Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: a genetic association study


lethargy), which frequently lead to poor psychosocial 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...
Inflammatory Bowel Disease
Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, Australia (Prof J M Andrews MD); School of Medicine, University of Adelaide, Adelaide, Australia (Prof J M Andrews); Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCS-CSS) Hospital, San Giovanni Rotondo, Italy (V Annessi MD); Azienda Ospedaliero Universitaria (AOU) Careggi, Unit of Gastroenterology, SOD2, Florence, Italy (V Annessi); Department of Medicine II, University Hospital Munich-Grosshadern, Ludwig-Maximilians-University, Munich, Germany (Prof S Brand MD); Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA (S R Brant MD); Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA (S R Brant); Department of Genetics, Yale School of Medicine, New Haven, CT, USA (Prof H Cho MD); Broad Institute of MIT and Harvard, Cambridge, MA, USA (Prof M J Daly PhD); Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, CA, USA (M Dubinsky MD); Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA (Prof R H Duerer MD); Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA (Prof R H Duerer); School of Medical Sciences, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand (Prof L R Ferguson PhD); Department of Medicine, University of Otago, Christchurch, New Zealand (Prof L R Ferguson); Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand (Prof L R Ferguson); Center for Applied Genomics, University of Auckland, Auckland, New Zealand; Institute of MIT and Harvard, Department of Human Genetics, Cambridge, MA, USA (Prof S Brand); Department of Medicine, University of Otago, Christchurch, New Zealand (Prof R H Geary PhD); Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand (Prof R H Geary); University of Auckland, Auckland, New Zealand (Prof L R Ferguson); and University of Auckland, Auckland, New Zealand (Prof R H Geary). 

Studies and patients
We acquired phenotype data for 34819 patients, including 19713 with Crohn’s disease and 14463 with ulcerative colitis. The cohort included all patients in different centres over the years who had inflammatory bowel disease as per Leonard-Jones’ criteria.16 All inclusion criteria are included in appendix A. After quality control (appendix A), the primary analysis included 29838 patients (16902 with Crohn’s disease, 12597 with ulcerative colitis, 255 with indeterminate colitis, and 84 missing an exact diagnosis). This study includes patients from population-based registries, and 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).59 For behaviour and surgery in Crohn’s disease, and colectomy in ulcerative colitis, Kaplan-Meier

Research in context
Evidence before this study
We searched PubMed for genotype-phenotype association studies in inflammatory bowel disease, published between Jan 1, 1996, and Oct 17, 2014, with the search terms “inflammatory bowel disease AND "genotype phenotype” AND “genetic association” AND (“disease course” or “disease extent” or “location behaviour”). We found 31 studies with sample sizes of 66 to 2804 patients, and studying between one and 163 genetic variants. Most of these studies implicated NOD2 and HLA in subphenotypes of inflammatory bowel disease. Many studies, especially the early ones, studied only NOD2 and HLA, and none included genetic variants not previously implicated in risk for inflammatory bowel disease. 

Added value of this study
This study is the largest genotype-phenotype study of inflammatory bowel disease by at least a factor of ten, and is among the largest studies of genetic determinants of clinical subphenotypes of any complex disease. We have refined the known associations (for instance, the effect of NOD2 on disease location is entirely driven by its association with disease location) and discovered one new associated locus (3p21/MST1 with age at diagnosis). We have explored the genetic relations between subtypes of inflammatory bowel disease with genetic risk scores for the first time, and have shown that ileal and colonic Crohn’s disease are at least as genetically distinct from each other as they are from ulcerative colitis.

Implications of all the available evidence
Established genetic factors can only explain a small fraction of the variability in subphenotype of inflammatory bowel disease, but genetic risk scores that capture all this information could be used to identify misdiagnosed patients. Future translational and clinical research should move away from a binary classification of inflammatory bowel disease into ulcerative colitis and Crohn’s disease, instead considering ileal and colonic Crohn’s disease as separate disease entities.

Methods
Study design and patients
We acquired phenotype data for 34819 patients, including 19713 with Crohn’s disease and 14463 with ulcerative colitis. The cohort included all patients in different centres over the years who had inflammatory bowel disease as per Leonard-Jones’ criteria.16 All inclusion criteria are included in appendix A. After quality control (appendix A), the primary analysis included 29838 patients (16902 with Crohn’s disease, 12597 with ulcerative colitis, 255 with indeterminate colitis, and 84 missing an exact diagnosis). This study includes patients from population-based registries, and 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).59 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.

