Genetic risk variants in intestinal inflammatory disorders

Dubois, Patrick Charles Alexander

Genetic risk variants in intestinal inflammatory disorders

Patrick Charles Alexander Dubois

PhD Thesis
2010

Centre for Digestive Diseases
Barts and the London School of Medicine & Dentistry
Queen Mary, University of London

Statement of Originality

The work in this thesis is entirely my own unless stated otherwise. The genome wide association studies presented in **Chapters 3** and **4** have involved collaboration with large numbers of other researchers whose contributions, including analyses and figures are acknowledged in the text.

For the genetic study on coeliac disease (**Chapter 3**), I coordinated all aspects of the project, together with David van Heel. I assisted in UK2 sample recruitment. I extracted DNA from blood and saliva samples from the majority of the 1,922 case individuals in theUK2 collection; some samples had been extracted for a previous study. I performed GWAS genotyping for the majority of UK2 case samples and oversaw all parts of the genotyping process. I was assisted in genotyping by Dr Alex Curtotti at The Genome Centre, Barts and the London School of Medicine & Dentistry. I coordinated follow-up genotyping of samples for USA, Spanish, Irish and Hungarian collections. I performed genotyping for these samples together with Karen A Hunt, Nicolas Bockett and Vanisha Mistry at Barts and the London and with Muddassar Mirza and Efterpi Papouli at King's College London School of Medicine. Genotyping and initial quality controls of Polish and Italian follow-up samples were performed at the University of Groningen. I performed genotype calling (using a customized algorithm supplied by Lude Franke, University of Groningen and Barts and the London) for UK1, UK2 and Finnish collections in the GWAS. I performed genotyping quality control and case: control association analyses for UK1, UK2, Finnish collections separately and for the whole study after including Italian and Dutch GWAS collections, where initial genotyping quality controls were performed by Gosia Trynka, University of Groningen. I performed all quality controls and association analyses for combined datasets in both GWAS and follow-up stages. Lude Franke (University of Groningen and Barts and the London) performed an expression quantitative trait meta-analysis: analysis for the coeliac project was performed in collaboration with David van Heel and I.

I conceived the study of azathioprine and 6-mercaptopurine induced pancreatitis (**Chapter 4**). I designed the study together with David van Heel. I recruited individuals from Barts and the London and extracted DNA from these individuals. Samples from other centres were recruited locally and DNA supplied. For all samples I quantified DNA and performed GWAS genotyping. I performed all quality control and association analyses.

Experiments in **Chapter 5** were conceived and designed by me together with David van Heel**.** Experiments and data analyses are my own work unless stated. Whole genome gene expression microarray assays were performed at Barts and the London Genome centre as a service. The expression quantitative trait meta-analysis data for **Chapter 5** was supplied by Lude Franke, University of Groningen on request.

Abstract

This thesis includes work on the genetics of intestinal inflammatory disorders, concentrating on coeliac disease and Crohn's disease. It explores how common genetic variants influence risk of complex phenotypes including immunological intolerance to gluten (coeliac disease) and intolerance to therapeutic agents (azathioprine and mercaptopurine) used in the treatment of intestinal inflammatory diseases. Finally it presents work aiming to move from genetic associations with complex phenotypes to understanding of how these variants modulate immunological processes.

Results of a large genome wide association study that identified more than 13 new genetic risk regions influencing susceptibility to coeliac disease are presented. Results of a genome wide association study of azathioprine and 6-mercaptopurine-induced pancreatitis in inflammatory bowel disease-affected individuals are presented. Finally, a cell cytokine release assay for the prostaglandin EP4 receptor was developed, with a view to investigating how SNPs associated with Crohn's disease in the 5p13.1 region influence EP4 receptor signalling and contribute to disease pathogenesis. This work highlights some of the challenges in moving from SNP-disease associations identified in GWASs to understanding how genetic variants change biological processes.

Table of Contents

Tables List

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Figures List

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Acknowledgements

This research was supported by a Clinical Research Training Fellowship awarded by the Medical Research Council. The coeliac disease genome wide association study (**Chapter 3**) was funded by a grant from the Wellcome Trust. The study was supported by Coeliac UK.

I am very grateful to all who have collaborated with me on this research, both in the UK and internationally. Thank you to Graham Heap and Karen Hunt in the lab for all their help and advice throughout my PhD studies and to Nick Bockett and Vanisha Mistry who joined the lab more recently for their help. I am grateful to Parveen Kumar for her expert clinical guidance and Tom Macdonald, my second PhD supervisor for his encouragement and constructive criticism, particularly with research presented in **Chapter 5**. Finally, a big thank you to my supervisor, David van Heel, for his invaluable guidance over the last three years and for the opportunity to work on such an interesting subject.

Chapter 1 Genetics of complex traits

Over the last five years research in complex disease genetics has been dominated by the publication of more than 600 genome wide association studies (GWASs), leading to the identification in many complex traits of tens of independent susceptibility loci (Hindorff 2010). These studies represent a culmination of several key international collaborative research efforts. The sequencing and assembly of a reference human genome, published in 2001, set the foundations for efforts to build databases of common genetic variation mapped to the human genome (Lander, Linton et al. 2001; Venter, Adams et al. 2001). The International HapMap project, in particular, genotyped a reference set of common single nucleotide variants, initially in 4 human populations, and showed which combinations of these variants were commonly inherited together (The International HapMap Consortium 2005). This key advance demonstrated that common haplotype variation could be captured by (inferred from) a much reduced set of haplotype-tagging SNPs. Thus, while there are an estimated 8 million SNPs with minor allele frequency greater than 5% in humans, a set of just 550,000 SNPs are highly correlated with 88% of these SNPs in individuals of northern European descent (Frazer, Murray et al. 2009). The ability to genotype hundreds of thousands (and currently millions) of SNPs, in parallel, became possible with advances in genotyping microarray technologies over a similar time period. Thus, by 2005 genome-wide ascertainment of a substantial fraction of common genetic variation became available with genotyping arrays that included assays for hundreds of thousands of SNPs. First generation GWASs had immediate success in identifying novel susceptibility loci (Klein, Zeiss et al. 2005; Duerr, Taylor et al. 2006). An early realisation was that for many disorders GWAS findings would lead to radical re-evaluations of the pathogenesis of these conditions. Thus one of the very first GWASs associated variants in the complement factor H gene with age-related macular degeneration (Klein, Zeiss et al. 2005). This was a startling insight, suggesting that complement mediated inflammation was involved in the pathogenesis of a condition previously thought to occur through non-inflammatory pathways. Many of the loci identified by these early studies have been called the low hanging fruit: these are typically the disease loci harbouring common variants with the largest effect sizes. Subsequently, GWASs and meta-analyses using more samples and more genetic markers have penetrated more deeply to reveal a long tail of genetic susceptibility variants of progressively weaker effect size in many complex diseases.

Despite their tremendous insights, the limits of GWASs as tools for understanding complex diseases have become clear. Many researchers have expressed disappointment that the fraction of the genetically determined component of complex diseases that is accounted for by GWAS discoveries is low, up to 20% at best. This has been framed as a problem of "missing heritability", with several explanations mooted. GWASs have been designed to assay common genetic variation and as such may be blind to disease-causing rare variants. What is clear is that currently our understanding of the genetic basis of complex diseases remains far from complete. Strategies to move towards a more complete understanding are discussed in this chapter. In many cases GWASs have robustly identified a region of association, but have been unable to identify causal variants or even the causal genes. This is a clear current limitation on the biological interpretation of GWAS findings. SNPs directly assayed in genome wide association studies are a small fraction of all human genetic variations and therefore in most cases causal variants have not been directly assayed. Association studies have so far proven limited in their ability to differentiate between hundreds or thousands of variants in a genomic susceptibility region. This is mainly a consequence of the strong linkage disequilibrium that exists within regions identified by GWASs. This chapter discusses some of the genetics approaches that may advance this area, allowing the identification of causal genetic variants.

This thesis presents research that has applied genome wide association methods and functional approaches to investigate common genetic variants in two intestinal inflammatory disorders, Crohn's disease and coeliac disease. Both are classic, heritable complex disorders. In both diseases, known environmental factors are necessary for intestinal inflammation- dietary gluten in coeliac disease and the intestinal microbiota in Crohn's disease, but genetic factors are critical in determining the host response. Thirdly, a genome wide association study of azathioprine-induced pancreatitis is presented. Here also an environmental factor, the drug, triggers inflammation, and though the heritability of this condition is unknown, it was hypothesized that genetic variation would be an important determinant of the risk.

1.1 The causes of complex disorders

For the simplest genetic and environmentally determined traits (e.g. single gene disorders, drug overdose) a minimal set of conditions that constitute the complete and sufficient causes of the trait can be identified. However, most human traits and most common diseases are complex, arising from multiple genetic and environmental causes. It has been proposed that the complete set of causal mechanisms for complex disorders consists of not one or a few but many distinct combinations of risk factors that lead to disease development, with major risk factors emerging in multiple of these combinations. Under this model, there would exist combinations of environmental and genetic risk factors that in each case inevitably lead (i.e. are sufficient for) the development of disease (Rothman and Greenland 2005). The number of these possible combinations is likely to be extremely large, as indicated from the tens of genetic loci already implicated in many common disorders (Janssens and van Duijn 2008). For example, 12 hypothetical bi-allelic genetic loci that could form causal combinations, each with 3 genotypes produce 3^{12} combinations (531,441). As the number of genetic and environmental factors scales, it becomes unlikely that individuals sharing a complex phenotype also share exactly the same combination of causal factors (Janssens and van Duijn 2008). Thus, complex diseases are likely to be manifestations of multiple, only partially shared combinations of genetic and environmental causes. This heterogeneity and the large number of causal factors greatly complicate efforts to arrive at a complete understanding of the causes of complex diseases. This will particularly be true if, as seems likely, the effects of causal factors on risk are not always independent. For example, specific HLA-DQ gene combinations and gluten exposure are both necessary factors for coeliac disease (Karell, Louka et al. 2003). Multiple non-HLA susceptibility variants have also been identified, but logically their effects on disease causation are contingent on the presence of the necessary factors. When interactions between putative causal factors are not understood, the power to detect these factors will be reduced.

In the absence of knowledge of the causal combinations for complex diseases, we can hope instead to begin by identifying those factors that emerge in multiple causal combinations, i.e. those that show greatest differentiation between individuals with the disease and those without. These sorts of recurring genetic factors can reveal important pathogenetic mechanisms that can inform strategies to develop new and better treatments for complex

diseases. Thus insights into disease biology and pathogenesis are the major motivation for research into the causes of complex diseases.

In contrast, as discussed further below, the predictive value of genetic risk factors in complex diseases is much less certain. Already several authors have warned that even a complete understanding of the heritable fraction of complex diseases would in most cases not enable very impressive risk prediction (Janssens and van Duijn 2008; Clayton 2009; Daly, Donaldson et al. 2009; Kraft, Wacholder et al. 2009). On the other hand, the value of genetic risk profiling may be greater in certain settings (e.g. prediction of drug responses) and is also determined by the nature and effectiveness of available interventions. Thus for certain conditions (e.g. some cancers) even a modest improvement in prediction may be sufficient to warrant changes in practice (e.g. starting surveillance programs at an earlier age). Prospective evaluation of genetic risk models will be required in multiple independent populations to establish the value of these applications.

1.2 Searching for the causes of complex disorders

1.2.1 Theoretical advantages of studying genetic causes over environmental causes

Environmental variation (the differences in all environmental factors to which individuals are exposed from the point of formation of the zygote) is practically infinite. Study of the environmental causes of diseases therefore has the potential to remain forever incomplete, with researchers seeking to form and test hypotheses without ever being able to comprehensively assay all of the environmental variation to which individuals are exposed. Many environmental variables can not be easily measured (e.g. *in utero* variables) and are not constant over time. Furthermore, the study of the environmental causes of disease is also limited by the difficulty of performing prospective studies of sufficient duration, particularly for diseases that may develop over many years.

In contrast genetic variation is both finite and knowable. Inherited (germline) genetic variations precede even the earliest developmental events and remain almost entirely constant throughout life. Thus, even in adulthood we can assay genetic variants and know that these exposures have occurred prior to and throughout the development of the individual. This fact confers a significant advantage to researchers, enabling us to infer causality for genetic variants that correlate with disease phenotypes, as it avoids completely those retrospective biases that are prevalent in many environmental studies. Genetic variants also conveniently manifest as quantum measures (0,1,2 or more copies of an allele) enabling the testing of dose-response relationships (as well as other models of inheritance): this further strengthens the test of causality. Thus purely as a tool for identifying causal factors in complex diseases, the study of genetics has many attractions and may be the most efficient strategy for diseases where aetiological factors are largely unknown.

1.2.2 The benefits of genetic discoveries in complex diseases

By identifying genes involved in disease, biological processes that underlie these disorders may be highlighted; in this way hitherto unsuspected biological processes may sometimes be implicated in a disease. Such discoveries advance understanding of disease pathogenesis and also cast new light on the interpretation of existing functional (e.g. immunological)

observations about a disease. For example, the association of genetic variants in the *IRGM* and *ATG16L1* genes with Crohn's disease focussed attention on autophagy, a biological process that had not previously been implicated in Crohn's pathogenesis (Mathew 2007). Crohn's-associated autophagy gene variants have since been linked to defective innate immune capture of intracellular pathogens and other defective innate immune responses, thereby supporting previous immunological observations of defective innate immunity in Crohn's disease (Nakagawa, Amano et al. 2004; Cadwell, Liu et al. 2008; Kuballa, Huett et al. 2008; Rahman, Marks et al. 2008).

Discoveries of genetic risk variants in complex diseases have also led to greater understanding of shared pathogenic factors in distinct diseases. For example, autoimmune diseases show appreciable overlap in genetic susceptibility variants (Smyth, Plagnol et al. 2008; Zhernakova, van Diemen et al. 2009). The patterns of overlap offer insights into shared autoimmune and disease –specific processes. More surprising have been examples of genetic risk variants identifying biological processes common to disorders thought previously to have completely distinct causes and not known to show epidemiological overlap. For example, genetic variants in leucine rich repeat kinase 2 (*LRRK2*) confer susceptibility to Parkinson's disease and Crohn's disease, although the function of this gene in both disorders is uncertain (Paisan-Ruiz, Jain et al. 2004; Barrett, Hansoul et al. 2008).

Identifying genetic risk variants also has the potential to point to environmental factors that are important in disease causation. This may prove to be a particularly valuable application, overcoming the problem of infinite environmental candidate causes. A striking example was the discovery of genetic variants in an enteroviral response gene, *IFIH1*, that predispose to type 1 diabetes, implicating enteroviral infections in diabetes pathogenesis (Smyth, Cooper et al. 2006; Nejentsev, Walker et al. 2009). Research presented in this thesis (chapter 3) similarly suggests a role for RNA viruses in coeliac disease pathogenesis. Cadwell et al. provided an example of how a very common susceptibility allele could contribute to a much less common disease by interacting with an environmental factor (Cadwell, Patel et al. 2010). The *ATG16L1* T300A SNP has a risk allele frequency ~ 0.5 in European populations. Cadwell et al. previously showed that mutant mice engineered to have hypomorphic ATG16L1 protein expression displayed Paneth cell abnormalities and that Crohn's patients with the *ATG16L1* risk allele showed similar abnormalities (Cadwell, Liu et al. 2008). In the more recent study they used the hypomorphic ATG16L1 mouse model to demonstrate that Paneth cell abnormalities and

increased ileal inflammation in response to dextran sodium sulphate occurred only in the context of intestinal infection with a specific strain of murine norovirus (Cadwell, Patel et al. 2010). This insight demonstrates how a common risk allele and a specific environmental factor can interact to contribute to intestinal inflammation. It shows that while common genetic risk alleles might only have weak effect sizes when assessed in unselected populations, stronger effects may be contingent on the presence of environmental exposures that occur only in a subset of the population. Such interactions have been proposed as one of the explanations for the 'missing heritability' of complex diseases.

Finally genetic risk variants in complex diseases have the potential to identify novel targets for drug development. These translational benefits will, in most cases, take many years to reach clinical practice. However, evidence that risk variants can identify genes that constitute efficacious targets for drug therapies already exists. *PPARG* and *KCNJ11* gene variants have been associated with type 2 diabetes and encode proteins that are targets of thiazoledinediones and sulphonylureas respectively. *IL12B* variants predispose to psoriasis (Nair, Duffin et al. 2009). *IL12B* encodes a subunit of the interleukin-12 and interleukin-23 cytokines and an antibody against this protein has already shown efficacy in a phase II trial in psoriasis (Krueger, Langley et al. 2007). These examples illustrate the possibilities and it is notable that even genetic variants with modest effect sizes may map to genes which when targeted by drugs have large therapeutic effects.

1.3 Personalized medicine

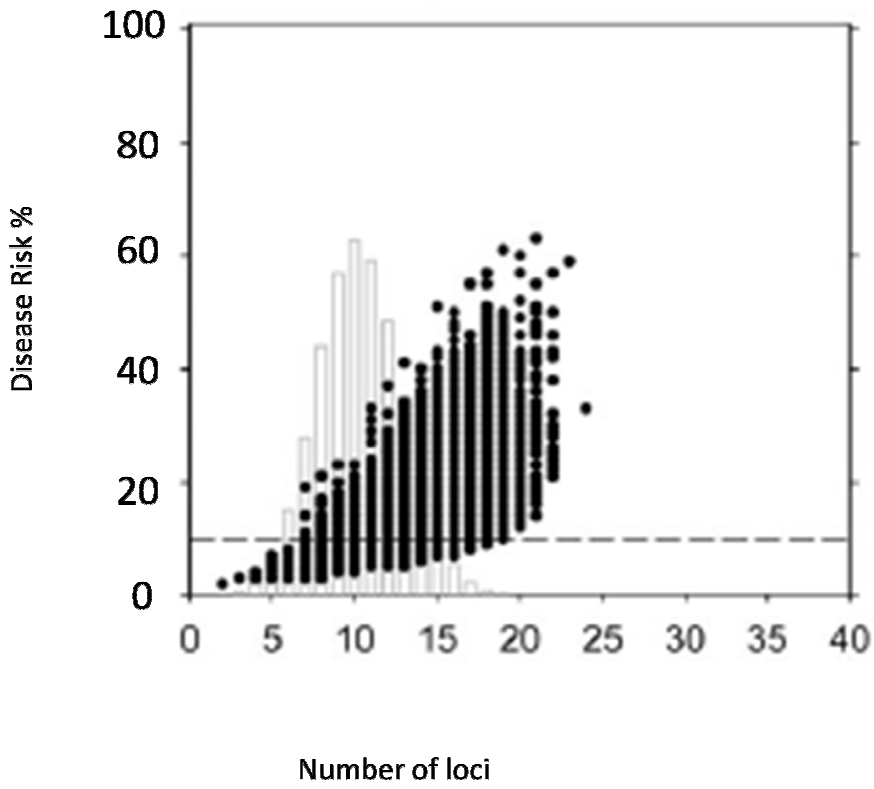
There has been hope that the identification of genetic risk variants in complex diseases will herald an era of personalized medicine. By understanding the genetic component of inter-individual variation in diseases, in disease sub-phenotypes and in drug responses, it may be possible to tailor preventive and therapeutic interventions to individuals based on their genetic profiles. Personalized medicine already exists for single gene disorders, for example familial adenomatous polyposis (FAP) where genetic testing provides a basis for near-perfect prediction of the phenotype and motivates the recommendation of colectomy and gastro-intestinal surveillance in affected individuals. In single gene disorders, the genetic screening test has very high sensitivity and specificity, characteristics which limit the number of false negatives and false positives. In contrast, for complex diseases, prediction accuracy is limited by a number of factors, notably the fact that these diseases are only partially heritable.

1.3.1 Genetic risk modelling

Given a set of genetic risk variants, ideally the predictive value of the genotypes of these variants could be determined by calculating the percentage of individuals with each genotype combination who get the disease. This would require a very large population study given that even for 10 variants there would be 3^{10} (59,049) genotype combinations. Instead researchers have opted to take one of two approaches, both of which overcome the exponential rising number of combinations problem by assuming gene-gene interactions have negligible effects on risk. The first strategy assigns a score to each individual, for example based on the number of risk alleles present. This therefore assumes that each allele has additive and equal effects on risk. The second strategy uses either logistic or Cox proportional hazards regression analyses to assign risk scores for each variant weighted according to effect size. Again this strategy assumes no gene-gene interaction effects. In theory, both these methods can be used to assign risk scores, which can then be evaluated for their predictive performance. Ideally, this should be performed in an independent cohort to that used for risk variant discovery or else predictive value is likely to be overestimated. Similarly, risk estimates for genetic scores ought ideally to be presented in terms of absolute risk as this is the measure most valuable to the individual. Very few studies have evaluated absolute risk, in part because of the use of case control data, where absolute risk can only be estimated with assumptions about disease incidence. Similarly, very few studies have conducted external validation, i.e. validation in an

independent population and therefore performance of these risk models is likely to be overstated. Many studies have evaluated relative risks or odds ratios in relation to a reference group of lowest allele score. These risks are not of great interest to the individual, who typically would like to know his risk compared to the average risk. From the clinicians perspective it would be helpful to know the distribution of risk related to the genetic variant score. This might be used to identify individuals who fall above a threshold above which intervention is justified (**Figure 1.1**). Again this data has been rarely presented.

Figure 1.1 **Distribution of risk according to number of disease risk genotypes. Adapted from (Janssens & van Duijn 2008) (Janssens and van Duijn 2008)**



Simulated data for complex disease with population prevalence of 10%, risk genotype frequencies between 1 and 60% and odds ratios varying from 1.05 – 2.0. Columns represent frequency distribution of the number of risk genotypes and illustrate that the majority of individuals have near average (10%) risk.

A useful and frequently used method of summarizing the discriminatory accuracy of a risk model in classifying individuals as cases versus controls is by plotting a receiver operating characteristic (ROC) curve. The ROC curve here plots the sensitivity against (one minus) the

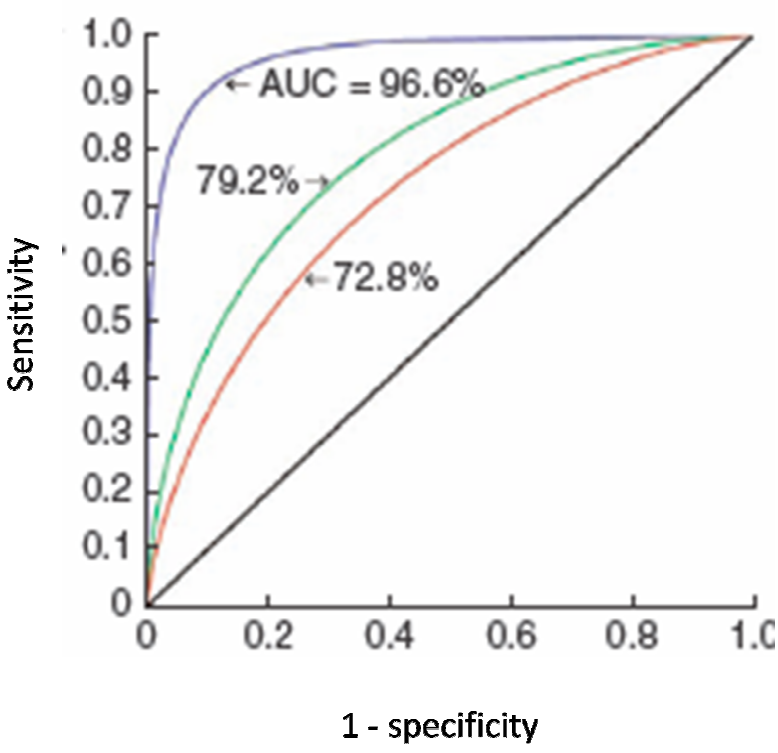
specificity for all possible thresholds for the prediction score (**Figure 1.2**). A good prediction model would show a high proportion of those developing the disease in individuals with the highest risk scores. The area under the receiver operating characteristic curve (AUC) is one commonly used summary measure of the discriminative power of the model. AUC ranges from 0.5 (no discriminative power) to 1 (perfect prediction). An AUC of 0.75 has been suggested as a lower limit for a useful test screening individuals already at increased risk of disease with an AUC of 0.99 for screening population individuals (Janssens, Moonesinghe et al. 2007).

Early attempts at risk prediction models using variants discovered in genome wide association studies have shown limited predictive value. For example, a model using 18 variants in type 2 diabetes had an AUC of 0.60 (van Hoek, Dehghan et al. 2008). For Crohn's disease the AUC for a similar prediction model was estimated at 0.73 using 30 known variants (Daly, Donaldson et al. 2009). This latter study estimated an upper limit for the performance of a risk model in Crohn's disease corresponding to an AUC of 0.966, assuming all heritable variation was incorporated. In a scenario that is achievable with current resources, Park et al. estimated that a model incorporating 142 independent Crohn's variants, hypothesized to be identifiable with larger GWASs, would have an AUC of 0.792 (**Figure 1.2**). This may still fall short of a useful prediction model for Crohn's disease, particularly in the absence of known preventive interventions. In general the predictive value of these models is limited fundamentally by the heritability and prevalence of the disease – less heritable and less common diseases are harder to predict from genetic profiles even if all causal genetic variation is known. Using known heritability and prevalence estimates, several authors have predicted that for most common disorders risk prediction models will never be very accurate and will only approach performance of models using traditional risk factors (Janssens and van Duijn 2008; Clayton 2009; Daly, Donaldson et al. 2009). It is also relevant that existing models have used data from GWASs that have selected cases using stringent criteria and therefore may tend towards the more extreme examples of the phenotype under study (this improves power to detect variants, but overestimates the effect size of discovered variants). Thus the performance of models using these variants may overestimate performance in the real world (Janssens and van Duijn 2009).

Despite these reservations it should be noted that even a test with relatively weak overall prediction performance, as assessed by AUC, may nevertheless for some conditions be of

sufficient value to warrant changes in preventive measures. Moreover these models may find merit for the identification of individuals at the extreme end of the risk distribution, where positive predictive value may be high enough that specific preventative measures or further diagnostic or surveillance tests are strongly justified. Genomic risk profiling, currently available in the form of array-based genotyping of disease-associated variants but eventually envisaged as whole-genome sequencing, will allow simultaneous profiling of risk for tens or hundreds of common diseases. Thus, one would expect that while most individuals will have near average risk for most diseases, most individuals will also be at the high end of genetically-determined risk for one or two of these disorders. For these diseases, positive predictive value may then be high enough to warrant personalized disease-specific interventions. Thus, the value of genetic profiling may lie in identifying the diseases for which individuals carry very high risk, rather than in accurately predicting whether an individual will get a specific disease.

Figure 1.2 **Receiver operating characteristic curves for genetic risk models in Crohn's disease (Reproduced from Park et al. 2010)**



Blue curve – theoretical model incorporating all genetic factors contributing to Crohn's risk.
Green curve – theoretical model incorporating 142 variants estimated to exist within the range of effect sizes seen in current Crohn's GWASs
Red curve – model incorporating variants at 30 known Crohn's loci
Black line – model with no discriminatory power
AUC=Area under the ROC curve

The value of genetic risk prediction for personalized medicine also depends on what preventive measures are available. Thus for some diseases, it is already of proven value for individuals of average risk to undergo preventive measures (e.g. screening in colorectal cancer). The addition of genetic profiling to existing risk factors (e.g. age, family history), might in these cases be sufficient to warrant subtle changes in the age at which screening begins or in the frequency of surveillance investigations for most individuals. However, in other diseases, e.g. Crohn's disease, there are few known protective interventions (stopping smoking is one) and it is unclear whether being assigned a high risk would be of real clinical value.

1.3.2 Pharmacogenomics

Another area where genomics shows promise is in the prediction of drug responses and drug adverse effects. Early experience suggests that common genetic variants, sometimes conferring large effects on risk, account for an important fraction of variation in drug responses. Certainly human populations have not been exposed to most drugs for periods of time sufficient to allow negative selection of deleterious variants. Genome wide association studies have recently begun to identify common variants with large effects that may in time form the basis of pharmacogenetic testing (Link, Parish et al. 2008; Daly, Donaldson et al. 2009; Ge, Fellay et al. 2009). Pharmacogenomics is discussed more fully in **Chapter 4**.

The potential benefits outlined above provide some of the principal motivations for the search for genetic risk variants in complex diseases. This search has produced spectacular successes over the last five years, with the pace of discovery showing no sign of abating. The methodological and technological advances in genetics, the attendant expansion in our knowledge of human genetic variation and the theoretical (e.g. population genetics) considerations that impact profoundly on the design and interpretation of studies aiming to identify genetic risk variants are now discussed.

1.4 Complex disease genetics

One can describe a spectrum for human diseases between the extremes of those whose risk is almost entirely genetically determined and those who risk is almost entirely environmentally determined. The former include chromosomal disorders (e.g. trisomy 21, Turner's syndrome) and single gene disorders (e.g. cystic fibrosis, sickle cell anaemia). Single gene (Mendelian) disorders are the simplest of genetic characters and occur where the genotypes at a single locus are sufficient to account for the character, given a normal environmental and genetic background. Mendelian diseases are rare in the population, due to low mutation rates and the negative or balancing selection pressures on single gene variants that have highly deleterious effects on reproductive fitness. These disorders produce characteristic pedigree patterns (autosomal dominant, autosomal recessive etc) and are therefore easily recognized. The discovery of genetic variants causing Mendelian disorders has had relative success, exploiting a variety of strategies (linkage, positional cloning, identification of homologues from model organisms, functional candidacy) that culminate in the testing of candidate gene variants in individuals with the disorder (Strachan 2004).

In contrast, complex disorders fall further along the spectrum of genetic vs. environmental susceptibility. Most have heritabilities estimated to be greater than 50% and usually exhibit familial clustering but do not manifest characteristic pedigree patterns (Boomsma, Busjahn et al. 2002). One of the first geneticists to outline the genetic characteristics of complex disorders was Cedric Carter, who wrote in 1969 that "the genetic element in most common disorders, is neither chromosome abnormality nor mutant gene of large effect, but very probably an underlying polygenically determined and continuously distributed genetic predisposition with a threshold beyond which individuals are at risk." (Carter 1969). The key distinction is the need to invoke a polygenic inheritance, where disease-predisposing genetic variants at many independent loci act together to determine risk. Carter's quote alludes to R.A. Fisher's observation that a number of independent genetic factors, each inherited in a Mendelian fashion would produce continuously distributed quantitative traits, and to Falconer's use of Sewell Wright's threshold model to show that dichotomous complex traits may manifest where an underlying continuous trait reaches a critical liability threshold (Fisher 1918; Wright 1934; Falconer 1965).

It was realised that in contrast to Mendelian disorders, a polygenic inheritance for complex disorders could include common variants (polymorphisms) since their relatively weak individual effects on risk protect them from strong negative selection. However, the true nature of the allelic spectra at common disease susceptibility loci can not be inferred from patterns of disease in the population (in contrast to single gene disorders). This question has been the subject of much conjecture, but can only be answered by empirical data. We are now on the cusp of answering these questions in complex diseases with the advent of an era of whole genome sequencing of large numbers of individuals.

1.4.1 Evidence of heritability in complex disorders

Familial clustering in complex diseases is the primary evidence of their heritable basis. Twin studies enable an estimation of the relative importance of genetic and environmental factors. Monozygotic twins share all germline genetic variants whereas dizygotic twins share only half of their genetic variants *(identical by descent)*. In contrast both monozygotic and dizygotic twins are expected to share similar proportions of their environments. Thus higher phenotypic concordance in monozygotic twins than dizygotic twins is evidence that genetic variants contribute to disease risk. Conversely, diseases that are dominated by environmental risk factors show similar monozygotic and dizygotic concordance.

The sibling relative recurrence risk (λ_s) is the risk in a disease sibling relative to that of the general population disease risk. It provides an estimate of familial clustering independent of population prevalence and therefore enables comparisons of the degree of heritability between diseases.

1.4.2 Common and rare variant hypotheses for complex diseases

Two polarised hypotheses illustrate the possibilities for the allelic spectra of complex disease loci. The **common disease – common variant** hypothesis proposes that the genetic basis of complex (common) disease is conferred by multiple common variants with individually weak effects on disease susceptibility (Lander 1996). The **common disease – rare variant** hypothesis proposes that common disease susceptibility is accounted for by a large number of variants that each occur only rarely within the population (Bodmer and Bonilla 2008; Schork, Murray et al. 2009). Thus two unrelated individuals with the same disease will tend to share fewer

disease-causing variants in this model. It is hypothesised that rare disease-predisposing variants will have stronger effects on susceptibility. This is so far supported by the few available data that have identified common and rare variants in complex diseases (**Figure 1.3**).

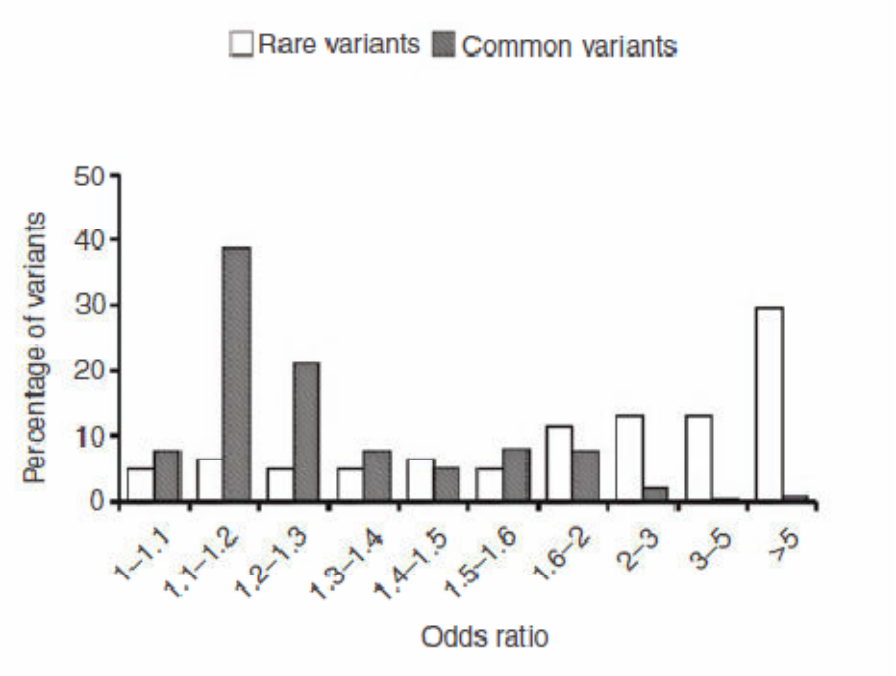
Importantly, both these hypotheses have implications for the choice of strategy for variant discovery in complex diseases. Genome-wide association studies have been designed primarily to test the common variant hypothesis and typically do not assay variants with population minor allele frequencies less than 5%. The capacity to test the rare variant hypothesis for common diseases is less mature. In contrast to the comprehensive catalogues of common variants (e.g. International HapMap project), cataloguing of rare variants in human populations is in a relatively early stage (e.g. 1000 Genomes project).

In time, the rapidly decreasing cost of whole genome sequencing will enable comprehensive ascertainment of all forms of genetic variation in individuals with common disease. As an intermediate strategy to look for rare variants, researchers have begun to focus resequencing efforts on genome regions most likely to be enriched for causal variants. These include regions identified by common variant associations in GWASs and the 30Mb protein- coding portion of the genome (the exome). An example of this approach was the re-sequencing of 10 type 1 diabetes candidate genes identified by common variant associations (Nejentsev, Walker et al. 2009). This study identified causal rare variants in one of these genes, *IFIH1*, providing support for the idea that the allelic spectra at at least some disease loci will include both common and rare variants. Similarly resequencing of the *NOD2* gene identified an allelic spectrum in which 3 protein-changing variants with allele frequencies between 1.2 and 4.3% in European ancestry populations accounted for 80% of Crohn's associated variants, with many additional rare variants in the *NOD2* gene also contributing to diseases risk (Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001; Hugot, Zaccaria et al. 2007).

Hirschprung's disease, a complex (though not very common) disease with known oligogenic inheritance also illustrates this point. A recent study examined multiple disease-predisposing variants of both common and rare frequencies in a single gene (receptor tyrosine kinase-*RET*) (Emison, Garcia-Barcelo et al. 2010). Hirschprung's disease (congenital intestinal aganglionosis) is phenotypically sub-divided into short segment (extending to upper sigmoid), long segment (extending to splenic flexure) and total colonic forms. Firstly the common variants were less penetrant and associated with lower disease recurrence rates in relatives

compared to the rare variants they identified, consistent with predictions of higher effect sizes for rare variants. Secondly, the profiles of *RET* variants differed between sub-phenotypes, with coding variants more frequent in the more extensive forms of the disease and a common regulatory variant found more frequently in the short segment form. It may be anticipated that other diseases that exhibit distinct sub-phenotypes such as Crohn's disease, may similarly have a genetic basis determined partly by the allelic spectrum at disease loci, with rare, as yet undiscovered, coding variants contributing to more severe and extensive forms of disease.

Figure 1.3 **Distribution of odds ratios for common and rare variants (Reproduced from Bodmer and Bonilla, 2008)**



61 rare and 217 common variants from the literature used in the analysis

Many complex disorders include a subset that follows Mendelian or near-Mendelian inheritance. For example, *BRCA1* and *BRCA2* mutations cause strongly familial forms of breast cancer, but do not contribute substantially to sporadic breast cancer. Loss-of-function mutations in the *IL10RA* and *IL10RB* genes encoding subunits of the interleukin 10 receptor have been shown to cause early, severe familial cases of inflammatory bowel disease (Glocker, Kotlarz et al. 2009). Again, common genetic variation in these genes has not been associated with sporadic forms of inflammatory bowel disease in large association studies (Barrett, Hansoul et al. 2008). Extending from these near-Mendelian examples, it is possible that complex disorders contain further subsets of individuals with distinct oligogenic aetiologies. Studies of multiply-affected families offer the best opportunity of identifying the next tiers of such genetic variants.

It is likely that novel study designs will be helpful in the identification of rare variants in complex diseases. For example, family-based and extreme-trait study designs may have power advantages compared to simple case-control designs (Cirulli and Goldstein 2010). Additional novel analytic techniques are also being developed to help overcome the fall-off in power to detect genetic risk variant associations that occurs as minor allele frequency falls. Thus, weighting according to predicted functional effects of variants or pooling of rare variants within genes for combined association testing may be helpful (Price, Kryukov et al. 2010).

1.4.3 Using intermediate traits in complex disease genetics

Disease definitions at their most basic begin with an observation of symptoms and signs that recur together in a recognizable pattern (co-vary) in different individuals. When knowledge of the cause(s) of these phenotypes is lacking, the disease is defined according to what is observable (symptoms, signs, supporting tests) and what is of clinical relevance (for purposes of prognosis or treatment). Thus since aetiologies are poorly understood for many common (complex) diseases, their classifications are at best only loosely rooted in an understanding of the underlying biology. This poses a potential problem for efforts to investigate the causes of these diseases.

For example ileal Crohn's disease appears to have at least partly distinct causation to colonic Crohn's disease, with some genetic variants only conferring risk to ileal Crohn's and vice versa

(Barrett, Hansoul et al. 2008). Conversely, many autoimmune diseases have been shown to share a surprising number of genetic susceptibility variants raising the possibility that these disorders might be better understood as alternative manifestations of a common autoimmune tendency (Zhernakova, van Diemen et al. 2009).

It has been argued that investigation of complex traits would be aided by the definition and measurement of continuous traits (so-called intermediate phenotypes or endophenotypes) that more closely correlate with polygenic liability (Plomin, Haworth et al. 2009). This follows the liability threshold model of complex disease traits. It has been argued that relying on clinical disease definitions of dichotomous disorders may limit our power to uncover the relevant genetic susceptibility factors. Certainly, if common dichotomous disorders are manifestations of single continuous traits, comparing individuals affected by disease with those who are unaffected means the unaffecteds will include some individuals who have a liability that is close to the threshold for disease (near-diseased individuals). The inclusion of such individuals in a case-control design study would be expected to reduce power to detect the genetic variants responsible for the disease liability compared to study designs that either sampled individuals from both extremes of the liability distribution or identified an appropriate endophenotype for quantitative trait analysis. However, the little available data that exist have not provided great support for the superiority of either approach compared to traditional case-control designs. For example the use of cognitive traits, considered excellent endophenotypes for schizophrenia, has not been more successful in helping discover genetic risk variants for schizophrenia liability than studies of schizophrenia itself (Cirulli, Kasperaviciute et al. 2010). Moreover, for many disorders it is far from clear whether an appropriate and measurable endophenotype could be identified. Use of an endophenotype risks missing some of the important liability that contributes to the dichotomous trait. It may be that intermediate phenotypes will prove of greatest value in testing their correlation with genetic variants that have firstly been identified in case-control studies. This will be a step towards testing hypotheses about the biology and pathogenesis of a disorder. For example, in Crohn's disease, the correlation of Crohn's associated genetic variants with intermediate traits such as macrophage function or innate immune clearance of bacteria may reveal hitherto under-appreciated roles for Crohn's genes. Crohn's disease-associated *NOD2* variants for example were associated with defects in autophagy induction, bacterial trafficking and antigen presentation (Cooney, Baker et al. 2010). In coeliac disease, one could envisage testing gene variants for correlation with T cell responses. For example, it has been shown that healthy HLA

DQ2+ individuals show T cell Ifni responses to gliadin peptides that are intermediate between individuals with coeliac disease and healthy individuals lacking HLA DQ2 or DQ8 (Anderson, van Heel et al. 2005). Many coeliac disease-associated genes are strongly expressed in T cells, but have poorly defined functions (Dubois and van Heel 2008). Therefore correlation of coeliac-associated variants in these genes with T cell responses may be instructive.

1.5 Human Genetic Variation

1.5.1 Origins of human genetic diversity

Mutation (change in DNA sequence) is the origin of all genetic variation. This wellspring of new variation occurs continuously and when affecting the germline can be passed on to the next generation. Mutations occur spontaneously, as a result of inherent instabilities of DNA chemistry and the mechanism of replication (see table) but may also be facilitated by exposure to environmental factors (e.g. viruses, radiation- ultraviolet light and ionizing radiation, chemicals). Defective DNA repair is also an important source of replication errors.

Causes of spontaneous mutation

- Chemical instability
  - Tautomerism
  - Depurination
  - Deamination
- Replication errors
  - Slipped strand mispairing
  - Non-homologous recombination
  - Mismatch repair failures
- Transposable elements

Mutation rates (the number of mutations occurring per base per generation) vary between species, due to multiple factors including differences in the efficiency of DNA repair mechanisms. In humans, a widely cited mutation rate estimate of 2.5×10^{-8} was based on a comparison of sequence data from humans and chimpanzees, using assumptions of ancestral population size and time to species divergence(Nachman and Crowell 2000). Recently, the mutation rate was directly observed at a lower rate of 1.1×10^{-8} from whole genome sequences of a family quartet (2 parents and 2 children) corresponding to 70 new mutations per diploid genome (Roach, Glusman et al. 2010).

Rates of mutation also vary across the genome. For example, there is an approximately 10fold higher mutation rate for CpG dinucleotides observed both in humans and chimpanzees (Chimpanzee-Sequencing-and-Analysis-Consortium 2005; Roach, Glusman et al. 2010). Highly

repetitive sequences (e.g. Alu repeats and segmental duplications) are more prone to non-allelic homologous recombination events, which can lead to multi-nucleotide mutations, including copy number variations spanning thousands to millions of nucleotides(Cooper, Nickerson et al. 2007; de Smith, Walters et al. 2008).

Mutation therefore introduces genetic variation, but how the frequencies of these variants change through generations depends on at least two other major factors. Natural selection operates to increase or reduce the frequencies of variants that have non-neutral effects on fitness (i.e. confer a reproductive advantage or disadvantage). However, the majority of mutations produce variants that have weak or neutral effects on fitness and so are not affected by natural selection. Genetic drift is the process whereby random allocation of alleles from parents to offspring causes changes in allele frequencies over time. These chance effects mean that even neutral mutations may show changes in allele frequencies over time. In smaller breeding populations genetic drift has a larger effect on allele frequencies than in larger populations as sampling error is proportionately greater. Thus in human populations, when small founding populations (population bottlenecks) have occurred, genetic drift initially dominates natural selection in changing allele frequencies, but as populations become larger the reverse is true and natural selection predominates.

1.5.2 Human population ancestry and effects on current population genetic variation

Non-African human populations originated from a small founding population migrating from east Africa to the Arabian peninsula c. 60-100,000 years ago (the recent single origin hypothesis) (Liu, Prugnolle et al. 2006). This has well-known consequences including a reduced genetic diversity of non-African populations compared to African populations. Indeed genetic diversity in human populations decreases with geographical distance from Africa (Coop, Pickrell et al. 2009). Moreover linkage disequilibrium is lower in African than non-African populations, consistent with non-African populations arising from a founder population(s) that contained only a subset of African human genetic variation. The different haplotype structure in different populations, may potentially allow susceptibility regions discovered in GWASs of European ancestry to be narrowed-down, by comparison with GWASs in, for example, African populations.

Genome wide SNP data from 3 of the 4 human HapMap phase II populations (CEU,YRI,CHB) and the CEPH-Human Genetic Diversity Panel (53 human populations) confirms the reduced genetic diversity seen with geographical distance from Africa (Coop, Pickrell et al. 2009). This study also examined signatures of selection in the genome and the global patterns of these signatures. One finding was that the global pattern of most putatively selected (highly population differentiated) SNPs was in fact predicted by ancestry, with three broad regions (Africa, West Eurasia and East Asia) capturing most of the differences. This suggests that neutral processes such as population migration and genetic drift play a large part in determining the distribution of population differentiated alleles. For example, the geographic distribution of alleles at a skin pigmentation locus, *SLC24A5*, that show high between population differentiation, nevertheless follow geographical patterns predicted by ancestry rather than latitude or climate(Coop, Pickrell et al. 2009).

This data therefore illustrates the balance between selection and neutral factors that have shaped genetic diversity within and between populations. It also demonstrates that for most SNPs, ancestry, rather than selection is the dominant determinant of differences in allele frequencies between populations.

1.5.3 Forms of DNA sequence variant

Human genetic variation is observable at all scales from single nucleotides to whole chromosomes. Genetic variation in humans has so far been most comprehensively defined at both extremes of this size spectrum. Karyotyping and fluorescent in situ hybridization (FISH) have identified most germ-line variants affecting genomic segments > ~3Mb (Feuk, Carson et al. 2006). These variants typically arise as *de novo* mutations in the germ line and due to strongly deleterious effects on fitness do not persist in the germ line for many generations. Well known examples include trisomy 21 (Down's syndrome), monosomy X (Turner's syndrome) and a 3 megabase deletion on chromosome 22q11.2 (diGeorge syndrome). Advances in molecular biology (especially PCR amplification and Sanger DNA sequencing) have facilitated the discovery of fine scale variants, particularly well catalogued at the single nucleotide level. Heritable variation at the intermediate, sub-microscopic scale (1kb-~ 3Mb) has not been as well defined, historically suffering from the relative inferiority of assays available to study these forms of variation.

1.5.3.1 Single Nucleotide Polymorphisms

A single nucleotide polymorphism refers to a nucleotide substitution that occurs in more than 1% of the population. The National Centre for Biotechnology Information (NCBI) in the United States provides a repository for the deposition of new SNP sequences (database of SNPs-dbSNP). dbSNP does not restrict variants based on allele frequency and therefore includes single base substitution variants with less than 1% minor allele frequency (i.e. not technically polymorphisms). dbSNP also records short insertion/deletion variants. In total, 23.6 million reference single nucleotide variants are included in the latest database release (dbSNP Human Build 131) of which 14.6million are validated. It is estimated that there are around 10 million common SNPs (MAF > 0.05) in most human populations, so that on average 1 of every 300 bases is expected to be polymorphic in the ~3 Gigabase haploid human genome. Over 7 million reference SNPs in dbSNP have so far been identified with minor allele frequency greater than 5% (Frazer, Murray et al. 2009). Short (<10bp) insertion/deletions (indels) appear to be surprisingly poorly catalogued despite their amenability to discovery by similar PCR/Sanger sequencing methods as used with SNPs. For example, sequencing and assembly of the diploid genome from an Asian individual showed that 86.4% of SNPs were present in dbSNP whereas only 40.9% of short indels (</= 3bp) were present in dbSNP (Wang, Wang et al. 2008). Similar results were observed for the Watson and Venter genomes (Levy, Sutton et al. 2007; Wheeler, Srinivasan et al. 2008).

1.5.3.1.1 The origin of SNPs and haplotypes

SNPs exist because of ancestral mutations that have usually occurred only once in human history. New single base variants are generated infrequently (the new mutation rate is roughly 10^{-8} per generation per base or 30 new variants per haploid gamete) (Manolio, Brooks et al. 2008). As only around 10^4 generations separate currently living individuals and their most recent common ancestor, the low mutation rate makes it likely that a SNP allele shared by apparently unrelated individuals actually has the same ancestral origin. In general the more common a variant is within a population, the more ancient it is, since with each generation the number of descendents carrying the variant allele can potentially increase. Genetic drift and natural selection are important processes influencing changes in variant frequency over time

in populations, but do not detract from the principle that the frequency of a variant correlates with the age of the founding mutation.

Haplotypes refer to combinations of variant alleles that occur together. They arise as a consequence of sexual reproduction and the history of the species. During meiosis, crossing over and recombination of segments of the maternal and paternal chromosome occurs leading to a hybrid chromosome in the gamete. These crossing over events occur non-randomly in the genome, with much greater probability in small regions called recombination hotspots. 80% of allelic recombination is confined to hotspots covering 10%–20% of the genome (Myers, Bottolo et al. 2005). The consequence is that the boundaries of the segments of chromosomes that are swapped during recombination vary mostly only between these hotspots, with intervening segments usually preserved *en bloc*. Over generations, due to repeated shuffling of segments by recombination events occurring at different hotspots, the segments of chromosome shared between the last descendant and the original ancestors get smaller. Similarly the lengths of these segments (haplotypes) that are shared between individuals within the descendant population are smaller when the founding ancestors were ancient compared to when they were recent. Indeed the presence of extended shared haplotypes between two individuals is a sign of recent shared ancestry. For this reason, non-African populations, which underwent a population bottleneck during migration of the founders from Africa, have longer haplotypes than African populations where founding populations were larger and have had more time for recombination.

By understanding the haplotype structure of common genetic variants in human populations, it was anticipated that SNPs could be selected from each common haplotype that would tag (be highly correlated with) other common variants on the haplotype. In this way common genetic variation could be summarised by a much smaller set of tag SNPs, enabling researchers to assay a large fraction of common genetic variation by assaying only a much proportion of the 10 million or so common SNPs. This was one of the principle motivations for the International HapMap project (section 1.5.4.2). Another benefit that has emerged is SNP imputation, where haplotypes containing dense sets of SNP alleles defined in the HapMap populations allow imputation of missing (non-genotyped) SNPs by reconstructing the haplotypes from those SNPs in the region that have been genotyped.

In many regions of the human genome, common variation comprises only a small number of different haplotypes. This means that SNP alleles within such regions tend to be highly correlated when examined in individuals from a single population. Linkage disequilibrium refers to this non-random association of SNPs in the population. By correlating SNP alleles within the HapMap populations, it has been possible to define these linkage disequilibrium blocks. This has greatly facilitated the design of genetic association studies by allowing the selection of a single tag SNP from each of these LD bins. Phase II of the HapMap Project determined that the vast majority of SNPs with a MAF of at least 5% could be tagged by single SNPs from each of ~550,000 LD bins in CEU/CHB/JPT populations or 1,100,000 LD bins for the more ancient YRI(Frazer, Murray et al. 2009).

1.5.3.2 Beyond SNPs- multinucleotide genetic variants

For simplicity, human genetic variation can be divided into single nucleotide sequence variation (base substitutions) and structural variation (affecting >1 contiguous base) (**Figure 1.4**) (Frazer, Murray et al. 2009; Conrad, Pinto et al. 2010). Other authors have preferred to further distinguish fine scale variants (<1 kb) from structural variants (>1kb) (Feuk, Carson et al. 2006; Redon, Ishikawa et al. 2006). Either classification has the virtue of being non-redundant and is largely agnostic to the mechanisms causing these variations. In contrast other descriptions of structural variants have reflected historical factors such as the techniques used for discovery, or the evolutionary origins of variants. For example many fine scale polymorphisms were discovered through the search for linkage markers (Nakamura, Leppert et al. 1987; Weber and May 1989). Variable number of tandem repeats (VNTRs) are one such class of polymorphism, and can be sub-classified into microsatellites (tandem repeats of less than 5 base pairs) and minisatellites (>5bp). Retrotransposon elements are common and include short and long interspersed nuclear elements (SINEs and LINEs) which are insertions/deletions in the 0.3kb and 6kb length range.

Structural variant discovery has lagged behind the discovery of SNPs and short indels (<~10bp). At the sub-microscopic scale (~<3Mb) this has been limited by a historic inferiority of techniques for both discovery and genotyping. This is rapidly changing with the advent of whole-genome sequencing of multiple individuals and other structural variant discovery strategies (Levy and Strausberg 2008; Conrad, Pinto et al. 2010; Pang, MacDonald et al. 2010). The sequencing of the first four diploid human genomes (Craig Ventner, James Watson, an

individual of Chinese and an individual of west African descent) together with other existing data suggest common SNPs occur in the millions, insertion/deletion polymorphisms in the hundreds of thousands and structural variants (>1kb) in the thousands (Levy, Sutton et al. 2007; Bentley, Balasubramanian et al. 2008; Wang, Wang et al. 2008; Wheeler, Srinivasan et al. 2008). Although less numerous than SNPs, structural variants account for far more variant bases between individual genomes due to their larger size (Levy, Sutton et al. 2007; Levy and Strausberg 2008; Frazer, Murray et al. 2009). Thus, structural variation has long been postulated to account for a substantial fraction of the genetic variance of human phenotypes including common diseases.

Figure 1.4 **Structural variant classification (adapted from Frazer et al. 2008)**

Structural variants may be balanced (inversions, translocations) or imbalanced (copy number variants- CNVs). Array based techniques allow interrogation of genomic DNA for variations in copy number, but are blind to balanced variants. Whole genome sequencing might be

expected to enable comprehensive ascertainment of all forms of variation, but with current sequencing technologies and computational approaches and even with deep sequence coverage, annotation of structural variants remains very challenging (Pang, MacDonald et al. 2010). The most extensive array-based survey of copy number variation used comparative genomic hybridisation microarrays comprising 42 million tiled probes that covered almost all the assayable genome (median spacing 56bp) (Conrad, Pinto et al.). The dense probe coverage was designed to enable sensitive detection of CNVs greater than 500 bp in length. Conrad et al. screened 41 individuals from the CEU and YRI HapMap populations and identified 8,599 validated copy number variations (CNVs) of greater than 443 base pairs (Conrad, Pinto et al. 2010). The majority of these CNVs were intergenic, with a paucity overlapping genes relative to random permutations, suggesting genic CNVs can have deleterious effects on fitness. On average per comparison of two diploid genomes, they found around 1000 CNVs > 500bp. They observed that these CNVs led to alterations in the coding sequence of 1.2% of all gene messenger RNAs on average between two diploid genomes. Thus, the potential for alterations in gene function and contributions to disease susceptibility is clear.

Most recently Pang et al. sought to combine array-based and whole genome sequencing approaches to more fully annotate structural variants in a single diploid human genome. This study showed that a previous *de novo* assembly of a diploid genome sequence using Sanger sequencing failed to detect many structural variants (especially larger variants, ~ > 1kb). However by applying additional computational and the array-based techniques they observed thousands of additional variants, confirming a relatively smooth inverse correlation between the size of variants and their frequency in an individual (Pang, MacDonald et al. 2010). Thus, in this study, which is to date the most comprehensive survey of genetic variation in a single individual, nearly 50Mb (1.6%) of sequence in the haploid genome was structurally variant and only 3 Mb (0.1%) were single nucleotide variants.

Both the Conrad et al. and Pang et al. studies estimated that around 75% of structural variants were imputable from SNPs. As such if copy number polymorphisms and other structural variants contribute to complex disease traits most should generate association signals in GWASs. The Wellcome Trust Case Control Consortium (WTCCC) directly assayed 3,432 copy number polymorphisms in 16,000 cases from eight common diseases and 3,000 shared controls (Craddock, Hurles et al. 2010). They estimated that these CNPs constitute approximately half of all autosomal CNVs > 500 bp long with minor allele frequency > 0.05 and

had 80% power to detect CNPs conferring odds ratios > 1.4 in their study. Association was observed for only 3 loci (*IRGM*-Crohn's disease, *HLA*-Crohn's, Rheumatoid arthritis and type 1 diabetes and *TSPAN8* for type 2 diabetes), compared to the 24 loci identified in the similarly sized SNP GWAS of seven diseases (Wellcome Trust Case Control Consortium 2007). Association at all of these loci had been previously identified by genome wide association studies using SNP arrays. Thus while this was a less complete survey of CNP variation (~50% of CNPs) compared to most SNP GWASs (typically >70% common SNP variation assayed) it nevertheless confirmed that CNPs are unlikely to account for a large fraction of the genetic basis of common diseases. It further confirms that causal CNPs are usually effectively tagged by SNPs included on currently available SNP genotyping platforms used for genome wide association studies. An efficient approach (for common genetic variants) may therefore be to search genomic regions showing association in SNP GWASs for structural variants present in genomic variation databases, and to consider direct genotyping only where structural variants are found. In time, comprehensive catalogues of structural variants and knowledge of their population frequencies and correlations with SNP haplotypes will further help this process.

1.5.4 Cataloguing human genetic variation

1.5.4.1 The Human Genome Project

In 2001 the first draft sequences of the human genome were published (Lander, Linton et al. 2001; Venter, Adams et al. 2001). Both of these draft sequences were haploid, approximately 3 billion bases in length and assembled from multiple donors. In 2004 a near-complete sequence of the human genome was reported, containing > 99% of euchromatic regions (Economou, Trikalinos et al. 2004). The remaining 1% consists of ~300 small gaps, many close to recently duplicated sequence and containing DNA that can not easily be propagated in bacteria prior to Sanger sequencing (Bentley 2006). Furthermore, ~ 200Mb of heterochromatic regions were not sequenced, again due to highly repetitive sequence. Much of this sequence is located in centromeres.

1.5.4.2 The International HapMap Project

The International HapMap project was designed to provide a public resource that would advance medical genetic research (The International HapMap Consortium 2005). In particular the initial aims have been to determine the frequencies and population correlations (linkage

disequilibrium) of common SNPs in 4 human populations of diverse ancestry. 270 individuals were initially selected, 30 mother, father, child Yoruban trios of from Ibadan, Nigeria (YRI), 30 trios from Utah, USA (CEU), 45 unrelated Han Chinese from Beijing (CHB) and 45 unrelated Japanese from Tokyo (JPT). In Phase I, approximately 1.3 million evenly distributed SNPs (approximately 1 every 5 kilobases) were genotyped in 269 individuals (1 JPT individual excluded with low quality DNA) leading to the publication of the first human haplotype map in 2005 (The International HapMap Consortium 2005). Phase 2 extended this to 3.4 million SNPs (~1 every kilobase) in the same individuals (Frazer, Ballinger et al. 2007). In some regions of the genome SNP genotyping assays are more difficult to design (segmental duplications, centromeres, telomeres, gaps in genome sequence – the "unSNPable genome") and therefore the HapMap has poor coverage in these regions. A follow-on from this is that current genome wide association studies also have relatively low coverage of variants in these regions. The HapMap phase III is now genotyping individuals from 7 additional populations, including 2 Kenyan populations, Gujarati in Houston, Texas, Mexican ancestry in California, Chinese in Denver, Colorado, Tuscans in Italy and African ancestry in South-west USA (Altshuler, Gibbs et al. 2010). Phasing of haplotypes was achieved by the inclusion of parent-offspring trios and also computational methods that take advantage of the fact that due to LD only relatively few of the large number of possible haplotypes consistent with the genotype data actually occur in population samples (The International HapMap Consortium 2005). By sequencing ten 500 kb regions in 48 individuals and genotyping all discovered SNPs in all the HapMap samples, it was observed that even though a high percentage of SNPs (46% MAF <0.05) are rare in any given individual 90% of heterozygous sites are due to common variants. Thus, the majority of SNP variation between individuals is indeed due to common variants. The HapMap project has enabled this variation to be summarized by a limited fraction of the 10 million or so common SNPs that tag haplotypes defined by the project. Furthermore, other forms of common genetic variation such as copy number polymorphisms (see section 1.5.3.2) appear to be inherited on these common haplotypes and can therefore also be tagged by common SNPs (McCarroll 2008).

1.5.4.3 The 1000 Genomes Project

The goal of the 1000 Genomes Project is to find most genetic variants that have frequencies of at least 1% in a selection of diverse human populations (www.1000genomes.org). To achieve this, the project is employing a variety of whole-genome sequencing strategies to maximize

the efficiency of variant discovery. Sequencing costs scale with the number of bases sequenced. If the aim is to accurately sequence an individual's diploid genome, this requires an average read depth of around 28 times (28x) using currently available sequencing platforms. However, if the goal is to discover all variants (with frequency > 1%) in the population rather than all variants in a single individual, a more efficient strategy is to sequence large numbers of individuals at low read depth (e.g. 4x). This therefore, is the principal strategy planned for the 1000 Genomes project. Pilot projects that have been completed and that have helped optimize sequencing methods included the sequencing of two mother-father-child trios at high read depth and 180 individuals at low read depth. The full project aims to sequence 2,500 individuals from 27 populations at low read depth.

The 1000 Genomes project will build on the catalogue of human genetic variation from the International HapMap project that has proved so valuable in the design of genome wide association studies. In two respects the 1000 genomes project is expected to advance the catalogue of human genetic variation. Firstly, it will identify less common variants (down to a minor allele frequency of 1%). Secondly, it will significantly advance the catalogue of structural variants compared to currently available databases (e.g. database of genomic variants). This data will allow incorporation of new variants on the next generation of genotyping platforms for association studies. In rare diseases, it will provide a valuable database to determine whether putative disease-causing mutations are disease-specific or in fact found in the wider population. For the more immediate purposes of building on the findings presented in Chapters 3 and 4 in this thesis, it will help the search for causal variants in regions of the genome associated with complex diseases. These regions can be screened for structural and sequence variants with obvious or predicted functional effect and such variants may be imputable in the GWAS data.

1.6 Approaches for identifying causal genetic variants in complex diseases

Two main methods have been used to identify genetic risk variants in complex diseases: association and linkage studies (**Table 1.1**).

Genetic association studies test correlations between genetic variants and the phenotype of interest. Correlations of quantitative traits with genotypes or alleles at a locus (quantitative trait loci) can be tested using linear regression or analysis of variance (Arking, Pfeufer et al. 2006; Weedon, Lettre et al. 2007). Association studies for dichotomous traits like coeliac disease and Crohn's disease have most frequently used a case-control design. These simply compare the frequencies of genetic variants (e.g. alleles or genotypes) in cases and controls (e.g. using Fisher's exact test, Pearson's chi-square, and Cochran-Armitage genotypic trend test). Variant alleles or genotypes that differ in frequency between cases and controls more than expected by chance show evidence of disease association. Candidate gene studies, for example, select a few variants close to or within a gene hypothesized to play an important role in disease pathogenesis, either as a result of known biological function or because the gene maps to an interval implicated in the disease by a linkage study. More recently, the HapMap project and advances in high throughput SNP genotyping methods have enabled the design of genome wide association studies. These studies assay hundreds of thousands of SNPs distributed across the genome and aim to assay a large proportion of common genomic variation (Pearson and Manolio 2008).

The second method employed is genetic (non-parametric) linkage analysis. This technique studies multiply affected families, testing co-segregation of polymorphic genetic markers (e.g. microsatellites) with disease. Markers that reside close to a disease-causing variant will tend to co-segregate with disease in multiply affected families. Co-segregation will tend to occur more frequently if the disease-causing variant has a strong effect on susceptibility. Rare variants of large effect ought to therefore be more tractable to linkage analysis than association studies. A single disease-causing variant may be rare in the general population and in sporadic cases, but within a single multiply affected family it may be common. Thus family-based studies offer advantages to detect rare variants compared to association studies.

Table 1.1 Gene finding approaches in complex disease

Study type	Method	Advantages	Disadvantages
Linkage studies	Test co-segregation of genetic markers with disease phenotype in affected relatives to establish broad regions of genome within which causal variants reside	Able to detect rare variants, and structural variants, if highly penetrant	Low power to detect weakly penetrant alleles Low genomic resolution Require large numbers of affected families
Candidate gene association studies	Compare frequencies of variants in candidate genes chosen on biological grounds or from knowledge of linkage regions	May pinpoint genes from regions of linkage Greater power to detect weakly penetrant alleles	Low power to detect rare variants Historically generated many false positives
Genome wide association studies	Compare frequencies of ~10^5 single nucleotide polymorphisms distributed throughout the genome between cases and controls	High resolution- able to pinpoint small region of genome Power to detect weakly penetrant alleles	Low power to detect rare alleles Low power to detect structural variants Expensive

A further advantage of linkage and family-based studies is that they effectively control for population stratification, a potential confounder of association studies. Genomic regions identified by linkage studies are however typically large, limited by the size of families (number of meioses) available for study. Such regions typically span several megabases. Moreover, in practice, for complex diseases, obtaining sample sizes great enough to resolve associated regions has proven difficult.

In Crohn's disease and coeliac disease, candidate gene and linkage studies have had only modest success in identifying disease genes. Many reported associations have failed replication in follow-up studies. The failures of these approaches reflect a number of limitations. Linkage studies have lacked power to detect all but the most highly penetrant variants (e.g. *NOD2* in Crohn's disease). Candidate gene studies are limited to testing a single hypothesis of association for one gene and are prone to type 1 errors. Moreover even in instances where the gene of interest has subsequently been confirmed in large GWASs (e.g. *CTLA4* in coeliac disease), the original studies were underpowered, making them susceptible to type 2 errors and inconsistent results (van Heel, Hunt et al. 2005).

1.6.1 Genome wide association studies

A genome-wide association study is an association study that assays a major fraction of common variation across the genome. For SNP GWASs this requires upwards of 100,000 markers at a bare minimum (see **Table 1.1**). A case-control design is commonly used for dichotomous traits. Genotyping in SNP GWASs uses SNP microarrays that incorporate assays for SNPs that have been chosen either to be evenly spaced across the genome (e.g. Affymetrix microarrays) or based on haplotype-tagging efficiency using HapMap data (e.g. some Illumina microarrays). The Illumina microarrays used in research presented in this thesis consist of millions of 2 or 3 micron silica beads each coated with SNP-specific oligonucleotide probes and a coding oligonucleotide probes. These beads are randomly assembled into microwells on a glass slide, forming the array. The positions of each bead are determined by a sequential hybridization of the coding oligonucleotides, which enables decoding of the SNP assay positions. Each SNP assay is represented by 20-30 beads. The Illumina technology involves hybridization of amplified whole genomic DNA fragments to the glass beads, with single base extension at the polymorphic nucleotide position using chain-terminating dideoxynucleotides incorporating either a biotin or 2,4 dinitrophenol (DNP). Fluorescently labelled antibodies to the biotin and DNP-modified nucleotides are used to generate and amplify an allele specific signal. The array is then scanned and intensity measurements for each SNP allele generated and normalised using an Illumina proprietary algorithm. Raw data is therefore available in the form of normalized intensity measurements (X_{Norm} , Y_{Norm}) corresponding to each allele.

Genotype calling: Genotype assignment from normalised intensities is usually performed by automated calling algorithms that seek to assign each individual to one of three genotype

cluster positions. While these procedures are efficient and help control for variables specific to the study data, visual inspection of SNP cluster plots (e.g. plot X_{Norm} vs. Y_{Norm} for all samples) is recommended for SNPs showing evidence of association to help exclude genotyping bias.

Genotyping quality controls: The success of these studies relies on accurate genotyping and a low rate of missing data. Stringent quality controls are necessary to ensure that genotype data from individual samples is of high quality, with low probability of genotyping error. Individuals are therefore excluded if systematic differences in assay intensity characteristics are observed across many SNPs. This is most readily determined from the SNP call rate, the proportion of all SNPs for which genotypes can be assigned for an individual. After removal of low quality samples, individual SNP assays must also be assessed for genotyping quality. Commonly used quality assessment steps include the proportion of samples for which genotype can be assigned, for example required to be >95%. Similarly some investigators assess differential SNP genotype missingness between cases and controls, a step which is particularly important if cases and controls have been genotyped under different conditions (e.g. different platforms). The distribution of genotypes and in particular deviation from hardy-weinberg equilibrium occurs when genotyping bias is present, but can occur for disease-associated variants for example under recessive or dominant inheritance models. Extreme deviations from hardy-weinberg equilibrium in controls are therefore commonly used to exclude SNPs as another indicator of genotyping bias.

Relatedness: Individuals that share recent ancestors will tend to share more genetic variants than unrelated individuals. Relatives within a case or control cohort can therefore increase or decrease the frequency of some variants compared to an equivalent unrelated cohort. Thus relatedness can produce spurious genetic associations. Close relatives and duplicate samples are relatively easily identified and excluded. More distant relatives are more difficult to identify and pose a problem of cryptic relatedness. The presence of cryptic relatedness may be suggested by general inflation of association test statistics. Genomic control – adjustment of individual SNP association test statistics for the degree of overall statistic inflation, is sometimes used to attempt to correct for this.

Population stratification: This occurs due to the presence of individuals from multiple source populations in the sample collection. When allele frequency differences between cases and controls arise due to differences in population origins, population stratification is said to exist

(Cardon and Palmer 2003). Allele frequencies may vary widely between populations independently of disease status, as noted previously due to unique population histories, founder effects, the effects of genetic drift and natural selection. Population stratification is one of the most common and problematic confounders of genome wide association studies. Nevertheless, consideration of this problem usually permits effective controls. A variety of methods for controlling for population stratification have been advocated, including genomic control, principal components correction and population-stratified analysis. Many researchers include parent-child trios, assessing non-random transmission of putative disease alleles to affected versus unaffected offspring (transmission disequilibrium test). This test is free from relatedness and population stratification confounding. Replication of positive associations in independent sample collections also provides some re-assurance that these effects are not driving the associations.

1.6.1.1 Genetic variation assayed by SNP genotyping chips

SNP genotyping platforms have progressively increased SNP content. For example, the Illumina Hap300 beadchip, a genotyping platform used in some earlier GWASs included more than 300,000 SNPs with minor allele frequency (MAF) > 0.05 that together tag 77% of SNPs (r^2 > 0.8) present in the HapMap CEU samples (Phase I and II). Thus a further ~23% of independent common variation discovered in HapMap CEU samples, is not well captured by this platform. Furthermore the HapMap project has not generated a complete survey of common variants, due to difficulty designing reliable SNP assays in certain genomic regions (centromeres, telomeres, gaps in genome sequence and segmental duplications)(Manolio, Brooks et al. 2008). Therefore the amount of true common variation captured is likely to be lower and moreover biased against variants in these genomic regions.

Newer genotyping platforms incorporate more SNPs and therefore provide greater coverage (see **Table 1.2**).

Available SNP genotyping platforms used in GWASs do not directly assay rare single nucleotide variants, even though rare variants can contribute to common disease susceptibility (Hugot, Chamaillard et al. 2001; Bodmer and Bonilla 2008; Ji, Foo et al. 2008).
Future platforms will supplement common SNPs with SNPs with lower minor allele frequencies, drawing from the expanding catalogues of these variants (e.g. 1000 Genomes Project).

Table 1.2 Common SNP coverage for commercially available SNP genotyping platforms. Data represent % of HapMap Phase II SNPs tagged at $r^2 > 0.8$. adapted from Manolio et al. *JCI* 2008 (Manolio, Brooks et al. 2008)

	HapMap population sample		
Genotyping platform	**CEU**	**YRI**	**CHB + JPT**
Affymetrix GeneChip 500k	68	46	67
Affymetrix SNP Array 6.0	82	66	81
Illumina HumanHap300	77	33	63
Illumina HumanHap550	88	55	83
Perlegen 600k	92	47	84
Illumina 1M	93	73	92

1.7 Glossary of genetics terms

Allelic spectrum: The number and population frequency of disease-predisposing alleles at a locus

Common, rare and private variants: These define ranges for the minor allele frequencies of variants, though consensus on precise frequencies is lacking. Approximately, common variants include those with minor allele frequency (MAF) greater than 1-5%, rare variants those with MAF between ~0.1% and 1-3% and private variants those restricted to probands and immediate relatives.

Complex disease: Synonyms include common disease, polygenic disease and multifactorial disease. Refers to diseases that, in contrast to single gene disorders, are typically common in the population and whose susceptibility is determined by both environmental factors and genetic factors. The genetic component of susceptibility for these disorders is polygenic, i.e. determined by disease-predisposing variants at multiple independent loci.

Copy number variant (and polymorphism): A form of structural variant, where a multinucleotide sequence varies in number between individuals. Usually used to refer to variants of this type, of size greater than 1 kilobase or 500 bp. Polymorphism indicates that the variant is common (minor allele frequency > 1%)

Effect size: The increase in risk conferred by a causal genetic variant

Genome wide association study: Genetic association study which tests for association between a phenotype and hundreds of thousands of genetic variants mapping across the genome. Variants are selected with the intent to capture a major proportion of population genetic variation, minimizing bias towards or against any particular genomic region. Most studies have genotyped hundreds of thousands of SNPs on commercially available SNP genotyping microarrays. Microarrays vary in SNP numbers (the latest chips include millions of SNPs) and in SNP selection, with some arrays taking advantage of HapMap determined patterns of linkage disequilibrium between SNPs to minimize redundancy and others seeking even spacing across the genome

Human Genome: This is the genetic information (DNA sequence) inherited by humans. It can be represented as a template or reference sequence formed of 22 autosomes, 2 sex chromosomes and mitochondrial DNA

Haplotype: A combination of alleles at different loci on a chromosome that are inherited together

Heritability: The proportion of phenotypic (trait) variation that is due to genetic variation. This is usually calculated by comparing phenotype correlations in individuals of varying degrees of relatedness (e.g. monozygotic and dizygotic twins)

Indel: A small insertion or deletion of nucleotides compared to a reference sequence

Linkage disequilibrium: Non-random association of alleles at different loci. Occurs when some combinations of alleles (haplotypes) occur more or less frequently in a population than would be expected based on formation of random haplotypes from their allele frequencies alone

Mendelian Disease: A disease transmitted through generations in a family in a dominant or recessive manner, typically determined by variants of large effect in a single gene.

Minor allele frequency: In the population studied, this is the proportion of alleles at a locus that are the less frequent allele (ranges from 0-50%)

Odds ratio: A relative measure of risk. In the case of an allelic odds ratio as used in genetic case-control studies this is calculated from a 2 x 2 contingency table as the ratio of allele 1 in cases versus controls (Allele 1 $count_{cases}$/Allele 1 $count_{controls}$) divided by the ratio of allele 2 in cases versus controls (Allele 2 $count_{cases}$/Allele 2 $count_{controls}$). The odds ratio is not a particularly intuitive measure, but approximates relative risk when the value is close to 1

Personalized medicine: Tailoring of preventive and therapeutic interventions for diseases on the basis of genetic profiles

Pharmacogenomics: The study of the effects of genome-wide genetic variation on drug responses.

Polymorphism: Variant with minor allele frequency greater than 1%. This threshold distinguishes polymorphisms from clearly deleterious mutations (as found in single gene disorders) which usually have frequencies less than 0.1% and even in completely recessive models rarely exceed 2-3%

Single Nucleotide Polymorphism: A nucleotide substitution that occurs in more than 1% of the population

Chapter 2 Immunogenetics and clinical aspects of coeliac disease and Crohn's disease

2.1 Coeliac Disease

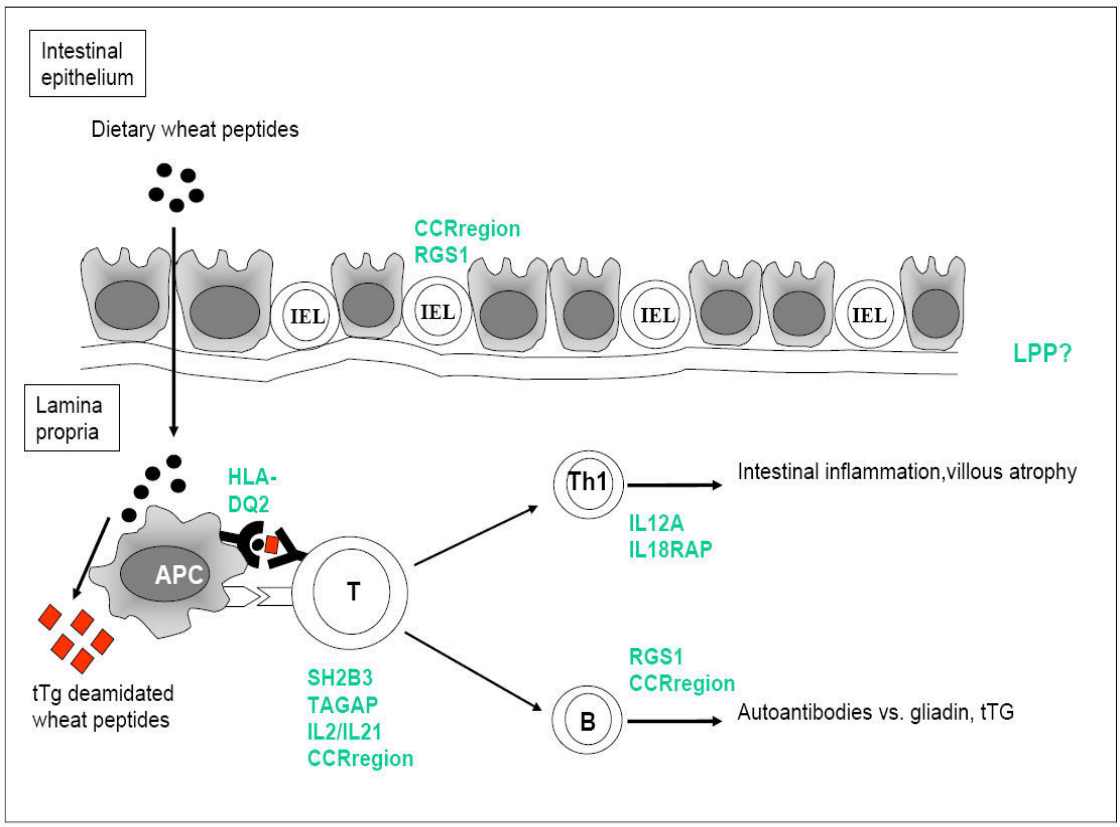
The content in this section reviews publications up to November 2009, prior to the full analysis and results of the research reported in chapter 3. Part of the text is adapted from two review publications (Dubois and van Heel 2008; Dubois, Trynka et al. 2010).

Coeliac disease is a common (~1% prevalence) inflammatory disorder of the small intestine occurring in both children and adults. Specific proteins in dietary wheat, rye and barley (gliadin, secalins, hordeins, usually referred to as "gluten") induce T cell responses restricted by HLA-DQ2 or -DQ8. These responses are central to the intestinal inflammation and loss of villous architecture that characterizes the disease (**Figure 2.1**). Now that serological testing is widespread, symptoms observed in diagnosed individuals vary greatly and are often absent. Classical malabsorption is now infrequent, and only the most florid of the spectrum of presentations seen in coeliac disease. Strict avoidance of dietary wheat, rye and barley (a gluten-free diet) usually induces remission. Disease reappears on re-challenge and dietary treatment is lifelong.

Many of the immunological mechanisms by which dietary wheat (and to a lesser extent rye and barley) induce coeliac disease are now understood (Sollid 2002). Wheat gluten is partially digested, but key toxic protein sequences are resistant to intestinal proteases - in part due to high proline (P) and glutamine (Q) content. Tissue transglutaminase in the intestinal epithelium deamidates critical peptide sequences such as the dominant HLA-DQ2 restricted wheat epitope sequence PQPQLPY to PQPELPY, and (cross-linked to critical wheat peptides during the deamidation step) is the antigen detected by current diagnostic serological tests such as the anti-endomysial or tissue transglutaminase antibody assays. It is unclear if these antibodies have a pathological role in coeliac disease. Work using intestinal T cell clones, intestinal biopsy culture, and peripheral blood T cells in wheat antigen challenged coeliac patients, has shown that wheat peptides are presented by HLA-DQ2 (or in a few patients –DQ8) to CD4+ helper T cells. Immuno-dominant wheat (and rye, barley) epitopes that are capable of inducing T cell responses in almost all coeliac patients have been defined, and the crystal structure of these epitopes bound to HLA-DQ2 or -DQ8 has been elucidated. Activated T cells secrete interferon-

gamma, and other cytokines. Interleukin-15, expressed by intestinal epithelial cells and lamina propria macrophages, appears to activate intra-epithelial lymphocytes and leads to epithelial cell killing. Multiple pathways lead to intestinal inflammation, villous atrophy and subsequent malabsorption.

Figure 2.1 **Model of gluten induced immune response in coeliac disease, and the sites of action of coeliac susceptibility genes.**
The most likely gene from each region is shown, although note that causality of a genetic variant in any one gene has not yet been proven.



The full HLA-DQ2 heterodimer (encoded at the DNA level by the combination of HLA-DQA1*0501 and DQB1*0201) is found in ~90% of coeliac disease patients, compared to ~30% of white European population controls. The remaining 10% of coeliac disease individuals either carry HLA-DQ8, or part of the HLA-DQ2 heterodimer. Carriage of one of these HLA types is therefore necessary but not sufficient to develop coeliac disease.

The HLA only explains around 30% of the heritable risk of coeliac disease, other genetic and environmental risk factors play a major role. Genetic risk variants on chromosome 4 (in a region containing the genes for the T cell cytokines interleukin-2 and interleukin-21) as well as variants in other immune system genes have been identified more recently (van Heel, Franke et al. 2007; Hunt, Zhernakova et al. 2008). Several of these have independently been shown to influence risk to other autoimmune diseases, especially type 1 diabetes mellitus (Smyth, Plagnol et al. 2008). The timing of the introduction of wheat during infant feeding is probably important, some studies suggesting that continued breastfeeding whilst weaning is protective (Ivarsson, Hernell et al. 2002; Ivarsson 2005). Whether gastrointestinal infections (e.g. rotavirus) in infancy are important triggers remains unclear (**section 2.1.1**).

2.1.1 Epidemiology

Serological screening of populations in Europe and regions with a high proportion of European descendents (North and South America, Australasia) suggests a coeliac disease prevalence of approximately 0.5-1% in adults (West, Logan et al. 2003; van Heel and West 2006). More limited data from North Africa and South West Asia suggest similar high prevalence of coeliac disease in these areas(Accomando and Cataldo 2004). In central Africa and the Far East there have been no large seroprevalence studies but overt coeliac disease is extremely rare (Bonamico, Mariani et al. 1994; Fasano, Berti et al. 2003; Freeman 2003). A study from Burkina Faso screened 600 individuals, all of whom ate wheat, but found no individuals with positive coeliac serology. Furthermore no individuals carried HLA-DQ2 and only one HLA-DQ8 (Cataldo, Lio et al. 2002). The Saharawi population of North Africa have the highest reported prevalence of coeliac disease worldwide(5.4%) mirrored by a very high carriage of the coeliac susceptibility marker HLA DQ2, whereas the prevalence of HLA DQ2 is very low in the far east (Catassi, Doloretta Macis et al. 2001; Fasano and Catassi 2001). Genetic differences across populations (particularly in HLA types) clearly contribute to the different observed population prevalences.

Some human populations have been exposed to gluten for around 10,000 years (Accomando and Cataldo 2004). Regions of the world where indigenous people have cultivated wheat, rye and barley (North Africa, Europe, South West Asia) paradoxically have higher population frequencies of coeliac disease-predisposing HLA-DQ alleles than regions where people have cultivated other grains (e.g. central Africa- millet and sorghum, South East Asia – rice, America

– maize). The reason is unknown. However, coeliac diseases and other autoimmune diseases appear to have become common relatively recently with incidence still clearly on the increase. This rapid increase can only be due to changing environmental factors, and implies that coeliac disease and other autoimmune diseases may not have been common for enough time to allow negative selection of deleterious HLA alleles.

Grain consumption broadly parallels coeliac prevalence, being low in the far east and sub-Saharan Africa(Fasano and Catassi 2001). Furthermore there is some evidence that the dose of gluten, particularly in early childhood, may be an important determinant of lifetime susceptibility. Countries in which infant gluten consumption is low (Denmark, Estonia, Finland) report a lower infant (and adult) incidence of coeliac disease than countries with a high infant gluten consumption (Sweden) (Weile, Cavell et al. 1995; Mitt and Uibo 1998).

Adult coeliac disease prevalence has been increasing over the last few decades (van Heel and West 2006). Improved clinical ascertainment contributes (especially in the USA), although some studies suggest a true increase in seroprevalence (Lohi, Mustalahti et al. 2007). Similar increases in prevalence have occurred in other chronic immune-mediated diseases, particularly type 1 diabetes, implicating recent changes in shared environmental factors (EUORDIAB ACE Study Group 2000). These factors remain unknown, although interest has focused logically on exposures occurring in early childhood, which might be critical in determining lifetime autoimmune disease risk. In coeliac disease, onset can occur at any age but the peak incidence is between 9 and 24 months, following the introduction of gluten into the diet (van Heel and West 2006). Breast feeding during gluten introduction has been shown to reduce susceptibility, suggesting that tolerance to gluten can be influenced by factors in breast milk(Ivarsson, Hernell et al. 2002). Tolerance to gluten might also be influenced by the context in which it is encountered by the mucosal immune system in early life. Childhood intestinal infections have been proposed as a factor that could promote loss of tolerance to gluten, possibly due to disrupted intestinal epithelial barrier function. Furthermore, inflammation up-regulates tissue transglutaminase, a key enzyme in coeliac disease required for the generation of immunogenic epitopes from gluten (Ientile, Caccamo et al. 2007). There are no animal models of coeliac disease to test this hypothesis and direct evidence for the role of intestinal infections is lacking. However, epidemiological studies have shown that coeliac disease is more common in children born in summer months, possibly due to the higher incidence of viral enteritis in winter months when these children start eating gluten(Ivarsson,

Hernell et al. 2003) Case-control studies have also suggested that increased exposure to infant enteral infections may confer modest increased susceptibility (odds ratios of 1.4-1.5) (Sandberg-Bennich, Dahlquist et al. 2002; Ivarsson 2005). Finally, one prospective study measured episodes of rotavirus infection by serology and found a modest increase in coeliac autoantibody incidence in infants exposed to multiple infections (Stene, Honeyman et al. 2006).

Although the development of coeliac disease has been considered a permanent gluten sensitive enteropathy, needing life-long treatment, recent reports suggest some children can at least partially resolve this intolerance when kept on a gluten-containing diet (Matysiak-Budnik, Malamut et al. 2007; Simell, Hoppu et al. 2007). These children may have normal small intestinal histology in adulthood, suggesting that coeliac disease can remit or enter a quiescent phase, with immunological tolerance to gluten, following initial clinically overt disease. How frequently this phenomenon occurs is unclear, much more research in this area is necessary – including whether such remission might be therapeutically induced.

2.1.2 Evidence for genetic susceptibility

The largest twin study, from Italy, reported that monozygotic twins have disease concordance rates of 75% compared to 11% in dizygotic twins (Greco, Romino et al. 2002). The authors note that the monozygotic twin concordance is likely in fact to be an underestimate, due to the fact that some of the discordant twin pairs were children and may therefore become concordant as they age. Interestingly, in 90% of concordant twin pairs the age at diagnosis differed by less than 2 years (median 1 month) (Nistico, Fagnani et al. 2006). If this does indeed reflect the timing of disease onset rather than just the timing of testing, the lack of difference between MZ and DZ twins would be consistent with age of onset being determined by environmental exposures rather than genetic factors. For coeliac disease, the best estimates have used modern serological screening for coeliac disease antibodies (endomysial antibodies, tissue transglutaminase antibodies) and confirmatory biopsy in siblings of coeliac index cases. In European ancestry populations sibling risk is around 10%, with a recurrence ratio (λ_s) of 10-20 based on population prevalence estimates of 0.5-1% (Bourgey, Calcagno et al. 2007; Rubio-Tapia, Van Dyke et al. 2008). This is similar to other polygenic immune mediated disorders

such as type 1 diabetes (λ_s=15), rheumatoid arthritis (λ_s=2-8) and Crohn's disease (λ_s=27) (Lewis, Whitwell et al. 2007).

The heritability of coeliac disease has been estimated as 87% using the Italian twin study data cited above and assuming a disease prevalence of around 1% (Nistico, Fagnani et al. 2006). As of 2009, prior to research presented in this thesis, it was estimated that around 40% of the heritable fraction of coeliac disease risk could be accounted for by known genetic variants, around 35% attributable to HLA DQ variants alone (Hunt, Zhernakova et al. 2008).

2.1.3 Immunogenetics of the HLA

The Human Leukocyte Antigen (HLA) complex is a highly polymorphic 4 Mb region on chromosome 6p21, containing more than 200 genes and over 3000 known alleles (Robinson, Waller et al. 2003). HLA class II genes (DP, DQ, and DR) are involved in exogenous peptide antigen presentation to T cells. The first reports of association with coeliac disease used serological methods to identify B8 and later DR3 as susceptibility variants (Falchuk, Rogentine et al. 1972; Keuning, Pena et al. 1976). The B8 and DR3 molecules are encoded by alleles on a 6Mb extended haplotype (A1-B8-DR3-DQ2) present in 10 % of Northern Europeans(Price, Witt et al. 1999). Interestingly, other autoimmune diseases are associated with this haplotype, including type 1 diabetes and autoimmune thyroid disease. Subsequent studies have pinpointed DQ2 and in particular the combination of HLA-DQA1*0501 and DQB1*0201 encoding the HLA-DQ2 (α1*0501,β1*0201) heterodimer as the cause of the coeliac disease association (Tosi, Vismara et al. 1983). This heterodimer can be encoded both in *cis* (by alleles on the same haplotype) or in *trans* (one subunit each from paternal and maternal haplotypes) (**Table 2.1, Figure 2.2**). Moreover several studies show that homozygosity for the *cis* haplotype or possessing a second DQB1*02 allele increases coeliac disease susceptibility further (Louka, Nilsson et al. 2002; van Belzen, Koeleman et al. 2004). The second B1*02 allele is usually inherited on the DR7-DQ2 haplotype carrying DQB1*0202 and DQA1*0201 (DQ2.2) but possession of this haplotype alone does not confer coeliac susceptibility (**Table 2.1**).

An explanation for the HLA gene-dosage effect was provided by an *in vitro* study demonstrating that the level of proliferation and cytokine responses of gluten-reactive T cell clones depends on DQ type and gene dose (Vader, Stepniak et al. 2003). Vader et al. used allogeneic peripheral blood mononuclear cells to present gluten epitopes to gluten-specific T

cell clones and showed that T cell responses were highest for DQ2.5 homozygotes, intermediate for DQ2.5/2.2, lower for DQ2.5/x heterozygotes and lowest for DQ2.2. Thus DQ2.2 in the presence of DQ2.5 can augment T cell stimulation through DQ2 mediated antigen presentation. Interestingly DQ2.2 alone, which is not associated with coeliac disease, was able to elicit strong T cell responses but only through presentation of a restricted subset of the gluten epitopes tested. This suggests that the DQ2 contribution to coeliac disease depends on its ability to present multiple closely related gluten epitopes- the ability of DQ2.2 molecules to present a small subset of epitopes exerts effects too weak to cause disease.

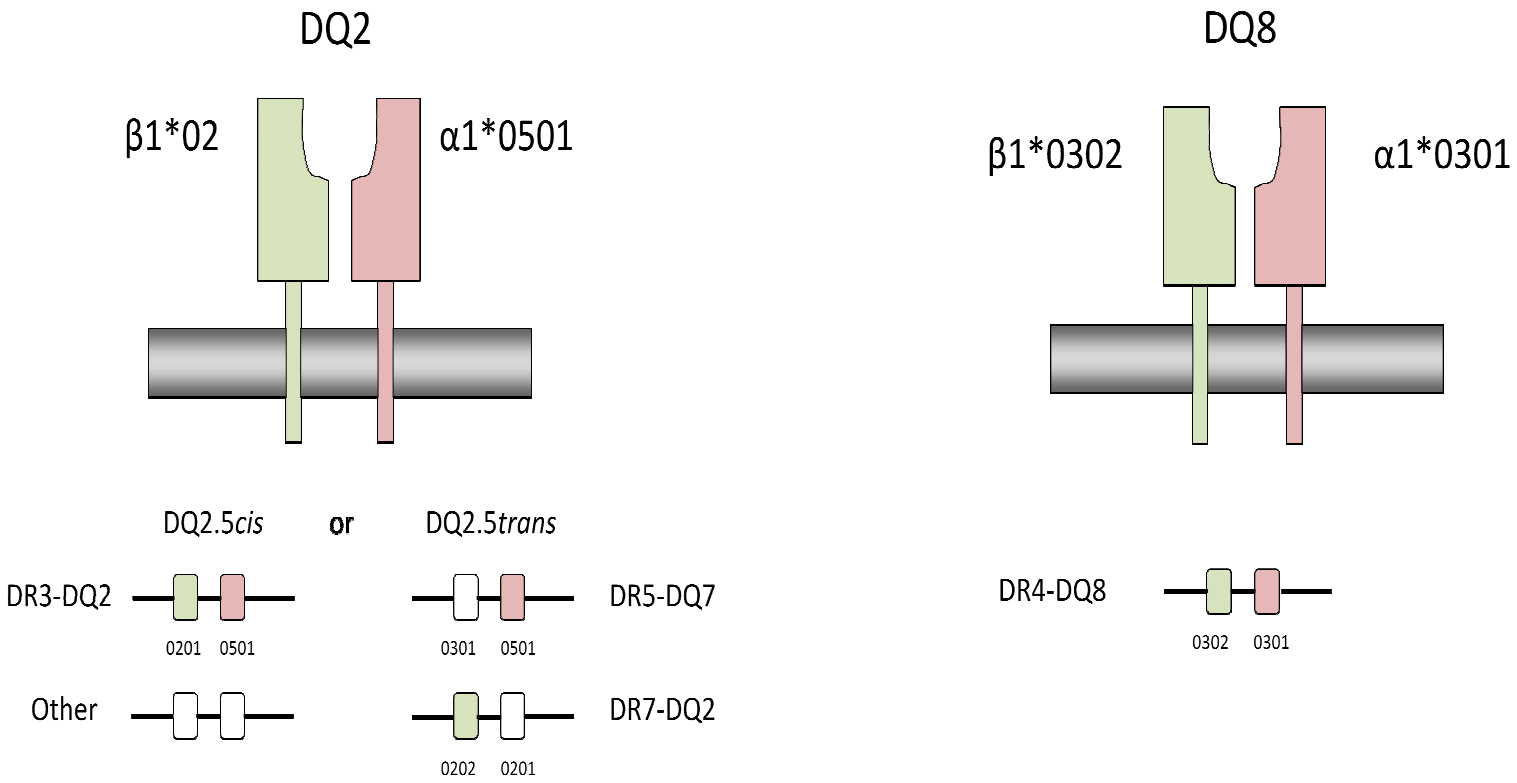
The HLA-DQ2.5 molecule encoded either in *cis* or in *trans* is present in around 90% of coeliac patients of Northern European origin(Sollid, Markussen et al. 1989). The majority of the remainder carry HLA-DQ8 (genetically DQA1*03, DQB1*0302) (Spurkland, Sollid et al. 1992; Karell, Louka et al. 2003). A large European collaborative study found that of those that lack both DQ2 and DQ8, only 4 of 1008 coeliac patients had neither the alpha or beta chain of the DQ2 heterodimer (Karell, Louka et al. 2003) This has led to a model of coeliac disease pathogenesis in which HLA DQ2/8 is necessary but not sufficient, since HLA-DQ2 is present in 30% of healthy Caucasian populations(Sollid, Markussen et al. 1989). The proportion of sibling relative risk attributable to known HLA variants is estimated to be between 30 and 40%, indicating non-HLA DQ variants contribute to coeliac disease susceptibility (Petronzelli, Bonamico et al. 1997; Bevan, Popat et al. 1999; Hunt, Zhernakova et al. 2008). Within the HLA complex itself there are many other genes with immune functions which might also contribute to the observed association signal. However, the high linkage disequilibrium (LD) that exists between genetic variants in this region is an obstacle to teasing out the true causal associations (Louka and Sollid 2003). Two studies that have controlled for LD to DQ have not found evidence of additional HLA risk variants, although statistical power was limited (Karell, Louka et al. 2003; Louka, Moodie et al. 2003).

Table 2.1 **Classical HLA DQ genotypes associated with coeliac disease and gene dosage effects**

Serological type	Chromosome Copy	DQ2 Genotype	DQ type	Coeliac susceptibility
DR3-DQ2/ DR3-DQ2	i ii	DQA1*0501-DQB1*0201/ DQA1*0501-DQB1*0201	DQ2.5*cis* homozygote	High
DR3-DQ2/ DR7-DQ2	i ii	DQA1*0501-DQB1*0201/ DQA1*0201-DQB1*0202	DQ2.5*cis* + DQ2.2	High
DR3-DQ2/ Other	i ii	DQA1*0501-DQB1*0201/ Other	DQ2.5*cis* heterozygote	Moderate
DR5-DQ7/ DR7-DQ2	i ii	DQA1*0501-DQB1*0301/ DQA1*0201-DQB1*0202	DQ2.5*trans*	Moderate
DR7-DQ2/ Other	i ii	DQA1*0201-DQB1*0202/ Other	DQ2.2	Nil
DR4-DQ8/ other	i ii	DQA1*0301- DQB1*0302/ Other	DQ8	Moderate

Disease causing alleles highlighted (see also **Figure 2.2**)
Adapted from van Heel et al. 2005 (van Heel, Hunt et al. 2005)
DQ2 type naming after Vader et al. (Vader, Stepniak et al. 2003)

Figure 2.2 **Haplotype combinations encoding the HLA-DQ2 and –DQ8 heterodimers.**
Adapted from Sollid, 2000. HLA proteins at the cell surface, and structure of the protein encoding DNA region, are shown.



The genetic loci harbouring variants that account for the remaining 70% or so of unexplained familial clustering in coeliac disease are the targets of gene finding studies. Two complementary approaches have been used: genetic linkage and association studies (**Table 2.1**).

In general, findings from linkage and candidate gene studies in coeliac disease, with the exception of the HLA region, have not been replicated consistently. Linkage regions identified include 5q31-33 and 19p13.1, although these remain tentative and lack robust replication (Greco, Babron et al. 2001; Van Belzen, Meijer et al. 2003). *MYO9B*, encoding the myosin IXB protein has emerged as a candidate gene from further studies of the 19p13.1 linkage region, although replication of this finding has been inconsistent (Monsuur, de Bakker et al. 2005; Amundsen, Monsuur et al. 2006; Hunt, Monsuur et al. 2006; Sanchez, Alizadeh et al. 2007). A candidate gene approach identified an association in the *CTLA4* region, a gene on chromosome 2q encoding cytotoxic T lymphocyte antigen 4 (Djilali-Saiah, Schmitz et al. 1998). CTLA-4 is expressed on T cells and is a receptor for B7 molecules that inhibit T cell activation. Replication studies of the *CTLA4* association have been somewhat inconclusive (van Heel, Hunt et al. 2005). Therefore, prior to the first genome wide association study in 2007, despite intensive efforts, no genetic susceptibility loci other than HLA DQ had been definitively identified.

2.1.4 HLA-DQ restricted T cells

Coeliac disease has multi-systemic features, but the predominant lesion mirrors the exposure of the small intestine to dietary gluten. Several lines of evidence implicate a T cell-orchestrated immunopathogenesis. Upon gluten challenge of small intestinal biopsies from treated (i.e. on a gluten-free diet) coeliac disease patients, infiltration of the lamina propria with (predominantly CD4+ αβ) T cells occurs within hours, followed by crypt hyperplasia and villous atrophy(Anand, Piris et al. 1981). This temporal sequence alludes to the central importance of T cells in coeliac disease. In untreated disease Th1 cytokines are highly expressed in the intestinal mucosa, particularly interferon (IFN)-γ, supporting the concept of a Th1 driven T cell mediated disorder (Nilsen, Jahnsen et al. 1998). Analysis of lamina propria (LP) infiltrating lymphocytes confirms not only IFN-γ expression in a high proportion, but also expression of the Th1 transcription factor T-bet (Monteleone, Monteleone et al. 2004). The Th1 bias of CD4+ T cells probably depends less on IL-12 in coeliac disease than in other inflammatory conditions. IL-12 is present in very low levels in coeliac disease mucosa (Salvati,

MacDonald et al. 2002; Di Sabatino, Pickard et al. 2007) although other Th1 inducing cytokines (IL-18 and IFN-α) are increased (Monteleone, Pender et al. 2001; Salvati, MacDonald et al. 2002; Leon, Garrote et al. 2006). Dendritic cells isolated from the intestinal mucosa in coeliac disease also express increased levels of IL-18 and IFN-α but lack IL12p40 (Di Sabatino, Pickard et al. 2007). Immunophenotyping of DQ2+ antigen presenting cells in treated versus untreated coeliac disease intestinal biopsies suggest a large increase in CD11+ myeloid dendritic cells in active disease (Raki, Tollefsen et al. 2006; Di Sabatino, Pickard et al. 2007). These cells efficiently present gluten peptides to CD4 T cells inducing proliferation and IFN-γ responses (Raki, Tollefsen et al. 2006).

The gluten-responsiveness of CD4 T cells in coeliac disease was first demonstrated in T cell lines and clones isolated from intestinal mucosa (Lundin, Scott et al. 1993; Lundin, Scott et al. 1994). These cells are not found in non-coeliac DQ2- or DQ8- controls but in coeliac disease proliferate and secrete IFN-γ when co-cultured with antigen presenting cells in the presence of a variety of peptides derived from gluten. These studies show that gluten peptides activate T cells in the intestinal mucosa exclusively through presentation by the disease-associated DQ 2- or DQ8- αβ heterodimers (Lundin, Scott et al. 1993; Lundin, Scott et al. 1994).

2.1.5 Gluten epitopes and the role of tissue transglutaminase

While there is heterogeneity between patients with coeliac disease in the gluten epitopes to which their T cells respond, some epitopes are immunodominant and elicit T cell activation in almost all coeliac individuals (Anderson, Degano et al. 2000; Arentz-Hansen, Korner et al. 2000). These responses have been demonstrated both in intestine-derived T cell lines or clones and in primary T cells isolated from peripheral blood following gluten challenge, supporting their contribution to disease *in vivo*(Anderson, Degano et al. 2000; Anderson, van Heel et al. 2005). T cell epitopes identified to date are derived from various gluten proteins, including α-gliadins, γ-gliadins and low molecular weight glutenins (Sjostrom, Lundin et al. 1998; Arentz-Hansen, Korner et al. 2000; Arentz-Hansen, McAdam et al. 2002; Vader, Kooy et al. 2002). The peptide binding groove structure of DQ2 and DQ8 dimers has been characterized and some of the constraints this places on selection of epitopes for binding DQ2 or 8 is known (Tollefsen, Arentz-Hansen et al. 2006) Both DQ2 and DQ8 dimers have preferences for negatively charged residues at key positions in the core peptide binding groove(Vartdal, Johansen et al. 1996; Godkin, Friede et al. 1997; van de Wal, Kooy et al. 1997). Negatively charged residues are

uncommon in gluten peptide sequences, but deamidation of glutamine residues to negatively-charged glutamate can drastically increase the immunogenicity of gliadin peptides (Sjostrom, Lundin et al. 1998). X-ray crystallographic analysis of DQ2- peptide interactions supports the importance of selective deamidation of glutamine residues in favouring peptide binding for gluten peptides (van de Wal, Kooy et al. 1998; Kim, Quarsten et al. 2004). Tissue transglutaminase (tTG), an enzyme first linked to coeliac disease by the discovery that it is the target of autoantibodies used in diagnosis, can catalyse this deamidation (Dieterich, Ehnis et al. 1997; Molberg, McAdam et al. 1998). tTG is likely to perform this function *in vivo* as it highly expressed in the small intestine, up-regulated in inflammation and favours deamidation of glutamine residues rather than transamidation under the acidic conditions which exist in the proximal small intestine (Fleckenstein, Molberg et al. 2002). For example, the immunodominant gluten peptide epitope PQPQLPY is deamidated to PQPELPY by tissue transglutaminase (Arentz-Hansen, Korner et al. 2000). More recently, a direct pathogenic contribution of tTG antibodies has been proposed, with *in vitro* studies suggesting that these antibodies can both activate monocytes by binding toll-like receptor 4 and inhibit angiogenesis by altering tTG function (Zanoni, Navone et al. 2006; Myrsky, Kaukinen et al. 2008). Such effects, if substantiated, may be a mechanism driving extra-intestinal manifestations in coeliac disease, since tTG autoantibody deposits have been observed in affected organs (e.g. liver, brain) remote from the site of gluten exposure in the intestine (Korponay-Szabo, Halttunen et al. 2004; Hadjivassiliou, Maki et al. 2006).

A further important characteristic of gluten epitopes is a high proline content(Arentz-Hansen, McAdam et al. 2002). This reflects the inability of human digestive enzymes to break amide bonds between proline residues and adjacent bulky hydrophobic amino acids, such that gluten peptides can reach the intestinal mucosa intact (Arentz-Hansen, McAdam et al. 2002; Hausch, Shan et al. 2002)

2.1.6 The innate immune system in coeliac disease

Both *in vivo* studies and studies of gluten challenge of intestinal biopsies have shown that effects on the mucosa begin within a few hours (Sturgess, Day et al. 1994; Maiuri, Picarelli et al. 1996; Fraser, Engel et al. 2003). This rapid onset cannot easily be accounted for by the (presumably slower) mechanism of gluten peptide presentation to CD4+ T lymphocytes and

has led to interest in a role for the innate immune system in coeliac disease. Further support for this hypothesis came from the observation that some gliadin peptides (p31-p43 α gliadin) that do not elicit classical DQ-restricted CD4+ T cell responses, can exert toxic effects on the epithelium (Maiuri, Ciacci et al. 2003). Interleukin-15 (IL-15), which is highly expressed in lamina propria macrophages and intestinal epithelium, appears to be a crucial intermediary of these effects. IL-15 enhances intraepithelial lymphocyte (IEL) proliferation, cytotoxicity (vs. epithelial cells) and cytokine release, with increases in IFN-γ and granzyme B (Mention, Ben Ahmed et al. 2003; Di Sabatino, Ciccocioppo et al. 2006) Furthermore, exogenous application of IL-15 partly reproduces the effects of gliadin challenge whereas anti-IL-15 antibodies abrogate the effects of gliadin (Mention, Ben Ahmed et al. 2003).

