Genome-wide association studies and Crohn’s disease

James C. Lee and Miles Parkes

Abstract
The development of genome-wide association scanning (GWAS) has revolutionized the search for genetic loci associated with complex diseases. Crohn’s disease (CD), together with ulcerative colitis, has been a principal beneficiary of this technology with a recent meta-analysis from the International IBD Genetics Consortium increasing the number of confirmed CD susceptibility loci to 71. When one considers that prior to the development of GWAS only three susceptibility loci had been identified, the degree of progress becomes obvious. In this article we will summarize the principal discoveries that have been made in CD genetics and explain how these have contributed to our improved understanding of disease pathogenesis.

Keywords: Crohn’s disease; GWAS; IBD; Autophagy; NOD2; IL23 pathway

Crohn’s disease (CD) is a chronic inflammatory intestinal disease, which affects 3–4 individuals per 1000 population. It typically presents in young adults and produces significant morbidity—requiring immunosuppressive therapy and resection of diseased areas of the gut in up to 80% of patients. CD is thought to arise as a result of an inappropriate immunological response to commensal flora within the gut, although the specific environmental triggers that cause an individual to develop disease are poorly understood [1]. It has long been suspected, however, that environmental factors alone are not sufficient to result in disease development and that an underlying genetic susceptibility is also required. Indeed, early studies confirmed a role for a genetic contribution to disease pathogenesis by demonstrating increased concordance in monozygotic, compared with dizygotic, twins [2, 3, 4]. These and other studies in multiplex families demonstrated that CD has a lambda s ($\lambda_s$) of 20–35 and estimate the degree of the disease variance accountable by genetics as $\sim$50% [5, 6]. Interestingly, the estimated contribution attributable to the human leukocyte antigen (HLA) region on chromosome 6 is modest at most [7]. Consequently, even before the development of genome-wide association scanning, there was a high prior suspicion that CD would be a fruitful area for hypothesis-free genomic approaches.

THE PRE-GWAS ERA

Initial attempts to identify genes that might be contributing to the genetic susceptibility to CD were made using linkage studies and fine mapping. These studies identified several regions of linkage, subsequently named IBD1–7 according to their date of reporting and independent replication. In 2001, three independent groups reported identification of the gene at IBD1 on chromosome 16q12 as NOD2 ($CARD15$) [8–10]. These reports were a notable and rare success for linkage techniques, not only in CD, but in complex diseases in general. NOD2 encodes a cytoplasmic pathogen-associated molecular pattern receptor that binds to muramyl dipeptide, a component of both gram-positive and...
gram-negative cell walls. Of the numerous genetic variants within NOD2, three mutations have been shown to account for 81% of NOD2’s contribution to CD risk [11]. All three variants are individually rare but lie on the common ‘SNP5’ haplotype—making its detectability by GWAS truly a ‘synthetic association’. The causal variants all affect the portion of the gene encoding the C-terminal third of the protein, containing the leucine-rich repeat region. The reason this gene was successfully identified was significantly due to the large effect size of these mutations. Heterozygotes for one of these mutations have a 2–4-fold increased risk of CD, while homozygotes or compound heterozygotes have a 15–40-fold increased risk. Nonetheless, it is noteworthy that 8–15% of healthy subjects are heterozygous for one of these mutations, and it is evident that only a small proportion of individuals who carry two copies of ‘high-risk’ NOD2 genotypes develop CD [12]. It is also notable that 60–70% of CD patients do not have any of the commonly associated NOD2 variants; further highlighting the complexity. Subphenotype analyses have shown a replicable association between these NOD2 variants and stricturing, ileal disease [13].

One major advance that arose from the identification of NOD2 as a CD susceptibility gene was the subsequent focus of attention upon the role of innate immunity in CD pathogenesis. Interestingly, despite the large effect size, the exact mechanism by which the NOD2 variants contribute to CD pathogenesis remains the subject of debate and may turn out to be multifactorial. Indeed, NOD2 has been reported to have numerous roles in various cellular processes that might account for the association, including regulating Paneth cell function [14], viral sensing [15], affecting apoptosis in regulatory T cells [16] and influencing autophagy [17, 18].

