Inflammatory bowel disease

Fransen, Karin

Inflammatory bowel disease:
the genetic background and beyond

Karin Fransen

The studies presented in this thesis were supported by the Jan Kronelis de Cockstichting, Groningen University Istitute for Drug Exploration (GUIDE), Ubbo Emmius Fonds en Junior scientific masterclass

Printing of this thesis was financially supported by: University of Groningen, University Medical Centre Groningen, Groningen University Istitute for Drug Exploration (GUIDE), Nederlandse vereniging voor Gastroenterologie (NVGE), Scheperziekenhuis Emmen, Ferring Pharmaceuticals, and Dr. Falk Pharma Benelux B.V.

K. Fransen

Inflammatory bowel disease: the genetic background and beyond

PhD thesis – Department of genetics and Department of gastroenterology

University of Groningen, University medical Centre Groningen

Layout & printing by:  Lovebrid design & printing solutions
www.lovebird-design.com

Cover by: P.A. Fransen

ISBN: 978-90-367-6881-8 (print)

978-90-367-6882-5 (ebook)

rijksuniversiteit groningen

Inflammatory bowel disease

The genetic background and beyond

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken
en volgens besluit van het College van Promoties

De openbare verdediging zal plaatsvinden op

maandag 16 juni 2014 om 14.30uur

door

Karin Fransen
geboren op 29 maart 1984
te Emmen

Promotores

Prof. dr. C. Wijmenga

Prof. dr. R. K. Weersma

Co-promotor

Dr. C.C. van Diemen

Beoordelingscommissie

Prof. dr. K.N. Faber

Prof. dr. H.M. Boezen

Prof. dr. rer. nat. A. Franke

Paranimfen

Drs. S. Kloen

Drs. A.N. van der Meer

Table of contents

Chapter 1	Preface and outline of the thesis	**9**
Chapter 2	The quest for genetic risk factors for Crohn's disease in the post-GWAS era *Genome Med. 2011 Feb 25;3(2):13.*	**17**
Chapter 3	Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease *Hum Mol Genet. 2010 Sep 1;19(17):3482-8.*	**39**
Chapter 4	Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis. *Inflamm Bowel Dis. 2011 Nov;17(11):2287-94.*	**59**
Chapter 5	Limited evidence for parent-of-origin effects in Inflammatory Bowel Disease associated loci *PLoS ONE. Sep 2012. 7(9):e45287.*	**83**
Chapter 6	Correlation of genetic risk and mRNA expression in a Th17/IL23 pathway analysis in Inflammatory Bowel Disease *Inflamm Bowel Dis – in press*	**105**
Chapter 7	Summary and future perspectives	**133**
Appendices	English summary	**153**
	Samenvatting	**157**
	Dankwoord	**161**
	Curriculum Vitae	**166**
	List of publications	**167**

CHAPTER 1

Preface and outline of the thesis

Inflammatory bowel disease (IBD) presents in mainly two forms: Crohn's disease (CD) and ulcerative colitis (UC). They are characterized by a chronic relapsing inflammatory response likely to commensal microbes of the gut in a genetically susceptible host. CD can occur in the entire gastro-intestinal tract and the inflammation can be transmural and patchy, whereas UC by definition is restricted to the colon, starting from the distal end of the colon to a certain extension and is restricted to the mucosal layer. Given the transmural inflammation CD patients thus have more risk of fistuling and stenosing disease, whereas UC patients suffer more from bloody diarrhea. In both diseases anemia and weight loss occur. Also extra-intestinal manifestations like uveitis, primary sclerosing cholangitis and psoriasis are common in both entities. With a cumulative prevalence of up to 800 per 100,000 in Europe and 570 in North America [1], it is considered one of the most common immune-related diseases worldwide. The high prevalence combined with a peak age of onset in the second and third decade of life where career choices and starting a family are important steps make IBD a disease with a high impact on the quality of life of patients and leads to a high health care expenditure. Treatments are costly, often ineffective and may have severe side effects such as leucopenia. The pathogenesis is largely unknown, but from twin studies it has become apparent that IBD is a complex genetic disease meaning that multiple heritable factors and environmental factors contribute.

The heritable factors reside in the human blue-print which is discovered in 1869 by the biochemist Johann Friedrich Miescher. This desoxyribonucleic acid, or DNA, was first isolated from white blood cells. Almost 70 years later in 1952 Alfred Hershey en Martha Chase discovered that DNA contains the heritable properties of cells. Not even one year later the chemical structure of DNA was discovered by Rosalind franklin, James D. Watson and Francis Crick. During the following period research of genetics and disease focused mainly on Mendelian diseases, in which mutations in one (or sometimes a few) gene(s) cause disease. Identification of genomic regions associated to Mendelian diseases is done by linkage analysis in families, a powerful tool for the identification of highly penetrant genetic

variants in affected individuals. For CD the most replicated and strongest associated locus, *NOD2*, was identified by linkage analysis. However, complex genetic diseases, such as CD, are not suitable for linkage analysis since multiple genes underlie their pathogenesis. The number of involved variants in complex genetic diseases is by some experts estimated to be over a hundred indicating that the associated genetic variants individually are much less penetrant than in Mendelian diseases and not easy to pick up in linkage analysis. Thus new techniques were needed to identify these common variants associated to complex traits.

Because of technical limitations, it took scientists almost half a century after the discovery of the existence of DNA to correctly determine the full DNA sequence of the human genome. This lead to the identification of hundreds of thousands of single nucleotide polymorphisms (SNPs), one base pair variations in the genome sequence that occur in >1% of the population. With this knowledge it was possible to produce SNP arrays on which the base pair sequence of hundreds of thousands of those SNPs can be determined. In a Genome wide association study (GWAS) the allele frequency of all SNPs present on such a SNP array in cases is compared to the frequency in controls, which makes it is possible to identify associations between certain SNPs and a trait. GWAS have led to the discovery of an overwhelming amount of such associated loci for complex traits.

Inflammatory bowel disease (IBD) is one of the success stories of GWAS studies in complex genetic diseases. At the start of the research described in the current thesis 99 genetic loci were identified to be associated with IBD, many of which are overlapping for CD and UC but some are disease specific. Additional large scale studies using a custom made GWAS platform focused on immune-mediated diseases (Immunochip) increased this number to 163. Despite the great success accomplished with GWAS the total disease variance explained by these loci is for CD 13.6% and for UC 7.5%. Since heritability estimates are inconsistent between studies it is difficult to predict the variance in heritability that is currently explained. However, given the most frequently reported heritability estimate of 50%, for CD 26%

and for UC 15% of the total heritability is explained by the current GWA findings. Hence, despite the large number of independent associated loci the quest for the missing heritability still continues.

Post-hoc analyses of GWAS results have led to solid clues of the involvement of several pathways in the pathogenesis and new causative insights have been created. Since associated loci may contain none, one or multiple genes, the causality remains uncertain. Rivas *et al*. have shown coding (causal) variants in some instances by sequencing some of the associated loci in a large number of patients but for the fast majority of associated loci the causal gene is lacking, which complicates functional follow-up.

In this thesis, my aim is to give insight in the steps undertaken in the GWAS era and potential next steps in the post-GWAS era to elucidate the hidden heritability question and to gain insight in the underlying pathogenesis.

Chapter 2, 'The quest for genetic risk factors for Crohn's disease in the post-GWAS era' gives an overview of potential sources of the as above discussed, hidden heritability. Here we focus on Crohn's disease and provide potential next steps to be taken in studying the source of the remaining heritability.

Chapter 3 and 4 will focus on further exploration of GWAS results. In chapter 3 a method is described to select SNPs of GWAS results for replication that did not reach the genome significant threshold of $p<10^{-8}$. We hypothesized that many of these SNPs can be truly associated variants but were discarded due to the stringent correction for multiple testing in GWA studies. We performed a replication study by prioritizing those SNPs that are influencing gene-expression of nearby genes. (*cis*-eQTL SNPs)

It is known that immune-mediated diseases share considerable parts of their genetic risk loci. For instance type 1 diabetes and IBD or celiac disease and IBD share much of their genetic background. Hence, variants associated to one immune-mediated disease known to share genetic background with another are candidate variants for association testing in the other disease. *PTPN22* is a likely candidate gene for involvement in IBD since it is known to be associated

to multiple immune-mediated diseases. Chapter 4 will focus on the role of *PTPN22* in IBD and specific disease sub-phenotypes.

In chapter 5 epigenetic effects are discussed as a potential source of the hidden heritability of IBD. It is known that children from mothers with CD have more chance of developing the disease themselves compared to children from fathers with CD. This implies a role for parent of origin effects for IBD associated genes. Here we used novel statistical methodology that has previously been successful in identifying parent of origin effects in type 1 diabetes, to study parent of origin effects, in IBD associated loci.

Another important step in studying the role of genetics in complex disease is to determine the functional consequences of associated loci. In IBD there is an overrepresentation of genes involved in Th17 signaling within the associated loci. Therefore we focused on the influence of genetic variants on gene expression, given both functional and genetic importance of the Th17 pathway, we focused In chapter 6 I will discuss a method with which we extensively investigate the IBD associated T helper 17 IL23R pathway. In our study we determine the effect of risk load (i.e. a measure of number of risk alleles in an individual) on this differential gene-expression in peripheral blood mononuclear cells and we perform a co-expression analysis.

In Chapter 7 an overview of the results of the studies is given and future perspectives discussed.

CHAPTER 2

The quest for genetic risk factors for Crohn's disease in the post-GWAS era

Karin Fransen, Mitja Mitrovic*, Cleo C van Diemen and Rinse K Weersma*

Genome Med. 2011 Feb 25;3(2):13.

**These authors contributed equally*

Abstract

Multiple genome-wide association studies (GWASs) and two large scale meta-analyses have been performed for Crohn's disease and have identified 71 susceptibility loci. These findings have contributed greatly to our current understanding of the disease pathogenesis. Yet, these loci only explain approximately 23% of the disease heritability. One of the future challenges in this post-GWAS era is to identify potential sources of the remaining heritability. Such sources may include common variants with limited effect size, rare variants with higher effect sizes, structural variations, or even more complicated mechanisms such as epistatic, gene-environment and epigenetic interactions. Here, we outline potential sources of this hidden heritability, focusing on Crohn's disease and the currently available data. We also discuss future strategies to determine more about the heritability; these strategies include expanding current GWAS, fine-mapping, whole genome sequencing or exome sequencing, and using family-based approaches. Despite the current limitations, such strategies may help to transfer research achievements into clinical practice and guide the improvement of preventive and therapeutic measures.

Background

Crohn's disease (CD) is one of the two main forms of inflammatory bowel disease (IBD), the other being ulcerative colitis (UC). It is a chronic disease characterized by recurring inflammation of the gut, and is thought to arise in response to the commensal microflora in a genetically susceptible host [1]. It can affect the entire gastrointestinal tract, although the most common locations are the terminal ileum and the colon. Symptoms can be diffuse, and include (bloody) diarrhea, abdominal discomfort, weight loss and anemia, and there may also be extra-intestinal symptoms such as arthritis, and eye and skin disorders. Complications such as strictures often occur in CD, and since the inflammation is transmural, fistulas and abscesses can develop, and these eventually require surgical treatment [2]. Most of the medications have significant side effects, and they are expensive, and often ineffective. CD is a major burden on healthcare services, with a prevalence of 100 to 150 cases per 100,000 persons per year in the western world and with a peak age of onset between 10 and 30 years of age [3]. CD is partly heritable; this is reflected in the higher concordance rate in monozygotic twins compared with dizygotic twins. The concordance for CD in dizygotic twins is 4%, and for monozygotic twins it is as high as 56% [4].

Prior to the introduction of genome-wide association studies (GWASs), only a few genetic factors (for example, *NOD2*, which encodes nucleotide binding oligomerization domain 2) had unequivocally been associated with CD. However, multiple GWASs have now been performed for CD, and a recent meta-analysis carried out by Franke *et al.* [5] has unveiled 71 genetic variants as associated with CD; Table 1 highlights some noteworthy genes from that study.

Table 1. Notable genes within regions associated with Crohn's disease

Gene	Odds ratio (95% CI)	Function
Innate immunity		
NOD2 (nucleotide binding oligomerization domain 2)	2.2-4.0 [58]	Involved in pattern recognition
ATG16L1 (ATG16 autophagy related 16-like 1)	1.34 (1.29-1.40) [5]	Involved in autophagy
IRGM (immunity-related GTPase family, M)	1.37 (1.28-1.47) [5]	Involved in autophagy
TLR4 (Toll-like receptor 4)	1.29 (1.08-1.54) [59]	Involved in pattern recognition
CARD9 (caspase recruitment domain family, member 9)	1.18 (1.13-1.22) [5]	Involved in pattern recognition
VAMP3 (vesicle-associated membrane protein 3)	1.05 (1.01-1.10) [5]	Involved in autophagy and TNF-α metabolism
REL (reticuloendotheliosis viral oncogene homolog)	1.14 (1.09-1.19) [5]	Transcriptional activator of NF-kB
ERAP2 (endoplasmic reticulum aminopeptidase 2)	1.05 (1.02-1.09) [5]	Involved in peptide trimming upon NF-kB stimulation; required for the generation of HLA binding peptides
UBE2L3 (ubiquitin-conjugating enzyme E2L 3)	0.70 [15]	Ubiquitinates, among others, the NF-kB precursor
Adaptive immunity		
IL23R (IL-23 receptor)	2.66 (2.36-3.00) [5]	Activates Th17 cells
IL12B (IL-12β)	1.18 (1.13-1.24) [5]	Stimulates Th0 differentiation to Th1 cells
CCR6 (chemokine (C-C motif) receptor 6)	1.17 (1.12-1.22) [5]	Chemoattractant receptor of immune cells
HLA-DQA2 (major histocompatibility complex, class II, DQα2)	1.19 (1.13-1.25) [5]	Antigen presenting to Th0
TNFSF11 (tumor necrosis factor super family 11)	1.10 (1.05-1.15) [5]	Augments the ability of dendritic cells to stimulate naive T-cell proliferation
TNFSF15 (tumor necrosis factor super family 15)	1.21 (1.15-1.27) [5]	Mediates activation of NF-kB
ICOSLG (inducible T-cell co-stimulator ligand)	1.18 (1.13-1.23) [5]	Acts as a costimulatory signal for T-cell proliferation and cytokine secretion
IL2RA (IL receptor α)	1.11 (1.05-1.16) [5]	Th0 activation

TAGAP (T-cell activation GTPase-activating protein)	1.10 (1.05-1.14) [5]	May function as a GTPase activating protein and may play important roles during T-cell activation
IL10 (IL-10)	1.12 (1.07-1.17) [5]	Inhibits synthesis of pro-inflammatory cytokines
IL18RAP (IL-18 receptor accessory protein)	1.19 (1.14-1.26) [5]	Protein required for NF-kB activation
TYK2 (tyrosine kinase 2)	1.12 (1.06-1.19) [5]	Probably involved in intracellular signal transduction by initiation of IFN signaling
JAK2 (Janus kinase 2)	1.18 (1.13-1.23) [5]	Involved in JAK/STAT pathway; mediates signal transduction of many cytokines
STAT3 (signal transducer and activator of transcription 3)	1.15 (1.10-1.21) [5]	Involved in JAK/STAT pathway; mediates signal transduction of many cytokines
SMAD3 (SMAD family member 3)	1.12 (1.07-1.16) [5]	Involved in Treg activation through TGF-β signal transduction
ICAM1,3 (intercellular adhesion molecule)	1.12 (1.06-1.19) [5]	Homing of leukocytes to inflammation
Other genes of interest		
MUC1,19 (mucin)	1.74 (1.55-1.95) [5]	Involved in mucus production, to protect the epithelial barrier
FUT2 (fucosyltransferase 2)	1.07 (1.04-1.11) [5]	Involved in the A and B antigen synthesis pathway
PUS10 (pseudouridylate synthase 10)	1.16 [19]	Post-transcriptional nucleotide modification of structural RNAs, including tRNA, rRNA and sRNAs

Genes that we consider to be noteworthy in the Crohn's disease associated loci. Further investigation is necessary to identify the causal variants.
CI, confidence interval; HLA, human leukocyte antigen; IFN, interferon; IL, interleukin; JAK, Janus kinase; NF, nuclear factor; rRNA, ribosomal RNA; sRNA, splicing RNA; STAT, signal transducer and activator of transcription; TGF, transforming growth factor; Th, T helper cell; TNF, tumor necrosis factor; Treg, regulatory T cell; tRNA transferRNA.

Many of the genes cluster in several different molecular pathways and gene networks. In particular, results from GWASs have indicated the importance of the immune system in disease pathogenesis by identifying genes involved in innate and adaptive immunity. Hence, the association of *IRGM*, encoding immunity-related GTPase family M, and *ATG16L1*, encoding autophagy-related 16-like 1, with CD has

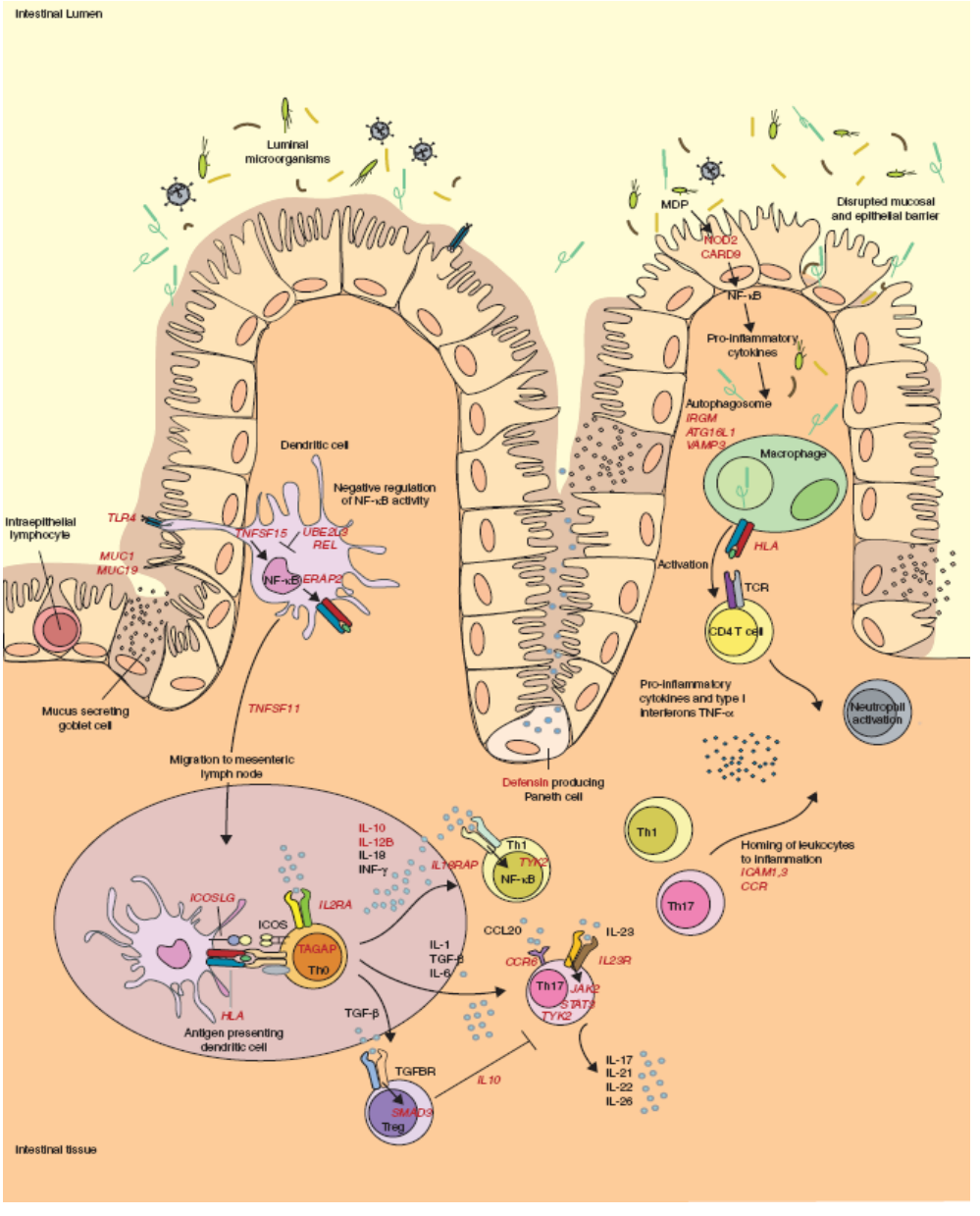


Figure 1. Schematic representation of the genes and pathways associated with Crohn's disease pathogenesis. The ongoing inflammatory response in the gastrointestinal tract in patients with Crohn's disease (CD) is thought to be caused by an aberrant immune response to commensal microflora in the gut. In patients with CD, defects in first defense mechanisms (that is, disrupted epithelial and mucosal barrier) contribute to increased bacterial penetration (MUC1 and MUC19). Genes involved in pattern recognition (NOD2, TLR4 and CARD9) suggest an increased response of antigen-presenting cells to commensal microbes. Consequently, the NF-kB cascade is activated (TNFSF15), leading to production of pro-inflammatory cytokines.

Association of REL and UBE2L3 suggest an impaired NF-kB negative feedback. Antigen-presenting cells migrate to Peyer's patches (intestinal mesenteric lymph nodes) (TNFSF11) to present antigens and stimulate T-cell proliferation (IL2RA and TAGAP) and differentiation. T cells of patients with CD, in turn, respond more intensely. Th0 cells are stimulated to differentiate into T-cell subtypes regulated by a variety of the produced cytokines and their receptors. Th17 cells are involved in many immune-related diseases, and they are activated through IL-23R, which, in turn, activates the JAK-STAT-TYK (Janus kinase-signal transducer and activator of transcription-tyrosine kinase) pathway that enhances pro-inflammatory cytokine production (JAK2, STAT3 and TYK2). Th1 and Th17 cells are pro-inflammatory, whereas Treg cells downregulate the immune response. Another major contribution to CD pathogenesis comes from autophagy. In autophagosomes, intracellular components, including phagocytosed microbes, are degraded, after which their antigens are presented to CD4+ cells. Autophagy is at least partly regulated by the CD risk genes ATG16L1, IRGM and VAMP3. The activation of CD4+ cells leads to the production of pro-inflammatory cytokines and the maintenance of the inflammation. All the displayed processes could finally lead to homing of leukocytes to inflammation sites (ICAM1,3, CCR cluster), and neutrophil recruitment. Consequently, chronic inflammation, ulceration and deeper microbial penetrance occur. The known associated genes are shown in red. Table 1 summarizes the associated loci shown here. CCL20, chemokine (C-C motif) ligand 20; ICOS, inducible T-cell co-stimulator; MDP, muramyl dipeptide; NF, nuclear factor; TCR, T-cell receptor; TGF, transforming growth factor; TGFBR, TGF ß receptor; Th, T helper cell; TNF, tumor necrosis factor; Treg, regulatory T cell.

implicated the process of autophagy [6]. The association of *NOD2*, *CARD9*, which encodes caspase recruitment domain family member 9, and *TLR4*, which encodes Toll-like receptor 4, indicates the involvement of pattern recognition mechanisms of the innate immune system [7]. Other genes are involved in pro-inflammatory pathways (T helper 1 cells and T helper 17 cells) and in anti-inflammatory pathways (regulatory T cells and IL-10), indicating that adaptive immunity also plays a role in CD pathogenesis (Figure 1) [8].

Another interesting association mapped to the *FUT2* gene, which encodes secretor type fucosyltransferase and regulates secretion of A and B blood group antigens in intestinal mucosa [9]. Recent functional studies have suggested that fucosylation of mucin proteins is involved in interception and exclusion of bacteria; thus, association of *FUT2* with CD might imply a role for the functional state of mucin in CD pathogenesis [10]. Although 5 years of GWASs have identified a substantial number of CD susceptibility loci, as much as 77% of the estimated heritability for CD is still considered to be unexplained [5].

Thus, one of the current challenges in the study of CD, like other complex diseases, is to identify potential sources of this hidden heritability. These might be additional common variants with

very limited effect size, or rare variants with a higher effect size. Part of the hidden heritability may lie in structural variations such as copy number variations (CNVs; a type of structural DNA sequence alteration, including deletions, duplications, insertions and inversions, that results in varying numbers of copies of a particular gene or DNA sequence from one person to the next) or even more complicated mechanisms, such as epistatic, gene-environment and epigenetic interactions. In this review, we discuss the known genetic risk factors for CD, the potential sources of the hidden heritability, and strategies to investigate these.

Further exploration of GWAS results

Thus far, the GWASs performed for CD have implicated many genes, and have thereby provided valuable insights into the etiology of CD. However, there are several ways to explore GWAS results in more depth that might lead to solving a part of the hidden heritability puzzle. The design of GWASs holds several limitations, with the first being the extensive correction needed for multiple testing. Hence, many true-positive findings are discarded because of the stringent significance thresholds, and large amounts of data are therefore ignored. Several methods have been applied successfully to overcome this statistical power issue. A major step to overcoming this problem has been taken by the International IBD Genetics Consortium (IIBDGC) [11], which performed a novel meta-analysis of six index GWASs and a follow-up study in independent cohorts. This study increased the number of confirmed CD loci to 71, although the explained heritability only increased from 20% to 23% [5].

Another way to overcome the lack of power inherent in GWASs is to follow-up specific SNPs (variation in a single base in the DNA sequence; the most common type of variation in the human genome) identified by them. Following up the top 1,000 less-strongly associated loci, for example, could yield new true associations. Meta-analysis of these results with the results from the index GWASs leads to a gain of power, as shown by a study of celiac disease [12]. Another approach

is to prioritize genes from the top associated loci based on interaction or functional analyses. This has proven to be a successful strategy in rheumatoid arthritis, where genes were prioritized based on network analysis or interaction analysis [13]. For CD, Wang et al. [14] used a different prioritizing criterion based on pathway analysis and they uncovered a significant association between susceptibility to CD and the IL-12/IL-23 pathway, harboring 20 genes. Prioritizing SNPs based on their effect on gene expression (for example, expression quantitative trait locus, a locus at which genetic allelic variation(s) correlates with variation in gene expression) led to identification of potentially novel associations of CD with UBE2L3, encoding ubiquitin-conjugating enzyme E2L 3 (involved in ubiquitinating the NF-kB precursor), and BCL3, encoding B-cell lymphoma 3-encoded protein (involved in down regulation of the NF-kB pathway) [15].

Results of GWASs and their meta-analyses have revealed that multiple autoimmune diseases have a common genetic architecture [16]. Several studies have been successful in identifying new CD risk variants by testing previously established loci for other immune-related diseases [17,18]. Festen et al. [19] developed a new method to identify shared risk loci of two immune-mediated diseases with a partially shared genetic background, namely celiac disease and CD. To increase the statistical power, they performed a combined analysis of GWAS results from celiac disease and CD, and identified TAGAP, which encodes T-cell activation GTPase-activating protein, and PUS-10, which encodes tRNA pseudouridylate synthase, as new shared loci [19].

The second limitation of the GWAS design is that it does not lead to the identification of causal variants, since the tested SNPs are merely tagging SNPs in linkage disequilibrium (LD; a non-random association of alleles at two or more loci as a result of a recent mutation, genetic drift, selection, or non-random mating) with the causal variants. Therefore, the effect sizes of known CD loci may be an underestimation of their actual relative risk. To further investigate the known risk loci and identify new SNPs, either as causal or close-to-causal variants, extensive fine-mapping is currently being performed by the IIBDGC using a custom-made GWA chip. In addition, cross-

ethnicity fine-mapping has proven successful in exploring conserved haplotype structures (that is, LD blocks) [20]. The most common LD blocks occur in all populations; however, their frequencies vary among different ethnicities [20]. For example, common *NOD2* and *IL23R* variants that are well established in Caucasians could not be replicated in an Indian population, implying that additional variants in these or other candidate genes may play a role in the pathogenesis of CD in Indians [21]. This principle was also successfully applied in analyzing the IL2/IL21 LD block, which is strongly conserved in Caucasians as opposed to Han Chinese, in which the IL2 and IL21 genes reside on two distinct LD blocks. Both *IL2* and *IL21* could be identified as separate UC risk loci in Han Chinese [22].

Park et al. [23] proposed a method to evaluate statistical power and risk prediction of future GWASs. They estimated that there are, in total, 142 CD susceptibility loci with effect sizes similar to the loci reported in the current GWASs, and that a sample size of approximately 50,000 would be needed to uncover them. However, even if a GWAS with hundreds of thousands of cases were to provide new CD susceptibility loci and explain more of the genetic variance, it seems unlikely that it would capture even half of the estimated heritability since 142 loci only explain 20% of the sibling relative risk for CD. We can speculate that identification of the true causal variants could amplify the effect size for some of the known loci and could consequently increase the discriminatory power of risk models.

Another potential source of hidden heritability could lie in sample mix-ups that occur accidentally during sample collection, genotyping or data management. Some genetic variants influence gene expression phenotypes (expression quantitative trait loci); this allows checking for concordance between phenotypic measurements and genetic variants that affect these phenotypes. Westra et al. (personal communication) found that 3% of sample mix-ups decrease the number of loci normally discovered by 23% for a trait with a heritability of 50% and 500 loci explaining the total heritability. Thus, sample mix-ups may explain part of the hidden heritability and it will be possible to detect them as long as databases encompass sufficient numbers of phenotypes that are strongly determined by known genetic variants.

GWASs are most likely to remain an important approach for investigating the hidden heritability, since the potential of their results can be enhanced by: performing meta-analyses (for example, between multiple GWASs or between similar disease phenotypes); following-up prioritized SNPs based on pathway, functional or interaction analyses; studying SNPs that have been associated with other immune-related diseases; and expanding the design of GWASs to include samples from non-Caucasians.

Low frequency and rare variants

Common variants identified by GWASs represent only a small fraction of the phenotypic variation. Thus, much speculation about the hidden heritability has focused on the contribution of variants with low allele frequencies, defined as 0.5% < minor allele frequency (MAF; proportion of the less common of two alleles in a population) < 5%, or from rare variants with MAF <0.5%, that are not sufficiently frequent to be captured by current GWA arrays, nor sufficiently penetrant to be captured by traditional, family-based linkage studies [24]. Detecting such variants will be facilitated by advances in high-throughput sequencing technologies and by the wide-ranging catalog of variants with MAF >1% generated by the 1000 Genomes Project [25]. Current efforts to identify rare variants by sequencing are likely to focus on the regions of most significant GWAS SNPs and around genes already implicated in CD pathogenesis or treatment. Resequencing of selected susceptibility loci has led recently to the discovery of three IL23R (the gene encoding (IL-23 receptor) coding variants that offer protection against CD [26]. The results of this particular study confirmed an increase in effect size with decreasing variant frequency, although rare variants explained less of the heritability than common variants.

In addition to resequencing efforts, whole-genome/exome sequencing will be needed to detect rare high-risk variants beyond the LD reach of tag SNPs. Although the costs of next-generation sequencing remain high, they are dropping fairly rapidly as the technologies improve and the process time per sample is becoming

shorter; so this method is becoming more and more feasible and accessible for researchers. Evaluating such signals and determining the real causal variant will, however, be a difficult task. Feng and Zhu [27] developed an alternative method for searching for rare variants in previously published GWAS datasets. Their method relies on haplotype analysis across the genome and the hypothesis that multiple rare variants can be captured by many haplotypes. Using this method, they confirmed nine previously established loci and also discovered four new CD susceptibility loci [27].

Another approach that may prove to be important is performing resequencing studies of individuals with extreme phenotypes in lipid levels; these studies have shown that such individuals seem more likely to be the carriers of rare, yet non-synonymous, variants [28]. A large number of rare variants may have distinct effects on the phenotype. Therefore, pooling variants of similar effect and locus-specific matching of cases with specific CD subphenotypes and controls throughout the genome may help to reveal some of the hidden heritability [29].

Structural variation

It has been estimated that chromosomal rearrangements (that is, duplications, deletions, insertions and inversions), collectively named CNVs, comprise 12% of the human genome [30]. Currently, more than 15,000 CNV loci are catalogued in the Database of Genomic Variants [31]. Some CNVs have been linked to complex disorders, such as autism, neuroblastoma and systematic lupus erythematosus [32-34]. A recent study suggested that CNVs are enriched in genomic regions containing genes that influence immunity [35]. In particular, low and high copy numbers of the ß-defensin gene (*HBD2*), which acts as an antimicrobial peptide and as a cytokine, have been found to predispose to colonic CD [36,37]. Yet, in a recent study, Aldhous et al. [38] failed to replicate both of the previously published associations. Moreover, they argued that these two associations could be due to measurement error because of a general deficiency of real-time PCR

to distinguish multiple CNV clusters. In addition to the ß-defensins, a fine-mapping study of the *IRGM* susceptibility locus revealed a 20-kb deletion polymorphism immediately upstream of *IRGM* that was associated with CD risk and *IRGM* expression [39]. Furthermore, a recent GWAS of CNVs from the Wellcome Trust Case Control Consortium has confirmed these CNVs for CD, and also discovered new CNVs in the *IRGM* and human leukocyte antigen (5.1 kb) regions [40]. The Wellcome Trust Case Control Consortium study also showed that the most common CNVs are well tagged by SNPs in current GWAS chips, and that they are unlikely to make much contribution to the hidden heritability in common diseases. More work is needed to elucidate the functional consequences and impact of high copy-number repeats (for example, long interspersed nuclear elements), and of rare CNVs on clinical phenotypes, such as CD.

Family-based approaches

Since the possibility of chip-based GWASs became available linkage analysis and family-based approaches have been largely discarded. However, now that the opportunities for gene detection by conventional GWASs have been almost exhausted, researchers are shifting back towards family-based approaches. These approaches can be helpful when GWASs fail to detect signals from rare variants and are biased by population stratification, which is defined as a presence of subpopulations in a supposedly homogeneous population. Subpopulations arise from differences in allele frequencies between individuals as a consequence of distinct ancestral and/or demographic origin. Family-based studies may also be advantageous since the low frequency risk alleles (SNPs with MAF <5%) are likely to be more prevalent in large families with several affected members and should therefore be easier to detect. By assessing GWAS data in such families, large regions of identity-by-descent may be identified and found to include genes associated with CD; this approach has already proved to be a powerful tool in classical linkage analysis. However, the shared environment of family members is an alternative

explanation for familial clustering that should be taken into account. Glocker et al. [41] identified loss-of-function mutations in two loci by considering early onset colitis as a monogenic trait in two consanguineous families. They performed a genetic linkage analysis followed by candidate gene sequencing and identified the IL10RA (the gene encoding IL-10 receptor α) and IL10RB (the gene encoding IL-10 receptor ß) loci as being associated with early-onset enterocolitis. However, it is most likely that in this particular case a private variant, not present in the general population, is responsible for the disease.

Akolkar et al. [42] found that CD is subject to a parent-of-origin effect, indicating that loci affected by genomic imprinting play a role in CD pathogenesis. In genomic imprinting, the expression of an inherited variant is determined by the parent from whom that variant is inherited. If the maternal allele, for instance, is inactivated by genomic imprinting, then expression of the locus is determined by the paternal allele only. If this effect is not taken into account, a significant loss in the statistical power of the study might develop [43]. Family-based approaches may be useful in the search for the hidden heritability since low-frequency variants accumulate in families with multiple affected individuals; moreover, low-frequency variants are not affected by population stratification and they also include parent-of-origin effects. However, the causal variants identified in such families may prove to be private variants or the shared environment may play a major role.

GWAS aftermath: epistatic, gene-environment and epigenetic interactions

Given that a large proportion of the heritability of CD and its complex architecture is as yet unexplained, one might speculate other aspects of inheritance, such as epistasis, gene-environment interactions or epigenetic effects, might be involved. GWASs may be missing higher-order genetic effects that arise from the interaction of two or more SNPs [44]. The underlying idea for such epistatic effects is that a significant proportion of the hidden heritability is not due

to single common variants, nor to single rare variants, but rather to rare combinations of common variants. Since typical GWASs examine the association of single SNPs with a phenotype, SNPs that contribute epistatically will not be revealed by such an analysis. A recent pair-wise analysis of variants related to the IL17-IL23 pathway showed an increasing odds ratio for CD when the 'risk' haplotypes for these genes were combined [45]. Analysis of epistatic interactions in better-powered datasets, and the use of more efficient computational approaches that can account for the complex nature of biomolecular networks, may yield new genetic risk factors for CD [46,47].

