Best Practice & Research Clinical Gastroenterology 28 (2014) 373–386

Contents lists available at ScienceDirect

Best Practice & Research Clinical Gastroenterology

2

Genetic studies of Crohn's disease: Past, present and future

Jimmy Z. Liu, BSc, PhD Student,
Carl A. Anderson, PhD, Group Leader *

The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK

Keywords:
Crohn's
Genetics
Genomics
Genotyping
Sequencing

A B S T R A C T

The exact aetiology of Crohn's disease is unknown, though it is clear from early epidemiological studies that a combination of genetic and environmental risk factors contributes to an individual's disease susceptibility. Here, we review the history of gene-mapping studies of Crohn's disease, from the linkage-based studies that first implicated the *NOD2* locus, through to modern-day genome-wide association studies that have discovered over 140 loci associated with Crohn's disease and yielded novel insights into the biological pathways underlying pathogenesis. We describe on-going and future gene-mapping studies that utilise next generation sequencing technology to pinpoint causal variants and identify rare genetic variation underlying Crohn's disease risk. We comment on the utility of genetic markers for predicting an individual's disease risk and discuss their potential for identifying novel drug targets and influencing disease management. Finally, we describe how these studies have shaped and continue to shape our understanding of the genetic architecture of Crohn's disease.

Introduction

Crohn's disease is a chronic inflammatory disease of the gastrointestinal tract affecting 26–200 per 100,000 in European populations [1]. Along with ulcerative colitis, it is one of the two major forms of inflammatory bowel disease (IBD). The exact causes of Crohn's disease are unknown, though it is likely

* Corresponding author.
E-mail address: carl.anderson@sanger.ac.uk (C.A. Anderson).

http://dx.doi.org/10.1016/j.bpg.2014.04.009

to involve a disrupted immunological response to gut microbiota in genetically susceptible individuals [2]. There is currently no known cure and disease is managed by a combination of immune-suppressing medications, dietary changes or surgery.

Family and twin studies

Epidemiological observations in the 1930s first suggested an inherited component to Crohn's disease risk. Familial clustering showed that 2–14% of patients have a family history of CD [3], while estimates of the sibling relative risk ratio (the ratio of disease risk among siblings of patients compared with that in the general population, i.e. the population prevalence) ranged from 15–42 [3]. The variation in these estimates highlights the difficulty in obtaining accurate heritability measures for relatively rare diseases such as Crohn's. Confounders also include inconsistent study design (e.g. only counting first degree relatives rather than all relatives), sample selection bias (e.g. using hospital cases that are likely to have a more severe form of the disease than the general Crohn's population), and variation in disease prevalence rates, both between different populations and over time [3–7]. Moreover, these observations do not in themselves suggest a role for genetics in disease risk because familial resemblance can also be a consequence of shared environmental factors.

Twin studies have now provided compelling evidence for a significant genetic component to Crohn's disease risk. The twin design assumes that the environmental component to phenotypic variation is the same between monozygotic (MZ) and dizygotic (DZ) twins, and thus the difference in disease concordance rates between sets of MZ and DZ twin pairs can be used to estimate the additive genetic, shared environmental and unique environmental components of disease risk. A meta-analysis of six twin studies with a combined set of 112 MZ and 196 DZ twin pairs reported concordance rates of 30.3% and 3.6% respectively [8], indicating that a large component of Crohn's disease risk is indeed genetic. Together, these family and twin studies provided the motivation for the first wave of gene-mapping studies throughout the mid-1990s aimed at identifying the regions of the genome that contribute to Crohn's disease risk.

Linkage studies

A linkage study identifies regions of the human genome underlying disease susceptibility by testing a series of marker alleles for cosegregation (linkage) with disease status across a number of families (or a single large family with multiple affected members). Owing to the large size of chromosomal segments segregating within a typical family, around 300 evenly distributed microsatellite markers are usually sufficient to capture the majority of positions where the chromosomes of the parents crossed over during meiosis (recombination events). The evidence for linkage in a region is evaluated by metrics such as a LOD (logarithm of odds) score, which compares the probability that the genotyped marker and the hypothetical disease locus are inherited together in the observed data versus the probability of observing the cosegregation pattern purely by chance. A typical linkage study will report all loci with LOD scores greater than three, which corresponds to the data being 1000 times more likely to arise due to cosegregation with disease than by chance [9]. By the mid-nineties, linkage studies had proven to be a robust means of identifying highly penetrant loci underlying monogenic disease such as cystic fibrosis [10] and Huntington's disease [11] and the utility of the method for mapping complex disease loci was increasingly being explored. From 1996 to 2004, 11 linkage studies were performed for Crohn's disease (reviewed here [7]), the largest of which was a meta-analysis consisting of 1068 affected relative pairs [12].

The first Crohn's disease linkage study in 1996 identified a significant disease susceptibility locus on chromosome 16 (dubbed IBD1) [13]. This result was confirmed in subsequent studies [14–20] and in 2001 the specific causal mutations that underlie risk were localised to three low frequency coding variants (R702W, G908R and L1007fs) within the *NOD2* gene (at that time, also known as *CARD15*) [21–25]. These three variants individually had odds ratios (ORs) of 2–4 in heterozygotes and 20–40 for homozygotes, and at least one mutation was present in 30–40% of Crohn's disease cases compared with 6–7% in European controls [7].

Spurred on by the discovery of *NOD2*, additional linkage studies of Crohn's disease (and other common complex diseases) were undertaken; The results of these studies were largely disappointing,

with few loci being consistently replicated [7]. This lack of success suggested that complex diseases, in contrast to Mendelian diseases, were unlikely to be driven by the highly penetrant risk loci that linkage is well powered to detect. In 1996 a seminal paper was published in Science proposing that complex diseases are underpinned by common variants of modest effect [26]. The authors demonstrated that, for a risk allele of 50% frequency and OR of 1.5, around 18,000 affected sib-pairs would be needed to detect the locus via linkage. In contrast, they reported that less than 1000 trios would be needed to detect such a locus adopting the transmission/disequilibrium association test of Spielman et al. [27]. Technological limitations at the time restricted the immediate uptake of the association study design; such studies require that a causal variant (or another variant in high linkage disequilibrium (LD) to the causal variant) is directly genotyped in order to detect a significant signal of association.

