Inherited determinants of Crohn's disease and ulcerative colitis phenotypes

Citation for published version:

Digital Object Identifier (DOI):
10.1016/S0140-6736(15)00465-1

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
The Lancet

Publisher Rights Statement:
Open Access article distributed under the terms of CC BY.
Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: a genetic association study


Summary

Background Crohn’s disease and ulcerative colitis are the two major forms of inflammatory bowel disease; treatment strategies have historically been determined by this binary categorisation. Genetic studies have identified 163 susceptibility loci for inflammatory bowel disease, mostly shared between Crohn’s disease and ulcerative colitis. We undertook the largest genome association study, to date, in widely used clinical subphenotypes of inflammatory bowel disease with the goal of further understanding the biological relations between diseases.

Methods This study included patients from 49 centres in 16 countries in Europe, North America, and Australasia. We applied the Montreal classification system of inflammatory bowel disease subphenotypes to 34 819 patients (19713 with Crohn’s disease, 14 683 with ulcerative colitis) genotyped on the Immunochip array. We tested for genotype–phenotype associations across 156 154 genetic variants. We generated genetic risk scores by combining information from all known inflammatory bowel disease associations to summarise the total load of genetic risk for a particular phenotype. We used these risk scores to test the hypothesis that colonic Crohn’s disease, ileal Crohn’s disease, and ulcerative colitis are all genetically distinct from each other, and to attempt to identify patients with a mismatch between clinical diagnosis and genetic risk profile.

Findings After quality control, the primary analysis included 29 838 patients (16 902 with Crohn’s disease, 12 597 with ulcerative colitis). Three loci (NOD2, MHC, and MST1 3p21) were associated with subphenotypes of inflammatory bowel disease, mainly disease location (essentially fixed over time; median follow-up of 10·5 years). Little or no genetic association with disease behaviour (which changed dramatically over time) remained after conditioning on disease location and age at onset. The genetic risk score representing all known risk alleles for inflammatory bowel disease showed strong association with disease subphenotype (p=1·65 × 10–78), even after exclusion of NOD2, MHC, and 3p21 (p=9·23 × 10–18). Predictive models based on the genetic risk score strongly distinguished colonic from ileal Crohn’s disease. Our genetic risk score could also identify a small number of patients with discrepant genetic risk profiles who were significantly more likely to have a revised diagnosis after follow-up (p=6·8 × 10–4).

Interpretation Our data support a continuum of disorders within inflammatory bowel disease, much better explained by three groups (ileal Crohn’s disease, colonic Crohn’s disease, and ulcerative colitis) than by Crohn’s disease and ulcerative colitis as currently defined. Disease location is an intrinsic aspect of a patient’s disease, in part genetically determined, and the major driver to changes in disease behaviour over time.

Funding International Inflammatory Bowel Disease Genetics Consortium members funding sources (see Acknowledgments for full list).

Introduction Crohn’s disease and ulcerative colitis, the two major forms of inflammatory bowel disease, affect about one in 200 people in developed countries, with a rising incidence and prevalence in developing countries. Many patients with inflammatory bowel disease have a lifetime of debilitating physical symptoms (eg, urgent diarrhoea, rectal bleeding, vomiting, anorexia, and lethargy), which frequently lead to poor psychosocial wellbeing with wide ranging consequences for academic attainment, employment, relationships, and sexual health. Furthermore, the financial costs of inflammatory bowel disease are substantial and are estimated at more than US$2·2 billion per year in the USA alone. Inflammatory bowel disease is characterised by an exaggerated mucosal immune response to luminal gut contents in genetically susceptible individuals. In Crohn’s disease, inflammation can occur in any part of the gastrointestinal tract, whereas ulcerative colitis is typically confined to the colon. The universally adopted Montreal
classification distinguishes clinical subphenotypes in Crohn’s disease by disease location and behavior, and age of onset, and in ulcerative colitis by disease extent and age of onset.1–11 Molecular studies have suggested that ileal and colonic Crohn’s disease are distinct entities because variants in NOD2 are associated with small bowel disease and HLA alleles with colonic disease.12–15 However, current recommendations do not advocate the use of these established markers in making treatment decisions, nor for choosing patients for clinical trials.16–23

The natural history and clinical course of inflammatory bowel disease is very heterogeneous: up to 20% of patients with ulcerative colitis need colectomy for medically refractory disease, and more than 50% of patients with Crohn’s disease need surgery within 10 years of diagnosis;30 however, up to 50% of patients with ulcerative colitis and 30% with Crohn’s disease will have a fairly indolent disease course without the need for immunosuppression or surgery.17–19

Inflammatory bowel disease has been at the vanguard of progress in understanding the genetic framework of complex diseases, with 163 susceptibility loci identified so far.20 Most inflammatory bowel disease loci confer risk on both ulcerative colitis and Crohn’s disease, but typically show distinct effect sizes in the two disorders. These findings suggest that genetic variation might define molecular subtypes independent of traditional and clinically defined diagnostic entities, allowing new insights into the molecular basis of these subphenotypes.

Our international study of around 30 000 patients with inflammatory bowel disease genotyped by microarray is the largest genotypic–subphenotype study in the disease done so far. We have used genetic risk scores to study genetic heterogeneity underpinning the natural history of inflammatory bowel disease. This analysis rejects the current binary classification of Crohn’s disease and ulcerative colitis as distinct and homogenous clinical entities in favour of a continuum of illness better fit by a three-category model (ie, ileal Crohn’s disease, colonic Crohn’s disease, and ulcerative colitis). We show that these risk scores have clinical potential (although they are currently only weak predictors), and believe they might have widespread applicability in other diseases.

