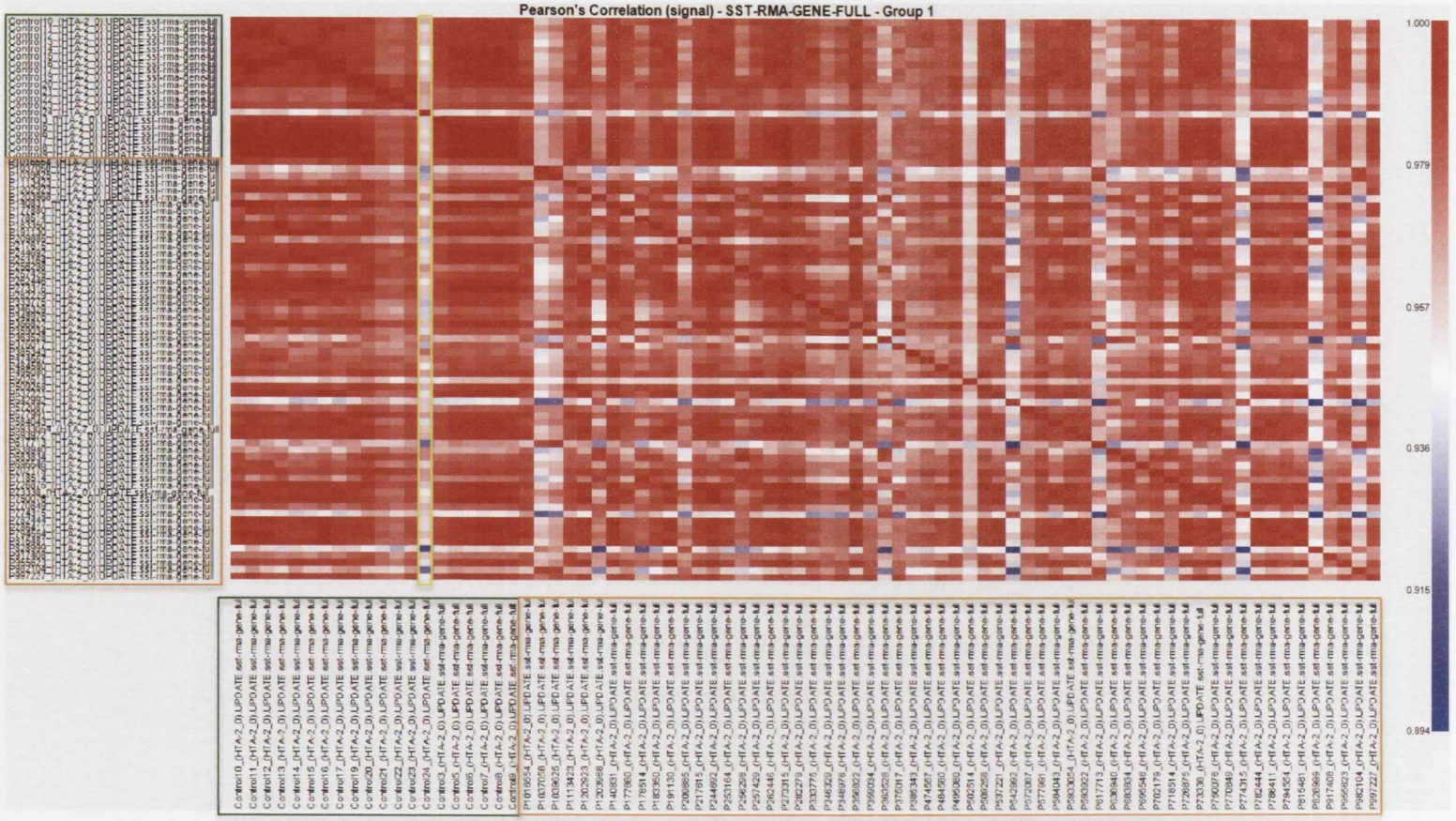


Figure 4.6: A heat map displaying the pairwise Pearson's correlation (r^2) of the signal values from each array in our study.

4.3.11.5 Principle Component Analysis

Expression Console enables Principle Component Analysis (PCA) to be carried out on the data. PCA identifies sets of variables (PCA1, PCA2, PCA3) that account for a majority of the variance in the original data set. PCA1 explains as much of the variability in the data as possible, followed by PCA2 which captures as much of the remaining variability, not accounted for by PCA1, as possible. PCA3 then captures as much of the remaining variability (not accounted for by PCA2) as possible and so on. 50,000 probe sets were used for this analysis.

PCA analysis (Fig 4.7) shows that our cases do not form separate or distinct groups, as one would expect. The control samples appear to cluster tightly together with the patient samples spread diffusely across the PCA space. This indicates that the overall gene expression pattern of a number of patients do not differ significantly between case and control groups in this dataset, however there are considerable differences in the gene expression in the extremes of the patient distribution. To determine if specific variables, such as gender, age at time of blood collection (Fig 4.8), age of diagnosis (Fig 4.9) or other factors (Table 4.1) were influencing these results; we subdivided the dataset into specific groups, extracted these individuals and performed separate PCA analysis on them. The protein coding genes and the non-protein coding genes were extracted separately and a number of comparisons were performed between these different variables to determine if the cases and controls would cluster into distinct groups under specific factors. Despite comparing a wide range of

different variables, no distinctive patterns emerged between our cases and controls or within the population of patients themselves.

Table 4.1: Variables examined using PCA analysis

Control Age >35-55 years	Vs	Patient Age >35-55 years
Control Age <35	Vs	Patient Age <35
Patient Age <35	Vs	Patient Age >35-55
Patient diagnosed <35 years	Vs	Patient diagnosed from 35->55 years
Patient Male	Vs	Patient Female
Patient	Vs	Control

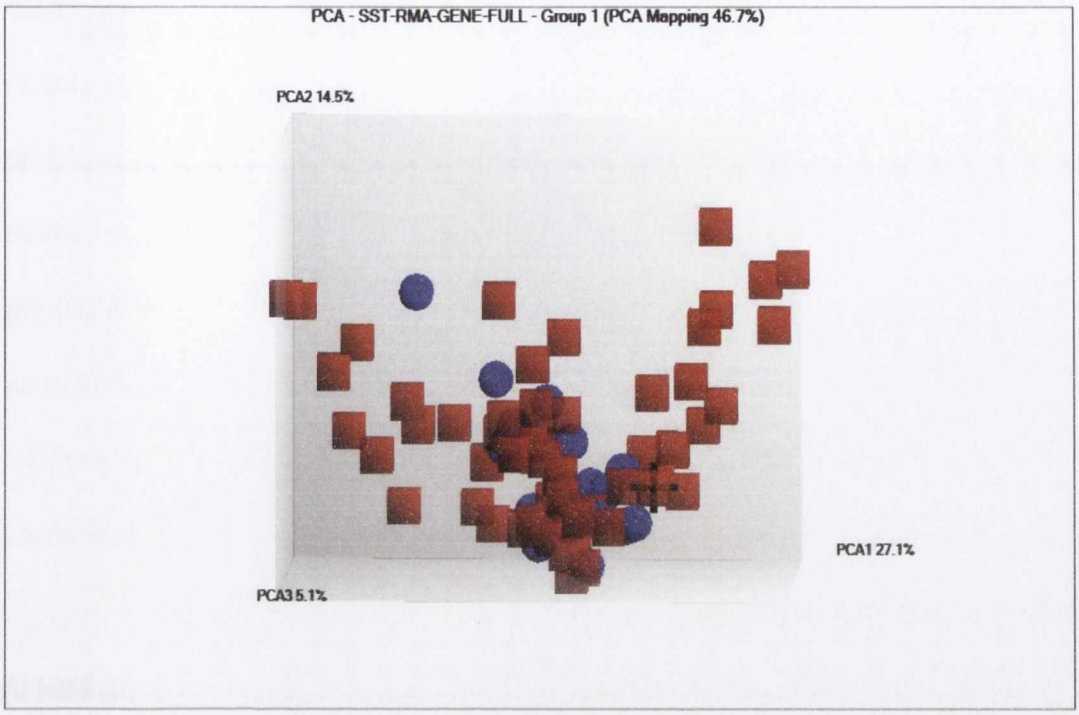


Figure 4.7: PCA plot of our coeliac samples versus controls. The red squares represent coeliac cases whilst the blue circles represent control samples. The plot shows the distribution of points on a principal plane, with the x-axis showing the projection of the point on the first principal component and the y-axis showing the projection of the point on the second principal component.

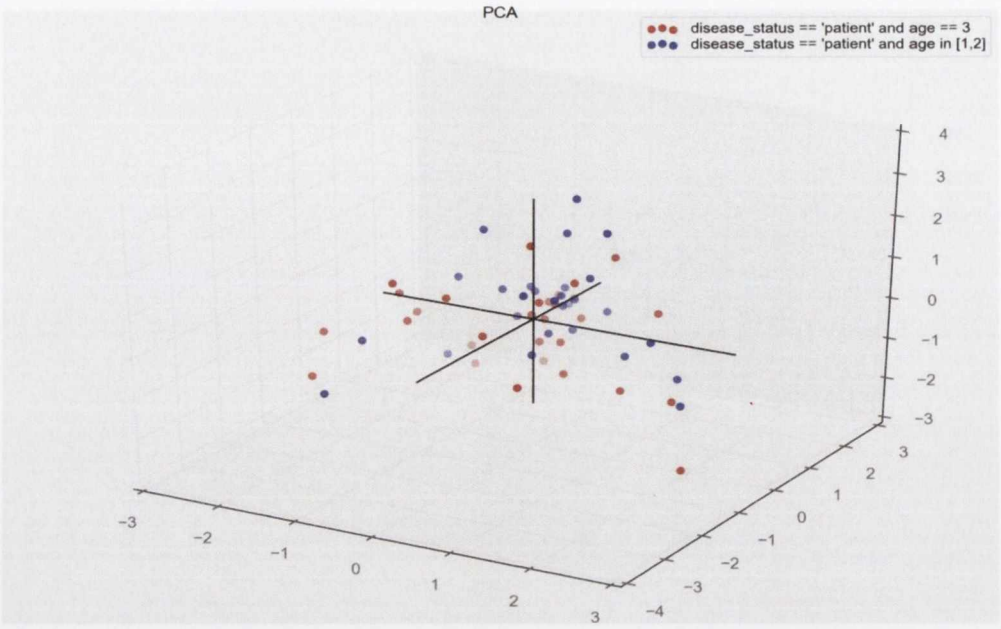


Figure 4.8: PCA plot of protein coding genes only, from coeliac patients separated by age at the time of blood collection. Blue = Age <35, Red = Age >35-55.

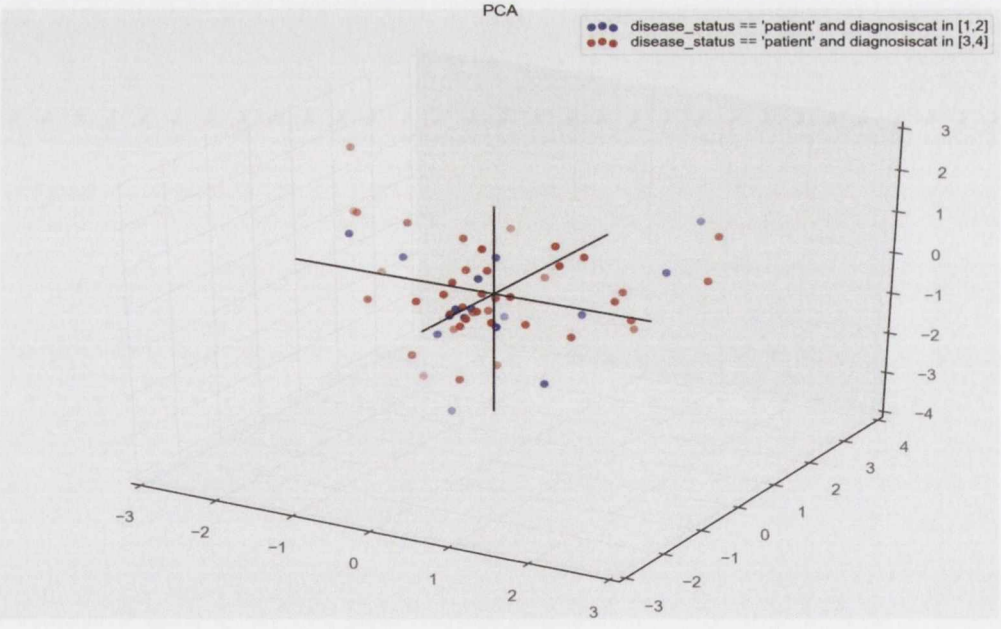


Figure 4.9: PCA plot of the protein coding genes only from coeliac patients seperated based on age of diagnosis. Blue = <35 years, Red = >35-55.

4.3.11.6 Removal of samples following QC

At this point 1 control sample was removed from the dataset. Control 24 failed to perform as well as we would have expected throughout all of the QC steps. Control 24 demonstrates less array intensity when compared to the other control samples both pre (Fig 4.3) and post normalisation (Fig 4.4) which could be indicative of a poor quality array and might not be representative of that individuals genotype. Moreover, control 24 does not correlate well with the other control samples as shown by a Pearson's correlation statistic of 0.94, highlighted by a yellow box in Fig 4.6.

A small proportion of coeliac samples did not perform as well as was hoped, which was evident from Pearson's correlation coefficient values (r^2) of <0.95. However none of the samples performed consistently poorly and the PCA plot did not show any obvious outliers, so none of the coeliac samples were excluded based on their QC performance. Two coeliac samples (718514 and 828999) were however removed for other reasons. 718514 was a duplicate of sample 178514, which was retained. We could also not gather any phenotypic information about sample 828999 therefore it was also removed. Thus there remained 19 control samples and 58 coeliac samples in the dataset for analysis post quality control.

4.3.12 Transcriptome Analysis Console (TAC) 2.0 software

TAC 2.0 software was used for the identification of differentially expressed genes, exons and visualization of alternative splicing events leading to the generation of

different transcript isoforms. It runs its analysis using EC generated CHP files. We identified differentially expressed genes using the default TAC filters, FDR P-value <0.05 and a linear fold change of < -2 and > 2. We also examined our results excluding both default criteria and with an FDR <0.05 excluding the linear fold change.

4.3.13 qRT-PCR

Quantitative real-time PCR was performed on five genes, *IFNG, IL18RAP, SOCS1, NOG* and *BACH2* using the Illumina ECO Real-Time PCR System. All primer sequences were pre-designed KiCqStart™ SYBR Green primers designed by Sigma-Aldrich. PCR was performed in a 10µl volume using sensiFAST SYBR PCR Master Mix (Medical Supply Company (MSC) LTD).

Amplification conditions were as follows:

Polymerase Activation: 95°C for 15 mins.

PCR Cycling: 40 cycles at 94°C for 1 min, 58°C for 1 min and 72°C for 1 min.

Melt Curve: 95°C for 15 secs, 55°C for 15 secs and 95°C for 15 secs.

Results were normalised by the expression of the housekeeping gene GAPDH and expressed as log2 fold changes compared with gene expression in healthy control subjects.

4.3.13.1 Calculating gene expression using threshold amplification cycle data (Ct)

The ΔΔCt method was used to calculate the log2FC from the gene expression data. GAPDH was used as the normaliser and the Ct value of the normaliser was subtracted from the Ct value from each sample in duplicate (ΔCt). The average ΔCt was then calculated. The ΔΔCt was calculated from the average ΔCt from all the cases and controls and the log2FC was calculated ($2^{\wedge \Delta\Delta Ct}$).

4.4 Results

4.4.1 Differential gene expression

TAC software supplied by Affymetrix was used to carry out differential expression (DE) analysis. The default filter criteria (linear fold change of < -2 and > 2 and FDR P-value <0.05) was used for initial analysis of DE genes. We found 188 DE genes, 142 coding genes and 46 non-coding. 26 genes (21 coding and 5 non-coding) were up regulated in coeliac cases vs control samples and 162 genes were down regulated (121 coding and 41 non-coding) (Fig 4.10 and Fig 4.11). Table 4.2 shows the top 25 most significantly up regulated and down regulated genes. A small number of these genes have either been found to be associated with CD in previous studies such as *IL2* (van Heel et al., 2007) or been investigated for there role in CD in previous studies, *IL12B* (Seegers et al., 2003). No genes overlapped between RNA-seq and microarray data at the default filter criteria (linear fold change of < -2 and > 2 and FDR P-value <0.05).

Table 4.2: The top 25 most significantly up regulated and down regulated genes. (linear fold change of < -2 and > 2 and FDR P-value <0.05).

Transcript Cluster ID	Fold Change (linear)	ANOVA p-value	FDR p-value	Gene Symbol	Group
Upregulated					
TC6_cox_hap2000091.hg.1	4.7	4.00E-06	1.70E-03	HLA-DQA1	Coding
TC6_qbl_hap6000082.hg.1	4.7	4.00E-06	1.70E-03	HLA-DQA1	Coding
TC04002704.hg.1	2.99	3.59E-03	9.88E-02		NonCoding
TC08002271.hg.1	2.91	6.62E-03	1.23E-01		NonCoding
TC19001524.hg.1	2.79	2.59E-02	1.95E-01	CLC	Coding
TC06000737.hg.1	2.78	2.21E-03	8.16E-02	CD109	Coding
TC01000226.hg.1	2.74	2.46E-04	2.77E-02		Coding
TC6_cox_hap2000193.hg.1	2.59	5.75E-04	4.28E-02	HLA-DRB1; HLA-DRB3	Coding
TC04001520.hg.1	2.51	6.23E-04	4.43E-02	IL2	Coding
TC06003152.hg.1	2.44	1.07E-03	5.89E-02		NonCoding
TC6_qbl_hap6000182.hg.1	2.44	8.55E-04	5.27E-02	HLA-DRB1; HLA-DRB3	Coding
TC04001570.hg.1	2.36	2.27E-02	1.86E-01	SLC7A11	Coding
TC08001099.hg.1	2.35	1.51E-02	1.63E-01	DUSP4	Coding
TC11001818.hg.1	2.33	2.82E-03	9.01E-02	MPEG1	Coding
TC08000482.hg.1	2.31	1.71E-03	7.32E-02	RDH10	Coding
TC01003490.hg.1	2.28	1.63E-02	1.67E-01	XCL2	Coding
TC06001677.hg.1	2.27	5.60E-04	4.21E-02		Coding
TC06001126.hg.1	2.25	4.31E-04	3.64E-02	SNX9	Coding
TC05002656.hg.1	2.19	1.20E-05	4.07E-03		NonCoding
TC06000946.hg.1	2.14	1.53E-04	2.07E-02	RNU1-18P	Coding
TC22001130.hg.1	2.1	2.96E-03	9.18E-02		NonCoding
TC01001469.hg.1	2.08	1.99E-02	1.79E-01	XCL1	Coding
TC05001997.hg.1	2.07	3.29E-02	2.12E-01	IL12B	Coding
TC05001064.hg.1	2.06	3.00E-05	7.83E-03	RNU1-17P	Coding
TC11002344.hg.1	2.06	4.00E-06	1.96E-03	AMICA1	Coding
Downregulated					
TC16000291.hg.1	-6.33	8.14E-08	1.62E-04	FLJ21408	Coding
TC12000885.hg.1	-5.06	9.60E-05	1.59E-02	OAS3	Coding
TC10000639.hg.1	-4.86	1.11E-03	5.98E-02	IFIT1	Coding
TC17000077.hg.1	-4.74	2.32E-04	2.68E-02	XAF1	Coding
TC01000795.hg.1	-4.63	7.40E-03	1.28E-01	IFI44	Coding
TC07001606.hg.1	-4.59	2.42E-04	2.76E-02	SAMD9L	Coding
TC01000794.hg.1	-4.45	2.75E-03	8.89E-02	IFI44L	Coding
TC12002572.hg.1	-4.44	3.91E-03	1.01E-01		NonCoding
TC01001767.hg.1	-4.32	6.10E-05	1.22E-02	DTL	Coding
TC19002491.hg.1	-4.1	3.00E-06	1.54E-03		NonCoding
TC06001798.hg.1	-3.9	1.14E-02	1.49E-01	IL17F	Coding
TC01002412.hg.1	-3.81	2.26E-03	8.24E-02	IFI6	Coding
TC04000484.hg.1	-3.81	7.00E-05	1.30E-02	HERC6	Coding
TC19002132.hg.1	-3.81	6.00E-06	2.60E-03		NonCoding
TC22000029.hg.1	-3.7	1.69E-02	1.69E-01	USP18	Coding
TC04001718.hg.1	-3.67	1.96E-04	2.40E-02	DDX60	Coding
TC01000835.hg.1	-3.66	5.40E-09	3.90E-05	GBP6	Coding
TC11003507.hg.1	-3.66	3.00E-06	1.54E-03	CARD17	Coding
TC19000554.hg.1	-3.6	8.00E-06	3.05E-03	517A; OTTHUMG0000014	Coding
TC22000647.hg.1	-3.59	1.08E-02	1.46E-01	LIF	Coding
TC01002500.hg.1	-3.51	6.35E-07	5.28E-04	CLSPN	Coding
TC18000006.hg.1	-3.47	1.10E-07	1.78E-04	TYMS	Coding
TC08002095.hg.1	-3.42	3.91E-04	3.47E-02		NonCoding
TC17002726.hg.1	-3.39	1.07E-02	1.46E-01		NonCoding
TC10000098.hg.1	-3.3	1.43E-04	2.02E-02	MCM10	Coding

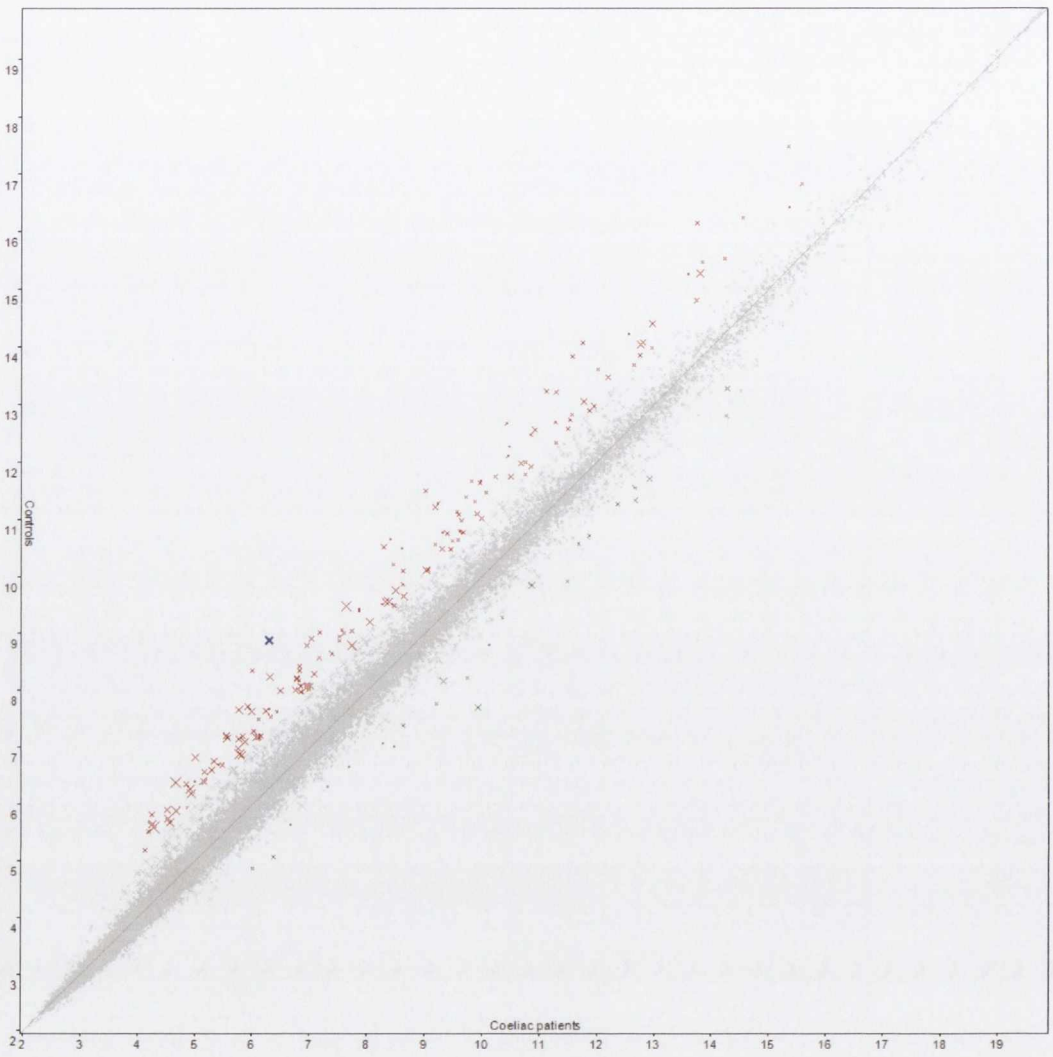


Figure 4.10: A scatter plot of controls versus coeliac cases. Coeliac cases are represented on the x-axis and controls are represented on the y-axis. The red X's represent those genes that are up regulated in controls (condition 1) vs. cases (condition 2). The green X's represent those genes that are down regulated in controls (condition 1) vs. cases (condition 2). The size of the X is based on P-value, the smaller the P-value, the bigger the X. The grey dots represent genes not meeting our threshold linear FC < -2 and > 2 and FDR <0.05.

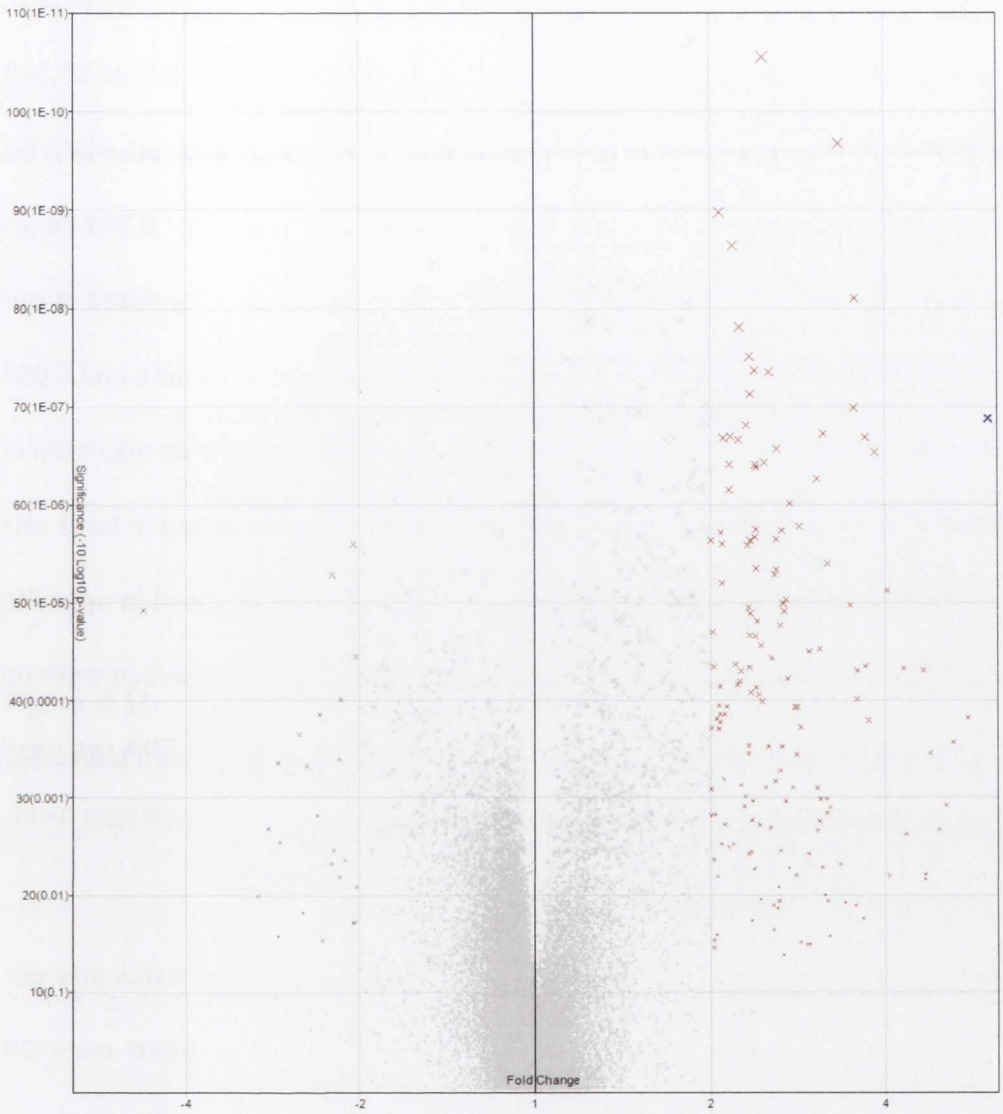


Figure 4.11: A Volcano plot of controls versus coeliac cases. Significance is represented on the y-axis as the -10log10 P-value of the FDR P-values versus the linear fold change on the x-axis. The red X's represent those genes that are up regulated in controls (condition 1) vs. cases (condition 2). The green X's represent those genes that are down regulated in controls (condition 1) vs. cases (condition 2). The grey dots represent genes not meeting our threshold linear FC < -2 and > 2 and FDR <0.05.