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 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<5×10⁻⁸) was required for individual single nucleotide polymorphisms (SNPs) and HLA types.

All signals that showed suggestive association (p<1×10⁻³) 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 variants was about OR 1.1, with a median minor allele frequency of roughly 30%. The sample size of our study gave us 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, 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”, 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 Children’s Hospital of Philadelphia, Philadelphia, PA, USA (Prof H Halpernson PhD); Department of Gastroenterology, Faculty of Medicine and Health, Örebro University, Sweden (Prof J Halfvarson MD); School of Health and Medical Sciences, Örebro University, Örebro, Sweden (Prof J Halfvarson); Norwegian FSC Research Center, Research Institute of Internal Medicine and Department of Transplantation Medicine, Oslo University Hospital and University of Oslo, Oslo, Norway (Prof L Fjellanger PhD); Gastrointestinal Unit, Institute of Genetics and Molecular Medicine (N A Kennedy PhD, Prof J Satsangi FRSE, CW Lees PhD), and Child Life and Health (Prof D C Wilson PhD), University of Edinburgh, Edinburgh, UK; Department of Gastroenterology, Lithuanian University of Health Sciences, Kaunas, Lithuania (Prof L Kupininskas MD); Centre for Inflammatory Bowel Diseases, Saint John of God Hospital, Sosiaus WA and School of Medicine and Pharmacology, University of Western Australia, Harry Perkins Institute for Medical Research, Murdoch, WA, Australia (Prof C Lawrence PhD); Inflammatory Bowel Disease Research Group, Addenbrooke’s Hospital, University of Cambridge, Cambridge, UK (J C Lee PhD, M Parkes PhD); Unit of Animal Genomics, Groupe interdisciplinaire de Genétophènomique Appliquée (GIGA-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 (Prof I C Lawrance PhD); Department of Interdisciplinary de Genétophènomique Appliquée (GIGA-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 (J R Hov PhD); 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, Edinburgh, UK; Department of Gastroenterology, Lithuanian University of Health Sciences, Kaunas, Lithuania (Prof L Kupininskas MD); Centre for Inflammatory Bowel Diseases, Saint John of God Hospital, Sosiaus WA and School of Medicine and Pharmacology, University of Western Australia, Harry Perkins Institute for Medical Research, Murdoch, WA, Australia (Prof C Lawrence PhD); Inflammatory Bowel Disease Research Group, Addenbrooke’s Hospital, University of Cambridge, Cambridge, UK (J C Lee PhD, M Parkes PhD); Unit of Animal Genomics, Groupe interdisciplinaire de Genétophènomique Appliquée (GIGA-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 (Prof I C Lawrance PhD); 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>Phenotypes</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>
</table>

### Demographics

#### Sex
- Male: 7,227 (44%)
- Female: 9,257 (56%)
- Missing: 418 (3%)

#### Age at diagnosis (years)
- Median (quartiles): 25 (19–36)
- <17 (A1): 2,568 (18%)
- 17–40 (A2): 9,166 (64%)
- ≥40 (A3): 2,626 (18%)

#### Smoking status
- Non-smoker: 6,752 (58%)
- Smoker: 3,319 (28%)

#### Family history
- Yes: 3,471 (27%)
- No: 9,575 (73%)

#### Smoking status
- Yes: 3,553 (82%)
- No: 7,130 (18%)

#### Disease location
- Ileal (L1): 3,786 (22%)
- Colorectal (L2): 2,533 (14%)
- Ileocolonic (L3): 5,520 (43%)
- Colon (L2): 2,933 (17%)
- Ileal (L1): 3,858 (23%)
- Extensive disease (E3): 5,166 (31%)

#### Disease extent
- Proctitis (E1): 1,271 (12%)
- Left-sided (E2): 4,087 (38%)
- Extensive (E3): 2,321 (18%)
- Missing: 2,777 (18%)

### Disease behaviour
- Inflammatory (B1): 6,196 (50%)
- Stenosing (B2): 3,250 (26%)
- Penetrating (B3): 3,054 (24%)
- Surgery (E4): 3,250 (26%)
- Missing: 3,040 (18%)