A feature of coeliac disease is expansion of the IEL population, as well as an inflammatory cell infiltrate deeper in the intestinal lamina propria. The IELs in coeliac disease comprise increased populations of both CD8+ TCRαβ lymphocytes as well as γδ (CD4-CD8- or CD8+) T cells that can directly induce enterocyte apoptosis (Jabri, de Serre et al. 2000). Some intra-epithelial T cells have been shown to demonstrate aberrant expression of NK lineage receptors and can perform NK-like functions including T cell receptor- independent killing of enterocytes in active coeliac disease (Jabri, de Serre et al. 2000; Hue, Mention et al. 2004; Meresse, Curran et al. 2006). These effects are stimulated by gluten peptides including p31-43 α-gliadin and include the induction of expression of the cell surface stress molecule MICA on enterocytes and its receptor NKG2D on IELs (Hue, Mention et al. 2004). Mechanistic details of the recognition of these apparently 'innate' peptides are unclear.

2.1.7 Genetic risk variants in coeliac disease

The first GWAS in coeliac disease tested 310,605 SNPs for association in 778 individuals with coeliac disease and 1422 controls (van Heel, Franke et al. 2007). Coeliac cases were recruited from hospital outpatient clinics in the United Kingdom and genotyped on the Illumina Human Hap300 v1 Beadchip (Illumina inc., San Diego, USA). Control data was obtained from individuals within the 1958 birth cohort who had been genotyped on the Illumina Human Hap550 v1 Beadchip as part of a collaboration with the Wellcome Trust Case Control Consortium (WTCCC). Association (strongest for the HLA DQ2.5*cis* tagging SNP rs2187668, $P < 10^{-19}$) mapping to the HLA region was confirmed. Outside of the HLA, a 480 kilobase region of strong linkage disequilibrium on chromosome 4q27 showed the strongest association

(rs13119723, p = 2.0 x 10^{-7}). Replication was confirmed for SNPs in this region in Dutch and Irish cohorts totalling 991 coeliacs and 1489 controls. A follow-up study tested 1020 of the most strongly associated SNPs in the GWAS in a further 1643 cases and 3406 controls, comprising additional UK cases and controls, as well as Dutch and Irish collections(Hunt, Zhernakova et al. 2008). This study identified a further 7 genomic risk regions harbouring SNPs with strong evidence of association ($P < 5 \times 10^{-7}$) (table 1). Of eight non-HLA regions identified in these studies, seven regions harbour genes with known immune functions. These genes implicate T cell signalling (*IL2/IL21, IL18RAP, IL12A, TAGAP, SH2B3*) and the control of lymphocyte trafficking (*CCR* gene region*, RGS1*) (Figure 1).

A second follow-up study to the first coeliac GWAS, tested 458 SNPs showing more modest association in the GWAS in the same UK, Dutch and Irish cohorts used in the first follow-up study(Trynka, Zhernakova et al. 2009). This study identified two new coeliac susceptibility regions, 6q23.3 (*OLIG3-TNFAIP3*) and 2p16.1 (*REL*), both of which reached genome-wide significance in the combined analysis of all 2987
cases and 5273 controls (rs2327832 p = 1.3x10-08, and rs842647 p = 5.2x10-07).

A number of the coeliac susceptibility regions contain variants influencing susceptibility with other autoimmune diesases, particularly type 1 diabetes (Table 1). SNPs from 18 loci associated with type 1 diabetes were tested in 2560 UK individuals with coeliac disease and 9339 controls(Smyth, Plagnol et al. 2008). In this study new evidence for association with coeliac disease of type 1 diabetes variants at the *CTLA4* (p=1.26 x 10^{-6}) and *PTPN2* (p=2.61 x 10^{-4}) loci was observed.

Finally in a follow-up study of the coeliac GWAS data in indviduals from the United States, strongly suggestive evidence for a further new coeliac risk locus on chromosome 2q31.3 that includes the *ITGA4* (integrin alpha 4) gene was observed (Garner, Murray et al. 2009).

Thus, as of June 2009, the coeliac loci with strong evidence of disease association include HLA-DQ, eight loci from the first GWAS and follow-up, 3 loci from 2 further follow-up studies and 2 loci from the type 1 diabetes- coeliac study. This is a current total of 14 genomic regions in which variants influence susceptibility to coeliac disease.

2.1.8 Function of non-HLA coeliac genes

2.1.8.1 IL2-IL21 region

Outside of the HLA region, the strongest marker from the first UK coeliac disease GWAS mapped to chromosome 4q27 (P =2 x 10^{-7}), a finding replicated in further UK, Dutch and Irish cohorts (Hunt, Zhernakova et al. 2008). The associated SNP tags a ~700kb linkage disequilibrium (LD) block encompassing 4 genes (*ADAD1, KIAA1109, IL2 and IL21*); the genetic association signal cannot differentiate between these genes. This region is emerging from other studies as a common autoimmune disease locus, with association to type 1 diabetes, Rheumatoid arthritis, psoriasis and Graves' disease (Todd, Walker et al. 2007; Liu, Helms et al. 2008; Raychaudhuri, Remmers et al. 2008; Barrett, Clayton et al. 2009). The most compelling biological candidates within the LD block are *IL2* and *IL21*.

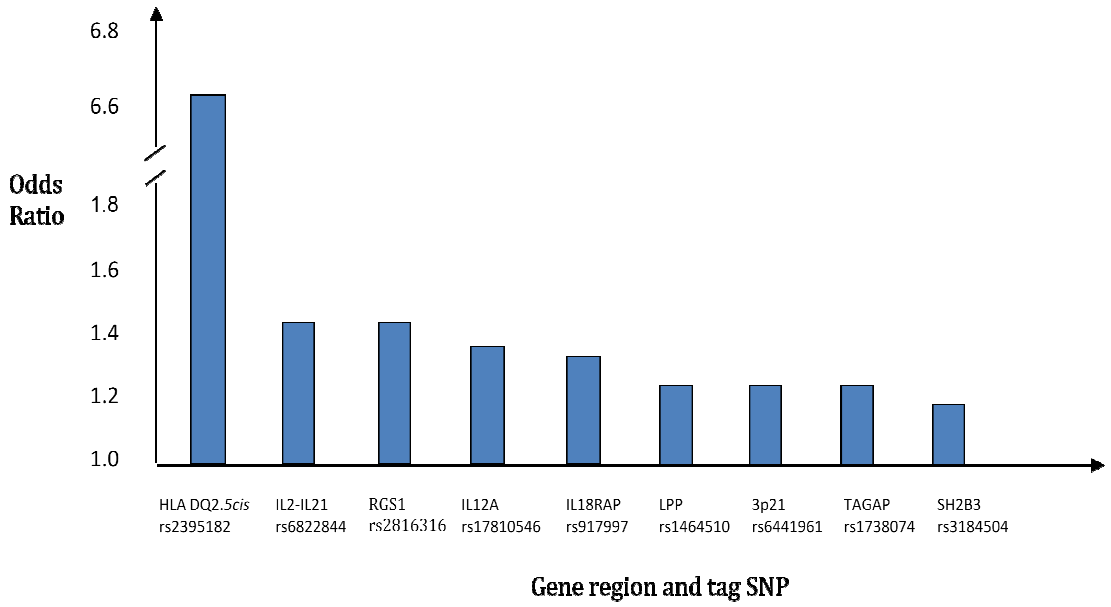
Interleukin-2 and interleukin-21 are members of the same cytokine family, sharing the same γ chain subunit in their receptors (Waldmann 2006). These cytokines have multiple and diverse roles in the immune response, posing a challenge in identifying the precise biological mechanisms relevant to coeliac disease. Interleukin-2 has a well defined autocrine function in stimulating T cell activation and proliferation, but can also stimulate natural killer (NK) cell proliferation and immunoglobulin production from B cells. This cytokine has a unique role in activation induced cell death, a process that eliminates self-reactive T cells, and in maintenance of CD4+ CD25+ regulatory T (T_{Reg}) cells (Lenardo 1996; Fontenot, Rasmussen et al. 2005; Maloy and Powrie 2005). In the non-obese diabetic (NOD) mouse model the region syntenic to human 4q27 determines susceptibility to multiple autoimmune diseases through an *IL2*-dependent mechanism (Yamanouchi, Rainbow et al. 2007). In this model, the murine risk variants were associated with reduced *IL2* gene expression, lower proportions of CD4+ CD25+ T_{Reg} cells in mesenteric lymph nodes and impaired function of these cells (Yamanouchi, Rainbow et al. 2007). It is thus possible that the *IL2-IL21* region risk variants in human coeliac disease might also exert their susceptibility effects through the CD4+ CD25+ T_{Reg} cell subset, for example by impairing tolerance to gluten peptides. However, in humans, there is as yet no comparable data of the effects of variants on gene expression or function. *IL21* remains a candidate gene in this region and expression is known to be increased in the small intestinal mucosa in untreated coeliac disease (Fina, Sarra et al. 2007). IL-21 is secreted mainly from CD4+ T cells and has proinflammatory effects including enhancement of B, T and NK cell

proliferation (Leonard and Spolski 2005). Anti-IL-21 antibodies in an *ex-vivo* intestinal biopsy culture model reduced T-bet and IFN-γ expression suggesting that interleukin-21 may be important in sustaining Th1 activity in coeliac disease (Fina, Sarra et al. 2007). The disease-predisposing SNP of rs6822844 (the most strongly associated SNP in the region) is correlated with serum IL-21 and is associated with induction of Graves' disease, suggesting it may play a role in induction of autoimmunity(Jones, Phuah et al. 2009).

The largest follow-up study of the first coeliac GWAS tested over 1000 of the most strongly associated non-HLA SNPs from the original UK GWAS in a large independent cohort (1643 new coeliac cases and 3406 controls) (Hunt, Zhernakova et al. 2008). The added power of this study yielded strong evidence ($P<5*10^{-7}$) for a further 7 new genomic regions, six of which harbour genes with immune functions (**Table 3.1 –chapter 3, Figure 2.1**). It was estimated in this follow-up study that the newly identified variants accounted for only 3-4% of the genetic susceptibility of coeliac disease, suggesting that many other true associations remain undetected. Effect sizes of the SNPs on disease susceptibility are modest in line with findings from genome wide association studies in other complex diseases (**Figure 2.3**) (Mathew 2007; Todd, Walker et al. 2007). The allele that is more frequent in cases can confer either protective or risk effects with odds ratios of all detected variants between 0.7 and 1.4. Given that there are an estimated 8 million SNPs with MAF > 5% in the human genome and only 300,000 SNPs were tested in the original GWAS, in most cases associated SNPs are unlikely to be causal, but instead will show variable levels of correlation with the true causal variants. Identification of the true causal variants is a priority of further research and will depend on fine-mapping and/or deep re-sequencing of the regions identified. Indications from other diseases suggest discovery of the true causal variants may lead to a significant upwards revision of both effect sizes and the estimated proportion of genetic susceptibility accounted for (Mathew 2007). In the interim, the primary significance of the genome wide association study findings is in providing new insights into the biological pathways relevant to the pathogenesis of coeliac disease.

All 8 non-HLA regions identified in the first coeliac GWAS and a large follow-up were identified by SNPs showing association and a Wellcome Trust Case Control Consortium advocated significance genome-wide threshold of $P<5 \times 10^{-7}$ (Wellcome Trust Case Control Consortium 2007). Causal variants in these regions are unknown.

Figure 2.3 **Estimates of effect size conferred by coeliac disease associated risk variants identified from the first GWAS and follow-up study (March 2008).**
Allelic odds ratios are shown for the best tag markers from GWAS, along with the most likely candidate gene(s) from each region. It is probable that the effect of the causal variants at non-HLA loci, once identified, will be larger.



2.1.8.2 RGS1 region

The strongest association ($p=2.58 \times 10^{-11}$) outside of the HLA region and *IL2-IL21* was for a SNP 8kb distal to the 5' end of *RGS1*. *RGS1* is of particular interest in coeliac disease because of its selective expression in the intestinal intra-epithelial lymphocyte compartment, but not conventional splenic or thymic T cells (Pennington, Silva-Santos et al. 2003; Hunt, Zhernakova et al. 2008). f*RGS1* regulates G protein signalling activity and is implicated in mice in regulating chemokine receptor signalling and B cell trafficking to lymph nodes (Han, Moratz et al. 2005).

2.1.8.3 3p21

Another strong association mapped to a chemokine receptor gene cluster on 3p21 including *CCR1, CCR2, CCRL2, CCR3, CCR5* and *XCR1* again hinting at the importance that chemokine receptor signalling and recruitment of effector immune cells to sites of inflammation may have in coeliac disease. The disease associated genetic variants may subtly influence these pathways.

2.1.8.4 *IL12A* and *IL18RAP*

Strong association ($p=10^{-9}$) of SNPs in a 70Kb LD block immediately 5' of *IL12A* implicate this gene, which encodes IL12p35, the subunit that forms one half of the interleukin 12 heterodimer with IL12p40. Interleukin-12 is expressed by antigen presenting cells and has a broad range of biological activities including induction of interferon-γ secreting Th1 cells. Although coeliac disease is characterized by a strong Th1 response, surprisingly IL12p40 is not expressed in coeliac disease mucosa after gluten challenge and both IL12p40 and IL12p35 expression were not found to be increased in dendritic cells isolated from untreated coeliac disease mucosa (Nilsen, Jahnsen et al. 1998; Di Sabatino, Pickard et al. 2007). It might well be in coeliac disease that IL12 signalling is important at an alternative site (mesenteric lymph nodes?) – attempting to make sense of these findings really highlights our limited knowledge of the primary underlying immunopathogenic mechanisms.

There is evidence for the importance of IFN-α and IL-18 in promoting a Th1 phenotype in CD4 T cells in coeliac disease (see below). *IL18* transcripts are very strongly expressed in the human small intestine. In this regard, another candidate gene identified from the GWAS (*IL18RAP)* encodes the β chain of the IL-18 receptor. Hunt et *al.* showed that the coeliac disease associated SNPs correlated with *IL18RAP* gene expression in peripheral blood. The risk alleles, found more commonly in individuals with coeliac disease, correlated with lower levels of *IL18RAP* mRNA suggesting that variants reduce gene expression. This might suggest a loss of function of IL-18 receptor signalling, a puzzling finding given the up-regulation of IL-18 and strong Th1 bias in coeliac disease. Again, these findings underline the limitations of current immunological models of coeliac and other immune mediated diseases, but also provide clues to inform the design of new functional studies.

2.1.8.5 *SH2B3 region*

SH2B3 is expressed in immune cells, up-regulated in coeliac mucosa and thought to function in regulation of T cell receptor, growth factor and cytokine receptor mediated signalling (Li, He et al. 2000; Velazquez, Cheng et al. 2002). A non-synonymous SNP (rs3184504) in *SH2B3* was associated with coeliac disease in the follow-up study. The same SNP is associated with type 1 diabetes, accounting entirely for the association in the latter disease (Todd, Walker et al.

2007). This SNP, in exon 3 of *SH2B3* leads to an amino acid substitution (R262W) in the pleckstrin homology (PH) domain of the SH2B3 protein. Pleckstrin homology domains are involved in targeting proteins to plasma membranes through binding phosphoinositides (Lemmon 2008). Mutations in PH domains in other proteins have been associated with disease by impairing phosphoinositide binding and membrane localisation (X-linked agammaglobulinaemia) or through causing constitutive membrane association (breast, colorectal and ovarian cancers) (Lindvall, Blomberg et al. 2005; Carpten, Faber et al. 2007). Functional studies of the effects of the R262W variant are needed to determine how this impacts on the biology of coeliac disease.

2.1.8.6 *TAGAP* and *LPP*

T cell activation GTPase activating protein-*TAGAP* is a gene expressed in activated T cells, whose function in immune cells is not well characterized but may modulate cytoskeletal changes (Mao, Biery et al. 2004). *LPP* is strongly expressed in the small intestine but the significance in relation to coeliac disease is unknown. *LPP* has more recently been associated with rheumatoid arthritis and vitiligo (Coenen, Trynka et al. 2009; Jin, Birlea et al. 2010).

2.1.8.7 Other coeliac candidate genes

Further follow-up studies and a study of type 1 diabetes SNPs in coeliac disease have provided support for a further 5 disease regions, some at lesser levels of statistical significance. Candidate genes in these regions include *CTLA4, PTPN2, REL, TNFAIP3 and ITGA4.*

CTLA4 was originally postulated as a coeliac susceptibility gene in candidate gene studies. More convincing evidence has since been acquired from testing of type 1 diabetes-associated SNPs in the *CTLA4* gene region in 2560 UK individuals with coeliac disease and 9339 controls ($P = 1.26 \times 10^{-6}$) (Smyth, Plagnol et al. 2008). Cytotoxic T lymphocyte protein 4 has an essential role in Treg cell function in mice and may be implicated in loss of tolerance in autoimmunity (Wing, Onishi et al. 2008). Analysis of CTLA4 haplotypes conferring risk to type 1 diabetes has shown correlation with reduced expression of the soluble splice isoform of CTLA-4 (Ueda, Howson et al. 2003). *PTPN2* association with coeliac disease was also through testing of type 1 diabetes associated SNPs in UK coeliacs (Smyth, Plagnol et al. 2008). This gene encodes T-cell

protein tyrosine phosphatase, little studied but thought to be important in T cell activation (Ounissi-Benkalha and Polychronakos 2008)

TNFAIP3 (Tumour necrosis factor, alpha-induced protein 3) implicated in the study by Trynka et al. is a strong autoimmune disease candidate, required for termination of NF-κB signalling (Boone, Turer et al. 2004). Gene knockout in mice leads to multi-organ inflammation (Lee, Boone et al. 2000). The other candidate gene identified in this study was *REL*, a component of the NF-κB transcription complex (Trynka, Zhernakova et al. 2009).

A region on chromosome 2q31 showed association in an American follow-up study that successfully genotyped 975 of the SNPs genotyped by Hunt et al. in 928 American coeliac cases and compared them to Illumina genotype data on 3905 European controls (Illumina iControlDB). 5 of the eight regions identified in Hunt et al showed strong association in this study. In addition, rs6433894 showed the strongest association in the 2q31 region (P=0.00066; $P_{Hunt\_combined}=1.32 \times 10^{-5}$). The closest gene to this SNP is *UBE2E3*, an ubiquitin-conjugating enzyme involved in ubiquitinization a mechanism that targets abnormal or short-lived proteins for degradation. *ITGA4* maps ~325kb from the associated SNPs in this region. This gene encodes integrin alpha 4, an integrin subunit that contributes to the alpha4-beta7 integrin implicated in targeting T cells to the intestine.

2.2 Crohn's Disease

Crohn's disease is a chronic relapsing-remitting intestinal inflammatory disease with a median population prevalence of in 140 per 100,000 (range 10-199) in populations of European ancestry (Loftus 2004). Inflammation may affect any part of the gastrointestinal tract, but most commonly affects the ileum and proximal colon. Over time the natural history is of progression from an inflammatory phase, with chronic mucosal inflammation and ulceration, to stricturing and finally penetrating disease which can cause perforation, fistulation and abscesses (Cosnes, Cattan et al. 2002).

Ulcerative colitis is the other main type of inflammatory bowel disease (IBD), with similar prevalence, again characterized by relapsing-remitting intestinal inflammation. In contrast to Crohn's disease, ulcerative colitis manifests with chronic rectal mucosal inflammation and may extend proximally in the colon. For unknown reasons, inflammation in ulcerative colitis is confined to the mucosa, thus stricturing and penetration are not characteristic features.

2.2.1 Epidemiology

The highest incidence and prevalence rates of Crohn's disease have been reported from Northern Europe, the United Kingdom and North America, with a north-south gradient of disease incidence within Europe and North America reported (Loftus 2004; Baumgart and Carding 2007). However, incidence of Crohn's disease is reported to be increasing in most areas of the world, including Southern Europe, Asia, Africa and Latin America with some narrowing of the incidence gap between European ancestry and developing world populations (Loftus 2004). Summarising these trends, Loftus observed that IBD incidence appears to be low in developing countries, but increases with development and westernization. Initially this increase is dominated by ulcerative colitis, but eventually Crohn's disease incidence matches that of ulcerative colitis in developed nations (Loftus 2004). While this may in part reflect ascertainment biases, the increase in IBD incidence and the relative increase in Crohn's disease over the last few decades have been confirmed in well-studied population cohorts such as that from Olmsted County (North America).

Crohn's disease onset can occur at any age, with a median of around 30 years (Loftus, Silverstein et al. 1998). Previous studies have suggested a peak in the second and third

decades, and another in later life, though this has not been a consistent finding from recent epidemiological studies (Loftus 2004). The male: female ratio is close to one for Crohn's disease, with slight female preponderance particularly for disease diagnosed in late adolescence and early adulthood.

Smoking confers a definite but modest increased risk (OR ~2) of Crohn's disease, but the mechanism is uncertain (Calkins 1989; Garcia Rodriguez, Gonzalez-Perez et al. 2005). Smoking is a protective factor for ulcerative colitis (Harries, Baird et al. 1982; Calkins 1989). Appendicectomy is a protective factor for ulcerative colitis, and most studies have suggested it acts as a risk factor for Crohn's disease, though again mechanisms are unclear (Koutroubakis, Vlachonikolis et al. 2002; Radford-Smith 2008). The evidence for other environmental factors is relatively weak, but includes a suggestion of modestly increased risk of Crohn's disease associated with oral contraceptive use (Godet, May et al. 1995).

2.2.2 Treatment

The clinical management of individuals with Crohn's disease is complex and individualised, reflecting substantial inter-individual variation in clinical phenotype (**Table 2.2**) and disease severity and trade-offs between the therapeutic benefits and adverse effects of available interventions. The mainstays of management of Crohn's disease are medical therapies and surgery, neither of which are curative, but aim rather to limit the impact of the disease. At least 50% of individuals require surgery within 10 years, rising to around 80% over much longer time periods, mostly for stricturing ileal disease (Bernell, Lapidus et al. 2000; Carter, Lobo et al. 2004). Multi-disciplinary management may also include dietetics (interventions either aimed at optimizing nutritional status or nutritional immunotherapy for reducing disease activity), psychological interventions (e.g. stress management, interventions to promote smoking cessation), pharmacist-led monitoring of drug toxicities and immunization programs in those exposed to immunosuppressants. A full discussion of management is beyond the scope of this chapter. However, medical therapies are briefly discussed, as one of these options, azathioprine, is the subject of investigation in chapter 4.

Table 2.2 Montreal classification of Crohn's disease phenotype

Age at diagnosis (years)[a]	<16	A1
	17-40	A2
	>40	A3
Location	Ileal	L1
	Colonic	L2
	Ileocolonic	L3
	Upper gastrointestinal	L4[a]
Behaviour	Non-stricturing, non-penetrating	B1
	Stricturing	B2
	Penetrating	B3
	Perianal	P[b]

[a]Early onset disease associated with a more severe disease course
L4 can be added to L1-L3
[b]'p' is added to B1-B3 when concomitant perianal disease is present

Medical therapies (e.g. corticosteroids) are initially reserved for disease flares, with maintenance therapies (e.g. azathioprine, methotrexate) added in those with frequent relapses. However, there is increasing evidence that intervention with newer agents such as the anti-TNFα antibody infliximab as well as immunosuppressants may alter the progression of disease, with reduced rates of relapse and surgery (Markowitz, Grancher et al. 2000; Hanauer, Feagan et al. 2002; Lichtenstein, Yan et al. 2005; Vernier-Massouille, Balde et al. 2008). This has helped drive a trend towards earlier use of immunosuppressants and biologics, particularly in those with prognostic markers suggestive of a more aggressive disease course (D'Haens 2009; Dignass, J.F. Colombel et al. 2010). However, these benefits must be weighed in each case against the toxicities of these drugs, which include increased rates of serious infections and lymphoma. Better prognostic factors for disease behaviour and response to treatment would greatly aid Crohn's disease management. It is hoped that genetic risk variants in Crohn's disease will offer better risk stratification than clinical risk factors alone. *NOD2* variants, which confer the greatest known effects of any genetic or environmental factors on Crohn's disease risk (homozygote OR = 17), predict earlier onset of ileal structuring disease and need for surgery (Alvarez-Lobos, Arostegui et al. 2005; Annese, Lombardi et al. 2005). Risk profiling using multiple markers from recent genome-wide association studies, currently has weak prediction performance (chapter 1). Reports of associations between variants and selected disease outcomes should be interpreted with caution at this stage: multiple tests

require appropriate statistical significance threshold and validation in independent sample collections (Henckaerts, Van Steen et al. 2009; Weersma, Stokkers et al. 2009). So far no studies have applied the new genome wide association methods to investigate genetic effects on drug response and adverse effects for any of the agents commonly used in Crohn's disease. However, this area has considerable promise, with examples from other pharmacogenomic studies suggesting that genetic variants with large effects on risk may be identifiable (Link, Parish et al. 2008; Daly, Donaldson et al. 2009). In chapter 4, results from a genome wide association study of azathioprine/mercaptopurine induced pancreatitis are presented.

The thiopurine antimetabolites mercaptopurine and azathioprine are used as first line immunosuppressants, as steroid-sparing agents in active disease and for maintenance of remission. Azathioprine is a prodrug of mercaptopurine and both drugs have similar efficacy and toxicities. Both azathioprine and mercaptopurine have a slow onset of action, taking up to 16 weeks to achieve their full therapeutic effect. The numbers needed to treat (NNT) to induce and maintain remission with azathioprine in Crohn's disease are 5 and 3 respectively (Sandborn, Sutherland et al. 2000; Prefontaine, Sutherland et al. 2009). Discontinuation of thiopurine treatment in individuals in remission is associated with high rates of relapse even after 5 years or more of treatment (Lemann, Mary et al. 2005; Treton, Bouhnik et al. 2009). The benefits of thiopurines are considered to outweigh their toxicities in most patients, a Cochrane review of clinical trial data estimating a number needed to harm of 14 (compared to NNT of 5) for induction of remission in Crohn's disease (Sandborn, Sutherland et al. 2000). A 4-5fold increased risk of non-Hodgkin's lymphoma is observed in individuals with inflammatory bowel disease exposed to azathioprine, although absolute incidence is low (1 per 1000 patient-years) (Kandiel, Fraser et al. 2005; Beaugerie, Brousse et al. 2009). 10-20% of individuals discontinue thiopurines due to adverse effects, and careful monitoring is required for ~5% individuals experiencing severe adverse effects (e.g. neutropaenia, pancreatitis). The mechanisms and short-term toxicities of thiopurines are discussed more fully in chapter 4.

2.2.3 **Crohn's disease aetiopathogenesis: the intestinal microbiota**

The intestinal microbiota appears to be a necessary factor in the pathogenesis of Crohn's disease. Evidence for this includes the fact that inflammation occurs most commonly in regions of the gastrointestinal tract with the highest concentrations of bacteria (ileum and colorectum). Secondly, diversion of the faecal stream improves inflammation in the diverted

segment. Thirdly, antibiotics have clinical benefit in reducing inflammation in some patients with Crohn's disease, although benefits are usually not sustained. Finally several animal model s of intestinal inflammation require the presence of commensal intestinal bacteria. For example, the *IL10* knockout mouse raised in a germ-free environment only develops colitis on re-colonisation with commensal bacteria (Sellon, Tonkonogy et al. 1998). Thus intestinal microbes appear necessary for gut inflammation in Crohn's disease. A different question is whether variation in the intestinal microbiota between individuals accounts for a proportion of Crohn's disease susceptibility.

2.2.4 **Evidence for an abnormal microbiota in Crohn's disease**

Many studies have attempted to compare the microbiota in patients with Crohn's disease to normal healthy controls. A prevailing obstacle is the fact that most bacterial species residing in the human intestine cannot be cultured *in vitro* ((Suau, Bonnet et al. 1999; Tannock 2000)). Thus, culture-based studies are biased towards the around 30% of organisms that may be cultured. For example, there have been reports of an increased prevalence of an adherent-invasive strain of *Escherichia Coli* in Crohn's disease ileal mucosa that is potentially pathogenic and invasive in *in vitro* studies (Darfeuille-Michaud, Neut et al. 1998). However, the true prevalence of these strains in healthy individuals may in fact be closer to that seen in Crohn's disease and appears to be similar in colon cancer patients (Martin, Campbell et al. 2004). Another body of research has explored the hypothesis that *Mycobacterium Avium* subspecies *Paratuberculosis* (MAP) may be a cause of Crohn's disease. MAP causes a chronic granulomatous inflammatory disease of the intestine in cattle (Johne's disease) and other mammals and has been identified in the intestine and mesenteric lymph nodes of some Crohn's patients (Feller, Huwiler et al. 2007; Frank 2008). A subset of individuals with Crohn's also shows T cell-reactivity to MAP (Olsen, Tollefsen et al. 2009). However, the association with Crohn's is contentious and even if proven it is unclear whether the presence of MAP is a consequence of Crohn's disease (impaired immunity) or a cause. A randomized trial of anti-MAP chemotherapy did not show evidence of sustained benefit (Selby, Pavli et al. 2007). Thus there remains no convincing evidence of the role of this bacterium and indeed any other single strain in disease pathogenesis (Strober, Fuss et al. 2007).

Metagenomic approaches aim to capture microbial diversity more fully through studying microbial genetic diversity. One approach involves sequencing or hybridization of microbial

ribosomal RNA from cloned DNA fragments derived from faeces or mucosal tissue. This enables a phylogenetic classification of the human microbiome. In Crohn's disease there is evidence of a reduced diversity and numbers among the phyla *Bacteriodetes* and *Firmicutes* (Manichanh, Rigottier-Gois et al. 2006; Frank, St Amand et al. 2007). However, whether these differences in microbiota precede the development of Crohn's disease and have a causal role is unknown.

The intestinal mucosal immune system in health exists in a state of fine balance, exposed continually to high concentrations of intestinal bacteria. Sensory components of the innate immune system, including epithelial toll like receptors and cytoplasmic NOD-like receptors continuously sample microbial antigens. The factors that tip the balance of the intestinal immune response towards inflammation are many and various, but genetics has been helpful in highlighting some of the pathways which may be dysregulated.

2.2.5 Defective innate immune responses in Crohn's disease

In this hypothesis defective innate immune responses to commensal bacteria disturb the normal homeostasis of the mucosal immune system (Marks, Miyagi et al. 2009). This might lead to defective clearance of normally sub-pathogenic bacteria and ensuing secondary immune responses (Smith, Rahman et al. 2009). Functional studies have demonstrated impairment of neutrophil chemotaxis and acute inflammation both in the intestine and at extra-intestinal sites, providing some support for this hypothesis (Marks, Harbord et al. 2006). Moreover, the granulomatous inflammation of Crohn's disease has similarities to the inflammation observed in the primary neutrophil immunodeficiency, chronic granulomatous disease (Marks, Miyagi et al. 2009). Again, the question of whether these abnormalities are primary disturbances contributing to Crohn's pathogenesis, or acquired abnormalities caused by the development of Crohn's disease is unanswered.

2.2.6 Evidence of for genetic susceptibility in inflammatory bowel disease

Twin studies have reported concordance of 60-70% in monozygotic twins and ~10% in dizygotic twins (Orholm, Binder et al. 2000; Halfvarson, Bodin et al. 2003). Familial clustering is also evident in Crohn's disease (Peeters, Nevens et al. 1996) with sibling relative risk (λ_s) estimated at between 25 and 35 (Satsangi, Parkes et al. 1998; Lewis, Whitwell et al. 2007).

Ulcerative colitis (UC) is less heritable than Crohn's disease (λ_s=10-15 UC) (Satsangi, Parkes et al. 1998). Only a few extremely rare Mendelian forms of inflammatory bowel disease (autosomal recessive inheritance) have ever been reported but the phenotype appears distinct from both Crohn's disease and ulcerative colitis, with onset of severe disease in the first few months of life (Fried and Vure 1974; Megarbane and Sayad 2007). Mutations in genes encoding the IL-10 receptor were identified as the cause of early-onset entero-colitis in two families (Glocker, Kotlarz et al. 2009). Heritability in Crohn's disease has been estimated as 50% with 20% of this accounted for by known variants (Tysk, Lindberg et al. 1988; Barrett, Hansoul et al. 2008). Ulcerative colitis, a related inflammatory bowel disease is more frequent in relatives of Crohn's disease, suggesting that some predisposing variants are shared in both diseases (Satsangi, Grootscholten et al. 1996).

2.2.7 Susceptibility variants in Crohn's disease

The first susceptibility gene identified in Crohn's disease was *NOD2*, discovered through linkage analysis and resequencing of the *NOD2* gene in 2001 (Hugot, Chamaillard et al. 2001; Ogura, Bonen et al. 2001). The 3 major disease-causing mutants were identified in these studies after resequencing the *NOD2* gene in only 62 individuals, with limited follow-up genotyping in cases and controls confirming association. These variants include two amino-acid substituting variants (R702W, G908R) and one frame-shift mutation (1007fs, which causes a truncated peptide) and account for 80 % of disease-causing variants (Aslan, Karaveli et al. 2007). Allele frequencies of these variants vary between 1.2 and 4.3 % in healthy controls with a meta-analysis showing that these variants confer odds ratios (ORs) for heterozygotes of 2.4 (2.0-2.9) and homozygote/compound heterozygote 17.1 (10.7-27.2) (Economou, Trikalinos et al. 2004). A notable feature of one of the original *NOD2* reports was the enrichment of many additional rare variants in Crohn's cases versus controls (Hugot, Chamaillard et al. 2001). Thus *NOD2* provides some support for both the common disease-common variant and common disease-rare variant hypotheses.

2.2.8 Genome wide association studies in Crohn's disease

Eight single nucleotide polymorphism (SNP) GWASs have been performed since the discovery of *NOD2* (**Table 2.3**). These have collectively generated strong evidence for disease association for over 30 genomic regions (Barrett, Hansoul et al. 2008). A further 18 susceptibility loci have

been reported from an extension of this meta-analysis by the International IBD Genetics Consortium, that analysed GWAS data from a total of 6,324 Crohn's disease cases and 15,054 controls of European ancestry (Parkes 2010). In all cases the associated variants confer modest effects on disease risk with odds ratios less than those of the causal variants in *NOD2,* typically allelic odds ratios < 1.5 (**Table 2.4**). In rare instances these studies have identified genetic variants that are likely to be causal (*IL23R, ATG16L1*). Inferring causality is supported by SNPs with obvious predicted functional effects (e.g. amino-acid sequence changing SNPs), and where association of surrounding SNPs can be entirely accounted for by the candidate SNP. However confirmation of causality relies also on functional studies demonstrating that these SNPs may directly alter biological processes in ways relevant to the disease.

In most instances, regions of association encompass 10s or 100s of kilobases of sequence (median size of linkage disequilibrium blocks around the top regional Crohn's disease-associated SNP was 165 kilobases) and the causal variants are unknown (Barrett, Hansoul et al. 2008). These regions typically contain hundreds of known common variants across a block of high linkage disequilibrium (LD). The common variants are therefore highly correlated and determining which variants are responsible for the disease association has usually not been possible from analyses of GWAS data. Early fine-mapping studies of some of these regions performed by the WTCCC (data not yet published), in which much larger numbers of variants (including some rare variants discovered through regional resequencing) are genotyped in cases and controls, have achieved some successes either in narrowing the region of association or identifying causal variants, but most frequently this approach has been insufficient to identify causal variants.

Some of the GWAS-identified Crohn's loci contain many genes while others contain no known genes at all (gene deserts). At one extreme the 3p21 locus associated with Crohn's disease, contains 35 genes while the 5p13.1 Crohn's locus has no genes (Barrett, Hansoul et al. 2008). Thus, a primary challenge in understanding the new GWAS associations is to define the causal variants within these regions and a second is to move towards a functional understanding of the impact of these variants on disease pathogenesis. These functional analyses are mostly in their infancy, but successes have been reported, particularly for genes where the causal variants have been identified (e.g. *NOD2, ATG16L1*).

Table 2.3 Genome Wide Association Studies in Crohn's Disease published before June 2010

Year of publication	Reference: first author	Sample population	Genotyping platform (no. SNPs post QC)	GWAS Sample size: cases vs. controls (post QC)	Novel risk loci[a]	Top SNP GWAS *P* value for each gene
2005	Yamazaki	Japanese	Multiplex PCR based (72,738)	94 vs. 752	*TNFSF15*	1.71×10^{-14}
2006	Duerr	North American (NIDDK)	Illumina Hap300 (308,332 SNPs)	547 vs. 548	*IL23R*	5.1×10^{-9}
2007	Hampe	German	SNPlex Genotyping System (19,779 nsSNPs)	735 vs. 368	*ATG16L1*	4.0×10^{-8}
2007	Rioux	North American (NIDDK)	Illumina Hap300 (304,413)	946 vs. 977	*ATG16L1*	6.4×10^{-8}
2007	Libioulle	French/Belgian	Illumina Hap300 (302,451)	547 vs. 928	*5p13.1 (PTGER4)*	4.1×10^{-8}
2007	WTCCC	United Kingdom	Affymetrix 500K Genechip (469,557 SNPs)	1748 vs. 2938	*10q24 (NKX2-3)* *PTPN2* *IRGM* *10q21 (?ZNF365)*	1.4×10^{-8} 4.6×10^{-8} 5.1×10^{-8} 2.7×10^{-7}
2007	Franke	German	Affymetrix 100k Genechip (92,387 SNPs)	393 vs. 399	*Nil*	
2007	Raelson	French-Canadian Founder population	Perlgen Platform (164,279 SNPs)	382 Parent-offspring trios	*Nil*	
2010	McGovern	United States	Illumina 610Quad and Illumina 370Duo (304,825 SNPs post-QC)	896 vs. 3204	*FUT2* *CCR6*[b] *IL12B*[b]	2×10^{-8} 6×10^{-8} 7×10^{-8}

Studies listed in order of publication date. SNP number and sample size indicate post-Quality Control numbers. Studies with ~100,000 or more genome-wide SNPs considered as GWASs: note Hampe et al. study does not meet this criterion. WTCCC-Wellcome Trust Case Control Consortium. Association *P* value from GWAS, not including replication.

[a]Loci with SNPs showing association with CrD in the GWAS phase at $P < 5 \times 10^{-7}$. Does not include loci identified in follow-up genotyping and combined analyses. Gene is the best candidate at each locus showing association. *NOD2* and 'IBD5' loci were previously identified by linkage studies and were strongly replicated in these studies.

[b]*CCR6* and *IL12B* previously identified in follow-up study and GWAS meta-analyses.
Duerr et al. GWAS samples were ileal CrD only. Samples were from the North American National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC)

Hampe et al. included a genome-wide set of non-synonymous SNPs, but genomic coverage does not constitute that usually accepted for a genome-wide association study

Rioux et al. GWAS samples from the NIDDK-IBDGC, includes samples analysed in Duerr et al.

Franke – cases selected for severe CrD phenotype (severe disease, onset < age 25, positive family history)

Raelson – analysed both single SNPs and haplotypes of 3-9 adjacent SNPs.

In general we may anticipate that both common and rare causal variants exist, but in the cases of common variants we should expect ORs will be lower and perhaps also that perturbations in biological function would be more subtle (Bodmer and Bonilla 2008). Genetic variants identified in Crohn's GWASs, that are causal, rather than just correlated with unknown causal variants, include the common allele (population frequency 0.93) of the *IL23R* missense SNP, rs11209026 (R381Q) and a missense SNP (rs2241880, T300A) in the autophagy gene *ATG16L1* (Duerr, Taylor et al. 2006; Hampe, Franke et al. 2007). The *ATG16L1* risk allele (frequency ~0.45 in Caucasian populations) appears to account for all disease association in the region. Hampe et al. resequenced ATG16L1 exons, promoters and splice sites in 47 Crohn's individuals and found no additional coding or splice site variants. They also performed conditional regression and haplotype analyses suggesting this variant accounted for the whole GWAS association signal (Cuthbert, Fisher et al. 2002). A meta-analysis has estimated heterozygote and homozygote ORs for Crohn's disease in Caucasian populations of 1.39 and 1.87 respectively, confirming a more modest effect size than for *NOD2* (Zhang, Qiu et al. 2009). A genomic region associated with Crohn's disease that contains the autophagy gene *IRGM,* contains a 20 kilobase deletion polymorphism upstream of *IRGM,* in perfect linkage disequilibrium with the most strongly associated GWAS SNP. This deletion polymorphism correlates with *IRGM* expression (McCarroll, Huett et al. 2008) and has been postulated as a causal variant. In other GWAS-discovered Crohn's disease regions, causal variants have not yet been convincingly demonstrated.

A major finding from the first wave of genome wide association studies in Crohn's disease has been the identification of independent loci implicating multiple genes within the same biological pathways. The two outstanding pathways to have emerged are Th17 cell/IL-23 receptor signalling and autophagy. Genes linked to Crohn's disease that participate in Th17 cell signalling include *IL23R, IL12B, CCR6, STAT3* and *JAK2* (Barrett, Hansoul et al. 2008). Crohn's-associated regions containing genes linked to autophagy include *ATG16L1, IRGM* and *ATG5* (Barrett, Hansoul et al. 2008). The significance of these pathways to Crohn's pathogenesis is discussed below.

Table 2.4 Meta-analysis *P* values, risk allele frequencies and odds ratios for most strongly associated SNPs at loci reported in individual GWASs

Locus	P value	Risk Allele Frequency	Odds ratio (case-control)
IL23R	6.66 x 10-63	0.933	2.50
ATG16L1	2.36 x 10-32	0.533	1.28
PTGER4	6.82 x 10-27	0.125	1.32
10q21 (?ZNF365)	4.46 x 10-20	0.387	1.25
5q31	2.32 x 10-18	0.425	1.25
PTPN2	5.10 x 10-17	0.152	1.35
NKX2-3	3.06 x 10-16	0.478	1.20
IRGM	3.40 x 10-16	0.090	1.33
MST1	1.15 x 10-12	0.271	1.20
TNFSF15	2.60 x 10-10	0.677	1.22

Adapted from Barrett et al. 2008(Barrett, Hansoul et al. 2008). Meta-analysis combined data from Duerr/Rioux, WTCCC and Libioulle GWASs. P values are for top SNP from each region calculated from GWAS meta-analysis and replication in 3664 cases and a mixture of population and family based controls. NOD2 not shown (not assayed in all GWASs). A further 19 loci were identified in a combined follow-up genotyping and GWAS meta-analysis

A further important feature is the recognition that current SNP associations account for only 20% of disease risk due to genetic variation (Barrett, Hansoul et al. 2008). Furthermore most of this risk remains dominated by *NOD2* and *IL23R* variants. Thus the remaining causal genetic variation in Crohn's, as in other complex diseases, lies elsewhere in variants not well-captured by SNPs tested in GWASs (see chapter 6).

2.2.9 Function of Crohn's disease genetic variants

2.2.9.1 *NOD2*

NOD2 is expressed strongly in Paneth cells (located at the base of intestinal crypts in the small intestine), as well as intestinal epithelial cells, macrophages and dendritic cells (Cho and Abraham 2007). It functions as a bacterial pattern recognition receptor that can bind muramyl dipeptide (MDP) a component of bacterial peptidoglycan. Disease-associated *NOD2* mutations in humans lead to diminished cytokine responses to MDP (van Heel, Hunt et al. 2005) and are associated with reduced alpha-defensin production from Paneth cells (Wehkamp, Harder et al. 2004). *NOD2* is required for MDP-induced autophagy in dendritic cells. Dendritic cells from individuals with Crohn's disease with *NOD2* mutations (1007fsinsC, R702W, G908R) or the *ATG16L1* T300A variant showed defective autophagy induced by MDP and defective antigen presentation to CD4+ T cells (Cooney, Baker et al. 2010). This study links *NOD2* with autophagy which has emerged as a key mechanism for bacterial handling and immune clearance in the intestine in Crohn's disease.

2.2.9.2 Autophagy genes: *ATG16L1, IRGM*

Genome wide-association studies in Crohn's disease have implicated autophagy proteins (*IRGM and ATG16L1).* Autophagy is an evolutionarily conserved process, originally described as having a role in the clearance of degraded organelles and long-lived proteins. The process of autophagy involves the cytoplasmic formation of a double-membrane bound vacuole, the autophagosome, which then fuses with lysosomes leading to the degradation of its contents (Xie and Klionsky 2007). Autophagy also exists as a mechanism for clearance of microbes, and has immune functions including defence against intracellular bacteria such as *salmonella typhi* (Birmingham, Smith et al. 2006; Singh, Davis et al. 2006) and a role in delivery of microbial peptides to MHC class II loading compartments (Schmid, Pypaert et al. 2007). The Crohn's disease gene *IRGM* has specifically been implicated in induction of autophagy and is required for autophagic clearance of *mycobacteria (Singh, Davis et al. 2006).* Human *IRGM* variants correlated with those associated with Crohn's disease have also been associated with susceptibility to pulmonary tuberculosis, suggesting that these variants may both cause defective innate immunity to tuberculosis and promote Crohn's disease (Intemann, Thye et al. 2009). However, autophagy has extremely diverse immune functions, including functions in adaptive immunity (Deretic 2010). It plays a key role in positive and negative selection during CD4 T cell development in the thymus, most likely through endogenous antigen presentation by thymic stromal epithelial cells and has been proposed to be critical in tolerance induction (Nedjic, Aichinger et al. 2008).

The ATG16L1 protein is a core autophagic protein and the Crohn's disease genetic variants have been linked to impaired autophagic internalisation of *Salmonella typhi (Kuballa, Huett et al. 2008)*. More recently, the Crohn's disease *ATG16L1* variant was linked to defective granule exocytosis in Paneth cells (Cadwell, Liu et al. 2008; Cadwell, Patel et al. 2009). This group were able to differentiate Paneth cells from patients with the *ATG16L1* variant from those without it on the basis of granule abnormalities visible in standard haematoxylin and eosin stained ileal sections (Cadwell, Liu et al. 2008). More recently, Cadwell et al. demonstrated that Paneth cell abnormalities in mice expressing hypomorphic ATG16L1, only emerged in the presence of infection with a specific strain of murine norovirus (Cadwell, Patel et al. 2010).

2.2.9.3 *IL23R*

IL23R variants were first linked to Crohn's disease by a north American GWAS but replicated strongly in subsequent GWASs with the strongest association of all genomic regions in the Barrett GWAS meta-analysis (though *NOD2* disease-causing variants were not directly tested)(Rioux, Xavier et al. 2007; Barrett, Hansoul et al. 2008). The most strongly associated SNP in the original study was rs11209026, a missense SNP (Arg381Gln) where the risk variant had a frequency of 0.93 in controls and 0.98 in cases. This is likely to be one of the causal variants but more than one association signal exists in the region.

IL23R is a compelling functional candidate for Crohn's disease, expressed in memory, but not naive T lymphocyte and NK cells. IL-23 receptor signalling promotes Th17 effector subtype differentiation, of relevance as this T cell subset plays a critical role in causing inflammation in animal models of intestinal inflammation and other autoimmune disease(Hue, Ahern et al. 2006; Uhlig, Coombes et al. 2006; Annunziato, Cosmi et al. 2007) . So far, there has been minimal published progress linking Crohn's *IL23R* variants with IL23-receptor function. A small study investigating serum levels of the Th17 cell secreted cytokine IL-22, found higher levels in Crohn's patients carrying *IL23R* risk alleles, suggesting that these variants may be cause gain of function with respect to IL23R signalling (Schmechel, Konrad et al. 2008). However, studies in individuals without the potential confounder of ongoing inflammation, specifically testing individual cell subset cytokine production are required to confirm these findings.

2.3 Ulcerative colitis susceptibility variants and overlap with Crohn's disease

Prior to genome wide association studies in inflammatory bowel disease, no susceptibility variants for ulcerative colitis had been identified. However, in the last year 4 GWASs have been completed (Silverberg, Cho et al. 2009; Franke, Balschun et al. 2010; McGovern, Gardet et al. 2010). A meta-analysis of 3 of these studies, comprising 2693 cases and 9791 controls of European descent identified around 30 susceptibility loci, with common risk variants of modest effect (McGovern, Gardet et al. 2010). The authors also tested SNPs from 30 Crohn's loci and together found that around half of Crohn's susceptibility loci are shared in ulcerative colitis. Notably Th17-IL23 receptor pathway genes (*IL23R, JAK2, IL12B, STAT3, IL17REL*) confer susceptibility in ulcerative colitis but autophagy genes (*ATG16L1, IRGM,* ATG5) and *NOD2* do not and appear therefore to be specific to Crohn's disease.

Chapter 3 Genome wide association study in coeliac disease

3.1 Introduction

Prior to the first coeliac disease genome wide association study (GWAS) in 2007, only specific alleles of the *HLA-DQA1* and *HLA-DQB1* genes had been convincingly identified as influencing coeliac disease susceptibility (van Heel, Hunt et al. 2005; Dubois and van Heel 2008). Positive reports of non-HLA genomic regions discovered through linkage and candidate gene association studies proved inconsistent in follow-up studies (van Heel, Hunt et al. 2005; Dubois and van Heel 2008). The first GWAS in coeliac disease and 3 follow-up studies that tested the most strongly associated SNPs in further independent sample collections identified a total of 11 new non-HLA coeliac risk regions (van Heel, Franke et al. 2007; Hunt, Zhernakova et al. 2008; Garner, Murray et al. 2009; Trynka, Zhernakova et al. 2009). Strong support for a further two risk loci was reported in a large association analysis of type 1 diabetes risk variants in coeliac disease (Smyth, Plagnol et al. 2008). Early replication studies of these new variants support the robustness of the coeliac disease associations in further independent European populations (Adamovic, Amundsen et al. 2008; Dema, Martinez et al. 2009; Koskinen, Einarsdottir et al. 2009; Romanos, Barisani et al. 2009; Amundsen, Rundberg et al. 2010). Together these studies have identified SNP variants at 14 loci (including the HLA) with convincing or strongly suggestive association to coeliac disease in populations of European ancestry (**Table 3.1**).

These findings have provided strong validation of the GWAS method for identifying new susceptibility variants in coeliac disease. In line with other early GWAS results in common diseases, newly identified risk variants conferred modest effects on disease risk, with odds ratios between 1.19 and 1.41 (**Table 3.1**). Other features of the GWAS-derived findings include a bias towards higher minor allele frequencies among these newly identified variants, again consistent with findings from GWASs in other common diseases(McCarthy, Abecasis et al. 2008). This skew towards higher minor allele frequency SNPs could reflect a genuine enrichment for very common variants in the underlying coeliac genetic architecture, but is more probably explained by reduced power to detect SNPs with low minor allele frequencies (**Figure 3.1**). Finally, new variants accounted for a small minority of non-HLA related heritability. The eight non-HLA variants identified from the first GWAS and a large follow-up

study were estimated to add only 5% to the 35% of heritability accounted for by known HLA risk variants (Hunt, Zhernakova et al. 2008).

A larger GWAS was therefore designed to increase statistical power to discover additional common genetic variation contributing to the around 60% of unexplained heritability in coeliac disease. The earlier GWAS findings, including those from other common diseases hinted at a genetic architecture in which the contribution of common variant risk was distributed across 10s or hundreds of loci of weaker effect (Cooper, Nickerson et al. 2007). The previous GWAS (778 cases and 1422 controls) had > 99% power to detect common variants (MAF>0.05) of large effect (OR > 3) but almost zero power to detect variants with the magnitude of effect sizes seen in the follow-up studies (ORs 1.19-1.41), using a Wellcome Trust Case Control Consortium (WTCCC)-advocated genome-wide-significance threshold of $P<5 \times 10^{-7}$(Korbel, Urban et al. 2007). A large increase in power was therefore desirable for a new GWAS to facilitate the discovery of further common variants of modest effect.

Table 3.1 **Non human leucocyte antigen (HLA) susceptibility loci for coeliac disease identified in the first coeliac GWAS and follow-up studies (van Heel, Franke et al. 2007; Hunt, Zhernakova et al. 2008; Smyth, Plagnol et al. 2008; Garner, Murray et al. 2009; Trynka, Zhernakova et al. 2009)**

Chromosome locus	SNP with strongest coeliac disease association[a]	Minor allele frequency (HapMap CEU)	Odds Ratio for risk allele	Genes of interest	Publication of coeliac disease association	Other diseases with association to same region
1q31	rs2816316	0.217	1.41	*RGS1*	Hunt et al. 2008	Type 1 diabetes (Smyth, Plagnol et al. 2008) Multiple sclerosis (De Jager, Jia et al. 2009)
2p16	rs842647	0.300	1.19	*REL*	Trynka et al. 2009	Rheumatoid arthritis, Ulcerative colitis (Gregersen, Amos et al. 2009; McGovern, Gardet et al. 2010).f
2q11-2q12	rs917997	0.233	1.27	*IL1RL1, IL18R1, IL18RAP, SLC9A4*	Hunt et al. 2008	Crohn's disease (Barrett, Hansoul et al. 2008)
2q31	rs6433894	0.403	1.16	*ITGA4, UBE2E3*	Garner et al. 2009	Ankylosing spondylitis (Reveille, Sims et al. 2010)
2q33	rs3087243	0.458	1.18	*CTLA4*	Smyth et al. 2008	Type 1 diabetes(Smyth, Plagnol et al. 2008), Rheumatoid arthritis (Gregersen, Amos et al. 2009), Primary biliary cirrhosis (Hirschfield, Liu et al. 2009)
3p21	rs6441961	0.317	1.21	*CCR1,CCR2,CCRL2, CCR3, CCR5, XCR1*	Hunt et al. 2008	Type 1 diabetes (Barrett, Clayton et al. 2009), Rheumatoid arthritis (Gregersen, Amos et al. 2009)

Table 3.1 (cont.)

3q25-3q26	rs17810546	0.100	1.34	*IL12A, SCHIP1*	Hunt et al. 2008	Primary biliary cirrhosis (Hirschfield, Liu et al. 2009) Multiple sclerosis (De Jager, Jia et al. 2009)
3q28	rs1465150	0.082	1.21	*LPP*	Hunt et al. 2008	
4q27	rs6822844	0.203	1.41	*IL2, IL21*	van Heel et al. 2007	Type 1 diabetes (Barrett, Clayton et al. 2009), Rheumatoid arthritis (Weersma, Zhernakova et al. 2007), Graves' disease (Todd, Walker et al. 2007), Psoriasis (Liu, Helms et al. 2008)
6q23	rs2327832	0.175	1.25	*TNFAIP3*	Trynka et al. 2009	Multiple sclerosis, Psoriasis, Rheumatoid arthritis, SLE, Ulcerative colitis (Plenge, Cotsapas et al. 2007; Graham, Cotsapas et al. 2008; De Jager, Jia et al. 2009; Nair, Duffin et al. 2009; Wang, Baldassano et al. 2010).
6q25	rs1738074	0.492	1.21	*TAGAP*	Hunt et al. 2008	Type 1 diabetes (Smyth, Plagnol et al. 2008)
12q24	rs653178	0.408	1.19	*SH2B3*	Hunt et al. 2008	Type 1 diabetes (Smyth, Plagnol et al. 2008), Hypertension (Newton-Cheh, Johnson et al. 2009), haematocrit (Ganesh, Zakai et al. 2009)
18p11	rs45450798	0.153	1.18	*PTPN2*	Smyth et al. 2008	Type 1 diabetes (Barrett, Clayton et al. 2009), Crohn's disease (Parkes, Barrett et al. 2007)

[a]Most strongly associated SNP reported.

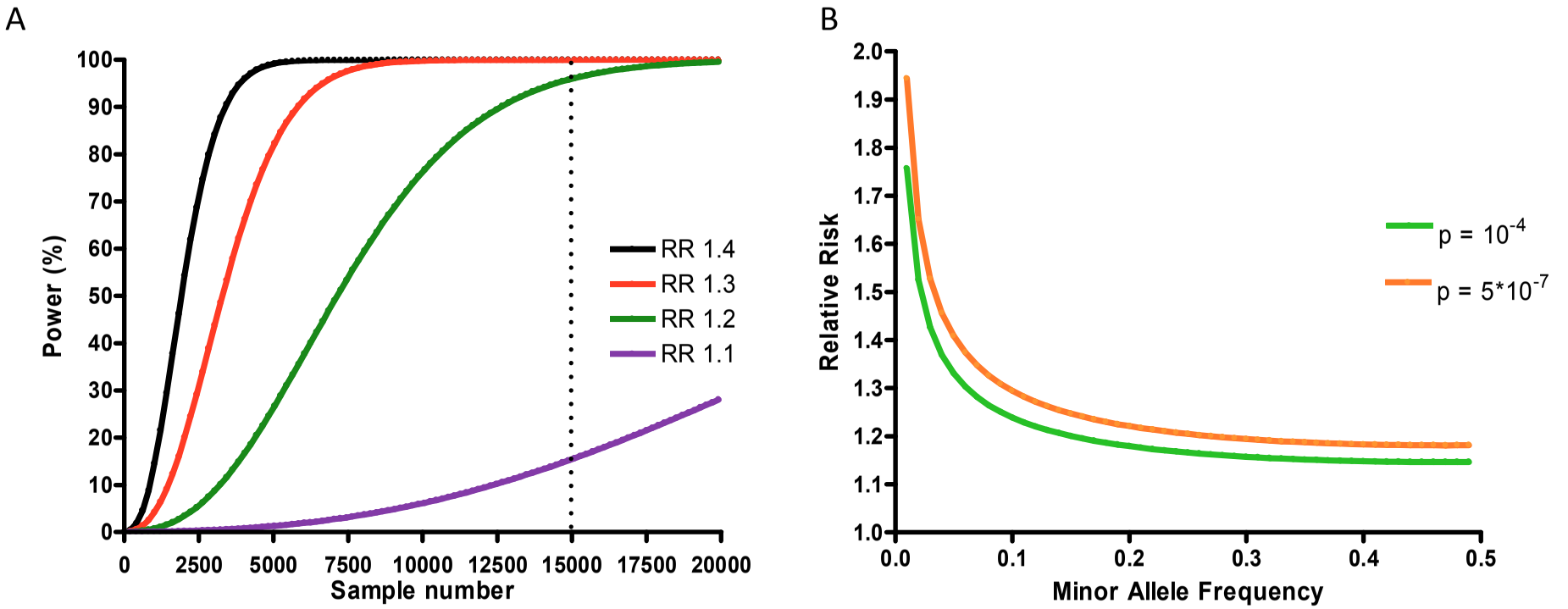
3.2 Power considerations

Increasing both sample size and genome-wide marker content enables an increase in power to detect common genetic variants(Balding 2006). At the time of study design the largest GWAS in common diseases (around 2000 cases and 3000 controls for each of 7 diseases) provided some validation of this approach, with the discovery of multiple common variants of weak effect ($P_{GWAS}<5 \times 10^{-7}$, median odds ratio = 1.34) (Wellcome Trust Case Control Consortium 2007). During the current study, other large GWASs in chronic immune-mediated diseases reported tens of new loci in single diseases (Barrett, Hansoul et al. 2008; Barrett, Clayton et al. 2009). In considering how to maximize power in a second generation GWAS, statistical modelling suggested that increasing sample size would be of relatively greater importance than increasing marker density beyond the approximately 300,000 SNPs used in the first GWAS (Burton, Clayton et al. 2007; Spencer, Su et al. 2009).

In the study reported here, a sevenfold increase in overall sample size (n= 15,283) compared to the first GWAS in coeliac disease (n=2,200) was achieved. To assay additional common genetic variation not captured by the Illumina Hap300 platform used in the first GWAS, new samples were genotyped to a minimum Illumina Hap550 density (562,495 SNPs, capturing ~88% of known HapMap CEU SNPs with MAF>0.05 at $r^2 > 0.8$). To assay common copy number polymorphisms an additional 97,952 monomorphic CNV probes were included, designed to capture 5,000 common heritable CNVs on the Illumina Human 670Quad-Custom chip. CNV probes were designed by Illumina in collaboration with Matthew Hurles at the Wellcome Trust Sanger Institute using deCode, Database of Genomic Variants, and Wellcome Trust Sanger Institute data.

Figure 3.1 illustrates power considerations for the study and highlights the limits of power in the current study with respect to SNP minor allele frequency and variant relative risk. It was apparent that there was rapid loss of power for SNPs with minor allele frequencies less than approximately 0.05 and relative risk less than 1.2 in a study of this size.

Figure 3.1 **Power to detect SNPs associated with coeliac disease.** A. The effect of sample size on power for variants conferring relative risks between 1.4 and 1.1[a]. B. The effect of SNP minor allele frequency on detectable relative risk of coeliac disease-associated variants[b]



a- Assumptions- co-dominant model, disease prevalence 1%, MAF 0.05, genotyped SNP to causal SNP r^2 0.88, control: case ratio 2:1, $P_{GWAS} = 10^{-4}$. RR = Relative risk. Dashed line indicates proposed sample size for the GWAS

b- Assumptions- co-dominant model, disease prevalence 1%, power 90%, no. cases 4533, no. controls 10,750, genotyped SNP to causal SNP r^2 0.88. Green line $P_{GWAS} = 10^{-4}$, Orange line $P_{GWAS} = 5 \times 10^{-7}$

3.3 Study Design

The study was performed in two stages (**Table 3.2**). In stage 1, GWAS, after quality controls 4,533 coeliac disease cases and 10,750 controls from four populations of European ancestry (2 collections from the UK, 1 each from the Netherlands, Italy and Finland) were included for analysis. Case-control association analysis was performed independently for 295,453 SNPs passing quality controls in all 5 sample collections, and a further set of 231,516 SNPs genotyped only in UK2, Dutch, Italian and Finnish collections (3,796 cases and 8,154 controls).

In stage 2, follow-up, after quality controls, 3,796 coeliac disease cases and 8,154 controls from 7 populations of European ancestry were analysed for 131 SNPs from 94 genomic loci, selected from the stage 1 analysis (**Table 3.2**).

3.3.1 Stage 1 – GWAS genotyping and SNP-calling

Cases from the first coeliac GWAS (coeliac GWAS1) were integrated in the stage 1 analysis, contributing to the UK1 collection (**Table 3.2**). The 1958 birth cohort individuals that had been used as controls in coeliac GWAS1, had since been genotyped at a higher marker density (1,252,158 markers) on the Illumina 1.2M-DuoCustom platform by the Wellcome Trust Case Control Consortium (WTCCC). These samples (labelled 1958bc-WTCCC) were selected instead as controls for the second UK coeliac collection (UK2) to make use of the fact that they had been genotyped for all markers genotyped in UK2 cases (660,447 markers present on the Illumina 670-Quadcustom beadchip). UK1 cases were instead paired with 2,596 additional population controls from the 1958 birth cohort (labelled 1958bc-T1DGC) who had been genotyped on the Illumina Human Hap550 platform, originally for the type 1 diabetes genetics consortium (T1DGC)(Barrett, Clayton et al. 2009).

All new case samples (UK2, Dutch, Italian, and Finnish) included in the GWAS were genotyped on the Illumina 670-Quadcustom platform. Dutch and Italian controls were also genotyped using the Illumina 670-Quadcustom platform. However, Finnish controls had been genotyped for another study, using the Illumina 610-Quad platform. This platform has nearly identical SNP content to the Quad670 platform, but has reduced CNV marker content.

Table 3.2 Sample collections and genotyping platforms

Collection	Country	Coeliac disease cases			Controls		
		Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]
Stage 1: Genome wide association							
1[e f]	UK	767	737	Illumina Hap300v1-1	2,596[j]	2,596	Illumina Hap550-2v3
2[eg]	UK	1,922	1,849	Illumina 670-QuadCustom_v1	5,069[j]	4,936	Illumina 1.2M-DuoCustom_v1
3[e]	Finland	674	647	Illumina 670-QuadCustom_v1	1,839[j]	1,829	Illumina 610-Quad
4[h]	Netherlands	876	803	Illumina 670-QuadCustom_v1	960	846	Illumina 670-QuadCustom_v1
5[e]	Italy	541	497	Illumina 670-QuadCustom_v1	580	543	Illumina 670-QuadCustom_v1
Analysis of Hap300 markers			**4,533**			**10,750**	
Analysis of additional Hap550 markers			**3,796**			**8,154**	
Stage 2: Follow-up							
6	USA	987	973	Illumina GoldenGate	615	555	Illumina GoldenGate
7	Hungary	979	965	Illumina GoldenGate	1,126	1,067	Illumina GoldenGate
8[i]	Ireland	653	597	Illumina GoldenGate	1,499	1,456	Illumina GoldenGate
9	Poland	599	564	Illumina GoldenGate	745	716	Illumina GoldenGate
10	Spain	558	550	Illumina GoldenGate	465	433	Illumina GoldenGate
11[e]	Italy	1,056	1,010	Illumina GoldenGate	864	804	Illumina GoldenGate
12[e]	Finland	270	259	Illumina GoldenGate	653[j]	653	Illumina 610-Quad[d]
Subtotal			**4,918**			**5,684**	
Analysis of Hap300 markers, and follow-up (91 SNPs)			**9,451**			**16,434**	
Analysis of additional Hap550 markers, and follow-up (40 SNPs)			**8,714**			**13,838**	

[a]Sample numbers attempted for genotyping, before any quality control (QC) steps were applied.

[b]Sample numbers after all quality control (QC) steps (used in the association analysis).

[c]All platforms contain a common set of Hap300 markers; the Hap550, 610-Quad, 670-Quad and 1.2M contain a common set of Hap550 markers.

[d]Finnish stage 2 controls were individuals within the Finrisk collection for whom Illumina 610-Quad genotype data became available after the completion of stage 1.

[e]As an additional quality control step, case-case and control-control comparisons for collection 1 versus 2, and collection 3 versus 12, for the 40 SNPs in **Table 2** were performed and no markers observed with $P<0.01$. We did observe (as expected) differences for collection 5 versus 11, from Northern and Southern Italy, respectively.

[f]All 737 post-QC cases reported in a previous GWAS(van Heel, Franke et al. 2007).

[g]690 of the post-QC cases and 1150 of the post-QC controls were included in a previous GWAS follow-up study (Hunt, Zhernakova et al. 2008).

[h]498 of the post-QC cases and 767 of the post-QC controls were included in a previous GWAS follow-up study (Hunt, Zhernakova et al. 2008).

[i]352 of the post-QC cases and 921 of the post-QC controls were included in a previous GWAS follow-up study (Hunt, Zhernakova et al. 2008).

[j]Some of these data were generated elsewhere, and some prior quality control steps (information not available) had been applied.

3.3.1.1 Genotyping bias considerations

A variety of factors may introduce systematic differences in genotyping between cases and controls (genotyping bias). Batch effects result from differences in assay performance under different conditions (e.g. different laboratories, different users, protocol variations, different assay batches or assay expiry times). Platform-specific biases may arise if assays for the same genetic marker perform differently on different genotyping platforms. This source of bias was observed in preliminary analyses of the data, arising from differences in SNP assay probe intensities between platforms and between versions of the same platform (see 3.4.1.1, **Figures 3.2 & 3.3**). Genotype calling biases can arise also for low minor allele frequency SNPs when sample numbers are low. This is a consequence of difficulty assigning the correct genotype to the minor allele homozygote which may have only a few or no observations if the sample size is small.

In order to minimize genotyping bias, taking account of these considerations, genotype calling from normalized SNP assay probe intensity data was performed in matched pools with the following objectives (**Table 3.3**).

1. Ensure samples within a calling pool have similar assay intensity data characteristics.
    - It was aimed to match genotype calling assay chemistries, platforms and genotyping facility. Random SNP intensity plots from each cohort were inspected to determine whether SNP intensity characteristics were similar for pooled cohorts.

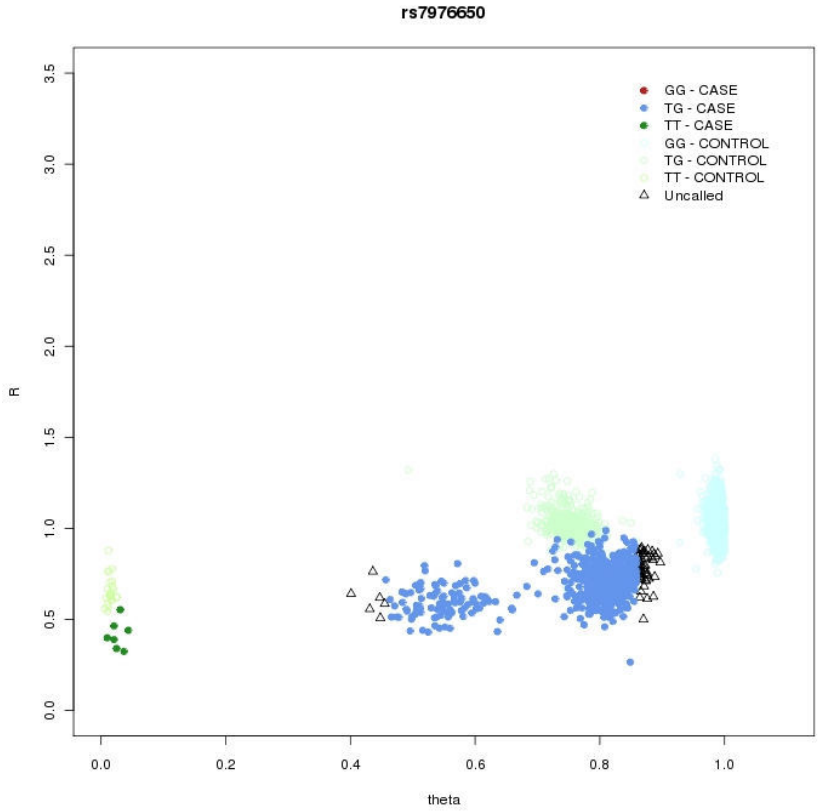
2. Ensure sufficient sample numbers for reliable calling of low minor allele frequency SNPs.
    - It was estimated that reliable SNP calling would require a minimum of 5 samples per genotype. Assuming perfect hardy-weinberg equilibrium, 2000 samples are required to observe 5 minor allele homozygotes for a SNP with a minor allele frequency of 0.05 (number of minor all homozygotes = 0.05^2 x 2000).

Table 3.3 Genotype calling pools for the GWAS

Genotype calling pool (sample numbers)[a]	Sample cohort (sample numbers)[a]	Genotyping facility	Genotyping platform[b]	Assay chemistry[c]	Number of shared markers for genotype calling
ONE: UK1 cases (2188)	UK1 cases (767)	Wellcome Trust Sanger Institute	Hap300_v1-1	Infinium II	313,481
	1958 birth cohort-WT (1421)[d]	Wellcome Trust Sanger Institute	Hap550_v1-1	Infinium II	
TWO: UK1 controls[e] (2596)	1958 birth cohort-T1DGC (2596)	Wellcome Trust Sanger Institute	Hap550-2_v3	Infinium II	561,466
THREE: UK2 collection (6963)	UK2 cases (1894)	Barts & the London Genome Centre	670-QuadCustom_v1	Infinium HD	660,447
	UK2 controls (5069)	Sanger Institute	1.2M-DuoCustom_v1	Infinium HD	
FOUR: Dutch and Italian collections (2957)	Dutch cases & controls (1836)	University Medical Centre Groningen	670-QuadCustom_v1	Infinium HD	660,447
	Italian cases & controls (1121)	University Medical Centre Groningen	670-QuadCustom_v1	Infinium HD	
FIVE: Finnish collection (6760)	Finnish cases (674)	Wellcome Trust Sanger Institute	670-QuadCustom_v1	Infinium HD	562,831
	Finnish controls (1839- 912 Finrisk, 927 Health 2000)	Wellcome Trust Sanger Institute	610-Quad_v1	Infinium HD	
	Additional Finrisk & Health2000 samples (4247)[f]	Wellcome Trust Sanger Institute	610-Quad_v1	Infinium HD	

[a]Sample numbers are numbers of samples entered for genotype calling [b]All genotyping platforms were Illumina microarrays using the patented Illumina Infinium assays and BeadArray technologies [c]Assay chemistry refers to the Illumina protocol used for sample preparation and genotyping [d]These samples were used for genotype calling but not included in association analyses [e]SNP intensity characteristics for this cohort showed significant differences to UK1 case data and so this cohort was called in a separate pool. An attempt to call UK1 cases and controls together led to multiple false positive SNP associations and genotyping bias.[f]Additional Finrisk and Health2000 samples included individuals with coronary artery disease and metabolic syndrome. These samples were used for genotype calling but not included in association analyses

Figure 3.2 **SNP genotyping error arising from automated genotype-calling. Differences in SNP assay probe intensity characteristics (R and theta[a]) between UK1 cases (Illumina Hap300_v1 array) and UK1 controls (Illumina Hap550_v3 array) cause non-overlap of genotype cluster positions**



[a]R = sum of normalised intensities of assay probes corresponding to the two alleles (X_{norm} and Y_{norm}).
Theta ("copy angle") = ratio of X_{norm} and Y_{norm} normalised to between 0 and 1 (theta = (2/pi) x arctan2 (Y_{norm} ,X_{norm}))
X_{norm} and Y_{norm} are normalised Cy3 and Cy5 probe intensities corresponding to the 2 SNP alleles

Automated calling of this SNP (rs7976650) generated MAF_{cases} = 0.51, $MAF_{controls}$ = 0.09, P_{assoc} = 2.9 x 10^{-290}. SNP intensity data from the UK1 cases and controls were called separately to avoid this bias.

3. Call genotypes on cases and controls together.
    - This was designed to minimize systematic differences in genotype calling between cases and controls.

Automated genotype calling was performed for each pool using a custom SNP genotype calling algorithm as used in coeliac GWAS1 (van Heel, Franke et al. 2007). A modification of the algorithm was introduced (Lude Franke – personal communication) such that SNP assay probes would be consistently identified with SNP alleles called in a standardised labelling format ("Illumina TOP"). This modification resolved potential inconsistencies in allele labelling for data generated at different facilities. The earlier version of the algorithm called alleles from normalised intensity data in the format provided in the output file (FinalReport file) from BeadStudio. Using the earlier version, inconsistencies arose frequently due to a lack of consensus on preferred allele labelling format in the FinalReport files generated for data from collaborating genotyping facilities. For example, 1958bc-T1DGC genotype data was available with alleles labelled only in "dbSNP forward strand" format. Attempting to merge data called in this format with datasets in other formats, led to difficulties assigning the DNA strand correctly for each SNP across all datasets. To validate automated genotype calling, cluster plots for all SNPs showing case-control association in the GWAS at a *P* value of less than 10^{-4} were generated and inspected in each sample collection (**Figure 3.3 for examples**).

3.4 Results

Quality control steps and SNP association results for stage 1, GWAS are first discussed in sections 3.4.1 and 3.4.2. Follow-up genotyping results for 131 selected SNPs showing association in stage 1 and combined (stage 1 and stage 2) association results are discussed in section 3.4.3.

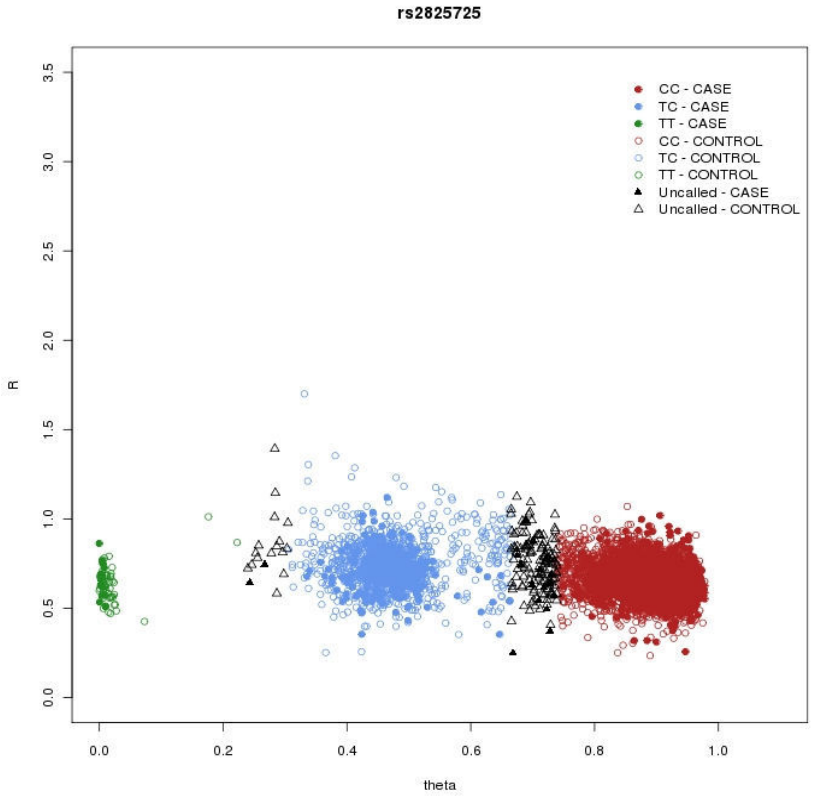
3.4.1 Stage 1- GWAS quality controls

Bead intensity data was processed and normalized for each sample in BeadStudio v3.2, an Illumina software application for visualizing and analyzing genotype data from SNP arrays. Samples showing call rate < 0.95 in BeadStudio using default (empirically determined on HapMap samples by Illumina) genotype cluster positions were excluded. R and theta values were exported for remaining samples and genotype calling performed using a modified algorithm (van Heel, Franke et al. 2007; Franke, de Kovel et al. 2008). R and theta describe the normalized 2-channel (Cy3 and Cy5) probe intensity data for each SNP assay (see **Figure 3.2** for detail).

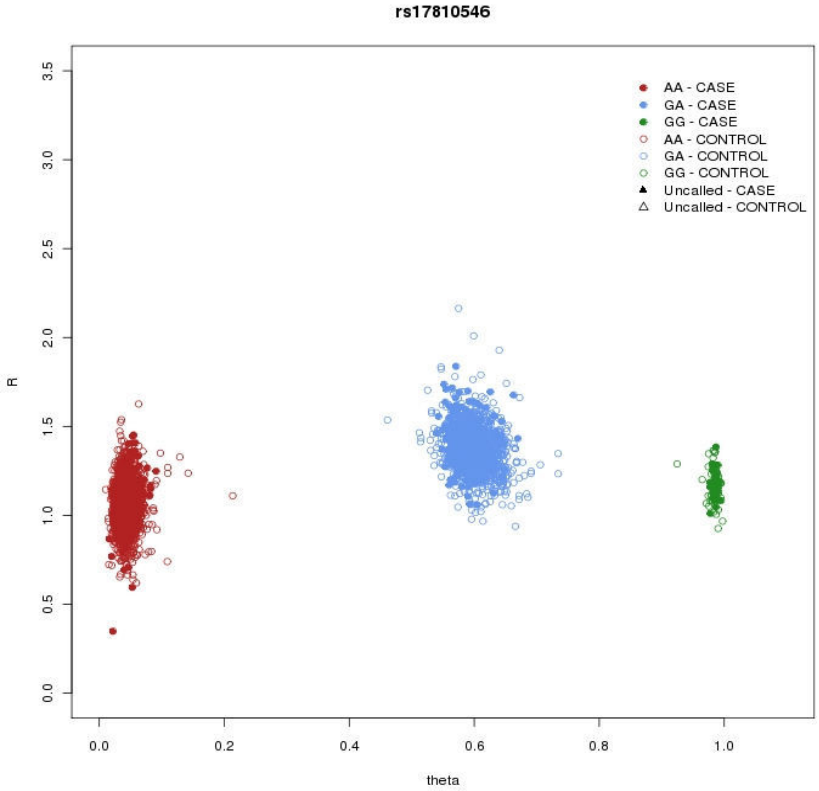
Further quality control steps were performed in the following order: CNV markers (monomorphic assays) were excluded. For the UK1 collection alone, where SNP intensity characteristics differed significantly between cases and controls and genotypes were called in separate pools, low (<0.95) call-rate SNPs were first excluded separately from cases and controls. All subsequent quality controls were performed on merged case and control datasets for each case-control sample collection. Samples were excluded for call rate <98%, incompatible recorded gender and genotype inferred gender, duplicates and first degree relatives and ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap Phase II data) (**Table 3.4)**. SNPs were excluded for call rates less than 95% or deviation from Hardy-Weinberg equilibrium (P<0.0001) in controls (**Table 3.5**).

Figure 3.3 SNP cluster plots for UK2 data. A. Example of a SNP, excluded after plot inspection, with probable genotype calling bias confounding case-control association (P_{assoc} UK2 = 1.55 x 10^{-03}). B. Example of a SNP, passing quality controls, P_{assoc} UK2 = 9.83 x 10^{-06}

A.



B.



After performing quality control steps within each sample collection, data from all 5 collections were merged for association analyses. Association analyses were performed separately for SNPs passing quality controls in all five collections (295,453 SNPs – "Hap300k") and for an additional set of 231,516 SNPs ("Hap250k") passing quality controls in UK2, Dutch, Italian and Finnish collections. 22 of 417 SNPs showing association ($P_{GWAS} < 10^{-4}$) were excluded after visual inspection of R theta plots in each sample collection suggested possible bias (**Figure 3.3**).

Figure 3.4 shows the allele frequency distribution of SNPs passing quality controls.

3.4.1.1 Exclusion of duplicate and closely related samples

The genome wide average proportion of alleles shared *identical by state* (IBS) can be used in homogeneous samples to estimate the proportion of alleles shared *identical by descent* (IBD). This analysis was implemented in PLINK v1.05 (Purcell, Neale et al. 2007). The "PI_HAT" metric estimates for each sample pair, the proportion of all SNPs where both alleles are shared *identical by descent*. This analysis has good sensitivity for detection of duplicates (PI_HAT = 1) and first degree relatives (PI_HAT = 0.5), though loses sensitivity to detect more distant relatives (**Figure 3.5**). Duplicates and first degree relatives were excluded from all datasets.

Table 3.4 Sample exclusions by sample collection

Sample collection		Inferred sex- gender incompatibility[a]	Call rate <0.98	Ethnic outlier	Related samples		Total (%)
					Duplicates	1st degree relative	
UK1	Cases[b]	0	6	0	14	10	30 (3.91%)
	Controls[c]	0	0	0	0	0	0
UK2	Cases	0	36	5	15	17	73 (3.80%)
	Controls	8	61	18	0	47	134 (2.64%)
Dutch	Cases & controls[d]	3	1	97	54		155 (8.44%)
Italian	Cases & controls[d]	6	0	8	58		72 (6.42%)
Finnish	Cases	0	0	0	6	12	18 (2.67%)
	Controls	0	0	8	1	1	10 (0.54%)
All samples		17	75	136	235		492(3.11%)

[a]Sex inferred from X chromosome genotype was compared to gender documented in sample records [b]Where related sample pairs were found between UK1 and UK2 collections, UK1 sample was removed [c]Data for UK1 controls had undergone prior quality controls for the Type 1 diabetes genetics consortium [d]Quality controls performed by Gosia Trynka, Netherlands. Breakdown by case and controls not available. Breakdown of relateds (duplicates versus 1st degree relatives not available)

Table 3.5 SNP exclusions by sample collection

Sample collection		SNPs genotyped in cases & controls and included for QC	Call rate < 95%	Hardy Weinberg $P < 10^{-4}$ in controls[a]	Suspected genotyping error (SNP plot inspection for SNPs $P_{assoc} < 10^{-4}$)[b]
UK1	Cases Controls	307,798[c]	1,416 676	1,073	29 (31.2%)
UK2	Cases Controls	562,831[d]	23,658 24,019	4,165	44 (25.6%)
Dutch	Cases & controls	562,831[d]	21,642	3,891	Not done
Italian	Cases & controls	562,831[d]		3,791	Not done
Finnish	Cases Controls	562,831[d]	12,790	2,399	43 (28.9%)

[a]Indicates deviation of observed genotype frequencies from hardy-weinberg equilibrium

[b]Additional inspection of SNP plots for each sample collection was carried out for SNPs with $P_{GWAS} < 10^{-4}$. This led to exclusion of a further 22 out of 417 SNPs (5.27%).

[c]SNPs genotyped in both cases and controls.

[d]SNPs genotyped in UK2,Dutch,Italian and Finnish collections and common to the genotyping platforms used (670Quad-Custom_v1, 1.2MDuo-Custom_v1, 610Quad_v1)

Figure 3.4 Minor allele frequency distributions of SNPs passing quality controls in the Hap300k SNP set (all collections) and Hap250k SNP set (UK2,Dutch, Italian, Finns only)

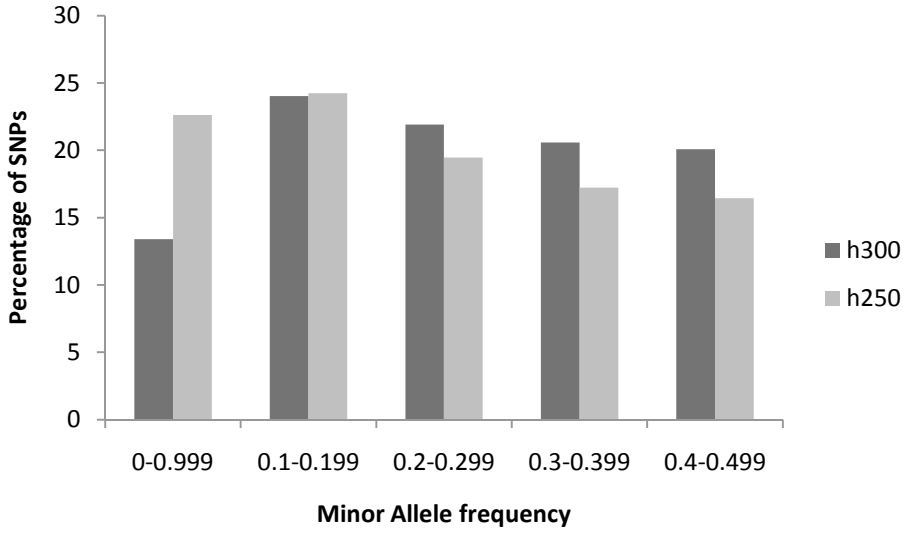
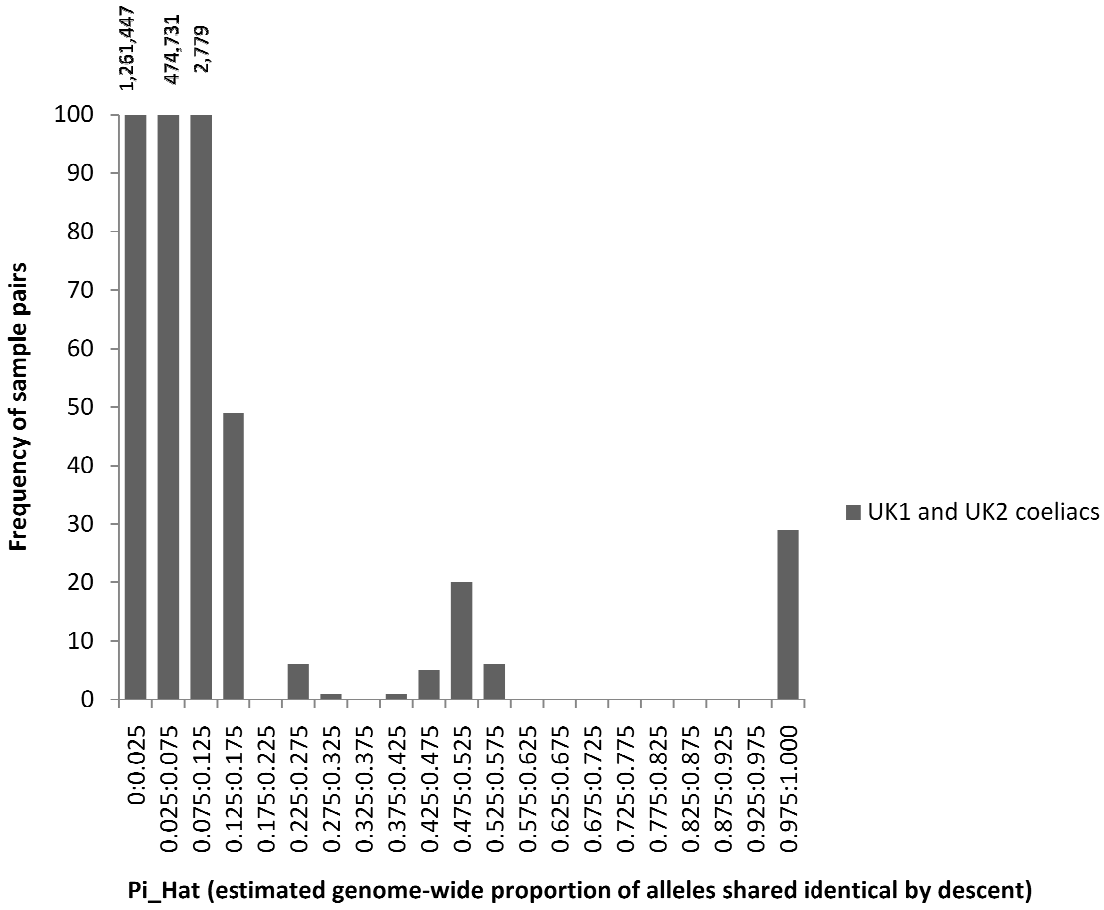


Figure 3.5 **Pairwise genome-wide SNP genotype *identity by descent* estimation for identifying related samples**



IBD estimates calculated in PLINK for 2587 UK (UK1 & UK2) coeliac samples, using 12,366 ancestry-informative SNPs, prior to related sample exclusions. The sample frequency peaks at Pi_HAT ~1, 0.5 and 0.25 correspond to duplicate, 1st degree relative and 2nd degree relative pairs. Pi_HAT does not resolve less closely related sample pairs (e.g. 3rd degree relatives with Pi_HAT ~0.125 overlap with tail of distribution for unrelated sample pairs).

3.4.1.2 Ethnic outlier analysis

Ethnic outliers were detected as described in general methods. A pairwise IBS matrix for a genome-wide SNP set was used for classical multi-dimensional scaling, a statistical method similar to principal components analysis that calculates values for each sample on orthogonal dimensions describing variation in the *IBS* metric. Using this method ancestral variation between 3 HapMap populations (CEU, YRI and CHB/JPT) is readily distinguished by plotting samples for the first two MDS dimensions (Wellcome Trust Case Control Consortium 2007). This technique was used in all sample collections, and sample outliers (greater than 4 standard deviations from the mean on either 1st or 2nd dimension) removed (**Figure 3.6**).

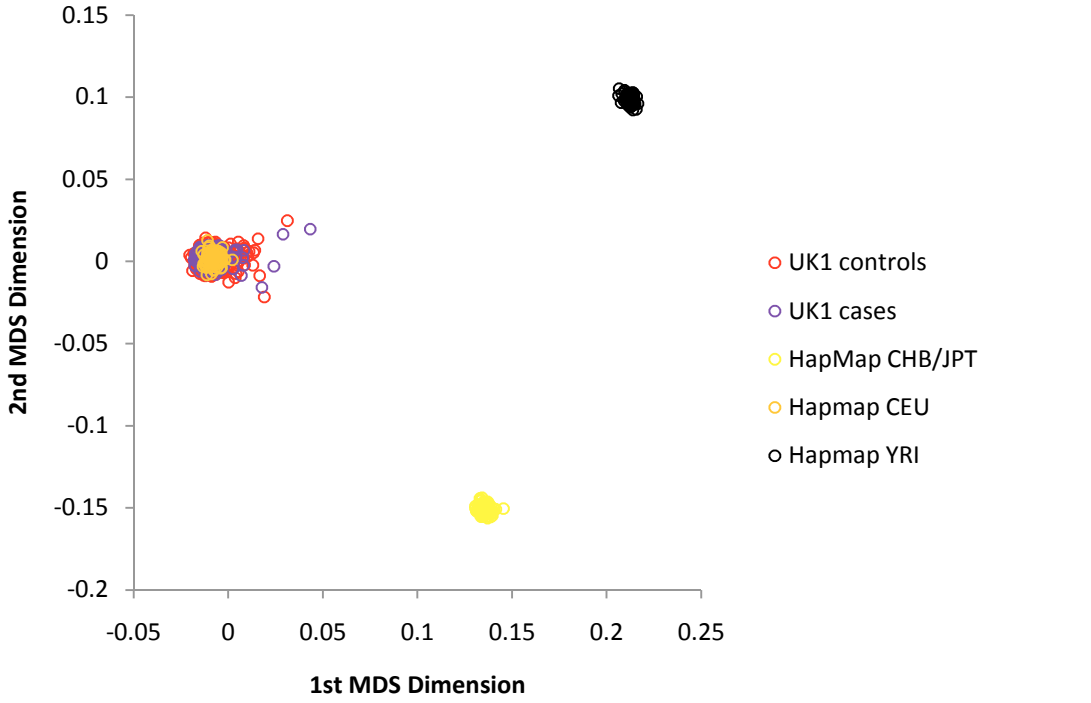
3.4.1.3 Identifying and controlling for population structure in cases and controls

Where relatedness within cases and within controls is greater than between cases and controls, case-control allele frequency differences can occur due to ancestry differences, independently of affectation status. In order to identify genetic variants that are causally associated with the disease, it is therefore important to ensure that ancestry differences between cases and controls are minimized since such differences can otherwise confound genetic association studies.

As a first step, cases and controls were matched for geographical region (by country of origin, **Table 3.2**). Association analyses were performed independently within these strata, to avoid confounding by ancestry differences between countries. The importance of this step was explored by comparing genome-wide SNP allele frequencies between control cohorts from each of the four countries (UK, Netherlands, Italy and Finland). **Figure 3.7** illustrates these differences by plotting control samples from each country on the first two multi-dimensional scaling dimensions. These first two dimensions capture genome-wide variation in SNPs due to ancestry, for example enabling easy distinction of the HapMap phase II populations (**Figure 3.6**) (Wellcome Trust Case Control Consortium 2007). To determine the extent to which the genetic differences observed between cohorts from different countries might confound case-control association analyses, control cohorts from each country were compared pairwise by genome-wide SNP association testing. In this analysis, the genomic inflation factor ($lambda_{GC}$), calculated as the observed median association test statistic / expected median association test statistic under the null hypothesis of no association, provides an indication of the genome-wide average degree of confounding present. Inflation ($lambda_{GC} > 1$) suggests genome-wide allele frequency differences. **Table 3.6** presents $lambda_{GC}$ for between-country control cohort comparisons. As expected, and in line with differences illustrated in **Figure 3.7** Italy and Finland appear as relative outliers, with the UK and Netherlands much more similar. The inflation of $Lambda_{GC}$ increases with sample size, since association test statistic values increase for the same absolute allele frequency difference as sample size increases. For this reason, in **Table 3.6**, $lambda_{GC}$ 1000 ($lambda_{GC}$ scaled for a sample size of 1000) is presented to aid comparison of genetic differences between association sample sets of varying size.

Figure 3.6 **Ethnic outliers visualised through multi-dimensional scaling plots. A. UK1 collection. B. UK2 collection**

A.



B.

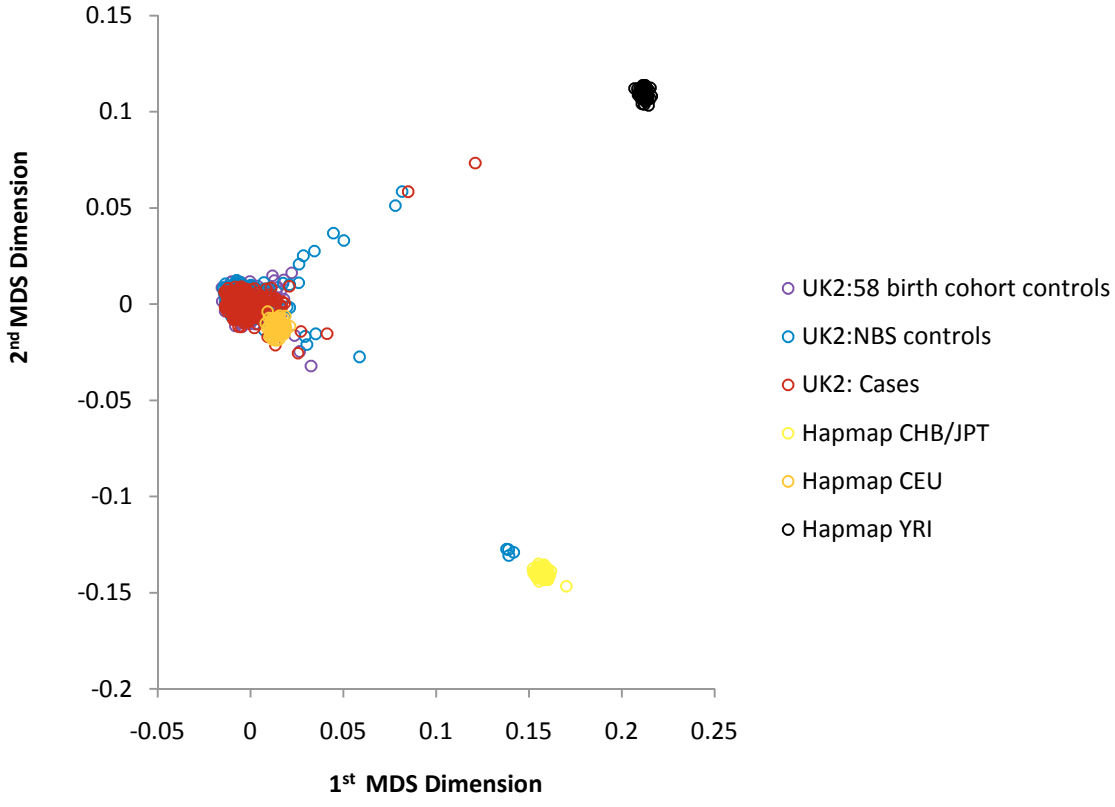
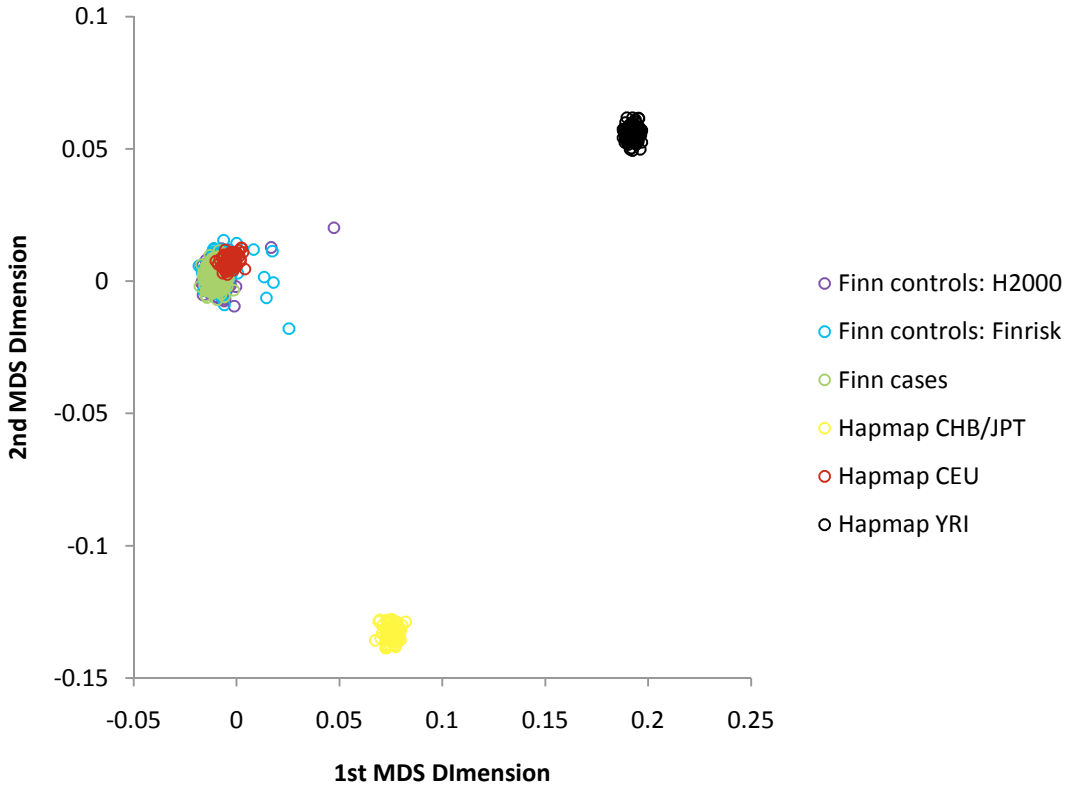


Figure 3.6 (cont.) Ethnic outliers visualised from multi-dimensional scaling plots. C. Finnish collection

C.



The comparisons between controls from the different countries also provide some insight into whether genome-wide allele frequency differences occur generally in the data due to genotyping bias or ancestry differences. If genotyping bias is prevalent, inflation of test statistics should be expected for comparisons of cohorts genotyped and called in differing ways (e.g. distinct genotyping platforms, genotyping facility, calling pool etc), even if ancestry is known to be similar. UK1 and UK2 controls, of known similar ancestry, offer this comparison since they were genotyped on different platforms and called independently. Thus the observed minimal inflation is re-assuring that no major genotyping bias exists. Conversely, ancestry differences should be expected to inflate test statistics in line with known differences between populations and this is also observed with the greatest inflation for comparisons between Italians and Finns and lesser inflation between UK and Dutch.

As a second step towards reducing confounding by relatedness and ancestry, duplicates, first degree relatives and ethnic outliers were removed. However, the analysis used to detect relatives lacks sensitivity for individuals of 2nd degree and lower relatedness (**Figure 3.5**). To determine whether hidden relatedness within each case- control collection might confound association analysis, lambda$_{GC}$ was calculated again, this time for cases versus controls. Since disease-causing variants were expected to be a small fraction of all assayed genomic variants, particularly after excluding the HLA region, significant inflation of the median chi-square statistic (lambda$_{GC}$) is not expected to occur due to disease causing variants alone(Balding 2006). **Table 3.7** presents lambda$_{GC}$ values before and after adjustment for the top ten principal components (section 3.4.1.3.1)

Table 3.6 **Genome-wide SNP allele frequency differences between European population control cohorts included in the GWAS and the effects on inflation of association test statistics.** Unadjusted (lambda$_{GC}$) and sample-size adjusted (lambda$_{GC}$ 1000) genomic inflation factors for SNP allele frequencies compared between different European populations control cohorts

Controls compared	Sample size	Lambda$_{GC}$	Lambda$_{GC}$ 1000
UK2- UK1	7532	1.014	1.002
UK2-DUTCH	5782	2.118	1.193
UK2-ITALIAN	5479	7.642	2.212
UK2 –FINNISH	6765	36.467	6.264
DUTCH-ITALIAN	1389	5.496	4.237
DUTCH-FINNISH	2675	16.250	6.700
ITALIAN-FINNISH	2372	21.044	9.450

Lambda$_{GC}$ 1000 is the inflation factor adjusted for a sample size of 1000, calculation as described by de Bakker et al. (de Bakker, Ferreira et al. 2008)

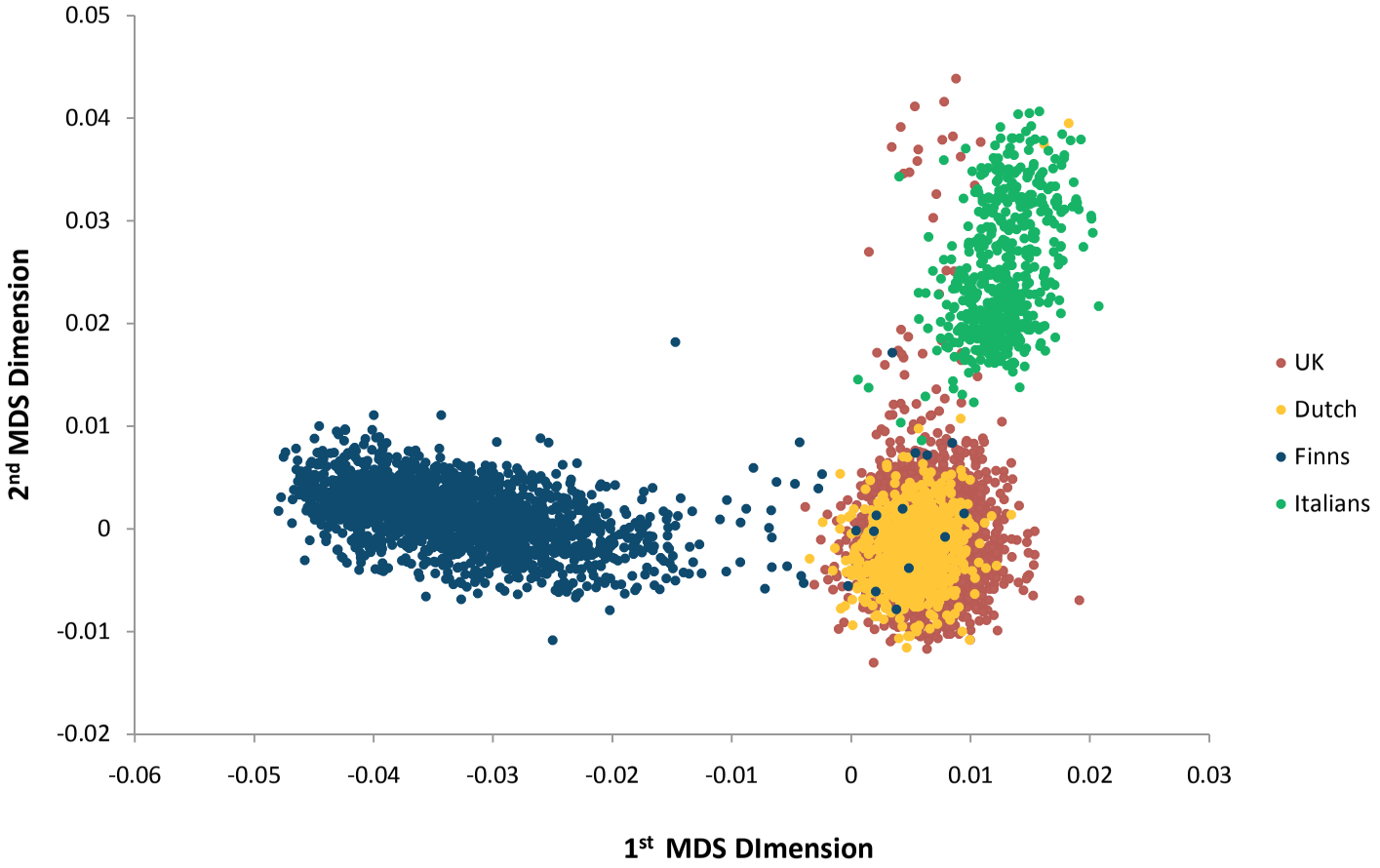
UK2-FINNISH, UK2-DUTCH, UK2-ITALIAN comparisons for 528,969 post-QC SNPs. UK2-UK1 comparisons for 295,453 post-QC SNPs

Table 3.7 **Genomic inflation factor (λ) by sample collection**

Case-control collection	λ unadjusted	λ adjusted
UK1	1.036	1.029
UK2	1.043	1.040
Dutch	1.045	1.042
Italian	1.020	1.016
Finnish	1.140	1.028

λ was calculated for Cochran-Armitage trend test association statistics (unadjusted) and after correction using the EIGENSTRAT method (Price, Patterson et al. 2006).

Figure 3.7 **Ancestral variation in genome-wide SNP data, visualised for four European population control cohorts**



Samples from the four GWAS control cohorts plotted for first two MDS dimensions. UK is pool of UK1 and UK2 controls. Sample collections after quality controls including ethnic outlier removal. 12,366 ancestry-informative non-HLA SNPs used for *identity by state* and multi-dimensional scaling analyses.