4.4.2 Comparison of microarray results with differentially expressed RNA-seq genes

When we looked at both sets of results we found that 13,319 genes out of 67,528 genes (~20%) from the Affymetrix array, overlapped between both datasets when no filters or thresholds were applied to the data. Of these 13,319 genes - 6,743 genes were up regulated in our RNA-seq data and 6,576 genes were down regulated in our RNA-seq data, whilst 7,225 genes were up regulated in our microarray data and 6,094 genes were down regulated in our microarray data. 3,546 genes were up regulated in both RNA-seq and microarray results. 2,897 genes were down regulated in both sets of data. However, a large number of genes were differentially expressed in opposite directions in both datasets; 3,679 genes that were down regulated in RNA-seq were up regulated in microarray data and 3,197 genes that were up regulated in RNA-seq were showing down regulation in our microarray data (Fig 4.12)

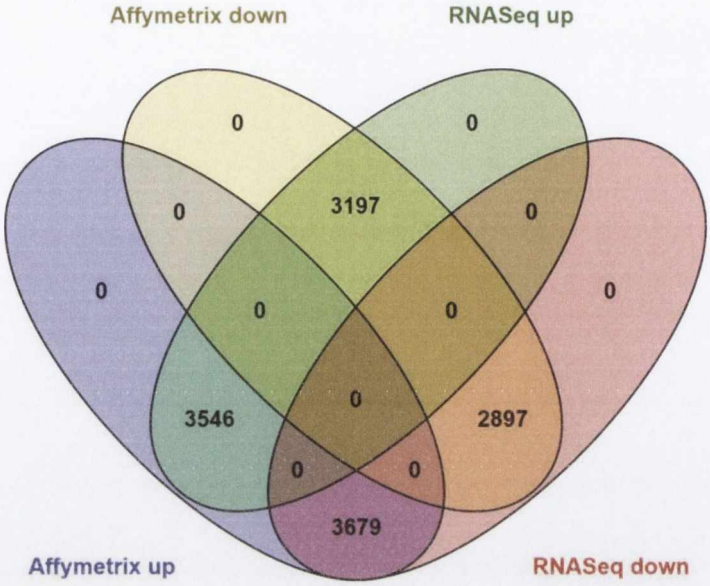


Figure 4.12: Venn diagram showing comparison of differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data. RNA-seq up regulated = green, down regulated = pink. Microarray up regulated = blue, down regulated = yellow)

We also extracted anti-CD3/CD28 significantly differentially expressed genes from our RNA-seq study (log2FC >1 & FDR of 5%) and compared these with our microarray results (no filters applied). 955 genes that were significantly differentially expressed in our RNA-seq data were also found in our microarray data and thus were extracted for comparison. 597 of these genes were up regulated in our RNA-seq and 358 genes were down regulated. 412 genes were up regulated in our microarray data, whilst 543 genes were down regulated. 207 genes were up regulated in both groups, whilst 153 genes were down regulated in both groups. 390 genes that were up regulated in RNA-seq were down regulated in our microarray results while 205 genes that were up

regulated in our microarray data were down regulated in our RNA-seq results at the specified thresholds (Fig 4.13).

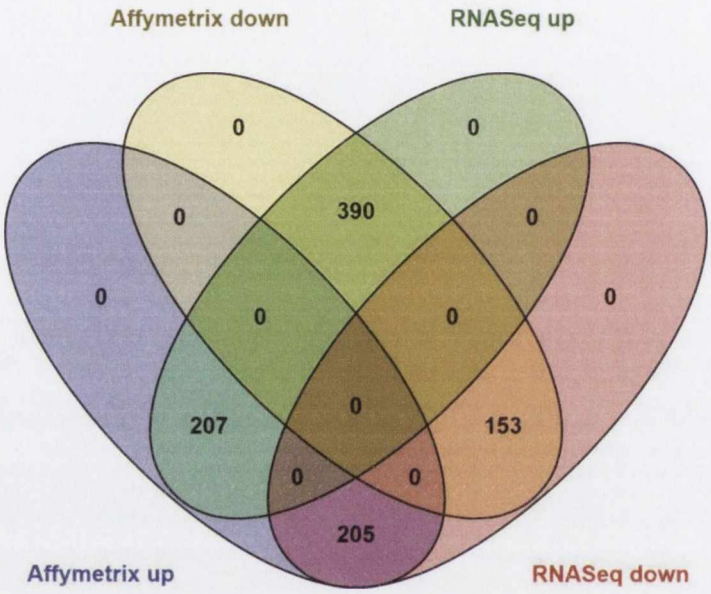


Figure 4.13: Venn diagram showing comparison of anti-CD3/CD28 stimulated $CD4^+$ T-cells significantly differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data.

We next applied an FDR of 5% to the microarray data as this was the threshold applied to our RNA-seq data and compared our results. Following this, a total of 94 out of 955 genes from our microarray results overlapped with our RNA-seq data at an FDR of 5%. However the majority of these genes did not correlate in their direction of expression. 87 out of 94 genes were down regulated in our microarray data but up regulated in the RNA-seq data (Fig 4.14). 4 genes show up regulation in our microarray data and down

regulation in the RNA-seq data whilst only 3 genes show differential expression in the same direction; *GK3P, CABLES1* and *CHN1* are up regulated in both datasets. TAC software displays the changes in gene expression from our microarray study as a linear fold change (FC) rather than log2FC, so we converted the log2FC used to measure differential expression in the RNA-seq data to linear fold change to make both datasets comparable.

CHN1 was the most strongly correlated gene, with a FC of 1.49 ($P_{adj} = 1.01x10^{-2}$) in the array results and FC= 2.26 ($P_{adj} = 1.33x10^{-2}$) in the RNA-seq data. CHN1 plays an important role in the early development of nerve cells. *GK3P* had a FC of 1.27 ($P_{adj} = 2.8x10^{-2}$) in the array results and FC= 3.85 ($P_{adj} = 4.44x10^{-5}$) in the RNA-seq results. *GK3P* is a glycerol kinase 3 pseudogene. *CABLES1* had a FC of 1.15 ($P_{adj} = 3.71x10^{-2}$) in the array data and FC= 3.68 ($P_{adj} = 3.98x10^{-3}$) in RNA-seq results. CABLES1 is a cell cycle regulatory protein, which interacts with several cyclin-dependent kinases.

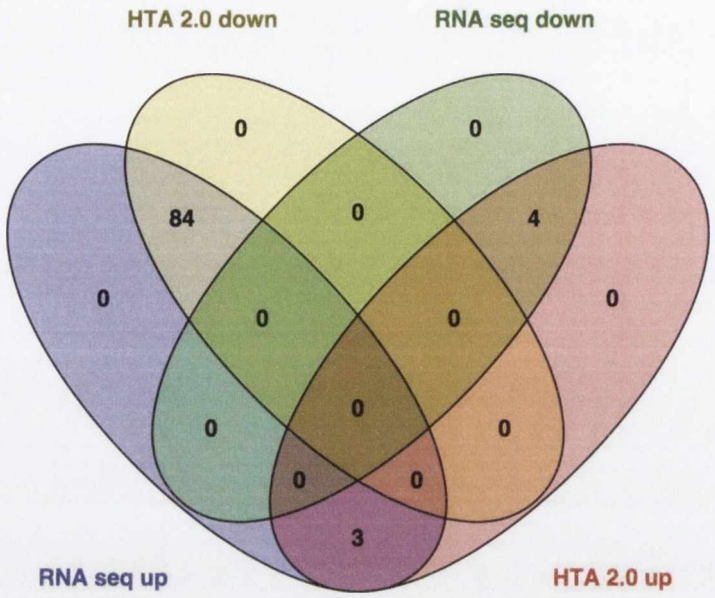


Figure 4.14: Venn diagram showing comparison of anti-CD3/CD28 significantly differentially expressed genes identified from RNA-seq and HTA 2.0 microarray data at an FDR of 5%. RNA-seq up regulated = blue, down regulated = green. Microarray up regulated = pink, down regulated = yellow).

4.4.3 Comparison of microarray DE genes with CD associated genes

We compared genes that were associated with coeliac disease in genomic studies with DE genes from the microarray data; we identified 75 coeliac associated genes when no filters were applied to the data (Fig 4.15). We also compared the 188 DE expressed genes (linear fold change of < -2 and > 2 and FDR P-value <0.05) from the microarray data with coeliac associated genes and identified 2 coeliac associated genes, *CCR2* (FC = -2.75, $P_{adj} = 5.31x10^{-3}$) and *IL2* (FC = 2.51, $P_{adj} = 4.43x10^{-2}$) (Fig 4.16).

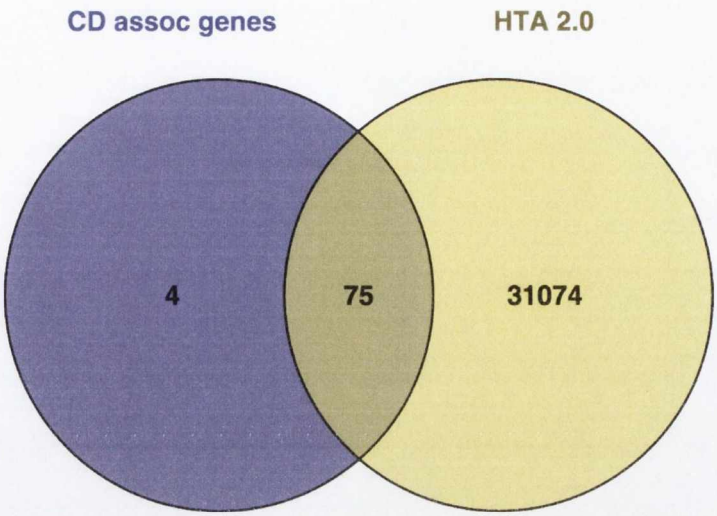


Figure 4.15: **Venn diagram showing the overlap of genes DE in our microarray data (no filters applied) with previously identified CD susceptibility genes.**

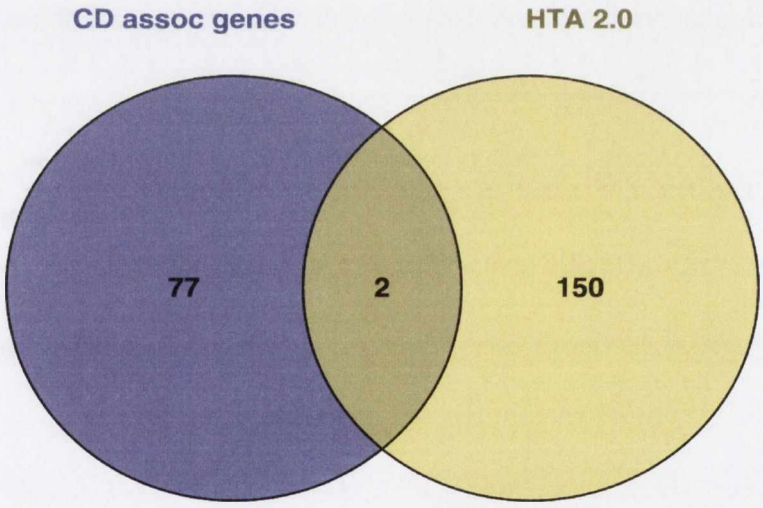


Figure 4.16: Venn diagram showing the overlap of genes DE in our microarray data (linear fold change of < -2 and > 2 and FDR P-value <0.05) with previously identified CD susceptibility genes.

4.4.4 Differential exon analysis

We also used the HTA 2.0 to examine the extent of alternative splicing on transcript regulation of $CD4^+$ T-cells in both patients with coeliac disease (58) and unaffected controls (19) by applying the differential exon tool in expression console and TAC 2.0 software. We identified 837 exons from 300 genes with alternative splicing events; 287 genes showed splicing in the coding regions and 13 genes in the non-coding regions (FDR<5%). As previously mentioned, we identified 647 exons from 542 genes with differential usage between the coeliacs and controls in the anti-CD3/CD28 stimulated samples (FDR<5%) in our RNA-seq study. We extracted this gene list and overlaid it with our alternative splicing results from the HTA 2.0. We found 7 genes *(HSH2D, MCM2, ATP2B4, TGIF1, ZNF367, IL7R* and *MCM4)* that overlapped between both datasets containing significantly differentially expressed exons but were not DE at gene level in our microarray data.

4.4.5 Comparison of microarray DE exons with CD associated genes

Interestingly, when we compared genes that were associated with coeliac disease in genomic studies with genes showing alternative splicing in the microarray data, we identified three coeliac associated genes, *IL18R1, CCR2* and *LPP* that also contained significantly differentially expressed exons but were not DE at the gene level (Fig 4.17).

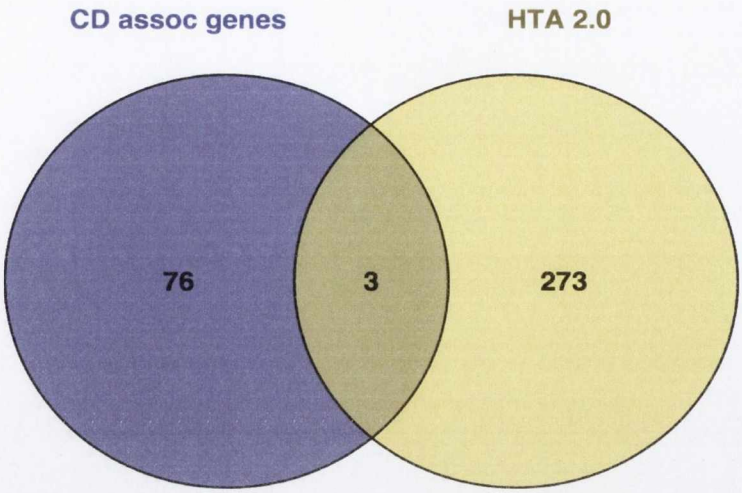


Figure 4.17: Venn diagram showing the overlap of genes showing evidence of alternative splicing in our microarray data with previously identified CD susceptibility genes.

4.4.6 qRT-PCR

Our RNA-seq results failed to be replicated using the HTA 2.0 array, with 92% of DE genes showing expression in opposite directions on both platforms and only 3% of genes showing differential expression in the same direction. Key genes such as *IFNG* and *IL18RAP* that one could expect to be up regulated in CD were down regulated on the HTA 2.0 array (although this down regulation was not significant) and in fact there was very little change in gene expression across all of the genes on the array. Therefore this raised concerns related to the reliability of the microarray results. In an attempt to investigate further these discordant findings, we used qPCR to validate mRNA expression for five specific genes, *IFNG, IL18RAP, SOCS1, NOG* and *BACH2* using RNA taken from 9 patients and 6 controls which had been run on the HTA 2.0 array.

IFNG, IL18RAP and *SOCS1* were chosen because both genes were significantly up regulated in our RNA-seq coeliac samples and have also been shown to be up regulated in additional coeliac studies also (Bodd et al., 2010; Plaza-Izurieta et al., 2011). *NOG* and *BACH2* were chosen because both genes were significantly down regulated in our RNA-seq coeliac samples and BACH2 has been shown to be down regulated in other coeliac studies (Dubois et al., 2010). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as the internal standard reference gene to normalize mRNA levels between different samples for an exact comparison of mRNA transcript level.

The gene expression qPCR results did not replicate their respective expression patterns from the HTA 2.0 array for four out of five genes - *IFNG, IL18RAP, NOG* and *BACH2* - but instead correlated with the direction of expression in our RNA-seq results although the extent of change in expression was less in all cases in the qPCR data. However *SOCS1* showed the same direction of expression as it did in the HTA 2.0 gene array data (Table 4.3 and Fig 4.18).

Table 4.3: Log2FC values of five genes analysed via qRT-PCR across all three platforms.

Gene	RNAseq Log2FC	HTA 2.0 Log2FC	RT-PCR Log2FC
IFNG	4.65	-0.189	2.13
IL18RAP	1.97	-0.378	1.02
SOCS1	1.19	-0.31	-0.292
NOG	-3.4	0.084	-0.15
BACH2	-1.75	0.298	-0.265

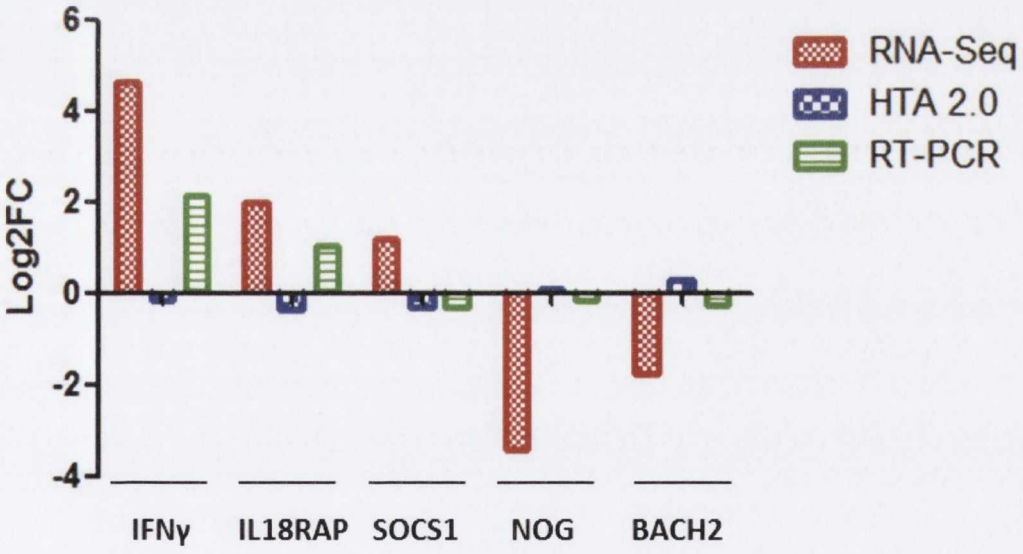


Figure 4.18: Bar chart showing levels of gene expression, represented by log2FC across all three platforms for all five chosen genes.

4.4.7 Genotyping

We had reserved several Immunochips so that we could fully genotype all individuals included in our RNA expression studies and use this data to correlate genotype and gene expression. Unfortunately however recruitment of patients for the gene expression studies took considerably longer than anticipated and the Immunochips were past their usable date by a significant margin by the time they were used, and as production had been discontinued, they were not replaceable. We nonetheless decided to try genotyping twenty-five samples, which had not previously been genotyped on the Immunochip. Unfortunately, all of the Immunochips failed and we could not collect genotype information from these patients. Therefore, this meant that we could no longer carry out a global genome wide eQTL study as we had initially planned. Illumina state that a new version of the Immunochip is due to be released in the third quarter of this year and we hope that this data can be collected at that time; however this time frame puts it beyond the scope of this thesis.

4.5 Discussion

Gene expression studies are a highly informative means of revealing a global picture of genomic activity and cellular function under specific conditions. Such insights provide improved understanding and allow for the generation of hypotheses about the underlying biology relating to disease processes. Using microarrays and more recently, next generation sequencing (NGS) technologies, gene expression studies have revolutionised our understanding of the human transcriptome.

As the first developed method of transcriptome profiling, gene expression microarrays have been used in the characterisation of many different cell types and tissues, developmental states and disease (Alizadeh et al., 2000). Microarrays have been used in previous coeliac studies (Diosdado et al., 2004; Bracken et al., 2008) to identify differentially expressed genes potentially contributing to disease pathology. There have been many successful applications of this technology, however it also has several well-acknowledged shortcomings such as poor sensitivity, high background noise, cross-hybridisation between genes, the inability to detect and quantify previously uncharacterised coding and non-coding transcript isoforms, and poor dynamic range making it difficult to confidently detect and quantify lowly expressed genes (Shendure, 2008).

Thus RNA-seq has surpassed microarray technology as the method of choice, with its superior inter-platform reproducibility, greater dynamic range, and the power to detect alternative splicing events to name a few. RNA-seq has the limitation of being

more expensive and choices need to be made in the set up stages for example whether non-poly A or small transcripts should be included which might impinge on overall power of the study. Nonetheless, RNA-seq has been the preferred method for characterising gene expression in recent years and a method we carried out successfully in our coeliac case-control transcriptome study described in chapter 3. Unfortunately, to carry out transcriptome analysis on a larger scale using RNA-seq was not an option at the time of this study but by taking advantage of the newly developed Affymetrix GeneChip® HTA 2.0 array we hoped to gather broader range of information to what we derived using RNA-seq. This high-resolution array was claimed to be superior than previous microarray platforms by minimising the limitations discussed previously and also claimed to rival RNA-seq by not only interrogating all coding transcripts in the cell but it also covered non-coding transcripts and exon-exon spice junctions for a fraction of the cost. Thus, this array would not only provide us with a gene expression profile of the entire transcriptome in an expanded population, it would also, in contrast to our RNA-seq study, give us the opportunity to detect non-coding RNAs and alternative splicing events in a more effective manner, something our particular RNA-seq study was not designed to achieve.

As previously mentioned, recent GWAS and fine mapping studies have identified over 40 CD susceptibility loci to date. However, detecting causal genes and specific variants within these signals and identifying their role in CD pathogenesis is still proving to be a challenge. The majority of SNPs (90%) are located in non-coding regions of the genome, suggesting that they are eQTLs playing a regulatory role in the pathogenesis

of CD. eQTL mapping has successfully identified SNPs that are associated with altered gene expression in a number of studies of CD. Hence, one of the objectives of this phase of the project was to use this array to gather transcriptome data on a larger study population in order to perform an eQTL study.

Numerous eQTLs have been found in CD associated loci to date (Dubois et al., 2010; Amundsen et al., 2014; Fairfax et al., 2014). An important progression in the design of these studies has been towards greater tissue/cell specificity to investigate key cellular players in the disease process. Another important consideration as shown by Fairfax and colleagues (2014), is the activation status of cells. Thus, we felt carrying out an eQTL study in activated $CD4^+$ T-cells specifically would be beneficial to help achieve a much more defined picture of the genetic architecture underlying a cell type with a critical involvement in CD.

Using both Expression Console and Transcriptome Analysis Console 2.0 software 188 significantly DE genes were identified between cases and controls using Affymetrix default filters (FC >2 and FDR 5%). A number of these genes such as *IL2* (FC=2.51, P = $4.43X10^{-2}$) have been implicated in CD in previous studies. *IL2* is in our 'top 10' most differentially expressed genes (Table 4.2). Van Heel *et al* (2007) discovered CD risk variants in the genomic region harbouring the genes *IL2* and *IL21* thus implicating *IL2* in CD pathogenesis. *IL2* is a key cytokine involved in the activation and proliferation of T-cells and has also been associated with other autoimmune diseases such as Type 1 Diabetes and Rheumatoid Arthritis (Zhernakova et al., 2007). *IL12B* is another gene

that is significantly up regulated in our microarray data. *IL12B* has previously been studied for its involvement in CD pathogenesis due to the key role it plays in IFNγ secretion, one of the pivotal inflammatory mediators of CD. However no link has been found between *IL12B* and CD to date (Seegers et al., 2003). We found significant down regulation of *IL17F*, a Th17 T-cell marker in patients with CD compared to controls (Log2FC -3.9, $P_{adj} = 1.49x10^{-1}$). This down regulation correlates in the direction of effect with our RNA-seq results, however the down regulation in the RNA-seq data was not significant (Log2FC -0.55, $P_{adj} = 0.178$). Th17 cells play a role in gut homeostasis and inflammation and whilst Th17 cells have been implicated in CD (Fernández et al., 2011), this observation might indicate decreased activity of Th17 cells which may be counter intuitive in an inflammatory condition such as CD.

With the HTA 2.0 array boasting such a high sensitivity and comparability to RNA-seq one would expect a high concordance between significantly DE genes on both platforms. Unfortunately however, the quantity and identity of significantly DE genes in the HTA 2.0 array data showed minimal concordance with our RNA-seq results. 188 genes were significantly DE in the array data at a default fold change of >2 and FDR of 5% in comparison to 1069 significantly differentially expressed genes in our RNA-seq study. When we compared these gene sets, we found that no genes overlapped. Removing the fold change filter from our microarray results left us with 955 overlapping genes, however there was very little consistency in the direction of effect with only 3 genes, *GK3P*, *CABLES1* and *CHN1* showing up regulation across both platforms.

None of these 3 genes have been associated with coeliac disease in previous studies. The *CHN1* gene encodes for two proteins, α1 chimaerin and α2 chimaerin, both of which are involved in neuronal signalling pathways. Mutations in this gene have previously been associated with Duane's retraction syndrome 2 (DURS2) (Miyake et al., 2008). However, in a recent study *CHN1* has been shown to be up regulated in the autoimmune disease cutaneous lupus erythematosus (CLE) (Kunz, 2013). The link between coeliac disease and systemic lupus erythematosus (SLE) has already been established, with individuals with CD at a greater risk of developing SLE (Ludvigsson et al., 2012; Park et al., 2013). Therefore it is possible there may also be a connection between CLE and coeliac disease and perhaps a role for *CHN1* in both diseases.

CABLES1 is a cyclin dependent kinase (CDK) binding protein that is involved in control of the cell cycle (Sakamoto et al., 2008). It serves as a negative regulator of cell proliferation and plays a role in apoptosis. Loss of *CABLES1* has been seen in many cancers such as ovarian (Dong et al., 2003), colon (Wu et al., 2001) and lung (Tan et al., 2003), where it may function as a tumour suppressor protein. Although *CABLES1* has not been implicated in CD in previous studies, it has been found to be associated with rheumatoid arthritis (RA) (Lu et al., 2012) where it was shown to be up regulated in patients and was linked to higher rates of $CD4^+$ T-cell apoptosis. Increased levels of apoptosis have been found in PBMCs in CD (Di Sabatino et al., 2001), however, there is no evidence of an increase in apoptosis of $CD4^+$ T-cells in patients in our study.

Despite 188 genes showing significant DE in our microarray data, the majority of the genes on the array looked 'flat' showing little variation in gene expression between cases and controls. PCA analysis showed that our coeliac patients are dispersed diffusely within the PCA space, with the controls in general forming a tighter cluster within this space (Fig 4.7). There was no evidence of distinct clustering within the patient group, however the diffuse spread of patients compared to controls within the 3 principal components indicates significant difference of gene expression between the extremes of the patient distribution. We attempted to correlate the patient distribution with phenotypic characteristics including age at sample collection, age at diagnosis and gender, however none of these were significant. There was also no association with batch effect.

As discussed in chapter 3, we found that 5 coeliac patients clustered tightly with controls upon both anti-CD3/CD28 and PMA stimulation and phenotypic assessment of a subset of the patient characteristics, such as age, gender and clinical presentation did not discriminate between groups. However, in a *post hoc* analysis of the RNA-seq dataset we observed that patients with a 'coeliac-like' expression profile had a tendency towards higher than average tTG levels and those patients that displayed a 'control-like' response had lower levels. This was unanticipated given that all patients were self reportedly adhering to a gluten free diet (GFD) and analysis of tTG levels was not used as an enrolment criterion in our expanded study.

However, the results from the HTA 2.0 array show that whilst there is some overlap between cases and controls, the cases represent a much less defined group indicating wide variability of gene expression in the patient sample, which is not seen in the controls. The lower tTG levels may explain some of this variability, as could different patterns of susceptibility gene inheritance or other environmental factors. As previously discussed, variability has been documented in coeliac disease by Anderson *et al* (2006) who discovered a subset of coeliac patients that were unresponsive to gliadin stimulation which presumably would be reflected in different gene expression profiles.

However, another possible explanation as to why our coeliac cases fail to exhibit a disease specific profile or cluster into separate groups could be due to the performance of the array itself. As mentioned above, there is no correlation between genes that are above the thresholds (FC >2, FDR 0.05) and when these thresholds are removed there is still no correlation in the direction of effect; genes that were showing up regulation on the array were down regulated in our RNA-seq study and vice versa. For instance IFNγ is a well established pro-inflammatory gene that has been shown to be up regulated not only in our RNA-seq study (Log2FC=4.6, $P_{adj} = 2.40x10^{-11}$) but also in previous celiac disease studies (Olaussen et al., 2002;Wapenaar et al., 2004; Plaza-Izurieta et al., 2014). However, IFNγ shows a non-significant decrease in expression in our microarray results (Log2FC = -0.189, $P_{adj} = 0.327$). *IL18RAP* is another gene that has shown significant up regulation in our RNA-seq study (log2FC=2, $P_{adj} = 2.05x10^{-6}$) and in multiple published coeliac studies (Lotta L. E. Koskinen et al., 2009; Plaza-Izurieta et al.,

2011; Pascual et al., 2015) but is non-significantly down regulated in our microarray results (log2FC= -0.378, P_{adj} = 0.381). These unexpected results raised serious issues regarding the quality or reliability of the array results.

However, since the release of the HTA 2.0 array in 2014, there have been several studies published reporting both gene expression and alternative splicing analysis in various diseases, such as multiple sclerosis (Hecker et al., 2015), breast cancer (Rogers et al., 2014) and hepatitis C (Daoud, 2014). A study by Ujvari *et al* (Ujvari et al., 2014) used qRT-PCR to verify their Affymetrix GeneChip HTA 2.0 results and showed excellent correlation between both platforms. To our knowledge, there have been no studies published comparing RNA-seq and the HTA 2.0 array.