An even more complex source for the hidden heritability might lie in gene-environment interactions, which are defined as the joint effect of one or more genes with one or more environmental factors that cannot be readily explained by their separate marginal effects [48]. The strongest and best replicated environmental risk factor for CD is smoking, which increases both the risk and severity of CD. However, a recent, moderately sized study found remarkable differences in associated loci between smoking and non-smoking CD patients, thereby implying that a complex gene-environment interaction must be at work [49]. Another example of the complex interaction between genetic and environmental factors was shown in a study by Cadwell et al. [50] where Atg16L1-deficient mice infected with a specific strain of norovirus developed CD-like phenotypes in a model of intestinal injury induced by dextran sodium sulfate. In particular, structural Paneth cell abnormalities and decreased production of antimicrobial granules in the mice resembled those found in CD patients who are homozygous carriers of the ATG16L1 risk alleles. Remarkably, the severity of intestinal injury induced by dextran sodium sulfate was not only dependent on aberrant Atg16L1 function and norovirus infection, but also on the timing of infection, secretion of the pro-inflammatory cytokines TNF-α and IFN-γ, and the presence of commensal bacteria in the mouse intestine.

Other environmental factors, such as appendectomy, diet and domestic hygiene habits, may also play a role in CD, but the evidence for each of these factors is much weaker. To study gene-environment

interactions will require careful consideration of the epidemiologic study design, exposure assessment, and methods of analysis, paying particular attention to ways of harmonizing these features across consortia.

An additional source of the hidden heritability might not lie in the genome sequence itself, but in subtle mechanisms interfering with genome functions, such as gene expression. These mechanisms include histone modification, methylation and gene inactivation, and are covered by the study of epigenetics. However, there is much controversy on this topic. Its role in CD is unknown, but there are some hints that methylation plays a role in other complex diseases: type 2 diabetes, rheumatoid arthritis and neurodegenerative diseases [51-53]. Epigenetics is also correlated with age, gender and nutrition, and it is likely that there are other environmental factors to be discovered [54,55]. It has been shown that changes in DNA methylation in mice can be provoked by dietary alterations and subsequently transmitted across generations [56]. Thus, sequence-independent epigenetic effects (beyond imprinting) that might be environmentally induced and transmitted across several generations [57] could represent a revolutionary glimpse into the enigmatic world of the heritability of complex diseases.

Conclusions

CD is a complex genetic disorder with an estimated heritability of 50% and it is characterized by a recurring inflammation of the gastrointestinal tract. Two decades of research have led to the discovery of 71 risk loci, which have improved our understanding of the disease pathogenesis.

At the moment, approximately 23% of the heritability can be explained. To fully understand the disease pathogenesis and link current insights to clinically relevant knowledge, it is important to continue our quest to identify more genetic risk factors in CD. In this review, we have presented various potential sources for the hidden heritability of complex diseases given the current knowledge on CD.

It is unlikely that conventional GWASs alone can solve the puzzle of the hidden heritability. They are not powerful enough to detect signals from common variants with low impact, nor extensive enough to capture rarer variants with high impact. The resources of GWASs are expected to be exhausted fairly soon, although new loci have recently been identified by replicating prioritized SNPs and meta-analysis of GWAS results.

Identification of causal variants may elucidate a substantial part of the hidden heritability; however, current GWASs are insufficient for the purpose of identifying causal variants since the identified SNPs are merely the surrogates for causal variants. However, fine-mapping can uncover SNPs closer to the causal variants, since SNPs can then be tested beyond the scope of GWASs. The true causal variants might be identified by whole genome sequencing or exome sequencing. More sources than the linear DNA sequence have to be investigated to unravel the total heritability. Epigenetics and gene-environment studies have been shown to be worthwhile, but the study of epistatic effects in CD is still needed, and results from other complex genetic diseases seem to be promising.

To fully unravel the hidden heritability of CD, collaborations between genome research centers are crucial, since the solutions to identify the hidden heritability are either costly or require a huge number of cases and controls. The IIBDGC is a good example of what can be achieved by performing large meta-analyses, and it is currently performing dense fine-mapping and replication studies to identify causal variants and additional risk loci in CD.

References

1. Nell S, Suerbaum S, Josenhans C: The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. Nat Rev Microbiol 2010, 8:564-577.
2. Baumgart DC, Sandborn WJ: Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet 2007, 369:1641-1657.
3. Logan I, Bowlus CL: The geoepidemiology of autoimmune intestinal diseases. Autoimmun Rev 2010, 9:A372-A378.
4. Brant SR: Update on the heritability of inflammatory bowel disease: the importance of twin studies. Inflamm Bowel Dis 2011, 17:1-5.
5. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, Anderson CA, Bis JC, Bumpstead S, Ellinghaus D, Festen EM, Georges M, Green T, Haritunians T, Jostins L, Latiano A, Mathew CG, Montgomery GW, Prescott NJ, Raychaudhuri S, Rotter JI, Schumm P, Sharma Y, Simms LA, Taylor KD, Whiteman D, et al.: Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010, 42:1118-1125.
6. Stappenbeck TS, Rioux JD, Mizoguchi A, Saitoh T, Huett A, Darfeuille-Michaud A, Wileman T, Mizushima N, Carding S, Akira S, Parkes M, Xavier RJ: Crohn's disease: A current perspective on genetics, autophagy and immunity. Autophagy 2010, 7:1-20.
7. Abraham C, Cho J: Inflammatory bowel disease. N Engl J Med 2009, 361:2066-2078.
8. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, et al.: Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008, 40:955-962.
9. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, et al.: Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet 2010, 19:3468-3476.
10. Linden SK, Sutton P, Karlsson NG, Korolik V, McGuckin MA: Mucins in the mucosal barrier to infection. Mucosal Immunol 2008, 1:183-197.
11. International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) [http://www.ibdgenetics.org]
12. Hunt KA, Zhernakova A, Turner G, Heap GA, Franke L, Bruinenberg M, Romanos J, Dinesen LC, Ryan AW, Panesar D, Gwilliam R, Takeuchi F, McLaren WM, Holmes GK, Howdle PD, Walters JR, Sanders DS, Playford RJ, Trynka G, Mulder CJ, Mearin ML, Verbeek WH, Trimble V, Stevens FM, O'Morain C, Kennedy NP, Kelleher D, Pennington DJ, Strachan DP, McArdle WL, et al.: Newly identified genetic risk variants for celiac disease related to the immune response. Nat Genet 2008, 40:395-402.
13. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, Guiducci C, Catanese JJ, Xie G, Stahl EA, Chen R, Alfredsson L, Amos CI, Ardlie KG; BIRAC Consortium, Barton A, Bowes J, Burtt NP, Chang M, Coblyn J, Costenbader KH, Criswell LA, Crusius JB, Cui J, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TW, et al.: Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat Genet 2009, 41:1313-1318.

14. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van Limbergen J, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H: Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet 2009, 84:399-405.

15. Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EA, Stokkers PC, van Bodegraven AA, Crusius JB, Hommes DW, Zanen P, de Jong DJ, Wijmenga C, van Diemen CC, Weersma RK: Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. Hum Mol Genet 2010, 19:3482-3488.

16. Zhernakova A, van Diemen CC, Wijmenga C: Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat Rev Genet 2009, 10:43-55.

17. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, Kugathasan S, Annese V, Dubinsky M, Rotter JI, Russell RK, Bradfield JP, Sleiman PM, Glessner JT, Walters T, Hou C, Kim C, Frackelton EC, Garris M, Doran J, Romano C, Catassi C, Van Limbergen J, Guthery SL, Denson L, Piccoli D, Silverberg MS, Stanley CA, Monos D, Wilson DC, Griffiths A, et al.: Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effect. Hum Mol Gen 2010, 19:2059-2067.

18. Danoy P, Pryce K, Hadler J, Bradbury LA, Farrar C, Pointon J; Australo-Anglo-American Spondyloarthritis Consortium, Ward M, Weisman M, Reveille JD, Wordsworth BP, Stone MA; Spondyloarthritis Research Consortium of Canada, Maksymowych WP, Rahman P, Gladman D, Inman RD, Brown MA: Association of variants at 1q32 and STAT3 with Ankylosing Spondilitis suggests genetic overlap with Crohn's disease. PLoS Genet 2010, 6:e1001195.

19. Festen EAM, Goyette P, Green T, Beauchamp C, Boucher G, Trynka G: A meta-analysis of genome wide association scans identifies TAGAP and PUS10 as shared risk loci for Crohn's disease and celiac disease. PLoS Genet 2011, 7:e1001283.

20. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D: The structure of haplotype blocks in the human genome. Science 2002, 296:2225-2229.

21. Mahurkar S, Banerjee R, Rani SV, Thakur N, Guduru VR, Duvvuru NR, Chandak GR: Common variants in NOD2 and IL23R are not associated with inflammatory bowel disease in Indian patients. J Gastroenterol Hepatol 2010, in press. doi: 10.1111/j.1440-1746.2010.06533.x

22. Shi J, Lu Z, Zhernakova A, Qian J, Zhu F, Sun G, Zhu L, Ma X, Dijkstra G Wijmenga C, Faber KN, Lu X, Weersma RK: Haplotype-based analysis of ulcerative colitis risk loci identifies both IL2 and IL21 as susceptibility genes in Han Chinese. Inflamm Bowel Dis 2010, in press.

23. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N: Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 2010, 42:570-575.

24. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex disease. Nature 2009, 461:747-753.

25. 1000 Genomes - A Deep Catalog of Human Genetic Variation [http://www.1000genomes.org]

26. Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L: Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet 2011, 43:43-47.
27. Feng T, Zhu X: Genome-wide searching of rare genetic variants in WTCCC data. Hum Genet 2010, 128:269-280.
28. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC: Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 2007, 39:513-516.
29. Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 2008, 83:311-321.
30. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, et al.: Global variation in copy number in the human genome. Nature 2006, 444:444-454.
31. Database of Genomic Variants - a curated catalogue of structural variation in the human genome [http://projects.tcag.ca/variation/]
32. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, et al.: Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 2009, 459:569-573.
33. Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K, Cole K, Mossé YP, Wood A, Lynch JE, Pecor K, Diamond M, Winter C, Wang K, Kim C, Geiger EA, McGrady PW, Blakemore AIF, London WB, Shaikh TH, Bradfield J, Grant SFA, Li H, Devoto M, Rappaport ER, Hakonarson H, Maris JM: Copy number variation at 1q21.1 associated with neuroblastoma. Nature 2009, 459:987-991.
34. Willcocks LC, Lyons PA, Clatworthy MR, Robinson JI, Yang W, Newland SA, Plagnol V, McGovern NN, Condliffe AM, Chilvers ER, Adu D, Jolly EC, Watts R, Lau YL, Morgan AW, Nash G, Smith KG: Copy number of FCGR3B, which is associated with systematic lupus erythematosus, correlates with protein expression and immune complex uptake. J Exp Med 2008, 205:1573-1582.
35. Schaschl H, Aitman TJ, Vyse TJ: Copy number variation in the human genome and its implication in autoimmunity. Clin Exp Immunol 2009, 156:12-16.
36. Fellermann K, Stange DE, Schaeffeler E, Schmalzl H, Wehkamp J, Bevins CL, Reinisch W, Teml A, Schwab M, Lichter P, Radlwimmer B, Stange EF: A chromosome 8 gene-cluster polymorphism with low human β-defensin 2 gene copy number predisposes to Crohn's disease of the colon. Am J Hum Genet 2006, 79:439-448.
37. Bentley R, Pearson J, Gearry R, Barclay M, McKinney C, Merriman T, Roberts R: Association of higher DEFB4 genomic copy number with Crohn's disease. Am J Gastroenterol 2010, 105:354-359.
38. Aldhous MC, Abu Bakar S, Prescott NJ, Palla R, Soo K, Mansfield JC, Mathew CG, Satsangi J, Armour JA: Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. Hum Mol Gen 2010, 19:4930-4938.
39. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, Zody MC, Hall JL, Brant SR, Cho JH, Duerr RH, Silverberg MS, Taylor KD, Rioux JD, Altshuler D, Daly MJ, Xavier RJ: Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 2008, 40:1107-1112.

40. The Wellcome Trust Consortium: Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010, 464:713-720.
41. Glocker EO, Kotlarz D, Boztug K, Gertz EM, Schäffer AA, Noyan F, Perro M, Diestelhorst J, Allroth A, Murugan D, Hätscher N, Pfeifer D, Sykora KW, Sauer M, Kreipe H, Lacher M, Nustede R, Woellner C, Baumann U, Salzer U, Koletzko S, Shah N, Segal AW, Sauerbrey A, Buderus S, Snapper SB, Grimbacher B, Klein C: Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. N Engl J Med 2009, 361:2033-2045.
42. Akolkar PN, Gulwani-Akolkar B, Heresbach D, Lin XY, Fisher S, Katz S, Silver J: Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. Am J Gastroenterol 1997, 92:2241-2244.
43. Hanson RL, Kobes S, Lindsay RS, Knowler WC: Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am J Hum Genet 2001 68:951-962.
44. Moore JH, Williams SM: Epistasis and its implications for personal genetics. Am J Hum Genet 2009, 85:309-320.
45. McGovern DP, Rotter JI, Mei L, Haritunians T, Landers C, Derkowski C, Dutridge D, Dubinsky M, Ippoliti A, Vasiliauskas E, Mengesha E, King L, Pressman S, Targan SR, Taylor KD: Genetic epistasis of IL23/IL17 related genes in Crohn's disease. Inflamm Bowel Dis 2009, 15:883-889.
46. Marchini J, Donnelly P, Cardon LR: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005, 37:413-417.
47. Cordell HJ: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 2009, 10:392-404.
48. Thomas D: Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 2010, 11:259-272.
49. Van der Heide F, Nolte IM, Kleibeuker JH, Wijmenga C, Dijkstra G, Weersma RK: Differences in genetic background between active smokers, passive smokers, and non-smokers with Crohn's disease. Am J Gastroenterol 2010, 105:1165-1172.
50. Cadwell K, Patel KK, Maloney NS, Liu TC, Ng AC, Storer CE, Head RD, Xavier R, Stappenbeck TS, Virgin HW: Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. Cell 2010, 141:1135-1145.
51. Maier, S and Olek A: Diabetes: a candidate disease for efficient DNA methylation profiling. J Nutr 2002, 132:2440S-2443S.
52. Kim, YI Logan JW, Mason JB, Roubenoff R: DNA hypomethylation in inflammatory arthritis: reversal with methotrexate. J Lab Clin Med 1996, 128:165-172.
53. Cara Terribas CJ, Gonzalez Guijarro L: Hypomethylation and multiple sclerosis, the susceptibility factor? Neurologia 2002, 17:132-135.
54. Issa JP: Epigenetic variation and human disease. J Nutr 2002, 132:2388S-2392S. 55. Ahuja N, Issa JP: Aging, methylation and cancer. Histol Histopathol 2000, 15:835-842.
56. Nadeau JH: Transgenerational genetic effects on phenotypic variation and disease risk. Hum Mol Genet 2009, 18:202-210.
57. Morgan HD, Sutherland HG, Martin DI, Whitelaw E: Epigenetic inheritance at the agouti locus in the mouse. Nat Genet 1999, 23:314-318.
58. Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP: Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. Am J Gastroenterol 2004, 99:2393-2404.
59. Shen X, Shi R, Zhang H, Li K, Zhao Y, Zhang R: The Toll-like receptor 4 D299G and T399I polymorphisms are associated with Crohn's disease and ulcerative colitis: a meta-analysis. Digestion 2010, 81:69-77.

CHAPTER 3

Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease

Karin Fransen, Marijn C. Visschedijk*, Suzanne van Sommeren, Jinyuan Y. Fu, Lude Franke, Eleonora A.M. Festen, Pieter C.F. Stokkers, Adriaan A. van Bodegraven, J. Bart A. Crusius, Daniel W. Hommes, Pieter Zanen, Dirk J. de Jong, Cisca Wijmenga, Cleo C. van Diemen and Rinse K. Weersma*

Hum Mol Genet. 2010 Sep 1;19(17):3482-8.

**These authors contributed equally*

Abstract

Genome-wide association studies (GWAS) for Crohn's disease (CD) have identified loci explaining ~20% of the total genetic risk of CD. Part of the other genetic risk loci is probably partly hidden among signals discarded by the multiple testing correction needed in the analysis of GWAS data. Strategies for finding these hidden loci require large replication cohorts and are costly to perform. We adopted a strategy of selecting SNPs for follow-up that showed a correlation to gene expression [cis-expression quantitative trait loci (eQTLs)] since these have been shown more likely to be trait-associated. First we show that there is an overrepresentation of cis-eQTLs in the known CD-associated loci. Then SNPs were selected for follow-up by screening the top 500 SNP hits from a CD GWAS data set. We identified 10 cis-eQTL SNPs. These 10 SNPs were tested for association with CD in two independent cohorts of Dutch CD patients (1539) and healthy controls (2648). In a combined analysis, we identified two cis-eQTL SNPs that were associated with CD rs2298428 in *UBE2L3* ($P = 5.22 * 10^{-5}$) and rs2927488 in *BCL3* ($P = 2.94 * 10^{-4}$). After adding additional publicly available data from a previously reported meta-analysis, the association with rs2298428 almost reached genome-wide significance ($P = 2.40 * 10^{-7}$) and the association with rs2927488 was corroborated ($P = 6.46 * 10^{-4}$). We have identified *UBE2L3* and *BCL3* as likely novel risk genes for CD. *UBE2L3* is also associated with other immune-mediated diseases. These results show that eQTL based pre-selection for follow-up is a useful approach for identifying risk loci from a moderately sized GWAS.

Introduction

Crohn's disease (CD) is a common, chronic, gastrointestinal inflammatory disorder with a prevalence of 100–200 per 100 000 in developed countries (1). The aetiology of CD is complex and is believed to originate in an aberrant immune response to the commensal intestinal bacterial flora in a genetically susceptible host (2).

Genome-wide association studies (GWAS) have already identified over 30 loci that convey risk for CD (3–8), representing 20% of the total genetic risk for this disease (8). The remaining 80% of genetic risk is probably partly made up by highly prevalent loci with very modest effect sizes and by rare loci with strong effect sizes. These remaining loci are hard to identify with a GWAS, in part because of the extensive multiple testing correction needed in GWAS analyses. This multiple testing correction is necessary to exclude false-positive loci, but simultaneously it discards many true-positive risk loci. Strategies for extricating these hidden true-positive loci include: increasing the GWAS sample size, performing a meta-analysis of GWAS data sets and replicating hundreds to thousands of GWAS signals in a larger cohort. Unfortunately, all of these methods still need substantial multiple testing correction and most are expensive to perform (9).

To cut down on the size of the follow-up study for a GWAS, and thus on the costs and need for multiple testing correction, we considered selecting SNPs for follow-up on the basis of a functional effect. In this study, we focus on the effect of SNPs on human gene expression levels which have been shown to have a strong heritable component (10). By treating gene expression as a quantitative trait, it is possible to correlate gene transcription levels with SNPs (expression quantitative trait loci, eQTLs) (10). SNPs can be correlated with the expression of genes located very near the SNP itself (cis-eQTL) or with the expression of genes located further away, even on other chromosomes (trans-eQTL). In this study, the maximal distance of a cis-eQTL SNP to a gene is 250 kb. Since the trans-eQTL effects are difficult to detect due to severe multiple testing issues, we chose to study cis-eQTL effects.

We hypothesized that SNPs affecting gene expression are more likely to be associated with CD than SNPs without such an effect,

Table 1a. Five out of 30 established Crohn's disease associated SNPs are *Cis*-eQTL SNPs

SNP	Chromosome	Risk allele	Expression Effect Risk allele	Effected gene	eQTL p-value
rs2301436	6q27	T	-	*RNASET2*	$6.52*10^{-05}$
rs2872507	17q12	A	-	*GSDML*	$5.20*10^{-09}$
rs3197999	3p21	A	+	*UBE1L*	$9.79*10^{-04}$
rs2872507	17q12	A	-	*ORMDL3*	$6,94*10^{-11}$
rs2188962	5q31	T	-	*SLC22A5*	$5.18*10^{-09}$

The 30 CD-associated loci in the meta-analysis conducted by Barrett *et al.* were tested on their *cis*-eQTL effects in the publicly available expression database used in our study (8,15) rs2872507 is correlated to the expression of two genes *ORMDL3* and *GSDML* (26)

Table 1b. *Cis*-eQTL effect of 13 identified SNPs within the top 500 of a publicly available GWAS dataset from the US-NIDDK Consortium.

SNP	Chromosome	Risk allele	Expression Effect Risk allele	Effected gene	eQTL p-value
rs6512121	19	G	+	*ZNF266*	$5.61*10^{-19}$
rs243323	16	G	+	*C16orf75*	$2.07*10^{-05}$
rs2298428	22	T	-	*UBE2L3*	$4.03*10^{-09}$
rs2066843	16	T	+	*CARD15*	$7.94*10^{-04}$
rs2927488	19	A	+	*BCL3*	$1.11*10^{-04}$
rs1156287	17	G	+	*COX11*	$5.35*10^{-06}$
rs9303363	17	A	+	*COX11*	$8.65*10^{-06}$
rs725660	19	C	+	*SYMPK*	$1.24*10^{-04}$
rs7142206	14	A	+	*ENTPD5*	$2.00*10^{-05}$
rs1005564	14	T	+	*ENTPD5*	$1.28 *10^{-05}$
rs3118663	9	G	-	*SURF1*	$5.37 *10^{-06}$
rs10278590	7	G	+	*RARRES2*	$2.10*10^{-04}$
rs359457	5	T	+	*CPEB4*	$6.12 *10^{-08}$

which provides a basis for selecting SNPs for replication. Cis-eQTLs have already been associated with several diseases, such as celiac disease and asthma (11,12). Our hypothesis is further supported by results from a recent GWAS in celiac disease in which a cis-eQTL

effect was seen in 20 out of 38 risk loci identified for celiac disease. Permutations showed that 50% of SNPs being cis-eQTLs were very unlikely to occur by chance and were not due to a bias of the genotyping platform used, nor to differences in minor allele frequency (MAF) (13). In a recent paper, Nicolae et al. (14) found that SNPs associated with complex traits are more likely to be eQTLs and that, by using this information, the discovery of complex disease-associated genes can be enhanced.

For this study, we first validated our hypothesis that cis-eQTL SNPs are overrepresented among the currently known CD-associated SNPs by comparing the amount of established CD-associated SNPs that are cis-eQTLs with the number of cis-eQTL SNPs expected by chance (Table 1) (8).

Next, a set of SNPs was selected for follow-up. We did this by comparing a list of CD risk SNPs with an cis-eQTL SNP database and aimed to identify novel CD-associated loci by selecting cis-eQTL SNPs from the top 500 hits from a publicly available CD GWAS (4). This resulted in 13 putative CD-associated eQTL SNPs, 10 of them were selected and studied in two independent cohorts of Dutch CD patients and controls (Fig. 1).

A combined analysis was performed using the data from the discovery GWAS and both our replication cohorts (4,15). A second separate meta-analysis was then performed using the data of another publicly available database of the CD meta-analysis conducted by Barrett et al. (8) and both our replication cohorts.

Results

CD-associated SNPs are more likely to be cis-eQTLs

To confirm our hypothesis that SNPs associated with CD are more likely to be eQTLs, we compared the amount of eQTL SNPs in the 30 established CD SNP with the amount expected by chance. Among the 30 top SNPs, five eQTLs were found ($P < 0.05$ corrected for FDR). We found after 100 permutations that this was higher than expected by chance ($P = 0.01$).

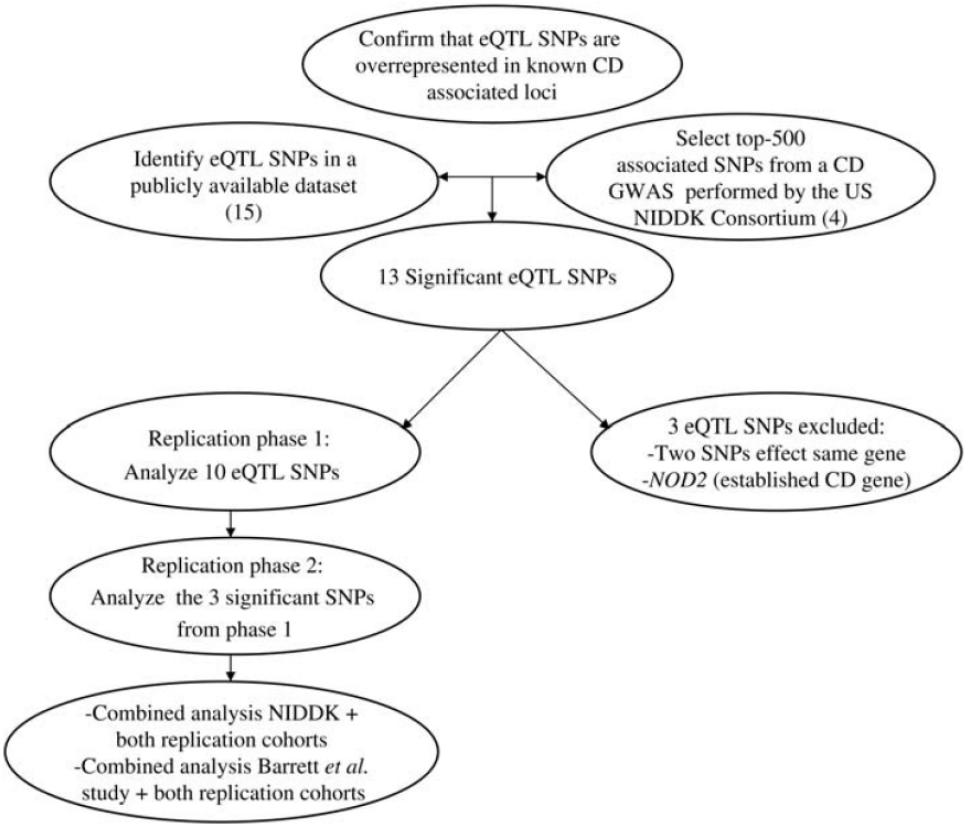


Fig.1 Study design.

Allelic association analysis

Results for the allelic association analysis for replication phases 1 and 2 are depicted in Tables 2 and 3. In the first replication phase, 10 SNPs were tested in a Dutch cohort of 777 CD cases and 964 healthy controls and we observed a significant association with CD for three SNPs. rs2298428 in UBE2L3 [P = 4.6 × 1024, odds ratio (OR) = 0.73, confidence interval (CI) 0.61–0.87], SNP rs2927488 in BCL3 (P = 0.011, OR = 0.80, CI 0.68–0.95) and rs725660 in SYMPK (P = 0.029, OR = 1.16, CI 1.01–1.32). In the second replication phase, we performed a follow-up analysis of these three SNPs in an independent cohort of 762 cases and 1648 controls. In this second cohort, we did not find any association for these SNPs (P = 0.70, 0.68, 0.50).

Combined analysis

The risk-increasing effect could be confirmed in a combined analysis including the original CD NIDDK GWAS data set and both our replication cohorts for two SNPs (UBE2L3 P = 5.22×1025, BCL3 P = 2.79×1024). The risk-increasing effect could not be confirmed for SYMPK with a P-value of 0.25. In a second combined analysis containing data of the CD meta-analysis by Barrett et al. (8) and both our replication cohorts, the risk-increasing effect could be confirmed for both SNPs (UBE2L3 P = 2.40 × 1027 and BCL3 P = 6.46 × 1024). For SYMPK the risk-increasing effect was not significant (P = 0.06) (Table 3). The meta-analysis performed by Barrett et al. contains the data of the GWAS used in the first combined analysis; to prevent overlap, this GWAS was excluded from the second combined analysis.

Risk alleles and expression

The eQTL SNP alleles associated with increased risk for CD had diverse effects on the expression of their correlated genes in a publicly available expression data set (15). For UBE2L3, the gene most strongly associated with CD, the minor allele that conferred risk was correlated with a higher expression of UBE2L3 (P = 4.21×1029) (Fig. 2A). In contrast, the risk variant of the BCL3-associated eQTL SNP was correlated with the lower expression of BCL3 P = 5.0×1025 (Fig. 2B).

Discussion

We have identified two novel potential risk genes for CD: UBE2L3 and BCL3. The SNPs that correlated with the expression of these genes were among the top 500 SNPs in the original GWAS but were not followed up (4,15). The association was strengthened in a combined analysis with two independent Dutch replication cohorts, although this could not be confirmed in all replication cohorts. By adding extracted data from a publicly available meta-analysis, the association of UBE2L3 with CD is even further strengthened and almost reaching genome-wide significance (P ¼ 2.40 × 1027), whereas the association of BCL3

was corroborated. In addition, we have shown that prioritizing eQTL SNPs from the top nominally associated SNPs of a GWAS for follow-up is a potentially promising strategy for identifying novel risk loci. This hypothesis is supported by the fact that NOD2, an established CD risk allele, is among the selected cis-eQTL SNPs in the top 500, although not in the top regions that were selected for follow-up in the original US NIDDK GWAS.

UBE2L3, the most significantly associated gene, encodes a protein involved in ubiquitination. This is the process in which abnormal or short-lived proteins are modified with ubiquitin to mark them for degradation. The protein encoded by UBE2L3 ubiquitinates, among others, the NF-kB precursor p105. The risk allele of the UBE2L3 eQTL SNP correlates with a higher expression of the UBE2L3 gene. Theoretically, overexpression of UBE2L3 could lead to a quicker degradation of the NF-kB precursor and thus to a lower production of NF-kB and consequently a diminished innate immune response. A similar effect is seen for the CD risk variants of NOD2, the strongest CD risk locus. The CD-associated NOD2 variants also lead to an inadequate innate immune response because of a lack of the NF-kB precursor (16). Moreover, the protein encoded by UBE2L3 has been shown in vitro to be involved in natural killer cell cytotoxic function, which is an important part of the innate immune response (17). SNPs in UBE2L3 have also been found to be associated with celiac disease, rheumatoid arthritis and systemic lupus erythematosus (13,18,19), three immune-related diseases known to share risk loci with CD. Our study suggests that UBE2L3 is yet another shared risk locus (20).

BCL3, the second likely novel CD risk gene, plays a role in mediating bacteria-induced colitis. Impaired Bcl3 expression in dendritic cells from Il10-/- mice leads to an increased expression of IL23 in reaction to bacterial lipopolysaccharides. BCL3 also diminishes the inflammatory response induced by bacterial lipopolysaccharides in macrophages (21). The risk variant associated with CD in our study is correlated with a low expression of BCL3. This could point to an increased adaptive immune response in CD patients mediated by the increased expression of IL23. Indeed,

IL23 appears to play an important role in the aberrant immune response that underlies CD (22).

Although the association for UBE2L3 was strengthened in the combined analysis, we could not confirm it in the individual second replication phase. This might have several explanations; the first is possible lack of power. The more recently associated SNPs have lower ORs than the already established associations, so in order to detect new associations, the power of the studies needs to increase. As replication cohorts get exhausted, implementation is difficult. Secondly, there might be true heterogeneity in the populations we genotyped. For example NOD2, the most established risk allele for CD, cannot be confirmed in all populations (23). In favour of the association is that P-values become more significant after performing a combined analysis.

Our results show that selecting SNPs with an eQTL effect for replication is a potentially useful strategy for identifying novel CD risk genes. One disadvantage is that it will only detect risk loci which effect gene expression, whereas not all consistently replicated disease susceptibility loci have such eQTL effects. Therefore, selecting loci for follow-up on additional criteria (i.e. other functional effects) could further improve the yield of this follow-up strategy.

Newly identified CD risk loci can only improve our understanding of the disease mechanism if the effect of the risk causing variant is known. This method of prioritizing eQTLs for replication not only improves the chances of finding relevant associations, but also provides a lead to functional studies. Since the eQTL SNP variants correlate with the expression of nearby genes, we would expect to see a difference in the expression of these genes in relevant tissues taken from patients and healthy controls. After measuring the expression of such genes in tissues relevant to the disease, assessing the functional effects of the differences in model systems might increase better understanding of CD pathogenesis.

We might have missed associated SNPs because we used gene expression data of celiac patients and HapMap data for finding cis-eQTL SNPs. It would be relevant to confirm the eQTL effect

of SNPs on the expression level of UBE2L3 and BCL3 in blood or colonic mucosal biopsies of CD patients. Since CD is characterized by an aberrant immune response, causal variants are probably in the immune cells, e.g. blood.

In summary, we have identified two novel potential risk genes for CD, UBE2L3 and BCL3, by prioritizing cis-eQTL SNPs for follow-up from the top 500 SNPs of a CD GWAS. UBE2L3 is shared between several immune-related diseases (21), but both loci fit with the proposed role of aberrant immune responses in CD pathogenesis. This strategy for following up GWAS data provides both an effective and cost-efficient way of finding new risk loci and leads for functional studies.

Materials and methods

CD-associated SNPs are more likely to be eQTLs.

We first assessed the 30 SNPs that recently have been reported to be associated with CD (8). We used two genetical genomics data sets in a meta-analysis setting, as reported by Heap et al. (15). These data sets comprise 109 celiac disease samples and 90 HapMap CEU samples. As the 109 celiac disease samples had been genotyped using Illumina HumanHap300 arrays, we attempted to impute all HapMap SNPs using Impute v2 and HapMap CEU release 23a. For 29 of the 30 SNPs, genotype data were eventually available, each having an MAF of at least 0.05, a call rate of at least 95% and exact HWE $P > 0.0001$. We investigated the 12 013 expression probes that were present in both genetical genomics data sets. We conducted a cis-eQTL analysis (SNP–probe distance <250 kb, 1000 permutations) and identified five significant cis-eQTLs (FDR controlled at 0.05) (Supplementary Material, Fig. S1). We subsequently assessed whether the five cis-eQTLs we had detected were higher than expected by chance. For each of the 29 included SNPs, we determined the MAF and assessed how many probes mapped within 250 kb distance. We then selected a random set of 29 SNPs, but ensured that each randomly selected SNP had an MAF and number of probes in its vicinity that matched the original SNP. We subsequently assessed how many significant cis-eQTLs could

be identified in this permuted set of SNPs (using identical settings as in the original cis-eQTL analysis). We ran 100 permutations and observed that none of the permutations identified at least five cis-eQTLs for the random set of matched SNPs (four cis-eQTLs were found at most, occurring in nine out of 100 permutations). This indicates that the top 30 CD SNPs are significantly enriched for cis-eQTLs ($P < 0.01$).