Candidate gene association studies

While it was infeasible to test for association at markers across the entire genome, markers within individual genes were often tested for association. Genes were selected based on *a priori* knowledge of biological function or because the lay within a region implicated through linkage analysis. These candidate gene studies typically involved genotyping a set of markers within a gene of interest in a sample of disease cases and controls, and testing for statistically significant differences in allele frequencies between the two groups. Other study designs such as transmission disequilibrium tests in parent-offspring trios were also often used.

Results from the majority of candidate gene studies of Crohn's disease were disappointing, with initial findings often failing to replicate in subsequent studies [28–31]. A combination of small sample sizes, false-positive association, publication bias and failure to account for multiple comparisons meant that as many as 95% of findings from candidate gene studies of complex traits during this era were false [32,33]. In some cases, the lack of power in these studies meant that variants in genes that later became established risk loci were missed altogether (for instance, *IL10* [34,35]). Ultimately however, it would take a combination of technological advances and a greater appreciation of the need for much larger sample sizes to make the identification of bona fide risk loci routine.

Genome-wide association studies

In the early 2000s, along with the closing phases of Human Genome Project, concurrent efforts were underway to gauge the extent of human genetic variation at the population level. Projects such as the SNP Consortium and dbSNP had catalogued over 1.4 million single nucleotide polymorphisms (SNPs) by 2001 [36,37]. It was found that common SNPs in physical proximity formed LD blocks punctuated by hotspots of recombination [38]. These correlation patterns were further characterised through the International Hapmap Project, which by 2007 had identified a further 3.1 million SNPs across 270 individuals from three distinct ancestry groups [39]. At the same time, technological advances in microarray technologies made possible the cost-effective genotyping of hundreds of thousands of SNPs spread throughout the genome [40]. The patterns of LD meant that these arrays could effectively survey the majority of common genetic variation in a population by directly genotyping only a fraction of the total number of variants in the genome. In Europeans and East Asians, around five million common SNPs (those with minor allele frequency greater than 5%) can be almost entirely tagged by a selection of approximately 500,000 SNPs [41,42]. Together, these advances paved the way for researchers to perform genome-wide association studies (GWAS) in order to identify loci associated with complex traits or disease risk.

Genome-wide association studies typically look for statistically significant differences in allele (or genotype) frequencies between a large number of diseased individuals and population controls across hundreds of thousands of SNPs spread throughout the genome. The SNPs that show significant association with disease status point to regions of the genome likely to harbour disease relevant genes. Unlike linkage studies, GWAS are not restricted to sibling pairs and families, and also have generally greater statistical power to detect associated loci of small to moderate effect sizes (Fig. 1) [26]. Due to patterns of LD, SNPs that are associated with disease are unlikely to be the true causal variant, but rather are correlated with ('tag') an untyped causal variant. In addition, genotypes at SNPs that were

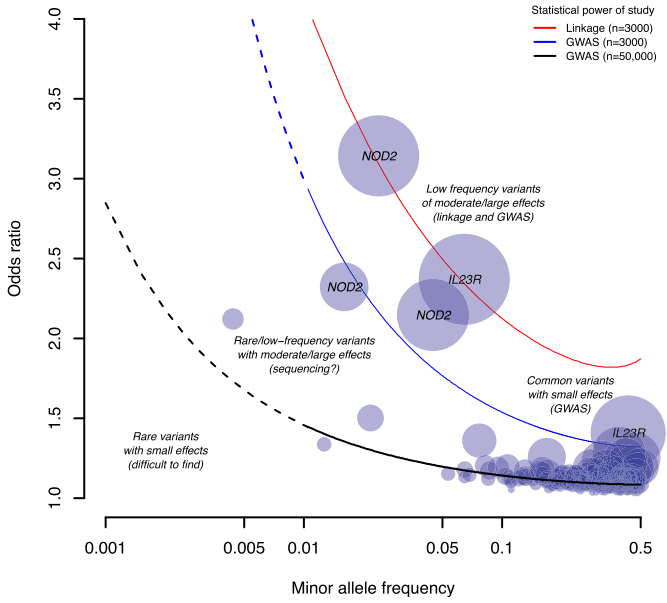


Fig. 1. The genetic architecture of Crohn's disease. Known Crohn's disease risk variants are plotted according to their minor allele frequency and odds ratio (OR) [59]. The size of the circles represents the amount of variance in Crohn's disease liability explained by that variant. The red, blue and black lines represent the minimum OR and allele frequency for a locus for which a linkage study with 3000 individuals, GWAS with 3000 individuals and GWAS with 50,000 individuals respectively will have >80% statistical power to detect [26,108]. *P*-value thresholds for power calculations were set to $P < 10^{-4}$ for linkage and $P < 5 \times 10^{-8}$ for GWAS. The dashed lines represents the allele frequency spectrum of variants that are typically poorly captured on GWAS microarrays (minor allele frequencies less than 1%).

not directed assayed can be inferred through imputation algorithms [43,44] based on the genotypes from a representative reference set of haplotypes [39,45,46], allowing for individual studies using different genotyping platforms to be effectively combined into meta-analyses.