To try to determine if the arrays were giving a reliable reading of gene expression we re-tested a small set of patients and controls using qRT-PCR. A small selection of candidate genes (*IFNG, IL18RAP, SOCS1, NOG* and *BACH2*), both up and down regulated in our RNA-seq data but nominally showing the opposite direction of effect on the array were chosen for analysis.

The gene expression patterns of *IFNG, IL18RAP, NOG* and *BACH2* correlated with our RNA-seq results and showed the opposite direction of effect from our microarray results. The only gene concordant with our microarray results was *SOCS1.* These findings are broadly supportive of our RNA-seq data, however, they are not significant and the size of the change is small, warranting further investigation. It is also worth mentioning that these patient samples were chosen at random and therefore many of

the patients could display 'control-like' expression patterns, which is a potential problem with such a small sample size.

Differential expression of 837 exons was determined in 300 genes using the HTA 2.0 array. However, there was the same lack of consistency between these results and the differentially expressed exons observed in the RNA-seq dataset, with only 7 genes overlapping both datasets, with none of these DE at gene level in the microarray data. Again the same pattern emerged when these DE exons from the HTA 2.0 array were compared with CD associated genes, only 3 coeliac associated genes, *IL18R1, CCR2* and *LPP* contained significantly differentially expressed exons but were not DE at the gene level.

In order to perform an eQTL study, both gene expression and genotype information are required from the same individual, thus running in parallel to our transcriptome profiling study we also carried out genotyping on the majority of individuals in our study. Due to clinical constraints our study relied on sampling individuals attending the clinic, the majority of whom had not been previously genotyped. During the period of collection we managed to collect 25 patients with known genotypes leaving 35 patients for whom we had to generate this data. Controls also needed to be genotyped as anonymous bio bank controls were used for the Immunochip study. To this end, we had reserved several Immunochip arrays at the conclusion of our previous study. Unfortunately, this collection phase took much longer than anticipated, by which time the Immunochips were beyond their useable date. Having consulted with

colleagues experienced in running microarrays, we felt it was worthwhile assaying these samples using the Immunochips, however this failed. The only alternative to genotyping these patients using the Immunochip at the time was to genotype a select number of candidate SNPs via qPCR. However, this was not regarded as an efficient or indeed viable approach due to time constraints at this time. Using qPCR or preferably the Immunochip 2 when it is released (due Q3 2015) to genotype our coeliac cohort and thus perform our eQTL study was our preferred plan for the future.

4.6 Conclusion

In summary, the objectives of this study were to replicate our RNA-seq transcriptome analysis study in a larger population of 60 coeliac cases and 20 unaffected controls, to gather genotype information on these individuals and to combine both genotype data with gene expression to perform eQTL analysis. However, we found that the microarray platform chosen for this transcriptome analysis provided very limited and inconsistent gene expression information. It failed to support any extensive gene expression differences from our RNA-seq dataset and previously published studies. Running in parallel to transcriptome profiling, we carried out genotyping analysis on the remaining individuals with CD and controls to allow us to perform eQTL analysis, but this was not successful.

Therefore in addition to the missing genotype data, to complete this study would require a reliable gene expression dataset. However the cause of the conflicting results between the RNA-seq and the array has yet to be resolved. As discussed major concerns attend to the reliability of the array data and thus its utility in conducting an

eQTL study. Therefore future work would require genotyping these individuals, collecting more reliable and informative gene expression data and carrying out an eQTL study.

Chapter V

General Discussion

The aim of the research presented in this thesis is to further the understanding of the underlying genetic causes of Coeliac Disease (CD). The last decade has seen a substantial increase in our knowledge of the genetic variants that contribute to CD pathogenesis but the precise mechanisms by which they exert their effects largely remain to be elucidated. We wished to exploit recent advances in knowledge and technology to uncover new susceptibility loci, gene expression profiles, pathways and regulatory networks for CD.

The HLA locus on chromosome 6p21 was discovered to be involved in the pathogenesis of CD in the 1970s. It has since been recognized as the main genetic predisposing factor for CD. Up to 95% of patients with CD express the class II HLA-DQ2 heterodimer with the remainder expressing the HLA-DQ8 heterodimer. However in the region of 35-40% of the general population also carry these alleles, which implies that the HLA is necessary, but not sufficient for CD pathogenesis. A recent study has suggested that the HLA locus accounts for ~41% of the genetic risk of CD (Gutierrez-Achury et al., 2015).

Given the HLA locus was not solely responsible for susceptibility to CD, this encouraged researchers to look outside the HLA locus for additional CD susceptibility variants. Genome-wide linkage studies (GWLS), followed by genome-wide association studies (GWAS) lead to an explosion of CD genetic information which has both confirmed existing theory but more importantly, has added new detail and indeed previously unsuspected vistas of the molecular mechanisms involved. GWAS allow for the

unbiased identification of genes and pathways associated with traits and complex diseases. A small number of GWAS have been carried out in CD to date (van Heel et al., 2007; Hunt et al., 2008) culminating in a study performed by Dubois *et al* (2010), where they brought the total number of non-HLA CD susceptibility loci to 26.

However, most of the loci identified contain more than one associated SNP and due to low marker density of these early GWAS SNP arrays, identifying the exact causal SNP responsible for disease pathogenesis has remained difficult. The development of the high-density Immunochip platform in 2010, aimed to improve the discovery of new genetic regions and detailed assessment of existing regions in CD and across a range of autoimmune diseases.

In 2011, Trynka *et al* (2011) used the Immunochip to carry out the largest (in terms of population size) association analysis of CD to date. They identified 13 new CD risk loci bringing the total number of known CD loci to 40, Including the HLA locus, comprising 57 independent non-HLA signals. A subsequent study by Garner and colleagues (2014) confirmed a previous suggestive association within *FRMD48* as a CD locus bringing the total number of CD associated loci to 41. Thus, GWAS have been successful in identifying a large number of CD risk loci however replication of such findings in independent populations is regarded as a necessary validation step.

One aim of our study was to replicate the results from the Trynka *et al* (2011) study in an independent Irish case-control sample, which we successfully achieved. We found

evidence of replication for 11 out of 57 independent non-HLA SNPs identified by Trynka *et al*. This was almost 4 times greater than would be expected by chance at the $P<0.05$ level. We also showed highly significant evidence of concordance in direction of effect for the majority (77%) of SNPs that were GWS in the Trynka *et al* study. Importantly, our findings provide the first replication of the 12 loci newly identified in the study of Trynka *et al,* strengthening their designation as true coeliac susceptibility loci.

Given the opportunity that combining our data with the previous study would contribute sufficient power to identify new genome-wide significant variants, a meta-analysis was performed. This resulted in the elevation of two additional loci beyond the GWS threshold, *NFIA* and *ZNF335,* and brought the total number of confirmed CD loci to 43, including the HLA locus.

Of itself our study was underpowered to detect any SNPs reaching established levels of GWS. However, given that the ability to identify causal variants is highly dependent on population size, odds ratios and allele frequencies, which are the major components of power, it is inevitable that a proportion of the causal genetic burden will continue to be missed. Therefore, given the practical limitations of amassing very large numbers of samples, new methods to model genetic risk without the constraints of reaching GWS have been developed. Polygenic Risk Score (PRS) analysis is one such method, and this allowed us to combine the effects of non-GWS, non-HLA susceptibility variants as well as the known HLA variants to determine whether a significant proportion of

variance in our CD cohort could be explained by non-GWS SNPs. We could use PRS to compare our data with that of Trynka *et al.,* to determine if the genetic architecture was similar in the two datasets despite the differences in size. We estimated that up to 35% of the genetic risk could be explained when all markers are included (HLA and non-HLA). This is comparable with previous observations where Abraham and colleagues (2014), explained 30-35% of the genetic burden in CD and three other complex diseases. When this is limited to non-HLA loci, we explained a maximum of 4.5% of the genetic variance, which is not dissimilar to observations in schizophrenia and multiple sclerosis where the non-HLA loci account for approximately 3% of the variance in disease risk (Purcell et al., 2009; The International Multiple Sclerosis Genetics Consortium (IMSGC), 2010).

In our study, most of the genetic burden of CD was explained by SNPs at a *P*-value threshold of ≤0.05. This implies that current approaches of only considering SNPs that reach the GWS threshold of $P<5x10^{-8}$ as being relevant to disease means that a number of true disease associated SNPs would be missed simply because they did not meet this threshold.

The PRS and similar approaches have the practical benefit of improving the accuracy of CD risk prediction and could be used in a clinical setting to genetically diagnose CD (Romanos et al., 2014). Despite improvements and the widespread availability of serological and histological testing, there are still a large number of individuals with CD that remain undiagnosed due to practical limitations in testing. Therefore genetic

testing, more specifically HLA-typing, is now currently also being used to diagnose CD in the clinic. Currently Individuals with a 'known' or a suspected higher risk of CD are screened using periodic serological testing and genetic testing (HLA-DQ2 and HLA-DQ8 typing) according to ESPGHAN recommendations. However, HLA-DQ2 and HLA-DQ8 have a poor positive predictive value although they are useful at identifying individuals with very low risk of CD.

Using the PRS approach, we were able to generate a genomic risk score including SNPs that failed to reach GWS, which incrementally improved the accuracy of CD status prediction compared to scores based on the HLA only. Our data showed the highest predictive power of 0.83 (AUC) at a SNP threshold $P<0.05$ when the HLA was included compared to 0.82 for the HLA alone (at *P*-value threshold of <0.5) and 0.67 (SNPs threshold $P<0.01$) using the non-HLA markers only. This confirms what we already know, that is, that the risk of developing CD is dominated by the HLA. However, the observation that when the HLA is excluded we have predictive power of 0.67 (AUC) may be clinically useful in discriminating between carriers of HLA risk alleles.

Despite our small sample size we obtained results comparable with Romanos *et al* (2014) who, using a much larger sample size of >2000 cases, showed a predictive power of AUC=0.854 and Abraham and colleagues (2014) who showed an AUC=0.86-0.9 when combining both the HLA and non-HLA susceptibility variants. Speed *et al* (2014) utilised a different genomic risk score method and showed a predictive power

of AUC=0.79 in >15,000 coeliac cases. The Framingham risk score for the prediction of coronary heart disease (Wilson et al., 1998) is considered a benchmark, and they have proven a predictive power of AUC=~0.8 to be clinically useful. Therefore, there is potential for our results to also be clinically beneficial as discussed above, especially since the Immunochip platform is very cost effective even compared to HLA genotype testing. As discussed, separating the HLA and non-HLA components may be of additional benefit in individuals considered to have a higher risk of developing the disease.

In terms of the polygenicity of CD susceptibility, our results suggest that there are possibly many tens or even hundreds of risk variants with small effects remaining to be identified. However PRS are limited in that only narrow sense heritability (i.e. a purely additive model of association) is examined. This method does not take into account the proportion of variation that is attributable to broad sense heritability, i.e. additive, non-additive, multiplicative and epistatic interactions between genetic variants. Epistasis is when the effect of one gene is dependent on the presence of one or more other genes or variants and it is believed to be a component of the heritability of complex traits including diseases. Research into epistatic interactions has been difficult due to computational and statistical challenges originating from the large number of variables that need to be considered for this analysis. However, studies have found strong evidence of epistasis contributing to disease risk in a number of complex autoimmune diseases such as psoriasis (Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, 2010), multiple sclerosis

(Lincoln et al., 2009) and type 1 diabetes (Barrett et al., 2009). More recent studies have shown evidence for epistasis in CD (Goudey et al., 2014). Goudey *et al* show an epistatic effect from the well-established HLA risk haplotypes for CD as well as evidence of epistasis between HLA class III genes. They also show that models using only epistatic pairs explained almost as much CD variance as the additive models alone. Notwithstanding the fact that much of this epistatic variance overlaps with the additive variance, this result supports the idea that combining both additive and epistatic models will better explain the heritability than using either model alone.

On the other hand, there is considerable debate currently as to whether the degree of missing heritability is in fact over-estimated. This has recently been shown for phenotypic traits such as height and BMI in a study by Yang and colleagues (Yang et al., 2015) and may imply that previous coeliac studies have explained much more of the total heritability then originally thought.

In any event, GWAS has succesfully identified 43 CD risk loci which harbour over 70 candidate genes. Interpretation of these genetic associations is one of the biggest challenges that now face researchers. As mentioned previously, approximately 90% of the CD risk loci identified map to non-coding regions, thus suggesting that they exert their effect by regulating the expression of specific genes rather than changing the nature of the protein coding regions. Recent studies have also highlighted the importance of cell and stimulus specificity in regulating gene expression. Therefore in

order to investigate this further, we examined gene expression in $CD4^+$ T-cells, a cell type crucial for the pathogenesis of CD, using three independent treatments.

We show that a substantial number of genes are differentially expressed (DE) between individuals with CD and controls. These include a large number of genetically linked genes and also genes known to be functionally involved in the coeliac lesion, for example *IL21, IL15R* and *IL23R*.

Interestingly *IFNG* is the most significantly up regulated gene, displaying a ~25 fold increase in expression in cases compared to controls. This up regulation correlates well with the high IFNγ production previously established in the coeliac lesion. Furthermore, WGCNA network analysis identified several modules of co-expressed genes tightly associated with coeliac disease. *IFNG* is a key hub gene within one of the most highly correlated modules (M3). M3 was the largest module identified and contained a large number of DE genes, indicating that these were co-expressed with IFNγ in $CD4^+$ T-cells. This indicates that these expression networks likely collectively contribute to the disease state and contain a number of genes that promote a more aggressive immune response. These findings are in agreement with Kumar et al (2014) who found evidence of co-expression of genetically associated coeliac genes with IFNγ, a subset of which were also shown to be controlled by IFNγ in monocytes by Fairfax et al (2014).

This all points to the centrality of *IFNG* expression in CD and is consistent with the well-established high production of IFNγ in the coeliac lesion, where it plays a crucial role promoting the Th1 response characteristic of CD (Olaussen et al., 2002). However, there has been no genetic association between the *IFNG* locus and CD, implying that IFNγ expression is regulated indirectly by upstream genetic susceptibility variants. Our gene expression results have led to a plausible explanation for the increased IFNγ expression levels seen in CD – the corresponding up regulation of *IL18RAP*. IL18RAP is the signaling component of the IL18 receptor, which is composed of both IL18RAP and IL18Rα. IL18Rα is encoded by the IL18R1 gene, which is also up regulated in our data. This up regulation of the IL18 receptor is likely to facilitate IL18 signaling leading to increased IFNγ expression.

IL18 acts synergistically with IL12 in inducing IFNγ production in Th1 cells, and these interleukins reciprocally up regulate each other's receptors (Xu et al., 1998). We also see evidence of this in our RNA-seq data, with both *IL12RB2* and *IL12RB1* significantly overexpressed in anti-CD3/CD28 stimulated patient cells. Thus the upstream ingredients for increased *IFNG* expression are present in our anti-CD3/CD28 stimulated $CD4^+$ T-cells.

This up regulation of *IL18RAP* is interesting as it appears to conflict with observations made by Dubois *et al* (2010) who report a very strong *IL18RAP* CD associated eQTL which correlates with significant down regulation of *IL18RAP*. However, the Dubois

study was performed in un-stimulated peripheral blood leukocytes and the eQTL may perform differently in resting cells compared to stimulated cells as mentioned earlier.

Network analysis identified M6 as the most highly correlated module with CD, highlighting the previously identified coeliac susceptibility gene *BACH2* as a key hub gene. Recent studies have shown that BACH2 acts as a transcription factor promoting T-reg formation and acting as a negative regulator of T-effector cell (Th1, -2, -17) differentiation, thereby preventing inflammation (Kim et al., 2014). In our study, we found down regulation of *BACH2* in coeliacs compared to controls - which is in this instance consistent with eQTL studies conducted by Dubois *et al.*, who showed that CD risk alleles at *BACH2* are associated with decreased expression (Dubois et al., 2010). The down regulation of *BACH2* seen in individuals with CD is what one would expect given BACH2's key role in maintaining immune homeostasis. Deficiency of BACH2 could inhibit T-reg and promote excessive T-effector formation, leading to a prolonged inflammatory response and with the potential for the intestinal damage characteristic of the coeliac lesion. Polymorphisms in *BACH2* have also been found in numerous other inflammatory diseases, including Crohn's disease (Franke et al., 2008), multiple sclerosis (International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2013), ankylosing spondylitis (International Genetics of Ankylosing Spondylitis Consortium, 2013) and type 1 diabetes (Onengut-Gumuscu et al., 2015). Ten different SNP associations have been identified to date, with a single SNP (rs72928038) found to be associated in multiple sclerosis, type 1 diabetes and ankylosing spondylitis. The remaining SNPs were unique to each different condition, with each SNP potentially

identifying different genetic regulatory mechanisms in each individual disease relevant cell type. Despite the identification of various BACH2 associated SNPs and the prospect of each SNP exerting different functions, there is one constant and that is, that BACH2 is involved in all of these diseases and is very likely playing an important role in their pathogenesis.

A recent study by Vahedi *et al* (2015) has shown BACH2 exerts its effect by binding to super enhancers and controlling genes crucial for T-cell biology (it also plays an important role in the function of B-cells). For this reason BACH2 is known as a 'guardian' transcription factor. Super enhancers are powerful genetic switches that control genes important for cell identity. In T-cells, BACH2 bound super enhancers disproportionately regulate the transcription of genes encoding cytokines and cytokine receptors, which are vital for the control of cell state and management of the immune response. By extension, any disruption in super enhancer function could seriously disrupt the immune response, potentially leading to a wide range of inflammatory diseases. It has been shown that there is a great enrichment of disease-associated SNPs found in super enhancers of disease-relevant cells, for example type 1 diabetes (Hnisz et al., 2013), rheumatoid arthritis (RA) (Vahedi et al., 2015) and systemic lupus erythematosus (SLE) (Hnisz et al., 2013). This finding is important, as it suggests that disease associated genetic variants work in part by disrupting super enhancer function and that the subsequent altered gene expression is important in development of disease.

We show for the first time, a strong signature of BACH2 regulated genes differentially expressed in CD. Genes previously shown to be repressed by BACH2 were up regulated in individuals with CD and the opposite was true for BACH2 activated genes (Vahedi et al., 2015). Interestingly, *IL18RAP* is one of the genes repressed by BACH2 - as is *IFNG*, indicating that BACH2 may be a nexus for CD associated risk factors. Considering that *IL18RAP* is both genetically associated and controlled by another disease-associated locus (i.e. *BACH2*) would seem to indicate that it contributes importantly to pathogenesis.

The potential role for BACH2 in the pathogenesis of CD is further illustrated by the discovery that *TGFβR2* is also down regulated in our coeliacs compared to our controls. Co-expression network analysis has identified *TGFβR2* as a hub gene in the same network as *BACH2* implying that they are correlated with one another and indeed Vahedi *et al* show *TGFβR2* is upregulated by BACH2. TGFβ signals through its receptor TGFβR2 leading to the differentiation of T-regs and regulation of the immune response, thus TGFβR2 plays a vital role in this process. Therefore reduced expression of TGFβR2 in CD caused by BACH2 deficiency could potentially disrupt T-reg differentiation and immune suppressor activity. Taken together, these observations support the idea that BACH2-orchestrated, decreased immune suppression via T-reg / T-effector cell homeostasis, is an central mechanism in CD pathogenesis.

The majority of the coeliac associated loci identified contain multiple variants, thus determining the exact causal SNP or indeed SNPs responsible for disease pathogenesis

within them remains difficult. One way of filtering out causal SNPs from these loci is to prioritise SNPs that have a functional role. Hence, in order to determine if the disease associated variants were responsible for altered gene expression we wanted to combine the gene expression and genetic association data, in a process known as eQTL analysis. A small number of eQTL studies have been carried out in CD to date (Dubois et al., 2010) with results indicating that CD-associated SNPs are greatly enriched for *cis*-eQTLs. However, as mentioned, these were carried out in (unstimulated) whole blood samples and whilst they identified a high proportion of disease associated SNPs as affecting gene expression in this scenario, subsequent studies have unsurprisingly shown that many *cis*-eQTLs act in a cell and stimulus specific manner (Fu et al., 2012; Fairfax et al., 2014; Plaza-Izurieta et al., 2014). This highlights the importance of performing eQTL analysis in the cellular and environmental context of interest.

In order to have sufficient power to detect eQTLs, we needed to expand our gene expression study into a larger dataset. We attempted to do this using the Affymetrix GeneChip® Human Transcriptome Array (HTA) 2.0. We also used the Immunochip to gather genotype data on our sample population. However, due to the failure of the Immunochip this part of the project was unable to be completed successfully at this time.

Another method of filtering out causal variants is to examine the epigenome of disease relevant cells. Functional areas of the genome such as promoters, enhancers and silencers are marked with distinct epigenetic modifications, particularly DNA and

histone modifications and nucleosome positioning. These control genetic activation and repression by regulating access for DNA binding proteins, including transcription factors, which are necessary for controlling transcription, most likely in a highly cell specific manner. As the vast majority of disease associated variants lie in non-coding regions, they presumably reside in regulatory areas of the genome, and epigenetic mapping could be used to identify these important functional regions in T-cells. SNPs lying in these regions we consider are much more likely to play a biological role than those that don't.

Much of this data is publicly available through databases such as ENCODE, Epigenomics Roadmap, HaploReg and RegulomeDB that incorporate extensive regulatory information and epigenetic maps of chromatin states across numerous cell types and disease states. However an epigenomic map of $CD4^+$ T-cells in CD is not currently available. Therefore mapping chromatin marks in this cell type under physiological stimuli would be a beneficial step towards the fine mapping of causal variants in CD. A recent study by Seumois *et al* (Seumois et al., 2014) in asthma, profiled active and poised enhancers in CD4+ T-cells, successfully identifying a large number of cell type and disease specific enhancers.

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a method that has been developed to map the binding sites of epigenetic modifications, transcription factors and chromatin-associated proteins in a variety of organisms. As part of my project, we initially planned to map poised and active enhancer elements in both

CD3/CD28 stimulated and un-stimulated $CD4^+$ T-cells from coeliac patients and unaffected controls to characterise these regulatory elements and explore the chromatin state in these two groups. These epigenomic marks could then be layered over CD associated SNP information to provide an in depth look at the location of these SNPs and their regulatory/functional potential. However, due to time restraints this part of the project could not be completed.

ChIP-seq can also map transcription factor binding sites. The discovery that BACH2 is a transcription factor that binds to super-enhancers important in CD indicates it would be interesting to map disparities in BACH2 binding in cases and controls under various conditions.

The motif of epigenetic differences being important is supported by the finding of the *ZNF335* gene as a bona fide CD susceptibility gene. It forms part of a chromatin-remodeling complex with HCFC1, another transcriptional co-regulator which is encoded in a CD susceptibility locus on Xq28 (Trynka et al., 2011). Given that ZNF335 is part of an epigenetic remodeling complex, by definition any modulation of its expression in disease could have a profound impact on the transcriptome. Therefore to better understand how this chromatin remodeling protein may contribute to CD pathogenesis, mapping ZNF335's binding regions could be carried out via ChIP-seq. Other interacting partners that bind to this complex could also be identified in this manner facilitating our understanding of the role this chromatin-remodeling complex

plays in CD. Mapping of chromatin-remodeling complexes using ChIP-seq has been carried out successfully in other diseases such as cancer (Euskirchen et al., 2011).

In summary, we have found that the genetic architecture of CD in the Irish population is highly similar to larger European populations. We have successfully identified two new CD genetic loci, *NFIA* and *ZNF335*, bringing the total number of coeliac risk loci to 43 including the HLA locus as well as highlighting the importance non-genome wide significant loci contribute to disease pathogenesis. *ZNF335* forms part of a chromatin remodeling complex essential for gene regulation, thus any disruption in its expression might have a severe impact on transcription, a mechanism that could be explored in further studies. NFIA is also interesting in so far as it is a transcription regulator about which little is known in the context of the immune system. Transcriptome analysis identified numerous gene expression differences between individuals with CD and controls, with *IFNG* showing the most significant up regulation of gene expression validating existing knowledge and further substantiating its role in CD pathogenesis. We also identified a novel finding of decreased expression of the genetically implicated *BACH2* gene in individuals with CD. BACH2 has a demonstrated profound affect on the transcriptome where its deficiency leads to an abnormal repression of T-regs and expansion of T-effector cells with a much more destructive phenotype. This transcriptomic signature is seen in CD but could also be an important mechanism of immune regulation in numerous other inflammatory diseases where BACH2 polymorphisms have also been identified. Our study also identified a subgroup of coeliac individuals that were transcriptomically (and perhaps genetically) distinctive,

further highlighting the complexity of CD, however the implications of which are not currently known.

As well as making a considerable contribution to our understanding of CD pathogenesis, the observations made in this study also highlight interesting avenues of further research that could provide further insights into the molecular mechanisms underlying CD and which may pave the way for potential diagnostic or therapeutic interventions in this disease.

Appendices

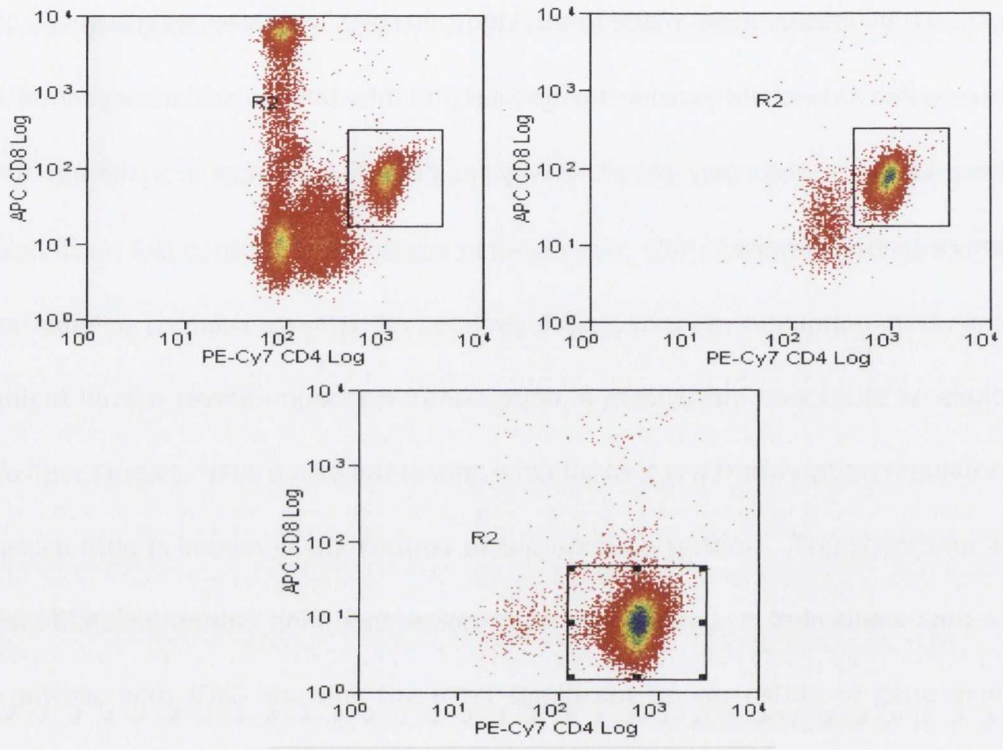
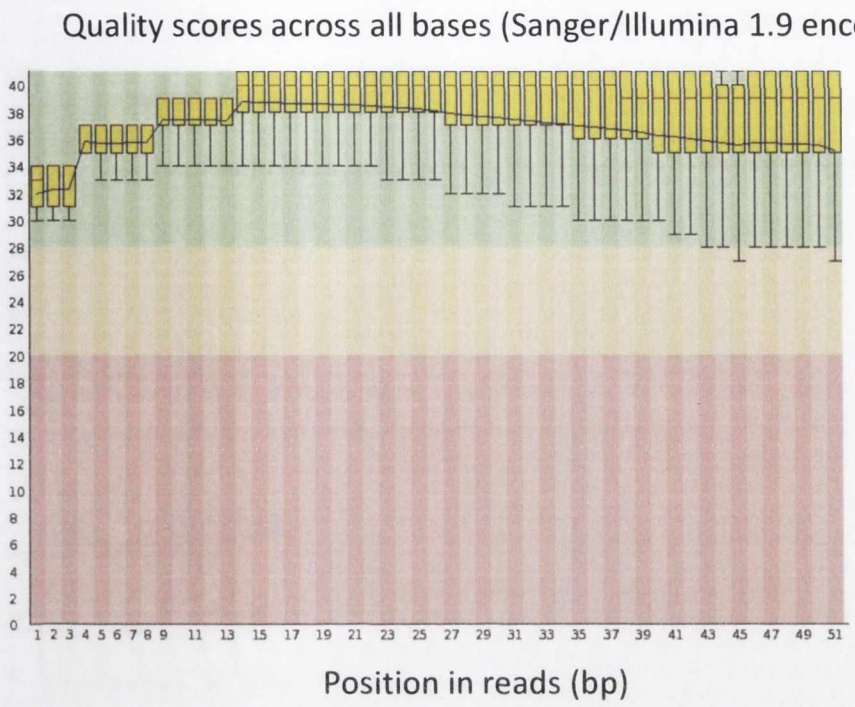


Figure A: Gating strategy for sorting CD4+ T-cells from PBMCs. Top left: Pre-enrichment (APC-CD8$^+$ vs PE-Cy7-CD4$^+$). Top right: Post-enrichment (APC-CD8$^+$ vs PE-Cy7-CD4$^+$). Bottom centre: Post culture (APC-CD8$^+$ vs PE-Cy7-CD4$^+$). R2 – CD4$^+$ T-cell gate. Post culture analysis showed >97% in all cases.