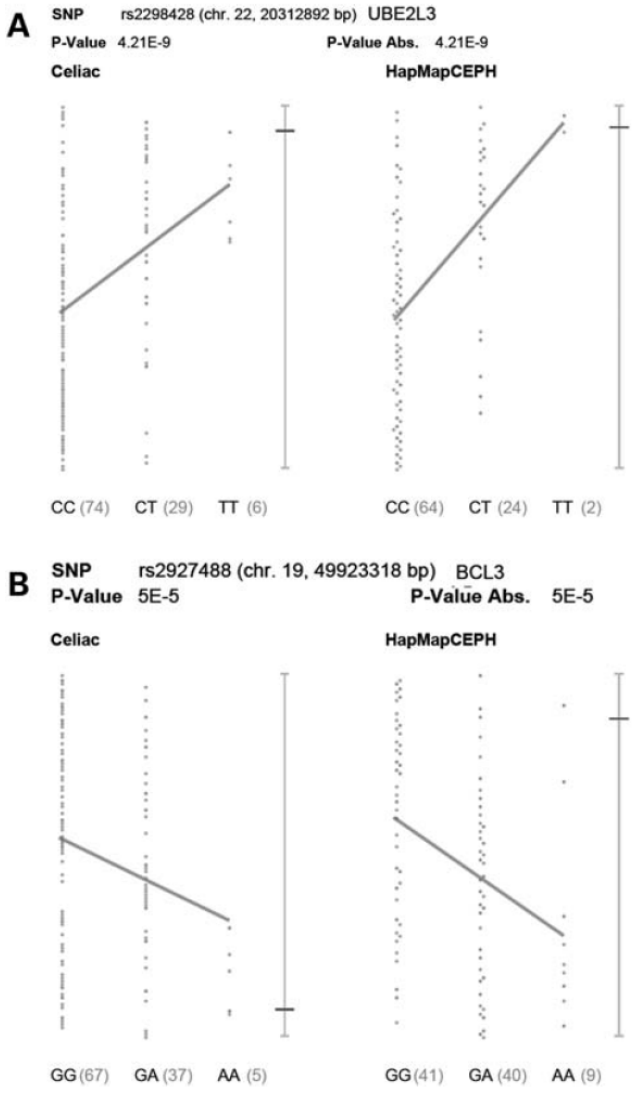


Figure 2. eQTL effects

Figure 2. (A) eQTL effect of rs2298428 on the expression of UBE2L3. On the X-axis, the three different genotypes for SNP rs2298428 are displayed and on the Y-axis the level of expression for UBE2L3. Each dot represents the expression level of UBE2L3 for one individual; the individuals are grouped per genotype. The level of expression of gene UBE2L3 is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap. (B) eQTL effect of rs2927488 on the expression of BCL3. On the X-axis, the three different genotypes for SNP rs2927488 are displayed and on the Y-axis the level of expression for BCL3. Each dot represents the expression level of BCL3 for one individual; the individuals are grouped per genotype. The level of expression of gene BCL3 is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap.

SNP selection

Based on these results, we reasoned that if a high-ranking SNP, but not reaching genome-wide significance, affect gene expression in cis, it is more likely to be a true disease association. We decided to investigate the top 500 SNPs of a publicly available GWAS performed by the US NIDDK Consortium (http://www.ncbi.nlm.nih.gov/gap) (4). Four hundred and ninety-eight of these 500 SNPs had been genotyped or imputed in our genetical genomics data sets. Four hundred and ninety-four SNPs out of 498 SNPs passed QC (having an MAF of at least 0.05, a call rate of at least 95% and an exact HWE P . 0.0001). Using identical eQTL analysis settings, we identified 13 significant cis-eQTLs. One of these SNPs correlated with the expression of NOD2. Since NOD2 is an established CD risk gene, it was not included in our independent replication study. For two genes, COX11 and ENTPD5, we had more than one eQTL SNP in our database, so we selected the SNP with the strongest eQTL effect for replication because this is more likely to be a causative variant. In total, we analysed the 10 remaining SNPs for replication in an initial cohort. The three SNPs that were significantly associated with CD (P , 0.05) were replicated in an independent second cohort (Fig. 1).

Subjects

Our initial analysis, in which we selected SNPs for follow-up, was done in a GWAS data set from a US-Canadian cohort of 946 CD patients and 977 healthy controls (4). The first replication analysis of the selected SNPs was then performed in a Dutch cohort of 777 CD patients and 964 healthy controls (Replication cohort I). The CD patients for this replication were collected by the University Medical Centre Groningen (n = 322) and by the Academic Medical Centre in Amsterdam (n = 455) (24). The 964 healthy controls were blood donors recruited from donor centers in Utrecht and Amsterdam (Table 2) (25).

The SNPs that were found to be associated with CD in Replication cohort I (P < 0.05) were genotyped in a second cohort (Replication cohort II) of 762 Dutch CD patients and 1684 Dutch controls.

The CD patients for the second cohort were collected by the University Medical Centre Leiden (n = 287), the VU University Medical Centre in Amsterdam (n = 317) and the Radboud University in Nijmegen (n = 158). The healthy controls were blood donors recruited from the donor centre in Groningen (n = 720) and healthy controls participating in a chronic obstructive pulmonary disease GWAS (n = 964).

Barrett et al. (8) performed a meta-analysis based on three GWAS performed by the US NIDDK consortium, The UK Wellcome Trust Case Control Consortium and a Belgian- French collaboration. This analysis contained a total of 3230 cases and 4829 controls. The results were used in a second combined analysis. Recruitment of participants was approved by the institutional review boards of each of the hospitals, and informed consent was obtained from all participants.

Genotyping

Genotyping of all CD cases from both replication cohorts was performed using TaqMan technology (Applied Biosystems, Foster City, CA, USA). SNP genotyping assays were obtained from Applied Biosystems and genotyping was carried out as recommended by the manufacturer. The patient DNA samples were processed in 384-well plates, each plate containing 16 genotyping controls [four duplicates of four DNA samples from the Centre d'Etude de Polymorphisme Humain (CEPH)]. All SNPs were successfully genotyped in more than 95% of all samples. We had 99% concordance between our genotype data and the CEPH data available from HapMap. Genotyping of the controls was performed on either the Illumina Human 610-Quad or 670-Quad-custom Beadchips, following the manufacturer's protocol (Illumina Inc., San Diego, CA, USA). Quality control on this data was performed by excluding all SNPs that were out of Hardy– Weinberg equilibrium (HWE) [P-value (HWE) < 0.001] and only including SNPs that were successfully genotyped in 99% of all the samples.
All selected SNPs in the control population were in HWE.

Statistical analysis

Differences in allele and genotype distribution between cases and controls of the individual cohorts were tested for significance by the x2 test. The significance threshold for P-values was set at 0.05. ORs were calculated and the CIs were approximated using Woolf's method with Haldane's correction. A combined analysis of the initial analysis and of the replication phases was performed with the METAL program (http://www.sph.umich.edu/csg/abecasis/metal). A second meta-analysis of both the replication phases and the publicly available Barrett et al. database was performed. Only P-values were available, so a weighted z-score meta-analysis was performed. This analysis was performed separately because the Barrett et al. database is based on a meta-analysis which contains the data of the GWAS performed by the NIDDK.

Acknowledgements: We thank all the patients and controls who participated in this study and Jackie Senior for correcting the manuscript.

Conflict of Interest statement. None declared.

Funding: This study was supported by a clinical fellowship grant (90.700.281) to R.K.W., E.A.M.F. is supported by a clinical traineeship research grant (92.003.533), a VICI grant (918.66.620) to C.W., a VENI grant (863.09.007) to J.F. and a VENI grant (916.10.135) to L.F., all from the Netherlands Organization for Scientific Research (NWO). Further support was provided by a grant from the Celiac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch Government, grant BSIK03009) to C.W. and a Horizon Breakthrough grant from the Netherlands Genomics Initiative (93519031) to L.F.

References

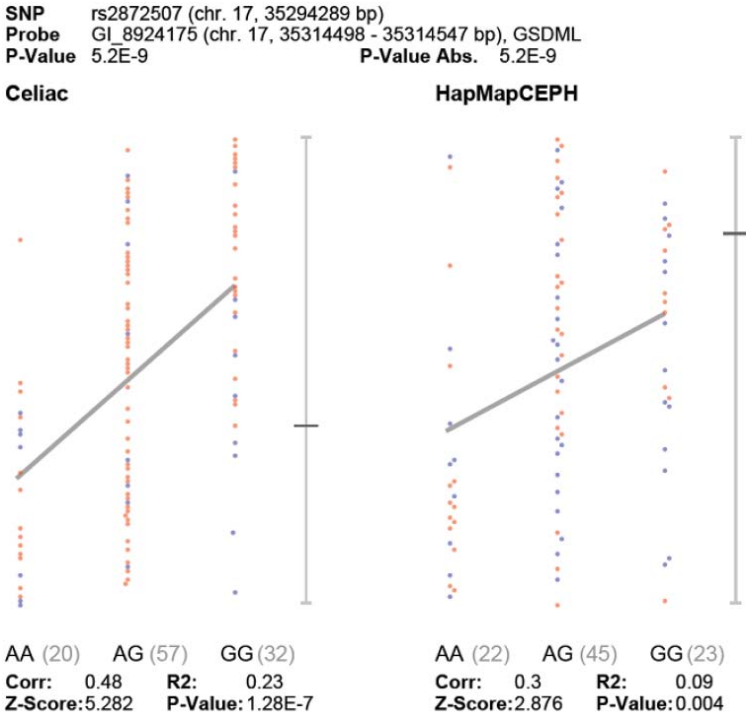
1. Loftus, E.V. Jr (2004) Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. Gastroenterology, 126, 1504–1517.
2. Baumgart, D.C. and Carding, S.R. (2007) Inflammatory bowel disease: cause and immunobiology. Lancet, 369, 1627–1640.
3. Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., Cardon, L., Takazoe, M., Tanaka, T., Ichimori, T. et al. (2005) Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum. Mol. Genet., 14, 3499–3506.
4. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, H.A., Abraham, C., Regueiro, M., Griffiths, A. et al. (2006) A genomewide association study identifies IL23R as an inflammatory bowel disease gene. Science, 314, 1461–1463.
5. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Wu Datta, L. et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet., 39, 596–604.
6. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A. 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, e58.
7. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447, 661–678.
8. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, L.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet., 40, 955–962.
9. Ioannidis, J.P., Thomas, G. and Daly, M.J. (2009) Validating, augmenting and refining genome-wide association signals. Nat. Rev. Genet., 10, 318–329.
10. Gilad, Y., Rifkin, S.A. and Pritchard, J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet., 24, 408–415.
11. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G.A.R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D. et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. Nat. Genet., 40, 395–402.
12. Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E. et al. (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature, 448, 470–473.
13. Dubois, P.C., Trynka, G., Franke, L., Hunt, K.A., Romanos, J., Curtotti, A., Zhernakova, A., Heap, G.A., Ada´ny, R., Aromaa, A. et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. Nat. Genet., 42, 295–302.
14. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E. and Cox, N.J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet., 4, e1000888.
15. Heap, G.A., Trynka, G., Jansen, R.C., Bruinenberg, M., Swertz, M.A., Dinesen, L.C., Hunt, K.A., Wijmenga, C., vanHeel, D.A. and Franke, L. (2009) Complex nature

of SNP genotype effects on gene expression in primary human leucocytes. BMC Med. Genomics, 2, 1.

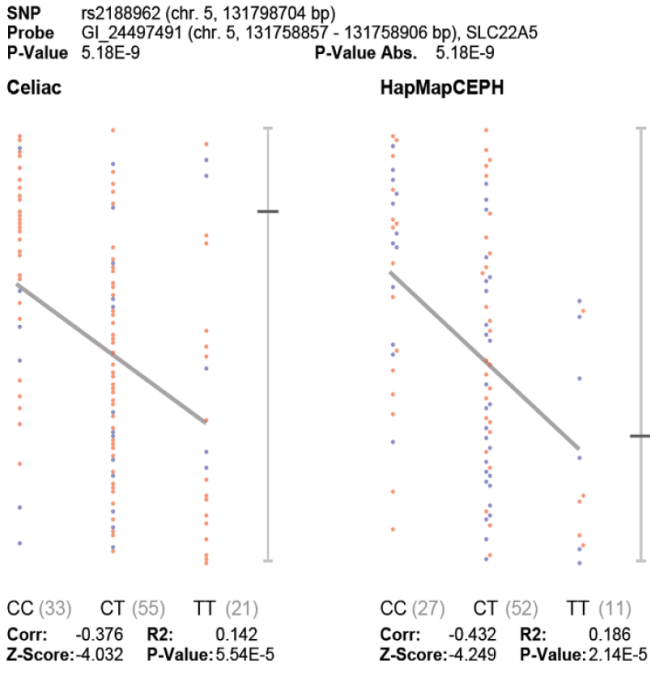
16. Rosenstiel, P., Sina, C., End, C., Renner, M., Lyer, S., Till, A., Hellmig, S., Nikolaus, S., Fo¨lsch, U.R., Helmke, B. et al. (2007) Regulation of DMBT1 via NOD2 and TLR4 in intestinal epithelial cells modulates bacterial recognition and invasion. J. Immunol., 178, 8203– 8211.
17. Fortier, J.M. and Kornbluth, J. (2006) NK lytic-associated molecule, involved in NK cytotoxic function, is an E3 ligase. J. Immunol., 176, 6454–6463.
18. Han, J.W., Zheng, H.F., Cui, Y., Sun, L.D., Ye, D.Q., Hu, Z., Xu, J.H., Cai, Z.M., Huang, W., Zhao, G.P. et al. (2009) Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nat. Genet., 41, 1234–1237.
19. Stahl, E.A., Raychaudhuri, S., Remmers, E.F., Xie, G., Eyre, S., Thomson, B.P., Li, Y., Kurreeman, F.A.S., Zhernakova, A., Hinks, A. et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat. Genet., 42, 508–514.
20. Zhernakova, A., van Diemen, C.C. and Wijmenga, C. (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nat. Rev. Genet., 10, 43–55.
21. Muhlbauer, M., Chilton, P.M., Mitchell, T.C. and Jobin, C. (2008) Impaired Bcl3 upregulation leads to enhanced lipopolysaccharide-induced interleukin (IL)-23P19 gene expression in IL-10(2/2) mice. J. Biol. Chem., 283, 14182–14189.
22. Kobayashi, T., Okamoto, S., Hisamatsu, T., Kamada, N., Chinen, H., Saito, R., Kitazume, M.T., Nakazawa, A., Sugita, A., Koganei, K. et al. (2008) IL23 differentially regulates the Th1/Th17 balance in ulcerative colitis and Crohn's disease. Gut, 57, 1682–1689.
23. Arnott, I.D.R., Ho, G.T., Nimmo, E.R. and Satsangi, J. (2005) Toll-like receptor 4 gene in IBD: further evidence for genetic heterogeneity in Europe. Gut, 54, 308–309.
24. Weersma, R.K., Stokkers, P.C., van Bodegraven, A.A., van Hogezand, R.A., Verspaget, H.W., de Jong, D.J., van der Woude, C.J., Oldenburg, B., Linskens, R.K., Festen, E.A.M. et al. (2009) Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. Gut, 58, 388–395.
25. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C.N.M., Bethel, G., Holmes, G.K.T. et al. (2007) A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet., 39, 827–829.
26. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolia, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl Acad. Sci. USA, 106, 9362–9367.

Supplementary figures

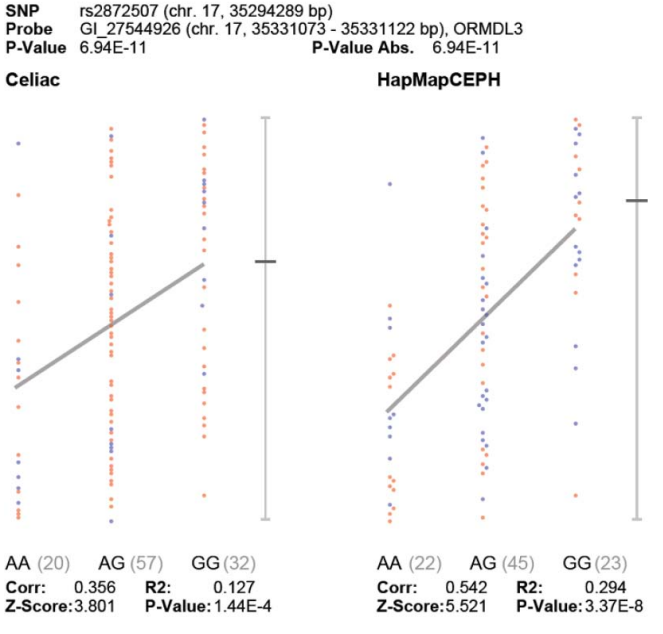
eQTL effects of five known Crohn's disease associated loci. (a) GSMDL (b) SLC22A5 (c) ORMDL3 (d) UBE1L (e) RNASET2 On the X-axis the three different genotypes are displayed, on the Y-axis the level of expression for each gene. Each dot represents the expression level of the gene for one individual; the individuals are grouped per genotype. The level of expression of each gene is correlated to the different genotypes. Data for this analysis were obtained from publicly available expression data from patients with celiac disease and HapMap.



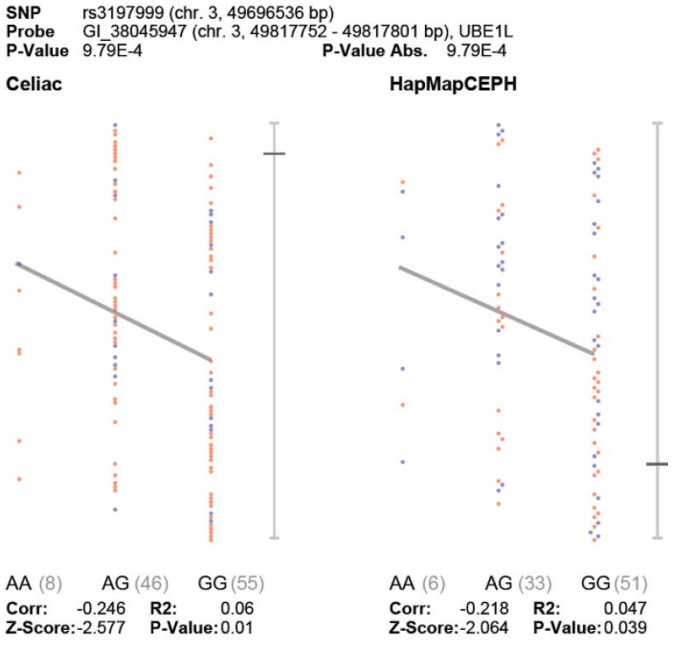
S1a. eQTL effects of known Crohn's disease associated locus on *GSMDL*



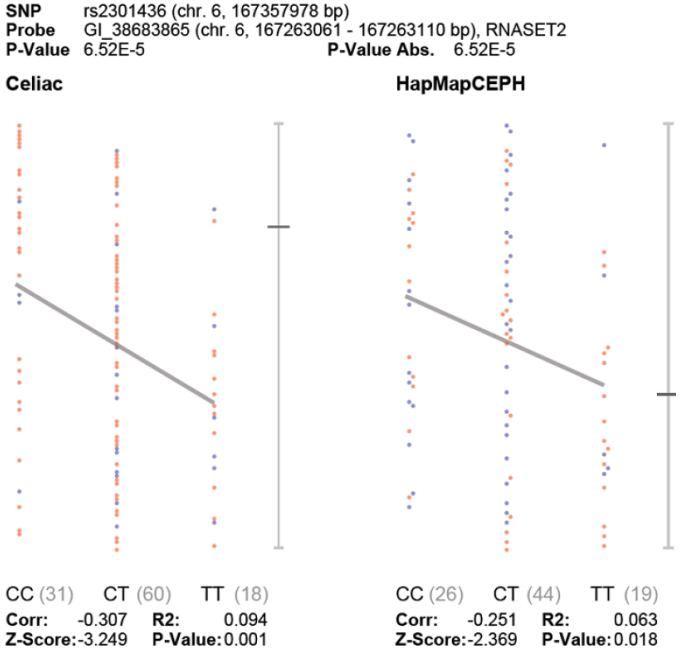
S1b. eQTL effects of known Crohn's disease associated locus on *SLC22A5*



S1c. eQTL effects of known Crohn's disease associated locus on *ORMDL3*



S1d. eQTL effects of known Crohn's disease associated locus on *UBE1L*



S1e. eQTL effects of known Crohn's disease associated locus on *RNASET2*

CHAPTER 4

Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis.

*Lina-Marcela Diaz-Gallo, Laura Espino-Paisán *, Karin Fransen *, Mariá Gómez Garciá, Suzanne van Sommeren, Carlos Cardña, Luis Rodrigo, Juan Luis Mendoza, Carlos Taxonera, Antonio Nieto, Guillermo Alcain, Ignacio Cueto, Miguel A. López-Nevot, Nunzio Bottini, Murray L. Barclay, J. Bart Crusius, Adriaan A. van Bodegraven, Cisca Wijmenga, Cyriel Y. Ponsioen, Richard B. Gearry, Rebecca L. Roberts, Rinse K. Weersma, Elena Urcelay, Tony R. Merriman, Behrooz Z. Alizadeh, and Javier Martin*

Inflamm Bowel Dis. 2011 Nov;17 (11):2287-94.

**These authors contributed equally*

Abstract

Background: The PTPN22 gene is an important risk factor for human autoimmunity. The aim of this study was to evaluate for the first time the role of the R263Q PTPN22 polymorphism in ulcerative colitis (UC) and Crohn's disease (CD), and to reevaluate the association of the R620W PTPN22 polymorphism with both diseases.

Methods: A total of 1677 UC patients, 1903 CD patients, and 3111 healthy controls from an initial case–control set of Spanish Caucasian ancestry and two independent sample sets of European ancestry (Dutch and New Zealand) were included in the study. Genotyping was performed using TaqMan SNP assays for the R263Q (rs33996649) and R620W (rs2476601) PTPN22 polymorphisms. Meta-analysis was performed on 6977 CD patients, 5695 UC patients, and 9254 controls to test the overall effect of the minor allele of R620W and R263Q polymorphisms.

Results: The PTPN22 263Q loss-of-function variant showed initial evidence of association with UC in the Spanish cohort (P = 0.026, odds ratio [OR] = 0.61, 95% confidence interval [CI]: 0.39–0.95), which was confirmed in the meta-analysis (P = 0.013 pooled, OR = 0.69, 95% CI: 0.51–0.93). In contrast, the 263Q allele showed no association with CD (P = 0.22 pooled, OR = 1.16, 95% CI: 0.91–1.47). We found in the pooled analysis that the PTPN22 620W gain-of-function variant was associated with reduced risk of CD (P = $7.4*10^{-06}$ pooled OR = 0.81, 95% CI: 0.75–0.89) but not of UC (P = 0.88 pooled, OR = 0.98, 95% CI: 0.85–1.15).

Conclusions: Our data suggest that two autoimmunity-associated polymorphisms of the PTPN22 gene are differentially associated with CD and UC. The R263Q polymorphism only associated with UC, whereas the R620W was significantly associated with only CD.

Introduction

Crohn's disease (CD) and ulcerative colitis (UC) are the main types of inflammatory bowel disease (IBD). They are relapsing and chronic inflammatory disorders that result from the complex interaction of genetic, immune, and environmental factors. It is estimated that the current number of loci associated with IBD only explain 10%–20% of the genetic risk attributed to UC and CD. Thus, additional genetic contributions clearly remain to be discovered [1–4].

The protein tyrosine phosphatase nonreceptor 22 (PTPN22) gene encodes the gatekeeper of T-cell receptor (TCR) signaling, protein tyrosine phosphatase (PTP, also known as LYP), and as such is a compelling candidate risk factor for IBD. In T cells, LYP (lymphoid tyrosine phosphatase) potently inhibits signaling through dephosphorylation of several substrates, including the Src-family kinases Lck and Fyn, as well as ZAP-70 and TCRzeta. Moreover, *PTPN22* has emerged as an important genetic risk factor for human autoimmunity [5–8]. Specifically, two missense single nucleotide polymorphisms (SNPs), both with functional influence, [6,8–12] have been associated with autoimmune diseases. The R620W (1858C>T, rs2476601) polymorphism in exon 14 of *PTPN22* was first associated with type 1 diabetes (T1D), and subsequently with autoimmune disorders such as rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), IBD, and other autoimmune diseases [13–16]. The R620W variation disrupts the interaction between Lck and LYP, leading to reduced phosphorylation of LYP, which ultimately contributes to gain-of-function inhibition of T-cell signaling [17]. The Q minor allele of R263Q (788G>A, rs33996649) in exon 10, within the catalytic domain of the enzyme, is a loss-of-function mutation that confers protection against development of SLE and RA [12,18]. In this study we sought first to determine whether the newly described amino acid substitution, R263Q (788G>A, rs33996649) is associated with altered susceptibility to CD and UC and, second, to reevaluate the influence of the R620W (1858C>T, rs2476601) polymorphism on these diseases by conducting a case–control study and meta-analysis.

Materials and methods

Case–Control Study

Study Population

A total of 1903 CD patients, 1677 UC patients, and 3111 healthy controls from an initial case–control set of Spanish Caucasian ancestry (699 CD patients, 658 UC patients, and 1685 healthy controls) and two independent sample sets of European ancestry from The Netherlands (694 CD patients, 548 UC patients, and 863 healthy controls) and New Zealand (510 CD patients, 471 UC patients, and 563 healthy controls) were included in the case–control study. All IBD patients were diagnosed according to standard clinical, endoscopic, radiologic, and histopathologic criteria [19–21]. Control individuals were matched by Caucasian origin, age, and gender. Written informed consent was obtained from all participants. The study was approved by the Ethics Committee of the Spanish and Dutch hospitals, and by the Upper (cases) and Lower (controls) South Regional Ethics Committees of New Zealand.

PTPN22 Genotyping

DNA from patients and controls was obtained using standard extraction methods. Samples were genotyped for SNP rs33996649 using a Custom TaqMan SNP Genotyping Assay (Applied Biosystems, Foster City, CA). The primer sequences were: forward 5' TTTGAACTAATGAAGGCCTCTGTGT 30' and reverse 5' ATTCCTGAGAACTTCAGTGTTTTCAGT 3'. The specific minor groove binder probe sequences were 5' TTGATCCGGGAAATG 3' and 5' TTGATCCAGGAAATG 3'. The samples were genotyped for rs2476601 polymorphism via TaqMan 50 allelic discrimination assay using a predesigned probe (Part number: C__16021387_20; Applied Biosystems). To verify the genotyping consistency 10% of samples from each studied cohort were genotyped twice. The concordance between original and repeat genotypes was 99%. The genotype call rate was >90% for all studied populations.

Data Analysis

Deviation from Hardy–Weinberg equilibrium (HWE) was tested by standard chi-square analysis. The differences in genotype distribution and allele frequency among cases and controls were calculated by contingency tables and when necessary by Fisher's exact test. An association was considered statistically significant if $P < 0.05$. Linkage disequilibrium (LD) measurements (r2) between rs33996649 and rs2476601 were estimated by the expectation-maximization algorithm using HAPLOVIEW v. 4.1 (VC Broad Institute of MIT and Harvard 2008, Cambridge, MA). Case–control association analysis was performed using PLINK (v. 1.07) (http://pngu.mgh.harvard.edu/purcell/plink/) to estimate odds ratios (OR) and 95% confidence intervals (CI) [22]. To test for associations of the PTPN22 polymorphisms with clinical features, a univariate analysis using chi^2 or Fisher's exact test was applied. The Montreal Classification [19] criteria were used to determine the clinical variables. We compare each variable with the healthy controls and within cases (see Supporting Information Tables 1–4). Multiple testing was corrected by false discovery rate control (pFDR). Analysis was conducted using PLINK (v. 1.07) and Stats Direct (v. 2.6.6 http://www.statsdirect.com) software.

Meta-analysis

Study Selection and Data Extraction

To estimate the common effect of the PTPN22 R620W polymorphism on IBD we conducted a search on MEDLINE and PUBMED electronic databases up to April 2010 to identify available articles in which this polymorphism was genotyped in patients with CD or UC and healthy controls. The search strategy included Medical Subject Heading (MeSH) terms and text words as follows: ''Inflammatory Bowel Disease'' [MeSH] OR''Crohn's Disease''[MeSH]) OR ''Colitis, Ulcerative'' [MeSH] AND ''PTPN22 protein, human ''[Substance Name] OR *PTPN22*. References in the studies were reviewed to identify additional studies not indexed by MEDLINE. Studies for the meta-analysis were selected

if they met the following conditions: 1) diagnosis and phenotype was established by means of the Vienna or Montreal Classifications [19–21]; 2) data were collected in Caucasian populations; 3) the study had a case–control design; 4) the SNPs genotyped were rs2476601 or rs6679677 (both are in complete linkage disequilibrium in Caucasian populations, http://www.hapmap.org); 5) the study supplied enough information to calculate the OR, or the authors provided the data by personal communication (the authors of articles which did not show complete data were contacted by email); 6) the study provided original data (independent of other studies included in the meta-analysis); and 7) the article was published in a peer-reviewed journal as a full article, not as an abstract or similar type of summary. Our systematic review of the literature identified 28 potential studies for the meta-analysis of R620W in IBD [13,16,23–47]. A total of 15 studies were not included in our analysis [13,27,28,31–36,38,39,41,44–46]. Five of these were not case– control studies[31,34,35,38,41] and three did not genotype rs2476601 or rs6679677 [27,28,36]. Another five did not supply enough information to calculate the OR [13,32,44–46]. One included some samples of our Spanish cohort [33] and another was carried out only on patients with ileal CD [39].

Data Analysis

The analysis of the combined data from all populations was performed using Stats Direct software, v. 2.6.6. The summarized ORs and CIs were obtained by means of both the random (DerSimonian-Laird) and the fixed (Mantel-Haenszelmeta-analysis) effect models. The heterogeneity of ORs among cohorts was calculated using Breslow-Day test. The statistical power of the R263Q and R620W meta-analysis was 97%, 99% for CD, and 96%, 99% for UC, respectively (assuming a P = 0.01; disease prevalence of 0.1% and allele frequency of 5%; done using CaTS software http://www.sph.umich.edu/csg/abecasis/CaTS/index.html).

Results

R263Q Polymorphism of PTPN22 Is Associated with Reduced Risk of UC

First we conducted an association study in a case–control set of Spanish Caucasian ancestry. The distribution of the allelic frequencies of the two polymorphisms, R263Q and R620W (Tables 1, 2) were in HWE both in patients and controls. As previously reported, [12,18] no LD between the PTPN22 R263Q and R620W genetic variants was observed in any population ($r2 < 0.03$ for each studied population). We observed that the 263Q allele was significantly associated with UC ($P = 0.026$, OR = 0.61, 95% CI: 0.39–0.95) but not with CD ($P = 0.07$, OR = 1.34, 95% CI: 0.97–1.85) (Table 1). We then conducted a follow-up study in two independent Caucasian populations. The case–control analysis in the Dutch and New Zealand cohorts did not show significant association with the R263Q polymorphism in either the CD (Dutch: $P = 0.98$, OR = 0.99 95%, CI:0.64–1.55, New Zealand: $P = 0.87$, OR = 0.95, 95% CI: 0.52–1.74) or the UC sample sets (Dutch: $P = 0.58$, OR = 0.87, 95% CI: 0.53–1.43, New Zealand: $P = 0.17$, OR = 0.61, 95% CI: 0.30–1.24) (Table 1). Our combined analysis of the three studied Caucasian sample sets did not reveal a significant association between the R263Q polymorphism and CD ($P = 0.22$ pooled, OR = 1.16, 95% CI: 0.91–1.47) but it did strengthen the initial association observed in UC in the Spanish sample set ($P = 0.013$ pooled, OR = 0.69, 95% CI: 0.51–0.93) (Table 1; Fig. 1), suggesting that the 263Q variant of the PTPN22 gene may reduce the risk of UC.

620W Allele of PTPN22 Is Associated with Reduced Risk of CD

In order to reevaluate the role of the R620W polymorphism of the PTPN22 gene on IBD, we conducted a case–control study in the three Caucasian cohorts. We did not observe a significant difference in genotype or in the minor allele frequency (MAF) between CD patients and healthy controls in the Spanish sample set ($P = 0.11$, OR = 0.81, 95% CI: 0.62–1.1). In contrast, we observed that the R620W variant was

associated with reduced risk of CD in the Dutch sample set (P = 0.036, OR = 0.76, 95% CI: 0.58–0.98) and in the New Zealand sample set (P = 0.014, OR = 0.67, 95% CI: 0.49–0.92) (Table 2). For the UC analysis, we did not observe a significant difference in either the Spanish or the New Zealand sample sets for the R620W polymorphism (Spanish: P = 0.68, OR = 1.05, 95% CI: 0.82–1.35, New Zealand: P = 0.93, OR = 0.99, 95% CI: 0.73–1.32). However, the 620W allele was associated with a reduced risk of UC in the Dutch sample set (P = 0.015, OR = 0.70, 95% CI: 0.52–0.93) (Table 2). We performed a meta-analysis to reevaluate the role of the R620W polymorphism in IBD. From the remaining 13 studies, three studies fulfilled inclusion criteria for meta-analysis of the R620W PTPN22 polymorphism in UC,[24,30,37] and Silverberg *et al.* [40] provided the minor allele frequencies of R620W in their initial cohort by personal communication.

TABLE 1. Genotype and Allele Frequencies for the R263Q *PTPN22* (rs33996649) Polymorphism in Healthy Controls and IBD Patients from Three Different Populations

Population		GG	%	GA	%	AA	%	Allele G	%	Allele A	%	*P*-value	OR	(95 % CI)	
Spanish	CD patients ($n = 699$)	640	91.6	59	8.4	0	0.0	1339	95.8	59	4.2	0.073	1.34	0.97	1.85
	UC patients ($n = 658$)	632	96.0	26	4.0	0	0.0	1290	98.0	26	2.0	0.026	0.61	0.39	0.95
	Controls ($n = 1685$)	1580	93.8	103	6.1	2	0.1	3263	96.8	107	3.2				
Dutch	CD patients ($n = 694$)	658	94.8	36	5.2	0	0.0	1352	97.4	36	2.6	0.98	0.99	0.64	1.55
	UC patients ($n = 548$)	523	95.4	25	4.6	0	0.0	1071	97.7	25	2.3	0.58	0.87	0.53	1.43
	Controls ($n = 863$)	818	94.8	45	5.2	0	0.0	1681	97.4	45	2.6				
New Zealand	CD patients ($n = 510$)	490	96.1	20	3.9	0	0.0	1000	98.0	20	2.0	0.87	0.95	0.52	1.74
	UC patients ($n = 471$)	459	97.5	12	2.5	0	0.0	930	98.7	12	1.3	0.17	0.61	0.30	1.24
	Controls ($n = 559$)	536	95.9	23	4.1	0	0.0	1095	97.9	23	2.1				
Pooled	CD patients ($n = 1903$)	1788	94.0	115	6.0	0	0.0	3691	97.0	115	3.0	0.22[a]	1.16	0.91	1.47
	UC patients ($n = 1677$)	1614	96.2	63	3.8	0	0.0	3291	98.1	63	1.9	0.013[b]	0.69	0.51	0.93
	Controls ($n = 3107$)	2934	94.4	171	5.5	2	0.1	6039	97.2	175	2.8				

CD, Crohn's disease'. UC, ulcerative colitis. P-value for the minor allele.
A) Meta-analysis calculated through the fixed effects model. Breslow-Day P = 0.44.
b) Meta-analysis calculated through the fixed effects model. Breslow-Day P = 0.54.

