UEG News

UNITED EUROPEAN GASTROENTEROLOGY ueg journal

United European Gastroenterology Journal
2019, Vol. 7(10) 1426–1430

DOI: 10.1177/2050640619891120
journals.sagepub.com/home/ueg
SAGE

# How to approach understanding complex trait genetics – inflammatory bowel disease as a model complex trait

## Introduction

Monogenic disorders are individually rare, run in families based on classic modes of inheritance termed Mendelian inheritance (autosomal dominant or recessive, X-linked), and are caused by variation within a single gene. The involved variants are rare and typically disrupt protein-coding genes thereby causing disease. Monogenic traits are however rare. Many – if not most – traits are associated with a familial risk, without demonstrating a typical Mendelian inheritance pattern. The lack of a Mendelian inheritance pattern however does not exclude a genetic origin. These common traits are classified as complex – or multifactorial – traits, caused by variation within multiple genes (polygenic) and environmental factors.

Here, we explain how complex trait genetics can be studied, some key concepts, and whether the current findings can be used in clinical decision-making. We will focus on inflammatory bowel disease (IBD), a prototypical complex disease, although the explanations may apply also to other complex diseases such as coeliac disease, primary sclerosing cholangitis, irritable bowel syndrome and sporadic colorectal cancer.

## *Genetic architecture of a complex trait*

In contrast to monogenic disorders which are caused by variants that strongly influence the function or stability of a single protein, the functional effect(s) of the variants involved in complex traits are much more subtle and complicated. The variants can be located in coding sequences, although experience has taught us that the majority are non-coding. Some are known to influence gene expression, stability of the mRNA or protein, or splicing, but for many we currently do not know what their function is, or which gene(s) they influence. What is more, the variants identified in the genetic studies are not necessarily the ones providing the functional link with the underlying biological mechanism, but merely flag genomic regions (loci) where the functionally causal variant(s)/gene(s) are located.

As a result of the subtle functional effect(s), in complex traits there is only weak genotype-phenotype correlation, with low effect sizes for the individually associated genetic variants. All genetic and environmental factors that contribute to a complex trait are collectively described as the 'liability'. Individuals for which the liability score is above a certain threshold value develop the disease, and if the score is below the threshold they will not develop the disease. Hence the term 'liability threshold model' in the context of complex traits.[1] The specific combination of genetic and environmental factors to reach a given threshold can differ among individuals.

Another important concept in understanding complex trait genetics is heritability. This is a measure of how much of the phenotypic variability in a trait is due to differences in peoples' genes. If heritability is zero, trait variation in a given population is fully dependent on environmental factors; if heritability is one, trait variation is only determined genetically and environmental factors play no role (as in monogenic disorders). For complex traits, heritability is somewhere between 0–1, and often estimated from twin studies. Phenotypic resemblance within monozygotic twin pairs and dizygotic pairs is compared based on sharing a complete genome or only half of it. A higher concordance in monozygotic than in dizygotic twins indicates an influence of genetic components.[2] The estimated heritability from twin studies for Crohn's disease (CD) is 75%, for ulcerative colitis it is 67%.[3]

## *How to study genetics of a complex trait?*

Much effort has been put towards finding the genetic determinants of complex traits. At first, researchers turned to methods known for studying monogenic traits which travel predictably through families. They were studied by family-based linkage analysis, in which tracing is through genetic markers segregated together with disease through the family, hereby identifying chromosomal regions carrying the causal gene or mutation. Due to the deterministic relationship between mutation and disease, linkage studies were very effective in identifying that genomic region for monogenic disorders.[4] The same family-based linkage approach was applied to common diseases. The results of these studies were however very disappointing, with only a handful regions identified, including the IBD1 risk locus on chromosome 16 for CD, later fine-mapped to *NOD2* as causal gene.[5,6] In hindsight, these disappointing results are not so surprising or difficult to

understand: in common complex diseases it is not one gene that determines disease, but many genes that exert a small influence.