Figure

B: Per-base sequence quality. Individual bases for the reads are mapped along the X-axis. The Y-axis on the graph represents the quality score for every base call. Values above 20 are considered to be of good quality.

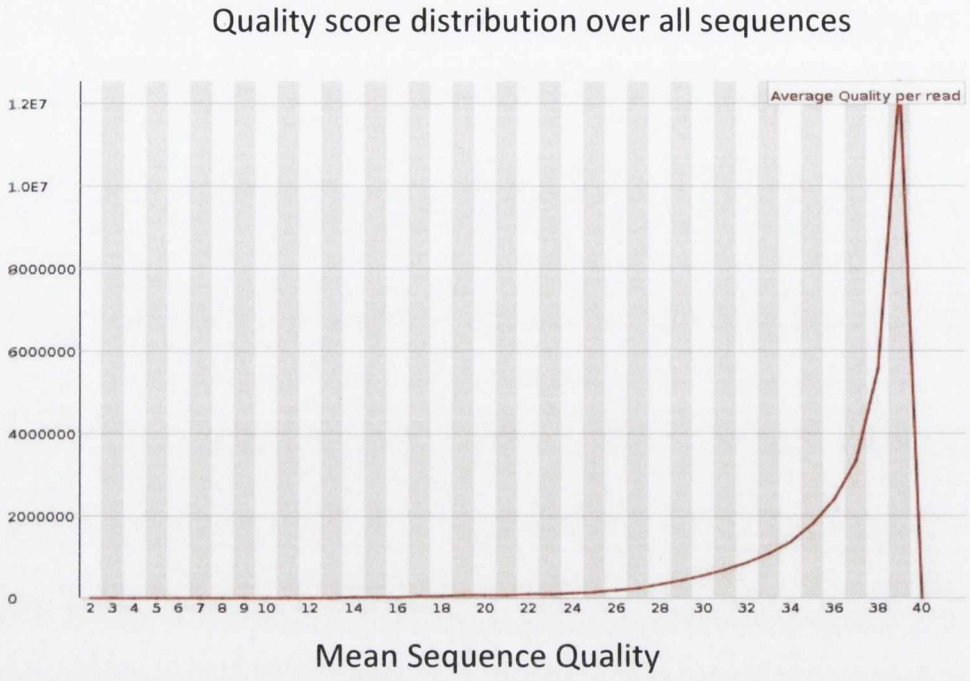


Figure C: Per-base sequence quality scores. Taking every sequence into account, the mean score across all the bases in a particular sequence is calculated and the distribution of these means is plotted on a graph.

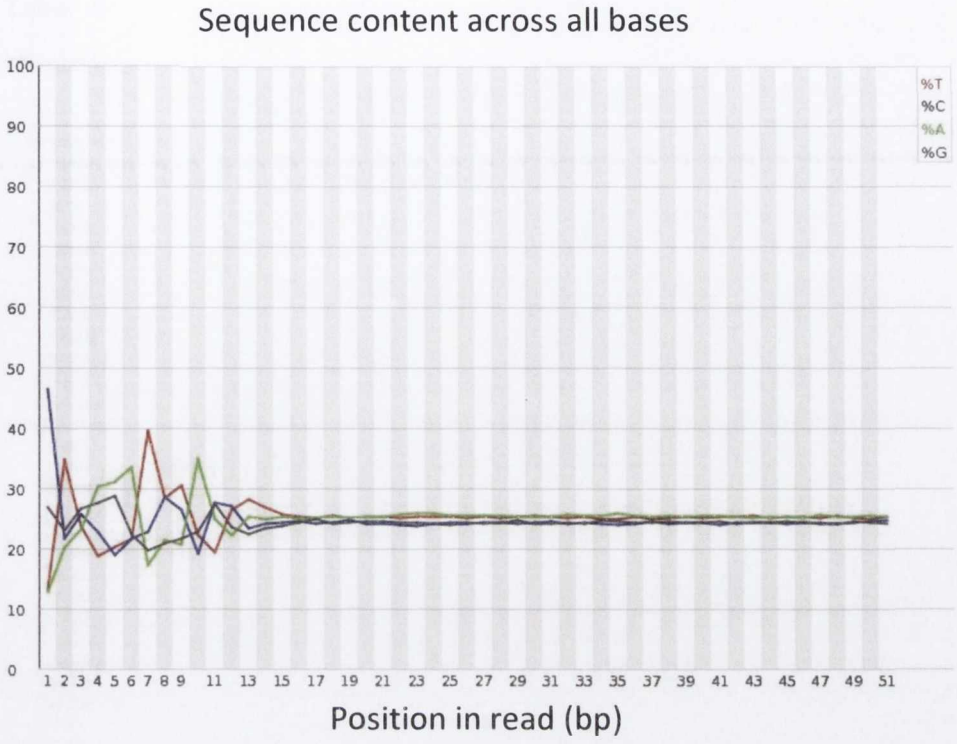


Figure D: Per-base sequence content. In a random library from an RNA-seq sample all four bases (G, A, T, C) should be evenly distributed across a sequence run; all lines should run parallel with each other. The nature of primers for RNA-seq generates an uneven distribution across the first 10 bases that is very common and normally observed with this type of data.

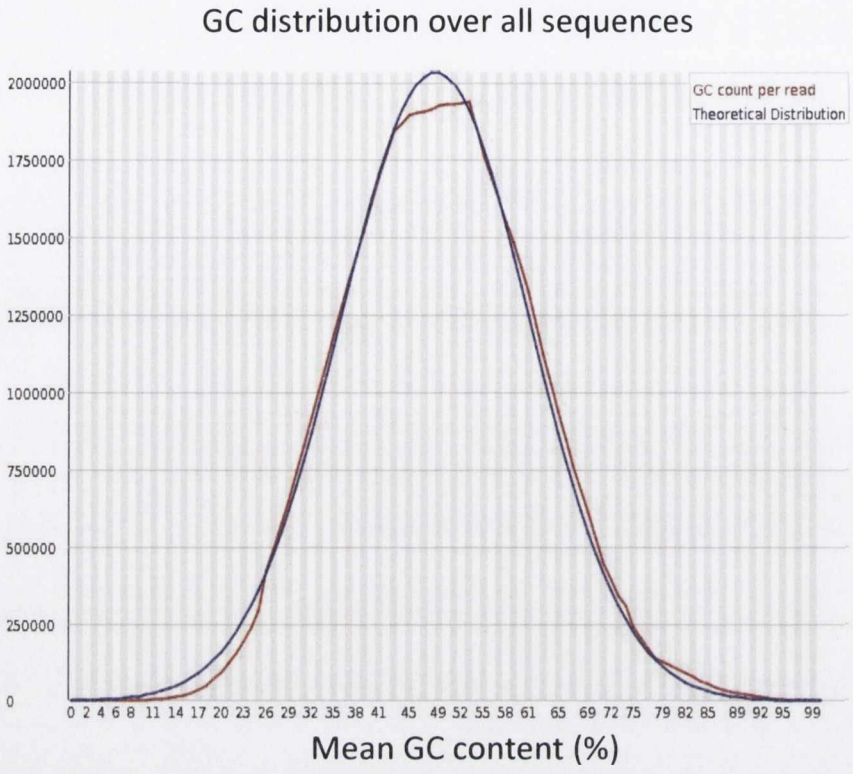


Figure E: Per sequence GC content. There should be a normal distribution between the data sample (red) and the theoretical normal distribution (blue). The normal distribution has the same mean and SD as the sample library.

Table A: The top 25 most significantly differentially regulated genes in the PMA stimulated group.

Up-regulated				Down-regulated			
Gene	basemean	log2foldchange	padj	Gene	basemean	log2foldchange	padj
P2RX7	527.3345	3.446152	8.15E-10	RASD2	25.59627	-2.843043	0.0000165
SETBP1	51.68241	4.413535	8.15E-10	RIPK4	31.92967	-1.917989	0.000041
HOPX	956.4151	2.978577	5.76E-09	DEPDC7	60.61403	-1.63975	0.0002045
AOAH	254.2935	3.30808	2.45E-08	STX1B	169.2576	-2.138324	0.0002225
ATP9A	140.9889	2.724517	3.19E-08	PPP2R2D	3555.327	-1.312261	0.0003299
GBP4	15153.23	3.033737	3.56E-08	BNIP3	1272.18	-1.210514	0.0003393
PSMB9	2981.985	2.413836	6.71E-08	CLN8	1491.871	-1.413576	0.00036
EPSTI1	887.6121	2.734552	9.56E-08	LOC729799	58.00986	-1.083702	0.00036
TNFSF13B	475.6063	2.774429	9.56E-08	SLC35E1	5236.029	-1.050966	0.0003715
STAMBPL1	467.122	2.775907	2.22E-07	IGF2BP3	22.48832	-1.122258	0.0004094
LOC441869	71.57745	2.575777	2.70E-07	HEY2	12.21415	-1.678984	0.0004481
LOC154761	644.9479	2.515748	2.71E-07	HLA-DQB1	920.9199	-1.96188	0.0004879
SNX33	88.53368	2.931861	2.71E-07	RAB3B	23.59017	-1.145977	0.0005315
ETV7	265.8132	3.452858	2.71E-07	SULF2	30.1859	-1.859356	0.0006621
CDCA7L	414.7128	3.39702	2.83E-07	RMRP	38.49173	-1.489556	0.0006928
TRIM69	621.5307	2.672154	3.18E-07	ABHD4	1164.469	-1.042756	0.0007342
ADAP1	424.1365	3.128896	3.69E-07	TNFRSF10B	4344.449	-1.06839	0.0007762
GBP5	39240.18	3.165066	5.93E-07	LYPD3	245.9689	-2.126681	0.0008336
NCALD	90.59918	2.713588	6.61E-07	IFNGR2	3106.584	-1.551813	0.0009807
SECTM1	463.5337	3.252774	6.61E-07	HOOK1	1612.078	-1.573512	0.0011531
TEC	61.19174	1.843125	8.78E-07	ARRDC3	6612.329	-1.666185	0.0012223
GALM	777.549	2.699599	8.78E-07	FLJ90757	131.7793	-1.399132	0.0012529
DLGAP1	21.24908	3.244879	8.78E-07	WNT9A	102.0716	-2.209025	0.0013017
TMEM14A	143.2035	3.559472	8.95E-07	TNFRSF10D	1047.654	-1.768244	0.0013179
HCAR1	150.9922	3.016419	1.13E-06	ISCA1	2969.435	-1.083854	0.0013734

Table B: The top 25 most significantly differentially up regulated genes and the 24 down regulated genes in the unstimulated group.

Up-regulated				Down-regulated			
Gene	basemean	log2foldchange	padj	Gene	basemean	log2foldchange	padj
GRAMD1C	91.59019	2.417436	0.0000845	RASL10B	75.81314	-2.63022	7.43E-06
CEBPD	97.06924	2.368168	0.0000879	PTPRK	115.2312	-1.876058	0.0000845
GZMA	517.1524	2.430692	0.0011236	PFKFB3	8225.38	-1.14592	0.0114833
LGALS3	3425.05	1.558282	0.0014231	MMP28	265.6923	-1.7095	0.0155153
ADAT3	75.20262	1.634881	0.001455	RLN1	33.83278	-1.362726	0.019296
SLC39A10	1969.988	1.458406	0.0036016	ADAMTSL2	30.33822	-1.960916	0.0227195
LOC154092	21.61805	2.194709	0.0036339	CSPG5	14.3946	-1.746099	0.0269038
TRIM69	123.9756	1.376469	0.0087631	KIAA1274	17.20379	-2.542871	0.0298966
HIST1H2AC	76.16026	1.260124	0.0095843	PTK2	725.425	-1.2058	0.0298966
SLAMF7	1981.226	2.010164	0.0109331	LOC100506388	35.42038	-2.193511	0.0310184
FCER2	37.96173	2.532108	0.0109331	LOC100272216	37.4335	-2.121348	0.0310184
CLEC2B	2286.742	1.14296	0.0114833	TGFA	532.8828	-1.677551	0.0310184
KCNK5	34.728	1.412526	0.0114833	TNFAIP8L1	1437.756	-1.066492	0.0310184
LYAR	2077.492	1.540474	0.0114833	FBLN7	2176.627	-1.241018	0.0316919
RNF5P1	50.71958	2.600693	0.0114833	EDA2R	58.65601	-1.732643	0.0332526
CDCA8	83.09666	1.167932	0.0169726	ZNF853	316.1896	-1.039771	0.0347738
NPC1	3287.46	1.391483	0.0169726	TMIGD2	903.1075	-1.115335	0.0353829
FAM110C	144.0647	1.632273	0.0169726	C5orf20	22.06011	-2.033616	0.0358681
RDH10	337.7897	1.06645	0.0176976	SMAD7	2194.718	-1.20758	0.0358681
TIMP1	2015.612	1.03961	0.019296	MAN1C1	2687.8	-1.096694	0.0360642
SPINK1	77.83007	2.31086	0.019296	CDK5R1	1965.004	-1.488351	0.0382462
C4orf48	92.6394	1.30665	0.0202839	FMNL3	3381.604	-1.027261	0.0386665
CST7	1287.665	2.012467	0.021675	DACH1	11.63398	-1.765343	0.0445356
MELK	23.31125	1.643205	0.0227195	ABCC4	1015.862	-1.03497	0.0445356
ZNF296	57.03985	1.777231	0.0227195				

Table C: 98 genes regulated by BACH2 were DE between coeliac cases and controls in our dataset.

Gene	Expression pattern	Gene	Expression pattern	Gene	Expression pattern
CD101	Bach2KO_Down	TSPAN2	Bach2KO_Up	FES	Bach2KO_Up
MAML3	Bach2KO_Down	DAB2	Bach2KO_Up	HBEGF	Bach2KO_Up
ID3	Bach2KO_Down	FAM102A	Bach2KO_Up	TNFRSF18	Bach2KO_Up
ANKRD55	Bach2KO_Down	GEM	Bach2KO_Up	MFSD2A	Bach2KO_Up
BCL9L	Bach2KO_Down	JUN	Bach2KO_Up	DDIT4	Bach2KO_Up
BACH2	Bach2KO_Down	IRF2BP2	Bach2KO_Up	ZEB2	Bach2KO_Up
YPEL2	Bach2KO_Down	IL15RA	Bach2KO_Up	PDCD1LG2	Bach2KO_Up
NT5E	Bach2KO_Down	TBX21	Bach2KO_Up	BATF3	Bach2KO_Up
FAIM3	Bach2KO_Down	SLCO4A1	Bach2KO_Up	TFRC	Bach2KO_Up
ITPKB	Bach2KO_Down	KCNK5	Bach2KO_Up	IL2RA	Bach2KO_Up
SMAD7	Bach2KO_Down	PLAGL2	Bach2KO_Up	BCL2L1	Bach2KO_Up
IL16	Bach2KO_Down	CTH	Bach2KO_Up	TNFRSF4	Bach2KO_Up
GIMAP1	Bach2KO_Down	PRDM1	Bach2KO_Up	SLAMF7	Bach2KO_Up
SBK1	Bach2KO_Down	LRRC32	Bach2KO_Up	IL18RAP	Bach2KO_Up
SKIL	Bach2KO_Down	RGS1	Bach2KO_Up	ACOT7	Bach2KO_Up
TBC1D16	Bach2KO_Down	CCND2	Bach2KO_Up	NFIL3	Bach2KO_Up
ACVR2A	Bach2KO_Down	BHLHE40	Bach2KO_Up	FRMD4B	Bach2KO_Up
TGFBR2	Bach2KO_Down	LMNB1	Bach2KO_Up	CCL1	Bach2KO_Up
CD27	Bach2KO_Down	BATF	Bach2KO_Up	IL3	Bach2KO_Up
MYB	Bach2KO_Down	SOCS1	Bach2KO_Up	IL12RB2	Bach2KO_Up
LEF1	Bach2KO_Down	ADAM8	Bach2KO_Up	AURKA	Bach2KO_Up
FGD3	Bach2KO_Down	ASB2	Bach2KO_Up	MCOLN2	Bach2KO_Up
TMEM66	Bach2KO_Down	MAP3K8	Bach2KO_Up	IL1RL1	Bach2KO_Up
IZUMO4	Bach2KO_Down	GPR171	Bach2KO_Up	CSF2	Bach2KO_Up
MAF	Bach2KO_Down	SLC2A1	Bach2KO_Up	LTA	Bach2KO_Up
PTS	Bach2KO_Down	TNF	Bach2KO_Up	IL4	Bach2KO_Up
NIPAL1	Bach2KO_Down	NME1	Bach2KO_Up	OSM	Bach2KO_Up
ZNRF1	Bach2KO_Down	CISH	Bach2KO_Up	TNFRSF8	Bach2KO_Up
QPRT	Bach2KO_Down	NAMPT	Bach2KO_Up	CCL4	Bach2KO_Up
CCR1	Bach2KO_Down	HK2	Bach2KO_Up	ZBTB32	Bach2KO_Up
G0S2	Bach2KO_Down	CD274	Bach2KO_Up	LIF	Bach2KO_Up
NEURL3	Bach2KO_Down	DUSP6	Bach2KO_Up	CCL3	Bach2KO_Up
		IFNG	Bach2KO_Up	GZMB	Bach2KO_Up

Table D: List of our BACH2 regulated genes and their ranking in the Vahedi *et al* dataset.

NAME	PROBE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	CORE ENRICHMENT
row_0	BATF3	6	1.172385931	0.0153442	Yes
row_1	TNFRSF4	7	1.172268271	0.03094624	Yes
row_2	LTA	13	1.132090926	0.04579737	Yes
row_3	ZBTB32	14	1.116472721	0.06065681	Yes
row_4	SOCS1	19	1.079076409	0.0748456	Yes
row_5	TNF	25	1.04335773	0.08851576	Yes
row_6	TNFRSF8	30	1.01940763	0.1019104	Yes
row_7	GZMB	32	1.005106688	0.1152444	Yes
row_8	CSF2	33	1.004811287	0.1286177	Yes
row_9	MAF	36	0.996374607	0.14179227	Yes
row_10	IL2RA	37	0.996121883	0.15504992	Yes
row_11	TFRC	39	0.986865938	0.16814116	Yes
row_12	TNFRSF18	44	0.983739197	0.18106107	Yes
row_13	IL15RA	49	0.978856504	0.19391601	Yes
row_14	IL18RAP	51	0.97518307	0.20685175	Yes
row_15	OSM	64	0.952778876	0.21901374	Yes
row_16	AURKA	78	0.922769129	0.23073308	Yes
row_17	IFNG	82	0.916872442	0.24280629	Yes
row_18	IL4	111	0.882847488	0.2533458	Yes
row_19	IL12RB2	136	0.863319278	0.26379833	Yes
row_20	MCOLN2	147	0.856616974	0.27476692	Yes
row_21	BATF	167	0.844063401	0.28517935	Yes
row_22	CCND2	213	0.821396351	0.29416603	Yes
row_23	HK2	219	0.818353951	0.30484155	Yes
row_24	MFSD2A	239	0.807636797	0.31476915	Yes
row_25	CCL4	258	0.794753492	0.32456854	Yes
row_26	NEURL3	276	0.786242306	0.33429787	Yes
row_27	FES	321	0.760252953	0.342514	Yes
row_28	ACOT7	344	0.751672089	0.35156706	Yes
row_29	NFIL3	351	0.74856627	0.36127052	Yes
row_30	BCL2L1	377	0.739208996	0.37002802	Yes
row_31	FRMD4B	379	0.737847626	0.379805	Yes
row_32	DUSP6	385	0.734828591	0.38936886	Yes
row_33	QPRT	387	0.734093904	0.3990959	Yes

row_34	ZEB2	427	0.72307313	0.40703332	Yes
row_35	LMNB1	438	0.72055441	0.41619104	Yes
row_36	DDIT4	467	0.71047604	0.42443642	Yes
row_37	ZNRF1	471	0.70892632	0.43374202	Yes
row_38	CTH	481	0.70655334	0.44275662	Yes
row_39	CCL3	496	0.70365804	0.4515165	Yes
row_40	PTS	515	0.69607264	0.46000254	Yes
row_41	SLCO4A1	525	0.6924454	0.46882936	Yes
row_42	CD274	537	0.6900546	0.47753793	Yes
row_43	CISH	553	0.68534786	0.4860109	Yes
row_44	SLAMF7	571	0.68022025	0.49432915	Yes
row_45	PLAGL2	574	0.67978311	0.5032901	Yes
row_46	IZUMO4	579	0.67801249	0.51214105	Yes
row_47	SLC2A1	604	0.67291594	0.5200594	Yes
row_48	NME1	695	0.64469427	0.5247488	Yes
row_49	MAP3K8	778	0.62704742	0.5295492	Yes
row_50	G0S2	797	0.62338793	0.5370678	Yes
row_51	LIF	881	0.60592943	0.54154384	Yes
row_52	LRRC32	882	0.6058113	0.54960674	Yes
row_53	ADAM8	906	0.60109198	0.5566125	Yes
row_54	ASB2	1254	0.54636508	0.548882	Yes
row_55	HBEGF	1279	0.54276532	0.55506825	Yes
row_56	PRDM1	1443	0.51711673	0.55490357	Yes
row_57	IL3	1460	0.51413477	0.5610546	Yes
row_58	PDCD1LG2	1520	0.50666052	0.56524706	Yes
row_59	TBX21	1814	0.47184321	0.5588594	No
row_60	NAMPT	1853	0.4674916	0.5634385	No
row_61	RGS1	2112	0.43953475	0.558134	No
row_62	CCL1	2144	0.43592098	0.5625956	No
row_63	GPR171	2449	0.40573722	0.55485255	No
row_64	CCR1	2643	0.38763744	0.5516676	No
row_65	IL1RL1	2849	0.37341958	0.54777455	No
row_66	BHLHE40	2863	0.3720229	0.5521639	No

row_67	NIPAL1	3327	0.3381336	0.5366469	No
row_68	KCNK5	3703	0.31361309	0.52460814	No
row_69	SBK1	21929	-0.5294169	-0.2562835	No
row_70	DAB2	21957	-0.533365	-0.2503521	No
row_71	FGD3	22011	-0.5449469	-0.2453906	No
row_72	MYB	22222	-0.5860325	-0.2466701	No
row_73	GEM	22444	-0.633794	-0.2477894	No
row_74	TBC1D16	22592	-0.669375	-0.2452359	No
row_75	IRF2BP2	22681	-0.6919311	-0.2398314	No
row_76	TSPAN2	22716	-0.7037675	-0.2319347	No
row_77	SKIL	22749	-0.7145669	-0.2238078	No
row_78	ID3	22766	-0.7178465	-0.2149456	No
row_79	JUN	22769	-0.718105	-0.2054746	No
row_80	ANKRD55	22781	-0.7227693	-0.1963306	No
row_81	CD27	22864	-0.7587839	-0.1897769	No
row_82	ITPKB	22904	-0.7804166	-0.1810763	No
row_83	ACVR2A	22909	-0.7820888	-0.1708402	No
row_84	FAM102A	22929	-0.7929533	-0.161108	No
row_85	BCL9L	22984	-0.8264884	-0.1524427	No
row_86	CD101	23044	-0.8566818	-0.1435917	No
row_87	GIMAP1	23047	-0.8579355	-0.1322596	No
row_88	NT5E	23060	-0.8675966	-0.1212314	No
row_89	YPEL2	23063	-0.8687891	-0.1097549	No
row_90	MAML3	23136	-0.9666377	-0.1000025	No
row_91	SMAD7	23149	-0.9848332	-0.0874139	No
row_92	BACH2	23173	-1.0352906	-0.0746293	No
row_93	IL16	23182	-1.0557615	-0.0609237	No
row_94	TGFBR2	23189	-1.0723158	-0.0469114	No
row_95	FAIM3	23196	-1.0912763	-0.0326467	No
row_96	LEF1	23222	-1.2716615	-0.0168027	No
row_97	TMEM66	23223	-1.2754661	1.73E-04	No

Bibliography

Abadie, V., Sollid, L.M., Barreiro, L.B., Jabri, B., 2011. Integration of Genetic and Immunological Insights into a Model of Celiac Disease Pathogenesis. Annu. Rev. Immunol. 29, 493–525. doi:10.1146/annurev-immunol-040210-092915

Abraham, G., Tye-Din, J.A., Bhalala, O.G., Kowalczyk, A., Zobel, J., Inouye, M., 2014. Accurate and Robust Genomic Prediction of Celiac Disease Using Statistical Learning. PLoS Genet 10, e1004137. doi:10.1371/journal.pgen.1004137

Agostoni, C., Decsi, T., Fewtrell, M., Goulet, O., Kolacek, S., Koletzko, B., Michaelsen, K.F., Moreno, L., Puntis, J., Rigo, J., Shamir, R., Szajewska, H., Turck, D., van Goudoever, J., 2008. Complementary Feeding: A Commentary by the ESPGHAN Committee on Nutrition. J. Pediatr. Gastroenterol. Nutr. 46, 99–110. doi:10.1097/01.mpg.0000304464.60788.bd

Akhabir, L., Bérubé, J.-C., Bossé, Y., Laviolette, M., Hao, K., Nickle, D.C., Timens, W., Sin, D.D., Paré, P.D., Postma, D.S., Sandford, A.J., 2014. Lung expression quantitative trait loci data set identifies important functional polymorphisms in the asthma-associated IL1RL1 region. J. Allergy Clin. Immunol. 134, 729–731. doi:10.1016/j.jaci.2014.02.039

Akobeng, A.K., Ramanan, A.V., Buchan, I., Heller, R.F., 2006. Effect of breast feeding on risk of coeliac disease: a systematic review and meta-analysis of observational studies. Arch. Dis. Child. 91, 39–43. doi:10.1136/adc.2005.082016

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511. doi:10.1038/35000501

Altshuler, D., Daly, M.J., Lander, E.S., 2008. Genetic Mapping in Human Disease. Science 322, 881–888. doi:10.1126/science.1156409

Amital, H., Govoni, M., Maya, R., Meroni, P.L., Ori, B., Shoenfeld, Y., Tincani, A., Trotta, F., Sarzi-Puttini, P., Atzeni, F., 2008. Role of infectious agents in systemic rheumatic diseases. Clin. Exp. Rheumatol. 26, S27–32.

Amundsen, S.S., Viken, M.K., Sollid, L.M., Lie, B.A., 2014. Coeliac disease-associated polymorphisms influence thymic gene expression. Genes Immun. 15, 355–360. doi:10.1038/gene.2014.26

Anderson, C., Frazer, A.C., French, J.M., Gerrard, J.W., Sammons, H.G., Smellie, J.M., 1952. CŒLIAC DISEASE : GASTRO-INTESTINAL STUDIES AND THE EFFECT OF DIETARY WHEAT FLOUR. The Lancet, Originally published as Volume 1, Issue 6713 259, 836–842. doi:10.1016/S0140-6736(52)90795-2

Anderson, R.P., van Heel, D.A., Tye - Din, J.A., Jewell, D.P., Hill, A.V.S., 2006. Antagonists and non - toxic variants of the dominant wheat gliadin T-cell epitope in coeliac disease. Gut 55, 485–491. doi:10.1136/gut.2005.064550

Anders, S., Pyl, P.T., Huber, W., 2014. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics btu638. doi:10.1093/bioinformatics/btu638

Anders, S., Reyes, A., Huber, W., 2012. Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017. doi:10.1101/gr.133744.111

Armstrong, D.L., Zidovetzki, R., Alarcón-Riquelme, M.E., Tsao, B.P., Criswell, L.A., Kimberly, R.P., Harley, J.B., Sivils, K.L., Vyse, T.J., Gaffney, P.M., Langefeld, C.D., Jacob, C.O., 2014. GWAS identifies novel SLE susceptibility genes and explains the association of the HLA region. Genes Immun. 15, 347–354. doi:10.1038/gene.2014.23

Aromaa, A., 1989. The Execution of the Mini-Finland Health Survey. Part 1: Aims, Methods, and Study Population [in Finnish with English summary]. Hels. Finl. Publ. Soc. Insur. Inst. Publication ML: 88.