**Methods**

**Study design and patients**

We acquired phenotype data for 34 819 patients, including 17 713 with Crohn’s disease and 14 683 with ulcerative colitis. The cohort included all patients in different centres over the years who had inflammatory bowel disease as per Lennard-Jones’ criteria.26 All inclusion criteria are included in appendix A. After quality control (appendix A), the primary analysis included 29 838 patients (16 902 with Crohn’s disease, 12 597 with ulcerative colitis, 255 with indeterminate colitis, and 84 missing an exact diagnosis). This study includes patients from population-based registries, secondary and tertiary-referral centres at 49 sites in 16 countries in Europe, North America, and Australasia, most of which have been previously described (appendix B). Confirmation of diagnosis of inflammatory bowel disease and assignment of clinical subphenotypes were done by clinicians specialising in inflammatory bowel disease or trained phenotypers through case note reviews of clinical, radiological, histopathological, and endoscopic reports, and classified per the Montreal classification criteria (see appendix A for details).44 For behaviour and surgery in Crohn’s disease, and colectomy in ulcerative colitis, Kaplan-Meier
survival curves, stratified by location of Crohn’s disease and extent of ulcerative colitis, were drawn to estimate time to first event (see appendix A for details).

The ethical boards of each separate recruiting centre approved the study. All patients included in this study gave written informed consent.

Procedures
All cases were genotyped with the Immunochip array (Illumina, San Diego, CA, USA; appendix B) as previously described. Briefly, the Immunochip is a 195 806-polymorphism genotyping platform comprising variants identified from association studies of immune-related disorders including Crohn’s disease and ulcerative colitis. Extensive quality control was performed on the dataset (appendix A), leaving 29 838 cases and 156 154 markers available for analyses. Variants in the MHC, including 23 HLA alleles that have been implicated in inflammatory bowel disease, were imputed as described in appendix A.35

All association tests were done on all genotyped variants, conditional on the first five principal components to account for population structure. Age of onset was analysed for Crohn’s disease and ulcerative colitis separately and then meta-analysed; time to surgery was analysed with parametric survival-time regression models; and upper gastrointestinal involvement and perianal disease were analysed with binary logistic regression (see appendix A). For multicategory phenotypes (disease location, behaviour, and extent) we used model selection to pick the most appropriate genetic model for the phenotype (appendix A). The model selection indicated a multinomial model for Crohn’s disease location (ie, three unordered categories), an ordinal logistic model for Crohn’s disease behaviour (three ordered categories, B3 penetrating>B2 stricturing>B1 inflammatory), and a binary model for disease extent of ulcerative colitis (two categories: E3 extensive disease vs E2 left-sided disease and E1 proctitis). To distinguish direct associations from indirect (ie, driven by an association with a correlated phenotype), we also adjusted all regression models for the other phenotypes (age of onset, location, and behaviour for Crohn’s disease; age of onset and disease extent for ulcerative colitis). Genome-wide significance (p<5×10⁻⁸⁸) was required for individual single nucleotide polymorphisms (SNPs) and HLA types.

All signals that showed suggestive association (p<1×10⁻⁵) with any of the disease subphenotypes were assessed in an independent cohort genotyped on a range of different genome-wide association study (GWAS) chips. These samples have also undergone rigorous quality control and imputation. Phenotype data for an additional and independent 2453 patients with Crohn’s disease and 3729 patients with ulcerative colitis were available for these analyses. See appendix A for additional information about the replication cohort.

To learn about the relative phenotypic variance explained by different risk factors in adult inflammatory bowel disease, we fitted a model to predict Crohn’s disease location that included both demographic predictors (smoking status, age at diagnosis, and year) and genetic predictors (SNPs at NOD2, MST1, and the HLA cluster as well as the genetic risk score). Variance explained on the logit scale by each predictor was calculated with the McKelvey–Zavoina pseudo R². Centres with a high proportion (>60%) of missing data for smoking status were removed. To reduce the effect of changes in clinical practice and smoking rates, only patients born between 1955 and 1985 were included.

In addition to looking at single SNPs, we also combined information from 193 SNPs and 23 HLA types previously associated with inflammatory bowel disease to generate genetic risk scores (appendix A), which provide better predictive accuracy than individual SNPs. To assess classification accuracy, we re-ran the risk score analyses with a cross-validation strategy, in which models were fitted in non-UK origin samples and assessed by how well they classified UK samples.

To assess if the risk score can be used to identify misclassified patients, we selected 97 outlier patients that fell in the extreme tail of the scores for the opposite phenotype [log (Crohn’s disease versus ulcerative colitis [CD vs UC]) score ≤−2 for Crohn’s disease outliers and log CD vs UC score ≥2 for ulcerative colitis outliers], as well as 95 randomly selected cases with non-outlier scores matched by recruitment centre. Clinicians from each centre were then asked to re-phenotype both outlier and non-outlier patients in a masked fashion. The CD versus UC risk score was chosen for this experiment because it had the strongest association with Crohn’s disease location and behaviour.

Statistical analysis
The median effect size of known inflammatory bowel disease risk variants13 was about OR 1.1, with a median minor allele frequency of roughly 30%. The sample size of our study gave high power to detect an effect of equivalent magnitude of Crohn’s disease location (power of 67% for ileal vs non-ileal disease), Crohn’s disease behaviour (94% for complicated vs non-complicated disease) and ulcerative colitis disease extent (84% for extensive vs non-extensive disease) at genome-wide significance. Binary and linear genotype–phenotype analyses were done with PLINK version 1.07,10 and multinomial and ordinal regression with a custom program, Trinculo version 0.4 (appendix A). Survival analysis and risk prediction were done with R-2.15.1 using the packages “survival” and “Mangrove”,11 respectively. Data handling and plotting was done with R.