TABLE 2. Genotype and Allele Frequencies for R620W PTPN22 (rs2476601) Polymorphism in Healthy Controls and IBD Patients from 14 Different Populations

Population		CC	%	CT	%	TT	%	Allele C	%	Allele T	%	*P*-value	OR	(95 % CI)	
Spanish	CD patients ($n = 699$)	626	89.6	69	9.9	4	0.6	1321	94.5	77	5.5	0.11	0.81	0.62	1.1
	UC patients ($n = 658$)	571	86.8	81	12.3	6	0.9	1223	92.9	93	7.1	0.68	1.05	0.82	1.35
	Controls ($n = 1685$)	1467	87.1	209	12.4	9	0.5	3143	93.3	227	6.7				
Dutch	CD patients ($n = 672$)	575	85.6	94	14.0	3	0.4	1244	92.6	100	7.4	0.036	0.76	0.58	0.98
	UC patients ($n = 539$)	468	86.8	67	12.4	4	0.7	1003	93.0	75	7.0	0.015	0.7	0.52	0.93
	Controls ($n = 834$)	683	81.9	142	17.0	9	1.1	1508	90.4	160	9.6				
New Zealand	CD patients ($n = 477$)	414	86.8	60	12.6	3	0.6	888	93.1	66	6.9	0.014	0.67	0.49	0.92
	UC patients ($n = 448$)	366	81.7	76	17.0	6	1.3	808	90.2	88	9.8	0.93	0.99	0.73	1.32
	Controls ($n = 563$)	454	80.6	106	18.8	3	0.5	1014	90.1	112	9.9				
Anderson et al. (2009) British	UC patients ($n = 2471$)	2024	81.9	425	17.2	22	0.9	4473	90.5	469	9.5	0.74	0.98	0.86	1.12
	Controls ($n = 2483$)	2025	81.6	435	17.5	23	0.9	4485	90.3	481	9.7				
De Jager et al (2006) Canadian	CD patients ($n = 249$)	225	90.0	23	9.6	1	0.4	468	94.8	30	6.0	0.33	1.33	0.71	2.50
	Controls ($n = 207$)	191	92.3	16	7.7	0	0.0	398	95.9	16	3.9				
Duerr et al. (2006) Caucasian European	CD patients ($n = 541$)	473	87.4	68	12.6	0	0.0	1014	93.7	68	6.3	0.003	0.63	0.46	0.86
	Controls ($n = 541$)	441	81.5	95	17.6	5	0.9	977	90.3	105	9.7				
Hradsky et al (2008) Czech	CD patients ($n = 345$)	275	79.7	66	19.1	4	1.2	616	89.3	74	10.7	0.92	1.02	0.74	1.39
	Controls ($n = 501$)	398	79.4	100	20.0	3	0.6	896	89.4	106	10.6				
Latiano et al. (2007)Italian	CD patients ($n = 301$)	283	94.0	18	6.0	0	0.0	584	97.0	18	3.0	0.31	0.73	0.39	1.37
	UC patients ($n = 306$)	278	90.8	28	9.2	0	0.0	584	95.4	28	4.6	0.70	1.10	0.63	1.96
	Controls ($n = 256$)	235	91.8	21	8.2	0	0.0	491	95.9	21	4.1				
Morgan et al. (2010) New Zealand	CD patients ($n = 315$)	260	82.5	52	16.5	3	1.0	572	90.8	58	9.2	0.33	0.85	0.60	1.19
	Controls ($n = 472$)	379	80.3	85	18.0	8	1.7	843	89.3	101	10.7				
Prescott et al. (2005) British	CD patients ($n = 294$)	254	86.4	37	12.6	3	1.0	545	92.7	43	7.3	0.46	0.86	0.58	1.29
	UC patients ($n = 220$)	192	86.9	26	12.2	2	0.9	410	92.8	30	6.8	0.38	0.83	0.53	1.30
	Controls ($n = 374$)	312	83.4	61	16.3	1	0.3	685	91.6	63	8.4				
Silverberg et al. (2009) Caucasian European	UC patients ($n = 1052$)	852	81.0	189	18.0	11	1.0	1893	90.0	211	10.0	0.008	1.27	1.06	1.50
	Controls ($n = 2571$)	2171	84.4	383	14.9	17	0.7	4725	91.9	417	8.1				

Van Oene et al. (2005) Canadian	CD patients ($n = 455$)	389	85.5	63	13.8	3	0.7	841	92.4	69	7.6	0.55	0.91	0.66	1.25			
	Controls ($n = 603$)	508	84.2	90	14.9	5	0.8	1106	91.7	100	8.3							
Wagenleiter et al (2005) German	CD patients ($n = 146$)	122	83.6	23	15.8	1	0.7	267	91.4	25	8.6	0.390	0.82	0.49	1.34			
	Controls ($n = 254$)	204	80.3	47	18.5	3	1.2	455	89.6	53	10.4							
WTCC (2007) Caucasian	CD patients ($n = 2005$)	1703	84.9	291	14.5	11	0.5	3697	92.2	313	7.8	0.001	0.79	0.69	0.91			
	Controls ($n = 3004$)	2447	81.5	533	17.7	24	0.8	5427	90.3	581	9.7							
Pooled	CD patients ($n = 6977$)	6013	86.2	925	13.3	39	0.6	12951	92.8	1003	7.2	7.4E-06a	0.81	0.75	0.89			
	Controls ($n = 9254$)	7718	83.4	1467	15.9	69	0.7	16903	91.3	1605	8.7							
	UC patients ($n = 5695$)	4751	83.4	893	15.7	51	0.9	10395	91.3	995	8.7	0.88[b]	0.98	0.85	1.15			
	Controls ($n = 8766$)	7347	83.8	1357	15.5	62	0.7	16051	91.6	1481	8.4							

CD, Crohn's disease. UC, ulcerative colitis. *P*-value for the minor allele
[a]Meta-analysis calculated through the fixed effects model. Breslow-Day $P = 0.18$.
[b]Meta-analysis calculated through the random effects model. Breslow-Day $P = 0.03$.

In CD, eight studies fulfilled inclusion criteria for meta-analysis of the R620W PTPN22 polymorphism, [23,25,29,30,37,42,43,47] and Duerr *et al.* [26] provided the minor allele frequencies of R620W in their initial cohort by personal communication. A strong association between the 620W variant and CD was demonstrated ($P = 7.4*10^{-06}$ pooled, OR = 0.81, 95% CI: 0.75–0.89) (Table 2; Fig. 2). This confirms the association of the reduced risk observed between this allele and CD in our initial case–control study in the Dutch and New Zealand sample sets and in the previous meta-analysis reported by Barrett *et al.* [13]. In contrast, no association was observed between the 620W allele and UC (P = 0.88 pooled, OR = 0.98, 95% CI: 0.85–1.15) (Table 2; Fig. 2).

620W Allele of PTPN22 Is Associated with Reduced Risk of Ileal Location in CD

We evaluated the possible associations of the R263Q and R620W variants of PTPN22 with the clinical phenotypes of UC and CD (Supplementary Tables 1–4). Meta-analysis revealed the 620W variant was significantly associated with reduced risk of ileal location of CD when compared to healthy controls (P FDR corrected = $9*10^{-03}$) pooled OR = 0.64, 95% CI = 0.49-0.84, Supplementary Table 2). We observed no significant association of the R263Q polymorphism with CD or UC clinical manifestations.

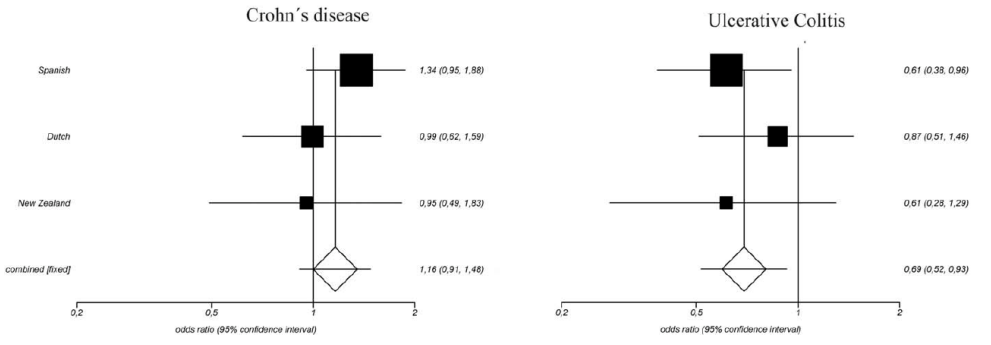


Figure 1. Forest plots for the meta-analyses of the PTPN22 R263Q (G788A; rs33996649) polymorphism in CD and UC. The analyses correspond to the frequency of the minor (A) allele in the three Caucasian IBD sample sets.

Discussion

This article reports for the first time the role of the newly identified R263Q polymorphism of PTPN22 in IBD. In addition, we performed a case–control study in Spanish, Dutch, and New Zealand populations and a meta-analysis to assess the role of the R620W PTPN22 polymorphism with CD and UC. Our results indicate that there is a differential association of the R263Q and R620W polymorphisms with IBD. On the one hand, the PTPN22 263Q loss-of function variant is a protective factor for UC, with no relationship to CD; on the other hand, the 620W gain-of-function variant confers protection against CD, while showing no association with UC. The effect size observed between the R263Q polymorphism and UC (0.69) is similar to that reported for SLE (i.e., 0.63) by Orru *et al.* [12,18] suggesting that this polymorphism could be another common genetic component in autoimmunity. In addition, we confirmed in the Dutch and New Zealand CD cohorts, together with a combined analysis, the previously reported protective role of the 620W allele in CD but not in UC [12,13,23,26,32,39,45–47]. Thus, there is support for the hypothesis that both outcomes of IBD have a partially different genetic component. On the other hand, we have reported evidence of a reduced risk factor of the 620W allele in the ileal location of CD. Nevertheless, these result should be taken cautiously, since we observed no significant difference when

comparing the ileal location against colonic/ileocolonic location of the disease. This may be an artifact of low statistical power of these stratified analyses (i.e., 50%–65% power). Replication studies are needed to confirm this new finding. Increased emphasis has been placed in the recent years on predictive biomarkers to predict the onset or future course of disease.48 In this regard, the present report supports the idea that subtle genetic differences combined with assessment of the pattern of critical mediators (i.e., presence of autoantibodies) may be useful for tracing progression of the disease.

To determine the immunological implications of the differential association of R263Q and R620W PTPN22 polymorphisms with CD and UC, functional approaches are required. Nevertheless, there is strong evidence to suggest that the 263Q allele is a loss-of-function variant which is less effective in reducing TCR signaling than 263R[12]. This supports the hypothesis that positive modulation of the TCR helps in reestablishing tolerance in at least a subset of autoimmune patients [6,49]. This functional evidence, together with the significant association that we observed with UC, suggests that TCR signaling is more important in this disease than in CD. Actually, autoantibodies are more often detected in UC than in CD patients. It is estimated that 60%–70% of UC patients are positive for atypical antineutrophilic cytoplasmic antibodies, whereas only few CD patients present autoantibodies (atypical antineutrophilic cytoplasmic antibodies 5%–25%, pancreatic autoantibodies 27%–37%, and thrombophilia-associated antibodies 3%–37%) [50].

The present study confirms that the 620W allele is associated with a reduced risk of developing CD, in contrast to the increasing risk that this genetic variant confers to other autoimmune diseases such as T1D, SLE, and RA [6,14–16]. Several authors have shown that 620W PTPN22 is a gain-of-function variant that reduces TCR signaling leading to decreased elimination of potentially autoreactive T cells and/or decreased production of natural regulatory T cells (Treg) (reviewed [6]). This could explain the loss of tolerance that takes place in autoimmune diseases like T1D, SLE, and RA, but not the protective role 620W allele appears to confer against CD. A possible explanation could be that IBD

may represent an inappropriate immune response to the commensal microbiota in a genetically predisposed host[3], mimicking an infection process. This hypothesis is supported by the fact that the 620W allele confers protection towards some highly prevalent infectious diseases [6]. Previous studies have reported a significant protective role of the 620W allele in tuberculosis (TB) [51,52]. Moreover, the R263Q polymorphism has been associated with increasing risk to develop TB [52], the opposite of the reported associations with SLE [12] and RA [18] and UC in the present study. Our findings suggest that many of the genetic loci involved in autoimmunity may be under balanced selection due to antagonistic pleiotropic effects. Genetic variants such as R620W and R263Q with opposite effects in different diseases may facilitate the maintenance of common susceptibility alleles in human populations [6,46,53]. Moreover, our results also support the idea that CD and UC differ in some genetic risk factors, thereby suggesting the involvement of different immunological mechanisms with a related nature [24,45,46,54,55].

Acknowledgements: We thank Sofia Vargas and Sonia Garcia for excellent technical assistance. We thank all the donors, patients, and controls.

References

1. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. Lancet. 2007;369:1627–1640.
2. Budarf ML, Labbe C, David G, et al. GWA studies: rewriting the story of IBD. Trends Genet. 2009;25:137–146.
3. Kaser A, Zeissig S, Blumberg RS. Inflammatory bowel disease. Annu Rev. Immunol. 2010;28:573–621.
4. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461:747–753.
5. Cloutier JF, Veillette A. Cooperative inhibition of T-cell antigen receptor signaling by a complex between a kinase and a phosphatase. J Exp Med. 1999;189:111–121.
6. Stanford SM, Mustelin TM, Bottini N. Lymphoid tyrosine phosphatase and autoimmunity: human genetics rediscovers tyrosine phosphatases. Semin Immunopathol. 2010;32:127–36.
7. Wu J, Katrekar A, Honigberg LA, et al. Identification of substrates of human protein-tyrosine phosphatase PTPN22. J Biol Chem. 2006;281: 11002–11010.
8. Yu X, Sun JP, He Y, et al. Structure, inhibitor, and regulatory mechanism of Lyp, a lymphoid-specific tyrosine phosphatase implicated in autoimmune diseases. Proc Natl Acad Sci U S A. 2007; 104:19767–19772.
9. Arechiga AF, Habib T, He Y, et al. Cutting edge: the PTPN22 allelic variant associated with autoimmunity impairs B cell signaling. J Immunol. 2009;182:3343–3347.
10. Bottini N, Musumeci L, Alonso A, et al. A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. Nat Genet. 2004;36:337–338.
11. Liu Y, Stanford SM, Jog SP, et al. Regulation of lymphoid tyrosine phosphatase activity: inhibition of the catalytic domain by the proximal interdomain. Biochemistry. 2009;48:7525–7532.
12. Orru V, Tsai SJ, Rueda B, et al. A loss-of-function variant of PTPN22 is associated with reduced risk of systemic lupus erythematosus. Hum Mol Genet. 2009;18:569–579.
13. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008;40:955–962.
14. Bottini N, Vang T, Cucca F, et al. Role of PTPN22 in type 1 diabetes and other autoimmune diseases. Semin Immunol. 2006;18:207–213.
15. Gregersen PK, Lee HS, Batliwalla F, et al. PTPN22: setting thresholds for autoimmunity. Semin Immunol. 2006;18:214–223.
16. Lee YH, Rho YH, Choi SJ, et al. The PTPN22 C1858T functional polymorphism and autoimmune diseases—a meta-analysis. Rheumatology (Oxford). 2007;46:49–56.
17. Fiorillo E, Orru V, Stanford SM, et al. Autoimmune-associated PTPN22 R620W variation reduces phosphorylation of lymphoid phosphatase on an inhibitory tyrosine residue. J Biol Chem. 2010;285: 26506–25618.
18. Rodriguez-Rodriguez L, Wan Taib WR, Topless R, et al. The PTPN22 R263Q polymorphism is a risk factor for rheumatoid arthritis in Caucasian case-control samples. Arthritis Rheum. 2010, Nov 15. [Epub ahead of print].
19. Gasche C, Scholmerich J, Brynskov J, et al. A simple classification of Crohn's disease: report of the Working Party for the World Congresses of Gastroenterology, Vienna 1998. Inflamm Bowel Dis. 2000;6:8–15.
20. Satsangi J, Silverberg MS, Vermeire S, et al. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut.

2006;55:749–753.

21. Silverberg MS, Satsangi J, Ahmad T, et al. Toward an integrated clinical, molecular and serological classification of inflammatory bowel disease: report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. Can J Gastroenterol. 2005;19(suppl A):5–36.
22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–575.
23. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447:661–678.
24. Anderson CA, Massey DC, Barrett JC, et al. Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. Gastroenterology. 2009;136:523–529 e523.
25. De Jager PL, Sawcer S, Waliszewska A, et al. Evaluating the role of the 620W allele of protein tyrosine phosphatase PTPN22 in Crohn's disease and multiple sclerosis. Eur J Hum Genet. 2006;14:317–321.
26. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science. 2006;314:1461–1463.
27. Franke A, Balschun T, Karlsen TH, et al. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat Genet. 2008;40:713–715.
28. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet. 2007;39:207–211.
29. Hradsky O, Lenicek M, Dusatkova P, et al. Variants of CARD15, TNFA and PTPN22 and susceptibility to Crohn's disease in the Czech population: high frequency of the CARD15 1007fs. Tissue Antigens. 2008;71:538–547.
30. Latiano A, Palmieri O, Valvano MR, et al. Evaluating the role of the genetic variations of PTPN22, NFKB1, and FcGRIIIA genes in inflammatory bowel disease: a meta-analysis. Inflamm Bowel Dis. 2007;13:1212–1219.
31. Lettre G, Rioux JD. Autoimmune diseases: insights from genomewide association studies. Hum Mol Genet. 2008;17:R116–121.
32. Libioulle C, Louis E, Hansoul S, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007;3:e58.
33. Martin MC, Oliver J, Urcelay E, et al. The functional genetic variation in the PTPN22 gene has a negligible effect on the susceptibility to develop inflammatory bowel disease. Tissue Antigens. 2005;66:314–317.
34. Massey DC, Parkes M. Genome-wide association scanning highlights two autophagy genes, ATG16L1 and IRGM, as being significantly associated with Crohn's disease. Autophagy. 2007;3:649–651.
35. Mathew CG. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. Nat Rev Genet. 2008;9:9–14.
36. Parkes M, Barrett JC, Prescott NJ, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet. 2007;39:830–832.
37. Prescott NJ, Fisher SA, Onnie C, et al. A general autoimmunity gene (PTPN22) is not associated with inflammatory bowel disease in a British population. Tissue Antigens. 2005;66:318–320.
38. Raelson JV, Little RD, Ruether A, et al. Genome-wide association study for Crohn's

disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A. 2007;104: 14747–14752.

39. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet. 2007;39:596–604.
40. Silverberg MS, Cho JH, Rioux JD, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. Nat Genet. 2009;41:216–220.
41. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics. 2008;92:265–272.
42. van Oene M, Wintle RF, Liu X, et al. Association of the lymphoid tyrosine phosphatase R620W variant with rheumatoid arthritis, but not Crohn's disease, in Canadian populations. Arthritis Rheum. 2005;52:1993–1998.
43. Wagenleiter SE, Klein W, Griga T, et al. A case-control study of tyrosine phosphatase (PTPN22) confirms the lack of association with Crohn's disease. Int J Immunogenet. 2005;32:323–324.
44. Yamazaki K, McGovern D, Ragoussis J, et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet. 2005;14:3499–3506.
45. McGovern DP, Gardet A, Torkvist L, et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet. 2010;42:332–337.
46. Wang K, Baldassano R, Zhang H, et al. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. Hum Mol Genet. 2010;19:2059–67.
47. Morgan AR, Han DY, Huebner C, et al. PTPN2 but not PTPN22 is associated with Crohn's disease in a New Zealand population. Tissue Antigens. 2010;76:119–125.
48. Rose NR. Predictors of autoimmune disease: autoantibodies and beyond. Autoimmunity. 2008;41:419–428.
49. Chatenoud L. CD3-specific antibodies as promising tools to aim at immune tolerance in the clinic. Int Rev Immunol. 2006;25:215–233.
50. Behr MA, Divangahi M, Lalande JD. What's in a name? The (mis)labeling of Crohn's as an autoimmune disease. Lancet. 2010;376:202–203.
51. Gomez LM, Anaya JM, Martin J. Genetic influence of PTPN22 R620W polymorphism in tuberculosis. Hum Immunol. 2005;66:1242–1247.
52. Lamsyah H, Rueda B, Baassi L, et al. Association of PTPN22 gene functional variants with development of pulmonary tuberculosis in Moroccan population. Tissue Antigens. 2009;74:228–232.
53. Dean M, Carrington M, O'Brien SJ. Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet. 2002;3:263–292.
54. Brant SR. Exposed: the genetic underpinnings of ulcerative colitis relative to Crohn's disease. Gastroenterology. 2009;136:396–399.
55. Diaz-Gallo LM, Palomino-Morales RJ, Gomez-Garcia M, et al. STAT4 gene influences genetic predisposition to ulcerative colitis but not Crohn's disease in the Spanish population: a replication study. Hum Immunol. 2010;71:515–519.

Supplementary Table 1. Allele frequencies distribution of the R263Q polymorphism in three Caucasian cohorts according to the clinical classification variables of CD[18-20] and healthy controls.

Population	Clinical Variable (N)[b]	2N	G	%	A	%	p value	OR	95% CI	
Spain	**Diagnosis Age A (248)**									
	<16 (A1)	66	64	96.97	2	3.03	0.94	1.39	0.39	5
	17 - 40 (A2)	332	317	95.48	15	4.52	0.19	1.52	0.88	2.62
	>40 (A3)	98	95	96.94	3	3.06	0.95	1.26	0.43	3.72
	Disease Location L (482)									
	Ileal (L1)	416	395	94.95	21	5.05	0.41[c]	1.68	1.04	2.7
	Colonic (L2)	178	171	96.07	7	3.93	0.58	1.41	0.66	2.9
	Ileocolonic (L3)	370	356	96.22	14	3.78	0.53	1.27	0.73	2.22
	Upper GI Tract (L4)	40	40	100.00	0	0.00	-	-	-	-
	Disease Behavior B (340)	680	651	95.74	29	4.26				
	Perforating (B3)	198	191	96.46	7	3.54	0.78	1.26	0.59	2.68
	Stricturing (B2)	90	88	97.78	2	2.22	0.61	1.02	0.29	3.63
	Inflammatory (B1)	392	372	94.90	20	5.10	0.29[c]	1.76	1.1	2.85
	Controls (1685)	3370	3263	96.82	107	3.18				
	Diagnosis Age A (452)									
	<16 (A1)	152	147	96.71	5	3.29	0.61	1.48	0.6	3.65
	17 - 40 (A2)	584	573	98.12	11	1.88	0.33	0.76	0.39	1.47
	>40 (A3)	168	164	97.62	4	2.38	0.86	1.1	0.42	2.96
	Disease Location L (425)									
	Ileal (L1)	226	220	97.35	6	2.65	0.97	1.16	0.5	2.67

Dutch	Colonic (L2)	206	203	98.54	3	1.46	0.32	0.72	0.24	2.15
	Ileocolonic (L3)	418	409	97.85	9	2.15	0.6	0.89	0.44	1.81
	Upper GI Tract (L4)	92	91	98.91	1	1.09	0.37	0.79	0.15	4.1
	Disease Behavior B (446)									
	Perforating (B3)	194	189	97.42	5	2.58	0.98	1.15	0.47	2.83
	Stricturing (B2)	372	364	97.85	8	2.15	0.61	0.9	0.43	1.89
	Inflammatory (B1)	326	319	97.85	7	2.15	0.63	0.91	0.42	1.99
	Controls (863)	1726	1681	97.39	45	2.61				
	Diagnosis Age A (497)									
	<16 (A1)	112	111	99.11	1	0.89	0.39	0.42	0.06	3.19
	17 - 40 (A2) + >40 (A3)	882	862	97.73	20	2.27	0.75	1.1	0.61	2.02
	Disease Location L(493)									
New Zealand[a]	Ileal (L1)	576	568	98.61	8	1.39	0.33	0.72	0.33	1.59
	Colonic (L2)	410	397	96.83	13	3.17	0.2	1.61	0.81	3.17
	Disease Behavior B (497)									
	Inflammatory (B1)	564	549	97.34	15	2.66	0.43	1.33	0.69	2.54
	(B2+B3)	430	424	98.60	6	1.40	0.38	0.75	0.31	1.81
	Controls (559)	1118	1095	97.94	23	2.06				

[a]Only showed the available data. [b]The meta-analysis calculated through the fixed effects model for each clinical variable did not show significant associations. [c]FDR correction p-value (based on nine comparisons)

Supplementary Table 2. Allelic frequencies of the R620W polymorphism in three Caucasian cohorts according to the clinical classification of CD[18-20] and healthy controls.

Population	Clinical Variable (N)	2N	C	%	T	%	p value	OR	95% CI	
	Diagnosis Age A (248)									
	<16 (A1)	66	60	90.9	6	9.1	0.45	1.58	0.69	3.59
	17 - 40 (A2)	332	316	95.2	16	4.8	0.17	0.74	0.44	1.24
	>40 (A3)	98	95	96.9	3	3.1	0.15	0.57	0.2	1.68
	Disease Location L (482)									
	Ileal (L1)	416	394	94.7	22	5.3	0.26	0.8	0.51	1.25
	Colonic (L2)	178	170	95.5	8	4.5	0.24	0.73	0.36	1.46
Spain	Ileocolonic (L3)	370	349	94.3	21	5.7	0.44	0.87	0.55	1.37
	Upper GI Tract (L4)	40	39	97.5	1	2.5	0.29	0.69	0.13	3.54
	Disease Behavior B (340)									
	Perforating (B3)	198	187	94.4	11	5.6	0.52	0.88	0.48	1.62
	Stricturing (B2)	90	84	93.3	6	6.7	0.98	1.14	0.51	2.55
	Inflammatory (B1)	392	373	95.2	19	4.8	0.15	0.74	0.46	1.19
	Controls (1685)	3370	3143	93.3	227	6.7				
	Diagnosis Age A (697)									
	<16 (A1)	654	620	94.8	34	5.2	$5.4*10^{-3c}$	0.53	0.36	0.77
	17 - 40 (A2)	578	551	95.3	27	4.7	$1.8*10^{-3c}$	0.48	0.31	0.72
	>40 (A3)	162	146	90.1	16	9.9	0.91	1.1	0.63	1.85
	Disease Location L (417)									
	Ileal (L1)	220	211	95.9	9	4.1	0.06[c]	0.44	0.23	0.86
	Colonic (L2)	206	194	94.2	12	5.8	0.08	0.62	0.34	1.13
Dutch	Ileocolonic (L3)	408	374	91.7	34	8.3	0.43	0.87	0.59	1.29
	Upper GI Tract (L4)	84	80	95.2	4	4.8	0.13	0.58	0.22	1.52
	Disease Behavior B (435)									
	Perforating (B3)	186	175	94.1	11	5.9	0.1	0.64	0.34	1.18

	Stricturing (B2)	364	343	94.2	21	5.8	0.18[c]	0.6	0.38	0.95
	Inflammatory (B1)	320	294	91.9	26	8.1	0.41	0.86	0.56	1.32
	Controls (834)	1668	1508	90.4	160	9.6				
	Diagnosis Age A (465)									
	<16 (A1)	106	96	90.6	10	9.4	0.87	0.94	0.48	1.86
	17 - 40 (A2)	-	-	-	-	-	-	-	-	-
	>40 (A3)	-	-	-	-	-	-	-	-	-
	Disease Location L(461)									
New Zealand[a]	Ileal (L1)	542	504	93.0	38	7.0	0.44[c]	0.68	0.47	1
	Colonic (L2)	380	351	92.4	29	7.6	0.18	0.75	0.49	1.45
	Ileocolonic (L3)	-	-	-	-	-	-	-	-	-
	Upper GI Tract (L4)	-	-	-	-	-	-	-	-	-
	Disease Behavior B (465)									
	Inflammatory (B1)	528	492	93.2	36	6.8	0.36[c]	0.66	0.45	0.98
	(B2+B3)	402	371	92.3	31	7.7	0.19	1.32	0.87	2
	Controls (563)	1126	1014	90.1	112	9.9				
Pooled[b,c]	**Diagnosis Age**									
	<16 (A1)	826	776	93.9	50	6.1	0.038[c]	0.64	0.47	0.88
	Disease Location									
	Ileal (L1)	1178	1109	94.1	69	5.9	$9*10^{-3}$[c]	0.64	0.49	0.84

[a]Only showed the available data. [b]Meta-analysis calculated through the fixed effects model and only showed the most relevant results. [C]FDR correction p-value (based on nine comparisons). We did not observe significant associations after the pooled analysis between the different subsets of the diseases (i.e. A1 vs. A2+A3).

Supplementary Table 3. Allele frequencies distribution of the R263Q polymorphism in three Caucasian cohorts according to the clinical classification variables of UC[18-20] and healthy controls.

Population	Clinical Variable (N)[b]	2N	G	%	A	%	p value	OR	95% CI	
	Diagnosis Age A (212)									
	<16 (A1)	12	12	100	0	0	-	-	-	-
	17 - 40 (A2)	232	229	98,71	3	1,29	0,11	0,53	0,18	1,54
	>40 (A3)	180	176	97,78	4	2,22	0,47	0,85	0,33	2,22
Spain	**Disease Extension E (388)**									
	Ulcerative Proctitis (E1)	64	63	98,44	1	1,56	0,46	0,94	0,19	4,282
	Left-side UC (E2)	398	391	98,24	7	1,76	0,12	0,62	0,29	1,3
	Extensive UC (E3)	314	306	97,45	8	2,55	0,54	0,89	0,44	1,8
	Controls (1685)	3370	3263	96,82	107	3,18				
	Diagnosis Age A (406)									
	<16 (A1)	76	74	97,37	2	2,63	0,99	1,46	0,4	5,34
	17 - 40 (A2)	546	535	97,99	11	2,01	0,44	0,82	0,43	1,57
	>40 (A3)	200	194	97	6	3	0,74	1,31	0,57	3,03
Dutch	**Disease Extension E (364)**									
	Ulcerative Proctitis (E1)	126	125	99,21	1	0,79	0,21	0,58	0,11	2,98
	Left-side UC (E2)	204	196	96,08	8	3,92	0,28	1,67	0,79	3,53
	Extensive UC (E3)	398	390	97,99	8	2,01	0,49	0,84	0,4	1,77
	Controls (834)	1726	1681	97,39	45	2,61				
	Diagnosis Age A (481)									
	<16 (A1)	52	52	100	0	0	-	-	-	-
	17 - 40 (A2) + >40 (A3)	910	899	98,79	11	1,21	0,14	0,61	0,29	1,24
New Zealand[a]	**Disease Extension E (476)**									
	Ulcerative Proctitis (E1) + Left-side UC (E2)	384	380	98,96	4	1,04	0,2	0,5	0,17	1,46
	Extensive UC (E3)	568	561	98,77	7	1,23	0,23	0,65	0,28	1,49
	Controls (559)	1118	1095	97,94	23	2,06				

[a]Only showed the available data. [b]The meta-analysis calculated through the fixed effects model for each clinical variable did not show significant associations. [C]FDR correction p-value (based on nine comparisons)

Supplementary Table 4. Allele frequencies distribution of the R620W polymorphism in three Caucasian cohorts according to the clinical classification variables of UC[18-20] and healthy controls.

Population	Clinical Variable (N)[b]	2N	C	%	T	%	p value	OR	95% CI	
	Diagnosis Age A (212)									
	<16 (A1)	12	10	83,33	2	0,33	0,17	3,8	0,94	15
	17 - 40 (A2)	232	217	93,53	15	0,13	0,87	1,01	0,59	1,7
	>40 (A3)	180	163	90,56	17	0,19	0,16	1,5	0,91	2,52
Spain	**Disease Extension E (388)**									
	Ulcerative Proctitis (E1)	64	62	96,88	2	0,06	0,25	0,66	0,18	2,34
	Left-side UC (E2)	398	372	93,47	26	0,13	0,88	1	0,66	1,51
	Extensive UC (E3)	314	296	94,27	18	0,11	0,49	0,88	0,54	1,44
	Controls (1685)	3370	3143		227					
	Diagnosis Age A (406)									
	<16 (A1)	76	68	89,47	8	0,21	0,79	1,2	0,59	2,54
	17 - 40 (A2)	540	511	94,63	29	0,11	0.018[c]	0,55	0,37	0,82
	>40 (A3)	196	186	94,90	10	0,10	0.351[c]	0,55	0,29	1,05
Dutch	**Disease Extension E (359)**									
	Ulcerative Proctitis (E1)	122	111	90,98	11	0,18	0,83	1	0,54	1,88
	Left-side UC (E2)	200	188	94,00	12	0,12	0,097	0,64	0,36	1,17
	Extensive UC (E3)	396	377	95,20	19	0,10	0.018[c]	0,5	0,31	0,8
	Controls (834)	1668	1508		160					
	Diagnosis Age A (431)									
	<16 (A1)	50	48	96,00	2	0,08	0,16	0,38	0,09	1,57
	17 - 40 (A2) + >40 (A3)	812	729	89,78	83	0,20	0,84	1,03	0,77	1,39
New Zealand[a]	**Disease Extension E (426)**									
	Ulcerative Proctitis (E1) + Left-side UC (E2)	532	476	89,47	56	0,21	0,52	1,15	0,75	1,78
	Extensive UC (E3)	320	292	91,25	28	0,18	0,52	0,89	0,58	1,37
	Controls (563)	1126	1014		112					

[a]Only showed the available data. [b]The meta-analysis calculated through the fixed effects model for each clinical variable did not show significant associations. [c]FDR correction p-value (based on nine comparisons)

CHAPTER 5

Limited evidence for parent-of-origin effects in Inflammatory Bowel Disease associated loci

Karin Fransen MD, Mitja Mitrovic*, Cleo C. van Diemen, PhD, Thelma BK, PhD, Ajit Sood, MD, PhD, Andre Franke, PhD, Stefan Schreiber, MD, PhD, Vandana Midha, MD, Garima Juyal, PhD, Uros Potocnik, PhD, Jingyuan Fu, PhD, Ilja Nolte, PhD and Rinse K. Weersma, MD, PhD.*

PLoS ONE 2012 sept. 7(9):e45287.

**These authors contributed equally*

Abstract

Background: Genome-wide association studies of two main forms of inflammatory bowel diseases (IBD), Crohn's disease (CD) and ulcerative colitis (UC), have identified 99 susceptibility loci, but these explain only ~23% of the genetic risk. Part of the 'hidden heritability' could be in transmissible genetic effects in which mRNA expression in the offspring depends on the parental origin of the allele (genomic imprinting), since children whose mothers have CD are more often affected than children with affected fathers. We analyzed parent-of-origin (POO) effects in Dutch and Indian cohorts of IBD patients.

Methods: We selected 28 genetic loci associated with both CD and UC, and tested them for POO effects in 181 Dutch IBD case-parent trios. Three susceptibility variants in NOD2 were tested in 111 CD trios and a significant finding was re-evaluated in 598 German trios. The UC-associated gene, BTNL2, reportedly imprinted, was tested in 70 Dutch UC trios. Finally, we used 62 independent Indian UC trios to test POO effects of five established Indian UC risk loci.

Results: We identified POO effects for NOD2 (L1007fs; OR = 21.0, P-value = 0.013) for CD; these results could not be replicated in an independent cohort (OR = 0.97, P-value = 0.95). A POO effect in IBD was observed for IL12B (OR = 3.2, P-value = 0.019) and PRDM1 (OR = 5.6, P-value = 0.04). In the Indian trios the IL10 locus showed a POO effect (OR = 0.2, P-value = 0.03).

Conclusions: Little is known about the effect of genomic imprinting in complex diseases such as IBD. We present limited evidence for POO effects for the tested IBD loci. POO effects explain part of the hidden heritability for complex genetic diseases but need to be investigated further.

Introduction

Crohn's disease (CD) and ulcerative colitis (UC) are the two main forms of chronic relapsing inflammatory bowel diseases (IBD). With a cumulative prevalence of up to 800 per 100,000 in Europe and 570 in North America [1], it is considered one of the most common immune-related diseases worldwide. Typically, from their second or third decade on, patients suffer from a chronic relapsing inflammation of the gut, which is often accompanied by extra-intestinal manifestations and complications that can be extremely debilitating and severe. Treatments are costly and often insufficient and can be accompanied by severe side-effects [2]. Hence, there is an urgent need for new therapeutic targets and curative medication. The pathogenesis is largely unknown and it is currently thought that an aberrant immune response to commensal microflora in a genetically susceptible host underlies the disease [3].

Prior to the introduction of genome-wide association studies (GWAS), only three loci had been consistently associated with either form of IBD. Over the past six years, multiple GWAS and meta-analyses have yielded a lengthening list of variants associated with CD (71 confirmed independent genetic risk loci) and UC (47 loci) [4,5]. Nevertheless, despite this encouraging progress, as much as 77% of the estimated heritability for CD and 72% for UC is still considered to be unexplained [4,5]. Thus, one of the challenges in the post-GWAS era is to identify potential sources of this 'hidden heritability' [6], which may reside in associated variants with lower odds ratios, gene-gene interactions, gene-environment interactions, and/or structural variation.