3.4.1.3.1 Principal Components Analysis

Principal components analysis (PCA) is a method that seeks to describe variation in a dataset along independent (orthogonal) axes (principal components). Principal components are ranked such that the first principal component describes the largest proportion of variation, the second describes the next largest proportion, and so on. With regard to SNP genotype data this method has been used to identify components of ancestry variation in the data(Price, Patterson et al. 2006). Where adjustment of allele frequencies using principal components causes lambda to fall back towards neutral (i.e. 1), this is evidence that inflation of lambda in the uncorrected data is due largely to correlated differences in SNP allele frequencies. Although theoretically genotyping bias (e.g. batch effects) can contribute to such correlated differences, population structure was a greater concern in the current GWAS analysis (see 3.4.1.3). Principal components were calculated for each sample collection separately and the genomic inflation factor calculated from association test statistics both before and after adjustment for the first 10 principal components (**Table 3.7**). The HLA region was excluded prior to these association analyses to avoid confounding by multiple strong coeliac associations mapping to this region. Inflation of association test statistics was modest in all collections, with the exception of the Finnish collection (**Table 3.7**).

3.4.1.3.2 Controlling bias due to population stratification

The genomic inflation factor (λ_{GC}) for the GWAS meta-analysis calculated for 15,288 samples was 1.11, suggestive of modest overall inflation and consistent with other studies of this size (Barrett, Hansoul et al. 2008; Barrett, Clayton et al. 2009). However, appreciable inflation was observed within the Finnish collection (lambda = 1.14, sample size = 2,503, **Table 3.7**). Finnish population substructure is well-known and has been observed in genome wide SNP data, reflecting population migrations from both the east and south/west into Finland during its settlement history (Jakkula, Rehnstrom et al. 2008; McEvoy, Montgomery et al. 2009). In the Finnish sample collection cases were mainly from southern Finland around Helsinki. Controls comprised two cohorts – Finrisk, mainly from southern Finland and Health 2000 from all over Finland. However, even within a region such as southern Finland, substantial ancestry-related genetic differences have been observed (Jakkula, Rehnstrom et al. 2008). The geographical

origins of the Finnish samples therefore did not provide re-assurance that cases and controls were well-matched for ancestry.

To further characterize genetic differences between Finnish cases and controls, principal components were calculated using a subset of 12,898 SNPs showing low linkage disequilibrium in the HapMap CEU population(Yu, Wang et al. 2008). Since linkage disequilibrium is a source of SNP allele correlations, and mostly reflects distant, including pre-human ancestry, application of principal components analysis to SNP data without prior LD-pruning tends to extract principal components partly capturing these patterns of local linkage disequilibrium. Longer distance SNP correlations, informative of more recent ancestry are therefore confounded by local linkage disequilibrium when performing principal components analysis on unfiltered genome-wide SNP data.

After exclusion of the HLA region 12,366 SNPs were used for analysis. In the Finnish dataset, cases and controls were strongly differentiated along the top four principal components ($P < 0.001$, ANOVA). However, most of the case-control variation is captured by the top two principal components. **Figure 3.8** plots eigenvalues for the first two principal components for each sample in the Finnish sample collection.

In order to control this variation, which was a confounding factor for the Finnish case-control association analysis, three methods were explored.

1) Removal of 70 individuals that were outliers on either of the first two eigenvectors (E1> 0.35, E2 > 0.40). After these exclusions $lambda_{GC}$ was unchanged (1.14), suggesting that significant structure remained.

2) Using a pairwise *identity by state* (IBS) matrix, samples were sequentially clustered, based on pairwise genetic similarity, to form 100 clusters. Association testing was repeated stratifying the analysis within each cluster. $Lambda_{GC}$ was again 1.14 suggesting that this method could not adjust for population structure.

3) Principal components were used as covariates to adjust association test statistics on the whole SNP dataset. Applying this method, using the top 10 eigenvectors, $lambda_{GC}$ fell to 1.028 (**Methods**). This demonstrated that principal components, calculated from the LD-filtered

12,366 SNP set, captured the majority of case-control SNP allele variance in the genome-wide SNP data. This is further evidence for underlying population structure and indicates that this may be controlled using a principal components based method.

Principal components-adjusted association test statistics for the GWAS meta-analysis are presented in **Table 3.11**, calculated using a sample size weighted Z score method.

3.4.2 SNP association results in the GWAS (Stage 1)

SNP association analysis was performed independently for the human "hap300k" marker set (295,453 SNPs), genotyped in all five sample collections and for the human "hap250k" set (an additional 231,516 SNPs) genotyped only in UK2, Dutch, Italian and Finnish collections. After performing analyses of the HLA region, all subsequent case-control association analysis was performed after exclusion of SNPs within the broad HLA region (Chr 6, 20-40Mb-**Methods)**.

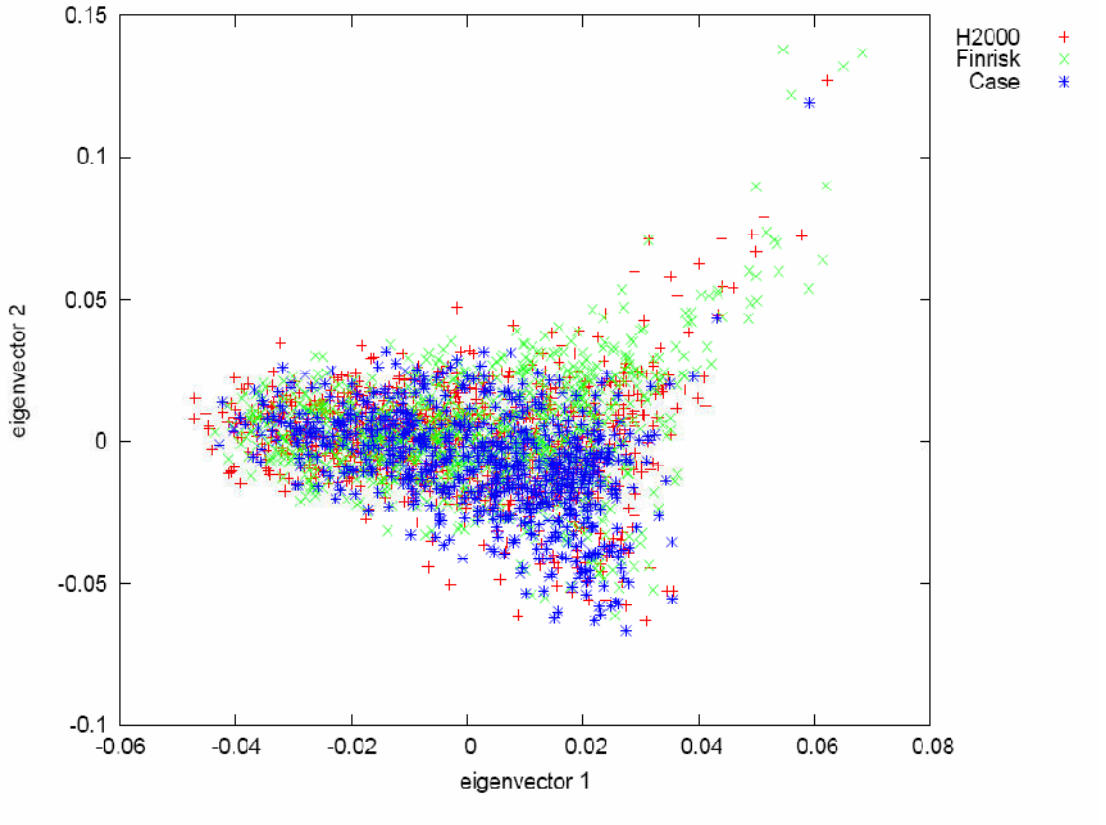
3.4.2.1 HLA association with coeliac disease

Strong association with coeliac disease was observed within the HLA region, the strongest overall association obtained for rs2187668, which is a near-perfect tag SNP for HLA-DQ2.5*cis* (the *HLA-DQA1*0501-HLA-DQB1*0201* haplotype encoding the DQ2 heterodimer)(Monsuur, de Bakker et al. 2008). **Table 3.8** shows inferred DQ2.5*cis* carriage rates for each of the five GWAS sample collections. These frequencies are consistent with reported data, including a known lower DQ2 frequency in Italian coeliacs, and an overall north-south European gradient in DQ2 frequencies in coeliac disease (Karell, Louka et al. 2003).

Comparing the two UK coeliac collections, UK2 DQ2.5*cis* frequency (0.844) was lower than UK1 DQ2.5*cis* frequency (0.883). This difference could reflect chance sampling variation from the same UK coeliac population (P_{Fisher} = 0.0002). However, differences in recruitment of UK1 and UK2 coeliacs may also have increased the chances of individuals without coeliac disease being inadvertently included in the UK2 coeliac cohort. Among UK2 coeliacs, 1415 out of 1849 individuals had been recruited by direct advertisement through Coeliac UK. For these individuals, hospital records were not readily available and diagnostic criteria could not be independently verified. Case status was established on the basis of Coeliac UK membership,

self-confirmed coeliac disease diagnosis and self-confirmed intestinal biopsy or coeliac antibody test.

Figure 3.8 **Ancestry differences between Finnish cases and controls, visualised by plotting eigenvalues for the first two principal components**



Principal components were calculated using 12,366 ancestry-informative SNPs. The first two principal components capture the majority of nationwide ancestral variation in Finland (Jakkula, Rehnstrom et al. 2008). Cases and controls were significantly differentiated on both eigenvector 1 (P_{ANOVA}= 5.94 x 10^{-04}) and eigenvector 2 (P_{ANOVA} = 1.67 x 10^{-15}).

As HLA-DQ2.*5cis* is the most strongly differentiated genetic marker between coeliacs and controls, its frequency can be used as a surrogate marker of the relative purity of a coeliac cohort. UK coeliacs recruited using standard objective criteria including intestinal biopsy from hospital outpatients (n= 1171) and UK coeliacs recruited by direct advertisement (n=1415) were compared directly. Inferred HLA-DQ2.*5cis* frequency was lower in direct advertisement-recruits (0.835), than in UK hospital recruits (0.873; P_{Fisher} – 0.0063). Assuming a true UK coeliac DQ2.*5cis* frequency of 0.873 and a UK control frequency of 0.266 (UK1 and UK2 control

combined DQ2.5*cis* frequency), the number of non-coeliacs among the direct advertisement recruits was estimated as 89 (6.3%). As such the effect on overall association results is likely to be small. The allelic case-control odds ratio for rs2187668, for example, is 6.58 (95% CI: 5.79-7.49) in the UK1 collection and 6.20 (95% CI: 5.65-6.80) in direct advertisement-recruited coeliacs versus UK2 controls.

Tag SNPs allowing imputation of other common coeliac-associated DQ types (HLA-DQ2*trans*, HLA-DQ8) were not available and so the frequencies of these heterodimers were not assessed (Monsuur, de Bakker et al. 2008).

3.4.2.1.1 Non-HLA-DQ coeliac disease associations in the HLA gene region

Extensive linkage disequilibrium within the HLA region has confounded previous attempts to test for HLA associations that are independent of known coeliac disease associated alleles (Louka, Moodie et al. 2003). In order to control for HLA-DQ association with coeliac disease, HLA SNP association analysis was performed in a GWAS subset of 764 cases and 196 controls homozygous for the DQ2.5*cis* tagging SNP rs2187668 (TT homozygotes). Association analysis was performed using Cochran-Mantel-Haenszel (sample collection-stratified) chi-square test of SNP allele counts and stratified logistic regression, after excluding SNPs with minor allele frequency < 0.05. Low MAF SNPs were excluded due to poor performance of the chi-square test for low allele counts, which was relevant in this small case-control sample. Residual association signal was observed within the region containing the *HLA-DQA1* and *HLA-DQB1* genes. The peak association in this region (within a 1 megabase window around rs2187668) was at rs3117582 (P_{CMH}= 1.35 x 10^{-4}, $P_{Logistic}$ = 1.08 x 10^{-4}), with multiple SNPs showing modest association (P<10^{-3}) throughout this region. Since multiple immune genes are located within this region (e.g. *HLA-DRB1, HLA-DRB5*) residual association signals may indicate independent non-DQ association. Alternatively, it was considered whether rs2187668 TT homozygotes imperfectly inferred DQ2.*5cis* homozygotes in the GWAS samples, leading to a residual HLA-DQ association signal in this analysis. For example, if linkage disequilibrium between the rs2187668 T allele and the DQ2.5*cis* haplotype varied between different European populations, this might occur. However, rs2187668 was originally established as a near-perfect tag for DQ2.5*cis* in each of UK, Dutch, Italian and Spanish populations (sensitivity 1.000, specificity 0.999) (Monsuur, de Bakker et al. 2008). Furthermore rs2187668-inferred DQ2.5*cis* frequencies in the GWAS coeliac collections were consistent with known DQ2

frequencies in these populations (**Table 3.8, (Karell, Louka et al. 2003)**). Finally there was no evidence of heterogeneity of odds ratios between sample collections for associated SNPs in the region, including for the most strongly associated SNP, rs3117582 ($P_{Breslow\text{-}Day}$ = 0.46). Inspection of rs2187668 cluster plots in GWAS collections did not suggest genotyping bias. No adequate proxy SNPs for rs2187668 (SNP in strongest LD was rs3129763, r^2 = 0.586 in 4936 UK2 controls) were available to perform an alternative selection of DQ2.5*cis* homozygotes. These factors provided re-assurance that *HLA-DQ* association was controlled in the analysis.

Within the broad HLA region (chromosome 6, 20-40 Mb), 20 SNPs were associated with coeliac disease in the DQ2.5*cis* homozygote analysis at a significance threshold of P_{CMH} or $P_{logistic}$ $<10^{-4}$. All 20 SNPs mapped to either of two distinct loci within the more narrowly defined HLA region (chromosome 6, 29-34Mb (Howson, Walker et al. 2009)). The most strongly associated SNP was rs9277554 (P_{CMH} = 9.27 x 10^{-6}, $P_{logistic}$ = 8.62 x 10^{-6} **Table 3.9**). This SNP lies within a region containing the MHC class II genes, *HLA-DPB1* and *HLA-DPA1.* Genetic variants in the *HLA-DPB1* gene region were recently associated with type 1 diabetes in a large analysis of HLA-DQ-, HLA-DRB1- independent associations. Association with another, closely related autoimmune disease increases the probability that variants identified in this region also independently influence susceptibility to coeliac disease (Howson, Walker et al. 2009). The SNP reported in type 1 diabetes, rs439121, was not genotyped on the Human 670-Quad custom chip and no adequate proxy SNP (r^2 > 0.7) was available in the GWAS data. However, rs439121 had been genotyped in UK2 controls on the Illumina 1.2MDuo-custom beadchip. Therefore pairwise SNP LD was estimated in 4936 UK2 controls between rs439121 and rs9277554 (r^2 = 0.395, D' = 0.698). These two SNPs map only 137kb apart, but with the current data, it is unclear whether the association signal is caused by the same variants in these two diseases.

The second most strongly associated SNP within the HLA region, rs9263715 (P_{CMH} = 1.42 x 10^{-5}, $P_{Logistic}$ = 3.86 x 10^{-5} **Table 3.9**) identified a region containing the class I MHC gene *HLA-C* . Association in the *HLA-C* gene region is also observed in type 1 diabetes, but in this disease is attributable to *HLA-DRB1*and *HLA-DQA1* alleles (Howson, Walker et al. 2009). Whether *HLA-DRB1* alleles contribute to coeliac disease risk is unknown. **Figure 3.9** displays association results across the HLA region (Chr 6, 29-34Mb).

Table 3.8 Estimated DQ2.5*cis* frequencies in each sample collection

Sample collection	DQ2.5*cis*[a] in cases (%)	DQ2.5*cis*[a] in controls (%)	Allelic odds ratio (95 % Confidence Interval)[b]
UK1	88.3	26.8	6.58 (5.79 - 7.49)
UK2	84.4[c]	26.4	6.29 (5.77 - 6.85)
Dutch	86.7	29.0	6.77 (5.75 – 7.98)
Italian	61.6	16.4	6.46 (5.03 – 8.30)
Finnish	87.3	18.8	9.19 (7.86 – 10.72)
Overall	**50.55**	**13.38**	**6.77 (6.38 – 7.18)**

[a]inferred from carriage of rs2187668 T allele

[b]odds ratio for rs2187668 T vs C (allelic chi-square test)

[c]DQ2.5*cis* 83.5% in individuals recruited through advertisement in Coeliac UK (n=434), 85.7% in individuals recruited from hospital outpatients (n=1415).P_{fisher} = 0.019.

Table 3.9 Strongest SNP associations in the HLA region in DQ2.5*cis* homozygotes

Position[a]	SNP	Minor Allele[b]	Minor Allele Frequency[b]	$P_{GWAS-CMH}$[c]	$P_{GWAS-logistic}$[d]	Odds ratio [95% CI][e]
33163516	rs9277554[f]	G	0.4421	9.27×10^{-6}	8.62×10^{-6}	0.60 [0.47-0.75]
31203780	rs9263715	T	0.0844	1.42×10^{-5}	3.86×10^{-5}	0.40 [0.26-0.61]

[a]NCBI build 36 coordinates on chromosome 6

[b]Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls.

[c] Cochran-Mantel-Haenzel test of allelic chi-square test

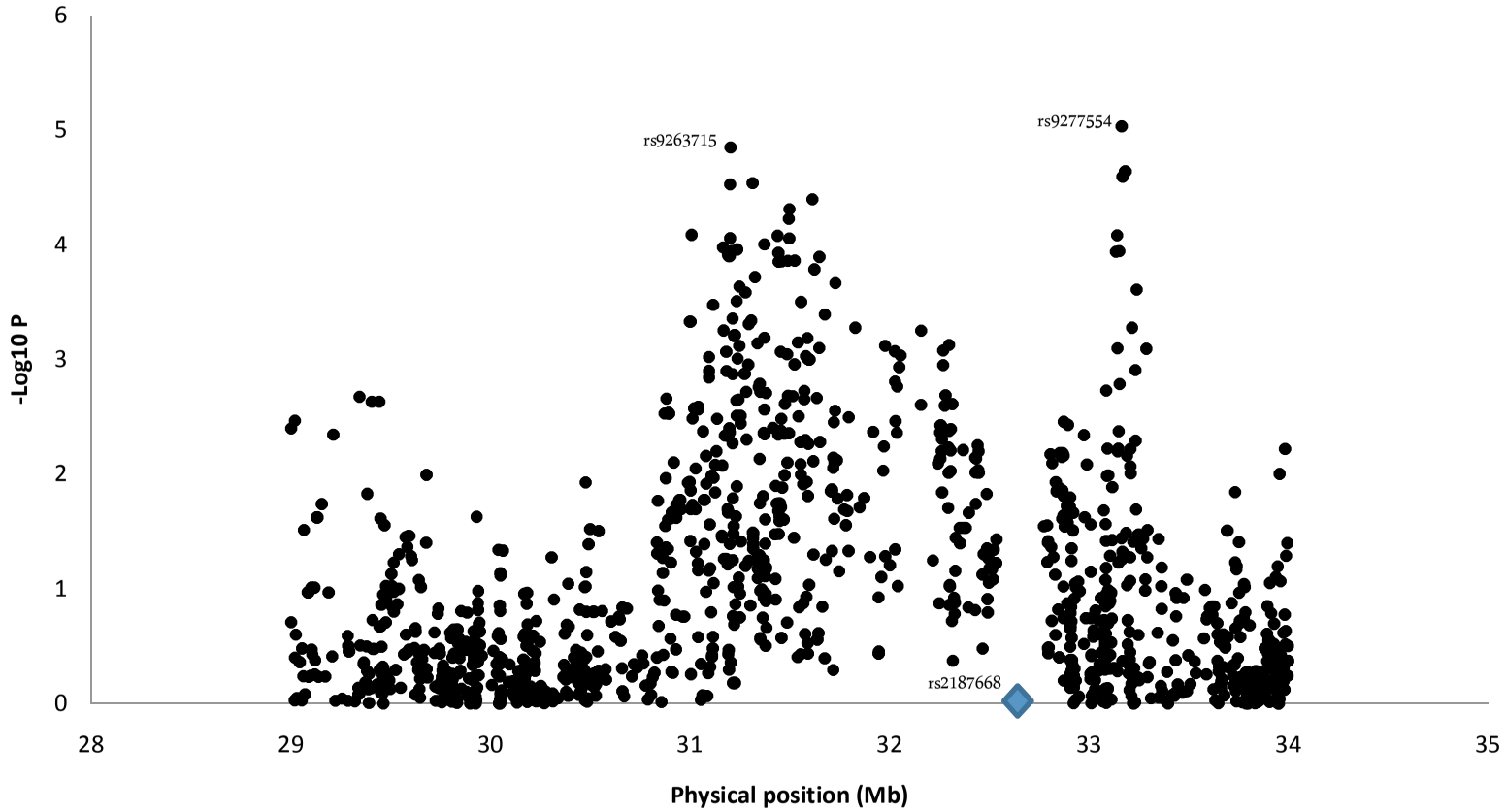
[d]Logistic regression performed on posterior genotype probabilities, with group membership included as a factorized covariate

[e]Odds ratio calculated with respect to minor allele in controls

[f]rs9277554 was not genotyped in UK1 individuals. Association P values are for 622 cases and 140 controls from UK2,Dutch,Italian and Finnish collections.

[g]rs9263715 genotyped in all samples (764 cases and 196 controls from the 5 GWAS sample collections

Figure 3.9 **Association plot of 1522 SNPs genotyped within the HLA region on chromosome 6 (29-34Mb) in DQ2.5*cis* homozygotes[a]**



[a]DQ2.5*cis* homozygotes inferred from rs2187668 TT homozygotes. Association results are for Cochran-Mantel-Haenzel allelic chi-square tests for 1072 SNPs from the "Hap300k" set, genotyped in 764 cases and 196 controls from the five GWAS sample collections and a further 450 SNPs within the "Hap250k" set genotyped in a subset of 622 cases and 140 controls (no UK1 samples).

3.4.2.2 Non-HLA associations in stage 1

After exclusion of HLA SNPs there were 292,387 SNPs from the Illumina Hap300 marker ("Hap 300k") set available for association testing in 4,533 celiac disease cases and 10,750 controls of European descent (**Table 3.2**). A further 231,362 additional non-HLA markers from the Illumina Hap550 marker ("Hap250k") set were tested for association in a subset of 3,796 celiac disease cases and 8,154 controls. All markers were from autosomes or the X chromosome.

Genotype call rates were >99.9% in both datasets. Findings were not substantially altered by imputation of missing genotypes for 737 coeliac disease cases genotyped on the Hap300 BeadChip and corresponding controls (**Table 3.2**, UK1 collection, imputation performed by Dr Jeff Barrett, Wellcome Trust Sanger Institute). In addition only directly genotyped data was analysed as around half the additional Hap550 markers cannot be accurately imputed from Hap300 data(Anderson, Pettersson et al. 2008).

All 13 previously reported non-HLA coeliac disease risk loci showed evidence for association in stage 1. Among these, 10 of 10 loci previously reported as meeting the WTCCC 2007 advocated genome-wide significance threshold ($P_{GWAS} < 5 \times 10^{-7}$) met the more conservative recently advocated genome wide significance threshold of $P<5x10^{-8}$ in stage 1 (**Table 3.10**)(Dudbridge and Gusnanto 2008; Pe'er, Yelensky et al. 2008). The 3 other previously reported loci, containing the *CTLA4/ICOS, PTPN2* and *ITGA4* genes had previously met less stringent significance thresholds (i.e. $P<10^{-4}$ for regions having been convincingly associated in another autoimmune disease(Smyth, Plagnol et al. 2008) or $P<10^{-3}$ for a strong functional candidate gene, *ITGA4* (Garner, Murray et al. 2009)). One of these regions met the $P<5 \times 10^{-8}$ threshold (ITGA4) in stage 1 while the other 2 showed strong association $P_{GWAS} = 8.80 \times 10^{-8}$ (*CTLA4/ICOS*) and $P_{GWAS} = 5.53 \times 10^{-7}$ (*PTPN2*) below this threshold.

Examination of the quantile-quantile plots of association test statistics in Hap300k (used in coeliac GWAS1) and Hap250k (unique to coeliac GWAS2) marker sets, showed that there was residual inflation of the tail of SNP association test statistics after removing 14 previously identified loci (**Figures 3.10 & 3.11**). However, applying this filter, only 3 SNPs, mapping within the *ETS1* gene locus, met criteria for genome-wide significance ($P_{GWAS}<5x10^{-8}$). Furthermore these SNPs were only present within the Hap250k set, and indeed no good Hap300 tag SNP for this locus exists. This result is an example of the value of denser genome-wide marker sets in

the identification of common disease variants. Indeed, of 13 new coeliac risk regions with SNPs meeting criteria for genome-wide significance in the final (stage 1 and stage 2 combined) analysis, three loci (regions containing *PLEK, CCR4* and CD80) would not have met this criterion if genotyping had been restricted to Illumina Hap300 SNPs. Applying, the less stringent WTCCC 2007 $P<5 \times 10^{-7}$ threshold, 7 new loci were identified in the GWAS phase. These include loci containing the *TNFRSF14, CD247, MYNN, ZMIZ1, ETS1* and *YDJC* genes (**Table 3.10**).

395 SNPs, from 113 non-*HLA* loci met a lower significance criterion of $P_{GWAS}<10^{-4}$, representing an excess of associated SNPs compared to the null distribution (**Figures 3.10 & 3.11**). SNPs from these loci were chosen for follow-up genotyping in a further 7 independent sample collections comprising 4,918 celiac disease cases and 5,684 controls of European descent. Within genotyping platform and cost constraints, it was possible to design assays for 144 SNPs, available on a 144-plex Illumina custom SNP genotyping array (**Methods**). The aim was to test SNPs from each of 89 regions with $P_{GWAS}<5 \times 10^{-5}$, following the use of this threshold in a successful GWAS meta-analysis of similar size performed in Crohn's disease (Barrett, Hansoul et al. 2008). Additionally, selected loci with $5x10^{-5}< P_{GWAS} <10^{-4}$ were chosen for the remaining SNP assays on the 144-plex array, according primarily to the presence of immune genes within these regions. Two SNPs were selected per region for: regions with stronger association; regions with possible multiple independent associations; and/or containing genes of obvious biological interest.

Markers that passed design and genotyping quality control included:
a) 18 SNPs from all 14 previously identified celiac disease risk loci (including a tag SNP for the major celiac disease associated *HLA-DQ2.5cis* haplotype(van Heel, Franke et al. 2007))
b) 13 SNPs from all 7 novel regions with $P_{GWAS}<5x10^{-7}$
c) 86 SNPs from 59 of 68 novel regions with $5x10^{-7}> P_{GWAS} >5x10^{-5}$ in stage 1
d) 14 SNPs from 14 of 30 novel regions with $5x10^{-5}> P_{GWAS} >10^{-4}$ in stage 1.

3.4.3 Combined stage 1 and stage 2 association results

131 SNPs were successfully genotyped in the 7 follow-up sample collections. Genotype call rates were >99.9% in each collection. The regions most strongly associated with coeliac disease are presented in **Table 3.10**. Complete follow-up association results are shown in **Table 3.13**.

All 18 SNPs from 14 previously reported coeliac risk regions were associated with coeliac disease in the follow-up collections ($P_{Follow\text{-}up}$<0.01, **Table 3.10**) and all 14 regions met the criterion for genome-wide significance ($P_{combined}$< 5 x 10^{-8}). A further 35 SNPs from 23 of 80 (28.8%) putatively new coeliac risk regions also showed evidence of coeliac disease association in follow-up collections ($P_{follow\text{-}up}$<0.01, **Table 3.13**).

Primary association analyses of the combined GWAS and follow-up data were performed with a two-sided 2x2x12 Cochran-Mantel-Haenszel test. 13 new coeliac risk regions obtained overall genome-wide significant evidence ($P_{combined}$ <5x10^{-8}) of association, further supported by both evidence of association in stage 1 (P_{GWAS}<10^{-4}) and stage 2 ($P_{follow\text{-}up}$<0.01). These regions included those containing the *BACH2, CCR4*, *CD80*, *CIITA*/*SOCS1*/*CLEC16A, ETS1, ICOSLG*, *RUNX3, THEMIS, TNFRSF14*, and *ZMIZ1* genes which have known immunological function (**Table 3.10**). A further 13 regions met 'suggestive' criteria for association (either 10^{-6}> $P_{combined}$ >5x10^{-8} and/or P_{GWAS}<10^{-4} and $F_{ollow\text{-}up}$<0.01). These regions also contain multiple genes of obvious immunological function, including *CD247*, *FASLG*/*TNFSF18*/*TNFSF4*, *IRF4*, *TLR7*/*TLR8*, *TNFRSF9* and *YDJC*. Four of the 39 non-HLA regions show evidence for the presence of multiple independently associated variants in a conditional logistic regression analysis (**Table 3.12**).

40 SNPs with the strongest association (**Table 3.10**) from each of the known genome-wide significant, new genome-wide significant, and new suggestive loci, were tested for evidence of heterogeneity across the 12 collections studied. Only the HLA region was significant (Breslow-Day test *P*<0.05 / 40 tests, rs2187668 *P*=4.8x10^{-8}) which is consistent with the well described North-South gradient in HLA allele frequency in European populations, and more specifically for HLA-DQ in celiac disease(Karell, Louka et al. 2003).

There was no evidence of epistasis assessed by deviation from a model of multiplicative effects (*P*<10^{-4}) for any of the non-HLA SNPs (**Table 3.10**) meeting genome wide significance criteria.

Table 3.10 Genomic regions with the strongest association signals for coeliac disease

Chr	Position (bp)	SNP	LD block[ab] (Mb)	Min or allele	Minor allele freq[c]	P_{GWAS} 4533 cases, 10750 controls	$P_{follow-up}$ 4918 cases, 5684 controls	$P_{combined}$ 9451 cases, 16434 controls	Odds ratio[c] [95% CI]	Multiple independent association signals[d]	RefSeq Genes in LD block	Genes of Interest and GRAIL annotation[e]
Previously reported risk variants												
1	190803436	rs2816316	190.73-190.81	C	0.160	1.45×10^{-12}	1.56×10^{-6}	2.20×10^{-17}	0.80 [0.76-0.84]		1	*RGS1*
2	61040333	rs13003464	60.78-61.74	G	0.401	4.92×10^{-8}	1.57×10^{-6}	3.71×10^{-13}	1.15 [1.11-1.20]	yes	8	*REL, AHSA2*
2	102437000	rs917997	102.22-102.57	A	0.236	5.97×10^{-15}	7.83×10^{-4}	1.11×10^{-15}	1.19 [1.14-1.25]		5	*IL18RAP, IL18R1, IL1RL1, IL1RL2*
2	181704290	rs13010713	181.50-181.97	G	0.448	2.02×10^{-8}	3.21×10^{-4}	4.74×10^{-11}	1.13 [1.09-1.18]		1	*ITGA4, UBE2E3*
2	204510823	rs4675374	204.40-204.52	A	0.223	8.80×10^{-8}	4.94×10^{-3}	5.79×10^{-9}	1.14 [1.09-1.19]		2	*CTLA4, ICOS, CD28*
3	46210205	rs13098911	45.90-46.57	A	0.097	2.53×10^{-11}	1.96×10^{-7}	3.26×10^{-17}	1.30 [1.23-1.39]	yes	11	*CCR1, CCR2, CCRL2, CCR3, CCR5, CCR9*
3	161147744	rs17810546	161.07-161.23	G	0.125	4.56×10^{-18}	9.57×10^{-12}	3.98×10^{-28}	1.36 [1.29-1.44]	yes	1	*IL12A*
3	189595248	rs1464510	189.55-189.62	A	0.485	9.49×10^{-24}	3.63×10^{-18}	2.98×10^{-40}	1.29 [1.25-1.34]		1	*LPP*
4	123334952	rs13151961	123.19-123.78	G	0.142	6.31×10^{-18}	4.45×10^{-11}	2.18×10^{-27}	0.74 [0.70-0.78]		4	*IL2, IL21*
6	32713862	rs2187668	gene identified	A	0.258	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	6.23 [5.95-6.52]	(yes)	6	*HLA-DQA1, HLA-DQB1*
6	138014761	rs2327832	137.92-138.17	G	0.216	1.41×10^{-14}	1.97×10^{-6}	4.46×10^{-19}	1.23 [1.17-1.28]		0	*TNFAIP3*
6	159385965	rs1738074	159.24-159.45	A	0.434	3.14×10^{-8}	1.56×10^{-8}	2.94×10^{-15}	1.16 [1.12-1.21]		2	*TAGAP*
12	110492139	rs653178	110.19-111.51	G	0.495	6.03×10^{-14}	1.47×10^{-8}	7.15×10^{-21}	1.20 [1.15-1.24]		13	*SH2B3*
18	12799340	rs1893217	12.73-12.91	G	0.165	5.52×10^{-7}	1.04×10^{-4}	2.52×10^{-10}	1.17 [1.12-1.23]		1	*PTPN2*
New loci, genome-wide significant evidence ($P_{combined}$ <5 x 10^{-8})												
1	2516606	rs3748816	2.40-2.78	G	0.339	4.93×10^{-7}	1.17×10^{-3}	3.28×10^{-9}	0.89 [0.85-0.92]		4	*TNFRSF14, MMEL1*
1	25176163	rs10903122	25.11-25.18	A	0.480	3.21×10^{-5}	8.44×10^{-7}	1.73×10^{-10}	0.89 [0.85-0.92]		1	*RUNX3*

Table 3.10 (cont.)

1	199158760	rs296547	199.12-199.31	A	0.357	6.46×10^{-5}	1.34×10^{-5}	4.11×10^{-9}	0.89 [0.86-0.92]		2	*?*
2	68452459	rs17035378[f]	68.39-68.54	G	0.278	1.34×10^{-5}	1.41×10^{-4}	7.79×10^{-9}	0.88 [0.84-0.92]		2	*PLEK*
3	32990473	rs13314993[f]	32.90-33.06	C	0.464	6.87×10^{-6}	1.09×10^{-4}	3.27×10^{-9}	1.13 [1.08-1.17]		2	*CCR4*
3	120601486	rs11712165[f]	120.59-120.78	C	0.394	5.40×10^{-7}	1.72×10^{-3}	8.03×10^{-9}	1.13 [1.08-1.17]		5	*CD80, KTELC1*
6	90983333	rs10806425	90.86-91.10	A	0.397	9.46×10^{-6}	9.25×10^{-6}	3.89×10^{-10}	1.13 [1.09-1.17]		1	*BACH2, MAP3K7*
6	128320491	rs802734	127.99-128.38	G	0.311	1.36×10^{-6}	1.70×10^{-9}	2.62×10^{-14}	1.17 [1.12-1.22]	yes	2	*PTPRK, THEMIS*
8	129333771	rs9792269	129.21-129.37	G	0.238	8.14×10^{-6}	1.00×10^{-4}	3.28×10^{-9}	0.88 [0.84-0.91]		0	*?*
10	80728033	rs1250552	80.69-80.76	G	0.466	5.80×10^{-8}	1.81×10^{-3}	9.09×10^{-10}	0.89 [0.86-0.92]		1	*ZMIZ1*
11	127886184	rs11221332[f]	127.84-127.99	A	0.237	4.74×10^{-11}	9.98×10^{-7}	5.28×10^{-16}	1.21 [1.16-1.27]	yes	1	*ETS1*
16	11311394	rs12928822	11.22-11.39	A	0.161	1.07×10^{-5}	7.59×10^{-4}	3.12×10^{-8}	0.86 [0.82-0.91]		4	*CIITA, SOCS1, CLEC16A*
21	44471849	rs4819388	44.42-44.47	A	0.280	3.42×10^{-5}	1.66×10^{-5}	2.46×10^{-9}	0.88 [0.84-0.92]		2	*ICOSLG*
New loci, suggestive evidence (either A. $10^{-6}<P_{combined}<5x10^{-8}$ and/or B. P_{GWAS}<10-4 and $P_{follow-up}$<0.01)												
1	7969259	rs12727642	7.84-8.13	A	0.185	3.06×10^{-5}	8.21×10^{-4}	9.11×10^{-8}	1.14 [1.09-1.20]		4	*PARK7, TNFRSF9*
1	61564451	rs6691768	61.52-61.62	G	0.378	2.63×10^{-5}	1.16×10^{-3}	1.19×10^{-7}	0.90 [0.87-0.94]		1	*NFIA*
1	165678008	rs864537	165.43-165.71	G	0.391	1.01×10^{-7}	9.25×10^{-2}	3.80×10^{-7}	0.91 [0.87-0.94]		1	*CD247*
1	170977623	rs859637	170.87-171.20	A	0.486	8.15×10^{-5}	5.68×10^{-3}	1.75×10^{-6}	1.10 [1.06-1.14]		1	*FASLG, TNFSF18, TNFSF4*
3	69335589	rs6806528[f]	69.27-69.37	A	0.097	4.84×10^{-5}	7.66×10^{-4}	1.46×10^{-7}	1.19 [1.12-1.27]		1	*FRMD4B*
3	170974795	rs10936599	170.84-171.09	A	0.252	2.99×10^{-7}	6.63×10^{-2}	4.57×10^{-7}	1.12 [1.07-1.16]		3	*?*
6	328546	rs1033180[g]	0.32-0.40	A	0.080	9.14×10^{-6}	1.48×10^{-3}	5.58×10^{-8}	1.21 [1.13-1.29]	yes	1	*IRF4*[g]
7	37341035	rs6974491	37.32-37.41	A	0.170	1.37×10^{-5}	2.63×10^{-3}	1.56×10^{-7}	1.14 [1.09-1.20]		1	*ELMO1*
13	49733716	rs2762051	49.63-49.96	A	0.184	3.35×10^{-5}	5.06×10^{-3}	6.64×10^{-7}	1.13 [1.08-1.18]		0	*?*
14	68347957	rs4899260	68.24-68.39	A	0.263	4.55×10^{-5}	2.21×10^{-3}	3.92×10^{-7}	1.12 [1.07-1.16]		2	*ZFP36L1*
17	42220599	rs2074404	41.40-42.25	C	0.250	5.03×10^{-5}	5.96×10^{-3}	1.23×10^{-6}	0.90 [0.86-0.94]		10	*?*
22	20312892	rs2298428	20.14-20.35	A	0.201	2.49×10^{-7}	4.13×10^{-2}	1.84×10^{-7}	1.13 [1.08-1.19]		6	*UBE2L3, YDJC*
X	12881445	rs5979785	12.82-12.93	G	0.263	6.32×10^{-6}	2.18×10^{-3}	6.36×10^{-8}	0.88 [0.84-0.92]		1	*TLR7, TLR8*

[a]The most significantly associated SNP from each region is shown. [b]LD regions were defined by extending 0.1 cM to the left and right of the focal SNP as defined by the HapMap3 recombination map. All chromosomal positions are based on NCBI build-36 coordinates. [c]Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls. [d]Evidence from logistic regression at a genome-wide significant or suggestive level of significance after conditioning on other associated SNPs (see **Table 11**). *HLA* region not tested, but previously known. [e]Selected named genes within or adjacent to the same LD block as the associated SNPs, causality is not proven. In particular, other genes and other causal mechanisms may exist. Gene names underlined are identified from GRAIL(Raychaudhuri, Plenge et al. 2009; Raychaudhuri, Thomson et al. 2009) analysis

(**Methods**) with P_{text}<0.01. [f]These markers were present on the Hap550 but not Hap300 SNP sets, and are not genotyped for 737 cases and 2596 controls in the stage I GWAS, and combined dataset analyses. Only minor changes in *P* values were observed when these genotypes were imputed and included in analysis.
[g]An *IRF4* marker (rs9738805, r^2=0.08 with rs1033180 in HapMap CEU) was previously identified as a 'highly differentiated SNP' showing large allele frequency differences across Great Britain(Wellcome Trust Case Control Consortium 2007). Further analysis was therefore performed using 'ancestry-informative' SNP based principal components analysis (section 3.4.1.2.3 and **Table 10**) to adjust association statistics: uncorrected $P_{GWAS\text{-}Armitage}=7.06 \times 10^{-6}$ and principal components-adjusted $P_{GWAS\text{-}Eigenstrat}=2.28 \times 10^{-5}$. Definitive exclusion of population stratification would require family based association studies.

Table 3.11 GWAS (stage 1) associations after correction for first 10 principal components for 40 coeliac risk regions from table 9

Chr	Position (bp)	SNP	LD block (Mb)	Minor allele	Minor allele freq	$P_{GWAS\text{-}Armitage}$ [a] 4533 cases, 10750 controls	$P_{GWAS\text{-}Eigenstrat}$ [b] 4533 cases, 10750 controls	Principal components-adjusted correction factor [c]	Genes of Interest and GRAIL annotation
Previously reported risk variants									
1	190803436	rs2816316	190.73-190.81	C	0.160	1.42×10^{-12}	4.87×10^{-12}	-0.53	*RGS1*
2	61040333	rs13003464	60.78-61.74	G	0.401	4.93×10^{-08}	1.80×10^{-07}	-0.56	*REL, AHSA2*
2	102437000	rs917997	102.22-102.57	A	0.236	6.33×10^{-15}	1.47×10^{-14}	-0.36	*IL18RAP, IL18R1, IL1RL1, IL1RL2*
2	181704290	rs13010713	181.50-181.97	G	0.448	9.05×10^{-09}	5.43×10^{-09}	0.22	*ITGA4, UBE2E3*
2	204510823	rs4675374	204.40-204.52	A	0.223	3.37×10^{-08}	4.14×10^{-08}	-0.09	*CTLA4, ICOS, CD28*
3	46210205	rs13098911	45.90-46.57	A	0.097	2.49×10^{-11}	1.41×10^{-10}	-0.75	*CCR1, CCR2, CCRL2, CCR3, CCR5, CCR9*
3	161147744	rs17810546	161.07-161.23	G	0.125	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	*IL12A*
3	189595248	rs1464510	189.55-189.62	A	0.485	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	*LPP*
4	123334952	rs13151961	123.19-123.78	G	0.142	$<10^{-16}$	$<10^{-16}$	$<10^{-16}$	*IL2, IL21*
6	32713862	rs2187668	gene identified	A	0.258	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	*HLA-DQA1, HLA-DQB1*
6	138014761	rs2327832	137.92-138.17	G	0.216	8.99×10^{-15}	2.03×10^{-13}	-1.35	*TNFAIP3*
6	159385965	rs1738074	159.24-159.45	A	0.434	4.44×10^{-08}	3.49×10^{-09}	1.10	*TAGAP*
12	110492139	rs653178	110.19-111.51	G	0.495	4.30×10^{-14}	1.07×10^{-13}	-0.40	*SH2B3*
18	12799340	rs1893217	12.73-12.91	G	0.165	4.46×10^{-07}	1.46×10^{-07}	0.48	*PTPN2*
New loci, genome-wide significant evidence ($P_{combined}$ <5 x 10^{-8})									
1	2516606	rs3748816	2.40-2.78	G	0.339	2.61×10^{-07}	4.59×10^{-07}	-0.24	*TNFRSF14, MMEL1*
1	25176163	rs10903122	25.11-25.18	A	0.480	2.55×10^{-05}	5.59×10^{-05}	-0.34	*RUNX3*
1	199158760	rs296547	199.12-199.31	A	0.357	3.95×10^{-05}	2.85×10^{-05}	0.14	*?*
2	68452459	rs17035378 [f]	68.39-68.54	G	0.278	5.83×10^{-06}	1.10×10^{-05}	-0.28	*PLEK*
3	32990473	rs13314993 [f]	32.90-33.06	C	0.464	2.62×10^{-06}	7.83×10^{-06}	-0.48	*CCR4*
3	120601486	rs11712165 [f]	120.59-120.78	C	0.394	5.97×10^{-07}	5.60×10^{-07}	0.03	*CD80, KTELC1*
6	90983333	rs10806425	90.86-91.10	A	0.397	6.29×10^{-06}	4.96×10^{-06}	0.10	*BACH2, MAP3K7*

Table 3.11 (cont.)

6	128320491	rs802734	127.99-128.38	G	0.311	1.03×10^{-06}	9.58×10^{-07}	0.03	*PTPRK, THEMIS*
8	129333771	rs9792269	129.21-129.37	G	0.238	4.76×10^{-06}	1.35×10^{-05}	-0.45	*?*
10	80728033	rs1250552	80.69-80.76	G	0.466	3.50×10^{-08}	6.88×10^{-08}	-0.29	*ZMIZ1*
11	127886184	rs11221332[f]	127.84-127.99	A	0.237	4.11×10^{-11}	2.26×10^{-10}	-0.74	*ETS1*
16	11311394	rs12928822	11.22-11.39	A	0.161	5.98×10^{-06}	2.56×10^{-05}	-0.63	*CIITA, SOCS1, CLEC16A*
21	44471849	rs4819388	44.42-44.47	A	0.280	4.55×10^{-05}	3.66×10^{-05}	0.09	*ICOSLG*
New loci, suggestive evidence (either A. $10^{-6}<P_{combined}<5\times10^{-8}$ and/or B. $P_{GWAS}<10\text{-}4$ and $P_{follow\text{-}up}<0.01$)									
1	7969259	rs12727642	7.84-8.13	A	0.185	1.08×10^{-05}	5.81×10^{-06}	0.27	*PARK7, TNFRSF9*
1	61564451	rs6691768	61.52-61.62	G	0.378	1.33×10^{-05}	1.09×10^{-05}	0.08	*NFIA*
1	165678008	rs864537	165.43-165.71	G	0.391	3.77×10^{-08}	6.52×10^{-08}	-0.24	*CD247*
1	170977623	rs859637	170.87-171.20	A	0.486	2.77×10^{-05}	7.00×10^{-05}	-0.40	*FASLG, TNFSF18, TNFSF4*
3	69335589	rs6806528	69.27-69.37	A	0.097	2.82×10^{-05}	8.29×10^{-05}	-0.47	*FRMD4B*
3	170974795	rs10936599	170.84-171.09	A	0.252	2.47×10^{-07}	1.31×10^{-06}	-0.73	*?*
6	328546	rs1033180	0.32-0.40	A	0.080	7.06×10^{-06}	2.28×10^{-05}	-0.51	*IRF4*[g]
7	37341035	rs6974491	37.32-37.41	A	0.170	8.87×10^{-06}	1.12×10^{-05}	-0.10	*ELMO1*
13	49733716	rs2762051	49.63-49.96	A	0.184	1.89×10^{-05}	9.66×10^{-06}	0.29	*?*
14	68347957	rs4899260	68.24-68.39	A	0.263	2.31×10^{-05}	6.95×10^{-05}	-0.48	*ZFP36L1*
17	42220599	rs2074404	41.40-42.25	C	0.250	2.63×10^{-05}	1.18×10^{-05}	0.35	*?*
22	20312892	rs2298428	20.14-20.35	A	0.201	1.15×10^{-07}	4.62×10^{-07}	-0.60	*UBE2L3, YDJC*
X	12881445	rs5979785	12.82-12.93	G	0.263	6.10×10^{-06}	1.94×10^{-06}	0.50	*TLR7, TLR8*

SNPs, LD block definition and association category all defined as for **Table 3.2**.

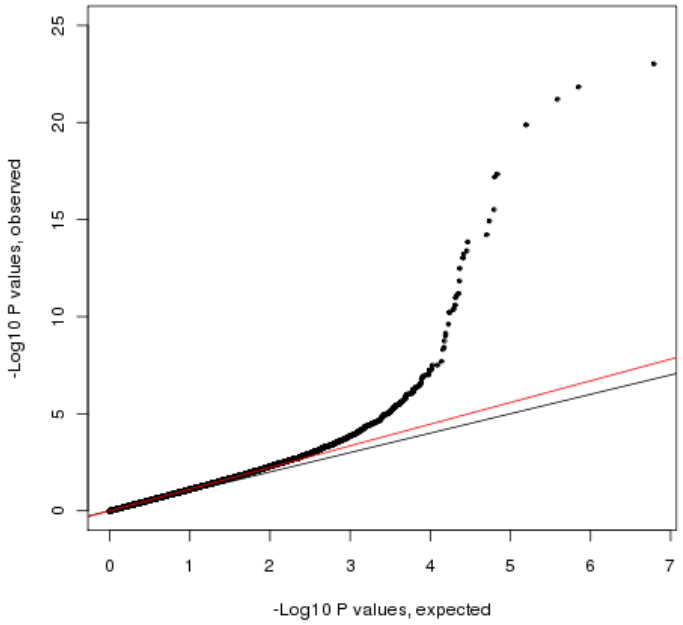
[a] $P_{GWAS\text{-}Armitage}$ calculated from Cochran-Armitage genotype trend test performed for each stage 1 sample collection and combined by weighted z score method – $P<10^{-16}$ is below limit of available *P* calculation from z scores

[b] $P_{GWAS\text{-}Eigenstrat}$ calculated from principal components-adjusted Cochran-Armitage genotype trend tests for each stage 1 sample collection, combined by weighted z score method

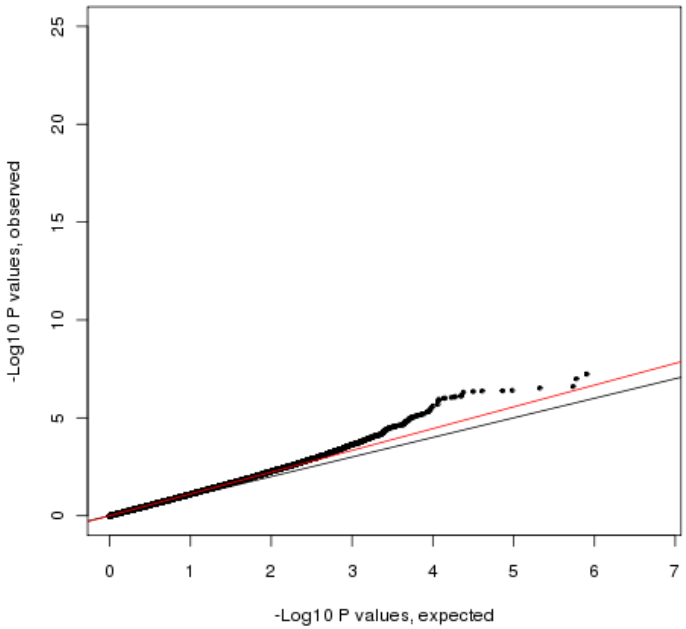
[c] Principal components-adjusted correction factor calculated as $-\text{Log10}(P_{GWAS\text{-}Eigenstrat}/P_{GWAS})$. Out of 39 non-HLA loci, 16 showed strengthened association and 23 showed weakened association after principal components adjustment

Figure 3.10 **Quantile-quantile plots of GWAS case-control association *P* values for "Hap300k" SNP marker set. A. Excluding SNPs from the *HLA* gene region. B. Excluding SNPs from 2 Megabase windows around each of the most strongly associated SNPs for 14 previously reported coeliac risk regions**

A.



B.

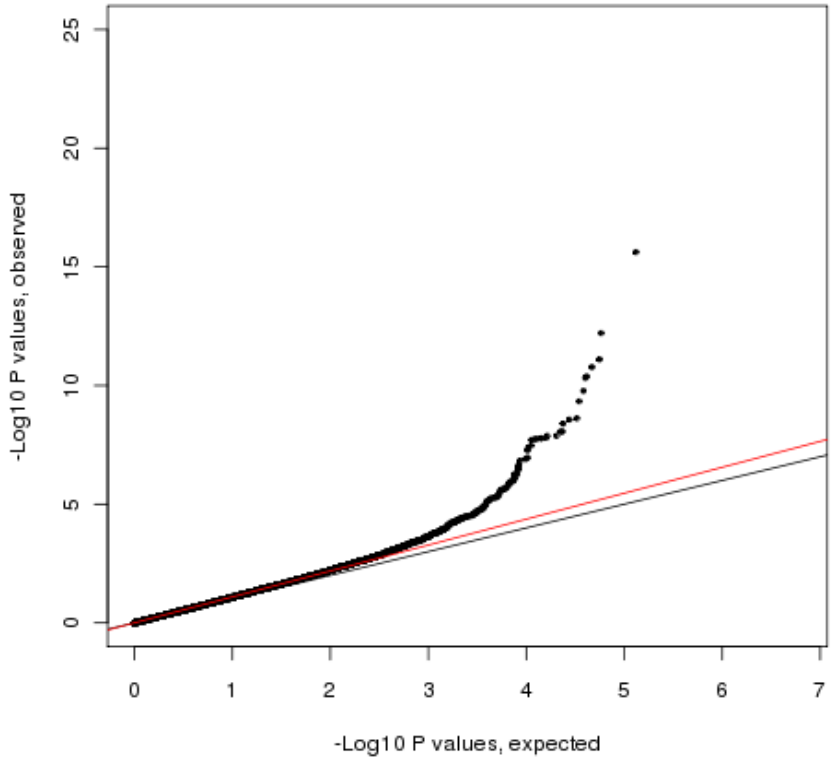


Black line indicates the null distribution.

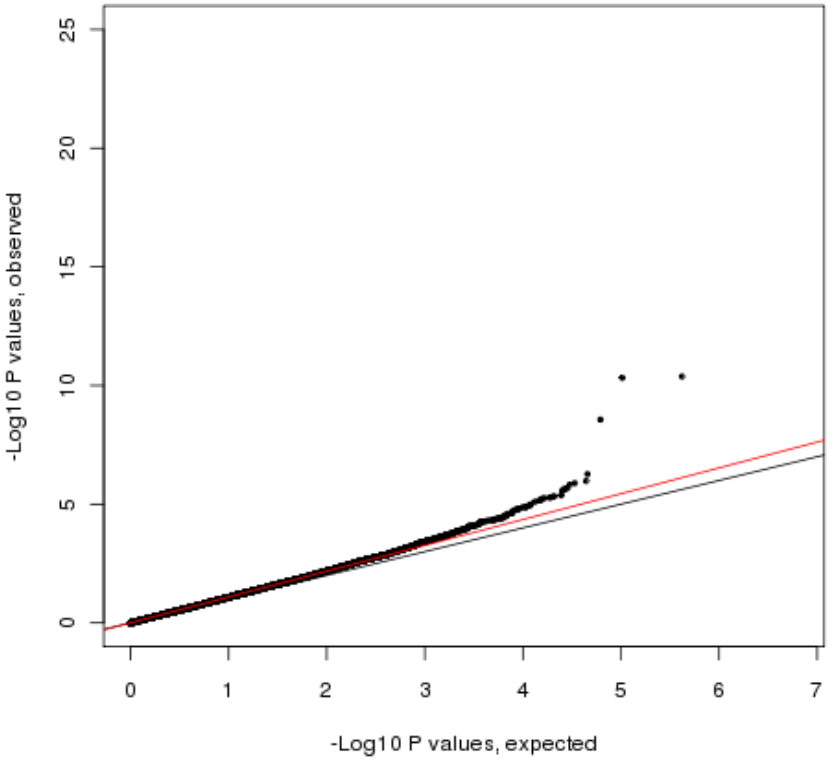
Red line indicates the overdispersion of test statistics corresponding to λ_{GC}

Figure 3.11 Quantile-quantile plots of GWAS case-control association *P* values for "Hap250k" SNP marker set. A. Excluding SNPs from the *HLA* gene region. B. Excluding SNPs from 2 Megabase windows around each of the most strongly associated SNPs for 14 previously reported coeliac risk regions.

A.



B.



3.4.3 Functional relatedness analysis

The boundaries of the genomic risk regions identified in the above analysis could not be defined precisely on the basis of the data obtained. Follow-up genotyping was restricted to one or two SNPs at each locus, so fine-mapping of the association signal in the combined dataset was not possible. However, since each true SNP association is expected to occur through linkage disequilibrium with one or more causal variants, regions can be broadly defined according to the boundaries of the linkage disequilibrium block within which the associated SNP resides. Following this, coeliac risk regions were defined by extending 0.1centiMorgan to the left and right of the most strongly associated SNP (LD region calculations performed by Jeff Barrett, Wellcome Trust Sanger Institute) (Dubois, Trynka et al. 2010). In most instances these boundaries corresponded to recombination hot spots in the genome, as defined in the HapMap Phase II data (Myers, Bottolo et al. 2005; Myers, Bowden et al. 2010). The mean size of genomic regions defined in this way was 281.1 kilobases (range 38.5Kb – 1.33Mb). The number of validated RefSeq genes mapping within these intervals ranged between 0 and 13 (mean =3).

Most coeliac risk regions contain genes with known immunological functions and some are clearly strong biological candidates for coeliac disease (e.g. *CD247, ICOSLG, CD80, TNFRSF14)*. In some regions multiple genes of possible immunological relevance are found (e.g. *SOCS1-CIITA, UBE2L3-YDJC*) and in others no obvious immune candidate genes exist (e.g. chr1-rs296547, chr8-rs9792269). To obtain more insight into the genes of relevance in these regions and to explore the functional relatedness of the coeliac loci, a statistical tool that utilizes text mining of PubMed abstracts (GRAIL) was used to annotate candidate genes from loci associated with common disease risk(Raychaudhuri, Plenge et al. 2009; Raychaudhuri, Thomson et al. 2009).

In the GRAIL analysis each of the 27 genome-wide significant coeliac disease loci (including *HLA-DQ*) was tested for functional relatedness to the other 26 regions using the other 26 regions as seed. GRAIL scores of P_{text}<0.01 (suggesting text-based functional relatedness of a gene within the locus to other coeliac loci genes – **Methods**) were obtained for 9 loci (33.3% sensitivity, **Table 3.10**). Factors that limit the sensitivity of GRAIL include biological pathways being both known (a 2006 dataset is used to avoid GWAS era studies), and published in the literature.

GRAIL analysis was then used, using the 27 known regions as a seed, to all 49 regions (49 SNPs) with $10^{-3} > P_{combined} > 5x10^{-8}$ and obtained GRAIL $P_{text} < 0.01$ for 9 regions (18.4%). As a control, only 5.5% (279 of 5033) of randomly selected Hap550 SNPs reached this threshold. In addition to the five 'suggestive' loci shown in **Table 3.2**, GRAIL annotated four further interesting gene regions of lower significance in the combined association results: rs944141/*PDCD1LG2* ($P_{combined}=4.4x10^{-6}$), rs976881/*TNFRSF8*($P_{combined}=2.1x10^{-4}$), rs4682103/*CD200*/*BTLA*($P_{combined}=6.8x10^{-6}$) and rs4919611/*NFKB2* ($P_{combined}=6.1x10^{-5}$). There appeared to be further enrichment for genes of immunological interest which are not GRAIL annotated in the $10^{-3}>P_{combined}>5x10^{-8}$ significance window, including rs3828599/*TNIP1* ($P_{combined}=1.55x10^{-4}$), rs8027604/*PTPN9* ($P_{combined}=1.4x10^{-6}$), rs944141/*CD274* ($P_{combined}=4.4x10^{-6}$). Some of these findings, for which neither genome-wide significant nor suggestive association is achieved, are likely to comprise part of a longer tail of disease predisposing common variants, of weaker effect sizes. Definitive assessment of these biologically plausible regions would require genotyping and association studies using much larger sample collections than the present study.

3.4.4 Autoimmune disease overlap

The extent to which new coeliac risk regions had been reported as risk regions in other autoimmune diseases was assessed. By searching two databases, 'A Catalog of Published Genome Wide Association Studies' (18 Nov 2009)(Hindorff, Sethupathy et al. 2009) and the HuGE database(Yu, Clyne et al. 2009) and using a threshold of $P<10^{-5}$ in other diseases, 18 of 27 genome-wide significant coeliac risk regions showed association with other autoimmune diseases. At only three of the 18 shared regions were associations across all diseases with the same SNP or a proxy SNP in $r^2>0.8$ in HapMap CEU. Nine regions appeared coeliac disease specific at the time of writing, including the regions containing rs296547 and rs9792269, and the regions around *CCR4*, *CD80, ITGA4*, *LPP*, *PLEK*, *RUNX3* and *THEMIS*. This may point to distinct coeliac pathogenetic factors in these regions. However, sharing with other autoimmune diseases is probably greater, due to both random variation in results between studies contingent on sample size limitations, and regions with a genuinely stronger effect size in one disease and weaker effect size in another.

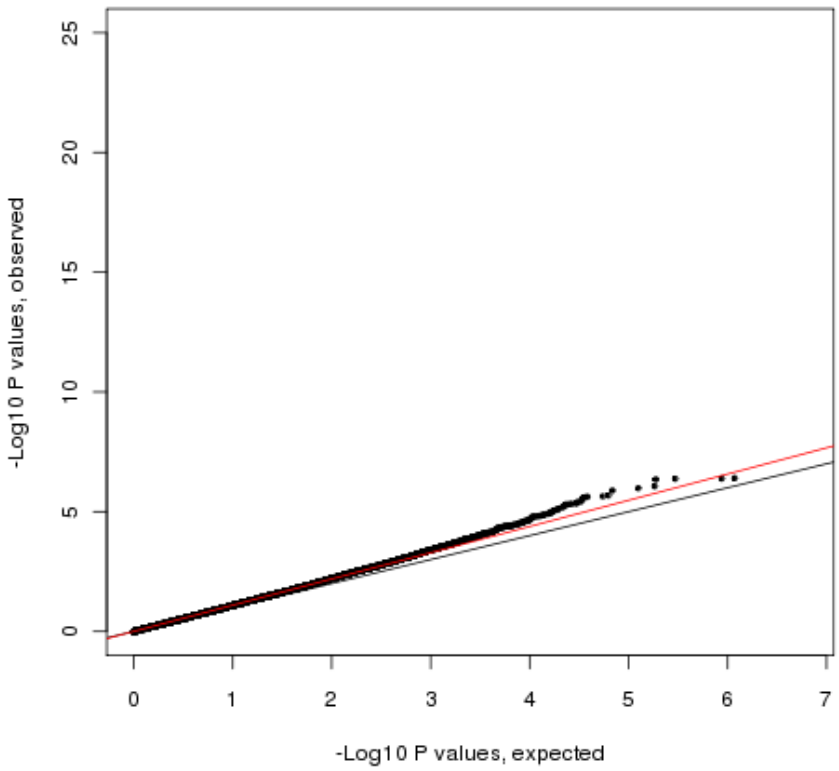
3.5 Discussion

3.5.1 Advancing understanding of the genetic architecture of coeliac risk

This study extends the number of convincingly identified coeliac risk regions from 14 to 27. In addition, there was strongly suggestive evidence for a substantial further tail of non-HLA risk regions. In the above analysis, a tier of the most convincing of these probable coeliac associations is highlighted ("suggestive evidence", **Table 3.10**). Within the HLA region, an analysis controlling for HLA-DQ association, suggested a further two possible new coeliac risk regions, one of which has been associated with type 1 diabetes. Neither of these regions obtained genome-wide significance and larger studies are planned, including denser fine-mapping of the region to help determine whether HLA-DQ independent HLA associations with coeliac disease are real. More sophisticated statistical methods to control for DQ coeliac association may also be useful to take advantage of the full case-control dataset: the analysis reported here excluded 96% of GWAS samples on the basis of non-homozygosity for DQ2.5*cis*.

After excluding all 40 convincing and "suggestive" loci, including the HLA, the distribution of association test statistics shows a small persisting excess of positive associations compared to the null distribution (**Figure 3.12**). This suggests additional coeliac risk loci among the long tail of positive associations. Complementary evidence for further loci is provided by the GRAIL analysis, which annotated an excess of loci in the $10^{-3} > P_{combined} > 5x10^{-8}$ range. Indeed applying a P_{text} threshold of <0.05, GRAIL annotates 201 loci obtaining $0.01 > P_{GWAS} > 10^{-4}$, with locus by locus inspection confirming obvious enrichment for immune pathway genes. Confirmation of disease association for a further swathe of these loci is feasible with further increases in sample size or combined analysis with similar large, but independent sample sets. To this end, SNPs from all loci obtaining $P_{GWAS} < 10^{-4}$, SNPs from 201 GRAIL-annotated loci obtaining $0.01 > P_{GWAS} > 10^{-4}$ and 19 SNPs from non-GRAIL annotated selected loci containing immune genes were submitted for SNP assay design and inclusion on the Illumina Immunochip, a collaboratively designed genotyping array with SNP content contributed from multiple immune disease investigators. This will enable relatively inexpensive follow-up genotyping of a large portion of the tail of positive associations in the GWAS. This experiment is currently underway, with genotyping extended to an additional roughly 5000 new Coeliac UK-recruited case samples.

Figure 3.12 **Quantile-quantile plot of GWAS case-control association *P* values for all SNPs ("Hap300k" and "Hap250k" SNP marker sets combined) after exclusion of SNPs from 2 Megabase regions around the most strongly associated SNP from each of 40 coeliac regions identified in the study (table 9)**



Despite these advances, it appears likely than some common genetic variants exert effects on risk too weak to be within the power of currently achievable association studies. In the current study, among genome-wide significant loci, odds ratios for non-HLA risk alleles varied between 1.12 and 1.36. Among newly identified risk alleles, odds ratios (including those in the suggestive evidence category) ranged between 1.10 and 1.21. As expected, there was under-representation of lower minor allele frequency SNPs (MAF < 0.10) among these associations (8.1% SNPs obtaining $P_{GWAS} < 10^{-4}$ had MAF < 0.10 compared to 17.5% SNPs included for analysis), likely reflecting lower power to detect these associations compared to higher MAF SNPs. Thus it is probable that the allelic risk spectrum in coeliac disease extends to include both further high frequency alleles with even lower odds ratios and lower frequency alleles with risk effect sizes extending across a broader range.

Together with previously reported variants, known variants can now explain around half of coeliac heritability (Dubois, Trynka et al. 2010). The 3% contributed by new SNP associations identified in this study is likely to be an underestimate since, in most cases, SNPs showing association in this study will not be perfectly correlated with the causal variants that drive the association signal detected in the GWAS. For example, the best tag *NOD2* SNP (rs17221417) in the WTCCC GWAS of seven common diseases observed a heterozygote odds ratio of only 1.29, whereas a *NOD2* causal variant (rs2066847) analysed in a large Crohn's disease meta-analysis was observed to have an allelic odds ratio of 3.99 (Wellcome Trust Case Control Consortium 2007; Barrett, Hansoul et al. 2008).

3.5.2 Function of Coeliac risk variants

The functional interpretation of complex disease association signals arising from GWASs is difficult. In the current study there were only four non-synonymous SNPs with evidence for coeliac disease association ($P_{GWAS}<10^{-4}$) from the 26 genome-wide significant associated non-*HLA* regions (rs3748816/*MMEL1*, rs3816281/*PLEK*, rs196432/*RUNX3*, rs3184504/*SH2B3*). Comprehensive regional resequencing is required to fully ascertain genetic variation in these regions and to test the possibility that coding variants contribute to the observed association signals. However, it is uncertain, even with this information, whether pure genetic association studies can resolve causal variants from a set of highly correlated candidates. Complementary approaches, that test intermediate phenotypes such as gene expression, protein expression or some aspect of gene function, offer an opportunity to compare disease association signals within a risk region, with genetic variants affecting the intermediate phenotype in the same interval (Barrett, Clayton et al. 2009; Dendrou, Plagnol et al. 2009). By exploiting knowledge of the local correlation structure of the genomic risk region, methods are being developed that test whether elimination of the genetic effect of a GWAS SNP lead to elimination of putative causal variant- intermediate phenotype correlations(Plagnol, Smyth et al. 2009; Nica, Montgomery et al. 2010). Teasing out these relationships is difficult and it appears that use of simple measures of LD (r-square and D') will often be inadequate in determining whether a local variant contributes to the association signal (Nica, Montgomery et al. 2010).

Effects of genetic variation on gene expression have been proposed as more likely to explain complex disease association signals than coding variants(Cookson, Liang et al. 2009). Indeed, an excess of *cis*-eQTLs has been reported mapping within complex disease gene loci, above that expected by chance (Nica, Montgomery et al. 2010). In order to investigate these effects in coeliac disease, a meta-analysis of expression quantitative trait loci datasets obtained from whole blood (PAXgene) samples was performed in collaboration with the GWAS. This work was performed by Dr Lude Franke (University Medical Center and Groningen University). In this analysis, Illumina Hap300 genotype data was available for 1,469 human whole blood samples comprising 7 sample collections, for which whole-genome gene expression array data was also available (Illumina Ref8 and HT12 arrays). For 34 of 39 coeliac loci, the top coeliac risk SNP had been genotyped (Hap300 SNP). For 4 of the 5 other loci, a proxy SNP (r-square > 0.5 in HapMap CEU) was available. For 6 loci showing evidence of a second independent association signal (**Table 3.12**), a second SNP was also assessed. In total, 44 SNPs from 38 loci were assessed for correlation with *cis* gene expression within a 1Mb window. After correction for multiple testing by controlling the false discovery rate at 5% (equating to Spearman rank correlation $P<0.0028$), SNPs from 20 of 38 (52.6%) non-HLA coeliac loci obtained significant eQTLs (**Table 3.14**). Since eQTL SNPs had a substantially higher average MAF than non-eQTL SNPs in the 294,767 Hap300 SNPs tested, 44 random SNPs of equal MAF distribution were selected to determine the frequency of eQTLs detected in random SNPs. This analysis found that coeliac SNPs were highly enriched for eQTLs (22 observed eQTL SNPs vs. 7.8 expected by chance eQTL SNPs; $P=9.3 \times 10^{-5}$, 10^6 permutations). Thus it appears that a significant proportion of coeliac risk variants influence coeliac disease susceptibility through a mechanism of altered gene expression. The eQTL analysis was extended to map the co-localization of the GWAS case-control association signal and eQTL signals in the region (**Figure 3.13**). Where the peak eQTL and case-control association signals are similar, it is more likely that the genetic association signal is caused by the observed regulation of gene expression (Plagnol, Smyth et al. 2009). This co-localisation was observed for a number of biologically plausible candidate genes including *CD247, IL18RAP* (previously reported (Hunt, Zhernakova et al. 2008)*), PARK7, PLEK, TAGAP* and *ZMIZ1.* However, in other instances co-localisation of eQTL and genetic association signals highlighted what was considered a less plausible candidate gene (rs3748816-*MMEL1* in a region containing *TNFRSF14,* rs4077924-*UBE2E3* in region containing *ITGA4*). In other regions, eQTL and genetic association signals showed poor co-localisation (e.g. *CCR3* gene region) suggesting that overlap of eQTL in this region might be co-incidental. Methods (such as those referred to above) that can fully account for the local LD structure in

these regions will need to be applied to further determine the relevance of eQTLs in each of these regions. It is currently unclear whether SNPs tagging coding variants would be more likely to fall within regions showing eQTLs, since, for example, they may be more likely to have greater proximity to genes and therefore regulatory regions. Thus comparison of eQTL frequencies in coeliac regions with random genomic SNPs may not control for this factor. Disentangling the contributions of coding variants from gene expression regulatory variants remains a big future challenge. Further genetic association analyses, informed by regional resequencing of coeliac regions, studies testing association of genetic variants with intermediate biological phenotypes and studies aiming to understand the function of regional genes in immune pathways relevant to coeliac disease will all be required.

Table 3.12 Coeliac risk loci with evidence of multiple independent associations

Locus	Chr	Pos_B36	LD block (Mb)	SNP	Pairwise SNP[a] r^2	Pairwise SNP[a] D'	$P_{combined}$ (CMH analysis)	$P_{combined}$[b] (logistic regression)	$P_{combined}$ conditioned on alternate SNP[c] (logistic regression)
Evidence of multiple independent SNPs at a locus ($p<10^{-3}$, both before and after conditioning on alternate SNP)									
CCR1/CCR3	3	46210205	45.90-46.57	rs13098911	0.135	0.781	3.26×10^{-17}	4.81×10^{-17}	5.71×10^{-09}
		46327388		rs6441961			2.93×10^{-15}	3.30×10^{-15}	2.97×10^{-07}
IL12A[d]	3	161147744	161.07-161.23	rs17810546	0.160	0.957	4.56×10^{-18}	4.29 x 10-^{18}	1.05×10^{-10}
		161179692		rs9811792			1.03×10^{-11}	9.59×10^{-12}	1.71×10^{-04}
IRF4	6	328546	0.32-0.40	rs1033180	0.079	0.880	5.58×10^{-08}	6.62×10^{-08}	2.05×10^{-05}
		356064		rs872071			8.22×10^{-07}	1.00×10^{-06}	3.05×10^{-04}
THEMIS / PTPRK	6	128320491	127.99-128.32	rs802734	0.077	0.993	2.62×10^{-14}	3.03×10^{-14}	6.60×10^{-10}
		128337195		rs7738609			2.88×10^{-10}	4.30×10^{-10}	7.78×10^{-06}

[a] pairwise LD estimated for 6785 individuals from UK 2 GWAS
[b] logistic regression performed on posterior genotype probabilities, with group membership included as a factorized covariate
[c] logistic regression perfomed as for b, conditioned on alternate SNP at locus
[d] no follow-up data for rs9811792, all SNP analyses confined to GWAS collections

Table 3.13 Association results for 131 SNPs from 94 genomic regions genotyped in stage 2

Locus number[a]	Chr	SNP	BP	Minor allele[b]	Minor allele frequency[b]	P_{GWAS}	$P_{Follow-up}$	$P_{Combined}$	OR	Gene(s) of interest
Previously reported coeliac risk regions										
1	1	rs2816316	190803436	C	0.171	1.45×10^{-12}	1.56×10^{-6}	2.20×10^{-17}	0.80	*RGS1*
2	2	rs842647	60972975	G	0.325	4.40×10^{-7}	7.97×10^{-3}	2.88×10^{-8}	0.89	*REL, PUS10*
2	2	rs13003464	61040333	C	0.385	4.92×10^{-8}	1.57×10^{-6}	3.71×10^{-13}	1.15	*REL, PUS10*
3	2	rs917997	102437000	A	0.222	5.97×10^{15}	7.83×10^{-4}	1.11×10^{-15}	1.19	*IL18RAP*
4	2	rs13010713	181704290	C	0.439	2.02×10^{-8}	3.21×10^{-4}	4.74×10^{-11}	1.13	*UBE2E3, ITGA4*
4	2	rs4667121	181758779	T	0.331	8.88×10^{-7}	5.95×10^{-3}	3.91×10^{-8}	0.89	*UBE2E3, ITGA4*
5	2	rs4675374	204510823	A	0.210	8.80×10^{-8}	4.94×10^{-3}	5.79×10^{-9}	1.14	*ICOS, CTLA4*
6	3	rs13098911	46210205	T	0.089	2.53×10^{-11}	1.96×10^{-7}	3.26×10^{-17}	1.30	*CCR1,CCR2, CCR5*
6	3	rs6441961	46327388	T	0.307	4.81×10^{-9}	1.18×10^{-7}	2.93×10^{-15}	1.17	*CCR3*
7	3	rs17810546	161147744	G	0.114	4.56×10^{-18}	9.57×10^{-12}	3.98×10^{-28}	1.36	*IL12A*
8	3	rs1464510	189595248	T	0.464	9.49×10^{-24}	3.63×10^{-18}	2.98×10^{-40}	1.29	*LPP*
9	4	rs13151961	123334952	C	0.156	6.31×10^{-18}	4.45×10^{-11}	2.18×10^{-27}	0.74	*IL2,IL21*
9	4	rs13119723	123437763	G	0.143	3.05×10^{-16}	3.75×10^{-10}	1.02×10^{-24}	0.74	*IL2,IL21*
10	6	rs2187668	32713862	T	0.134	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	6.23	*DQ2.5cis*
11	6	rs2327832	138014761	C	0.205	1.41×10^{-14}	1.97×10^{-6}	4.46×10^{-19}	1.23	*TNFAIP3*
12	6	rs1738074	159385965	T	0.423	3.14×10^{-8}	1.56×10^{-8}	2.94×10^{-15}	1.16	*TAGAP*
13	12	rs653178	110492139	C	0.476	6.03×10^{-14}	1.47×10^{-8}	7.15×10^{-21}	1.20	*SH2B3*
14	18	rs1893217	12799340	C	0.160	5.52×10^{-7}	1.04×10^{-4}	2.52×10^{-10}	1.17	*PTPN2*
New P_{GWAS} **<5x10-7**										
1	1	rs3748816	2516606	C	0.352	4.93×10^{-7}	1.17×10^{-3}	3.28×10^{-9}	0.89	*MMEL1, TNFRSF14*
1	1	rs3890745	2543484	G	0.331	8.16×10^{-7}	1.43×10^{-3}	6.40×10^{-9}	0.89	*MMEL1, TNFRSF14*
2	1	rs864537	165678008	C	0.398	1.01×10^{-7}	9.25×10^{-2}	3.80×10^{-7}	0.91	*CD247*
2	1	rs2056626	165687049	C	0.400	9.93×10^{-7}	6.85×10^{-2}	1.19×10^{-6}	0.91	*CD247*
3	3	rs10936599	170974795	A	0.246	2.99×10^{-7}	6.63×10^{-2}	4.57×10^{-7}	1.12	*MYNN*
3	3	rs1997392	170992346	T	0.267	1.20×10^{-5}	0.29	7.80×10^{-5}	1.09	*LRRC34*
4	10	rs1250552	80728033	C	0.482	5.80×10^{-8}	1.81×10^{-3}	9.09×10^{-10}	0.89	*ZMIZ1*
4	10	rs1250539	80707235	T	0.480	7.03×10^{-6}	9.44×10^{-3}	3.76×10^{-7}	0.91	*ZMIZ1, PPIF*
5	11	rs11221332	127886184	A	0.222	4.74×10^{-11}	9.98×10^{-7}	5.28×10^{-16}	1.21	*ETS1*
5	11	rs11221335	127891116	C	0.222	4.16×10^{-11}	2.27×10^{-6}	1.23×10^{-15}	1.21	*ETS1*
5	11	rs4245079	127926136	T	0.441	7.88×10^{-7}	1.49×10^{-4}	5.34×10^{-10}	0.89	*ETS1*
6	22	rs2298428	20312892	T	0.195	2.49×10^{-7}	4.13×10^{-2}	1.84×10^{-7}	1.13	*YDJC*
7	23	rs12687129	3659902	G	0.315	4.48×10^{-7}	0.65	4.98×10^{-5}	1.10	*PRKX*
P_{GWAS}**<5x10-5**										
1	1	rs12122754	4598419	A	0.175	4.94×10^{-6}	8.29×10^{-2}	4.79×10^{-2}	0.95	*AJAP1*
1	1	rs16839450	4637391	T	0.206	2.50×10^{-6}	1.55×10^{-2}	0.11	0.96	*AJAP1*

Table 3.13 (cont.)

2	1	rs12727642	7969259	T	0.182	3.06×10^{-5}	8.21×10^{-4}	9.11×10^{-8}	1.14	*PARK7, TNFRSF9*
3	1	rs976881	12156341	T	0.321	1.57×10^{-5}	0.36	2.05×10^{-4}	0.92	*TNFRSF1B, TNFRSF8, VPS13D*
3	1	rs6679088	12596858	C	0.223	1.36×10^{-5}	0.71	8.43×10^{-4}	0.92	*DHRS3*
3	1	rs3010928	12601328	A	0.235	1.61×10^{-5}	0.31	1.61×10^{-4}	0.91	*DHRS3*
4	1	rs195712	24691086	A	0.492	8.30×10^{-5}	3.59×10^{-2}	1.95×10^{-5}	0.92	*RCAN3(DSCR1L 2), NPAL3*
4	1	rs10903122	25176163	A	0.491	3.21×10^{-5}	8.44×10^{-7}	1.73×10^{-10}	0.89	*RUNX3, CLIC4*
5	1	rs12081664	60278068	A	0.116	1.36×10^{-5}	0.93	1.78×10^{-3}	1.10	*C1orf87*
6	1	rs10489912	61539729	A	0.429	2.98×10^{-5}	4.90×10^{-3}	6.16×10^{-7}	0.91	*NFIA*
6	1	rs6691768	61564451	G	0.385	2.63×10^{-5}	1.16×10^{-3}	1.19×10^{-7}	0.90	*NFIA*
7	1	rs2094219	91897051	G	0.025	4.66×10^{-5}	0.92	5.30×10^{-3}	1.19	*HSP90B3P*
8	1	rs17021444	104695375	C	0.156	4.63×10^{-6}	0.97	8.92×10^{-4}	0.91	*RP11-364B6.3, gene desert*
9	1	rs1772415	157269402	T	0.169	3.15×10^{-5}	4.87×10^{-2}	1.57×10^{-5}	1.12	*IFI16*
9	1	rs1061511	158456124	C	0.433	4.18×10^{-5}	0.45	1.73×10^{-2}	1.05	*PEA15, COPA, WDR42A,PEX20*
10	1	rs12041565	243839664	A	0.184	3.50×10^{-5}	0.43	4.68×10^{-4}	0.91	*KIF26B*
11	2	rs1355208	30298826	A	0.372	2.91×10^{-5}	9.49×10^{-2}	2.43×10^{-5}	1.09	*AC104698.2,YP EL5*
12	2	rs17035378	68452459	G	0.287	1.34×10^{-5}	1.41×10^{-4}	7.79×10^{-9}	0.88	*PLEK*
12	2	rs3816281	68461451	A	0.260	2.85×10^{-5}	1.05×10^{-3}	1.13×10^{-7}	0.89	*PLEK*
13	3	rs655754	32432202	T	0.384	6.64×10^{-5}	9.39×10^{-2}	4.10×10^{-5}	0.92	*CMTM7(CKLFSF 7)*
13	3	rs13314993	32990473	G	0.445	6.87×10^{-6}	1.09×10^{-4}	3.27×10^{-9}	1.13	*KRT5;CCR4*
13	3	rs12167	33013187	G	0.288	6.30×10^{-5}	5.74×10^{-3}	1.34×10^{-6}	0.90	*GLB1*
14	3	rs1050592	39281788	C	0.275	1.90×10^{-5}	0.44	1.37×10^{-2}	1.06	*CX3CR1, CCR8*
14	3	rs3732379	39282260	T	0.277	4.68×10^{-5}	0.45	1.22×10^{-2}	1.05	*CX3CR1, CCR8*
15	3	rs6806528	69335589	A	0.091	4.84×10^{-5}	7.66×10^{-4}	1.46×10^{-7}	1.19	*FRMD4B, UBE1C*
16	3	rs13081814	97031445	G	0.129	2.79×10^{-6}	4.67 x 1001	1.45×10^{-4}	1.12	*AC117432.4, EPHA6*
17	3	rs4682103	113538483	A	0.471	1.57×10^{-5}	5.39×10^{-2}	6.76×10^{-6}	1.09	*CD200, BTLA*
17	3	rs9842650	113552082	G	0.357	2.04×10^{-5}	0.20	8.78×10^{-5}	0.92	*CD200, BTLA*
18	3	rs11712165	120601486	G	0.386	5.40×10^{-7}	1.72×10^{-3}	8.03×10^{-9}	1.13	*CD80*
18	3	rs3755579	120697239	C	0.268	2.95×10^{-5}	3.62 x 1001	1.97×10^{-4}	1.08	*KTELC1, CD80*
18	3	rs1599796	120726624	A	0.198	7.82×10^{-6}	8.12×10^{-3}	3.42×10^{-7}	1.13	*CD80*
19	3	rs4679166	126118202	A	0.295	4.06×10^{-5}	1.93×10^{-2}	5.16×10^{-6}	0.90	*MUC13*
20	4	rs3774867	5805016	A	0.090	2.45×10^{-5}	0.51	7.47×10^{-4}	1.12	*EVC*
21	5	rs9324871	141535699	C	0.088	1.11×10^{-5}	0.89	1.22×10^{-3}	1.11	*NDFIP1, SPRY4, RNF14*
21	5	rs2961693	141544829	C	0.077	3.70×10^{-5}	0.37	1.16×10^{-2}	1.09	*NDFIP1, SPRY4, RNF14*
22	5	rs3828599	150381989	A	0.245	9.39×10^{-5}	0.18	1.55×10^{-4}	1.09	*GPX3*
22	5	rs2303038	150515206	T	0.221	1.10×10^{-5}	0.12	1.57×10^{-5}	1.10	*ANXA6*
23	6	rs1033180	328546	A	0.076	9.14×10^{-6}	1.48×10^{-3}	5.58×10^{-8}	1.21	*N/A*
23	6	rs872071	356064	A	0.489	7.22×10^{-5}	2.27×10^{-2}	8.22×10^{-7}	0.91	*IRF4*
23	6	rs1933650	694311	A	0.106	4.65×10^{-5}	0.29	1.56×10^{-4}	0.89	*EXOC2*
24	6	rs3734665	14244352	G	0.211	4.73×10^{-5}	0.74	8.55×10^{-3}	0.94	*CD83*
25	6	rs207270	90885603	G	0.446	1.65×10^{-5}	5.94×10^{-2}	1.19×10^{-5}	1.09	*BACH2, CX62, CASP8AP4*
25	6	rs10806425	90983333	A	0.385	9.46×10^{-6}	9.25×10^{-6}	3.89×10^{-10}	1.13	*BACH2, CX62, CASP8AP5*
26	6	rs9386829	110065962	C	0.234	8.04×10^{-5}	0.26	3.14×10^{-2}	0.95	*C6orf199*

Table 3.13 (cont.)

27	6	rs802734	128320491	G	0.301	1.36×10^{-6}	1.70×10^{-9}	2.62×10^{-14}	1.17	*PTPRK*
27	6	rs7738609	128337195	T	0.148	1.94×10^{-5}	2.51×10^{-6}	2.88×10^{-10}	0.84	*PTPRK*
28	6	rs9403998	149275688	C	0.041	4.09×10^{-5}	0.93	2.24×10^{-3}	0.85	*UST*
28	6	rs7761698	149483938	G	0.200	4.13×10^{-5}	0.79	4.19×10^{-3}	1.07	*RP11-365H23.1*
29	7	rs10215905	36769488	T	0.128	7.95×10^{-5}	0.12	1.13×10^{-4}	1.12	*AOAH*
29	7	rs6974491	37341035	A	0.165	1.37×10^{-5}	2.63×10^{-3}	1.56×10^{-7}	1.14	*ELMO1*
30	7	rs17664027	49899465	C	0.243	8.17×10^{-5}	0.42	7.83×10^{-4}	0.92	*VWC2*
31	7	rs874355	50205347	T	0.249	3.68×10^{-5}	0.28	1.52×10^{-4}	0.92	*AC020743.7*
32	8	rs10093096	65070255	C	0.431	3.56×10^{-5}	0.98	$2.^{-14} \times 10^{-3}$	1.06	*AC013492.5*
33	8	rs893225	97047924	A	0.141	4.86×10^{-6}	0.99	8.37×10^{-4}	0.91	*AC007992.12*
34	8	rs9792269	129333771	G	0.243	8.14×10^{-6}	1.00×10^{-4}	3.28×10^{-9}	0.88	*AC007860.6*
35	9	rs540909	263160	T	0.232	1.02×10^{-5}	8.17×10^{-2}	1.24×10^{-5}	1.11	*DOCK8, FOXD4*
36	9	rs944141	5480522	T	0.229	6.20×10^{-6}	6.80×10^{-2}	4.41×10^{-6}	1.11	*CD274*
37	9	rs10908921	91511109	C	0.260	2.28×10^{-6}	0.48	1.12×10^{-4}	1.09	*GADD45G, SEMA4D*
37	9	rs11265889	91532610	A	0.275	8.53×10^{-5}	1.00	3.38×10^{-3}	1.06	*GADD45G, SEMA4D*
38	9	rs4743150	99779945	A	0.222	1.48×10^{-5}	0.21	4.76×10^{-5}	0.91	*ANP32B,HEMGN, FOXE1*
38	9	rs874610	99822516	A	0.222	2.05×10^{-5}	0.25	7.79×10^{-5}	0.91	*ANP32B, HEMGN*
39	10	rs10823120	69390088	A	0.330	3.10×10^{-5}	0.88	2.57×10^{-3}	1.06	*HERC4*
40	10	rs4919611	103884929	G	0.112	2.29×10^{-5}	0.17	6.08×10^{-5}	1.12	*PPRC1, HPS6, LDB1, ELOVL4*
41	11	rs11043097	11092371	G	0.165	2.60×10^{-5}	0.81	9.96×10^{-4}	1.09	*GALNTL4*
42	11	rs1354329	16441877	T	0.303	8.72×10^{-7}	0.12	2.57×10^{-6}	1.10	*SOX6, SMAP, PLEKHA7*
42	11	rs11024021	16692067	A	0.416	2.37×10^{-5}	0.31	1.31×10^{-4}	1.08	*SOX6, SMAP, PLEKHA10*
43	11	rs224619	32021832	T	0.486	4.55×10^{-5}	0.18	8.62×10^{-5}	1.08	*AL078612.8*
44	11	rs12788589	132881660	C	0.183	4.37×10^{-5}	0.99	3.63×10^{-3}	0.93	*OPCML*
45	13	rs9316483	21079334	T	0.276	4.39×10^{-5}	0.49	6.36×10^{-4}	0.92	*EFHA1*
46	13	rs2305100	43346934	T	0.100	3.96×10^{-5}	0.73	6.52×10^{-3}	0.92	*CCDC122*
47	13	rs2762051	49733716	A	0.179	3.35×10^{-5}	5.06×10^{-3}	6.64×10^{-7}	1.13	*RP11-480P3.1*
48	14	rs11851414	68329255	G	0.218	5.53×10^{-6}	8.51×10^{-3}	4.40×10^{-7}	1.13	*ZFP36L1, ACTN1*
48	14	rs4899260	68347957	T	0.252	4.55×10^{-5}	2.21×10^{-3}	3.92×10^{-7}	1.12	*C14orf181*
49	14	rs1667515	84763657	C	0.365	8.71×10^{-6}	0.94	8.06×10^{-4}	0.94	*AL357172.3*
50	15	rs4411464	61782476	G	0.138	9.34×10^{-6}	0.71	4.37×10^{-4}	0.90	*HERC1, USP3*
51	15	rs8027604	73234826	C	0.350	4.70×10^{-6}	3.24×10^{-2}	1.37×10^{-6}	1.10	*AC113208.5*
52	16	rs6498114	10871619	G	0.257	9.41×10^{-5}	1.95×10^{-2}	1.02×10^{-5}	1.11	*CIITA*
53	16	rs12928822	11311394	A	0.171	1.07×10^{-5}	7.59×10^{-4}	3.12×10^{-8}	0.86	*SOCS1*
53	16	rs12927773	11311464	A	0.168	2.53×10^{-5}	9.43×10^{-4}	9.78×10^{-8}	0.86	*SOCS1*
54	17	rs11078559	5203473	T	0.484	3.82×10^{-5}	0.18	7.27×10^{-5}	1.08	*RABEP1*
55	18	rs1394466	49159204	A	0.461	2.47×10^{-6}	8.48×10^{-2}	1.81×10^{-2}	1.05	*DCC*
56	20	rs1535253	19630909	A	0.349	2.11×10^{-6}	0.42	7.54×10^{-5}	0.92	*SLC24A3, RIN2*
57	21	rs2822590	14633438	G	0.151	2.42×10^{-5}	0.16	3.44×10^{-2}	1.06	*ABCC13*
58	21	rs4819388	44471849	T	0.284	3.42×10^{-5}	1.66×10^{-5}	2.46×10^{-9}	0.88	*ICOSLG*
59	23	rs1947953	12877811	C	0.298	1.08×10^{-5}	2.21×10^{-3}	1.05×10^{-7}	0.88	*TMSB4X;TMSL2;TMSL1, TLR8*
59	23	rs5979785	12881445	G	0.270	6.32×10^{-6}	2.18×10^{-3}	6.36×10^{-8}	0.88	*TMSB4X;TMSL2;TMSL1, TLR8*
P_{GWAS}<10-4										
1	1	rs6684553	58775554	G	0.096	7.28×10^{-5}	0.91	5.35×10^{-3}	0.91	*OMA1, TACSTD2, JUN*

Table 3.13 (cont.)

2	1	rs859637	170977623	A	0.484	8.15×10^{-5}	5.68×10^{-3}	1.75×10^{-6}	1.10	*FASLG*
3	1	rs2274065	181826327	C	0.081	9.73×10^{-5}	0.18	1.30×10^{-4}	0.87	*NCF2;p67-PHOX*
4	1	rs296547	199158760	A	0.367	6.46×10^{-5}	1.34×10^{-5}	4.11×10^{-9}	0.89	*C1orf106*
5	2	rs1429248	155060456	T	0.153	9.94×10^{-5}	4.03×10^{-2}	2.12×10^{-5}	0.89	*GALNT13*
6	5	rs1020388	55595784	G	0.474	7.23×10^{-5}	0.65	1.11×10^{-3}	0.94	*AC016638.9, ANKRD55*
7	6	rs4446534	92886209	A	0.358	7.97×10^{-5}	0.12	6.18×10^{-2}	1.04	*AL590814.5*
8	10	rs1539234	6316749	G	0.420	8.89×10^{-5}	5.38×10^{-2}	2.61×10^{-5}	1.08	*PFKFB3, IL2RA*
9	10	rs10857580	49356390	A	0.098	8.67×10^{-5}	0.16	1.36×10^{-4}	0.88	*ARHGAP22, MAPK8*
10	12	rs10466829	9767358	C	0.487	8.78×10^{-5}	0.31	3.95×10^{-2}	1.04	*CLECL1*
11	12	rs988606	66767678	A	0.220	8.37×10^{-5}	0.59	1.66×10^{-3}	1.08	*IFNG, IL26, IL22*
12	16	rs477639	87517011	A	0.320	5.41×10^{-5}	0.53	9.37×10^{-4}	1.07	*CBFA2T3*
13	17	rs2074404	42220599	G	0.252	5.03×10^{-5}	5.96×10^{-3}	1.23×10^{-6}	0.90	*WNT3*
14	18	rs3809983	54353748	G	0.488	8.71×10^{-5}	0.70	7.86×10^{-3}	1.05	*ALPK2, MALT1*

[a]Regions numbered within each GWAS association category, as defined for follow-up SNP selection

[b]Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls.

Table 3.14 **Coeliac risk variants correlated with *cis* gene expression (adapted from Dubois et al., 2010). Analysis performed by Lude Franke, University of Groningen**

SNP[a]	Chr	SNP position[b]	Probe Centre Position[b]	Illumina ArrayAddressID	Expression dataset[c]	Gene name	eQTL *P* value[d]
Loci with genome-wide significant evidence ($P_{combined}$ <5 x 10^{-8})							
rs3748816	1	2516606	2412221	650452	HT-12	*PLCH2*	1.66 x 10^{-5}
rs3748816	1	2516606	2482955	6520725	Ref-8v2 + HT-12	*TNFRSF14*	1.30 x 10^{-3}
rs3748816	1	2516606	2510429	6250338	Ref-8v2	*C1orf93*	1.16 x 10^{-4}
rs3748816	1	2516606	2533115	2070246	Ref-8v2 + HT-12	*MMEL1*	1.03 x 10^{-20}
rs296547	1	199158760	198880146	1300279	Ref-8v2 + HT-12	*DDX59*	2.45 x 10^{-5}
rs842647	2	60972975	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	3.30 x 10^{-10}
rs13003464[e]	2	61040333	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	6.39 x 10^{-11}
rs3816281[f]	2	68461451	68461957	4810020	Ref-8v2 + HT-12	*PLEK*	7.97 x 10^{-26}
rs917997	2	102437000	102418571	6520180	Ref-8v2 + HT-12	*IL18RAP*	7.35 x 10^{-87}
rs13010713	2	181704290	181593865	1780433	HT-12	*UBE2E3*	4.93 x 10^{-5}
rs13098911	3	46210205	45964449	6550333	Ref-8v2 + HT-12	*CXCR6*	9.66 x 10^{-6}
rs13098911	3	46210205	46255176[g]	2190671	HT-12	*CCR3*	5.50 x 10^{-10}
rs13098911	3	46210205	46255176[g]	7570670	Ref-8v2	*CCR3*	5.69 x 10^{-4}
rs6441961[d]	3	46327388	46255176[h]	2190671	HT-12	*CCR3*	2.87 x 10^{-19}
rs6441961[d]	3	46327388	46255176[h]	7570670	Ref-8v2	*CCR3*	1.02 x 10^{-4}
rs11922594[f]	3	120608512	120683364[i]	6550288	Ref-8v2 + HT-12	*KTELC1*	5.09 x 10^{-17}
rs11922594[f]	3	120608512	120683364[i]	3850161	Ref-8v2 + HT-12	*KTELC1*	7.34 x 10^{-6}
rs10806425	6	90983333	90878075	3520349	HT-12	*BACH2*	1.92 x 10^{-3}
rs1738074	6	159385965	159380068	5890739	Ref-8v2 + HT-12	*TAGAP*	1.99 x 10^{-3}
rs1738074	6	159385965	159381094[j]	5360364	HT-12	*TAGAP*	3.23 x 10^{-4}

Table 3.14 (cont.)

rs1738074	6	159385965	159381094 [j]	4860242	HT-12	*TAGAP*	2.18×10^{-3}
rs1250552	10	80728033	80622540	2450131	Ref-8v2 + HT-12	*ZMIZ1*	1.80×10^{-3}
rs653178	12	110492139	110399552	6560301	Ref-8v2 + HT-12	*SH2B3*	9.24×10^{-12}
rs653178	12	110492139	110710447	840253	Ref-8v2 + HT-12	*ALDH2*	1.44×10^{-4}
rs653178	12	110492139	110894406 [k]	2070736	HT-12	*TMEM116*	3.68×10^{-4}
rs653178	12	110492139	110894406 [k]	3190129	Ref-8v2	*TMEM116*	1.51×10^{-3}
rs12928822	16	11311394	11335627	4540072	Ref-8v2 + HT-12	*C16orf75*	1.02×10^{-8}
rs4819388	21	44471849	44049567	7200373	Ref-8v2	*RRP1*	2.62×10^{-3}
Loci with suggestive evidence (either A. $10^{-6}<P_{combined}<5\times10^{-8}$ and/or B. $P_{GWAS}<10\text{-}4$ and $P_{follow\text{-}up}<0.01$)							
rs12727642	1	7969259	7956138	610193	Ref-8v2 + HT-12	*PARK7*	9.76×10^{-15}
rs864537	1	165678008	165710482 [l]	6290400	Ref-8v2 + HT-12	*CD247*	1.77×10^{-9}
rs864537	1	165678008	165710482 [l]	3890689	HT-12	*CD247*	2.93×10^{-7}
rs6974491	7	37341035	37157761	2750154	Ref-8v2 + HT-12	*ELMO1*	5.40×10^{-6}
rs2074404	17	42220599	41824345	3520672	Ref-8v2 + HT-12	*LRRC37A*	1.17×10^{-4}
rs2074404	17	42220599	42106695 [m]	5260138	Ref-8v2 + HT-12	*NSF*	1.20×10^{-5}
rs2074404	17	42220599	42106695 [m]	1410484	HT-12	*NSF*	4.28×10^{-4}
rs2074404	17	42220599	42223012	4070615	HT-12	*WNT3*	2.77×10^{-3}
rs2074404	17	42220599	42485154	4880037	HT-12	*LOC388397*	1.78×10^{-9}
rs2298428	22	20312892	20308188	1230242	Ref-8v2 + HT-12	*UBE2L3*	1.96×10^{-90}
rs5979785	X	12881445	12842944 [n]	6480360	Ref-8v2 + HT-12	*TLR8*	3.88×10^{-13}
rs5979785	X	12881445	12842944 [n]	3390612	Ref-8v2 + HT-12	*TLR8*	1.07×10^{-7}

[a]The SNP with the strongest association from 34 of 39 non-HLA loci ($P_{combined}<10^{-6}$, **Table 3.2**), Hap300 proxy SNPs for 4 further loci, and a second independently associated SNP from 6 loci, were tested for correlation with gene expression in PAXgene blood RNA in up to 1,349 individuals. 1 locus (containing *ETS1)* where an adequate proxy SNP was not available was not included for the eQTL analysis. SNP-gene expression correlations were tested for probes within a 1Mb window. Results are presented for SNPs showing significant correlations with *cis* gene expression after controlling false discovery rate at 5% (corresponding to $P<0.0028$). [b]All chromosomal positions are based on NCBI build-36 coordinates. Probe centre position was determined by re-mapping probe sequences to the human transcriptome and calculated from the mid-point of the transcript start and transcript end positions in genomic co-ordinates.

[c]'HT-12' comprise 1240 individuals with blood gene expression assayed using Illumina Human HT-12v3 arrays, 'Ref-8v2' comprise 229 individuals with blood gene expression assayed using Illumina Human-Ref-8v2 arrays. [d]Spearman rank correlation of genotype and residual variance in transcript expression. Meta-analysis eQTL *P* value shown if both datasets had identical probes. [e]Second, independently associated SNP from this locus. [f]Proxy SNP, r^2=0.61 in HapMap CEU with most associated SNP rs11712165.
[g,h,i,j,k,l,m,n]Different Illumina probe sequences with the same Probe Centre Position.

3.5.2.1 Function of coeliac loci candidate genes

As with previously reported coeliac risk loci, new loci identified in the GWAS mostly contain genes with known immune function. However, the larger number of immune genes now implicated is enabling the definition of specific pathways in the immune system that alter coeliac disease susceptibility.

3.5.2.1.1 T and B cell co-stimulation/ co-inhibition

Immunological studies suggest that coeliac disease is a T_{h1} –driven disorder, in which gluten peptides presented on DQ2 or DQ8 heterodimers activate $CD4^+$ T cells. The importance of *HLA-DQA1* and *HLA-DQB1* gene variants is well understood. The current study adds to this understanding, highlighting the role of T and B cell co-stimulatory molecules (*CTLA4/ICOS/CD28, TNFRSF14, CD80, ICOSLG, TNFRSF9, TNFSF4*) in addition to *CD247*, encoding the zeta subunit of the T cell receptor. Since DQ2 is carried by 30% of the population, and is therefore not sufficient to explain gluten toxicity, a model in which the threshold for T cell activation by gluten peptides depends also on inherited variation in T cell co-stimulatory molecules is attractive.

3.5.2.1.2 T cell development in the thymus

The thymus is implicated in coeliac disease pathogenesis by association signals mapping to 4 regions containing genes with prominent known roles in T cell development in the thymus (*THEMIS, RUNX3, ETS1, and TNFRSF14*). The thymus gland's key role in establishing tolerance to self-antigens through thymocyte selection makes it a prime candidate for involvement in autoimmune disease pathogenesis, but it has not previously been implicated in coeliac disease. Coeliac disease onset occurs at all ages, but the disposition appears to be present from childhood, since the prevalence in children is as high as in adults(van Heel and West 2006). Thymic T cell development and output to the periphery is most prodigious before adolescence, and therefore selection events at this stage are strong candidates for determining loss of tolerance to gluten early in life. Indeed, exogenous antigen presentation and selection can occur in the thymus via migratory dendritic cells – this has been demonstrated for skin and hypothesized for food antigens (Bonasio, Scimone et al. 2006; Klein, Hinterberger et al. 2009). In type 1 diabetes, disease associated genetic variation in the

insulin gene *INS* causes altered thymic insulin expression and subsequent T cell tolerance for insulin as a self-protein (Vafiadis, Bennett et al. 1997).

The rs802734 LD block contains the recently identified gene *THEMIS* 'THymus-Expressed Molecule Involved in Selection'. This gene shows relatively selective expression in the thymus, especially in immature double positive thymocytes, compared to peripheral lymphoid tissues and is not expressed at all in non-lymphoid tissues (Fu, Vallee et al. 2009; Patrick, Oda et al. 2009). *THEMIS* plays a key role in T-cell selection during late thymocyte development, such that its elimination leads to profound defects in T cell development (Fu, Vallee et al. 2009; Johnson, Aravind et al. 2009; Lesourne, Uehara et al. 2009; Patrick, Oda et al. 2009). The most prominent of these defects is impaired differentiation of both $CD4^+$ and $CD8^+$ T cells, through impaired positive selection. The role of *THEMIS* appears contingent on T cell receptor signalling, although T cell receptor stimulation can reverse some of the defects observed (Lesourne, Uehara et al. 2009). *RUNX3,* mapping within the rs10903122 LD block has been proposed to play a role as a master regulator of $CD8^+$ T lymphocyte development in the thymus since it potentiates CD8 and abrogates CD4 expression through binding to the CD8 gene enhancer and CD4 silencer regions, respectively (Woolf, Xiao et al. 2003; Sato, Ohno et al. 2005). *ETS1* (rs11221332 LD block) has also been shown to play a key role in thymic $CD8^+$ lineage differentiation in part through promoting *RUNX3* expression (Zamisch, Tian et al. 2009). Finally, *TNFRSF14* (LIGHTR, rs3748816 LD block) has a critical role in promoting thymocyte apoptosis in response to HLA/self-peptide TCR interactions (negative thymocyte selection) *(Wang and Fu 2003).*

Distinguishing the relative importance of these genes in thymocyte development versus peripheral leucocyte function, where all of these genes have also been shown to function, will require further immunological studies. Of the genes highlighted, *THEMIS* is most selectively thymus-expressed and clearly plays a critical role in thymocyte selection. *RUNX3* and *ETS1* are transcription factors that control a variety mature T cell processes in addition to their thymocyte functions ,such as regulation of expression of T cell receptor alpha and beta (*ETS1* (Ho, Bhat et al. 1990; Wotton, Prosser et al. 1993)) and co-stimulatory molecules (Arman, Aguilera-Montilla et al. 2009). *TNFRSF14* has well-defined roles in T cell co-stimulation and co-inhibition (see 3.5.2.1.1). Nevertheless, together these findings have suggested new pathways, particularly involved in thymocyte development and selection, with important roles in autoimmune disease pathogenesis.

3.5.2.1.3 Innate immune detection of viral RNA.

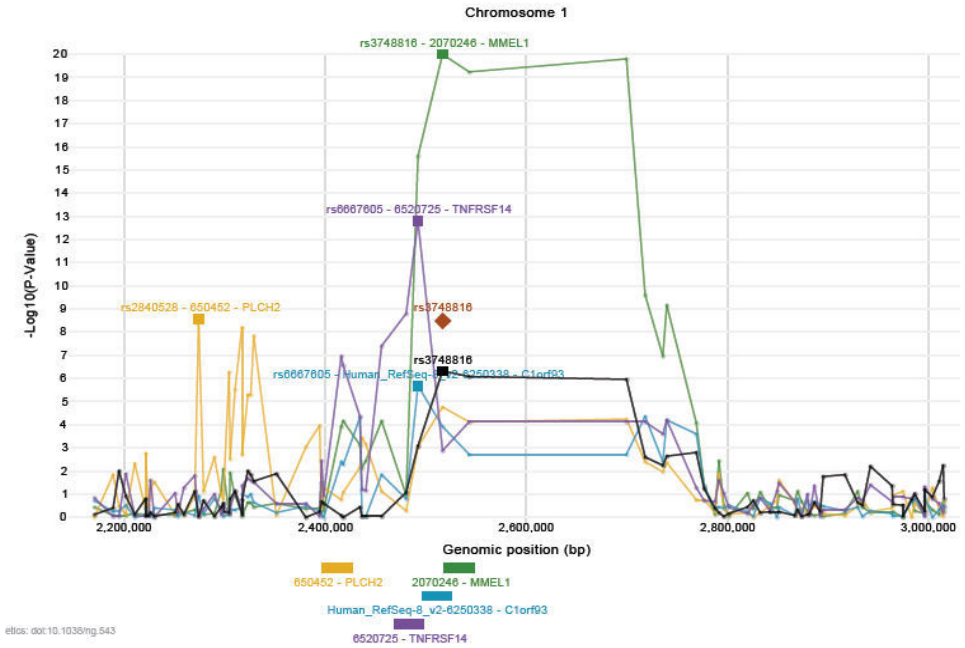
TLR7 and *TLR8* map within a coeliac risk region tagged by rs5979785 ($P_{combined}=6.36x10^{-8}$). Both these genes encode endosomal toll-like receptors that are activated by viral single stranded RNA (Diebold, Kaisho et al. 2004; Krieg and Vollmer 2007). The GRAIL analysis annotates *TLR7* whereas as *TLR8*'s relevance to the association signal is supported by a strong cis effect on gene expression of rs5979785 in whole blood (Dubois, Trynka et al. 2010). At face value these associations suggest that TLR responses to viral infection may play a role in coeliac pathogenesis and provide some support for epidemiological evidence that viral infections are a pre-disposing environmental trigger (Sandberg-Bennich, Dahlquist et al. 2002; Ivarsson, Hernell et al. 2003; Ivarsson 2005) . The recent observation of rare loss of function mutations in the enteroviral response gene IFIH1 in type 1 diabetes, provide additional support for the role of viral infection (and the nature of the host response to infection) as a putative environmental trigger common to these autoimmune diseases (Nejentsev, Walker et al. 2009). However, an alternative explanation for *TLR7/TLR8* gene involvement in coeliac disease is that genetic variation influencing these molecules alters the usually muted TLR7 or TLR8 responses to host nucleic acids, driving autoimmunity (Krieg 2002; Leadbetter, Rifkin et al. 2002; Viglianti, Lau et al. 2003).