<table>
<thead>
<tr>
<th>Disease behaviour</th>
<th>Crohn’s disease</th>
<th>Ulcerative colitis</th>
<th>Inflammatory bowel disease*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inflammatory (B1)</td>
<td>6,196 (50%)</td>
<td>-</td>
<td>8,360 (28%)</td>
</tr>
<tr>
<td>Stenosing (B2)</td>
<td>3,250 (26%)</td>
<td>-</td>
<td>5,260 (18%)</td>
</tr>
<tr>
<td>Penetrating (B3)</td>
<td>3,054 (24%)</td>
<td>-</td>
<td>4,662 (16%)</td>
</tr>
<tr>
<td>Surgery (E4)</td>
<td>3,250 (26%)</td>
<td>-</td>
<td>3,250 (11%)</td>
</tr>
</tbody>
</table>

*Includes 255 patients with indeterminate colitis and 84 patients with missing exact diagnosis.

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). Refining 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<0.001; figure 1; appendix A). Overall, 7,257 (52%) of 13,862 patients with Crohn’s disease had undergone surgery by the time of the last 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^-4; figure 1).

We tested genetic variants for 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^-6) 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...
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>
<thead>
<tr>
<th>Loci</th>
<th>MAF Age at diagnosis of IBD</th>
<th>Age at diagnosis of Crohn’s disease</th>
<th>Age at diagnosis of ulcerative colitis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3p21 (MST1)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs25261698</td>
<td>0.306</td>
<td>6.34 x 10^-14 -0.06 (0.01)*</td>
<td>3.65 x 10^-14 -0.06 (0.01)</td>
</tr>
<tr>
<td>rs2172252</td>
<td>0.288</td>
<td>1.35 x 10^-14 -0.06 (0.01)*</td>
<td>2.93 x 10^-14 -0.07 (0.01)*†</td>
</tr>
<tr>
<td>rs3197999</td>
<td>0.281</td>
<td>2.73 x 10^-14 -0.06 (0.01)*</td>
<td>2.37 x 10^-14 -0.07 (0.01)*</td>
</tr>
<tr>
<td><strong>6p21 (MHC)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3155674</td>
<td>0.116</td>
<td>3.42 x 10^-14 -0.03 (0.01)</td>
<td>3.35 x 10^-14 -0.04 (0.02)</td>
</tr>
<tr>
<td>rs431651</td>
<td>0.034</td>
<td>...</td>
<td>1.15 x 10^-14 -0.07 (0.03)</td>
</tr>
<tr>
<td>rs3129931</td>
<td>0.209</td>
<td>1.15 x 10^-14 -0.05 (0.01)</td>
<td>1.43 x 10^-14 -0.09 (0.02)*†</td>
</tr>
<tr>
<td>rs926832</td>
<td>0.393</td>
<td>7.42 x 10^-14 -0.05 (0.01)*</td>
<td>2.19 x 10^-14 -0.04 (0.01)</td>
</tr>
<tr>
<td>rs482044</td>
<td>0.401</td>
<td>...</td>
<td>1.51 x 10^-14 0.03 (0.01)</td>
</tr>
<tr>
<td><strong>16q12 (NOD2)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2066844 (p.R702W)</td>
<td>0.045</td>
<td>3.58 x 10^-14 -0.08 (0.02)</td>
<td>1.21 x 10^-14 -0.1 (0.02)</td>
</tr>
<tr>
<td>rs2066845 (p.G908R)</td>
<td>0.016</td>
<td>2.10 x 10^-14 -0.1 (0.03)</td>
<td>5.50 x 10^-15 -0.09 (0.03)</td>
</tr>
<tr>
<td>rs2066847 (p.L1007fsX)</td>
<td>0.024</td>
<td>6.64 x 10^-14 -0.16 (0.02)*†</td>
<td>2.04 x 10^-14 -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.

Table 2: Associations between genotype and age at diagnosis achieving genome-wide significance
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 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 0.29–0.34; 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). rs7700575 was independently associated with Crohn’s disease behaviour (p=1.56×10⁻⁹; figure 2; appendix A). Notably, alleles associated with susceptibility to ulcerative colitis, ileal involvement, and ileocolonic involvement 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 (3.50×10⁻⁹; 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, MHC, 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.