The first Crohn's disease GWAS was conducted in a Japanese population in 2005, and identified *TNFSF15* as a susceptibility locus [47]. This was followed by a rush of studies from 2006 to 2008 [48–55], each including approximately 500–2000 Crohn's disease cases and a similar number of controls genotyped at 100,000–600,000 SNPs. Unlike linkage studies, the development of standardised quality control protocols along with strict statistical criteria for claiming association and replication [55,56] meant the vast majority of SNPs that achieved genome-wide statistical significance (association *p*-value $<5 \times 10^{-8}$) were true positives. The genes and pathways identified by these early GWAS provided many insights into the biological processes underlying Crohn's disease. Most notably, associations at *ATG16L1* and *IRGM* first suggested a role for autophagy in disease pathogenesis [2,48,52]. Other genes involved in both the innate (*TLR4*, *CARD9*, *IL23R*, *STAT3*) and adaptive immune system (*HLA*, *TNFSF15*, *IRF5*, *PTPN22*) pathways were also implicated [57]. GWAS have also shed light on the genetic overlap between Crohn's and other immune-related diseases. Around 30% of associated variants in these initial studies were shared with ulcerative colitis, while close to 50% of loci are shared with at least one other immune-mediated disease such as type 1 diabetes, coeliac disease or rheumatoid arthritis [58]. Unlike many of these diseases, genes in the human leucocyte antigen (HLA) region only confer a modest effect on Crohn's disease risk (ORs 1.1–1.2). This is in contrast to ulcerative colitis, where several variants in *HLA-B* make the largest contribution to genetic risk (ORs 1.4–1.5) [59].

These early GWAS showed that, with the exception of *NOD2*, the typical effect size of a Crohn's susceptibility locus was modest (OR < 1.3), such that the loci identified only explain a fraction of the known genetic component of Crohn's disease risk (highlighting the 'missing heritability problem'

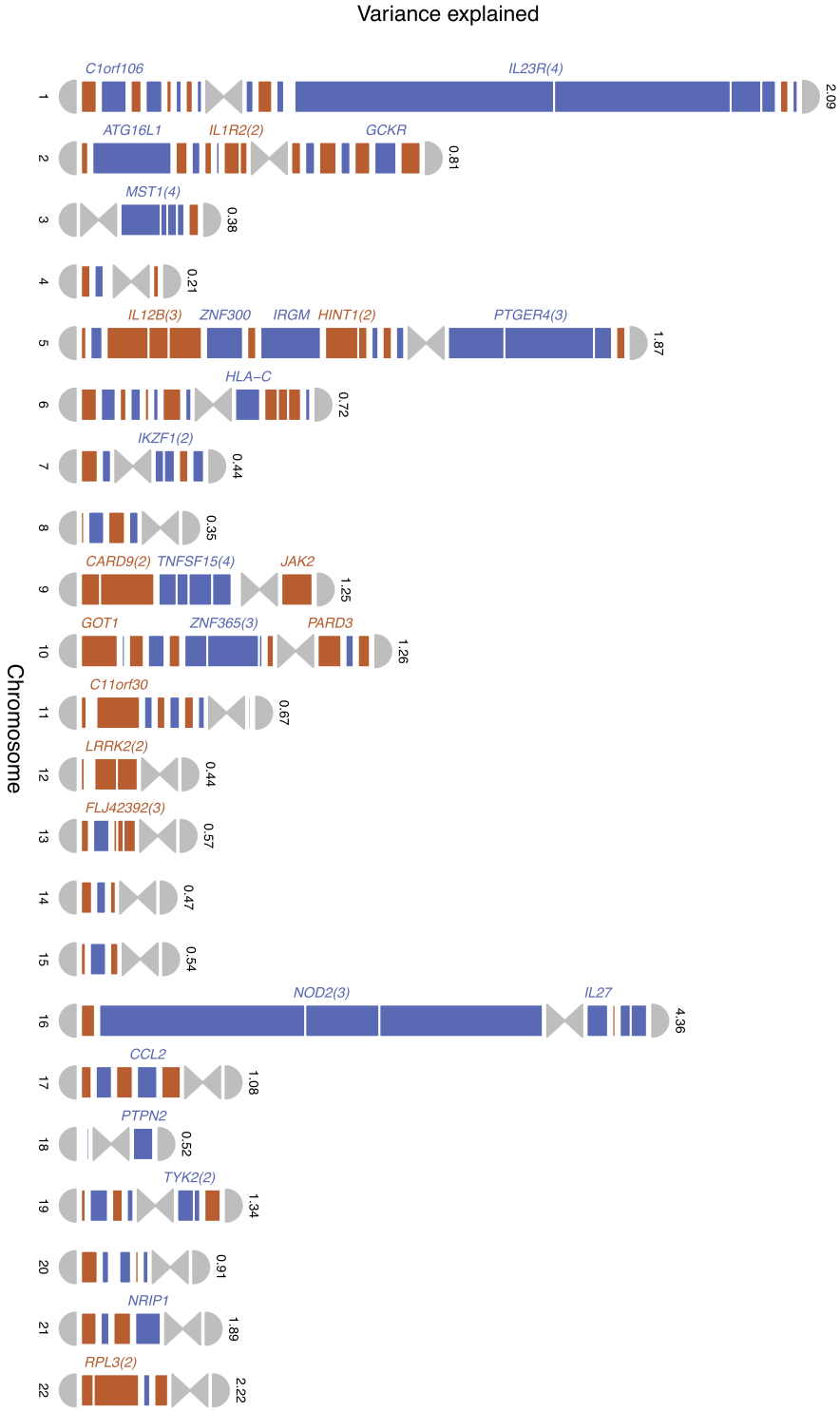
[60,61]). While it is likely that a proportion of this missing heritability is due to rare (minor allele frequency less than 1%) and structural variants that are not well-captured on the current generation of GWAS microarrays, a substantial number of common variants will have even smaller effects than those identified, requiring much larger sample sizes to detect [62]. Indeed, for Crohn's disease, it has been estimated that 22% of the differences in individual disease risk seen in the population (variance in disease liability) can be explained by common variants tagged on microarrays [63] – more than double that explained by known risk loci [64].