Thanks to fundamental scientific (such as the Human Genome Project, the HapMap project) and technological (such as microarrays) advances, systematic and large-scale genome-wide approaches have became a possibility. Hundreds of thousands of predefined variants can be screened simultaneously, and analysed for differences in frequency between large sets of cases and controls in genome-wide association studies (GWAs). The vast majority of work goes into collecting samples, and quality control (QC) of the genetic data to make sure all factors that could influence the analysis are removed or corrected for. Best practices for all QC steps have been extensively described elsewhere.[8,9] Associations are tested using as simple a model as possible. Typically a logistic regression model is used, where an additive effect of the risk allele is assumed (0/1/2 coding for the different genotypes: being homozygous for the risk allele gives you double the risk compared to being heterozygous), and correcting for population structure and other confounders if necessary.[9] Since you are testing for so many different variants, a multiple testing correction has to be applied. For GWAs, the threshold for calling an association genome-wide significant is when it reaches a $p$-value $\leq 5 \times 10^{-8}$. Even after applying strict QC measures and significance threshold, false positive associations are possible. Therefore, each reported association has to be validated in an independent dataset.

Since typical effect sizes are small – the majority of associated variants have effect sizes in the order of 1.1–1.57 – huge sample sizes are needed in these studies to achieve statistically significant results. The first individual GWAs were typically done on a few 100–1000 cases and controls and gave rise to a couple of interesting findings; for example the association with genes like IL23R and ATG16L1, thereby implicating a link between innate and adaptive immunity and autophagy in CD.[10–12] Thanks to international collaborations, larger and larger studies were possible in which individual GWAs were analysed collectively in so-called meta-analyses.[13–17] The latest study in IBD included almost 60,000 subjects and brought the counter to over 240 IBD-associated loci.[18] Very similar stories are seen for other complex diseases, although the exact number of as yet identified loci can differ. For example, currently there are 57 loci identified for coeliac disease, and 20 for primary sclerosing cholangitis.[19,20] Together, these genome-wide significant loci explain 20–25% of variance in disease liability for IBD.[7,16] The gap between this percentage and the estimated heritability (from twin studies) is referred to as missing heritability. A number of possible causes could be considered and are discussed in detail elsewhere.[21,22] There are some causes that we would like to highlight here. First, the estimated heritability from twin studies might be overestimated due to difficulties with excluding shared environmental factors. Also, the heritability estimates lack precision, with broad confidence intervals. Second, in GWAs we use arrays that cover only a portion of the genome (e.g. common variants). We thus have not yet considered in much detail low-frequency (minor allele frequency (MAF) between 1–5%) and especially rare variants (MAF <1%). Newer arrays are better for covering lower frequency and rare variants and, with ever bigger sample sizes, future GWAs will allow identification of associations with these lower frequency variants.

Thanks to progress in high-throughput sequencing technologies over the last two decades, whole exome (WES) and whole-genome sequencing (WGS) have become cheaper and more widespread. The first applications of massive parallel sequencing in the context of complex diseases were limited to mostly coding sequences. These studies included targeted re-sequencing of known loci to find independently associated low-frequency or rare variants,[23–25] or analyses of cases with a very severe phenotype within the first six years of life and in whom a monogenic cause was expected (also see below).[26,27] With prices dropping, and increasing technological experience, we are currently looking at the first large-scale WGS studies on 10,000s of cases being conducted by the International IBD Genetics Consortium (IIBDGC). These sample sizes are needed, because we are expecting low-frequency and (ultra)rare variants with low-medium risk, and thus power will be otherwise limited. Computation and interpretation, also of non-coding variants, will be the major challenges in these studies.

## Can we use the genetic findings to improve clinical decision making?

The stable nature of a DNA sequence makes genomics attractive for the identification of biomarkers to guide clinical decision-making. Although clinical applications still lag behind, emerging data point to a role for clinical genetics and genomics in IBD.

*Very early onset IBD.* In contrast to the great bulk of IBD patients in whom inflammation is caused by complex interactions between multiple genes and environmental factors, several rare monogenic disorders with complete penetrance have been identified among children with very early onset (VEO) IBD (defined as diagnosed before six years of age). The gastrointestinal phenotype of these syndromes is often indistinguishable from polygenic IBD, but many of these patients are refractory to

both conventional treatment and standard targeted therapies. More than 50 individual genes have been associated with VEO IBD and whole-exome sequencing (or targeted sequencing) is currently being implemented in clinical practice to provide guidance on causal genetic defects, possible pathway-specific treatment strategies and when to use haematopoietic stem-cell transplantation. In addition to these aspects of clinical decision-making, genomic medicine may allow family counselling and the possibility of screening for other disease syndromes.

*Prediction of diagnosis.* Among patients with a polygenic disease, analyses of individual risk loci as predictors of IBD are less informative, even for loci with the strongest effect size such as the *NOD2* variants (associated with a 2–3 times increased risk for CD). Based on an average life-time risk for CD of 0.4% in the background population, carriage of one of these *NOD2* variants only translates into a 0.8–1.2% life-time risk of developing CD. The low life-time risk of CD and the fact that most carriers of the *NOD2* variants never develop the disease, demonstrate the poor applicability of genetic testing for single susceptibility variants to predict who will develop IBD.