THE GWAS ERA
Since the application of GWAS technologies to complex disease genetics, CD has been at the forefront of a rapidly moving field. Its success has been aided not only by the high heritability and robust diagnostic criteria for the phenotype, but also by the collaborative approach adopted by investigators in assembling large sample sets for study for example within the International IBD Genetics Consortium.

The first genome-wide studies in CD used arrays which surveyed only a small proportion of the genome but, consistent with the prior hypothesis that CD represented a fruitful area for such research, succeeded in identifying susceptibility genes that had not previously been implicated. The first such report in CD was from Japan [19], and identified association between CD and variants in TNFSF15—a gene encoding a cytokine belonging to the tumour necrosis factor (TNF) family. Association at this locus was subsequently also confirmed in European populations [20]. Functional studies have since demonstrated that allele-specific differences exist in the expression of TNFSF15 transcripts such that increased amounts of TNFSF15 are produced upon stimulation of T cells in individuals that have the disease-associated variant [21]. Furthermore, the protein encoded by TNFSF15 (TL1A) has been reported to induce both Th1 and Th17 immune responses [22], which later GWAS studies would confirm to be important components of the adaptive immune response that contributes to CD pathogenesis.

Another early genome-wide study in CD focused upon non-synonymous SNPs (nsSNPs) and identified association with a variant in ATG16L1, a gene whose protein is an integral component in the autophagy pathway [23]. This association was subsequently replicated in several full GWAS studies [24, 25]. One advantage of only using non-synonymous SNPs, despite this not truly being a ‘genome-wide’ approach, is that there is a higher probability that any associated variant will be causal rather than a proxy for the true causal variant. Indeed, in this case the association is fully explained by the T300A nsSNP. This study was also noteworthy as it was the first to implicate a known autophagy gene in CD pathogenesis. Subsequent GWAS studies have corroborated the association of CD pathogenesis with autophagy by detecting associated variants in other genes implicated in this pathway, including IRGM [25, 26]. This clearly illustrates the power of such hypothesis-free approaches, as autophagy had not previously been considered to play a role in CD pathogenesis, and has further emphasized the importance of defects in innate immune pathways in predisposing to CD.

Autophagy has long been recognized as an evolutionarily conserved process by which cytoplasmic organelles can be degraded in response to starvation. More recently, it has become clear that this process is also an important mechanism for the handling of intracellular micro-organisms. Consistent with this, and the original concept that CD arises due to an inappropriate response to the gut microbiota, several groups have subsequently reported that impaired
autophagy (through knock-down of proteins including ATG16L1 and IRGM) is associated with prolonged survival of intracellular pathogens [24, 27]. Furthermore, in a development that highlights the benefit of identifying novel pathways in disease pathogenesis, two groups have subsequently shown that NOD2 is also involved in autophagy [17, 18]. However, the complexity of environmental interaction with genetic susceptibility should not be underestimated. In a recent report, which nicely illustrates the type of multifactorial pathogenic process likely to pertain in IBD, a CD-like illness was noted to develop in ATG16-hypomorphic mice, but only in the presence of a dietary toxin (dextran sulphate sodium), intact gut flora and concurrent norovirus infection [28].

Another association reported by one of the earliest GWAS studies in CD was with IL23R, which encodes the receptor for interleukin (IL)-23 [29]. The IL-23 cytokine receptor actually comprises two subunits: IL12Rb1, which it shares with the IL-12 receptor, and the subunit encoded by IL23R, IL23R1. Notably, IL-12 had previously been thought to play an important role in the perpetuation of intestinal inflammation, as antibody-mediated neutralization of IL-12 resulted in abrogation of experimentally induced colitis in mice [30]. However, shortly before the association with genetic variants in IL23R was reported, experiments had demonstrated that this effect was actually due to IL-23 rather than IL-12 [31]. Interestingly, the risk variant in this gene accounting for most of the association signal is very common, with the minor allele being protective against CD. Regression analysis has since demonstrated that other variants also contribute to the signal at this locus [32]. Furthermore, Tremelling and colleagues [32] also demonstrated that, unlike the association with autophagy genes, this locus is also associated with ulcerative colitis; the other common form of inflammatory bowel disease.

