Review article

Understanding inflammatory bowel disease via immunogenetics

Katrina M. de Lange a, Jeffrey C. Barrett a, b, c, *

a Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, United Kingdom
b European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, United Kingdom
c Centre for Therapeutic Target Validation, Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, United Kingdom

A R T I C L E   I N F O

Article history:
Received 23 July 2015
Accepted 23 July 2015
Available online 7 August 2015

Keywords:
Inflammatory bowel disease
Genetics
Genome wide association study
Therapeutic targets
Microbial sensing

A B S T R A C T

The major inflammatory bowel diseases, Crohn's disease and ulcerative colitis, are both debilitating disorders of the gastrointestinal tract, characterized by a dysregulated immune response to unknown environmental triggers. Both disorders have an important and overlapping genetic component, and much progress has been made in the last 20 years at elucidating some of the specific factors contributing to disease pathogenesis. Here we review our growing understanding of the immunogenetics of inflammatory bowel disease, from the twin studies that first implicated a role for the genome in disease susceptibility to the latest genome-wide association studies that have identified hundreds of associated loci. We consider the insight this offers into the biological mechanisms of the inflammatory bowel diseases, such as autophagy, barrier defence and T-cell differentiation signalling. We reflect on these findings in the context of other immune-related disorders, both common and rare. These observations include links both obvious, such as to pediatric colitis, and more surprising, such as to leprosy. As a changing picture of the underlying genetic architecture emerges, we turn to future directions for the study of complex human diseases such as these, including the use of next generation sequencing technologies for the identification of rarer risk alleles, and potential approaches for narrowing down associated loci to casual variants. We consider the implications of this work for translation into clinical practice, for example via early therapeutic hypotheses arising from our improved understanding of the biology of inflammatory bowel disease. Finally, we present potential opportunities to better understand environmental risk factors, such as the human microbiota in the context of immunogenetics.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Contents

1. Introduction .................................................................................................................. 92
2. The early days of IBD genetics ...................................................................................... 92
   2.1. Twin studies .............................................................................................................. 92
   2.2. Linkage studies ........................................................................................................ 92
3. The GWAS era .............................................................................................................. 94
   3.1. Technological developments that made GWAS possible ..................................... 94
   3.2. GWAS: A revolution in IBD genetics ............................................................... 94
   3.3. Meta-analyses and the importance of sample size .............................................. 94
   3.4. IBD genetics in the context of other diseases .................................................. 96
   3.5. Expanding into non-European populations .................................................... 96
4. Beyond GWAS ............................................................................................................. 96
   4.1. Rare and low frequency variation ....................................................................... 97
   4.2. Identifying the casual mutations ........................................................................ 97

* Corresponding author. Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1HH, United Kingdom.
E-mail address: jb26@sanger.ac.uk (J.C. Barrett).

http://dx.doi.org/10.1016/j.jaut.2015.07.013
0896-8411/© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Clinical and epidemiological features of the two major inflammatory bowel disease subtypes, Crohn's disease and ulcerative colitis [12].

<table>
<thead>
<tr>
<th></th>
<th>Crohn's disease</th>
<th>Ulcerative colitis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Incidence patterns</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>More common in women than men</td>
<td>Equal rates in men and women</td>
</tr>
<tr>
<td>Prevalence rates</td>
<td>CD is more prevalent than UC in developed countries</td>
<td>UC emerged before CD in developed countries, and is more prevalent in still-developing countries</td>
</tr>
<tr>
<td><strong>Disease localisation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Affected areas</td>
<td>Entire gastrointestinal tract (from mouth to anus)</td>
<td>Colon, plus some potential backwash ileitis</td>
</tr>
<tr>
<td>Inflammation pattern</td>
<td>May occur as patchy, discontinuous inflammation</td>
<td>Continuous inflammation in the affected area (though sometimes a separate cecal patch)</td>
</tr>
<tr>
<td><strong>Histopathology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penetration</td>
<td>Transmural inflammation of the entire bowel wall</td>
<td>Inflammation restricted to the mucosal and submucosal layers (other than in fulminant colitis)</td>
</tr>
<tr>
<td>Appearance</td>
<td>Thickened colon wall with granulomas, deep fissures and a cobblestone appearance</td>
<td>Distorted crypt architecture, with shallow erosions and ulcers; granulomas, if present, only around crypts</td>
</tr>
<tr>
<td><strong>Serological markers</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Anti-Saccaromyces cerevisiae antibodies</td>
<td>Anti-neutrophil cytoplasmic antibodies</td>
</tr>
<tr>
<td><strong>Complications</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fistulas, abdominal mass (typically lower right quadrant), colonic and small-bowel obstructions, stomatitis</td>
<td>Haematochezia, passage of mucus or pus, fulminant colitis and toxic megacolon</td>
</tr>
</tbody>
</table>
measured the patterns of inheritance within families: around 300 markers evenly distributed across the genome were sufficient to capture the pattern of DNA inheritance within a family [9]. By tracing the DNA segments that segregate with disease status (such as variant alleles only seen in affected individuals, and not in their unaffected relatives), sections of the genome that confer risk to the disease can be identified. This linkage analysis approach is good for detecting highly penetrant variants (i.e. those that are extremely likely to cause disease whenever present) that segregate well with disease status.