To go beyond the impact of individual risk loci and estimate the overall genetic effect, genetic risk score models have been developed. The 'unweighted' genetic risk score represents the simplest model, which is generated by summing up the total risk allele dosage of an individual patient's genome. The more sophisticated 'weighted' genetic risk score takes into account information on the effect size of each genetic variant, by multiplying the allele dosage at each variant by the logarithm of the odds ratio from the discovery GWAs. Even though the 'weighted' method is considered as more appropriate, the models often generate similar results, since most susceptibility variants are associated with similar relative risk estimates. The performance of these risk scores has improved with time, as more and more IBD-associated variants have been identified. However, their clinical utility has been limited by their low specificity, as there is a large overlap between patients and controls with respect to the distribution of the genetic risk score.

To overcome some of the limitations with traditional genetic risk scores, polygenic risk scores (PRSs) have recently been introduced. PRSs do not only take the IBD-risk variants (identified at a genome-wide significance level) into account, but may be calculated by using genetic variants that have been identified at various levels of *p*-values. Interestingly, PRS that include variants with lower *p*-values seem to improve disease prediction, since they tend to explain more of the phenotypic variance. These novel PRSs may help to identify individuals who are at high risk of developing IBD. However, the utility of PRSs as a screening tool may be questioned from a clinical and ethical perspective as long as no valid preventive measures of IBD have been identified. Also, PRSs calculated from GWAs based on European ancestry samples (i.e. most GWAs published) do not apply to individuals of non-European ancestry. Large-scale GWAs thus will have to be performed on diverse human populations.

*Prediction of future disease course.* There have been numerous attempts to advance this field by trying to identify genotype-phenotype associations. The largest effort has been undertaken by the IIBDGC.[28] They analysed an international cohort of 29,838 IBD patients, and identified only a handful associations (*NOD2*, *MHC* and *MST1* 3p21) with certain phenotypes, mainly age at onset and disease location. Notably, little or no genetic association with CD behaviour remained after conditioning on disease location and age at onset.

To identify possible prognostic markers, patients with an indolent disease have been compared with those with a poor disease course. Even though individual genetic markers such as FOXO3, XACT and the HLA region have been linked to future disease course in this type of within-cases GWAs, none of these markers have progressed to clinical applications. Up until now, attempts to use genetic risk scores instead of individual risk loci for the prediction of future disease course have also failed.

*Prediction of treatment response and immunogenicity.* Genotyping of the most common thiopurine methyl transferase gene (*TPMT*) variants represents the only currently established genetic analysis in IBD implemented in clinical decision-making. The test provides guidance on thiopurine doses to be used and allows prediction of drug intolerance, even though it does not replace the need for regular monitoring of blood tests. Recently, variants in nudix hydrolase 15 (*NUDT15*) have also been linked to the risk of thiopurine-induced myelosuppression,[29,30] and *HLA-DQA1–HLA-DRB1* variants with pancreatitis.[31]

Numerous studies have failed to identify and validate genetic markers of response to anti-tumor necrosis factor (anti-TNF) treatment. However, treatment response has often been subjectively defined, and it is reasonable to believe that the absence of objectively defined outcomes has hampered progress within the field. In contrast to clinical outcomes, immunogenicity represents a more objective outcome. Attempts to associate specific genotypes with anti-drug antibody development have been more successful.[32] Recent data from the Personalising Anti-TNF Therapy in Crohns Disease

(PANTS) cohort demonstrate that *HLA-DQA1*05* is associated with development of antibodies to both infliximab and adalimumab.[33] Interestingly, the immunogenicity in carriers is attenuated by concomitant treatment with an immunomodulatory drug. These findings point to a possible clinical application, even though prediction of clinical outcomes, e.g. loss of response, is still to be shown.

## Future role of clinical genomics in complex traits

Even though current examples of clinical applications are few, progress within the genomics of IBD has played a key role in our current understanding of disease aetiology, revealed novel disease mechanisms such as autophagy, and helped us to identify key pathways of relevance to drug development. Advancements in sequencing technologies and reduction of costs associated with the use of these tools are expected to provide novel insight into disease mechanisms. Various ethnic groups will have to be included in future large-scale studies. In addition, to speed up the identification of clinically useful markers, there is an urgent need to develop objective and robust definitions of outcomes with respect to disease course and response to treatment.