IL-23 is a cytokine that plays a key role in the development and maintenance of Th17 cells. This subset of CD4 T cells, which are distinct from the conventional Th1 and Th2 subsets, are a class of effector T cells that produce IL-17 (hence the name) and have been associated with several other autoimmune and inflammatory conditions including psoriasis, ankylosing spondylitis, multiple sclerosis and celiac disease. Indeed, subsequent GWAS studies in CD [25, 33, 34] have not only replicated the association at the IL23R locus but also identified many other genes that lie within the Th17 pathway (e.g. TYK2, JAK2, STAT3, CCR6 and ICOS-L).

One of the reasons for seeking to better understand the pathogenesis of complex diseases, such as CD, is that it may be possible to develop novel therapies that could target pathways implicated in disease once they are identified. In this regard, GWAS studies in CD have spurred direct clinical progress as drugs targeting both the IL-23 pathway and autophagy are now being trialled as therapies [35, 36].

A key theme that has arisen from GWAS studies in CD, but which also applies more generally in complex disease genetics, is that most associated variants are not in coding regions of the genome and may therefore exert their effects by altering transcriptional regulation. One of the earliest examples of this was the reported association of a large gene desert on chromosome 5p13 with CD [25, 26, 33]. Variants at this locus have been reported to correlate with expression of PTGER4 [33], which encodes a prostaglandin whose receptor had already been implicated in a murine model of IBD [37]. Many other loci associated with CD also lie in non-coding regions of the genome and subsequently the notion has emerged that regulatory variation, impacting on gene transcription, may be at least as important as coding variation [38].

In total, GWAS studies have identified genetic variants within 71 genes and loci that are associated with CD [39] and this has lead to our appreciation of the integral role of pathways within both the innate and adaptive immune systems in the pathogenesis of CD. Given generally small effect sizes, the successful detection of these variants has relied significantly on large-scale, international collaborations [38, 39] in order to generate sufficiently powered sample sets. Indeed, the most recent meta-analysis performed by the International IBD Genetics Consortium had >20 000 cases and 25 000 matched controls [39], a sample size much larger than any individual centre would be able to easily acquire. The benefit of such collaborations is highlighted by the recent detection of several new associations. These include SMAD3, a gene involved in TGF-β signalling which has been linked to inflammation and fibrosis; FUT2, a gene which can influence susceptibility to norovirus infection; and DNMT3a, which encodes one of three human DNA methyltransferases that can directly influence gene transcription via methylation of CpG islands within the genome—further highlighting the
importance of epigenetic regulation in disease pathogenesis. However, it is true to say that the mechanisms by which many of these polymorphisms result in increased disease susceptibility is not well understood, and represents an area of considerable interest for future study.

Interestingly, with regard to CD pathogenesis there appears to be only a limited role of the HLA region, which has been associated predominantly with isolated colonic CD—at the same variant that is associated with severe ulcerative colitis [40]. Besides this, and the association of NOD2 variants with ileal disease, it is notable, and perhaps surprising, that none of the other CD-associated variants have shown association with specific subphenotypes of disease. This does not mean that subphenotypes are not heritable, as there is already evidence that they can be [6, 41], but perhaps implies that comparing a heterogenous group of disease cases with controls is not the optimal method to detect polymorphisms that may only be enriched in a subset of these cases.

One striking feature that has emerged as more susceptibility genes have been detected is the degree of overlap in genetic susceptibility between CD and other autoimmune/inflammatory conditions. The greatest overlap is, unsurprisingly, seen with ulcerative colitis. It has been interesting to note that the loci that appear to be shared mainly involve the adaptive immune system (e.g. IL23R, IL10, PTPN22, ICOSL, HLA and others), and those that appear to be disease specific predominantly affect different aspects of innate immunity, such as autophagy genes in CD and epithelial barrier genes in UC. There is also considerable overlap with other inflammatory diseases, such as type 1 diabetes, multiple sclerosis and psoriasis. More surprising is the apparent overlap at multiple loci with type 2 diabetes and also with infectious disease, notably leprosy [42]. The latter observation may reflect distinct selection pressures acting in different populations through evolutionary history.