Role of the funding source
The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors had full access to all the

www.thelancet.com Published online October 19, 2015 http://dx.doi.org/10.1016/S0140-6736(15)00465-1 3

Children’s Hospital of Philadelphia, Philadelphia, PA, USA (Prof H Hakonarson PhD); Department of Gastroenterology, Faculty of Medicine and Health, Örebro University, Sweden (J Halfvarson MD); School of Health and Medical Sciences, Örebro University, Örebro, Sweden (J Halfvarson); Norwegian PSC Research Center, Research Institute of Internal Medicine and Department of Transplantation Medicine, Osaka University Hospital and University of Osaka, Osaka, Norway (J R Hov PhD); Gastrointestinal Unit, Institute of Genetics and Molecular Medicine (N A Kennedy PhD, Prof Satsangi FRSE, C W Lees PhD), and Child Life and Health (Prof C Wilson PhD), University of Edinburgh, Edinburgh, UK; Department of Gastroenterology, Lithuanian University of Health Sciences, Kaunas, Lithuania (Prof L Kuprinskas MD); Centre for Inflammatory Bowel Diseases, Saint John of God Hospital, Sobioxa WA and School of Medicine and Pharmacology, University of Western Australia, Harry Perkins Institute for Medical Research, Murdoch, WA, Australia (Prof I C Lawrance PhD); Inflammatory Bowel Disease Research Group, Addenbrooke’s Hospital, University of Cambridge, Cambridge, UK (C Lee PhD, M Parkes PhD); Unit of Animal Genomics, Groupe Interdisciplinaire de Genoproteomique Appliquee (GICA-R) and Faculty of Veterinary Medicine, University of Liege, Liege, Belgium (E Théâtre PhD); Division of Gastroenterology, Centre Hospitalier Universitaire, Universite de Liege, Liege, Belgium (E Théâtre); Department of Gastroenterology and Hepatology, Leiden University Medical Center, Leiden, Netherlands (A E van der Meulen-de Jong MD); Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, Netherlands (Prof R K Weersma PhD); Royal Hospital for Sick Children,
Table 1: Phenotype distribution of primary cohort

<table>
<thead>
<tr>
<th>Disease location†</th>
<th>Crohn’s disease (n=16 902)</th>
<th>Ulcerative colitis (n=12 597)</th>
<th>Inflammatory bowel disease* (n=29 838)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease extent†</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proctitis (E1)</td>
<td>−</td>
<td>1271 (12%)</td>
<td>−</td>
</tr>
<tr>
<td>Left-sided (E2)</td>
<td>−</td>
<td>4087 (38%)</td>
<td>−</td>
</tr>
<tr>
<td>Extensive (E3)</td>
<td>−</td>
<td>5212 (48%)</td>
<td>−</td>
</tr>
<tr>
<td>Other</td>
<td>−</td>
<td>205 (2%)</td>
<td>−</td>
</tr>
<tr>
<td>Missing</td>
<td>−</td>
<td>1822 (14%)</td>
<td>−</td>
</tr>
<tr>
<td><strong>Disease behaviour†</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflammatory (B1)</td>
<td>6196 (50%)</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Strictures (B2)</td>
<td>3250 (26%)</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Penetrating (B3)</td>
<td>3054 (24%)</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Surgery</td>
<td>2762 (18%)</td>
<td>−</td>
<td>−</td>
</tr>
</tbody>
</table>

Phenotypes

- **Disease location†**: ileal (L1); colorectal (L2); ileocolonic (L3); and upper GI (L4).
- **Disease extent†**: proctitis, left-sided, extensive, other, and missing.
- **Disease behaviour†**: inflammatory, stricture, penetrating, and surgery.

*Includes 255 patients with indeterminate colitis and 84 patients with missing exact diagnosis.
†Excludes data obtained with patient questionnaires (2658 patients).

**Results**

Our primary analyses were done on matched genotype and phenotype data from 29 838 patients of European ancestry (appendix A) with inflammatory bowel disease (16 902 with Crohn’s disease, 12 597 with ulcerative colitis; table 1) with a total of 217 195 patient-years of follow-up (median per patient: 11 years for Crohn’s disease and 10 years for ulcerative colitis). Demographic features of the study population agreed with previously published results: patients with Crohn’s disease were more likely to be younger at diagnosis, female, smokers, and have affected family members than were patients with ulcerative colitis (table 1). Extensive disease was more common in those diagnosed at a younger age in both Crohn’s disease and ulcerative colitis, whereas disease behaviour was relatively unaffected by age at diagnosis (appendix A). Reaffirming the progressive nature of Crohn’s disease, the proportion of patients with strictureing (B2) or penetrating (B3) disease increased from less than 30% (n/N) at diagnosis to 43% (n/N) at 5 years, 56% (n/N) at 10 years, and 74% (n/N) at 30 years (figure 1, which shows the progression in B1, B2, and B3 disease individually [smoothed estimates over intervals]). By contrast, disease location showed little variation during the same period (figure 1). With the exception of the population-based cohorts from Scandinavia, survival analyses of time to development of complicated disease (B2, B3) or first surgery in Crohn’s disease were highly consistent across the different countries of origin despite different health-care systems and methods of sampling (appendix A). In Crohn’s disease, time from diagnosis to progression (complicated disease or surgical intervention) was significantly shorter in purely ileal (L1) compared with ileocolonic (L3) or colonic (L2) disease (p<10^-100; figure 1; appendix A). Overall, 7257 (52%) of 13 862 patients with Crohn’s disease had undergone surgery by the time of first follow-up. In ulcerative colitis, in which the overall rate of colectomy was 22% 10 years after diagnosis, time to surgery was shorter in patients with extensive disease (E3) than in those with left-sided disease (E2) or proctitis (E1; p=8×10^-44; figure 1).