Linkage studies successfully identified many such highly penetrant variants for rare disorders [10–16], and were subsequently applied to a range of more common diseases. In 1996, the first such study in IBD linked a portion of chromosome 16 (dubbed ‘IBD1’) with Crohn’s disease [17], which was successfully replicated in subsequent studies [18–23]. This finding was followed up using more closely packed markers within a small number of genes, and the IBD1 region on chromosome 16 was found to be caused by three disease risk alleles in the gene NOD2, whose role in the recognition of bacterial peptidoglycans and subsequent stimulation of an immune response supports its likely association with the development of CD (Fig. 1A) [24–26]. These variants were especially common in Ashkenazi Jews, partially explaining the increased burden of CD in that group. Unfortunately, however, successes like NOD2 were rare: it remained one of the few robustly replicated genetic risk loci discovered via linkage, not just in IBD, but across common diseases.

2.3 Limitations of linkage studies and the common disease, common variant hypothesis

The widespread disappointment from linkage results among common diseases reflected a fundamental property of their genetic architecture: these diseases did not have a single, highly penetrant genetic cause. Instead, it was proposed by Risch and Merikangas (1996) that complex diseases were driven by the accumulation of many risk factors of only modest effect (the ‘common disease, common variant’ hypothesis). Finding associations via linkage under this scenario is difficult, as the genetic risk may be spread throughout the genome rather than concentrated in a single locus.
An alternative association analysis approach (which tests if the population-level allele frequencies of cases and controls are statistically different) is much more powerful, provided it is possible to choose the right variant to test among the millions known to exist in the human population. Risch and Merikangas (1996) calculated that 17,997 affected sibling pairs would be needed to detect a risk allele with 50% frequency and an odds ratio of 1.5 using linkage, as opposed to just 484 using an association analysis.

3. The GWAS era

3.1. Technological developments that made GWAS possible

Case-control association studies could detect signals too weak to show linkage, but suffered the drawback of needing to know which variants to test. One solution to this problem was to select candidate genes, based on prior biological hypotheses, but this produced a deluge of association claims with weak statistical evidence that did not replicate in subsequent studies [28]. This failure to find robust genetic associations in biological candidates highlighted a repeated theme of genetic studies that can scan the entire genome in an unbiased way producing unexpected insights.

Three developments upended this stasis in gene discovery, and fundamentally changed gene mapping. First, by 2005, the public database of the most common type of genetic variant, single nucleotide polymorphisms (SNPs, where a single letter of DNA is variable) contained 9.2 million sites that had been catalogued by projects such as the SNP Consortium and the International HapMap Consortium [29,30]. Second, these catalogues of population-level genetic variation had also shown that variants common in the general population (minor allele frequency [MAF] > 5%), and in physical proximity, were highly correlated, or in linkage disequilibrium (LD), with each other. Human population history had left a pattern of long LD blocks of high correlation, separated by small ‘hotspots’ where most historical recombination events tended to cluster [31]. This uneven LD pattern meant that it was possible to test the majority of common variants by carefully selecting markers in each long LD block. Appropriately 500,000 well chosen SNPs could capture nearly 5 million common SNPs in Europeans and East Asians; unsurprisingly, the more genetically diverse African populations required almost twice as many markers to capture the same amount of variation [32]. Finally, in the mid-2000s, it became economically feasible to genotype hundreds of thousands of variants using new microarray technologies. These key advances opened the way for genome-wide association studies (GWAS) that could be used to detect the diverse genomic loci associated with a given complex trait. GWAS combined the hypothesis-free ability to scan the whole genome of linkage with the statistical power to detect associations of smaller effect size.