3.5.2.1.4 Cytokines, chemokines and their receptors

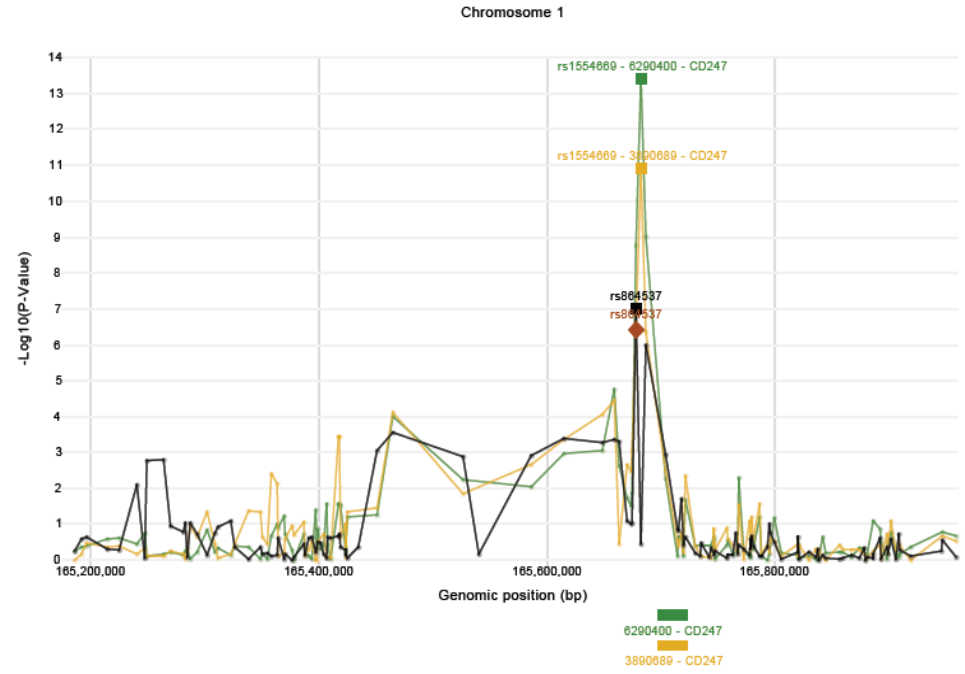
Genes in this category include the previously reported 2q11-12 interleukin receptor cluster (*IL18RAP,IL18R1,iL1R1,IL1R2,IL1RL1,IL1RL2),* the 3p21 chemokine receptor cluster *(CCR3,CCR5 etc),* the *IL2-IL21* region and *IL12A* (Hunt, Zhernakova et al. 2008). Loci containing *CCR4* and the cytokine *TNFRSF18* can now be added to this. These genes, together with *ITGA4*, which encodes the integrin alpha 4 subunit expressed on gut-homing T cells as one half of the α4β7 integrin, suggest variation affecting recruitment of immune cells to and within the intestinal mucosa is important in coeliac disease.

Figure 3.13 **Co-localization of case-control association and genotype-expression correlation (eQTL) signals within coeliac risk regions. A. Region containing *MMEL1* and *TNFRSF14.* B. Region containing *CD247.* Figure created by Lude Franke, reproduced from Dubois et al. (Dubois, Trynka et al. 2010)**

A.



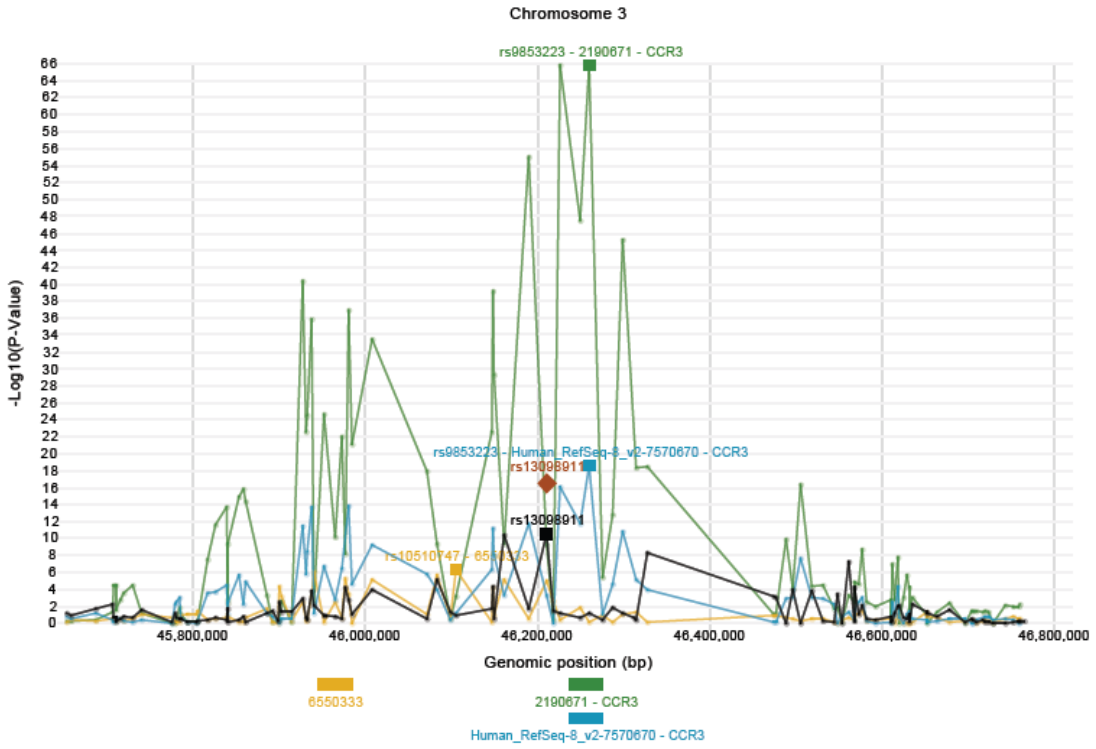
B.



P_{GWAS} values (black points and black line) from 4,533 cases and 10,750 controls. $P_{combined}$ values (red points) from 9,451 cases and 16,434 controls. Genotype-expression correlation P values for SNP positions across the genome

at each tested eQTL are shown in a different colour for each probe (annotated with Illumina ArrayAddressID and gene name).
Figures made by Lude Franke and reproduced from Dubois et al. (Dubois, Trynka et al. 2010).

Figure 3.13 (cont). C. Region containing *CCR3*



3.6 Methods

Details of genotyping and analysis methods are given in general methods. However, methods specific to this study are presented below.

3.6.1 Ethical approval

Written informed consent was obtained for all subjects participating in the study with local Ethics Committee / Institutional Review Board approval (3.9.2 provides more detail). Unless stated, affected coeliac individuals were diagnosed according to standard clinical, serological and histopathological criteria including small intestinal biopsy. DNA samples were from blood, saliva or lymphoblastoid cell lines.

3.6.2 Study participants

3.6.2.1 Study participants: GWAS (stage 1)

UK1

Coeliac disease patients were recruited from adult outpatient clinics at seven UK hospital sites (Barts and the London, London; Derbyshire Royal Infirmary, Derby; Hammersmith Hospital, London; John Radcliffe Hospital, Oxford; Leeds University Hospitals, Leeds; Llandough Hospital, Cardiff; Sheffield University Hospitals, Sheffield). Inclusion criteria were based on presence of villous atrophy at diagnosis and (since test introduction) positive anti-endomysial/tissue transglutaminase antibody (van Heel, Franke et al. 2007). Population-based controls were analysed from the 1958 British Birth Cohort. Ethics committee (Oxfordshire REC B) and local approval were obtained for all cohorts. Genomic DNA was extracted from peripheral blood, or from immortalised peripheral blood lymphocyte cell lines (1958 British Birth Cohort). All individuals were unrelated and of white northern European ethnic origin.

UK2

Coeliac disease cases were recruited from adult outpatient clinics as described for UK1 (434 individuals) and by advertisement through Coeliac UK (1415 individuals). Screening

questionnaires were used to verify coeliac disease status in Coeliac UK members responding to advertisement placed in the quarterly magazine, *Crossed Grain.* For inclusion individuals were required to report that they had been diagnosed with coeliac disease and had either had a positive coeliac antibody test or intestinal biopsy. Genomic DNA was extracted from peripheral blood (individuals recruited from hospital outpatients) and saliva (Oragene) or from immortalised peripheral blood lymphocyte cell lines (1958 British Birth Cohort).

Dutch

795 unrelated Dutch individuals with coeliac disease were diagnosed according to the revised ESPGAN criteria (UEGW Working Group 2001). The cohort encompassed individuals that showed a Marsh II or Marsh III lesion in the initial diagnostic small-bowel biopsy specimens upon re-evaluation by one of two experienced pathologists, or presented with dermatitis herpetiformis and were HLA-DQ2 positive. The control cohort comprised Dutch blood bank donors (*n* = 833) and NELSON controls (*n* = 850). The NELSON project—an ongoing population-based, randomized multi-centre lung cancer screening trial recruits male smokers (van Iersel, de Koning et al. 2007). These controls were collected from the north and centre of the Netherlands (Groningen, Utrecht and Drenthe, The Netherlands). All the control subjects were heavy smokers or ex-smokers (a minimum of 16 cigarettes/day for 25 years or 11 cigarettes/day for 30 years), but did not develop airway obstruction or emphysema suggesting chronic obstructive pulmonary disease (COPD) until the end of a 4 year observation period. The current study was approved by the local ethics committees and all the patients and controls gave their written informed consent.

Italian

DNA isolated from whole blood was available from 538 patients diagnosed by a referral centre for coeliac disease (Centro per la prevenzione e diagnosi della malattia celiaca, Fondazione IRCCS Ospedale Maggiore Policlinico) and from 593 healthy controls from the north of Italy. The average age of onset was 24.7 years (range 1–78 years). All the affected individuals were diagnosed according to the revised ESPGHAN criteria showing a Marsh III lesion (UEGW Working Group 2001). In addition, patients' serum samples tested positive for both anti-transglutaminase and anti-endomysium antibodies. Only 1.3% of the affected individuals had no HLA-DQ2 and/or HLA-DQ8 risk alleles, which is in accordance with published data (Louka,

Moodie et al. 2003). Written informed consent was obtained from all individuals before enrolment in the study. The study was approved by the ethics committee of the Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy.

Finnish

Finnish affected individuals (sporadic cases, or unrelateds from affected families across Finland) were recruited at the Universities of Tampere and Helsinki (Koskinen, Einarsdottir et al. 2009; Koskinen, Einarsdottir et al. 2009). Finnish population controls comprised 904 samples from Finrisk (Corogene, excluding coronary heart disease) and 925 samples from Health 2000 (excluding metabolic syndrome and positive coeliac disease serology). Finrisk controls (912 individuals without coronary artery disease) come mainly from a broad area of southern Finland of mixed ancestry that captures the ancestral mix of the whole country (Jakkula, Rehnstrom et al. 2008; Kathiresan, Voight et al. 2009). The Health2000 cohort: a subset of 2138 individuals (out of the full ~6000 cohort) with metabolic syndrome (n=1213) or clearly without metabolic syndrome (n=925) was available. These samples were geographically from the whole of Finland.

The study was approved by the ethical committees of the Tampere and Helsinki University Hospitals

3.6.2.2 Study Participants: Follow-up (stage 2)

Follow-up: USA comprised 525 coeliac cases and 340 controls from the Mayo Clinic (Minnesota), and 448 coeliac cases and 215 controls from the University of California Irvine(Garner, Murray et al. 2009). Polish coeliac cases were diagnosed in hospital clinics, and controls from donors at the Children's Memorial Health Institute (Warsaw), excluding coeliac serology positive samples. Italian samples comprised 377 coeliac cases and 94 controls from Rome, and 637 coeliac cases and 711 coeliac serology negative controls from Naples(Megiorni, Mora et al. 2008). Irish coeliac cases and controls were as described, with additional samples (Hunt, Franke et al. 2007). 259 Finnish coeliac cases were recruited similarly to GWAS samples, and controls were an additional 653 population controls from the Finrisk study. 965 Hungarian coeliac cases were collected from Budapest and Debrecen children clinic, and 1067 controls representative of the Hungarian population were selected from an epidemiological study. Part of the Hungarian cohorts have been described earlier(Koskinen, Einarsdottir et al. 2009).

Spanish coeliac cases were recruited in Madrid hospitals, controls were donors and hospital employees(Dema, Martinez et al. 2009).

3.6.3 DNA extraction and quantification

3.6.3.1 DNA extraction from blood samples

For samples processed at Barts and the London School of Medicine and Dentistry, genomic DNA was obtained from whole blood, collected in EDTA Vacuette tubes (17U/10ml)(Greiner Bio-one, Gloucestershire UK, 455036) and stored at -80^{0}C, using Puregene^{TM} (Flowgen, Nottingham, UK, D50K1 D50K2, D50K3) DNA extraction solutions according to manufacturer's instructions. Briefly 9ml of thawed EDTA blood was added to 27ml of red blood cell lysis solution. Sample was inverted to mix and incubated for 5 minutes at room temperature before undergoing centrifugation (Mistral 3000i centrifuge, MSE, UK) for 5mins at 3100rpm (2000g). Supernatant was poured away and pellet re-suspended in 9ml cell lysis solution before shaking using a vortex (Vortex Genie 2, Scientific Industries, New York, USA). 3ml protein precipitate solution was added and sample vortexed for 20 seconds before re-centrifuging for 5 mins at 3200rpm (2000g). Supernatant was poured into a clean tube containing 9mls propan-2-ol, after which 9ml of 70% ethanol (VWR, Leicester UK, 10107) were added to pellet. The tube was inverted once and centrifuged. Supernatant was poured away and tube inverted on absorbent paper until the pellet was dry, The pellet was re-suspended in 1ml deionised, DNase and RNase free water (Sigma Aldrich, Dorset UK, W4202)

3.6.3.2 DNA extraction from saliva samples

Saliva samples were collected using Oragene^{TM} DNA Self-Collection kits (OG-250, DNA Genotek Inc, Canada). Saliva is collected into an Oragene collection container and DNA preserving solution is added on sealing the container. This allows long-term storage of samples at room temperature without DNA degradation.

DNA was purified from 0.5ml of saliva/Oragene DNA purifying solution according to the Oragene laboratory protocol for manual purification of DNA (Issue 3.7). Briefly, samples were incubated at 50^{0}C for 1 hour to release DNA from cells and inactivate nucleases. 500µl sample was transferred to a 1.5ml microcentrifuge tube and 20µl of Oragene DNA purifier (OG-L2P) added. Samples were mixed by vortexing for 5 seconds and incubated on ice for 10 minutes.

Samples were centrifuged at room temperature for 5 minutes at 15,000g (12,700rpm) using an Eppendorf Centrifuge 5415D. The clear supernatant was transferred using a pipette into a fresh 1.5ml microcentrifuge tube containing 500µl of 97-100% ethanol (Ethanol, absolute, Fisher Scientific, E/650DF/P17). The mixture was mixed by gently inverting 10 times and then incubated at room temperature for 10 minutes to allow DNA precipitation. Samples were then centrifuged at room temperature for 2 minutes at 15,000g (12,700rpm). The supernatant was carefully removed and 250µl of 70% ethanol added and left for 1 minute to wash. The ethanol was carefully removed and DNA pellet left to air dry for 5 minutes. 20µl of DNase and RNase free water (Sigma Aldrich, Dorset UK, W4202) was added to dissolve the DNA pellet and samples were left an orbital shaker (KCH-Vibrax VXR, Kinematic, Switzerland) for 12 hours (overnight) to ensure complete re-hydration of DNA. DNA samples were stored at -80^{0}C.

3.6.3.3 DNA quantification

DNA was quantified for genotyping microarray experiments using a Quanti-iT PicoGreen dsDNA assay kit (Invitrogen, UK, P11496) according to the manufacturer's instructions. DNA samples were assayed in duplicate in 96 well plate format (40 samples per plate). Briefly, DNA was diluted by adding 1µl of DNA sample to 999µls Tris-EDTA (1x)(Sigma Aldrich, Dorset UK, T9285) in a deep (1.2ml) 96well plate (ABgene, Epsom, UK, AB-0564). Plates were sealed using adhesive PCR films (Thermo Scientific, AB-0558) and shaken for 5 seconds using a vortex (Vortex Genie 2, Scientific Industries, New York, USA). Plates were then also inverted twice to mix. Lambda DNA was serially diluted in 1x Tris-EDTA. Stock lambda (100µg/ml) DNA, supplied with the Quanti-iT^{TM} PicoGreen dsDNA assay kit and stored at -20^{0}C, was thawed to room temperature and 20µl was transferred into a 1.5ml tube containing 980µl Tris-EDTA solution (Sigma Aldrich, Dorset UK, T9285). Serial 1 in 4 dilutions were performed y transferring 250µl from the top standard to a second 1.5ml tube containing 750µl Tris-EDTA. The tube was shaken for 2 seconds using a vortex and further serial dilutions continued in the same way. 100µl of duplicate sample and standards were transferred to corresponding wells of a 96 well Pico plate (Costar^{TM} 3610 96 well assay plate, Corning Inc., USA). Fluorescent dye solution was prepared by transferring 50µl fluorescent dye to 11 ml of Tris-EDTA (1x) solution in a foil-covered 15 ml tube. The tube was shaken on a vortex for 10 seconds to mix. 100µl fluorescent dye solution was added to each well of the Pico plate containing DNA samples, and the plate was then loaded onto the fluorometer (BMG Labtech FLUOstar OPTIMA, Germany). Filter settings were 485nm(excitation) and 520nm(emission) for fluorescence measurement. Sample

DNA concentrations were calculated from the mean of measurement from the pair of sample duplicates by fitting a standard curve. Samples showing poor sample pair measurement concordance were re-assayed from the stock solution.

3.6.4 Genotyping

3.6.4.1 GWAS genotyping

Genotyping platforms used in the study are listed in **Table 3.8**. The locations where genotyping was performed are listed in **Table 3.8**. Genotyping assay protocols for each platform are described in general methods.

UK cases were genotyped at The Genome Centre, Barts and the London School of Medicine and Dentistry. 200ng DNA diluted at 50ng/µl for each sample was used. Genotyping on Illumina Human 670-Quad Customv1 BeadChips was performed according to the Illumina Infinium^{TM} HD Assay Super Protocol Guide Revision C (Illumina Inc, San Diego, USA). Briefly, DNA was whole-genome amplified, fragmented and resuspended in hybridization buffer prior to loading onto Beadchips. After hybridization, BeadChips were washed and single base extension and staining performed. BeadChips were loaded onto the BeadArray Reader, a two channel high resolution laser imager, using an Illumina Autoloader.

Normalized raw intensities for red and green channels corresponding to each SNP allele is performed using Illumina's proprietary normalization procedure. Data was outputted as X_{Norm} and Y_{Norm} or R (X_{Norm} + Y_{Norm}) and theta ("copy angle"), the ratio of X_{norm} and Y_{norm} normalised to between 0 and 1 (theta = (2/pi) x arctan2 (Ynorm, Xnorm)) using Illumina BeadStudio 2.0 software. Genotype calling was performed using R and theta data, merging datasets for calling in 5 calling pools as described in **section 3.3.3.1**. A modified version of a custom-designed algorithm created by Lude Franke (Groningen and Barts and the London) was used to call genotypes for all SNPs (van Heel, Franke et al. 2007; Franke, de Kovel et al. 2008). This algorithm seeks to assign samples to one of three biallelic SNP genotypes by comparing sample intensity values (corresponding to R-theta position on a SNP cluster plot) with the mean and standard deviations of the three genotype clusters. SNP cluster plots for individual SNPs were generated by plotting R versus theta for all samples.

3.6.4.2 Follow-up genotyping

Finnish controls (12) were genotyped on the Human 610-Quad BeadChip at the Wellcome Trust Sanger Institute. All other samples were genotyped using the Illumina GoldenGate assay on the Veracode/BeadXpress platform at Barts and The London Genome Centre; King's College London; and University Medical Centre Groningen.

Genotyping was performed using 144 SNPplex custom designed Illumina VeraCode™ GoldenGate assays. The genotyping was performed following the VeraCode Assay Guide (Revision A). Briefly, in the Illumina GoldenGate Assay, DNA is activated through biotinylation. Assay oligonucleotides (oligos) are added and hybridized to the sample DNA and the mixture bound to streptavidin-conjugated magnetic particles. After oligo hybridization, mis- and non-hybridized oligos are washed off and allele-specific extension of hybridized oligos is performed. The extended and ligated products form a template that is transferred to a PCR reaction and amplified. The strand containing the fluorescent signal in the PCR products is isolated and hybridized to the VeraCode beads via the address sequence. After the hybridization, the VeraCode beads are washed and scanned on the Illumina
BeadXpress Reader, a two channel laser fluorometer.

Genotype calling was performed in BeadStudio by visual inspection of SNP cluster plots and manual adjustment of genotype cluster positions. Calling was performed separately for combined cases and controls in each collection, with the exceptions of the Finnish collection, and whole genome amplified samples (89 Irish cases and 106 Spanish controls). Finnish samples were called from Human 670-Quad-custom data together with samples used in the GWAS. Whole genome amplified samples were called manually in BeadStudio separately from non-WGA samples due to observed differences in assay intensities. Sample and SNP quality control steps were performed as for the GWAS (with the exception of ethnic outlier analysis which was not possible with only 144 SNPs). 131 of 144 SNPs passed quality controls.

3.7 Statistical analysis

3.7.1 Case-control association analysis

Most analyses were performed using PLINK v1.05(Purcell, Neale et al. 2007).

Quality control steps performed using PLINK included sex-estimation for X chromosome SNP zygosity and hardy-weinberg equilibrium testing. A subset of 12,344 non-HLA SNPs, selected due to low pairwise linkage disequilibrium as informative of ancestry were used for ethnic outlier detection and relatedness detection analyses in PLINK(Yu, Wang et al. 2008). These analyses were performed using pairwise SNP *identity by state* calculations, multi-dimensional scaling analyses and *identity by descent* estimation for detection and exclusion of relatives. Linkage disequilibrium estimation (r-square and D') for putatively independent SNPs at a single risk locus was calculated in PLINK using 4936 UK2 controls. Determination of linkage disequilibrium among 395 SNPs with evidence of association in stage 1 ($P_{GWAS} < 10^{-4}$), was facilitated by use of Haploview (http://www.broadinstitute.org/haploview/haploview) to aid selection of SNPs tagging putatively independent associations within a single genomic region. 4936 UK2 controls were used as input for this analysis.

Quantile-quantile plots were made using the R statistical package (http://www.r-project.org/). SNPs from non-HLA coeliac risk loci were excluded for some plots, by removing SNPs within a 2 megabase window centred on the most strongly associated SNP from each region.

The Cochran-Mantel-Haenszel extension of the chi-square test of SNP allele counts (1 degree of freedom) was used for most association analyses. Logistic regression analyses were used to define the independence of association signals within the same linkage disequilibrium block, with group membership included as a factorized covariate. $Lambda_{GC}$ was estimated for the GWAS meta-analysis from the Cochran-Mantel-Haenszel allelic chi-square test statistics. The broad HLA region, excluded in most SNP association analyses, was defined as chromosome 6: 20,000,000 to 39,999,999 base pairs (Human Genome NCBI build 36 coordinates).

Pairwise SNP epistasis was assessed in PLINK utilizing a logistic regression model.

Principal components analysis was performed using Eigensoft v3.0 software (http://genepath.med.harvard.edu/~reich/Software.htm)(Price, Patterson et al. 2006). 12,344 non-HLA ancestry-informative SNPs were used for calculation of the top ten principal components in each GWAS sample collection. Case-control difference on each of the principal components was calculated using an ANOVA, 2 degrees of freedom test with the Eigensoft smartPCA application. Estimates of $lambda_{GC}$ for individual collections before and after adjustment for the top ten principal components was calculated for Cochran-Armitage trend association statistics using EIGENSTRAT software application(Price, Patterson et al. 2006). A weighted Z score method, using z scores obtained from Cochran-Armitage trend p values (adjusted and unadjusted) in each collection was used to calculate a stratified meta-analysis P value before and after EIGENSTRAT correction for principal components. This method weights the z scores according to sample collection size as recommended by de Jager et al. (De Jager, Jia et al. 2009).

The proportion of the total genetic variance of coeliac disease accounted for by non-HLA SNPs identified in the study was estimated using the logistic regression method implemented with the INPower software tool of Park et al. (http://dceg.cancer.gov/bb/tools/INPower) (Park, Wacholder et al. 2010). Effect sizes for SNPs were estimated from the stage 2 sample collections. The total genetic variance for coeliac disease was estimated from a sibling recurrence risk of 10, based on a log-normal distribution of genetic risk for polygenic traits ($\lambda_{sibling}^{2} = e^{variance}$) (Pharoah, Antoniou et al. 2002).

3.7.2 GRAIL analysis

Gene Relationships Among Implicated Loci (GRAIL) analysis (http://www.broadinstitute.org/mpg/grail/grail.php) was performed using HG18 and Dec2006 PubMed datasets, default settings for SNP reference sequence (rs) number submission, and the 27 genome-wide significant celiac disease risk loci (most associated SNP) as seeds(Raychaudhuri, Plenge et al. 2009). As a query we used either associated SNPs, or 101 x 50 randomly chosen Hap550 SNP datasets (5050 SNPs, of which 5033 mapped to the GRAIL database).

Briefly, the GRAIL method first defines the boundaries of the linkage disequilibrium block around the submitted SNP and identifies genes mapping within this region. All other human

genes are then ranked for text-based similarity with each gene in the region using pre-2006 Pubmed abstracts. The region is then compared to other disease (seed) regions, calculating a score based on the number of seed regions with related genes. The score of the most related gene in the region is used to calculate the significance of the region (P_{text}) and to annotate the most related gene in the region.

3.8 Bioinformatics and software resources

Bioinformatics resources and software used extensively in this chapter and throughout this thesis include:

University of California at Santa Cruz Genome Browser

Available at: http://genome.ucsc.edu/cgi-bin/hgGateway

NCBI dbSNP (Sherry, Ward et al. 2001)

Available at: http://www.ncbi.nlm.nih.gov/projects/SNP/

NCBI Gene

Available at: http://www.ncbi.nlm.nih.gov/gene

The Human HapMap project (Consortium 2005)

Available at: http://hapmap.ncbi.nlm.nih.gov/

Haploview v4.0 (Barrett, Fry et al. 2005)

Available from: http://www.broadinstitute.org/haploview/haploview

Genetics Power Calculator (Purcell, Cherny et al. 2003)

Available at: http://pngu.mgh.harvard.edu/~purcell/gpc/

Power for Genetic Association Analyses (PGA) (Menashe, Rosenberg et al. 2008)

Available from: http://dceg.cancer.gov/bb/tools/pga

WGA Viewer: software for genomic annotation of whole genome association studies (Link, Parish et al. 2008)

Available from: http://people.genome.duke.edu/~dg48/WGAViewer/

PLINK v1.05 Whole genome association analysis toolset (Purcell, Neale et al. 2007)

Available from: http://pngu.mgh.harvard.edu/~purcell/plink/

Eigensoft: Eigenstrat and smartPCA applications (Price, Patterson et al. 2006)

Available from: http://genepath.med.harvard.edu/~reich/Software.htm

R statistical package

Available from: http://www.r-project.org/

Gene Relationships Among Implicated Loci (GRAIL) (Raychaudhuri, Plenge et al. 2009)

Available at: http://www.broadinstitute.org/mpg/grail/grail.php

INPower (Park, Wacholder et al. 2010)

R package available from: http://dceg.cancer.gov/bb/tools/INPower

Figure acknowledgements

Figures 3.2 and **3.3** were created using a Perl script written for this study by Graham A Heap, Barts & the London School of Medicine and Dentistry
Figure 3.13 was created by Lude Franke, Barts & the London School of Medicine and Dentistry and University of Groningen

Chapter 4 Genome wide association study (GWAS) of azathioprine and mercaptopurine-induced pancreatitis

4.1 Introduction

The thiopurines, 6-mercaptopurine and azathioprine, are used in clinical practice as immunosuppressants and anti-leukaemic agents. As treatments for intestinal inflammatory diseases, their principal indications are in maintaining remission in ulcerative colitis and Crohn's disease (Timmer, McDonald et al. 2007; Prefontaine, Sutherland et al. 2009). Azathioprine is also less commonly used treat individuals with refractory coeliac disease (Maurino, Niveloni et al. 2002; Goerres, Meijer et al. 2003; Rubio-Tapia and Murray 2010). The undoubtedly valuable efficacy of these drugs is offset by both dose-dependent (type A) and idiosyncratic (type B) adverse effects. 10-20% of exposed individuals experience adverse effects leading to drug discontinuation (Present, Meltzer et al. 1989; de Jong, Derijks et al. 2003; Warman, Korelitz et al. 2003). Discontinuation occurs most frequently for mild, dose-dependent effects including nausea, malaise, myalgias, arthralgias and headache. In addition, dose-dependent myelo- and hepato-toxicity necessitate regular monitoring blood tests during both initiation and maintenance of thiopurine therapy. Mild leucopoenia is common but clinically significant myelosuppression requiring dose reduction or discontinuation occurs in around 2% of individuals (Present, Meltzer et al. 1989; Warman, Korelitz et al. 2003). Dose-dependent hepatotoxicity occurs in between 3-10% of individuals, though this includes mild, asymptomatic elevations in liver biochemistry tests that rarely require drug discontinuation (Present, Meltzer et al. 1989; Bastida, Nos et al. 2005; Gisbert, Gonzalez-Lama et al. 2007). Dose-independent reactions occur in less than 5% of individuals and include a hypersensitivity-like syndrome with fever, rash and diarrhoea (Present, Meltzer et al. 1989; de Jong, Derijks et al. 2003). Acute pancreatitis is a clinically important risk, occurring as an idiosyncratic reaction in around 3% of thiopurine-exposed individuals in the setting of inflammatory bowel disease (Haber, Meltzer et al. 1986; Bermejo, Lopez-Sanroman et al. 2008). There is some evidence that pancreatitis occurs less frequently in other disease settings (for unknown reasons). However, an increase in the risk of pancreatitis for all disease indications is supported by a large population-based case-control study where the overall risk of pancreatitis was increased approximately 8fold in users of azathioprine for any indication, even after adjustment for

inflammatory bowel disease diagnosis and other known causes of pancreatitis (Floyd, Pedersen et al. 2003). The investigations presented in this chapter aimed to determine whether common genetic variants are a major cause or contributor to pancreatitis risk in thiopurine-exposed individuals.

4.1.1 History and clinical uses of thiopurines

Thiopurines are purine compounds that have been modified by the substitution of sulphur for oxygen atoms (**Figure 4.1**). 6-mercaptopurine and azathioprine are the two thiopurines in common clinical use today. Thioguanine, while still used as a treatment for leukaemia, is rarely used as an immunosuppressant - its use has been abandoned in inflammatory bowel disease due to concerns over hepatotoxicity and in particular nodular regenerative hyperplasia which may cause irreversible portal hypertension (Dubinsky, Vasiliauskas et al. 2003). The major clinical indications for thiopurines are as immunosuppressants in the treatment of chronic inflammatory diseases and after allograft transplantation and as cytotoxic agents in the treatment of haematological malignancies, particularly acute lymphoblastic leukaemia (**Table 4.1**).

The development of thiopurines arose from a search for purine antimetabolites in the 1950s, where the substitution of the oxygen by sulphur at the carbon 6 position of hypoxanthine and guanine was found to produce compounds (6-mercaptopurine and 6-thioguanine) that inhibited purine utilization and were active against tumours and leukaemias (Elion, Hitchings et al. 1951). This discovery led to the first clinical trial of 6-mercaptopurine (6-MP) in acute leukaemias in 1953 (Burchenal, Murphy et al. 1953). Azathioprine was subsequently developed from efforts to synthesize thiopurines that were protected from some of the metabolic inactivating steps that were recognized to occur for 6-MP (Elion 1989). Azathioprine was subsequently shown to act as a pro-drug of 6-MP, reacting with glutathione in red cells to release 6-MP. It was found to have similar anti-leukaemic efficacy and toxicity to 6-MP.

The immunosuppressive effects of thiopurines, distinct from direct myelotoxicity, were first demonstrated for 6-MP, after it was shown to suppress the antibody response to foreign antigen in rabbits in 1958 (Schwartz, Stack et al. 1958). The first direct comparison of azathioprine and 6-MP as immunosuppressants showed superior graft survival in dogs after renal transplant for azathioprine (Calne 1960). This led to the first successful trial of

azathioprine as an immunosuppressant for human renal allograft transplantation(Murray, Merrill et al. 1963). The first reported use of thiopurines in inflammatory bowel disease was for 6-MP in the successful treatment of ulcerative colitis in 1962 (Bean 1962). The first randomized controlled trial supporting the use of thiopurines in IBD was reported for 6-MP in Crohn's disease in 1980 (Present, Korelitz et al. 1980). Multiple randomized controlled trials have established the efficacy of azathioprine and 6-MP in the induction and maintenance of remission of Crohn's disease (Sandborn, Sutherland et al. 2000; Prefontaine, Sutherland et al. 2009). In ulcerative colitis, azathioprine and 6-MP have efficacy in maintaining remission and as steroid-sparing agents (Timmer, McDonald et al. 2007).

Figure 4.1 **Chemical structure of azathioprine, mercaptopurine and the naturally occurring purine, hypoxanthine from which mercaptopurine was derived**

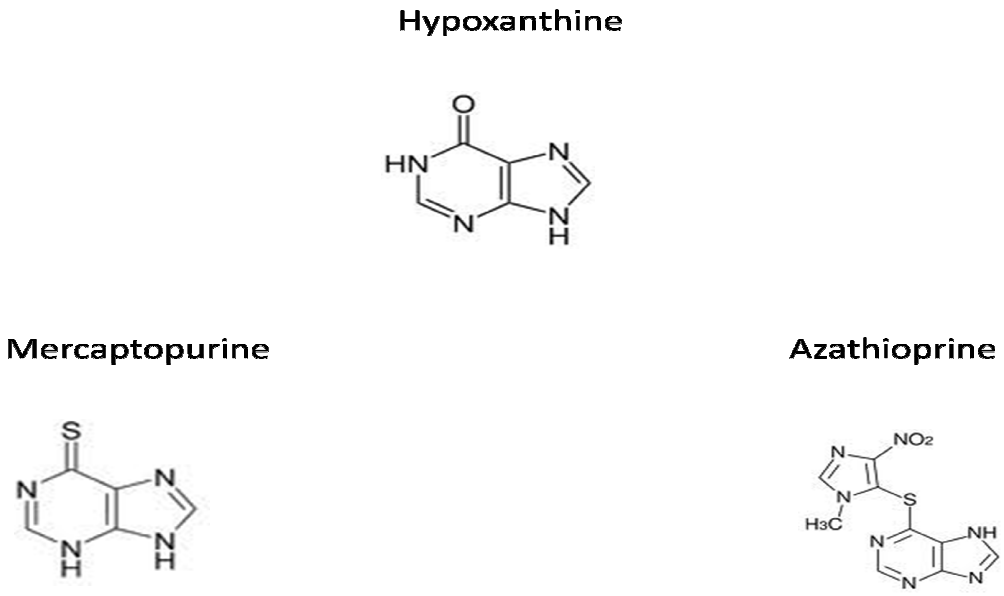


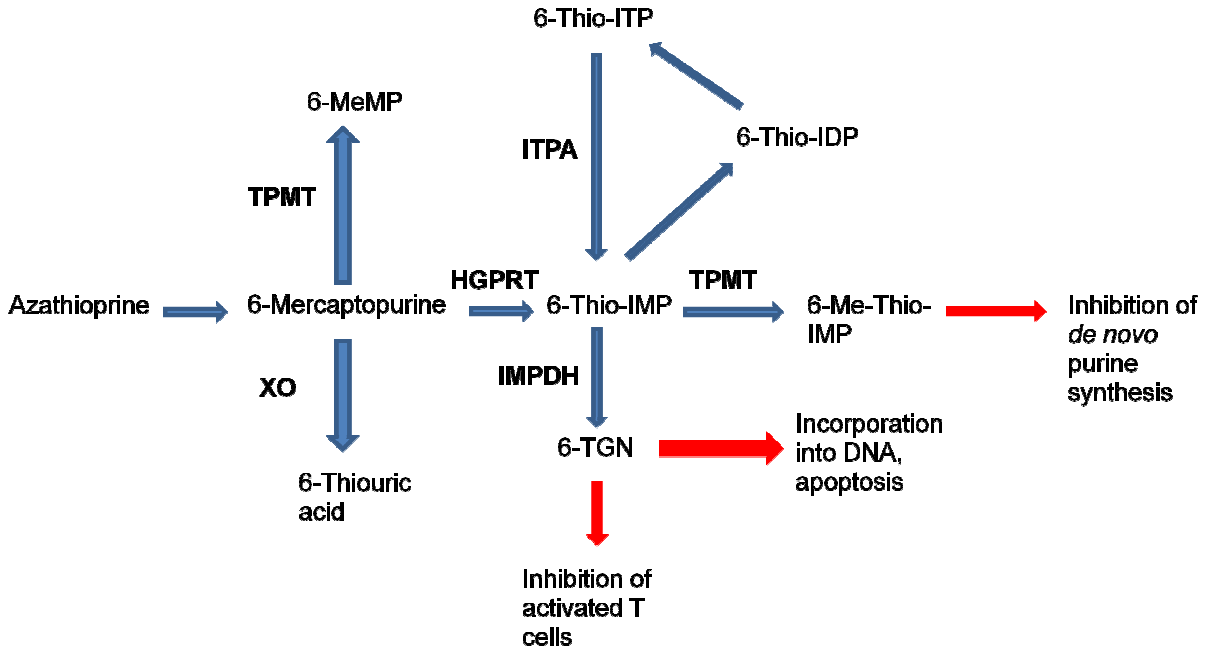
Table 4.1 Clinical indications for azathioprine and mercaptopurine

Indications	Reference
Childhood Acute lymphoblastic leukaemia	(Chessells, Bailey et al. 1995)
Solid organ transplantation	(Ponticelli, Tarantino et al. 1999; Germani, Pleguezuelo et al. 2009)
Crohn's Disease	(Sandborn, Sutherland et al. 2000; Prefontaine, Sutherland et al. 2009)
Ulcerative Colitis	(Timmer, McDonald et al. 2007)
Refractory coeliac disease	(Goerres, Meijer et al. 2003)
Autoimmune hepatitis	(Pratt, Flavin et al. 1996)
Rheumatoid arthritis	(Heurkens, Westedt et al. 1991)
Systemic lupus erythematosus	(Abu-Shakra and Shoenfeld 2001)
Atopic eczema	(Meggitt, Gray et al. 2006)

4.1.2 Metabolism and mechanism of action of azathioprine and 6-MP.

The metabolism and pharmacology of thiopurines is complex and the mechanisms of efficacy and toxicity remain only partly understood. Thiopurines are inactive pro-drugs that exert their cytotoxicity and immunosuppressive effects after they have been metabolized intra-cellularly. The cytotoxic effects are mediated by metabolites, principally thioguanine nucleotides which, by incorporation into DNA trigger cell cycle arrest and apoptosis. This mechanism may also mediate some of the immunosuppressive effects of azathioprine and mercaptopurine by inhibiting lymphocyte proliferation (Lennard 1992). In activated T cells, thioguanine nucleotides inhibit the expression of genes involved in inflammation such as *TRAIL* (TNF-related apoptosis-inducing ligand), *TNFRS7* (TNF-receptor superfamily 7) and *ITGA4* (alpha4 integrin) (Thomas, Myhre et al. 2005). Moreover 6-thio-GTP generated from azathioprine can block Rac1 GTPase activation that occurs on CD28-mediated T co-stimulation. This can promote T cell apoptosis or inhibit antigen-presenting cell-T cell conjugation (Tiede, Fritz et al. 2003; Poppe, Tiede et al. 2006). An alternative mechanism of immunosuppression is through the generation of 6-methyl-thioinosine 5' monophosphate (**Figure 4.2**) which acts as a potent inhibitor of de novo purine biosynthesis. The immunosuppressive effects of this action include blockade of lymphocyte proliferation.

Figure 4.2 **Azathioprine and 6-mercaptopurine metabolism and mechanism of action**



TPMT Thiopurine methyl transferase; XO Xanthine oxidase; 6-MeMP 6-Methyl mercaptopurine; HGPRT Hypoxanthine-guanine phosphoribosyltransferase; 6-Thio-IMP 6-thio-inosine monophosphate; ITPA Inosine triphosphate pyrophosphatase; 6-Thio-IDP 6-thio inosine diphosphate; 6-Thio-ITP 6-thio inosine triphosphate; IMPDH Inosine monophosphate dehydrogenase; 6-TGN: 6-thioguanine nucleotide; 6-Me-Thio-IMP: S-methyl-thioinosine 5'-monophosphate

4.2 Pharmacogenetics of drug adverse effects

In contrast to common diseases, sufficiently large family and in particular twin studies of drug adverse effects to enable estimates of heritability have not been reported. Nevertheless there have been a few reports of familial clustering of idiosyncratic drug reactions (though none specifically for thiopurine-induced pancreatitis) (Gennis, Vemuri et al. 1991; Pellicano, Silvestris et al. 1992; Johnson-Reagan and Bahna 2003). Monozygotic twin concordance for carbamazepine hypersensitivity has also been reported (Edwards, Hubbard et al. 1999). Together these limited reports provide some support for a heritable component to at least some idiosyncratic drug reactions.

A distinction between dose-dependent and dose-independent adverse effects may be helpful in considering the likely genetic contributions to these effects. Dose-dependent toxicities are likely to be influenced mainly by genetic variation in genes involved in the drug metabolic or transportation pathways (affecting pharmacokinetics). Conversely dose-independent, idiosyncratic adverse effects are likely to be influenced mainly by genetic variation outside these pathways, influencing drug and drug metabolite interactions with target organs or the immune system (pharmacodynamics) (Daly, Donaldson et al. 2009). As such, focussing on genetic variation within the key metabolic enzyme genes of thiopurine metabolism may be the most efficient approach to identify variants influencing dose-dependent effects. In contrast, for dose-independent effects, a much broader (ideally whole-genome) assay of genetic variation should be considered. In the case of thiopurine-induced pancreatitis the evidences cited for dose-independence include the observations that pancreatitis can occur at low as well as high doses, that pancreatitis does not correlate with other dose-dependent toxicities and that once established, it does not reverse on dose reduction. Furthermore pancreatitis recurs on re-exposure even at low doses (Haber, Meltzer et al. 1986; Sandborn, Sutherland et al. 2000; Weersma, Peters et al. 2004). However, while these observations separate pancreatitis from typical dose-dependent thiopurine toxicities (e.g. myelosuppression, hepatotoxicity) they do not entirely preclude dose (and pharmacokinetics) being an important factor influencing risk. This possibility has not been systematically investigated and is an unresolved question. At least some other idiosyncratic drug reactions occur more frequently at higher doses (e.g. statin-induced myopathy) (Pirmohamed 2010). Thus, pharmacokinetic mechanisms might still play a lesser role in the pathogenesis of idiosyncratic reactions including thiopurine-induced pancreatitis.

4.2.1 Pharmacogenetics of thiopurine dose-dependent toxicity: the example of TPMT polymorphisms

Genetic variation affecting activity of thiopurine methyl transferase (TPMT) is an important cause of dose-dependent adverse effects. TPMT is a cytosolic enzyme which can catalyse S-methylation of 6-MP and 6-TG as well as other aromatic and heterocyclic thio compounds (**Figure 4.2**). Erythrocyte TPMT activity, which mirrors activity in leucocytes and other tissues, follows a trimodal distribution, with approximately 1 in 300 individuals showing low/undetectable activity and 10% showing intermediate activity(Weinshilboum and Sladek 1980). This activity distribution is primarily determined by *TPMT* gene polymorphisms (Tai, Krynetski et al. 1996). In 80-95% of Caucasians, reduced TPMT activity is attributable to 3 common polymorphisms: TPMT*2(238 G > C), TPMT*3A (460 G > A, 719A>G) and TPMT*3C (719A > G) (Schaeffeler, Fischer et al. 2004; Sahasranaman, Howard et al. 2008).

TPMT activity is an important determinant of azathioprine and mercaptopurine toxicity. Low activity is associated with increased production of thioguanine nucleotides and increased cytotoxicity including myelosuppression (Chocair, Duley et al. 1992; Soria-Royer, Legendre et al. 1993). Activity is also a predictor of clinical response, with high activity associated with lower response rates and lower TGN levels (Cuffari, Dassopoulos et al. 2004). However, TPMT activity and genetic polymorphisms do not correlate well with other adverse effects including the most commonly encountered effects (nausea, myalgias and arthralgias, rash, abdominal pain and pancreatitis) (Marinaki, Ansari et al. 2004).

4.3 Thiopurine-induced acute pancreatitis

Acute pancreatitis has been reported in between 1.4 and 5% of individuals with inflammatory bowel disease during treatment with azathioprine or 6-mercaptopurine, usually within the first few weeks after initiation of therapy (Haber, Meltzer et al. 1986; Present, Meltzer et al. 1989; Kirschner 1998; Fraser, Orchard et al. 2002; Floyd, Pedersen et al. 2003; Weersma, Peters et al. 2004). 6-thioguanine appears to be exempt from this risk and individuals experiencing pancreatitis on azathioprine or mercaptopurine do not develop pancreatitis after switching to 6-thioguanine (Bonaz, Boitard et al. 2003; Dubinsky, Feldman et al. 2003). This suggests that the risk of pancreatitis is due to mercaptopurine or one of its unique metabolites, rather than on the thioguanine nucleotides (generated via 6-thioguanosine-5'-monophosphate) that are shared end products of all three drugs.

Azathioprine and mercaptopurine-induced pancreatitis is a potentially serious adverse effect, usually necessitating hospital admission. However, in contrast to acute pancreatitis in other settings, where mortality is around 5%, thiopurine-induced pancreatitis is relatively mild and cases of severe or life-threatening pancreatitis are very rare (Pitchumoni, Rubin et al. 2010). In one of the largest series of 46 cases of azathioprine or mercaptopurine induced pancreatitis, all were mild (Bermejo, Lopez-Sanroman et al. 2008).

The risk of pancreatitis does not appear to be related to dose and is not correlated with other dose-dependent toxicities (liver injury, bone marrow suppression) (Weersma, Peters et al. 2004). Pancreatitis recurs rapidly on re-exposure, even at lower doses (Haber, Meltzer et al. 1986; Cappell and Das 1989; Present, Meltzer et al. 1989). These features have been cited as evidence that the mechanism may involve an immune-meditated hypersensitivity reaction.

The risk of pancreatitis appears to be elevated in all individuals exposed to thiopurines, regardless of disease indication (Floyd, Pedersen et al. 2003). However, there is evidence that thiopurine-induced pancreatitis risk is higher in individuals with Crohn's disease than in other autoimmune conditions and indeed it has been very rarely reported in studies of azathioprine use in other autoimmune diseases and after renal transplantation (Weersma, Peters et al. 2004). At The Royal London hospital's renal unit, a search of the renal transplant database revealed no documented cases of azathioprine-induced pancreatitis (personal communication). A possible explanation for the elevated risk in Crohn's disease is that the risk

of acute pancreatitis is elevated generally in individuals with inflammatory bowel disease, independent of thiopurine exposure. In this model, thiopurine exposure might elevate further the risk of the "idiopathic" component of increased pancreatitis observed in inflammatory bowel disease. The largest case-control study to examine this question found 4-fold and 8-fold increased risks in ulcerative colitis and Crohn's disease respectively (Rasmussen, Fonager et al. 1999). It has been hypothesized that low-grade pancreatic inflammation may exist in Crohn's disease and interact with thiopurine risk to explain the increased risk in Crohn's disease in particular (Weersma, Peters et al. 2004). Pancreatic autoantibodies have been observed in around 30% of individuals with Crohn's disease but in only 0-3.7% of controls (Seibold, Hufnagl et al. 1999; Joossens, Vermeire et al. 2004; Koutroubakis, Drygiannakis et al. 2005). A single study measuring these autoantibodies in individuals with a history of Crohn's disease-associated azathioprine-induced pancreatitis found 2/8 (25%) individuals with autoantibodies, this was a higher frequency and at higher titres than in unaffected Crohn's patients (7.7%) in this study(Weersma, Batstra et al. 2008). Whether these autoantibodies play a role in pathogenesis is uncertain, but it is an intriguing possibility that a subset of Crohn's patients may have immune-sensitized or mildly inflamed pancreas prior to thiopurine exposure and may therefore be most at risk of pancreatitis on exposure.

4.3.1 Genetics of thiopurine-induced pancreatitis

Only a few small candidate gene studies have been reported in the field of azathioprine or mercaptopurine induced pancreatitis. These studies have focused on genes encoding enzymes catalyzing the conversion of potentially toxic thiopurine metabolites. Thiopurine methyltransferase (*TPMT*) gene polymorphisms influence the production of cytotoxic thioguanine nucleotides and affect the risk of thiopurine-induced myelotoxicity (Weinshilboum and Sladek 1980; Lennard, Lilleyman et al. 1990; Schaeffeler, Fischer et al. 2004). However, TPMT polymorphisms do not influence the risk of pancreatitis (Marinaki, Ansari et al. 2004). One UK study has suggested that polymorphisms in the inosine triphosphate pyrophosphatase (ITPA) gene, encoding an enzyme that converts the mercaptopurine metabolite, 6-thio-ITP to 6-thio-IMP, may influence the risk of pancreatitis. Deficiencies in this enzyme are predicted to cause accumulation of the potentially toxic metabolite, 6-thio-ITP (**Figure 4.2**). In this study a polymorphism, 94A>C, present in 4% of controls and associated with reduced levels of ITPA activity was associated with an increased risk of pancreatitis (OR 6.2, CI 1.1–32.6) (Marinaki,

Ansari et al. 2004; Marinaki, Duley et al. 2004). However this analysis included only 8 patients with pancreatitis and the association has not been replicated in other studies (Gearry, Roberts et al. 2004; van Dieren, van Vuuren et al. 2005). As discussed, genes involved in the thiopurine metabolic pathways are good candidates for dose-dependent toxicities (e.g. myelotoxicity, hepatotoxicity) but are less plausible candidates for dose-independent toxicities like pancreatitis. The sensitization that occurs (rapid recurrence on drug re-exposure) with thiopurine-induced pancreatitis has given rise to the hypothesis that the mechanism involves an immune-mediated hypersensitivity reaction. Whatever the mechanism, the rapid resolution after drug withdrawal suggests that continued drug exposure is required to drive pancreatic toxicity. It is possible that one or more thiopurine metabolites are directly toxic to the pancreas, but that genetic variation influences susceptibility to these metabolites, for example through variation in membrane solute transporters expressed in the pancreas. Given the relative lack of understanding of the pathogenesis of thiopurine-induced pancreatitis, a comprehensive and well-powered genome-wide survey of genetic variation offered the opportunity to derive new insights into mechanisms driving this clinical outcome.

4.4 Genome-wide association studies of drug adverse effects

Human populations have not been exposed to most drugs, including azathioprine, over sufficient time to enable the negative selection of genetic variants that reduce fitness in the context of drug exposure. Therefore variants which are highly deleterious in the context of thiopurines may have reached significant population frequencies by chance or due to selection favouring other functions (Cirulli and Goldstein 2010). Thus, in contrast to common disease, where nearly all risk variants identified confer modest effects on disease risk and have therefore required very large association studies for their identification, much smaller studies may have sufficient power to detect risk variants influencing susceptibility to drug adverse effects.

Thiopurine-induced pancreatitis is common and therefore a common genetic variant conferring a substantially increased risk of pancreatitis might form the basis of a useful screening test. Proof of principle for this idea has come from the example of abacavir hypersensitivity. Abacavir is a human immunodeficiency virus-1 (HIV-1) reverse transcriptase inhibitor. A hypersensitivity syndrome of skin rash, gastrointestinal symptoms (nausea, vomiting, diarrhoea, abdominal pain) and respiratory disturbance (cough, pharyngitis, dyspnoea) occurs in around 4% of abacavir treated individuals (Hetherington, McGuirk et al. 2001). In 90% of cases the syndrome occurs within the first 6 weeks of therapy with the median time to onset 11 days in a large retrospective review of 1803 cases (Hetherington, McGuirk et al. 2001). Moreover, symptoms resolve on drug withdrawal and return (often with greater severity) on re-exposure. Thus the frequency and phenotype has some similarities to thiopurine-induced pancreatitis. In 2002, the MHC class 1 allele HLA-B*5701 was identified as a strong risk factor for abacavir hypersensitivity (odds ratio = 117) in an investigation of MHC alleles in hypersensitivity cases versus drug-tolerant controls (Mallal, Nolan et al. 2002). The HLA-B*5701 allele occurs in around 6% of Caucasian individuals. The mechanism of this effect may involve HLA-B*5701-restricted CD8 T cell activation by an abacavir metabolite (Chessman, Kostenko et al. 2008). A randomized controlled trial of HLA-B*5701 screening in HIV infected individuals prior to treatment with abacavir showed that screening could reduce the incidence of immunologically-confirmed hypersensitivity from 3.1 % to 0%(Mallal, Phillips et al. 2008). This test had a 100% negative predictive value and 47.9% positive predictive value in this trial and is now mandatory in Europe prior to commencing treatment with abacavir (Mallal, Phillips et al. 2008; Daly, Donaldson et al. 2009). A variant of this effect size in thiopurine-induced

pancreatitis would clearly be identifiable in a genome-wide association study including only a small numbers of cases. Nelson et al. estimated that a GWAS genotyping 500,000 SNPs would need only 15 cases and 200 population controls to detect the HLA-B*5701 effect (Nelson, Bacanu et al. 2009).

Two recent reports have provided further support for the GWAS approach, successfully identifying common genetic variants with large effects on susceptibility to idiosyncratic drug toxicities (Link, Parish et al. 2008; Daly, Donaldson et al. 2009). The first of these studies genotyped 316,184 SNPs in 85 individuals who developed myopathy while taking the cholesterol-lowering medication simvastatin and 90 controls. Statin-induced myopathy occurs in around 1 person per 10,000 per year, but the risk is increased at higher statin doses (Daly, Donaldson et al. 2009). While most cases occur within the first year of statin exposure, a significant proportion can occur later. Thus, while statin-induced myopathy has features of an idiosyncratic reaction (rare, severe, unpredictable), dose-dependency is observed. A genome-wide association study was performed between the myopathy cases and the drug-exposed but unaffected controls. This study identified a single genome-wide significant SNP (rs4363657). No other SNPs showed association at $P_{GWAS}<10^{-5}$. The myopathy-associated SNP risk allele had a frequency of 0.46 in myopathy cases and 0.13 in drug-exposed, non-myopathy controls (0.15 in the healthy population controls) conferring an increased risk of myopathy (allelic odds ratio 4.5, $P_{\text{GWAS-genotypic trend}}=2.5 \times 10^{-8}$))(Link, Parish et al. 2008). The finding was replicated in a smaller sample collection that included 21 myopathy cases. Rs4363657 maps to an intron in the *SLCO1B1* gene and is in strong linkage disequilibrium with a non-synonymous SNP in *SLCO1B1* (rs4140956), which accounted for the association in this study. *SLCO1B1* encodes the organic anion–transporting polypeptide OATP1B1, which mediates the hepatic uptake of various drugs, including most statins. The myopathy risk allele is associated with lower uptake and higher serum statin concentrations (Konig, Seithel et al. 2006).

The second study genotyped 866,399 SNPs in 51 individuals with flucloxacillin induced liver injury and 282 unaffected controls. Flucloxacillin-induced liver injury is rare with incidence estimated at 8.5 per 100,000 new users in the United Kingdom (Russmann, Kaye et al. 2005). As with thiopurine-induced pancreatitis, flucloxacillin-induced liver injury occurs within the first few weeks after drug exposure (mean 25 days) (Daly, Donaldson et al. 2009). This study identified a SNP whose risk allele conferred an allelic odds ratio in favour of liver injury of 45. This variant is present in 5% of healthy controls and 84% of cases and was reported to be in

complete linkage disequilibrium with *HLA-B*5701,* a human leucocyte antigen class I gene variant, suggesting that this HLA variant plays a critical role in facilitating the immune mediated liver injury. The mechanism here is currently unknown, although the same allele is associated with abacavir hypersensitivity. Since flucloxacillin lacks structural similarity to abacavir and its metabolites, it has been postulated that the shared association may rather be due to a missense SNP (rs2395029) in *HCP5,* which is in complete linkage disequilibrium with the *HLA-B*5701* allele. This gene is expressed in immune cells - the drug-hypersensitivity associated SNP is associated with protection against HIV infection. Thus the mechanism of flucloxacillin-induced liver injury could depend on the immune modulatory effects of this gene rather than a classical HLA-B: CD8 interaction(Daly, Donaldson et al. 2009).

Certainly the phenotype here, although much rarer than thiopurine-induced pancreatitis, is similar in time of onset and dose-independence. This study further raised the possibility that variants with very large effects on risk of thiopurine-induced pancreatitis might also exist, tractable to a study of similar size.

4.5 Aims and power calculations

It was hypothesized that the risk of thiopurine-induced pancreatitis is influenced by common genetic variants of large effect. The current study was powered based on the odds ratios and population frequencies of variants discovered in the abacavir, flucloxacillin and simvastatin studies. A primary motivation was to ensure sufficient power to detect variants that would be useful in a screening test (cf. abacavir). Thus, this study was not powered to detect moderate or low-effect size variants. For example, assuming an incidence of pancreatitis of 3% in azathioprine-exposed individuals, there is 80% power to detect a common variant allele (MAF >/= 0.05 in controls) conferring an odds ratio of 4 or more in an additive genetic model, with 55 cases and 5000 unselected controls, with alpha=$5x10^{-7}$(Purcell, Cherny et al. 2003). For more common variants (MAF>0.25), this study has 80% power to detect variants with ORs>/= 3 with alpha =$5x10^{-7}$. However, as variant effect size or minor allele frequency fall, power rapidly tails off. For example the same sized study would only have 80% to detect a variant with MAF=0.05 and OR=3 at an alpha = 10^{-3}. Thus we have good power to detect common variants conferring allelic odds ratios greater than 4, a reasonable threshold for a clinically useful variant.

4.6 Study populations

Genotyping and analyses were performed for two sample collections, one from the Netherlands and one from the United Kingdom (**Table 4.2**).

Cases were recruited from collaborators in the United Kingdom and the Netherlands (Groningen). Cases were individuals with a diagnosis of azathioprine or mercaptopurine-induced acute pancreatitis (**Table 4.3- Clinical characteristics)**. All cases were screened using a phenotype questionnaire (**Appendix 1**). For inclusion, a diagnosis of pancreatitis was established through the presence of typical clinical features (including severe upper abdominal pain) together with biochemical or radiological supporting features (raised serum or urine amylase, radiological features of acute pancreatitis). The likelihood that the thiopurine was the cause of pancreatitis was assessed on a case by case basis by participating gastroenterologists. This included an assessment of the timing of thiopurine exposure, resolution on drug withdrawal and exclusion of other common causes of pancreatitis. In addition, phenotyping questionnaires were completed by contributing investigators and independently assessed by PCAD to validate or refute case inclusion. In total 15 cases were identified and included from the Netherlands and 42 cases were identified from 9 centres in England and Scotland. All cases were of documented European ancestry.

Population controls were used for both sample collections. For the Netherlands and United Kingdom collections, population controls used in the coeliac genome wide association study (chapter 3) were included (Dubois, Trynka et al. 2010). It was considered that the theoretical advantages of matching controls for disease (Crohn's disease or Ulcerative Colitis) or thiopurine exposure would be offset by reduced power arising from lower available sample size. Power estimates using thiopurine-exposed, but tolerant controls, suggested only modest improvement in detectable odds ratios. For example, even assuming a pancreatitis incidence in thiopurine exposed individuals of 5%, detectable odds ratio for a 5% MAF SNP would improve only from 4 to 3.8 using assumptions as stated above and a control sample size of 2000.

Our study was powered only to detect variants of large effect (up to allelic OR >3 for more common SNPs). IBD disease-associated SNPs discovered in large genome wide association studies have much more modest effects and are therefore unlikely to confound the detection

of large effect variants. It was considered very unlikely that variants reaching thresholds for statistical significance in this study could be caused by association with inflammatory bowel disease (due to modest effects of known IBD risk variants). It was anticipated that in instances where putative thiopurine-induced pancreatitis SNPs fall within known IBD-associated regions, methods to test for confounding by association to known IBD SNPs could be used (e.g. conditional logistic regression). Similarly, SNPs mapping to regions not previously associated with inflammatory bowel disease were even more unlikely to be confounded by IBD association. In general, it was anticipated that not matching controls for disease or thiopurine exposure would have negligible effects on allele frequency differences between cases and controls and minimal effects on type 1 and type 2 error rates.

Table 4.2 Sample collections and genotyping platforms

	Cases		Controls		
	Sample size	Platform	Sample size (pre-QC)	Platform	SNPs (post-QC)[a]
United Kingdom (UK)	40 (41 pre-QC)	Illumina 1Mv3	4936 (5069 pre-QC)	Illumina 1.2Mv1	920,266
Netherlands (NL)	15 (15 pre-QC)	Illumina 670-Quad	846 (960 pre-QC)	Illumina 670-Quad	535,753

[a]Number of SNPs genotyped in both cases and controls and passing quality controls

Table 4.3 Case clinical characteristics

	UK[a]	Netherlands[b]	All cases[c]
Total case number	40	15	55
Crohn's disease	8 (62%)	14 (93%)	22 (79%)
Ulcerative colitis	5 (38%)	1 (7%)	6 (21%)
Male	7 (54%)	4 (27%)	11 (39%)
Female	6 (46%)	11 (73%)	17 (61%)
Azathioprine	13 (100%)	15 (100%)	28 (100%
Mercaptopurine	0	0	0
Azathioprine dose-median mg (range)[d]	125 (100-150)	138 (75-175)	138 (75-175)
Serum amylase - median xULN (range)[e]	7.4 (1.9-21.8)	3.3 (1.2-24.5)	5.6 (1.2-24.5)
Mean age at diagnosis (range)	45.4 (21-74)	29.4 (17-67)	37.7 (17-74)

[a]data available from 13 of 40 UK samples unless stated (missing or incomplete for 27 UK samples)
[b]data available from 15 of 15 NL samples
[c]data available from 28 of 55 samples
[d]data only available for 6 UK and 14 NL samples
[e]data only available for 12 UK and 13 NL samples

4.7 Results

4.7.1 Quality control steps

Genotypes were called directly from the Illumina SNP intensity data using the same genotype-calling algorithm used in the coeliac GWAS (chapter 3). Genotypes were called separately for Dutch (cases and controls) and UK (cases and controls) samples. Controls failing quality controls in the coeliac GWAS were excluded.

GWAS quality control filters included: removing related samples (1st degree relatives or duplicates), ethnic outliers, and samples with genotype call rate <98%, SNPs with call rate <95% or hardy-weinberg p<0.0001 in controls as applied in Chapter 3. Since case numbers were small relative to controls, application of a SNP call rate filter was applied separately to cases and controls to ensure that SNP assays performing badly only in cases (for example due to batch effects or genotyping platform differences) were excluded. In addition, a differential missingness filter to exclude SNPs with lower call rate in either cases or controls was assessed but did not lead to exclusion of any of the top associated SNPs ($P_{GWAS}<10^{-4}$). SNP intensity cluster plots were inspected to assess genotype-calling accuracy for all SNPs showing association at $P_{GWAS}<10^{-4}$. SNPs were excluded using a low threshold for possible calling bias. Sample exclusions: 1 individual was excluded from the UK collection as an ethnic outlier. All other individuals passed quality controls.

SNP exclusions: 204,344 SNPs were excluded from the UK data. 26,742 SNPs were excluded from the Dutch data (**Table 4.4**)

Table 4.4 SNP numbers passing quality controls in the GWAS

	UK	**Netherlands**
Pre- quality controls	1,124,600	562,495
Genotype call rate < 0.95	934,629 (189,971)	541,250 (21,245)
Hardy-weinberg P < 0.0001	920,266 (14,363)	535,753 (5,497)

4.7.2 Primary association analysis and identification of false positive SNP associations

After application of quality controls, case-control association analysis was performed. SNP case-control association was assessed primarily by comparison of SNP allele frequencies between cases and controls using Fisher's exact test. This test is appropriate in the absence of knowledge of the model of inheritance of risk variants (additive, dominant or recessive) and is relatively unbiased in favour of any one model (Balding 2006). Moreover Fisher's exact test is superior to Pearson's chi-square test when the numbers of observations in one or more category are small. This arises for rare SNPs and low sample numbers. Association testing for SNPs between UK and Dutch controls indicated significant genome-wide allele frequency differences between UK and Dutch population samples (genomic inflation factor 2.15; see also chapter 3). Case-control association testing stratified by UK and Dutch collections was therefore performed to minimize confounding by known ancestry differences. Meta-analysis of association test statistics within each sample collection was performed using a sample size-weighted z score method(de Bakker, Ferreira et al. 2008).

Allelic tests of association, while being relatively unbiased towards any one mode of inheritance, have less power to detect associations compared to tests that fit the model of inheritance appropriately (Balding 2006). Thus, in the absence of knowledge of the mode of inheritance, another suggested approach has been to select the test (modelling additive, dominant and recessive effects) that gives the highest association test statistic for each SNP (Balding 2006). Fisher's exact test can be used to assess not only differences in allele counts, but also binary comparisons modelling dominant and recessive effects. Under additive assumptions, a Cochran-Armitage trend test is widely used and has greater power in this scenario than the allelic comparison. By calculating association test statistics using case-control tests that model dominant, recessive and additive inheritance, the best model of inheritance was estimated for each SNP showing allelic association < 10^{-4} (**Table 4.5**). This analysis was primarily undertaken to test whether any SNPs showed much more compelling evidence of case-control association under these models than had been identified by applying the conservative allelic Fisher's exact test.

Table 4.5 **Most strongly associated SNPs from 39 loci with $P_{GWAS}<10^{-4}$. Loci ordered by strength of association.**

SNP	Chr	BP (HG18)	Minor allele[a]	MAF-cases (UK) n=40	MAF-controls (UK) n=4396	MAF-cases (NL) n=15	MAF-controls (NL) n=846	P_{GWAS}[b]	OR [95%CI][c]	Best inheritance mode (P_{GWAS})[d]	SNP included on Immuno chip	Closest gene & gene(s) of interest
rs4943552	13	37402142	C	0.638	0.394	0.607	0.407	2.46×10^{-6}	2.58 [1.75-3.83]	REC (4.82×10^{-7})	Yes	*TRPC4*
rs7788583	7	47835563	C	0.400	0.177	nd	Nd	2.87×10^{-6}	3.10 [3.10-1.97]	ADD (3.05×10^{-6})	*Yes*	*PKD1L1*
rs2346996	16	26418533	G	0.013	0.179	nd	Nd	5.05×10^{-6}	0.06 [0.01-0.42]	DOM (3.97×10^{-6})	*No*	*Nil*
rs11211587	1	47906280	G	0.238	0.075	nd	Nd	5.98×10^{-6}	3.86 [2.29-6.50]	DOM (2.69×10^{-6})	*Yes*	*FOXD2*
rs7723119	5	139391887	G	0.588	0.341	nd	Nd	8.32×10^{-6}	2.75 [1.76-4.30]	REC (3.87×10^{-6})	*Yes*	*NRG2; SLC4A9, CXXC5*[i]
rs6928830	6	84276031	C	0.359	0.166	0.267	0.142	8.84×10^{-6}	2.64 [1.76-3.95]	ADD (8.29×10^{-6})	*Yes*	*PRSS35*
rs6849889	4	120405796	C	0.113	0.332	nd	Nd	1.13×10^{-5}	0.25 [0.13-0.51]	ADD (3.69×10^{-5})	*Yes*	*USP53*
rs17187115	14	86449033	A	0.150	0.038	0.100	0.036	1.83×10^{-5}	4.03 [2.32-7.02]	ADD (2.17×10^{-5})	*Yes*	*Nil*
rs2405316	13	45797944	C	0.175	0.060	0.133	0.036	1.93×10^{-5}	3.49 [2.32-7.02]	ADD (1.76×10^{-5})	*Yes*	*C13orf18*
rs17124166	14	87712056	T	0.050	0.232	nd	Nd	2.27×10^{-5}	0.17 [0.06-0.48]	DOM (3.02×10^{-5})	*No*	*KCNK10*
rs12423591	12	79994653	T	0.238	0.132	0.400	0.113	2.31×10^{-5}	2.68 [1.77-4.08]	ADD (8.52×10^{-6})	*Yes*	*ACSS3*
rs2647087	6	32789027	C	0.475	0.252	nd	Nd	2.34×10^{-5}	2.69 [1.77-4.08]	ADD (1.35×10^{-5})	Yes	HLA-DQA2[i]; HLA-DQA1, HLA-DQB1
rs9616333	22	48436851	C	0.613	0.377	nd	nd	2.43×10^{-5}	2.61 [1.66-4.10]	ADD (1.98×10^{-5})	*Yes*	*C22orf34*
rs28399637	19	50015978	T	0.538	0.310	nd	nd	2.86×10^{-5}	2.58 [1.66-4.02]	DOM (1.87×10^{-5})	*Yes*	*BCAM*
rs2305350	2	46673319	A	0.250	0.479	nd	nd	3.96×10^{-5}	0.36 [0.22-0.60]	ADD (3.88×10^{-5})	*No*	*PIGF*
rs7688988	4	47587092	C	0.125	0.054	0.267	0.053	3.96×10^{-5}	3.42 [2.05-5.71]	ADD (5.26×10^{-5})	*Yes*	*NFXL1*
rs709873	14	84473543	A	0.250	0.477	nd	nd	4.12×10^{-5}	0.37 [0.22-0.61]	ADD (4.52×10^{-5})	*Yes*	*Nil*
rs9329070	5	178550468	A	0.113	0.275	0.100	0.282	4.15×10^{-5}	0.32 [0.18-0.58]	DOM (1.30×10^{-4})	*Yes*	*ADAMTS2*
rs1917179	10	54351221	A	0.150	0.317	0.100	0.302	4.50×10^{-5}	0.35 [0.20-0.60]	DOM (1.10×10^{-4})	*Alt SNP*	*MBL2*[i]
rs6578065	8	141142984	T	0.225	0.082	0.167	0.087	4.52×10^{-5}	2.91 [1.83-4.64]	REC (4.79×10^{-6})	*No*	*TRAPPC9*
rs1435649	3	179077632	A	0.113	0.313	nd	nd	4.65×10^{-5}	0.28 [0.14-0.56]	DOM (3.91×10^{-5})	*Yes*	*Nil*

Table 4.5 (cont.)

rs11779957	8	12919286	G	0.138	0.289	0.067	0.285	4.68×10^{-5}	0.33 [0.19-0.59]	ADD (1.40×10^{-4})	*Yes*	*C8orf79*
rs1782962	21	32893902	T	0.150	0.042	0.100	0.040	5.17×10^{-5}	3.65 [2.10-6.34]	DOM (2.38×10^{-5})	*Yes*	*C21orf59*
rs2322157	3	71961721	A	0.300	0.127	nd	nd	5.62×10^{-5}	2.95 [1.82-4.77]	ADD (3.70×10^{-5})	*Yes*	*PROK2*
rs2026589	9	24831259	G	0.438	0.288	0.567	0.296	5.64×10^{-5}	2.19 [1.51-3.20]	ADD (6.23×10^{-5})	*Yes*	*Nil*
rs11744322	5	91543421	A	0.125	0.027	nd	nd	6.21×10^{-5}	5.22 [2.66-10.24]	REC (1.43×10^{-9})	*Yes*	*Nil*
rs4883418	12	7644562	T	0.188	0.076	0.200	0.073	6.27×10^{-5}	2.90 [1.79-4.69]	ADD (6.48×10^{-5})	*Alt SNP*	*APOBEC1; CD163L1*[i]
rs932311	1	175460320	T	0.238	0.484	0.433	0.470	6.65×10^{-5}	0.44 [0.29-0.67]	ADD (6.38×10^{-5})	*Yes*	*FAM5B*
rs5949978	23	95838982	A	0.193	0.076	0.269	0.080	6.84×10^{-5}	3.31 [1.95-5.62]	ADD (6.57×10^{-4})	*Yes*	*DIAPH2*
rs6589652	11	117563568	G	0.425	0.224	nd	nd	7.00×10^{-5}	2.56 [1.64-4.00]	ADD (5.08×10^{-5})	*Yes*	*AMICA1; TMPRSS4, IL10RA*[i]
rs11793373	9	128447364	A	0.088	0.271	nd	nd	7.06×10^{-5}	0.26 [0.12-0.56]	DOM (2.38×10^{-5})	*Alt SNP*	*LMX1B*
rs7704592	5	71331087	C	0.350	0.187	0.367	0.198	7.10×10^{-5}	2.34 [1.58-3.48]	DOM (4.56×10^{-5})	*Yes*	*MAP1B*
rs1001990	10	130298844	T	0.113	0.259	0.067	0.263	7.10×10^{-5}	0.32 [0.17-0.59]	ADD (1.64×10^{-4})	*No*	*Nil*
rs11801594	1	5325970	T	0.475	0.366	0.767	0.368	7.34×10^{-5}	2.15 [1.47-3.14]	REC (3.98×10^{-5})	*Yes*	*Nil*
rs11840483	13	114106015	C	0.125	0.027	nd	nd	7.89×10^{-5}	5.06 [2.58-9.93]	ADD (7.87×10^{-5})	*Yes*	*ZNF828*
rs11253892	10	16372788	C	0.075	0.019	0.133	0.020	8.07×10^{-5}	5.09 [2.62-9.91]	ADD (1.00×10^{-4})	*Yes*	*PTER*
rs6696562	1	100051191	A	0.150	0.351	nd	nd	8.45×10^{-5}	0.33 [0.18-0.60]	DOM (1.54×10^{-4})	*No*	*AGL*
rs8072153	17	64481493	C	0.100	0.032	0.167	0.034	8.62×10^{-5}	3.95 [2.19-7.13]	ADD (9.50×10^{-5})	*Yes*	*ABCA9*
rs10509423	10	82737765	A	0.050	0.212	nd	nd	8.86×10^{-5}	0.20 [0.07-0.54]	DOM (2.17×10^{-5})	*No*	*SH2D4B*

[a]Minor allele in UK and Dutch (NL) controls. [b]Fisher's exact test of SNP allele counts, stratified by UK and Dutch collections and meta-analysed using case sample number-weighted z score method. [c]Odds ratio with reference to minor allele [95% confidence intervals]. Odds ratios for SNPs genotyped in both UK and Dutch collections estimated from Cochran-Mantel-Haenzel meta-analysis of allele counts [d]Best mode of inheritance here refers to model under which the strongest SNP association was observed, Recessive and dominant models assessed using Fisher's exact test stratified by sample collection. Additive model assessed using exact Armitage trend test stratified by sample collection. [e]SNPs submitted to and included on final version of Illumina Immunchip for replication effort. [f]Closest validated RefSeq gene (within 750Kb window). Genes of particular interest underlined. Immune genes denoted with [i]
TRPC4 – Transient receptor potential cation channel 4: calcium channel expressed in pancreas *SLC4A9* – Solute carrier 4, sodium bicarbonate: expressed in pancreas *TMPRSS4* – transmembrane protease, serine 4: over-expressed in pancreatic carcinoma

Preliminary association results showed no detectable overall inflation of test statistics in both Dutch (λ= 1.000) and UK (λ=1.000) collections, suggesting results are unlikely to be majorly confounded by genotyping bias or population stratification. Inspection of SNP (R vs. theta) cluster plots led to exclusion of 8 out of 69 SNPs with $P_{GWAS}<10^{-4}$ (Fisher's exact test of allele counts, meta-analysis of 2 sample collections by weighted z score method). 7 SNPs obtained genome-wide significant evidence of association after standard quality controls. However, for each of these SNPs no other SNPs within a broad genomic interval around the SNP were observed with $P_{GWAS}<10^{-4}$. Four of the seven SNPs had been genotyped in both Dutch and UK collections; for each of these SNPs, case (but not control) allele frequencies were discordant between Dutch and UK. For all seven SNPs association was driven by the UK data. These features suggested that genotyping error in the UK cases may have caused spurious associations. The low numbers of samples genotyped on the Illumina 1M-Duov3 platform (40 UK cases) meant that standard quality controls (including inspection of SNP cluster plots) may have been inadequate to reliably detect bias in UK cases for some SNPs. In contrast, all other samples had been genotyped as part of large collections on a single platform and were therefore more amenable to quality assessment (**Table 4.2**).

These 7 top SNP associations were designated "singleton SNP associations". Although genotyping error was suspected, genuine case-control association could not yet be excluded. For example, in the simvastatin myopathy study, the only positive (and subsequently replicated) association was for a similar singleton SNP. In an effort to assess SNP assay quality on the Illumina 1M-Duov3 platform, assay intensity data for 170 unrelated human HapMap samples genotyped on the Illumina 1M-Duov3 platform was obtained from Illumina for assessment. This data was used to generate SNP cluster plots for all SNPs showing association in the GWAS at $P<10^{-4}$. Using this approach only one (rs7503953) of the seven singleton SNPs showing genome-wide significant association appeared prone to probable genotype-calling bias. All SNPs included in the final association results (**Table 4.5**) had well-separated genotype clouds amenable to accurate automated calling and this provided some re-assurance that genotyping bias was not prevalent among these SNPs.

As a second approach, a search for proxy (i.e. highly correlated) SNPs was undertaken for each of the 7 singleton SNPs. For each SNP, the best proxy (highest r-square) SNP was identified with LD calculations performed using 4936 UK controls in PLINKv1.07. 2 SNPs had excellent proxies (r-square > 0.99), 1 SNP a more modest proxy (r-square = 0.766), and the other 4 SNPs

only weakly correlated SNPs (r-square 0.356-0.685) that were considered inadequate to serve as proxies for association testing. Each of the 3 adequate proxy SNPs showed no evidence of case-control association (P_{GWAS}> 0.05), suggesting that the associations in these regions were false positives. However, these approaches were still considered inadequate to rule out genuine case-control association for at least 4 of these SNPs (those without proxies). Genotyping was therefore repeated for all 7 singleton SNPs using an alternative SNP genotyping assay. Custom-designed KASPar SNP genotyping assays (KBioscience, Herts) were obtained for all seven singleton SNPs and a further non-singleton positive control SNP (rs7788583). Repeat genotyping was performed in 54 of 55 cases for which DNA was available and an additional 94 new controls. Since DNA was not readily available for controls used for GWAS genotyping, controls were randomly chosen individuals with coeliac disease, for which DNA was easily available.

SNP genotype calling on KASPar intensity data was performed from inspection of intensity cluster plots. The mean SNP call rate for 149 genotyped samples was 99.0%, with individual SNP call rates ranging from 98.0% to 100%. All SNP assays showed adequate separation of genotype clouds to enable calling, as assessed independently by two individuals.

Genotypic concordance for samples with non-missing data was 100% for all 8 SNPs genotyped in Dutch cases (Human 670Quad-customv1 vs. KASPar). Genotypic concordance for the positive control SNP (rs7788583) in UK cases (Human 1M-Duov3 vs. KASPar) was 100%. However, for the 7 singleton SNPs, concordance was 49.3% (Human 1M-Duov3 vs. KASPar), suggesting genotyping error in the Illumina 1M-Duov3 UK case data for these SNPs. Using KASPar assayed genotypes, there was no evidence of association between UK cases and controls (P>0.05) for any of the 7 singleton SNPs (**Table 4.6**). These SNPs were therefore excluded from further consideration and analysis.

Table 4.6 7 top singleton SNP associations in the GWAS and exclusion of associations on re-genotyping

SNP	Chr	BP (HG18)	Minor allele [b]	MAF-controls (Infinium genotyping) [b]	MAF-cases (Infinium genotyping) [c]	MAF-cases (KASPar genotyping) [d]	MAF-KASPar controls [e]	$P_{GWAS-Infinium}$ [f]	$P_{GWAS-KASPar}$ [g]
rs7267722	20	22,735,622	C	0.1531	0.85	0.1341	0.1237	3.47×10^{-34}	0.803
rs17002187	21	26,762,664	T	0.09828	0.5132	0.07317	0.0806	1.47×10^{-19}	0.576
rs7503953	17	6,082,401	T	0.1632	0.6795	0.1098	0.179	2.02×10^{-14}	0.050
rs7054078	23	19,137,253	T	0.01756	0.2456	0.03571	0.0158	2.00×10^{-12}	0.261
rs7630157	3	169,797,194	C	0.01479	0.175	0.01282	0.0319	7.03×10^{-10}	0.631
rs9322193	6	149,960,836	G	0.342	0.0625	0.3293	0.367	6.75×10^{-09}	0.907
rs4862110	4	183,988,023	C	0.1694	0.5	0.1463	0.1862	1.85×10^{-07}	0.358

[a] Minor allele in GWAS controls

[b] Minor allele frequency in 4936 UK controls, assayed by Illumina Human-1.2MDuo-customv1.

[c] Minor allele frequency in 40 UK cases, assayed by Illumina Human-1MDuov3. Dutch data not shown, as KASPar genotyping concordance with GWAS data was 100%

[d] Minor allele frequency in 39 UK cases, assayed by KASPar SNP genotyping

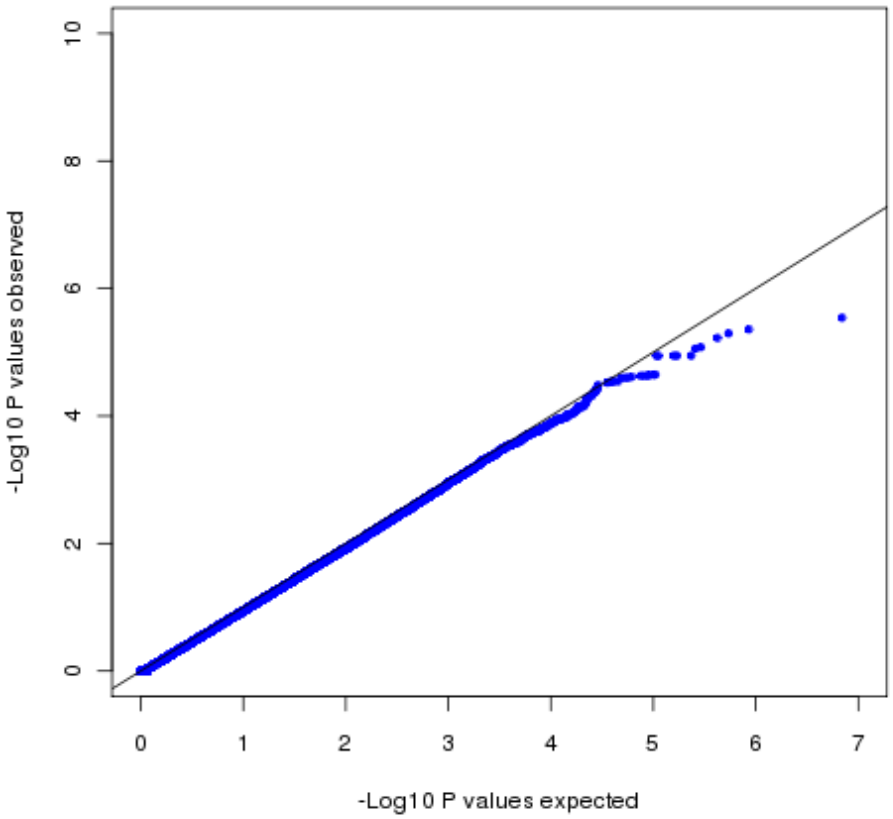
[e] Minor allele frequency in 94 UK coeliacs (new controls), assayed by KASPar SNP genotyping

[f] Meta-analysis allelic *P* value for UK and Dutch collections using GWAS genotype data (Fisher's exact test, sample-weighted z score meta-analysis)

[g] Meta-analysis allelic *P* value for UK and Dutch collections substituting KASPar genotypes for 39 UK and 15 Dutch cases

After exclusion of singleton SNPs whose associations did not withstand KASPar assay re-genotyping, and SNPs ($P_{GWAS}<10^{-4}$) with suspected genotyping error, the distribution of test statistics for all SNPs closely followed the distribution expected under the null hypothesis (**Figure 4.3, Q-Q plot**), with no excess of positive associations.

Figure 4.3 **Quantile-quantile plot of GWAS case-control association *P* values after removal of SNPs with suspected genotyping bias**



In total, 73 SNPs from 39 loci showed association at a threshold of $P_{GWAS}<10^{-4}$. None of these associations reached the threshold of genome-wide significance (P_{GWAS} <5 x 10^{-8}) and the number of SNPs with $P_{GWAS}<10^{-4}$ does not constitute a significant excess compared to that expected under the null hypothesis of no association. The study had approximately 80% power to detect variants with MAF >/= 0.05 and OR>/=4 and therefore variants of this

frequency and magnitude are unlikely to have been missed. Certainly variants conferring effects on risk of strength similar to those identified for abacavir hypersensitivity or flucloxacillin liver injury have not been identified in this study and are unlikely to exist for thiopurine-induced pancreatitis. However, the study was not well powered to detect common genetic variants of intermediate or modest effect (SNPs with ORs <3). Such variants, although less likely to be of immediate clinical relevance for genetic screening might nevertheless offer insights into thiopurine pancreatitis pathogenesis. Examination of the function of genes in the regions around SNPs with $P_{GWAS}<10^{-4}$ was undertaken to identify genes of possible functional relevance to azathioprine/mercaptopurine-induced pancreatitis. Possible candidate genes are discussed in section 4.9. Of note, the strongest association in the GWAS was for rs4943552 ($P_{GWAS}=2.46 \times 10^{-6}$ OR= 2.59) mapping 60 kilobases from the closest gene, *TRPC4,* a calcium channel expressed in the pancreas. Multiple SNPs showed association in the *HLA* gene region on chromosome 6 and this was the 12th most strongly associated locus overall (**Figure 4.4**).

4.7.3 Supplementary case-control association analyses

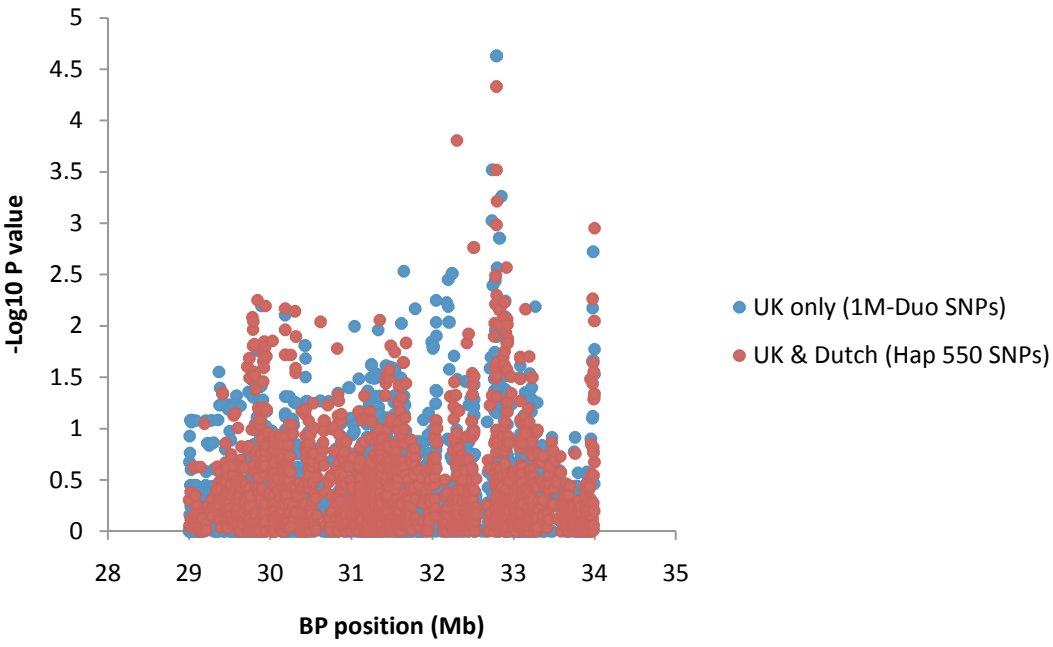
Genetic risk variants identified in previous pharmacogenomic genome wide association studies have mostly fitted an additive mode of inheritance. This is true for the simvastatin *SCOL1B* SNP and the *HLA-B*5701* tagging variant identified in the flucloxacillin drug-induced liver injury study (Link, Parish et al. 2008; Daly, Donaldson et al. 2009). In addition, *TPMT* polymorphisms associated with azathioprine and mercaptopurine myelotoxicity also show additive effects on risk. Thus, although any of the three modes of inheritance could in principle most closely model a true thiopurine-induced pancreatitis variant, it was considered that additive modes were most likely. The allelic Fisher's exact test used in the primary association analysis is conservative compared to tests that fit the mode of inheritance appropriately. Thus, association test statistics for additive, recessive and dominant modes of inheritance were examined for all SNPs showing $P_{GWAS}<10^{-4}$. In this analysis, a single SNP (rs11744322, $P_{GWAS\text{-}recessive} = 1.43 \times 10^{-9}$) showed genome-wide significant evidence of association under a recessive model of inheritance.

This SNP had been genotyped in the UK collection only (**Figure 4.5**). Genotyping error must be considered a possible cause of this association, even though this SNP assay showed good cluster characteristics assessed in 170 HapMap individuals. There were no other regional SNPs

with association, but none of the genotyped SNPs were in strong linkage disequilibrium to rs11744322 (r-square> 0.5). Rs11744322 maps to a genomic region on chromosome 5 that has no known genes close by (within 500kb up or downstream) and no genes of good biological candidacy in the broader region. Assessment of this association will require genotyping in additional sample collections and/or re-genotyping in UK cases.

All other SNPs analysed under dominant, recessive and additive assumptions did not reach the genome-wide significance threshold of $P_{GWAS}= 5 \times 10^{-8}$ (**Table 4.5**)

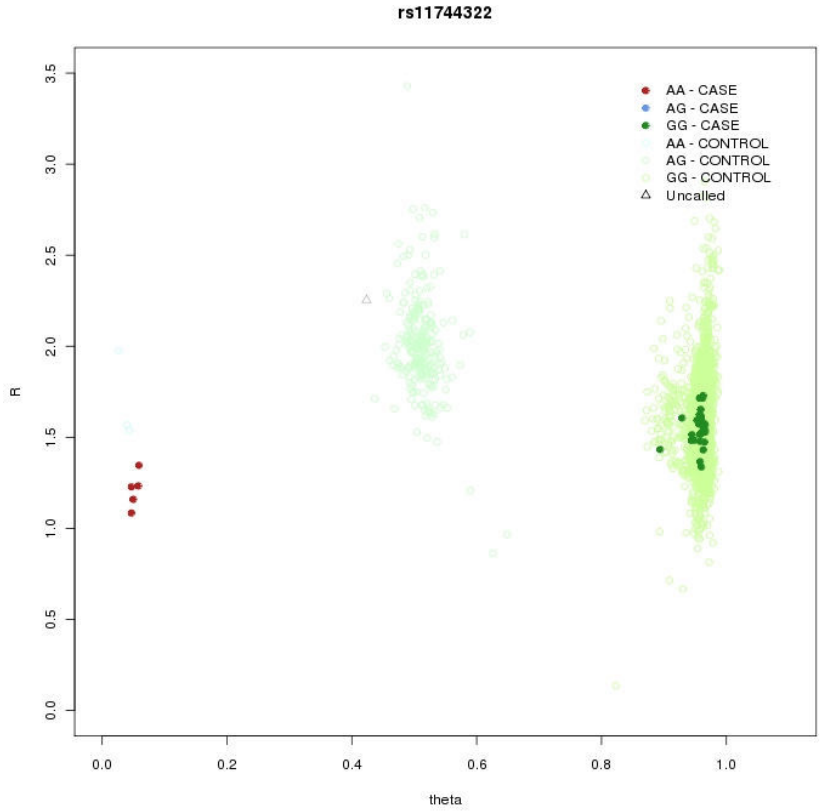
Figure 4.4 GWAS SNP associations within the HLA gene region (Chr 6, 29-34Mb)



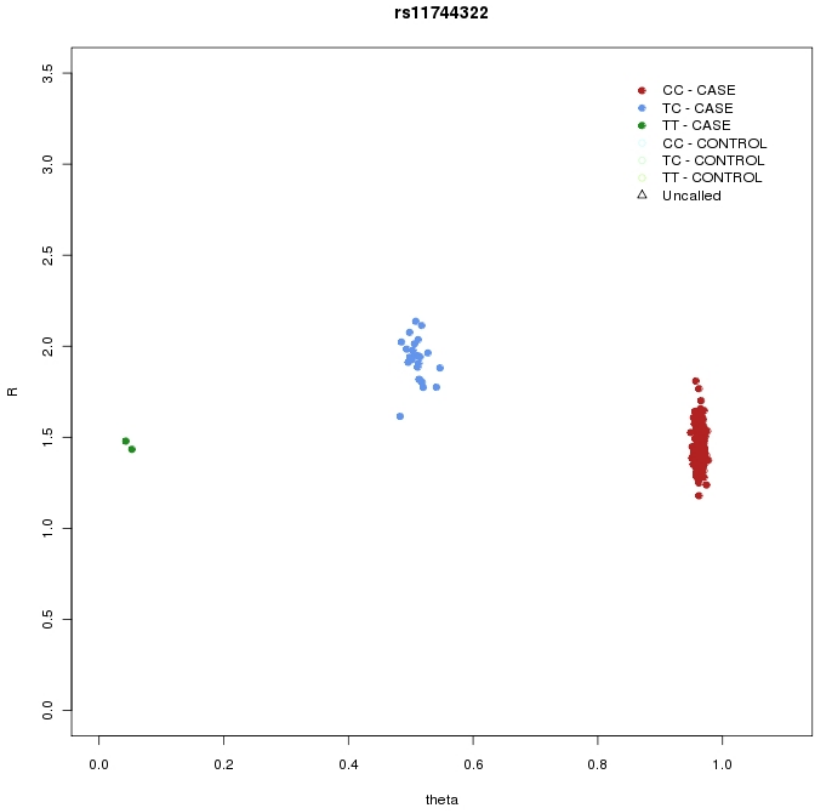
.

Figure 4.5 SNP cluster plots for rs11744322. A. 40 cases (Illumina 1MDuov3) and 4936 controls (Illumina 1.2MDuov1) in the UK collection. B. 170 HapMap individuals (Illumina 1MDuov3)

A.



B.



4.7.4 Association in known Inflammatory Bowel Disease risk regions

Data was also inspected to determine whether known IBD loci showed evidence of association in the GWAS. Regions containing SNPs reaching genome-wide significance in the largest Crohn's and Ulcerative colitis meta-analyses were searched for case-control association in the current GWAS. There was no inflation of association statistics for SNPs within 30 known Crohn's disease loci (Barrett, Hansoul et al. 2008)) or 12 known ulcerative colitis loci, excluding the HLA (McGovern, Gardet et al. 2010). This finding was consistent with the fact that genetic variants associated with inflammatory bowel disease were below the power of this study to detect. Genetic variation within the HLA region has been associated with both colonic Crohn's disease and ulcerative colitis. The strongest HLA association in the largest Crohn's disease meta-analysis was for rs3763313. This SNP is also weakly associated with ulcerative colitis (McGovern, Gardet et al. 2010). This SNP showed no association in the current study (P_{GWAS}=0.710). Similarly, rs2395185, which was the most strongly ulcerative colitis-associated SNP within the HLA region in the McGovern et al. meta-analysis, showed no association in the current study (P_{GWAS}=0.679). The HLA associations in the current study are for variants of putatively much greater risk effect (odds ratio = 2.69), than those influencing the risk of inflammatory bowel disease (e.g. rs3763313 - OR for Crohn's disease = 1.19). Thus the allelic distribution of HLA variants in the thiopurine-induced pancreatitis samples was not expected to be significantly altered by Crohn's disease of ulcerative colitis status and cannot explain the current moderate HLA association observed in this study.

4.7.5 Association in TPMT and ITPA gene regions

Association with common SNPs in regions around the *TPMT* and *ITPA* genes was assessed. *TPMT*3A* (rs1800460), the commonest polymorphism associated with low TPMT enzyme activity in Caucasian populations, showed no association (P_{GWAS}=0.586). Similarly, rs6909725, a proxy for *TPMT*3C* (r-sq = 0.316, D' = 1) showed no association (P_{GWAS}=1). *TPMT*2* was not genotyped and had no adequate proxies in the GWAS data. In a 1Mb window around the *TPMT* gene region, there was mild overall test statistic inflation for 336 SNPs, with peak association for rs2842938 (P_{GWAS}= 0.00201), a synonymous SNP in *TPMT*. Definitive assessment of common SNP association in and around the *TPMT* gene with pancreatitis would require replication genotyping in a much larger sample collection.

The *ITPA* IVS2+21A>C polymorphism (rs7270101) showed no association (P_{GWAS}=0.252). The *ITPA* 94 C>A polymorphism was not genotyped in the GWAS or in HapMap phase II and therefore proxies could not be assessed. Within the *ITPA* gene region (1Mb window, 331SNPs), there was no overall inflation of test statistics and no association for SNPs within the ITPA gene (P_{GWAS}>0.05). Again, genotyping of common SNPs in larger sample collections would be required to determine whether any real association with azathioprine or mercaptopurine-induced pancreatitis is present.

4.7.6 Association in known idiopathic and hereditary pancreatitis risk regions

Genetic studies in thiopurine-induced pancreatitis have been small, few in number and restricted to investigating gene candidates of the thiopurine metabolic pathway (*TPMT* and *ITPA*). In hereditary and idiopathic acute and chronic pancreatitis, variants in several genes with known pancreatic function affect risk. These genes have not been tested in thiopurine-induced pancreatitis but implicate pathways that may be common to all forms of pancreatitis (Whitcomb; Whitcomb 2004).

Hereditary pancreatitis is an autosomal dominant condition, characterized by early onset recurrent acute pancreatitis that progresses to chronic pancreatitis in around 50% of cases. A combination of linkage and candidate gene approaches led to the identification of mutations in the *PRSS1* gene, encoding cationic trypsinogen, as the main cause of this disorder (Gorry, Gabbaizedeh et al. 1997). The *PRSS1* associations have been replicated and cause gain-of-function with increased activation of trypsinogen (Witt, Luck et al. 1999; Rebours, Boutron-Ruault et al. 2009). *PRSS1* gene mutations have also been associated with some cases of idiopathic chronic pancreatitis in the absence of the classical hereditary pancreatitis phenotype (Gorry, Gabbaizedeh et al. 1997; Liu, Gao et al. 2008). It has been suggested that chronic pancreatitis occurs through recurrent episodes of acute pancreatitis and that the mechanisms causing both acute and chronic pancreatitis are mostly shared(Whitcomb 2004). Thus, it is possible that some of the mechanisms underlying thiopurine-induced pancreatitis are also shared with those underlying idiopathic or hereditary pancreatitis. Candidate gene studies in familial and idiopathic forms of chronic pancreatitis have implicated a number of genes in the broader trypsin activity pathway. Most implicated genes either promote trypsinogen activation or impair clearance of trypsin suggesting that trypsin activation and

clearance mechanisms are central to the pathogenesis of pancreatitis(Whitcomb) . **Table 4.7** lists the genes implicated in different forms of pancreatitis to date.

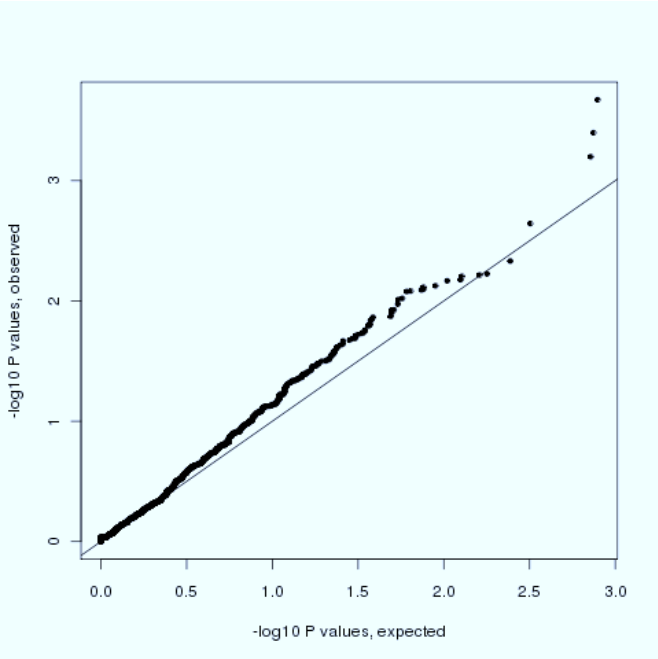
Indirect support for the idea that these genes could play a role in the risk of thiopurine-induced pancreatitis comes from studies of analogous patient groups at high risk of acute pancreatitis from other well-defined causes. For example, protein coding-variants in the cystic fibrosis transmembrane regulator (*CFTR*) were strongly associated (P<0.0001) with acute pancreatitis in an association study of 126 individuals with hypertriglyceridaemia (Chang, Chang et al. 2008). Hypertriglyceridaemia may be a reasonable model for the genetics of drug-induced pancreatitis. In both cases most individuals at risk (due either to hypertriglyceridaemia or thiopurine exposure) do not develop acute pancreatitis. The *CFTR* gene association in hypertriglyceridaemic suggests that risk factors could combine independently to determine pancreatitis risk. Thus, it is possible that thiopurine-induced pancreatitis risk is also determined not by genes interacting with thiopurines or their metabolites, but by factors that independently increase the risk of acute pancreatitis.

For this reason, SNPs in 1Mb windows around each of the known pancreatitis genes (**Table 4.7**) were assessed for association in the current study. Modest inflation of association test statistics was observed for 1889 SNPs from these 5 regions (*PRSS1* and *PRSS2* within same region, **Figure 4.6**). The strongest associations ($10^{-3}<P_{GWAS}<10^{-2}$) mapped to the *CFTR* gene region (P_{GWAS} = 0.00139), the *PRSS1* and *PRSS2* gene regions (P_{GWAS} = 0.00410) and the *CTRC* region (P_{GWAS} = 0.00313) all between 125 and 212kb from any of the genes of interest.

Table 4.7 Genes associated with pancreatitis in candidate gene studies

Gene	Condition (references)
***PRSS1* (Cationic trypsinogen)**	Hereditary pancreatitis and idiopathic chronic pancreatitis (Liu, Gao et al. 2008; Rebours, Boutron-Ruault et al. 2009)
***PRSS2* (Anionic trypsinogen)**	Idiopathic chronic pancreatitis(Witt, Sahin-Toth et al. 2006; Santhosh, Witt et al. 2008)
***CTRC* (Chymotrypsin C)**	Hereditary and idiopathic chronic pancreatitis (Masson, Chen et al. 2008; Rosendahl, Witt et al. 2008)
***CASR* (Calcium sensing receptor)**	Hereditary and idiopathic chronic pancreatitis (Felderbauer, Hoffmann et al. 2003; Felderbauer, Karakas et al. 2008; Muddana, Lamb et al. 2008)
***SPINK1* (Serine protease inhibitor, Kazal type 1)**	Chronic pancreatitis, recurrent acute pancreatitis (Aoun, Chang et al. 2008; Aoun, Muddana et al. 2010)
***CFTR* (Cystic fibrosis transmembrane regulator)**	Chronic pancreatitis, recurrent acute pancreatitis, hypertriglyceridaemic acute pancreatitis (Whitcomb; Audrezet, Dabricot et al. 2008; Chang, Chang et al. 2008; Segal, Yaakov et al. 2008)

Figure 4.6 Quantile-quantile plot of association tests statistics in the GWAS for SNPs within 6 pancreatitis gene regions



4.8 Selection of SNPs for follow-up genotyping in an independent sample collection

A strategy that prioritized SNPs showing the strongest evidence of association in the GWAS for replication in independent samples was considered to be the most economical and effective strategy to detect any truly associated variants.

With this in mind, SNPs from loci with $P_{GWAS} < 10^{-4}$ were selected for inclusion on the Illumina Immuno- Beadchip. The Illumina Immuno Beadchip ("Immunochip") is a custom designed SNP genotyping array that uses the same Illumina Infinium HD and BeadArray technologies that are the basis of the Quad670-custom, 1M-Duo and 1.2M-Duo-custom BeadChips used in the GWAS. The Illumina Immunochip incorporated assays for 196,524 SNPs submitted by investigators for 8 chronic immune mediated diseases. The deadline for SNP submissions for Immunochip design occurred very soon after completion of GWAS genotyping. Consequently, SNPs were selected for inclusion based on a preliminary association analysis. This analysis was performed on pooled UK and Dutch samples, without stratification by sample collection. In addition, genotypes were called using default Illumina cluster positions in BeadStudio for the UK case collection. Thus SNPs with $P_{GWAS} < 10^{-4}$ in this analysis differed slightly from those in the final, stratified meta-analysis. SNP associations were considered to be non-independent (tagging the same associated LD block) if they fell within 1 Mb of each other. Where multiple SNPs were observed at the same locus, the two most strongly associated SNPs were selected for Immunochip. 92 SNPs from 50 loci showed association in the GWAS ($P_{GWAS\text{-}preliminary}<10^{-4}$). In total 79 SNPs from 50 loci showing $P_{GWAS\text{-}preliminary}<10^{-4}$ *were* selected for Immunochip submission. 72 of these 79 SNPs passed Immuno Beadchip design quality controls. After exclusion of bad SNPs through automated genotype calling, full GWAS quality controls and KASPar genotyping, 62 Immunochip-included SNPs remained for re-assessment in the final association analysis.