An appreciation of the need for larger sample sizes led to the creation of the International IBD Genetics Consortium (IIBDGC) (http://www.ibdgenetics.org/) to bring together investigators and GWAS datasets from IBD genetics groups around the world. From 2008 to 2012 the IIBDGC published three genome-wide association study meta-analyses [59,64,65]. The first of these in 2008 combined data for ~13,000 individuals from three previously published GWAS and identified 21 new Crohn's susceptibility loci [64]. This was followed two years later by a meta-analysis of six GWAS with a total sample size of ~50,000 individuals where 30 new loci were identified, bringing the total count to 71 [65]. The most recent meta-analysis in 2012 included 75,000 individuals and doubled the number of known Crohn's susceptibility loci to 140 (Fig. 2) [59]. Along with the 23 loci associated with ulcerative colitis, the total number of 163 inflammatory bowel disease associated loci represents the most for any complex disease to date. These loci were enriched for genes involved in primary immunodeficiencies and this enrichment was even more striking for genes harbouring Mendelian susceptibility to mycobacterial disease (MSMD) risk variants, where six of the eight genes linked to MSMD overlap with IBD. Similarly, seven of the eight genes known to be associated with leprosy are also shared with IBD, and altogether, 66 IBD loci are shared with other immune-mediated diseases. These overlaps suggest that selection pressures driven by mycobacterial infection may have shaped the genetic architecture of Crohn's disease.

The results of Jostins et al. [59] also highlighted the role of noncoding variation in disease risk. Many of these variants are likely to affect the amount that a gene is expressed (gene regulation) rather than code the protein product. Only nine of the 140 associated loci exclusively harbour variants within coding regions of genes (*IL23R*, *GPR35*, *CD6*, *MUC19*, *GPR65*, *ZNF831*, *ADAM30*, *NOD2*, *FUT2*) while an additional 13 have variants encompassing both coding and noncoding regions (*UBQLN4*, *ITLN1*, *FCGRA2A*, *MST1/BSN*, *SLC22A4*, *REV3L*, *CARD9*, *ZPBP2/GSDMB*, *TUBD1*, *CD226*, *YDJC*, *ATG16L1*, *LACC1*). Altogether, 51 loci overlap variants with a known effect on gene expression (eQTLs – expression quantitative trait loci; Fig. 3). Most studies that detect eQTLs have been limited to only a few hundred individuals in only a small number of cell types (liver, brain, fibroblasts, monocytes, T cells and lymphoblastoid cell lines) [66,67]. The overlap between eQTLs and disease-associated variants will increase as more eQTL studies are performed across larger samples sizes and across different cell types, especially those involved in the immune system in the case of Crohn's disease.

Targeted genotype arrays

A feature of the latest IIBDGC meta-analysis was the use of the Immunochip custom genotyping array for replicating signals identified in the original GWAS meta-analysis. The Immunochip was designed after the first wave of GWAS meta-analyses to aid in the replication, fine-mapping and discovery of loci associated with inflammatory and autoimmune diseases [68]. To take advantage of the pervasive genetic overlap between many of these diseases, the Immunochip contains a dense panel of ~130,000 SNPs located in 186 regions with known association with one or more of 12 immune-related diseases. SNPs within the regions were ascertained via dbSNP, the 1000 Genomes project (February 2010 release), and autoimmune disease resequencing projects. While not all SNPs passed the Illumina design process and made it onto the microarray, the Immunochip provides unprecedented coverage of common, low-frequency and rare variants across these 186 genomic regions. A further 50,000 SNPs that were suggestively significant in the original GWAS studies were also included. This panel served as the replication set of SNPs in Jostins et al., where over 40,000 IBD cases and controls were genotyped. The cost-effectiveness of the Immunochip (at ~20% that of a GWAS microarray at the time) allows for studies with much larger sample sizes than GWAS and also enables powerful disease subphenotype and cross-disease comparisons [69].


Variance explained
Chromosome
1
C1orf106
IL23R(4)
2.09
2
ATG16L1
IL1R2(2)
GCKR
0.81
3
MST1(4)
0.38
4
0.21
5
IL12B(3)
ZNF300
IRGM
HINT1(2)
PTGER4(3)
1.87
6
HLA−C
0.72
7
IKZF1(2)
0.44
8
0.35
9
CARD9(2)
TNFSF15(4)
JAK2
1.25
10
GOT1
ZNF365(3)
PARD3
1.26
11
C11orf30
0.67
12
LRRK2(2)
0.44
13
FLJ42392(3)
0.57
14
0.47
15
0.54
16
NOD2(3)
IL27
4.36
17
CCL2
1.08
18
PTPN2
0.52
19
TYK2(2)
1.34
20
0.91
21
NRIP1
1.89
22
RPL3(2)
2.22

Fine-mapping associated loci

The causal variants that underlie the majority of loci discovered through GWAS remain unidentified. An associated locus will often consist of dozens of correlated SNPs in high LD spanning across many genes, with very similar association signals. In the 140 loci associated with Crohn's risk, the number of SNPs that are tagged ($r^2 > 0.8$) by the main GWAS SNP range from 1 to 306 per locus (median 13). Narrowing these down to a single causal variant is difficult and will initially require a combination of many complementary approaches. Firstly, much larger sample sizes will be required to differentiate statistical signals at causal variants over their highly correlated neighbours. Secondly, as patterns of LD differ between different ancestral groups, obtaining samples from multiple populations can narrow the associated region for risk loci that are shared across populations. Thirdly, combining functional genetic information with association results allows variants with relevant annotations to be up-weighted in association analyses. Data from projects such as ENCODE [70] and GTEx [71] provide rich functional genomic information that can be readily integrated with GWAS results. In addition to providing functional candidates, these functional annotations can also uncover biological mechanisms through which variants act, either through the specific cell type or functional element [72–74]. Under the auspices of the IIBDGC, efforts to fine-map associated loci using the Immunochip are underway, along with transethnic studies of IBD across European and non-European populations. Ultimately, the direct modelling of these variants in cell lines and model organisms may be required for final confirmation of causality. Emerging technologies such as DNA editing through CRISPR and engineering induced pluripotent stem cells are likely to play an important role [75–77].

The *IRGM* locus exemplifies some of the challenges in identifying causal variants. The SNP initially associated with disease was later found to be in perfect LD with a 20 kb deletion upstream of *IRGM* [52,78]. This deletion was thought to be causal because it affects the expression of *IRGM*, which in turn regulates the efficiency of autophagy. A later study showed, however, that this deletion is one of several highly correlated Crohn's disease associated variants in the region that affect *IRGM* expression, none of which can reasonably be ruled out as causal [79]. Furthermore, the variants are also not associated with Crohn's disease in the Japanese population, suggesting either European-specific gene-environment interactions or the presence of an untyped causal variant that arose after the European-Asian population split [79].