In addition, parent-of-origin effects (POO) may comprise a piece of the missing heritability puzzle in IBD, as suggested by Akolkar et al. [7]. They show that offspring of mothers with CD are at higher risk for CD than when fathers are affected. More recently, Zelinkova et al. showed there was maternal imprinting and female predominance in familial Crohn's disease [8]. This could be explained with at least two distinct types of POO mechanisms (Fig. 1). If the paternal allele is inactivated by genomic imprinting, then expression of the locus is determined only by the maternal allele (Fig. 1a). If this effect is not

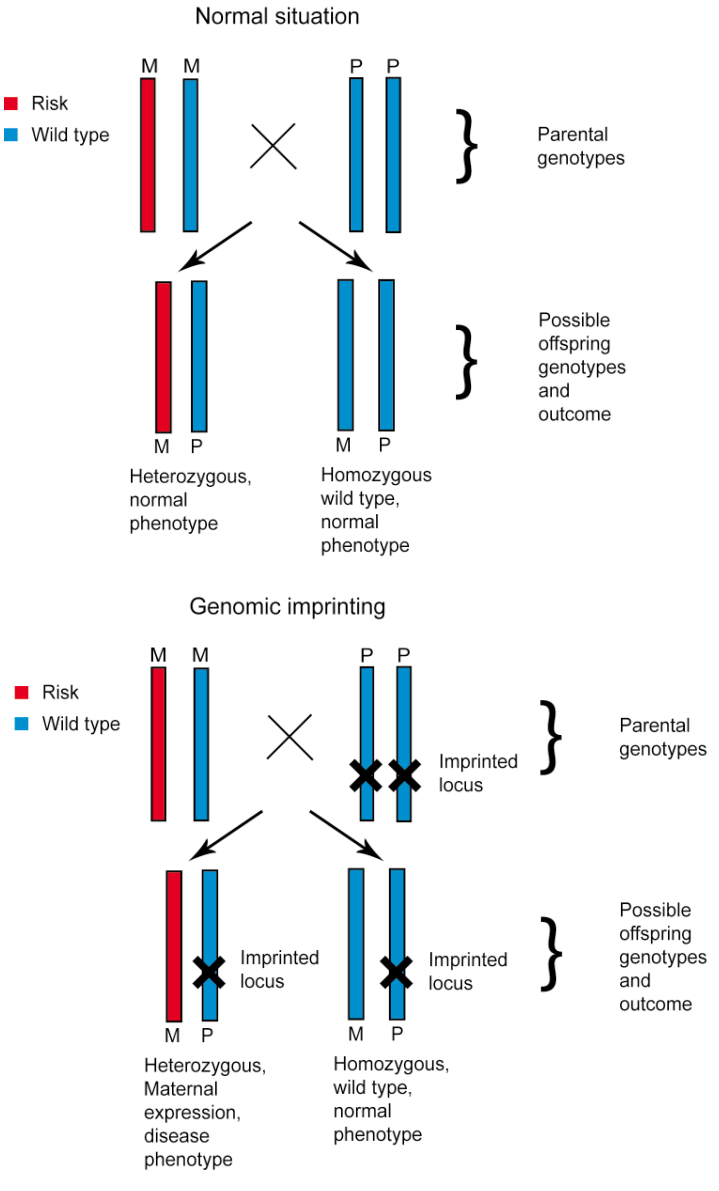


Figure 1. Distinct types of parent-of-origin mechanisms tested in this study.

Fig. 1a. Genomic imprinting. Genomic imprinting is characterized by consequent silencing of one allele, depending on the parental origin. In the example shown above a normal situation is displayed on the left and the genomic imprinting is shown on the right; red is the risk allele and bleu is the wild type allele. The maternal genotype is heterozygous, the father's genotype is homozygous wild-type. Offspring in the left scenario have a normal phenotype since the paternal wild-type allele is expressed in the heterozygous offspring and the mutated allele of the mother is thus rescued by the paternal allele. On the right genomic imprinting is shown, reflecting the □-term in the method used to test for parent of origin effects. In this example there is a significant genomic imprinting effect and the OR >1 so the paternal allele is silenced (see materials and methods section statistical analysis). We assume an additive or recessive model of inheritance. Two possible outcomes are listed, if the offspring inherits the risk allele from the mother and the wild-type allele from the father is subjected to genomic imprinting, then only the risk allele is expressed, thus the offspring is affected by the mutated allele from the mother.

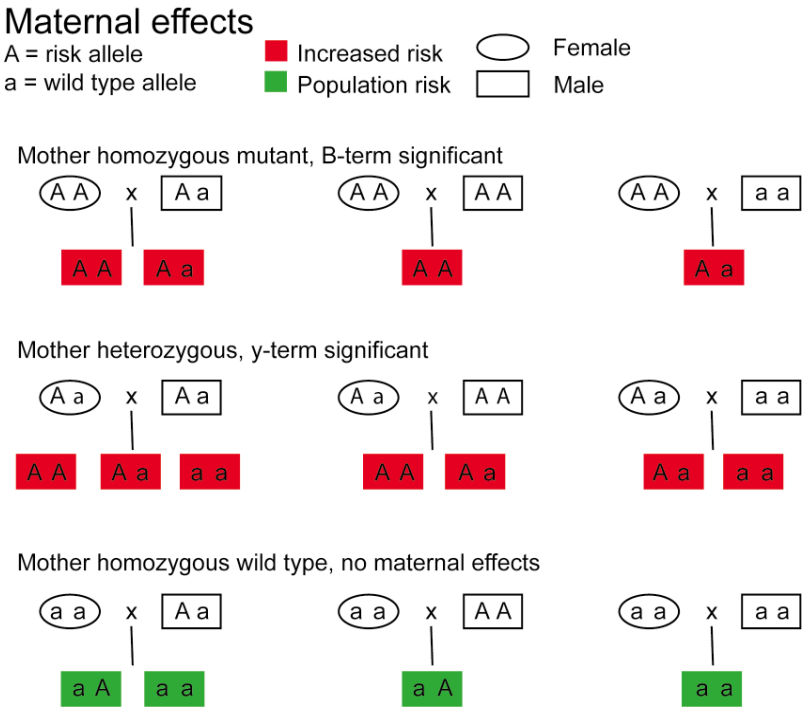


Fig. 1b. Maternal effects. Maternal effects are effects of the maternal genotype on the fetal phenotype, irrespective of the fetal genotype, these effects are reflected by the β- and γ-terms in the likelihood ratio test that was used to test for parent of origin effects in our study. In the example given above, the β- and γ-terms are significant with an OR >1, meaning that the risk of disease is higher if the mother carries two or one risk allele respectively. A recessive or co-dominant model is assumed, and higher expression of the mutant allele leads to disease. If the genotype of the offspring is red, then maternal effects cause increased disease risk and if it is green than the normal population risk applies. If the mother is homozygous wild-type, no maternal effects occur. If she is homozygous mutant or heterozygous for the risk allele, the offspring is subjected to maternal effects and thus has an increased disease risk. Note that the wild-type homozygous offspring has a higher disease risk if both parents are heterozygous.

taken into account, there may be a significant loss in the statistical power of genetic association studies [9,10]. Secondly, maternal effects such as diet or genotype affect the environment for the developing fetus (Fig. 1b). It is thought that maternal proteins or circulating RNA passes the placental barrier and may cause changes in the epigenome of fetal DNA, thereby influence its phenotype.

In this study, we tested for these two types of POO effects in IBD by using a likelihood ratio test developed by Weinberg [11]. A previous version of this method (parental asymmetry test) has already successfully identified a POO effect in another complex genetic disease, type 1 diabetes [12].

Materials and methods

Ethical considerations

This study was approved by the institutional review boards (Institutional ethical committee, Dayanand Medical College and Hospital, Ludhiana and Institutional ethical committee, University of Delhi South Campus; Ethical Review Board of the Medical Faculty of the Christian-Albrechts-University of Kiel; Institutional review board, University Medical Centre Groningen, The Netherlands) of each of the hospitals and written informed consent was obtained from all subjects personally.

Table 1. Phenotypic characterization of subjects with Crohn's disease.

Cohort	Number of trios	No. of males (%)	AOO	Ileal	Colonic	Ileocolon	Upper GI
Dutch CD	115	39 (34%)	24	23	28	64	11
German CD	598	N/A	N/A	N/A	N/A	N/A	N/A

Table 2. Phenotypic characterization of subjects with ulcerative colitis.

Cohort	Number of trios	No. of males (%)	AOO	Proctitis	Left-sided	Extended	Unknown
Dutch UC	72	30 (42%)	25	8	19	38	5
Indian UC	62	45 (73%)	28	22	15	20	5

N/A not available, AOO average age of onset. Cases and disease location are given according to the Montreal classification. for CD L1, L2, L3 and L4; for UC E1, E2, E3. No phenotypic information was available for the German cohort.

Genotyping and quality control

All Dutch and Indian subjects were genotyped using the Illumina Immunochip (iCHIP) (Illumina Inc., San Diego, California, United

States of America), which is a custom-made genotyping array that contains ~200,000 single nucleotide polymorphisms (SNPs) focusing on immune-mediated diseases [15,16]. Genotyping was performed according to the manufacturer's protocol. Genotyping clusters of the SNPs included in the current analysis were checked manually using GenomeStudio software by Illumina Inc. [15,16]. Individuals with a call rate < 95% and/or discordant gender information, and SNPs with a call rate < 98% were removed from further analysis. Identity-by-descent analysis by Plink software was used to test for incorrect family relations (Mendelian errors) in the trios, but no mismatches were identified.

SNP selection

For this study we used several strategies to select SNPs. First, to gain power we pooled the Dutch UC and CD trios and tested for POO effects in 28 established IBD loci that are both associated to CD and UC [4]. To avoid losing significance due to multiple testing correction we tested the 28 overlapping loci instead of all 99 associated risk loci. Variant rs736289 is not present on iCHIP and no proxy (r2 > 0.5) could be found. Two variants (rs12261843, rs181359) were also not captured by iCHIP, but we identified perfect proxies using SNAP software: rs12261843 was represented by rs12254167 (r2 = 1; D' = 1) and rs181359 was represented by rs2266961 (r2 = 1; D' = 1) [17]. Second, we aimed to test for the existence of POO effects in the UC risk SNPs established in Indians [18]. In addition, we aimed to include SNPs from known imprinted genes in our analysis. For this, a publicly available database of known imprinted genes was compared with all 99 IBD-associated loci [19]. The associated locus was defined as the region of r2 > 0.5 flanking the most significantly associated SNP, then extended to the nearest recombination hot-spot, and from there for an extra 100 kb. Comparison of IBD-associated genes with the known imprinted genes resulted in the inclusion of one extra gene, BTNL2; since this is a UC-specific locus it was only analyzed in the Dutch and Indian UC cohorts. SNP rs9268853 is the reported UC risk SNP in the

Caucasians and was tested in the Dutch trios, rs3763313 was tested for POO effect in the Indian trios since this is the reported risk SNP in the Indian population. Lastly, we included NOD2 since it is the most strongly associated gene and is most replicated in association studies of CD in populations of western European descent. Three common disease-susceptibility variants (G809R, R702W, and L1007fs) were therefore tested for POO effects in the Dutch CD trios, and subsequently the L1007fs variant was tested in the German replication cohort [20].

Statistical analysis

A power analysis was performed with Quanto software [21] and showed that in Dutch trios (n = 181) we had more than 80% power to detect POO effects of OR ≥ 3 in variants with MAF ≥ 0.025. In Indian trios (n = 62) we had 80% power to detect POO effects of OR ≥ 3 in variants with MAF ≥ 0.075 (see Fig. S1). POO effects were calculated by a log-likelihood ratio test, which is a statistical test used to compare the fit of the null hypothesis (i.e. no evidence/ presence of POO in our case) and the alternative hypothesis. The test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other and can be used to decide whether to reject the null model in favor of the alternative model. Weinberg et al. [11] have developed a log-linear model when a case-parents triad is genotyped and jointly classified according to the number of copies of a particular allele carried by the mother, father, and child (denoted as “M,” “P,” and “C,” respectively), there are 15 possible outcomes (i.e. mating types). The family-specific outcomes (i.e., the cell into which a particular triad is classified) are independent, provided that each family contributes only one case. The counts based on classification of the triads studied can therefore be thought of as distributed according to a 15-cell multinomial. The method is based on consideration of mating types in which the mother and father carry unequally many copies of the variant allele, with further stratification on the number of inherited copies of the allele, C. This second level of conditioning (on C) effectively removes

any effects related jointly to the inherited number of copies and the parental-allele counts M, P. In short, the method is valid if inheritance of allele is Mendelian, if there is parental symmetry within mating types in the population studied, and if the gene under study is not in linkage disequilibrium with another disease-susceptibility gene. First, the relative penetration of the risk allele of the child is established by determining the parental origin of the risk allele. In the latter, the difference in disease risk is compared for the varying amounts of risk alleles carried by the mothers; the genotype of the child is not relevant. The likelihood ratio test calculates α-, ß-, and γ-terms. The α-term indicates the significance level for genomic imprinting effects: if OR > 1, the risk allele is transmitted more often from the mother to the patient and if OR < 1 then it is transmitted more often from the father. The í- and γ-terms indicate the prenatal effect of the maternal genotype when the mother carries two risk alleles or one risk allele, respectively. When the ß- or γ-term is significant and OR > 1 then the child has more chance of getting the disease due to maternal effects. If OR < 1 then the child has less chance of developing the disease as a consequence of this prenatal effect. Bonferroni multiple testing corrections were applied to the four different analyses.

Results

DNA of 249 complete IBD trios was available for our study, of which four CD (4/115) and two Dutch UC trios (2/72) did not pass the quality control. Therefore 243 IBD trios (111 CD, 70 Dutch UC & 62 Indian UC) were available for the discovery phase of the study. Our findings were then replicated in an independent replication cohort consisting of 598 German CD trios.

Parent-of-origin analysis in Dutch IBD trios

A nominally significant genomic imprinting effect was found in the IL12B gene (α term: P = 0.019; OR = 3.2), with OR > 1 indicating that the risk allele is more often transmitted from the mother to the child.

Table 3. Results of the parent-of-origin (POO) analysis of Dutch IBD Trios (n = 181) for the 28 known SNPs shared between ulcerative colitis and Crohn's disease.

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs11209026	*IL23R*	G	0.6	0.6	1.0	1.6	1.0	1.0
rs7554511	*KIF21B*	C	0.2	1.7	0.5	0.7	0.6	0.8
rs3024505	*IL10*	A	0.4	1.5	0.4	0.5	0.3	1.5
rs7608910	*REL*	G	0.4	0.7	0.7	0.8	0.8	0.9
rs2310173	*IL1R2*	T	0.5	1.4	0.3	0.5	0.7	0.9
rs3197999	*MST1*	A	0.5	0.8	0.6	1.4	0.3	1.4
rs6451493	*PTGER4*	T	0.1	2	0.5	1.6	0.4	1.6
rs6871626	***IL12B*§**	A	**0.019**	**3.2**	**0.003**	**0.2**	0.2	0.6
rs6556412	*IL12B*§	A	1.0	1.0	0.9	1.0	0.9	0.9
rs6908425	*CDKAL1*	C	0.5	0.7	0.8	0.8	0.6	0.7
rs6911490	***PRDM1***	T	0.2	0.5	**0.04**	**5.6**	0.6	1.2
rs10758669	*JAK2*	C	1.0	1.0	0.2	0.5	0.8	0.8
rs4246905	*TNFSF15*	C	0.8	0.9	0.8	0.8	0.8	0.9
rs10781499	*CARD9*	A	0.8	1.0	0.5	0.7	0.4	0.8
rs12254167	*CREM CCNY**	N/A	0.1	0.5	0.6	1.4	0.3	1.4
rs10761659	*ZNF365*	G	0.4	1.4	0.8	0.9	0.9	0.9
rs6584283	*NKX2-3*	T	0.3	0.6	0.4	1.6	0.1	1.9
rs2155219	*C11orf30*	T	0.1	0.5	0.3	1.8	0.6	1.2
rs17293632	*SMAD3*	T	0.2	1.8	0.1	0.4	0.3	0.7
rs2872507	*ORMDL3*	A	0.8	0.9	0.3	1.7	0.6	1.2
rs1893217	*PTPN2*	G	0.1	2.1	0.7	1.2	0.3	0.7
rs12720356	*TYK2*	C	0.6	1.4	0.5	0.5	0.9	0.9
rs2297441	*RTEL1-SLC2A4RG*	A	0.7	0.8	0.9	1.1	0.9	0.9
rs1297265	*intergenic*	A	0.9	1.1	0.3	1.7	0.3	1.5
rs2836878	*intergenic*	G	0.6	1.3	0.9	0.9	0.6	1.3
rs2838519	*ICOSLG*	G	0.7	0.8	0.3	1.7	0.8	0.9
rs2266961	*YDJC**	N/A	0.9	1.0	0.8	0.8	0.5	0.8

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-terms. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. N/A not available. Significant associations are shown in bold. P-values displayed in the table are not corrected for multiple testing. *reported SNP not present/captured by the Immunochip, a proxy was used, therefore no risk allele could be reported. r2 = 1; §r2 = 0.03; two independent hits in one gene.

In addition, the β term was nominally significant (p-value = 0.003; OR = 0.2) with OR < 1, indicating that offspring have less chance of getting the disease if their mothers carry two risk alleles. The PRDM1 gene showed a nominally significant maternal effect if the mother carried two risk alleles (β term: p-value = 0.04; OR = 5.6), with OR >1 indicating that the offspring have more chance of getting the disease. The other tests did not result in significant POO effects. However, none of these associations were significant after the Bonferroni correction (table 3).

NOD2 in Dutch and German CD trios

Three established CD variants in NOD2 (G809R, R702W, L1007fs) were tested for POO effect in the 111 CD trios [20]. After correcting for multiple testing, a significant genomic imprinting effect was detected for the L1007fs variant (α term: p-value = 0.013; OR = 21.0). The risk allele was transmitted more often from the mother than the father. Given the high OR we aimed to replicate this finding in an independent German cohort for which NOD2 genotyping data was available. Unfortunately, our results could not be replicated in this cohort (α term: p-value = 0.95; OR = 0.97) (table 4).

Table 4. Results of the parent-of-origin (POO) analysis for the NOD2 variants in Dutch Crohn's disease trios (n = 111) and replication in German Crohn's disease trios (n = 598).

SNP	Gene	p-α	p-α replication	OR-α	p-ß	p-ß replication	OR-ß	p-γ	p-γ replication	OR-γ
G908R	*NOD2*	0.1		9.0	N/A*		N/A*	0.2		0.3
R702W	*NOD2*	0.4		2.3	1.0		0	0.2		0.4
L1007fs	***NOD2***	**0.01**◊	0.9	**21.0**	1.0	1.0	7.7	0.1	0.8	0.2

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-terms. P-value of the replication study (p- α; -ß; - γ replication) of the alpha-, beta-, and gamma-terms. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold. ◊Significant after Bonferroni multiple testing correction. P-values displayed in the table are not corrected for multiple testing. *No homozygous mothers are available for beta-term analysis.

Known imprinted gene BTNL2 in Dutch UC trios

No significant POO effects were detected the established UC SNP in the BTNL2 locus (rs9268853) in 72 Dutch UC trios (table 5).

Table 5. Results of the parent-of-origin (POO) analysis in the BTNL2 locus in Dutch UC ulcerative colitis trios (n = 72).

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs9268853	*BTNL2*	T	1.0	1.0	0.7	0.7	0.6	0.6

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-term. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold. P-values displayed in the table are not corrected for multiple testing.

Indian UC analysis

The established Indian UC SNPs were tested for POO effects in 62 Indian trios [18]. We found a nominally significant genomic imprinting effect in the IL10 locus (p-value = 0.03; OR = 0.16) where the OR < 1 indicates that the risk allele is more often transmitted from the father. This association does not, however, pass the multiple testing correction. This SNP was also tested in the population of western European descent and we could not detect any significant imprinting effect. The NOD2 variant that showed association in the Indian population could not be tested for POO effects since only homozygous wild-type fathers were available, hence all the trios were uninformative (table 6).

Discussion

For the first time parent-of-origin effects have been tested in IBD on a genetic level for the overlapping IBD-associated loci. We found limited evidence that POO effects exist in IBD in the Dutch population for IL12B, PRDM1 and NOD2 in our discovery cohort, but the large POO effect for NOD2 could not be replicated in an independent German replication cohort. Moreover, we found a nominally significant POO

Table 6. Results of the parent-of-origin (POO) analysis in Indian UC trios (n = 62).

SNP	Gene	RA	p-α	OR-α	p-ß	OR-ß	p-γ	OR-γ
rs6426833	*RNF186*	A	0.2	0.3	0.2	4.3	0.1	4.8
rs3024505	***IL10***	A	**0.03**	**0.16**	0.8	1.3	0.5	1.5
rs3763313	*BTNL2*	T	0.8	0.8	1.0	2.0	1.0	4.0
rs2395185	*HLA-DRA*	A	0.3	2.3	0.8	1.3	0.4	3.0

P-value (p-α; ß; γ) and odds ratio (OR-α; ß; γ) of the alpha-, beta-, and gamma-term. Alpha-term indicates the genomic imprinting effect; Beta-term and gamma-term indicate the maternal effect in case the mother carries respectively two and one risk alleles. Significant associations are in bold. P-values displayed in the table are not corrected for multiple testing.

effect in IL10 in our Indian population. Although the results from the Dutch trios might be false-positive, they imply that the paternal allele has been silenced and thus does not increase the disease risk in all genes for which we found a POO effect. This is consistent with results from epidemiological studies that show that IBD is transmitted to offspring more often from the mother than the father.

NOD2 is the most strongly associated and most consistently replicated CD gene. Here we observed a genomic imprinting effect for the L1007fs mutation in Dutch CD trios, yet we failed to replicate this in an independent German cohort. The results in our initial analysis might be a false-positive finding, although we had sufficient power to detect effects in the Dutch trios. Although both cohorts were of western European descent, we question whether population specific and environmental factors might play a major role in POO effects and explain part of the lack of replication. We will further elaborate on this later in the discussion. TheL1007fs mutation seems to have a predominant role in CD families since recently it has been shown in a case report that all family members were carrying the mutation and had CD [22]. Moreover in cases of homozygosity this variant will lead to ileal stenosis [23,24], implying a strong effect of the L1007fs mutation on the disease phenotype.

Our results suggest a possible contradictory effect of the two types of POO effects we studied for the IL12B gene both with

a nominal significance. The genomic imprinting analysis (α-term) showed inheritance of the disease risk from the mother, while the analysis of independent maternal effect showed protection for disease if the mother carries two risk alleles (β-term). This suggests that the maternal risk allele is expressed and causes a higher disease risk, but simultaneously and independently, if the mother carries two risk alleles the child has a lower risk for IBD due to in utero effects on the fetus. IL12B resides on the established and consistently replicated IBD locus on chromosome 5q33 and it encodes a sub-unit of IL23, which is involved in Th17/IL23R signaling. This pathway has been implicated in several chronic, immune-related diseases such as psoriasis, rheumatoid arthritis, and ankylosing spondylitis [25,26,27].

The PRDM1 gene showed a nominally significant maternal effect in the POO analysis in the population of western European descent when mothers carried two risk alleles; the OR of 5.6 supports results from others that IBD is more often transmitted from the mother than the father. The environment in which the fetus develops causes changes that increase the risk for CD. PRDM1 has been associated to several immune-related diseases, but also to various types of lymphomas [28,29,30,31]. It encodes a protein that represses the expression of the β-interferon gene. Hypothetically, if the maternal effect causes altered expression of PRDM1, an aberrant immune response could increase CD risk.

In the Indian trio study we found a nominally significant genomic imprinting effect for the IL10 locus. In contrast to the findings in the population of western European descent, the risk allele was more often transmitted from the father to the child. No epidemiological studies of the Indian population are available to validate the paternal transmission. We could not replicate the POO effects from the Dutch population in the Indian trios nor vice versa. This might indicate that POO effects are population specific. Later in the discussion we will discuss this in further detail.

Weinberg's method to detect POO effect is a robust one. Moreover it takes genomic imprinting and maternal effects into consideration simultaneously. It therefore has less power than the standard parental

asymmetry test (PAT) that only tests for genomic imprinting, but PAT is invalid if maternal effects are present [11]. The importance of these maternal effects has been shown in mice, with a knockout of the serotonin 1A receptor gene leading to an anxiety-like phenotype. Implantation of wild-type embryos into knockout mothers and cross-fostering of the pups with wild-type mothers showed the full anxiety-phenotype, indicating that the maternal genotype influences the phenotype and that this effect persists after birth [32]. This is supported by our evidence for maternal effects in the PRDM1 and IL12B loci.

We do not know why POO effects occur. Humans are diploid organisms and as such, can survive the, on average, 500 recessive mutations that are present in every human being, since most deleterious effects are rescued by the other allele [33]. Genomic imprinting significantly deduces diploidy by consequently inactivating one haplotype depending on its parental origin and thus impairing the rescue mechanism. The most cited and best supported hypothesis for the existence of this counter-intuitive phenomenon is the parental conflict hypothesis, in which both sexes have a need to pass on their genetic information to the next generation. Yet this does not explain the existence of genomic imprinting in immune-related genes, for example [34]. Hypothetically, to prevent adverse reactions passing from mother to fetus, it is important that the immune responses are alike and thus preferably maternal immune genes are expressed.

In our study we had sufficient power (>80%) to detect POO effects with an OR of three or higher for each SNP in our study. By adding more tests the significance level must be adjusted accordingly and the power to detect differences is lower. Therefore we chose to only test the 28 overlapping IBD risk loci in a pooled cohort of Dutch CD and UC trios instead of all 99 risk loci. Consequently, bigger cohorts are needed to test the remaining IBD loci for POO effects.

None of our findings in one population could be replicated in another population. At least three reasons could explain this fact. First, it might be that the initial findings are false positive findings: the cohorts have a limited size and thus more variation around the

mean, resulting in a higher chance of false positive findings. Second, it is unknown for how long genomic imprinting effects are stable in humans. It has been shown from mouse studies that genomic imprinting is stable for at least 3 generations [35]. No data is available in human studies. Moreover , genomic imprinting was shown to be influenced by environmental factors [36,37], which could mean that although the imprinting mechanism is global, distinct genes may be imprinted in different populations because they were exposed to distinct environmental effects. The latter two could indicate that even within populations different imprinting effects occur.

In conclusion, we aimed to identify genomic imprinting effects and maternal effects acting on the risk alleles of IBD and we showed, for the first time, that IL12B, NOD2 and PRDM1 might be involved in these phenomena in Dutch IBD trios. It has already been shown that POO effects exist in type 1- and type 2 diabetes, which like IBD are complex genetic disorders and show a substantial overlap of disease susceptibility loci with IBD [10,12]. Given the high OR in NOD2 we sought to replicate our findings, but could not confirm the POO effect for NOD2 in an independent German replication cohort. In the Indian population we did identify POO effects in the IL10 gene. We could neither replicate our findings from the Dutch trios in the Indian or in the German population nor our findings from the Indian population in the Dutch cohort. This suggests that POO effects are either false-positive findings or prone to be population specific. We anticipate that future investigations, using larger, multi-ethnical cohorts will help to shed light on these complex and currently little known relationships. Parent-of-origin effects can take various forms and are not restricted to imprinting, but may involve a variety of mechanisms including gender effects, epistasis, epigenetic effects, and environmental influences during pre- or postnatal development. Better understanding of such effects will probably require detailed studies of model organisms in which breeding and environment can be carefully controlled. Given that alleles identified through GWAS account for a relatively small fraction of heritability, parent of origin effects may underlie some of the missing heritability problem. With appropriate family-

based study designs data analysis methods and international collaborative efforts it will be possible to screen for parent of origin effects across the entire genome. In addition, epigenetic profiling on a genome scale will likely lead to the identification of novel epigenetic marks in a variety of disorders that may provide a bridge among the parental genome, parental environment, and offspring phenotype. We anticipate that the investigations of alternative models of inheritance, appropriate study design and application of novel technologies will enable a more complete picture of heritability in human traits, leading to new insights in the field of genetics of complex diseases.

Acknowledgments: We thank all the patients who participated in this study and Jackie Senior for editing the manuscript.

Funding: K. Fransen was supported by a MD/PhD grant from GUIDE at University Medical Center Groningen. M. Mitrovic was supported by a PhD grant from Slovenian Research Agency (Grant No. 630-39/2012-1). U. Potocnik was supported by a grant from Slovenian Research Agency (Grant No. J3-2175). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. RKW is supported by a clinical fellowship grant (90.700.281) from the Netherlands Organization for Scientific Research (NWO) and the Broad Medical Research Program of The Broad Foundation.

Competing interests: The authors have declared that no competing interests exist.

References

1. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, et al. (2012) Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 142: 46-54 e42.
2. Di Sabatino A, Liberato L, Marchetti M, Biancheri P, Corazza GR (2011) Optimal use and cost-effectiveness of biologic therapies in inflammatory bowel disease. Intern Emerg Med 6 Suppl 1: 17-27.
3. Nell S, Suerbaum S, Josenhans C (2010) The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. Nat Rev Microbiol 8: 564-577.
4. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43: 246-252.
5. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118-1125.
6. Fransen K, Mitrovic M, van Diemen CC, Weersma RK (2011) The quest for genetic risk factors for Crohn's disease in the post-GWAS era. Genome Med 3: 13.
7. Akolkar PN, Gulwani-Akolkar B, Heresbach D, Lin XY, Fisher S, et al. (1997) Differences in risk of Crohn's disease in offspring of mothers and fathers with inflammatory bowel disease. Am J Gastroenterol 92: 2241-2244.
8. Zelinkova Z, Stokkers PC, van der Linde K, Kuipers EJ, Peppelenbosch MP, et al. (2012) Maternal imprinting and female predominance in familial Crohn's disease. J Crohns Colitis.
9. Hanson RL, Kobes S, Lindsay RS, Knowler WC (2001) Assessment of parent-of-origin effects in linkage analysis of quantitative traits. Am J Hum Genet 68: 951-962.
10. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. Nature 462: 868-874.
11. Weinberg CR (1999) Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. Am J Hum Genet 65: 229-235.
12. Wallace C, Smyth DJ, Maisuria-Armer M, Walker NM, Todd JA, et al. (2010) The imprinted DLK1-MEG3 gene region on chromosome 14q32.2 alters susceptibility to type 1 diabetes. Nat Genet 42: 68-71.
13. Podolsky DK (2002) Inflammatory bowel disease. N Engl J Med 347: 417-429.
14. Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, et al. (2007) Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. Proc Natl Acad Sci U S A 104: 14747-14752.
15. Cortes A, Brown MA (2011) Promise and pitfalls of the Immunochip. Arthritis Res Ther 13: 101.
16. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43: 1193-1201.
17. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938-2939.
18. Juyal G, Prasad P, Senapati S, Midha V, Sood A, et al. (2011) An investigation of genome-wide studies reported susceptibility loci for ulcerative colitis shows limited replication in north Indians. PLoS One 6: e16565.
19. Luedi PP, Dietrich FS, Weidman JR, Bosko JM, Jirtle RL, et al. (2007) Computational and experimental identification of novel human imprinted genes. Genome Res

17: 1723-1730.

20. Mitrovic M, Potocnik U (2011) High-resolution melting curve analysis for high-throughput genotyping of NOD2/CARD15 mutations and distribution of these mutations in Slovenian inflammatory bowel diseases patients. Dis Markers 30: 265-274.
21. Gauderman WJ (2002) Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21: 35-50.
22. Schnitzler F, Seiderer J, Stallhofer J, Brand S (2012) Dominant disease-causing effect of NOD2 mutations in a family with all family members affected by Crohn's disease. Inflamm Bowel Dis 18: 395-396.
23. Jurgens M, Brand S, Laubender RP, Seiderer J, Glas J, et al. (2010) The presence of fistulas and NOD2 homozygosity strongly predict intestinal stenosis in Crohn's disease independent of the IL23R genotype. J Gastroenterol 45: 721-731.
24. Brand S (2012) Homozygosity for the NOD2 p.Leu1007fsX1008 variant is the main genetic predictor for fibrostenotic Crohn's disease. Inflamm Bowel Dis 18: 393-394.
25. Australo-Anglo-American Spondyloarthritis C, Reveille JD, Sims AM, Danoy P, Evans DM, et al. (2010) Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. Nat Genet 42: 123-127.
26. Hollis-Moffatt JE, Merriman ME, Rodger RA, Rowley KA, Chapman PT, et al. (2009) Evidence for association of an interleukin 23 receptor variant independent of the R381Q variant with rheumatoid arthritis. Ann Rheum Dis 68: 1340-1344.
27. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41: 199-204.
28. Raychaudhuri S, Thomson BP, Remmers EF, Eyre S, Hinks A, et al. (2009) Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat Genet 41: 1313-1318.
29. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, et al. (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat Genet 41: 1228-1233.
30. Sokol L (2011) Fox and Blimp in NK-cell lymphoma. Blood 118: 3192-3193.
31. Best T, Li D, Skol AD, Kirchhoff T, Jackson SA, et al. (2011) Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin's lymphoma. Nat Med 17: 941-943.
32. Gleason G, Liu B, Bruening S, Zupan B, Auerbach A, et al. (2010) The serotonin1A receptor gene as a genetic and prenatal maternal environmental factor in anxiety. Proc Natl Acad Sci U S A 107: 7592-7597.
33. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227-1234.
34. Guilmatre A, Sharp AJ (2011) Parent of origin effects. Clin Genet.
35. Yazbek SN, Spiezio SH, Nadeau JH, Buchner DA (2010) Ancestral paternal genotype controls body weight and food intake for multiple generations. Hum Mol Genet 19: 4134-4144.
36. Thompson SL, Konfortova G, Gregory RI, Reik W, Dean W, et al. (2001) Environmental effects on genomic imprinting in mammals. Toxicol Lett 120: 143-150.
37. Wang S, Yu Z, Miller RL, Tang D, Perera FP (2011) Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. Hum Hered 71: 196-208.

Supplementary figures

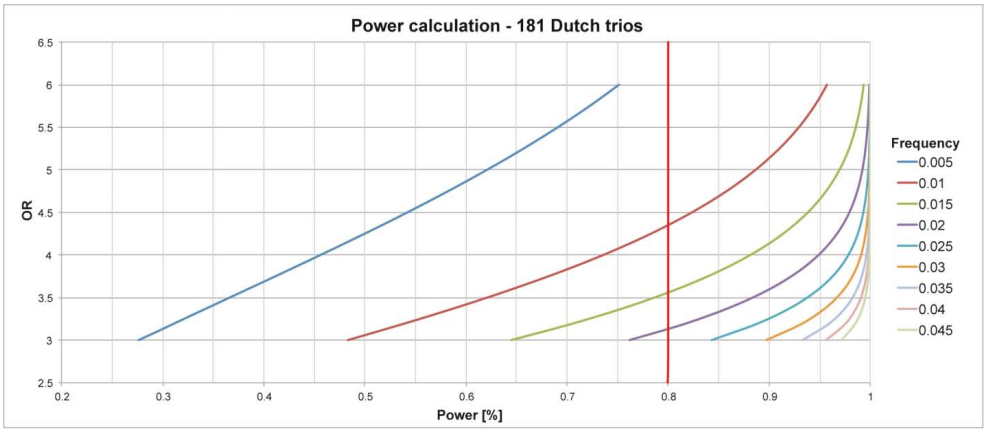
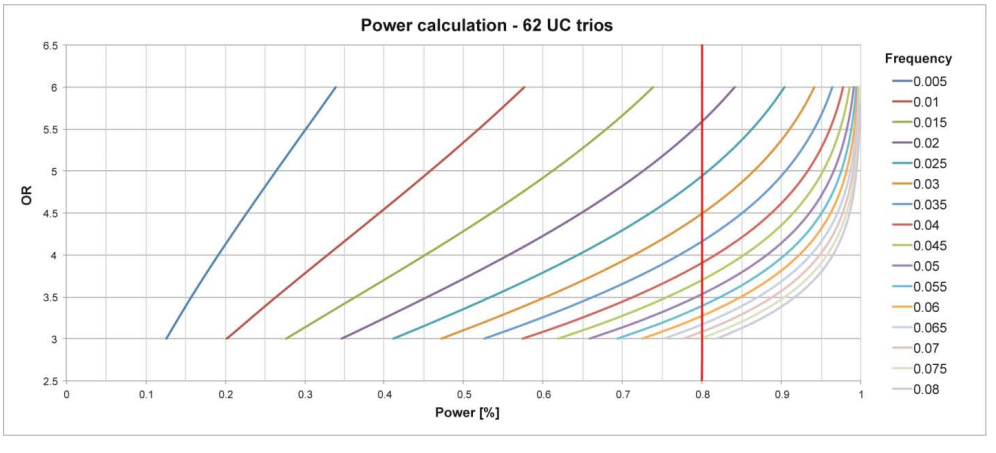


Figure S1. Power analysis of a. the Dutch trio analysis (181 trios) and b. the Indian trios (62 trios).