We tested genetic variants with association with age at diagnosis and time to surgery in all patients with inflammatory bowel disease; disease location and behaviour in Crohn’s disease; and disease extent in ulcerative colitis (table 2 and table 3). Across all analyses, three loci achieved genome-wide significance (p<5×10^-8): 3p21 (MST1), NOD2, and the MHC. No additional signals were noted after replication of suggestive loci (p<1×10^-8) in an independent GWAS cohort (appendix B). Although NOD2 was strongly associated with Crohn’s disease location, behaviour, and age at diagnosis, adjustment for the other phenotypes showed that the data in the study and had final responsibility for the decision to submit for publication.
Figure 1: Evolution of clinical subphenotypes

(A) Proportion of patients with Crohn’s disease who have inflammatory (Montreal classification B1), stricturing (B2), or penetrating (B3) disease over time from diagnosis to most recent follow-up. (B) Proportion of patients with Crohn’s disease who have ileal (L1), colonic (L2), or ileocolonic (L3) disease over time from diagnosis to most recent follow-up. (C) Survival plot of time from diagnosis of Crohn’s disease to resectional surgery stratified by disease location. (D) Survival plot of time from diagnosis of ulcerative colitis to colectomy stratified by disease extent (extensive disease, E3; non-extensive disease, E1 and E2).

Table 2: Associations between genotype and age at diagnosis achieving genome-wide significance

<table>
<thead>
<tr>
<th>MAF</th>
<th>Age at diagnosis of IBD p value</th>
<th>P (SE)</th>
<th>Age at diagnosis of Crohn’s disease p value</th>
<th>P (SE)</th>
<th>Age at diagnosis of ulcerative colitis p value</th>
<th>P (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3p21 (MST1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs35261698</td>
<td>0.306</td>
<td>6.34e-10</td>
<td>-0.06 (0.01)*</td>
<td></td>
<td>3.65e-10</td>
<td>-0.06 (0.01)</td>
</tr>
<tr>
<td>rs2172252</td>
<td>0.288</td>
<td>1.35e-12</td>
<td>-0.06 (0.01)*</td>
<td></td>
<td>2.93e-10</td>
<td>-0.07 (0.01)*</td>
</tr>
<tr>
<td>rs2197999</td>
<td>0.281</td>
<td>2.73e-10</td>
<td>-0.06 (0.01)*</td>
<td></td>
<td>2.37e-10</td>
<td>-0.07 (0.01)*</td>
</tr>
<tr>
<td>6p21 (MHC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs315674</td>
<td>0.116</td>
<td>3.42e-05</td>
<td>-0.03 (0.01)</td>
<td></td>
<td>3.35e-10</td>
<td>-0.04 (0.02)</td>
</tr>
<tr>
<td>rs431651</td>
<td>0.034</td>
<td></td>
<td></td>
<td></td>
<td>1.50e-04</td>
<td>-0.07 (0.03)</td>
</tr>
<tr>
<td>rs3129891</td>
<td>0.097</td>
<td>1.15e-06</td>
<td>-0.05 (0.01)</td>
<td></td>
<td>1.43e-10</td>
<td>-0.09 (0.02)*</td>
</tr>
<tr>
<td>rs9268832</td>
<td>0.393</td>
<td>7.42e-10</td>
<td>-0.05 (0.01)*</td>
<td></td>
<td>4.56e-10</td>
<td>-0.06 (0.01)</td>
</tr>
<tr>
<td>rs482044</td>
<td>0.401</td>
<td></td>
<td></td>
<td></td>
<td>1.51e-06</td>
<td>0.03 (0.01)</td>
</tr>
<tr>
<td>16q12 (NOD2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2066844</td>
<td>0.045</td>
<td>3.58e-07</td>
<td>-0.08 (0.02)</td>
<td></td>
<td>1.21e-10</td>
<td>-0.1 (0.02)</td>
</tr>
<tr>
<td>rs2066845</td>
<td>0.016</td>
<td>2.10e-04</td>
<td>-0.1 (0.03)</td>
<td></td>
<td>5.50e-10</td>
<td>-0.09 (0.03)</td>
</tr>
<tr>
<td>rs2066847</td>
<td>0.024</td>
<td>6.64e-10</td>
<td>-0.16 (0.02)</td>
<td></td>
<td>2.04e-10</td>
<td>-0.17 (0.02)*</td>
</tr>
</tbody>
</table>

Loci are listed by single nucleotide polymorphism. Age at diagnosis assessed by linear regression analysis on normalised data for Crohn’s disease and ulcerative colitis; IBD assessed by meta-analysis of Crohn’s disease and ulcerative colitis data. Effect size is given as standard deviation unit (standard error of effect). MAF = minor allele frequency. IBD = inflammatory bowel disease. * = non-significant associations (p nominal < 0.05). † = genome-wide significant associations. # = The most significant association per subphenotype, if genome-wide significant.
association of NOD2 with behaviour was driven almost entirely by its phenotypic correlation with location and age at diagnosis (figure 2).

We noted complex and correlated HLA signals for susceptibility to inflammatory bowel disease overall and age at onset, as well as Crohn’s disease location and behaviour, extent of ulcerative colitis, and surgery (figure 2; appendix A; appendix B). In-depth analysis of the MHC region including classical HLA alleles showed that the strongest signal for disease location was a colonic association with HLA-DRB1*01:03 (p=1·47 × 10⁻²³; ileal vs colonic odds ratio [OR] 0·32, 95% CI 0·29–0·41; ileocolonic vs colonic OR 0·47, 0·39–0·57), which is also the strongest shared risk allele for Crohn’s disease and ulcerative colitis, followed by HLA-DRB1*07:01 (figure 2; appendix A). rs77005575 was independently associated with Crohn’s disease behaviour (p=1·56 × 10⁻⁰⁹; figure 2; appendix A). Notably, alleles associated with susceptibility to ulcerative colitis were better predictors of colonic disease location in Crohn’s disease than alleles associated with susceptibility to Crohn’s disease (appendix A). The top signal for extent of ulcerative colitis was rs3115674 (p=5·11 × 10⁻¹⁷; OR 0·70, 0·64–0·76; appendix A), which correlates with HLA-B*08 (R²=0·66), found mostly on the ancestral 8.1 HLA haplotype. HLA-DRB1*13:01 was the top signal for age at diagnosis of ulcerative colitis (3·50 × 10⁻⁰⁹; figure 2; appendix A).