Next generation sequencing

The role of rare variants in complex diseases is currently an important area of focus in human genetics. High-throughput discovery and accurate genotyping of rare variants has recently been made feasible through large reductions in the cost of next-generation sequencing. Often cited as a possible explanation for missing heritability, rare variants are in theory likely to have much larger effect sizes than common variants due to purifying selection maintaining damaging alleles at low frequencies [61]. Indeed, loci that are associated with complex disease are enriched for rare variants that cause known Mendelian disorders and it has been suggested that recessive variants confer risk to related complex diseases when the carrier is heterozygote [80]. Independent rare variant associations are also often found in genes with known common associated variants [81–83].

Since the rare allele of individual rare variants are observed so infrequently, single variant tests of association will be underpowered for all but the most highly penetrant alleles. For instance, for an allele that doubles disease risk (OR = 2) and has a frequency of 0.1%, nearly 60,000 cases and a similar number of controls will be required for the variant to reach genome-wide significance. To increase power to detect association, rare variants are often aggregated based on characteristics such as their

Fig. 2. The Crohn's disease genome. Known Crohn's disease risk loci are shown according to their location on the long or short arms of chromosomes. The size of each locus indicates the proportion of variance in Crohn's disease liability explained by that locus [59]. Several notable genes are marked. Parentheses next to gene names denote the number of independent risk variants within the locus. The number above each chromosome is the ratio of the total amount of variance explained by that chromosome vs the expected number given the chromosome's size. Sex chromosomes were excluded as no loci have been conclusively implicated, largely due to these chromosomes being overlooked from most GWAS.

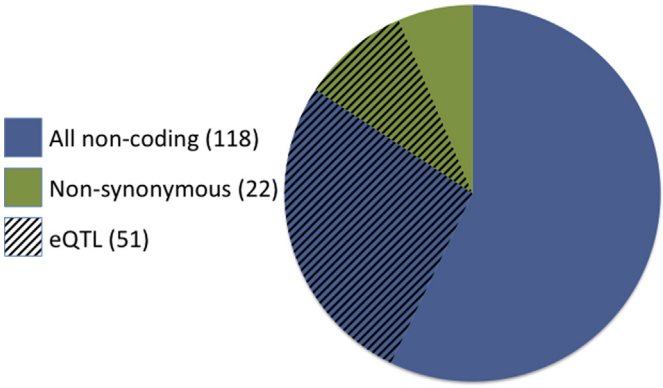


Fig. 3. Proportion of noncoding, nonsynonymous and known eQTLs among known 140 Crohn's disease risk loci [59]. A locus is labelled as nonsynonymous if the lead SNP is in high linkage disequilibrium (LD; $r^2 > 0.8$) with a nonsense or missense mutation. Similarly, loci with eQTLs were marked if the lead SNP is in high LD with a known eQTL in studies of liver, brain, fibroblasts, monocytes, T cells and lymphoblastoid cell lines [66,67].

position within genes, functional features and allele frequencies [84]. Dozens of these burden tests have been proposed [84–87] along with methods for meta-analysis and replication [88–90].

The degree to which such variants contribute to Crohn's disease heritability is unclear, and the results from early large scale sequencing studies targeted at known susceptibility genes have been disappointing [81,91,92]. These studies typically involved sequencing the coding regions of several candidate genes in a few hundred cases and controls followed by the direct genotyping of putatively associated variants in a much larger replication cohort. Coding regions are targeted because the functional consequences of variants in these regions are much better understood than those in non-coding parts of the genome. In addition, coding variants are hypothesized to have larger effect sizes given their direct impact on protein product and are generally more evolutionarily conserved than noncoding variants [93]. Momozawa et al. [81] initially sequenced 63 candidate genes in 112 Crohn's disease cases and 112 controls with replication in an additional 288 to 928 cases and 288 to 1216 controls, and identified four independent associations in *IL23R*, although only one of these exceeded genome-wide significance. Similarly, Rivas et al. [92] sequenced 56 genes in 350 cases and 350 controls with follow-up genotyping in 16,054 cases and 17,575 controls, and identified 12 independent rare variant associations across seven genes, of which two (coding variants in *NOD2* and *CARD9*) exceeded genome-wide significance. These three genome-wide significant variants were included on the Immunochip and subsequently confirmed in Jostins et al. [59] using around 75,000 samples. However, a recent sequencing study of 25 candidate genes across 41,911 individuals, of which 3271 were Crohn's disease cases, failed to identify any novel associations [91]. A natural extension for candidate gene sequencing studies is to sequence the entire exome of cases and controls. A recent exome sequencing study in 42 Crohn's cases with follow-up genotyping in 9348 cases and 14,567 controls found suggestive rare variant associations in *PRDM1* [94]. Again, the variant failed to reach genome-wide significance and other whole exome studies with much larger sample sizes are currently underway.

The sobering results from these studies highlight the challenges in rare variant association studies. As it is currently not economically feasible to perform high coverage whole-genome sequencing in a large number of cases and controls, compromises often need to be made in terms of the number of genomic regions covered and the number of individuals. The majority of loci identified in GWAS lie in noncoding regions (Fig. 3), which have been overlooked by the current generation of sequencing studies. A large number of rare noncoding variants will play a role in gene regulation, though it remains to be seen whether their effects are large enough to be a major contributor to disease susceptibility. Performing burden tests across rare variants in regulatory regions such as promoters and enhancers may show promise. Most importantly, the sample sizes used in these sequencing studies have thus far simply been insufficient to robustly identify rare variant associations. Under certain assumptions about

the effect size distribution of rare variants and selection pressures, cohorts of more than 25,000 cases may be required in order to find these signals, along with an equally large number for replication [95].