S1a. Power calculation of the Dutch trio analysis.
The power is shown on the x-axis, the different odds ratios (OR) are shown on the y-axis. The different lines represent SNPs with different minor allele frequencies. In red, the regular 80% power cut-off is shown. With an OR of 3, we have sufficient power to detect parent-of-origin effects in SNPs with a MAF of 2.5%.



S1b. Power calculation of the Indian trio analysis.
The power is shown on the x-axis, the different odds ratios (OR) are shown on the y-axis. The different lines represent SNPs with different minor allele frequencies. In red, the 80% power cut-off is shown. With an OR of 4, we have sufficient power to detect parent-of-origin effects in SNPs with a MAF of 4.0%.

CHAPTER 6

Correlation of genetic risk and mRNA expression in a Th17/IL23 pathway analysis in Inflammatory Bowel Disease

Karin Fransen[1, 2],MD, Suzanne van Sommeren[1, 2]*,MD, Harm-Jan Westra[1], Monique Veenstra[1], Titia Lamberts[1],MD, Rutger Modderman[1], Gerard Dijkstra[2], MD, PhD, Jingyuan Fu[1], PhD, Cisca Wijmenga[1], PhD, Lude Franke[1], PhD, Rinse K. Weersma[2],MD, PhD, Cleo C. van Diemen[1*], PhD.*

Inflamm Bowel Dis. 2014 in press.

**These authors contributed equally*

Abstract

Background: The Th17/IL23 pathway has both genetically and biologically been implicated in the pathogenesis of the Inflammatory Bowel Diseases (IBD), Crohn's disease (CD) and ulcerative colitis (UC). So far, it is unknown whether and how associated risk variants affect expression of the genes encoding for Th17/IL23 pathway proteins.

Methods: 10IBD associated SNPs residing near Th17/IL23 genes were used to construct a genetic risk model in 753 Dutch IBD cases and 1045 controls. In an independent cohort of 40 CD, 40 UC and 40 controls the genetic risk load and presence of IBD were correlated to qPCR generated mRNA expression of nine representative Th17/IL23 genes in both unstimulated and PMA/CaLo stimulated peripheral blood mononuclear cells (PBMCs). In 1240 individuals with various immunological diseases with whole genome genotype and mRNA-expression data we also assessed correlation between genetic risk load and differential mRNA-expression and sought for SNPs affecting expression of all currently known Th17/IL23 pathway genes (*cis*-eQTLs).

Results: The presence of IBD, but not the genetic risk load, was correlated to differential mRNA expression for *IL6* in unstimulated PBMCs and to *IL23A* and *RORC* in response to stimulation. The *cis*-eQTL analysis showed little evidence for correlation between genetic risk load and mRNA expression of Th17/IL23 genes, since we identified for only two out of 22Th17/IL23 genes a *cis*-eQTL SNP that is also associated to IBD (*STAT3* and *CCR6*).

Conclusion: Our results suggest that only the presence of IBD and not the genetic risk load alters mRNA expression levels of IBD associated Th17/IL23 genes.

Introduction

Inflammatory bowel diseases (IBD) can be divided in two main forms – Crohn's disease (CD) and ulcerative colitis (UC). They are disabling diseases characterized by a chronic relapsing inflammatory response to commensal microflora of the gut. The underlying pathogenesis is largely unknown but there is a clear genetic susceptibility. [1,2]

Genome wide association (GWA) studies have been extremely successful in identifying risk loci for IBD; single nucleotide polymorphisms (SNPs) in 163 independent risk loci have been associated to IBD. [3] Functional consequences of the associated loci remain largely unknown while unraveling the functional implication of genetic associations is essential for the clarification of the pathogenesis. Since many loci contain multiple genes, one of the first steps in this process is prioritization of candidate genes. Different gene prioritization methods have been applied to GWAS results, such as identifying protein altering coding SNPs, checking expression quantitative trait locus (eQTL) data, protein-protein interactions (such as DAPPLE), text-mining (such as GRAIL) and co-expression of genes with known implicated genes. [4,5] Because these analyses use interconnectivity between genes they are also able to highlight disease associated pathways. These combined methods resulted in the genetic implication of, amongst others, the Th17/IL23 pathway in the pathogenesis of IBD. [6]

The Th17/IL23 pathway acts in Th17 cells, which are suggested to play a role in the chronic inflammatory processes. Next to the genetic associations, functional studies highlighted the role of the Th17/IL23 pathway in IBD. First, multiple studies showed elevated messenger RNA (mRNA) expression of genes involved in the Th17/IL23 pathway in colonic biopsies of cases compared to controls and inflamed compared to non-inflamed tissue. [7-11] Second, elevated numbers of mucosal Th17 cells have been measured in active versus quiescent Crohn's disease. [12]

The Th17/IL23 pathway contains at least 22 proteins, 10 of the encoding genes reside in loci that are associated to IBD ('IBD associated Th17 genes': *IL23R*, *TYK2, RORC*, *IL21*, *IL12B*, *CCR6*, *JAK2*,

IFNy, *SMAD3* and *STAT3*, figure 1). Despite these breakthroughs, it still remains unclear how the genetic risk variants in the Th17/IL23 pathway exactly contribute to this aberrant function of the Th17/Il23 pathway. The current view is that genetic risk variants alter mRNA expression levels of nearby genes and in this way disturb the function of a pathway. [13-16]

In this study we investigated the correlation between the (combined) genetic risk (load) of IBD associated Th17 genes and the mRNA expression profile of the Th17/IL23 pathway.

Material and methods

More extensive descriptions are available in the supplementary material and methods section.

Samples

All patient samples were collected at the outpatient clinic of the University Medical Centre in Groningen, the Netherlands. IBD patients were diagnosed according to the standard clinical criteria by endoscopy, radiology and histopathology [17] and gave written informed consent.

The 'genetic risk model cohort', consisting of 1798 individuals, was used to construct and test a genetic risk model for the IBD associated Th17 genes. A separate cohort of 118 individuals (39 CD, 40 UC, 39 controls), the 'IBD-PBMC cohort', was used to determine mRNA-expression of nine Th17 representative genes in peripheral blood mononuclear cells (PBMCs), which was then correlated to both disease status and genetic risk load. To prevent influence of immunomodulators and disease activity on the mRNA-expression profile of the patients in the IBD-PBMC cohort, IBD patients did not use any systemic anti-inflammatory drug; only topical mesalasine treatment was allowed, and patients were in clinical remission. [18] Characteristics of these two cohorts are listed in table 1.

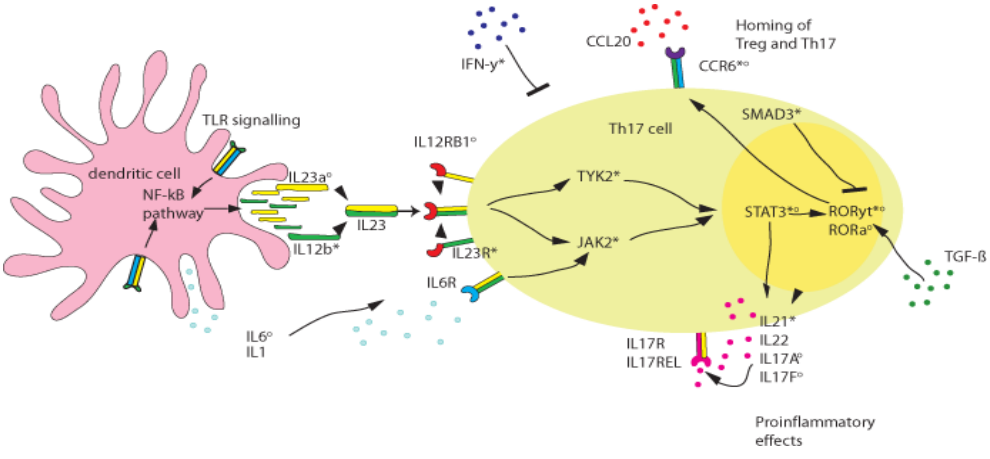


Figure 1.Schematic representation of the T helper 17/IL23 pathway. Antigen presenting cells (APC) are activated by extracellular microbes, e.g. fungi, via the Toll Like Receptor (TLR), and in response produce several cytokines. First, Interleukin-6 (IL6) and Interleukin-1 (IL1) promote the differentiation of naïve T helper cells into Th17 cells via the 9P130 and the IL6 receptor (IL6R); second, Interleukin-23 alpha **(***IL23a)* and Interleukin-12-subunit p40 **(***IL12b)* are produced and form the cytokine complex Interleukin-23 (IL23), which promotes not only differentiation but also proliferation of Th17 cells via de IL23 receptor (IL23R) on Th17 cells. The IL23R consists of two subunits, interleukin-12-receptor-beta-1 **(***IL12RB1)* and *IL23R* and activates together with IL6R and 9p130 the Janus Kinase 2 **(***JAK2)* gene. Simultaneously, Tyrosine Kinase 2 (*TYK2)* is stimulated via the IL23R. *TYK2* and *JAK2* stimulate transcriptionfactors signal transducer and activator of transcription 3 (*STAT3*) and RAR-related orphan receptor C (*RORC, RORyt*) to produce the proinflammatory cytokines Interleukin-21 (IL21), Interleukin-22 (IL22), Interleukin-17 (IL17). These cytokines stimulate Th17 cells to maintain the inflammatory response. Transforming Growth Factor Beta (TGF-β) stimulates *RORC* directly to produce cytokines and to express chemokine (C-C motif) receptor 6 (CCR6). CCR6 responds to the chemo-attractant chemokine (C-C motif) ligand 20 (*CCL20*)which is produced to home other Th17 cells [36]. Interferon gamma (IFN-y) down regulates the Th17/IL23 pathway. °genes are included in the mRNA-expression analysis ('qPCR-Th17'); *genes are included in the genetic risk model.

A third in-house cohort of 1240 patients with various immunological diseases with whole genome genotype and mRNA-expression data from whole blood was used to replicate correlations between genetic risk load and mRNA-expression. Most patients suffered from amyotrophic lateral sclerosis (n=733) and from chronic

obstructive pulmonary disease (n=452) and only few of UC (n=48). Prior to analysis of this dataset, a stringent normalization step was performed by correcting gene expression for the first fifty principal components to account for any effects of batches, patient, laboratory specificity, disease specificity and so forth. These components explain most of the variation of the data. This cohort will be referred to as the 'eQTLcohort', details of the cohort have been published previously. [19]

Table 1. Phenotypes of the genetic risk model cohort and the IBD-PBMC cohort.

	Genetic risk model cohort			IBD-PBMC cohort		
General characteristics	Crohn's disease	ulcerative colitis	controls	Crohn's disease	ulcerative colitis	controls
Number	423	330	1045	39	40	39
Average age at inclusion	46	49	unknown	55	51	34
Number of males (percentage)	161(38%)	181(54%)	561(53%)	17 (43%)	18 (45%)	19 (47%)

Genetic risk model

We hypothesize that the combined effect of multiple SNPs in genes of a pathway has a stronger effect on gene expression of the members of that pathways than single SNPs. Therefore we constructed a genetic risk model for the Th17/IL23 pathway. From the 163 IBD associated loci, the top SNPs of the loci containing genes involved in the Th17/IL23 pathway were selected for the genetic risk model. This resulted in inclusion of 10independent IBD associated SNPs, [3] residing in 10 loci, all containing one Th17 gene (*IL23R*, *RORC*, *IL21*, *IL12B*, *CCR6*, *JAK2*, *IFNy*, *SMAD3*, *STAT3*, *TYK2*,figure 1, supplementary table 2). Genotypes of the 10IBD associated SNPs were weighted for the magnitude of their association with IBD and included in a risk load per individual. A one-way ANOVA test was used to test the difference in genetic risk load between controls and IBD patients of the genetic risk model cohort. Subsequently a t-test was used to determine differences between CD, UC and controls separately.

Effect of disease status and genetic risk load on Th17 mRNA-expression

Nine genes from different stages in the Th17/IL23 pathway were selected for mRNA-measurements('qPCR-Th17 genes': *CCR6*, *STAT3*, *IL17F*, *IL23A*, *IL12RB1*, *RORA*, *IL6*, *RORC*, *IL17A* (figure 1)). They serve as representatives for the different stages of the pathway. We speculated that disease status and differences in genetic risk load not only might give differential gene expression in 'resting' Th17 cells, but can also influence the degree of response of Th17 cells to *ex vivo* stimulation. Therefore, mRNA expression was determined in both unstimulated and T cell specific stimulated (PMA/CaLo) PBMCs from the IBD-PBMC cohort using qPCR measurements.

To test correlations between disease status (IBD and both subtypes separately) or genetic risk score and differential unstimulated mRNA expression, we performed a linear regression analysis with ΔCT values (normalized with *GAPDH*) as outcome variable and either genetic risk load or disease status as a covariate. To test whether disease status or genetic risk load influenced response to stimulation, a linear regression analysis was performed with the ΔCT values of the qPCR-Th17 genes in stimulated PBMCs as outcome variable, and baseline ΔCT values and respectively genetic risk load or disease status as covariates. This is a hypothesis driven analysis and p-values are therefore not corrected for multiple testing.

For replication purpose, we used the eQTL cohort to correlate genetic risk load to mRNA expression of the nine Th17 genes analyzed by qPCR.

eQTL analysis Th17/IL23 pathway

In the eQTL cohort all 22 genes involved in the Th17/IL23 pathway were assessed for SNPs close by (i.e. within 250 Kb, *cis*-eQTL) or far away (i.e> 5Mb away, *trans*-eQTL) affecting expression these genes. If an eQTL-SNP for a Th17 gene was identified, the genetic association with IBD was also assessed.

Results

Genetic risk model

The genetic risk model cohort showed a significant difference in genetic risk load of UC cases and CD cases and controls (mean: controls = 3.0; CD = 3.2; UC = 3.2, p-value one-way ANOVA = $4.8*10^{-8}$). Subsequently, a t-test revealed a significant difference of IBD cases compared to controls for genetic risk load (p-value = $1.4*10^{-5}$), but not between CD and UC cases (p-value=0.9) suggesting that the Th17/IL23 is genetically important for both subtypes of IBD.

Effect of disease status and genetic risk load on Th17 mRNA-expression

In unstimulated PBMCs we only observed a correlation between disease status and mRNA expression levels of *IL6*, *IL23A*, *STAT3* and *RORC*, but no effect of genetic risk load on expression was observed. All correlations between disease status or genetic risk load and mRNA expression of the nine qPCR-Th17 genes are listed in table 2. *IL6* was significantly lower expressed in IBD cases compared to the healthy controls. A trend was observed for higher expression of *RORC* and lower expression of *IL23A* and *STAT3* in cases compared to controls. When CD and UC were analyzed separately these effects were predominantly seen in CD (supplementary table 4).

We also only observed effects of disease presence, and not the genetic risk load on the effect of stimulation. For *IL23A* the difference in expression after stimulation is significantly larger in IBD cases compared to controls, and for *RORC* the stimulation effect was significantly smaller in IBD cases compared to controls. A trend toward a less strong response for cases was observed for *IL17F* (table 2).

In the eQTL cohort, the genetic risk load of each individual was correlated to mRNA expression of all genes present on the expression array in the eQTL cohort. No significant association was observed for any of our nine qPCR-Th17 genes, confirming the lack of association in the IBD-PBMC cohort. Moreover, only one probe, coding for the *RNASET2* gene, of all approximately 23,000 probes on the array was significantly correlated to genetic risk load (p-value = $1,78*10^{-15}$).

Table 2. Effect of disease presence and genetic riskload on mRNA-expression in unstimulated and stimulated PBMCs.

	Unstimulated		Stimulated#	
Gene	Disease status	Genetic risk load	Disease status	Genetic risk load
CCR6	NS	NS	NS	NS
IL23A	lower (p=0.053)	NS	smaller(p=0.002)	NS
IL12RB1	NS	NS	NS	NS
STAT3	lower (p=0.076)	NS	NS	NS
IL17F	NS	NS	larger (p=0.085)	NS
IL17A	NS	NS	NS	NS
IL6	lower (p=0.002)	NS	NS	NS
RORC	higher (p=0.093)	NS	larger(p=0.003)	NS
RORA	NS	NS	NS	NS

Gene: Th17 gene. For the unstimulated PBMCs the correlation between disease status and differential mRNA expression is reported, reported P-values from linear regression. 'Lower' and 'higher' indicate respectively lower and higher mRNA expression in cases compared to controls. For stimulated PBMCs the correlation between disease presence and differential mRNA expression is reported. The linear regression model was adjusted for baseline mRNA expression. 'larger' implies that cases had a larger difference in expression in response to stimulation than healthy controls. 'smaller' implies that cases had a smaller difference in expression in response to stimulation than healthy controls (supplementary figure 1). P-values are not corrected for multiple testing and only p-values below 0.10 are reported. No significant correlation between genetic risk load and baseline mRNA expression or response to stimulation was found. NS= not significant.

Concluding, these analyses suggest that only the presence of IBD, and not genetic risk, influences mRNA-expression of Th17 genes.

eQTL analysis Th17/IL23 pathway

In the eQTL cohort we assessed whether *cis*- or *trans*- eQTL SNPs exist for all 22 genes in the Th17/IL23 pathway (figure 2, supplementary table 5). Out of 22 genes involved in the Th17/IL23 pathway, eight genes had one or more *cis*-eQTL SNP(s). Of these eight genes with *cis*-eQTL SNPs, four genes have previously been associated with IBD. This means that seven from the eleven IBD associated Th17

genes do not have an *cis*-eQTL SNP. These results confirm the limited support for genetic effects in Th17 gene expression in IBD as shown by our qPCR analyses.

Both *STAT3* and *TYK2* have *cis*-eQTL SNPs that are associated with IBD, indicating that the genetic risk effect of these SNPs are, at least partly, driven by eQTL effects. A *cis*-eQTL SNP for *CCR6* is also associated to IBD. However, this SNP also has a much more significant correlation with *RNASET2* expression (p-value = $1{,}78*10^{-15}$), making it questionable if *CCR6* is actually the candidate gene in this locus. A *cis*-eQTL SNP of *RORC* is in high LD with a reported IBD SNP, indicating that this eQTL SNP drives the association of *RORC* to IBD. However, the *cis*-eQTL SNP is only borderline significantly associated to IBD, which contradicts this assumption. Interestingly, we found a SNP that has a strong *cis*-eQTL effect on *IL12RB1*, and is moderately associated with CD (p-value=$7.5*10^{-5}$, supplementary table 5). Considering the biological function of this gene in the Th17/IL23 pathway, this might be an interesting new implication of *IL12RB1* in the genetics of IBD. No *trans*-eQTL effects were detected.

Discussion

Currently it is hypothesized that, in complex diseases, a substantial amount of associated genetic variants influence gene expression of nearby genes, and in this way contribute to disease pathogenesis [13-16]. This study is the first to convert all currently known Th17/IL23 IBD risk variants into a risk load and correlate this to mRNA expression of genes in this pathway. For this study we analyzed mRNA expression levels of nine representative genes in PBMCs of 80 IBD cases and 40 controls, but we did not observe any significant correlations with genetic risk load. In a much larger cohort of 1240 individuals with various immune related diseases we performed the same analyses. There we identified a strong correlation with *RNASET2* mRNA expression with IBD risk load, but not with any of the Th17/IL23 genes. Moreover we sought for individual eQTL SNPs for the 10 IBD associated Th17/IL23 genes in a larger cohort of non-IBD individuals.

We observed that only two out of the 10IBD associated Th17/IL23 genes had a convincing *cis*-eQTL SNP that also increases IBD risk.

We were surprised not to find any correlation between genetic risk load and mRNA expression in our initial cohort. The genetic risk model, constructed from IBD associated Th17/IL23 genetic variants, was highly significant associated to IBD, underscoring the additive effect of the variants on disease risk. Furthermore, since we used PBMCs from patients that did not use any systemic anti-inflammatory treatments, expression patterns were not influenced by this. Additionally, for these genes no eQTL has been described in other publicly available eQTL datasets. [3]

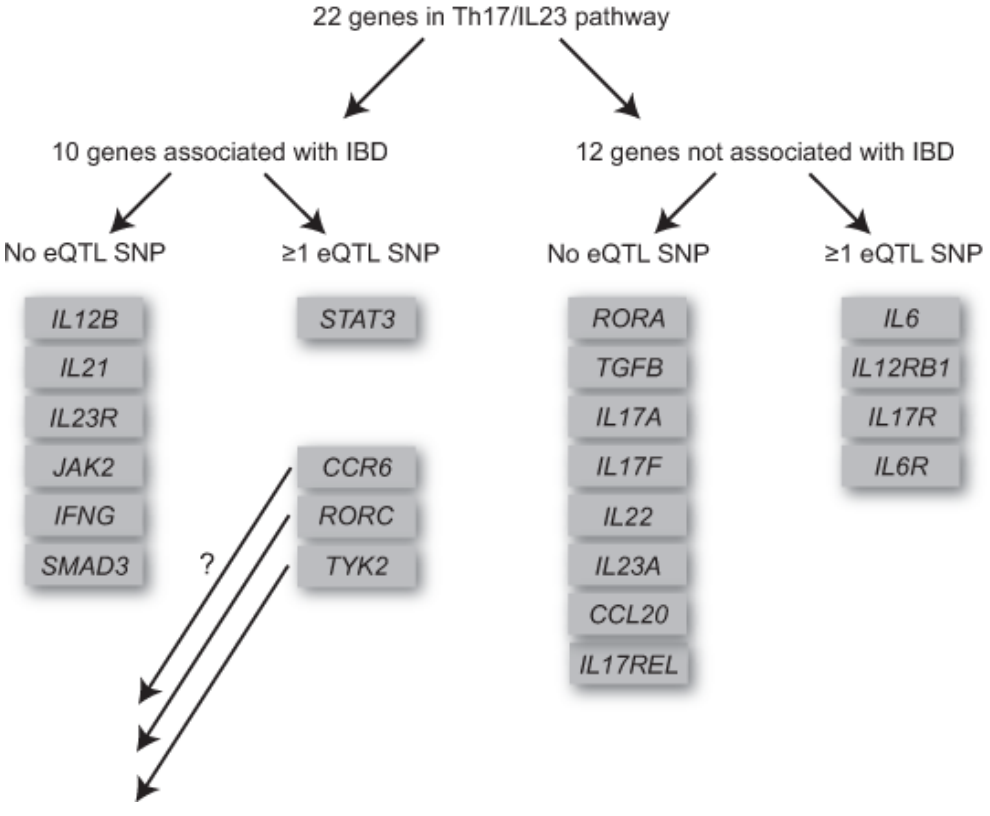


Figure 2. Overview of classification of Th17/IL23 pathway genes based on the presence of an eQTL SNP and association to IBD.
The Th17/IL23 pathway contains at least 22 genes, of which 10are associated with IBD. Eight genes have at least one *cis*-eQTL SNP, of which four are associated with IBD.

One could speculate that our findings result because of lack of power in the initial IBD-PBMC cohort. Although we had sufficient power to find that mRNA expression levels were influenced by IBD status, the genetic effects may be smaller and thus more difficult to pick up. More convincingly, we also did not observe significant correlations between genetic risk load or single SNPs and mRNA expression in the much larger eQTL cohort of 1240 individuals. This cohort already proved to be sufficiently large to detect single SNP effects as shown by Fehrmann*et al.* [19]

Another explanation for lack of correlation resides in the tissue in which mRNA expression was measured in both the IBD-PBMC and eQTL cohort. The Th17 /IL23 pathway is highly specific to the Th17 cells, which have a very low frequency in peripheral blood. Therefore it is possible that the Th17/IL23 mRNA expression effects are diluted to undetectable levels in the heterogeneous cell populations of PBMCs and whole blood. To test this, we analyzed the probe-intensity values of the genes with and without eQTL SNPs in the microarray data of the eQTL cohort. It appeared that indeed the genes without eQTL SNPs have significantly lower expression compared to genes for which an eQTL SNPs could be detected (data not shown). It is also possible that peripheral Th17 cells are phenotypically and expression-wise not the same as mucosal Th17 cells, and that we have investigated the wrong cells.

With this finding in mind, we have to be careful when interpreting eQTL-data. Fu *et al.* found that eQTL effects can be highly tissue specific, which might result in omission of effects when expression is measured in the wrong tissue. [20] eQTL data is usually based on whole blood mRNA expression and the pitfall is that this might result in omission of tissue specific effects. In other words, if mRNA expression is measured in the wrong tissue type, the correlation between the genetic risk variant and the mRNA expression of the true culprit gene cannot be picked up, and correlations with mRNA expression of other genes might falsely be interpreted as being relevant for disease pathogenesis. This might also be the case for the locus containing *CCR6* and *RNASET2*. When analyzing eQTL data

in whole blood this SNP has a very strong *cis*-eQTL effect on *RNASET2*, indicating that this might be the culprit gene. However, *CCR6* is prioritized based on its role in the Th17/IL23 pathway in literature. Hence, prioritizing genes in risk loci based on eQTL data should always be combined with other tools like DAPPLE or GRAIL.

Our risk model includes all SNPs for which Th17/IL23 pathway genes have been prioritized. This inclusion strategy might have resulted in inclusion of SNPs in which the culprit gene is not involved in the Th17/IL23 pathway; and vice versa we might have omitted SNPs and culprit genes involved in the Th17/IL23 pathway with up until now unknown function. The answer to this question will only be definitely answered if causal variants and genes are discovered in all IBD associated loci.

Although at first perhaps surprising, it is very well possible that genetic risk variants do not contribute to disease pathogenesis through altered gene expression. What other mechanisms can account for this? One of the possibilities is an altered protein structure. This can be caused by rare, yet unidentified variants in exomes. Rivas *et al*. identified two such variants in the *IL23R* gene by targeted resequencing of IBD loci. [21] Though the contribution of such rare variants to the missing heritability is probably relatively small. [22] Another possibility is abrogated phosphorylation of genes, as recently shown for compound heterozygous missense mutations in the *IL10*locus which caused IL-10 induced abrogated IL-10R1 phosphorylation. [23] Furthermore, the ENCODE project has shown that regulatory elements of genes can be located at great distance of target genes. [24] This implies that the culprit genes can be at a great distance from the associated loci, in which case we should look for trans-eQTL effects. In the current study we could not identify such *trans*-eQTL effects; this is likely due to lack of power.

In conclusion, our findings show only limited influence of genetic risk variants on gene expression of genes involved in the Th17/IL23 pathway in PBMCs. Moreover prioritizing genes in associated loci based on their eQTL effects in whole blood should be interpreted with caution.

Acknowledgements

We thank all the patients who participated in this study. KF and SvS are supported by a MD/PhD grant from GUIDE at University Medical Center Groningen. LF is supported by a VENI grant (916.10.135) from the Netherlands Organization for Scientific Research (NWO). RKW is supported by a clinical fellowship grant (90.700.281) from NWO and the Dutch Digestive Foundation (WO 11-72). CCD is supported by a NWO Horizon grant (93.511.022). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Nell S, Suerbaum S, Josenhans C. The impact of the microbiota on the pathogenesis of IBD: lessons from mouse infection models. Nat Rev Microbiol. 2010;8:564-577.
2. Molodecky NA, Soon IS, Rabi DM, *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology. 2012; 142:46-54.
3. Jostins L, Ripke S, Weersma RK, *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119-124.
4. Rossin EJ, Lage K, Raychaudhuri S, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genetics. 2011;7:e1001273.
5. Raychaudhuri S, Plenge RM, Rossin EJ, et al. Identifying relationships among genomics disease regions : Predicting genes at pathogenic SNP associations and rare deletions. PLoS Genetics. 2009;5:e1000534.
6. Abraham C, Cho J. Interleukin-23/Th17 pathways and inflammatory bowel disease.Inflamm Bowel Dis. 2009 Jul;15:1090-1100.
7. Pidasheva S, Trifari S, Phillips A, et al. Functional studies on the IBD susceptibility gene IL23R implicate reduced receptor function in the protective genetic variant R381Q. PLoS One. 2011;6:e25038.
8. Olsen T, Rismo R, Cui G, *et al.*TH1 and TH17 interactions in untreated inflamed mucosa of inflammatory bowel disease, and their potential to mediate the inflammation. Cytokine. 2011;56:633-640.
9. Bogaert S, Laukens D, Peeters H,*et al.*Differential mucosal expression of Th17-related genes between the inflamed colon and ileum of patients with inflammatory bowel disease. BMC Immunol. 2010;11:61.
10. Sugihara T, Kobori A, Imaeda H, *et al.*The increased mucosal mRNA expressions of complement C3 and interleukin-17 in inflammatory bowel disease. Clin ExpImmunol.2010;160:386-393.
11. Kobayashi T, Okamoto S, Hisamatsu T, *et al.*IL23 differentially regulates the Th1/Th17 balance in ulcerative colitis and Crohn's disease. Gut. 2008;57:1682-1689.
12. Dige A, Støy S, Rasmussen TK, *et al.* Increased levels of circulating Th17 cells in quiescent versus active Crohn's disease. J Crohns Colitis. 2012 ePub ahead of print.
13. Nicolae DL, Gamazon E, Zhang W, *et al.*Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet. 2010;6:e1000888.
14. Schadt EE, Lamb J, Yang X, *et al.*An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet. 2005; 37:710–717.
15. Emilsson V, Thorleifsson G, Zhang B, *et al.*Genetics of gene expression and its effect on disease. Nature. 2008;452:423-U422.
16. Naukkarinen J, Surakka I, Pietilainen KH, *et al.*Use of Genome-Wide Expression Data to Mine the "Gray Zone" of GWA Studies Leads to Novel Candidate Obesity Genes. PLoS Genet. 2010;6:e1000976.
17. Podolsky DK. Inflammatory bowel disease. N Engl J Med. 2002;347:417-429.
18. Hölttä V, Sipponen T, Westerholm-Ormio M, *et al.*In Crohn's Disease, Anti-TNF-α Treatment Changes the Balance between Mucosal IL-17, FOXP3, and CD4 Cells. ISRN Gastroenterol. 2012;2012:505432.
19. Fehrmann RS, Jansen RC, Veldink JH, *et al.*Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate

genes, with a major role for the HLA. PLoS Genet. 2011;7:e1002197.

20. Fu J, Wolfs MG, Deelen P, *et al*.Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. PLoS Genet. 2012;8:e1002431.
21. Rivas MA, Beaudoin M, Gardet A, *et al*. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43:1066-1073.
22. Hunt KA, Mistry V, Bockett NA,*et al*. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability.Nature. 2013;13;232-235
23. Mao H, Yang W, Lee PP, *et al*. Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. Genes Immun. 2012;13:437-442.
24. The ENCODE project consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57-74.

Supplementary Methods

Samples

Extensive information about the 'genetic risk model cohort' and the 'IBD-PBMC cohort' is described in supplementary table 1.

Genotyping

A custom-made genotyping chip, the Illumina Immunochip(Illumina Inc., San Diego, California, United States of America), was used to genotype all individuals included in the genetic risk model cohort and in the IBD-PBMC cohort. The Immunochip is designed to densely fine-map the approximately 180 immune disease related loci and contains approximately 200,000 SNPs. DNA was hybridized according to the manufactures protocol. Standard quality control was performed. [1] Two out of 120 individuals of the IBD-PBMC cohort failed quality control and were excluded for further analysis.

Genetic risk model

Genotypes of the 10IBD associated SNPs [2] were extracted from the Immunochip data. To create a genetic risk model, the best fitting inheritance models for each of the 10SNPs needed to be determined in the genetic risk model cohort. According to the best fitting inheritance model, numeric genotype scores were assigned for each of the SNPs. For a dominant model homozygous wild type genotypes were scored "null", and homozygous risk and heterozygous SNPs were scored as "two". For a recessive model homozygous wild type and heterozygous genotypes were scored "null" and homozygous risk genotypes were scored "two". In the additive model, homozygous wild type genotypes were scored "null", heterozygous SNPs were scored "one" and homozygous risk SNPs were scored "two". Genotype frequencies of included SNPs can be found in supplementary table 3. Logistic regression analysis was performed for all SNPs for all models using SPSS (PASW statistics 18). The best fitting model was chosen by determining the lowest log likelihood ratio for the three models per SNP. For all SNPs the additive model was the best fitting model.

To incorporate the magnitude of the association for each of the SNPs, we determined the beta-estimate per SNP in a logistic regression model (supplementary table 2) and multiplied the risk score per SNP per individual with this beta-estimate to get a weighted risk score per SNP and per individual. The total genetic risk score per individual was calculated by adding up all weighted risk score for all SNPs. [3,4]

PBMC isolation and gene-expression measurements in the IBD-PBMC cohort

PBMCs were isolated from 20 ml whole blood EDTA tubes using the Ficoll method (Ficoll-Paque™ PLUS, GE Healthcare Europe GmbH, Diegem, Belgium) from 40 UC, 40 CD patients and 40 controls. $10*10^6$ PBMCs were stored in RNA stabilizing agent (RNA later, QIAGEN Benelux B.V, Venlo, Netherlands) ("unstimulated PBMCs") and $10*10^6$ cells were cultured in RPMI supplemented with 0.01ug/ml Phorbol 12-myristate 13-acetate (PMA) and 1ug/ml Calcium ionophore (CaLo) for 16 hours at 37˚C, 5%CO2, and stored in RNAlater at -20˚C ("stimulated PBMCs"). RNA was isolated using the RNeasy Mini Kit (QIAGEN Benelux B.V, Venlo, Netherlands) according to the manufacturers protocol. RNA quanties were measured using Nanodrop and were standardized for further processes. To better assess the quality of the samples we randomly performed gel-based electrophoresis measurements on Experion. mRNA was converted into cDNA using the Revert Aid H minus first strand cDNa kit (Fermentas GmbH, St. Leon-Rot, Germany) for qPCR analysis. mRNA sequences of the nine 'qPCR-Th17 genes' *(CCR6*, *STAT3*, *IL17F*, *IL23A*, *IL12RB1*, *RORA*, *IL6*, *RORC*, *IL17A)* were retrieved from ensemble genome browser (www.ensembl.org) and primers were designed using primer3. [5] Primer sequences are available upon request. SYBR Green/ROX qPCR Master Mix was used for qPCR reactions on mRNA from unstimulated and stimulated PBMCs according to protocol (Maxima™ SYBR Green/ROX qPCR Master Mix, Fermentas GmbH, St. Leon-Rot, Germany) and run on the ABI 7900HT (Applied Biosystems, Foster City, California, USA). CT levels were normalized using CT levels of the housekeeping gene *GAPDH*, thereby correcting for sample and plate-plate variability. Stability

of *GAPDH* expression was observed for the different experimental set-ups. Three individuals had one or more qPCRs that failed and were excluded from analysis when necessary. Per qPCR-Th17 gene (both stimulated and unstimulated) the normality of the distribution of mRNA expression levels was measured by comparing the skewness to the standard error of the skewness. ΔCT values of all qPCR-Th17 genes were normally distributed.

Effect of disease status and genetic risk load on Th17 mRNA-expression in the IBD-PBMC cohort

Since all ΔCT values of all qPCR-Th17 genes were normally distributed, linear regression analysis was used for the different correlation analyses between either disease status or genetic risk load and unstimulated mRNA-expression. To test whether disease status or genetic risk load influenced response to stimulation, a linear regression analysis was performed with the ΔCT values of the qPCR-Th17 genes in stimulated PBMCs as outcome variable, and baseline ΔCT values and respectively genetic risk load or disease status as covariates. We prefer this method over calculating the more widely used ΔΔCT values, since the latter calculates the absolute difference and not the relative difference in ΔCT which is more relevant in our opinion.