In the final GWAS association analysis, 73 SNPs from 39 loci showed association $P_{GWAS}<10^{-4}$ (sample collection-stratified meta-analysis of allelic Fisher's exact test, using weighted Z score method). Of the 62 Immunochip SNPs 39 SNPs from 29 of 39 loci had $P_{GWAS}<10^{-4}$. A further 3 of the submitted SNPs with lesser P_{GWAS} mapped to 2 regions obtaining $P_{GWAS}<10^{-4}$. In addition, exploring all SNPs present on Immunochip identified a SNP showing modest association in the GWAS ($P_{GWAS}=0.0083$) for one other $P_{GWAS}<10^{-4}$ region. Thus in total, SNPs showing association in the GWAS from 32 of 39 loci ($P_{GWAS}<10^{-4}$) are present on Immunochip. Of the other 20 SNPs

showing $P_{GWAS\text{-}preliminary}<10^{-4}$ that mapped to other loci, only 2 showed no association in the final analysis after automated calling and full quality controls. The other 18 SNPs submitted and passing design quality controls on Immunochip mapped to 16 regions of lesser significance ($10^{-4}<P_{GWAS}<0.0073$), with the majority showing $10^{-4}< P_{GWAS}<10^{-3}$. Exploring additional content submitted by other researchers for other phenotypes, a total of 941 Immunochip SNPs had been genotyped in the GWAS and mapped within non-HLA regions showing $P_{GWAS}<10^{-4}$ (defined by extending 500kb from the first and last SNP with $P_{GWAS}<10^{-4}$ in each region). Within the HLA region (Chr 6, 29Mb-34Mb), there are a total of 10,082 SNPs on Immunochip, potentially enabling fine-mapping of the most compelling candidate region identified in the GWAS. **Table 4.5** lists loci with SNPs showing evidence of association $P<10^{-4}$ and SNPs available for follow-up genotyping on Immunochip.

4.9 Discussion

All loci, with SNPs $P_{GWAS}<10^{-4}$, were searched for genes with good functional candidacy for thiopurine-induced pancreatitis. Specifically, genes with known roles in thiopurine metabolism, immune system genes, and genes with known roles in pancreatic function or pancreatitis were searched for (**Table 4.7**). No genes with known roles in thiopurine metabolism or pancreatic function were identified. The top SNP, rs4943552 ($P_{GWAS}=2.46 \times 10^{-6}$ OR= 2.59) mapped to an intergenic region on chromosome 13. The closest gene is *TRPC4,* 60 kilobases downstream. Transient receptor potential cation channel 4 (*TRPC4*), is known to be expressed in the pancreas. It functions as a calcium channel in other tissues, although within the pancreas its role and cellular expression is unknown. Genes regulating calcium entry to pancreatic acinar cells are good candidates for thiopurine-induced pancreatitis. Acinar cell calcium entry promotes trypsinogen activation and has been implicated in the pathogenesis of pancreatitis(Whitcomb). *CASR*, which plays an important role in regulating calcium influx in pancreatic acinar cells has been associated with hereditary and sporadic forms of chronic pancreatitis (**Table 4.7**)(Felderbauer, Hoffmann et al. 2003). Thus, although little is known about *TRPC4* in the pancreas, it is here considered a possible candidate gene for thiopurine-induced pancreatitis, and is the strongest association in the current GWAS.

Among other associations, the strongest biological candidates for a role in thiopurine-induced pancreatitis are the multiple HLA genes within the HLA region on chromosome 6 (**Figure 4.3**). The strongest association here was for rs2647087 ($P_{GWAS}=2.34 \times 10^{-5}$ OR=2.69). HLA gene variants are strong candidates for causative roles in idiosyncratic drug reactions and have been associated with abacavir hypersensitivity and flucloxacillin and co-amoxiclav induced liver injury (Mallal, Nolan et al. 2002; Daly, Donaldson et al. 2009; Pirmohamed 2010). An HLA-restricted immunopathogenesis for thiopurine-induced pancreatitis is consistent with onset a few days to weeks after drug exposure and rapid recurrence on drug re-exposure. The peak association in the HLA region, maps between several HLA genes, the closest of which is *HLA-DQA2*. Extensive linkage disequilibrium in the region has prevented more precise understanding of the *HLA* alleles that might cause this putative association.

4.10 Conclusion

A genome-wide association study was performed to test the hypothesis that common genetic variants of large effect (OR>/=4) influenced the risk of azathioprine and mercaptopurine-induced pancreatitis. The findings do not support this hypothesis, suggesting that any common genetic contributions to risk are more modest and therefore differ from the large effect variants seen for some other idiosyncratic drug reactions (abacavir hypersensitivity, flucloxacillin-induced liver injury). Among the most significant SNP associations in this study, variants in the *HLA* gene region showed moderate association and warrant further study, particularly since HLA variants have been associated frequently with other idiosyncratic drug reactions. Follow-up genotyping of the top SNP associations in this study ($P_{GWAS}<10^{-4}$) is planned in a similar sized case-control collection. It is anticipated that this will provide adequate power to test the validity of SNP associations within the *HLA* and within other moderately associated gene regions. These experiments will take place after this thesis is submitted.

4.11 Methods

4.11.1 Study participants

UK: Cases were recruited from 6 centres in England and 2 in Scotland. All individuals had been prescribed azathioprine or mercaptopurine for Crohn's disease or ulcerative colitis. The diagnosis of azathioprine or mercaptopurine-induced pancreatitis was determined by the recruiting physician using standard clinical, biochemical and radiological criteria for acute pancreatitis. Controls were population individuals genotyped for the Wellcome Trust Case Control Consortium recruited from the National Blood Service or 1958 birth cohort and passing quality controls in a coeliac GWAS as described in chapter 3.

Dutch: Cases were recruited by Dr Rinse Weersma at the University of Groningen, the Netherlands. Azathioprine or mercaptopurine-induced pancreatitis was determined by the recruiting physician using standard clinical, biochemical and radiological criteria for acute pancreatitis. All individuals had been prescribed azathioprine for either Crohn's disease or ulcerative colitis. The control cohort comprised Dutch blood bank donors and NELSON controls passing quality controls for a GWAS in coeliac disease as described in chapter 3. The NELSON project—an ongoing population-based, randomized multi-centre lung cancer screening trial recruits male smokers (van Iersel, de Koning et al. 2007). These controls were collected from the north and centre of the Netherlands (Groningen, Utrecht and Drenthe, The Netherlands).All the control subjects were heavy smokers or ex-smokers (a minimum of 16 cigarettes/day for 25 years or 11 cigarettes/day for 30 years), but did not develop airway obstruction or emphysema suggesting chronic obstructive pulmonary disease (COPD) until the end of a 4 year observation period. The study was approved by the local ethics committees and all the patients and controls gave their written informed consent.

4.11.2 Genotyping

4.11.2.1 GWAS genotyping

GWAS genotyping was performed using Illumina Human 1M-Duo version 3 Beadchips. 800ng genomic DNA (50ng/µl) was used as input and samples genotyped in accordance with the Infinium™ HD Gemini Assay Guide (Revision A). Briefly, samples were whole genome amplified, fragmented, precipitated and re-suspended in hybridization buffer. Re-suspended

samples were hybridized to 1M-Duo Beadchips. Beadchips were washed. Single base extension and staining steps were performed. Beadchips were loaded onto an Illumina iScan System, containing a two channel laser imager and intensity data subsequently generated using Illumina BeadStudio 2.0 software. All steps other than scanning were performed at The Genome Centre (Barts and the London School of Medicine and Dentistry). Scanning was performed at the University College London Microarray centre (institute of Child Health, London).

4.11.2.2 Singleton SNP repeat genotyping

Genotyping of 7 "singleton SNPs" was performed using fluorescence-based allele specific (KASPar™) custom-designed SNP genotyping assays (KBiosciences, Hoddeston, UK). SNP DNA sequences were submitted to KBiosciences for SNP assay design. The KASPar SNP genotyping system is a competitive allele specific PCR assay. The assay uses allele specific primer pairs, with one common (reverse) primer shared for both SNP alleles (i.e. 3 primers in total). The primers are used in a PCR reaction with FAM and VIC fluor labelling of the allele-specific PCR products. PCR amplification was performed with a PTC-225 Peltier Thermal Cycler (MJ Research, USA) with thermocycler conditions as specified in the KASPar SNP Genotyping System Reagent Manual (KBioscience, UK). A total of 25ng DNA was used as input (5µl at 2.5ng/µl for each duplicate reaction). An ABI Prism 7900HT Sequence Detection System plate reader was used to measure fluorescence data, with excitation and emission values as stated in the KASPar SNP Genotyping System Reagent Manual (KBioscience, UK). KASPar genotyping was performed at The Genome Centre (Barts and the London School of Medicine and Dentistry) according to the manufacturer's instructions (KASP genotyping QuickStart Guide). Control samples used for KASPar genotyping were blood-extracted DNA samples from 94 individuals with coeliac disease that had been included in the UK1 sample collection for the coeliac GWAS **Chapter 3**). Samples were arrayed in 384 well plates and assayed in duplicate. Genotypes were called by visual inspection of SNP cluster plots (plotting intensities from both fluor- labelled PCR products) and manual adjustment of genotype cluster positions.

4.11.3 Statistical analysis and bioinformatics resources

Quality controls and most case-control association analyses were performed in PLINKv1.07 (Purcell, Neale et al. 2007). Meta-analysis of association p values from UK and Dutch

collections was performed using METAL software designed by Gonzalo Abecasis and colleagues (www.sph.umich.edu/csg/abecasis/metal/).

Exact *P* values for Cochran-Armitage trend tests were calculated using StatXactv9 (Cytel Inc., Cambridge, MA, USA), which uses a proprietary permutation-based method to determine exact test statistics.

Proxy SNPs genotyped in HapMap CEU samples were explored using the SNP annotation and search facility (SNAP v2.1, Broad Institute, Boston, MA, USA; http://www.broadinstitute.org/mpg/snap/). Other bioinformatics resources are listed in **Chapter 3.8.**

Chapter 5 Functional investigation of Crohn's disease-associated single nucleotide polymorphisms at 5p13.1

This work was started in December 2007 and completed in September 2008.

5.1 Introduction

5.1.1 SNPs in a gene desert on chromosome 5 (5p13.1) are associated with Crohn's disease, ulcerative colitis and multiple sclerosis

SNPs in a 250 kilobase (kb)region of strong linkage disequilibrium on chromosome 5p13.1 were first reported to show association with Crohn's disease in a genome wide-association study (GWAS) of 547 Crohn's disease (CrD) cases and 928 controls from Belgium and France (Libioulle, Louis et al. 2007). The association was strongly replicated in later European and North American Crohn's disease GWASs (Hampe, Franke et al. 2007; Rioux, Xavier et al. 2007; WTCCC 2007). In the largest meta-analysis of Crohn's disease GWASs reported to date, SNPs in this region showed the third strongest association with Crohn's disease after the *IL23R* and *ATG16L1* gene regions (top SNP = rs4613763, $MAF_{controls}$ 0.125, OR 1.32, $P_{Combined}= 6.82 \times 10^{-27}$) (Barrett, Hansoul et al. 2008). McGovern et al. found moderate association for rs4613763 with ulcerative colitis (UC) in a meta-analysis and replication study of 4702 UC cases and 8371 controls ($P_{Combined}= 4.2 \times 10^{-4}$). While this suggests that the 5p13.1 locus may be a shared IBD susceptibility region, effects are currently more clearly established for Crohn's disease than UC. Of interest, McGovern et al. observed ulcerative colitis association in a region on chromosome 1p36, containing *PLA2G2E*, a secretory form of phospholipase A2 involved in prostaglandin synthesis, providing additional indirect support for the importance of prostaglandin signalling in inflammatory bowel disease. Crohn's disease associated SNPs at 5p13.1 have also been associated with multiple sclerosis susceptibility in a GWAS meta-analysis, indicating that this is not a Crohn's disease or IBD-specific locus (rs9292777 $P_{MS\text{-}meta}=2.2 \times 10^{-7}$) (De Jager, Jia et al. 2009). This suggests that the relevant biological perturbations are in (probably immune) pathways common to the pathogenesis of both multiple sclerosis and Crohn's disease.

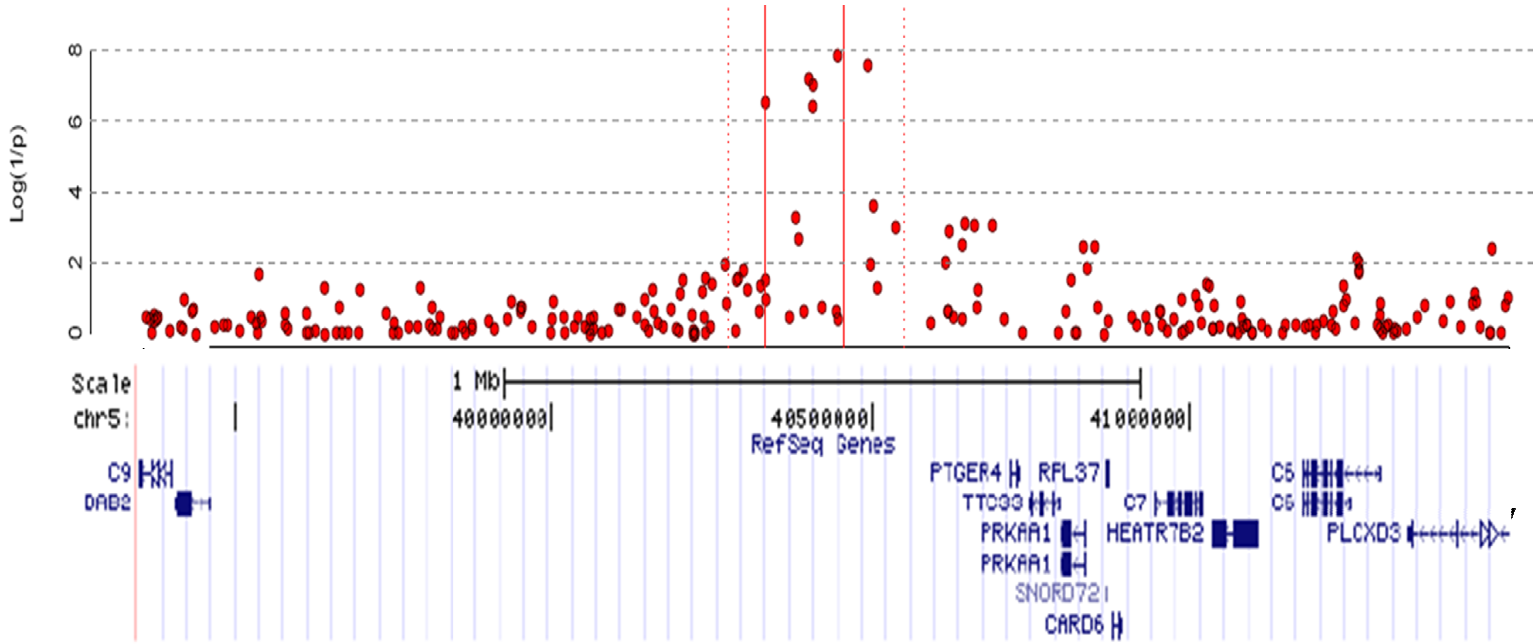
No known genes map within the 250 kb linkage disequilibrium block showing disease association- this was one of the first observations that genetic variants in regions devoid of known genes could be associated with common diseases. A hypothesis that has been widely

advanced to explain disease associations in these so-called gene deserts is that the causal variants driving these associations affect regulatory elements in the genome that alter local or distant gene expression (McCarthy, Abecasis et al. 2008). Regulatory sequences in the human genome are relatively poorly annotated, despite ongoing efforts (e.g. ENCODE) and therefore bioinformatic approaches are not yet robust in predicting whether SNPs will impact on regulatory functions in the way that protein-coding alterations can be predicted from exonic SNPs (Birney, Stamatoyannopoulos et al. 2007). Libioulle et al. tested for correlations between SNPs in the 5p13.1 linkage disequilibrium block and local gene expression in lymphoblastoid cell lines generated from 378 individuals. Prostaglandin E Receptor 4 (*PTGER4*) is the closest known gene in the region, 270kb away from the association peak, although several other genes reside nearby, including some with other known immune functions (*CARD6, C5, C6, C7*) (**Figure 5.1)**. A strong *cis*-expression Quantitative Trait Locus (*cis*-eQTL) with *PTGER4* expression was observed for several SNPs in the region including some of the most strongly Crohn's disease-associated SNPs. Precise co-localisation of the peak Crohn's disease association and peak expression correlation association was not observed. The top *cis*-eQTLs in the region were observed for rs7720838 and rs4495224, 55.7 Kb distant and 46.3 Kb distant from rs1373692, the strongest CrD-associated SNP. Rs7720838 and rs4495224 show only modest LD with rs1373692, the top CrD associated SNP in this study (rs7720838 r-sq=0.24; rs4495224 r-sq=0.45). However the two top eQTL SNPs did show evidence of CrD association (rs4495224 P_{GWAS} =2.2 x 10^{-7}) and are correlated (r-square= 0.67, D'= 0.85 HapMap CEU). Correlation of CrD-associated 5p13.1 SNPs with the expression of other genes in the region was not observed.

These observations supported the hypothesis that the 5p13.1 Crohn's disease association was due to regulation of expression of *PTGER4* and the assignment of *PTGER4* as the causal gene influencing Crohn's disease susceptibility in this region has been widely adopted (Van Limbergen, Wilson et al. 2009; Perdigones, Martin et al. 2010). However, while this evidence was suggestive, it was also clear that correlations of SNP genotypes with local gene expression are much more prevalent in the human genome than was previously appreciated and affect around 30% of loci in the human genome (Ge, Pokholok et al. 2009; Pastinen 2010). In a large eQTL analysis of human peripheral blood samples (discussed in **Chapter 3**) 18% of randomly selected hap300 SNPs had evidence of correlation with local gene expression (Dubois, Trynka et al. 2010). Furthermore, it has been recently proposed that GWAS associations may arise from rare protein-coding causal variants residing in genes at distances >1Mb away, beyond the

boundaries of local linkage disequilibrium of the GWAS association (Dickson, Wang et al. 2010). The proposed mechanism is that GWAS associations could arise due to 'loading' of rare variants on an extended haplotype (sometimes spanning one or two megabases) of recent ancestry. While this proposal is controversial and awaits empirical confirmation, clearly the possibility that the *PTGER4 cis*-eQTL reflects a causal mechanism driving the 5p13.1 Crohn's disease association requires further evidence. Such evidence should include firstly, a demonstration that *PTGER4* is a compelling biological candidate for Crohn's disease pathogenesis; secondly a demonstration that the disease-associated SNP risk alleles correlate with changes in prostaglandin E receptor 4 signalling that promote intestinal inflammation and Crohn's disease and thirdly the identification of the causal genetic variants, and understanding of how these influence regulatory mechanisms causing changes in expression. As a first step towards testing these suppositions, the experiments presented in this chapter aimed to test whether CrD SNPs in the region correlated with prostaglandin E receptor 4 signalling activity in primary human immune cells.

Figure 5.1 **SNP associations (-Log10(P)) in a 1.8 Mb region (Chr 5p13.1) from the Libioulle et al. genome wide association study of 547 Crohn's cases and 928 controls. Figure adapted from Libioulle et al, 2007**



Dashed vertical lines delimit ~250Kb region of linkage disequilibrium, solid vertical lines delimit region of stronger LD within this. Genes with known immunological function include *C5, C6, C7, C9* (Complement component precursor genes), *CARD6* (Caspase recruitment domain 6, a positive regulator of NF-κB) and *PTGER4.*

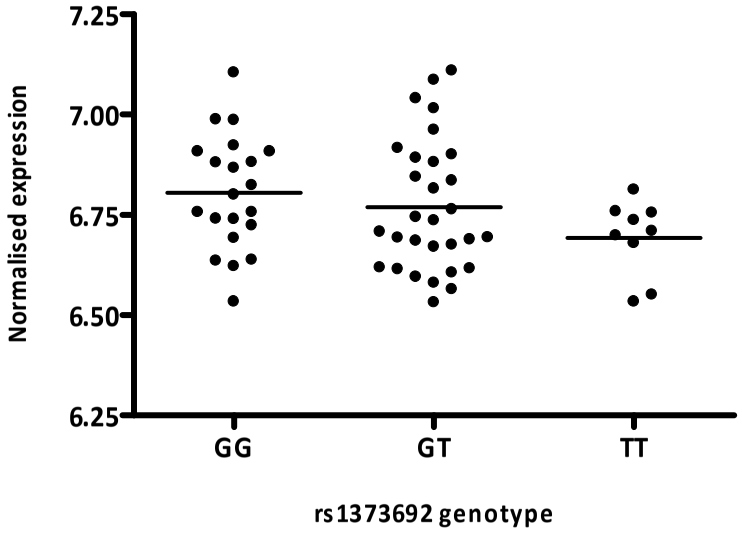
5.1.2 Crohn's disease associated SNPs correlate with expression of *PTGER4*

It was hypothesized that CrD-associated variants in this region exert susceptibility effects by increasing expression of *PTGER4* and augmenting prostaglandin E_2 – EP4 mediated signalling in immune cells.

Gene expression correlations with Crohn's disease associated SNPs were assessed in an independent publically available dataset of array-based gene expression data from lymphoblastoid cell lines (LCLs) generated from 90 HapMap CEU individuals (Stranger, Nica et al. 2007). Several of the most strongly associated SNPs identified in Crohn's disease GWASs correlate with *PTGER4* expression levels, with the CrD risk allele associated with higher gene expression (**Figure 5.2**). There was no evidence of correlation of these variants with levels of expression of the other genes in the region (data not shown).

Figure 5.2 ***PTGER4*** **expression in lymphoblastoid cell lines from 90 HapMap CEU individuals by SNP genotype**

A. **rs1373692 (*P* = 0.039 GG vs TT). Top ranked 5p13.1 SNP in Libioulle et al. GWAS (P_{GWAS}=4.1 x 10^{-8}). G is CrD risk allele**



B. **rs1002922 (*P* = 0.032 CC vs TT). 2nd ranked SNP in Libioulle et al. GWAS (P_{GWAS}=9.1 x 10^{-8}). T is CrD risk allele**

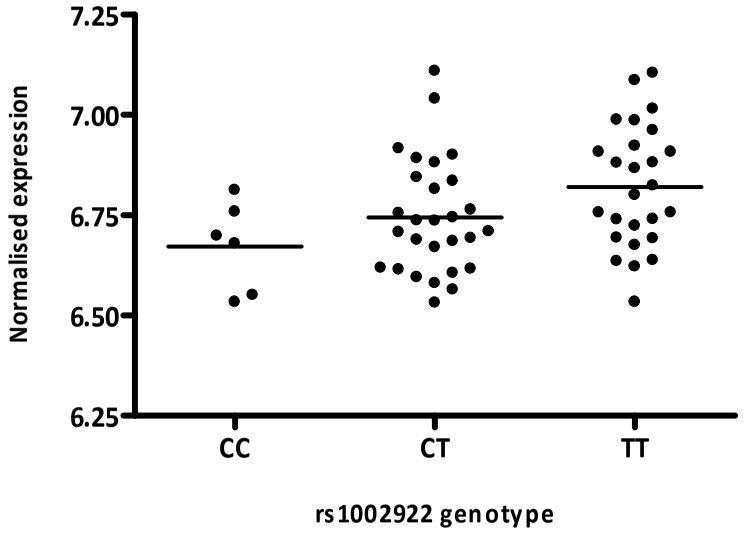
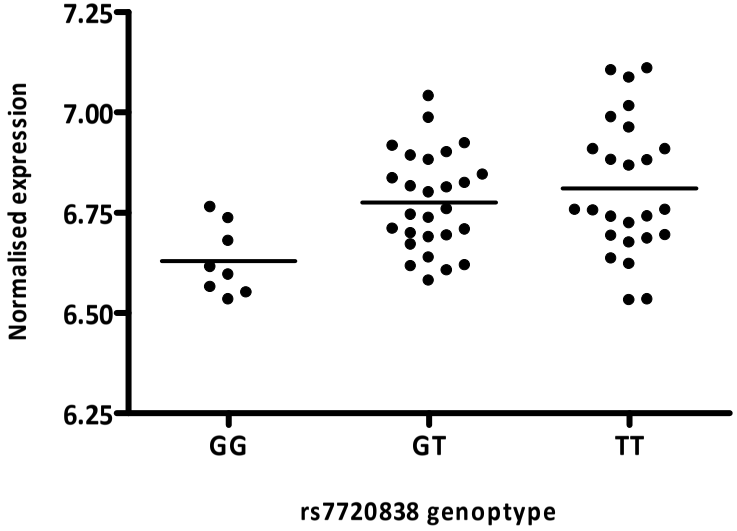
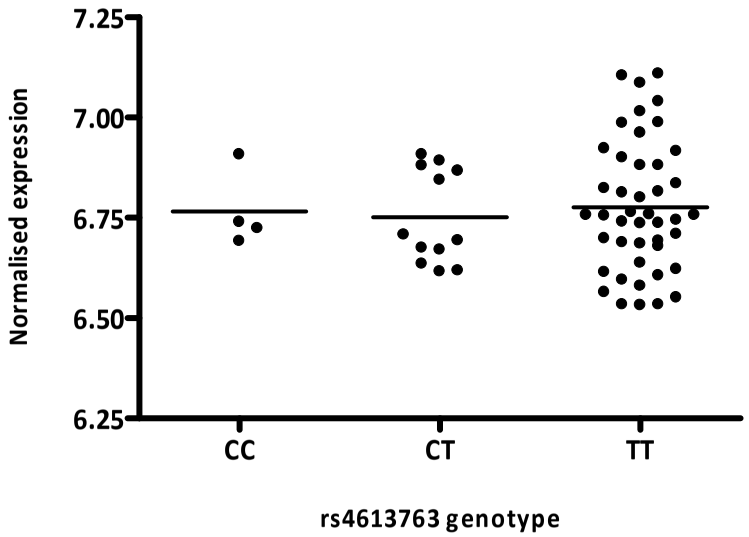


Figure 5.2. *PTGER4* expression in lymphoblastoid cell lines from 90 HapMap CEU individuals by SNP genotype.

C. rs7720838 (*P* = 0.0073 GG vs TT). Strongest 5p13.1 SNP showing association with *PTGER4* expression in Libioulle et al. T is CrD risk allele



D. rs4613763 (*P* = ns). Top ranked 5p13.1 SNP in Barrett et al. GWAS meta-analysis ($P_{meta+replication}$= 6.8 x 10^{-27})



A meta-analysis dataset of whole-genome (array-assayed) gene expression measurements for 1,469 whole blood samples (reflecting mainly leucocyte gene expression) was also searched (Dubois, Trynka et al. 2010). Genotypes were available for 18 SNPs in the 260Kb CrD-associated LD block (Chr 5, 40.32-40.48Mb as defined in Barrett et al. (Barrett, Hansoul et al. 2008)). These SNPs had been genotyped in all 1,469 samples on the Illumina Hap300 platform. This analysis dataset included gene expression measurements from primary human leucocytes and may therefore be less susceptible to artefacts present in immortalised cell lines. Such artefacts can arise due to variables in the immortalisation of B cells used to generate cell lines: the subpopulation of B cells chosen may influence cell line phenotype including gene expression, the amount of and individual response to EBV virus varies, as can the history of cell culture conditions for each cell line. This has led to concern that increased variability in gene expression between cell lines may arise from these factors and confound efforts to assess and detect QTLs (Choy, Yelensky et al. 2008). The peripheral blood meta-analysis dataset had substantially higher sample size than the LCL datasets which ought to confer much greater power to detect eQTLs in the region.

This eQTL meta- analysis was performed by Lude Franke (University of Groningen, the Netherlands). Within this dataset SNP genotype - *cis* gene expression correlations for genes residing within 500 kb of each tested SNP were calculated using Spearman's rank correlation. After correction for multiple testing (equal to false discovery rate of 0.05), 5 out of 18 SNPs in the region showed significant correlations with *PTGER4* expression (**Table 5.1**). However, SNPs did not show correlation with expression of any other genes within 500kb.

The strongest eQTL was observed for rs12514679 with *PTGER4* (P=8.06 x 10^{-7}; Spearman rank correlation meta-analysis P value; **Table 5.1**). The second strongest eQTL was observed for rs10512734 (P=9.39 x 10^{-5}), which shows moderate linkage disequilibrium with the top disease-associated SNP reported by Libioulle et al. (rs1373692, r-square = 0.75, D'=1.00 in HapMap CEU) and strong association in a recent updated meta-analysis of Crohn's GWASs (**Table 5.1**). Rs1373692 itself showed moderate correlation with *PTGER4* expression (P= 0.00659), that just fails correction for multiple testing (FDR=0.05). The most strongly associated Crohn's disease SNP from the Barrett et al. meta-analysis (rs4613763) showed no correlation with expression of *PTGER4* (P= 0.68) or any other genes within 500kb in this analysis supporting the finding observed in the Stranger data. The peripheral blood eQTL meta-analysis data therefore support the LCL data of CrD SNPs at 5p13.1 correlating with

PTGER4 expression. Moreover they add weight to the observation that not all of the CrD SNPs correlate with *PTGER4* expression. These features underline the potential complexity of genetic effects on gene expression. The absence of precise co-localisation of the expression correlation and disease association signals may be indicative of the independence of these two phenomena (i.e. the observed partial co-localisation of these associations is co-incidental and disease association is actually driven by mechanisms other than regulation of *PTGER4* expression). Alternatively, stochastic variation in the association signals (generated from two different sets of individuals) might cause variation in peak associations. Such complexity reinforces the value of looking beyond gene expression at gene function. By assaying *PTGER4* gene function, effects of disease associated SNPs on function could be tested more directly.

Table 5.1 Illumina Hap300 SNPs from the 250 kb region on 5p13.1 associated with Crohn's disease showing significant correlation with *PTGER4* expression in whole blood samples from 1469 individuals (Dubois, Trynka et al. 2010).

SNP	SNP position[a]	Probe centre position[b]	CrD risk allele expression effect[c]	eQTL P value[d]	CrD assoc P value[e]
rs12514679	40402560	2940438	unknown	8.06×10^{-7}	unknown
rs10512734	40429362	2940438	Increase	5.39×10^{-5}	2.9×10^{-29}
rs1002922	40422312	2940438	Increase	1.23×10^{-4}	5.4×10^{-29}
rs7725523	40407980	2940438	Increase	5.06×10^{-4}	3.5×10^{-18}
rs6880934	40464949	2940438	Increase	6.11×10^{-4}	0.023

[a]Base pair coordinates on chromosome 5, NCBI build 36

[b] Probe centre position was determined by re-mapping probe sequences to the human transcriptome and calculated from the mid-point of the transcript start and transcript end positions in genomic co-ordinates

[c]Direction of effect of Crohn's disease risk allele on *PTGER4* expression

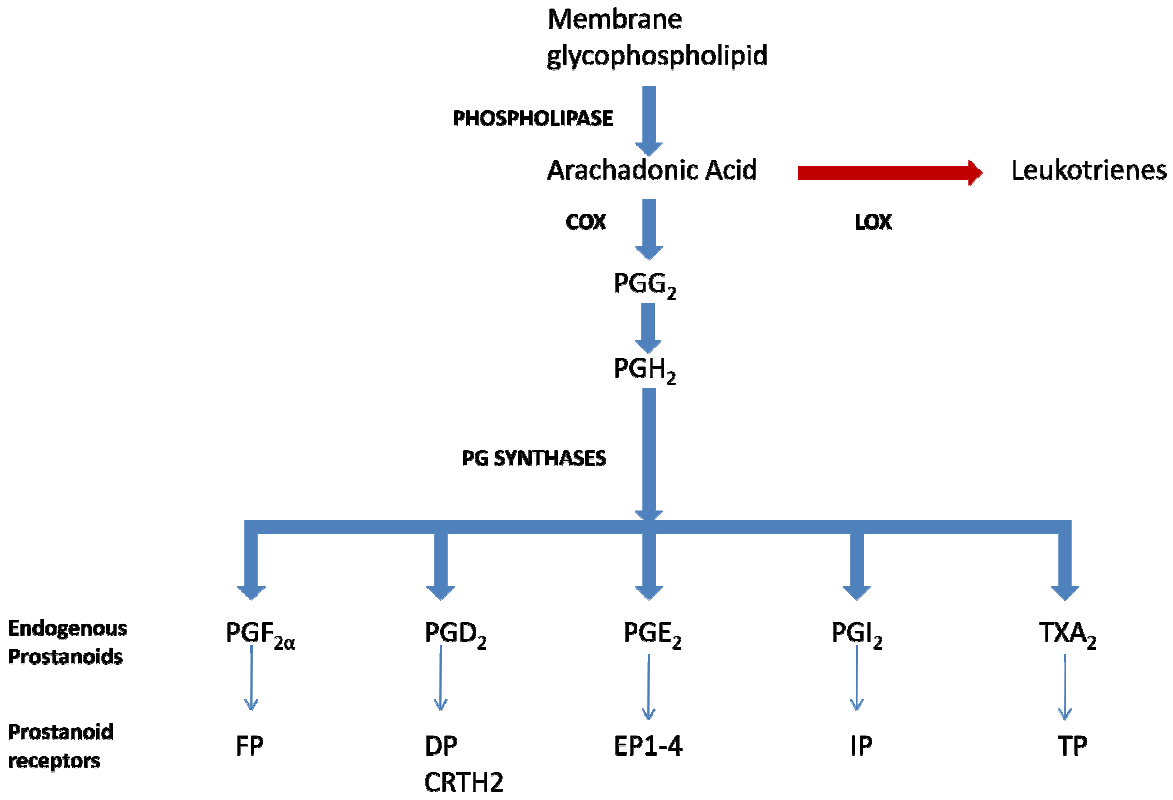
[e]Crohn's disease case-control association p value from updated meta-analysis of 3 GWASs (personal communication, Jeff Barrett & (Barrett, Hansoul et al. 2008))

5.1.3 Overview of prostaglandins

Prostaglandins are so-called following their original identification in 1935 in seminal fluid, believed at the time to represent secretions of the prostate gland (Goldblatt 1935). Prostaglandins are a family of lipid mediators within the broader class of eicosanoids. All eicosanoids are formed by the oxidation of 20-carbon essential fatty acids released from plasma and nuclear membranes by the action of phospholipases. The oxygenation step is

catalysed by cyclooxygenase (COX) to generate prostanoids or by lipoxygenase (LOX) to form leukotrienes. Prostaglandins are one family of prostanoids, the others being thromboxanes and prostacyclins (**Figure 5.3**). The prostanoids produced from arachidonic acid are termed series 2 prostanoids, and are the dominant endogenous prostanoids. Series 1 and series 3 prostanoids (e.g. PGE_1, PGE_3) are generated from alternative essential fatty acids (gamma-linolenic acid and eicosapentaenoic acid respectively).

Figure 5.3 Prostanoid synthesis from membrane phospholipids and their receptors



COX Cyclooxygenase; LOX Lipoxygenase; PGI_2 Prostacyclin; TXA_2 Thromboxane; CRTH2 Chemoattractant receptor homologous molecule expressed on Th2 cells

Prostaglandins can be synthesized in nearly all nucleated cells studied, with the possible exception of lymphocytes (Simmons, Botting et al. 2004). Endogenous synthesis depends on constitutively expressed COX1 and inducible COX2, a key enzyme in inflammation. After synthesis within the cytoplasm, prostaglandin release from cells is mediated at least in part by a specific transporter, multidrug resistance protein 4 (MRP4). Cellular uptake is also active and mediated by the prostaglandin transporter (*SLCO2A1*). Among immune cell types, monocytes

and macrophages can produce large quantities of PGE_2, moderate amounts are released from neutrophils in response to inflammatory stimuli, but no PGE_2 was detected from lymphocytes despite expression of COX isoenzymes in these cells (Pablos, Santiago et al. 1999; Simmons, Botting et al. 2004). PGE_2 is also synthesized throughout the gastrointestinal tract (Simmons, Botting et al. 2004). Prostaglandin signalling depends both on the pattern of tissue expression of terminal prostaglandin synthases and prostanoid receptors (**Figure 5.3**). Prostaglandin E_2 is the endogenous ligand of the prostaglandin E receptor 4 (EP4), one of four G-protein coupled E prostanoid receptors. Prostaglandin E_2 (PGE_2) can have both pro- and anti-inflammatory effects depending on receptor expression and cell type(Hata and Breyer 2004).

5.1.4 Prostanoid receptors

5.1.4.1 Prostaglandin EP receptors - pharmacology

PTGER4 encodes the prostaglandin EP4 receptor, one of four human prostaglandin E_2 receptors (EP1-4). The EP4 receptor is a 488 amino acid, plasma membrane localised, G-protein coupled receptor linked to adenylate cyclase. Both G_s and G_i can couple the EP4 receptor to adenylate cyclase although the most frequent effect in most cell types studied is coupling via G_s with stimulation of adenylate cyclase and an intracellular rise in cyclic AMP (Narumiya, Sugimoto et al. 1999). Some studies have also shown that the EP4 receptor can activate an alternative second messenger enzyme, phosphatidylinositol 3 kinase (Fujino, Xu et al. 2003).

Of the endogenous prostanoids, prostaglandin E_2 has by far the strongest affinity (receptor association constant < 1nM) for the EP4 receptor (Abramovitz, Adam et al. 2000). The relative potencies of other endogenous prostanoids (PGD_2, PGI_2, $PGF_{2\alpha}$) at the EP4 receptor are around 1000-10,000fold lower (Wilson, Rhodes et al. 2004). However, PGE_2 is not a selective EP4 agonist since it binds all EP receptors (EP1-4) with high affinity. Radioligand binding assays using human embryonic kidney stem cells selectively expressing EP receptor subtypes show that PGE_2 binds all human EP receptor subtypes with high affinity in the low nanomolar range (Abramovitz, Adam et al. 2000). The affinity of PGE_2 for EP receptors is in the following order EP3>EP4>EP2>EP1 (Abramovitz, Adam et al. 2000). Pharmacological assays of EP4 function (cAMP assays) in similar receptor-subtype specific expressing cell lines suggest that the full range of pharmacological response occurs over 10^{-11}-10^{-9} Molar range of PGE_2 concentrations (Wilson, Rhodes et al. 2004). In this study the potency of PGE_2 at the EP4 receptor (EC_{50}:

concentration at which 50% of maximal pharmacological response is observed) was 5 x 10^{-11} Molar. At the EP2 receptor, PGE_2 had a 1000 fold lower potency (3 x 10^{-8} Molar). Whether these pharmacological data from recombinant expressed receptors in embryonic stem cells hold in primary human cells is uncertain. While these data suggest that PGE_2 is a more potent agonist of the EP4 receptor than the EP2 receptor, the degree to which the physiological effects of PGE_2 are mediated by EP4 versus other EP receptors in primary human cells may be influenced by many other factors, including cell type and importantly the relative expression levels of different EP receptor subtypes. Of the other EP receptors, the most important potential confounder of EP4 signalling is the EP2 receptor. This receptor also couples to adenylate cyclase, leading to increased cAMP on receptor activation (Narumiya, Sugimoto et al. 1999). This contrasts with the second messenger signals for EP1 (increased calcium) and EP3 (reduced cAMP). Of all the EP receptor subtypes, the EP2 receptor shows the most similar pattern of tissue expression compared to the EP4 receptor and is co-expressed in many of immune cells (see below).

5.1.4.2 Prostaglandin EP receptor: tissue expression

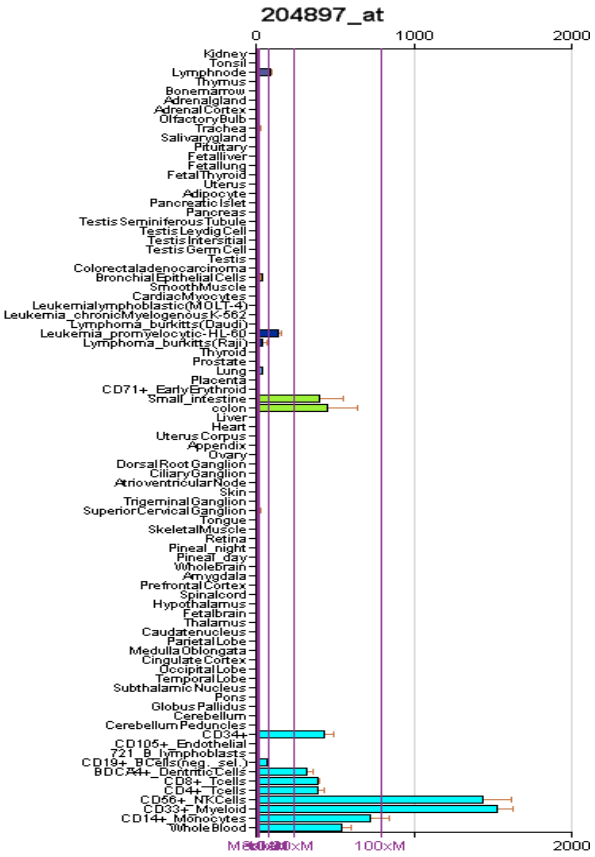
PTGER4 is expressed in myeloid and lymphoid cell lineages and intestinal epithelium (**Figure 5.4**) (Cosme, Lublin et al. 2000; Su, Wiltshire et al. 2004). *PTGER4* is moderately expressed in PBMCs (Mori, Tanaka et al. 1996). It is also expressed in intestinal cells throughout the gastrointestinal tract, and is thought to be the subtype responsible for intestinal chloride secretion in response to PGE_2 (Bukhave and Rask-Madsen 1980; Narumiya, Sugimoto et al. 1999). *PTGER4* is the only EP receptor gene showing significant expression in intestinal epithelium. *PTGER1* and *PTGER3* are not strongly expressed in leucocytes or intestine and are therefore unlikely to contribute to PGE_2 signalling in inflammation (Su, Wiltshire et al. 2004). In contrast, *PTGER2* is co-expressed with *PTGER4* in leucocytes (**Figure 5.4)**, and is therefore the likely major alternative E-prostanoid signalling mechanism in immune cells**.** *PTGER4* is basally expressed at higher levels in most immune cell types than *PTGER2* (Takayama, Garcia-Cardena et al. 2002). In human lamina propria mononuclear cells (LPMCs) and intestinal epithelial cells EP4, but not EP1-3, is present by immunostaining and northern blot (Cosme, Lublin et al. 2000). EP receptor expression is also influenced by inflammatory stimuli. For example in intestinal epithelium, *PTGER4* expression was higher in ulcerative colitis inflamed colonic tissue than in uninflamed controls (Cosme, Lublin et al. 2000). Conversely, EP2 receptor expression has been shown to increase in mouse peritoneal macrophages in response to

lipopolysaccharide, while EP4 expression was suppressed by LPS (Ikegami, Sugimoto et al. 2001).

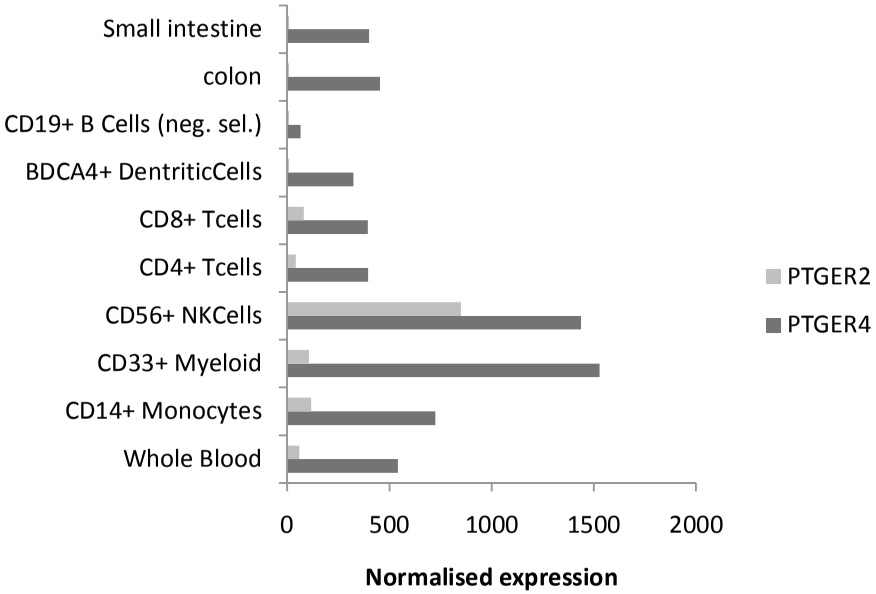
As PGE_2 can act as an agonist of all EP receptor subtypes, and as (at least) EP4 and EP2 receptors are co-expressed in primary human immune cells, the effects of PGE_2 on these cells are likely to represent a combination of EP2- and EP4- mediated signals. Moreover both receptors signal through adenylate cyclase with increased cAMP as a second messenger and might therefore mediate similar cell responses. Selective EP4 receptor agonists and antagonists should offer advantages over PGE_2 in assaying specific EP4 receptor-mediated functions.

Figure 5.4 **Human tissue distribution of gene expression for E-prostanoid receptors EP2 and EP4 assayed by microarray profiling (data from Novartis gene atlas (Su, Wiltshire et al. 2004)) A. Normalised expression of *PTGER4* in human tissues. B. Expression of *PTGER2* and *PTGER4* compared in myeloid and lymphoid lineages and intestine**

A. *PTGER4*



B. *PTGER2* and *PTGER4*



***PTGER1* and *PTGER3* not expressed at detectable levels in leucocytes or intestine**

5.1.4.3 Prostaglandin EP4 Receptor Function

PGE_2 has historically been considered a predominantly a pro-inflammatory mediator (Narumiya, Sugimoto et al. 1999; Simmons, Botting et al. 2004). In rheumatoid arthritis, PGE_2 is present in synovial fluid, and has pro-inflammatory effects at least in part dependent on the EP4 receptor (McCoy, Wicks et al. 2002; Karouzakis, Neidhart et al. 2006; Sheibanie, Yen et al. 2007). Anti-PGE_2 antibodies suppress inflammation in a rat model of arthritis (Portanova, Zhang et al. 1996). In humans COX inhibitors (non-steroidal anti-inflammatory drugs- NSAIDs) are established therapies that reduce joint inflammation, partly through reduction in PGE_2 (Simmons, Botting et al. 2004). However, NSAIDs are relatively contra-indicated in inflammatory bowel disease due to concerns that they can exacerbate intestinal inflammation. This suggests that PGE_2 may have predominantly anti-inflammatory effects in the intestine (Wang and Dubois 2010). Reconciling these opposing effects has proved difficult, but there is evidence that the EP4 receptor may mediate pro- or anti- inflammatory effects depending on the tissue cellular context.

Mouse studies highlight the potential for both pro- and anti- inflammatory effects of EP4 receptor activation. EP4 knockout mice show disrupted intestinal epithelial repair with increased susceptibility to DSS colitis(Kabashima, Saji et al. 2002). Similarly, EP4 agonists appear to protect against epithelial disruption in the DSS model suggesting that EP4 signalling may be necessary for the maintenance and repair of intestinal epithelium(Jiang, Nieves et al. 2007).

In contrast, selective EP4 agonists in wild-type mice can exacerbate the TNBS colitis model of intestinal inflammation (Sheibanie, Yen et al. 2007). TNBS does not directly damage the colonic epithelium, but rather induces a chronic immune cell infiltration of colonic mucosa, and it may be that the key effects here are on immune cells rather than mechanisms for maintaining intestinal epithelial integrity. In an important study by Sheibanie et al., EP4 receptor stimulation in LPS-treated dendritic cells promoted interleukin 23 (IL-23) secretion and inhibited IL-12/IL-27 secretion. This was observed to lead to induction of IL-17 in activated T cells (Sheibanie, Yen et al. 2007). More recently Yao et al. found that PGE_2-EP4 signalling can promote Th-17 cell amplification via IL-23, but also that EP4 agonists could promote Th1 differentiation(Yao, Sakata et al. 2009). In this study the administration of EP4 antagonists suppressed disease progression in experimental autoimmune encephalitis. Chen et al. observed similar effects of EP4 stimulation on Th1 differentiation, IL-23 secretion from

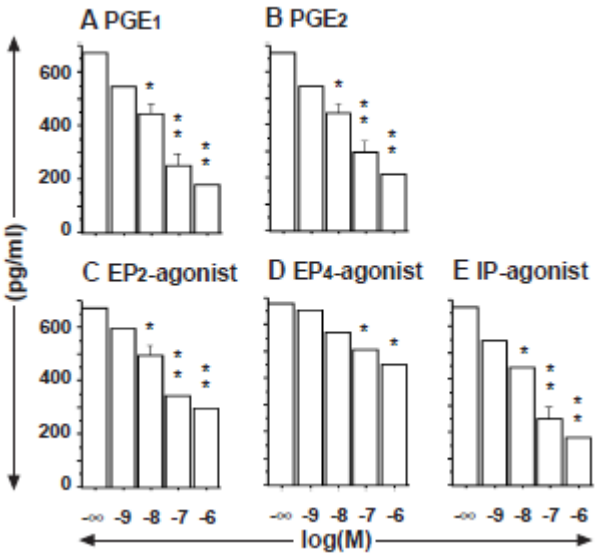
dendritic cells and Th17 cell expansion (Franke, de Kovel et al. 2008). A novel selective EP4 antagonist suppressed these effects and suppressed disease in mouse model of rheumatoid arthritis. Together these studies suggest both a possible link to IL-23/Th17 signalling, clearly of high relevance to Crohn's disease and also a mechanism of pro-inflammatory effect for EP4 signalling. Other pro-inflammatory effects attributed to EP4 signalling include stimulation of release of the pro-inflammatory cytokines interleukin-6 (IL-6) and interleukin-8 (IL-8) from macrophages and monocytes and T cells by EP4 agonists (Standiford, Kunkel et al. 1992; Caristi, Piraino et al. 2005; Maloy and Powrie 2005).

Countering the reported pro-inflammatory effects of EP4 signalling in human leucocytes have been observations that EP4 agonists can suppress production of pro-inflammatory cytokines from peripheral blood mononuclear cells (PBMCs), monocytes and macrophages in certain settings. These reports have demonstrated suppression of cytokines from lipopolysaccharide co-incubated PBMCs, monocytes and macrophages (TNFα, IFNγ, IL-12, IL-18) by both PGE_2 and selective EP4 agonists (van der Pouw Kraan, Boeije et al. 1995; Meja, Barnes et al. 1997; Takayama, Garcia-Cardena et al. 2002; Takahashi, Iwagaki et al. 2005; Takahashi, Iwagaki et al. 2005). Selective EP4 agonists used in these experiments reproduce part of the effect of PGE_2 though EP2 and IP agonists also reproduced the effect (**Figure 5.5**). With regard to T cells, Okano et al. showed that both EP2 and EP4 agonists suppressed T cell proliferation and cytokine release in antigen-stimulated CD4 T cell lines (Okano, Sugata et al. 2006).

It was hypothesized that the dominant pathogenic effects of CrD genetic variants on EP4 signalling occur in immune cells rather than intestinal epithelium or other tissues. This hypothesis has been supported (since design of the study) by the finding that 5p13.1 Crohn's disease associated SNPs are also associated with multiple sclerosis (De Jager, Jia et al. 2009). This shared disease risk implicates mechanisms common to both disorders (i.e. the immune system). Therefore, the (mainly anti-inflammatory) effects of EP4 signalling on intestinal epithelium are less likely to be relevant to the genetic association. The link of EP4 receptor signalling to the IL23R pathway and Th17 cell pathways also suggests this is the key mechanism, since multiple Crohn's disease variants are known to affect this pathway (Barrett, Hansoul et al. 2008). Finally, the risk alleles of 5p13.1 SNPs correlate with increased *PTGER4* expression. If this confers an increase in EP4 signalling, as hypothesized here, the pro-inflammatory effects of EP4 signalling on immune cells via Th1/Th17 promotion provide a comparatively attractive model for Crohn's disease as opposed to the promotion of intestinal

epithelial repair anticipated for increased EP4 signalling in the intestinal epithelium. Mechanisms of EP4 signalling are clearly complex and not yet well understood, and furthermore gain of EP4 receptor function in immune cells could promote pro or anti-inflammatory effects depending on factors such as immune cell type, cellular differentiation and cytokine environment. Thus teasing out the relevant effects for Crohn's disease is a formidable challenge.

Figure 5.5 **PGE_1, PGE_2 and selective prostanoid receptor agonists suppress TNF-α from PBMCs incubated with Lipopolysaccharide. Figure reproduced from Takahashi et al. *Eur J. Pharm* 2005**



Prostanoid concentration on x axis. TNF-α concentration in cell supernatants assayed by ELISA (y axis). ONO-AE1-329 was used as the EP4 agonist. PBMCs (1 x 10^6/well) incubated for 48 hours with LPS (1ng/ml) and prostanoid agonists.

The first aim of this study was to develop a cytokine assay of EP4 function in human peripheral blood mononuclear cells (PBMCs). The aim was not to model precisely EP4 receptor signalling events as they might contribute to Crohn's pathogenesis, but rather to develop as selective an assay of EP4 receptor function in primary human immune cells as possible. Once optimized, this assay could be used to measure EP4 function in PBMCs from healthy individuals and

stratify responses by CrD-associated SNP genotype. This would enable a test of the hypothesis that Crohn's disease-associated 5p13.1 SNPs correlate not only with increased *PTGER4* expression but also with *PTGER4* function. A comparison of the CrD-associated SNPs at 5p13.1 that correlate with *PTGER4* expression with the 5p13.1 CrD-associated SNPs that do not should aid this investigation and help determine whether non-expression correlated SNPs are associated with any alteration in EP4 function. The absence of such an effect could suggest that the CrD association in this region is driven by effects on other genes than *PTGER4*.

5.1.5 Sample size calculation for SNP genotype-EP4 receptor function correlation experiments

The number of randomly selected population individuals required to test the hypothesis that CrD-associated 5p13.1 SNPs correlated with EP4 receptor function was considered prior to the study. The sample size calculation depends particularly on the frequencies of genotypes for the CD- associated SNPs and the expected effect size of the variants on function in this assay. rs9292777 had the second strongest association in the largest GWAS in CrD (the Welcome Trust Case Control Consortium- WTCCC) and has also been associated with multiple sclerosis. This SNP is a near-perfect proxy for the top SNP in the Libioulle et al. GWAS (rs1373692, r-square = 1 in 90 HapMap CEU samples), but is in weak LD (r-square = 0.13) with the top 5p13.1 SNPs from the WTCCC GWAS (rs17234657) and Barrett et al. meta-analysis (rs4613763). The CrD-associated SNPs appear to tag two distinct regions of LD, with the former rs1373692-rs9292777 association being that which is correlated with *PTGER4* expression. The minor allele frequency for rs9292777 was 0.394 in British population controls used in the WTCCC study (Parkes, Barrett et al. 2007). The second region of LD associated with CrD comprises the peak associations at rs17234657 and rs4613763, which are in perfect LD in the HapMap CEU collection and have minor allele frequencies of 0.125 in the WT British population controls. Since the rs1373692-rs9292777 SNPs correlate with *PTGER4* expression, these SNPs were of major interest for the *PTGER4* functional assay experiments. Assuming Hardy-Weinberg equilibrium the frequencies of genotypes aa, Aa and AA (a= minor allele, A = major allele) are 0.155, 0.478 and 0.367 respectively. Thus, a minimum sample size for the study was determined by considering how many minor allele homozygote individuals for this SNP would be required to detect a genotype effect on EP4 function. Rs9292777 was reported to have an odds ratio for Crohn's disease of 1.34 (Parkes, Barrett et al. 2007). However, the size of the odds ratio was not expected to be a good predictor of the size of the effect on gene function. The *NOD2* 1007fs variant confers an allelic odds ratios of 4 for disease, but produced almost

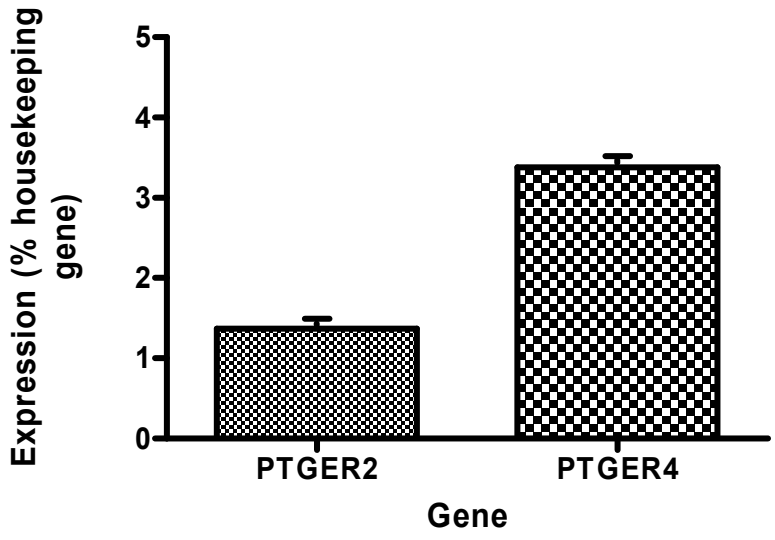
100% abrogation of PBMC cytokine responses in a similar assay to that proposed here (van Heel, Ghosh et al. 2005). For the purposes of estimating sample size here, a 30% difference in EP4 assay measurements between rs9292777 minor allele (aa) homozygotes and major allele (AA) homozygotes was assumed. Under conservative assumptions of within-group assay variation (standard deviation = 25%), 11 individuals in each group would be needed for 80% power to detect a 30% difference (t test of means between aa and AA homozygotes). Under assumptions of hardy Weinberg equilibrium, 71 individuals would be required to obtain 11 minor allele homozygotes. Given the uncertainty of the underlying assumptions, the power and sample size calculations necessarily have wide confidence intervals. However, it was estimated that around 100 randomly selected British population individuals would provide sufficient power to detect 30% differences in EP4 assay responses. Under Hardy-Weinberg equilibrium this sample collection would be expected to include 16 minor allele homozygotes, 48 heterozygotes and 37 major allele homozygotes. For the less common CrD associated SNPs (rs17234657-rs461376, MAF 0.125) 100 samples would be expected to include 2 minor allele homozygotes, 22 heterozygotes and 76 major allele homozygotes.

5.2 Results

5.2.1 *PTGER4* expression in PBMCs and monocyte-enriched subsets

PTGER4 gene expression in peripheral blood mononuclear cells (PBMCs) and PBMC subsets, separated using CD14 antibody-coated magnetic beads was assessed using qPCR. As PBMCs were obtained from several sources (leucofilters and buffy coats from the National Blood Service, fresh blood from volunteers) *PTGER4* expression was assessed in cells from each of these sample sources. *PTGER2* expression, which has been reported to be expressed in some PBMC subsets, has similar pharmacological function (increased cAMP) and may mediate similar effects on PBMC cytokine production (**Figure 5.5**) was also assayed (Fedyk, Ripper et al. 1996; Mori, Tanaka et al. 1996; Narumiya, Sugimoto et al. 1999; Sugimoto and Narumiya 2007). **Figure 5.6** shows results from these experiments and confirms *PTGER2* and *PTGER4* expression in PBMCs. Similar expression was observed for *PTGER2* and *PTGER4* in monocyte enriched (CD14+) and depleted (CD14-) subsets (data not shown, from three individuals). Expression did not differ significantly between monocyte-enriched and depleted fractions. CD14 +ve cells were selected using a positive antibody selection method (methods). Expression relative to housekeeping genes was calculated from the mean of 3 biological replicates.

Figure 5.6 *PTGER2* and *PTGER4* expression in PBMCs



5.2.2 Cytokine assays

5.2.2.1 Prostaglandin E_2 Pilot assays

To determine the feasibility of a cell supernatant cytokine assay for measuring biological responses to prostaglandins, pilot experiments were conducted with PGE_2, the major endogenous ligand of the EP4 receptor. These experiments were performed prior to obtaining selective EP4 agonists and antagonists that were used in later experiments. In these assays fresh PBMCs (2×10^5 cells/well unless stated) from healthy volunteers were incubated with PGE_2. Cytokine concentrations in cell supernatants were measured by enzyme-linked immunosorbent assay (ELISA). Conditions were varied to determine the optimum conditions under which the PGE_2 dose-response was maximal. Cell numbers, incubation times and PGE_2 concentrations were all tested over different ranges. In some experiments cells were co-incubated with lipopolysaccharide, following reports that PGE_2 and EP4 agonists can suppress LPS-induced secretion of some cytokines (TNFα, IFNγ, IL-18) (Takahashi, Iwagaki et al. 2005; Takahashi, Iwagaki et al. 2005). Similar effects have also been demonstrated in monocytes and macrophages with PGE_2 (Takayama, Garcia-Cardena et al. 2002; Takahashi, Iwagaki et al. 2005). These assays were tested at a range of LPS concentrations (0.1-100 ng/ml).

Dose-dependent suppression of TNFα release by PGE_2 from PBMCs incubated with lipopolysaccharide was observed (**Figure 5.7**). PGE_2 at 10^{-6} Molar concentration produced 90% suppression of TNFα in supernatants from PBMCs incubated for 24 hours with LPS (1ng/ml). These effects were optimum at LPS 1ng/µl and at 24h (earlier time points showed less pronounced effects). Similar effects were observed for PGE_2 on IFNγ production, with 10^{-6} Molar PGE_2 producing 85% suppression of IFNγ in supernatants from PBMCs incubated for 24 hours with LPS (1ng/ml) (**Figure 5.7)**.

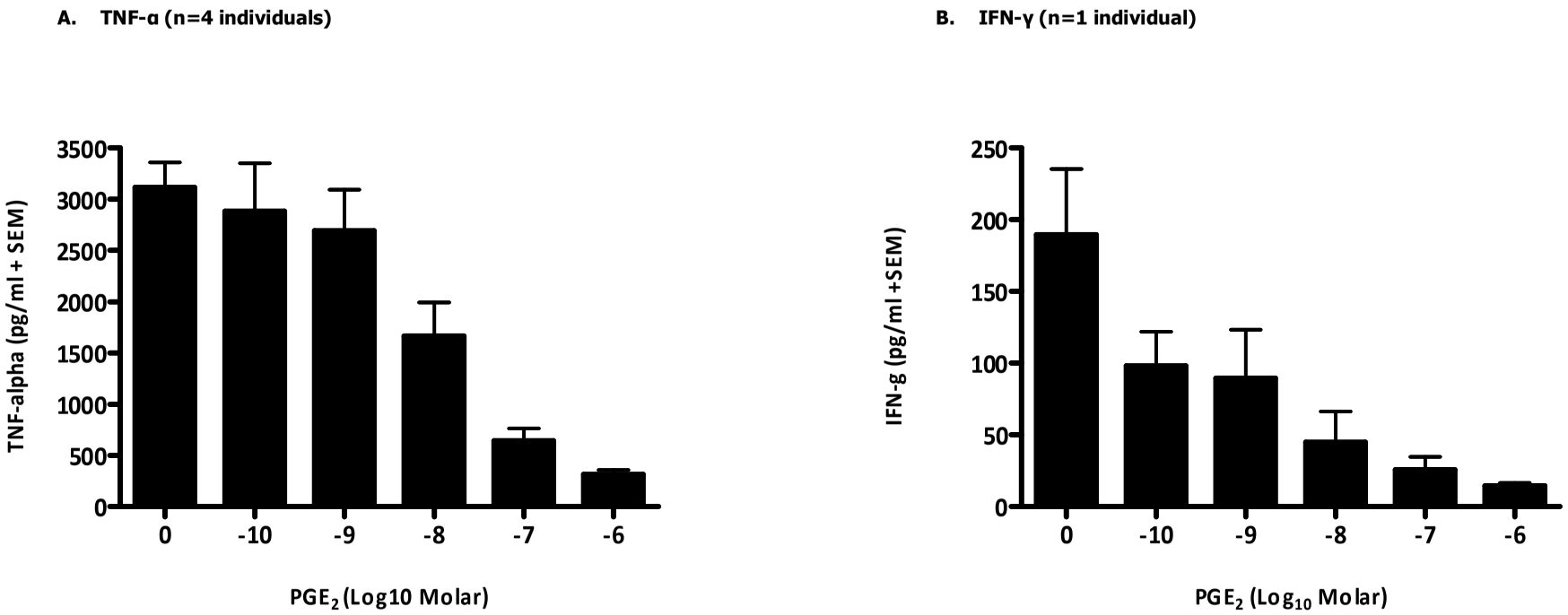
A modest increase in IL-1β in supernatants from PBMCs incubated with LPS (1ng/ml) for 24 hours was observed with increasing doses of PGE_2 (**Figure 5.8**). LPS 1 ng/ml was the best concentration to observe a dose-response to PGE_2 IL-1β was not detectable in the absence of LPS, and at a higher dose (100 ng/ml) IL-1β concentrations appeared saturated across the range of PGE_2 concentrations.

TNFα, IFNγ and IL-1 β were not detectable in cell supernatants incubated with PGE_2 in the absence of LPS. While these cytokines exhibited strong PGE_2 dose-dependence in cells incubated with LPS, cytokines released in response to PGE_2 in the absence of LPS or other

stimuli were sought, due to concerns over the effects of LPS on EP receptor expression and inter-individual variability in LPS responsiveness. Bothe IL-8 and particularly IL-6 showed dose-dependent release with increasing PGE_2 concentration (**Figure 5.9**). IL-6 was of particular interest as a candidate cytokine for EP4 responsiveness, due to the wide IL-6 concentration range observed with increasing physiological concentrations of PGE_2. There have not been previous reports of IL-6 release from PBMCs, though IL-6 up-regulation in response to PGE_2 has been observed in macrophages. Moreover the EP4 receptor is thought to mediate these effects (Fiebich, Schleicher et al. 2001; Maloy and Powrie 2005).

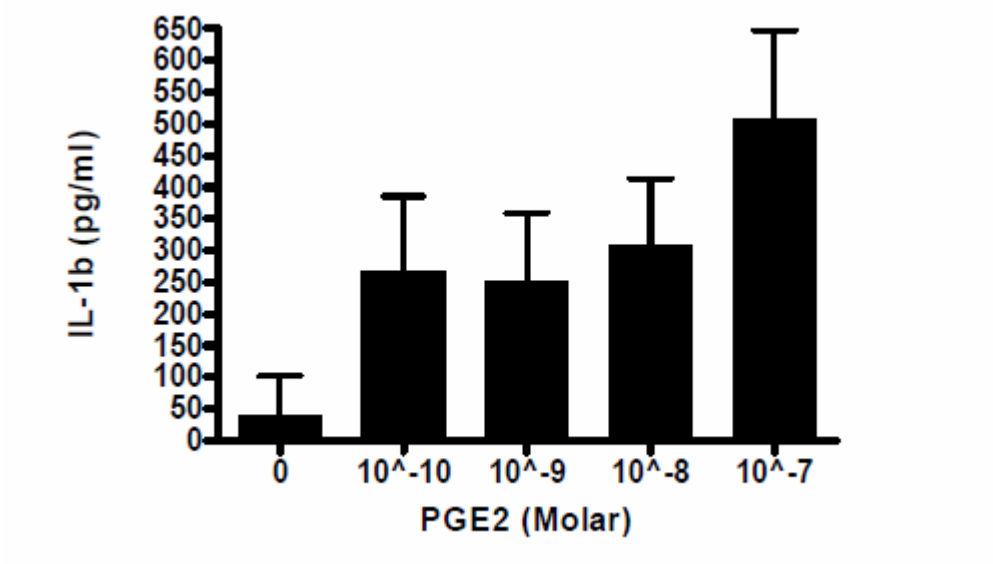
IL-17A was present at very low levels or was undetectable by ELISA in PBMC supernatants in similar experiments with PGE_2 and LPS.

Figure 5.7 PGE_2 suppresses TNF-α and IFN-γ release from PBMCs incubated with lipopolysaccharide



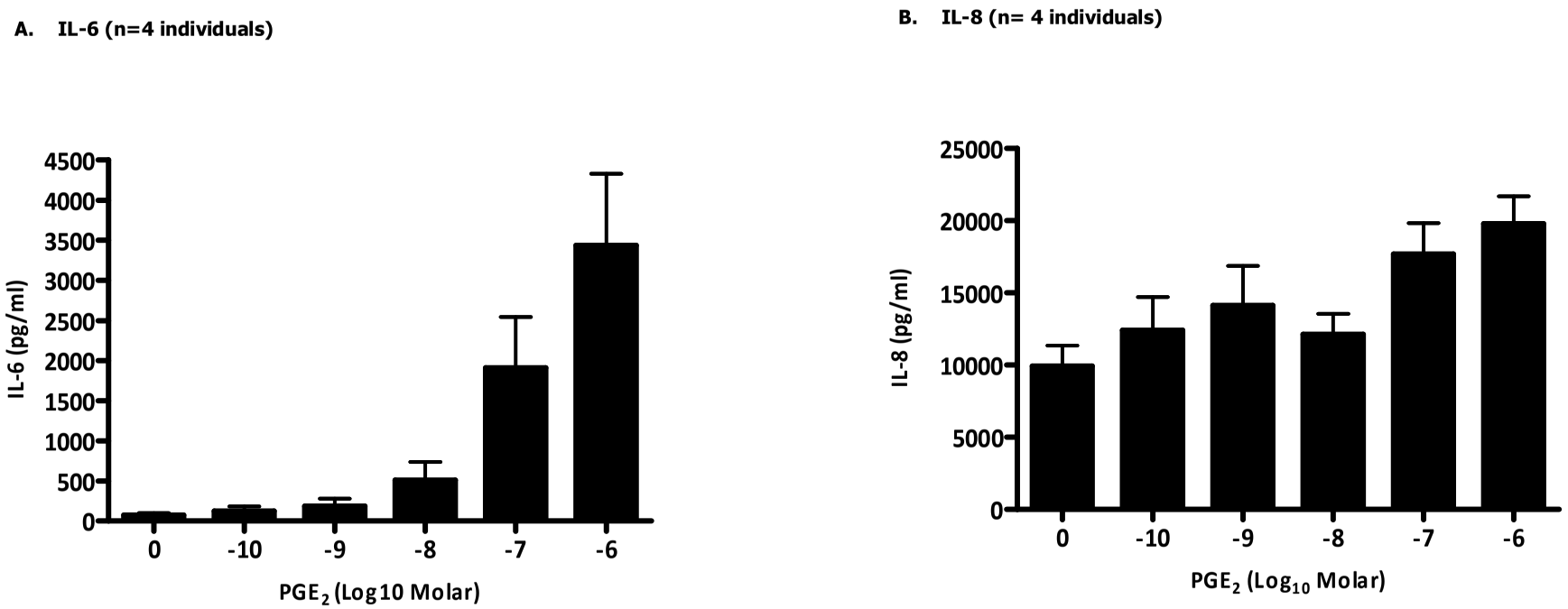
Cytokine concentrations measured in PBMC supernatants after 18-24 hours incubation in media with LPS 1 ng/ml and PGE_2. 2 x 10^5 cells in 200 µl volume per well. Error bars indicate standard error.

Figure 5.8 PGE_2 augments lipopolysaccahride induced IL-1β release from PBMCs



Cytokine concentrations measured in PBMC supernatants after 18-24 hours incubation in media with LPS 1 ng/ml and PGE_2. 2 x 10^5 cells in 200 µl volume per well. Error bars indicate standard error. data from n=2 individuals

Figure 5.9 PGE_2 stimulates IL-6 and IL-8 release from PBMCs



Cytokine concentrations measured in PBMC supernatants after 18-24 hours incubation in media with LPS 1 ng/ml and PGE_2, 2 x 10^5 cells in 200 µl volume per well. Error bars indicate standard error.

5.2.2.2 Assays using selective EP4 agonists/antagonists

As PGE_2 is a non-selective agonist at all prostaglandin EP subtypes, any of the EP receptors may mediate PGE_2 responses in the above experiments. EP4-selective agonists were therefore obtained to improve the specificity of this assay for EP4. No commercially available EP4 agonists are available. The development of selective agents has proved difficult despite efforts from major pharmaceutical companies (e.g. Merck, GlaxoSmithKline) and the field has been limited by a lack of truly selective agonists and antagonists (Wilson, Rhodes et al. 2004). Most published studies in humans have used an EP4 agonist developed by ONO pharmaceuticals (ONO-AE1-329, Osaka, Japan) (Yamamoto, Maruyama et al. 1999). Other prostanoid agonists and antagonists, showing some selectivity for the EP4 receptor have been developed by Merck and GlaxoSmithKline (Billot, Chateauneuf et al. 2003). All three of these companies were approached for use of these agents in the current study. A GlaxoSmithKline (GSK) EP4 agonist (GSK324202A) and antagonist (GW627378X) were obtained, as gifts, in January 2008. Pharmacological data have been published for GW627378X, but not for GSK324202A (Wilson, Giblin et al. 2006). In May 2008 the ONO EP4 agonist (ONO-AE1-329) and an antagonist (ONO-AE3-208) were obtained as gifts from ONO Pharmaceuticals Co. (Osaka, Japan).

5.2.2.3 Experiments using GSK324202A (EP4 agonist) and GW627378X (EP4 antagonist)

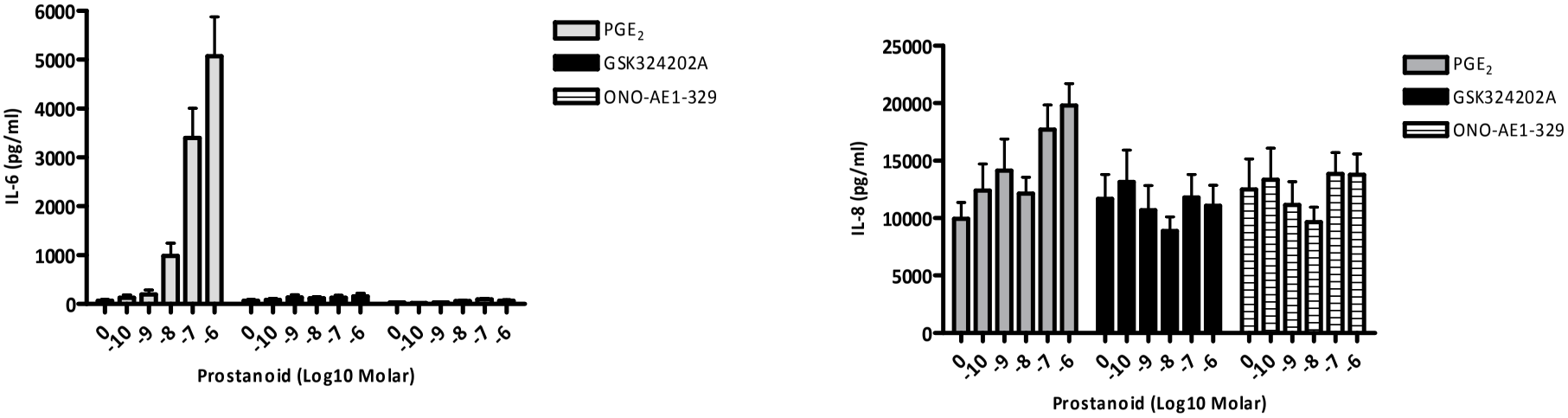
The effect of GSK324202A was tested in parallel with PGE_2 in the PBMC cytokine assays. GSK324202A did not reproduce the effect of PGE_2 on PBMC cytokine secretion in experiments with 6 individuals. This EP4 receptor agonist did not have any discernible effect on IL-6, IL-8 (without LPS) or TNFα production (**figures 5.10, 5.11**) across a large dose range (10^{-12}M to 10^{-5}M). GSK have not tested this agent in primary human cells and have only limited data showing partial efficacy for EP4 receptors in transfected human embryonic kidney cells expressing recombinant human EP4 (intrinsic activity versus PGE_2 of 62% using cAMP and calcium influx assays, unpublished data). While these studies also suggested that GSK324202A showed selectivity as an EP4 agonist compared to other EP receptor subtypes (pEC_{50} 7.5 for EP4 vs. <5 for EP1-EP3) the concern is that as GSK324202A is a partial EP4 agonist, it is uncertain whether it has sufficient efficacy at human EP4 receptors to generate observable effects on cytokine release from PBMCs. GSK324202A was therefore abandoned for use in subsequent experiments.

GW627378X is an effective EP4 antagonist, devoid of antagonist activity at EP2 and EP3 receptors, but with modest EP1 antagonism and significant antagonist activity at prostanoid TP receptors (Wilson, Giblin et al. 2006). Effects on EP1 and TP receptors were not thought to be relevant to the cytokine responses in the current assay, due to low expression in PBMCs and therefore this agent was thought to be an adequate antagonist of the EP4 receptor-mediated component of prostanoid modulation of cytokine responses in these assays.

PBMCs were first incubated with variable doses of GW627378X in the absence of LPS or PGE_2 to determine whether this agent reversed possible endogenous PGE_2 effects on cytokine release. GW627378X (up to 10^{-5}M) had no effect on LPS-induced TNFα or basally released IL-8 in PBMC supernatants (data not shown). These results suggested that basal or LPS-induced prostanoids did not contribute a major component of the observed TNFα or IL-8 production via EP4 receptor signalling. Secondly TNFα responses to PGE_2 were assayed with and without the addition of GW627378X (at 10^{-6}M at 10^{-5}M) to determine whether this produced the expected rightward displacement of the PGE_2 dose response curve. This was not observed, suggesting either that the PGE_2 dose-response effects were non-EP4 mediated or that GW627278X lacked EP4 receptor antagonism in these experiments. GW627378X produced modest rightward displacement of PGE_2- induced IL-6 release in one individual, consistent with the PGE_2-induced IL-6 being a partially EP4 receptor dependent effect (**Figure 5.12**).

The most likely explanations for these results were either that GSK324202A and GW627368X do not have significant efficacy at the EP4 receptor at the concentrations tested or that the effects on cytokine production observed for PGE_2 are mediated by non-EP4 receptor mechanisms.

Figure 5.10 EP4 agonists GSK324202A and ONO-AE1-329 do not reproduce effects of PGE_2 on IL-6 and IL-8 release from PBMCs



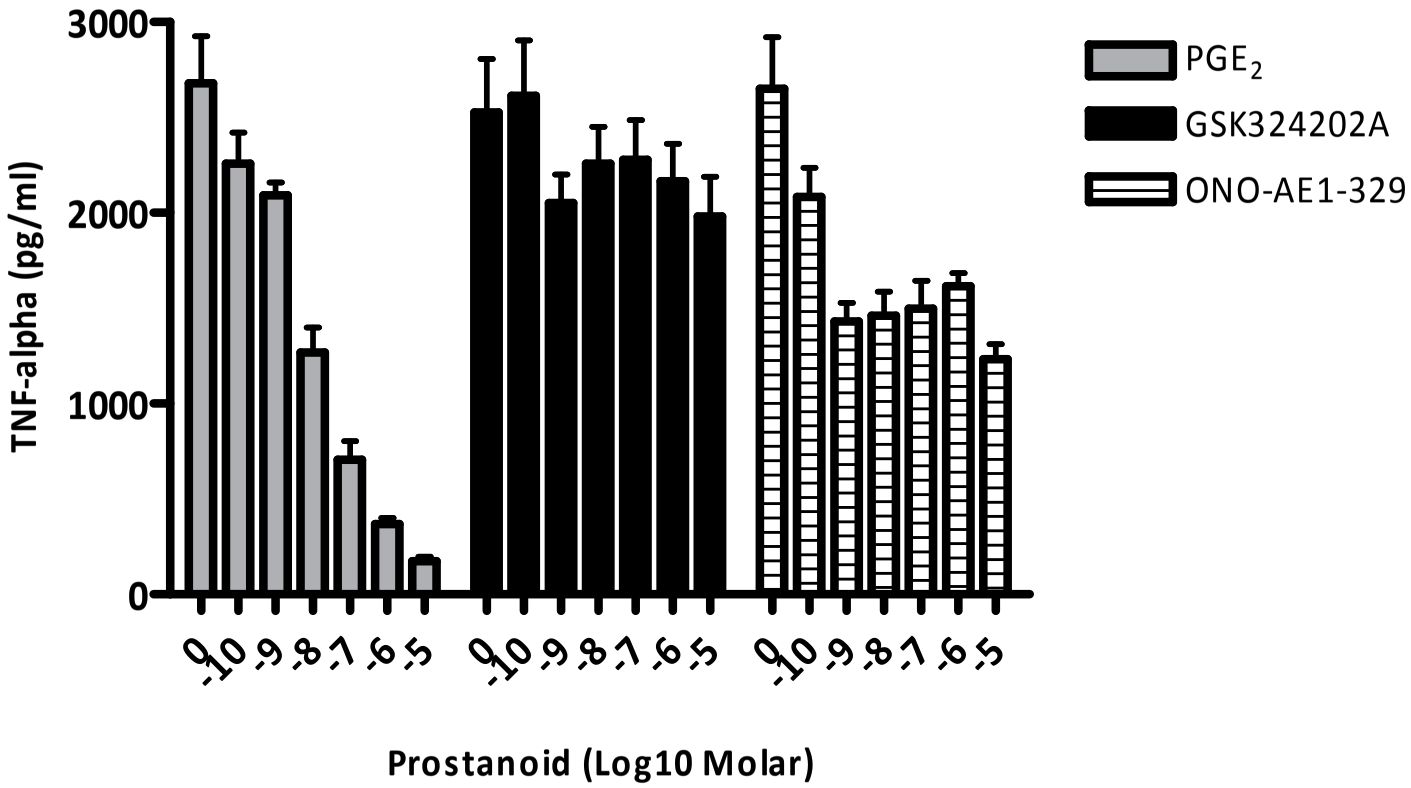
Cytokine concentrations measured in PBMC supernatants after 18-24 hours incubation in media with LPS 1 ng/ml and PGE_2. 2 x 10^5 cells in 200 µl volume per well. Error bars indicate standard error. Data from n= 4 individuals per group, each prostanoid tested in parallel in the same experiment.

5.2.2.4 Experiments using ONO-AE1-329 (EP4 agonist)

Prostanoid agonist PBMC incubation experiments with and without LPS were repeated using ONO-AE1-329, using PGE_2 as a positive control. ONO-AE1-329 is a selective EP4 agonist (EC_{50} 3.1nM (Yamamoto, Maruyama et al. 1999)). Moreover, ONO-AE1-329 has been reported to suppress LPS- induced TNFα secretion from PBMCs (Takahashi, Iwagaki et al. 2005). This observation was replicated in fresh blood derived PBMCs (**Figure 5.11**). The finding was replicated also in mononuclear cells separated (by Ficoll density gradient centrifugation) from leucofilters (n=4) obtained from the National Blood Service (data not shown).

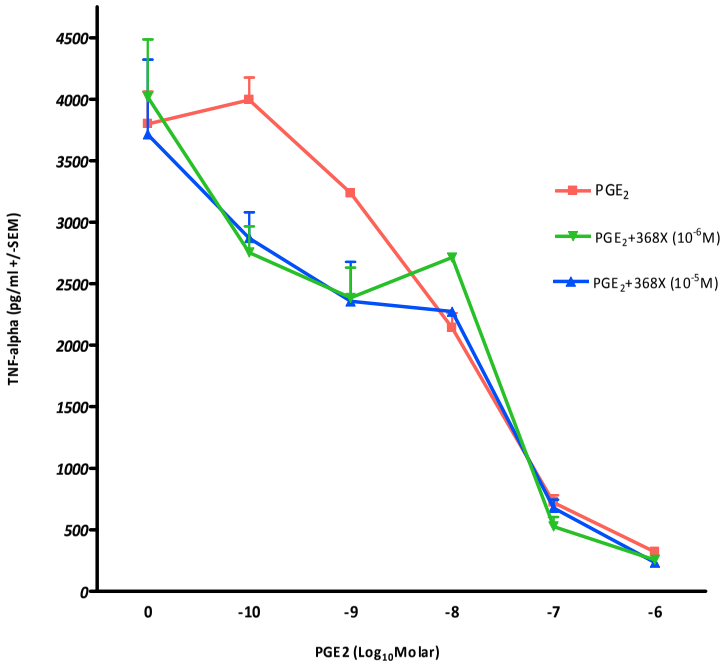
ONO-AE1-329 had no effect on IL-6 secretion from resting PBMCs and did not mimic the stimulation seen with PGE_2 (**Figure 5.10**). No consistent effects on IL-8 release were observed. For some individuals dose-dependant increases in IL-8 were observed for both PGE_2 and ONO-AE1-329 whereas for others no or a reverse effect was observed. It was considered that these opposite effects might be consistent with opposing effects of prostaglandins on different PBMC cell types. There have been reports that PGE_2 suppresses IL-8 production from monocytes(Standiford, Kunkel et al. 1992) but up-regulates IL-8 in T cells, though at much lower levels than produced by monocytes(Caristi, Piraino et al. 2005).

Figure 5.11 EP4 agonists GSK324202A and ONO-AE1-329 effects on TNF-α release from PBMCs

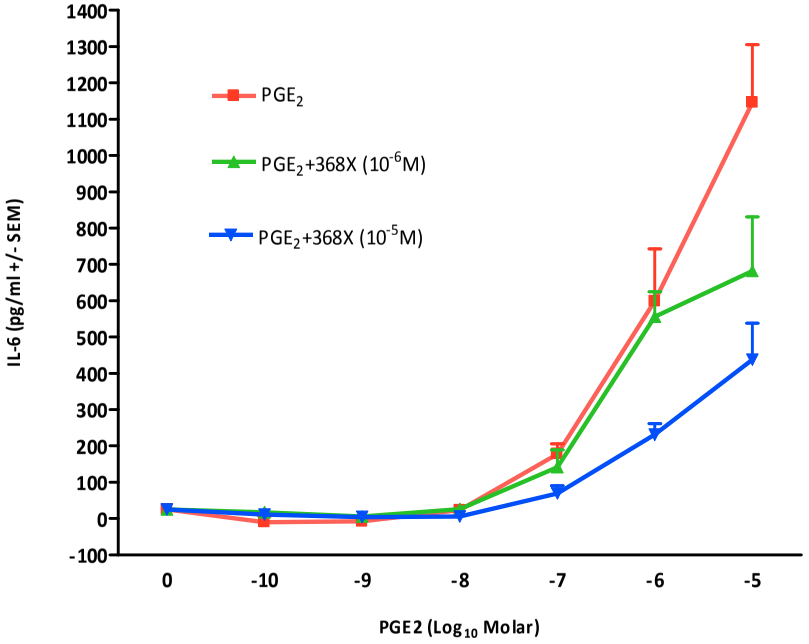


Cytokine concentrations measured in PBMC supernatants after 24 hours incubation in media using 2 x 10^5 cells in 200 µl volume per well with prostanoids added. Error bars indicate standard error. Data from n= 3 individuals, each prostanoid tested in parallel in the same experiment.

Figure 5.12 GW627368X (EP4 antagonist) effects on PGE_2 mediated TNF-α and IL-6 release from PBMCs



Cytokine concentrations measured in PBMC supernatants after 22 hours incubation in media with LPS (1ng/ml) using 2 x 10^5 cells in 200 µl volume per well with prostanoids added. Error bars indicate standard error. Data from n=1 individual.



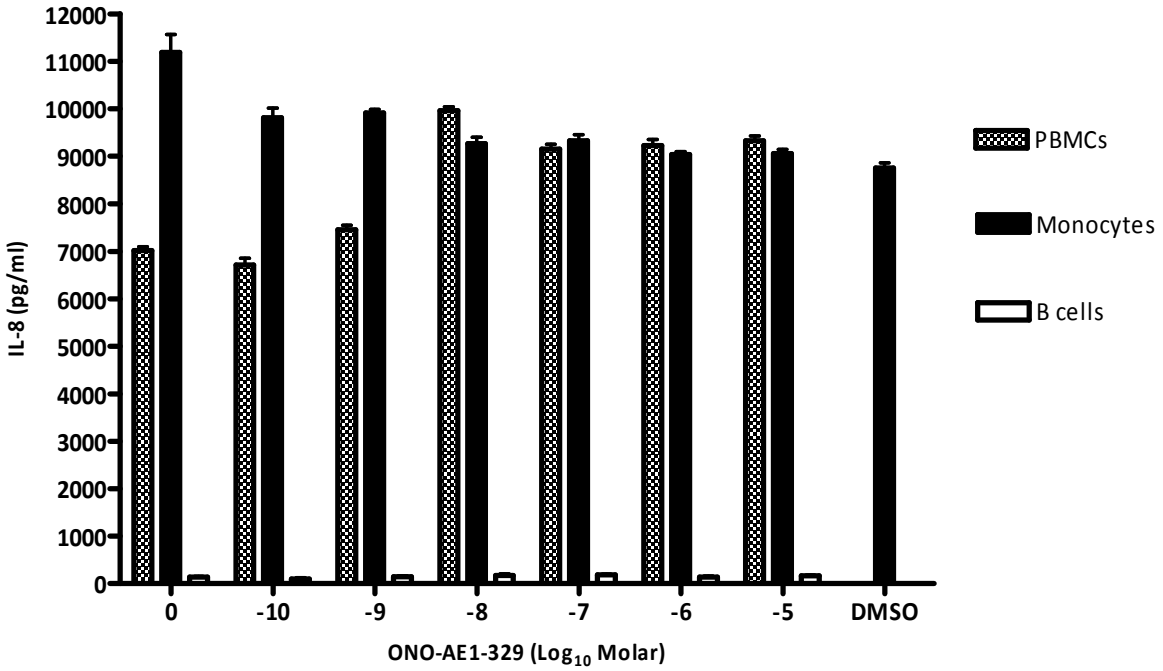
Cytokine concentrations measured in PBMC supernatants after 22 hours incubation in media with LPS (1ng/ml) using 2 x 10^5 cells in 200 µl volume per well with prostanoids added. Error bars indicate standard error. Data from n=1 individual.

In order to determine whether a PBMC subset would provide a more consistent ONO-AE1-329 dose response, experiments were performed using monocytes, B cells and T cells separated by RosetteSep^{TM} (Stem Cell Technologies, Canada), an antibody-based negative selection technique, from leucofilter-derived PBMCs. The purity of the enriched cell populations is reported to be 91% for (CD4) T cells, 89% for B cells and 73% for monocytes (Stem Cell Technologies, product information).

Results from these experiments suggested that IL-8 in PBMC supernatants is accounted for by monocyte secretion. IL-8 secretion from T cells was undetectable (data not shown). There was very little or undetectable IL-8 in B cell supernatants (**Figure 5.13**). Thus, cell subset separation was considered unhelpful for improving variation in the PGE_2 or ONO_AE1-329 induced IL-8 response. As such the IL-8 response was considered too variable to form the basis of a *PTGER4* specific assay.

The above experiments suggested that the most promising assay to observe a dose-response for the EP4 receptor agonist ONO-AE1-329 was that of TNFα in supernatants from LPS-treated PBMCs or monocytes, as previously reported by Takahashi et al. (Takahashi, Iwagaki et al. 2005). The data in these experiments suggested under optimum conditions a maximum 50% suppression of TNFα with ONO-AE1-329 could be observed. However, the usefulness of this assay was limited by significant inter-individual variability in cell supernatant TNFα concentrations and it was considered unlikely that the hypothesized genotype effects (30% differences) would be detectable in this assay. A major concern was that these assays required LPS co-incubation, introducing inter-individual variation due to LPS responsiveness. Moreover previously reported EP2 expression up-regulation and EP4 down-regulation by LPS could contribute further variation and increases the possibility of EP2 receptors mediating some of the observed effects of ONO-AE1-329 (Ikegami, Sugimoto et al. 2001). LPS also stimulates PGE_2 release from macrophages and monocytes, raising concerns that PGE_2 release from monocytes might confound these experiments (Ikegami, Sugimoto et al. 2001; Simmons, Botting et al. 2004). ONO-AE1-329 has 400fold lower agonist activity at EP2 receptors compared to EP4 receptors, and therefore significant EP2 agonism would be expected at the higher ONO-AE1-329 concentrations studied. In conclusion, none of the cytokine assays tested were thought to offer a sufficiently accurate and selective assay of *PTGER4* (EP4) function to take forward to assay in 100 individuals for genotype comparisons of EP4-mediated cytokine responses.

Figure 5.13 IL-8 in cell supernatants from PBMC subsets



Cytokine concentrations measured in PBMC supernatants after 42h using 7.5 x 10^4 cells/well in 200 µl volume. IL-8 undetectable in T cells – data not shown

5.2.3 Whole genome gene expression

The cytokines measured in supernatants from PBMCs incubated with ONO-AE1-329 were selected based on cytokines that had previously been reported to be modulated by PGE_2 and EP4 agonists in PBMCs and immune cell subsets. However, in order to undertake an unbiased survey of the effects of the EP4 receptor agonist ONO-AE1-329 on PBMCs, whole genome gene expression profiling was undertaken using gene expression microarrays. It was hypothesized that genes, particularly cytokine genes, showing several fold differences in expression in response to ONO-AE1-329 would be suitable for testing in the PBMC cytokine assay.

This experiment (n=2 healthy individuals) compared gene expression in PBMCs cultured with ONO-AE1-329 (10^{-7} Molar) or medium alone for 3 hours (without LPS). Whole genome gene

expression was assayed using the Illumina Human WG-6 chip, which assays 48,000 transcripts from > 24,000 known genes.

No genes in this experiment showed greater than 50% differential expression (**Figure 5.14**). Only 5 genes showed differential expression after Bonferroni correction (**Table 5.2**). The top differentially expressed cytokine was CCL22 (32% differential gene expression, $P = 7.37 \times 10^{-8}$ **Table 5.2**), ranked as the 3rd most strongly differentially expressed gene overall. In general the lack of genes showing large differential gene expression at 3 hours with ONO-AE1-329 suggested that this agent has only weak effects on gene expression in PBMCs.

Although 32% differential expression is modest, the up-regulation of CCL22 in unstimulated PBMCs prioritized it as the most attractive candidate for assessment in the cytokine assay experiments. CCL22 (C-C Chemokine motif 22) is chemotactic for monocytes, dendritic cells, natural killer cells and activated T lymphocytes. It binds chemokine receptor 4 and has a role in trafficking activated T lymphocytes to inflammatory sites. Chemokine receptor 4 is found mainly on Th2 cells and is predominant on these cells in individuals with autoimmune diseases including Crohn's disease (Jo, Matsumoto et al. 2003). Furthermore PGE_2 dose-dependently up-regulates CCL22 in monocyte derived dendritic cells, supporting the claim that the observed effects of ONO-AE1-329 on *CCL22* expression are prostaglandin EP-receptor mediated (McIlroy, Caron et al. 2006).
CCL22 was assayed by ELISA in PBMC supernatants after incubation of PBMCs for 18 hours with PGE_2 and ONO-AE1-329 (**Figure 5.15**). This experiment did not confirm any effect of either PGE_2 or ONO-AE1-329 on CCL22 release from PBMCs.

Table 5.2 Significantly differentially expressed genes in PBMCs withstanding Bonferroni correction

Gene	Differential expression P Value[a]	Bonferroni corrected Diff exp P value[a]	Norm expression (media only)	Norm expression (ONO-AE1-329 10^{-7}M)	Gene name &description
BX105338	1.38×10^{-14}	6.72×10^{-10}	361.30	237.38	Function unknown
ZFP36L1	1.13×10^{-08}	5.50×10^{-04}	1328.63	1839.11	Zinc finger protein 36, a putative nuclear transcription factor
CCL22	$7.37 \times ^{-08}$	3.59×10^{-03}	259.13	177.29	Chemokine (C-C motif) ligand 22
LOC644931	1.59×10^{-07}	7.76×10^{-03}	186.84	138.53	Hypothetical gene, unknown function
RPL7	6.08×10^{-07}	2.97×10^{-02}	691.01	876.90	Ribosomal protein L7
RBPJ	8.50×10^{-07}	4.15×10^{-02}	557.39	705.76	Homo sapiens recombination signal binding protein for immunoglobulin kappa J region, transcript variant 4.

[a]P values calculated on Illumina BeadStudio gene expression module v3.3.8 using Illumina's proprietary expression analysis algorithms

PBMCs (3×10^6 cells/well) were cultured for 3 hours at $37\ ^0$C prior to RNA extraction. Data from 2 individuals.

Figure 5.14 **Differential whole genome normalised mRNA transcript intensities. Pooled data from 2 individuals, PBMCs cultured with (ONO AVG_Signal) vs. without (Neg AVG_Signal) ONO-AE1-329 (10^{-7}M) for 3 hours**

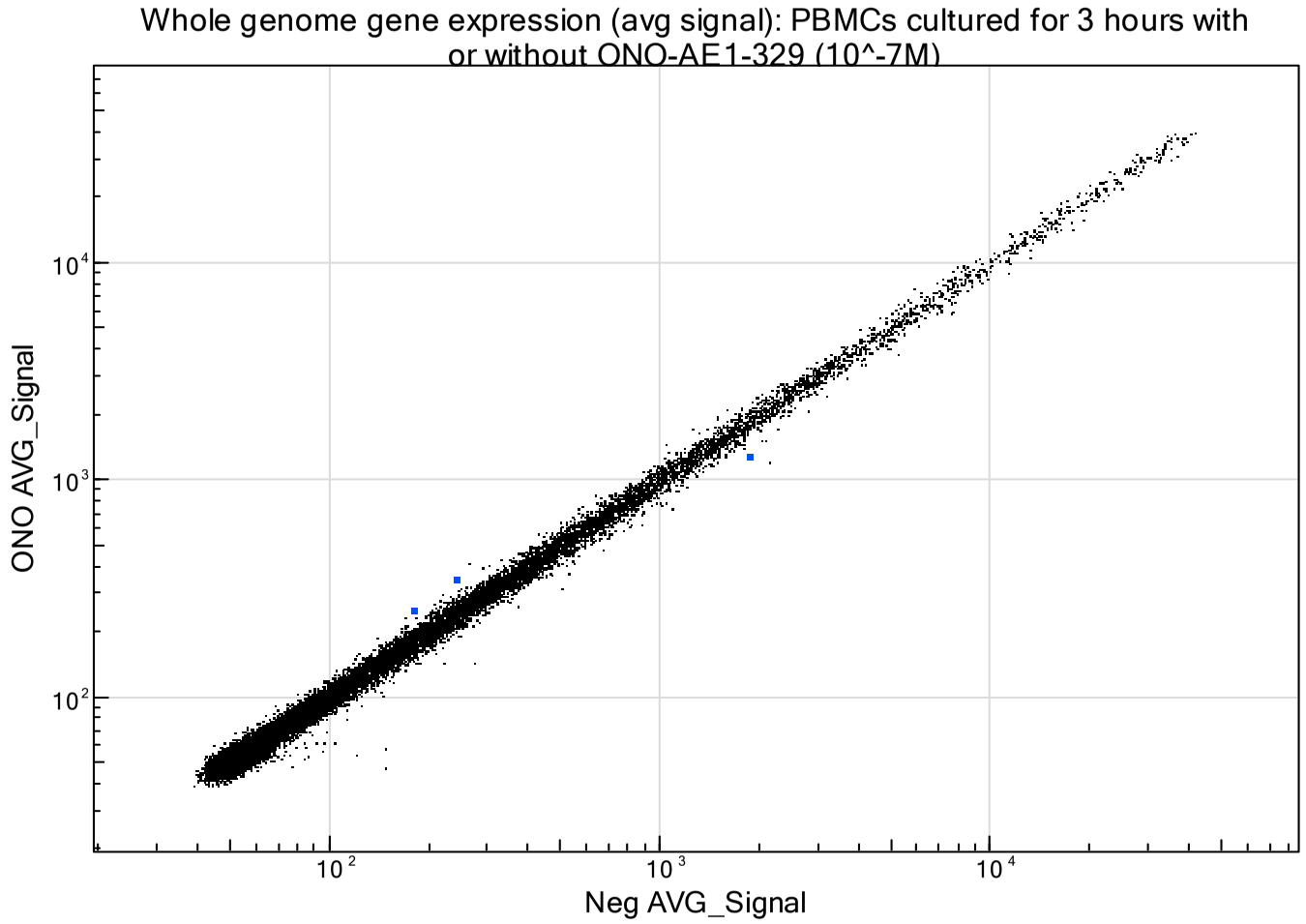
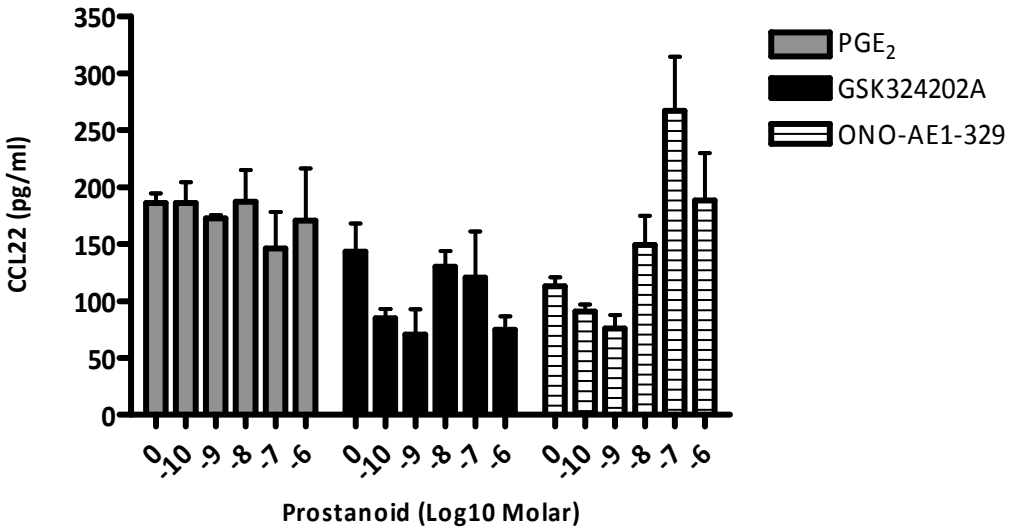


Figure 5.15 CCL22 concentrations in PBMC supernatants after culture with PGE_2 or ONO-AE1-329



Cytokine concentrations measured in PBMC supernatants after 18 hours incubation in media using 2 x 10^5 cells in 200 µl volume per well with prostanoids added. Error bars indicate standard error. Data from n= 1 individual, each prostanoid tested in parallel in the same experiment.

5.3 Discussion and Conclusion

5.3.1 Limitations of available EP4 agonists

Of the two selective EP4 agonists investigated, only ONO-AE1-329 showed clear prostaglandin E_2-like effects in the PBMC assays. In particular ONO-AE1-329 produced around 50% suppression of TNFα in supernatants from PBMCs incubated with LPS (1ng/ml) for 24 hours at a dose of 10^{-8} Molar (**Figure 5.11**). At this concentration, pharmacological data using recombinant human EP receptor subtypes suggest that ONO-AE1-329 (EC_{50} = 3.1 x 10^{-9} Molar) acts as a selective EP4 agonist with negligible agonist activity at other EP receptors (Yamamoto, Maruyama et al. 1999). In contrast, GSK324202A was unable to suppress TNFα in similar PBMC experiments, despite testing over a wide range of dose concentrations. The lack of effect of GSK324202A in all of the assays may reflect the lower efficacy of this compound at EP4 receptors (intrinsic activity 62% compared to PGE_2- unpublished data supplied by GlaxoSmithKline).

The most promising results in the PGE_2 pilot experiments had shown a strong dose-response effect on IL-6 production from non-LPS treated PBMCs (**Figure 5.10**). This was not reproduced by either EP4 agonist. In some other cell types studied (e.g. astrocytes, macrophages) PGE_2-induction of IL-6 appears to be EP4 receptor mediated (Fiebich, Schleicher et al. 2001; Ma and Quirion 2005). However, in mouse EP4 receptor deficient neutrophils PGE_2 can augment IL-6 release, an effect replicated with EP2 agonists. In contrast ONO-AE1-329 was unable to reproduce this effect in EP2 knockout mouse-derived neutrophils (Yamane, Sugimoto et al. 2000). These observations raise the possibility that PGE_2-dependent IL-6 release in PBMCs may also be EP2-dependent, although this has not been studied. An EP2-mediated mechanism for PGE_2 induced IL-6 production from PBMCs would be consistent with the data we observed for ONO-AE1-329 and GSK324202A. On the other hand, the modest rightward displacement of the PGE_2 –IL-6 dose response curve observed with the addition of the EP4 antagonist GW627378X is at odds with this, apparently indicating a component of EP4 receptor dependency of the IL-6 response. The lack of any effect of GW627378X on PGE_2 mediated suppression of TNFα from LPS-treated PBMCs does not necessarily suggest that this agent lacks EP4 antagonist efficacy as PGE_2 is known also to suppress TNFα through other EP receptor subtypes (Takahashi, Iwagaki et al. 2005). Thus while the available EP4 agonists were ineffective for the purposes of generating a strong IL-6 response in PBMCs that would be

suitable for an EP4-selective assay, further studies of PBMCs are required to determine the mechanism of PGE_2-induced IL-6. These experiments could include repeat experiments with EP2 and other EP receptor subtype agonists and antagonists or with small interfering RNA (siRNA) knockdown of EP4 and other EP receptor expression in PBMCs.

The search for the ideal cytokine for measurement in a simple PBMC cell culture assay was expanded through the use of array-based whole genome gene expression profiling. However, only modest differential gene expression was observed for ONO-AE1-329, suggesting that in unstimulated PBMCs this agent has little activity. Thus, if EP4 receptor mediated effects in unstimulated PBMCs are modest and only become prominent in modulating the response to other inflammatory stimuli, the prospects for a selective EP4 assay, where extraneous variables are minimized (e.g. LPS or other inflammatory stimulus inter-individual variation) would appear to be bleak. Testing of the most strongly differentially expressed cytokine, CCL22 showed that the small differential effects on gene expression did not translate to large effects on protein levels assayed by ELISA in cell supernatants.

The development of a truly selective EP4 assay, in a biological context relevant to the pathogenesis of Crohn's disease has proven a major challenge. Currently, we may be limited by a lack of agents that act as potent but selective EP4 agonists. The opposing pro and anti-inflammatory effects of PGE_2 – EP4 signalling in different cell types and biological contexts adds further complexity. Moving from an observation that Crohn's Disease associated variants correlate with *PTGER4* gene expression to understanding the functional contribution of this increased expression may require better understanding of PGE_2 – EP4 signalling in inflammation in general. Based on these data, it was decided not to pursue these experiments further in this thesis.

5.3.2 Limitations of correlating gene function with GWAS SNP associations

Moving from GWAS associations to understanding the mechanisms by which genetic variants alter gene function and contribute to disease susceptibility has proven difficult not just in Crohn's disease, but in all common diseases studied. There have only been rare exceptions. In Crohn's disease the *ATG16L1* T300A variant has been associated with impaired autophagic control of *salmonella typhimurium* in intestinal epithelial cells and with secretory granule exocytosis abnormalities in Paneth cells (Cadwell, Liu et al. 2008; Kuballa, Huett et al. 2008).

Similarly *NOD2* protein-altering SNPs have been shown to have loss of function effects in human immune cells (van Heel, Ghosh et al. 2005). It may not be coincidental that the genes for which successes have been observed have so far been those where causal genetic variants had been identified in the primary genetic studies. Unfortunately, for the vast majority of GWAS SNP associations, the causal variants are currently unknown. This adds uncertainties to attempts to correlate SNP genotypes with gene function. Firstly, as for the 5p13.1 SNPs here, the gene(s) whose function is altered is often uncertain. Secondly, the SNPs showing disease association are likely to be only partially correlated with the causal variants under study. This not only weakens the strength of the case-control GWAS association but is likely to dilute the strength of SNP-gene function correlations, since only some individuals carrying the GWAS SNP risk allele will carry a causal variant risk allele. The most extreme scenario that may mitigate against the discovery of GWAS SNP- gene function associations is that in which GWAS associations arise from rare variants, potentially in genes that reside megabases from the peak GWAS association. In this scenario even if the studied gene turns out to be the correct one, only a small proportion of individuals carrying the common GWAS SNP risk allele may harbour the (more highly penetrant) rare variant(s). Nevertheless, it is also currently far from clear whether pure genetic approaches (e.g. genetic fine mapping and re-sequencing efforts with subsequent case-control association testing) will have sufficient discriminatory power to define causal variants, particularly for regulatory variants which may be difficult to recognize. Indeed early experiences of these approaches have not been promising (ref-WTCCC?). It is therefore likely that functional experiments, that link genetic variants to alterations in biological functions will have to be embarked upon in many cases without the prior knowledge of causal variants and these approaches indeed may be critical for identifying the causal variants driving GWAS associations in common diseases.