WHAT NEXT?

It is clear that GWAS studies have greatly advanced our understanding of CD pathogenesis, and that the jigsaw puzzle that is the molecular genetic architecture of CD is gradually becoming clearer. However, despite this, an important caveat of the progress to date is that most large-scale efforts have focused on populations of European origin. It is well known that different patterns of linkage disequilibrium exist in different ethnic groups, and studies in non-European populations have already identified significant differences in susceptibility loci. For example, CD is not associated with NOD2, IL23R or ATG16L1 variants in East Asian populations [43, 44]. Understanding how the genetic aetiology and disease mechanisms differ between ethnic groups will be an important challenge for the IBD genetics community in the future.

It will also be important for future research to embrace complementary genetic approaches, as it is increasingly clear that GWAS will not explain all of the heritability of CD. Indeed, it has become apparent that the newly associated loci are conferring an ever decreasing effect size. Consequently, while the first 32 CD loci identified were estimated to explain 20% of the genetic risk [38], the next 39, identified using a sample set twice as large, has increased this to just 25% [39]. Moreover, one can extrapolate this observation to conclude that even with sample sets 10-fold larger than those used in the most recent meta-analysis, the amount of heritability explained would not exceed 30% [39]. The reasons for this are multiple but include the following:

(i) rare variants, which may confer a relatively large effect size, are not well tagged by GWAS arrays and hence have received little study to date;
(ii) many identified association intervals may contain more than one causal allele;
(iii) 30% of the genome is poorly covered by GWAS arrays; and
(iv) most GWAS SNPs are proxies for the true causal variant and hence are likely to significantly underestimate the true attributable risk at each locus.

Many of these limitations can be addressed with alternative approaches—which will also yield additional mechanistic insights. For example, fine mapping of IBD susceptibility loci to detect the causal variant(s) will become possible through the use of custom arrays based upon large-scale sequencing efforts; whole genome sequencing will increasingly identify rare variants for interrogation; and techniques of CHIP-seq and RNA-seq in specific cell types will increasingly illuminate the impact of putative causal variants on the regulation of gene transcription.
The other major challenge for future work in this field will be to determine the functional role of the variants that have been discovered and explain how these interact with environmental factors. This represents a major challenge and one that will require collaboration with other fields of science outside genetics to be truly successful. Some progress is already being made in this direction—with recent examples in CD coming from the interrogation of autophagy pathways. In one such example, an animal model of IBD was developed which required not only disruption of the ATG16L1 gene and an intact gut flora but also a persisting noroviral infection and an environmental stressor to elaborate the full phenotype. Such complex interactions are likely to be the reality of CD in humans. Their accommodation within animal models of IBD can only help us to better understand the pathogenic mechanisms by which a host of genetic variants predispose to chronic relapsing bowel inflammation in CD-affected individuals.

FUNDING

James Lee is a Wellcome Trust Clinical PhD Programme Fellow. We also acknowledge the NIHR Biomedical Research Centre awards to Addenbrooke’s Hospital/University of Cambridge School of Clinical Medicine.

Key Points

- Genome-wide association studies have identified 71 susceptibility loci in CD.
- Key discoveries include the role of autophagy in CD pathogenesis and the shared role of the IL-23 pathway with ulcerative colitis.
- Despite this success, only ~25% of the genetic variance of CD has been accounted for so far.
- For most loci the causal variant remains unknown—the identification of these variants, elucidation of their effect and discovery of the ‘missing heritability’ represent the major challenge for IBD genetics in the next 5 years.

References

1. Xavier RJ, Podolsky DK. Unravelling the pathogenesis of inflammatory bowel disease. 