Effect of genetic risk load on Th17 mRNA-expression in the eQTL cohort

For replication purpose, we used the large eQTL cohort, to correlate genetic risk load to mRNA expression of the nine qPCR-Th17 genes. In the eQTL cohort two out of 10SNPs from our initial genetic risk model were not present in the GWAS platform, could not be imputed and no perfect proxy could be identified. Therefore these SNPs were not included in the analysis. They include rs6871626 for *IL12B* and rs11879191 for *TYK2*. For three SNPs a proxy was identified, rs1819333 (*CCR6*) was substituted by rs415890 (r^2;D' = 1.0;1.0), rs7657746 (*IL21*) was substituted by rs11734090 (r^2;D' = 0.9;0.9) and rs12942547 (*STAT3*) was substituted by rs744166

(r^2;D' = 1.0;1.0). Correlation between genetic risk load and genome-wide mRNA-expression was tested with the non-parametric spearman correlation test.

eQTL analysis Th17/IL23 pathway

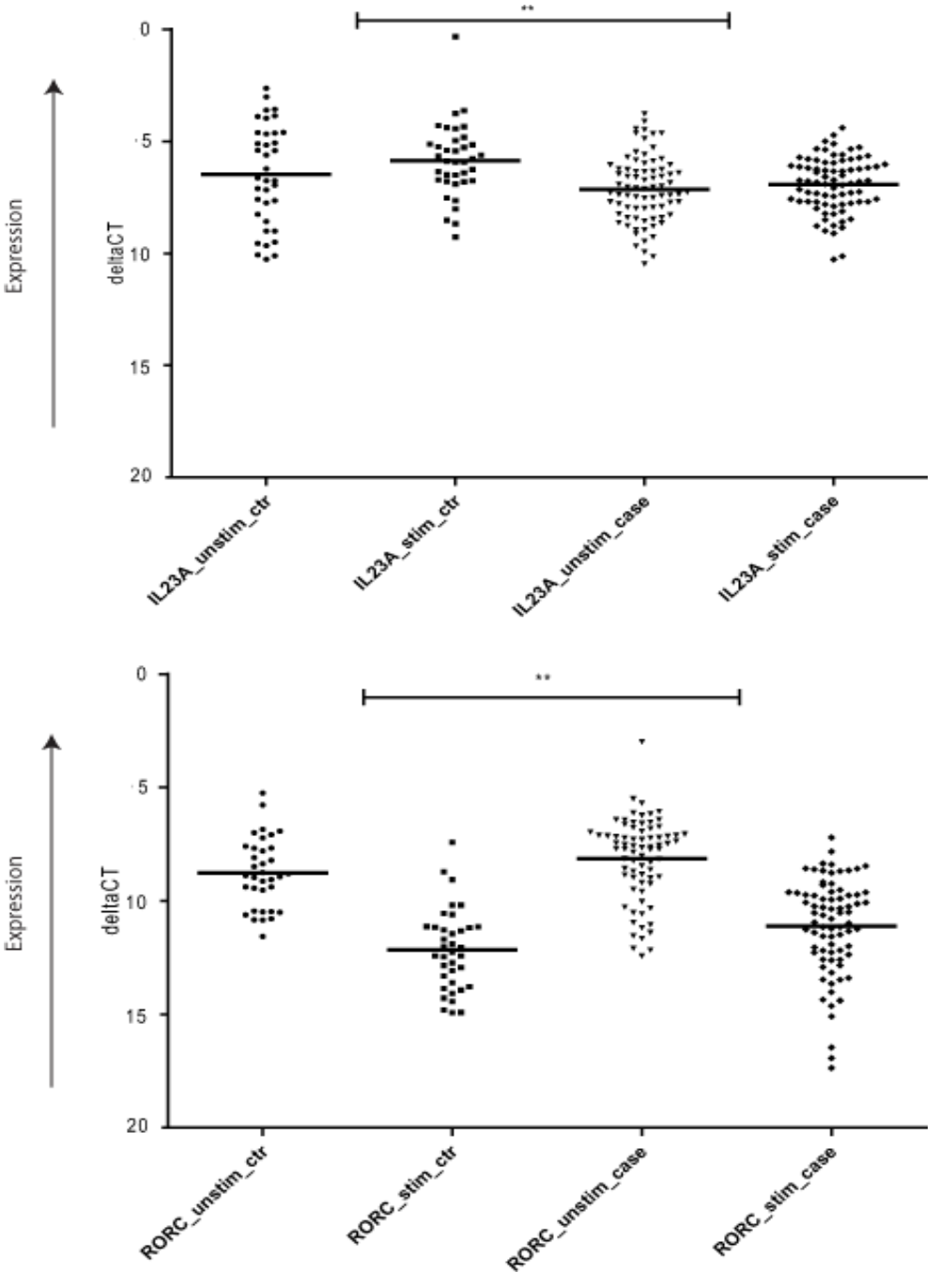
In the eQTL cohort all 22 genes involved in the Th17/IL23 pathway were assessed for SNPs with *cis*- and/or *trans*-eQTL effects on these genes according to the method developed by Fehrmann*et al*. [6] In brief, for all 22 genes expression probes were identified. For the *cis*-eQTL analysis all genotyped and imputed SNPs within a 250Kb window of the gene were included, in the case of associated genes all IBD risk SNP reside within this 250Kb window. For the *trans*-eQTL analysis all SNPs residing more than 5Mb away from the gene were included. Non-parametric spearman rank testing is used to test for eQTL effects. A correction is applied for the first 50 principal components, these are components that explain the largest part of variation in the data (such as technical variation and batch-effects, physiological status and environmental factors), and overshadow true biological differences. When eQTL SNPs were identified for a Th17 gene, the p-value for the association to IBD was checked for that SNP in the 1000genomes imputed CD and UC meta-analysis data using the on-line database, Ricopili (www.broadinstitute.org/mpg/ricopili). When the eQTL SNP was not the SNP with the strongest association to IBD, the LD between those two SNPs was assessed using SNAP pair wise LD software. [7]

Ethical considerations

The study was approved by the institutional review board of the hospital and written informed consent was obtained from all subjects.

Supplementary references

1. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. Nature Protocols. 2010;5:1564-1573.
2. Jostins L, Ripke S, Weersma RK, *et al*. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491:119-124.
3. Romanos J, van Diemen CC, Nolte IM, *et al.*Analysis of HLA and non-HLA alleles can identify individuals at high risk for celiac disease. Gastroenterology. 2009;137:834-840.
4. Weersma RK, Stokkers PC, van Bodegraven AA, *et al*. Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. Gut. 2009;58:388-395.
5. Rozen S, Skaletsky HJ. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000;132:365-386. Available at: http://frodo.wi.mit.edu.
6. Fehrmann RS, Jansen RC, Veldink JH, *et al.*Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. PLoS Genet. 2011;7:e1002197.
7. Johnson AD, Handsaker RE, Pulit S, *et al*. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24:2938-2939.



Supplementary figure 1. Gene expression results for A) IL23A and B) RORC for cases and controls in response to stimulation. **indicates significant difference ($p<0.01$) in expression (presented as ΔCT after normalization) after stimulation between cases and controls in linear regression analyses.

Supplementary table 1. Phenotypes of the genetic risk model cohort and the IBD-PBMC cohort.

	Genetic risk model cohort			IBD-PBMC cohort		
General characteristics	Crohn's disease	ulcerative colitis	controls	Crohn's disease	ulcerative colitis	controls
Number	423	330	1045	39	40	39
Average age at inclusion	46	49	unknown	55	51	34
Number of males (percentage)	161(38%)	181(54%)	561(53%)	17 (43%)	18 (45%)	19 (47%)
Age of onset						
Average age of onset	31	34		33	36	
<17 (A1)	35 (8%)	23 (7%)		2 (5%)	2 (5%)	
>17 and <40 (A2)	277 (65%)	194 (58%)		22 (56%)	26 (65%)	
>40 (A3)	89 (21%)	89 (26%)		11 (28%)	12 (30%)	
unknown	22 (5%)	24 (7%)		4 (10%)	0 (0%)	
Disease extent						
Crohn's disease						
Ileum (L1)	117 (27%)			10 (25%)		
Colon (L2)	89 (21%)			7 (17%)		
Ileum and colon (L3)	212 (50%)			22 (56%)		
Upper Gastro Intestinal (L4)	38 (8,9%)			3 (8%)		
Unknown	4 (1%)			0 (0%)		
Ulcerative Colitis						
Proctitis (E1)		41 (12%)			5 (13%)	
Leftsided (E2)		94 (28%)			14 (35%)	
Extended (E3)		150 (45%)			18 (45%)	
Unknown		45 (13%)			3 (8%)	
Disease behavior for Crohn's disease						
Non stenosing non penetrating (B1)	166 (39%)			9 (23%)		
Stenosing (B2)	117 (28%)			14 (35%)		
Penetrating (B3)	161 (38%)			20 (51%)		
Unknown	7 (2%)			0 (0%)		

Supplementary table 2. SNPs included in the genetic risk model.

SNP	Chr.	Gene	Risk allele	P-value	Beta-estimate	Location
rs11209026	1	*IL23R*	G	$2.8*10^{-06}$	0.80	Coding:Arg381Gln in *IL23R*
rs4845604	1	*RORC*	G	0.02	0.23	Intronic: *RORC*
rs7657746	4	*IL21*	A	0.08	0.14	Intronic: *KIAA1109*
rs6871626	5	*IL12B*	A	0.02	0.16	Intergenic: *LOC285626 -*- LOC285627*
rs1819333	6	*CCR6*	C#	0.12	0.11	Intergenic*RNASET2 -*- FGFR1OP*
rs10758669	9	*JAK2*	C	$1.8*10^{-04}$	0.26	Intergenic*RCL1 -*- JAK2*
rs7134599	12	*IFNG*	A	0.08	0.12	Intergenic*IFNy-AS1 -*- IFNy*
rs17293632	15	*SMAD3*	T	0.36	0.07	Intronic:*SMAD3*
rs12942547	17	*STAT3*	G#	0.32	0.07	Intronic:*STAT3*
rs11879191	19	*TYK2*	G	0.004	0.26	Intronic: *CDC37*

SNPs included in the risk model. Chr: chromosome. Gene: Th17 gene as candidate gene from Jostins *et al*. P-value: p-value of logistic regression analysis in the 'genetic risk model cohort'. Beta-estimate: beta-estimates of logistic regression analysis. Risk allele: risk allele of logistic regression analysis. Location: category of SNP, -*- marks the location of the SNP with respect to the environment. #risk allele is opposite to the risk allele reported by Jostins *et al*.

Supplementary table 3. Genotype frequencies per SNP.

		Genetic risk model cohort			IBD-PBMC cohort			
Gene	SNP	Wild type	Heterozygous	Mutant	Wild type	Heterozygous	Mutant	MAF 1000 genomes
IL23R	rs11209026	0.02	0.10	0.88	0.01	0.09	0.90	0.033
RORC	rs4845604	0.02	0.28	0.70	0.01	0.31	0.68	0.153
IL21	rs7657746	0.06	0.37	0.57	0.08	0.27	0.65	0.225
IL12B	rs6871626	0.41	0.46	0.13	0.43	0.46	0.11	0.383
CCR6	rs1819333	0.19	0.50	0.31	0.20	0.48	0.32	0.483
JAK2	rs10758669	0.41	0.45	0.14	0.47	0.43	0.10	0.309
IFNG	rs7134599	0.35	0.48	0.17	0.40	0.44	0.16	0.219
SMAD3	rs17293632	0.57	0.37	0.06	0.59	0.38	0.03	0.108
STAT3	rs12942547	0.19	0.50	0.31	0.15	0.51	0.34	0.390
TYK2	rs11879191	0.03	0.29	0.68	0.03	0.30	0.67	0.208

Gene: Th17 gene as candidate gene from Jostins et al. Wildtype: homozygous for protective allele. Heterozygous: one risk allele, one protective allele. Mutant: homozygous for risk allele. MAF 1000 genomes: minor allele frequency of SNP in 1000 genomes database (www.1000genomes.org).

Supplementary table 4. Effect of CD and UC status on mRNA-expression in unstimulated PBMCs and on stimulated PBMCs.

	Unstimulated		Stimulated#	
Gene	CD	UC	CD	UC
CCR6	NS	NS	NS	NS
IL23A	lower (p=0.062)	NS	smaller (p=0.001)	smaller (p=0.0441)
IL12RB1	NS	NS	NS	NS
STAT3	lower (p=0.050)	NS	NS	NS
IL17F	NS	NS	NS	NS
IL17A	NS	NS	NS	NS
IL6	lower (p=0.002)	lower (p=0.020)	NS	NS
RORC	higher (p=0.038)	NS	larger (p=0.005)	larger (p=0.019)
RORA	NS	NS	NS	NS

Gene: Th17 gene. For the unstimulated PBMCs the correlation between disease status and differential mRNA expression is reported, reported P-values from linear regression. 'Lower' and 'higher' indicate respectively lower and higher mRNA expression in cases compared to controls. For stimulated PBMCs the correlation between disease presence and differential mRNA expression is reported. The linear regression model was adjusted for baseline mRNA expression. 'larger' implies that cases had a larger difference in expression in response to stimulation than healthy controls. 'smaller' implies that cases had a smaller difference in expression in response to stimulation than healthy controls (supplementary figure 1). P-values are not corrected for multiple testing and only p-values below 0.10 are reported. No significant correlation between genetic risk load and baseline mRNA expression or response to stimulation was found. NS= not significant.

Supplementary table 5. eQTL analysis of all Th17/IL23 associated genes in 1240 whole blood samples of healthy controls.

Gene	eSNP	eQTL p-value	Direction	Ricopili p-value CD	Ricopili p-value UC	gSNP	r^2;D'	Likelihood that eSNP explains IBD association	Comments
IBD associated genes									
IL23R	-	-	-	-	-	rs11209026	-	-	-
RORC	rs11801866	$5.7*10^{-5}$	Down	0.01	0.005	rs4845604	0.7; 1	uncertain	Moderate ass. of eSNP
IL21	-	-	-	-	-	rs17388568	-	-	-
IL12B	-	-	-	-	-	rs6556412	-	-	-
CCR6	rs1358882	$4.9*10^{-5}$	Up	$5.3*10^{-8}$	0.005	rs415890	0.5; 0.9	uncertain	*RNASET2* much stronger eQTL
JAK2	-	-	-	-	-	rs10758669	-	-	-
IFNG	-	-	-	-	-	rs7134599	-	-	-
SMAD3	-	-	-	-	-	rs17293632	-	-	-
STAT3	rs744166	$1.4*10^{-7}$	Down	$3.1*10^{-8}$	$9.9*10^{-6}$	rs12942547	1	certain	-
STAT3	rs744166	$1.4*10^{-7}$	Down	$3.1*10^{-8}$	$9.9*10^{-6}$	rs11871801	0.2; 0.6	uncertain	Co-effect of other *STAT3* risk SNP
TYK2	rs281423	$7.7*10^{-5}$	Down	$3.8*10^{-6}$	0.07	rs12720356	0.02; 1	uncertain	High MAF eSNP, low MAF gSNP
IL17REL	-	-	-	-	-	rs5771069	-	-	-
Non-IBD associated genes									
TGFB	-	-	-	-	-	N/A	-	-	-
IL6	rs11772019	$1.6*10^{-4}$	-	N/A#	N/A#	N/A	-	unknown	No proxy of eSNP in Ricopili
RORA	-	-	-	-	-	N/A	-	-	-
IL17A	-	-	-	-	-	N/A	-	-	-
IL17F	-	-	-	-	-	N/A	-	-	-
IL22	-	-	-	-	-	N/A	-	-	-
IL12RB1	rs438421	$7.5*10^{-8}$	Up	$2.0*10^{-5}$	0.5	N/A	-	uncertain	Potential novel CD gene?
IL23A	-	-	-	-	-	N/A	-	-	-
IL6R	rs4845623	$1.5*10^{-6}$	-	0.05	0.2	N/A	-	not	No IBD association
IL17R	rs971768	$6.6*10^{-11}$	-	0.6	0.6	N/A	-	not	No IBD association
CCL20	-	-	-	-	-	N/A	-	-	-

eSNP: the eQTL SNP with the most significant association with IBD in the Ricopili database. eQTL p-value: p-value derived from non-parametric correlation of the eQTL effect of the eSNP. Direction: directional effect of risk allele of eSNP on mRNA-expression level. Ricopili p-value CD and UC: the p-value of the eSNP in the Ricopili database for respectively Crohn's Disease and ulcerative colitis. gSNP: the associated GWAS hit for the indicated locus. r^2;D': LD measures between gSNP and eSNP as reported by SNAP software. # eSNP not present in Ricopili database.

CHAPTER 7

Conclusions and future perspectives

Inflammatory bowel diseases (IBD) are comprised of mostly two forms; Crohn's disease (CD) and Ulcerative colitis (UC). IBD is a complex genetic disease characterized by a chronic relapsing inflammatory response of the gut immune system against commensal microbes in a genetically susceptible host. Based on twin studies, the heritability of IBD is estimated to be approximately 50%; the most recent twin study, performed in the Swedish twin registry, reports that due to a short inclusion time and short follow-up, twins might still go on to develop either form of IBD later in life, leading to an underestimation of the heritability [3]. The International IBD Genetics Consortium (IIBDGC) has performed the largest genotyping meta-analysis thus far, including more than 75,000 cases and controls [36]. They used a custom-made genotyping array, Immunochip, covering approximately 180 immune-related risk loci with approximately 180,000 single nucleotide polymorphisms (SNPs) and additional content focusing on deep replication of genome-wide association study (GWAS) results. Using Immunochip, 163 independent risk SNPs for IBD were identified that explain 7 and 13% of the disease variance for UC and CD, respectively. Assuming the widely used heritability estimate of approximately 50% is correct, only 15-26% of the total heritability can be explained by these common variants.

Numerous studies have been performed to shed light on the question of the missing heritability, as described in Chapter 2 of this thesis. It is defined as the difference between the heritability estimate and the heritability that can be explained by additive models applied to the number of risk SNPs. There are three potential sources for this difference:

1. The heritability of IBD is overestimated and thus a higher percentage of the heritability can actually be explained by already-identified loci.

2. New associations need to be identified because GWAS studies are still underpowered (e.g. rare variants with large effect size are not captured by GWAS, or common variants with low effect size are not reaching the genome-wide significance threshold).

3. The proportion of heritability explained by the identified SNPs is larger but hidden by, for example, epigenetic effects or gene-environment interactions not yet accounted for.

The overestimation of heritability

One of the potential explanations of the hidden heritability problem is that the heritability estimates may be exaggerated. Several methods to calculate heritability are available [1]. Each of these methods introduces its own form of bias that, combined with different sampling strategies, often produces more questions than answers. A common bias is that the strategies to calculate heritability do not account for gene-gene interactions and gene-environment interactions. It is stated that this omission leads to an overestimation of heritability, and creates so-called "phantom-heritability" [2]. However this statement is based on the assumption that gene-gene interactions and gene-environment interactions always strengthen each other's effect, even though this has yet to be proven. Several studies have estimated the heritability of IBD and the scientific community seems to have reached consensus in using a heritability estimate of 50%. The most recent twin study from the Swedish twin registry reports that the heritability estimates have to be decreased [3]. Concordance rates in monozygotic (MZ) CD twins were lowered from 58% to 38%, and in dizygotic (DZ) twins the rate was lowered from 4% to 2%. For UC, the concordance rate dropped from 19% to 15% in MZ UC twins, but rose from 0% to 8% in DZ twins. All these percentages are lower than those reported by the Danish twin registry in 2000 and 2005>: 58.3% for MZ CD twins and 0% for DZ and 18.2% for MZ UC twins and 4.5% for DZ [4,5]. Twin studies may overestimate heritability because environmental factors are relatively similar for twins, so it is difficult to make distinctions and gene-environment interactions are neglected. Moreover, it is assumed that MZ and DZ twins share the same environment, but MZ twins usually tend to show more alike behavior than DZ twins. The heritability estimate debate is likely to continue in the future and may never fully be resolved. It is likely worthwhile to explore other sources of the hidden heritability question rather than focusing only on the over-estimation of the total heritability. In the next sections I will focus on some of these sources.

Approaches for identifying new risk variants

The recent results of the meta-analysis of the IIBDGC, which identified many new risk SNPs, are encouraging [36]. They increased the number of risk SNPs by more than 50% using the largest cohort in the history of IBD and by establishing excellent worldwide collaboration. Moreover, Stahl [6] and colleagues suggest that in the complex genetic diseases, such as rheumatoid arthritis, celiac disease, cardiovascular disease and type 2 diabetes, many hundreds of additional common SNP associations remain to be identified. Their results further suggest that common causal variants of weak effect underlie the vast majority of these genetic contributions. It is likely that these results can be extrapolated to all complex genetic diseases and GWAS will thus remain fruitful. Log-linear models have been used to estimate the proportion of variation in disease liability that is captured in GWAS by considering all SNPs simultaneously [7]. For CD it is estimated at 22%, which is significantly higher than thus far explained by GWAS results. Unfortunately, neither of these methods account for epistatic effects and gene-environment interactions, and thus the effect of the single SNPs may be underestimated because such interactions may strengthen the effect on the pathophysiology, leading to an overestimate of the number of SNPs contributing to disease. In short, there are probably more risk SNPs to be discovered, but the number should be estimated with caution.

Selection of SNPs for replication

GWAS have yielded insight into the associated areas on the genome and the underlying pathways in disease pathogenesis, but, as suggested previously, we have probably not uncovered all the associated SNPs, making further exploration worthwhile [6]. Due to the large number of tests performed in GWAS, the significance threshold is stringent. This stringency leads to variants with a so-called type 1 error in the “sub-top” non-genome wide associated loci. These are SNPs that did not reach the threshold for statistical significance due to a lack of power, but that are truly associated

to disease risk. Power can be gained by increasing the sample size or by lowering the number of tests performed and thus the need to correct for multiple testing. In Chapter 3 of this thesis, we demonstrate that new hits can be identified by selecting for replication those SNPs in the sub-top non-genome-wide associated SNPs of GWAS results that influence gene-expression levels (expression quantitative trait loci (eQTLs)). By using this prioritization method, we identified ten potential risk loci for CD. Moreover, we have shown that this is more than would be expected by chance. In the replication-phase and meta-analysis, we identified *UBE2L3* and *BCL3* as new risk loci for CD. The association of *UBE2L3* was later confirmed in a large cohort meta-analysis [8]. Currently, even larger cohorts for the identification of eQTLs are available providing more potential SNPs for follow-up.

Expanding cohorts by including non-IBD cases

Given the cohort size of the latest IIBDGC's effort to identify new risk loci, one can envision that all the Caucasian IBD cohorts have been exhausted and that a gain in power by increasing the size of this cohort is virtually impossible. We need to find new strategies to further uncover this part of the hidden heritability problem. GWAS have shown that multiple immune-related diseases show overlap in their genetic background. For instance, type 1 diabetes and celiac disease show major overlap in association signals, but celiac disease and IBD also show a large overlap. This knowledge was the basis for the construction of the Immunochip, with the cooperation of genetic consortia for various immune-related diseases, like celiac disease, IBD and rheumatoid arthritis. Indeed, many of the 163 IBD loci discovered by Immunochip are also implicated in other immune-related disorders. This was seen most prominently for ankylosing spondylitis and psoriasis. One current hypothesis is that there are two different types of associated loci: 1) common loci, implicated in all immune-mediated diseases, and 2) disease-specific loci, only associated to one disease. By grouping different diseases together, the statistical power to detect the first type of loci increases, but the power

for detecting the second type of loci decreases. In Chapter 3, we report a test for the association of multiple known variants in the *PTPN22* gene to both forms of IBD. *PTPN22* is known to be associated to many immune-related diseases. The *PTPN22* 263Q loss-of-function variant showed evidence of association with UC but not with CD. In contrast, we found that the *PTPN22* 620W gain-of-function variant was associated with reduced CD risk, but that was not so with UC. Studies have successfully gained power for GWAS analysis by performing a cross-disease meta-analysis by grouping traits known to share part of their genetic background. For instance, four new shared loci for CD and celiac disease and for rheumatoid arthritis and celiac disease have been identified [9, 10]. Several efforts are currently underway that include all the Immunochip results for many immune-mediated diseases.

Expanding cohorts by including non-Caucasian cases

Another approach to increasing cohort size, and thereby increasing power, is to include other non-Caucasian populations. This is a statistical challenge given population stratification: differing genetic drifts caused by different environmental factors and different epidemiology can lead to altered linkage disequilibrium (LD) and genomic structure. Genotyping platforms have thus far only been constructed for Caucasian populations. However, because the SNPs on the arrays are selected based on Caucasian LD block structure, they are not ideal for non-Caucasian cohorts, given the differences in minor allele frequencies and LD blocks between ethnicities. Despite this bias, GWAS platforms have been used to study other ethnicities, for example the study of UC in the Japanese population in 2009 [11]. In addition to the expected strong association of the HLA locus, this study identified three more risk loci using approximately 1,300 UC and 3,000 control samples. While they attempted to replicate the Caucasian UC risk loci, the researchers could only replicate association of a few loci. The HLA locus was replicated in the Japanese cohort, although SNPs other than the Caucasian SNPs in this locus were

the most significantly associated. In addition, 1q36 and *JAK2* could be replicated. These results indicate that the top-hit SNP in one population is not necessarily the same in another population, although the same locus might well be involved. This study could not replicate other associations that had been identified in Caucasians and found that some SNPs in the *IL10* and *IL23R* loci from the GWAS platform were monomorphic in the Japanese population, and thus non-informative.

A current standard in GWAS is imputation: increasing the density of analyzed SNPs on the genotyping platform by predicting adjacent SNPs based on knowledge of sequence data and underlying LD structures. Fortunately, an increasing number of non-Caucasian populations have been sequenced, enabling imputation of more population-specific SNPs and thus improving GWAS analysis in non-Caucasian cohorts. Population stratification should be accounted for by thoroughly checking and searching for matched case-control cohorts to include in the GWAS analysis of the different populations, followed by a meta-analysis.

Structural variations and missing heritability

Structural variants in the genome, such as copy number variants (CNVs), could contribute to the missing heritability question for complex genetic diseases. The biggest effort to assess these types of variation in the genome was performed by the WTCCC [12]. They designed an array to measure the majority of CNVs from an inventory of CNV compiled from an extensive discovery cohort, and they typed 3,000 controls and 2,000 cases of eight complex genetic diseases including CD. They confirmed the association of the CNV in *IRGM* and identified a new CNV in the HLA region [12]. A limitation named by the authors was a lack of power to detect the associations for the CNVs, because they only tested ~50% of all existing CNVs with more than 500 copies, so half of the large CNVs were neglected and all smaller CNVs were neglected. Moreover, false positive hits were due to differences in CNVs detected in the DNA derived from cell

lines and original blood samples. Association testing resulted in the confirmation of loci already associated to disease risk identified by conventional GWAS. Thus, their findings suggest that the majority of CNVs currently captured by CNV arrays will be tagged by common SNPs used in GWAS. The lack of associations for CNVs to complex genetic diseases might still reside in an inadequacy of the detection of these variants or they may simply not contribute. Large-scale, whole-genome sequencing might elucidate more associations for CNVs to complex genetic diseases since only then will we be able to uncover more and smaller CNVs.

In conclusion, for GWAS to remain fruitful, large consortia are necessary to generate sufficiently large cohorts o provide the power needed to detect more variants. The IIBDGC is currently setting up such a major cross-ethnicity project to identify new shared-risk variants.

Examining the known associations

The third source of the hidden heritability may be that more of the heritability can be explained by already known associations. Additive models are used to calculate the explained heritability by associated variants. These models add up odds ratios of associated variants while ignoring that these are not the culprits and are merely tagging SNPs, thus the effect size of the causal variant may be underestimated. Moreover, this method ignores other effects like epigenetics, gene-environment and gene-gene interactions.

Identifying causal variants for IBD

Including non-Caucasian cohorts has greater advantages than just the gain in power due to increased cohort size. The genetic drift causing differences in LD blocks discussed earlier can be used to narrow down the region around the associated SNP. It is likely that different ethnicities have different genetic backgrounds but the same disease causing genes for the same disease, and this information

can be used to narrow down the region around the culprit gene. For example, the tight LD block around the *IL2/IL21* genes found in Caucasians is split into two parts in the Han Chinese population [13]. Because the LD is so tight in Caucasians, it was difficult to disentangle the signal and identify the culprit gene. Both LD blocks have been investigated separately in the Han Chinese, and both appeared to be associated to UC, indicating that both genes play a role in the pathogenesis of this disease [13]. This principle is the basis for a current project in trans-ethnicity fine-mapping, performed by the IIBDGC and involving five non-Caucasian ethnicities: Japanese, Iranian, Chinese, Korean and Indian. Trans-ethnic fine-mapping has already been applied successfully to metabo-chip data [14], a chip comparable to the Immunochip, and specially designed for metabolic diseases and traits, like cardiovascular disease, type 2 diabetes and blood lipid levels. In several loci, two or more independent signals are associated to blood lipid levels using metabo-chip data, thereby explaining a 1.3- to 1.8- fold increase in the phenotypic variance explained. Moreover, the signals of several variants associated to blood lipid levels could be narrowed down using this method. Hopefully these methods will also be fruitful in IBD and help explain more of the hidden heritability.

Low frequency and rare variants

GWAS target SNPs with minor allele frequencies (MAF) of >5%. It is hypothesized that (coding) low frequency and rare variants, defined as MAF< 5% and <0.5%, respectively, contain part of the hidden heritability because they are not sufficiently frequent and penetrant enough for GWAS to capture [15]. In an attempt to identify rare coding variants, 56 genes in CD-associated loci were sequenced. This identified additional risk and protective variants in *NOD2* and *IL23R,* and a splice site variant in *CARD9,* as well as coding variants in five other genes implicated in CD [16]. A similar effort in 55 UC-associated genes identified three associated rare risk variants [17]. Both studies started with a discovery cohort of 200-350 cases and controls. Given that the MAF of rare variants is defined as <0.5%, statistically there

should only be one case with such a mutation in their discovery cohort. To overcome this problem, pooled exome sequencing and simultaneous genotyping of 25 GWAS risk genes for six auto-immune diseases was also performed in a much larger cohort of 24,000 cases and 41,000 controls. Even then, only limited evidence for a role of these rare variants in disease pathogenesis was found [18].

Given the fact that the low frequency variants are not detected by conventional GWAS, true associated areas can remain undiscovered. It might therefore still be worthwhile to select genes or regulatory regions (hypothesis-driven selection) for targeted sequencing and assess them for associated low frequency and rare variants. However, selecting such areas remains troublesome and large-scale sequencing studies are expensive due to the low *a priori* chance of discovery with such low frequencies. An alternative approach is to use consanguineous and other families with an extreme CD phenotype to discover private variants, such as in a study into the *IL10R* locus [19].

Epigenetics

Another potential source of the hidden heritability may reside in altered cellular function caused by effects other than the 2D DNA structure. These mitotically heritable changes in gene function that are not explained by the DNA sequence are defined as epigenetic effects. It is known that some epigenetic effects are stable for several generations, but they can also be influenced by environmental influences, and they are tissue-specific. Dietary intake during pregnancy was shown to affect the epigenome in offspring in the Second World War and remained stable for two generations [20-22]. This partially reversible process of switching on or off part of the genome is primarily controlled by DNA methylation, histone modification, RNA interference and chromatin structure. DNA is methylated by DNMT enzymes and the methyltransferase gene *DNMT3a* has also been identified as a CD susceptibility gene, hinting at a role for epigenetics in the disease pathogenesis [8]. A methylation chip, on which the methylation status of ~28,000 CpG sites can be measured, has been applied to whole

blood samples of CD cases and controls [23]. *S*ignificantly different levels of methylation of 50 sites between cases and controls were identified. However, the relation between histone modification and IBD has been less thoroughly studied, with only studies on murine and rat models and biopsies of CD patients are available. These data suggest a role for epigenetics in the disease pathogenesis of IBD. However, one of the major challenges remains to determine whether this altered methylation status is a cause or a consequence of disease. From epidemiological studies, it has become apparent that children of mothers with CD have a higher chance of getting the disease than children of fathers with CD, suggesting that epigenetic effects inherited from the mother alter the genomic function. We tested all overlapping UC and CD loci, the *NOD2* variants and a gene implicated in UC pathogenesis that is known to be imprinted for such a parent-of-origin effect, but found limited evidence for such effects in these genes. A more thorough study using more trios and testing more SNPs should be performed to answer this question.

Gene-environment interactions

Little is known about the environmental factors contributing to the pathogenesis of IBD and providing any proof of association has been a major challenge. Smoking is the first and only consistently associated risk factor for IBD. Current and ex-smokers have a higher risk of developing CD [24-26], although current smokers seem protected against UC [24, 25, 27, 28]. Some enteric infections, such as salmonella and campylobacter exposure [29] have been implicated and modest evidence has recently been found that childhood exposure to antibiotics is associated with development of IBD at a later age [30]. These latter findings may indicate that changes in the gut's microbial environment, the microbiome, influence the disease development and course. An imbalance in the microbiome of IBD patients has been observed in many studies [31-34]: decreased biodiversity, a higher proportion of *Gammaproteobacteria*, and a decrease of *Firmicutes* have been consistently observed.

The current hypothesis is that environmental factors alter the epigenome, thereby altering cell function and contributing to the pathogenesis. Detailed information on dietary intake, microbiome, and current or former use of antibiotics needs to be collectively analyzed to unravel interactions and elucidate part of the pathogenesis of IBD. Gene-environment interactions can also be studied in healthy individuals, making it easier to collect the large sample sizes needed for sufficient power to perform the necessary analyses, and without confounding factors such as disease activity or drug use. A recently initiated population-based cohort containing 175,000 individuals from the three northern provinces of the Netherlands, the Lifelines study, may shed light on this matter [35]. This cohort contains detailed information on health, dietary and lifestyle habits, and medical history collected using questionnaires, blood samples and DNA from (preferably) three generations. Recently, the investigators started a sub-study, LifeLines DEEP, in which they also collect exhaled breath and stool samples from approximately 1,500 individuals. The data collected from these samples may prove to be a treasure trove of new information, but it will be challenging to interpret, both computationally and methodologically.

Translation into function

GWAS have uncovered many associated regions for IBD leading to greater insight into the disease pathogenesis. For instance, the role of autophagy in inflammation in IBD patients was discovered by GWAS. Recently, the overlap between susceptibility loci for mycobacterial infection and IBD was also discovered [36]. Gene co-expression network analysis emphasizes this relationship, with pathways shared between host responses to mycobacteria and those predisposing to IBD, suggesting that host-microbe encounters have shaped genetic predisposition to IBD.

One of the topics in IBD research is identifying the culprit gene in a locus. For many associated regions this still is a challenge, as loci may contain none, one or many genes. Common denominators

for genes in associated regions can be used to prioritize genes in loci. Several tools have been developed for prioritization based on gene characteristics (Gene-ontology; GO-terms), on co-occurrence in the literature (Gene Relationship Across Implicated Loci; GRAIL), on co-expression of genes, and on protein-protein interaction data. The T helper 17/interleukin 23 (Th17/IL23) pathway has been implicated in the pathogenesis of IBD using these methods. In Chapter 6, we sought to unravel the correlation between genetic risk load, i.e. the number of genetic risk loci, and gene-expression levels for genes in the Th17/IL23 pathway. Surprisingly, we found only limited evidence for such a correlation. Furthermore, we found that only two out of 23 genes in the pathway eQTL-SNPs also showed association to IBD. In our study, we used peripheral blood mononuclear cells, but ideally Th17 cells isolated from both an inflamed part of the gut and from a non-inflamed part of the gut from the same individual should be tested for differential expression to test for disease status effects. Further, the gene expression of Th17 cells from non-inflamed gut biopsies in individuals with a low genetic risk should be compared to Th17 cells isolated from non-inflamed gut biopsies in high genetic risk individuals to test for the effect of genetics on expression levels.