In addition to candidate gene and whole-exome sequencing, the next few years will also see the emergence of low-coverage whole-genome sequencing studies. While sequencing individuals at deep coverage (>30×) is required to obtain accurate individual genotype calls, low coverage sequencing (less than 6×) and jointly calling shared (non-private) variants across many thousands of individuals is a cost effective method for discovering rare variants. For instance, for a SNP with frequency 0.2% to be discovered, over 2000 individuals need to be sequenced at 30× coverage (60,000 genomes). In contrast, the same SNP can be identified in ~3000 individuals sequenced at 4× (12,000 genomes) – a five-fold reduction in sequencing cost [96] and, with more sequenced individuals, greater power to detect associations. The obvious advantage of these study designs over targeted or whole-exome sequencing is that they survey the entire genome rather than individual regions. Additionally, large cohorts of sequenced individuals can be used as reference panels to impute rare variants into new and existing GWAS datasets at much greater accuracy than existing panels. Over the course of 2014–15, it is expected that over 30,000 individuals, of which ~5000 are IBD cases, will be sequenced at low-coverage. Imputing the millions of new variants discovered from this set into ~25,000 Crohn's disease cases (of which ~15,000 have already been genotyped as part of GWAS) along with sufficient replication will, for the first time, enable studies with sufficient power to begin detecting associations at SNPs with frequencies in the order of 0.1–1% and ORs of 2–3 (Fig. 2).

Genetic prediction

In addition to gaining a better understanding of disease biology, genetic information can also potentially be used for disease risk prediction. Prediction methods for complex diseases typically involve assigning a risk score to an individual based on their genotypes and previously estimated effect sizes (for instance, ORs from GWAS) across risk alleles. Risk alleles can be assigned not only based on known associations, but also include nominally associated variants. Prediction accuracy can be evaluated by methods such as the receiver operating characteristic curve (ROC), which estimates the true and false positive rates of the predictor at various risk score cut-offs [97]. The area under the ROC (AUC) is the probability that for a randomly selected pair of diseased and healthy individuals, the diseased individual will have a higher risk score. An AUC of 0.5 means that the prediction method is no better than chance, while a value of one means that the method perfectly discriminates between diseased and healthy individuals.

In Crohn's disease, genetic risk prediction is still in its infancy and does not currently offer much in terms of clinical utility. Estimates of AUC using just family history of disease, genetic risk loci or the two together range from 0.56 to 0.74 [98,99]. Including risk factors such as smoking and age into the risk model may improve the AUC. Nevertheless, given its high heritability, the theoretical maximum possible AUC assuming that all Crohn's disease risk loci have been identified and effect sizes are accurately measured is estimated to lie between 0.96 and 0.98 [100,101]. However, while this figure seems high, the utility of genetic prediction is limited given the low prevalence of Crohn's disease. Even assuming a generous disease prevalence estimate of 1% and AUC of 0.98, less than 12% of individuals who test positive (using a sensitivity cut-off of 0.93) will develop disease [100]. Increasing the threshold will increase the proportion of positively identified individuals but also exclude a higher number of cases from being identified. Genetic prediction may offer better value in existing patients through informing best course of treatment, though greater knowledge of how risk loci affect these subphenotypes will be required [102].

Crohn's disease subphenotypes

Genetic studies have begun to shed light on loci associated with Crohn's disease subphenotypes such as disease location and clinical course. These studies often focus on one or more candidate genes with a subphenotype of interest [102–104] or can survey the whole genome in the same way as a GWAS [105]. One of the strongest signals associated with clinical outcome has been coding variants in *NOD2*, which are strongly predictive for ileal disease, stenosis, fistula and Crohn's related surgery [102]. It has been

suggested that patients with these *NOD2* mutations respond poorly to bacterial antigens [106]. A non-coding variant in *FOXO3A* was also recently implicated with Crohn's disease prognosis [104]. The minor allele was found to be significantly more common in indolent patients (defined as having disease for longer than four years but with no immunomodular or intestinal resections required) than patients with frequent flaring or a complicated course of treatment (where two or more immunomodular therapies/intestinal resections were required). Despite not being associated with Crohn's disease susceptibility itself, *FOXO3A* variants are thought to affect disease outcome via regulation of IL7 and IL2 signalling pathways, whose expression patterns correlate with clinical course of autoimmune diseases [107]. The study demonstrated how integrating GWAS data into functional genomic, model organism and clinical information can uncover basic biological pathways that are associated with disease outcome.

The genetic architecture of Crohn's disease

Putting together the results from linkage, genome-wide association and sequencing studies, the genetic architecture of Crohn's disease represents that of a typical multifactorial complex trait where a combination of multiple genes, along with the environment, lead to disease. With few exceptions, individual risk loci confer only a modest effect on disease susceptibility and together, the known loci explain ~13% of variation in disease liability [59]. The majority of the genetic contribution to disease risk remains to be explained. Nevertheless, it is perhaps safe to say that that nearly all variants with frequency greater than 5% and ORs greater than 1.2 in individuals of European ancestry have been identified and the remaining genetic contribution will arise from a combination of common variants with ever-smaller effect sizes and rare variants [62]. In addition, all variants with large effects ($OR > 3$) and frequency greater than 1% have also been uncovered by GWAS and linkage studies (Fig. 1). Future sequencing studies will shed light on the exact effect size distribution of rare variants. Finally, it should be emphasised that locus discovery is not an end in itself. Challenges remain in taking what we've learned from genetic studies to build more complete models of disease pathogenesis and ultimately translating these into better patient outcomes.

Practice points

- Early familial observations and twin studies demonstrated that there is a significant genetic contribution to Crohns disease risk.
- Identifying the specific genes responsible for disease will provide insights into disease biology and potential therapeutic targets.
- The discovery of NOD2 variants demonstrated the potential of linkage and candidate gene studies to identify risk loci, though subsequent efforts were largely unsuccessful.
- Since 2006, genome-wide association studies have established over 140 loci associated with Crohns disease risk.

Research agenda

- Fine-mapping studies and direct experimental work will be required to identify the causal variants and biological mechanisms that underlie Crohns disease risk loci.
- Whole-genome sequencing studies will help elucidate the contribution of rare and low-frequency variants to disease risk.