Ayalew, M., Le-Niculescu, H., Levey, D.F., Jain, N., Changala, B., Patel, S.D., Winiger, E., Breier, A., Shekhar, A., Amdur, R., Koller, D., Nurnberger, J.I., Corvin, A., Geyer, M., Tsuang, M.T., Salomon, D., Schork, N.J., Fanous, A.H., O'Donovan, M.C., Niculescu, A.B., 2012. Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. Mol. Psychiatry 17, 887–905. doi:10.1038/mp.2012.37

Bai, D., Brar, P., Holleran, S., Ramakrishnan, R., Green, P.H.R., 2005. Effect of gender on the manifestations of celiac disease: Evidence for greater malabsorption in men. Scand. J. Gastroenterol. 40, 183–187. doi:10.1080/00365520510011498

Barrett, J.C., Clayton, D.G., Concannon, P., Akolkar, B., Cooper, J.D., Erlich, H.A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D.J., Stevens, H., Todd, J.A., Walker, N.M., Rich, S.S., 2009. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707. doi:10.1038/ng.381

Bodd, M., Ráki, M., Tollefsen, S., Fallang, L.E., Bergseng, E., Lundin, K.E.A., Sollid, L.M., 2010. HLA-DQ2-restricted gluten-reactive T-cells produce IL-21 but not IL-17 or IL-22. Mucosal Immunol. 3, 594–601. doi:10.1038/mi.2010.36

Botstein, D., Risch, N., 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. 33, 228–237. doi:10.1038/ng1090

Bracken, S., Byrne, G., Kelly, J., Jackson, J., Feighery, C., 2008. Altered gene expression in highly purified enterocytes from patients with active coeliac disease. BMC Genomics 9, 377. doi:10.1186/1471-2164-9-377

Brophy, K., Ryan, A.W., Thornton, J.M., Abuzakouk, M., Fitzgerald, A.P., McLoughlin, R.M., O'Morain, C., Kennedy, N.P., Stevens, F.M., Feighery, C., Kelleher, D., McManus, R., 2005. Haplotypes in the CTLA4 region are associated with coeliac disease in the Irish population. Genes Immun. 7, 19–26. doi:10.1038/sj.gene.6364265

Buchanan, C.C., Torstenson, E.S., Bush, W.S., Ritchie, M.D., 2012. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. J. Am. Med. Inform. Assoc. JAMIA 19, 289–294. doi:10.1136/amiajnl-2011-000652

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J., Todd, J.A., Donnelly, P., Barrett, J.C., Burton, P.R., Davison, D., Donnelly, P., Easton, D., Evans, D., Leung, H.-T., Marchini, J.L., Morris, A.P., Spencer, C.C.A., Tobin, M.D., Cardon, L.R., Clayton, D.G., Attwood, A.P., Boorman, J.P., Cant, B., Everson, U., Hussey, J.M., Jolley, J.D., Knight, A.S., Koch, K., Meech, E., Nutland, S., Prowse, C.V., Stevens, H.E., Taylor, N.C., Walters, G.R., Walker, N.M., Watkins, N.A., Winzer, T., Todd, J.A., Ouwehand, W.H., Jones, R.W., McArdle, W.L., Ring, S.M., Strachan, D.P., Pembrey, M., Breen, G., Clair, D.S., Caesar, S., Gordon-Smith, K., Jones, L., Fraser, C., Green, E.K., Grozeva, D., Hamshere, M.L., Holmans, P.A., Jones, I.R., Kirov, G., Moskvina, V., Nikolov, I., O'Donovan, M.C., Owen, M.J., Craddock, N., Collier, D.A., Elkin, A., Farmer, A., Williamson, R., McGuffin, P., Young, A.H., Ferrier, I.N., Ball, S.G., Balmforth, A.J., Barrett, J.H., Bishop, D.T., Iles, M.M., Maqbool, A., Yuldasheva, N., Hall, A.S., Braund, P.S., Burton, P.R., Dixon, R.J., Mangino, M., Stevens, S., Tobin, M.D., Thompson, J.R., Samani, N.J., Bredin, F., Tremelling, M., Parkes, M., Drummond, H., Lees, C.W., Nimmo, E.R., Satsangi, J., Fisher, S.A., Forbes, A., Lewis, C.M., Onnie, C.M., Prescott, N.J., Sanderson, J., Mathew, C.G., Barbour, J., Mohiuddin, M.K., Todhunter, C.E., Mansfield, J.C., Ahmad, T., Cummings, F.R., Jewell, D.P., Webster, J., Brown, M.J., Clayton, D.G., Lathrop, G.M., Connell, J., Dominiczak, A., Samani, N.J., Marcano, C.A.B., Burke, B., Dobson, R., Gungadoo, J., Lee, K.L., Munroe, P.B., Newhouse, S.J., Onipinla, A., Wallace, C., Xue, M., Caulfield, M., Farrall, M., Barton, A., (braggs), T.B. in R.G. and G., Bruce, I.N., Donovan, H., Eyre, S., Gilbert, P.D., Hider, S.L., Hinks, A.M., John, S.L., Potter, C., Silman, A.J., Symmons, D.P.M., Thomson, W., Worthington, J., Clayton, D.G., Dunger, D.B., Nutland, S., Stevens, H.E., Walker, N.M., Widmer, B., Todd, J.A., Frayling, T.M., Freathy, R.M., Lango, H., Perry, J.R.B., Shields, B.M., Weedon, M.N., Hattersley, A.T., Hitman, G.A., Walker, M., Elliott, K.S., Groves, C.J., Lindgren, C.M., Rayner, N.W., Timpson, N.J., Zeggini, E., McCarthy, M.I., Newport, M., Sirugo, G., Lyons, E., Vannberg, F., Hill, A.V.S., Bradbury, L.A., Farrar, C., Pointon, J.J., Wordsworth, P., Brown, M.A., Franklyn, J.A., Heward, J.M., Simmonds, M.J., Gough, S.C.L., Seal, S., (uk), B.C.S.C., Stratton, M.R., Rahman, N., Ban, M., Goris, A., Sawcer, S.J., Compston, A., Conway, D., Jallow, M., Newport, M., Sirugo, G., Rockett, K.A., Kwiatkowski, D.P., Bumpstead, S.J., Chaney, A., Downes, K., Ghori, M.J.R., Gwilliam, R., Hunt, S.E., Inouye, M., Keniry, A., King, E., McGinnis, R., Potter, S., Ravindrarajah, R., Whittaker, P., Widden, C., Withers, D., Deloukas, P., Leung, H.-T., Nutland, S., Stevens, H.E., Walker, N.M., Todd, J.A., Easton, D., Clayton, D.G., Burton, P.R., Tobin, M.D., Barrett, J.C., Evans, D., Morris, A.P., Cardon, L.R., Cardin, N.J., Davison, D., Ferreira, T., Pereira-Gale, J., Hallgrimsdóttir, I.B., Howie, B.N., Marchini, J.L., Spencer, C.C.A., Su, Z., Teo, Y.Y., Vukcevic, D., Donnelly, P., Bentley, D., Brown, M.A., Cardon, L.R., Caulfield, M., Clayton, D.G., Compston, A., Craddock, N., Deloukas, P., Donnelly, P., Farrall, M., Gough, S.C.L., Hall, A.S., Hattersley, A.T., Hill, A.V.S., Kwiatkowski, D.P., Mathew, C.G., McCarthy, M.I., Ouwehand, W.H., Parkes, M., Pembrey, M., Rahman, N., Samani, N.J., Stratton, M.R., Todd, J.A., Worthington, J., 2007.

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678. doi:10.1038/nature05911

Burwick, R.M., Ramsay, P.P., Haines, J.L., Hauser, S.L., Oksenberg, J.R., Pericak-Vance, M.A., Schmidt, S., Compston, A., Sawcer, S., Cittadella, R., Savettieri, G., Quattrone, A., Polman, C.H., Uitdehaag, B.M.J., Zwemmer, J.N.P., Hawkins, C.P., Ollier, W.E.R., Weatherby, S., Enzinger, C., Fazekas, F., Schmidt, H., Schmidt, R., Hillert, J., Masterman, T., Hogh, P., Niino, M., Kikuchi, S., Maciel, P., Santos, M., Rio, M.E., Kwiecinski, H., Zakrzewska-Pniewska, B., Evangelou, N., Palace, J., Barcellos, L.F., 2006. APOE epsilon variation in multiple sclerosis susceptibility and disease severity Some answers. Neurology 66, 1373–1383. doi:10.1212/01.wnl.0000210531.19498.3f

Cadwell, K., Patel, K.K., Maloney, N.S., Liu, T.-C., Ng, A.C.Y., Storer, C.E., Head, R.D., Xavier, R., Stappenbeck, T.S., Virgin, H.W., 2010. Virus-Plus-Susceptibility Gene Interaction Determines Crohn's Disease Gene Atg16L1 Phenotypes in Intestine. Cell 141, 1135–1145. doi:10.1016/j.cell.2010.05.009

Cammarota, G., Cuoco, L., Cianci, R., Pandolfi, F., Gasbarrini, G., 2000. Onset of coeliac disease during treatment with interferon for chronic hepatitis C. The Lancet 356, 1494–1495. doi:10.1016/S0140-6736(00)02880-4

Catassi, C., Ratsch, Ii.-M., Gandolfi, L., Pratesi, R., Fabiani, E., Asmar, R.E., Frijia, M., Bearzi, I., Vizzoni, L., 1999. Why is coeliac disease endemic in the people of the Sahara? The Lancet 354, 647–648. doi:10.1016/S0140-6736(99)02609-4

Challar, M., Jouma, M., Sitzmann, F., Seferian, V., Shahin, E., 2004. Prevalence of asymptomatic celiac disease in a Syrian population sample. JABMS 6.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 1–16. doi:10.1186/s13742-015-0047-8

Chin, M.W., Mallon, D.F., Cullen, D.J., Olynyk, J.K., Mollison, L.C., Pearce, C.B., 2009. Screening for coeliac disease using anti-tissue transglutaminase antibody assays, and prevalence of the disease in an Australian community. Med. J. Aust. 190, 429–432.

Christophersen, A., Ráki, M., Bergseng, E., Lundin, K.E., Jahnsen, J., Sollid, L.M., Qiao, S.-W., 2014. Tetramer-visualized gluten-specific CD4+ T-cells in blood as a potential diagnostic marker for coeliac disease without oral gluten challenge. United Eur. Gastroenterol. J. 2, 268–278. doi:10.1177/2050640614540154

Clarke, C.A., Edwards, J.W., Haddock, D.R.W., Howel-Evans, A.W., McConnell, R.B., Sheppard, P.M., 1956. ABO Blood Groups and Secretor Character in Duodenal Ulcer. Br. Med. J. 2, 725–731.

Coleman, C., Quinn, E.M., Ryan, A.W., Conroy, J., Trimble, V., Mahmud, N., Kennedy, N., Corvin, A.P., Morris, D.W., Donohoe, G., O'Morain, C., MacMathuna, P., Byrnes, V., Kiat, C., Trynka, G., Wijmenga, C., Kelleher, D., Ennis, S., Anney, R.J., McManus, R., 2015. Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci. Eur. J. Hum. Genet. doi:10.1038/ejhg.2015.87

Colombo, P.-E., Milanezi, F., Weigelt, B., Reis-Filho, J.S., 2011. Microarrays in the 2010s: the contribution of microarray-based gene expression profiling to breast

cancer classification, prognostication and prediction. Breast Cancer Res. 13, 1–15. doi:10.1186/bcr2890

Consortium, T. 1000 G.P., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073. doi:10.1038/nature09534

Cooper, J.D., Simmonds, M.J., Walker, N.M., Burren, O., Brand, O.J., Guo, H., Wallace, C., Stevens, H., Coleman, G., Franklyn, J.A., Todd, J.A., Gough, S.C.L., 2012. Seven newly identified loci for autoimmune thyroid disease. Hum. Mol. Genet. 21, 5202–5208. doi:10.1093/hmg/dds357

Cooper, J.D., Smyth, D.J., Smiles, A.M., Plagnol, V., Walker, N.M., Allen, J.E., Downes, K., Barrett, J.C., Healy, B.C., Mychaleckyj, J.C., Warram, J.H., Todd, J.A., 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. Nat. Genet. 40, 1399–1401. doi:10.1038/ng.249

Corazza, G.R., Villanacci, V., 2005. Coeliac disease. J. Clin. Pathol. 58, 573–574. doi:10.1136/jcp.2004.023978

Corrao, G., Corazza, G.R., Bagnardi, V., Brusco, G., Ciacci, C., Cottone, M., Guidetti, C.S., Usai, P., Cesari, P., Pelli, M.A., Loperfido, S., Volta, U., Calabró, A., Certo, M., 2001. Mortality in patients with coeliac disease and their relatives: a cohort study. The Lancet 358, 356–361. doi:10.1016/S0140-6736(01)05554-4

Cortes, A., Brown, M.A., 2011. Promise and pitfalls of the Immunochip. Arthritis Res. Ther. 13, 101. doi:10.1186/ar3204

Costello, C.M., Mah, N., Häsler, R., Rosenstiel, P., Waetzig, G.H., Hahn, A., Lu, T., Gurbuz, Y., Nikolaus, S., Albrecht, M., Hampe, J., Lucius, R., Klöppel, G., Eickhoff, H., Lehrach, H., Lengauer, T., Schreiber, S., 2005. Dissection of the Inflammatory Bowel Disease Transcriptome Using Genome-Wide cDNA Microarrays. PLoS Med 2, e199. doi:10.1371/journal.pmed.0020199

Crosnier, C., Stamataki, D., Lewis, J., 2006. Organizing cell renewal in the intestine: stem cells, signals and combinatorial control. Nat. Rev. Genet. 7, 349–359. doi:10.1038/nrg1840

Daoud, S.S., 2014. Genome Wide Identification of FGFR2 Alternative Splicing in Hepatitis C: Potential Roles in Malignant Transformation. Pharmacogenomics Pharmacoproteomics.

DeWan, A., Liu, M., Hartman, S., Zhang, S.S.-M., Liu, D.T.L., Zhao, C., Tam, P.O.S., Chan, W.M., Lam, D.S.C., Snyder, M., Barnstable, C., Pang, C.P., Hoh, J., 2006. HTRA1 Promoter Polymorphism in Wet Age-Related Macular Degeneration. Science 314, 989–992. doi:10.1126/science.1133807

Dicke, W.K., 1950. Coeliac disease. Investigation of the harmful effects of certain types of cereal on patients with coeliac disease. Dissertation.

Dieli-Crimi, R., Cénit, M.C., Núñez, C., n.d. The genetics of celiac disease: A comprehensive review of clinical implications. J. Autoimmun. doi:10.1016/j.jaut.2015.07.003

Dieterich, W., Ehnis, T., Bauer, M., Donner, P., Volta, U., Riecken, E.O., Schuppan, D., 1997. Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat. Med. 3, 797–801. doi:10.1038/nm0797-797

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Arcelus, M.G., Sekowska, M., Gagnebin, M., Nisbett, J.,

Deloukas, P., Dermitzakis, E.T., Antonarakis, S.E., 2009. Common regulatory variation impacts gene expression in a cell type dependent manner. Science 325, 1246–1250. doi:10.1126/science.1174148

Dinarello, C.A., Novick, D., Kim, S., Kaplanski, G., 2013. Interleukin-18 and IL-18 Binding Protein. Front. Immunol. 4. doi:10.3389/fimmu.2013.00289

Diosdado, B., Wapenaar, M.C., Franke, L., Duran, K.J., Goerres, M.J., Hadithi, M., Crusius, J.B.A., Meijer, J.W.R., Duggan, D.J., Mulder, C.J.J., Holstege, F.C.P., Wijmenga, C., 2004. A microarray screen for novel candidate genes in coeliac disease pathogenesis. Gut 53, 944–951. doi:10.1136/gut.2003.018374

Di Sabatino, A., Corazza, G.R.,. Coeliac disease. The Lancet 373, 1480–1493. doi:10.1016/S0140-6736(09)60254-3

Di Sabatino, A., D'Alò, S., Millimaggi, D., Ciccocioppo, R., Parroni, R., Sciarra, G., Cifone, M.G., Corazza, G.R., 2001. Apoptosis and peripheral blood lymphocyte depletion in coeliac disease. Immunology 103, 435–440. doi:10.1046/j.1365-2567.2001.01245.x

Dominguez-Villar, M., Baecher-Allan, C.M., Hafler, D.A., 2011. Identification of T helper type 1-like, Foxp3+ regulatory T-cells in human autoimmune disease. Nat. Med. 17, 673–675. doi:10.1038/nm.2389

Dong, Q., Kirley, S., Rueda, B., Zhao, C., Zukerberg, L., Oliva, E., 2003. Loss of Cables, a Novel Gene on Chromosome 18q, in Ovarian Cancer. Mod. Pathol. 16, 863–868. doi:10.1097|[sol]|01.MP.0000084434.88269.0A

Dubois, P.C.A., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A.R., Adány, R., Aromaa, A., Bardella, M.T., van den Berg, L.H., Bockett, N.A., de la Concha, E.G., Dema, B., Fehrmann, R.S.N., Fernández-Arquero, M., Fiatal, S., Grandone, E., Green, P.M., Groen, H.J.M., Gwilliam, R., Houwen, R.H.J., Hunt, S.E., Kaukinen, K., Kelleher, D., Korponay-Szabo, I., Kurppa, K., MacMathuna, P., Mäki, M., Mazzilli, M.C., McCann, O.T., Mearin, M.L., Mein, C.A., Mirza, M.M., Mistry, V., Mora, B., Morley, K.I., Mulder, C.J., Murray, J.A., Núñez, C., Oosterom, E., Ophoff, R.A., Polanco, I., Peltonen, L., Platteel, M., Rybak, A., Salomaa, V., Schweizer, J.J., Sperandeo, M.P., Tack, G.J., Turner, G., Veldink, J.H., Verbeek, W.H.M., Weersma, R.K., Wolters, V.M., Urcelay, E., Cukrowska, B., Greco, L., Neuhausen, S.L., McManus, R., Barisani, D., Deloukas, P., Barrett, J.C., Saavalainen, P., Wijmenga, C., van Heel, D.A., 2010. Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet. 42, 295–302. doi:10.1038/ng.543

Dudbridge, F., 2013. Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9, e1003348. doi:10.1371/journal.pgen.1003348

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I.H., Nyberg, S., Wolf, M., Borresen-Dale, A.-L., Kallioniemi, O., 2011. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol. 12, R6. doi:10.1186/gb-2011-12-1-r6

Einarsdottir, E., Bevova, M.R., Zhernakova, A., Monsuur, A., Koskinen, L.L., Slot, R. van't, Mulder, C., Mearin, M.L., Korponay-Szabo, I.R., Kaukinen, K., Kurppa, K., Kere, J., Mäki, M., Wijmenga, C., Saavalainen, P., 2011. Multiple independent variants in 6q21-22 associated with susceptibility to celiac disease in the Dutch,

Finnish and Hungarian populations. Eur. J. Hum. Genet. 19, 682–686. doi:10.1038/ejhg.2011.2

Einarsdottir, E., Koskinen, L.L., Dukes, E., Kainu, K., Suomela, S., Lappalainen, M., Ziberna, F., Korponay-Szabo, I.R., Kurppa, K., Kaukinen, K., Ádány, R., Pocsai, Z., Széles, G., Färkkilä, M., Turunen, U., Halme, L., Paavola-Sakki, P., Not, T., Vatta, S., Ventura, A., Löfberg, R., Torkvist, L., Bresso, F., Halfvarson, J., Mäki, M., Kontula, K., Saarialho-Kere, U., Kere, J., D'Amato, M., Saavalainen, P., 2009. IL23R in the Swedish, Finnish, Hungarian and Italian populations: association with IBD and psoriasis, and linkage to celiac disease. BMC Med. Genet. 10, 8. doi:10.1186/1471-2350-10-8

Euskirchen, G.M., Auerbach, R.K., Davidov, E., Gianoulis, T.A., Zhong, G., Rozowsky, J., Bhardwaj, N., Gerstein, M.B., Snyder, M., 2011. Diverse Roles and Interactions of the SWI/SNF Chromatin Remodeling Complex Revealed Using Global Approaches. PLoS Genet 7, e1002008. doi:10.1371/journal.pgen.1002008

Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., Amos, C.I., Padyukov, L., Toes, R.E.M., Huizinga, T.W.J., Wijmenga, C., Trynka, G., Franke, L., Westra, H.-J., Alfredsson, L., Hu, X., Sandor, C., de Bakker, P.I.W., Davila, S., Khor, C.C., Heng, K.K., Andrews, R., Edkins, S., Hunt, S.E., Langford, C., Symmons, D., Concannon, P., Onengut-Gumuscu, S., Rich, S.S., Deloukas, P., Gonzalez-Gay, M.A., Rodriguez-Rodriguez, L., Ärlsetig, L., Martin, J., Rantapää-Dahlqvist, S., Plenge, R., Raychaudhuri, S., Klareskog, L., Gregersen, P.K., Worthington, J., 2012. High density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. Nat. Genet. 44, 1336–1340. doi:10.1038/ng.2462

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., Knight, J.C., 2014. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science 343, 1246949. doi:10.1126/science.1246949

Fasano A, Berti I, Gerarduzzi T, et al, 2003. Prevalence of celiac disease in at-risk and not-at-risk groups in the united states: A large multicenter study. Arch. Intern. Med. 163, 286–292. doi:10.1001/archinte.163.3.286

Fazi, F., Rosa, A., Fatica, A., Gelmetti, V., De Marchis, M.L., Nervi, C., Bozzoni, I., 2005. A Minicircuitry Comprised of MicroRNA-223 and Transcription Factors NFI-A and C/EBPα Regulates Human Granulopoiesis. Cell 123, 819–831. doi:10.1016/j.cell.2005.09.023

Fernández, S., Molina, I.J., Romero, P., González, R., Peña, J., Sánchez, F., Reynoso, F.R., Pérez-Navero, J.L., Estevez, O., Ortega, C., Santamaría, M., 2011. Characterization of Gliadin-Specific Th17 Cells from the Mucosa of Celiac Disease Patients. Am. J. Gastroenterol. 106, 528–538. doi:10.1038/ajg.2010.465

Festen, E.A.M., Goyette, P., Green, T., Boucher, G., Beauchamp, C., Trynka, G., Dubois, P.C., Lagacé, C., Stokkers, P.C.F., Hommes, D.W., Barisani, D., Palmieri, O., Annese, V., van Heel, D.A., Weersma, R.K., Daly, M.J., Wijmenga, C., Rioux, J.D., 2011. A Meta-Analysis of Genome-Wide Association Scans Identifies IL18RAP,

PTPN2, TAGAP, and PUS10 As Shared Risk Loci for Crohn's Disease and Celiac Disease. PLoS Genet 7, e1001283. doi:10.1371/journal.pgen.1001283

Festen, E.A.M., Goyette, P., Scott, R., Annese, V., Zhernakova, A., Lian, J., Lefèbvre, C., Brant, S.R., Cho, J.H., Silverberg, M.S., Taylor, K.D., De Jong, D.J., Stokkers, P.C., McGovern, D., Palmieri, O., Achkar, J.-P., Xavier, R.J., Daly, M.J., Duerr, R.H., Wijmenga, C., Weersma, R.K., Rioux, J.D., 2009. Genetic Variants in the Region Harbouring IL2/IL21 Associated to Ulcerative Colitis. Gut 58, 799–804. doi:10.1136/gut.2008.166918

Fisher, R.., 1919. The Correlation between Relatives on the Supposition of Mendelian Inheritance.. Trans. R. Soc. Edinb. 52, 399–433.

Fløyel, T., Brorsson, C., Nielsen, L.B., Miani, M., Bang-Berthelsen, C.H., Friedrichsen, M., Overgaard, A.J., Berchtold, L.A., Wiberg, A., Poulsen, P., Hansen, L., Rosinger, S., Boehm, B.O., Ram, R., Nguyen, Q., Mehta, M., Morahan, G., Concannon, P., Bergholdt, R., Nielsen, J.H., Reinheckel, T., Herrath, M. von, Vaag, A., Eizirik, D.L., Mortensen, H.B., Størling, J., Pociot, F., 2014. CTSH regulates β -cell function and disease progression in newly diagnosed type 1 diabetes patients. Proc. Natl. Acad. Sci. 111, 10305–10310. doi:10.1073/pnas.1402571111

Franke, A., Balschun, T., Karlsen, T.H., Hedderich, J., May, S., Lu, T., Schuldt, D., Nikolaus, S., Rosenstiel, P., Krawczak, M., Schreiber, S., 2008. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat. Genet. 40, 713–715. doi:10.1038/ng.148

Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., Anderson, C.A., Bis, J.C., Bumpstead, S., Ellinghaus, D., Festen, E.M., Georges, M., Green, T., Haritunians, T., Jostins, L., Latiano, A., Mathew, C.G., Montgomery, G.W., Prescott, N.J., Raychaudhuri, S., Rotter, J.I., Schumm, P., Sharma, Y., Simms, L.A., Taylor, K.D., Whiteman, D., Wijmenga, C., Baldassano, R.N., Barclay, M., Bayless, T.M., Brand, S., Büning, C., Cohen, A., Colombel, J.-F., Cottone, M., Stronati, L., Denson, T., De Vos, M., D'Inca, R., Dubinsky, M., Edwards, C., Florin, T., Franchimont, D., Gearry, R., Glas, J., Van Gossum, A., Guthery, S.L., Halfvarson, J., Verspaget, H.W., Hugot, J.-P., Karban, A., Laukens, D., Lawrance, I., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mowat, C., Newman, W., Panés, J., Phillips, A., Proctor, D.D., Regueiro, M., Russell, R., Rutgeerts, P., Sanderson, J., Sans, M., Seibold, F., Steinhart, A.H., Stokkers, P.C.F., Torkvist, L., Kullak-Ublick, G., Wilson, D., Walters, T., Targan, S.R., Brant, S.R., Rioux, J.D., D'Amato, M., Weersma, R.K., Kugathasan, S., Griffiths, A.M., Mansfield, J.C., Vermeire, S., Duerr, R.H., Silverberg, M.S., Satsangi, J., Schreiber, S., Cho, J.H., Annese, V., Hakonarson, H., Daly, M.J., Parkes, M., 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. 42, 1118–1125. doi:10.1038/ng.717

Fransen, K., Visschedijk, M.C., Sommeren, S. van, Fu, J.Y., Franke, L., Festen, E.A.M., Stokkers, P.C.F., Bodegraven, A.A. van, Crusius, J.B.A., Hommes, D.W., Zanen, P., Jong, D.J. de, Wijmenga, C., Diemen, C.C. van, Weersma, R.K., 2010. Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as

potential new risk genes for Crohn's disease. Hum. Mol. Genet. 19, 3482–3488. doi:10.1093/hmg/ddq264

Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., James, M., Liu, P., Tichelaar, J.W., Vikis, H.G., You, M., Mills, I.G., 2011. Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. 43, 513–518. doi:10.1038/ng.840

Fujimoto, M., Tsutsui, H., Xinshou, O., Tokumoto, M., Watanabe, D., Shima, Y., Yoshimoto, T., Hirakata, H., Kawase, I., Nakanishi, K., Kishimoto, T., Naka, T., 2004. Inadequate induction of suppressor of cytokine signaling - 1 causes systemic autoimmune diseases. Int. Immunol. 16, 303–314. doi:10.1093/intimm/dxh030

Fu, J., Wolfs, M.G.M., Deelen, P., Westra, H.-J., Fehrmann, R.S.N., te Meerman, G.J., Buurman, W.A., Rensen, S.S.M., Groen, H.J.M., Weersma, R.K., van den Berg, L.H., Veldink, J., Ophoff, R.A., Snieder, H., van Heel, D., Jansen, R.C., Hofker, M.H., Wijmenga, C., Franke, L., 2012. Unraveling the Regulatory Mechanisms Underlying Tissue-Dependent Genetic Variation of Gene Expression. PLoS Genet 8, e1002431. doi:10.1371/journal.pgen.1002431

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D., 2002. The Structure of Haplotype Blocks in the Human Genome. Science 296, 2225–2229. doi:10.1126/science.1069424

Garapaty, S., Xu, C.-F., Trojer, P., Mahajan, M.A., Neubert, T.A., Samuels, H.H., 2009. Identification and Characterization of a Novel Nuclear Protein Complex Involved in Nuclear Hormone Receptor-mediated Gene Regulation. J. Biol. Chem. 284, 7542–7552. doi:10.1074/jbc.M805872200

Garner, C., Ahn, R., Ding, Y.C., Steele, L., Stoven, S., Green, P.H., Fasano, A., Murray, J.A., Neuhausen, S.L., 2014. Genome-Wide Association Study of Celiac Disease in North America Confirms FRMD4B as New Celiac Locus. PLoS ONE 9, e101428. doi:10.1371/journal.pone.0101428

Gee, S., 1888. On the coeliac affection. St Bartholomews Hosp Rep. 24:17–20.