### Declaration of conflicting interests

All authors: no conflicts.

### Funding

No applicable funding.

### References

1. Lee SH, Wray NR, Goddard ME, et al. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* 2011; 88: 294–305.
2. Gordon H, Trier Moller F, Andersen V, et al. Heritability in inflammatory bowel disease: From the first twin study to genome-wide association studies. *Inflamm Bowel Dis* 2015; 21: 1428–1434.
3. Chen GB, Lee SH, Brion MJ, et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. *Hum Mol Genet* 2014; 23: 4710–4720.
4. Glazier AM, Nadeau JH and Aitman TJ. Finding genes that underlie complex traits. *Science* 2002; 298: 2345–2349.
5. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; 411: 599–603.
6. Hugot JP, Laurent-Puig P, Gower-Rousseau C, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. *Nature* 1996; 379: 821–823.
7. Mirkov MU, Verstockt B and Cleynen I. Genetics of inflammatory bowel disease: Beyond NOD2. *Lancet Gastroenterol Hepatol* 2017; 2: 224–234.
8. Anderson CA, Pettersson FH, Clarke GM, et al. Data quality control in genetic case-control association studies. *Nat Protoc* 2010; 5: 1564–1573.
9. Clarke GM, Anderson CA, Pettersson FH, et al. Basic statistical analysis in genetic case-control studies. *Nat Protoc* 2011; 6: 121–133.
10. 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: 661–678.
11. Hampe J, Franke A, Rosenstiel P, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat Genet* 2007; 39: 207–211.
12. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet* 2007; 39: 596–604.
13. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008; 40: 955–962.
14. Ellinghaus D, Jostins L, Spain SL, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat Genet* 2016; 48: 510–518.
15. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* 2011; 42: 1118–1125.
16. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 2012; 491: 119–124.
17. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* 2015; 47: 979–986.
18. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* 2017; 49: 256–261.
19. Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* 2011; 43: 1193–1201.
20. Ji SG, Juran BD, Mucha S, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet* 2017; 49: 269–273.
21. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008; 456: 18–21.
22. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009; 461: 747–753.
23. Beaudoin M, Goyette P, Boucher G, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLoS Genet* 2013; 9: e1003723.
24. Momozawa Y, Mni M, Nakamura K, et al. Resequencing of positional candidates identifies low

frequency IL23R coding variants protecting against inflammatory bowel disease. *Nat Genet* 2011; 43: 43–47.
25. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* 2011; 43: 1066–1073.
26. Uhlig HH. Monogenic diseases associated with intestinal inflammation: Implications for the understanding of inflammatory bowel disease. *Gut* 2013; 62: 1795–1805.
27. Uhlig HH, Schwerd T, Koletzko S, et al. The diagnostic approach to monogenic very early onset inflammatory bowel disease. *Gastroenterology* 2014; 147: 990–1007.
28. Cleynen I, Boucher G, Jostins L, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* 2016; 387: 156–167.
29. Yang SK, Hong M, Baek J, et al. A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. *Nat Genet* 2014; 46: 1017–1020.
30. Walker GJ, Harrison JW, Heap GA, et al. Association of genetic variants in NUDT15 with thiopurine-induced myelosuppression in patients with inflammatory bowel disease. *JAMA* 2019; 321: 773–785.
31. Heap GA, Weedon MN, Bewshea CM, et al. HLA-DQA1-HLA-DRB1 variants confer susceptibility to pancreatitis induced by thiopurine immunosuppressants. *Nat Genet* 2014; 46: 1131–1134.
32. Billiet T, Vande Casteele N, Van Stappen T, et al. Immunogenicity to infliximab is associated with HLA-DRB1. *Gut* 2015; 64: 1344–1345.
33. Sazonovs A, Kennedy NA, Moutsianas L, et al. HLA-DQA1*05 Carriage Associated With Development of Anti-Drug Antibodies to Infliximab and Adalimumab in Patients With Crohn's Disease. *Gastroenterology* 2019 2019/10/11. DOI: 10.1053/j.gastro.2019.09.041.

**Isabelle Cleynen[1] and Jonas Halfvarsson[2]**
*[1]Department of Human Genetics, KU Leuven, Leuven, Belgium*
*[2]Department of Gastroenterology, Örebro University, Örebro, Sweden*

*Corresponding author: Isabelle Cleynen*
*E-mail: isabelle.cleynen@kuleuven.be*