Acknowledgements

J.Z.L. and C.A.A. are supported by a grant from the Wellcome Trust (098051).

References

[1] Loftus Jr EV. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology 2004;126:1504–17.
[2] Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature 2011;474:307–17.
[3] Halme L, Paavola-Sakki P, Turunen U, Lappalainen M, Farkkila M, Kontula K. Family and twin studies in inflammatory bowel disease. World J Gastroenterol 2006;12:3668–72.
[4] Shivananda S, Lennard-Jones J, Logan R, Fear N, Price A, Carpenter L, et al. Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). Gut 1996;39:690–7.
[5] Hiatt RA, Kaufman L. Epidemiology of inflammatory bowel disease in a defined northern California population. West J Med 1988;149:541–6.
[6] Farrokhyar F, Swarbrick ET, Irvine EJ. A critical review of epidemiological studies in inflammatory bowel disease. Scand J Gastroenterol 2001;36:2–15.
[7] Mathew CG, Lewis CM. Genetics of inflammatory bowel disease: progress and prospects. *Hum Mol Genet*. 13 Spec No 2004;1:R161–8.
[8] Brant SR. Update on the heritability of inflammatory bowel disease: the importance of twin studies. Inflamm Bowel Dis 2011;17:1–5.
[9] Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–7.
[10] Tsui L, Buchwald M, Barker D, Braman J, Knowlton R, Schumm J, et al. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. Science 1985;230:1054–7.
[11] Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature 1983;306:234–8.
[12] van Heel DA, Fisher SA, Kirby A, Daly MJ, Rioux JD, Lewis CM, et al. Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. Hum Mol Genet 2004;13:763–70.
[13] Hugot JP, Laurent-Puig P, Gower-Rousseau C, Olson JM, Lee JC, Beaugerie L, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature 1996;379:821–3.
[14] Brant SR, Fu Y, Fields CT, Baltazar R, Ravenhill G, Pickles MR, et al. American families with Crohn's disease have strong evidence for linkage to chromosome 16 but not chromosome 12. Gastroenterology 1998;115:1056–61.
[15] Cho JH, Nicolae DL, Gold LH, Fields CT, LaBuda MC, Rohal PM, et al. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: Evidence for epistasis between 1p and IBD1. Proc Natl Acad Sci 1998;95:7502–7.
[16] Mirza MM, Lee J, Teare D, Hugot JP, Laurent-Puig P, Colombel JF, et al. Evidence of linkage of the inflammatory bowel disease susceptibility locus on chromosome 16 (IBD1) to ulcerative colitis. J Med Genet 1998;35:218–21.
[17] Ohmen JD, Yang H-Y, Yamamoto KK, Zhao H-Y, Ma Y, Bentley LG, et al. Susceptibility Locus for Inflammatory Bowel Disease on Chromosome 16 has a Role in Crohn's disease, but Not in Ulcerative Colitis. Hum Mol Gen 1996;5:1679–83.
[18] Curran ME, Lau KF, Hampe J, Schreiber S, Bridger S, Macpherson§ AJS, et al. Genetic analysis of inflammatory bowel disease in a large European cohort supports linkage to chromosomes 12 and 16. Gastroenterology 1998;115:1066–71.
[19] Cavanaugh JA, Callen DF, Wilson SR, Stanford PM, Sraml ME, Gorska M, et al. Analysis of Australian Crohn's disease pedigrees refines the localization for susceptibility to inflammatory bowel disease on chromosome 16. Ann Hum Genet 1998;62:291–8.
[20] Cavanaugh J. International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: crohn disease and chromosome 16. Am J Hum Genet 2001;68:1165–71.
[21] Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cezard J-P, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 2001;411:599–603.
[22] Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 2001;411:603–6.
[23] Cuthbert AP, Fisher SA, Mirza MM, King K, Hampe J, Croucher PJP, et al. The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. Gastroenterology 2002;122:867–74.
[24] Vermeire S, Wild G, Kocher K, Cousineau J, Dufresne L, Bitton A, et al. CARD15 Genetic variation in a Quebec population: prevalence, genotype-phenotype relationship, and haplotype structure. Am J Hum Genet 2002;71:74–83.
[25] Hampe J, Cuthbert A, Croucher PJP, Mirza MM, Mascheretti S, Fisher S, et al. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet 2001;357:1925–8.
[26] Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–7.
[27] Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–16.
[28] Stoll M, Corneliussen B, Costello CM, Waetzig GH, Mellgard B, Koch WA, et al. Genetic variation in DLG5 is associated with inflammatory bowel disease. Nat Genet 2004;36:476–80.
[29] Yamazaki K, Takazoe M, Tanaka T, Ichimori T, Saito S, Iida A, et al. Association analysis of SLC22A4, SLC22A5 and DLG5 in Japanese patients with Crohn disease. J Hum Genet 2004;49:664–8.
[30] Gewirtz AT, Vijay-Kumar M, Brant SR, Duerr RH, Nicolae DL, Cho JH. Dominant-negative TLR5 polymorphism reduces adaptive immune response to flagellin and negatively associates with Crohn's disease. Am J Physiol Gastrointest Liver Physiol 2006;290:G1157–63.
[31] Brand S, Staudinger T, Schnitzler F, Pfennig S, Hofbauer K, Dambacher J, et al. The role of Toll-like receptor 4 Asp299Gly and Thr399Ile polymorphisms and CARD15/NOD2 mutations in the susceptibility and phenotype of Crohn's disease. Inflamm Bowel Dis 2005;11:645–52.
[32] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001;29:306–9.