3.2. GWAS: A revolution in IBD genetics

Crohn’s disease was among the first diseases studied using GWAS, beginning in 2006. In addition to confirming the established NOD2 association, these early studies [33–36] identified four new loci at genome-wide levels of statistical significance (p < 5 × 10⁻⁸), demonstrating the power of the GWAS approach. The strongest new association was a protective variant in IL23R, which encodes a receptor protein embedded in the cell membrane of many different types of immune cells and, upon binding of IL23, starts a signalling cascade that promotes inflammation and coordinates an adaptive immune response (Fig. 1B) [33]. A more surprising discovery was an association to a protein-coding variant in ATG16L1, which encodes a protein involved in the autophagosome pathway (Fig. 1A) [34], and provided the first evidence for the importance of autophagy in CD. This pathway is responsible for processing intracellular bacteria, and so the ATG16L1 association contributed to further understanding of dysfunction of the intestinal barrier in Crohn’s disease. Finally, these early studies discovered a pair of associations on chromosomes 5p13 and 10q21 that were far from any genes [35,36]. Unlike the previous associations, these new results highlighted the important role of regulatory and non-coding elements in complex disease. Motivated by these early successes, further GWAS used increasingly larger sample sizes to implicate both the innate (NOD2, CARD9) and adaptive (TUNFSF15, PITPN2, IL-12B) immune response pathways in inflammatory bowel disease, and recapitulate the role of autophagy and intracellular bacteria management (NOD2, ATG16L1, IRGM) in Crohn’s disease [37].

These initial CD studies also suggested a partial overlap of genetic risk for ulcerative colitis: of the Crohn’s disease associations discovered, about 30% were also found to be associated with UC via replication studies [9]. Additional GWAS in ulcerative colitis cohorts lead to the discovery of multiple novel UC-specific loci [38–41]. Three loci with roles in intercellular interactions were identified, suggesting epithelial barrier defects may play a role in UC-specific pathogenesis: HNF4A, which encodes a transcription factor regulating the expression of components in the cell–cell junction; CDH1, which produces a transmembrane glycoprotein that is crucial to the adherens junction compartmentation and has also been associated with colorectal cancer susceptibility, and; LAMB1, which encodes the β1 subunit of a laminin, the major noncollagenous components of basement membranes [41].

These UC-specific studies also confirmed the long-established association between UC and the classic human leukocyte antigen (HLA) locus [42], which contains genes encoding antigen-presenting proteins on the surface of the cell, and plays a crucial role in the regulation of the adaptive immune system. Despite the HLA being strongly associated with many other chronic inflammatory and autoimmune disorders, the association with CD is much weaker [43]. The pattern of association to IBD in the HLA region is the most complicated in the genome. While the most recent study of HLA in IBD conclusively showed that the HLA-DRB1*01:03 allele is the most strongly associated in both CD and UC, it also identified more than ten additional risk alleles associated with one or both diseases [44]. Most of these associations are disease-specific; HLA class I and class II variation contributes equally to CD, class II variation is more important in ulcerative colitis. In addition, evidence of decreased heterozygosity in HLA genes was observed for ulcerative colitis only. This non-additive effect (similar to that observed in HLA alleles associated with Type 1 diabetes [45]) highlights the importance of being able to detect a wide range of antigens for protective immunity.