5.4 Methods

Ethical approval for recruitment of healthy volunteers and use of fresh blood samples for the experiments below was granted by the EAST LONDON & THE CITY RESEARCH ETHICS COMMITTEE (REC number: P/03/229). Informed consent was taken and documented for each participant. Leucofilters and buffy coat fractions were obtained anonymously from the British National Blood Service and were used for cell culture experiments and genetic research with approval from the Oxfordshire REC (05/Q1605/89).

5.4.1 Isolation of peripheral blood mononuclear cells by density gradient centrifugation.

50-60ml peripheral venous blood was collected from healthy volunteers into 10 ml Lithium-heparinised tubes (Becton Dickinson, UK, 367874). In some experiments leucofilter or buffy coat fractions were obtained from the National Blood Service (Tooting, London, UK) as a surplus blood product for research. Cell separation was performed on the same day as blood collection.

Per 30 ml blood/ blood product: 15 ml Lymphoprep (Axis-Shield Diagnostics) was added to a 50ml LeucoSep tube (Greiner Labs) and centrifuged for 30 seconds at 1000g (2200 rpm) using a Heraeus Megafuge 10R with a BS4402/A (3360) bucker rotor at room temperature. 30ml blood from Li-Heparin tubes was added to the Lymphoprep in the LeucoSep tube and centrifuged at 1000g for 10 mins at room temperature without the brake engaged. The mononuclear cell monolayer suspension above the LeucoSep filter was removed to a separate 50ml Falcon tube, and washed by addition of 30ml Phosphate Buffered Saline and centrifugation at 250g (1100rpm) for 10 mins. Supernatant was discarded, and the cell pellet resuspended and incubated in 20ml of red cell lysis solution (155mM NH_4Cl (Sigma Aldrich, UK, A0171), 10mM $KHCO_3$ (Sigma Aldrich, UK, P7682), 0.1mM EDTA (Gibco, Invitrogen, UK, 155575-038), tissue culture grade H_2O (Sigma Aldrich, UK) was added and left for 10 minutes before centrifuging for 250g (1100rpm) for 10 minutes. Red cell lysis buffer was removed the cell pellet and re-suspended in 50 ml sterile PBS. This was re-centrifuged at 250g for 10 minutes and the wash repeated once. Pelleted cells were resuspended in 1ml of X-Vivo-15 serum free media (Lonza

Group, Switzerland, 04-418). Cells were counted using a standard method with Trypan Blue staining and manual laboratory haemocytometer.

5.4.2 Cell culture experiments with Prostaglandin E_2, EP4 agonists/antagonists and lipopolysaccharide (LPS).

Prostaglandin E_2 was obtained from Cayman Chemicals (Michigan, USA, 14010). GSK324202A (EP4 agonist) and GW627368X (EP4 antagonist) were gifts from Glaxo-SmithKline (Stevenage, UK). ONO-AE1-329 (EP4 agonist) was a gift from ONO Pharmaceuticals (Osaka, Japan). X-Vivo-15 serum free media (Lonza Group, Switzerland, 04-418) was used for reagent dilution, cell suspension and culture.

Cells were cultured in triplicate at a density varied between 7.5 x 10^4/ml and 3 x 10^6/ml in 96 well cell culture plates suspended in 250µl total volume, including media and ligands at 37 ^{0}C. After culture, 150µl of cell culture supernatant was removed from each well and transferred to a separate 96 well cell culture plate and frozen at -80^{0}C.

5.4.3 Enzyme linked immunosorbent assay (ELISA) for quantification of cytokines and chemokines in cell supernatants

Sandwich ELISA was performed on cell supernatants.

IL-1β,TNFα, IL-8 and IL-6 ELISAs were performed using matched monoclonal antibodies at 1:2000 dilutions (Immunotools, Friesoythe, Germany), streptavidin-horseradish peroxidise (R&D systems, Abingdon, UK) and TMB-H_20_2 (BD Bioscience, Oxford, UK). IFNγ and IL-17A ELISAs were performed using Human Interferon-γ and IL-17A ELISA Ready-Set-GO kits (Ebioscience, Hatfield, UK 88-7316, 88-7176). CCL22 ELISA was performed using matched antibody pairs and recombinant human CCL22 protein for standards from R&D, UK (catalogue no. DY336). Recombinant proteins for standards used in ELISA assays with matched antibody pairs were obtained from Immunotools (IL-8: catalogue number 11340080; TNFα: 11343013; IL-6 11340060) and Firstlink, UK (IL-1β, catalogue number hrIL-1β).

Cell culture supernatants were diluted at part 1 in 4 for TNFα, part 1 in 4 for IL1-β, part 1 in 20 for IL-8, part 1 in 4 for IL-17A, part 1 in 4 for IFNγ, part 1 in 10 for CCL22, part 1 in 4 for IL-6 in assay buffer (5%BSA (Sigma Aldrich, UK, A7030), 0.05% Tween (Sigma Aldrich, UK, P1379), PBS) to make a total volume of 100µl.

For ELISA experiments using matched monoclonal antibody pairs, the capture antibody was diluted in 0.05mM carbonate-bicarbonate buffer (Sigma Aldrich, UK, C3041) according to manufacturer's instructions and 100µl added to a 96 well plate (Nunc Immuno^{TM} F96 Maxisorp, VWR, 442404) for 24 hours at 4^0C. The plate was washed 5 times with >200µl of ELISA wash solution (PBS, 0.05% Tween) using a 12 channel wash station (Nunc Immuno^{TM} Wash, VWR, 735-0057). The plate was patted dry, and cell culture supernatants added.

A recombinant protein, serially diluted 1:2 in assay buffer, was added in duplicate to plates to create a standard curve. Samples were incubated for one hour on a shaking platform (KCH-VIBRAX). After washing samples five times with >200µl of ELISA wash, 100µl of secondary biotinylated antibody from the matched pair, diluted in assay diluents at appropriate dilutions according to manufacturer's instructions was added, followed by a 1 hour incubation at room temperature on a shaking platform. After five further washes 100µl of streptavidin conjugated HRP (R&D Systems, UK, DY998) diluted 1:200 in assay diluents was added and incubated for 30 minutes at room temperature on a shaking platform. After five final washes, 100µl of mixed BD OptEIA^{TM} TMB-H_2O_2 substrate (BD Bioscience, UK, 555214) was added to each well and incubated at room temperature for 15-20 minutes on the shaking platform. The reaction was terminated with 50µl of 1mM H_4PO_4 (Sigma Aldrich, UK, P6560) followed by dual absorbance measurement at 450nm and subtraction of 570nm background (680 Microplate Reader, BioRad). For IL-17A and IFNγ the same principle was followed using pre-formed kits according to manufacturer's instructions

5.4.4 Cell subset separation

PBMC fractions were separated using the BD Biosciences IMag^{TM} cell separation system (BD Biosciences, Oxford, UK) for qPCR experiments according to the manufacturer's instructions. Briefly, PBMCs were washed and re-suspended at 10^7 cells/ml in BD IMag^{TM} Buffer. 50µl BD anti human CD14 antibody coated magnetic particles were added per 10^7 cells of the PBMC suspension, mixed by pipetting up and down and incubated at room temperature for 30 minutes. The PBMC/magnetic particle mixture was placed in a magnetic field using the BD IMagnet^{TM} for 10 minutes; CD14 -ve cell suspension mixture was transferred to a second tube while the PBMC/magnetic particle mixture was in the magnetic field- to generate the CD14 -ve cell/ supernatant mix. CD14+ve cells were retained in the original tube and resuspended after 2 washes with BD IMag buffer.

Rosette-Sep (StemCell Technologies, Sheffield, UK) was used for Monocyte (RosetteSep Human Monocyte enrichment cocktail, 15068) and B cell (RosetteSep Human Monocyte enrichment cocktail, 15024) fraction separation from PBMC/leucofilter samples for ELISA and microarray experiments. Briefly the Rosette-Sep monocyte or B cell enrichment cocktail was added at 50ul/ml of whole blood or leucofilter sample prior to addition of Lymphoprep and density gradient centrifugation (DGC). A modified DGC protocol was followed where LeucoSep tubes were not used and enriched cells were isolated from the plasma-Lymphoprep interface by Pasteur pipette following the recommended Rosette-Sep protocol.

5.4.5 RNA extraction

Cells were cultured at a density of $3x10^6$/well per condition in 24 well cell culture plates (VWR, UK, 734-0020). After 22 hours of culture at 37^{o}C, 5% CO_2, cell culture supernatants were removed and spun down in 1.5ml microcentrifuge tubes (Axygen, VWR,UK, 525-0231) on a Heraeus Biofuge Pico with a PP1/96 rotor (Heraeus, VWR, UK). Supernatants were discarded and 1ml of TRIzol reagent (Invitrogen UK, 15596-018) added to the pellet. After a 5 minute incubation and agitation by pipetting, the 1ml of TRIzol solution was added to the cell culture well and incubated for a further 5 minutes. The TRIzol and cell lysate solution was pipetted up and down to ensure all cells were lysed, transferred to a microcentrifuge tube and stored at -80^{o}C. Once thawed, 200ml of chloroform (Sigma-Aldrich, UK, C2432) was added to the lysate mixture and vortexed vigorously (Genie2, Scientific Industries, USA) for 15 seconds. The solution was spun at 11,000 rpm at 4^{o}C using an Eppendorf 5417R MiniCentrifuge (Eppendorf, UK). The aqueous layer was transferred to a microcentrifuge tube containing 500µl 100% 2-propanol (Propan-2-ol AnalaR, VWR, UK, 102246L) and incubated at room temperature for 15 minutes to precipitate the nucleic acids, followed by a 15 minute spin at 15,000 rpm (4^{o}C) on the Eppendorf MiniCentrifuge. The supernatant was discarded, 1 ml 70% ethanol added and the tube spun at 12,000 rpm for 5 minutes (at 4^{o}C). The supernatant was discarded and the pellet allowed to air dry. The RNA pellet was re-suspended in 25µl RNase free molecular biology grade H_2O (Sigma Aldrich, UK, W4502) at 60^{o}C for 5 minutes. RNA was quantified by absorbance 280nm on a Nanodrop* ND 1000 Spectrophotometer (Nanodrop technologies, USA). RNA was stored at -80^{o}C.

5.4.6 RNeasy RNA cleanup

For use on the gene expression microarray, RNA extracted by TRIzol was purified on Qiagen Spin RNeasy Mini Columns (Qiagen, UK, 74104) with the on-column DNase digestion (Qiagen, UK, 79254). 100 µl TRIzol isolated RNA in RNase free H_2O (Sigma Aldrich, UK, W4502) was combined with 350 µl of Buffer RLT (Qiagen, UK) and 250 µl ethanol (VWR, UK, 10107) and added to an RNeasy Mini Spin Column (Qiagen, UK). The column was spun at 9000g on an Eppendorf 5418 centrifuge with a FA-45-18-11 rotor at room temperature. 350 µl of Buffer RW1 (Qiagen, UK) was added to the column and spun for 15 seconds at 9000g for 15 seconds followed by the addition of 15 µl DNase (Qiagen, UK) in 70µl Buffer RDD (Qiagen, UK) for 10 minutes. 350 µl of Buffer RW1 was added to the column and spun for a further 15 seconds. 500µl Buffer RPE (Qiagen, UK) was added to the membrane for 1 minute followed by a further spin for 15 seconds. This process was repeated with an additional two minute spin and a further 2 minute spin without addition of a buffer to dry the membrane. RNA was eluted into a 1.5ml microcentrifuge tube with 30µl molecular biology grade water and spun for 1 minute at 9000g. RNA was stored at -80^{o}C.

5.4.7 Reverse transcription PCR (RT-PCR)

Complementary DNA (cDNA) was generated from total RNA using the cDNA High Capacity Reverse transcription kit (Applied Biosystems, USA, 4368814) according to manufacturer's instructions. Briefly, a master mix, comprising 2.0µl 2x RT Buffer (ABI, USA), 0.8µl 25x dNTPs (100mM) (ABI, USA), 1µl RNase inhibitor (Roche, Ambion, USA, AM2682), 3.2µl Nuclease-free water per reaction (plus 10% overage) was made. 10µl of master mix was added to 10xl (50ng/ml) RNA in 0.2ml PCR tubes (VWR, UK, 732-0548) and mixed by pipetting. Tubes were briefly spun (Mini Galaxy, VWR, UK) prior to reverse transcription under the following conditions on a MJ Research DNA Tetrad 2 PCR Machine (MJ Research, USA):

25^{o}C – 10 minutes
37^{o}C – 120 minutes
85^{o}C – 5 seconds

cDNA was stored at -20C prior to use in PCR reactions.

Double dye quantitative PCR assays

Quantitative PCR Taqman (qPCR) primers and probes were obtained from Applied Biosystems for *PTGER4, PTGER2, ACTB* (Beta actin) and *GUSB* (Glucuronidase beta). Probes contained a FAM reporter dye and a non-fluorescent quencher (TAMRA). A qPCR master mix was constructed with 2x absolute QPCR Rox Mix (ABGene, UK AB-1139), 40x Taqman primers and Probe Mix and deionized H_2O up to the required volume. Reactions were set up in 96 well optical plates (Applied Biosystems, USA, N8010560). 96 well PCR was performed on an Applied Biosystems 7500 Real Time PCR machine (Applied Biosystems, USA), under the following conditions:

Denature for 20 seconds at 95^{o}C
40 cycles of:
95^{o}C – 15 seconds
60^{o}C – 60 seconds
Detection range was from cycles 1-40

Results were visualised in SDS software v2.3 (7900HT) or v1.4 (7500) (Applied Biosystems, USA).

qPCR data is expressed relative to a housekeeping gene based on the ΔCT method. The geometric mean of triplicate measurements with standard deviations (SD) less than 0.3 (or from duplicates if one measurement outside 0.3 SD) was calculated. Values were considered usable if duplicate values with <0.3SD and Ct values <35 were obtained.

5.4.7.1 qPCR Calculations

qPCR calculations were performed relative to housekeeping genes using the ΔΔCT method as recommended by Applied Biosystems User Bulletin #1 guide to Performing Relative Quantification of Gene expression using real-time Quantitative PCR. The geometric mean of triplicate values of the target was subtracted from the geometric mean of the endogenous control to give ΔCT. ΔΔCT values were calculated by comparison to a reference group. Expression of *PTGER2* and *PTGER4* was calculated by deltaCT method as a percentage of mean *ACTB/GUSB* expression.

5.4.8 Expression Microarrays

Gene expression microarray experiments were carried out using the Illumina Sentrix Gene expression System (Illumina Inc, San Diego, USA; WG-6 v3.0 expression Beadchip). 1µg of RNA was supplied for the assay, quantified using Nanodrop (Nanodrop Technologies, USA) and subsequently RNA integrity and quantity assessed using the 2100 BioAnalyzer (Agilent Technologies, Santa Clara, USA). Briefly, RNA was reverse transcribed and amplified, labelled, hybridized to the Human WG-6 v3.0 expression Beadchip, stained and Beadchip was scanned using Illumina's BeadArray Reader. These experiments were performed as a service by Barts and the London Genome Centre (London, UK). Data was analysed using Illumina BeadStudio 2.0 software (gene expression module).

Chapter 6 Discussion

6.1 Summary of research

This thesis has investigated the contributions of common genetic variants to common disease susceptibility and pathogenesis. The focus of the research has been the intestinal inflammatory disorders, coeliac disease and Crohn's disease together with an inflammatory adverse effect (pancreatitis) of one of the major drug classes used in their treatment. Each phenotype arises from interactions between the host immune system and environmental factors present in the gut. In coeliac disease, the intestinal immune system interacts with well-defined dietary gluten peptides to cause inflammation. In Crohn's disease the immune system interacts with resident intestinal microbiota to cause inflammation. Finally, reports emerging during the period of this PhD regarding genetic susceptibility to idiosyncratic drug reactions, stimulated interest in pancreatitis occurring in response to azathioprine or 6-mercaptopurine, drugs used in the treatment of intestinal inflammatory disorders. For this phenotype, the environmental exposure is precisely defined; pancreatitis is hypothesized to occur through interaction of the immune system with the drugs, their metabolites or with a neo-autoantigen formed by the interaction of these drugs with pancreatic tissues. Genetic risk variant discovery has been the focus of the research into coeliac disease and azathioprine-induced pancreatitis in this thesis whereas investigations into the mechanisms by which DNA sequence variants alter cell immunobiology were the subject of research in Crohn's disease. Together the work uses complementary genetics approaches to understand the genetic basis of these complex gastrointestinal phenotypes.

6.1.1 New genetic risk variants in coeliac disease

The largest part of the work identified multiple new genetic risk variants in coeliac disease, providing insight into how genetic liability is distributed across multiple genomic loci. This research has also added detail to the immunogenetic understanding of the causes of coeliac disease, by confirming the importance of some well known coeliac immunological pathways, but also by drawing attention to others whose role was previously not appreciated. This large collaborative study included 4,533 cases and 10,750 controls in a genome wide association study (GWAS) phase and 4,918 cases and 5,684 controls in a follow-up phase, comprising a

total of 12 sample collections from 10 nations, all of European ancestry. The study replicated associations in all 14 coeliac susceptibility regions reported in a prior UK coeliac GWAS and follow-up studies. In addition, genome-wide significance ($P_{Combined}$<5 x 10^{-8}) was observed in the combined analysis (GWAS + follow-up) for SNPs from 13 susceptibility regions. In 10 of these 13 regions strong candidate genes are found, with known roles in the immune system and mostly with T cell functions. In the other 3 regions, no genes mapped to the associated linkage disequilibrium block in one case and in the other two cases genes mapping in the region do not have proven immune or intestinal functions. A further 13 risk regions were identified from the combined analysis with lesser significance (10^{-6}<$P_{combined}$<5x10^{-8} and/or P_{GWAS}<10-4 and $P_{follow-up}$<0.01). Again these regions mostly contain genes with known immune functions. While these regions represent the next tier of associations in the GWAS, a number of findings point to additional coeliac risk variants among regions obtaining significance levels below this cut-off. Firstly, after excluding 40 coeliac loci including those reported for the first time in this study, there was residual, albeit modest inflation of the tail of the distribution of association test statistics in the GWAS data (**Figure 3.12**). Secondly, a PubMed abstract mining algorithm (GRAIL), using the 27 genome-wide significant regions as a seed, suggested enrichment for coeliac loci among SNPs of lesser significance. Among 49 regions (49 SNPs) with 10^{-3} >$P_{combined}$ >5x10^{-8} GRAIL P_{text} <0.01 was observed for 9 regions (18.4%). As a control, only 5.5% (279 of 5033) of randomly selected Hap550 SNPs reached this threshold. Moreover, 201 loci (~10%) have GRAIL P_{text} <0.05 among loci showing lesser GWAS phase association (10^{-4}<P_{GWAS}<0.01). The vast majority of these GRAIL-annotated loci contain genes with known immune functions, including some known to be associated with other autoimmune diseases.

The 39 non-HLA variants with genome-wide or suggestive levels of significance as defined in the study were estimated to account for around 6% of coeliac heritability, supplementing the 35-40% of heritability accounted for by HLA alleles. Thus this research provides major insight into the genetic risk architecture of coeliac disease. Risk is determined by large effect variants in the HLA, necessary for coeliac disease and tens or more likely hundreds or thousands of variants at non-HLA loci.

6.1.1.1 Missing heritability in coeliac disease

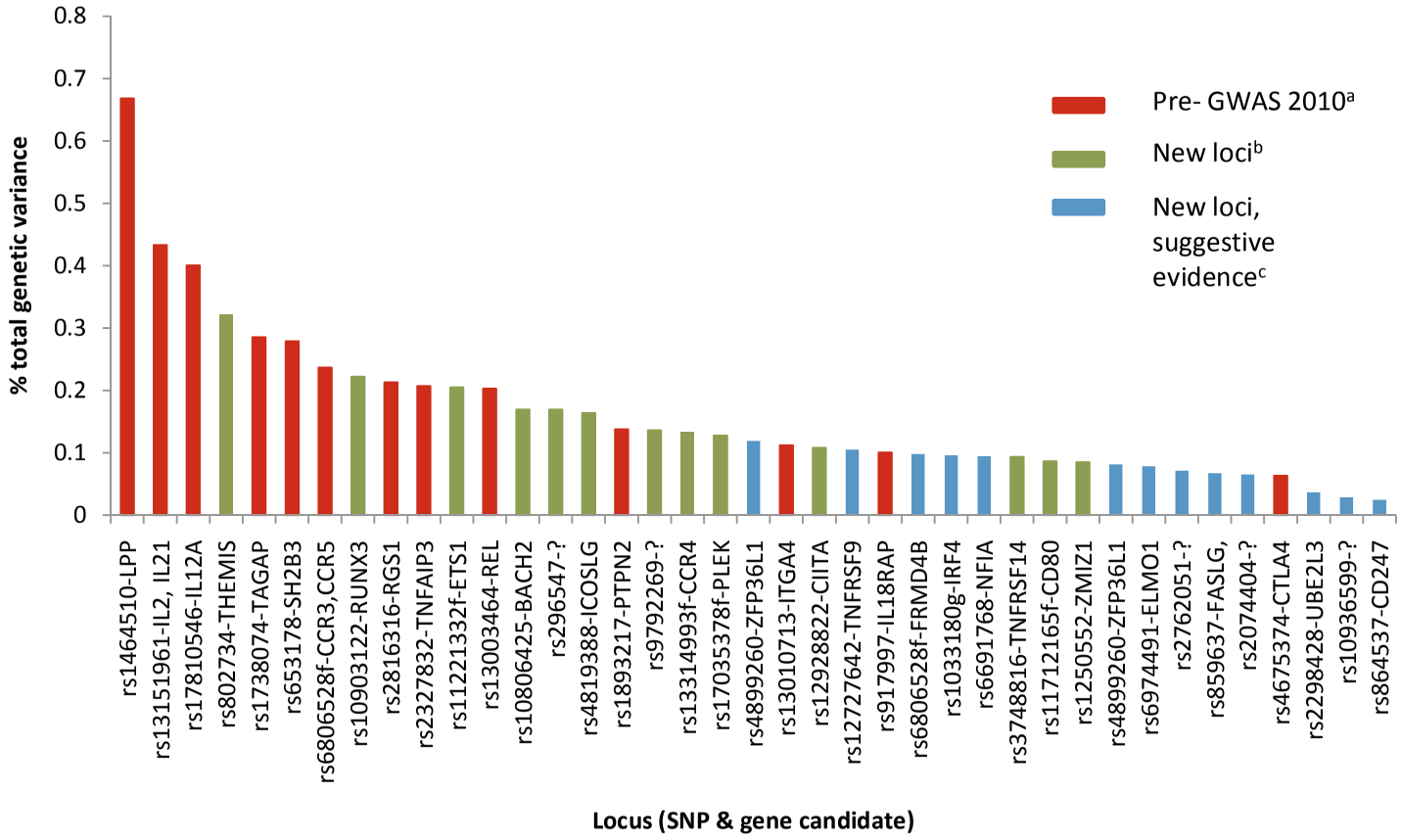
HLA and non-HLA coeliac risk variants identified in the GWAS are estimated to account for up to 50% of the heritable fraction of coeliac disease occurrence. However, this is dominated by

the contribution of HLA alleles. 39 non-HLA loci account for only around 6% of the total genetic variance of coeliac disease (**Figure 6.1**). In Crohn's disease less than 20% of genetic variance is explained by known variants using a similar estimation method (Park, Wacholder et al. 2010). Even when restricting the definition of heritability of complex traits to that percentage of phenotypic variation due to additive genetic effects, there is a clear "missing heritability" problem (Yang, Benyamin et al. 2010). Under this definition, non-additive effects (e.g. gene-gene or gene-environment interactions) are excluded. The explanation for the missing additive genetic effects is either that many genetic variants have such weak effects on phenotypic variation that they have been below the power of genome wide association studies to detect, or that GWAS SNPs are in incomplete linkage disequilibrium with causal variants (Yang, Benyamin et al. 2010). By considering the collective effects on phenotypic variation of a much larger proportion of GWAS SNPs, including those that do not reach conservative levels of significance (e.g. all GWAS SNPs or SNPs with P_{GWAS}<0.5), variants of very weak effect can be included in estimating the proportion of total genetic variance accounted for by GWAS SNPs. Purcell et al., for example, showed that around a third of schizophrenia liability was accounted for by several thousand GWAS SNPs, selected using very liberal association statistic thresholds (P_{GWAS}<0.5). These SNPs predicted schizophrenia liability in independent GWAS sample collections but not in GWAS data from other (non-psychiatric) common diseases (Purcell, Wray et al. 2009). Similarly, 45% of human height variance (more than half of the heritable fraction) could be accounted for by considering all ~300,000 SNPs in an analysis of height GWAS data, compared to just 5% explained by 40 genome wide significant loci (Yang, Benyamin et al. 2010). Both studies suggest that the remaining variance could be accounted for by incomplete linkage disequilibrium between SNPs that have been genotyped in the GWASs and causal variants. A major implication of these studies is that genetic liability to common diseases consists of thousands, rather than tens or hundreds of susceptibility loci. The true number of susceptibility loci with SNPs in the range of effect sizes identified in the coeliac GWAS (**Chapter 3**) can be estimated based on the distribution of effect sizes of SNPs already identified, and on the power to detect these associations in the original studies. This analysis was performed for Crohn's disease by Park et al., who estimated that 142 independent loci exist within the range of effect sizes seen in current GWASs, accounting for 20% of genetic variance for the trait (Park, Wacholder et al. 2010). Applying the same analysis to coeliac disease here, 253 non-HLA loci are here estimated to exist accounting for 16% of the genetic variance of coeliac disease. In this analysis, effect sizes for coeliac SNPs were calculated from logistic regression of the 4,918 follow-up cases and 5,684 controls with sample collection membership as a factorized

covariate. The follow-up collections were used here to avoid potential over-estimation of effect sizes in the GWAS discovery data set ("winner's curse"). The total genetic variance for coeliac disease was estimated from a sibling recurrence risk of 10, based on a log-normal distribution of genetic risk for polygenic traits ($\lambda_{sibling}^2 = e^{variance}$) (Pharoah, Antoniou et al. 2002).

The modest effect sizes of individual variants reflect the effect size averaged across the study population as a whole. Each of these variants accounts for tiny proportions of the heritable basis of the disease, but could nevertheless, in the right environmental and genetic context contribute much more substantially to disease risk in the individual. It is possible that many of these factors will therefore prove to have larger effects on risk in combination with other relevant genetic or, particularly environmental factors. The potential for such effects is illustrated by Cadwell et al.'s study of *ATG16L1* effects on mouse intestinal inflammation, contingent on infection with specific murine norovirus (Cadwell, Patel et al. 2010). An alternative is that genetic risk factors act largely independently, contributing truly small effects on risk in the individual, with disease arising once a liability threshold is crossed through the accumulation of sufficient genetic and environmental risk factors. This is Sewall Wright's liability threshold model of the polygenic basis of binary traits (Wright 1934). The truth perhaps, will lie somewhere in between for diseases like coeliac disease and Crohn's disease. In coeliac disease, it is plausible that multiple genes influencing T cell activation and the immunological synapse will act additively, but these effects can only translate to disease if the correct HLA-DQ molecule is present on antigen presenting cells.

Figure 6.1 **Contributions to the total genetic variance of coeliac disease of 39 non-HLA loci**



Effect sizes were estimated from the stage 2 (follow-up) sample collections comprising 4,918 cases and 5,684 controls by logistic regression with collection membership as a factorized covariate. The SNP most strongly associated with coeliac disease in the combined GWAS+follow-up analysis was chosen at each locus. [a]Loci identified in the first coeliac GWAS and 3 follow-up studies prior to the current coeliac GWAS (van Heel, Franke et al. 2007; Hunt, Zhernakova et al. 2008; Smyth, Plagnol et al. 2008; Garner, Murray et al. 2009; Trynka, Zhernakova et al. 2009) [b]Loci with SNPs at $P_{Combined}<5 \times 10^{-8}$ [c]Loci with SNPs showing suggestive association (either A. 10-6>*P*combined>5x10-8 and/or B. *P*GWAS<10-4 and *P*follow-up<0.01)

6.1.1.2 Strategies for resolving the allelic spectra in GWAS-identified regions

The above studies have little bearing on the nature of the allelic spectra at GWAS-identified loci. GWAS associations might arise from multiple variants of both rare and more common frequencies (Dickson, Wang et al. 2010). Firstly, genetic variation in these regions will be more comprehensively defined by ongoing sequencing projects, including the 1000 Genomes Project. In addition, array-based or in-solution sequence capture and amplification methods can enrich genomic DNA for regions of interest (e.g. regions showing GWAS association or the exome). Resequencing of these regions in cases is hoped to be an efficient strategy for the discovery of novel, and in particular, rare sequence variants. Exome sequencing has already proven valuable in identifying causal variants in rare Mendelian disorders, with filtering of candidate variants based on DNA segments shared by between affected individuals, by functional weighting of variants and by exclusion of common variants (Ng, Bigham et al. 2010; Ng, Buckingham et al. 2010).

Recently, sequencing of GWAS-identified genes associated with hypertriglyceridaemia led to the discovery of an excess of rare variants in all 4 tested genes in cases compared to controls (Johansen, Wang et al. 2010). For this phenotype a model incorporating both common and rare variants at the GWAS loci was best able to account for phenotypic variation. This study adds to previous examples of common and rare disease-causing variants occurring in GWAS-associated genomic regions (e.g. *NOD2* – Crohn's*, IFIH1*- type 1 diabetes) and provides hope that rare causal variants may be discovered by sequencing of coeliac GWAS-identified genes.

Fine mapping, where a much increased marker density in GWAS-identified regions is used to genotype samples in an effort to refine the region of association, has not proven to be of great value so far. Unpublished data from the WTCCC, where fine-mapping was undertaken for GWAS-identified regions originally identified in the WTCCC SNP GWAS in 2007, found that for most regions, the region of association could not be significantly reduced (unpublished data presented at Genomics of common diseases conference, Boston, 2008). Imputation of markers not directly genotyped in association studies may similarly allow some refinement of association signals and testing of whether untyped variants could account for GWAS associations. The 1000 Genomes project is providing phased data in some HapMap populations for variants with lower minor allele frequencies (down to ~1%) thus potentially

enabling imputation of rare variants and testing of whether these variants account for GWAS associations.

An alternative approach is to test for association in populations of different ancestry to the original GWASs, where the linkage disequilibrium patterns between SNPs differ from the discovery GWAS populations. Most GWASs to date have used European ancestry populations; use of African ancestry populations, where linkage disequilibrium blocks are on average smaller, may help refine associations. This strategy would be expected to offer a modest reduction in the size of the region of association. Helgason et al., for example, refined the association signal for common variants in the type 2 diabetes *TCF7L2* gene region, by replication in a West African sample collection (Helgason, Palsson et al. 2007). Another benefit of this approach arises from the fact that variants with similar effect sizes in different populations may have markedly different allele frequencies, leading to differences in power to detect risk variants. Thus variants in *KCNQ1,* a gene that encodes the target of sulphonylureas, were detected in two type 2 diabetes GWASs of East Asians (Unoki, Takahashi et al. 2008; Yasuda, Miyake et al. 2008). A meta-analysis of European GWASs showed that these variants confer similar risk in Europeans, but due to lower minor allele frequencies of these SNPs in Europeans the SNPs did not obtain genome-wide significance (Rosenberg, Huang et al. 2010). Thus studies in non-European populations may increase power to detect risk variants that have low allele frequencies in Europeans but higher frequencies in non-Europeans.

6.1.1.3 Beyond GWAS- alternative strategies for finding coeliac disease variants

For some common diseases, particularly quantitative traits, selection of samples from the extreme ends of a trait distribution (super-cases) may be helpful in enriching for disease-causing variants. For binary traits like coeliac disease and Crohn's, defining super cases and super controls is more difficult. For these traits, multiply affected families may offer the richest hunting ground for novel variant discovery, since extreme familial clustering implies disease risk variants of higher penetrance (Bodmer and Bonilla 2008). Exome or whole-genome resequencing in multiply affected families and comparison of the most distantly related affected individuals, on the assumption that they will share rare disease-causing variants, is an attractive approach, enabling exclusion of large parts of the genome not shared by these relatives. This reduces the set of candidate variants, but additional filters are still needed to

reduce the still very large number of variants identified by this approach. For example, selection of variants at sites that show high evolutionary conservation, or variants predicted to have functional (e.g. amino acid changing) effects could be used. Exclusion of known common variants (e.g. SNPs present in dbSNP) might also be used. Multiply affected families allow testing of co-segregation of candidate variants with disease and offer a potentially powerful approach for testing disease association for these variants (Cirulli and Goldstein 2010). However, the small numbers of individuals in even the largest multiply affected families mean that prior filtering of candidate variants will be necessary. As yet the power of these strategies to identify variants in common diseases is unproven.

6.1.1.4 Implications for understanding the immunopathogenesis of coeliac disease

New GWAS-identified coeliac risk regions have provided greater detail of the biological pathways involved in the pathogenesis of this disorder by highlighting candidate genes. The dominant theme remains T cell function, re-emphasizing current immunological models of the disease, where the key event is presentation of gluten peptides by HLA-DQ2 or -DQ8 expressing antigen presenting cells to CD4+ T cells. A number of genes in coeliac risk regions are well-placed to modulate this interaction, with roles in T cell co-stimulation and co-inhibition (*CTLA4/ICOS/CD28, TNFRSF14, CD80, ICOSLG, TNFRSF9, TNFSF4*) (**Table 6.1**). However, also of interest are pathways notably absent from GWAS findings. These include the IL-15/NK cell mediated responses to gluten peptides and regulation of tight junctions and intestinal permeability (e.g. zonulin) highlighted in some frequently cited immunological studies (Clemente, De Virgiliis et al. 2003; Maiuri, Ciacci et al. 2003). The lack of genes participating in these processes may indicate that these pathways are of less importance than previously supposed or that they have secondary roles that occur after the induction of loss of adaptive immune tolerance to gluten. It is theoretically possible that population genetic variation affecting the function of these processes is relatively limited and therefore does not translate to variation in coeliac susceptibility, at least within the power of the current study to detect. In this scenario, these processes may still have causal roles in coeliac pathogenesis if more significant perturbations are driven by environmental factors. For example, viral enteritis causes increased intestinal permeability and this might be an important event in coeliac pathogenesis. However, despite these caveats the genetic evidence should be interpreted as indicating that coeliac susceptibility is determined primarily by perturbations in

T cell function, presumably affecting immunological tolerance to gluten peptides. Coeliac disease requires antigen presenting cells to express DQ2- or DQ8- together with factors that permit loss of immunological tolerance to gluten. The site of T cell gluten encounter that leads to loss of tolerance in coeliac disease is unknown. However, this research has drawn attention to the thymus, the site of T cell development and tolerance induction for auto reactive T cells. *THEMIS* and *RUNX3* are genes with key roles in T cell development in the thymus. Several other coeliac-associated genes are expressed by developing thymocytes (e.g. *ETS1, TNFRSF14*) but also function in mature peripheral T cells. Understanding the tissue context in which these genes contribute to coeliac disease risk is a challenge that will require immunological study and a challenge that is currently hampered by the lack of a satisfactory animal model of coeliac disease.

Table 6.1 **T cell co-stimulatory and co-inhibitory genes from the immunoglobulin and TNFR superfamilies and associations with coeliac disease (chapter 3). Genes with genome-wide significant or suggestive evidence of association highlighted in bold. Table adapted from (Murphy, Nelson et al. 2006)**

Molecule (gene name)	Top SNP association[a] (P_{GWAS}, $P_{Combined}$)	Expression	Ligand (gene name)	Top SNP association (P_{GWAS}, $P_{Combined}$)	Expression
Co-stimulatory immunoglobulin domain containing receptors					
CD28	8.8×10^{-8}; 5.79×10^{-9}	T cells (constitutively expressed)	**CD80**, CD86	5.4×10^{-7}; 8.03×10^{-9}	B cells, monocytes, T cells, inducible somatic tissues
ICOS	8.8×10^{-8}; 5.79×10^{-9}	Activated T cells, activated DCs	**ICOSL**	3.42×10^{-5}; 2.46×10^{-9}	B cells, monocytes, T cells, inducible somatic tissues
Inhibitory immunoglobulin domain containing receptors					
CTLA4	8.8×10^{-8}; 5.79×10^{-9}	Activated T cells	**CD80**, CD86	5.4×10^{-7}; 8.03×10^{-9}	B cells, monocytes, T cells, inducible somatic tissues
PD1 (CD279)	>0.01; nd	Activated T cells, activated B cells, activated DCs	**PDL1 (CD274)**	6.20×10^{-6}; 4.41×10^{-6}	B cells, T cells, some somatic tissues, inducible in monocytes and DCs
			PDL2	>0.01	
BTLA (CD272)	1.57×10^{-5}; 6.76×10^{-6}	T cells, B cells, DCs, myeloid cells	**HVEM (TNFRSF14)**	4.93×10^{-7}; 3.28×10^{-9}	T cells, B cells, NK cells, DCs, myeloid cells, inducible in somatic tissues
Co-stimulatory TNFRs					
4-1BB (TNFRSF9)	3.06×10^{-5}; 9.11×10^{-8}	Activated T cells, activated B cells, activated DCs	4-1BBL (TNFSF9)	>0.01; nd	Activated T cells, activated B cells, activated DCs and activated monocytes
OX40 (TNFRSF4)	1.65×10^{-3}; nd	Activated T cells, activated B cells, activated DCs	**OX40L (TNFSF4)**	8.15×10^{-5}; 1.75×10^{-6}	Activated T cells, activated B cells, activated DCs and activated monocytes

Table 6.1 (cont.)

CD27 (TNFRSF7)	1.16 x 10^{-3}; nd	T cells, activated B cells	CD70 (TNFSF7)	>0.01; nd	Activated T cells, activated B cells, activated DCs and activated monocytes
CD30 (TNFRSF8)	1.36 x 10^{-5}; 8.43 x 10^{-4}	Activated T cells, activated B cells, activated DCs	CD30L (TNFSF8)	7.51 x 10^{-4}; nd	Activated T cells, activated B cells, activated monocytes
CD40 (TNFRSF5)	2.51 x 10^{-3}; nd	B cells, DCs	CD40L (TNFSF5)	>0.01; nd	Activated T cells, activated DCs
HVEM (TNFRSF14)	4.93 x10^{-7}; 3.28 x 10^{-9}	T cells, B cells, NK cells, DCs, myeloid cells, inducible in somatic tissues	LIGHT (TNFSF14)	>0.01; nd	Immature DCs, monocytes, activated T cells

[a] SNP with strongest association in gene region

nd- not done. No SNPs from locus genotyped in follow-up sample collections

Another intriguing finding has been the coeliac GWAS association implicating genes with roles in innate immune responses to viruses (*TLR7, TLR8, BACH2*). *BACH2,* associated also with type 1 diabetes, encodes a B cell-specific transcription factor and has been shown to be important in mediating innate immune responses to viral nucleic acids, though whether this is the important pathway for autoimmune disease pathogenesis is unclear (Hong, Kim et al. 2008; Todd 2010). The *TLR7/TLR8* association with coeliac disease, however, more clearly points towards a role for RNA viruses in coeliac disease. Epidemiological data exist suggesting that rotavirus infection is more common in children developing coeliac disease (Stene, Honeyman et al. 2006). Rotavirus infection and enteroviruses have been implicated in type 1 diabetes and recent data implicate viruses in Crohn's disease (Ballotti and de Martino 2007; Cadwell, Patel et al. 2010; Hober and Sauter 2010). This includes the observation that mice expressing hypomorphic *ATG16L1* have abnormal intestinal inflammatory responses contingent on infection with a specific murine norovirus strain. This has provided an example of how common variant: environmental interactions may work to generate intestinal inflammatory (Cadwell, Patel et al. 2010). Investigating how the viral response genes that have been associated with these diseases (e.g. *IFIH1*, *TLR7/TLR8, BACH2*) function on viral interaction is likely to be informative in understanding the pathogenesis of these conditions. It may be possible to identify specific viruses that are important in causing coeliac disease by serological testing of monozygotic twins who are discordant for coeliac disease for viruses known to infect

the human gastrointestinal tract. Viral exposures segregating with disease-affection status in twins would be strong candidates for disease causation. Prospective follow-up of the unaffected twin with further interval serology, might identify viral exposures preceding onset of the disease and by using monozygotic twins, genetic variation is controlled. Alternatively, viral exposures have also been postulated to be protective against the development of autoimmunity ('the hygiene hypothesis') and in this model would be expected to cluster in unaffected twins.

6.1.2 Function of 5p13.1 genetic variants in Crohn's disease

A second phenotype investigated in this thesis was Crohn's disease (**Chapter 5**). In this work, a genetic risk region previously identified in genome wide association studies was explored. Here the aim was to move beyond the association of genetic variants with disease to understanding how variants altered gene function and contributed to disease susceptibility. The region of Crohn's disease association on 5p13.1 was devoid of genes. Similar associations with *gene deserts* have been frequent in GWASs for complex disease traits but causal variants accounting for these associations have not been identified. The work here sought to explore the hypothesis that the association was caused by *cis*-acting regulatory sequence variants that influenced expression of nearby genes and in particular *PTGER4*. The work also aimed to determine whether these variants caused changes in the function of the *PTGER4* gene product, the prostaglandin EP4 receptor and how this might contribute to Crohn's pathogenesis.

This work validated published observations that Crohn's disease-associated SNPs correlated with expression of the closest known gene, *PTGER4,* in immune cells. This has now been shown in larger datasets of both lymphoblastoid cell lines and primary human leucocytes. It may in future be helpful to examine *cis* effects of Crohn's SNPs on *PTGER4* expression in immune cell subsets, since it has previously been shown that quantitative trait loci can exert tissue-specific effects with opposite positive and negative expression correlations in different tissues (McCarroll, Huett et al. 2008). However, since expression quantitative trait loci are prevalent in the human genome, such correlations could be co-incidental rather than indicating that the Crohn's association is due to the presence of expression regulatory variants. Thus a further aim was to test whether 5p13.1 Crohn's variants influenced *PTGER4* function in a biological context relevant to Crohn's disease. This is a crucial step in proving that *PTGER4* is the causal gene in the region. A pharmacological approach aiming to develop a *PTGER4*-

specific assay in primary human immune cells was used here. Experiments under a variety of conditions, assaying a variety of prostanoid-induced cytokine responses, showed the limitations of this approach. The available pharmacological agents appear to lack sufficient potency and/or receptor selectivity to assay the prostaglandin EP4 receptor effectively. Secondly, cell cytokine responses showed major variation that can not be attributed simply to 5p13.1 genetic variation. This variation was sufficient to prevent use of the assay in genetic variant – cytokine response correlation studies.

These experiments have illustrated some of the difficulties of moving from genetic associations in GWASs to biological interpretations of these associations. Subtle effects on function are very difficult to distinguish in the context of large background variation in responses. This variation is commonly observed in human studies, in contrast to studies of highly inbred murine strains. Controlling this variation, while retaining a biological context that is relevant to the disease in question, is extremely challenging. Large variation in prostanoid responses (with PGE_2 and EP4 agonists) was observed between individuals, but also between experiments on cells from the same individual obtained on different days. Within-individual variation includes unintended variation in experimental conditions (e.g. cell numbers, relative proportions of cell subsets in mixed cell populations, cell viability and receptor expression). Between-individual variation also includes genetic variation in genes and pathways that contribute to prostanoid responses apart from *PTGER4* variation (e.g. LPS responses, EP receptor second messengers, cytokine genes etc). To overcome this variation, one future approach would be to use very large numbers of individuals to increase power to detect subtle effects. Investigating the relationship between the 5p13.1 Crohn's SNPs, *PTGER4* expression and EP4 function in the same individuals may help to determine whether the genetic risk variants do indeed act through increased expression and a proportionate increase in EP4 signalling. However, the true causal relevance of this mechanism would require identification of causal regulatory variants that influence transcription factor binding and activation. Moreover, the possibility that rare, protein-altering variants in *PTGER4* or other local genes are responsible for the Crohn's GWAS association cannot be excluded. Until a fuller ascertainment of the allelic spectrum at this locus has been undertaken, this possibility can not be discounted. Thus, proceeding without knowledge of causal variants at a locus is likely to leave many questions unanswered. Experiments attempting to correlate gene function with tag SNP genotype, where causal variants are unknown, are not recommended on the basis of the experience presented in this thesis. Finding causal variants in GWAS regions remains a high priority.

6.1.3 Common genetic risk variants for azathioprine/6-mercaptopurine-induced pancreatitis

The third phenotype investigated in this thesis was acute pancreatitis triggered by azathioprine or 6-mercaptopurine, drugs used in the treatment of intestinal inflammatory disorders. There are both theoretical reasons and emerging empirical data to suggest that some drug-induced idiosyncratic reactions have a unique genetic risk architecture that includes common, highly penetrant causal variants. Such variants are found infrequently for common diseases, since deleterious variants have usually been kept at low frequencies in populations by natural selection. Exceptions may occur where environmental exposures in human populations change rapidly, limiting the time for selection to operate, or where balancing selection occurs. Thus HLA risk alleles in autoimmune diseases like coeliac disease and type 1 diabetes are common and confer substantial relative risks (ORs > 5 for some alleles) (Cucca, Lampis et al. 2001; Margaritte-Jeannin, Babron et al. 2004). These alleles are highly population differentiated, variation perhaps indicative of natural selection on populations exposed to different and rapidly changing patterns of infections.

Drug-induced idiosyncratic reactions offer an extreme model for this process. Here introduction of a novel environmental agent to a naive population can reveal genetically-determined variation in the host response. HLA-B*5701 –restricted abacavir hypersensitivity represents an extreme example that has led to pre-treatment genetic screening. For azathioprine-induced pancreatitis, moderate HLA region association was observed, but will require follow-up in additional case and control samples to test whether it is a true association. It is possible to conclude already that, even assuming that causal alleles may have effects on risk exceeding those of tag SNPs assayed in the GWAS, common predisposing variants will have only moderate effects on risk (allelic ORs less than ~ 5). This falls well short of the odds ratios observed for abacavir and flucloxacillin variants and is likely to limit the utility of HLA genotyping for clinical prediction of pancreatitis in individuals considered for azathioprine or mercaptopurine therapy. This alone is an important implication of the research and constitutes evidence against the hypothesis that common variants of very large effect influence the risk of this idiosyncratic reaction. Thus, it is anticipated that the main benefit of this research will be in identifying important biological pathways causing a phenotype whose pathogenesis has hitherto been unknown.

6.2 Prospects for genetic risk modelling of common human phenotypes

Risk prediction for complex diseases using GWAS-identified genetic risk variants is currently of modest value. Simple models that assume multiplicative effects of risk variants are likely to fall short for accurate disease risk prediction, even if all heritable genetic variation was identified (Clayton 2009). Understanding how genetic and environmental factors combine to cause disease in individuals may eventually lead to more sophisticated models that incorporate these interactions and have greater prediction accuracy, but such models are currently a distant prospect. Arguably the greatest area of promise for personalized medicine is pharmacogenomics. Several recent examples have identified common genetic variants conferring large effects on risk of drug responses. The next few years are likely to see a wealth of genomics studies of drug responses and it is anticipated that some variants identified will have sufficiently large effects on risk that they can be used in clinical practice. Those areas of clinical practice where severe drug toxicities and sub-optimal effectiveness are currently accepted because of a lack of alternatives and the severity of the underlying disease may benefit most from pharmacogenomic insights. Genetic variation influencing treatment responses to pegylated interferon-alpha and ribavirin in individuals chronically infected with hepatitis C virus has provided an early illustration of such benefits. This treatment is prolonged, of variable effectiveness and associated with substantial limiting adverse effects. Common genetic variation in *IL28B* affecting sustained clearance of hepatitis C following pegylated Interferon-α and ribavirin treatment was identified by genome wide association studies (Ge, Fellay et al. 2009). Prospective replication confirms that the effect size is sufficiently large that it may be clinically useful in predicting treatment response (Thompson, Muir et al. 2010). More recently still, an inosine triphosphatase (*ITPA*) inactivating gene variant was found to be protective for ribavirin induced haemolytic anaemia (Fellay, Thompson et al. 2010). At present, these genetic findings have not yet reached clinical practice and we can only speculate on their eventual utility. Individuals with poor response *IL28B* genotypes might benefit from more intensive pegylated interferon and ribavirin, alternative treatments or an earlier discontinuation of therapy in the absence of an early virological response. Ribavirin dose might also be usefully adjusted according to *ITPA* genotype. It may prove possible to pharmacologically inhibit inosine triphosphatase to protect against haemolytic anaemia in individuals with functional *ITPA* receiving ribavirin. These findings illustrate the promise of pharmacogenomics. It is hoped that in inflammatory bowel disease, larger

pharmacogenomic studies of azathioprine and other immunosuppressant and biologic therapies will identify similarly helpful genetic risk variants influencing drug efficacy and adverse effects.

6.3 Overlap between genetic risk variants in intestinal inflammatory diseases and between autoimmune diseases

Extensive sharing of genetic risk regions between chronic immune mediated diseases has been a major finding of the genome wide association study era. The true extent of sharing is not yet clear. Sampling variation between studies undoubtedly contributes to variation in loci reaching defined levels of significance. This can be illustrated by retrospectively estimating the power of an association study to detect the SNP associations identified the original discovery study, using the effect sizes (odds ratio) observed in an independent collection. For coeliac disease, for example, two new loci approached genome-wide significance in the GWAS phase (rs11221335-*ETS1* P_{GWAS}= 4.16 x 10^{-11}; rs1250552-*ZMIZ1* P_{GWAS}=5.80 x 10^{-8}). Using odds ratios from the follow-up collections, it was estimated that the power to detect these associations at genome-wide significance was 0.97 for *ETS1* but only 0.27 for *ZMIZ1* in a study of the size of the combined GWAS and follow-up (25,000 samples). This, and differences in the importance of individual loci in one disease versus another is likely to contribute to underestimation of the true degree of genetic sharing between autoimmune diseases. Instances where regions showing association in one disease have been systematically tested in a second disease offer the best opportunity to estimate the degree of sharing. Using this approach in a coeliac disease-type 1 diabetes study, around half of 28 risk regions studied showed strong or suggestive evidence of association in both diseases (Smyth, Plagnol et al. 2008). Similarly, testing Crohn's disease SNPs in ulcerative colitis led to an estimate of sharing at around half of the loci tested (McGovern, Gardet et al. 2010). In both cases the degree of sharing is greater than the degree of epidemiological overlap, suggesting perhaps that environmental variation plays a relatively greater role in determining disease type than genetics. Under this model genetics may contribute core autoimmune/chronic immune disease susceptibility, with additional mainly environmental factors determining which disease(s) develop.

A further interesting finding emphasized in the type 1 diabetes-coeliac study, was that associated variants at risk regions quite often differed between diseases and in some cases a SNP allele conferring risk to one disease conferred protection to the other (e.g. rs917997-*IL18RAP* and *r*s1738074-*TAGAP*). This adds further complexity, suggesting that while genetic

variation at a locus may influence more than one disease, the mechanism may be different, perhaps again relating to different environmental exposures interacting with the gene in the different diseases. Such opposite effect associations may arise from a causal variant conferring risk or protection of disease depending on environmental exposure or may be due to different causal variants in each disease being tagged by the same SNP.

In coeliac disease examples of truly coeliac specific genetic risk regions have not been easy to identify. For example, the *LPP* gene region which seemed to be coeliac disease specific following the first wave of autoimmune disease GWASs, has recently been strongly associated with vitiligo (same SNP- rs1464510) with additional suggestive evidence of association in juvenile and rheumatoid arthritis since the publication of the coeliac GWAS (chapter 3) (Coenen, Trynka et al. 2009; Hinks, Martin et al. 2010; Jin, Birlea et al. 2010). Nearly all regions examined contain genes that function in the immune system, with no genes with obvious intestinal specificity. Even the region containing *ITGA4* encoding alpha 4 integrin, one half of the α4β7 gut-homing T-cell expressed integrin, does not appear specific to intestinal diseases. This region has been associated with ankylosing spondylitis and it is also notable that *ITGA4* is not necessarily the causal gene in the region (Reveille, Sims et al. 2010). These findings therefore support the idea that the non-HLA component of the genetic predisposition to coeliac disease contributes a T-cell orientated general autoimmune tendency, with specific HLA alleles and environmental factors (gluten, viruses, breastfeeding) required to trigger the coeliac phenotype. Understanding of the precise causal variants in GWAS regions and their effects in different diseases may in future allow refinement of this model and show important differences in the way causal variants at shared risk regions act in the pathogenesis of different autoimmune diseases.

6.4 Sex bias in coeliac disease.

The cause of the female preponderance of coeliac disease is unknown and has not been explained by the genetic findings presented in this thesis. A female preponderance of HLA-DQ2/-DQ8 has previously been reported, but was not observed in our data (Megiorni, Mora et al. 2008; Dubois, Hunt et al. 2009). Female sex hormones are known to modulate immune responses and this may explain increased female coeliac incidence after puberty does not easily explain the childhood female coeliac excess seen in population based studies (Bingley, Williams et al. 2004; Fish 2008).

6.5 Epigenetics

Epigenetics refers to heritable non-DNA sequence variations. This includes changes that persist across cell replications and less commonly trans-generationally. Such changes include methylation of CpG dinucleotides and histone modifications that affect DNA folding and availability for transcription factor binding. CpG islands are 0.5 to 5kb long sequences of GC dinucleotide repeats, often found in gene promoters. Methylation of CpG islands correlates negatively with gene expression (Feil and Berger 2007). Human diseases can result from heritable epigenetic modifications, the classic examples involving specific gene methylations and aberrant transcription in Beckwith-Wiedemann, Prader-Willi and Angelman syndromes. However, in common diseases, the importance of epigenetics is less certain.

There has been discussion in the literature as to whether heritable epigenetic changes might contribute to the 'missing heritability' discussed above. However, most epigenetic changes are not inherited across generations, and indeed those that do and persist for several generations ought to be highly correlated with genetic markers used in GWASs and therefore would not be 'missing' from heritability estimates (McCarthy and Hirschhorn 2008). On the other hand epigenetic changes that decay rapidly across generations ought not to contribute significantly to overall heritable risk (Slatkin 2009). Thus, it seems unlikely that trans-generational epigenetic modifications contribute a large part of the missing heritability observed from GWAS data.

6.6 Concluding remarks

New tools and bioinformatics resources in genetics have enabled a leap forward in the study of complex human phenotypes. These advances extend not just to increased understanding of the genetic basis of these disorders, but to fundamental insights into their biology and pathogenesis. The introduction of whole-genome assays has led to a shift away from hypothesis driven experiments, to relatively unbiased screening approaches. Human genetic variation is finite and not rapidly changing. For the first time we have had tools to assay a significant proportion of this variation in large numbers of individuals using whole genome SNP genotyping microarrays. These technologies continue to advance, with newer platforms containing millions rather than hundreds of thousands of SNP assays. Using Park et al.'s INPower method, extrapolating from the number of loci detected in the current coeliac GWAS,

it was estimated that 105 loci would obtain genome wide significance ($P < 5 \times 10^{-8}$) in a 50,000 sample GWAS or 197 loci in a 100,000 sample GWAS (Park, Wacholder et al. 2010). Larger GWASs in coeliac disease (and IBD) are therefore likely to enable further new variant discovery, but achieveable sample size inevitably will be a major factor limiting the identification of further susceptibility loci through the GWAS approach.

Over the time period of the research presented in this thesis, rapid improvements in DNA sequencing technologies and reductions in cost have steered geneticists towards experiments that offer the hope of much more comprehensively assaying genetic variation contributing to common diseases. Initially, these efforts will be restricted to high yield targets in the genome (e.g. GWAS-identified regions, the exome), but eventually whole genome sequencing will become a viable strategy for disease association studies. One hope is that these approaches will identify causal genetic variants with larger effect sizes than those observed in GWASs and therefore will overcome some of the power limitations imposed by foreseeable sample sizes. In some cases, sequence level data is expected to reveal obvious causal variants that affect, for example, amino acid sequence and protein function. However, in other cases, we can anticipate that causal variants will not be easily recognized from DNA sequence alone. The consequences of sequence variation in regulatory elements, for example is not easily predicted. Inevitably therefore, researchers will need to again employ hypothesis driven approaches to identify the mechanisms by which sequence variants influence cell biology. At present, there are few obvious short cuts to the goal of understanding how genes change biology to cause complex diseases. Successes so far have employed a variety of molecular biology techniques based on hypotheses pertinent to the genes and diseases in question (Cadwell, Liu et al. 2008; Saitoh, Fujita et al. 2008; Cadwell, Patel et al. 2010; Cooney, Baker et al. 2010).

Future studies will explore the function of GWAS region candidate genes in a variety of models. Gene knockout and gene manipulation studies in mice can allow rapid evaluation of loss of gene function in a variety of tissues including the intestine (more difficult to study in humans). Inbred mice strains enable precise control of genetic variation that can confound studies in humans. On the other hand experience with *NOD2,* for example, cautions that functional interpretations of these mouse models does not always readily translate to humans: opposing effects have been observed in humans versus mice (Maeda, Hsu et al. 2005; van Heel, Ghosh et al. 2005). Thus, complementary approaches in humans will be critical in

supporting mouse findings. These studies may include gene silencing through siRNA both in the simplest models (e.g. human cell lines) and more biologically relevant, but complex settings (e.g. primary human immune cells, *ex vivo* intestinal mucosa). Non-diseased individuals offer the opportunity to study the function of these genes without the confounding effects of inflammation. However, in some cases it may be preferable to use tissues or cells from disease-affected individuals. For example, in coeliac disease, peripheral T cells and intestine-derived T cell lines and clones respond to stimulation with gluten peptides presented by antigen presenting cells (Lundin, Scott et al. 1993; Gjertsen, Sollid et al. 1994; Anderson, Degano et al. 2000). An immunodominant gluten peptide could be used to evaluate non-HLA coeliac gene function in this model, ensuring that HLA types are matched. In the absence of the knowledge of causal variants such studies could employ gene silencing (siRNA) or inhibition of gene products (e.g. monoclonal antibodies).

There is much work to be done. History shows that the journey from genetic association to full understanding of function may take many years: the *HLA* association with coeliac disease was first identified in 1972, but it was only in the last decade that the interaction between gluten peptides and HLA DQ2 and DQ8 heterodimers was been fully solved (Falchuk, Rogentine et al. 1972; Sollid 2002). Efforts to identify genetic variants causing the new disease associations and efforts to understand the function of implicated genes should proceed in parallel. This research presented here offers many new avenues for future enquiry in both these directions. There is real hope that, with serendipity no doubt playing its part, this will lead to a more thorough understanding of causes of these complex human disorders.

References

Abramovitz, M., M. Adam, et al. (2000). "The utilization of recombinant prostanoid receptors to determine the affinities and selectivities of prostaglandins and related analogs." Biochim Biophys Acta **1483**(2): 285-93.

Abu-Shakra, M. and Y. Shoenfeld (2001). "Azathioprine therapy for patients with systemic lupus erythematosus." Lupus **10**(3): 152-3.

Accomando, S. and F. Cataldo (2004). "The global village of celiac disease." Dig Liver Dis **36**(7): 492-8.

Adamovic, S., S. S. Amundsen, et al. (2008). "Association study of IL2/IL21 and FcgRIIa: significant association with the IL2/IL21 region in Scandinavian coeliac disease families." Genes Immun **9**(4): 364-7.

Alvarez-Lobos, M., J. I. Arostegui, et al. (2005). "Crohn's disease patients carrying Nod2/CARD15 gene variants have an increased and early need for first surgery due to stricturing disease and higher rate of surgical recurrence." Ann Surg **242**(5): 693-700.

Amundsen, S. S., A. J. Monsuur, et al. (2006). "Association analysis of MYO9B gene polymorphisms with celiac disease in a Swedish/Norwegian cohort." Hum Immunol **67**(4-5): 341-5.

Amundsen, S. S., J. Rundberg, et al. (2010). "Four novel coeliac disease regions replicated in an association study of a Swedish-Norwegian family cohort." Genes Immun **11**(1): 79-86.

Anand, B. S., J. Piris, et al. (1981). "The timing of histological damage following a single challenge with gluten in treated coeliac disease." Q J Med **50**(197): 83-94.

Anderson, C. A., F. H. Pettersson, et al. (2008). "Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms." Am J Hum Genet **83**(1): 112-9.

Anderson, R. P., P. Degano, et al. (2000). "In vivo antigen challenge in celiac disease identifies a single transglutaminase-modified peptide as the dominant A-gliadin T-cell epitope." Nat Med **6**(3): 337-42.

Anderson, R. P., D. A. van Heel, et al. (2005). "T cells in peripheral blood after gluten challenge in coeliac disease." Gut **54**(9): 1217-23.

Annese, V., G. Lombardi, et al. (2005). "Variants of CARD15 are associated with an aggressive clinical course of Crohn's disease--an IG-IBD study." Am J Gastroenterol **100**(1): 84-92.

Annunziato, F., L. Cosmi, et al. (2007). "Phenotypic and functional features of human Th17 cells." J Exp Med **204**(8): 1849-61.

Aoun, E., C. C. Chang, et al. (2008). "Pathways to injury in chronic pancreatitis: decoding the role of the high-risk SPINK1 N34S haplotype using meta-analysis." PLoS One **3**(4): e2003.

Aoun, E., V. Muddana, et al. (2010). "SPINK1 N34S is strongly associated with recurrent acute pancreatitis but is not a risk factor for the first or sentinel acute pancreatitis event." Am J Gastroenterol **105**(2): 446-51.

Arentz-Hansen, H., R. Korner, et al. (2000). "The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase." J Exp Med **191**(4): 603-12.

Arentz-Hansen, H., S. N. McAdam, et al. (2002). "Celiac lesion T cells recognize epitopes that cluster in regions of gliadins rich in proline residues." Gastroenterology **123**(3): 803-9.

Arking, D. E., A. Pfeufer, et al. (2006). "A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization." Nat Genet **38**(6): 644-51.

Arman, M., N. Aguilera-Montilla, et al. (2009). "The human CD6 gene is transcriptionally regulated by RUNX and Ets transcription factors in T cells." Mol Immunol **46**(11-12): 2226-35.
Aslan, A., C. Karaveli, et al. (2007). "Does noncomplicated acute appendicitis cause bacterial translocation?" Pediatr Surg Int **23**(6): 555-8.
Audrezet, M. P., A. Dabricot, et al. (2008). "Validation of high-resolution DNA melting analysis for mutation scanning of the cystic fibrosis transmembrane conductance regulator (CFTR) gene." J Mol Diagn **10**(5): 424-34.
Balding, D. J. (2006). "A tutorial on statistical methods for population association studies." Nat Rev Genet **7**(10): 781-91.
Ballotti, S. and M. de Martino (2007). "Rotavirus infections and development of type 1 diabetes: an evasive conundrum." J Pediatr Gastroenterol Nutr **45**(2): 147-56.
Barrett, J. C., D. G. Clayton, et al. (2009). "Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes." Nat Genet.
Barrett, J. C., B. Fry, et al. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-5.
Barrett, J. C., S. Hansoul, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." Nat Genet.
Bastida, G., P. Nos, et al. (2005). "Incidence, risk factors and clinical course of thiopurine-induced liver injury in patients with inflammatory bowel disease." Aliment Pharmacol Ther **22**(9): 775-82.
Baumgart, D. C. and S. R. Carding (2007). "Inflammatory bowel disease: cause and immunobiology." Lancet **369**(9573): 1627-40.
Bean, R. H. (1962). "The treatment of chronic ulcerative colitis with 6-mercaptopurine." Med J Aust **49(2)**: 592-3.
Beaugerie, L., N. Brousse, et al. (2009). "Lymphoproliferative disorders in patients receiving thiopurines for inflammatory bowel disease: a prospective observational cohort study." Lancet **374**(9701): 1617-25.
Bentley, D. R. (2006). "Whole-genome re-sequencing." Curr Opin Genet Dev **16**(6): 545-52.
Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-9.
Bermejo, F., A. Lopez-Sanroman, et al. (2008). "Acute pancreatitis in inflammatory bowel disease, with special reference to azathioprine-induced pancreatitis." Aliment Pharmacol Ther **28**(5): 623-8.
Bernell, O., A. Lapidus, et al. (2000). "Risk factors for surgery and postoperative recurrence in Crohn's disease." Ann Surg **231**(1): 38-45.
Billot, X., A. Chateauneuf, et al. (2003). "Discovery of a potent and selective agonist of the prostaglandin EP4 receptor." Bioorg Med Chem Lett **13**(6): 1129-32.
Bingley, P. J., A. J. Williams, et al. (2004). "Undiagnosed coeliac disease at age seven: population based prospective birth cohort study." Bmj **328**(7435): 322-3.
Birmingham, C. L., A. C. Smith, et al. (2006). "Autophagy controls Salmonella infection in response to damage to the Salmonella-containing vacuole." J Biol Chem **281**(16): 11374-83.
Birney, E., J. A. Stamatoyannopoulos, et al. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**(7146): 799-816.
Bodmer, W. and C. Bonilla (2008). "Common and rare variants in multifactorial susceptibility to common diseases." Nat Genet **40**(6): 695-701.
Bonamico, M., P. Mariani, et al. (1994). "Celiac disease in two sisters with a mother from Cape Verde Island, Africa: a clinical and genetic study." J Pediatr Gastroenterol Nutr **18**(1): 96-9.

Bonasio, R., M. L. Scimone, et al. (2006). "Clonal deletion of thymocytes by circulating dendritic cells homing to the thymus." Nat Immunol **7**(10): 1092-100.
Bonaz, B., J. Boitard, et al. (2003). "Tioguanine in patients with Crohn's disease intolerant or resistant to azathioprine/mercaptopurine." Aliment Pharmacol Ther **18**(4): 401-8.
Boomsma, D., A. Busjahn, et al. (2002). "Classical twin studies and beyond." Nat Rev Genet **3**(11): 872-82.
Boone, D. L., E. E. Turer, et al. (2004). "The ubiquitin-modifying enzyme A20 is required for termination of Toll-like receptor responses." Nat Immunol **5**(10): 1052-60.
Bourgey, M., G. Calcagno, et al. (2007). "HLA related genetic risk for coeliac disease." Gut **56**(8): 1054-9.
Bukhave, K. and J. Rask-Madsen (1980). "Saturation kinetics applied to in vitro effects of low prostaglandin E2 and F 2 alpha concentrations on ion transport across human jejunal mucosa." Gastroenterology **78**(1): 32-42.
Burchenal, J. H., M. L. Murphy, et al. (1953). "Clinical evaluation of a new antimetabolite, 6-mercaptopurine, in the treatment of leukemia and allied diseases." Blood **8**(11): 965-99.
Burton, P. R., D. G. Clayton, et al. (2007). "Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants." Nat Genet.
Cadwell, K., J. Y. Liu, et al. (2008). "A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells." Nature **456**(7219): 259-63.
Cadwell, K., K. K. Patel, et al. (2009). "A common role for Atg16L1, Atg5 and Atg7 in small intestinal Paneth cells and Crohn disease." Autophagy **5**(2): 250-2.
Cadwell, K., K. K. Patel, et al. (2010). "Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine." Cell **141**(7): 1135-45.
Calkins, B. M. (1989). "A meta-analysis of the role of smoking in inflammatory bowel disease." Dig Dis Sci **34**(12): 1841-54.
Calne, R. Y. (1960). "The rejection of renal homografts. Inhibition in dogs by 6-mercaptopurine." Lancet **1**(7121): 417-8.
Cappell, M. S. and K. M. Das (1989). "Rapid development of pancreatitis following reuse of 6-mercaptopurine." J Clin Gastroenterol **11**(6): 679-81.
Cardon, L. R. and L. J. Palmer (2003). "Population stratification and spurious allelic association." Lancet **361**(9357): 598-604.
Caristi, S., G. Piraino, et al. (2005). "Prostaglandin E2 induces interleukin-8 gene transcription by activating C/EBP homologous protein in human T lymphocytes." J Biol Chem **280**(15): 14433-42.
Carpten, J. D., A. L. Faber, et al. (2007). "A transforming mutation in the pleckstrin homology domain of AKT1 in cancer." Nature **448**(7152): 439-44.
Carter, C. O. (1969). "Genetics of common disorders." Br Med Bull **25**(1): 52-7.
Carter, M. J., A. J. Lobo, et al. (2004). "Guidelines for the management of inflammatory bowel disease in adults." Gut **53 Suppl 5**: V1-16.
Cataldo, F., D. Lio, et al. (2002). "Consumption of wheat foodstuffs not a risk for celiac disease occurrence in burkina faso." J Pediatr Gastroenterol Nutr **35**(2): 233-4.
Catassi, C., M. Doloretta Macis, et al. (2001). "The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence." Tissue Antigens **58**(6): 402-6.
Chang, Y. T., M. C. Chang, et al. (2008). "Association of cystic fibrosis transmembrane conductance regulator (CFTR) mutation/variant/haplotype and tumor necrosis factor (TNF) promoter polymorphism in hyperlipidemic pancreatitis." Clin Chem **54**(1): 131-8.
Chessells, J. M., C. Bailey, et al. (1995). "Intensification of treatment and survival in all children with lymphoblastic leukaemia: results of UK Medical Research Council trial UKALL X.

Medical Research Council Working Party on Childhood Leukaemia." Lancet **345**(8943): 143-8.
Chessman, D., L. Kostenko, et al. (2008). "Human leukocyte antigen class I-restricted activation of CD8+ T cells provides the immunogenetic basis of a systemic drug hypersensitivity." Immunity **28**(6): 822-32.
Chimpanzee-Sequencing-and-Analysis-Consortium (2005). "Initial sequence of the chimpanzee genome and comparison with the human genome." Nature **437**(7055): 69-87.
Cho, J. H. and C. Abraham (2007). "Inflammatory bowel disease genetics: Nod2." Annu Rev Med **58**: 401-16.
Chocair, P. R., J. A. Duley, et al. (1992). "The importance of thiopurine methyltransferase activity for the use of azathioprine in transplant recipients." Transplantation **53**(5): 1051-6.
Choy, E., R. Yelensky, et al. (2008). "Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines." PLoS Genet **4**(11): e1000287.
Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nat Rev Genet **11**(6): 415-25.
Cirulli, E. T., D. Kasperaviciute, et al. (2010). "Common genetic variation and performance on standardized cognitive tests." Eur J Hum Genet **18**(7): 815-20.
Clayton, D. G. (2009). "Prediction and interaction in complex disease genetics: experience in type 1 diabetes." PLoS Genet **5**(7): e1000540.
Clemente, M. G., S. De Virgiliis, et al. (2003). "Early effects of gliadin on enterocyte intracellular signalling involved in intestinal barrier function." Gut **52**(2): 218-23.
Coenen, M. J., G. Trynka, et al. (2009). "Common and different genetic background for rheumatoid arthritis and coeliac disease." Hum Mol Genet **18**(21): 4195-203.
Conrad, D. F., D. Pinto, et al. "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-12.
Conrad, D. F., D. Pinto, et al. (2010). "Origins and functional impact of copy number variation in the human genome." Nature **464**(7289): 704-12.
Consortium, T. H. (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-320.
Cookson, W., L. Liang, et al. (2009). "Mapping complex disease traits with global gene expression." Nat Rev Genet **10**(3): 184-94.
Cooney, R., J. Baker, et al. (2010). "NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation." Nat Med **16**(1): 90-7.
Coop, G., J. K. Pickrell, et al. (2009). "The role of geography in human adaptation." PLoS Genet **5**(6): e1000500.
Cooper, G. M., D. A. Nickerson, et al. (2007). "Mutational and selective effects on copy-number variants in the human genome." Nat Genet **39**(7 Suppl): S22-9.
Cosme, R., D. Lublin, et al. (2000). "Prostanoids in human colonic mucosa: effects of inflammation on PGE(2) receptor expression." Hum Immunol **61**(7): 684-96.
Cosnes, J., S. Cattan, et al. (2002). "Long-term evolution of disease behavior of Crohn's disease." Inflamm Bowel Dis **8**(4): 244-50.
Craddock, N., M. E. Hurles, et al. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-20.
Cucca, F., R. Lampis, et al. (2001). "A correlation between the relative predisposition of MHC class II alleles to type 1 diabetes and the structure of their proteins." Hum Mol Genet **10**(19): 2025-37.
Cuffari, C., T. Dassopoulos, et al. (2004). "Thiopurine methyltransferase activity influences clinical response to azathioprine in inflammatory bowel disease." Clin Gastroenterol Hepatol **2**(5): 410-7.

Cuthbert, A. P., S. A. Fisher, et al. (2002). "The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease." Gastroenterology **122**(4): 867-74.
D'Haens, G. R. (2009). "Top-down therapy for Crohn's disease: rationale and evidence." Acta Clin Belg **64**(6): 540-6.
Daly, A. K., P. T. Donaldson, et al. (2009). "HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin." Nat Genet **41**(7): 816-9.
Daly, A. K., P. T. Donaldson, et al. (2009). "HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin." Nat Genet.
Darfeuille-Michaud, A., C. Neut, et al. (1998). "Presence of adherent Escherichia coli strains in ileal mucosa of patients with Crohn's disease." Gastroenterology **115**(6): 1405-13.
de Bakker, P. I., M. A. Ferreira, et al. (2008). "Practical aspects of imputation-driven meta-analysis of genome-wide association studies." Hum Mol Genet **17**(R2): R122-8.
De Jager, P. L., X. Jia, et al. (2009). "Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci." Nat Genet.
de Jong, D. J., L. J. Derijks, et al. (2003). "Safety of thiopurines in the treatment of inflammatory bowel disease." Scand J Gastroenterol Suppl(239): 69-72.
de Smith, A. J., R. G. Walters, et al. (2008). "Small deletion variants have stable breakpoints commonly associated with alu elements." PLoS One **3**(8): e3104.
Dema, B., A. Martinez, et al. (2009). "Association of IL18RAP and CCR3 with coeliac disease in the Spanish population." J Med Genet **46**(9): 617-9.
Dendrou, C. A., V. Plagnol, et al. (2009). "Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource." Nat Genet **41**(9): 1011-5.
Deretic, V. (2010). "Autophagy in infection." Curr Opin Cell Biol **22**(2): 252-62.
Di Sabatino, A., R. Ciccocioppo, et al. (2006). "Epithelium derived interleukin 15 regulates intraepithelial lymphocyte Th1 cytokine production, cytotoxicity, and survival in coeliac disease." Gut **55**(4): 469-77.
Di Sabatino, A., K. M. Pickard, et al. (2007). "Evidence for the role of interferon-alfa production by dendritic cells in the Th1 response in celiac disease." Gastroenterology **133**(4): 1175-87.
Dickson, S. P., K. Wang, et al. (2010). "Rare variants create synthetic genome-wide associations." PLoS Biol **8**(1): e1000294.
Diebold, S. S., T. Kaisho, et al. (2004). "Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA." Science **303**(5663): 1529-31.
Dieterich, W., T. Ehnis, et al. (1997). "Identification of tissue transglutaminase as the autoantigen of celiac disease." Nat Med **3**(7): 797-801.
Dignass, A. V. A., G. Lindsay, J.O. Lémann, J. Söderholm,, S. D. J.F. Colombel, A. D'Hoore, M. Gassull, F. Gomollón, D.W. Hommes,, et al. (2010). "The second European evidence-based consensus on the diagnosis and management of Crohn's disease: Current management." Journal of Crohn's and Colitis **4**: 28-62.
Djilali-Saiah, I., J. Schmitz, et al. (1998). "CTLA-4 gene polymorphism is associated with predisposition to coeliac disease." Gut **43**(2): 187-9.
Dubinsky, M. C., E. J. Feldman, et al. (2003). "Thioguanine: a potential alternate thiopurine for IBD patients allergic to 6-mercaptopurine or azathioprine." Am J Gastroenterol **98**(5): 1058-63.
Dubinsky, M. C., E. A. Vasiliauskas, et al. (2003). "6-thioguanine can cause serious liver injury in inflammatory bowel disease patients." Gastroenterology **125**(2): 298-303.
Dubois, P., K. Hunt, et al. (2009). "Sex differences in HLA DQ in celiac disease." Am J Gastroenterol **104**(3): 784; author reply 784-5.
Dubois, P. C., G. Trynka, et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression." Nat Genet **42**(4): 295-302.

Dubois, P. C. and D. A. van Heel (2008). "Translational mini-review series on the immunogenetics of gut disease: immunogenetics of coeliac disease." Clin Exp Immunol **153**(2): 162-73.
Dudbridge, F. and A. Gusnanto (2008). "Estimation of significance thresholds for genomewide association scans." Genet Epidemiol **32**(3): 227-34.
Duerr, R. H., K. D. Taylor, et al. (2006). "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene." Science **314**(5804): 1461-3.
Economou, M., T. A. Trikalinos, et al. (2004). "Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis." Am J Gastroenterol **99**(12): 2393-404.
Edwards, S. G., V. Hubbard, et al. (1999). "Concordance of primary generalised epilepsy and carbamazepine hypersensitivity in monozygotic twins." Postgrad Med J **75**(889): 680-1.
Elion, G. B. (1989). "The purine path to chemotherapy." Science **244**(4900): 41-7.
Elion, G. B., G. H. Hitchings, et al. (1951). "Antagonists of nucleic acid derivatives. VI. Purines." J Biol Chem **192**(2): 505-18.
EUORDIAB ACE Study Group (2000). "Variation and trends in incidence of childhood diabetes in Europe. EURODIAB ACE Study Group." Lancet **355**(9207): 873-6.
Falchuk, Z. M., G. N. Rogentine, et al. (1972). "Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy." J Clin Invest **51**(6): 1602-5.
Falconer, D. S. (1965). "The inheritance of liability to certain diseases estimated from the incidence in relatives." Annals of Human Genetics **29**: 51-76.
Fasano, A., I. Berti, 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**(3): 286-92.
Fasano, A. and C. Catassi (2001). "Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum." Gastroenterology **120**(3): 636-51.
Fedyk, E. R., J. M. Ripper, et al. (1996). "A molecular analysis of PGE receptor (EP) expression on normal and transformed B lymphocytes: coexpression of EP1, EP2, EP3beta and EP4." Mol Immunol **33**(1): 33-45.
Feil, R. and F. Berger (2007). "Convergent evolution of genomic imprinting in plants and mammals." Trends Genet **23**(4): 192-9.
Felderbauer, P., P. Hoffmann, et al. (2003). "A novel mutation of the calcium sensing receptor gene is associated with chronic pancreatitis in a family with heterozygous SPINK1 mutations." BMC Gastroenterol **3**: 34.
Felderbauer, P., E. Karakas, et al. (2008). "Pancreatitis risk in primary hyperparathyroidism: relation to mutations in the SPINK1 trypsin inhibitor (N34S) and the cystic fibrosis gene." Am J Gastroenterol **103**(2): 368-74.
Fellay, J., A. J. Thompson, et al. (2010). "ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C." Nature **464**(7287): 405-8.
Feller, M., K. Huwiler, et al. (2007). "Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis." Lancet Infect Dis **7**(9): 607-13.
Feuk, L., A. R. Carson, et al. (2006). "Structural variation in the human genome." Nat Rev Genet **7**(2): 85-97.
Fiebich, B. L., S. Schleicher, et al. (2001). "Mechanisms of prostaglandin E2-induced interleukin-6 release in astrocytes: possible involvement of EP4-like receptors, p38 mitogen-activated protein kinase and protein kinase C." J Neurochem **79**(5): 950-8.
Fina, D., M. Sarra, et al. (2007). "Interleukin-21 Contributes To The Mucosal T Helper Cell Type 1 Response In Celiac Disease." Gut.
Fish, E. N. (2008). "The X-files in immunity: sex-based differences predispose immune responses." Nat Rev Immunol **8**(9): 737-44.

Fisher, R. A. (1918). "The Correlation Between Relatives on the Supposition of Mendelian Inheritance." Philosophical Transactions of the Royal Society of Edinburgh **52**: 399-433.
Fleckenstein, B., O. Molberg, et al. (2002). "Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. Role of enzyme specificity and pH influence on the transamidation versus deamidation process." J Biol Chem **277**(37): 34109-16.
Floyd, A., L. Pedersen, et al. (2003). "Risk of acute pancreatitis in users of azathioprine: a population-based case-control study." Am J Gastroenterol **98**(6): 1305-8.
Fontenot, J. D., J. P. Rasmussen, et al. (2005). "A function for interleukin 2 in Foxp3-expressing regulatory T cells." Nat Immunol **6**(11): 1142-51.
Frank, D. N. (2008). "Mycobacterium avium subspecies paratuberculosis and Crohn's disease." Lancet Infect Dis **8**(6): 345; author reply 345-6.
Frank, D. N., A. L. St Amand, et al. (2007). "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases." Proc Natl Acad Sci U S A **104**(34): 13780-5.
Franke, A., T. Balschun, et al. (2010). "Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL)." Nat Genet **42**(4): 292-4.
Franke, L., C. G. de Kovel, et al. (2008). "Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays." Am J Hum Genet **82**(6): 1316-33.
Fraser, A. G., T. R. Orchard, et al. (2002). "The efficacy of azathioprine for the treatment of inflammatory bowel disease: a 30 year review." Gut **50**(4): 485-9.
Fraser, J. S., W. Engel, et al. (2003). "Coeliac disease: in vivo toxicity of the putative immunodominant epitope." Gut **52**(12): 1698-702.
Frazer, K. A., D. G. Ballinger, et al. (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-61.
Frazer, K. A., S. S. Murray, et al. (2009). "Human genetic variation and its contribution to complex traits." Nat Rev Genet **10**(4): 241-51.
Freeman, H. J. (2003). "Biopsy-defined adult celiac disease in Asian-Canadians." Can J Gastroenterol **17**(7): 433-6.
Fried, K. and E. Vure (1974). "A lethal autosomal recessive entero-colitis of early infancy." Clin Genet **6**(3): 195-6.
Fu, G., S. Vallee, et al. (2009). "Themis controls thymocyte selection through regulation of T cell antigen receptor-mediated signaling." Nat Immunol **10**(8): 848-56.
Fujino, H., W. Xu, et al. (2003). "Prostaglandin E2 induced functional expression of early growth response factor-1 by EP4, but not EP2, prostanoid receptors via the phosphatidylinositol 3-kinase and extracellular signal-regulated kinases." J Biol Chem **278**(14): 12151-6.
Ganesh, S. K., N. A. Zakai, et al. (2009). "Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium." Nat Genet **41**(11): 1191-8.
Garcia Rodriguez, L. A., A. Gonzalez-Perez, et al. (2005). "Risk factors for inflammatory bowel disease in the general population." Aliment Pharmacol Ther **22**(4): 309-15.
Garner, C. P., J. A. Murray, et al. (2009). "Replication of celiac disease UK genome-wide association study results in a US population." Hum Mol Genet **18**(21): 4219-25.
Ge, B., D. K. Pokholok, et al. (2009). "Global patterns of cis variation in human cells revealed by high-density allelic expression analysis." Nat Genet **41**(11): 1216-22.
Ge, D., J. Fellay, et al. (2009). "Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance." Nature **461**(7262): 399-401.
Gearry, R. B., R. L. Roberts, et al. (2004). "Lack of association between the ITPA 94C>A polymorphism and adverse effects from azathioprine." Pharmacogenetics **14**(11): 779-81.

Gennis, M. A., R. Vemuri, et al. (1991). "Familial occurrence of hypersensitivity to phenytoin." Am J Med **91**(6): 631-4.
Germani, G., M. Pleguezuelo, et al. (2009). "Azathioprine in liver transplantation: a reevaluation of its use and a comparison with mycophenolate mofetil." Am J Transplant **9**(8): 1725-31.
Gisbert, J. P., Y. Gonzalez-Lama, et al. (2007). "Thiopurine-induced liver injury in patients with inflammatory bowel disease: a systematic review." Am J Gastroenterol **102**(7): 1518-27.
Gjertsen, H. A., L. M. Sollid, et al. (1994). "T cells from the peripheral blood of coeliac disease patients recognize gluten antigens when presented by HLA-DR, -DQ, or -DP molecules." Scand J Immunol **39**(6): 567-74.
Glocker, E. O., D. Kotlarz, et al. (2009). "Inflammatory bowel disease and mutations affecting the interleukin-10 receptor." N Engl J Med **361**(21): 2033-45.
Godet, P. G., G. R. May, et al. (1995). "Meta-analysis of the role of oral contraceptive agents in inflammatory bowel disease." Gut **37**(5): 668-73.
Godkin, A., T. Friede, et al. (1997). "Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2)." Int Immunol **9**(6): 905-11.
Goerres, M. S., J. W. Meijer, et al. (2003). "Azathioprine and prednisone combination therapy in refractory coeliac disease." Aliment Pharmacol Ther **18**(5): 487-94.
Goldblatt, M. W. (1935). "Properties of human seminal plasma." J Physiol **84**(2): 208-18.
Gorry, M. C., D. Gabbaizedeh, et al. (1997). "Mutations in the cationic trypsinogen gene are associated with recurrent acute and chronic pancreatitis." Gastroenterology **113**(4): 1063-8.
Graham, R. R., C. Cotsapas, et al. (2008). "Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus." Nat Genet **40**(9): 1059-61.
Greco, L., M. C. Babron, et al. (2001). "Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families." Ann Hum Genet **65**(Pt 1): 35-41.
Greco, L., R. Romino, et al. (2002). "The first large population based twin study of coeliac disease." Gut **50**(5): 624-8.
Gregersen, P. K., C. I. Amos, et al. (2009). "REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis." Nat Genet **41**(7): 820-3.
Haber, C. J., S. J. Meltzer, et al. (1986). "Nature and course of pancreatitis caused by 6-mercaptopurine in the treatment of inflammatory bowel disease." Gastroenterology **91**(4): 982-6.
Hadjivassiliou, M., M. Maki, et al. (2006). "Autoantibody targeting of brain and intestinal transglutaminase in gluten ataxia." Neurology **66**(3): 373-7.
Halfvarson, J., L. Bodin, et al. (2003). "Inflammatory bowel disease in a Swedish twin cohort: a long-term follow-up of concordance and clinical characteristics." Gastroenterology **124**(7): 1767-73.
Hampe, J., A. Franke, et al. (2007). "A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1." Nat Genet **39**(2): 207-11.
Han, S. B., C. Moratz, et al. (2005). "Rgs1 and Gnai2 regulate the entrance of B lymphocytes into lymph nodes and B cell motility within lymph node follicles." Immunity **22**(3): 343-54.
Hanauer, S. B., B. G. Feagan, et al. (2002). "Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial." Lancet **359**(9317): 1541-9.
Harries, A. D., A. Baird, et al. (1982). "Non-smoking: a feature of ulcerative colitis." Br Med J (Clin Res Ed) **284**(6317): 706.

Hata, A. N. and R. M. Breyer (2004). "Pharmacology and signaling of prostaglandin receptors: multiple roles in inflammation and immune modulation." Pharmacol Ther **103**(2): 147-66.
Hausch, F., L. Shan, et al. (2002). "Intestinal digestive resistance of immunodominant gliadin peptides." Am J Physiol Gastrointest Liver Physiol **283**(4): G996-G1003.
Helgason, A., S. Palsson, et al. (2007). "Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution." Nat Genet **39**(2): 218-25.
Henckaerts, L., K. Van Steen, et al. (2009). "Genetic risk profiling and prediction of disease course in Crohn's disease patients." Clin Gastroenterol Hepatol **7**(9): 972-980 e2.
Hetherington, S., S. McGuirk, et al. (2001). "Hypersensitivity reactions during therapy with the nucleoside reverse transcriptase inhibitor abacavir." Clin Ther **23**(10): 1603-14.
Heurkens, A. H., M. L. Westedt, et al. (1991). "Prednisone plus azathioprine treatment in patients with rheumatoid arthritis complicated by vasculitis." Arch Intern Med **151**(11): 2249-54.
Hindorff, L. A., P. Sethupathy, et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-7.
Hindorff, L. A. J., H.A. Hall, P.N. Mehta, J.P. and Manolio, T.A (2010). "A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies Accessed 22/07/2010."
Hinks, A., P. Martin, et al. (2010). "Investigation of type 1 diabetes and coeliac disease susceptibility loci for association with juvenile idiopathic arthritis." Ann Rheum Dis.
Hirschfield, G. M., X. Liu, et al. (2009). "Primary biliary cirrhosis associated with HLA, IL12A, and IL12RB2 variants." N Engl J Med **360**(24): 2544-55.
Ho, I. C., N. K. Bhat, et al. (1990). "Sequence-specific binding of human Ets-1 to the T cell receptor alpha gene enhancer." Science **250**(4982): 814-8.
Hober, D. and P. Sauter (2010). "Pathogenesis of type 1 diabetes mellitus: interplay between enterovirus and host." Nat Rev Endocrinol **6**(5): 279-89.
Hong, S. W., S. Kim, et al. (2008). "The role of Bach2 in nucleic acid-triggered antiviral innate immune responses." Biochem Biophys Res Commun **365**(3): 426-32.
Howson, J. M., N. M. Walker, et al. (2009). "Confirmation of HLA class II independent type 1 diabetes associations in the major histocompatibility complex including HLA-B and HLA-A." Diabetes Obes Metab **11 Suppl 1**: 31-45.
Hue, S., P. Ahern, et al. (2006). "Interleukin-23 drives innate and T cell-mediated intestinal inflammation." J Exp Med **203**(11): 2473-83.
Hue, S., J. J. Mention, et al. (2004). "A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease." Immunity **21**(3): 367-77.
Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.
Hugot, J. P., I. Zaccaria, et al. (2007). "Prevalence of CARD15/NOD2 mutations in Caucasian healthy people." Am J Gastroenterol **102**(6): 1259-67.
Hunt, K. A., L. Franke, et al. (2007). "Large scale replication of a genome-wide association study in celiac disease." American Society of Human Genetics Meeting: platform talk 26.
Hunt, K. A., A. J. Monsuur, et al. (2006). "Lack of association of MYO9B genetic variants with coeliac disease in a British cohort." Gut **55**(7): 969-72.
Hunt, K. A., A. Zhernakova, et al. (2008). "Newly identified genetic risk variants for celiac disease related to the immune response." Nat Genet **40**(4): 395-402.
Ientile, R., D. Caccamo, et al. (2007). "Tissue transglutaminase and the stress response." Amino Acids **33**(2): 385-94.

Ikegami, R., Y. Sugimoto, et al. (2001). "The expression of prostaglandin E receptors EP2 and EP4 and their different regulation by lipopolysaccharide in C3H/HeN peritoneal macrophages." J Immunol **166**(7): 4689-96.
Intemann, C. D., T. Thye, et al. (2009). "Autophagy gene variant IRGM -261T contributes to protection from tuberculosis caused by Mycobacterium tuberculosis but not by M. africanum strains." PLoS Pathog **5**(9): e1000577.
Ivarsson, A. (2005). "The Swedish epidemic of coeliac disease explored using an epidemiological approach--some lessons to be learnt." Best Pract Res Clin Gastroenterol **19**(3): 425-40.
Ivarsson, A., O. Hernell, et al. (2003). "Children born in the summer have increased risk for coeliac disease." J Epidemiol Community Health **57**(1): 36-9.
Ivarsson, A., O. Hernell, et al. (2002). "Breast-feeding protects against celiac disease." Am J Clin Nutr **75**(5): 914-21.
Jabri, B., N. P. de Serre, et al. (2000). "Selective expansion of intraepithelial lymphocytes expressing the HLA-E-specific natural killer receptor CD94 in celiac disease." Gastroenterology **118**(5): 867-79.
Jakkula, E., K. Rehnstrom, et al. (2008). "The genome-wide patterns of variation expose significant substructure in a founder population." Am J Hum Genet **83**(6): 787-94.
Janssens, A. C., R. Moonesinghe, et al. (2007). "The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases." Genet Med **9**(8): 528-35.
Janssens, A. C. and C. M. van Duijn (2008). "Genome-based prediction of common diseases: advances and prospects." Hum Mol Genet **17**(R2): R166-73.
Janssens, A. C. and C. M. van Duijn (2009). "Genome-based prediction of common diseases: methodological considerations for future research." Genome Med **1**(2): 20.
Ji, W., J. N. Foo, et al. (2008). "Rare independent mutations in renal salt handling genes contribute to blood pressure variation." Nat Genet **40**(5): 592-9.
Jiang, G. L., A. Nieves, et al. (2007). "The prevention of colitis by E Prostanoid receptor 4 agonist through enhancement of epithelium survival and regeneration." J Pharmacol Exp Ther **320**(1): 22-8.
Jin, Y., S. A. Birlea, et al. (2010). "Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo." N Engl J Med **362**(18): 1686-97.
Jo, Y., T. Matsumoto, et al. (2003). "CCR4 is an up-regulated chemokine receptor of peripheral blood memory CD4+ T cells in Crohn's disease." Clin Exp Immunol **132**(2): 332-8.
Johansen, C. T., J. Wang, et al. (2010). "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia." Nat Genet **42**(8): 684-7.
Johnson-Reagan, L. and S. L. Bahna (2003). "Severe drug rashes in three siblings simultaneously." Allergy **58**(5): 445-7.
Johnson, A. L., L. Aravind, et al. (2009). "Themis is a member of a new metazoan gene family and is required for the completion of thymocyte positive selection." Nat Immunol **10**(8): 831-9.
Jones, J. L., C. L. Phuah, et al. (2009). "IL-21 drives secondary autoimmunity in patients with multiple sclerosis, following therapeutic lymphocyte depletion with alemtuzumab (Campath-1H)." J Clin Invest **119**(7): 2052-61.
Joossens, S., S. Vermeire, et al. (2004). "Pancreatic autoantibodies in inflammatory bowel disease." Inflamm Bowel Dis **10**(6): 771-7.
Kabashima, K., T. Saji, et al. (2002). "The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut." J. Clin. Invest. **109**(7): 883-893.
Kandiel, A., A. G. Fraser, et al. (2005). "Increased risk of lymphoma among inflammatory bowel disease patients treated with azathioprine and 6-mercaptopurine." Gut **54**(8): 1121-5.

Karell, K., A. S. Louka, et al. (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**(4): 469-77.
Karouzakis, E., M. Neidhart, et al. (2006). "Molecular and cellular basis of rheumatoid joint destruction." Immunol Lett **106**(1): 8-13.
Kathiresan, S., B. F. Voight, et al. (2009). "Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants." Nat Genet **41**(3): 334-41.
Keuning, J. J., A. S. Pena, et al. (1976). "HLA-DW3 associated with coeliac disease." Lancet **1**(7958): 506-8.
Kim, C. Y., H. Quarsten, et al. (2004). "Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease." Proc Natl Acad Sci U S A **101**(12): 4175-9.
Kirschner, B. S. (1998). "Safety of azathioprine and 6-mercaptopurine in pediatric patients with inflammatory bowel disease." Gastroenterology **115**(4): 813-21.
Klein, L., M. Hinterberger, et al. (2009). "Antigen presentation in the thymus for positive selection and central tolerance induction." Nat Rev Immunol **9**(12): 833-844.
Klein, R. J., C. Zeiss, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-9.
Konig, J., A. Seithel, et al. (2006). "Pharmacogenomics of human OATP transporters." Naunyn Schmiedebergs Arch Pharmacol **372**(6): 432-43.
Korbel, J. O., A. E. Urban, et al. (2007). "Paired-end mapping reveals extensive structural variation in the human genome." Science **318**(5849): 420-6.
Korponay-Szabo, I. R., T. Halttunen, et al. (2004). "In vivo targeting of intestinal and extraintestinal transglutaminase 2 by coeliac autoantibodies." Gut **53**(5): 641-8.
Koskinen, L. L., E. Einarsdottir, et al. (2009). "Association study of the IL18RAP locus in three European populations with coeliac disease." Hum Mol Genet **18**(6): 1148-55.
Koskinen, L. L., E. Einarsdottir, et al. (2009). "Fine mapping of the CELIAC2 locus on chromosome 5q31-q33 in the Finnish and Hungarian populations." Tissue Antigens **74**(5): 408-16.
Koutroubakis, I. E., D. Drygiannakis, et al. (2005). "Pancreatic autoantibodies in Greek patients with inflammatory bowel disease." Dig Dis Sci **50**(12): 2330-4.
Koutroubakis, I. E., I. G. Vlachonikolis, et al. (2002). "Role of appendicitis and appendectomy in the pathogenesis of ulcerative colitis: a critical review." Inflamm Bowel Dis **8**(4): 277-86.
Kraft, P., S. Wacholder, et al. (2009). "Beyond odds ratios--communicating disease risk based on genetic profiles." Nat Rev Genet **10**(4): 264-9.
Krieg, A. M. (2002). "A role for Toll in autoimmunity." Nat Immunol **3**(5): 423-4.
Krieg, A. M. and J. Vollmer (2007). "Toll-like receptors 7, 8, and 9: linking innate immunity to autoimmunity." Immunol Rev **220**: 251-69.
Krueger, G. G., R. G. Langley, et al. (2007). "A human interleukin-12/23 monoclonal antibody for the treatment of psoriasis." N Engl J Med **356**(6): 580-92.
Kuballa, P., A. Huett, et al. (2008). "Impaired autophagy of an intracellular pathogen induced by a Crohn's disease associated ATG16L1 variant." PLoS ONE **3**(10): e3391.
Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-9.
Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
Leadbetter, E. A., I. R. Rifkin, et al. (2002). "Chromatin-IgG complexes activate B cells by dual engagement of IgM and Toll-like receptors." Nature **416**(6881): 603-7.
Lee, E. G., D. L. Boone, et al. (2000). "Failure to regulate TNF-induced NF-kappaB and cell death responses in A20-deficient mice." Science **289**(5488): 2350-4.

Lemann, M., J. Y. Mary, et al. (2005). "A randomized, double-blind, controlled withdrawal trial in Crohn's disease patients in long-term remission on azathioprine." Gastroenterology **128**(7): 1812-8.
Lemmon, M. A. (2008). "Membrane recognition by phospholipid-binding domains." Nat Rev Mol Cell Biol **9**(2): 99-111.
Lenardo, M. J. (1996). "Fas and the art of lymphocyte maintenance." J Exp Med **183**(3): 721-4.
Lennard, L. (1992). "The clinical pharmacology of 6-mercaptopurine." Eur J Clin Pharmacol **43**(4): 329-39.
Lennard, L., J. S. Lilleyman, et al. (1990). "Genetic variation in response to 6-mercaptopurine for childhood acute lymphoblastic leukaemia." Lancet **336**(8709): 225-9.
Leon, A. J., J. A. Garrote, et al. (2006). "Interleukin 18 maintains a long-standing inflammation in coeliac disease patients." Clin Exp Immunol **146**(3): 479-85.
Leonard, W. J. and R. Spolski (2005). "Interleukin-21: a modulator of lymphoid proliferation, apoptosis and differentiation." Nat Rev Immunol **5**(9): 688-98.
Lesourne, R., S. Uehara, et al. (2009). "Themis, a T cell-specific protein important for late thymocyte development." Nat Immunol **10**(8): 840-7.
Levy, S. and R. L. Strausberg (2008). "Human genetics: Individual genomes diversify." Nature **456**(7218): 49-51.
Levy, S., G. Sutton, et al. (2007). "The diploid genome sequence of an individual human." PLoS Biol **5**(10): e254.
Lewis, C. M., S. C. Whitwell, et al. (2007). "Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease." J Med Genet **44**(11): 689-94.
Li, Y., X. He, et al. (2000). "Cloning and characterization of human Lnk, an adaptor protein with pleckstrin homology and Src homology 2 domains that can inhibit T cell activation." J Immunol **164**(10): 5199-206.
Libioulle, C., E. Louis, et al. (2007). "Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4." PLoS Genet **3**(4): e58.
Lichtenstein, G. R., S. Yan, et al. (2005). "Infliximab maintenance treatment reduces hospitalizations, surgeries, and procedures in fistulizing Crohn's disease." Gastroenterology **128**(4): 862-9.
Lindvall, J. M., K. E. Blomberg, et al. (2005). "Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling." Immunol Rev **203**: 200-15.
Link, E., S. Parish, et al. (2008). "SLCO1B1 variants and statin-induced myopathy--a genomewide study." N Engl J Med **359**(8): 789-99.
Liu, H., F. Prugnolle, et al. (2006). "A geographically explicit genetic model of worldwide human-settlement history." Am J Hum Genet **79**(2): 230-7.
Liu, Q. C., F. Gao, et al. (2008). "Multisite mutations of the PRSS1 gene in a Chinese patient with chronic pancreatitis." Hepatobiliary Pancreat Dis Int **7**(3): 331-2.
Liu, Y., C. Helms, et al. (2008). "A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease Loci." PLoS Genet **4**(3): e1000041.
Loftus, E. V., Jr. (2004). "Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences." Gastroenterology **126**(6): 1504-17.
Loftus, E. V., Jr., M. D. Silverstein, et al. (1998). "Crohn's disease in Olmsted County, Minnesota, 1940-1993: incidence, prevalence, and survival." Gastroenterology **114**(6): 1161-8.
Lohi, S., K. Mustalahti, et al. (2007). "Increasing prevalence of coeliac disease over time." Aliment Pharmacol Ther **26**(9): 1217-25.

Louka, A. S., S. J. Moodie, et al. (2003). "A collaborative European search for non-DQA1*05-DQB1*02 celiac disease loci on HLA-DR3 haplotypes: analysis of transmission from homozygous parents." Hum Immunol **64**(3): 350-8.
Louka, A. S., S. Nilsson, et al. (2002). "HLA in coeliac disease families: a novel test of risk modification by the 'other' haplotype when at least one DQA1*05-DQB1*02 haplotype is carried." Tissue Antigens **60**(2): 147-54.
Louka, A. S. and L. M. Sollid (2003). "HLA in coeliac disease: unravelling the complex genetics of a complex disorder." Tissue Antigens **61**(2): 105-17.
Lundin, K. E., H. Scott, et al. (1994). "T cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8." Hum Immunol **41**(4): 285-91.
Lundin, K. E., H. Scott, et al. (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**(1): 187-96.
Ma, W. and R. Quirion (2005). "Up-regulation of interleukin-6 induced by prostaglandin E from invading macrophages following nerve injury: an in vivo and in vitro study." J Neurochem **93**(3): 664-73.
Maeda, S., L. C. Hsu, et al. (2005). "Nod2 mutation in Crohn's disease potentiates NF-kappaB activity and IL-1beta processing." Science **307**(5710): 734-8.
Maiuri, L., C. Ciacci, et al. (2003). "Association between innate response to gliadin and activation of pathogenic T cells in coeliac disease." Lancet **362**(9377): 30-7.
Maiuri, L., A. Picarelli, et al. (1996). "Definition of the initial immunologic modifications upon in vitro gliadin challenge in the small intestine of celiac patients." Gastroenterology **110**(5): 1368-78.
Mallal, S., D. Nolan, et al. (2002). "Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir." Lancet **359**(9308): 727-32.
Mallal, S., E. Phillips, et al. (2008). "HLA-B*5701 screening for hypersensitivity to abacavir." N Engl J Med **358**(6): 568-79.
Maloy, K. J. and F. Powrie (2005). "Fueling regulation: IL-2 keeps CD4+ Treg cells fit." Nat Immunol **6**(11): 1071-2.
Manichanh, C., L. Rigottier-Gois, et al. (2006). "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach." Gut **55**(2): 205-11.
Manolio, T. A., L. D. Brooks, et al. (2008). "A HapMap harvest of insights into the genetics of common disease." J Clin Invest **118**(5): 1590-605.
Mao, M., M. C. Biery, et al. (2004). "T lymphocyte activation gene identification by coregulated expression on DNA microarrays." Genomics **83**(6): 989-99.
Margaritte-Jeannin, P., M. C. Babron, et al. (2004). "HLA-DQ relative risks for coeliac disease in European populations: a study of the European Genetics Cluster on Coeliac Disease." Tissue Antigens **63**(6): 562-7.
Marinaki, A. M., A. Ansari, et al. (2004). "Adverse drug reactions to azathioprine therapy are associated with polymorphism in the gene encoding inosine triphosphate pyrophosphatase (ITPase)." Pharmacogenetics **14**(3): 181-7.
Marinaki, A. M., J. A. Duley, et al. (2004). "Mutation in the ITPA gene predicts intolerance to azathioprine." Nucleosides Nucleotides Nucleic Acids **23**(8-9): 1393-7.
Markowitz, J., K. Grancher, et al. (2000). "A multicenter trial of 6-mercaptopurine and prednisone in children with newly diagnosed Crohn's disease." Gastroenterology **119**(4): 895-902.
Marks, D. J., M. W. Harbord, et al. (2006). "Defective acute inflammation in Crohn's disease: a clinical investigation." Lancet **367**(9511): 668-78.

Marks, D. J., K. Miyagi, et al. (2009). "Inflammatory bowel disease in CGD reproduces the clinicopathological features of Crohn's disease." Am J Gastroenterol **104**(1): 117-24.
Martin, H. M., B. J. Campbell, et al. (2004). "Enhanced Escherichia coli adherence and invasion in Crohn's disease and colon cancer." Gastroenterology **127**(1): 80-93.
Masson, E., J. M. Chen, et al. (2008). "Association of rare chymotrypsinogen C (CTRC) gene variations in patients with idiopathic chronic pancreatitis." Hum Genet **123**(1): 83-91.
Mathew, C. G. (2007). "New links to the pathogenesis of Crohn disease provided by genome-wide association scans." Nat Rev Genet.
Matysiak-Budnik, T., G. Malamut, et al. (2007). "Long-term follow-up of 61 coeliac patients diagnosed in childhood: evolution toward latency is possible on a normal diet." Gut **56**(10): 1379-86.
Maurino, E., S. Niveloni, et al. (2002). "Azathioprine in refractory sprue: results from a prospective, open-label study." Am J Gastroenterol **97**(10): 2595-602.
McCarroll, S. A. (2008). "Extending genome-wide association studies to copy-number variation." Hum Mol Genet **17**(R2): R135-42.
McCarroll, S. A., A. Huett, et al. (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." Nat Genet.
McCarthy, M. I., G. R. Abecasis, et al. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet **9**(5): 356-69.
McCarthy, M. I. and J. N. Hirschhorn (2008). "Genome-wide association studies: potential next steps on a genetic journey." Hum Mol Genet **17**(R2): R156-65.
McCoy, J. M., J. R. Wicks, et al. (2002). "The role of prostaglandin E2 receptors in the pathogenesis of rheumatoid arthritis." J Clin Invest **110**(5): 651-8.
McEvoy, B. P., G. W. Montgomery, et al. (2009). "Geographical structure and differential natural selection among North European populations." Genome Res **19**(5): 804-14.
McGovern, D. P., A. Gardet, et al. (2010). "Genome-wide association identifies multiple ulcerative colitis susceptibility loci." Nat Genet **42**(4): 332-7.
McIlroy, A., G. Caron, et al. (2006). "Histamine and prostaglandin E up-regulate the production of Th2-attracting chemokines (CCL17 and CCL22) and down-regulate IFN-gamma-induced CXCL10 production by immature human dendritic cells." Immunology **117**(4): 507-16.
Megarbane, A. and R. Sayad (2007). "Early lethal autosomal recessive enterocolitis: report of a second family." Clin Genet **71**(1): 89-90.
Meggitt, S. J., J. C. Gray, et al. (2006). "Azathioprine dosed by thiopurine methyltransferase activity for moderate-to-severe atopic eczema: a double-blind, randomised controlled trial." Lancet **367**(9513): 839-46.
Megiorni, F., B. Mora, et al. (2008). "HLA-DQ and susceptibility to celiac disease: evidence for gender differences and parent-of-origin effects." Am J Gastroenterol **103**(4): 997-1003.
Meja, K. K., P. J. Barnes, et al. (1997). "Characterization of the prostanoid receptor(s) on human blood monocytes at which prostaglandin E2 inhibits lipopolysaccharide-induced tumour necrosis factor-alpha generation." Br J Pharmacol **122**(1): 149-57.
Menashe, I., P. S. Rosenberg, et al. (2008). "PGA: power calculator for case-control genetic association analyses." BMC Genet **9**: 36.
Mention, J. J., M. Ben Ahmed, et al. (2003). "Interleukin 15: a key to disrupted intraepithelial lymphocyte homeostasis and lymphomagenesis in celiac disease." Gastroenterology **125**(3): 730-45.
Meresse, B., S. A. Curran, et al. (2006). "Reprogramming of CTLs into natural killer-like cells in celiac disease." J Exp Med **203**(5): 1343-55.
Mitt, K. and O. Uibo (1998). "Low cereal intake in Estonian infants: the possible explanation for the low frequency of coeliac disease in Estonia." Eur J Clin Nutr **52**(2): 85-8.