On the basis of sample size, our primary analysis had similar power to detect associations to disease location (ileal vs colonic) as the first International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) meta-analysis on Crohn’s disease. However, with the exception of NOD2, HMC, and MST1, we do not report significant associations between subphenotypes and individual SNPs, including those robustly associated with disease susceptibility. We noted, however, that many known risk loci showed nominal evidence for association to a range of subphenotypes, so we posited that genetic risk scores representing the combined effect of many individually weak signals might be a more powerful approach to study the genetic underpinnings of subphenotypes for inflammatory bowel disease. We

<table>
<thead>
<tr>
<th>MAF</th>
<th>Crohn’s disease</th>
<th></th>
<th></th>
<th>Ulcerative colitis</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Disease location</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
<td>p value</td>
<td>OR (95% CI)</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td></td>
<td>ileocolonic vs colonic</td>
<td></td>
<td></td>
<td></td>
<td>ileal vs colonic</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3p21 (MST1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2127252</td>
<td>0.288</td>
<td>3.10 × 10⁻¹⁰</td>
<td>1.07</td>
<td>(1.00–1.13)</td>
<td>1.10</td>
<td>(1.02–1.19)</td>
</tr>
<tr>
<td>rs3197999</td>
<td>0.281</td>
<td>2.10 × 10⁻¹⁰</td>
<td>1.08</td>
<td>(1.02–1.15)</td>
<td>1.10</td>
<td>(1.02–1.19)</td>
</tr>
<tr>
<td>6p21 (MHC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3115674</td>
<td>0.316</td>
<td>3.00 × 10⁻¹⁰</td>
<td>0.88</td>
<td>(0.80–0.97)</td>
<td>0.81</td>
<td>(0.72–0.91)</td>
</tr>
<tr>
<td>rs4151651</td>
<td>0.034</td>
<td>2.42 × 10⁻¹⁰</td>
<td>0.71</td>
<td>(0.62–0.81)</td>
<td>0.58</td>
<td>(0.50–0.68)</td>
</tr>
<tr>
<td>rs9307007</td>
<td>0.112</td>
<td>8.13 × 10⁻¹⁰</td>
<td>0.68</td>
<td>(0.62–0.75)</td>
<td>0.58</td>
<td>(0.52–0.65)</td>
</tr>
<tr>
<td>rs3128991</td>
<td>0.209</td>
<td>3.10 × 10⁻¹⁰</td>
<td></td>
<td></td>
<td>8.00 × 10⁻¹⁰</td>
<td>0.92</td>
</tr>
<tr>
<td>rs9268832</td>
<td>0.393</td>
<td>3.40 × 10⁻¹⁰</td>
<td>1.03</td>
<td>(0.97–1.09)</td>
<td>0.94</td>
<td>(0.87–1.02)</td>
</tr>
<tr>
<td>rs482044</td>
<td>0.401</td>
<td>2.38 × 10⁻¹⁰</td>
<td>1.15</td>
<td>(1.08–1.22)</td>
<td>1.25</td>
<td>(1.16–1.35)</td>
</tr>
<tr>
<td>rs7005575</td>
<td>0.439</td>
<td>1.00 × 10⁻¹⁰</td>
<td>1.21</td>
<td>(1.16–1.26)</td>
<td>1.33</td>
<td>(1.23–1.44)</td>
</tr>
</tbody>
</table>

Loci are listed by single nucleotide polymorphism. Disease location assessed by multinomial logistic regression analysis; disease behaviour by ordinal logistic regression analysis (effect size is odds ratio [95% CI] for ileal versus B1, which is also equivalent to B3 vs B2–B1); and disease extent by binomial logistic analysis. Surgery and colectomy assessed by survival analysis under a Weibull distribution. MAF=minor allele frequency. OR=odds ratio. HR=hazard ratio. *Non-significant associations (p value >0.05). †Genome-wide significant associations. ‡The most significant association for locus per subphenotype, if genome-wide significant.

Table 3: Associations between genotype and disease location, behaviour, extent, surgery, and colectomy achieving genome-wide significance
calculated different inflammatory bowel disease risk scores constructed from all available data (strength and direction of association) on the lead SNPs from each of the 163 known inflammatory bowel disease susceptibility loci. Although all of the risk scores were associated with Crohn’s disease and ulcerative colitis subphenotypes, the most powerful score used the differences between Crohn’s disease and ulcerative colitis (CD vs UC score; figure 2). Importantly, this CD versus UC score retained significance even after NOD2, MHC, and MST1 were removed (appendix A), lending support to the notion that the genetic risk score offers more information about the genetic substructure of inflammatory bowel disease than individual SNP associations alone. The strongest correlations in our study were between the CD versus UC risk score and Crohn’s disease location and behaviour (figure 2; \( p = 1.65 \times 10^{-28} \), or \( p = 9.23 \times 10^{-18} \) after genome-wide significant loci were removed). Risk scores that incorporated imputed HLA types that have been implicated in risk for inflammatory bowel disease significantly improved the genetic risk scores compared with those using SNPs only (appendix A).