3.3. Meta-analyses and the importance of sample size

While this flurry of discoveries generated new biological hypotheses for IBD, it became clear that these relatively weak associations cumulatively explained only a fraction of the heritability expected from twin studies. This ‘missing heritability’ problem was universal amongst complex diseases during the early GWAS era, and was partially attributed to types of variation not captured by GWAS (e.g. non-European, rare and structural variants) [46,47]. As Fig. 2 shows, these early studies were actually poorly powered because the true genetic architecture of IBD includes many variants with odds ratios <1.2 or even 1.1. The International IBD Genetics Consortium (IIBDGC) was formed to pool thousands of already genotyped samples from previous GWAS to search for these small effects. Comparisons among different genotyping chips was enabled by imputation, which infers missing data by comparing known genotypes to those in a representative reference set with
The first of these IIBDGC meta-analyses effectively tripled the number of known Crohn’s disease susceptibility loci with the identification of 21 novel associations, including another autophagy gene (LRRK2) [49]. This was followed by a meta-analysis of ulcerative colitis studies, which identified 29 new UC risk loci, and a second Crohn’s disease meta-analysis that brought the number of CD susceptibility loci to 71 [50,51]. This rapid accumulation of IBD risk loci culminated in 2012 with a meta-analysis containing over 75,000 cases (including both CD and UC for the first time) and controls, that brought the total number of IBD loci to 163 [52].

Numerous pathways were implicated through multiple genetic associations, including those involved in innate mucosal defence, JAK/STAT signalling, cytokine production (particularly interferon-γ, interleukin (IL)-12, tumour-necrosis-factor-α and IL10 signalling) and lymphocyte activation.

This dramatic growth in the number of IBD-associated loci, together with the first large-scale joint analyses of CD and UC, revealed that genetic risk for CD and UC substantially overlap. Although early GWAS data had suggested quite disparate underlying pathways, of the 163 loci identified in the Jostins et al. (2012) study, numerous pathways in inflammatory bowel disease pathogenesis were implicated. Genes belonging to these pathways and falling within IBD-associated loci are indicated, and cases where these overlap with other immune-mediated disorders are marked. Note however that in some cases the specific genes have not yet been identified as causal and, as many loci contain multiple candidate genes, these should not be treated as confirmed [52,53,63,86,94–101].

### Table 2

<table>
<thead>
<tr>
<th>Pathway implicated</th>
<th>Pathway genes in IBD-associated loci</th>
<th>Overlap with other immune-mediated disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Innate immune response</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epithelial barrier function and repair</td>
<td>CDH1, ERRFI1, GNA12, HNF4A, ITLN1, MUC19, NKK2-3, PLAGL2E, PTGEB4, REL, STAT3</td>
<td>NKK2-3, PTGER4 REL REL REL - - PTGER4, STAT3</td>
</tr>
<tr>
<td>Innate mucosal defence</td>
<td>CARD9, FCGR2A, IL18RAP, ITLN1, NOD2, REL, SLC11A1</td>
<td>CARD9, FCGR2A REL IL18RAP, REL FCGR2A, REL FCGR2A, IL18RAP</td>
</tr>
<tr>
<td>Autophagy</td>
<td>ATG16L1, CUL2, DAP, IRGM, LRRK2, NOD2, PARK7</td>
<td>- - REL REL REL REL</td>
</tr>
<tr>
<td>Apoptosis/necroptosis</td>
<td>DAP, FASLG, MST1, PUS10, THADA</td>
<td>- - PUS10 - - -</td>
</tr>
<tr>
<td><strong>Activation of adaptive immune response</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IL23-R response pathway</td>
<td>CCR6, IL12B, IL21, IL23R, JAK2, STAT3, STAT4, TYK2</td>
<td>IL2B, IL23R, JAK2, TYK2 IL2B, IL23R, STAT3, TYK2 IL21, STAT4, TYK2</td>
</tr>
<tr>
<td>NF-κB</td>
<td>NFκB1, REL, TNFAIP3, TNIP1</td>
<td>REL, TNFAIP3 REL, TNFAIP3 REL, TNFAIP3</td>
</tr>
<tr>
<td>Aminopeptidases</td>
<td>ERAAD1, ERAAD2</td>
<td>IL2, IL21, IL2RA</td>
</tr>
<tr>
<td>IL2 and IL-21 T-cell activation</td>
<td>ICOSLG, IFNG, IL12B, IL2, IL21, IL23R, IL2RA, IL7R, NDFIP1, PIM3, PRDM1, TAGAP, TNFRSF9, TNFSF8</td>
<td>ICOSLG, TAGAP PRDM1, TAGAP, TAGAP PRDM1, IL12B, IL2RA, IL7R, TAGAP</td>
</tr>
<tr>
<td><strong>Regulation of adaptive immune response</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Th17 cell differentiation</td>
<td>AHR, CCR6, IL2, IL22, IL23R, IRF4, JAK2, RORC, STAT3, TNFSF15, TYK2</td>
<td>IL23R, JAK2, IL23R, TYK2 CCR6, TYK2 TYK2</td>
</tr>
<tr>
<td>T-cell regulation</td>
<td>ICOSLG, IFNG, IL12B, IL2, IL21, IL23R, IL2RA, IL7R, NDFIP1, PIM3, PRDM1, TAGAP, TNFRSF9, TNFSF8</td>
<td>ICOSLG, TAGAP PRDM1, TAGAP, TAGAP PRDM1, IL12B, IL2RA, IL7R, TAGAP</td>
</tr>
<tr>
<td>B-cell regulation</td>
<td>BACH2, IKZF1, IL5, IL7R, IRF5</td>
<td>BACH2</td>
</tr>
</tbody>
</table>
paper, 110 were associated with both phenotypes. Furthermore, of the 30 CD-specific and 23 UC-specific loci, 43 show the same direction of effect in the non-associated disease suggesting that only a tiny minority truly have zero effect in the other disease. This considerable overlapping genetic risk implies that the two diseases shared many biological mechanisms. The few loci that are CD- or UC-specific, as well as the relative size of effects at shared loci, might reveal clues about the distinct pathologies of the two diseases.