Molberg, O., S. N. McAdam, et al. (1998). "Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease." Nat Med **4**(6): 713-7.
Monsuur, A. J., P. I. de Bakker, et al. (2005). "Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect." Nat Genet **37**(12): 1341-4.
Monsuur, A. J., P. I. de Bakker, et al. (2008). "Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms." PLoS ONE **3**(5): e2270.
Monteleone, G., S. L. Pender, et al. (2001). "Role of interferon alpha in promoting T helper cell type 1 responses in the small intestine in coeliac disease." Gut **48**(3): 425-9.
Monteleone, I., G. Monteleone, et al. (2004). "Regulation of the T helper cell type 1 transcription factor T-bet in coeliac disease mucosa." Gut **53**(8): 1090-5.
Mori, K., I. Tanaka, et al. (1996). "Gene expression of the human prostaglandin E receptor EP4 subtype: differential regulation in monocytoid and lymphoid lineage cells by phorbol ester." J Mol Med **74**(6): 333-6.
Muddana, V., J. Lamb, et al. (2008). "Association between calcium sensing receptor gene polymorphisms and chronic pancreatitis in a US population: role of serine protease inhibitor Kazal 1type and alcohol." World J Gastroenterol **14**(28): 4486-91.
Murphy, K. M., C. A. Nelson, et al. (2006). "Balancing co-stimulation and inhibition with BTLA and HVEM." Nat Rev Immunol **6**(9): 671-81.
Murray, J. E., J. P. Merrill, et al. (1963). "Prolonged survival of human-kidney homografts by immunosuppressive drug therapy." N Engl J Med **268**: 1315-23.
Myers, S., L. Bottolo, et al. (2005). "A fine-scale map of recombination rates and hotspots across the human genome." Science **310**(5746): 321-4.
Myers, S., R. Bowden, et al. (2010). "Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination." Science **327**(5967): 876-9.
Myrsky, E., K. Kaukinen, et al. (2008). "Coeliac disease-specific autoantibodies targeted against transglutaminase 2 disturb angiogenesis." Clin Exp Immunol **152**(1): 111-9.
Nachman, M. W. and S. L. Crowell (2000). "Estimate of the mutation rate per nucleotide in humans." Genetics **156**(1): 297-304.
Nair, R. P., K. C. Duffin, et al. (2009). "Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways." Nat Genet **41**(2): 199-204.
Nakagawa, I., A. Amano, et al. (2004). "Autophagy defends cells against invading group A Streptococcus." Science **306**(5698): 1037-40.
Nakamura, Y., M. Leppert, et al. (1987). "Variable number of tandem repeat (VNTR) markers for human gene mapping." Science **235**(4796): 1616-22.
Narumiya, S., Y. Sugimoto, et al. (1999). "Prostanoid receptors: structures, properties, and functions." Physiol Rev **79**(4): 1193-226.
Nedjic, J., M. Aichinger, et al. (2008). "Autophagy in thymic epithelium shapes the T-cell repertoire and is essential for tolerance." Nature **455**(7211): 396-400.
Nejentsev, S., N. Walker, et al. (2009). "Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes." Science **324**(5925): 387-9.
Nelson, M. R., S. A. Bacanu, et al. (2009). "Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions." Pharmacogenomics J **9**(1): 23-33.
Newton-Cheh, C., T. Johnson, et al. (2009). "Genome-wide association study identifies eight loci associated with blood pressure." Nat Genet.
Ng, S. B., A. W. Bigham, et al. (2010). "Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome." Nat Genet.

Ng, S. B., K. J. Buckingham, et al. (2010). "Exome sequencing identifies the cause of a mendelian disorder." Nat Genet **42**(1): 30-5.
Nica, A. C., S. B. Montgomery, et al. (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations." PLoS Genet **6**(4): e1000895.
Nilsen, E. M., F. L. Jahnsen, et al. (1998). "Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease." Gastroenterology **115**(3): 551-63.
Nistico, L., C. Fagnani, et al. (2006). "Concordance, disease progression, and heritability of coeliac disease in Italian twins." Gut **55**(6): 803-8.
Ogura, Y., D. K. Bonen, et al. (2001). "A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease." Nature **411**(6837): 603-6.
Okano, M., Y. Sugata, et al. (2006). "E prostanoid 2 (EP2)/EP4-mediated suppression of antigen-specific human T-cell responses by prostaglandin E2." Immunology **118**(3): 343-52.
Olsen, I., S. Tollefsen, et al. (2009). "Isolation of Mycobacterium avium subspecies paratuberculosis reactive CD4 T cells from intestinal biopsies of Crohn's disease patients." PLoS ONE **4**(5): e5641.
Orholm, M., V. Binder, et al. (2000). "Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study." Scand J Gastroenterol **35**(10): 1075-81.
Ounissi-Benkalha, H. and C. Polychronakos (2008). "The molecular genetics of type 1 diabetes: new genes and emerging mechanisms." Trends Mol Med **14**(6): 268-75.
Pablos, J. L., B. Santiago, et al. (1999). "Cyclooxygenase-1 and -2 are expressed by human T cells." Clin Exp Immunol **115**(1): 86-90.
Paisan-Ruiz, C., S. Jain, et al. (2004). "Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease." Neuron **44**(4): 595-600.
Pang, A. W., J. R. MacDonald, et al. (2010). "Towards a comprehensive structural variation map of an individual human genome." Genome Biol **11**(5): R52.
Park, J. H., S. Wacholder, et al. (2010). "Estimation of effect size distribution from genome-wide association studies and implications for future discoveries." Nat Genet **42**(7): 570-5.
Parkes, M., J. C. Barrett, et al. (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility." Nat Genet **39**(7): 830-2.
Parkes, M. M., D. Franke, A. Vermiere, S. Louis, E. Ahmad, T. Mathew, T. Annese, V. Rioux, J. Rotter, J. Dubinsky, M. Taylor, K. Kugathasan, S. Brant, S. Duerr, R. Griffiths, A. Schreiber, S. Silverberg, M. Hakornarson, H., Satsangi, J. Daly, M. Cho, J. (2010). "New Crohn's disease susceptibility genes and loci identified by the International IBD Genetics Consortium." Gastroenterology **138**(5 (s1)): S-115.
Pastinen, T. (2010). "Genome-wide allele-specific analysis: insights into regulatory variation." Nat Rev Genet.
Patrick, M. S., H. Oda, et al. (2009). "Gasp, a Grb2-associating protein, is critical for positive selection of thymocytes." Proc Natl Acad Sci U S A **106**(38): 16345-50.
Pe'er, I., R. Yelensky, et al. (2008). "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants." Genet Epidemiol **32**(4): 381-5.
Pearson, T. A. and T. A. Manolio (2008). "How to interpret a genome-wide association study." Jama **299**(11): 1335-44.
Peeters, M., H. Nevens, et al. (1996). "Familial aggregation in Crohn's disease: increased age-adjusted risk and concordance in clinical characteristics." Gastroenterology **111**(3): 597-603.

Pellicano, R., A. Silvestris, et al. (1992). "Familial occurrence of fixed drug eruptions." Acta Derm Venereol **72**(4): 292-3.
Perdigones, N., E. Martin, et al. (2010). "Study of chromosomal region 5p13.1 in Crohn's disease, ulcerative colitis, and rheumatoid arthritis." Hum Immunol.
Pharoah, P. D., A. Antoniou, et al. (2002). "Polygenic susceptibility to breast cancer and implications for prevention." Nat Genet **31**(1): 33-6.
Pirmohamed, M. (2010). "Pharmacogenetics of idiosyncratic adverse drug reactions." Handb Exp Pharmacol(196): 477-91.
Pitchumoni, C. S., A. Rubin, et al. (2010). "Pancreatitis in inflammatory bowel diseases." J Clin Gastroenterol **44**(4): 246-53.
Plagnol, V., D. J. Smyth, et al. (2009). "Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13." Biostatistics **10**(2): 327-34.
Plenge, R. M., C. Cotsapas, et al. (2007). "Two independent alleles at 6q23 associated with risk of rheumatoid arthritis." Nat Genet **39**(12): 1477-82.
Plomin, R., C. M. Haworth, et al. (2009). "Common disorders are quantitative traits." Nat Rev Genet **10**(12): 872-8.
Ponticelli, C., A. Tarantino, et al. (1999). "Renal transplantation, past, present and future." J Nephrol **12 Suppl 2**: S105-10.
Poppe, D., I. Tiede, et al. (2006). "Azathioprine suppresses ezrin-radixin-moesin-dependent T cell-APC conjugation through inhibition of Vav guanosine exchange activity on Rac proteins." J Immunol **176**(1): 640-51.
Portanova, J. P., Y. Zhang, et al. (1996). "Selective neutralization of prostaglandin E2 blocks inflammation, hyperalgesia, and interleukin 6 production in vivo." J Exp Med **184**(3): 883-91.
Pratt, D. S., D. P. Flavin, et al. (1996). "The successful treatment of autoimmune hepatitis with 6-mercaptopurine after failure with azathioprine." Gastroenterology **110**(1): 271-4.
Prefontaine, E., L. R. Sutherland, et al. (2009). "Azathioprine or 6-mercaptopurine for maintenance of remission in Crohn's disease." Cochrane Database Syst Rev(1): CD000067.
Present, D. H., B. I. Korelitz, et al. (1980). "Treatment of Crohn's disease with 6-mercaptopurine. A long-term, randomized, double-blind study." N Engl J Med **302**(18): 981-7.
Present, D. H., S. J. Meltzer, et al. (1989). "6-Mercaptopurine in the management of inflammatory bowel disease: short- and long-term toxicity." Ann Intern Med **111**(8): 641-9.
Price, A. L., G. V. Kryukov, et al. (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-8.
Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet **38**(8): 904-9.
Price, P., C. Witt, et al. (1999). "The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases." Immunol Rev **167**: 257-74.
Purcell, S., S. S. Cherny, et al. (2003). "Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits." Bioinformatics **19**(1): 149-50.
Purcell, S., B. Neale, et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." Am J Hum Genet **81**(3): 559-75.
Purcell, S. M., N. R. Wray, et al. (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." Nature **460**(7256): 748-52.
Radford-Smith, G. L. (2008). "What is the importance of appendectomy in the natural history of IBD?" Inflamm Bowel Dis **14 Suppl 2**: S72-4.

Rahman, F. Z., D. J. Marks, et al. (2008). "Phagocyte dysfunction and inflammatory bowel disease." Inflamm Bowel Dis **14**(10): 1443-52.
Raki, M., S. Tollefsen, et al. (2006). "A unique dendritic cell subset accumulates in the celiac lesion and efficiently activates gluten-reactive T cells." Gastroenterology **131**(2): 428-38.
Rasmussen, H. H., K. Fonager, et al. (1999). "Risk of acute pancreatitis in patients with chronic inflammatory bowel disease. A Danish 16-year nationwide follow-up study." Scand J Gastroenterol **34**(2): 199-201.
Raychaudhuri, S., R. M. Plenge, et al. (2009). "Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions." PLoS Genet **5**(6): e1000534.
Raychaudhuri, S., E. F. Remmers, et al. (2008). "Common variants at CD40 and other loci confer risk of rheumatoid arthritis." Nat Genet **40**(10): 1216-23.
Raychaudhuri, S., B. P. Thomson, et al. (2009). "Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk." Nat Genet.
Rebours, V., M. C. Boutron-Ruault, et al. (2009). "The natural history of hereditary pancreatitis: a national series." Gut **58**(1): 97-103.
Redon, R., S. Ishikawa, et al. (2006). "Global variation in copy number in the human genome." Nature **444**(7118): 444-54.
Reveille, J. D., A. M. Sims, et al. (2010). "Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci." Nat Genet **42**(2): 123-7.
Rioux, J. D., R. J. Xavier, et al. (2007). "Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis." Nat Genet **39**(5): 596-604.
Roach, J. C., G. Glusman, et al. (2010). "Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing." Science.
Robinson, J., M. J. Waller, et al. (2003). "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex." Nucleic Acids Res **31**(1): 311-4.
Romanos, J., D. Barisani, et al. (2009). "Six new coeliac disease loci replicated in an Italian population confirm association with coeliac disease." J Med Genet **46**(1): 60-3.
Rosenberg, N. A., L. Huang, et al. (2010). "Genome-wide association studies in diverse populations." Nat Rev Genet **11**(5): 356-66.
Rosendahl, J., H. Witt, et al. (2008). "Chymotrypsin C (CTRC) variants that diminish activity or secretion are associated with chronic pancreatitis." Nat Genet **40**(1): 78-82.
Rothman, K. J. and S. Greenland (2005). "Causation and causal inference in epidemiology." Am J Public Health **95 Suppl 1**: S144-50.
Rubio-Tapia, A. and J. A. Murray (2010). "Classification and management of refractory coeliac disease." Gut **59**(4): 547-57.
Rubio-Tapia, A., C. T. Van Dyke, et al. (2008). "Predictors of family risk for celiac disease: a population-based study." Clin Gastroenterol Hepatol **6**(9): 983-7.
Russmann, S., J. A. Kaye, et al. (2005). "Risk of cholestatic liver disease associated with flucloxacillin and flucloxacillin prescribing habits in the UK: cohort study using data from the UK General Practice Research Database." Br J Clin Pharmacol **60**(1): 76-82.
Sahasranaman, S., D. Howard, et al. (2008). "Clinical pharmacology and pharmacogenetics of thiopurines." Eur J Clin Pharmacol **64**(8): 753-67.
Saitoh, T., N. Fujita, et al. (2008). "Loss of the autophagy protein Atg16L1 enhances endotoxin-induced IL-1beta production." Nature **456**(7219): 264-8.
Salvati, V. M., T. T. MacDonald, et al. (2002). "Interleukin 18 and associated markers of T helper cell type 1 activity in coeliac disease." Gut **50**(2): 186-90.
Sanchez, E., B. Z. Alizadeh, et al. (2007). "MYO9B gene polymorphisms are associated with autoimmune diseases in Spanish population." Hum Immunol **68**(7): 610-5.

Sandberg-Bennich, S., G. Dahlquist, et al. (2002). "Coeliac disease is associated with intrauterine growth and neonatal infections." Acta Paediatr **91**(1): 30-3.
Sandborn, W., L. Sutherland, et al. (2000). "Azathioprine or 6-mercaptopurine for inducing remission of Crohn's disease." Cochrane Database Syst Rev(2): CD000545.
Santhosh, S., H. Witt, et al. (2008). "A loss of function polymorphism (G191R) of anionic trypsinogen (PRSS2) confers protection against chronic pancreatitis." Pancreas **36**(3): 317-20.
Sato, T., S. Ohno, et al. (2005). "Dual functions of Runx proteins for reactivating CD8 and silencing CD4 at the commitment process into CD8 thymocytes." Immunity **22**(3): 317-28.
Satsangi, J., C. Grootscholten, et al. (1996). "Clinical patterns of familial inflammatory bowel disease." Gut **38**(5): 738-41.
Satsangi, J., M. Parkes, et al. (1998). "Genetics of inflammatory bowel disease." Clin Sci (Lond) **94**(5): 473-8.
Schaeffeler, E., C. Fischer, et al. (2004). "Comprehensive analysis of thiopurine S-methyltransferase phenotype-genotype correlation in a large population of German-Caucasians and identification of novel TPMT variants." Pharmacogenetics **14**(7): 407-17.
Schmechel, S., A. Konrad, et al. (2008). "Linking genetic susceptibility to Crohn's disease with Th17 cell function: IL-22 serum levels are increased in Crohn's disease and correlate with disease activity and IL23R genotype status." Inflamm Bowel Dis **14**(2): 204-12.
Schmid, D., M. Pypaert, et al. (2007). "Antigen-loading compartments for major histocompatibility complex class II molecules continuously receive input from autophagosomes." Immunity **26**(1): 79-92.
Schork, N. J., S. S. Murray, et al. (2009). "Common vs. rare allele hypotheses for complex diseases." Curr Opin Genet Dev **19**(3): 212-9.
Schwartz, R., J. Stack, et al. (1958). "Effect of 6-mercaptopurine on antibody production." Proc Soc Exp Biol Med **99**(1): 164-7.
Segal, I., Y. Yaakov, et al. (2008). "Cystic fibrosis transmembrane conductance regulator ion channel function testing in recurrent acute pancreatitis." J Clin Gastroenterol **42**(7): 810-4.
Seibold, F., R. Hufnagl, et al. (1999). "[Differential diagnosis of chronic inflammatory bowel diseases. Value of determination of autoantibodies pANCA, ASCA AND PAB (perinuclear antineutrophil cytoplasmic antibody, antibody against Saccharomyces cerevisiae, antibody against pancreas)]." Fortschr Med **117**(6): 42-3.
Selby, W., P. Pavli, et al. (2007). "Two-year combination antibiotic therapy with clarithromycin, rifabutin, and clofazimine for Crohn's disease." Gastroenterology **132**(7): 2313-9.
Sellon, R. K., S. Tonkonogy, et al. (1998). "Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice." Infect Immun **66**(11): 5224-31.
Sheibanie, A. F., J. H. Yen, et al. (2007). "The proinflammatory effect of prostaglandin E2 in experimental inflammatory bowel disease is mediated through the IL-23-->IL-17 axis." J Immunol **178**(12): 8138-47.
Sherry, S. T., M. H. Ward, et al. (2001). "dbSNP: the NCBI database of genetic variation." Nucleic Acids Res **29**(1): 308-11.
Silverberg, M. S., J. H. Cho, et al. (2009). "Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study." Nat Genet **41**(2): 216-20.
Simell, S., S. Hoppu, et al. (2007). "Fate of five celiac disease-associated antibodies during normal diet in genetically at-risk children observed from birth in a natural history study." Am J Gastroenterol **102**(9): 2026-35.

Simmons, D. L., R. M. Botting, et al. (2004). "Cyclooxygenase isozymes: the biology of prostaglandin synthesis and inhibition." Pharmacol Rev **56**(3): 387-437.
Singh, S. B., A. S. Davis, et al. (2006). "Human IRGM induces autophagy to eliminate intracellular mycobacteria." Science **313**(5792): 1438-41.
Sjostrom, H., K. E. Lundin, et al. (1998). "Identification of a gliadin T-cell epitope in coeliac disease: general importance of gliadin deamidation for intestinal T-cell recognition." Scand J Immunol **48**(2): 111-5.
Slatkin, M. (2009). "Epigenetic inheritance and the missing heritability problem." Genetics **182**(3): 845-50.
Smith, A. M., F. Z. Rahman, et al. (2009). "Disordered macrophage cytokine secretion underlies impaired acute inflammation and bacterial clearance in Crohn's disease." J Exp Med **206**(9): 1883-97.
Smyth, D. J., J. D. Cooper, et al. (2006). "A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region." Nat Genet **38**(6): 617-9.
Smyth, D. J., V. Plagnol, et al. (2008). "Shared and distinct genetic variants in type 1 diabetes and celiac disease." N Engl J Med **359**(26): 2767-77.
Sollid, L. M. (2002). "Coeliac disease: dissecting a complex inflammatory disorder." Nat Rev Immunol **2**(9): 647-55.
Sollid, L. M., G. Markussen, et al. (1989). "Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer." J Exp Med **169**(1): 345-50.
Soria-Royer, C., C. Legendre, et al. (1993). "Thiopurine-methyl-transferase activity to assess azathioprine myelotoxicity in renal transplant recipients." Lancet **341**(8860): 1593-4.
Spencer, C. C., Z. Su, et al. (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS Genet **5**(5): e1000477.
Spurkland, A., L. M. Sollid, et al. (1992). "HLA-DR and -DQ genotypes of celiac disease patients serologically typed to be non-DR3 or non-DR5/7." Hum Immunol **35**(3): 188-92.
Standiford, T. J., S. L. Kunkel, et al. (1992). "Regulation of human alveolar macrophage- and blood monocyte-derived interleukin-8 by prostaglandin E2 and dexamethasone." Am J Respir Cell Mol Biol **6**(1): 75-81.
Stene, L. C., M. C. Honeyman, et al. (2006). "Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study." Am J Gastroenterol **101**(10): 2333-40.
Strachan, T. R., A P. (2004). Human Molecular Genetics 3, Garland Science.
Stranger, B. E., A. C. Nica, et al. (2007). "Population genomics of human gene expression." Nat Genet **39**(10): 1217-24.
Strober, W., I. Fuss, et al. (2007). "The fundamental basis of inflammatory bowel disease." J Clin Invest **117**(3): 514-21.
Sturgess, R., P. Day, et al. (1994). "Wheat peptide challenge in coeliac disease." Lancet **343**(8900): 758-61.
Su, A. I., T. Wiltshire, et al. (2004). "A gene atlas of the mouse and human protein-encoding transcriptomes." Proc Natl Acad Sci U S A **101**(16): 6062-7.
Suau, A., R. Bonnet, et al. (1999). "Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut." Appl Environ Microbiol **65**(11): 4799-807.
Sugimoto, Y. and S. Narumiya (2007). "Prostaglandin E receptors." J Biol Chem **282**(16): 11613-7.
Tai, H. L., E. Y. Krynetski, et al. (1996). "Thiopurine S-methyltransferase deficiency: two nucleotide transitions define the most prevalent mutant allele associated with loss of catalytic activity in Caucasians." Am J Hum Genet **58**(4): 694-702.

Takahashi, H. K., H. Iwagaki, et al. (2005). "Prostaglandins E1 and E2 inhibit lipopolysaccharide-induced interleukin-18 production in monocytes." Eur J Pharmacol **517**(3): 252-6.
Takahashi, H. K., H. Iwagaki, et al. (2005). "Differential effect of prostaglandins E1 and E2 on lipopolysaccharide-induced adhesion molecule expression on human monocytes." Eur J Pharmacol **512**(2-3): 223-30.
Takayama, K., G. Garcia-Cardena, et al. (2002). "Prostaglandin E2 suppresses chemokine production in human macrophages through the EP4 receptor." J Biol Chem **277**(46): 44147-54.
Tannock, G. W. (2000). "The intestinal microflora: potentially fertile ground for microbial physiologists." Adv Microb Physiol **42**: 25-46.
The International HapMap Consortium (2005). "A haplotype map of the human genome." Nature **437**(7063): 1299-320.
Thomas, C. W., G. M. Myhre, et al. (2005). "Selective inhibition of inflammatory gene expression in activated T lymphocytes: a mechanism of immune suppression by thiopurines." J Pharmacol Exp Ther **312**(2): 537-45.
Thompson, A. J., A. J. Muir, et al. (2010). "Interleukin-28B polymorphism improves viral kinetics and is the strongest pretreatment predictor of sustained virologic response in genotype 1 hepatitis C virus." Gastroenterology **139**(1): 120-9 e18.
Tiede, I., G. Fritz, et al. (2003). "CD28-dependent Rac1 activation is the molecular target of azathioprine in primary human CD4+ T lymphocytes." J Clin Invest **111**(8): 1133-45.
Timmer, A., J. W. McDonald, et al. (2007). "Azathioprine and 6-mercaptopurine for maintenance of remission in ulcerative colitis." Cochrane Database Syst Rev(1): CD000478.
Todd, J. A. (2010). "Etiology of type 1 diabetes." Immunity **32**(4): 457-67.
Todd, J. A., N. M. Walker, et al. (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." Nat Genet **39**(7): 857-64.
Tollefsen, S., H. Arentz-Hansen, et al. (2006). "HLA-DQ2 and -DQ8 signatures of gluten T cell epitopes in celiac disease." J Clin Invest **116**(8): 2226-36.
Tosi, R., D. Vismara, et al. (1983). "Evidence that celiac disease is primarily associated with a DC locus allelic specificity." Clin Immunol Immunopathol **28**(3): 395-404.
Treton, X., Y. Bouhnik, et al. (2009). "Azathioprine withdrawal in patients with Crohn's disease maintained on prolonged remission: a high risk of relapse." Clin Gastroenterol Hepatol **7**(1): 80-5.
Trynka, G., A. Zhernakova, et al. (2009). "Coeliac disease associated risk variants in TNFAIP3 and REL implicate altered NF-{kappa}B signalling." Gut.
Tysk, C., E. Lindberg, et al. (1988). "Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking." Gut **29**(7): 990-6.
Ueda, H., J. M. Howson, et al. (2003). "Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease." Nature **423**(6939): 506-11.
UEGW Working Group (2001). "When is a coeliac a coeliac? Report of a working group of the United European Gastroenterology Week in Amsterdam, 2001." Eur J Gastroenterol Hepatol **13**(9): 1123-8.
Uhlig, H. H., J. Coombes, et al. (2006). "Characterization of Foxp3+CD4+CD25+ and IL-10-secreting CD4+CD25+ T cells during cure of colitis." J Immunol **177**(9): 5852-60.
Unoki, H., A. Takahashi, et al. (2008). "SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations." Nat Genet **40**(9): 1098-102.
Vader, W., Y. Kooy, et al. (2002). "The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides." Gastroenterology **122**(7): 1729-37.

Vader, W., D. Stepniak, et al. (2003). "The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses." Proc Natl Acad Sci U S A **100**(21): 12390-5.
Vafiadis, P., S. T. Bennett, et al. (1997). "Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus." Nat Genet **15**(3): 289-92.
van Belzen, M. J., B. P. Koeleman, et al. (2004). "Defining the contribution of the HLA region to cis DQ2-positive coeliac disease patients." Genes Immun **5**(3): 215-20.
Van Belzen, M. J., J. W. Meijer, et al. (2003). "A major non-HLA locus in celiac disease maps to chromosome 19." Gastroenterology **125**(4): 1032-41.
van de Wal, Y., Y. Kooy, et al. (1998). "Selective deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity." J Immunol **161**(4): 1585-8.
van de Wal, Y., Y. M. Kooy, et al. (1997). "Unique peptide binding characteristics of the disease-associated DQ(alpha 1*0501, beta 1*0201) vs the non-disease-associated DQ(alpha 1*0201, beta 1*0202) molecule." Immunogenetics **46**(6): 484-92.
van der Pouw Kraan, T. C., L. C. Boeije, et al. (1995). "Prostaglandin-E2 is a potent inhibitor of human interleukin 12 production." J Exp Med **181**(2): 775-9.
van Dieren, J. M., A. J. van Vuuren, et al. (2005). "ITPA genotyping is not predictive for the development of side effects in AZA treated inflammatory bowel disease patients." Gut **54**(11): 1664.
van Heel, D. A., L. Franke, et al. (2007). "A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21." Nat Genet **39**(7): 827-9.
van Heel, D. A., S. Ghosh, et al. (2005). "Muramyl dipeptide and toll-like receptor sensitivity in NOD2-associated Crohn's disease." Lancet **365**(9473): 1794-6.
van Heel, D. A., K. Hunt, et al. (2005). "Genetics in coeliac disease." Best Pract Res Clin Gastroenterol **19**(3): 323-39.
van Heel, D. A. and J. West (2006). "Recent advances in coeliac disease." Gut **55**(7): 1037-46.
van Hoek, M., A. Dehghan, et al. (2008). "Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study." Diabetes **57**(11): 3122-8.
van Iersel, C. A., H. J. de Koning, et al. (2007). "Risk-based selection from the general population in a screening trial: selection criteria, recruitment and power for the Dutch-Belgian randomised lung cancer multi-slice CT screening trial (NELSON)." Int J Cancer **120**(4): 868-74.
Van Limbergen, J., D. C. Wilson, et al. (2009). "The genetics of Crohn's disease." Annu Rev Genomics Hum Genet **10**: 89-116.
Vartdal, F., B. H. Johansen, et al. (1996). "The peptide binding motif of the disease associated HLA-DQ (alpha 1* 0501, beta 1* 0201) molecule." Eur J Immunol **26**(11): 2764-72.
Velazquez, L., A. M. Cheng, et al. (2002). "Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice." J Exp Med **195**(12): 1599-611.
Venter, J. C., M. D. Adams, et al. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.
Vernier-Massouille, G., M. Balde, et al. (2008). "Natural history of pediatric Crohn's disease: a population-based cohort study." Gastroenterology **135**(4): 1106-13.
Viglianti, G. A., C. M. Lau, et al. (2003). "Activation of autoreactive B cells by CpG dsDNA." Immunity **19**(6): 837-47.
Waldmann, T. A. (2006). "The biology of interleukin-2 and interleukin-15: implications for cancer therapy and vaccine design." Nat Rev Immunol **6**(8): 595-601.
Wang, D. and R. N. Dubois (2010). "The role of COX-2 in intestinal inflammation and colorectal cancer." Oncogene **29**(6): 781-8.

Wang, J. and Y. X. Fu (2003). "LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide." J Immunol **170**(8): 3986-93.
Wang, J., W. Wang, et al. (2008). "The diploid genome sequence of an Asian individual." Nature **456**(7218): 60-5.
Wang, K., R. Baldassano, et al. (2010). "Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects." Hum Mol Genet.
Warman, J. I., B. I. Korelitz, et al. (2003). "Cumulative experience with short- and long-term toxicity to 6-mercaptopurine in the treatment of Crohn's disease and ulcerative colitis." J Clin Gastroenterol **37**(3): 220-5.
Weber, J. L. and P. E. May (1989). "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction." Am J Hum Genet **44**(3): 388-96.
Weedon, M. N., G. Lettre, et al. (2007). "A common variant of HMGA2 is associated with adult and childhood height in the general population." Nat Genet **39**(10): 1245-50.
Weersma, R. K., M. R. Batstra, et al. (2008). "Are pancreatic autoantibodies associated with azathioprine-induced pancreatitis in Crohn's disease?" Jop **9**(3): 283-9.
Weersma, R. K., F. T. Peters, et al. (2004). "Increased incidence of azathioprine-induced pancreatitis in Crohn's disease compared with other diseases." Aliment Pharmacol Ther **20**(8): 843-50.
Weersma, R. K., P. C. Stokkers, et al. (2009). "Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort." Gut **58**(3): 388-95.
Weersma, R. K., A. Zhernakova, et al. (2007). "ATG16L1 and IL23R Are Associated With Inflammatory Bowel Diseases but Not With Celiac Disease in The Netherlands." Am J Gastroenterol.
Wehkamp, J., J. Harder, et al. (2004). "NOD2 (CARD15) mutations in Crohn's disease are associated with diminished mucosal alpha-defensin expression." Gut **53**(11): 1658-64.
Weile, B., B. Cavell, et al. (1995). "Striking differences in the incidence of childhood celiac disease between Denmark and Sweden: a plausible explanation." J Pediatr Gastroenterol Nutr **21**(1): 64-8.
Weinshilboum, R. M. and S. L. Sladek (1980). "Mercaptopurine pharmacogenetics: monogenic inheritance of erythrocyte thiopurine methyltransferase activity." Am J Hum Genet **32**(5): 651-62.
Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-78.
West, J., R. F. Logan, et al. (2003). "Seroprevalence, correlates, and characteristics of undetected coeliac disease in England." Gut **52**(7): 960-5.
Wheeler, D. A., M. Srinivasan, et al. (2008). "The complete genome of an individual by massively parallel DNA sequencing." Nature **452**(7189): 872-6.
Whitcomb, D. C. "Genetic aspects of pancreatitis." Annu Rev Med **61**: 413-24.
Whitcomb, D. C. (2004). "Mechanisms of disease: Advances in understanding the mechanisms leading to chronic pancreatitis." Nat Clin Pract Gastroenterol Hepatol **1**(1): 46-52.
Wilson, R. J., G. M. Giblin, et al. (2006). "GW627368X ((N-{2-[4-(4,9-diethoxy-1-oxo-1,3-dihydro-2H-benzo[f]isoindol-2-yl)phenyl] acetyl} benzene sulphonamide): a novel, potent and selective prostanoid EP4 receptor antagonist." Br J Pharmacol **148**(3): 326-39.
Wilson, R. J., S. A. Rhodes, et al. (2004). "Functional pharmacology of human prostanoid EP2 and EP4 receptors." Eur J Pharmacol **501**(1-3): 49-58.
Wing, K., Y. Onishi, et al. (2008). "CTLA-4 control over Foxp3+ regulatory T cell function." Science **322**(5899): 271-5.

Witt, H., W. Luck, et al. (1999). "A signal peptide cleavage site mutation in the cationic trypsinogen gene is strongly associated with chronic pancreatitis." Gastroenterology **117**(1): 7-10.
Witt, H., M. Sahin-Toth, et al. (2006). "A degradation-sensitive anionic trypsinogen (PRSS2) variant protects against chronic pancreatitis." Nat Genet **38**(6): 668-73.
Woolf, E., C. Xiao, et al. (2003). "Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis." Proc Natl Acad Sci U S A **100**(13): 7731-6.
Wotton, D., H. M. Prosser, et al. (1993). "Regulation of human T cell receptor beta gene expression by Ets-1." Leukemia **7 Suppl 2**: S55-60.
Wright, S. (1934). "An analysis of variability in the number of digits in an inbred strain of guinea pigs." Genetics **19**: 537-551.
WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-78.
Xie, Z. and D. J. Klionsky (2007). "Autophagosome formation: core machinery and adaptations." Nat Cell Biol **9**(10): 1102-9.
Yamamoto, H., T. Maruyama, et al. (1999). "Novel four selective agonists for prostaglandin E receptor subtypes." Prostaglandins & other Lipid Mediators **59**: 150.
Yamane, H., Y. Sugimoto, et al. (2000). "Prostaglandin E(2) receptors, EP2 and EP4, differentially modulate TNF-alpha and IL-6 production induced by lipopolysaccharide in mouse peritoneal neutrophils." Biochem Biophys Res Commun **278**(1): 224-8.
Yamanouchi, J., D. Rainbow, et al. (2007). "Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity." Nat Genet **39**(3): 329-37.
Yang, J., B. Benyamin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-9.
Yao, C., D. Sakata, et al. (2009). "Prostaglandin E(2)-EP4 signaling promotes immune inflammation through T(H)1 cell differentiation and T(H)17 cell expansion." Nat Med.
Yasuda, K., K. Miyake, et al. (2008). "Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus." Nat Genet **40**(9): 1092-7.
Yu, K., Z. Wang, et al. (2008). "Population substructure and control selection in genome-wide association studies." PLoS ONE **3**(7): e2551.
Yu, W., M. Clyne, et al. (2009). "Phenopedia and Genopedia: Disease-centered and Gene-centered Views of the Evolving Knowledge of Human Genetic Associations." Bioinformatics.
Zamisch, M., L. Tian, et al. (2009). "The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus." J Exp Med.
Zanoni, G., R. Navone, et al. (2006). "In celiac disease, a subset of autoantibodies against transglutaminase binds toll-like receptor 4 and induces activation of monocytes." PLoS Med **3**(9): e358.
Zhang, H. F., L. X. Qiu, et al. (2009). "ATG16L1 T300A polymorphism and Crohn's disease susceptibility: evidence from 13,022 cases and 17,532 controls." Hum Genet.
Zhernakova, A., C. C. van Diemen, et al. (2009). "Detecting shared pathogenesis from the shared genetics of immune-related diseases." Nat Rev Genet **10**(1): 43-55.

Appendix 1 Phenotyping form for azathioprine induced pancreatitis cases

Investigator name (person reviewing medical records):
Participant ID number:
Phenotype Information Form for possible Aza/6-MP-induced Pancreatitis cases

Participant Name/sex _____Participant Id Number_____
Date of Birth _____ Hospital/Centre_____

ESSENTIAL INFORMATION

1. What is the individual's ethnicity?
   - White European ☐
   - Non-White European- specify region of origin ☐
     _____
   - Mixed ethnicity-specify ☐
     _____

2. What was the indication for thiopurine treatment?
   - Crohn's Disease ☐
   - Ulcerative Colitis ☐
   - Other-Specify disease/indication ☐
     _____

3. Did the individual have clinical symptoms consistent with acute pancreatitis including acute, severe abdominal pain? **Y / N / unknown**

4. Was the individual taking aza/MP within 1 week of the start of symptoms? **Y / N / unknown**

5. What was the time interval between starting the thiopurine and onset of pancreatitis (start of symptoms)? **_____days**

6. Which thiopurine was suspected of causing pancreatitis?
   - Azathioprine ☐
   - Mercaptopurine ☐

   (give dose -total and mg/kg if known)
   _____

7. Did the individual have raised serum or urinary pancreatic enzymes (amylase or lipase) consistent with the timing of symptoms? **Y / N / not done / unknown**
   - If available please record the individual's highest serum/urinary amylase (and laboratory normal range) _____

8. Did the individual have imaging evidence of pancreatitis (e.g. CT or Ultrasound) consistent with the timing of symptoms? **Y / N / not done / unknown**
   - If yes, state modality and brief radiological findings _____

8. Did symptoms resolve on thiopurine withdrawal? **Y / N / unknown**

9. Was the individual ever re-challenged with a thiopurine? **Y / N / unknown**
   - and if so did the individual tolerate re-challenge?
     Y / N-recurrent pancreatitis / N-abdominal pain / N-other / unknown

Investigator name (person reviewing medical records):
Participant ID number:

10. Was a diagnosis of azathioprine or mercaptopurine -induced pancreatitis documented in the individual's medical records? **Y / N / unknown**

11. Was there evidence that might suggest an alternative cause of pancreatitis?

- Did the individual have a history of heavy alcohol use? **Y / N / unknown**
  If yes, amount (Units/week)_____
- Did the individual have a history of previous acute pancreatitis, unrelated to thiopurines? **Y / N / unknown**
- Did the individual have a history of chronic pancreatitis? **Y / N / unknown**
- Did the individual have evidence of gallstones on imaging? **Y / N / not done / unknown**
- If any other potential causes of pancreatitis were identified (e.g. other drugs including 5-ASAs) please specify

_____

12. In the opinion of the doctor now reviewing the medical records, was a thiopurine the most likely cause of pancreatitis (give reasons if other potential causes identified)? **Y / N**

_____

SUPPLEMENTARY INFORMATION

13. What is the individual's thiopurine methytransferase genotype/phenotype ?
Please state TPMT genotype or enzyme activity (with laboratory reference range)
_____

- If TPMT activity stated, was this result obtained before starting a thiopurine or during thiopurine treatment? **Before / During / unknown**

14. Has the individual experienced any other adverse effects attributable to azathioprine/mercaptopurine? **Y / N / unknown**

- Abnormal LFTs (give maximally abnormal LFT values inc ALT/AST) ☐
- Leucopenia (give minimum white count/ neutrophil count) ☐

  WCC ($x10^9$/L): _____Neutrophils ($x10^9$/L)_____
- Other (state) ☐

  _____

Appendix 2 Publications

1. **Dubois, P. C**. and D. A. van Heel (2008). "New susceptibility genes for ulcerative colitis." Nat Genet **40**(6): 686-8.

2. **Dubois, P. C**. and D. A. van Heel (2008). "Translational mini-review series on the immunogenetics of gut disease: immunogenetics of coeliac disease." Clin Exp Immunol **153**(2): 162-73.

3. **Dubois, P.C**, K Hunt and D.A. van Heel (2009). "Sex differences in HLA DQ in celiac disease." Am J Gastroenterol **104**(3): 784

4. **Dubois, P. C.**, G. Trynka, et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression." Nat Genet **42**(4): 295-302

5. **Dubois, P. C**. and D. A. van Heel (2010). Coeliac Disease. Oxford Textbook of Medicine (5th Edition), Oxford University Press. **Vol 2**, Chapter 15.10.3.

6. Smyth, D.J, V Plagnol, N Walker, J.D Cooper, K Downes, J.H.M Yang, J.M.M Howson, H. Stevens, R. McManus, C. Wijmenga, G.A. Heap, **P.C. Dubois**, D.G. Clayton, K.A. Hunt, D.A. van Heel, J.A. Todd (2008). Shared and distinct genetic variants in type 1 diabetes and celiac disease New England Journal of Medicine 359(26):2767-77

7. Koskinen, L.L, E Einarsdottir, E Dukes, G.A. Heap, **P Dubois,** I.R. Korponay-Szabo, K. Kaukinen, K Kurppa, F Ziberna, S Vatta, T Not, A Ventura, P Sistonen, R Adány, Z Pocsai, G Széles, M Mäki, J Kere, C Wijmenga, D.A. van Heel, P. Saavalainen (2009). Association study of the IL18RAP locus in three european populations with celiac disease. Human Molecular Genetics 18(6):1148-55

8. Heap GA, J.H. Yang, K Downes, B.C. Healy, K.A. Hunt, N Bockett, L Franke, **P.C. Dubois,** C.A. Mein, R.J. Dobson, T.J. Albert, M.J. Rodesch, D.G. Clayton, J.A. Todd, D.A. van Heel, V. Plagnol (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. Human Molecular Genetics**.** 19(1):122-34

Full copies of publications 2, 4 and 5 (uncorrected proof) are included below:
Publication 1 is a “News and Views” section commentary and publication 3 is a “Letter to the editor”. These and non-first author articles are not included in full here.

Clinical & Experimental Immunology
The Journal of Translational Immunology
British Society for immunology
Clinical and Experimental Immunology
IMMUNOGENETICS OF GUT DISEASE

Translational Mini-Review Series on the Immunogenetics of Gut Disease: Immunogenetics of coeliac disease

OTHER ARTICLE PUBLISHED IN THIS TRANSLATIONAL MINI-REVIEW SERIES ON THE IMMUNOGENETICS OF GUT DISEASE
Immunogenetics of Inflammatory Bowel Disease

P. C. Dubois and D. A. van Heel
Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, London, UK

Accepted for publication 9 May 2008
Correspondence: P. C. Dubois, Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, 4 Newark Street, London E1 2AT, UK.
E-mail: p.c.dubois@qmul.ac.uk

Summary

Recent advances in immunological and genetic research in coeliac disease provide new and complementary insights into the immune response driving this chronic intestinal inflammatory disorder. Both approaches confirm the central importance of T cell-mediated immune responses to disease pathogenesis and have further begun to highlight other relevant components of the mucosal immune system, including innate immunity and the control of lymphocyte trafficking to the mucosa. In the last year, the first genome wide association study in celiac disease led to the identification of multiple new risk variants. These risk regions implicate genes involved in the immune system. Overlap with autoimmune diseases is striking with several of these regions being shown to confer susceptibility to other chronic immune-mediated diseases, particularly type 1 diabetes.

Keywords: autoimmune, coeliac, genetics, immunogenetics, genome-wide association

Introduction

Coeliac disease is a common intestinal inflammatory condition with prevalence estimates of 0·5–1% in populations of European ancestry [1]. Dietary prolamins (storage proteins in grain) from wheat, rye and barley trigger inflammation in the small intestine in susceptible individuals. Heritable genetic variation is a major determinant of this susceptibility, with a greater genetic contribution than for many common complex diseases. Until very recently, human leucocyte antigen DQ (HLA-DQ) gene variants have dominated our understanding of this genetic susceptibility and their identification has led to an immunological appreciation of how DQ heterodimers present gluten epitopes and drive T cell reactivity in coeliac disease. The identification of non-HLA susceptibility genes has accelerated dramatically in the past year, following the first genome-wide association study (GWAS) in coeliac disease. This study and a follow-up identified at least eight new genomic regions with robust levels of disease association [2,3]. Seven of these regions harbour genes with known immune functions and many are also implicated in conferring susceptibility to other autoimmune diseases.

This review is a synthesis of our current understanding of the genetics and immunology of coeliac disease. Major advances have been achieved in coeliac disease, in part because the antigen is well defined (cereal gluten), target organ (small intestine) samples are readily obtained and a strong genetic component to susceptibility has enabled disease gene identification. These advances have many general relevant findings for other human chronic immune-mediated diseases.

Epidemiology

Serological screening of populations in Europe and regions with a high proportion of European descendents (North and South America, Australasia) suggests a coeliac disease prevalence of approximately 0·5–1% in adults [1,4]. More limited data from North Africa and South-west Asia suggest a similar high prevalence of coeliac disease in these areas [5]. In central Africa and the Far East there have been no large seroprevalence studies, but overt coeliac disease is extremely rare [6–8]. A study from Burkina Faso screened 600 individuals, all of whom ate wheat, but found no individuals with positive coeliac serology. Furthermore, no individuals carried HLA-DQ2 and only one HLA-DQ8 [9]. The Sawahari population of North Africa have the highest reported prevalence of coeliac disease worldwide (5·4%) mirrored by

a very high carriage of the coeliac susceptibility marker HLA DQ2, whereas the prevalence of HLA DQ2 is very low in the Far East [10,11]. Genetic differences across populations (particularly in HLA types) clearly contribute to the different observed population prevalences.

Grain consumption also broadly parallels coeliac prevalence, being low in the Far East and sub-Saharan Africa [11]. Furthermore, there is some evidence that the dose of gluten, particularly in early childhood, may be an important determinant of lifetime susceptibility. Countries in which infant gluten consumption is low (Denmark, Estonia, Finland) report a lower infant (and adult) incidence of coeliac disease than countries with a high infant gluten consumption (Sweden) [12,13].

Adult coeliac disease prevalence has been increasing over the last few decades [1]. Improved clinical ascertainment contributes (especially in the United States), although some studies suggest a true increase in seroprevalence [14]. Similar increases in prevalence have occurred in other chronic immune-mediated diseases, particularly type 1 diabetes, implicating recent changes in shared environmental factors [15]. These factors remain unknown, although interest has focused logically upon exposures occurring in early childhood, which might be critical in determining lifetime autoimmune disease risk. In coeliac disease, onset can occur at any age but the peak incidence is between 9 and 24 months, following the introduction of gluten into the diet [1]. Breast feeding during gluten introduction has been shown to reduce susceptibility, suggesting that tolerance to gluten can be influenced by factors in breast milk [16]. Tolerance to gluten might also be influenced by the context in which it is encountered by the mucosal immune system in early life. Childhood intestinal infections have been proposed as a factor that could promote loss of tolerance to gluten, possibly because of disrupted intestinal epithelial barrier function. Furthermore, inflammation up-regulates tissue transglutaminase (tTG), a key enzyme in coeliac disease required for the generation of immunogenic epitopes from gluten [17]. There are no animal models of coeliac disease to test this hypothesis and direct evidence for the role of intestinal infections is lacking. However, epidemiological studies have shown that coeliac disease is more common in children born in summer months, possibly because of the higher incidence of viral enteritis in winter months when these children start eating gluten [18] Case–control studies have also suggested that increased exposure to infant enteral infections may confer modest increased susceptibility [odds ratios (ORs) of 1·4–1·5] [19,20]. Finally, one prospective study measured episodes of rotavirus infection by serology and found a modest increase in coeliac autoantibody incidence in infants exposed to multiple infections [21].

Although the development of coeliac disease has been considered a permanent gluten-sensitive enteropathy, needing lifelong treatment, recent reports suggest that some children can resolve this intolerance at least partially when kept on a gluten-containing diet [22,23]. These children may have normal small intestinal histology in adulthood, suggesting that coeliac disease can remit or enter a quiescent phase, with immunological tolerance to gluten, following initial clinically overt disease. How frequently this phenomenon occurs is unclear; much more research in this area is necessary – including whether such remission might be induced therapeutically.

Evidence for genetic susceptibility

Closely related individuals with coeliac disease have a higher disease concordance than unrelated individuals (familial clustering). Monozygotic twins have disease concordance rates of 75% compared with 11% in dizygotic twins [24]. Sibling relative risk ratios (λ_s) provide the best estimates of familial clustering, controlling for population prevalence. For coeliac disease, sibling relative risk ratios of between 20 and 60 have been reported [25–27]. This is higher than for most other polygenic immune-mediated disorders such as type 1 diabetes ($\lambda_s = 15$), rheumatoid arthritis ($\lambda_s = 2–8$) or Crohn's disease ($\lambda_s = 27$) [28].

Immunogenetics of the HLA

The HLA complex is a highly polymorphic 4 Mb region on chromosome 6p21, containing more than 200 genes and over 3000 known alleles [29]. HLA class II genes (DP, DQ and DR) are involved in exogenous peptide antigen presentation to T cells. The first reports of association with coeliac disease used serological methods to identify B8 and later DR3 as susceptibility variants [30,31]. The B8 and DR3 molecules are encoded by alleles on a 6Mb extended haplotype (A1-B8-DR3-DQ2) present in 10% of northern Europeans [32]. Interestingly, other autoimmune diseases are associated with this haplotype, including type 1 diabetes and autoimmune thyroid disease. Subsequent studies have pinpointed DQ2 and in particular the combination of HLA-DQA1*0501 and DQB1*0201 encoding the HLA-DQ2 (α1*0501,β1*0201) heterodimer as the cause of the coeliac disease association [33]. This heterodimer can be encoded both in *cis* (by alleles on the same haplotype) or in *trans* (one subunit each from paternal and maternal haplotypes) (Table 1, Fig. 1). Moreover several studies show that homozygosity for the *cis* haplotype or possessing a second DQB1*02 allele increases coeliac disease susceptibility further [37,38]. The second B1*02 allele is usually inherited on the DR7-DQ2 haplotype carrying DQB1*0202 and DQA1*0201 (DQ2·2), but possession of this haplotype alone does not confer coeliac susceptibility (Table 1).

An explanation for the HLA gene-dosage effect was provided by an *in vitro* study demonstrating that the level of proliferation and cytokine responses of gluten-reactive T cell clones depends on DQ type and gene dose [35]. Vader *et al.* used allogeneic peripheral blood mononuclear cells to

Table 1. Classical human leucocyte antigen (HLA) DQ genotypes associated with coeliac disease and gene dosage effects.

Serological type	Chromosome copy	DQ2 genotype	DQ type	Coeliac susceptibility
DR3–DQ2/ DR3–DQ2	i ii	DQA1*0501–DQB1*0201/ DQA1*0501–DQB1*0201	DQ2·5 *cis* homozygote	High
DR3–DQ2/ DR7–DQ2	i ii	DQA1*0501–DQB1*0201/ DQA1*0201–DQB1*0202	DQ2·5 *cis* + DQ2·2	High
DR3–DQ2/ other	i ii	DQA1*0501–DQB1*0201/ other	DQ2·5 *cis* heterozygote	Moderate
DR5–DQ7/ DR7–DQ2	i ii	DQA1*0505–DQB1*0301/ DQA1*0201–DQB1*0202	*DQ2·5trans*	Moderate
DR7–DQ2/ other	i ii	DQA1*0201–DQB1*0202/ other	DQ2·2	Nil
DR4–DQ8/ other	i ii	DQA1*0301–DQB1*0302/ other	DQ8	Moderate

Disease causing alleles highlighted (see also Fig. 1). Adapted from van Heel *et al.* [34]; DQ2 type naming after Vader *et al.* [35].

present gluten epitopes to gluten-specific T cell clones and showed that T cell responses were highest for DQ2·5 homozygotes, intermediate for DQ2·5/2·2, lower for DQ2·5/x heterozygotes and lowest for DQ2·2. Thus DQ2·2 in the presence of DQ2·5 can augment T cell stimulation through DQ2-mediated antigen presentation. DQ2·2 alone, which is not associated with coeliac disease, was able to elicit strong T cell responses but only through presentation of a restricted subset of the gluten epitopes tested. This suggests that the DQ2 contribution to coeliac disease depends upon its ability to present multiple closely related gluten epitopes – the ability of DQ2·2 molecules to present a small subset of epitopes exerts effects too weak to cause disease.

The HLA-DQ2·5 molecule encoded either in *cis* or in *trans* is present in around 90% of coeliac patients of northern European origin [39]. The majority of the remainder carry HLA-DQ8 (genetically DQA1*03, DQB1*0302) [40,41]. A large European collaborative study found that of those that lack both DQ2 and DQ8, only four of 1008 coeliac patients had neither the alpha nor beta chain of the DQ2 heterodimer [41] This has led to a model of coeliac disease pathogenesis in which HLA DQ2/8 is necessary but not sufficient, as HLA-DQ2 is present in 30% of healthy Caucasian populations [39]. The proportion of sibling relative risk attributable to known HLA variants is estimated to be between 30 and 40%, indicating that non-HLA DQ variants contribute to coeliac disease susceptibility [3,25,26]. Within the HLA complex itself there are many other genes with immune functions which might also contribute to the observed association signal. However, the high linkage disequilibrium (LD) that exists between genetic variants in this region is an obstacle to teasing out the true causal associations [42]. Two studies that have controlled for LD to DQ have not found evidence of additional HLA risk variants, although statistical power was limited [41,43].

The genetic loci harbouring variants that account for the remaining 70% or so of unexplained familial clustering in coeliac disease are the targets of gene finding studies. Two complementary approaches have been used: genetic linkage and association studies (Table 2).

In general, findings from linkage and candidate gene studies in coeliac disease, with the exception of the HLA

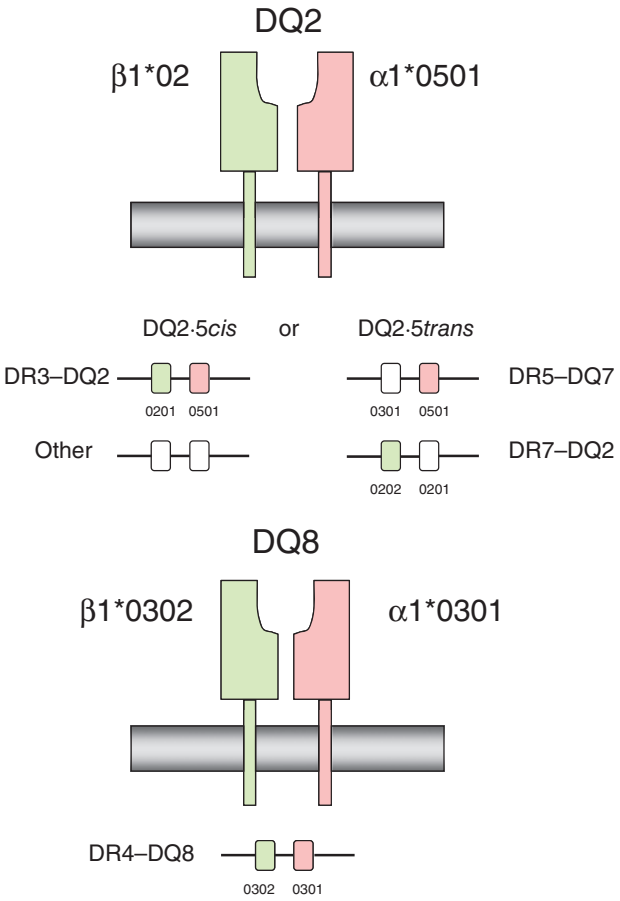


Fig. 1. Classical haplotype combinations encoding the human leucocyte antigen (HLA)-DQ2 and -DQ8 heterodimers. Adapted from Sollid [36]. HLA proteins at the cell surface, and structure of the protein encoding DNA region, are shown.

Table 2. Gene-finding approaches in coeliac disease.

Study type	Method	Advantages	Disadvantages	Examples
Linkage studies	Test co-segregation of genetic markers with disease phenotype in affected relatives to establish broad regions of genome within which causal variants reside	Able to detect rare variants, and structural variants, if highly penetrant	Low power to detect weakly penetrant alleles Low genomic resolution Require large numbers of affected families	5q31–33 [44,45] 19p13.1 (?*MYO9B*) [46]
Candidate gene association studies	Compare frequencies of variants in candidate genes chosen on biological grounds or from knowledge of linkage regions	May pinpoint genes from regions of linkage Greater power to detect weakly penetrant alleles	Low power to detect rare variants Historically generated many false positives	*CTLA4* [47]
Genome-wide association studies	Compare frequencies of ~10^5 single nucleotide polymorphisms distributed throughout the genome between cases and controls	High resolution: able to pinpoint small region of genome Power to detect weakly penetrant alleles	Low power to detect rare alleles Low power to detect structural variants Expensive	*IL2–IL21* region, *RGS1*, *IL18RAP*, *SH2B3* [2,3]

region, have not been replicated consistently. Linkage regions identified include 5q31–33 and 19p13.1, although these remain tentative and lack robust replication [44,46]. *MYO9B*, encoding the myosin IXB protein, has emerged as a candidate gene from further studies of the 19p13.1 linkage region, although replication of this finding has been inconsistent [48–51]. A candidate gene approach identified an association in the *CTLA4* region, a gene on chromosome 2q encoding cytotoxic T lymphocyte antigen 4 [47]. CTLA-4 is expressed on T cells and is a receptor for B7 molecules that inhibit T cell activation. Replication studies of the *CTLA4* association have been somewhat inconclusive [34]. Therefore, prior to the first GWAS in 2007, despite intensive efforts, no genetic susceptibility loci other than HLA DQ had been definitively identified.

Human leucocyte antigen-DQ-restricted T cells

Coeliac disease has multi-systemic features, but the predominant lesion mirrors the exposure of the small intestine to dietary gluten. Several lines of evidence implicate a T cell-orchestrated immunopathogenesis. Upon gluten challenge of small intestinal biopsies from treated (i.e. on a gluten-free diet) coeliac disease patients, infiltration of the lamina propria (LP) with (predominantly $CD4^+$ αβ) T cells occurs within hours, followed by crypt hyperplasia and villous atrophy [52]. This temporal sequence alludes to the central importance of T cells in coeliac disease. In untreated disease T helper 1 (Th1) cytokines are highly expressed in the intestinal mucosa, particularly interferon (IFN)-γ, supporting the concept of a Th1-driven T cell-mediated disorder [53]. Analysis of LP infiltrating lymphocytes confirms not only IFN-γ expression in a high proportion, but also expression of the Th1 transcription factor T-bet [54]. The Th1 bias of $CD4^+$ T cells probably depends less on interleukin (IL)-12 in coeliac disease than in other inflammatory conditions. IL-12 is present in very low levels in coeliac disease mucosa [55,56] although other Th1-inducing cytokines (IL-18 and IFN-α) are increased [56–58]. Dendritic cells isolated from the intestinal mucosa in coeliac disease also express increased levels of IL-18 and IFN-α but lack IL12p40 [55]. Immunophenotyping of $DQ2^+$ antigen-presenting cells in treated versus untreated coeliac disease intestinal biopsies suggest a large increase in $CD11^+$ myeloid dendritic cells in active disease [55,59]. These cells efficiently present gluten peptides to CD4 T cells inducing proliferation and IFN-γ responses [59].

The gluten-responsiveness of CD4 T cells in coeliac disease was first demonstrated in T cell lines and clones isolated from intestinal mucosa [60,61]. These cells are not found in non-coeliac $DQ2^-$ or $DQ8^-$ controls but in coeliac disease proliferate and secrete IFN-γ when co-cultured with antigen-presenting cells in the presence of a variety of peptides derived from gluten. These studies show that gluten peptides activate T cells in the intestinal mucosa exclusively through presentation by the disease-associated $DQ2^-$ or $DQ8^-$ αβ heterodimers [60,61].

Gluten epitopes and the role of tTG

While there is heterogeneity between patients with coeliac disease in the gluten epitopes to which their T cells respond, some epitopes are immunodominant and elicit T cell activation in almost all coeliac individuals [62,63]. These responses have been demonstrated both in intestine-derived T cell lines or clones and in primary T cells isolated from peripheral blood following gluten challenge, supporting

their contribution to disease *in vivo* [63,64]. T cell epitopes identified to date are derived from various gluten proteins, including α-gliadins, γ-gliadins and low molecular weight glutenins [62,65–67]. The peptide-binding groove structure of DQ2 and DQ8 dimers has been characterized and some of the constraints this places on selection of epitopes for binding DQ2 or DQ8 are known [68]. Both DQ2 and DQ8 dimers have preferences for negatively charged residues at key positions in the core peptide-binding groove [69–71]. Negatively charged residues are uncommon in gluten peptide sequences, but deamidation of glutamine residues to negatively charged glutamate can increase drastically the immunogenicity of gliadin peptides [67]. X-ray crystallographic analysis of DQ2-peptide interactions supports the importance of selective deamidation of glutamine residues in favouring peptide binding for gluten peptides [72,73]. tTG, an enzyme first linked to coeliac disease by the discovery that it is the target of autoantibodies used in diagnosis, can catalyse this deamidation [74,75]. tTG is likely to perform this function *in vivo*, as it highly expressed in the small intestine, up-regulated in inflammation and favours deamidation of glutamine residues rather than transamidation under the acidic conditions which exist in the proximal small intestine [76]. More recently, a direct pathogenic contribution of tTG antibodies has been proposed, with *in vitro* studies suggesting that these antibodies can both activate monocytes by binding Toll-like receptor 4 and inhibit angiogenesis by altering tTG function [77,78]. Such effects, if substantiated, may be a mechanism driving extra-intestinal manifestations in coeliac disease, because tTG autoantibody deposits have been observed in affected organs (e.g. liver, brain) remote from the site of gluten exposure in the intestine [79,80].

A further important characteristic of gluten epitopes is a high proline content [65]. This reflects the inability of human digestive enzymes to break amide bonds between proline residues and adjacent bulky hydrophobic amino acids, such that gluten peptides can reach the intestinal mucosa intact [65,81].

The innate immune system in coeliac disease

Both *in vivo* studies and studies of gluten challenge of intestinal biopsies have shown that effects on the mucosa begin within a few hours [82–84]. This rapid onset cannot be accounted for easily by the (presumably slower) mechanism of gluten peptide presentation to $CD4^+$ T lymphocytes and has led to interest in a role for the innate immune system in coeliac disease. Further support for this hypothesis came from the observation that some gliadin peptides (p31–p43 α gliadin) that do not elicit classical DQ-restricted $CD4^+$ T cell responses can exert toxic effects on the epithelium [85]. IL-15, which is highly expressed in LP macrophages and intestinal epithelium, appears to be a crucial intermediary of these effects. IL-15 enhances intra-epithelial lymphocyte (IEL) proliferation, cytotoxicity (*versus* epithelial cells) and cytokine release, with increases in IFN-γ and granzyme B [86,87] Furthermore, exogenous application of IL-15 partly reproduces the effects of gliadin challenge, whereas anti-IL-15 antibodies abrogate the effects of gliadin [87].

A feature of coeliac disease is expansion of the IEL population, as well as an inflammatory cell infiltrate deeper in the intestinal LP. The IELs in coeliac disease comprise increased populations of both $CD8^+$ TCRαβ lymphocytes as well as γδ ($CD4^-CD8^-$ or $CD8^+$) T cells that can induce enterocyte apoptosis directly [88]. Some intra-epithelial T cells have been shown to demonstrate aberrant expression of natural killer (NK) lineage receptors and can perform NK-like functions including T cell receptor-independent killing of enterocytes in active coeliac disease [88–90]. These effects are stimulated by gluten peptides including p31–43 α-gliadin and include the induction of expression of the cell surface stress molecule major histocompatibility complex class I chain related gene A on enterocytes and its receptor NKG2D on IELs [90]. Mechanistic details of the recognition of these apparently 'innate' peptides are unclear.

Genome-wide Association Studies in coeliac disease

Current models of complex disease estimate that that the majority of genetic variation contributing to disease susceptibility is carried by multiple variants of weak effect size. Variants with modest effects are below the threshold of detection of even large linkage studies and candidate gene association studies have rarely proved successful as a primary approach to gene finding. GWAS offer major advantages both in power to detect variants with modest effects and in defining smaller genomic regions in which causal variants reside [91]. Nevertheless, the power of GWAS depends upon many variables including sample size, number of single nucleotide polymorphisms (SNPs) tested, ORs conferred by associated SNPs, model of inheritance (e.g. dominant, recessive) and the minor allele frequency. The Wellcome Trust Case Control Consortium GWAS estimated power of 80% to detect SNPs with minor allele frequencies (MAFs) > 5% and OR = 1·5 using 2000 cases and 3000 controls [92]. Rare alleles with important effects may be missed, even in large studies, particularly as more than half of SNPs in the human genome are estimated to have MAFs < 5% [93]. Furthermore, structural variation, which may also account for a large proportion of human genetic variation, is not well captured by the first generation SNP arrays used in recent GWAS, which tag mainly common haplotypes [94,95].

The first GWAS in coeliac disease tested over 300 000 SNPs in 778 UK coeliac cases and 1422 controls [2]. This study confirmed the known association of coeliac disease with the HLA region, with the strongest association at a SNP tagging HLA DQ2·5 *cis.* There was weak evidence of association in the previously reported *CD28–CTLA4–ICOS* region ($P = 0{\cdot}007$), but not the *MYO9B* region.

Table 3. Non-human leucocyte antigen (HLA) susceptibility loci for coeliac disease from recent Genome-Wide Association Study [2,3].

Locus	Tag SNP with strongest association	Odds ratio (CI)	Candidate genes	Other diseases associated with the same region
4q27	rs6822844	0.71 (0.63–0.80)	*IL2*, IL21	Type 1 diabetes, rheumatoid arthritis, Graves' disease, psoriasis
1q31	rs2816316	0.71 (0.63–0.80)	*RGS1*	
2q11–2q12	rs917997	1.27 (1.15–1.40)	*IL1RL1*, *IL18R1*, *IL18RAP*, SLC9A4	Crohn's disease
3p21	rs6441961	1.21 (1.10–1.32)	*CCR1*, *CCR2*, *CCRL2, CCR3*, *CCR5*, XCR1	Type 1 diabetes
3q25–3q26	rs17810546	1.34(1.19–1.51)	*IL12A*, SCHIP1	
3q28	rs1465150	1.21 (1.11–1.31)	*LPP*	
6q25	rs1738074	1.21 (1.11–1.31)	*TAGAP*	
12q24	rs653178	1.19 (1.10–1.30)	*SH2B3*	Type 1 diabetes

CI, confidence interval; SNP, single nucleotide polymorphism.

New coeliac disease genes

IL2–IL21 region

Outside HLA, the strongest marker from the recent UK coeliac disease GWAS mapped to chromosome 4q27 ($P = 2 \times 10^{-7}$), a finding replicated in further UK, Dutch and Irish cohorts [3]. The associated SNP tags a ~700 kb LD block encompassing four genes (*ADAD1*, *KIAA1109*, *IL2* and *IL21*), such that variants in any of these genes could explain the genetic association. This region is emerging from other studies as a common autoimmune disease locus (see below). The most compelling biological candidates within the LD block are *IL2* and *IL21*.

Interleukin-2 and IL-21 are members of a cytokine family, sharing the same γ chain subunit in their receptors [96]. These cytokines have multiple and diverse roles in the immune response, posing a challenge in identifying the precise biological mechanisms relevant to coeliac disease. IL-2 has a well-defined autocrine function in stimulating T cell activation and proliferation, but can also stimulate NK cell proliferation and immunoglobulin production from B cells. This cytokine has a unique role in activation-induced cell death, a process that eliminates self-reactive T cells, and in maintenance of $CD4^+$ $CD25^+$ regulatory T (T_{reg}) cells [97–99]. In the non-obese diabetic mouse model the region syntenic to human 4q27 determines susceptibility to multiple autoimmune diseases through an *IL2*-dependent mechanism [100]. In this model, the murine risk variants were associated with reduced *IL2* gene expression, lower proportions of $CD4^+$ $CD25^+$ T_{reg} cells in mesenteric lymph nodes and impaired function of these cells [100]. It is thus possible that the *IL2–IL21* region risk variants in human coeliac disease might also exert their susceptibility effects through the $CD4^+$ $CD25^+$ T_{reg} cell subset, for example by impairing tolerance to gluten peptides. However, in humans, there are as yet no comparable data of the effects of variants on gene expression or function. *IL21* remains a candidate gene in this region and expression is known to be increased in the small intestinal mucosa in untreated coeliac disease [101]. IL-21 is secreted mainly from $CD4^+$ T cells and has proinflammatory effects including enhancement of B, T and NK cell proliferation [102]. Anti-IL-21 antibodies in an *ex-vivo* intestinal biopsy culture model reduced T-bet and IFN-γ expression, suggesting that IL-21 may be important in sustaining Th1 activity in coeliac disease [101].

The follow-up study from the first coeliac GWAS was reported recently [3]. This tested over 1000 of the most strongly associated non-HLA SNPs from the original UK GWAS in a large independent cohort (1643 new coeliac cases and 3406 controls). The added power of this study yielded strong, genome-wide significant results ($P < 5 \times 10^{-7}$) for a further seven new genomic regions, six of which harbour genes with immune functions (Table 3, Fig. 2). It was estimated in this follow-up study that the newly identified variants account for only 3–4% of the genetic susceptibility of coeliac disease, suggesting that many other true associations remain undetected. Effect sizes of the SNPs on disease susceptibility are modest, in line with findings from GWAS in other complex diseases (Fig. 3) [103,104]. The allele that is more frequent in cases can confer either protective or risk effects with ORs of all detected variants between 0·7 and 1·4. Given that there are an estimated 8 million SNPs with MAF > 5% in the human genome and only 300 000 SNPs were tested in the original GWAS, in most cases associated SNPs are unlikely to be causal, but instead will show variable levels of correlation with the true causal variants. Identification of the true causal variants is a priority of further research and will depend on fine-mapping and/or deep resequencing of the regions identified. Indications from other diseases suggest that discovery of the true causal variants may lead to a significant upwards revision of both effect sizes and the estimated proportion of genetic susceptibility accounted for [103]. In the interim, the primary significance of the GWAS study findings is in providing new insights into the biological pathways relevant to the pathogenesis of coeliac disease.

RGS1 region

The strongest association ($P = 2{\cdot}58 \times 10^{11}$) outside the HLA region and *IL2–IL21* was for a SNP 8 kb distal to the 5′ end

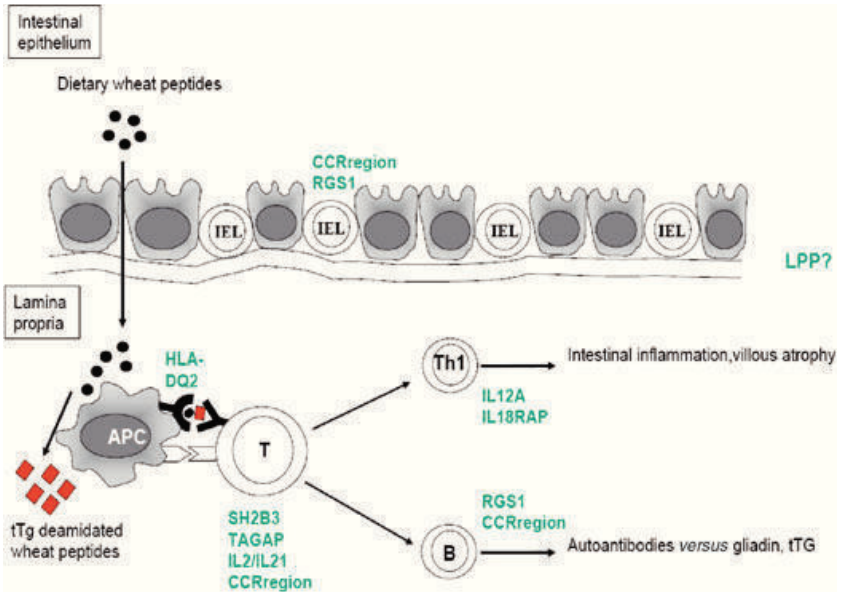


Fig. 2. Model of gluten induced immune response in coeliac disease, and the sites of action of coeliac susceptibility genes. The most likely gene from each region is shown, although note that causality of a genetic variant in any one gene has not yet been proved.

of *RGS1*. *RGS1* is of particular interest in coeliac disease because of its selective expression in the intestinal IEL compartment, but not conventional splenic or thymic T cells [3,105]. *RGS1* regulates G protein signalling activity and is implicated in mice in regulating chemokine receptor signalling and B cell trafficking to lymph nodes [106].

3p21

Another strong association mapped to a chemokine receptor gene cluster on 3p21 including *CCR1*, *CCR2*, *CCRL2*, *CCR3*, *CCR5* and *XCR1*, again hinting at the importance that chemokine receptor signalling and recruitment of effector immune cells to sites of inflammation may have in coeliac disease. The disease associated genetic variants may influence these pathways subtly.

IL12A and *IL18RAP*

Strong association ($P = 10^{-9}$) of SNPs in a 70 Kb LD block immediately 5′ of *IL12A* implicate this gene, which encodes IL12p35, the subunit that forms one-half of the IL-12 heterodimer with IL-12p40. IL-12 is expressed by antigen-presenting cells and has a broad range of biological activities, including induction of IFN-γ-secreting Th1 cells. Although coeliac disease is characterized by a strong Th1 response, surprisingly IL12p40 is not expressed in coeliac disease mucosa after gluten challenge and both IL-12p40 and IL-12p35 expression were not found to be increased in dendritic cells isolated from untreated coeliac disease mucosa [53,55]. It might well be in coeliac disease that IL-12 signalling is important at an alternative site (e.g. mesenteric lymph nodes) – attempting to make sense of these findings really

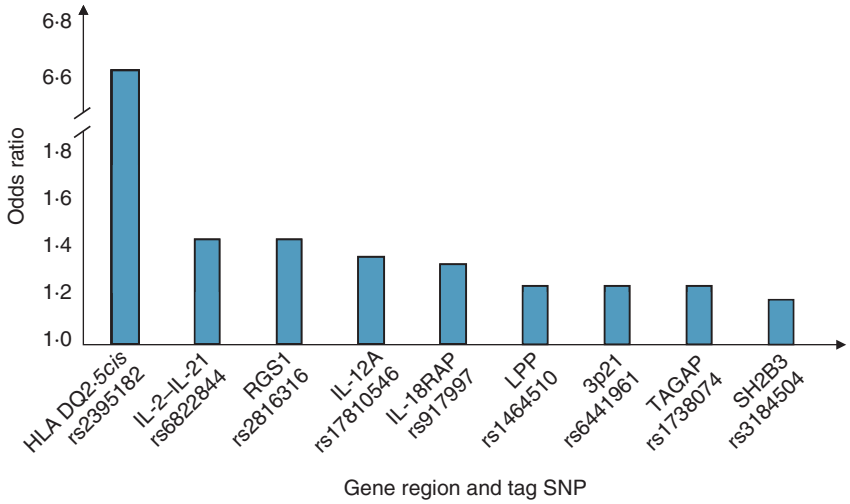


Fig. 3. Current estimates of effect size conferred by the coeliac disease-associated risk variants. Allelic odds ratios are shown for the best tag markers from the Genome-Wide Association Study, along with the most likely candidate gene(s) from each region. It is probable that the effect of the true causal variants, once identified, will be larger.

highlights our limited knowledge of the primary underlying immunopathogenic mechanisms.

There is evidence for the importance of IFN-α and IL-18 in promoting a Th1 phenotype in CD4 T cells in coeliac disease (see above). *IL18* transcripts are expressed very strongly in the human small intestine. In this regard, another candidate gene identified from the GWAS (*IL18RAP*) encodes the β chain of the IL-18 receptor. Hunt *et al.* showed that the coeliac disease associated SNPs correlated with *IL18RAP* gene expression in peripheral blood. The risk alleles, found more commonly in individuals with coeliac disease, correlated with lower levels of *IL18RAP* mRNA suggesting that variants reduce gene expression. This might suggest a loss of function of IL-18 receptor signalling, a puzzling finding given the up-regulation of IL-18 and strong Th1 bias in coeliac disease. Again, these findings underline the limitations of current immunological models of coeliac and other immune-mediated diseases, but also provide clues to inform the design of new functional studies.

SH2B3 region

SH2B3 is expressed in immune cells, up-regulated in coeliac mucosa and thought to function in regulation of T cell receptor, growth factor and cytokine receptor-mediated signalling [107,108]. A non-synonymous SNP (rs3184504) in *SH2B3* was associated with coeliac disease in the follow-up study. The same SNP is associated with type 1 diabetes, accounting entirely for the association in the latter disease [104]. This SNP, in exon 3 of *SH2B3*, leads to an amino acid substitution (R262W) in the pleckstrin homology (PH) domain of the SH2B3 protein. PH domains are involved in targeting proteins to plasma membranes through binding phosphoinositides [109]. Mutations in PH domains in other proteins have been associated with disease by impairing phosphoinositide binding and membrane localization (X-linked agammaglobulinaemia) or through causing constitutive membrane association (breast, colorectal and ovarian cancers) [110,111]. Functional studies of the effects of the R262W variant are needed to determine how this impacts on the biology of coeliac disease.

TAGAP and *LPP*

T cell activation GTPase activating protein-*TAGAP* is a gene expressed in activated T cells, whose function in immune cells is not well characterized but may modulate cytoskeletal changes [112]. *LPP* is strongly expressed in the small intestine but the significance in relation to coeliac disease is unknown.

Some coeliac disease-associated regions also influence susceptibility to other chronic immune-mediated conditions

An unexpected finding from the recent coeliac disease genetic studies was the identification of gene regions that have been associated with other chronic immune-mediated conditions (Table 3). Coeliac disease is associated with an increased prevalence of several autoimmune conditions, including type 1 diabetes, autoimmune thyroid disease and rheumatoid arthritis [113]. Comparison of GWAS data sets of coeliac disease and autoimmune diseases implicate a novel shared disease association between coeliac disease and type 1 diabetes in the *SH2B3* gene, 3p21 *CCR* gene region and *IL2–IL21* region, whereas variants in the *IL18RAP* region have also been identified in Crohn's disease [92,104]. *IL2–IL21* variants have been associated with Graves' disease, rheumatoid arthritis and psoriasis in addition to type 1 diabetes, suggesting that this may be a common autoimmune disease locus [104,114,115]. The association of at least four independent gene regions with both type 1 diabetes and coeliac disease (*HLA DQ*, *IL2–IL21*, *SH2B3* and *CCR* region) is particularly striking and points to shared mechanisms in the immunopathogenesis of these two conditions. These genes all have putative roles in $CD4^+$ T cell activation or recruitment, reinforcing the central importance of this cell in both diseases. Type 1 diabetes and coeliac disease have both shown rising incidence in recent years, with tantalizing, although still inconclusive, evidence for the role of early childhood intestinal infections, particularly rotavirus [19–21,116]. Thus a model emerges in which common genetic and environmental factors might drive a shared type 1 diabetes/coeliac predisposition, with further disease-specific genes or environmental factors biasing individuals towards one or both diseases.

Therapeutic prospects arising from coeliac gene discovery

Human leucocyte antigen-DQ remains the only coeliac disease locus in which the causal variants are known and their contribution to disease pathogenesis is understood (Box 1). Identification of causal variants within the new coeliac disease regions and functional investigation of these new candidate genes is a priority for future research. Therapeutic manipulation of the pathways identified in these studies may also prove fruitful. Despite modest effect sizes of the genes identified, more profound modulation of function can have important clinical benefits. In type 2 diabetes, where genetic variants in peroxisome proliferator-activated receptor-γ (*PPARG*) confer modest susceptibility (OR 1·1), thiazoledinediones, which act as agonists of PPAR-γ, have significant clinical benefit [117]. A variant in the ATP-sensitive potassium channel (*KCNJ11*), which is the pharmacological target of another class of type 2 diabetes medication (sulphonylureas), again shows a modest susceptibility effect, with a heterozygote *versus* homozygote ORs of only 1·1 [118]. Perhaps the most exciting prospect, given that a safe and effective treatment for coeliac disease already exists, is the possibility that these genes may reveal pathways that can be exploited for long-lasting immunomodulation in the prevention of coeliac and other

Box 1

Potential immunopathogenic functions of newly identified coeliac susceptibility genes

Newly identified coeliac susceptibility gene regions implicate pathways involved in both innate and adaptive immune responses

Chemokine signalling – two gene regions (*CCR* region and *RGS1*) have known roles in chemokine signalling, suggesting that mechanisms of immune cell recruitment to the intestinal epithelium/mucosa have a significance that has not been emphasized previously in immunological models. *RGS1* is expressed selectively in intra-epithelial lymphocytes, and can influence lymphocyte trafficking, possibly providing insight into why the intra-epithelial lymphocyte population is expanded in coeliac disease

T cell activation and differentiation – gluten peptide presentation via DQ2/8 activates a clinically significant $CD4^+$ T cell response in coeliac disease, but not in non-coeliac $DQ2/8^+$ individuals. Several of the new susceptibility genes have roles in T cell activation (*IL2–IL21*, *TAGAP*, *SH2B3*), T_{reg} function (*IL2*) and Th1 differentiation (*IL18RAP*, *IL12A*). These gene variants may subtly influence the outcome of antigen presenting cell–T cell gluten peptide interactions, shifting the balance towards gluten-specific effector Th1 cells (as in coeliac disease) rather than tolerance to gluten (non-coeliac DQ2/8 controls)

Box 2

Future priorities for coeliac disease immunogenetic research

1. Identification of additional susceptibility loci
   a. Extending current GWAS with larger sample sizes and study meta-analysis
   b. Development of high throughput methods to detect structural variants
2. Identification of causal variants in susceptibility regions
   a. Fine-mapping of associations with high-density SNPs
   b. Resequencing of involved areas to identify rare and causal variants
   c. Use of biological information to implicate causality (e.g. SNP/gene expression correlation studies, gene knockdown studies)
3. Detailed functional studies of new coeliac gene variants to understand molecular and physiological roles in disease pathogenesis
4. Development of an animal model to test immunological hypotheses, including initial responses to gluten and factors determining tolerance
5. Development of new diagnostic tests
6. Use of newly defined targets to develop therapeutic strategies

related immune-mediated conditions such as type 1 diabetes. Any such strategies must be safe, with minimal toxicity.

Concluding remarks

Our understanding of the immunogenetic pathogenesis of coeliac disease is well advanced in comparison with most other chronic immune-mediated conditions. The antigen (gluten) and many of the immunodominant epitopes that drive T cell responses in coeliac disease have been identified. The role of tTG in enhancing the immunogenicity of gluten peptides by deamidation of glutamine residues is known. The major causal variants in the HLA region are identified and this has led to functional understanding of how these molecules select and present immunogenic gluten peptides. Other aspects of the mucosal immune response in coeliac disease have been characterized, including the roles of IEL and non-T cell receptor-dependent mechanisms of gluten toxicity.

Genome-wide association studies are now rapidly adding information on primary genetic predisposing factors in coeliac disease, with eight new loci now identified, seven of which contain genes influencing immune function. Thus the genetic factors are directly relevant to, and may guide further, our immunological understanding of the disease (Box 2). Several of the coeliac risk loci are also implicated in type 1 diabetes, suggesting far greater similarity in the immunopathogenesis of these conditions than suspected previously. These new loci promise to provide insights into why not all individuals with HLA-DQ2 or DQ8 develop coeliac disease and point to factors that subtly modulate T cell activation and effector cell (Th1) differentiation. The precise causal variants remain to be determined, but their identification and functional studies will, in time, provide further insights into the pathogenesis of coeliac disease and related immune-mediated conditions.

Acknowledgements

We thank Professor Cisca Wijmenga for the concept of Figure 2. Dr P. C. Dubois holds an MRC Clinical Research Training Fellowship.

References

1 van Heel DA, West J. Recent advances in coeliac disease. Gut 2006; **55**:1037–46.
2 van Heel DA, Franke L, Hunt KA *et al*. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat Genet 2007; **39**:827–9.
3 Hunt KA, Zhernakova A, Turner G *et al*. Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 2008.
4 West J, Logan RF, Hill PG *et al*. Seroprevalence, correlates, and characteristics of undetected coeliac disease in England. Gut 2003; **52**:960–5.
5 Accomando S, Cataldo F. The global village of celiac disease. Dig Liver Dis 2004; **36**:492–8.
6 Freeman HJ. Biopsy-defined adult celiac disease in Asian-Canadians. Can J Gastroenterol 2003; **17**:433–6.
7 Bonamico M, Mariani P, Triglione P *et al*. Celiac disease in two sisters with a mother from Cape Verde Island, Africa: a clinical and genetic study. J Pediatr Gastroenterol Nutr 1994; **18**:96–9.

8 Fasano A, Berti I, Gerarduzzi T *et al*. Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study. Arch Intern Med 2003; **163**:286–92.
9 Cataldo F, Lio D, Simpore J, Musumeci S. Consumption of wheat foodstuffs not a risk for celiac disease occurrence in Burkina Faso. J Pediatr Gastroenterol Nutr 2002; **35**:233–4.
10 Catassi C, Doloretta Macis M, Ratsch IM, De Virgiliis S, Cucca F. The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence. Tissue Antigens 2001; **58**:402–6.
11 Fasano A, Catassi C. Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum. Gastroenterology 2001; **120**:636–51.
12 Weile B, Cavell B, Nivenius K, Krasilnikoff PA. Striking differences in the incidence of childhood celiac disease between Denmark and Sweden: a plausible explanation. J Pediatr Gastroenterol Nutr 1995; **21**:64–8.
13 Mitt K, Uibo O. Low cereal intake in Estonian infants: the possible explanation for the low frequency of coeliac disease in Estonia. Eur J Clin Nutr 1998; **52**:85–8.
14 Lohi S, Mustalahti K, Kaukinen K *et al*. Increasing prevalence of coeliac disease over time. Aliment Pharmacol Ther 2007; **26**:1217–25.
15 EURODIAB ACE Study Group. Variation and trends in incidence of childhood diabetes in Europe. Lancet 2000; **355**:873–6.
16 Ivarsson A, Hernell O, Stenlund H, Persson LA. Breast-feeding protects against celiac disease. Am J Clin Nutr 2002; **75**:914–21.
17 Ientile R, Caccamo D, Griffin M. Tissue transglutaminase and the stress response. Amino Acids 2007; **33**:385–94.
18 Ivarsson A, Hernell O, Nystrom L, Persson LA. Children born in the summer have increased risk for coeliac disease. J Epidemiol Commun Health 2003; **57**:36–9.
19 Ivarsson A. The Swedish epidemic of coeliac disease explored using an epidemiological approach – some lessons to be learnt. Best Pract Res 2005; **19**:425–40.
20 Sandberg-Bennich S, Dahlquist G, Kallen B. Coeliac disease is associated with intrauterine growth and neonatal infections. Acta Paediatr 2002; **91**:30–3.
21 Stene LC, Honeyman MC, Hoffenberg EJ *et al*. Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study. Am J Gastroenterol 2006; **101**:2333–40.
22 Matysiak-Budnik T, Malamut G, de Serre NP *et al*. Long-term follow-up of 61 coeliac patients diagnosed in childhood: evolution toward latency is possible on a normal diet. Gut 2007; **56**:1379–86.
23 Simell S, Hoppu S, Hekkala A *et al*. Fate of five celiac disease-associated antibodies during normal diet in genetically at-risk children observed from birth in a natural history study. Am J Gastroenterol 2007; **102**:2026–35.
24 Greco L, Romino R, Coto I *et al*. The first large population based twin study of coeliac disease. Gut 2002; **50**:624–8.
25 Petronzelli F, Bonamico M, Ferrante P *et al*. Genetic contribution of the HLA region to the familial clustering of coeliac disease. Ann Hum Genet 1997; **61**:307–17.
26 Bevan S, Popat S, Braegger CP *et al*. Contribution of the MHC region to the familial risk of coeliac disease. J Med Genet 1999; **36**:687–90.
27 Risch N. Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 1987; **40**:1–14.
28 Lewis CM, Whitwell SC, Forbes A, Sanderson J, Mathew CG, Marteau TM. Estimating risks of common complex diseases across genetic and environmental factors: the example of Crohn disease. J Med Genet 2007; **44**:689–94.
29 Robinson J, Waller MJ, Parham P *et al*. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res 2003; **31**:311–14.
30 Falchuk ZM, Rogentine GN, Strober W. Predominance of histocompatibility antigen HL-A8 in patients with gluten-sensitive enteropathy. J Clin Invest 1972; **51**:1602–5.
31 Keuning JJ, Pena AS, van Leeuwen A, van Hooff JP, van Rood JJ. HLA-DW3 associated with coeliac disease. Lancet 1976; **1**:506–8.
32 Price P, Witt C, Allcock R *et al*. The genetic basis for the association of the 8.1 ancestral haplotype (A1, B8, DR3) with multiple immunopathological diseases. Immunol Rev 1999; **167**:257–74.
33 Tosi R, Vismara D, Tanigaki N *et al*. Evidence that celiac disease is primarily associated with a DC locus allelic specificity. Clin Immunol Immunopathol 1983; **28**:395–404.
34 van Heel DA, Hunt K, Greco L, Wijmenga C. Genetics in coeliac disease. Best Pract Res 2005; **19**:323–39.
35 Vader W, Stepniak D, Kooy Y *et al*. The HLA-DQ2 gene dose effect in celiac disease is directly related to the magnitude and breadth of gluten-specific T cell responses. Proc Natl Acad Sci USA 2003; **100**:12390–5.
36 Sollid LM. Molecular basis of celiac disease. Annu Rev Immunol 2000; **18**:53–81.
37 Louka AS, Nilsson S, Olsson M *et al*. HLA in coeliac disease families: a novel test of risk modification by the 'other' haplotype when at least one DQA1*05-DQB1*02 haplotype is carried. Tissue Antigens 2002; **60**:147–54.
38 van Belzen MJ, Koeleman BP, Crusius JB *et al*. Defining the contribution of the HLA region to cis DQ2-positive coeliac disease patients. Genes Immun 2004; **5**:215–20.
39 Sollid LM, Markussen G, Ek J, Gjerde H, Vartdal F, Thorsby E. Evidence for a primary association of celiac disease to a particular HLA-DQ alpha/beta heterodimer. J Exp Med 1989; **169**:345–50.
40 Spurkland A, Sollid LM, Polanco I, Vartdal F, Thorsby E. HLA-DR and -DQ genotypes of celiac disease patients serologically typed to be non DR3 or non-DR5/7. Hum Immunol 1992; **35**:188–92.
41 Karell K, Louka AS, Moodie SJ *et al*. 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 2003; **64**:469–77.
42 Louka AS, Sollid LM. HLA in coeliac disease: unravelling the complex genetics of a complex disorder. Tissue Antigens 2003; **61**:105–17.
43 Louka AS, Moodie SJ, Karell K *et al*. A collaborative European search for non-DQA1*05–DQB1*02 celiac disease loci on HLA-DR3 haplotypes: analysis of transmission from homozygous parents. Hum Immunol 2003; **64**:350–8.
44 Greco L, Babron MC, Corazza GR *et al*. Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families. Ann Hum Genet 2001; **65**:35–41.
45 Greco L, Corazza G, Babron MC *et al*. Genome search in celiac disease. Am J Hum Genet 1998; **62**:669–75.
46 Van Belzen MJ, Meijer JW, Sandkuijl LA *et al*. A major non-HLA locus in celiac disease maps to chromosome 19. Gastroenterology 2003; **125**:1032–41.
47 Djilali-Saiah I, Schmitz J, Harfouch-Hammoud E CTLA-4 gene polymorphism is associated with predisposition to coeliac disease. Gut 1998; **43**:187–9.
48 Monsuur AJ, de Bakker PI, Alizadeh BZ *et al*. Myosin IXB variant

increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nat Genet 2005; **37**:1341–4.
49 Hunt KA, Monsuur AJ, McArdle WL *et al.* Lack of association of MYO9B genetic variants with coeliac disease in a British cohort. Gut 2006; **55**:969–72.
50 Amundsen SS, Monsuur AJ, Wapenaar MC *et al.* Association analysis of MYO9B gene polymorphisms with celiac disease in a Swedish/Norwegian cohort. Hum Immunol 2006; **67**:341–5.
51 Sanchez E, Alizadeh BZ, Valdigem G *et al.* MYO9B gene polymorphisms are associated with autoimmune diseases in Spanish population. Hum Immunol 2007; **68**:610–15.
52 Anand BS, Piris J, Jerrome DW, Offord RE, Truelove SC. The timing of histological damage following a single challenge with gluten in treated coeliac disease. Q J Med 1981; **50**:83–94.
53 Nilsen EM, Jahnsen FL, Lundin KE *et al.* Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. Gastroenterology 1998; **115**:551–63.
54 Monteleone I, Monteleone G, Vecchio Blanco DG *et al.* Regulation of the T helper cell type 1 transcription factor T-bet in coeliac disease mucosa. Gut 2004; **53**:1090–5.
55 Di Sabatino A, Pickard KM, Gordon JN *et al.* Evidence for the role of interferon-alfa production by dendritic cells in the Th1 response in celiac disease. Gastroenterology 2007; **133**:1175–87.
56 Salvati VM, MacDonald TT, Bajaj-Elliott M *et al.* Interleukin 18 and associated markers of T helper cell type 1 activity in coeliac disease. Gut 2002; **50**:186–90.
57 Monteleone G, Pender SL, Alstead E *et al.* Role of interferon alpha in promoting T helper cell type 1 responses in the small intestine in coeliac disease. Gut 2001; **48**:425–9.
58 Leon AJ, Garrote JA, Blanco-Quiros A *et al.* Interleukin 18 maintains a long-standing inflammation in coeliac disease patients. Clin Exp Immunol 2006; **146**:479–85.
59 Raki M, Tollefsen S, Molberg O, Lundin KE, Sollid LM, Jahnsen FL. A unique dendritic cell subset accumulates in the celiac lesion and efficiently activates gluten-reactive T cells. Gastroenterology 2006; **131**:428–38.
60 Lundin KE, Scott H, Hansen T *et al.* 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 1993; **178**:187–96.
61 Lundin KE, Scott H, Fausa O, Thorsby E, Sollid LM. T cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8. Hum Immunol 1994; **41**:285–91.
62 Arentz-Hansen H, Korner R, Molberg O *et al.* The intestinal T cell response to alpha-gliadin in adult celiac disease is focused on a single deamidated glutamine targeted by tissue transglutaminase. J Exp Med 2000; **191**:603–12.
63 Anderson RP, Degano P, Godkin AJ, Jewell DP, Hill AV. *In vivo* antigen challenge in celiac disease identifies a single transglutaminase-modified peptide as the dominant A-gliadin T-cell epitope. Nat Med 2000; **6**:337–42.
64 Anderson RP, van Heel DA, Tye-Din JA *et al.* T cells in peripheral blood after gluten challenge in coeliac disease. Gut 2005; **54**:1217–23.
65 Arentz-Hansen H, McAdam SN, Molberg O *et al.* Celiac lesion T cells recognize epitopes that cluster in regions of gliadins rich in proline residues. Gastroenterology 2002; **123**:803–9.
66 Vader W, Kooy Y, Van Veelen P *et al.* The gluten response in children with celiac disease is directed toward multiple gliadin and glutenin peptides. Gastroenterology 2002; **122**:1729–37.
67 Sjostrom H, Lundin KE, Molberg O *et al.* Identification of a gliadin T-cell epitope in coeliac disease: general importance of gliadin deamidation for intestinal T-cell recognition. Scand J Immunol 1998; **48**:111–15.
68 Tollefsen S, Arentz-Hansen H, Fleckenstein B *et al.* HLA-DQ2 and -DQ8 signatures of gluten T cell epitopes in celiac disease. J Clin Invest 2006; **116**:2226–36.
69 Vartdal F, Johansen BH, Friede T *et al.* The peptide binding motif of the disease associated HLA-DQ (alpha 1*0501, beta 1*0201) molecule. Eur J Immunol 1996; **26**:2764–72.
70 van de Wal Y, Kooy YM, Drijfhout JW *et al.* alpha 1*0201, beta 1*0202) molecule. Immunogenetics 1997; **46**:484–92.
71 Godkin A, Friede T, Davenport M *et al.* Use of eluted peptide sequence data to identify the binding characteristics of peptides to the insulin-dependent diabetes susceptibility allele HLA-DQ8 (DQ 3.2). Int Immunol 1997; **9**:905–11.
72 Kim CY, Quarsten H, Bergseng E, Khosla C, Sollid LM. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. Proc Natl Acad Sci USA 2004; **101**:4175–9.
73 van de Wal Y, Kooy Y, van Veelen P *et al.* Selective deamidation by tissue transglutaminase strongly enhances gliadin-specific T cell reactivity. J Immunol 1998; **161**:1585–8.
74 Dieterich W, Ehnis T, Bauer M *et al.* Identification of tissue transglutaminase as the autoantigen of celiac disease. Nat Med 1997; **3**:797–801.
75 Molberg O, McAdam SN, Korner R *et al.* Tissue transglutaminase selectively modifies gliadin peptides that are recognized by gut-derived T cells in celiac disease. Nat Med 1998; **4**:713–17.
76 Fleckenstein B, Molberg O, Qiao SW *et al.* Gliadin T cell epitope selection by tissue transglutaminase in celiac disease. Role of enzyme specificity and pH influence on the transamidation versus deamidation process. J Biol Chem 2002; **277**:34109–16.
77 Zanoni G, Navone R, Lunardi C *et al.* In celiac disease, a subset of autoantibodies against transglutaminase binds Toll-like receptor 4 and induces activation of monocytes. PLoS Med 2006; **3**:e358.
78 Myrsky E, Kaukinen K, Syrjanen M, Korponay-Szabo IR, Maki M, Lindfors K. Coeliac disease-specific autoantibodies targeted against transglutaminase 2 disturb angiogenesis. Clin Exp Immunol 2008; **152**:111–19.
79 Korponay-Szabo IR, Halttunen T, Szalai Z *et al. In vivo* targeting of intestinal and extraintestinal transglutaminase 2 by coeliac autoantibodies. Gut 2004; **53**:641–8.
80 Hadjivassiliou M, Maki M, Sanders DS *et al.* Autoantibody targeting of brain and intestinal transglutaminase in gluten ataxia. Neurology 2006; **66**:373–7.
81 Hausch F, Shan L, Santiago NA, Gray GM, Khosla C. Intestinal digestive resistance of immunodominant gliadin peptides. Am J Physiol 2002; **283**:G996–G1003.
82 Sturgess R, Day P, Ellis HJ *et al.* Wheat peptide challenge in coeliac disease. Lancet 1994; **343**:758–61.
83 Fraser JS, Engel W, Ellis HJ *et al.* Coeliac disease: *in vivo* toxicity of the putative immunodominant epitope. Gut 2003; **52**:1698–702.
84 Maiuri L, Picarelli A, Boirivant M *et al.* Definition of the initial immunologic modifications upon *in vitro* gliadin challenge in the small intestine of celiac patients. Gastroenterology 1996; **110**:1368–78.

85 Maiuri L, Ciacci C, Ricciardelli I *et al*. Association between innate response to gliadin and activation of pathogenic T cells in coeliac disease. Lancet 2003; **362**:30–7.
86 Di Sabatino A, Ciccocioppo R, Cupelli F *et al*. Epithelium derived interleukin 15 regulates intraepithelial lymphocyte Th1 cytokine production, cytotoxicity, and survival in coeliac disease. Gut 2006; **55**:469–77.
87 Mention JJ, Ben Ahmed M, Begue B *et al*. Interleukin 15: a key to disrupted intraepithelial lymphocyte homeostasis and lymphomagenesis in celiac disease. Gastroenterology 2003; **125**:730–45.
88 Jabri B, de Serre NP, Cellier C *et al*. Selective expansion of intraepithelial lymphocytes expressing the HLA-E-specific natural killer receptor CD94 in celiac disease. Gastroenterology 2000; **118**:867–79.
89 Meresse B, Curran SA, Ciszewski C *et al*. Reprogramming of CTLs into natural killer-like cells in celiac disease. J Exp Med 2006; **203**:1343–55.
90 Hue S, Mention JJ, Monteiro RC *et al*. A direct role for NKG2D/MICA interaction in villous atrophy during celiac disease. Immunity 2004; **21**:367–77.
91 Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev 2005; **6**:95–108.
92 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007; **447**:661–78.
93 Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 2008; **82**:100–12.
94 Korbel JO, Urban AE, Affourtit JP *et al*. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007; **318**:420–6.
95 Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet 2007; **3**:1787–99.
96 Waldmann TA. The biology of interleukin-2 and interleukin-15: implications for cancer therapy and vaccine design. Nat Rev 2006; **6**:595–601.
97 Lenardo MJ. Fas and the art of lymphocyte maintenance. J Exp Med 1996; **183**:721–4.
98 Fontenot JD, Rasmussen JP, Gavin MA, Rudensky AY. A function for interleukin 2 in Foxp3-expressing regulatory T cells. Nat Immunol 2005; **6**:1142–51.
99 Maloy KJ, Powrie F. Fueling regulation: IL-2 keeps CD4+ Treg cells fit. Nat Immunol 2005; **6**:1071–2.
100 Yamanouchi J, Rainbow D, Serra P *et al*. Interleukin-2 gene variation impairs regulatory T cell function and causes autoimmunity. Nat Genet 2007; **39**:329–37.
101 Fina D, Sarra M, Caruso R *et al*. Interleukin-21 contributes to the mucosal T helper cell type 1 response in celiac disease. Gut 2007; online early.
102 Leonard WJ, Spolski R. Interleukin-21: a modulator of lymphoid proliferation, apoptosis and differentiation. Nat Rev 2005; **5**:688–98.
103 Mathew CG. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat Rev Genet 2007; **9**:9–14.
104 Todd JA, Walker NM, Cooper JD *et al*. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007; **39**:857–64.
105 Pennington DJ, Silva-Santos B, Shires J *et al*. The inter-relatedness and interdependence of mouse T cell receptor gammadelta+ and alphabeta+ cells. Nat Immunol 2003; **4**:991–8.
106 Han SB, Moratz C, Huang NN *et al*. Rgs1 and Gnai2 regulate the entrance of B lymphocytes into lymph nodes and B cell motility within lymph node follicles. Immunity 2005; **22**:343–54.
107 Li Y, He X, Schembri-King J, Jakes S, Hayashi J. Cloning and characterization of human Lnk, an adaptor protein with pleckstrin homology and Src homology 2 domains that can inhibit T cell activation. J Immunol 2000; **164**:5199–206.
108 Velazquez L, Cheng AM, Fleming HE *et al*. Cytokine signaling and hematopoietic homeostasis are disrupted in Lnk-deficient mice. J Exp Med 2002; **195**:1599–611.
109 Lemmon MA. Membrane recognition by phospholipid-binding domains. Nat Rev Mol Cell Biol 2008; **9**:99–111.
110 Lindvall JM, Blomberg KE, Valiaho J *et al*. Bruton's tyrosine kinase: cell biology, sequence conservation, mutation spectrum, siRNA modifications, and expression profiling. Immunol Rev 2005; **203**:200–15.
111 Carpten JD, Faber AL, Horn C *et al*. A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 2007; **448**:439–44.
112 Mao M, Biery MC, Kobayashi SV *et al*. T lymphocyte activation gene identification by coregulated expression on DNA microarrays. Genomics 2004; **83**:989–99.
113 Farrell RJ, Kelly CP. Celiac sprue. N Engl J Med 2002; **346**:180–8.
114 Zhernakova A, Alizadeh BZ, Bevova M *et al*. 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 2007; **81**:1284–8.
115 Liu Y, Helms C, Liao W *et al*. A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci. PLoS Genet 2008; **4**:e1000041.
116 Ballotti S, de Martino M. Rotavirus infections and development of type 1 diabetes: an evasive conundrum. J Pediatr Gastroenterol Nutr 2007; **45**:147–56.
117 Altshuler D, Hirschhorn JN, Klannemark M *et al*. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 2000; **26**:76–80.
118 van Dam RM, Hoebee B, Seidell JC, Schaap MM, de Bruin TW, Feskens EJ. Common variants in the ATP-sensitive K+ channel genes KCNJ11 (Kir6.2) and ABCC8 (SUR1) in relation to glucose intolerance: population-based studies and meta-analyses. Diabet Med 2005; **22**:590–8.

nature genetics

ARTICLES

Multiple common variants for celiac disease influencing immune gene expression

Patrick C A Dubois[1,39*], Gosia Trynka[2,39], Lude Franke[1,2], Karen A Hunt[1], Jihane Romanos[2], Alessandra Curtotti[3], Alexandra Zhernakova[4], Graham A R Heap[1], Róza Ádány[5], Arpo Aromaa[6], Maria Teresa Bardella[7,8], Leonard H van den Berg[9], Nicholas A Bockett[1], Emilio G de la Concha[10], Bárbara Dema[10], Rudolf S N Fehrmann[2], Miguel Fernández-Arquero[10], Szilvia Fiatal[5,11], Elvira Grandone[12], Peter M Green[13], Harry J M Groen[14], Rhian Gwilliam[15], Roderick H J Houwen[16], Sarah E Hunt[15], Katri Kaukinen[17], Dermot Kelleher[18], Ilma Korponay-Szabo[19,20], Kalle Kurppa[17], Padraic MacMathuna[21], Markku Mäki[17], Maria Cristina Mazzilli[22], Owen T McCann[15], M Luisa Mearin[23], Charles A Mein[3], Muddassar M Mirza[13], Vanisha Mistry[1], Barbara Mora[22], Katherine I Morley[15], Chris J Mulder[24], Joseph A Murray[25], Concepción Núñez[10], Elvira Oosterom[2], Roel A Ophoff[26–28], Isabel Polanco[29], Leena Peltonen[15,30], Mathieu Platteel[2], Anna Rybak[31], Veikko Salomaa[6], Joachim J Schweizer[23], Maria Pia Sperandeo[32], Greetje J Tack[24], Graham Turner[18], Jan H Veldink[9], Wieke H M Verbeek[24], Rinse K Weersma[33], Victorien M Wolters[16], Elena Urcelay[10], Bozena Cukrowska[34], Luigi Greco[32], Susan L Neuhausen[35], Ross McManus[18], Donatella Barisani[36], Panos Deloukas[15], Jeffrey C Barrett[15], Paivi Saavalainen[37,38], Cisca Wijmenga[2] & David A van Heel[1]

We performed a second-generation genome-wide association study of 4,533 individuals with celiac disease (cases) and 10,750 control subjects. We genotyped 113 selected SNPs with $P_{GWAS} < 10^{-4}$ and 18 SNPs from 14 known loci in a further 4,918 cases and 5,684 controls. Variants from 13 new regions reached genome-wide significance ($P_{combined} < 5 \times 10^{-8}$); most contain genes with immune functions (*BACH2, CCR4, CD80, CIITA-SOCS1-CLEC16A, ICOSLG* and *ZMIZ1*), with *ETS1, RUNX3, THEMIS* and *TNFRSF14* having key roles in thymic T-cell selection. There was evidence to suggest associations for a further 13 regions. In an expression quantitative trait meta-analysis of 1,469 whole blood samples, 20 of 38 (52.6%) tested loci had celiac risk variants correlated ($P < 0.0028$, FDR 5%) with *cis* gene expression.