Genetic Analysis of Psoriasis Consortium & the Wellcome Trust Case Control Consortium 2, 2010. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat. Genet. 42, 985–990. doi:10.1038/ng.694

Gerrits, A., Li, Y., Tesson, B.M., Bystrykh, L.V., Weersing, E., Ausema, A., Dontje, B., Wang, X., Breitling, R., Jansen, R.C., de Haan, G., 2009. Expression Quantitative Trait Loci Are Highly Sensitive to Cellular Differentiation State. PLoS Genet 5, e1000692. doi:10.1371/journal.pgen.1000692

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., Tam, P.K.-H., Tsui, L.-C., Waye, M.M.Y., Wong, J.T.-F., Zeng, C., Zhang, Q., Chee, M.S., Galver, L.M., Kruglyak, S., Murray, S.S., Oliphant, A.R., Montpetit, A., Hudson, T.J., Chagnon, F., Ferretti, V., Leboeuf, M., Phillips, M.S., Verner, A., Kwok, P.-Y., Duan, S., Lind, D.L., Miller, R.D., Rice,

J.P., Saccone, N.L., Taillon-Miller, P., Xiao, M., Nakamura, Y., Sekine, A., Sorimachi, K., Tanaka, T., Tanaka, Y., Tsunoda, T., Yoshino, E., Bentley, D.R., Deloukas, P., Hunt, S., Powell, D., Altshuler, D., Gabriel, S.B., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Aniagwu, T., Marshall, P.A., Matthew, O., Nkwodimmah, C., Royal, C.D.M., Leppert, M.F., Dixon, M., Stein, L.D., Cunningham, F., Kanani, A., Thorisson, G.A., Chakravarti, A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Donnelly, P., Marchini, J., McVean, G.A.T., Myers, S.R., Cardon, L.R., Abecasis, G.R., Morris, A., Weir, B.S., Mullikin, J.C., Sherry, S.T., Feolo, M., Daly, M.J., Schaffner, S.F., Qiu, R., Kent, A., Dunston, G.M., Kato, K., Niikawa, N., Knoppers, B.M., Foster, M.W., Clayton, E.W., Wang, V.O., Watkin, J., Sodergren, E., Weinstock, G.M., Wilson, R.K., Fulton, L.L., Rogers, J., Birren, B.W., Han, H., Wang, H., Godbout, M., Wallenburg, J.C., L'Archevêque, P., Bellemare, G., Todani, K., Fujita, T., Tanaka, S., Holden, A.L., Lai, E.H., Collins, F.S., Brooks, L.D., McEwen, J.E., Guyer, M.S., Jordan, E., Peterson, J.L., Spiegel, J., Sung, L.M., Zacharia, L.F., Kennedy, K., Dunn, M.G., Seabrook, R., Shillito, M., Skene, B., Stewart, J.G., (chair), D.L.V., (co-Chair), E.W.C., (co-Chair), L.B.J., Cho, M.K., Duster, T., Jasperse, M., Licinio, J., Long, J.C., Ossorio, P.N., Spallone, P., Terry, S.F., (chair), E.S.L., (co-Chair), E.H.L., (co-Chair), D.A.N., Boehnke, M., Douglas, J.A., Hudson, R.R., Kruglyak, L., Nussbaum, R.L., 2003. The International HapMap Project. Nature 426, 789–796. doi:10.1038/nature02168

Gibson, G., 2012. Rare and common variants: twenty arguments. Nat. Rev. Genet. 13, 135–145. doi:10.1038/nrg3118

Glas, J., Wagner, J., Seiderer, J., Olszak, T., Wetzke, M., Beigel, F., Tillack, C., Stallhofer, J., Friedrich, M., Steib, C., Göke, B., Ochsenkühn, T., Karbalai, N., Diegelmann, J., Czamara, D., Brand, S., 2012. PTPN2 Gene Variants Are Associated with Susceptibility to Both Crohn's Disease and Ulcerative Colitis Supporting a Common Genetic Disease Background. PLoS ONE 7, e33682. doi:10.1371/journal.pone.0033682

Goudey, B., Abraham, G., Kikianty, E., Wang, Q., Rawlinson, D., Shi, F., Haviv, I., Stern, L., Kowalczyk, A., Inouye, M., 2014. Extensive epistasis within the MHC contributes to the genetic architecture of celiac disease. bioRxiv. doi:10.1101/002485

Granzotto, M., Bo, S. dal, Quaglia, S., Tommasini, A., Piscianz, E., Valencic, E., Ferrara, F., Martelossi, S., Ventura, A., Not, T., 2008. Regulatory T-Cell Function Is Impaired in Celiac Disease. Dig. Dis. Sci. 54, 1513–1519. doi:10.1007/s10620-008-0501-x

Greco, L., Corazza, G., Babron, M.-C., Clot, F., Fulchignoni-Lataud, M.-C., Percopo, S., Zavattari, P., Bouguerra, F., Dib, C., Tosi, R., Troncone, R., Ventura, A., Mantavoni, W., Magazzù, G., Gatti, R., Lazzari, R., Giunta, A., Perri, F., Iacono, G., Cardi, E., de Virgiliis, S., Cataldo, F., De Angelis, G., Musumeci, S., Ferrari, R., Balli, F., Bardella, M.-T., Volta, U., Catassi, C., Torre, G., Eliaou, J.-F., Serre, J.-L., Clerget-Darpoux, F., 1998. Genome Search in Celiac Disease. Am. J. Hum. Genet. 62, 669–675. doi:10.1086/301754

Greco, L., Romino, R., Coto, I., Cosmo, N.D., Percopo, S., Maglio, M., Paparo, F., Gasperi, V., Limongelli, M.G., Cotichini, R., D'Agate, C., Tinto, N., Sacchetti, L., Tosi, R., Stazi, M.A., 2002. The first large population based twin study of coeliac disease. Gut 50, 624–628. doi:10.1136/gut.50.5.624

Green, E.K., Hamshere, M., Forty, L., Gordon-Smith, K., Fraser, C., Russell, E., Grozeva, D., Kirov, G., Holmans, P., Moran, J.L., Purcell, S., Sklar, P., Owen, M.J., O'Donovan, M.C., Jones, L., Jones, I.R., Craddock, N., 2012. Replication of bipolar disorder susceptibility alleles and identification of two novel genome-wide significant associations in a new bipolar disorder case–control sample. Mol. Psychiatry. doi:10.1038/mp.2012.142

Gregory, S.G., Schmidt, S., Seth, P., Oksenberg, J.R., Hart, J., Prokop, A., Caillier, S.J., Ban, M., Goris, A., Barcellos, L.F., Lincoln, R., McCauley, J.L., Sawcer, S.J., Compston, D. a. S., Dubois, B., Hauser, S.L., Garcia-Blanco, M.A., Pericak-Vance, M.A., Haines, J.L., 2007. Interleukin 7 receptor α chain (IL7R) shows allelic and functional association with multiple sclerosis. Nat. Genet. 39, 1083–1091. doi:10.1038/ng2103

Guo, H., Fortune, M.D., Burren, O.S., Schofield, E., Todd, J.A., Wallace, C., 2015. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. Hum. Mol. Genet. 24, 3305–3313. doi:10.1093/hmg/ddv077

Gusella, J.F., Wexler, N.S., Conneally, M.P., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonilla, E., Martin, J.B., 1983. A polymorphic DNA marker genetically linked to Huntingtons disease. Landmarks Med. Genet.

Gutierrez-Achury, J., Coutinho de Almeida, R., Wijmenga, C., 2011. Shared genetics in coeliac disease and other immune-mediated diseases. J. Intern. Med. 269, 591–603. doi:10.1111/j.1365-2796.2011.02375.x

Gutierrez-Achury, J., Zhernakova, A., Pulit, S.L., Trynka, G., Hunt, K.A., Romanos, J., Raychaudhuri, S., van Heel, D.A., Wijmenga, C., de Bakker, P.I.W., 2015. Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease. Nat. Genet. 47, 577–578. doi:10.1038/ng.3268

Hecker, M., Fitzner, B., Koczan, D., Friess, J., Roch, L., Schröder, I., Flechtner, K., Thiesen, H.-J., Winkelmann, A., Zettl, U., 2015. Differential Effects of Fingolimod on Gene Expression of Helper and Cytotoxic T-Cells in the Blood of MS Patients (P1.156). Neurology 84, P1.156.

Hindorff, L.A., MacArthur, J., Morales, J., Junkins, H.A., Hall, P., Klemm, A., Manolio, T.A., 2015. A Catalog of Published Genome-Wide Association Studies.

Hmida, N.B., Ahmed, M.B., Moussa, A., Rejeb, M.B., Said, Y., Kourda, N., Meresse, B., Abdeladhim, M., Louzir, H., Cerf-Bensussan, N., 2012. Impaired Control of Effector T-cells by Regulatory T-cells: A Clue to Loss of Oral Tolerance and Autoimmunity in Celiac Disease? Am. J. Gastroenterol. 107, 604–611. doi:10.1038/ajg.2011.397

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H., Young, R.A., 2013. Transcriptional super-enhancers connected to cell identity and disease. Cell 155. doi:10.1016/j.cell.2013.09.053

Holopainen, P., Arvas, M., Sistonen, P., Mustalahti, K., Collin, P., Mäki, M., Partanen, J., 1999. CD28/CTLA4 gene region on chromosome 2q33 confers genetic susceptibility to celiac disease. A linkage and family-based association study. Tissue Antigens 53, 470–475. doi:10.1034/j.1399-0039.1999.530503.x

Hrdlickova, B., Kumar, V., Kanduri, K., Zhernakova, D.V., Tripathi, S., Karjalainen, J., Lund, R.J., Li, Y., Ullah, U., Modderman, R., Abdulahad, W., Lähdesmäki, H., Franke, L., Lahesmaa, R., Wijmenga, C., Withoff, S., 2014. Expression profiles of long non-coding RNAs located in autoimmune disease-associated regions reveal immune cell-type specificity. Genome Med. 6, 88. doi:10.1186/s13073-014-0088-0

Hrdlickova, B., Westra, H.-J., Franke, L., Wijmenga, C., 2011. Celiac disease: moving from genetic associations to causal variants. Clin. Genet. 80, 203–313. doi:10.1111/j.1399-0004.2011.01707.x

Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.-P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J.-F., Sahbatou, M., Thomas, G., 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411, 599–603. doi:10.1038/35079107

Hunt, K.A., Mistry, V., Bockett, N.A., Ahmad, T., Ban, M., Barker, J.N., Barrett, J.C., Blackburn, H., Brand, O., Burren, O., Capon, F., Compston, A., Gough, S.C.L., Jostins, L., Kong, Y., Lee, J.C., Lek, M., MacArthur, D.G., Mansfield, J.C., Mathew, C.G., 2013. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498, 232–235. doi:10.1038/nature12170

Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D., Gwilliam, R., Takeuchi, F., McLaren, W.M., Holmes, G.K.T., Howdle, P.D., Walters, J.R.F., Sanders, D.S., Playford, R.J., Trynka, G., Mulder, C.J.J., Mearin, M.L., Verbeek, W.H.M., Trimble, V., Stevens, F.M., O'Morain, C., Kennedy, N.P., Kelleher, D., Pennington, D.J., Strachan, D.P., McArdle, W.L., Mein, C.A., Wapenaar, M.C., Deloukas, P., McGinnis, R., McManus, R., Wijmenga, C., van Heel, D.A., 2008. Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet. 40, 395–402. doi:10.1038/ng.102

Husby, S., Koletzko, S., Korponay-Szabó, I.R., Mearin, M.L., Phillips, A., Shamir, R., Troncone, R., Giersiepen, K., Branski, D., Catassi, C., Lelgeman, M., Mäki, M., Ribes-Koninckx, C., Ventura, A., Zimmer, K.P., 2012. European Society for Pediatric Gastroenterology, Hepatology, and Nutrition Guidelines for the Diagnosis of Coeliac Disease: J. Pediatr. Gastroenterol. Nutr. 54, 136–160. doi:10.1097/MPG.0b013e31821a23d0

International Genetics of Ankylosing Spondylitis Consortium, 2013. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. Nat. Genet. 45, 730–738. doi:10.1038/ng.2667

Jiang, H., Wong, W.H., 2009. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25, 1026–1032. doi:10.1093/bioinformatics/btp113

Joshi, A., Smet, R.D., Marchal, K., Peer, Y.V. de, Michoel, T., 2009. Module networks revisited: computational assessment and prioritization of model predictions. Bioinformatics 25, 490–496. doi:10.1093/bioinformatics/btn658

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Philip Schumm, L., Sharma, Y., Anderson, C.A., Essers, J., Mitrovic, M., Ning, K., Cleynen, I., Theatre, E., Spain, S.L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A.N., Andersen, V., Andrews, J.M., Baidoo, L., Balschun, T., Bampton, P.A., Bitton, A., Boucher, G., Brand, S., Büning, C., Cohain, A., Cichon, S., D'Amato, M., De Jong, D., Devaney, K.L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L.R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T.H., Kupcinskas, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I.C., Lees, C.W., Louis, E., Mahy, G., Mansfield, J., Morgan, A.R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C.Y., Potocnik, U., Prescott, N.J., Regueiro, M., Rotter, J.I., Russell, R.K., Sanderson, J.D., Sans, M., Satsangi, J., Schreiber, S., Simms, L.A., Sventoraityte, J., Targan, S.R., Taylor, K.D., Tremelling, M., Verspaget, H.W., De Vos, M., Wijmenga, C., Wilson, D.C., Winkelmann, J., Xavier, R.J., Zeissig, S., Zhang, B., Zhang, C.K., Zhao, H., The International IBD Genetics Consortium (iibdgc), Silverberg, M.S., Annese, V., Hakonarson, H., Brant, S.R., Radford-Smith, G., Mathew, C.G., Rioux, J.D., Schadt, E.E., Daly, M.J., Franke, A., Parkes, M., Vermeire, S., Barrett, J.C., Cho, J.H., 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124. doi:10.1038/nature11582

Juran, B.D., Hirschfield, G.M., Invernizzi, P., Atkinson, E.J., Li, Y., Xie, G., Kosoy, R., Ransom, M., Sun, Y., Bianchi, I., Schlicht, E.M., Lleo, A., Coltescu, C., Bernuzzi, F., Podda, M., Lammert, C., Shigeta, R., Chan, L.L., Balschun, T., Marconi, M., Cusi, D., Heathcote, E.J., Mason, A.L., Myers, R.P., Milkiewicz, P., Odin, J.A., Luketic, V.A., Bacon, B.R., Bodenheimer, H.C., Liakina, V., Vincent, C., Levy, C., Franke, A., Gregersen, P.K., Bossa, F., Gershwin, M.E., deAndrade, M., Amos, C.I., Lazaridis, K.N., Seldin, M.F., Siminovitch, K.A., 2012. Immunochip analyses identify a novel risk locus for primary biliary cirrhosis at 13q14, multiple independent associations at four established risk loci and epistasis between 1p31 and 7q32 risk variants. Hum. Mol. Genet. 21, 5209–5221. doi:10.1093/hmg/dds359

Juuti-Uusitalo, K., Mäki, M., Kaukinen, K., Collin, P., Visakorpi, T., Vihinen, M., Kainulainen, H., 2004. cDNA microarray analysis of gene expression in coeliac disease jejunal biopsy samples. J. Autoimmun. 22, 249–265. doi:10.1016/j.jaut.2003.12.001

Kadarmideen, H.N., Watson-haigh, N.S., 2012. Building gene co-expression networks using transcriptomics data for systems biology investigations: Comparison of methods using microarray data. Bioinformation 8, 855–861. doi:10.6026/97320630008855

Kagnoff, M.F., 2005. Overview and pathogenesis of celiac disease. Gastroenterology 128, S10–S18. doi:10.1053/j.gastro.2005.02.008

Kagnoff, M.F., Austin, R.K., Hubert, J.J., Bernardin, J.E., Kasarda, D.D., 1984. Possible role for a human adenovirus in the pathogenesis of celiac disease. J. Exp. Med. 160, 1544–1557. doi:10.1084/jem.160.5.1544

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30. doi:10.1093/nar/28.1.27

Karell, K., Louka, A.S., Moodie, S.J., Ascher, H., Clot, F., Greco, L., Ciclitira, P.J., Sollid, L.M., Partanen, J., European Genetics Cluster on Celiac Disease, 2003. HLA types in celiac disease patients not carrying the DQA1*05-DQB1*02 (DQ2) heterodimer: results from the European Genetics Cluster on Celiac Disease. Hum. Immunol. 64, 469–477.

Kempthorne, O., 1968. The correlation between relatives on the supposition of mendelian inheritance. Am. J. Hum. Genet. 20, 402–403.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C., 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245, 1073–1080. doi:10.1126/science.2570460

Kim, C.-Y., Quarsten, H., Bergseng, E., Khosla, C., Sollid, L.M., 2004. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc. Natl. Acad. Sci. 101, 4175–4179. doi:10.1073/pnas.0306885101

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. doi:10.1186/gb-2013-14-4-r36

Kim, E.H., Gasper, D.J., Lee, S.H., Plisch, E.H., Svaren, J., Suresh, M., 2014. Bach2 Regulates Homeostasis of Foxp3+ Regulatory T-cells and Protects against Fatal Lung Disease in Mice. J. Immunol. 192, 985–995. doi:10.4049/jimmunol.1302378

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.-Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh., J., 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308, 385–389. doi:10.1126/science.1109557

Koskinen, L.L.E., Einarsdottir, E., Dukes, E., Heap, G.A.R., Dubois, P., Korponay-Szabo, I.R., Kaukinen, K., Kurppa, K., Ziberna, F., Vatta, S., Not, T., Ventura, A., Sistonen, P., Ádány, R., Pocsai, Z., Széles, G., Mäki, M., Kere, J., Wijmenga, C., Heel, D.A. van, Saavalainen, P., 2009. Association study of the IL18RAP locus in three European populations with coeliac disease. Hum. Mol. Genet. 18, 1148–1155. doi:10.1093/hmg/ddn438

Koskinen, L.L.E., Einarsdottir, E., Korponay-Szabo, I.R., Kurppa, K., Kaukinen, K., Sistonen, P., Pocsai, Z., Széles, G., Ádány, R., Mäki, M., Kere, J., Saavalainen, P., 2009. Fine mapping of the CELIAC2 locus on chromosome 5q31-q33 in the Finnish and Hungarian populations. Tissue Antigens 74, 408–416. doi:10.1111/j.1399-0039.2009.01359.x

Kumar, V., Wijmenga, C., Withoff, S., 2012. From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Semin. Immunopathol. 34, 567–580. doi:10.1007/s00281-012-0312-1

Kunz, M., 2013. Lupus erythematosus. Part I: epidemiology, genetics and immunology. JDDG J. Dtsch. Dermatol. Ges. 11, 709–720. doi:10.1111/ddg.12165

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., Bastide, M. de la, Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A., Morgan, M.J., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. doi:10.1038/35057062

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. doi:10.1186/1471-2105-9-559

Lähdeaho, M.-L., Lehtinen, M., Rissa, H.-R., Hyöty, H., Reunala, T., Mäki, M., 1993. Antipeptide Antibodies to Adenovirus E1b Protein Indicate Enhanced Risk of Celiac Disease and Dermatitis herpetiformis. Int. Arch. Allergy Immunol. 101, 272–276. doi:10.1159/000236457

Lawler, M., Humphries, P., O'Farrelly, C., Hoey, H., Sheils, O., Jeffers, M., O'Briain, D.S., Kelleher, D., 1994. Adenovirus 12 E1A gene detection by polymerase chain reaction in both the normal and coeliac duodenum. Gut 35, 1226–1232. doi:10.1136/gut.35.9.1226

Lee, M.N., Ye, C., Villani, A.-C., Raj, T., Li, W., Eisenhaure, T.M., Imboywa, S.H., Chipendo, P.I., Ran, F.A., Slowikowski, K., Ward, L.D., Raddassi, K., McCabe, C., Lee, M.H., Frohlich, I.Y., Hafler, D.A., Kellis, M., Raychaudhuri, S., Zhang, F., Stranger, B.E., Benoist, C.O., De Jager, P.L., Regev, A., Hacohen, N., 2014. Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. Science 343, 1246980. doi:10.1126/science.1246980

Lee, P.H., O'Dushlaine, C., Thomas, B., Purcell, S.M., 2012. INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics 28, 1797–1799. doi:10.1093/bioinformatics/bts191

Lehtonen, R., Kuusela, V., 1986. Statistical efficiency of the Mini-Finland Health Survey's sampling design. Part 5. The execution of the Mini-Finland Health Survey (in Finnish, English summary). Hels. Turku Publ. Soc. Insur. Inst. Finland ML 65.

Lewis, N.R., Scott, B.B., 2010. Meta-analysis: deamidated gliadin peptide antibody and tissue transglutaminase antibody compared as screening tests for coeliac disease. Aliment. Pharmacol. Ther. 31, 73–81. doi:10.1111/j.1365-2036.2009.04110.x

Lewis, N.R., Scott, B.B., 2006. Systematic review: the use of serology to exclude or diagnose coeliac disease (a comparison of the endomysial and tissue transglutaminase antibody tests). Aliment. Pharmacol. Ther. 24, 47–54. doi:10.1111/j.1365-2036.2006.02967.x

Li, B., Tsoi, L.C., Swindell, W.R., Gudjonsson, J.E., Tejasvi, T., Johnston, A., Ding, J., Stuart, P.E., Xing, X., Kochkodan, J.J., Voorhees, J.J., Kang, H.M., Nair, R.P., Abecasis, G.R., Elder, J.T., 2014. Transcriptome Analysis of Psoriasis in A Large Case-Control Sample: Rna-Seq Provides Insights Into Disease Mechanisms. J. Invest. Dermatol. doi:10.1038/jid.2014.28

Lincoln, M.R., Ramagopalan, S.V., Chao, M.J., Herrera, B.M., DeLuca, G.C., Orton, S.-M., Dyment, D.A., Sadovnick, A.D., Ebers, G.C., 2009. Epistasis among HLA-DRB1, HLA-DQA1, and HLA-DQB1 loci determines multiple sclerosis susceptibility. Proc. Natl. Acad. Sci. 106, 7542–7547. doi:10.1073/pnas.0812664106

Lindquist, K.J., Jorgenson, E., Hoffmann, T.J., Witte, J.S., 2013. The Impact of Improved Microarray Coverage and Larger Sample Sizes on Future Genome-Wide Association Studies. Genet. Epidemiol. 37, 383–392. doi:10.1002/gepi.21724

Lionetti, E., Castellaneta, S., Francavilla, R., Pulvirenti, A., Tonutti, E., Amarri, S., Barbato, M., Barbera, C., Barera, G., Bellantoni, A., Castellano, E., Guariso, G.,

Limongelli, M.G., Pellegrino, S., Polloni, C., Ughi, C., Zuin, G., Fasano, A., Catassi, C., 2014. Introduction of Gluten, HLA Status, and the Risk of Celiac Disease in Children. N. Engl. J. Med. 371, 1295–1303. doi:10.1056/NEJMoa1400697

Liu, J., Juo, S.-H., Holopainen, P., Terwilliger, J., Tong, X., Grunn, A., Brito, M., Green, P., Mustalahti, K., Mäki, M., Gilliam, T.C., Partanen, J., 2002. Genomewide Linkage Analysis of Celiac Disease in Finnish Families. Am. J. Hum. Genet. 70, 51–59. doi:10.1086/338453

Liu, J., Zhang, H.-X., 2013. CTLA-4 Polymorphisms and Systemic Lupus Erythematosus: A Comprehensive Meta-Analysis. Genet. Test. Mol. Biomark. 17, 226–231. doi:10.1089/gtmb.2012.0302

Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R., Kane, J.P., Saccone, S., Worthington, J., Bruce, I., Kwok, P., Menter, A., Krueger, J., Barton, A., Saccone, N.L., Bowcock, A.M., 2008. A Genome-Wide Association Study of Psoriasis and Psoriatic Arthritis Identifies New Disease Loci. PLoS Genet 4, e1000041. doi:10.1371/journal.pgen.1000041

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15. doi:10.1186/s13059-014-0550-8

Lu, C., Xiao, C., Chen, G., Jiang, M., Zha, Q., Yan, X., Kong, W., Lu, A., 2012. Cold and heat pattern of rheumatoid arthritis in traditional Chinese medicine: distinct molecular signatures indentified by microarray expression profiles in CD4-positive T-cell. Rheumatol. Int. 32, 61–68. doi:10.1007/s00296-010-1546-7

LUDVIGSSON, J.F., RUBIO-TAPIA, A., CHOWDHARY, V., MURRAY, J.A., SIMARD, J.F., 2012. Increased Risk of Systemic Lupus Erythematosus in 29,000 Patients with Biopsy-verified Celiac Disease. J. Rheumatol. 39, 1964–1970. doi:10.3899/jrheum.120493

Ludvigsson, J.F., Rubio-Tapia, A., van Dyke, C.T., Melton, L.J., Zinsmeister, A.R., Lahr, B.D., Murray, J.A., 2013. Increasing Incidence of Celiac Disease in a North American Population. Am. J. Gastroenterol. 108, 818–824. doi:10.1038/ajg.2013.60

Lundin, K.E.A., Sollid, L.M., 2014. Advances in coeliac disease: Curr. Opin. Gastroenterol. 30, 154–162. doi:10.1097/MOG.0000000000000041

Lundin, K.E., Scott, H., Hansen, T., Paulsen, G., Halstensen, T.S., Fausa, O., Thorsby, E., Sollid, L.M., 1993. Gliadin-specific, HLA-DQ(alpha 1*0501,beta 1*0201) restricted T-cells isolated from the small intestinal mucosa of celiac disease patients. J. Exp. Med. 178, 187–196.

Luo, W., Brouwer, C., 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics 29, 1830–1831. doi:10.1093/bioinformatics/btt285

Lu, W., Quintero-Rivera, F., Fan, Y., Alkuraya, F.S., Donovan, D.J., Xi, Q., Turbe-Doan, A., Li, Q.-G., Campbell, C.G., Shanske, A.L., Sherr, E.H., Ahmad, A., Peters, R., Rilliet, B., Parvex, P., Bassuk, A.G., Harris, D.J., Ferguson, H., Kelly, C., Walsh, C.A., Gronostajski, R.M., Devriendt, K., Higgins, A., Ligon, A.H., Quade, B.J., Morton, C.C., Gusella, J.F., Maas, R.L., 2007. NFIA Haploinsufficiency Is Associated with a

CNS Malformation Syndrome and Urinary Tract Defects. PLoS Genet. 3, e80. doi:10.1371/journal.pgen.0030080

Madonna, S., Scarponi, C., Pallotta, S., Cavani, A., Albanesi, C., 2012. Anti-apoptotic effects of suppressor of cytokine signaling 3 and 1 in psoriasis. Cell Death Dis. 3, e334. doi:10.1038/cddis.2012.69

Mahajan, M.A., Murray, A., Samuels, H.H., 2002. NRC-Interacting Factor 1 Is a Novel Cotransducer That Interacts with and Regulates the Activity of the Nuclear Hormone Receptor Coactivator NRC. Mol. Cell. Biol. 22, 6883–6894. doi:10.1128/MCB.22.19.6883-6894.2002

Makharia, G.K., Verma, A.K., Amarchand, R., Bhatnagar, S., Das, P., Goswami, A., Bhatia, V., Ahuja, V., Datta Gupta, S., Anand, K., 2011. Prevalence of celiac disease in the northern part of India: A community based study. J. Gastroenterol. Hepatol. 26, 894–900. doi:10.1111/j.1440-1746.2010.06606.x

Malone, J.H., Oliver, B., 2011. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 9, 34. doi:10.1186/1741-7007-9-34

Mankaï, A., Landolsi, H., Chahed, A., Gueddah, L., Limem, M., Ben Abdessalem, M., Yacoub-Jemni, S., Ghannem, H., Jeddi, M., Ghedira, I., 2006. Celiac disease in Tunisia: serological screening in healthy blood donors. Pathol. Biol. (Paris) 54, 10–13. doi:10.1016/j.patbio.2005.02.005

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753. doi:10.1038/nature08494

Marsh, M., 1992. Mucosal pathology in gluten sensitivity. Oxf. UK Blackwell Sci pp 136–191.

Marsh, M.N., 1992. Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). Gastroenterology 102, 330–354.

Mattner, F., Fischer, S., Guckes, S., Jin, S., Kaulen, H., Schmitt, E., Rüde, E., Germann, T., 1993. The interleukin-12 subunit p40 specifically inhibits effects of the interleukin-12 heterodimer. Eur. J. Immunol. 23, 2202–2208. doi:10.1002/eji.1830230923

Maya, R., Gershwin, M.E., Shoenfeld, Y., 2008. Hepatitis B virus (HBV) and autoimmune disease. Clin. Rev. Allergy Immunol. 34, 85–102. doi:10.1007/s12016-007-8013-6

M, B., Ap, D., Nb, M., 2009. Prevalence of celiac disease in north Indian children. Indian Pediatr. 46, 415–417.

McClymont, S.A., Putnam, A.L., Lee, M.R., Esensten, J.H., Liu, W., Hulme, M.A., Hoffmüller, U., Baron, U., Olek, S., Bluestone, J.A., Brusko, T.M., 2011. Plasticity of Human Regulatory T-cells in Healthy Subjects and Patients with Type 1 Diabetes. J. Immunol. 186, 3918–3926. doi:10.4049/jimmunol.1003099

McGowan, K.E., Lyon, M.E., Butzner, J.D., 2008. Celiac Disease and IgA Deficiency: Complications of Serological Testing Approaches Encountered in the Clinic. Clin. Chem. 54, 1203–1209. doi:10.1373/clinchem.2008.103606

Medhat, A., El Salam, N.A., Hassany, S.M., Hussein, H.I., Blum, H.E., 2011. Frequency of celiac disease in Egyptian patients with chronic diarrhea: Endoscopic, histopathologic and immunologic evaluation. J. Physiol. Pathophysiol. 2, 1–5.