### Table 3: Associations between genotype and disease location, behaviour, extent, surgery, and colectomy achieving genome-wide significance

<table>
<thead>
<tr>
<th>MAF Crohn’s disease</th>
<th>Ulcerative colitis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease location</strong></td>
<td><strong>Disease behaviour</strong></td>
</tr>
<tr>
<td>p value</td>
<td>OR (95% CI)</td>
</tr>
<tr>
<td><strong>3’p12 (MST1)</strong></td>
<td></td>
</tr>
<tr>
<td>rs2127552</td>
<td>0.288</td>
</tr>
<tr>
<td>rs3197999</td>
<td>0.281</td>
</tr>
<tr>
<td><strong>6’p21 (MHC)</strong></td>
<td></td>
</tr>
<tr>
<td>rs3115674</td>
<td>0.016</td>
</tr>
<tr>
<td>rs4151651</td>
<td>0.034</td>
</tr>
<tr>
<td>rs6937077</td>
<td>0.112</td>
</tr>
<tr>
<td>rs319891</td>
<td>0.209</td>
</tr>
<tr>
<td>rs9268832</td>
<td>0.393</td>
</tr>
<tr>
<td>rs482044</td>
<td>0.401</td>
</tr>
<tr>
<td>rs7700575</td>
<td>0.439</td>
</tr>
<tr>
<td><strong>3’q21 (NOD2)</strong></td>
<td></td>
</tr>
<tr>
<td>rs206844</td>
<td>0.045</td>
</tr>
<tr>
<td>rs206845</td>
<td>0.016</td>
</tr>
<tr>
<td>rs206847</td>
<td>0.024</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 B2 versus B1, which is also equivalent to B3 versus 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>0.05). *Genome-wide significant associations. †The most significant association per locus per subphenotype, if genome-wide significant.

6 www.thelancet.com Published online October 19, 2015 http://dx.doi.org/10.1016/S0140-6736(15)00465-1
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=1·65×10⁻⁷⁸, or p=9·23×10⁻¹⁸ 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. 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. 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 versus 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
under the curve of only 0·63 (0·59–0·66; appendix A). Comparison of the clinical characteristics of patients with Crohn’s disease that fell into the extreme tails of the genetic risk score (low: \(\log[CD \text{ vs } UC] \leq -2\); high: \(\log[CD \text{ vs } UC] \geq 3\)) showed significant differences in disease location and behaviour (appendix A), suggesting that although a genetic risk score might not be able to classify all patients, it could be informative at the extremes.

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 S), and 1–1% for disease extent in ulcerative colitis (table S).

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. Past studies in inflammatory bowel disease have established a genetic component of disease sub-phenotype, 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, 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 strictureing 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.

---

**Table 5: Variance explained by demographic and genetic factors for disease extent in adult onset of ulcerative colitis**

<table>
<thead>
<tr>
<th>Beta</th>
<th>SE</th>
<th>p value</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever smoking</td>
<td>0·1268</td>
<td>0·0856</td>
<td>1·38 (\times 10^{-1})</td>
</tr>
<tr>
<td>Current smoking</td>
<td>0·1229</td>
<td>0·1179</td>
<td>2·97 (\times 10^{-2})</td>
</tr>
<tr>
<td>Age at diagnosis</td>
<td>-0·0234</td>
<td>0·0055</td>
<td>2·41 (\times 10^{-4})</td>
</tr>
<tr>
<td>Year of birth</td>
<td>0·0068</td>
<td>0·0055</td>
<td>8·79 (\times 10^{-2})</td>
</tr>
<tr>
<td>rs3115674</td>
<td>0·3795</td>
<td>0·0832</td>
<td>5·08 (\times 10^{-6})</td>
</tr>
<tr>
<td>Genetic risk score*</td>
<td>0·0784</td>
<td>0·0502</td>
<td>1·18 (\times 10^{-1})</td>
</tr>
<tr>
<td>Genetic parameters</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Genetics and smoking</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>All parameters†</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*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.