Having shown the genetic risk score to be a useful measurement of inflammatory bowel disease subphenotype, we used it to study the genetic relation between ileal Crohn’s disease, colonic Crohn’s disease, and ulcerative colitis. We observed significant differences in the genetic risk score between ileal and colonic Crohn’s disease (figure 2; \( p < 0.003 \) for all phenotype-score combinations). These findings suggest that the genetic risk score can be used to identify subgroups of patients with different genetic susceptibilities to inflammatory bowel disease.
and ulcerative colitis. The CD versus UC risk score placed colonic Crohn’s disease between ileal Crohn’s disease and ulcerative colitis (figure 3). Several other risk scores supported this relation, and although partly driven by the highly location-specific NOD2 variants, a risk score with NOD2 removed showed a similar pattern (appendix A). Additionally, statistical model selection across SNPs and HLA types strongly favoured a model in which colonic Crohn’s disease is intermediate between ileal Crohn’s disease and ulcerative colitis over one that grouped both Crohn’s disease subphenotypes as a single category (appendix A). To test whether this finding extends to other subtypes of inflammatory bowel disease, we applied the genetic risk score to two intermediate forms of the disease: ileocolonic Crohn’s disease (L3), in which the disease affects both small and large bowel, and colonic inflammatory bowel disease unclassified, in which the clinical and histological appearances are indistinguishable between Crohn’s disease and ulcerative colitis. The CD versus UC score placed ileocolonic Crohn’s disease as intermediate between ileal (L1) and colonic (L2) Crohn’s disease, and colonic inflammatory bowel disease unclassified between ulcerative colitis and colonic Crohn’s disease (figure 3). Despite the statistical significance of the associations between genetic risk score and subphenotype, the small effect sizes translated into fairly low predictive accuracy when tested by cross-validation (appendix A). The risk score that was most significantly associated with location of Crohn’s disease in the primary analysis (CD vs UC) gave an area under the receiver-operating characteristic (ROC) curve of only 0.60 (95% CI 0.57–0.63) for distinguishing between colonic (L2) and ileal (L1) Crohn’s disease in cross-validation, and even a specifically constructed ileal versus colonic score achieved an area

<table>
<thead>
<tr>
<th>Beta</th>
<th>SE</th>
<th>p value</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever smoker</td>
<td>-0.041</td>
<td>0.108</td>
<td>7.00 × 10⁻¹</td>
</tr>
<tr>
<td>Smoker at diagnosis</td>
<td>0.473</td>
<td>0.117</td>
<td>5.34 × 10⁻¹</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>-0.033</td>
<td>0.005</td>
<td>1.50 × 10⁻⁴</td>
</tr>
<tr>
<td>Year of birth</td>
<td>-0.010</td>
<td>0.005</td>
<td>6.94 × 10⁻²</td>
</tr>
<tr>
<td>rs693077 (MHC)</td>
<td>-0.302</td>
<td>0.082</td>
<td>2.19 × 10⁻⁴</td>
</tr>
<tr>
<td>rs77005575 (MHC)</td>
<td>0.190</td>
<td>0.055</td>
<td>5.21 × 10⁻⁴</td>
</tr>
<tr>
<td>NOD2*</td>
<td>0.522</td>
<td>0.070</td>
<td>2.60 × 10⁻⁴</td>
</tr>
<tr>
<td>Genetic risk score†</td>
<td>0.165</td>
<td>0.038</td>
<td>1.61 × 10⁻³</td>
</tr>
</tbody>
</table>

*Number of risk alleles at the three NOD2 hits. †Crohn’s disease versus ulcerative colitis (CD vs UC) genetic risk score (without NOD2 and MHC). § The total R² for these parameters, excluding the principal component used to account for population stratification. In view of the correlation structure, this is not expected to be equivalent to the sum of R² obtained for each parameter.

Table 4: Variance explained by demographic and genetic factors for disease location in adult onset of Crohn’s disease

Figure 3: Violin plot showing the genetic substructure of inflammatory bowel disease location

The violin represents the range of the log CD versus UC score for the indicated subphenotype (calculated with the R package “vioplot”), with dots representing the mean of that group and error bars the 95% CIs. Although the effects are small compared with the variation within groups, the mean effects can still be measured accurately (right side of the figure). It can be seen on this figure that the Crohn’s disease versus ulcerative colitis (CD vs UC) risk score placed colonic Crohn’s disease between ileal Crohn’s disease and ulcerative colitis. The plot also shows the positioning of the intermediate phenotypes (ileocolonic Crohn’s disease and inflammatory bowel disease unclassified [IBD-U]) in between ileal and colonic Crohn’s disease, and ulcerative colitis and colonic Crohn’s disease, respectively.
inflammatory bowel disease has raised the exciting possibility of a more personalised approach to clinical management. In inflammatory bowel disease, this quest is particularly urgent because of the substantial heterogeneity in disease course, and individual response to therapy.\(^{39}\) Overall, we can only explain a little of the phenotypic variance in the adult population with either classical or genetic predictors. The combination of smoking and the strongest genetic predictors explains only 6·8% of the variance for disease location in Crohn’s disease (table 4 and table 5), and 1·1% for disease extent in ulcerative colitis (table 5).

To assess the possible clinical usefulness of the genetic risk score, we reassessed the cases of Crohn’s disease with a low CD versus UC risk score (more ulcerative colitis-like), and cases of ulcerative colitis with a high CD versus UC risk score (more Crohn’s disease-like; figure 4). Masked re-phenotyping of these cases raised doubts about the original diagnosis in 27% of the outlier cases compared with 8% of non-outlier cases (corrected for disease location, \(p=6·8\times 10^{-4}\); appendix A). This finding suggests that we can indeed use genetics to identify small numbers of misclassified patients.

**Discussion**

The successful identification of genetic variants associated with complex diseases such as inflammatory bowel disease has raised the exciting possibility of a more personalised approach to clinical management. In inflammatory bowel disease, this quest is particularly urgent because of the substantial heterogeneity in disease course, and individual response to therapy.\(^{19}\) Past studies in inflammatory bowel disease have established a genetic component of disease sub-phenotype,\(^{14,16,18,40,41}\) but these studies have been limited to a handful of candidate regions in modest numbers of patients. Our study, involving the universal application of standardised phenotyping by trained personnel on nearly 30 000 patients with inflammatory bowel disease from 49 centres worldwide, combined with matching genotypes from more than 150 000 variants, represents the definitive investigation to date into the genetic basis of subphenotypes of inflammatory bowel disease.