Meresse, B., Chen, Z., Ciszewski, C., Tretiakova, M., Bhagat, G., Krausz, T.N., Raulet, D.H., Lanier, L.L., Groh, V., Spies, T., Ebert, E.C., Green, P.H., Jabri, B., 2004. Coordinated Induction by IL15 of a TCR-Independent NKG2D Signaling Pathway Converts CTL into Lymphokine-Activated Killer Cells in Celiac Disease. Immunity 21, 357–366. doi:10.1016/j.immuni.2004.06.020

Meresse, B., Malamut, G., Cerf-Bensussan, N., 2012. Celiac Disease: An Immunological Jigsaw. Immunity 36, 907–919. doi:10.1016/j.immuni.2012.06.006

Miklos, G.L.G., Maleszka, R., 2004. Microarray reality checks in the context of a complex disease. Nat. Biotechnol. 22, 615–621. doi:10.1038/nbt965

Miyake, N., Chilton, J., Psatha, M., Cheng, L., Andrews, C., Chan, W.-M., Law, K., Crosier, M., Lindsay, S., Cheung, M., Allen, J., Gutowski, N.J., Ellard, S., Young, E., Iannaccone, A., Appukuttan, B., Stout, J.T., Christiansen, S., Ciccarelli, M.L., Baldi, A., Campioni, M., Zenteno, J.C., Davenport, D., Mariani, L.E., Sahin, M., Guthrie, S., Engle, E.C., 2008. Human CHN1 mutations hyperactivate α2-chimaerin and cause Duane's retraction syndrome. Science 321, 839–843. doi:10.1126/science.1156121

Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E., Heinzmann, A., Simma, B., Frischer, T., Willis-Owen, S.A.G., Wong, K.C.C., Illig, T., Vogelberg, C., Weiland, S.K., von Mutius, E., Abecasis, G.R., Farrall, M., Gut, I.G., Lathrop, G.M., Cookson, W.O.C., 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448, 470–473. doi:10.1038/nature06014

Montgomery, S.B., Dermitzakis, E.T., 2011. From expression QTLs to personalized transcriptomics. Nat. Rev. Genet. 12, 277–282. doi:10.1038/nrg2969

Moore, F., Colli, M.L., Cnop, M., Esteve, M.I., Cardozo, A.K., Cunha, D.A., Bugliani, M., Marchetti, P., Eizirik, D.L., 2009. PTPN2, a candidate gene for type 1 diabetes, modulates interferon-gamma-induced pancreatic beta-cell apoptosis. Diabetes 58, 1283–1291. doi:10.2337/db08-1510

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628. doi:10.1038/nmeth.1226

Munafo, M.R., Flint, J., 2004. Meta-analysis of genetic association studies. Trends Genet. 20.

Mustalahti, K., Catassi, C., Reunanen, A., Fabiani, E., Heier, M., McMillan, S., Murray, L., Metzger, M.-H., Gasparin, M., Bravi, E., Mäki, M., Coeliac EU Cluster, Project Epidemiology, 2010. The prevalence of celiac disease in Europe: results of a centralized, international mass screening project. Ann. Med. 42, 587–595. doi:10.3109/07853890.2010.505931

Myléus, A., Ivarsson, A., Webb, C., Danielsson, L., Hernell, O., Högberg, L., Karlsson, E., Lagerqvist, C., Norström, F., Rosén, A., Sandström, O., Stenhammar, L., Stenlund, H., Wall, S., Carlsson, A., 2009. Celiac Disease Revealed in 3% of Swedish 12-year-olds Born During an Epidemic. J. Pediatr. Gastroenterol. Nutr. 49, 170–176. doi:10.1097/MPG.0b013e31818c52cc

Nacu, S., Yuan, W., Kan, Z., Bhatt, D., Rivers, C.S., Stinson, J., Peters, B.A., Modrusan, Z., Jung, K., Seshagiri, S., Wu, T.D., 2011. Deep RNA-sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. BMC Med. Genomics 4, 11. doi:10.1186/1755-8794-4-11

Naluai, A.T., Nilsson, S., Gudjónsdóttir, A.H., Louka, A.S., Ascher, H., Ek, J., Hallberg, B., Samuelsson, L., Kristiansson, B., Martinsson, T., Nerman, O., Sollid, L.M., Wahlström, J., 2001. Genome-wide linkage analysis of Scandinavian affected sib-pairs supports presence of susceptibility loci for celiac disease on chromosomes 5 and 11. Eur. J. Hum. Genet. EJHG 9, 938–944. doi:10.1038/sj.ejhg.5200752

Nilsen, E.M., Jahnsen, F.L., Lundin, K.E.A., Johansen, F.-E., Fausa, O., Sollid, L.M., Jahnsen, J., Scott, H., Brandtzaeg, P., 1998. Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. Gastroenterology 115, 551–563. doi:10.1016/S0016-5085(98)70134-9

Nisticò, L., Buzzetti, R., Pritchard, L.E., Auwera, B.V. der, Giovannini, C., Bosi, E., Larrad, M.T.M., Rios, M.S., Chow, C.C., Cockram, C.S., Jacobs, K., Mijovic, C., Bain, S.C., Barnett, A.H., Vandewalle, C.L., Schuit, F., Gorus, F.K., Tosi, R., Pozzilli, P., Todd, J.A., 1996. The CTLA-4 Gene Region of Chromosome 2q33 Is Linked to, and Associated with, Type 1 Diabetes. Hum. Mol. Genet. 5, 1075–1080. doi:10.1093/hmg/5.7.1075

Nistico, L., Fagnani, C., Coto, I., Percopo, S., Cotichini, R., Limongelli, M.G., Paparo, F., D'Alfonso, S., Giordano, M., Sferlazzas, C., Magazzu, G., Momigliano-Richiardi, P., Greco, L., Stazi, M.A., 2006. Concordance, disease progression, and heritability of coeliac disease in Italian twins. Gut 55, 803–808. doi:10.1136/gut.2005.083964

Oberhuber, G., Granditsch, G., Vogelsang, H., 1999. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur. J. Gastroenterol. Hepatol. 11.

Okoniewski, M.J., Miller, C.J., 2006. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7, 276. doi:10.1186/1471-2105-7-276

Olaussen, R.W., Johansen, F.E., Lundin, K.E.A., Jahnsen, J., Brandtzaeg, P., Farstad, I.N., 2002. Interferon-γ-Secreting T-cells Localize to the Epithelium in Coeliac Disease. Scand. J. Immunol. 56, 652–664. doi:10.1046/j.1365-3083.2002.01195.x

Park, S.W., Park, S.J., Shin, J.I., 2013. Another Possible Underlying Mechanism for the Positive Association Between Celiac Disease and Systemic Lupus Erythematosus: The Role of Interleukin 21. J. Rheumatol. 40, 1619–1619. doi:10.3899/jrheum.130426

Pascual, V., Medrano, L., Bodas, A., Lopez-Palacios, N., Cuevas, D., Jimenez, N., Gonzalez-Perez, B., Salazar, I., Nunez, C., 2015. Gene expression studies in paediatric and adult celiac disease.

Pasinetti, G.M., 2001. Use of cDNA microarray in the search for molecular markers involved in the onset of Alzheimer's disease dementia. J. Neurosci. Res. 65, 471–476. doi:10.1002/jnr.1176

Pharoah, P.D.P., Tsai, Y.-Y., Ramus, S.J., Phelan, C.M., Goode, E.L., Lawrenson, K., Price, M., Fridley, B.L., Tyrer, J.P., Shen, H., Weber, R., Karevan, R., Larson, M.C., Song, H., Tessier, D.C., Bacot, F., Vincent, D., Cunningham, J.M., Dennis, J., Dicks, E., Aben, K.K., Anton-Culver, H., Antonenkova, N., Armasu, S.M., Baglietto, L., Bandera, E.V., Beckmann, M.W., Birrer, M.J., Bloom, G., Bogdanova, N., Brenton, J.D., Brinton, L.A., Brooks-Wilson, A., Brown, R., Butzow, R., Campbell, I., Carney, M.E., Carvalho, R.S., Chang-Claude, J., Chen, Y.A., Chen, Z., Chow, W.-H., Cicek, M.S., Coetzee, G., Cook, L.S., Cramer, D.W., Cybulski, C., Dansonka-Mieszkowska, A., Despierre, E., Doherty, J.A., Dörk, T., du Bois, A., Dürst, M., Eccles, D., Edwards, R., Ekici, A.B., Fasching, P.A., Fenstermacher, D., Flanagan, J., Gao, Y.-T., Garcia-Closas, M., Gentry-Maharaj, A., Giles, G., Gjyshi, A., Gore, M., Gronwald, J., Guo, Q., Halle, M.K., Harter, P., Hein, A., Heitz, F., Hillemanns, P., Hoatlin, M., Høgdall, E., Høgdall, C.K., Hosono, S., Jakubowska, A., Jensen, A., Kalli, K.R., Karlan, B.Y., Kelemen, L.E., Kiemeney, L.A., Kjaer, S.K., Konecny, G.E., Krakstad, C., Kupryjanczyk, J., Lambrechts, D., Lambrechts, S., Le, N.D., Lee, N., Lee, J., Leminen, A., Lim, B.K., Lissowska, J., Lubiński, J., Lundvall, L., Lurie, G., Massuger, L.F.A.G., Matsuo, K., McGuire, V., McLaughlin, J.R., Menon, U., Modugno, F., Moysich, K.B., Nakanishi, T., Narod, S.A., Ness, R.B., Nevanlinna, H., Nickels, S., Noushmehr, H., Odunsi, K., Olson, S., Orlow, I., Paul, J., Pejovic, T., Pelttari, L.M., Permuth-Wey, J., Pike, M.C., Poole, E.M., Qu, X., Risch, H.A., Rodriguez-Rodriguez, L., Rossing, M.A., Rudolph, A., Runnebaum, I., Rzepecka, I.K., Salvesen, H.B., Schwaab, I., Severi, G., Shen, H., Shridhar, V., Shu, X.-O., Sieh, W., Southey, M.C., Spellman, P., Tajima, K., Teo, S.-H., Terry, K.L., Thompson, P.J., Timorek, A., Tworoger, S.S., van Altena, A.M., Berg, D.V.D., Vergote, I., Vierkant, R.A., Vitonis, A.F., Wang-Gohrke, S., Wentzensen, N., Whittemore, A.S., Wik, E., Winterhoff, B., Woo, Y.L., Wu, A.H., Yang, H.P., Zheng, W., Ziogas, A., Zulkifli, F., Goodman, M.T., Hall, P., Easton, D.F., Pearce, C.L., Berchuck, A., Chenevix-Trench, G., Iversen, E., Monteiro, A.N.A., Gayther, S.A., Schildkraut, J.M., Sellers, T.A., 2013. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat. Genet. 45, 362–370e2. doi:10.1038/ng.2564

Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., Pritchard, J.K., 2010. Understanding mechanisms underlying human gene expression variation with RNA-sequencing. Nature 464, 768–772. doi:10.1038/nature08872

Plaza-Izurieta, L., Castellanos-Rubio, A., Irastorza, I., Fernández-Jimenez, N., Gutierrez, G., Bilbao, J.R., 2011. Revisiting genome wide association studies (GWAS) in coeliac disease: replication study in Spanish population and expression analysis of candidate genes. J. Med. Genet. 48, 493–496. doi:10.1136/jmg.2011.089714

Plaza-Izurieta, L., Fernandez-Jimenez, N., Irastorza, I., Jauregi-Miguel, A., Romero-Garmendia, I., Vitoria, J.C., Bilbao, J.R., 2014. Expression analysis in intestinal mucosa reveals complex relations among genes under the association peaks in celiac disease. Eur. J. Hum. Genet. doi:10.1038/ejhg.2014.244

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P., (Leader), S.M.P., Ruderfer, D.M., McQuillin, A., Morris, D.W., O'Dushlaine, C.T., Corvin, A., Holmans, P.A., O'Donovan, M.C., Macgregor, S., Gurling, H., Blackwood, D.H.R., Craddock, N.J., Gill, M., Hultman, C.M., Kirov, G.K., Lichtenstein, P., Muir, W.J., Owen, M.J., Pato, C.N., Scolnick, E.M., Clair, D.S., (Leader), P.S., Williams, N.M., Georgieva, L., Nikolov, I., Norton, N., Williams, H., Toncheva, D., Milanova, V., Thelander, E.F., Sullivan, P., O'Dushlaine, C.T., Kenny, E., Quinn, E.M., Choudhury, K., Datta, S., Pimm, J., Thirumalai, S., Puri, V., Krasucki, R., Lawrence, J., Quested, D., Bass, N., Crombie, C., Fraser, G., Kuan, S.L., Walker, N., McGhee, K.A., Pickard, B., Malloy, P., Maclean, A.W., Beck, M.V., Pato, M.T., Medeiros, H., Middleton, F., Carvalho, C., Morley, C., Fanous, A., Conti, D., Knowles, J.A., Ferreira, C.P., Macedo, A., Azevedo, M.H., Kirby, A.N., Ferreira, M.A.R., Daly, M.J., Chambert, K., Kuruvilla, F., Gabriel, S.B., Ardlie, K., Moran, J.L., 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460, 748–752. doi:10.1038/nature08185

Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., Imboywa, S., Korff, A.V., Okada, Y., Patsopoulos, N.A., Davis, S., McCabe, C., Paik, H., Srivastava, G.P., Raychaudhuri, S., Hafler, D.A., Koller, D., Regev, A., Hacohen, N., Mathis, D., Benoist, C., Stranger, B.E., Jager, P.L.D., 2014. Polarization of the Effects of Autoimmune and Neurodegenerative Risk Alleles in Leukocytes. Science 344, 519–523. doi:10.1126/science.1249547

Rampertab, S.D., Pooran, N., Brar, P., Singh, P., Green, P.H.R., 2006. Trends in the Presentation of Celiac Disease. Am. J. Med. 119, 355.e9–355.e14. doi:10.1016/j.amjmed.2005.08.044

Rashtak, S., Ettore, M.W., Homburger, H.A., Murray, J.A., 2008. Comparative Usefulness of Deamidated Gliadin Antibodies in the Diagnosis of Celiac Disease. Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc. 6, 426–370. doi:10.1016/j.cgh.2007.12.030

Rawashdeh, M.O., Khalil, B., Raweily, E., 1996. Celiac disease in Arabs. J. Pediatr. Gastroenterol. Nutr. 23, 415–418.

Ricaño-Ponce, I., Wijmenga, C., 2013. Mapping of Immune-Mediated Disease Genes. Annu. Rev. Genomics Hum. Genet. 14, 325–353. doi:10.1146/annurev-genom-091212-153450

Ricaño-Ponce, I., Wijmenga, C., Gutierrez-Achury, J., n.d. Genetics of celiac disease. Best Pract. Res. Clin. Gastroenterol. doi:10.1016/j.bpg.2015.04.004

Rivas, M.A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C.K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J.M., Kuruvilla, F., Lagacé, C., Neale, B., Lo, K.S., Schumm, P., Törkvist, L., Dubinsky, M., Brant, S.R., Silverberg, M., Duerr, R.H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M.,

McGovern, D.P.B., Cho, J.H., Rioux, J.D., Xavier, R.J., Daly, M.J., 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43, 1066–1073. doi:10.1038/ng.952

Rogers, L., Varadan, V., Palkar, A., Miskimen, K., Williams, N., Morton, M., Maximuk, S., Vinayak, S., Sloan, A., Barnholtz-Sloan, J., Gilmore, H., Harris, L., 2014. Bm-28biological Pathways Associated with Breast Cancer Metastasis to the Brain. Neuro-Oncol. 16, v38–v38. doi:10.1093/neuonc/nou240.28

Romanos, J., Rosén, A., Kumar, V., Trynka, G., Franke, L., Szperl, A., Gutierrez-Achury, J., Diemen, C.C. van, Kanninga, R., Jankipersadsing, S.A., Steck, A., Eisenbarth, G., Heel, D.A. van, Cukrowska, B., Bruno, V., Mazzilli, M.C., Núñez, C., Bilbao, J.R., Mearin, M.L., Barisani, D., Rewers, M., Norris, J.M., Ivarsson, A., Boezen, H.M., Liu, E., Wijmenga, C., Scerri, C., Koltai, T., Kolaček, S., Mišak, Z., Abdović, S., Koletzko, S., Osiander, G., Werkstetter, K., Mummert, E., Korponay-Szabo, I.R., Gyimesi, J., Shamir, R., Hartman, C., Bravi, E., Poles, M., Auricchio, R., Limongelli, G.G.G., Greco, L., Troncone, R., Villanacci, V., Bindels, J.G., Brand, R., Kupper, B.F., Esch, C.E.H., Hopman, E.G., Koning, F., Kooy-Winkelaar, Y., Marvelde, C. te, Putter, H., Stoopman, E., Vriezinga, S., Sollid, L.M., Ráki, M., Chmielewska, A., Dziechciarz, P., Pieścik-Lech, M., Szajewska, H., Szaflarska-Szczepanik, A., Castillejo, G., Capilla, A., Varea, V., Ribes-Koninckx, C., Lopez, A., Crespo, P., Martinez, E., Polanco, I., Högberg, L., Stenhammar, L., Carlsson, A., Webb, C., Hammarroth, S., Hernell, O., Lagerqvist, C., Myléus, A., Nordyke, K., Norström, F., Sandström, O., Wall, S., Karlsson, E., 2014. Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants. Gut 63, 415–422. doi:10.1136/gutjnl-2012-304110

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., Daly, M.J., International Inflammatory Bowel Disease Genetics Constortium, 2011. Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. PLoS Genet 7, e1001273. doi:10.1371/journal.pgen.1001273

Rossi, T.M., Albini, C.H., Kumar, V., 1993. Incidence of celiac disease identified by the presence of serum endomysial antibodies in children with chronic diarrhea, short stature, or insulin-dependent diabetes mellitus. J. Pediatr. 123, 262–264. doi:10.1016/S0022-3476(05)81699-3

Roychoudhuri, R., Hirahara, K., Mousavi, K., Clever, D., Bonelli, M., Klebanoff, C., Sartorelli, V., Kanno, Y., Gattinoni, L., Wang, E., Liu, H., Marincola, F., Kazuhiko, I., O'Shea, J., Restifo, N.P., 2013. BACH2 represses effector programmes to stabilize Treg-mediated immune homeostasis - a new target in tumor immunotherapy? J. Immunother. Cancer 1, 1–1. doi:10.1186/2051-1426-1-S1-O14

Ruggeri, C., La Masa, A.T., Rudi, S., Squadrito, G., Di Pasquale, G., Maimone, S., Caccamo, G., Pellegrino, S., Raimondo, G., Magazzù, G., 2008. Celiac disease and non-organ-specific autoantibodies in patients with chronic hepatitis C virus infection. Dig. Dis. Sci. 53, 2151–2155. doi:10.1007/s10620-007-0146-1

Ryder, L.P., Andersen, E., Sveigaard, A., 1979. HLA and Disease Registry. Third Rep. Munksgaard Cph.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.-Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Etten, W.J.V., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S., Altshuler, D., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928–933. doi:10.1038/35057149

Sakamoto, H., Friel, A.M., Wood, A.W., Guo, L., Ilic, A., Seiden, M.V., Chung, D.C., Lynch, M.P., Serikawa, T., Munro, E., Oliva, E., Orsulic, S., Kirley, S.D., Foster, R., Zukerberg, L.R., Rueda, B.R., 2008. Mechanisms of Cables 1 gene inactivation in human ovarian cancer development. Cancer Biol. Ther. 7, 180–188. doi:10.4161/cbt.7.2.5253

Scharl, M., Paul, G., Weber, A., Jung, B.C., Docherty, M.J., Hausmann, M., Rogler, G., Barrett, K.E., McCole, D.F., 2009. Protection of Epithelial Barrier Function by the Crohn's Disease Associated Gene Protein Tyrosine Phosphatase N2. Gastroenterology 137, 2030–2040.e5. doi:10.1053/j.gastro.2009.07.078

Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421–427. doi:10.1038/nature13595

Schunkert, H., König, I.R., Kathiresan, S., Reilly, M.P., Assimes, T.L., Holm, H., Preuss, M., Stewart, A.F.R., Barbalic, M., Gieger, C., Absher, D., Aherrahrou, Z., Allayee, H., Altshuler, D., Anand, S.S., Andersen, K., Anderson, J.L., Ardissino, D., Ball, S.G., Balmforth, A.J., Barnes, T.A., Becker, D.M., Becker, L.C., Berger, K., Bis, J.C., Boekholdt, S.M., Boerwinkle, E., Braund, P.S., Brown, M.J., Burnett, M.S., Buysschaert, I., Cardiogenics, Carlquist, J.F., Chen, L., Cichon, S., Codd, V., Davies, R.W., Dedoussis, G., Dehghan, A., Demissie, S., Devaney, J.M., Diemert, P., Do, R., Doering, A., Eifert, S., Mokhtari, N.E.E., Ellis, S.G., Elosua, R., Engert, J.C., Epstein, S.E., de Faire, U., Fischer, M., Folsom, A.R., Freyer, J., Gigante, B., Girelli, D., Gretarsdottir, S., Gudnason, V., Gulcher, J.R., Halperin, E., Hammond, N., Hazen, S.L., Hofman, A., Horne, B.D., Illig, T., Iribarren, C., Jones, G.T., Jukema, J.W., Kaiser, M.A., Kaplan, L.M., Kastelein, J.J.P., Khaw, K.-T., Knowles, J.W., Kolovou, G., Kong, A., Laaksonen, R., Lambrechts, D., Leander, K., Lettre, G., Li, M., Lieb, W., Loley, C., Lotery, A.J., Mannucci, P.M., Maouche, S., Martinelli, N., McKeown, P.P., Meisinger, C., Meitinger, T., Melander, O., Merlini, P.A., Mooser, V., Morgan, T., Mühleisen, T.W., Muhlestein, J.B., Münzel, T., Musunuru, K., Nahrstaedt, J., Nelson, C.P., Nöthen, M.M., Olivieri, O., Patel, R.S., Patterson, C.C., Peters, A., Peyvandi, F., Qu, L., Quyyumi, A.A., Rader, D.J., Rallidis, L.S., Rice, C., Rosendaal, F.R., Rubin, D., Salomaa, V., Sampietro, M.L., Sandhu, M.S., Schadt, E., Schäfer, A., Schillert, A., Schreiber, S., Schrezenmeir, J., Schwartz, S.M., Siscovick, D.S., Sivananthan, M., Sivapalaratnam, S., Smith, A., Smith, T.B., Snoep, J.D., Soranzo, N., Spertus, J.A., Stark, K., Stirrups, K., Stoll, M., Tang, W.H.W., Tennstedt, S., Thorgeirsson, G., Thorleifsson, G., Tomaszewski, M., Uitterlinden, A.G., van Rij, A.M., Voight, B.F.,

Wareham, N.J., Wells, G.A., Wichmann, H.-E., Wild, P.S., Willenborg, C., Witteman, J.C.M., Wright, B.J., Ye, S., Zeller, T., Ziegler, A., Cambien, F., Goodall, A.H., Cupples, L.A., Quertermous, T., März, W., Hengstenberg, C., Blankenberg, S., Ouwehand, W.H., Hall, A.S., Deloukas, P., Thompson, J.R., Stefansson, K., Roberts, R., Thorsteinsdottir, U., O'Donnell, C.J., McPherson, R., Erdmann, J., Consortium, the Cardi., 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat. Genet. 43, 333–338. doi:10.1038/ng.784

Schuppan, D., Junker, Y., Barisani, D., 2009. Celiac Disease: From Pathogenesis to Novel Therapies. Gastroenterology 137, 1912–1933. doi:10.1053/j.gastro.2009.09.008

Seegers, D., Borm, M.E.A., van Belzen, M.J., Mulder, C.J.J., Bailing, J., Crusius, J.B.A., Meijer, J.W.R., Wijmenga, C., Peña, A.S., Bouma, G., 2003. IL12B and IRF1 gene polymorphisms and susceptibility to celiac disease. Eur. J. Immunogenetics Off. J. Br. Soc. Histocompat. Immunogenetics 30, 421–425.

Seumois, G., Chavez, L., Gerasimova, A., Lienhard, M., Omran, N., Kalinke, L., Vedanayagam, M., Ganesan, A.P.V., Chawla, A., Djukanović, R., Ansel, K.M., Peters, B., Rao, A., Vijayanand, P., 2014. Epigenomic analysis of primary human T-cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. Nat. Immunol. 15, 777–788. doi:10.1038/ni.2937

Shahbazkhani, B., Malekzadeh, R., Sotoudeh, M., Moghadam, K.F., Farhadi, M., Ansari, R., Elahyfar, A., Rostami, K., 2003. High prevalence of coeliac disease in apparently healthy Iranian blood donors: Eur. J. Gastroenterol. Hepatol. 15, 475–478. doi:10.1097/01.meg.0000059118.41030.96

Shamir, R., Lerner, A., Shinar, E., Lahat, N., Sobel, E., Bar-or, R., Kerner, H., Eliakim, R., 2002. The use of a single serological marker underestimates the prevalence of celiac disease in Israel: a study of blood donors. Am. J. Gastroenterol. 97, 2589–2594. doi:10.1111/j.1572-0241.2002.06028.x

Shan, L., Molberg, Ø., Parrot, I., Hausch, F., Filiz, F., Gray, G.M., Sollid, L.M., Khosla, C., 2002. Structural Basis for Gluten Intolerance in Celiac Sprue. Science 297, 2275–2279. doi:10.1126/science.1074129

Shendure, J., 2008. The beginning of the end for microarrays? Nat. Methods 5, 585–587. doi:10.1038/nmeth0708-585

Shoenfeld, Y., Zandman-Goddard, G., Stojanovich, L., Cutolo, M., Amital, H., Levy, Y., Abu-Shakra, M., Barzilai, O., Berkun, Y., Blank, M., de Carvalho, J.F., Doria, A., Gilburd, B., Katz, U., Krause, I., Langevitz, P., Orbach, H., Pordeus, V., Ram, M., Toubi, E., Sherer, Y., 2008. The mosaic of autoimmunity: hormonal and environmental factors involved in autoimmune diseases--2008. Isr. Med. Assoc. J. IMAJ 10, 8–12.

Simon, R., Roychowdhury, S., 2013. Implementing personalized cancer genomics in clinical trials. Nat. Rev. Drug Discov. 12, 358–369. doi:10.1038/nrd3979

Smyth, D.J., Plagnol, V., Walker, N.M., Cooper, J.D., Downes, K., Yang, J.H.M., Howson, J.M.M., Stevens, H., McManus, R., Wijmenga, C., Heap, G.A., Dubois, P.C., Clayton, D.G., Hunt, K.A., van Heel, D.A., Todd, J.A., 2008. Shared and Distinct

Genetic Variants in Type 1 Diabetes and Celiac Disease. N. Engl. J. Med. 359, 2767–2777. doi:10.1056/NEJMoa0807917

Sollid, L.M., Jabri, B., 2013a. Triggers and drivers of autoimmunity: lessons from coeliac disease. Nat. Rev. Immunol. 13, 294–302. doi:10.1038/nri3407

Sollid, L.M., Jabri, B., 2013b. Triggers and drivers of autoimmunity: lessons from coeliac disease. Nat. Rev. Immunol. 13, 294–302. doi:10.1038/nri3407

Sollid, L.M., Lie, B.A., 2005. Celiac disease genetics: current concepts and practical applications. Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc. 3, 843–851.

Sollid, L.M., Qiao, S.-W., Anderson, R.P., Gianfrani, C., Koning, F., 2012. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. Immunogenetics 64, 455–460. doi:10.1007/s00251-012-0599-z

Song, G.G., Kim, J.-H., Kim, Y.H., Lee, Y.H., 2013. Association between CTLA-4 polymorphisms and susceptibility to Celiac disease: A meta-analysis. Hum. Immunol. 74, 1214–1218. doi:10.1016/j.humimm.2013.05.014

Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., Amos, C.I., Ardlie, K.G., Barton, A., Bowes, J., Brouwer, E., Burtt, N.P., Catanese, J.J., Coblyn, J., Coenen, M.J., Costenbader, K.H., Criswell, L.A., Crusius, J.B.A., Cui, J., de Bakker, P.I.W., De Jager, P.L., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L.J., Huizinga, T.W.J., Kastner, D.L., Ke, X., Lee, A.T., Liu, X., Martin, P., Morgan, A.W., Padyukov, L., Posthumus, M.D., Radstake, T.R., Reid, D.M., Seielstad, M., Seldin, M.F., Shadick, N.A., Steer, S., Tak, P.P., Thomson, W., van der Helm-van Mil, A.H.M., van der Horst-Bruinsma, I.E., van der Schoot, C.E., van Riel, P.L., Weinblatt, M.E., Wilson, A.G., Wolbink, G.J., Wordsworth, P., Wijmenga, C., Karlson, E.W., Toes, R.E.M., de Vries, N., Begovich, A.B., Worthington, J., Siminovitch, K.A., Gregersen, P.K., Klareskog, L., Plenge, R.M., 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet. 42, 508–514. doi:10.1038/ng.582

Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A.S., Consortium, D.G.R. and M., Consortium, M.I.G., Kathiresan, S., Wijmenga, C., Gregersen, P.K., Alfredsson, L., Siminovitch, K.A., Worthington, J., Bakker, P.I.W. de, Raychaudhuri, S., Plenge, R.M., 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat. Genet. 44, 483–489. doi:10.1038/ng.2232

Stene, L.C., Honeyman, M.C., Hoffenberg, E.J., Haas, J.E., Sokol, R.J., Emery, L., Taki, I., Norris, J.M., Erlich, H.A., Eisenbarth, G.S., Rewers, M., 2006. Rotavirus Infection Frequency and Risk of Celiac Disease Autoimmunity in Early Childhood: A Longitudinal Study. Am. J. Gastroenterol. 101, 2333–2340. doi:10.1111/j.1572-0241.2006.00741.x

Subramaniam, S., Napolitano, F., Mealer, R.G., Kim, S., Errico, F., Barrow, R., Shahani, N., Tyagi, R., Snyder, S.H., Usiello, A., 2012. Rhes, a striatal-enriched small G protein, mediates mTOR signaling and L-DOPA-induced dyskinesia. Nat. Neurosci. 15, 191–193. doi:10.1038/nn.2994

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102, 15545–15550. doi:10.1073/pnas.0506580102

Talley, N.J., M, V., Tm, P., Ha, C., 3rd, M.L., 1994. Epidemiology of celiac sprue: a community-based study. Am. J. Gastroenterol. 89, 843–846.