[33] Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. Lancet 2003;361:865–72.
[34] Castro-Santos P, Suarez A, Lopez-Rivas L, Mozo L, Gutierrez C. TNF[alpha] and IL-10 gene polymorphisms in inflammatory bowel disease. Association of -1082 AA low producer IL-10 genotype with steroid dependency. Am J Gastroenterol 2006; 101:1039–47.
[35] Parkes M, Satsangi J, Jewell D. Contribution of the IL-2 and IL-10 genes to inflammatory bowel disease (IBD) susceptibility. Clin Exp Immunol 1998;113:28–32.
[36] Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 1999;9:677–9.
[37] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928–33.
[38] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science 2004;304:581–4.
[39] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3. 1 million SNPs. Nature 2007;449:851–61.
[40] Syvanen A-C. Toward genome-wide SNP genotyping. Nat Genet; 2005:37.
[41] International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61.
[42] Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet 2006;38:659–62.
[43] Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499–511.
[44] Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387–406.
[45] Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
[46] International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52–8.
[47] Yamazaki K, McGovern D, Ragoussis J, Paolucci M, Butler H, Jewell D, et al. Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. Hum Mol Genet 2005;14:3499–506.
[48] Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 2007;39:207–11.
[49] Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 2006;314:1461–3.
[50] Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, 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.
[51] Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, 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.
[52] Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 2007;39:830–2.
[53] Franke A, Hampe J, Rosenstiel P, Becker C, Wagner F, Häsler R, et al. Systematic association mapping identifies *NELL1* as a Novel IBD Disease Gene. PLoS ONE 2007;2:e691.
[54] Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, 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: 14747–52.
[55] Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.
[56] Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet; 2011. Chapter 1:Unit1 19.
[57] Van Limbergen J, Wilson DC, Satsangi J. The genetics of Crohn's disease. Annu Rev Genomics Hum Genet 2009;10:89–116.
[58] 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.
[59] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119–24.
[60] Maher B. Personal genomes: The case of the missing heritability. Nature 2008;456:18–21.
[61] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–53.
[62] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42:565–9.
[63] Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 2011;88:294–305.
[64] Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–62.
[65] Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118–25.
[66] Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet 2012;8:e1002639.
[67] Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013;501:506–11.
[68] Cortes A, Brown M. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011;13:101.
[69] Parkes M, Cortes A, van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 2013;14:661–73.

[70] Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[71] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5.
[72] Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res 2012;22:1748–59.
[73] Trynka G, Raychaudhuri S. Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases. Curr Opin Genet Dev 2013;23:635–41.
[74] Liu JZ, Almarri MA, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, et al. Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. Nat Genet 2012;44:1137–41.
[75] Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science 2013;339:819–23.
[76] Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science 2013;339:823–6.
[77] Robinton DA, Daley GQ. The promise of induced pluripotent stem cells in research and therapy. Nature 2012;481:295–305.
[78] McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 2008;40:1107–12.
[79] Prescott NJ, Dominy KM, Kubo M, Lewis CM, Fisher SA, Redon R, et al. Independent and population-specific association of risk variants at the IRGM locus with Crohn's disease. Hum Mol Genet 2010;19:1828–39.
[80] Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. Cell 2013;155:70–80.
[81] Momozawa Y, Mni M, Nakamura K, Coppieters W, Almer S, Amininejad L, et al. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet 2011;43:43–7.
[82] Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387–9.
[83] Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, et al. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 2008;40:198–203.
[84] Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 2010;11:773–85.
[85] Asimit J, Zeggini E. Rare Variant Association Analysis Methods for Complex Traits. Ann Rev Genet 2010;44:293–308.
[86] Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genetic Epidemiol 2011;35:606–19.
[87] Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet 2012;44:623–30.
[88] Hu Y-J, Berndt Sonja I, Gustafsson S, Ganna A, Hirschhorn J, North KE, et al. Meta-analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics. Am J Hum Genet 2013;93:236–48.
[89] Lee S, Teslovich Tanya M, Boehnke M, Lin X. General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. Am J Hum Genet 2013;93:42–53.
[90] Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. Nat Genet 2014;46:200–4.
[91] Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 2013;498:232–5.
[92] Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 2011;43:1066–73.
[93] Chen CT, Wang JC, Cohen BA. The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet 2007;80:692–704.
[94] Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, et al. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. Gastroenterology 2013;145:339–47.
[95] Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: Designing rare variant association studies. Proc Natl Acad Sci USA. Published online January 17, http://www.pnas.org/content/early/2014/01/16/1322563111; 2014.
[96] Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 2011;21:940–51.
[97] Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform 2005;38:404–15.
[98] Ruderfer DM, Korn J, Purcell SM. Family-based genetic risk prediction of multifactorial disease. Genome Med 2010;2:2.
[99] Kang J, Kugathasan S, Georges M, Zhao H, Cho JH, Consortium NIG. Improved risk prediction for Crohn's disease with a multi-locus approach. Hum Mol Genet 2011;20:2435–42.
[100] Jostins L, Barrett JC. Genetic risk prediction in complex disease. Hum Mol Genet 2011;20:R182–8.
[101] Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 2010;6:e1000864.
[102] Cleynen I, Gonzalez JR, Figueroa C, Franke A, McGovern D, Bortlik M, et al. Genetic factors conferring an increased susceptibility to develop Crohn's disease also influence disease phenotype: results from the IBDchip European Project. Gut 2013;62:1556–65.
[103] Jurgens M, Brand S, Laubender RP, Seiderer J, Glas J, Wetzke M, et al. The presence of fistulas and NOD2 homozygosity strongly predict intestinal stenosis in Crohn's disease independent of the IL23R genotype. J Gastroenterol 2010;45:721–31.
[104] Lee JC, Espeli M, Anderson CA, Linterman MA, Pocock JM, Williams NJ, et al. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. Cell 2013;155:57–69.

[105] Dubinsky MC, Kugathasan S, Kwon S, Haritunians T, Wrobel I, Wahbeh G, et al. Multidimensional prognostic risk assessment identifies association between IL12B variation and surgery in Crohn's disease. Inflamm Bowel Dis 2013;19:1662–70.
[106] Brand S. Moving the genetics of inflammatory bowel diseases from bench to bedside: first steps towards personalised medicine. Gut 2013;62:1531–3.
[107] McKinney EF, Lyons PA, Carr EJ, Hollis JL, Jayne DR, Willcocks LC, et al. A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. Nat Med 2010;16:586–91. 1p following 591.
[108] Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003;19:149–50.