Translation to the clinic

GWAS have been promoted as the bridge between the genetic association of IBD and the clinical impact. However, for both IBD and for most of the other complex genetic diseases, the clinical impact of GWAS results has been very limited thus far. It could be of major health care-, personal- and economic-importance to be able to specify treatment and to be able to predict the disease course and drug response for each individual patient, for example. This would enable clinicians to apply personalized medicine, i.e., for every IBD patient the disease can be treated effectively from the beginning instead of wasting time and money following the classical step-up strategy to attain optimal therapy for a patient. To date, a few successes have been achieved. For example, patients who are treated with azathioprine

or 6-mercaptopurine are increasingly screened for a variant in the *TPMT* gene, for which low expression is associated to bone marrow toxicity as the marrow-toxic metabolite 6MP is not inactivated. More studies are also being performed to detect genetic prediction factors for mesalasine-induced nephrotoxicity and complications of thiopurines and anti-TNF therapies. To gain further insight, incorporation of genetic information with environmental information (such as the state of the microbiome or smoking status) and disease status, treatment, and disease course is of utmost importance. Building biobanks like the String of Pearls research initiative (*Parel Snoer Instituut, PSI*) in the Netherlands addresses this need and should contribute greatly to elucidating clinically important genetic factors [37]. PSI consists of eight research lines (pearls), including IBD, for which detailed and standardized patient and disease information and blood, feces, biopsy, resection tissue, and DNA sample are collected and stored in a biobank for major, integrated,'omics' analysis.

Conclusion

Three potential sources of the hidden heritability of IBD remain to be explored. First, the heritability estimate may be overestimated. Second, there may be more risk SNPs to be identified. Third, more of the heritability might be explained by the established risk SNPs and variants in LD. Large cohorts, extensive information on the phenotype and accessibility to various tissue types of IBD patients and controls are needed to further elucidate the genetic background and enable personalized medicine. To collect and process all this information, collaborations between consortia both at a national and international level will be essential.

References

1. The heritability of human disease: estimation, uses and abuses. A Tenesa, CS Haley. Nature Reviews Genetics. 2013: 14, 139-149.
2. The mystery of missing heritability: Genetic interactions create phantom heritability. O Zuk, E Hechter, SR Sunyaev *et al.* Proc Natl Acad Sci USA. 2012: 109, 1193-8.
3. Genetics in twins with Crohn's disease: less pronounced than previously believed? J Halfvarson. Inflamm Bowel Dis. 2011: 17, 6-12.
4. Disease concordance, zygosity, and NOD2/CARD15 status: follow-up of a population-based cohort of Danish twins with inflammatory bowel disease. T Jess, L Riis, C Jespersgaard, *et al.* Am J Gastroenterol. 2005: 100, 2486-92.
5. Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study. M Orholm, V Binder, TI Sørensen, *et al.* Scand J Gastroenterol. 2000: 35, 1075-81.
6. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Eli A Stahl, Daniel Wegmann, Gosia Trynka, *et al.* Nat Genet. 2012: 44, 483–489.
7. Estimating missing heritability for disease from genome-wide association studies. SH Lee, NR Wray, ME Goddard, *et al.* American J HumGen. 2011: 88, 294–305.
8. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. A Franke, DP McGovern, JC Barrett, *et al.* Nat Genet. 2010: 42, 1118–1125.
9. A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. EA Festen, P Goyette, T Green, *et al.* PLoS Genet. 2011: 27, 7.
10. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. A Zhernakova, EA Stahl, G Trynka, *et al.* PLoS Genet. 2011: 7(2):e1002004.
11. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. K Asano, T Matsushita, J Umeno, *et al.* Nat Genet. 2009: 41, 1325-1329.
12. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. N Craddock, ME Hurles, N Cardin, *et al.* Nature. 2010: 464, 713-20.
13. Haplotype-based analysis of ulcerative colitis risk loci identifies both IL2 and IL21 as susceptibility genes in Han Chinese. J Shi, L Zhou, A Zhernakova, *et al.* Inflamm Bowel Dis. 2011: 17, 2472-9.
14. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. Y Wu, LL. Waite, AU. Jackson, *et al.* PLoS Genet. 2013: 9(3):e1003379.
15. Rare variants create synthetic genome-wide associations. SP Dickson, K Wang, I Krantz, *et al. PLoS Biol.* 2010: 8(1): e1000294.
16. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. MA Rivas, M Beaudoin, A Gardet, *et al.* Nat Genet. 2011: 43, 1066-73.
17. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. M Beaudoin, P Goyette, G Boucher, *et al.* PLoS Genet. 2013: 9(9):e1003723
18. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. KA Hunt, V Mistry, NA Bockett, *et al.* Nature. 2013: 13, 232-5.

19. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. EO Glocker, D Kotlarz, K Boztug, *et al.* N Engl J Med. 2009: 361, 2033-45.
20. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. RA Waterland, RL Jirtle. Mol Cell Biol. 2003: 23, 5293-300.
21. The epigenome: archive of the prenatal environment. BT Heijmans, EW Tobi, LH Lumey, *et al.* Epigenetics. 2009: 16, 526-31.
22. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. EW Tobi, LH Lumey, RP Talens, *et al.* Hum Mol Genet. 2009: 18, 4046-53.
23. Genome-wide methylation profiling in Crohn's disease identifies altered epigenetic regulation of key host defense mechanisms including the Th17 pathway. ER Nimmo, JG Prendergast, MC Aldhous, *et al.* Inflamm Bowel Dis. 2012: 18(5), 889-99.
24. A prospective study of cigarette smoking and the risk of inflammatory bowel disease in women. LM Higuchi, H Khalili, AT Chan *et al.* Am J Gastroenterol. 2012: 107, 1399-406.
25. Smoking in inflammatory bowel diseases: Good, bad or ugly? PL Lakatos, T Szamosi, L Lakatos. World J Gastroenterol. 2007: 13, 6134-6139.
26. Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. T Birrenbach, U Böcker. Inflamm Bowel Dis. 2004: 10, 848-59.
27. Tobacco and IBD: relevance in the understanding of disease mechanisms and clinical practice. J Cosnes. Best Pract Res Clin Gastroenterol. 2004: 18, 481-96.
28. What is the link between the use of tobacco and IBD? J Cosnes. Inflamm Bowel Dis. 2008: 14, Suppl 2:S14-5.
29. Increased short- and long-term risk of inflammatory bowel disease after salmonella or campylobacter gastroenteritis. KO Gradel, HL Nielsen, HC Schønheyder, *et al.* Gastroenterology. 2009: 137, 495-501.
30. Association between the use of antibiotics in the first year of life and pediatric inflammatory bowel disease. SY Shaw, JF Blanchard, CN Bernstein. Am J Gastroenterol. 2010: 105, 2687-92.
31. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. DN Frank, AL St Amand, RA Feldman, *et al.* Proc Natl Acad Sci USA 2007: 104, 13780-13785.
32. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. C Manichanh, L Rigottier-Gois, E Bonnaud, *et al.* Gut 2006: 55, 205-211.
33. Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. SJ Ott, M Musfeldt, DF Wenderoth, *et al.* Gut 2004: 53,685-693.
34. Commensal bacteria, traditional and opportunistic pathogens, dysbiosis and bacterial killing in inflammatory bowel diseases. CD Packey, RB Sartor. Curr Opin Infect Dis 2009: 22,292-301.
35. LifeLines. https://www.lifelines.nl
36. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. L Jostins, S Ripke, RK Weersma, *et al.* Nature. 2012: 491, 119-24.
37. String of Pearls initiative [*Parelsnoer*]. http://www.parelsnoer.org/ (in Dutch and English)

APPENDICES

Summary

Chronic inflammatory bowel diseases (IBD) are mostly comprised of Crohn's disease (CD) and Ulcerative colitis (UC). Their prevalence in Western countries is approximately 100-200 per 100,000. They are complex genetic diseases, meaning that both environmental and genetic factors contribute to disease risk and the course of the disease. They are characterized by a chronic and recurring inflammation of the gut, which is triggered by an aberrant immune response to commensal gastrointestinal microbes in a genetically susceptible host. This leads to symptoms such as abdominal pain, bloody stools and diarrhea. Epidemiologic studies have uncovered some environmental risk factors, including smoking and intestinal infections, but this search has been problematic. Most research has focused instead on unraveling the genetic background of IBD by scanning the genome in tens of thousands of cases and controls for genetic variants – single nucleotide polymorphisms (SNPs) – associated to disease risk. During the work described in this thesis, the International IBD Genetics Consortium (IIBDGC) performed the largest risk SNP meta-analysis so far. It included more than 75,000 cases and controls, expanding the total number of known risk SNPs to 163 for IBD, the largest number of known risk loci for any of the complex genetic diseases. These 163 loci together explain 7% and 13% of the disease variance for UC and CD, respectively. Given a heritability estimate of 50%, approximately 15-26% of the total heritability can now be explained.

In this thesis, we aimed to elaborate on the partially uncovered genetic background of IBD. We performed an extensive literature search for potential sources of hidden heritabilities by identifying new risk SNPs, and we took the first steps towards identifying other potential sources for this newly identified genetic background, moving beyond pinpointing new risk SNPs in the genome. Moreover, we investigated the consequences of such risk loci for cell function.

This thesis has three parts. In the first part, we aimed to identify new risk loci for IBD by prioritizing specific loci known to influence gene-expression levels. Genome-wide association studies (GWAS) use genotyping platforms that determine the genetic code of hundreds

of thousands of SNPs and test for genetic association to a trait or disease. This approach requires a stringent correction for multiple testing, and therefore a true association might be discarded because of a lack of statistical power. Replication studies of the top-1000 loci have been successful in determining new risk loci, but are costly and time-consuming. By testing only a subset of the loci, the number of tests for which correction is needed decreases and thus the power to detect risk loci increases. We chose to test only risk loci that influence gene-expression levels – expression quantitative trait loci (eQTLs) – since they are known to be more-often trait-associated. We applied this strategy successfully to the top-1000 SNPs of a publicly available GWAS and identified 10 eQTL SNPs. Of these, one was an established SNP in the *NOD2* locus, a well-known and established risk locus for IBD, and another two risk SNPs had a significant association to CD in our study cohort, later confirmed in a meta-analysis.

A second strategy for identifying new risk SNPs was based on the insight that many immune-mediated diseases share part of their genetic background. The *PTPN22* locus is an example of such a common locus. In our study we tested two coding SNPs for association to CD and UC, R263Q and R620W. The R263Q SNP was only associated to UC, whereas the R620W SNP was only significantly associated to CD.

In the second part of the thesis, we focused on epigenetic effects as another potential source of hidden heritability. These are effects on gene-expression levels by, for instance, controlling the tightness of winding of the DNA and thereby controlling the availability of the DNA for expression. It is known from epidemiological studies that children of mothers with CD have a higher risk of developing CD themselves than children of fathers with CD. Epigenetic effects might underlie this phenomenon of genomic imprinting. We aimed to identify risk loci subjected to such effects in the overlapping CD and UC loci, in *NOD2* and in a risk locus for UC already known to be imprinted. However, we found only weak evidence for such effects in these loci in IBD patients using a parent-of-origin test, thus further research with larger cohorts is necessary to test the remaining IBD risk loci.

In the third part of the thesis, we investigated the functional

consequences of risk SNPs for a pathway implicated in the pathogenesis of IBD: the Th17/IL23 pathway. The Th17/IL23 pathway acts in Th17 cells, which are thought to play a role in chronic inflammatory processes. Next to the genetic associations, functional studies had also highlighted the role of the Th17/IL23 pathway in IBD. Using expression data of peripheral blood mononuclear cells from 40 healthy individuals, 40 CD and 40 UC patients for whom genotyping data was available, we looked for a correlation between the number of risk loci and gene-expression in the Th17/IL23 pathway. In 1,240 control individuals with whole-genome genotype and mRNA-expression data, we also assessed the correlation between genetic risk load and differential mRNA-expression and looked for *cis*-eQTL SNPs for all the currently known Th17/IL23 genes. Surprisingly, we found little evidence for such genetic-gene-expression correlations, but we did find a disease status-gene-expression correlation.

Samenvatting

Chronische inflammatoire darm ziekten (IBD) bestaan voornamelijk uit de ziekte van Crohn (CD) en colitis ulcerosa (CU). De prevalentie in de westerse landen is ongeveer 100-200 per 100.000 individuen. Het zijn complexe genetische ziekten, wat inhoudt dat zowel de omgeving en genetische factoren een rol spelen. Ze worden gekarakteriseerd door een chronische, terugkerende ontsteking van de darm, die ontstaat door een afwijkende reactie van het immuunsysteem op commensale gastrointestinale microben in genetisch vatbare individuen. Dit leidt tot symptomen als buikpijn, bloederige ontlasting en diarree. Epidemiologische studies hebben een aantal omgevingsrisicofactoren ontdekt, zoals roken en intestinale infecties, maar de zoektocht is problematisch gebleken. Het onderzoeksveld rondom het ontrafelen van de genetische achtergrond van IBD is succesvol geweest door het scannen van genetische varianten –zogeheten enkelvoudige nucleotide polymorfismen (SNPs) - in het hele genoom van tienduizenden patiënten en gezonde controles en zodoende het ziekte risico te bepalen van elk van deze SNPs. Tijdens het werk wat verricht is en beschreven staat in dit proefschrift heeft het International IBD Genetics Consortium (IIBDGC) de grootste meta-analyse tot nu toe verricht. Zij includeerden meer dan 75.000 patiënten en controles, en vergrootten daarmee het aantal bekende IBD risico SNPs tot 163. Dit is het grootst aantal bekende risico SNPs voor alle complexe genetische ziekten. De 163 SNPs samen verklaren respectievelijk 7% en 13% van de ziekte variantie voor UC en CD. Bij een geschatte erfelijkheid van ongeveer 50% kan ongeveer 15-26% van de totale erfelijkheid verklaard worden.

In dit proefschrift hebben we getracht de deels ontrafelde genetische achtergrond van IBD verder uit te werken. Er is een uitgebreid literatuuronderzoek gedaan om potentiële bronnen van de verborgen erfelijkheid voor IBD te ontrafelen. Hiernaast hebben we nieuwe risico SNPs voor IBD gezocht, en we namen de eerste stappen in het exploreren van andere potentiële bronnen van deze verborgen genetische achtergrond, daarmee gaan we verder dan het aanwijzen van nieuwe risico SNPs in het genoom. Tenslotte onderzochten we de consequenties van risico SNPs voor cel functie.

Dit proefschrift is opgedeeld in drie delen. In het eerste deel hebben we nieuwe risico SNPs voor IBD willen identificeren door specifieke SNPs te prioriteren die gen-expressie beïnvloeden. Genoombrede associatie studies (GWAS) maken gebruik van genotyperingsplatformen die de genetische code van honderdduizenden SNPs tegelijk bepalen in honderden individuen en die daarna getest worden voor associatie met een bepaalde eigenschap of ziekte. Deze aanpak heeft een zeer strenge statistische correctie nodig door het grote aantal testen wat uitgevoerd wordt, waardoor werkelijk geassocieerde SNPs gemist kunnen worden door een gebrek aan power. Replicatiestudies van de, bijvoorbeeld, top-1000 niet-geassocieerde SNPs zijn succesvol gebleken in het identificeren van nieuwe risico SNPs, maar deze studies zijn duur en tijdrovend. Door alleen een deel van deze SNPs te testen neemt de correctie die nodig is voor het aantal testen af waardoor de power om SNPs te vinden toeneemt. Wij hebben ervoor gekozen alleen die SNPs te selecteren die ook gen-expressie beïnvloeden - expressie kwantitatieve karakter eigenschap SNP (eQTLs) - aangezien die vaker geassocieerd zijn met een bepaalde eigenschap. Wij hebben deze strategie succesvol toegepast op de top-1000 SNPs van een publiekelijk beschikbare GWAS en hebben hierin 10 eQTL SNPs geïdentificeerd. Van deze SNPs is één een bekende SNP in het *NOD2* gen, een bekend risico gen voor IBD, en twee andere SNPs waren significant geassocieerd met CD in ons cohort en werden later bevestigd in een meta-analyse.

Een tweede strategie voor het identificeren van nieuwe risico SNPs is gebaseerd op het inzicht dat vele immuun-gemedieerde ziekten een deel van hun erfelijke achtergrond delen. *PTPN22* is een voorbeeld van een algemeen gedeelde SNP. In onze studie hebben we twee coderende SNPs getest voor associatie met CD en UC, R263Q en R620W. De R263Q SNP was geassocieerd met UC risico, terwijl de R620W SNP geassocieerd was met CD risico.

In het tweede deel van dit proefschrift focussen we op epigenetische effecten als potentiële bron voor de nog niet ontrafelde erfelijkheid van IBD. Dit zijn effecten op gen-expressie door bijvoorbeeld het beïnvloeden van hoe sterk het DNA opgerold is en daarmee de

beschikbaarheid van dat DNA om tot expressie gebracht te worden. Het is bekend vanuit epidemiologische studies dat kinderen van moeders met CD een grotere kans hebben op het krijgen van CD dan kinderen van vaders met CD. Epigenetische effecten zouden verantwoordelijk kunnen zijn voor dit fenomeen, zogenoemd genomische imprinting. Wij wilden risico SNPs met zulke effecten vinden in de overlappende CD en UC SNPs, in het *NOD2 gen,* en in een risico gen voor UC waarvan al bekend is dat het beïnvloed wordt door imprinting. We gebruikten een parent-of-origin test om deze hypothese te toetsen, maar vonden slechts zwak bewijs voor zulke effecten in deze SNPs in IBD patiënten. Verder onderzoek is nodig waarbij grotere cohorten gebruikt worden en waarin meer SNPs getest kunnen worden.

In het derde deel van dit proefschrift hebben we de functionele gevolgen van risico SNPs getest in een pathway wat geassocieerd is in de pathogenese van IBD: het Th17/IL23 pathway. Het Th17/IL23 pathway werkt in Th17 cellen, dit zijn cellen waarvan gedacht wordt dat ze een rol spelen in chronische ontstekingsprocessen. Naast genetische associatie hebben functionele studies ook de rol van het Th17/IL23 pathway in IBD belicht. Door gebruik te maken van expressie data van perifere bloed mononucleaire cellen van 40 gezonde individuen, 40 CD en 40 UC patiënten waarvan we ook genotype data hadden, hebben we gezocht naar een correlatie tussen het aantal risico SNPs en de gen-expressie in het Th17/IL13 pathway. In 1.240 controle individuen met genotype en gen-expressie data van het gehele genoom hebben we de correlatie tussen genetische risico zwaarte en differentiële gen-expressie eveneens onderzocht. Daarnaast hebben we gezocht naar cis-eQTL SNPs voor alle tot nu toe bekende Th17/IL23 genen. Verrassend genoeg vonden we maar weinig bewijs voor het bestaan van deze correlaties, behalve een ziektestatus effect op gen-expressie.

Dankwoord

Zoals ik al in één van mijn stellingen heb benadrukt, is het niet mogelijk een PhD af te ronden zonder een enthousiast team dat achter je staat, je steunt, corrigeert en bovenal stimuleert. Gelukkig heb ik daar gebruik van mogen maken.

Ten eerste wil ik mijn promotorenProf C. Wijmenga en Prof. Dr. R. K. Weersma en mijn co-promotor Dr. C.C. van Diemen bedanken. Hoe jullie elkaar in mijn supervisie aanvullen is uniek en geweldig om gebruik van te mogen maken.

Beste Rinse, jij bent een begeleider, coachen steun en toeverlaat op zowel professioneel als persoonlijke vlak, zoals elke PhD-student die zou moeten hebben. Ik heb ontzettend veel waardering voor je. Met name omdat jij, ondanks alle ballen die je in de lucht moet houden, nog steeds een warm en begripvol mens bent. Ik heb ontzettend veel van je geleerd en hoop dat nog vele jaren te mogen en kunnen doen, als dokter, wetenschapper en mens. Op z'n Drents/Gronings: 't had minder kent.'

Lieve Cleo, jij hebt me met alle geduld van de wereld de vele laboratoriumtechnieken geleerd en je hebt daarbij meer dan eens hartelijk kunnen lachen om mijn stomme acties. Ik dank je ontzettend voor alle tijd, aandacht en steun.Hoe vaak heb ik niet bij je aan het bureau gestaan met: "Uhm Cleo… ik heb een vraagje… , of "Wat vind jij hiervan", of "Kun je even (liefst binnen nu en vijf minuten) mijn abstract nakijken?". Altijd heb jij tijd voor mij, hoe druk je het zelf ook hebt. Ook heb je me nog een aantal heel belangrijke andere zaken geleerd, zoals Baileys maken ;-). Jij bent voor mij een voorbeeld in het balanceren tussen carrière maken en een prive leven houden.

Beste Cisca, voor een groot deel van de tijd zou jij mijn promotor zijn met Cleo en Rinse mijn copromotoren. Het is hun verdienste dat ik jou nooit heel erg op de voorgrond heb gezien. Maar ik weet dat jij altijd een oogje in het zeil hield om te zien hoe wij in "het IBD groepje" het deden. Je hebt het me zeker niet altijd makkelijk gemaakt, maar

ook de duivel heeft een advocaat nodig ;-). Ik heb dan ook heel veel aan je te danken. Van jou heb ik met name geleerd om te denken als een wetenschapper en altijd overal vragen bij te stellen.

Mijn meest naaste collega's.... S. van Sommeren, Dr. M. Mitrovic en Dr. E.A.M. Festen: bedankt voor alle steun!

Lieve Suzanne, lieve evilpumpkin, lieve jul, vanaf minuut 1 ben je een moordwijf geweest. Bek als een scheermes en nooit te beroerd om de voorzetjes die ik geef, in te koppen. Gelukkig geef je mij ook genoeg ballen om in te koppen. Zolang jij "aroundscrewed" en "are justgonneplaywithit" vermaak ik me opperbest met je. Ik ben blij dat jij ook MDL-arts gaat worden en ik hoop dat je nog heel lang mijn collega blijft! Ook hoop ik dat we op de opleidingweer jut en jul mogen zijn. Beviel me wel!

Dear Mitja, hey part, you were supposed to come and help me hybe the slo samples on the Ichip for 3 months. Gladly it turned out to be 18 moths cause I've gained a friend during this period. I still see us in the lab in the beginning. Both no clue what to do but with "a little less conversation and a little more action" we did it! And when that wouldn't help there was always you and your "no worries". The MiFra pipet still needs some work but we will design it! Thanksfor the great time together!! Ajde.

Beste Noortje, ook aan jou heb ik heel veel te danken! Je hebt direct tegen Cisca gezegd dat ik een PhD wilde worden en hebt me vanaf het begin daarin gestimuleerd. Ik was nog wat naïef en wist nog niet hoe leuk wetenschappelijk onderzoek was, maar jij hebt ervoor gezorgd dat ik daar snel genoeg achter kwam! Ik heb super veel bewondering voor je werklust, je gedrevenheid en hoe slim je bent, maar bovenal bewonder ik je onzelfzuchtigheid. Bedankt voor alles! Ik hoop als dokter en wetenschapper nog veel van je te mogen leren!

I would like to thank the members of the thesis evaluation committee, prof. dr. K.N. Faber, prof. dr. H.M. Boezen, and prof. dr. rer. nat. A. Franke for carefully reading and evaluating my thesis.

Beste Ilja, dank je voor je steun die ik nodig had met het parent of origin project. Ik heb ontzettend veel van je geleerd. En als ik nog eens een nieuw vakantieland zoek, weet ik bij wie ik om tips moet vragen.

Beste Marijn, bedankt voor de samenwerking met ons UBE project het waren af en toe wat lage avonden, maar met jou erbij was het erg gezellig!

Beste Gerard, bedankt voor je input en adviezen bij mijn projecten.

Beste Willem Thijs, misschien ben je verbaasd hierin te staan, misschien ook niet. Dat ik je noem heeft een aantal redenen. Jij hebt me voorgesteld aan Rinse, mijn promotor en aan de MDL-artsen in Zwolle, mijn toekomstige opleidingsziekenhuis. Maar bovenal heb je me laten zien hoe mooi het vak van MDL-arts is. Je sleepte me mee naar regioavonden, bijscholingsbijeenkomsten en tijdens mijn coschap op de spoed riep je me naar de scopie-kamer om daar te helpen. Ik kwam je bedanken met een flesje wijn, maar liever wilde je een kopie van mijn proefschrift. Bij dezen!

Beste kamergenoten, Helga, met jou en René mocht ik naar New York en daarna naar de ASHG in Washington. Dit was fantastisch! Ik zie ons nog slepen met onze koffers langs die weg in ongeveer de meest gevaarlijke buurt van Washington. Annemieke, met jou heb ik de meeste tijd doorgebracht op één kamer. Wat was het fijn om zo nu en dan flink te zeiken over van alles en nog wat. En leuk dat ik je Piet mocht zijn ;-) Gerben, tank, in jouw dankwoord voelde ik een uitdaging. Ik ben dan ook benieuwd wie deze titanenstrijd gaat winnen. Peter, gelukkig heb ik nooit aan een pubquiz mee hoeven doen zonder jou. Dan waren we echt glansrijk laatste geworden. Anna D, Anna P, Ettje, Yustina, Yunia, Christine thanks for all the support, chats and good moments we shared in our room.

Dear celiac Group people, Agata, Alex, Asia, Barbara, Dasha, Gosia, Harm-Jan, Isis, Javier, Jihane, Jingyuan, Juha, Lude, Patrick, Roan, Rodrigo, Sasha, Sebo, Senepati and Vinod, Marcel, I owe a lot to all of

you. Thanks for the intellectual input in this thesis and thanks for all the great times we had at defense parties, the Irish pub or just any other day. Dear Agata, drama queen and Asia, evil creature, thanks for all the great laughs and help withall my stupid questions. Asia thanks again for all the great things you did to make my thesis look like this! You totally rock! Lude, Harm-Jan thanks for all the help with eQTLs, Jing thanks for always answering my questions about whatever topic. Jihane, Gosia thanks for all the help with computerstuff, plinkstuff and all the rest.

Dear Dineke, Celine, Olga, Ester, Ellen, Robert, Christine, Paul, Rian, Olga and Mats thanks for all the good laughter during lunches, defense presentations, and other occasions.

Beste Mathieu, Monique, Pieter, Ron, Soesma, Rutger, Ludolf, Jan, Jelkje, Michiel, Marcel, Mariska, Astrid, en Bahram, zonder jullie liep ik nu nog verdwaasd door het lab te zoeken naar van alles, misschien was er wel geen lab meer omdat ik de boel had laten ontploffen.Dank voor al jullie hulp in het lab.

Beste Jacky, bedankt voor al het werk dat je hebt verricht om mijn subsidieaanvragen, manuscripten, andere stukjes tekst en mijn proefschrift te editen.

Beste Helene, Bote, Joke, Ria en Mentje, Natasja en Erna, dank voor alles wat jullie voor me gedaan hebben. Het is te veel om op te noemen.

Een deel van dit proefschrift is mogelijk gemaakt met behulp van DNA samples van het Nederlands Initiatief voor Crohn en Colitis (ICC). Graag wil ik alle leden hiervan bedanken voor hun inzet bij het verzamelen van DNA en fenotyperingen. Met name het parelsnoer project is een geweldige basis voor nieuw gezamenlijk onderzoek. Ook zonder de patiënten en gezonde controles die belangeloos bloed en DNA doneerden, was dit onderzoek en proefschrift nooit tot stand gekomen.

Dear co-authors, thank you for all the help with writing, collecting samples and analyzing data.

Lieve Saar, lieve An, jeetje wat zijn de jaren voorbijgevlogen. Ik ben super trots en blij dat jullie naast me staan als mijn paranimfen, ik kan me geen betere voorstellen. We hebben elkaar altijd door dik en dun gesteund. Ik hoop dat nog vele jaren te doen.

Lieve papa, lieve mama, lieve Ruud, lieve Dagmar, lieve Tom, bedankt! Pap en mam, ik ben ontzettend blij en trots dat jullie mijn ouders zijn en hoe geweldig jullie mij altijd hebben gesteund en gestimuleerd. Ik ben jullie ontzettend veel dank verschuldigd. Ik hoop door dit proefschrift een deel van mijn dankbaarheid te laten zien. Mam, wat heb je een geweldige cover gemaakt! Ik heb nooit aan je getwijfeld. Ruud, Dagmar en Tom, thuis is er altijd wat te doen met jullie! Wat hebben we het altijd heerlijk samen. Dagmar bedankt dat ik dankzij jou altijd een slaapadresje had in Groningen.

Lieve Raymond, dank je voor alles wat je doet en laat én deed en liet voor mij, en dat slechts in ruil voor een beetje respect ;-) Nodeloos te zeggen... zonder jou had ik het niet gekund. Je bent en blijft een topper!

Curriculum Vitae

Karin Fransen, geboren op 29 maart 1984 te Emmen, groeide op in Nieuw-Weerdinge. Hier ging ze ook naar de basisschool. In 2001 behaalde ze op de Regionale Scholengemeenschap (RSG) te Ter Apel haar HAVO diploma. Nadat ze in 2003, eveneens op de RSG Ter Apel, haar VWO diploma had behaald, werd ze direct ingeloot voor de opleiding Geneeskunde aan de Rijks Universiteit te Groningen. Haar Junior coschappen deed ze in het Universitair Medisch Centrum Groningen (UMCG), haar coschappen liep ze in het Scheper ziekenhuis te Emmen en haar senior coschap deed ze op de afdeling Maag-darm-leverziekten in de Isala Klinieken te Zwolle met een verdiepende stage op de Hepatobiliaire chirurgie in het UMCG. In haar klinische les stond het onderwerp 'levercysten' centraal. Haar wetenschappelijke stage vervulde ze op de afdelingen genetica en maag-darm-leverziekten, eveneens in het UMCG. Het onderwerp van haar onderzoek was mRNA expressie in T cellen bij patiënten met colitis ulcerosa.

Tijdens haar studie is Karin drie maanden in Ghana geweest om vrijwilligerswerk te doen in het WomenHospital of Kumasi. Daar heeft ze geholpen met de organisatie van een promotie- en voorlichtingscampagne tegen seksueel overdraagbare aandoeningen. Hiervoor heeft ze scholen bezocht, waar ze wonden van kinderen verzorgde. Ook heeft ze geholpen met een screeningsprogramma voor volwassenen met hoge bloeddruk.

Tijdens haar wetenschappelijke stage in het UMCG solliciteerde Karin met succes naar een MD/PhD positie bij het Groninger University research institutefor Drug exploration (GUIDE). Hierdoor kon ze na haar senior coschap in 2010 beginnen met haar promotietraject onder leiding van promotor Prof C. Wijmenga en Prof. Rinse K. Weersma en copromotor Cleo C. van Diemen, wat resulteerde in het huidige proefschrift.

Momenteel werkt Karin in het Scheper Ziekenhuis te Emmen als AIOS op de algemene interne afdeling. Dit doet zij in het kader van haar opleiding tot Maag-, Darm-, en Leverarts.

List of publications

Fransen K, van Sommeren S, Westra HJ, Veenstra M, Lamberts LE, Modderman R, Dijkstra G, Fu J, Wijmenga C, Franke L, Weersma RK, van Diemen CC.Correlation of Genetic Risk and Messenger RNA Expression in a Th17/IL23 Pathway Analysis in Inflammatory Bowel Disease. Inflamm Bowel Dis. 2014 Mar 20.

Smeekens SP, Ng A, Kumar V, Johnson MD, Plantinga TS, van Diemen C, Arts P, Verwiel ET, Gresnigt MS, **Fransen K**, van Sommeren S, Oosting M, Cheng SC, Joosten LA, Hoischen A, Kullberg BJ, Scott WK, Perfect JR, van der Meer JW, Wijmenga C, Netea MG, Xavier RJ. *Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans.* Nat Commun. 2013;4:1342.

Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Büning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, **Fransen K**, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H; International IBD Genetics Consortium (IIBDGC), Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. *Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.* Nature. 2012 Nov 1;491(7422):119-24.

Fransen K, Mitrovic M, van Diemen CC, Thelma BK, Sood A, Franke A, Schreiber S, Midha V, Juyal G, Potocnik U, Fu J, Nolte I, Weersma RK. *Limited evidence for parent-of-origin effects in Inflammatory Bowel Disease associated loci.* PLoS ONE 2012

Hunt KA, Smyth DJ, Balschun T, Ban M, Mistry V, Ahmed T, Anand V, Barrett JC, Bhaw-Rosun L, Bockett NA, Brand OJ, Brouwer E, Concannon P, Cooper JD, Dias KM, van Diemen C, Dubois PC, Edkins S, Fölster-Holst R, **Fransen K**, Glass D, Heap GAR, Hofmann S, Huizinga TWJ, Hunt S, Langford C, Lee J, Mansfield J, Marrosu MG, Mathew CG, Mein CA, Müller-Quernheim J, Nutland S, Onengut-Gumuscu S, Ouwehand W, Pearce K, Prescott N, Posthumus MD, Potter S, Rosati G, Sambrook J, Satsangi J, Schreiber S, Shtir C, Simmonds MJ, Sudman M, Thompson SD, Toes R, Trynka G, Vyse TJ, Walker NM, Weidinger S, Zhernakova A, Zoledziewska M, Type 1 Diabetes Genetics Consortium,

UK IBD Genetics Consortium, Wellcome Trust Case Control Consortium, Weersma RK, Gough SCL, Sawcer S, Wijmenga C, Parkes M, Cucca F, Franke A, Deloukas P, Rich SS, Todd JA, van Heel DA. *Rare and functional SIAE variants do not alter autoimmune disease risk in up to 66,924 samples.* Nat Genet. 2011 Dec 27;44(1):3-5

Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G, de la Concha EG, de Almeida RC, Dias KM, van Diemen CC, Dubois PCA, Duerr RH, Edkins S, Franke L, **Fransen K,** Gutierrez J, Heap GAR, Hrdlickova B, Hunt S, Izurieta LP, Izzo, V, Joosten LAB, Langford C, Mazzilli MC, Mein CA, Midah V, Mitrovic M, Mora B, Morelli M, Nutland S, Núñez C, Onengut-Gumuscu S, Pearce K, Platteel M, Polanco I, Potter S, Ribes-Koninckx C,Ricaño-Ponce I, Rich SS, Rybak A, Santiago JL, Senapati S, Sood A, Szajewska H, Troncone R, Varadé J, Wallace C, Wolters VM, Zhernakova A, CEGEC (Spanish Consortium on the Genetics of Coeliac Disease), PreventCD Study Group, Wellcome Trust Case Control Consortium, Thelma BK, Cukrowska B, Urcelay E, Bilbao JR, Mearin ML, Barisani D, Barrett JC, Plagnol V, Deloukas P, Wijmenga C, van Heel DA. *Dense genotyping reveals and localises multiple common and rare variant association signals in celiac disease.* Nat Genet. 2011 Nov 6;43(12):1193-201

Fransen K, Mitrovic M, van Diemen CC, Weersma RK. *The quest for genetic risk factors for Crohn's disease in the post-GWAS era*. Genome Med. 2011 Feb. 25;3(2):13

Diaz-Gallo LM, **Fransen K,** Espino-Paisán L, Gómez-García M, van Sommeren S, Cardeña C, Rodrigo L, Mendoza JL, Taxonera C, Nieto A, Alcain G, Cueto I, López-Nevot MA, Bottini N, Barclay ML, Crusius JB, van Bodegraven AA, Wijmenga C, Ponsioen CY, Gearry RB, Roberts RL, Weersma RK, Urcelay E, Merriman TR, Alizadeh BZ, Martin J. *Differential association of two PTPN22 coding variants with Crohn's disease and ulcerative colitis*.Inflamm Bowel Dis. 2011 Feb 1. doi: 10.1002/ibd.21630.

Roberts RL, Hollis-Moffatt JE, Gómez-García M, **Fransen K,** Ponsioen CY, Crusius BA, Wijmenga C, Martín J, Weersma RK, Merriman TR, Barclay ML, Gearry RB, Alizadeh BZ. *Association of the protein-tyrosine phosphatase nonreceptor type substrate 1 (PTPNS1) gene with inflammatory bowel disease.* Inflammatory Bowel Diseases 2011 feb. 17(2):19–21

Fransen K, Visschedijk MC, van Sommeren S, Fu JY, Franke L, Festen EA, Stokkers PC, van Bodegraven AA, Crusius JB, Hommes DW, Zanen P, de Jong DJ, Wijmenga C, van Diemen CC, Weersma RK. *Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease.* Hum Mol Genet. 2010 Sep 1;19(17):3482-8