Tan, D., Kirley, S., Li, Q., Ramnath, N., Slocum, H.K., Brooks, J.S., Wu, C.-L., Zukerberg, L.R., 2003. Loss of cables protein expression in human non–small cell lung cancer: A tissue microarray study. Hum. Pathol. 34, 143–149. doi:10.1053/hupa.2003.26

Tatar, G., Elsurer, R., Simsek, H., Balaban, Y.H., Hascelik, G., Ozcebe, O.I., Buyukasik, Y., Sokmensuer, C., 2004. Screening of Tissue Transglutaminase Antibody in Healthy Blood Donors for Celiac Disease Screening in the Turkish Population. Dig. Dis. Sci. 49, 1479–1484. doi:10.1023/B:DDAS.0000042250.59327.91

Tengah, D.S.N.A.P., Wills, A.J., Holmes, G.K.T., 2002. Neurological complications of coeliac disease. Postgrad. Med. J. 78, 393–398. doi:10.1136/pmj.78.921.393

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature 437, 1299–1320. doi:10.1038/nature04226

The International Multiple Sclerosis Genetics Consortium (IMSGC), 2010. Evidence for Polygenic Susceptibility to Multiple Sclerosis--The Shape of Things to Come. Am. J. Hum. Genet. 86, 621–625. doi:10.1016/j.ajhg.2010.02.027

Thomas, C.C., 1945. On the coeliac affection. in: Classic descriptions of disease 600–601.

Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B.J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S.I., Calle, E.E., Feigelson, H.S., Thun, M.J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F.R., Giovannucci, E., Willett, W.C., Cussenot, O., Valeri, A., Andriole, G.L., Crawford, E.D., Tucker, M., Gerhard, D.S., Fraumeni, J.F., Hoover, R., Hayes, R.B., Hunter, D.J., Chanock, S.J., 2008. Multiple loci identified in a genome-wide association study of prostate cancer. Nat. Genet. 40, 310–315. doi:10.1038/ng.91

Todd, J.A., Walker, N.M., Cooper, J.D., Smyth, D.J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S.F., Payne, F., Lowe, C.E., Szeszko, J.S., Hafler, J.P., Zeitels, L., Yang, J.H.M., Vella, A., Nutland, S., Stevens, H.E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L.J., Healy, B., Burren, O.S., Lam, A.A.C., Ovington, N.R., Allen, J., Adlem, E., Leung, H.-T., Wallace, C., Howson, J.M.M., Guja, C., Ionescu-Tîrgovişte, C., Simmonds, M.J., Heward, J.M., Gough, S.C.L., Dunger, D.B., Wicker, L.S., Clayton, D.G., 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat. Genet. 39, 857–864. doi:10.1038/ng2068

Tonutti, E., Bizzaro, N., 2014. Diagnosis and classification of celiac disease and gluten sensitivity. Autoimmun. Rev., Diagnostic criteria in Autoimmune diseases9th

International Congress on Autoimmunity 13, 472–476. doi:10.1016/j.autrev.2014.01.043

Torinsson Naluai, Å., Nilsson, S., Samuelsson, L., Gudjónsdóttir, A. h., Ascher, H., Ek, J., Hallberg, B., Kristiansson, B., Martinsson, T., Nerman, O., Sollid, L. m., Wahlström, J., 2000. The CTLA4/CD28 gene region on chromosome 2q33 confers susceptibility to celiac disease in a way possibly distinct from that of type 1 diabetes and other chronic inflammatory disorders. Tissue Antigens 56, 350–355. doi:10.1034/j.1399-0039.2000.560407.x

Trynka, G., Hunt, K.A., Bockett, N.A., Romanos, J., Mistry, V., Szperl, A., Bakker, S.F., Bardella, M.T., Bhaw-Rosun, L., Castillejo, G., de la Concha, E.G., de Almeida, R.C., Dias, K.-R.M., van Diemen, C.C., Dubois, P.C.A., Duerr, R.H., Edkins, S., Franke, L., Fransen, K., Gutierrez, J., Heap, G.A.R., Hrdlickova, B., Hunt, S., Plaza Izurieta, L., Izzo, V., Joosten, L.A.B., Langford, C., Mazzilli, M.C., Mein, C.A., Midah, V., Mitrovic, M., Mora, B., Morelli, M., Nutland, S., Núñez, C., Onengut-Gumuscu, S., Pearce, K., Platteel, M., Polanco, I., Potter, S., Ribes-Koninckx, C., Ricaño-Ponce, I., Rich, S.S., Rybak, A., Santiago, J.L., Senapati, S., Sood, A., Szajewska, H., Troncone, R., Varadé, J., Wallace, C., Wolters, V.M., Zhernakova, A., Spanish Consortium on the Genetics of Coeliac Disease (CEGEC), PreventCD Study Group, Wellcome Trust Case Control Consortium (WTCCC), Thelma, B.K., Cukrowska, B., Urcelay, E., Bilbao, J.R., Mearin, M.L., Barisani, D., Barrett, J.C., Plagnol, V., Deloukas, P., Wijmenga, C., van Heel, D.A., 2011. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat. Genet. 43, 1193–1201. doi:10.1038/ng.998

Trynka, G., Zhernakova, A., Romanos, J., Franke, L., Hunt, K.A., Turner, G., Bruinenberg, M., Heap, G.A., Platteel, M., Ryan, A.W., Kovel, C. de, Holmes, G.K.T., Howdle, P.D., Walters, J.R.F., Sanders, D.S., Mulder, C.J.J., Mearin, M.L., Verbeek, W.H.M., Trimble, V., Stevens, F.M., Kelleher, D., Barisani, D., Bardella, M.T., McManus, R., Heel, D.A. van, Wijmenga, C., 2009. Coeliac disease-associated risk variants in TNFAIP3 and REL implicate altered NF-κB signalling. Gut 58, 1078–1083. doi:10.1136/gut.2008.169052

Tsukumo, S., Unno, M., Muto, A., Takeuchi, A., Kometani, K., Kurosaki, T., Igarashi, K., Saito, T., 2013. Bach2 maintains T-cells in a naive state by suppressing effector memory-related genes. Proc. Natl. Acad. Sci. 110, 10735–10740. doi:10.1073/pnas.1306691110

Twine, N.A., Janitz, K., Wilkins, M.R., Janitz, M., 2011. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. PLoS ONE 6, e16266. doi:10.1371/journal.pone.0016266

Ujvari, A., Chamnongpol, S., Kourennaia, O., Myers, K., Bressler, V., Hubbard, J., Schweitzer, A., Kubu, C., 2014. Concordance of Affymetrix GeneChip® Human Transcriptome Array 2.0 and USB® VeriQuest™ real-time PCR data (LB207). FASEB J. 28, LB207.

Upchurch, H.F., Conway, E., Patterson, M.K., Maxwell, M.D., 1991. Localization of cellular transglutaminase on the extracellular matrix after wounding:

Characteristics of the matrix bound enzyme. J. Cell. Physiol. 149, 375–382. doi:10.1002/jcp.1041490304

Vahedi, G., Kanno, Y., Furumoto, Y., Jiang, K., Parker, S.C.J., Erdos, M.R., Davis, S.R., Roychoudhuri, R., Restifo, N.P., Gadina, M., Tang, Z., Ruan, Y., Collins, F.S., Sartorelli, V., O'Shea, J.J., 2015. Super-enhancers delineate disease-associated regulatory nodes in T-cells. Nature 520, 558–562. doi:10.1038/nature14154

Vaira, V., Roncoroni, L., Barisani, D., Gaudioso, G., Bosari, S., Bulfamante, G., Doneda, L., Conte, D., Tomba, C., Bardella, M.T., Ferrero, S., Locatelli, M., Elli, L., 2014. microRNA profiles in coeliac patients distinguish different clinical phenotypes and are modulated by gliadin peptides in primary duodenal fibroblasts. Clin. Sci. Lond. Engl. 1979 126, 417–423. doi:10.1042/CS20130248

Van Belzen, M.J., Meijer, J.W.., Sandkuijl, L.A., Bardoel, A.F.., Mulder, C.J.., Pearson, P.L., Houwen, R.H.., Wijmenga, C., 2003. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology 125, 1032–1041. doi:10.1016/S0016-5085(03)01205-8

Van Eerdewegh, P., Little, R.D., Dupuis, J., Del Mastro, R.G., Falls, K., Simon, J., Torrey, D., Pandit, S., McKenny, J., Braunschweiger, K., Walsh, A., Liu, Z., Hayward, B., Folz, C., Manning, S.P., Bawa, A., Saracino, L., Thackston, M., Benchekroun, Y., Capparell, N., Wang, M., Adair, R., Feng, Y., Dubois, J., FitzGerald, M.G., Huang, H., Gibson, R., Allen, K.M., Pedan, A., Danzig, M.R., Umland, S.P., Egan, R.W., Cuss, F.M., Rorke, S., Clough, J.B., Holloway, J.W., Holgate, S.T., Keith, T.P., 2002. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. Nature 418, 426–430. doi:10.1038/nature00878

van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C.N.M., Bethel, G., Holmes, G.K.T., Feighery, C., Jewell, D., Kelleher, D., Kumar, P., Travis, S., Walters, J.R., Sanders, D.S., Howdle, P., Swift, J., Playford, R.J., McLaren, W.M., Mearin, M.L., Mulder, C.J., McManus, R., McGinnis, R., Cardon, L.R., Deloukas, P., Wijmenga, C., 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827–829. doi:10.1038/ng2058

van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C.N.M., Bethel, G., Holmes, G.K.T., Feighery, C., Jewell, D., Kelleher, D., Kumar, P., Travis, S., Walters, J.R., Sanders, D.S., Howdle, P., Swift, J., Playford, R.J., McLaren, W.M., Mearin, M.L., Mulder, C.J., McManus, R., McGinnis, R., Cardon, L.R., Deloukas, P., Wijmenga, C., 2007. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. 39, 827–829. doi:10.1038/ng2058

van Heel, D.A., Hunt, K., Greco, L., Wijmenga, C., 2005. Genetics in coeliac disease. Best Pract. Res. Clin. Gastroenterol. 19, 323–339. doi:10.1016/j.bpg.2005.01.001

Vivas, S., Ruiz de Morales, J.M., Fernandez, M., Hernando, M., Herrero, B., Casqueiro, J., Gutierrez, S., 2008. Age-Related Clinical, Serological, and Histopathological Features of Celiac Disease. Am. J. Gastroenterol. 103, 2360–2365. doi:10.1111/j.1572-0241.2008.01977.x

Vriezinga, S.L., Auricchio, R., Bravi, E., Castillejo, G., Chmielewska, A., Crespo Escobar, P., Kolaček, S., Koletzko, S., Korponay-Szabo, I.R., Mummert, E., Polanco, I.,

Putter, H., Ribes-Koninckx, C., Shamir, R., Szajewska, H., Werkstetter, K., Greco, L., Gyimesi, J., Hartman, C., Hogen Esch, C., Hopman, E., Ivarsson, A., Koltai, T., Koning, F., Martinez-Ojinaga, E., te Marvelde, C., Pavic, A., Romanos, J., Stoopman, E., Villanacci, V., Wijmenga, C., Troncone, R., Mearin, M.L., 2014. Randomized Feeding Intervention in Infants at High Risk for Celiac Disease. N. Engl. J. Med. 371, 1304–1315. doi:10.1056/NEJMoa1404172

Walker, M.M., Murray, J.A., 2011. An update in the diagnosis of coeliac disease. Histopathology 59, 166–179. doi:10.1111/j.1365-2559.2010.03680.x

Walker, M.M., Murray, J.A., Ronkainen, J., Aro, P., Storskrubb, T., D'Amato, M., Lahr, B., Talley, N.J., Agreus, L., 2010. Detection of Celiac Disease and Lymphocytic Enteropathy by Parallel Serology and Histopathology in a Population-Based Study. Gastroenterology 139, 112–119. doi:10.1053/j.gastro.2010.04.007

Walker-Smith, J.A., Guandalini, S., Schmitz, J., Shmerling, D.H., Visakorpi, J.K., 1990. Revised criteria for diagnosis of coeliac disease. Arch. Dis. Child. U. K.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B., 2008. Alternative Isoform Regulation in Human Tissue Transcriptomes. Nature 456, 470–476. doi:10.1038/nature07509

Wang, K.S., Frank, D.A., Ritz, J., 2000. Interleukin-2 enhances the response of natural killer cells to interleukin-12 through up-regulation of the interleukin-12 receptor and STAT4. Blood 95, 3183–3190.

Wang, X.-B., Zheng, C.-Y., Giscombe, R., Lefvert, A.K., 2001. Regulation of Surface and Intracellular Expression of CTLA-4 on Human Peripheral T-cells. Scand. J. Immunol. 54, 453–458. doi:10.1046/j.1365-3083.2001.00985.x

Wang, Y., Ghaffari, N., Johnson, C.D., Braga-Neto, U.M., Wang, H., Chen, R., Zhou, H., 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 12, S5. doi:10.1186/1471-2105-12-S10-S5

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63. doi:10.1038/nrg2484

Wapenaar, M.C., van Belzen, M.J., Fransen, J.H., Sarasqueta, A.F., Houwen, R.H.J., Meijer, J.W.R., Mulder, C.J.J., Wijmenga, C., 2004. The interferon gamma gene in celiac disease: augmented expression correlates with tissue damage but no evidence for genetic susceptibility. J. Autoimmun. 23, 183–190. doi:10.1016/j.jaut.2004.05.004

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P.M., Baldassano, R.N., Hakonarson, H., 2013. Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. Am. J. Hum. Genet. 92, 1008–1012. doi:10.1016/j.ajhg.2013.05.002

Willer, C.J., Li, Y., Abecasis, G.R., 2010. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma. Oxf. Engl. 26, 2190–2191. doi:10.1093/bioinformatics/btq340

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J. 'an, Kutalik, Z., Amin, N., Buchkovich, M.L., Croteau-Chonka, D.C., Day, F.R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A.U., Karjalainen, J., Lo, K.S., Locke, A.E., Mägi, R., Mihailov, E., Porcu, E., Randall,

J.C., Scherag, A., Vinkhuyzen, A.A.E., Westra, H.-J., Winkler, T.W., Workalemahu, T., Zhao, J.H., Absher, D., Albrecht, E., Anderson, D., Baron, J., Beekman, M., Demirkan, A., Ehret, G.B., Feenstra, B., Feitosa, M.F., Fischer, K., Fraser, R.M., Goel, A., Gong, J., Justice, A.E., Kanoni, S., Kleber, M.E., Kristiansson, K., Lim, U., Lotay, V., Lui, J.C., Mangino, M., Leach, I.M., Medina-Gomez, C., Nalls, M.A., Nyholt, D.R., Palmer, C.D., Pasko, D., Pechlivanis, S., Prokopenko, I., Ried, J.S., Ripke, S., Shungin, D., Stancáková, A., Strawbridge, R.J., Sung, Y.J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S.W., van Setten, J., Van Vliet-Ostaptchouk, J.V., Wang, Z., Yengo, L., Zhang, W., Afzal, U., Ärnlöv, J., Arscott, G.M., Bandinelli, S., Barrett, A., Bellis, C., Bennett, A.J., Berne, C., Blüher, M., Bolton, J.L., Böttcher, Y., Boyd, H.A., Bruinenberg, M., Buckley, B.M., Buyske, S., Caspersen, I.H., Chines, P.S., Clarke, R., Claudi-Boehm, S., Cooper, M., Daw, E.W., De Jong, P.A., Deelen, J., Delgado, G., Denny, J.C., Dhonukshe-Rutten, R., Dimitriou, M., Doney, A.S.F., Dörr, M., Eklund, N., Eury, E., Folkersen, L., Garcia, M.E., Geller, F., Giedraitis, V., Go, A.S., Grallert, H., Grammer, T.B., Gräßler, J., Grönberg, H., de Groot, L.C.P.G.M., Groves, C.J., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hannemann, A., Hartman, C.A., Hassinen, M., Hayward, C., Heard-Costa, N.L., Helmer, Q., Hemani, G., Henders, A.K., Hillege, H.L., Hlatky, M.A., Hoffmann, W., Hoffmann, P., Holmen, O., Houwing-Duistermaat, J.J., Illig, T., Isaacs, A., James, A.L., Jeff, J., Johansen, B., Johansson, Å., Jolley, J., Juliusdottir, T., Junttila, J., Kho, A.N., Kinnunen, L., Klopp, N., Kocher, T., Kratzer, W., Lichtner, P., Lind, L., Lindström, J., Lobbens, S., Lorentzon, M., Lu, Y., Lyssenko, V., Magnusson, P.K.E., Mahajan, A., Maillard, M., McArdle, W.L., McKenzie, C.A., McLachlan, S., McLaren, P.J., Menni, C., Merger, S., Milani, L., Moayyeri, A., Monda, K.L., Morken, M.A., Müller, G., Müller-Nurasyid, M., Musk, A.W., Narisu, N., Nauck, M., Nolte, I.M., Nöthen, M.M., Oozageer, L., Pilz, S., Rayner, N.W., Renstrom, F., Robertson, N.R., Rose, L.M., Roussel, R., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F.R., Schunkert, H., Scott, R.A., Sehmi, J., Seufferlein, T., Shi, J., Silventoinen, K., Smit, J.H., Smith, A.V., Smolonska, J., Stanton, A.V., Stirrups, K., Stott, D.J., Stringham, H.M., Sundström, J., Swertz, M.A., Syvänen, A.-C., Tayo, B.O., Thorleifsson, G., Tyrer, J.P., van Dijk, S., van Schoor, N.M., van der Velde, N., van Heemst, D., van Oort, F.V.A., Vermeulen, S.H., Verweij, N., Vonk, J.M., Waite, L.L., Waldenberger, M., Wennauer, R., Wilkens, L.R., Willenborg, C., Wilsgaard, T., Wojczynski, M.K., Wong, A., Wright, A.F., Zhang, Q., Arveiler, D., Bakker, S.J.L., Beilby, J., Bergman, R.N., Bergmann, S., Biffar, R., Blangero, J., Boomsma, D.I., Bornstein, S.R., Bovet, P., Brambilla, P., Brown, M.J., Campbell, H., Caulfield, M.J., Chakravarti, A., Collins, R., Collins, F.S., Crawford, D.C., Cupples, L.A., Danesh, J., de Faire, U., den Ruijter, H.M., Erbel, R., Erdmann, J., Eriksson, J.G., Farrall, M., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N.G., Forrester, T., Gansevoort, R.T., Gejman, P.V., Gieger, C., Golay, A., Gottesman, O., Gudnason, V., Gyllensten, U., Haas, D.W., Hall, A.S., Harris, T.B., Hattersley, A.T., Heath, A.C., Hengstenberg, C., Hicks, A.A., Hindorff, L.A., Hingorani, A.D., Hofman, A., Hovingh, G.K., Humphries, S.E., Hunt, S.C., Hypponen, E., Jacobs, K.B., Jarvelin, M.-R., Jousilahti, P., Jula, A.M., Kaprio, J., Kastelein, J.J.P., Kayser, M., Kee, F.,

Keinanen-Kiukaanniemi, S.M., Kiemeney, L.A., Kooner, J.S., Kooperberg, C., Koskinen, S., Kovacs, P., Kraja, A.T., Kumari, M., Kuusisto, J., Lakka, T.A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lupoli, S., Madden, P.A.F., Männistö, S., Manunta, P., Marette, A., Matise, T.C., McKnight, B., Meitinger, T., Moll, F.L., Montgomery, G.W., Morris, A.D., Morris, A.P., Murray, J.C., Nelis, M., Ohlsson, C., Oldehinkel, A.J., Ong, K.K., Ouwehand, W.H., Pasterkamp, G., Peters, A., Pramstaller, P.P., Price, J.F., Qi, L., Raitakari, O.T., Rankinen, T., Rao, D.C., Rice, T.K., Ritchie, M., Rudan, I., Salomaa, V., Samani, N.J., Saramies, J., Sarzynski, M.A., Schwarz, P.E.H., Sebert, S., Sever, P., Shuldiner, A.R., Sinisalo, J., Steinthorsdottir, V., Stolk, R.P., Tardif, J.-C., Tönjes, A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., The Electronic Medical Records and Genomics (eMERGE) Consortium, The MIGen Consortium, The PAGE Consortium, The LifeLines Cohort Study, Amouyel, P., Asselbergs, F.W., Assimes, T.L., Bochud, M., Boehm, B.O., Boerwinkle, E., Bottinger, E.P., Bouchard, C., Cauchi, S., Chambers, J.C., Chanock, S.J., Cooper, R.S., de Bakker, P.I.W., Dedoussis, G., Ferrucci, L., Franks, P.W., Froguel, P., Groop, L.C., Haiman, C.A., Hamsten, A., Hayes, M.G., Hui, J., Hunter, D.J., Hveem, K., Jukema, J.W., Kaplan, R.C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N.G., März, W., Melbye, M., Moebus, S., Munroe, P.B., Njølstad, I., Oostra, B.A., Palmer, C.N.A., Pedersen, N.L., Perola, M., Pérusse, L., Peters, U., Powell, J.E., Power, C., Quertermous, T., Rauramaa, R., Reinmaa, E., Ridker, P.M., Rivadeneira, F., Rotter, J.I., Saaristo, T.E., Saleheen, D., Schlessinger, D., Slagboom, P.E., Snieder, H., Spector, T.D., Strauch, K., Stumvoll, M., Tuomilehto, J., Uusitupa, M., van der Harst, P., Völzke, H., Walker, M., Wareham, N.J., Watkins, H., Wichmann, H.-E., Wilson, J.F., Zanen, P., Deloukas, P., Heid, I.M., Lindgren, C.M., Mohlke, K.L., Speliotes, E.K., Thorsteinsdottir, U., Barroso, I., Fox, C.S., North, K.E., Strachan, D.P., Beckmann, J.S., Berndt, S.I., Boehnke, M., Borecki, I.B., McCarthy, M.I., Metspalu, A., Stefansson, K., Uitterlinden, A.G., van Duijn, C.M., Franke, L., Willer, C.J., Price, A.L., Lettre, G., Loos, R.J.F., Weedon, M.N., Ingelsson, E., O'Connell, J.R., Abecasis, G.R., Chasman, D.I., Goddard, M.E., Visscher, P.M., Hirschhorn, J.N., Frayling, T.M., 2014. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. 46, 1173–1186. doi:10.1038/ng.3097

Wright, H.L., Moots, R.J., Bucknall, R.C., Edwards, S.W., 2010. Neutrophil function in inflammation and inflammatory diseases. Rheumatology 49, 1618–1631. doi:10.1093/rheumatology/keq045

Wu, C.-L., Kirley, S.D., Xiao, H., Chuang, Y., Chung, D.C., Zukerberg, L.R., 2001. Cables Enhances Cdk2 Tyrosine 15 Phosphorylation by Wee1, Inhibits Cell Growth, and Is Lost in Many Human Colon and Squamous Cancers. Cancer Res. 61, 7325–7332.

Xu, D., Chan, W.L., Leung, B.P., Hunter, D., Schulz, K., Carter, R.W., McInnes, I.B., Robinson, J.H., Liew, F.Y., 1998. Selective Expression and Functions of Interleukin 18 Receptor on T Helper (Th) Type 1 but not Th2 Cells. J. Exp. Med. 188, 1485–1492. doi:10.1084/jem.188.8.1485

Yachha, S., Poddar, U., 2006. Celiac disease in India. Indian J. Gastroenterol. Off. J. Indian Soc. Gastroenterol. 26, 230–237.

Yamana, J., Yamamura, M., Okamoto, A., Aita, T., Iwahashi, M., Sunahori, K., Makino, H., 2004. Resistance to IL-10 inhibition of interferon gamma production and expression of suppressor of cytokine signaling 1 in CD4+ T-cells from patients with rheumatoid arthritis. Arthritis Res. Ther. 6, R567–R577. doi:10.1186/ar1445

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A.A.E., Lee, S.H., Robinson, M.R., Perry, J.R.B., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., The LifeLines Cohort Study, Esko, T., Milani, L., Mägi, R., Metspalu, A., Hamsten, A., Magnusson, P.K.E., Pedersen, N.L., Ingelsson, E., Soranzo, N., Keller, M.C., Wray, N.R., Goddard, M.E., Visscher, P.M., 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nat. Genet. advance online publication. doi:10.1038/ng.3390

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M., 2010. Common SNPs explain a large proportion of heritability for human height. Nat. Genet. 42, 565–569. doi:10.1038/ng.608

Yang, Y.J., Baltus, A.E., Mathew, R.S., Murphy, E.A., Evrony, G.D., Gonzalez, D.M., Wang, E.P., Marshall-Walker, C.A., Barry, B.J., Murn, J., Tatarakis, A., Mahajan, M.A., Samuels, H.H., Shi, Y., Golden, J.A., Mahajnah, M., Shenhav, R., Walsh, C.A., 2012. Microcephaly Gene Links Trithorax and REST/NRSF to Control Neural Stem Cell Proliferation and Differentiation. Cell 151, 1097–1112. doi:10.1016/j.cell.2012.10.043

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 11, R14. doi:10.1186/gb-2010-11-2-r14

Zanoni, G., Navone, R., Lunardi, C., Tridente, G., Bason, C., Sivori, S., Beri, R., Dolcino, M., Valletta, E., Corrocher, R., Puccetti, A., 2006. In Celiac Disease, a Subset of Autoantibodies against Transglutaminase Binds Toll-Like Receptor 4 and Induces Activation of Monocytes. PLoS Med. 3. doi:10.1371/journal.pmed.0030358

Zanzi, D., Stefanile, R., Santagata, S., Iaffaldano, L., Iaquinto, G., Giardullo, N., Lania, G., Vigliano, I., Vera, A.R., Ferrara, K., Auricchio, S., Troncone, R., Mazzarella, G., 2011. IL-15 Interferes With Suppressive Activity of Intestinal Regulatory T-cells Expanded in Celiac Disease. Am. J. Gastroenterol. 106, 1308–1317. doi:10.1038/ajg.2011.80

Zhernakova, A., Alizadeh, B.Z., Bevova, M., van Leeuwen, M.A., Coenen, M.J.H., Franke, B., Franke, L., Posthumus, M.D., van Heel, D.A., van der Steege, G., Radstake, T.R.D.J., Barrera, P., Roep, B.O., Koeleman, B.P.C., Wijmenga, C., 2007. Novel Association in Chromosome 4q27 Region with Rheumatoid Arthritis and Confirmation of Type 1 Diabetes Point to a General Risk Locus for Autoimmune Diseases. Am. J. Hum. Genet. 81, 1284–1288. doi:10.1086/522037

Zhernakova, A., Stahl, E.A., Trynka, G., Raychaudhuri, S., Festen, E.A., Franke, L., Westra, H.-J., Fehrmann, R.S.N., Kurreeman, F.A.S., Thomson, B., Gupta, N.,

Romanos, J., McManus, R., Ryan, A.W., Turner, G., Brouwer, E., Posthumus, M.D., Remmers, E.F., Tucci, F., Toes, R., Grandone, E., Mazzilli, M.C., Rybak, A., Cukrowska, B., Coenen, M.J.H., Radstake, T.R.D.J., van Riel, P.L.C.M., Li, Y., de Bakker, P.I.W., Gregersen, P.K., Worthington, J., Siminovitch, K.A., Klareskog, L, Huizinga, T.W.J., Wijmenga, C., Plenge, R.M., 2011. Meta-Analysis of Genome-Wide Association Studies in Celiac Disease and Rheumatoid Arthritis Identifies Fourteen Non-HLA Shared Loci. PLoS Genet 7, e1002004. doi:10.1371/journal.pgen.1002004

Zhong, F., McCombs, C.C., Olson, J.M., Elston, R.C., Stevens, F.M., McCarthy, C.F., Michalski, J.P., 1996. An autosomal screen for genes that predispose to celiac disease in the western counties of Ireland. Nat. Genet. 14, 329–333. doi:10.1038/ng1196-329

Zondervan, K.T., Cardon, L.R., 2004. The complex interplay among factors that influence allelic association. Nat. Rev. Genet. 5, 89–100. doi:10.1038/nrg1270

Zuk, O., Hechter, E., Sunyaev, S.R., Lander, E.S., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. 109, 1193–1198. doi:10.1073/pnas.1119675109