Celiac disease is a common heritable chronic inflammatory condition of the small intestine induced by dietary wheat, rye and barley, as well as other unidentified environmental factors, in susceptible individuals. Specific *HLA-DQA1* and *HLA-DQB1* risk alleles are necessary, but not sufficient, for disease development[1,2]. The well-defined role of HLA-DQ heterodimers encoded by these alleles is to present cereal peptides to $CD4^+$ T cells, activating an inflammatory immune response in the intestine. A single genome-wide association study (GWAS) has been performed in celiac disease, which identified the *IL2-IL21* risk locus[1]. Subsequent studies probing the GWAS information in greater depth have identified a further 12 risk regions. Most of these regions contain a candidate gene that functions in the immune system, although only in the case of *HLA-DQA1* and *HLA-DQB1* have the causal variants been established[3–5]. Many of the known celiac disease–associated loci overlap with those of other immune-related diseases[6]. To identify additional risk variants, particularly those with smaller effect sizes, we performed a second-generation GWAS using more than six times as many samples as the previous GWAS and a denser genome-wide SNP set. We followed up promising findings in a large collection of independent samples.

RESULTS

Overview of study design

The GWAS included five European celiac disease case and control sample collections, including the celiac disease dataset reported previously[1]. We performed stringent data quality control (see Online Methods), including calling genotypes using a custom algorithm on both large sample sets and, where possible, cases and controls together (see Online Methods). We tested 292,387 non-*HLA* SNPs from the Illumina Hap300 marker set for association in 4,533 individuals with celiac disease and 10,750 control subjects of European descent (**Table 1**). A further 231,362 additional non-*HLA* markers from the Illumina Hap550 marker set were tested for association in a subset of 3,796 individuals with celiac disease and 8,154 controls. All markers

*A full list of author affiliations appears at the end of the paper.

Received 4 December 2009; accepted 4 February 2010; published online 28 February 2010; corrected online 12 March 2010; doi:10.1038/ng.543

Table 1 Sample collections and genotyping platforms

		Celiac disease cases			Controls		
Collection	Country	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]	Sample size (pre-QC)[a]	Sample size (post-QC)[b]	Platform[c]
Stage 1: Genome-wide association							
1[d,e]	UK	778	737	Illumina Hap300v1-1	2,596[i]	2,596	Illumina Hap550-2v3
2[d,f]	UK	1,922	1,849	Illumina 670-QuadCustom_v1	5,069[i]	4,936	Illumina 1.2M-DuoCustom_v1
3[d]	Finland	674	647	Illumina 670-QuadCustom_v1	1,839[i]	1,829	Illumina 610-Quad
4[g]	The Netherlands	876	803	Illumina 670-QuadCustom_v1	960	846	Illumina 670-QuadCustom_v1
5[d]	Italy	541	497	Illumina 670-QuadCustom_v1	580	543	Illumina 670-QuadCustom_v1
Analysis of Hap300 markers[c]			4,533			10,750	
Analysis of additional Hap550 markers[c]			3,796			8,154	
Stage 2: Follow-up							
6	USA	987	973	Illumina GoldenGate	615	555	Illumina GoldenGate
7	Hungary	979	965	Illumina GoldenGate	1,126	1,067	Illumina GoldenGate
8[h]	Ireland	653	597	Illumina GoldenGate	1,499	1,456	Illumina GoldenGate
9	Poland	599	564	Illumina GoldenGate	745	716	Illumina GoldenGate
10	Spain	558	550	Illumina GoldenGate	465	433	Illumina GoldenGate
11[d]	Italy	1,056	1,010	Illumina GoldenGate	864	804	Illumina GoldenGate
12[d]	Finland	270	259	Illumina GoldenGate	653[j]	653	Illumina 610-Quad[j]
Subtotal			4,918			5,684	
Analysis of Hap300 markers and follow-up (91 SNPs)[c]			9,451			16,434	
Analysis of additional Hap550 markers and follow-up (40 SNPs)[c]			8,714			13,838	

[a]Sample numbers attempted for genotyping, before any quality control (QC) steps were applied. [b]Sample numbers after all quality control (QC) steps (used in the association analysis). [c]All platforms contain a common set of Hap300 markers; the Hap550, 610-Quad, 670-Quad and 1.2M contain a common set of Hap550 markers. [d]As an additional quality control step, we performed case-case and control-control comparisons for collection 1 versus 2, and collection 3 versus 12, for the 40 SNPs in **Table 2** and observed no markers with $P < 0.01$. We did observe (as expected) differences for collection 5 versus 11, from northern and southern Italy, respectively. [e]All 737 post-QC cases reported in a previous GWAS[1]. [f]690 of the post-QC cases and 1,150 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [g]498 of the post-QC cases and 767 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [h]352 of the post-QC cases and 921 of the post-QC controls were included in previous GWAS follow-up studies[22,32]. [i]Some of these data were generated elsewhere, and some prior quality control steps (information not available) had been applied. [j]Finnish stage 2 controls were individuals within the Finrisk collection for whom Illumina 610-Quad genotype data became available after the completion of stage 1.

were from autosomes or the X chromosome. Genotype call rates were >99.9% in both datasets. The overdispersion factor of association test statistics, $\lambda_{GC} = 1.12$, was similar to that observed in other GWASs of this sample size[7,8]. Findings were not substantially altered by imputation of missing genotypes for 737 cases with celiac disease genotyped on the Hap300 BeadChip and corresponding controls (**Table 1**, collection 1). Here we present results for directly genotyped SNPs, as around half the additional Hap550 markers cannot be accurately imputed from Hap300 data[9] (including the new *ETS1* locus reported in this study). Results for the top 1,000 markers are available in **Supplementary Data 1**; however, because of concerns regarding the detection of individuals' identities[10], results for all markers are available only on request to the corresponding author.

For follow-up, we first inspected genotype clouds for the 417 non-*HLA* SNPs that met $P_{GWAS} < 10^{-4}$, being aware that top GWAS signals might be enriched for genotyping artifact, and excluded 22 SNPs from further analysis using a low threshold for possible bias. We selected SNPs from 113 loci for replication. Markers that passed design and genotyping quality control included (i) 18 SNPs from all 14 previously identified celiac disease risk loci (including a tag SNP for the major celiac disease–associated *HLA-DQ2.5cis* haplotype[1]); (ii) 13 SNPs from all 7 newly discovered regions with $P_{GWAS} < 5 \times 10^{-7}$; (iii) 86 SNPs from 59 of 68 newly discovered regions with $5 \times 10^{-7} < P_{GWAS} < 5 \times 10^{-5}$ in stage 1; and (iv) 14 SNPs from 14 of 30 newly discovered regions with $5 \times 10^{-5} < P_{GWAS} < 10^{-4}$ in stage 1 (for this last category, we mostly chose regions with immune system genes). Two SNPs were selected per region for regions with stronger association, regions with possible multiple independent associations and/or regions containing genes of obvious biological interest. We successfully genotyped 131 SNPs in 7 independent follow-up cohorts comprising 4,918 individuals with celiac disease and 5,684 control subjects of European descent (**Table 1**). Genotype call rates were >99.9% in each collection. Primary association analyses of the combined GWAS and follow-up data were performed with a two-sided 2 × 2 × 12 Cochran-Mantel-Haenszel test. Finally, we examined associated risk loci for *cis* expression-genotype correlations; a summary of subjects used for expression quantitative trait locus (eQTL) analyses is reported in **Supplementary Table 1**.

Celiac disease risk variants

The *HLA* locus and all 13 other previously reported celiac disease risk loci showed evidence for association at a genome-wide significance threshold ($P_{combined} < 5 \times 10^{-8}$; **Table 2** and **Supplementary Fig. 1**). We note that some loci were previously reported using less stringent criteria (for example, the $P < 5 \times 10^{-7}$ recommended by the 2007 WTCCC study[11]); however, in the current, much larger sample set, all known loci meet recently proposed $P < 5 \times 10^{-8}$ thresholds[12,13].

We identified 13 new risk regions with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$) of association, including regions containing the *BACH2*, *CCR4*, *CD80*, *CIITA-SOCS1-CLEC16A*, *ETS1*, *ICOSLG*, *RUNX3*, *THEMIS*, *TNFRSF14* and *ZMIZ1* genes, which have obvious immunological functions (**Table 2** and **Supplementary Fig. 1**). A further 13 regions met 'suggestive' criteria for association ($10^{-6} > P_{combined} > 5 \times 10^{-8}$ and/or $P_{GWAS} < 10^{-4}$ and $P_{followup} < 0.01$; **Table 2** and **Supplementary Fig. 1**). These regions also contain multiple genes with immunological functions, including *CD247*, *FASLG-TNFSF18-TNFSF4*, *IRF4*, *TLR7-TLR8*, *TNFRSF9* and *YDJC*. Six of the 39 non-HLA regions show evidence for the presence of multiple independently associated variants in a conditional logistic regression analysis (**Supplementary Table 2**).

We tested the 40 SNPs with the strongest association (**Table 2**) from each of the known genome-wide significant, new genome-wide significant and new suggestive loci for evidence of heterogeneity across the 12 collections studied. Only the *HLA* region was significant (Breslow-Day test $P < 0.05$ per 40 tests, rs2187668 $P = 4.8 \times 10^{-8}$), which is consistent with the well-described North-South gradient in HLA allele frequency in European populations, and more specifically for *HLA-DQ* in celiac disease[14].

We observed no evidence for interaction between each of the 26 genome-wide significant non-*HLA* loci, which is consistent with what has been reported for other complex diseases so far. However, we did observe weak evidence for lower effect sizes at non-*HLA* loci in high risk *HLA-DQ2.5cis* homozygotes, similar to what has been observed in type 1 diabetes[7].

To obtain more insight into the functional relatedness of the celiac disease risk loci, we applied GRAIL, a statistical tool that uses text mining of PubMed abstracts to annotate candidate genes from loci associated with common disease risk[15,16]. To assess the sensitivity of this tool (using known loci as a positive control), we first

Table 2 Genomic regions with the strongest association signals for celiac disease

Chr.	Position (bp)	SNP	LD block[a,b] (Mb)	Minor allele	Minor allele freq[c]	P_{GWAS} 4,533 cases, 10,750 controls	$P_{follow-up}$ 4,918 cases, 5,684 controls	$P_{combined}$ 9,451 cases, 16,434 controls	Odds ratio[c] (95% CI)	Multiple independent association signals[d]	Refs.	RefSeq Genes in LD block	Genes of interest and GRAIL annotation[e]
Previously reported risk variants													
1	190803436	rs2816316	190.73–190.81	C	0.160	1.45×10^{-12}	1.56×10^{-6}	2.20×10^{-17}	0.80 (0.76–0.84)		22	1	*RGS1*
2	61040333	rs13003464	60.78–61.74	G	0.401	4.92×10^{-8}	1.57×10^{-6}	3.71×10^{-13}	1.15 (1.11–1.20)	Yes	32	8	*REL*, *AHSA2*
2	102437000	rs917997	102.22–102.57	A	0.236	5.97×10^{-15}	7.83×10^{-4}	1.11×10^{-15}	1.19 (1.14–1.25)		22	5	*IL18RAP*, *IL18R1*, *IL1RL1*, *IL1RL2*
2	181704290	rs13010713	181.50–181.97	G	0.448	2.02×10^{-8}	3.21×10^{-4}	4.74×10^{-11}	1.13 (1.09–1.18)		33	1	*ITGA4*, *UBE2E3*
2	204510823	rs4675374	204.40–204.52	A	0.223	8.80×10^{-8}	4.94×10^{-3}	5.79×10^{-9}	1.14 (1.09–1.19)		17	2	*CTLA4*, *ICOS*, *CD28*
3	46210205	rs13098911	45.90–46.57	A	0.097	2.53×10^{-11}	1.96×10^{-7}	3.26×10^{-17}	1.30 (1.23–1.39)	Yes	22	11	*CCR1*, *CCR2*, *CCRL2*, *CCR3*, *CCR5*, *CCR9*
3	161147744	rs17810546	161.07–161.23	G	0.125	4.56×10^{-18}	9.57×10^{-12}	3.98×10^{-28}	1.36 (1.29–1.44)	Yes	22	1	*IL12A*
3	189595248	rs1464510	189.55–189.62	A	0.485	9.49×10^{-24}	3.63×10^{-18}	2.98×10^{-40}	1.29 (1.25–1.34)		22	1	*LPP*
4	123334952	rs13151961	123.19–123.78	G	0.142	6.31×10^{-18}	4.45×10^{-11}	2.18×10^{-27}	0.74 (0.70–0.78)		1	4	*IL2*, *IL21*
6	32713862	rs2187668	Gene identified	A	0.258	$<10^{-50}$	$<10^{-50}$	$<10^{-50}$	6.23 (5.95–6.52)	(Yes)	1,3	6	*HLA-DQA1*, *HLA-DQB1*
6	138014761	rs2327832	137.92–138.17	G	0.216	1.41×10^{-14}	1.97×10^{-6}	4.46×10^{-19}	1.23 (1.17–1.28)		32	0	*TNFAIP3*
6	159385965	rs1738074	159.24–159.45	A	0.434	3.14×10^{-8}	1.56×10^{-8}	2.94×10^{-15}	1.16 (1.12–1.21)		22	2	*TAGAP*
12	110492139	rs653178	110.19–111.51	G	0.495	6.03×10^{-14}	1.47×10^{-8}	7.15×10^{-21}	1.20 (1.15–1.24)		22	13	*SH2B3*
18	12799340	rs1893217	12.73–12.91	G	0.165	5.52×10^{-7}	1.04×10^{-4}	2.52×10^{-10}	1.17 (1.12–1.23)		17	1	*PTPN2*
New loci with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$)													
1	2516606	rs3748816	2.40–2.78	G	0.339	4.93×10^{-7}	1.17×10^{-3}	3.28×10^{-9}	0.89 (0.85–0.92)			4	*TNFRSF14*, *MMEL1*
1	25176163	rs10903122	25.11–25.18	A	0.480	3.21×10^{-5}	8.44×10^{-7}	1.73×10^{-10}	0.89 (0.85–0.92)			1	*RUNX3*
1	199158760	rs296547	199.12–199.31	A	0.357	6.46×10^{-5}	1.34×10^{-5}	4.11×10^{-9}	0.89 (0.86–0.92)			2	?
2	68452459	rs17035378[f]	68.39–68.54	G	0.278	1.34×10^{-5}	1.41×10^{-4}	7.79×10^{-9}	0.88 (0.84–0.92)			2	*PLEK*
3	32990473	rs13314993[f]	32.90–33.06	C	0.464	6.87×10^{-6}	1.09×10^{-4}	3.27×10^{-9}	1.13 (1.08–1.17)			2	*CCR4*
3	120601486	rs11712165[f]	120.59–120.78	C	0.394	5.40×10^{-7}	1.72×10^{-3}	8.03×10^{-9}	1.13 (1.08–1.17)			5	*CD80*, *KTELC1*
6	90983333	rs10806425	90.86–91.10	A	0.397	9.46×10^{-6}	9.25×10^{-6}	3.89×10^{-10}	1.13 (1.09–1.17)			1	*BACH2*, *MAP3K7*
6	128320491	rs802734	127.99–128.38	G	0.311	1.36×10^{-6}	1.70×10^{-9}	2.62×10^{-14}	1.17 (1.12–1.22)	Yes		2	*PTPRK*, *THEMIS*
8	129333771	rs9792269	129.21–129.37	G	0.238	8.14×10^{-6}	1.00×10^{-4}	3.28×10^{-9}	0.88 (0.84–0.91)			0	?
10	80728033	rs1250552	80.69–80.76	G	0.466	5.80×10^{-8}	1.81×10^{-3}	9.09×10^{-10}	0.89 (0.86–0.92)			1	*ZMIZ1*
11	127886184	rs11221332[f]	127.84–127.99	A	0.237	4.74×10^{-11}	9.98×10^{-7}	5.28×10^{-16}	1.21 (1.16–1.27)	Yes		1	*ETS1*
16	11311394	rs12928822	11.22–11.39	A	0.161	1.07×10^{-5}	7.59×10^{-4}	3.12×10^{-8}	0.86 (0.82–0.91)			4	*CIITA*, *SOCS1*, *CLEC16A*
21	44471849	rs4819388	44.42–44.47	A	0.280	3.42×10^{-5}	1.66×10^{-5}	2.46×10^{-9}	0.88 (0.84–0.92)			2	*ICOSLG*
New loci with suggestive evidence (either $10^{-6} > P_{combined} > 5 \times 10^{-8}$ or $P_{GWAS} < 10^{-4}$ and $P_{follow-up} < 0.01$)													
1	7969259	rs12727642	7.84–8.13	A	0.185	3.06×10^{-5}	8.21×10^{-4}	9.11×10^{-8}	1.14 (1.09–1.20)			4	*PARK7*, *TNFRSF9*
1	61564451	rs6691768	61.52–61.62	G	0.378	2.63×10^{-5}	1.16×10^{-3}	1.19×10^{-7}	0.90 (0.87–0.94)			1	*NFIA*
1	165678008	rs864537	165.43–165.71	G	0.391	1.01×10^{-7}	9.25×10^{-2}	3.80×10^{-7}	0.91 (0.87–0.94)			1	*CD247*
1	170977623	rs859637	170.87–171.20	A	0.486	8.15×10^{-5}	5.68×10^{-3}	1.75×10^{-6}	1.10 (1.06–1.14)			1	*FASLG*, *TNFSF18*, *TNFSF4*
3	69335589	rs6806528[f]	69.27–69.37	A	0.097	4.84×10^{-5}	7.66×10^{-4}	1.46×10^{-7}	1.19 (1.12–1.27)			1	*FRMD4B*
3	170974795	rs10936599	170.84–171.09	A	0.252	2.99×10^{-7}	6.63×10^{-2}	4.57×10^{-7}	1.12 (1.07–1.16)			3	?
6	328546	rs1033180[g]	0.32–0.40	A	0.080	9.14×10^{-6}	1.48×10^{-3}	5.58×10^{-8}	1.21 (1.13–1.29)	Yes		1	*IRF4*[g]
7	37341035	rs6974491	37.32–37.41	A	0.170	1.37×10^{-5}	2.63×10^{-3}	1.56×10^{-7}	1.14 (1.09–1.20)			1	*ELMO1*
13	49733716	rs2762051	49.63–49.96	A	0.184	3.35×10^{-5}	5.06×10^{-3}	6.64×10^{-7}	1.13 (1.08–1.18)			0	?
14	68347957	rs4899260	68.24–68.39	A	0.263	4.55×10^{-5}	2.21×10^{-3}	3.92×10^{-7}	1.12 (1.07–1.16)			2	*ZFP36L1*
17	42220599	rs2074404	41.40–42.25	C	0.250	5.03×10^{-5}	5.96×10^{-3}	1.23×10^{-6}	0.90 (0.86–0.94)			10	?
22	20312892	rs2298428	20.14–20.35	A	0.201	2.49×10^{-7}	4.13×10^{-2}	1.84×10^{-7}	1.13 (1.08–1.19)			6	*UBE2L3*, *YDJC*
X	12881445	rs5979785	12.82–12.93	G	0.263	6.32×10^{-6}	2.18×10^{-3}	6.36×10^{-8}	0.88 (0.84–0.92)			1	*TLR7*, *TLR8*

[a]The most significantly associated SNP from each region is shown. [b]LD regions were defined by extending 0.1 cM to the left and right of the focal SNP as defined by the HapMap3 recombination map. All chromosomal positions are based on NCBI build-36 coordinates. [c]Minor allele in all samples in the combined dataset, odds ratios (shown for combined dataset) defined with respect to the minor allele in all controls. [d]Evidence from logistic regression at a genome-wide significant or suggestive level of significance after conditioning on other associated SNPs (see **Supplementary Table 2**). *HLA* region not tested, but previously known. [e]Selected named genes within or adjacent to the same LD block as the associated SNPs; causality is not proven. In particular, other genes and other causal mechanisms may exist. Gene names underlined are identified from GRAIL[15,16] analysis (see Online Methods) with $P_{text} < 0.01$. [f]These markers were present on the Hap550 but not Hap300 SNP sets, and are not genotyped for 737 cases and 2,596 controls in the stage 1 GWAS, and combined dataset analyses. Only minor changes in *P* values were observed when these genotypes were imputed and included in analysis. [g]The *IRF4* region (specifically rs9738805, $r^2 = 0.08$ with rs1033180 in HapMap CEU) was previously identified as showing strong geographical differentiation[11]. Association with celiac disease was still observed after correction for population stratification using either a structured association approach[34] (corrected $P_{GWAS} = 5.16 \times 10^{-6}$, 478 × 2 × 2 CMH test) or principal components correction (uncorrected $P_{GWAS} = 7.05 \times 10^{-6}$, corrected $P_{GWAS} = 2.28 \times 10^{-5}$, Cochran-Armitage trend tests combined using weighted *Z* scores; see Online Methods). However, definitive exclusion of population stratification would require family-based association studies.

performed a 'leave-one-out' analysis of the 27 genome-wide significant celiac disease loci (including *HLA-DQ*). GRAIL scores of $P_{text} < 0.01$ were obtained for 12 of the 27 loci (44% sensitivity; **Table 2**). Factors that limit the sensitivity of GRAIL include biological pathways being both known (a 2006 dataset is used to avoid GWAS-era studies) and published in the literature. We then applied GRAIL analysis, using the 27 known regions as a seed, to all 49 regions (49 SNPs) with $10^{-3} > P_{combined} > 5 \times 10^{-8}$ and obtained GRAIL $P_{text} < 0.01$ for 9 regions (18.4%). As a control, only 5.5% (279 of 5,033) of randomly selected Hap550 SNPs reached this threshold. In addition to the five

Table 3 Celiac risk variants correlated with *cis* gene expression

SNP[a]	Chr.	SNP position[b]	Probe center position[b]	Illumina ArrayAddressID	Expression dataset[c]	Gene name	eQTL P[d]
Loci with genome-wide significant evidence ($P_{combined} < 5 \times 10^{-8}$)							
rs3748816	1	2516606	2412221	650452	HT-12	*PLCH2*	1.66×10^{-5}
rs3748816	1	2516606	2482955	6520725	Ref-8v2 + HT-12	*TNFRSF14*	1.30×10^{-3}
rs3748816	1	2516606	2510429	6250338	Ref-8v2	*C1orf93*	1.16×10^{-4}
rs3748816	1	2516606	2533115	2070246	Ref-8v2 + HT-12	*MMEL1*	1.03×10^{-20}
rs296547	1	199158760	198880146	1300279	Ref-8v2 + HT-12	*DDX59*	2.45×10^{-5}
rs842647	2	60972975	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	3.30×10^{-10}
rs13003464[e]	2	61040333	61263810	1170220	Ref-8v2 + HT-12	*AHSA2*	6.39×10^{-11}
rs3816281[f]	2	68461451	68461957	4810020	Ref-8v2 + HT-12	*PLEK*	7.97×10^{-26}
rs917997	2	102437000	102418571	6520180	Ref-8v2 + HT-12	*IL18RAP*	7.35×10^{-87}
rs13010713	2	181704290	181593865	1780433	HT-12	*UBE2E3*	4.93×10^{-5}
rs13098911	3	46210205	45964449	6550333	Ref-8v2 + HT-12	*CXCR6*	9.66×10^{-6}
rs13098911	3	46210205	46255176[g]	2190671	HT-12	*CCR3*	5.50×10^{-10}
rs13098911	3	46210205	46255176[g]	7570670	Ref-8v2	*CCR3*	5.69×10^{-4}
rs6441961[d]	3	46327388	46255176[h]	2190671	HT-12	*CCR3*	2.87×10^{-19}
rs6441961[d]	3	46327388	46255176[h]	7570670	Ref-8v2	*CCR3*	1.02×10^{-4}
rs11922594[f]	3	120608512	120683364[i]	6550288	Ref-8v2 + HT-12	*KTELC1*	5.09×10^{-17}
rs11922594[f]	3	120608512	120683364[i]	3850161	Ref-8v2 + HT-12	*KTELC1*	7.34×10^{-6}
rs10806425	6	90983333	90878075	3520349	HT-12	*BACH2*	1.92×10^{-3}
rs1738074	6	159385965	159380068	5890739	Ref-8v2 + HT-12	*TAGAP*	1.99×10^{-3}
rs1738074	6	159385965	159381094[j]	5360364	HT-12	*TAGAP*	3.23×10^{-4}
rs1738074	6	159385965	159381094[j]	4860242	HT-12	*TAGAP*	2.18×10^{-3}
rs1250552	10	80728033	80622540	2450131	Ref-8v2 + HT-12	*ZMIZ1*	1.80×10^{-3}
rs653178	12	110492139	110399552	6560301	Ref-8v2 + HT-12	*SH2B3*	9.24×10^{-12}
rs653178	12	110492139	110710447	840253	Ref-8v2 + HT-12	*ALDH2*	1.44×10^{-4}
rs653178	12	110492139	110894406[k]	2070736	HT-12	*TMEM116*	3.68×10^{-4}
rs653178	12	110492139	110894406[k]	3190129	Ref-8v2	*TMEM116*	1.51×10^{-3}
rs12928822	16	11311394	11335627	4540072	Ref-8v2 + HT-12	*C16orf75*	1.02×10^{-8}
rs4819388	21	44471849	44049567	7200373	Ref-8v2	*RRP1*	2.62×10^{-3}
Loci with suggestive evidence (either $10^{-6} > P_{combined} > 5 \times 10^{-8}$ or $P_{GWAS} < 10^{-4}$ and $P_{follow\text{-}up} < 0.01$)							
rs12727642	1	7969259	7956138	610193	Ref-8v2 + HT-12	*PARK7*	9.76×10^{-15}
rs864537	1	165678008	165710482[l]	6290400	Ref-8v2 + HT-12	*CD247*	1.77×10^{-9}
rs864537	1	165678008	165710482[l]	3890689	HT-12	*CD247*	2.93×10^{-7}
rs6974491	7	37341035	37157761	2750154	Ref-8v2 + HT-12	*ELMO1*	5.40×10^{-6}
rs2074404	17	42220599	41824345	3520672	Ref-8v2 + HT-12	*LRRC37A*	1.17×10^{-4}
rs2074404	17	42220599	42106695[m]	5260138	Ref-8v2 + HT-12	*NSF*	1.20×10^{-5}
rs2074404	17	42220599	42106695[m]	1410484	HT-12	*NSF*	4.28×10^{-4}
rs2074404	17	42220599	42223012	4070615	HT-12	*WNT3*	2.77×10^{-3}
rs2074404	17	42220599	42485154	4880037	HT-12	*LOC388397*	1.78×10^{-9}
rs2298428	22	20312892	20308188	1230242	Ref-8v2 + HT-12	*UBE2L3*	1.96×10^{-90}
rs5979785	X	12881445	12842944[n]	6480360	Ref-8v2 + HT-12	*TLR8*	3.88×10^{-13}
rs5979785	X	12881445	12842944[n]	3390612	Ref-8v2 + HT-12	*TLR8*	1.07×10^{-7}

See **Supplementary Figures 2** and **3** for detailed results and **Supplementary Table 3** for more details of Illumina expression probes.

[a]We tested the SNP with the strongest association from 34 of 39 non-HLA loci ($P_{combined} < 10^{-6}$, **Table 2**), Hap300 proxy SNPs for 4 further loci, and a second independently associated SNP from 6 loci, for correlation with gene expression in PAXgene blood RNA in up to 1,349 individuals. One locus (containing *ETS1*) where an adequate proxy SNP was not available was not included for the eQTL analysis. SNP-gene expression correlations were tested for probes within a 1-Mb window. Results are presented for SNPs showing significant correlations with *cis* gene expression after controlling false-discovery rate at 5% (corresponding to $P < 0.0028$). [b]All chromosomal positions are based on NCBI build-36 coordinates. Probe center position was determined by re-mapping probe sequences to the human transcriptome and calculated from the midpoint of the transcript start and transcript end positions in genomic coordinates. [c]'HT-12' comprise 1,240 individuals with blood gene expression assayed using Illumina Human HT-12v3 arrays; 'Ref-8v2' comprise 229 individuals with blood gene expression assayed using Illumina Human-Ref-8v2 arrays (see Online Methods). [d]Spearman rank correlation of genotype and residual variance in transcript expression. Meta-analysis eQTL P value shown if both datasets had identical probes. [e]Second, independently associated SNP from this locus. [f]Proxy SNP, $r^2 = 0.61$ in HapMap CEU with most associated SNP rs11712165. [g–n]Different Illumina probe sequences with the same probe center position.

'suggestive' loci shown in **Table 2**, GRAIL annotated four further interesting gene regions with lower significance in the combined association results: rs944141-*PDCD1LG2* ($P_{combined} = 4.4 \times 10^{-6}$), rs976881-*TNFRSF8* ($P_{combined} = 2.1 \times 10^{-4}$), rs4682103-*CD200-BTLA* ($P_{combined} = 6.8 \times 10^{-6}$) and rs4919611-*NFKB2* ($P_{combined} = 6.1 \times 10^{-5}$). There appeared to be further enrichment for genes of immunological interest that are not GRAIL-annotated in the $10^{-3} > P_{combined} > 5 \times 10^{-8}$ significance window, including rs3828599-*TNIP1* ($P_{combined} = 1.55 \times 10^{-4}$), rs8027604-*PTPN9* ($P_{combined} = 1.4 \times 10^{-6}$) and rs944141-*CD274* ($P_{combined} = 4.4 \times 10^{-6}$). Some of these findings, for which neither genome-wide significant nor suggestive association is achieved, are likely to comprise part of a longer tail of disease-predisposing common variants with weaker effect sizes. Definitive assessment of these biologically plausible regions would require genotyping and association studies using much larger sample collections than the present study.

We previously showed that there is considerable overlap between risk loci for celiac disease and type 1 diabetes[17], as well as between risk loci for celiac disease and rheumatoid arthritis[18], and more generally, there is now substantial evidence for shared risk loci between the common chronic immune-mediated diseases[6]. To update these observations, we searched 'A Catalog of Published Genome Wide Association Studies' (accessed 18 November 2009)[19] and the HuGE database[20]. We found some evidence (requiring a published association report of $P < 1 \times 10^{-5}$) of shared loci with at least one other inflammatory or immune-mediated disease for 18 of the current 27 genome-wide significant celiac disease risk regions. We defined shared regions as the broad linkage disequilibrium block; however, different SNPs are often reported in different diseases, and at only 3 of the 18 shared regions are associations across all diseases with the same SNP or a proxy SNP in $r^2 > 0.8$ in HapMap CEU. Currently, nine regions seem to be specific to celiac disease and might reflect distinctive disease biology, including the regions containing rs296547 and rs9792269 and the regions around *CCR4*, *CD80*, *ITGA4*, *LPP*, *PLEK*, *RUNX3* and *THEMIS*. In fact, locus sharing between diseases is probably greater because of both stochastic variation in results from sample size limitations and regions that have a genuinely stronger effect size in one disease and weaker effect size in another.

Genetic variation in *ETS1* has recently been reported to be associated with systemic lupus erythematosus (SLE) in the Chinese population, although it is not associated with SLE in European populations[21]. The most strongly associated celiac disease (European population) SNP, rs11221332, and the most strongly associated SLE (Chinese population) SNP, rs6590330, map 70 kb apart. Inspection of the HapMap phase II data shows broadly similar linkage disequilibrium patterns between Chinese (CHB) and European (CEU) populations in this region, with the two associated SNPs in separate nonadjacent linkage disequilibrium blocks. Thus, distinct common variants within the same gene can predispose to different autoimmune diseases across different ethnic groups.

Exploring the function of celiac disease risk variants

Celiac disease risk variants in the *HLA* genes alter protein structure and function[4]. However, we identified only four nonsynonymous SNPs with evidence for association with celiac disease ($P_{GWAS} < 10^{-4}$) from the other 26 genome-wide significant associated regions (rs3748816-*MMEL1*, rs3816281-*PLEK*, rs196432-*RUNX3*, rs3184504-*SH2B3*). Although comprehensive regional resequencing is required to test the possibility that coding variants contribute to the observed association signals, more subtle effects of genetic variation on gene expression are the more likely functional mechanism for complex disease genes. With this in mind, we performed a meta-analysis of new and published genome-wide eQTL datasets comprising 1,469 human whole blood (PAXgene) samples reflecting primary leukocyte gene expression. We applied a new method, transcriptional components, to remove a substantial proportion of inter-individual nongenetic expression variation and performed eQTL meta-analysis on the residual expression variation (Online Methods).

We assessed 38 of the 39 genome-wide significant and suggestive celiac disease–associated non-*HLA* loci (**Table 2**) for *cis* expression-genotype correlations. We tested the SNP with the strongest association from each region. However, for five regions the most associated SNP was not genotyped in the eQTL samples (Hap300 data); instead, for four of these, we tested a proxy SNP ($r^2 > 0.5$ in HapMap CEU). In addition, for six loci showing evidence of multiple independent associations in conditional regression analyses, we tested a second SNP that showed independent association with celiac disease for eQTL analysis. In total, we assessed 44 independent non-*HLA* SNP associations in peripheral whole blood samples genotyped on the Illumina Hap300 BeadChip and either Illumina Ref8 or HT12 expression arrays, correlating each SNP with data from gene probes mapping within a 1-Mb window.

We identified significant (Spearman $P < 0.0028$, corresponding to 5% false-discovery rate) eQTLs at 20 of 38 (52.6%) non-*HLA* celiac loci tested (**Table 3** and **Supplementary Figs. 2** and **3**). Some loci had evidence of eQTLs with multiple probes, genes or SNPs (**Table 3**). We assessed whether the number of SNPs with *cis*-eQTL effects out of the 44 SNPs that we tested was significantly higher than expected. On average, eQTL SNPs had a substantially higher minor allele frequency (MAF) than non-eQTL SNPs in the 294,767 SNPs tested. To correct for this, we selected 44 random SNPs that had an equal MAF distribution and determined for how many of these MAF-matched SNPs eQTLs were observed. There were a significantly higher number of eQTL SNPs ($P = 9.3 \times 10^{-5}$, 10^6 permutations) among the celiac disease–associated SNPs than expected by chance (22 observed eQTL SNPs versus 7.8 expected eQTL SNPs). Therefore, the celiac disease–associated regions are greatly enriched for eQTLs. These data indicate that some risk variants might influence celiac disease susceptibility through a mechanism of altered gene expression. Candidate genes with a significant eQTL where the peak eQTL signal and peak case-control association signal are similar (**Supplementary Fig. 3**) include *MMEL1*, *NSF*, *PARK7*, *PLEK*, *TAGAP*, *RRP1*, *UBE2L3* and *ZMIZ1*.

We also assessed the coexpression of genes that mapped within 500 kb of SNPs that showed the strongest case-control association from the 40 genome-wide significant and suggestive celiac disease loci in an analysis of the 33,109 human Affymetrix Gene Expression Omnibus dataset. This analysis loses power to detect tissue-specific correlations from the use of numerous tissue types, but it greatly gains power from the large sample size. We detected several distinct coexpression clusters (Pearson correlation coefficient between genes >0.5), including four clusters of immune-related genes that contain at least one gene from 37 of the 40 genome-wide significant and suggestive loci (**Fig. 1**). These data further demonstrate that genes from celiac disease risk loci map to multiple distinct immunological pathways involved in disease pathogenesis.

DISCUSSION

We previously reported that most celiac genetic risk variants mapped near genes that are functional in the immune system[22], and this remains true for the 13 new genome-wide significant and 13 new suggestive risk variants from the current study. We can now refine

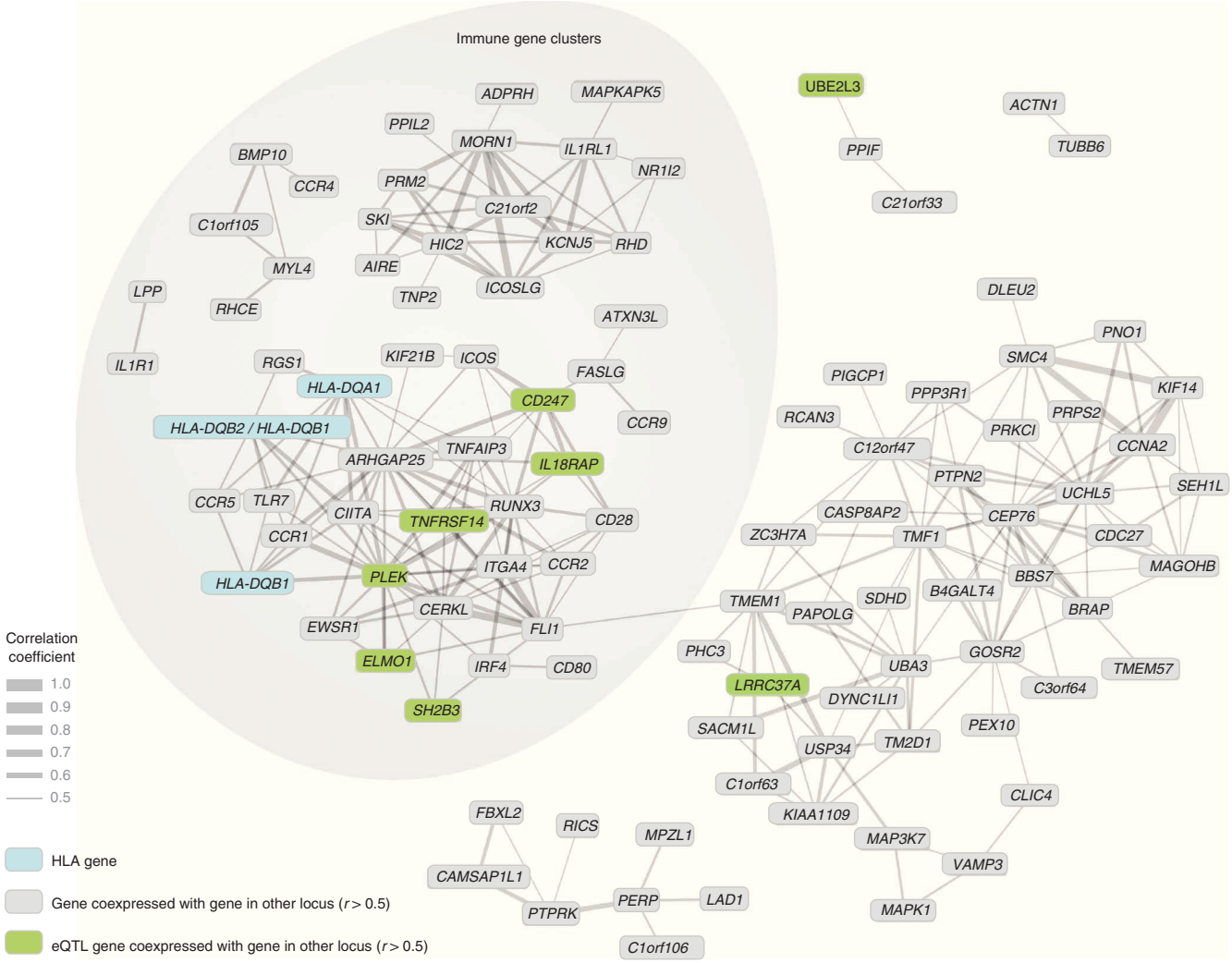


Figure 1 Coexpression analysis of genes mapping to 40 genome-wide significant and suggestive celiac disease regions in 33,109 heterogenous human samples from the Gene Expression Omnibus. Genes mapping within a 1-Mb window of associated SNPs (**Table 2**) were tested for interaction with genes from other loci. Interactions with Pearson correlation > 0.5 are shown ($P < 10^{-100}$). Only the genes known to contain causal mutations (*HLA-DQA1*, *HLA-DQB1*) were analyzed from the HLA region; HLA-DQB2/HLA-DQB1 is a single expression probeset mapping to both genes. No probe for *THEMIS* was present on the earlier version of the U133 array; however, in a subset analysis of U133 Plus2.0 data, *THEMIS* is coexpressed in the major immune gene cluster.

these observations and highlight specific immunological pathways that are relevant to the pathogenesis of celiac disease.

One key pathway worth highlighting is T-cell development in the thymus. The rs802734 linkage disequilibrium block contains the recently identified gene *THEMIS* (thymus-expressed molecule involved in selection). *THEMIS* has a key regulatory role in both positive and negative T-cell selection during late thymocyte development[23]. Furthermore, the rs10903122 linkage disequilibrium block contains *RUNX3*, a master regulator of $CD8^+$ T lymphocyte development in the thymus[24,25]. *TNFRSF14* (LIGHTR, rs3748816 linkage disequilibrium block) has widespread functions in peripheral leukocytes and a crucial role in promoting thymocyte apoptosis[26]. The *ETS1* transcription factor (rs11221332 linkage disequilibrium block) is also active in peripheral leukocytes; however, it is also a key player in thymic $CD8^+$ lineage differentiation, acting in part by promoting *RUNX3* expression[27].

The importance of the thymus in the pathogenesis of autoimmune diseases has been previously emphasized by the established role of thymectomy in the treatment of myasthenia gravis. In type 1 diabetes, disease-associated genetic variation in the insulin gene *INS* causes altered thymic insulin expression and subsequent T-cell tolerance for insulin as a self-protein[28]. However, the importance of thymic T-cell regulation in the etiology of celiac disease has not been previously recognized. It is conceivable that the associated variants might alter biological processes before thymic MHC-ligand interactions. Alternatively, it is now clear that exogenous antigen presentation and selection occurs in the thymus through migratory dendritic cells; this has been demonstrated for skin and has been hypothesized for food antigens[29,30]. These findings suggest that it would be worthwhile to investigate immunological and pharmacological modifiers of T-cell tolerance more generally in autoimmune diseases.

A second pathway worth noting is the innate immune detection of viral RNA. Although the association signal at rs5979785 ($P_{combined} = 6.36 \times 10^{-8}$) in the *TLR7-TLR8* region is just outside our genome-wide significance threshold, we observe a strong effect of rs5979785 on *TLR8* expression in whole blood. Both TLRs recognize viral RNA. Taken together with the recent observation that rare loss-of-function mutations in the enteroviral response gene *IFIH1* are protective against

type 1 diabetes[31], these findings implicate viral infection (and the nature of the host response to infection) as a putative environmental trigger that could be common to these autoimmune diseases.

A third pathway involves T- and B-cell co-stimulation (or co-inhibition). This class of molecules controls the strength and nature of the response to T-cell or B-cell (immunoglobulin) receptor activation by antigens. We observe multiple regions with genes (*CTLA4-ICOS-CD28*, *TNFRSF14*, *CD80*, *ICOSLG*, *TNFRSF9*, *TNFSF4*) from this class of ligand-receptor pairs, indicating that fine control of the adaptive immune response might be altered in individuals at risk of celiac disease.

A final pathway involves cytokines, chemokines and their receptors. Our previous report discussed the function of the 2q11–12 interleukin receptor cluster (*IL18RAP* and so on), the 3p21 chemokine receptor cluster (*CCR5* and so on) and the loci containing *IL2-IL21* and *IL12A*[22]. We now report additional loci containing *TNFSF18* and *CCR4*.

We estimate that the current celiac disease variants, including the major celiac disease-associated *HLA* variant, *HLA-DQ2.5cis*, less common celiac disease–associated haplotypes in the *HLA* (*HLA-DQ8*; *HLA-DQ2.5trans*; *HLADQ2.2*), and the additional 26 definitively implicated loci explain about 20% of total celiac disease variance, which would represent 40% of genetic variance, assuming a heritability of 0.5. A long tail of common variants with low effect size, along with highly penetrant rare variants (both at the established loci and elsewhere in the genome), might contribute substantially to the remaining heritability.

We observed different haplotypes within the *ETS1* region associated with celiac disease in Europeans and SLE in the Chinese population. For some autoimmune diseases studied in European origin populations, although the same linkage disequilibrium block has been associated, the association is with a different haplotype. In some cases, the same variants are associated, but the direction of association is opposite (for example, rs917997-*IL18RAP* in celiac disease versus type 1 diabetes). We believe further exploration of these signals might reveal critical differences in the nature of the immune system perturbation between these diseases.

Previously, investigators have observed that only a small proportion of GWAS signals involve coding variants and have suggested that these variants might instead influence regulation of gene expression. Here we show that over half the variants associated with celiac disease are correlated with expression changes in nearby genes. This mechanism is likely to explain the function of some risk variants for other common, complex diseases. Further research is needed to definitively determine at each locus both the variants that can cause celiac disease and their functional mechanisms.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession numbers. Expression data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/) as GSE20142 and GSE20332.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank Coeliac UK for assistance with direct recruitment of individuals with celiac disease, and UK clinicians (L.C. Dinesen, G.K.T. Holmes, P.D. Howdle, J.R.F. Walters, D.S. Sanders, J. Swift, R. Crimmins, P. Kumar, D.P. Jewell, S.P.L. Travis and K. Moriarty) who recruited the celiac disease blood samples described in our previous studies[1,22]. We thank the genotyping facility of the UMCG (J. Smolonska and P. van der Vlies) for generating part of the GWAS and replication data and the gene expression data; R. Booij and M. Weenstra for preparation of Italian samples; H. Ahola, A. Heimonen, L. Koskinen, E. Einarsdottir and K. Löytynoja for their work on Finnish sample collection, preparation and data handling; and E. Szathmári, J.B.Kovács, M. Lörincz and A. Nagy for their work with the Hungarian families. The Health2000 organization, Finrisk consortium, K. Mustalahti, M. Perola, K. Kristiansson and J. Koskinen are thanked for providing the Finnish control genotypes. We thank D.G. Clayton and N. Walker for providing T1DGC data in the required format. We thank the Irish Transfusion Service and Trinity College Dublin Biobank for control samples and V. Trimble, E. Close, G. Lawlor, A. Ryan, M. Abuzakouk, C. O'Morain and G. Horgan for celiac disease sample collection and preparation We acknowledge DNA provided by Mayo Clinic Rochester and thank M. Bonamico and M. Barbato (Department of Paediatrics, Sapienza University of Rome, Italy) for recruiting individuals. We thank Polish clinicians for recruitment of individuals with celiac disease (Z. Domagala, A. Szaflarska-Poplawska, B. Oralewska, W. Cichy, B. Korczowski, K. Fryderek, E. Hapyn, K. Karczewska, A. Zalewska, I. Sakowska-Maliszewska, R. Mozrzymas, A. Zabka, M. Kolasa and B. Iwanczak). We thank M. Szperl for isolating DNA from blood samples provided by the Children's Memorial Health Institute (Warsaw, Poland). Dutch and UK genotyping for the second celiac disease GWAS was funded by the Wellcome Trust (084743 to D.A.v.H.). Italian genotyping for the second celiac disease GWAS was funded by the Coeliac Disease Consortium, an Innovative Cluster approved by the Netherlands Genomics Initiative and partially funded by the Dutch Government (BSIK03009 to C.W.) and by the Netherlands Organisation for Scientific Research (NWO, VICI grant 918.66.620 to C.W.). E.G. is funded by the Italian Ministry of Healthy (grant RC2009). L.H.v.d.B. acknowledges funding from the Prinses Beatrix Fonds, the Adessium foundation and the Amyotrophic Lateral Sclerosis Association. L.F. received a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031) and a VENI grant from NWO (ZonMW grant 916.10.135). P.C.A.D. is an MRC Clinical Training Fellow (G0700545). G.T. received a Ter Meulen Fund grant from the Royal Netherlands Academy of Arts and Sciences (KNAW). The gene expression study was funded in part by COPACETIC (EU grant 201379). This study makes use of data generated by the Wellcome Trust Case-Control Consortium 2 (WTCCC2). A full list of the WTCCC2 investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the WTCCC2 project was provided by the Wellcome Trust under award 085475. This research utilizes resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Human Genome Research Institute (NHGRI), National Institute of Child Health and Human Development (NICHD) and Juvenile Diabetes Research Foundation International (JDRF) and supported by U01 DK062418. We acknowledge the use of BRC Core Facilities provided by the financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust. We acknowledge funding from the NIH: DK050678 and DK081645 (to S.L.N.), NS058980 (to R.A.O.); and DK57892 and DK071003 (to J.A.M.). The collection of Finnish and Hungarian subjects with celiac disease was funded by the EU Commission (MEXT-CT-2005-025270), the Academy of Finland, Hungarian Scientific Research Fund (contract OTKA 61868), the University of Helsinki Funds, the Competitive Research Funding of the Tampere University Hospital, the Foundation of Pediatric Research, the Sigrid Juselius Foundation and the Hungarian Academy of Sciences (2006TKI247 to R.A.). Funding for the collection and genotyping of the Polish samples was provided by UMC Cooperation Project (6/06/2006/NDON). R.M. is funded by Science Foundation Ireland. C. Núñez has a FIS contract (CP08/0213). The Dublin Centre for Clinical Research contributed to collection of samples from affected individuals and is funded by the Irish Health Research Board and the Wellcome Trust. Finally, we thank all individuals with celiac disease and control individuals for participating in this study.

AUTHOR CONTRIBUTIONS

D.A.v.H. and C.W. designed, co-ordinated and led the study. Experiments were performed in the labs of C.W., D.A.v.H., C.A.M., P.D. and P.M.G. Major contributions were: (i) DNA sample preparation: P.C.A.D., G.T., K.A.H., J.R., A.Z. and P.S.; (ii) genotyping: P.C.A.D., G.T., K.A.H., A.C., J.R. and R.G.; (iii) expression data generation: H.J.M.G., L.H.v.d.B., R.A.O., R.K.W. and L.F.; (iv) case-control association analyses: P.C.A.D., G.T., L.F., J.C.B. and D.A.v.H.; (v) expression analyses: L.F., G.A.R.H. and R.S.N.F.; (vi) manuscript preparation: P.C.A.D., G.T., L.F., R.S.N.F., G.A.R.H., J.C.B., C.W. and D.A.v.H. Other authors contributed variously to sample collection and all other aspects of the study. All authors reviewed the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring *IL2* and *IL21*. *Nat. Genet.* **39**, 827–829 (2007).
2. van Heel, D.A. & West, J. Recent advances in coeliac disease. *Gut* **55**, 1037–1046 (2006).
3. Sollid, L.M. *et al.* Evidence for a primary association of celiac disease to a particular HLA-DQ α/β heterodimer. *J. Exp. Med.* **169**, 345–350 (1989).
4. Kim, C.Y., Quarsten, H., Bergseng, E., Khosla, C. & Sollid, L.M. Structural basis for HLA-DQ2-mediated presentation of gluten epitopes in celiac disease. *Proc. Natl. Acad. Sci. USA* **101**, 4175–4179 (2004).
5. Henderson, K.N. *et al.* A structural and immunological basis for the role of human leukocyte antigen DQ8 in celiac disease. *Immunity* **27**, 23–34 (2007).
6. Zhernakova, A., van Diemen, C.C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55 (2009).
7. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
8. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
9. Anderson, C.A. *et al.* Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.* **83**, 112–119 (2008).
10. Jacobs, K.B. *et al.* A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nat. Genet.* **41**, 1253–1257 (2009).
11. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
12. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
13. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
14. Karell, K. *et al.* 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 (2003).
15. Raychaudhuri, S. *et al.* Genetic variants at *CD28, PRDM1* and *CD2/CD58* are associated with rheumatoid arthritis risk. *Nat. Genet.* **41**, 1313–1318 (2009).
16. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
17. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
18. Coenen, M.J. *et al.* Common and different genetic background for rheumatoid arthritis and coeliac disease. *Hum. Mol. Genet.* **18**, 4195–4203 (2009).
19. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
20. Yu, W., Clyne, M., Khoury, M.J. & Gwinn, M. Phenopedia and Genopedia: Disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* **26**, 145–146 (2010).
21. Han, J.W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41**, 1234–1237 (2009).
22. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40**, 395–402 (2008).
23. Allen, P.M. Themis imposes new law and order on positive selection. *Nat. Immunol.* **10**, 805–806 (2009).
24. Sato, T. *et al.* Dual functions of Runx proteins for reactivating CD8 and silencing CD4 at the commitment process into CD8 thymocytes. *Immunity* **22**, 317–328 (2005).
25. Woolf, E. *et al.* Runx3 and Runx1 are required for CD8 T cell development during thymopoiesis. *Proc. Natl. Acad. Sci. USA* **100**, 7731–7736 (2003).
26. Wang, J. & Fu, Y.X. LIGHT (a cellular ligand for herpes virus entry mediator and lymphotoxin receptor)-mediated thymocyte deletion is dependent on the interaction between TCR and MHC/self-peptide. *J. Immunol.* **170**, 3986–3993 (2003).
27. Zamisch, M. *et al.* The transcription factor Ets1 is important for CD4 repression and Runx3 up-regulation during CD8 T cell differentiation in the thymus. *J. Exp. Med.* **206**, 2685–2699 (2009).
28. Vafiadis, P. *et al.* Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* **15**, 289–292 (1997).
29. Bonasio, R. *et al.* Clonal deletion of thymocytes by circulating dendritic cells homing to the thymus. *Nat. Immunol.* **7**, 1092–1100 (2006).
30. Klein, L., Hinterberger, M., Wirnsberger, G. & Kyewski, B. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nat. Rev. Immunol.* **9**, 833–844 (2009).
31. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J.A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
32. Trynka, G. *et al.* Coeliac disease-associated risk variants in *TNFAIP3* and *REL* implicate altered NF-κB signalling. *Gut* **58**, 1078–1083 (2009).
33. Garner, C.P. *et al.* Replication of celiac disease UK genome-wide association study results in a US population. *Hum. Mol. Genet.* **18**, 4219–4225 (2009).
34. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).

[1]Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [2]Genetics Department, University Medical Center and Groningen University, Groningen, The Netherlands. [3]The Genome Centre, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [4]Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [5]Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary. [6]National Institute for Health and Welfare, Helsinki, Finland. [7]Fondazione IRCCS Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [8]Department of Medical Sciences, University of Milan, Milan, Italy. [9]Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Centre Utrecht, Utrecht, The Netherlands. [10]Clinical Immunology Department, Hospital Clínico San Carlos, Madrid, Spain. [11]Public Health Research Group of Hungarian Academy of Sciences, Medical & Health Science Center, University of Debrecen, Debrecen, Hungary. [12]Unitá di Aterosclerosi e Trombosi, I.R.C.C.S Casa Sollievo della Sofferenza, S. Giovanni Rotondo, Foggia, Italy. [13]NIHR GSTFT/KCL Comprehensive Biomedical Research Centre, King's College London School of Medicine, Guy's Hospital, London, UK. [14]Department of Pulmonology, University Medical Center and Groningen University, Groningen, The Netherlands. [15]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [16]Department of Paediatric Gastroenterology, University Medical Centre Utrecht, Utrecht, The Netherlands. [17]Paediatric Research Centre, University of Tampere Medical School and Tampere University Hospital, Tampere, Finland. [18]Department of Clinical Medicine, Institute of Molecular Medicine, Trinity College Dublin, Dublin, Ireland. [19]Heim Pal Childrens Hospital, Budapest, Hungary. [20]Department of Pediatrics, Medical and Health Science Center, University of Debrecen, Hungary. [21]Gastrointestinal Unit, Mater Misericordiae University Hospital, Dublin, Ireland. [22]Department of Experimental Medicine, Sapienza University of Rome, Rome, Italy. [23]Department of Paediatrics, Leiden University Medical Centre, Leiden, The Netherlands. [24]Department of Gastroenterology, VU Medical Center, Amsterdam, The Netherlands. [25]Division of Gastroenterology and Hepatology, Department of Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. [26]Department of Medical Genetics and [27]Rudolf Magnus Institute, University Medical Center Utrecht, Utrecht, The Netherlands. [28]Center for Neurobehavioral Genetics, University of California, Los Angeles, California, USA. [29]Pediatric Gastroenterology Department, Hospital La Paz, Madrid, Spain. [30]Institute for Molecular Medicine Finland (FIMM), Helsinki, Finland. [31]Department of Gastroenterology, Hepatology and Immunology, Children's Memorial Health Institute, Warsaw, Poland. [32]European Laboratory for Food Induced Disease, University of Naples Federico II, Naples, Italy. [33]Department of Gastroenterology and Hepatology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands. [34]Department of Pathology, Children's Memorial Health Institute, Warsaw, Poland. [35]Department of Population Sciences, Beckman Research Institute of the City of Hope, Duarte, California, USA. [36]Department of Experimental Medicine, Faculty of Medicine University of Milano-Bicocca, Monza, Italy. [37]Department of Medical Genetics and [38]Research Program for Molecular Medicine, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland. [39]These authors contributed equally to this work. Correspondence should be addressed to D.A.v.H. (d.vanheel@qmul.ac.uk) or regarding expression analyses to L.F. (lude@cleverfranke.com).

ONLINE METHODS

Subjects. Written informed consent was obtained from all subjects, with Ethics Committee/Institutional Review Board approval. All individuals are of European ancestry. Affected celiac individuals were diagnosed according to standard clinical, serological and histopathological criteria, including small intestinal biopsy. DNA samples were from blood, lymphoblastoid cell lines or saliva. A more detailed description of subjects is provided in a **Supplementary Note**.

GWAS genotyping. For an overview, see **Table 1**. UK(1) case and control genotyping has been described[1,7]. Illumina 670-Quad and 1.2M-Duo (custom chips designed for the WTCCC2 and comprising Hap550/1M and common CNV content) and 610-Quad genotyping was performed in London, Hinxton and Groningen. Bead intensity data was normalized for each sample in BeadStudio, *R* and theta values exported and genotype calling performed using a custom algorithm[1,35]. A detailed description of genotype calling steps is provided in a **Supplementary Note**.

Quality control steps were performed in the following order. First, very low call rate samples and SNPs were excluded. SNPs were excluded from all sample collections if any collection showed call rates <95% or deviation from Hardy-Weinberg equilibrium ($P < 0.0001$) in controls. Samples were excluded for call rate <98%, incompatible recorded gender and genotype-inferred gender, ethnic outliers (identified by multi-dimensional scaling plots of samples merged with HapMap Phase II data), duplicates and first-degree relatives. We excluded 22 of 417 SNPs showing apparent association ($P_{GWAS} < 10^{-4}$) after visual inspection of *R* theta plots suggested possible bias.

The over-dispersion factor of association test statistics (genomic control inflation factor), λ_{GC}, was calculated using observed versus expected values for all SNPs in PLINK.

Follow-up genotyping. For an overview, see **Table 1**. Finnish controls (12) were genotyped on the 610-Quad BeadChip; other samples were genotyped using Illumina GoldenGate BeadXpress assays in London and Groningen. Genotyping calling was performed in BeadStudio for combined cases and controls in each separate collection, with the exception of the Finnish collection, and whole genome amplified samples (89 Irish cases and 106 Spanish controls). Quality control steps were performed as for the GWAS. In total, 131 of 144 SNPs passed quality control and visual inspection of genotype clouds.

SNP association analysis. Analyses were performed using PLINK v1.07 (ref. 36), mostly using the Cochran-Mantel-Haenzel test. Logistic regression analyses were used to define the independence of association signals within the same linkage disequilibrium block, with group membership included as a factorized covariate.

Genotype imputation was performed for samples genotyped on the Hap300 using BEAGLE and CEU, TSI, MEX and GIH reference samples from HapMap3. Association analysis was performed using logistic regression on posterior genotype probabilities, with group membership included as a factorized covariate.

Structured association tests were performed using PLINK as described using genetically matched cases and controls within collections identified by identity by state similarity across autosomal non-HLA SNPs[34] (settings–ppc 0.001–cc, clusters constrained by the five collections). Principal components analysis was performed using EIGENSTRAT and a set of 12,810 autosomal non-HLA SNPs chosen for low LD and ancestry information[37,38]; association tests were corrected for the top 10 principal components and combined using weighted *Z* scores.

The fraction of additive variance was calculated using a liability threshold model[39] assuming a population prevalence of 1%. Effect sizes and control allele frequencies were estimated from the combined replication panel. Genetic variance was calculated assuming 50% heritability.

GRAIL analysis. We performed GRAIL analysis (http://www.broadinstitute.org/mpg/grail/grail.php) using HG18 and Dec2006 PubMed datasets, default settings for SNP rs number submission, and the 27 genome-wide significant celiac disease risk loci (most associated SNP) as seeds. As a query, we used either associated SNPs or 101 × 50 randomly chosen Hap550 SNP datasets (5,050 SNPs, of which 5,033 mapped to the GRAIL database).

Identification of transcriptional components. We noted that the power of eQTL studies in humans is limited by substantial observed inter-individual variation in expression measurements due to nongenetic factors, and therefore developed a method, 'transcriptional components', to remove a large component of this variation (manuscript in preparation). Expression data from 42,349 heterogeneous human samples hybridized to Affymetrix HG-U133A (GEO accession number: GPL96) or HG-U133 Plus 2.0 (GEO accession number: GPL570) Genechips were downloaded[40]. Samples missing data for >150 probes were excluded, and only probes available on both platforms were analyzed, resulting in expression data for 22,106 probes and 41,408 samples. We performed quantile normalization using the median rank distribution[41] and $\log_2$ transformed the data, ensuring an identical distribution of expression signals for every sample, discarding previous normalization and transformation steps.

Initial quality control (QC) was performed by applying principal component analysis (PCA) on the sample correlation matrix (pair-wise Pearson correlation coefficients between all samples). The first principal component (PC), explaining ~80–90% of the total variance[42,43], describes probe-specific variance. 6,375 samples with correlation $R < 0.75$ of the sample array with this PC were considered outliers of lesser quality and excluded from analysis. We excluded entire GEO datasets where >25% of the samples were outliers (probably expression ratios versus a reference, not absolute data). The final dataset comprised 33,109 samples (17,568 GPL96 and 15,541 GPL570 samples), and we repeated the normalization and transformation on the originally deposited expression values of these post-quality control samples.

We next applied PCA on the pairwise 22,106 × 22,106 probe Pearson correlation coefficient matrix assayed on the 33,109 sample dataset (our fast C++ tool, *MATool*, is available upon request), attempting to simplify the structure of the data. Here, PCA represents a transformation of a set of correlated probes into sets of uncorrelated linear additions of probe expression signals (eigenvectors) that we name transcriptional components (TCs). Each TC is a weighted sum of probe expression signals and eigenvector probe coefficients. These TC scores can be calculated for each observed expression array sample (reflecting the TC activity per sample).

Subjects for expression-genotype correlation. We obtained peripheral blood DNA and RNA (PAXgene) from Dutch and UK individuals who were disease cases or controls for GWAS studies (**Supplementary Table 1**). All samples had been genotyped for a common SNP set on Illumina platforms. Analysis was confined to 294,767 SNPs that had a MAF ≥ 5%, call-rate ≥ 95% and exact HWE $P > 0.001$. RNA from the samples was hybridized to either Illumina HumanRef-8 v2 arrays (229 samples, Ref-8v2) or Illumina HumanHT-12 arrays (1,240 samples, HT-12), and raw probe intensity extracted using BeadStudio. The Ref-8v2 samples were jointly quantile normalized and $\log_2$ transformed, as were the HT-12 samples. Subsequent analyses were also conducted separately for these datasets, up to the eventual eQTL mapping, which uses a meta-analysis framework, combining eQTL results from both arrays. HT-12 and Ref-8v2 arrays are different, but share many probes with identical probe sequences. Illumina sometimes use different probe identifiers for the same probe sequences; in meta-analysis and **Table 3**, the label HT-12 was used if both HT-12 and Ref-8v2 had the same sequence.

Re-mapping of probes. If probes mapped incorrectly or cross-hybridized to multiple genomic loci, it might be that an eQTL would be detected that would be deemed a *trans*-eQTL. To prevent this, we used a mapping approach versus a known reference that we developed for high-throughput short sequence RNAseq data[44]. We took the DNA sequence as synthesized for each cDNA probe and aligned it against a transcript masked gDNA genome combined with cDNA sequences. A more detailed description of probe re-mapping is provided in a **Supplementary Note**. Probes that did not map or that mapped to multiple different locations were removed.

Affymetrix transcriptional components applied to Illumina expression data. TC scores can be inferred in new (non-Affymetrix) datasets for every new individual sample. For the Illumina samples (used for the *cis*-eQTL mapping), only Illumina probes that could be mapped to any of our 22,106 Affymetrix

probes were used (www.switchtoi.com/probemapping.ilmn). The TC score of sample *i* for the j^{th} TC is defined as: $TCscore_{ij} = \sum_{t=1}^{t=n} a_{ti} \times v_{tj}$, where v_{tj} is defined as the t^{th} Affymetrix probe coefficient for the j^{th} TC; a_{ti} is the Illumina expression measurement for the t^{th} mapped probe for sample *i*. We inferred the Illumina TC scores for the top 1,000 TCs.

Removal of transcriptional component effects from Illumina expression data. Because our Illumina eQTL dataset (n = 1,469) is much less heterogeneous than the Affymetrix dataset (n = 33,109), we expect that some TCs will hardly vary. We therefore performed a PCA on the covariance matrix of the top 1,000 inferred TC scores for the Illumina dataset to effectively compress the TC data into a small set of 'aggregate TCs' (aTCs). As aTCs are orthogonal, we used linear regression to eliminate the effect of the top 50 aTCs. We correlated the TC-scores for each peripheral blood sample with probe expression levels. We then used the resulting residual gene expression data for subsequent *cis*-eQTL mapping.

***cis*-eQTL mapping.** We used the residual gene expression data in a meta-analysis framework, as described[45,46]. In brief, analyses were confined to those probe-SNP pairs for which the distance from probe transcript midpoint to SNP genomic location was less than 500 kb. To prevent spurious associations due to outliers, a nonparametric Spearman's rank correlation analysis was performed. When a particular probe-SNP pair was present in both the HT12 and H8v2 datasets, an overall, joint *P* value was calculated using a weighted *Z*-method (square root of the dataset sample number). To correct for multiple testing, we controlled the false-discovery rate (FDR). The distribution of observed *P* values was used to calculate the FDR, by permuting expression phenotypes relative to genotypes 1,000 times within the HT12 and H8v2 dataset. Finally, we removed any probes from analysis which contained a known SNP (1000Genomes CEU SNP data, April 2009 release).

35. Franke, L. *et al.* Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am. J. Hum. Genet.* **82**, 1316–1333 (2008).
36. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
37. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
38. Yu, K. *et al.* Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551 (2008).
39. Risch, N.J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
40. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
41. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
42. Sherlock, G. Analysis of large-scale gene expression data. *Brief. Bioinform.* **2**, 350–362 (2001).
43. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
44. Heap, G.A. *et al.* Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Hum. Mol. Genet.* **19**, 122–134 (2010).
45. Heap, G.A. *et al.* Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Med. Genomics* **2**, 1 (2009).
46. Franke, L. & Jansen, R.C. eQTL analysis in humans. *Methods Mol. Biol.* **573**, 311–328 (2009).

Other treatments

Prolonged antibiotic therapy poses potential clinical problems including diarrhoea, enterocolitis, patient intolerance, and bacterial resistance. A prokinetic agent that could help clear the small intestine of the overgrowth flora would be an attractive therapy, and experimental animal studies suggest that this might be helpful. There have been two small studies of these agents in patients with SBBO, one utilizing cisapride and one using octreotide, both leading to positive results. Another study utilizing octreotide and erythromycin in patients with scleroderma and SBBO attained positive responses. Large controlled trials of prokinetic therapy in patients with SBBO have yet to be completed.

Since the days of Metchkinoff, it has been thought that one could manipulate the intestinal flora by giving live 'probiotic' microbial supplements that would change the balance in the intestinal flora. Studies to date with probiotic therapy in subjects with SBBO have been disappointing. A placebo-controlled, randomized crossover trial compared norfloxacin, amoxicillin–clavulanic acid, and *Saccharomyces boulardii* in 10 symptomatic patients with SBBO. Both antibiotic treatments led to significant decreases in symptoms and a substantial improvement in the results of hydrogen breath testing, but the probiotic treatment did not result in any improvement in these parameters.

Nutritional support is an important part of treatment of SBBO and may be needed despite attempts to control the bacterial overgrowth by antimicrobial agents because of irreversible damage to the enterocytes. A lactose-free diet and substitution of a large proportion of dietary fat by medium-chain triglycerides may be necessary. Patients with cobalamin malabsorption should receive monthly injections of cobalamin (1000 µg). Deficiencies of other nutrients such as calcium and vitamin K should also be corrected.

Further reading

Attar A, *et al.* (1999). Antibiotic efficacy in small intestinal bacterial overgrowth-related chronic diarrhea: a cross-over, randomized trial. *Gastroenterology*, **117**, 794–7.

Bishop WP (1997). Breath hydrogen testing for small bowel bacterial overgrowth—a lot of hot air? *J Pediatr Gastroenterol Nutr*, **25**, 245–9.

Bouhnik Y, *et al.* (1999). Bacterial populations contaminating the upper gut in patients with small intestinal bacterial overgrowth syndrome. *Am J Gastroenterol*, **94**, 1327–9.

Bratten JR, Spanier J, Jones MP. (2008). Lactulose breath testing does not discriminate patients with irritable bowel syndrome from healthy controls. *Am J Gastroenterol*,**103**, 958–63.

Corazza GR, *et al.* (1990). The diagnosis of small bowel bacterial overgrowth. *Gastroenterology*, **98**, 302–5.

Fried M, *et al.* (1996). Duodenal bacterial overgrowth during treatment with omeprazole in outpatients. *Gut*, **35**, 23–7.

King CE, Toskes PP (1986). Comparison of the 1-gram [^{14}C]xylose, 10-gram lactulose-H_2, and 80-gram glucose-H_2 breath tests in patients with small intestine bacterial overgrowth. *Gastroenterology*, **91**, 1447–51.

Lin HC (2004). Small intestinal bacterial overgrowth; a framework for understanding irritable bowel syndrome. *JAMA*, **292**, 852–858.

Postsserud I, *et al.* (2007) Small intestinal bacterial overgrowth in patients with irritable bowel syndrome. *Gut*, **56**, 802–8.

Rana SV, Bhardwaj B (2008). Small intestinal bacterial overgrowth. *Scand J Gastroenterol*, **43**, 1030–7.

Saltsman J, *et al.* (1994). Bacterial overgrowth without clinical malabsorption in elderly hypochlorhydric subjects. *Gastroenterology*, **106**, 615–18.

Singh VV, Toskes PP. (2003). Small bowel bacterial overgrowth: Presentation, diagnosis and treatment. *Curr Gastroenterol Rep*, **5**, 365–72.

Soudah H, Hasler W, Owyang C (1991). Effect of octreotide on intestinal motility and bacterial overgrowth in scleroderma. *N Engl J Med*, **325**, 1461–7.

Walters B, Vanner JS. (2005).Detection of bacterial overgrowth in IBS using the lactulose H_2 breath test: Comparison with ^{14}C D-xylose and healthy controls. *Am J Gastroenterol*, **100**, 1566–1570.

15.10.3 Coeliac disease

Patrick C. Dubois and David A. van Heel

Essentials

Coeliac disease is a common disorder of the small intestine in which specific proteins in dietary wheat, rye, and barley (gliadin, secalins, hordeins, usually referred to as 'gluten') induce T-cell responses restricted by HLA DQ2 or DQ8 that are central to the subsequent intestinal inflammation and loss of villous architecture that characterize the disease.

The condition presents most commonly either in early childhood or in the third or fourth decade of life. A 'classical' malabsorption syndrome characterized by diarrhoea, steatorrhoea, weight loss, fatigue, and anaemia may occur in severe cases, but is now rare: most patients have a milder constellation of symptoms such as abdominal discomfort, bloating, indigestion or nongastrointestinal symptoms (e.g. dermatitis herpetiformis), and many have no symptoms at all.

Diagnosis is made by serological testing for antitissue transglutaminase/antiendomysial antibodies, which have excellent sensitivity and specificity. About 1% of the (white European origin) population have positive coeliac serology, but many are undiagnosed. Positive serological tests should be followed by small intestinal biopsy, whilst a normal (gluten containing) diet is continued, looking for histological features of intraepithelial lymphocytosis, chronic immune cell infiltration of the lamina propria, loss of villous height (villous atrophy), and crypt hyperplasia.

Treatment is by strict avoidance of dietary wheat, rye, and barley (a gluten-free diet), which is safe and usually effective, but constitutes a major challenge for some people. Most patients (but not all) can eat pure oats. Screening for osteoporosis, vitamin D deficiency, and osteomalacia is advised, with treatment if indicated.

Intestinal complications include enteropathy-associated T-cell lymphoma, which should be considered particularly in older patients experiencing a clinical relapse in symptoms, despite effective gluten exclusion, after a prolonged period of clinical response. The overall prognosis of coeliac disease is excellent, but requires lifelong commitment to a gluten-free diet to reduce the risk of complications.

Introduction

Coeliac disease is a common (*c.*1% prevalence) inflammatory disorder of the small intestine occurring in both children and adults. Specific proteins in dietary wheat, rye, and barley (gliadin, secalins, hordeins, usually referred to as 'gluten') induce T cell responses restricted by HLA DQ2 or DQ8. These responses are central to the subsequent intestinal inflammation and loss of villous architecture that characterizes the disease (Fig. 15.10.1). Now that serological testing is widespread, symptoms observed in diagnosed individuals vary greatly and are often absent. Classical malabsorption is now infrequent, and only the most florid of the spectrum of presentations seen in coeliac disease. Strict avoidance of dietary wheat, rye, and barley (a gluten-free diet) usually induces remission. Disease reappears on re-challenge and dietary treatment is lifelong.

Historical perspective

Aretaeus (2nd century AD) gave the first recognizable account of coeliac disease (Greek: *koiliakos*, abdominal) describing steatorrhoea, that disease occurred in both children and adults, and that it was more common in women than men. Samuel Gee presented the first clear modern description of coeliac disease in 1888. Willem Dicke (1950) in his doctoral thesis entitled 'Investigation of the harmful effects of certain types of cereal on patients suffering from coeliac disease' outlined the modern treatment of a gluten-free diet. Dicke came to these observations in part by noticing that when wheat flour (i.e. bread) became scarce in the wartime Netherlands, children with coeliac disease paradoxically improved. John Paulley (1954) demonstrated using surgical operative specimens that villous atrophy occurs in the small-intestinal mucosa in coeliac patients. A technique enabling small-bowel biopsy by the oral route was first developed by Margot Shiner (1956), refined as the 'Crosby capsule' (1957), and subsequently replaced in the 1980s by fibre optic endoscopy. Shiner and Doniach (1960) were then able to show using light and electron microscopy the identical histology of adult idiopathic steatorrhoea and childhood coeliac disease. Marsh described the sequence of changes in small-intestinal histology, and a classification system. Duhring (1884) was the first to describe dermatitis herpetiformis, and the often coexisting coeliac small bowel changes were described by Marks and Watson (1966).

The cultivation of wheat in Europe began about 5000 years ago, and (with rye) it became more common in the diet with the introduction of crop rotation in the Middle Ages. Serological diagnostic tests became available in the 1960s (antigliadin antibodies) and 1970s (antireticulin antibodies), although they lacked specificity until the development of the antiendomysial antibody test (1984). The HLA association was recognized in 1972. Dieterich and colleagues (1997) identified tissue transglutaminase as the endogenous target of antiendomysial antibodies and the key autoantigen in coeliac disease.

Aetiology

Many of the immunological mechanisms by which dietary wheat (and to a lesser extent rye and barley) induce coeliac disease are now understood. Wheat gluten is partially digested, but key toxic protein sequences are resistant to intestinal proteases—in part due to high proline (P) and glutamine (Q) content. Tissue transglutaminase in the intestinal epithelium deamidates critical peptide sequences such as the dominant HLA DQ2 restricted wheat epitope sequence PQPQLPY to PQPELPY, and (cross-linked to critical wheat peptides during the deamidation step) is the antigen detected by current diagnostic serological tests such as the antiendomysial or tissue transglutaminase antibody assays. It is unclear if these antibodies have a pathological role in coeliac disease. Work using intestinal T cell clones, intestinal biopsy culture, and peripheral blood T cells in wheat antigen challenged coeliac patients, has shown that wheat peptides are presented by HLA DQ2 (or in a few patients DQ8) to CD4+ helper T cells. Immunodominant wheat (and rye, barley) epitopes that are capable of inducing T cell responses in almost all coeliac patients have been defined, and the crystal structure of these epitopes bound to HLA DQ2 or DQ8 has been elucidated. Activated T cells secrete interferon-γ, and other cytokines. Interleukin-15, expressed by intestinal epithelial cells and lamina propria macrophages, appears to activate intraepithelial lymphocytes and leads to epithelial cell killing. Multiple pathways lead to intestinal inflammation, villous atrophy and subsequent malabsorption.

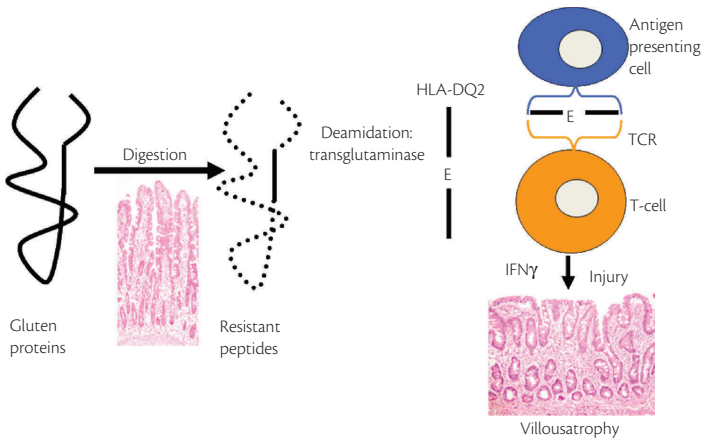


Figure 15.10.3.1 Model of gluten toxicity in coeliac disease. Toxic peptides in gluten are resistant to human digestive enzymes. Deamidation of key glutamine residues by mucosal tissue transglutaminase creates gluten epitopes with enhanced affinity for the peptide-binding groove of HLA DQ2. These gluten peptides are taken up by antigen presenting cells and presented by HLA DQ2 heterodimers to CD4+ T cells. Upon activation CD4+ T cells secrete interferon-γ and other cytokines and drive the intestinal inflammatory response.

The full HLA DQ2 heterodimer (encoded at the DNA level by the combination of HLA DQA1*0501 and DQB1*0201) is found in around 90% of coeliac disease patients, compared to around 30% of white European population controls. The remaining 10% of coeliac disease individuals either carry HLA DQ8, or part of the HLA DQ2 heterodimer. Carriage of one of these HLA types is therefore necessary but not sufficient to develop coeliac disease.

The HLA only explains around 30% of the heritable risk of coeliac disease; other genetic and environmental risk factors play a major role. Genetic risk variants on chromosome 4 (in a region containing the genes for the T-cell cytokines interleukin-2 and interleukin-21) as well as variants in other immune system genes have recently been identified. Several of these have independently been shown to influence risk to other autoimmune diseases, especially type 1 diabetes mellitus. The timing of the introduction of wheat during infant feeding is probably important, some studies suggesting that continued breastfeeding while weaning is protective. Whether gastrointestinal infections (e.g. rotavirus) in infancy are important triggers remains unclear.

Epidemiology

Prevalence estimates of clinically diagnosed coeliac disease (i.e. where symptoms lead to diagnostic testing) should be distinguished from population prevalence studies that employ serological screening. Most studies have been performed in populations of mainly white European origin, and used combined serological and intestinal biopsy testing. In these studies the prevalence of clinically diagnosed disease is around 0.1% (range 0.05% to 0.3%), whereas seroprevalence (including previously undiagnosed cases) in the general population is around 0.5 to 1% in both children and adults. Prevalence is even higher in close relatives of affected individuals; about 10% in first degree relatives. A large proportion of coeliacs in most populations remain undiagnosed—recently estimated at four out of five affected individuals in the United Kingdom. The highest population prevalence of 5% was found in Saharawi refugees living in Algeria. Coeliac disease occurs in Asians, but is extremely rare in individuals of tropical African, Japanese, and Chinese descent.

The similar United Kingdom population seroprevalence found in studies of children (1.0% in 5470 7-year olds) and adults (1.2% in 7550 over-45-year olds), suggests the coeliac trait is present from childhood in all cases, even those subsequently diagnosed as adults. Environmental trigger factors resulting in breakdown of oral tolerance to wheat, rye, and barley are therefore likely to occur in the first few years of life. The clinical observation that some adults suddenly develop symptoms in later life remains unexplained, but may reflect a later event in the control of immunological tolerance.

Clinical features

Although coeliac disease can be diagnosed at any age, it presents most commonly either in early childhood (between 9 and 24 months) or in the third or fourth decade of life. Coeliac disease is more common in females, with an approximately 2:1 sex ratio. Although the 'classical' gastrointestinal malabsorption syndrome characterized by diarrhoea, steatorrhoea, weight loss, fatigue, and anaemia may occur in severe cases, most patients nowadays have a milder constellation of symptoms such as abdominal discomfort, bloating, indigestion, or nongastrointestinal symptoms (or no symptoms at all). The clinical manifestation appears to be changing, with increasing numbers being diagnosed as a result of the investigation of iron deficiency (anaemia), fatigue and/or 'nonclassical' symptoms (Box 15.10.3.1).

Box 15.10.3.1 Clinical presentations in coeliac disease

With the advent of highly sensitive serological tests, coeliac disease is diagnosed in several settings.

- Classical: symptoms and clinical features of intestinal malabsorption—a relatively infrequent presentation in the developed world
- Atypical: minimal or no gastrointestinal symptoms. Coeliac disease suspected due to presence of associated features or conditions. Examples include iron and folate deficiencies, raised hepatic transaminases, osteoporosis, infertility, or short stature
- Silent: asymptomatic with no clinical manifestations of coeliac disease, diagnosed by serological screening or intestinal biopsy performed for another reason
- Latent: patients who may later develop coeliac disease, but who currently have normal intestinal mucosa on a gluten-containing diet. These include individuals with positive coeliac serology but normal intestinal biopsies

Although the natural history of the disease may be changing (possibly due to environmental factors), a more likely explanation for the current clinical manifestations is that the ability to make the diagnosis has improved (both better tests, and greater test accessibility) throughout the last 20 years with the development of accurate serological markers of the disease and increasing use of endoscopic biopsy techniques. Therefore a much broader spectrum of individuals are being investigated for coeliac disease and consequently being diagnosed (Fig. 15.10.3.2).

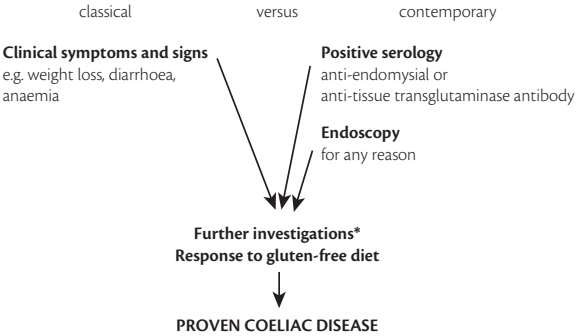


Fig. 15.10.3.2 Contemporary and classical diagnosis of coeliac disease. In the past, coeliac disease was mainly diagnosed after clinical presentation. Nowadays, many more patients are referred on the basis of positive serological tests. Endoscopy and 'routine' duodenal biopsy (without prior suspicion of coeliac disease) may also lead to diagnosis.
Adapted from Green PH, Rostami K, Marsh MN (2005). Diagnosis of coeliac disease. *Best Pract Res Clin Gastroenterol* 19, 389–400, and van Heel DA, West J (2006). Recent advances in coeliac disease. *Gut*, 55, 1037–46.

Intestinal complications

Refractory coeliac disease

This term is used for the small minority of patients (<5%) who show persistent histological features of coeliac disease with villous atrophy, despite apparently strict exclusion of gluten. In some individuals, this occurs due to the development of an aberrant, premalignant intraepithelial lymphocyte population. Immunohistochemistry is helpful in distinguishing these patients (see below) from those with persistent villous atrophy without aberrant lymphocytes, who have a very low risk of progression to lymphoma.

Enteropathy-associated T cell lymphoma (EATL)

This is a rare complication of coeliac disease but should be considered particularly in older patients experiencing a clinical relapse in symptoms, despite effective gluten exclusion, after a prolonged period of clinical response. Symptoms may include anorexia, weight loss, abdominal pain, fever, night sweats, and diarrhoea.

Ulcerative jejunitis

This presents with small intestinal ulcerations and stricturing—a high index of suspicion should be maintained for the presence of an EATL, as lymphoma may also cause similar appearances, including benign-appearing ulcerations.

Small-bowel adenocarcinoma

The risk of small-bowel adenocarcinoma is increased in coeliac disease, but the absolute risk of this rare cancer is still very small.

Extraintestinal manifestations and associated conditions

Coeliac disease shares similarities with autoimmune diseases, even though the trigger for inflammation in the intestine is not an autoantigen, but dietary gluten. Coeliac disease may have multisystemic effects, thought to be immune-mediated phenomena, although the pathophysiology is unproven in most cases.

Skin

Dermatitis herpetiformis is an inflammatory skin condition characterized by pruritic papules and vesicles over extensor surfaces and IgA deposition in the dermal papillae adjacent to lesions. Histological features of coeliac disease are present on intestinal biopsy in nearly all patients, but only 20% have intestinal symptoms. Dermatitis herpetiformis responds to gluten exclusion, but this may take months to years. Dapsone provides relief of the intense pruritus associated with dermatitis herpetiformis within 2 or 3 days and can lead to healing of the skin lesions, but not cure, as lesions recur rapidly on discontinuation of therapy.

Liver

Mild elevations of hepatic transaminases are common in untreated coeliac disease, which resolve in most cases within 6–12 months of starting a strict gluten-free diet. Separately, there are also associations between coeliac disease and autoimmune liver disorders including autoimmune hepatitis and primary biliary cirrhosis. The progression of these autoimmune disorders in the presence of coeliac disease is unaffected by subsequent gluten exclusion. Although accounting for a small minority of coeliac patients with abnormal liver function tests, these diagnoses should be considered in patients whose abnormal liver function tests do not improve despite prolonged gluten exclusion.

Neurological

Malabsorption may rarely lead to neurological sequelae from vitamin deficiency: vitamin B_{12} deficiency may cause peripheral neuropathy and myelopathy; vitamin E deficiency can cause cerebellar ataxia or myopathy. Tetany may be seen with severe hypocalcaemia or hypomagnesaemia. Associations with coeliac disease have also been reported for several neurological disorders, notably cerebellar ataxia, peripheral neuropathy, and epilepsy, although most studies have been small or inconsistent. A large Swedish study that retrospectively compared the frequency of several neurological diseases in 14 000 coeliac cases and population controls, found an increased risk of polyneuropathy, but not of other neurological diseases including ataxia.

Other immune-mediated diseases

There is an approximately fivefold increased risk of autoimmune disorders in coeliac disease. Definite associations include type 1 diabetes mellitus, autoimmune thyroid disease, Sjögren's syndrome, and Addison's disease.

Miscellaneous

Several cross-sectional studies have shown that the prevalence of coeliac disease is increased (approximately fivefold) in individuals with Down's syndrome. In untreated coeliac disease, rates of miscarriage and infertility are increased, possibly due to undernutrition, but rates return to near normal following diagnosis and institution of a gluten-free diet.

Differential diagnosis

Several other small-intestinal diseases can cause villous atrophy (Box 15.10.3.2). However, most conditions bear only partial resemblance to coeliac disease and can usually be distinguished either through the clinical history or histologically on careful review. Response to treatment (gluten exclusion) plays an important part in confirming the diagnosis of coeliac disease and excluding other causes. Patients who do not show a clinical or histological response to a strict gluten-free diet warrant consideration of alternative diagnoses and complications of coeliac disease. As well as other causes of villous atrophy, many comorbid conditions may mimic symptoms of coeliac disease and other causes of malabsorption should be excluded. Conditions occurring more frequently in coeliac disease, that may have similar symptoms, include small intestinal bacterial overgrowth, secondary lactase deficiency, microscopic colitis, Crohn's disease, and ulcerative colitis.

Clinical investigation

Pathology

The coeliac lesion occurs predominantly in the proximal small intestine, reflecting the distribution of gluten encounter. Changes may be mild and patchy and for this reason it is recommended that multiple (>4) biopsies are taken from separate sites, usually by

Box 15.10.3.2 Non-coeliac-related causes of villous atrophy

- Autoimmune enteropathy
- Chronic ischaemic enteritis
- Common variable immunodeficiency
- Crohn's disease
- Eosinophilic gastroenteritis
- Giardiasis
- Graft vs host disease
- HIV enteropathy
- Nonsteroidal anti-inflammatory drug enteropathy
- Peptic duodenitis
- Post-chemotherapy intestinal mucositis
- Radiation enteritis
- Tropical sprue

upper gastrointestinal endoscopy from the second part of the duodenum. The classic histological features are intraepithelial lymphocytosis, chronic immune cell infiltration of the lamina propria, loss of villous height (villous atrophy), and crypt hyperplasia. These features may be graded according to a commonly used classification proposed by Marsh. Intraepithelial lymphocytosis is the earliest change, but specificity for the diagnosis of coeliac disease increases with the presence of the other accompanying features, particularly villous atrophy.

Immunohistochemistry for T-cell markers (CD3, CD8) and the epithelial integrin CD103 are of value in refractory coeliac disease in detecting an aberrant intraepithelial T cell population that can precede the development of overt lymphoma.

Haematological abnormalities

A variety of haematological abnormalities may occur, arising from haematinic deficiencies, hyposplenism, and autoimmune phenomena. IgA deficiency (2–3%) and non-Hodgkin's lymphoma (see below) are also more common in coeliac disease.

Anaemia occurs frequently with microcytosis due to iron deficiency, but folate deficiency is also common and may cause macrocytosis. Vitamin B_{12} levels are usually preserved, except in severe, long-standing disease with involvement of the whole small intestine. Pancytopenia may occur in these cases as a result of folate or vitamin B_{12} deficiency.

Leucopenia and thrombocytopenia may also occur rarely as an autoimmune phenomenon.

Thrombocytosis is common in coeliac disease and can occur as a result of iron deficiency or hyposplenism, but usually resolves with gluten exclusion.

Morphological red cell changes characteristic of functional hyposplenism (Howell–Jolly bodies, target cells, acanthocytosis) may be apparent on blood film. Hyposplenism (based on sensitive research techniques, such as pitted red cell counting) is common in adult coeliac disease, but is rare in children and may be more frequent in patients with associated autoimmune disorders. The cause of hyposplenism in coeliac disease is unknown. Most studies suggest hyposplenism does not revert after treatment with a gluten-free diet. The risk of infection due to hyposplenism in coeliac disease is likely to be increased, but to date there have been only a few studies. A modest increased risk of infections in all patients with coeliac disease has been suggested by a large Swedish cohort study examining hospital inpatient episodes. The increased risk is partly accounted for by a 2.5-fold increase in the rate of pneumococcal infections. Immunization against the encapsulated organisms *Haemophilus influenza* type b, *Streptococcus pneumonia*, and *Neisseria meningitidis* should be considered in those with blood film evidence of hyposplenism. However, as yet no studies evaluating the effectiveness of this approach in coeliac disease have been performed. Immunization against influenza should also be considered in older patients because of the risk of secondary bacterial infections.

IgA deficiency

This occurs more commonly in coeliac disease, affecting 2 to 3% of patients. Conversely, the prevalence of coeliac disease in IgA deficiency is also increased and may be as high as 8%. IgA deficiency is important in coeliac disease as it may be a cause of false negative IgA endomysial or tissue transglutaminase tests.

Biochemistry

Fat malabsorption occurs in classical coeliac disease, leading to steatorhoea and malabsorption of vitamins A, D, E, and K. Hypocalcaemia and hypomagnesaemia may occur due to vitamin D deficiency. Rarely coagulopathy with prolonged prothrombin time is seen due to vitamin K malabsorption. Serum albumin can be low in the setting of intestinal inflammation, but systemic inflammatory markers such as C-reactive protein or ESR are not usually raised.

Antibody tests

Antiendomysial antibody (EMA) and human recombinant tissue transglutaminase (TTG) antibody tests have about 95% sensitivity and specificity in untreated coeliac disease. These tests have superseded both antigliadin and antireticulin antibody tests which have much lower diagnostic accuracy. The sensitivity and specificity estimates for EMA and TTG antibody tests were obtained in studies with patients with classical histological changes on biopsy including villous atrophy. Diagnostic difficulties therefore may arise in patients with mild disease, who may have negative serology and only mild inflammatory (infiltrative) changes on biopsy. Such patients may still have clinical manifestations that respond to gluten exclusion. Intestinal biopsy should therefore be obtained in all patients with unexplained features consistent with coeliac disease even if antibody tests are negative. EMA is assayed by indirect immunofluorescence (most commonly against monkey oesophagus) whereas TTG antibody titres are measured by ELISA and provide a quantitative measure that may be useful in assessing patients' compliance with a gluten-free diet.

Radiology

Barium radiology (barium follow-through, enteroclysis) lacks sensitivity in coeliac disease and is rarely used in diagnosis, but is of value when complications are suspected (lymphoma, ulcerative jejunitis) or alternative diagnoses such as Crohn's disease need to

be excluded. Intestinal lymphoma usually has a diffuse pattern of bowel involvement and can be particularly difficult to diagnose. Barium studies in uncomplicated disease may show thickening of mucosal folds and flocculation, segmentation or clumping of barium. CT or MR cross-sectional imaging with enteroclysis is superior when complications are suspected, enabling assessment of the intestinal wall but also regional lymphadenopathy and extra-intestinal disease.

Wireless capsule enteroscopy

This technique has good sensitivity and specificity for the diagnosis of coeliac disease and may be considered where upper gastrointestinal endoscopy and duodenal biopsies are nondiagnostic, but suspicion of small-bowel pathology remains (e.g. iron deficiency). Wireless capsule enteroscopy also has a role in investigation of patients with refractory sprue to help exclude complications such as lymphoma, small-bowel adenocarcinoma, and ulcerative jejunitis. This may lead on to targeted biopsies of suspicious areas by laparoscopy or double balloon enteroscopy.

HLA DQ typing

Genetic testing for HLA DQ2/8 is valuable, but only as an exclusionary test. The absence of genes encoding subunits of the HLA DQ2 or DQ8 heterodimers has almost 100% negative predictive value. However, local laboratories vary greatly in the format in which results are reported, making this a confusing area, and clinicians without experience are advised to refer back to the laboratory to ensure correct interpretation. The test is particularly useful in those in whom the diagnosis remains uncertain after serological testing and intestinal biopsy.

Criteria for diagnosis

Definitive diagnosis is based on intestinal biopsy and the finding of characteristic histological features of coeliac disease, together with clinical improvement on a gluten-free diet. Published guidelines on diagnosis and treatment are listed below (see 'Further reading'). Upper gastrointestinal endoscopy and distal duodenal biopsy can be undertaken as an outpatient with local throat anaesthetic spray or intravenous sedation. An improvement in symptoms and nutritional parameters, including micronutrient deficiencies, occurs in most patients within months after commencing a gluten-free diet and provides important confirmatory support for the diagnosis. Repeat intestinal biopsy after gluten exclusion to observe recovery of the intestinal mucosa is no longer considered necessary for diagnosis in adults, provided other objective indicators of response to gluten exclusion are observed (e.g. disappearance of positive coeliac antibody titres).

In patients with suspected coeliac disease who have commenced a gluten-free diet before a small-intestinal biopsy has been obtained and in whom serological tests and biopsies are nondiagnostic, biopsy after prolonged gluten challenge (equivalent to 4 slices of bread per day for at least 2 weeks) is helpful to confirm the diagnosis.

Treatment

Strict, lifelong gluten exclusion is the cornerstone of therapy and is effective in most individuals. The gluten-free diet is safe and usually effective, but constitutes a major challenge for some people because of the pervasiveness of these grains in modern diets and the paucity of palatable alternatives. Resolution of symptoms and nutrient deficiencies are the earliest markers of response. Bone density and other nutritional parameters such as body mass index and fat mass also increase, predominantly in the first year after starting a gluten-free diet. Subjective indices of well-being, such as self-reported vitality, may also improve. In children histological recovery is usually complete within a few months, but recovery in adults may be slower. Box 15.10.3.3 summarizes a typical course of treatment and additional investigation after diagnosis of coeliac disease.

Box 15.10.3.3 Action after diagnosis of coeliac disease

Initiate gluten-free diet

- Referral to a dietitian with suitable expertise
- Membership of a coeliac support society
- (In the United Kingdom: prescription of gluten-free foods)

Possible investigations for comorbid conditions

- Full blood count
- Iron studies, vitamin B_{12}, and folate
- Calcium, phosphate, patathyroid hormone, vitamin D
- Liver function tests
- Thyroid function tests
- Bone densitometry scan

Additional therapy

- Correct iron, vitamin B_{12}, folate deficiency
- Calcium and vitamin D supplements
- Pneumococcal, meningococcal, and *Haemophilus influenza* type b immunization in patients with hyposplenism

Resolution of positive EMA and TTG antibody titres provides a useful objective marker of response to gluten exclusion and usually occurs within 6 to 12 months. However, it should be remembered that these antibodies are commonly negative in the presence of low-grade histological abnormalities and are therefore limited markers of the extent of disease response. Monitoring of antibody tests, particularly quantitative TTG antibodies, is useful in patient follow-up to assess compliance. Major dietary indiscretions can lead to a rise in antibody levels, and can be helpful to reinforce efforts to improve compliance.

Compliance is also aided by joining a local coeliac society and by review with a dietitian with coeliac expertise. In the United Kingdom, Coeliac UK provides direct patient support and a comprehensive directory of gluten-free and gluten-containing food products. In general, wheat, rye, and barley should be avoided entirely. Feeding studies have established that pure oats are safe for most patients, but contamination of oat products with wheat gluten during harvesting or production is a common problem. A small number of patients appear to have a true coeliac intolerance to gluten-related avenins in oats. T-cell lines reactive to avenins in

oats can be generated from the intestinal mucosa of some of these patients.

It is unclear whether there is a safe amount of gluten that may be consumed without adverse effects, although for the majority even small amounts of gluten (50 mg/day) appear sufficient to cause ongoing intestinal inflammation. Individuals appear to vary considerably in their sensitivity to gluten. A few are exquisitely sensitive and even minimal amounts of gluten may provoke gastrointestinal symptoms and histological abnormalities. At the opposite end of the spectrum, some patients have no symptoms despite a normal gluten-containing diet.

Patients with coeliac disease show a modestly increased risk of osteoporosis and fractures. Hip fractures are increased nearly twofold, a significant concern given the high incidence of these fractures in ageing populations. The most effective intervention is the gluten-free diet, which improves bone density in coeliac disease, predominantly in the first year. Patients should be encouraged to undertake regular weight-bearing exercise, and advised on consuming adequate dietary calcium (*c.*1000 mg/day). Calcium supplements may be prescribed to meet these targets. Screening for osteoporosis with bone densitometry scanning should be considered, particularly in older patients who have the greatest risk of fractures and in those with other risk factors (low body mass index, weight loss, poor adherence to gluten-free diet). Patients at high risk of fractures, with osteoporosis determined by bone densitometry scanning, should receive appropriate supplementary therapies for osteoporosis including bisphosphonates.

Patients should be screened for vitamin D deficiency and osteomalacia. This may be suggested by hypocalcaemia, hypophosphataemia and raised alkaline phosphatase and is confirmed by serum 25-hydroxyvitamin D (calcidiol) assay. The British Society of Gastroenterology have produced guidelines on the management of low bone mineral density in coeliac disease. These guidelines recommend screening for secondary hyperparathyroidism as a surrogate marker of vitamin D deficiency, by measuring serum calcium and parathyroid hormone. Patients with a high parathyroid hormone level and normal calcium should receive supplementation with calcium and vitamin D (800–1000 units/day).

Persistent clinical symptoms

The commonest reason for recurrent or persistent clinical manifestations in coeliac disease is inadequate adherence to a gluten-free diet. This may be inadvertent, and a careful dietary review should be undertaken to assess presence of gluten in the diet. Symptoms may also commonly persist or recur due to the presence of comorbidities, which should be carefully sought and treated (see above).

Rarely patients have true refractory coeliac disease, if symptoms and histological features persist despite strict gluten exclusion over several months. Intestinal complications of coeliac disease, including enteropathy-associated T-cell lymphoma, should be considered and excluded in these patients (see above). It is worth remembering that the incidence of several gastrointestinal conditions that are not connected to coeliac disease, e.g. sporadic colorectal carcinoma greatly exceeds that of enteropathy-associated T-cell lymphoma and should also be excluded in patients with persisting symptoms.

Table 15.10.3.1 Estimates of relative and absolute risks in coeliac disease

	Relative risk	Absolute risk (incidence per 100 000 person-years)	
		General population	Coeliac
Any sepsis	2.6	58	139
Hip fracture	2.2	128	197
Any fracture	1.5	444	600
Lymphoma	5.9	8	45

Comparisons of events in the cohort of coeliacs (>10 000 cases) in the Swedish inpatient register with the general population. Includes first year after diagnosis.
Adapted from the analysis of Walters JRF, *et al.* (2008). Coeliac disease and the risk of infections. *Gut*, **57**, 1034–5.

Prognosis

Prognosis in coeliac disease is excellent, provided a prompt diagnosis is made and treatment instituted with strict adherence to a gluten-free diet. In long-term treated coeliac disease mortality is comparable to that of population controls.

The largest cohort studies point to an increased risk (*c.*twofold) of malignancy and mortality occurring within the first 2 or 3 years after diagnosis, although there is evidence for a sustained (*c.*sixfold) increased risk of lymphoproliferative disorders beyond this. It should be noted that absolute risks (i.e. at an individual patient level) of malignancy are small (Table 15.10.3.1). The increased risk appears to correlate with disease severity as it is highest in those with overt malabsorption but not detected in studies of patients with asymptomatic disease.

Screening

Screening for coeliac disease in asymptomatic individuals (including those at higher risk, e.g. with a family history or coexisting type 1 diabetes) remains controversial. The natural history of disease (especially risk of complications) in asymptomatic screening-detected cases is currently unknown, hence clear guidance on whether such individuals should commence a gluten-free diet cannot be given.

Likely developments over the next 5 to 10 years

Understanding of the heritable genetic risk factors predisposing to coeliac disease is rapidly increasing, driven by advances in genetics. Several new approaches to therapy are currently being developed or in early clinical trials. These include oral peptidase supplements designed to breakdown toxic cereal peptides, small molecules to inhibit various steps in pathogenesis (e.g. directed against transglutaminase, HLA DQ2, zonulin), and cereals genetically modified to reduce antigenicity.

Further reading

Halfdanarson TR, Litzow MR, Murray JA (2007). Hematologic manifestations of celiac disease. *Blood*, **109**, 412–21.

Hill ID, *et al.* (2005). Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. *J Pediatr Gastroenterol Nutr*, **40**, 1–19. [Provides paediatric guidance including information on presentation of disease in infancy and childhood, diagnostic and therapeutic approaches in children.]

Hunt KA, *et al.* (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet*, **40**, 395–402.

Kagnoff, MF (2006). AGA Institute Medical Position Statement on the Diagnosis and Management of Celiac Disease. *Gastroenterology*, **131**, 1977–80. [Provides practical clinical guidance for management of adults and children including internationally oriented dietary advice and list of useful websites.]

Ludvigsson JF, *et al.* (2008). Coeliac disease and risk of sepsis. *Gut*, **57**, 1074–80.

Rostom A, Murray JA, Kagnoff MF (2006). American Gastroenterological Association (AGA) Institute technical review on the diagnosis and management of celiac disease. *Gastroenterology*, **131**, 1981–2002. [Provides guidance in adults with coeliac disease. Strong focus on diagnosis, including difficulties encountered by physicians and use of serological tests.]

Scott, BB, Lewis NR (2007). *Guidelines for osteoporosis in inflammatory bowel disease and coeliac disease*. British Society of Gastroenterology (http://www.bsg.org.uk). [Provides practical guidance on targeted screening and treatment of osteoporosis in coeliac disease.]

Sollid LM (2002). Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol*, **2**, 647–55.

van Heel DA, West J (2006). Recent advances in coeliac disease. *Gut*, **55**, 1037–46.

WGO Celiac Disease Review Team (2007). *World Gastroenterology Organization Practice Guideline: celiac disease*. http://www.worldgastroenterology.org/assets/downloads/en/pdf/guidelines/04_celiac_disease.pdf

15.10.4 Gastrointestinal lymphoma

P.G. Isaacson

Essentials

Primary gastrointestinal lymphoma, which is the commonest extranodal lymphoma and almost exclusively of non-Hodgkin's type, is defined as lymphoma that has presented with the main bulk of disease in the gastrointestinal tract, with or without involvement of contiguous lymph nodes, and necessitating direction of treatment to that site.

MALT lymphoma describes a group of low-grade B-cell lymphomas whose histology recapitulates the features of mucosa-associated lymphoid tissue (MALT). It most commonly affects the stomach, presenting with nonspecific dyspepsia. Endoscopy typically shows inflamed or eroded mucosa rather than tumour mass. Many if not all cases appear to be driven by *Helicobacter pylori*, with 75% regressing following eradication of the organism with appropriate antibiotics. Deeply invasive lymphomas and those with adverse histological or cytogenetic features are unlikely to respond.

Enteropathy-associated T-cell lymphoma (EATL) is an intestinal tumour of intraepithelial T-lymphocytes that occurs most commonly in the jejunum or ileum and is sometimes associated with coeliac disease. It presents with abdominal pain, often due to intestinal perforation, and in some cases there is a prodromal period of refractory coeliac disease (sometimes accompanied by ulcerative jejunitis). The prognosis is usually poor, with death frequently resulting from abdominal complications in patients already weakened by uncontrolled malabsorption.

Burkitt's lymphoma is the most frequent childhood gastrointestinal lymphoma and is particularly common in the Middle East. B-cell lymphoproliferative conditions associated with immunodeficiency commonly present in the gastrointestinal tract and are increasingly important.

Introduction

The lymphomas that may arise in the gastrointestinal tract are listed in Box 15.10.4.1. Two of these, namely B-cell lymphoma of mucosa-associated lymphoid tissue (MALT) and enteropathy-associated T-cell lymphoma (EATL), do not arise in peripheral lymph nodes and will be discussed in more detail in this section. Any of the lymphomas that normally arise in lymph nodes may present as a primary gastrointestinal tumour, the most frequent being diffuse large B-cell lymphoma which, in fact accounts for the majority of primary gastrointestinal lymphomas, and mantle-cell lymphoma, which typically manifests in the gut as lymphomatous polyposis. Burkitt's lymphoma, is the commonest childhood gastrointestinal lymphoma, and is an especially common primary small intestinal lymphoma in the Middle East. The increasingly important group of B-cell lymphoproliferative conditions

Box 15.10.4.1 Primary gastrointestinal non-Hodgkin's lymphoma

B cell

- MALT lymphoma (including IPSID) with or without evidence of high-grade transformation
- Mantle-cell lymphoma (lymphomatous polyposis)
- Burkitt's lymphoma
- Other types corresponding to lymph node equivalents:

  follicular lymphoma
  lymphocytic lymphoma
  Diffuse large B-cell lymphoma
- Immunodeficiency-related lymphomas:

  post-transplant
  acquired (AIDS)
  congenital

T cell

- EATL
- Other types not associated with enteropathy

Rare types
(including conditions that may simulate lymphoma)