The only genome-wide significant associations we noted were between age of onset and disease location with variants at \(NOD2\), MHC, and \(3p21\). Although the associations between ileal Crohn’s disease and \(NOD2\), and those between colonic Crohn’s disease and the MHC, have been previously described,\(^{14–19}\) our study has dissected the phenotype–location associations for the first time. Importantly, in Crohn’s disease, we showed that \(NOD2\) is not associated with strictureting disease after accounting for disease location. These findings, and the rarity of long-term change in disease location compared with behaviour, suggest that location is a fundamental biological aspect of a patient’s disease, whereas behaviour (like surgery or treatment history) is a marker of disease progression.
Our results for ulcerative colitis accord with the previously reported independent associations between the MHC and both extensive disease and colectomy. Notably, the strongest associations with extensive ulcerative colitis are variants on the ancestral 8.1 HLA haplotype. This haplotype is a known recessive risk for primary sclerosing cholangitis, a disease often associated with an extensive but quiescent form of ulcerative colitis. Although variants associated with susceptibility to Crohn’s disease and those associated with susceptibility to ulcerative colitis are both predictive of disease location, the ulcerative colitis-associated variants are the most predictive. The variants associated with disease susceptibility are also slightly predictive of age at diagnosis.

Composite genetic risk scores from all 163 independent susceptibility signals were strongly associated with all our main subphenotypes, and these findings remained significant after excluding NOD2 and MHC. This result accords with a similar finding for genetic risk scores in bipolar disorder, and hints at the possibility that such approaches might be broadly applicable for studying clinical heterogeneity of common diseases. This finding suggests that many or most risk variants for inflammatory bowel disease do contribute weakly to subphenotype. The relative dearth of individual single-nucleotide polymorphism associations with subphenotypes in our study, by contrast with those reported in similarly powered studies of inflammatory bowel disease susceptibility, suggests that the genetic variants studied here have a small effect, and that environmental factors (such as diet, microbiota, and smoking) might be strong contributors to the subphenotypes. However, an intriguing possibility, supported by the notable absence of any functional or pathway enrichment in the components of the genetic risk scores, is that current phenotypic classifications do not correspond strongly to underlying molecular entities. Of particular note is the genetic distinction seen between ileal Crohn’s disease, colonic Crohn’s disease, and ulcerative colitis. These disease types were identified as equally distinct entities on a genetic continuum: on multiple risk scores, colonic Crohn’s disease was genetically intermediate between ileal Crohn’s disease and ulcerative colitis, a finding that remained significant after excluding NOD2. This result substantiates the view that colonic versus ileal disease, rather than disease extent, is the primary clinical unit of Crohn’s disease classification, and is further supported by the finding that both the genetic risk score and clinical complication rate in patients with ileocolonic disease is intermediate between that of patients with ileal disease and colonic disease. It will be of great interest, and potential clinical use, to see if application of these risk scores helps to classify the 10% of patients who are currently designated colonic inflammatory bowel disease unclassified.

The composite scores were also able to identify small numbers of patients with outlier scores who were much more likely to be misdiagnosed than a typical patient. These findings support the possible clinical usefulness of composite scores of multiple genetic variants, each of small effect. For example, genetic outliers could be excluded or specifically targeted for clinical trials to select more homogeneous groups. If these data were readily available, they might affect clinical decisions or inform risk–benefit discussions with patients. For example, the type of surgery offered to patients with refractory colitis crucially depends on whether they have ulcerative colitis or colonic Crohn’s disease; poor outcomes (including pelvic sepsis, incontinence, and sexual dysfunction) of ileal pouch–anal anastomosis reconstruction surgery are much more common in patients with colonic Crohn’s disease. Use of genetic risk scores to identify possible misdiagnoses in this group of patients could help to reduce this problem. Our data also suggest that genetic risk scores could augment biomarkers, such as faecal calprotectin, currently used for patient stratification in inflammatory bowel disease.

A limitation of our study is that the genetic variants tested were restricted to those present (and that passed quality control) on the Immunochip platform, designed for replication and fine mapping of potential immune-mediated disease loci. Therefore, there still might be important loci that determine disease behaviour, location, and age at onset but are independent of those that confer risk for inflammatory bowel disease (or other immune-mediated diseases), and had limited or absent coverage on the Immunochip. A high-powered GWAS, designed to assess possibly overlooked genetic determinants for these outcomes of phenotype expressivity, which uses our large collection of cooperatively phenotyped cases of inflammatory bowel disease with genomic DNA (and the pre-existing Immunochip genotypes), will likely be of great value.

In summary, our research represents the largest genotype–phenotype study in inflammatory bowel disease done so far. Associations achieving genome-wide significance were identified at only three loci, suggesting that new clinical phenotypic classifications might need to be explored for inflammatory bowel disease, and the relation examined between subphenotypes and other omic profiles and environmental factors, including the microbiota. However, our data suggest that on the basis of genetic factors, inflammatory bowel disease is better classified into three distinct groups (ileal Crohn’s disease, colonic Crohn’s disease, and ulcerative colitis), and we would recommend that clinicians adopt this nomenclature in regular practice. We also show that, although genetic risk scores do not yet have widespread clinical use, they are already valuable in some contexts in inflammatory bowel disease, and their study in other complex disease phenotypes is warranted.
Contributors
IC, GB, Lj, DPBM, IC, and CWL contributed equally to the manuscript. CWL, JCB, GR-S, MSS, JM, JDR, and MP conceived, designed, and managed the study and managed the funding. IC, GB, Lj, LPS, JDR, DPBM, JCB, and CWL were involved in manuscript preparation. IC, GB, Lj, PG, MJD, HHu, JDR, and JCB performed or supervised statistical and computational analyses. SZ, TA, VAnd, JMA, VAng, SB, SBR, JHC, MD, RHDr, LRF, RBG, HHa, HHu, JK, ICL, JCL, SS, ET, AEvM D-d, RKW, DCW, MP, SV, JDR, JM, MSS, GR-S, DPBM, and CWL were involved in patient recruitment and assembling phenotypic data. IC, LPS, SZ, RHD, AF, JCL, ET, RKW, and JDR, and DPBM established DNA collections, genotyping, and data management. LPS assembled, managed, and validated all submitted phenotype data. SZ, TA, VAnd, JMA, VAnn, SB, SBR, JHC, MD, RHD, LRF, RG, HHu, JK, ICL, JCL, SS, ET, AEvM, RKW, DCW, MP, SV, JDR, JM, MSS, GR-S, DPBM, and CWL were involved in patient recruitment and assembling phenotypic data. All authors read and approved the final manuscript before submission.