3.4. IBD genetics in the context of other diseases

Understanding both the shared and private genetics of related disorders can be useful for constructing hypotheses about the underlying biological pathways that may be driving each disease, and how distinct clinical phenotypes may arise. For example, known IBD loci are enriched for genes involved in primary immunodeficiencies, including those linked to reduced levels of circulating T cells (ADA, CD40, TAP1, TAP2, NBN, BLM, DNMT3B), and to T-helper cells responsible for producing Th17, memory, and regulatory T cells (STAT3, SP110, STAT5B). It thus became clear that the same genes can be affected by damaging protein coding variants that cause these severe disorders as well as much more subtle (presumed regulatory) variants that slightly affect risk of complex diseases like IBD.

Several studies have also extended these cross-disease genetic comparisons to potentially related complex diseases, such as the common immune-mediated disorders (ankylosing spondylitis, coeliac disease, multiple sclerosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, and type 1 diabetes). Early analysis of GWAS results from across these diseases (and IBD) suggested that the innate immune response, as well as the general immune pathways involved in T-cell differentiation and signalling, are shared between many of them (Table 2) [43]. This observed overlap of risk loci among common immune mediated diseases then motivated the design of a new genotype array, called Immunochip, which contained markers densely covering loci selected for their known association with at least one of 11 immune-mediated diseases, or their suggestive significance in the early immune-related GWAS studies. This targeted array, which cost approximately 20% of the price of contemporary GWAS chips, made the genotyping of large samples of immune-mediated disorders possible, and also paved the way for more extensive disease subphenotype and cross-disease studies [53]. Indeed, the Immunochip formed the basis of the Jostins et al. (2012) IBD meta-analysis, which showed that 70% (113 out of 163) of the IBD loci identified are also shared with other complex diseases or traits, including 66 loci shared with other immune-mediated disorders. Sharing is particularly strong between IBD and the other sero-negative diseases, ankylosing spondylitis and psoriasis. Interestingly, across the immune-mediated diseases those loci that are not shared tend to have large effect sizes, which would explain why the genetic underpinnings of CD and UC appeared so misleadingly disparate prior to the large meta-analysis efforts [53]. Overlap with more distantly related diseases can also provide useful insights. For example, Jostins et al. (2012) observed an enrichment in genes previously linked with Mendelian susceptibility to mycobacterial disease (MSMD) and leprosy (a complex mycobacterial disease): these overlaps suggest that the genetic architecture of IBD may have been shaped by selection pressures arising from mycobacterial infection [52].

3.5. Expanding into non-European populations

Up until this point, GWAS in IBD had largely focused on samples of European ancestry. One notable exception was a Crohn’s disease study in 2005 [54], performed in a Japanese population after it was noted that NOD2 did not appear to play a significant role in the pathogenesis of CD in Japan [55–57]. This study identified a strong association between the gene TNFSF15 and CD despite an initial sample size of fewer than 100 patients. Additional IBD genome wide association studies within Indian, Japanese and Korean populations showed that most IBD genetic risk is shared regardless of ancestry [58–62], but many of these studies were small, preventing informative comparisons across populations.

A large IBD study of multiple ancestries was conducted by the IIBDGC both to study IBD associations apparently unique to one population, and to boost power for detection in all populations using meta-analysis techniques that account for population stratification. GWAS and Immunochip data were analysed from 96,486 individuals of European, East Asian, Indian and Iranian descent, yielding a total of 200 IBD associated regions [63]. For the vast majority of these loci, the direction and magnitude of the effect is consistent between the European and non-European cohorts, implying that the underlying causal variants at these shared loci are likely to be common, as rare alleles are more likely to be population-specific. For the handful of associations that appear to be heterogeneous between populations, nearly all are due to differences in allele frequency between populations. For example, NOD2 is not biologically less relevant in Japan, but rather the IBD risk variants are simply absent in that population. Only TNFSF15 (which exhibits microbial-induced expression [64]) and the autophagy gene ATG16L1 are common in all populations but appear to have different effect sizes, possibly reflecting differences in gene–environment interactions between the populations.

4. Beyond GWAS

Both the meta-analyses and trans-ancestry studies contributed to an almost 20-fold increase in the number of known IBD-associated loci (Fig. 3). However, as with many complex diseases, this approach of analysing ever-larger genotype array-based datasets still captures only the fraction of IBD heritability explained by common variants, mostly in European populations. In fact, the latest estimates by Chen et al. (2014) suggest that common variants explain only 26% of the heritability of Crohn’s disease, and 19% of the heritability of ulcerative colitis [65]. Some of this missing heritability may be found in regions sometimes
overlooked by GWAS, such as the sex chromosomes. A recent study by Chang et al. (2014) utilized X-chromosome data from existing datasets to identify a new IBD-associated gene, ARHGFP6, which interacts with a major surface protein on Helicobacter pylori (a gastric bacterium) [66]. Rare loss-of-function variants in the X-chromosome gene XIAP, which encodes a protein that inhibits apoptosis, have also been identified as strongly predisposing for early-onset Crohn’s disease in males [67,68]. However, uncovering rare variants associated with complex disease will require the development of new study designs, as they generally have low correlation to the marker SNPs used (which usually have much higher allele frequencies, MAF > 0.05, to better capture other common variation) and are therefore not well tagged [69].

4.1. Rare and low frequency variation

To successfully identify a rare or low frequency disease-associated allele it is necessary to directly test the variant site itself, as such variants are not in high LD with many others, preventing the capture of its signal by a proxy variant (the method which drove the success of GWAS). Furthermore, because such alleles are by definition observed infrequently in the population, even the largest catalogues of human variation are unlikely to contain all variants of interest. Instead, discovery tends to require sequencing of an entire region (not just the known variable sites), something that became possible with the emergence of next generation sequencing (NGS) technology.

In its infancy, NGS was still prohibitively expensive, so sequencing was limited to a handful of genes in small numbers of samples. One approach to maximize the effectiveness of IBD sequencing studies was to consider early-onset IBD, as the XIAP studies did. Early-onset IBD tends to be more severe, and may be more similar to single-gene, or Mendelian, disorders than adulthood IBD. Glocker et al. (2009) identified rare mutations affecting IL10R protein subunits using a combination of linkage analysis and candidate gene sequencing in early-onset IBD cases from unrelated consanguineous families [70]. Similarly, Blaydon et al. (2011) identified a rare loss-of-function mutation in the gene ADAM17 (necessary for the cleavage of the epithelial-cell mitogen TGF-α from the cell membrane) that was homozygous in a consanguineous sibling pair affected by inflammatory bowel disease and skin lesions [71]. As the cost of sequencing started to fall, several studies used NGS to search for rare and low frequency variation in candidate IBD loci across case control cohorts. One of the earliest such studies sequenced 56 candidate genes identified by GWAS in 350 CD cases and 350 controls (with follow up genotyping in tens of thousands of IBD patients), identifying four additional risk variants in NOD2, two protective variants in IL23R, and a protective splice variant in CARD9 [72]. A similar study of 55 candidate genes in 200 UC cases and 150 controls recapitulated the presence of rare variants in CARD9 and IL23R, and identified a new association in RNF186 [73].