Declaration of interests
IC, GB, Lj, LPS, SZ, TA, SBR, JHC, MD, AF, RBG, PG, HHu, JRH, HHu, JK, ICL, JCL, SS, ET, DCW, MP, SV, JDR, JM, GR-S, and DPBM declare no competing interests. VAnn reports personal fees from Merck (MSD; consulting, participation in advisory board), and other fees (participation in a clinical study) from Roche, outside the submitted work. JMA reports other fees (presence on advisory board) from Takeda, AbbVie, Janssen, Ferring, and Hospira; speakers fees from Shire, Abbvie, Janssen, Ferring, and AstraZeneca; grants from Ferring, AbbVie, Janssen, and Abbott; non-financial support from Janssen and AbbVie, outside the submitted work. VAnn reports grants and personal fees (advisory board) from AbbVie and MSD; personal fees (advisory board) from Takeda, Hospira, and Janssen; research grants from Sofia and Giuliano, outside the submitted work. RKW reports grants from the European Research Council (ERC); grants from the fryckling Foundation, National Health and Medical Research Council, Australia, and by the European Community (5th PCRDT). We acknowledge the following colectomy.

Work supported by Wellcome Trust grants 083948/Z/07/2, 085475/Z/08/Z, and 085475/Z/08/2. North American collections and data processing were supported by funds to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium, which is funded by the following grants: DK062431 (SRB), DK063422 and DK062429 (LPS, JHC), DK062420 (RHD), DK062432 (JDR), DK062423 (MSS), DK076984 and DK084554 (MJD). Additional funds were provided by funding to JHC (DK062429-S1 and Crohn’s & Colitis Foundation of America, Senior Investigator Award [S-2229]), and RHD (CA14743). IBD Research at Cedars-Sinai is supported by grant PO1DK046736 and the Cedars-Sinai F Widjaja Foundation Inflammatory Bowel and Immunology Research Institute Fund Research. DPBM is supported by DK062413, A0107608, and U54DK023798-01, grant HS021747 from the Agency for Healthcare Research and Quality, grant 303510 from the European Union, and the Leona M and Harry B Helmsley Charitable Trust. We also acknowledge the University of Pittsburgh Genomics and Proteomics Core Laboratories who undertook genotyping for the Immunophenotypic project. IC and SV are supported by an FWO grant from the Flemish Funds for Scientific Research (FWO) and received funding from the Interuniversity Attraction Poles programme of the Belgian federal scientific policy office, project P7/43 BeMGI. The University of Liege, Belgium is funded by the Walloon Region (IPSEQ, Croh’s & CIBLES projects), by the FEDER, by the Politique Scientifique Fédérale (IAF BeMGI), by the Fonds National de la Recherche Scientifique (FNRS), and by the Communauté Française de Belgique (ARC IBD@ULg). The University of Liege thanks the GIGA-R genotranscriptomic platform for their contribution to genotyping. RKW is supported by a VIDI grant (016.136.308) from the Netherlands Organisation for Scientific Research (NWO). This study was also supported by the German Ministry of Education and Research through the National Genome Research Network, the PopGen biobank, through the Deutsche Forschungsgemeinschaft (DFG) cluster of excellence “Inflammation at Interfaces” and DFG grant no. FR 2821/2-1. SB was supported by DFG BR 1912/6-1 and the Else-Kröner-Fresenius-Stiftung (Else Kröner-Exzellenzstipendium 2010_EKES.32). Italian case collections were supported by the Italian Group for IBD and the Italian Society for Paediatric Gastroenterology, Hepatology and Nutrition and funded by the Italian Ministry of Health GR-01/08 1144485. Activities in Sweden were supported by the Swedish Society of Medicine, Ilere Foundation, Orebro University Hospital Research Foundation, Karolinska Institute, the Swedish National Program for IBD Genetics, the Swedish Organization for IBD, and the Swedish Research Council (JH 521-2011-751). SB was supported by SHS Aabenraa, Denmark. We acknowledge funding provided by the Royal Brisbane and Women’s Hospital Foundation, National Health and Medical Research Council, Australia, and by the European Community (5th PCRDT). We acknowledge the following colectomy. Published biological samples or data for this study: the Inflammatory Bowel disease in South Eastern Norway (IBSEN) study group, including Bjørn Moum, Morten Vatn, and Marie Lie Hovik, is acknowledged for providing patients with inflammatory bowel disease and data updates; the Avon Longitudinal Study of Parents and Children; the Human Biological Data Interchange and Diabetes UK; and Banco Nacional de ADN, Salamanca. This research also uses resources provided by the Type 1 Diabetes Genetics Consortium, a collaborative clinical study sponsored by the NIDDK, National Institute of Allergy and Infectious Diseases, National Human Genome Research Institute, National Institute of Child Health and Human Development, and Juvenile Diabetes Research Foundation, and supported by U91 DK062418.

References