Just as was seen during the GWAS era, the logical next step is to scale these candidate-gene sequencing studies up to genome-wide projects: however, deep sequencing of whole genomes across sufficiently large case/control cohorts is still too expensive. Because the minor allele of a given rare variant is observed so infrequently, obtaining a significantly large difference in MAF between cases and controls is not possible with achievable sample sizes (Fig. 2). One approach is to use burden testing, which reduces the number of samples needed to detect a rare variant association by aggregating information across all variants in a given target region (such as a gene or exon). Every occurrence of a variant at any position in the region contributes to the overall count, and the difference in these counts between cases and controls is then tested as though they were from a single site of variation. In this way, rare variant associations can be detected with sample sizes that are more comparable to those used to test common variation.

Despite this, obtaining sufficiently large sequenced datasets is still difficult. Zuk et al., [2014] suggest at least 25,000 cases and an equivalent number of controls are needed for a well-powered study [74]. While ultimately whole genome sequencing will become affordable, two distinct intermediate approaches exist to sequence large numbers of individuals. First, borrowing the most popular approach in Mendelian genetics, is to only sequence the so-called exome (all exons, or coding regions, in the genome), as this represents just 1% of the complete genome [75]. However, the majority of IBD-associated loci identified during the GWAS era actually lie in non-coding regions, and it is likely that rare variants affecting gene regulatory pathways will be of interest. The second design is to spread a fixed amount of sequence data across the whole genomes of many individuals. This produces lower quality data per individual, but the increased sample size improves power to detect low frequency and rare variation in a fixed-cost study [76]. As an added advantage, such cohorts of sequenced individuals then provide useful disease-specific reference panels for imputing rare variants into new and existing GWAS datasets.

4.2. Identifying the casual mutations

With 200 loci associated with Crohn’s disease and ulcerative colitis over the past two decades, and the promise of more to come as next generation sequencing studies grow, attention is now turning to the identification of casual genes and variants within these loci (fine-mapping). Historically, follow-up of genetic associations has proceeded via time-consuming experimental validation of proposed genes using cellular or mouse models. While such functional evidence is essential to fully understanding the biology implicated by genetics, it is also possible to leverage the huge sample sizes put together for GWAS to improve fine-mapping before undertaking these experiments. Currently, attempts are being made to fine-map causal variants in a high-throughput way using the IBDGC’s large Immunochip cohorts, aiming to replicate the success seen in coeliac disease, where the densely packed markers on the Immunochip were used to narrow approximately half of the known signals to an individual gene (or even subregions of genes in some cases) [77]. Further prioritisation of candidate
SNPs can be improved by the availability of quality functional annotations from projects such as ENCODE [78], samples from multiple populations (as LD patterns differ between groups of differing ancestry), and combined datasets of huge sample size. Various algorithms have been developed to rank variants within a locus [79,80], but no definitive method for identifying the disease risk allele exists.

A recent study by Farh et al. (2015) highlights some of the potential challenges in fine-mapping loci given the current knowledge of the effects of different types of genetic variation, with the observation that as much as 90% of causal IBD variants may be non-coding. It was noted that, while casual variants often occurred near the binding sites of master regulators of immune differentiation and stimulus-dependent gene activation, only 10–20% alter a known transcription-factor binding motif [79]. Gaining a more complete understanding of this regulatory code remains an important challenge in both IBD and complex disease genetics more generally.

5. Prospects for translation

5.1. Informing treatment

IBD susceptibility genes have already been shown to have important applications in the development of new treatments. A notable case is the associated locus near SMAD7, which has been shown to reduce the activity of TGF-β1 (an immunosuppressive cytokine) when present at high levels. In a recent phase 2 trial of an oral SMAD7 antisense oligonucleotide, mongersen, Crohn’s disease patients receiving the drug had significantly higher remission rates than those given a placebo [81]. Similarly, the drug Efalizumab targets the product of ITGAL, an integrin αL subunit of lymphocyte function-associated antigen 1 (LFA-1), and has been used to treat psoriasis. A brief, open-label study of Efalizumab for treating Crohn’s disease showed evidence of a clinical response in the majority of subjects [82]. Notably, the effect sizes of these clinically relevant genes are relatively small (Fig. 4), highlighting the importance of continuing to catalogue IBD-associated loci to build up a complete picture of disease pathogenesis and susceptibility.

As well as uncovering potential targets for therapeutic development, identified genetic associations can also prove useful in determining clinical subphenotypes and predicting disease course. For example, in Crohn’s disease, associations have been found between the HLA and colonic CD [83], while NOD2 variants have been shown to predict ileal location and the need for CD-related surgery [84]. Similarly, for ulcerative colitis the HLA is associated with extensive disease and colectomy [85]. Such information can be used to construct individual genetic risk scores, which summarize predictions about disease risk and likely progression based on a patient’s specific genetic profile. Techniques like this can then help to identify misdiagnosed patients and drive more personalized treatment approaches.

5.2. Environmental factors: The microbiome

Another area that offers particular promise for the translation of genetic findings into clinical practice is investigation into the interaction between an individual’s genome and their environment. In the case of IBD, loci identified to date have provided strong evidence of a role for the gut microbiota in disease pathogenesis, with the epithelial barrier and autophagy pathways repeatedly implicated [86]. Microbiome studies in IBD have shown there are distinct differences in the composition of the gut flora in diseased and healthy individuals, such as a decrease in bacteroides, firmicutes, ruminococcaceae and bifidobacterium, and an increase in the presence of Escherichia coli and fusobacterium [87]. However, cause and effect are difficult to disentangle: did the disturbed microbiome arise as a result of the extensive inflammatory response, or did it trigger it? The effects of therapeutics on the intestinal environment further complicate such questions, as treatments such as antibiotics are known to affect the gut microbial community [88,89]. Finally, even amongst healthy individuals the precise composition of the microbiome is extremely sensitive to diet and other unknown environmental factors: family observations show that sharing both genetics and a living space is no guarantee of a completely shared microbiome, and even within the same individual temporal variations are observed [90].

The importance of understanding the role of the microbiome is reflected in the recent success of fecal microbiota transplants (FMTs) as a treatment for inflammatory bowel disease. FMTs aim to reduce dysbiosis in the bowel by modifying the microbiome using stool from a healthy donor. Although the idea was first introduced by Eiseman et al., in 1958 to treat pseudomembranous enterocolitis [91], it has only recently gained popular attention from the IBD community: An initial study by Suskind et al. (2015) showed temporary remission in seven of nine patients, and more extended remission in five of those cases. Efficacy of the FMT depended on whether it successfully engrafted or not, and on how similar the recipient’s original microbiome was to the donor one [92]. Despite this early success, further clinical studies are required to properly evaluate the safety and efficacy of this method.

6. Summary

For more than twenty years the study of genetics has held promise for understanding the causes of IBD. Over that span it became clear that its substantial heritability would not be explained by a handful of heavy hitters, like NOD2, but by the cumulative contribution of hundreds of genetic variants of small effect. While this means the prospect of genetic diagnosis of IBD (or indeed any complex auto-inflammatory disorder) is distant, the accumulation of these genetic clues has steadily improved our understanding of disease biology. The overlap of risk loci between CD and UC underscores their shared pathways and points to a subtle molecular classification of IBD that may only be partially represented by current clinical definitions. The factors that separate the two disorders, likely involving distinct genetic risk factors, tissue specific stimuli or differential response to environmental triggers remain unresolved. More broadly, genetics offers insight into the relationship of IBD to other immune and inflammatory disorders, both common and rare, including surprises such as the overlap with leprosy that may point to the evolutionary history of genes that must strike a balance between defence and self-damage. As the biological pathways implicated by genetics become clearer, and new studies inevitably lead to new discoveries of rarer risk alleles, the progress in understanding IBD genetics will provide the foundation for progress in developing IBD therapies.

Acknowledgements

KMdL is supported by a Woolf Fisher Trust scholarship. KMdL and JCB are supported by Wellcome Trust grant WT098051.

References