---

**Figure 4: Histograms of Crohn’s disease versus ulcerative colitis (CD vs UC) genetic risk score in patients with inflammatory bowel disease**

Risk scores created from the 163 known inflammatory bowel disease risk loci with per-locus contributions estimated to maximally distinguish all Crohn’s disease from ulcerative colitis. Distributions of ulcerative colitis samples are shown in blue, ileal Crohn’s disease samples in green, and colonic Crohn’s disease with hatched lines (middle area in dark green shows overlap of blue and green distributions). The overlap of all three distributions shows the shared genetic aetiology of inflammatory bowel disease, and the intermediate position of colonic Crohn’s disease between ulcerative colitis and ileal Crohn’s disease shows that it is genetically distinct from the others. Vertical dashed lines show boundaries for outlier analysis: ulcerative colitis cases above 2 were selected as being likely Crohn’s disease and Crohn’s disease cases below –2 as likely ulcerative colitis.
Our results for ulcerative colitis accord with the previously reported independent associations between the MHC and both extensive disease and colectomy.\textsuperscript{16,17} 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,\textsuperscript{44,45} a disease often associated with an extensive but quiescent form of ulcerative colitis.\textsuperscript{46} 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,\textsuperscript{47} 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.\textsuperscript{48}

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.\textsuperscript{49} 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.
phenotype data. SZ, TA, V, and DPBM established DNA collections, genotyping, and data management. IC, GC, LG, LPS, SZ, RHD, AF, JCL, ET, RKW, JDR, and DPBM established DNA collections, genotyping, and data management. LPS assembled, managed, and validated all submitted phenotype data. SZ, TA, VAnn, JMA, VAnn, SRB, SRB, JHC, MD, RHD, LRF, RG, HHu, JH, JRH, JRH, NK, LC, ICL, JCI, JS, ET, AEVdM-D, 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, G, LG, LPS, SZ, TA, SRB, JHC, MD, AF, RBG, HG, HHu, JRH, HHu, LC, ICL, JCI, SS, ET, DCW, MP, SV, JDR, JM, GR-S, and DPBM declare no competing interests. VAnn reports personal fees from Merck (MD: 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 Sofas and Giulianis, outside the submitted work. RHD reports personal fees from Genentech (Etrolizumab advisory committee), outside the submitted work. LRF reports grants for database collection and analysis from New Zealand Ministry for Business and Innovation, during the conduct of the study. JH reports grants from the Swedish Research Council, Örebro University Hospital Research Foundation, and the Ile Foundation, during the conduct of the study. NAK reports grants from the Wellcome Trust; personal fees and non-financial support from MSD; personal fees from Warner Chilcott and Ferring; non-financial support from Shire and AbbVie, outside the submitted work. JH reports grants from EC-FFP award, Medical Research Council, Wellcome Trust, CSO Scotland, Cunningham Trust, and AbbVie; and personal fees from Takeda, MSD, and Ferring UK, outside the submitted work. AEvD M-dj reports grants from AbbVie; personal fees from Tramedico and Takeda; and non-financial support from Ferring (eHEALTH program), outside the submitted work. RKW reports grants from the Netherlands Organisation for Scientific Research (NWO, VIDI grant 016.136.308), during the conduct of the study. MSS reports grants from National Institute of Diabetes and Digestive and Kidney Diseases, and Crohn’s and Colitis Canada, during the conduct of the study. JCB reports grants from the Wellcome Trust, during the conduct of the study; and personal fees from Genomics Ltd (consulting), outside the submitted work. CWL reports grants and personal fees from AbbVie; personal fees from MSD, Hospira, Vifor Pharma, Pharmacosmos, Takeda, Shire, and Janssen, outside the submitted work.

Acknowledgements
We thank all the individuals who contributed samples and the physicians and nursing staff who helped with recruitment worldwide. We also thank Prof Gil McVean (professor of Statistical Genetics, Wellcome Trust Centre of Human Genetics, University of Oxford, UK) for thorough review of all statistical methods. UK case collections were supported by the National Association for Colitis and Crohn’s disease; Wellcome Trust grants 098801 (JCB) and 098759 (LG); Medical Research Council UK; the Catherine Myfanwy Foundation; Peninsula College of Medicine and Dentistry, Exeter; the National Institute for Health Research, through the Comprehensive Local Research Network, and through Biomedical Research Centre awards to Guy’s and St Thomas’ National Health Service Trust, King’s College London, Addenbrooke’s Hospital, University of Cambridge School of Clinical Medicine and to the University of Manchester and Central Manchester Foundation Trust. The Wellcome Trust Case Control Consortium projects were supported by Wellcome Trust grants 083948/Z/07/Z, 085475/Z/08/Z, and 085475/Z/08/Z. 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 DK063420 (RHD), LB62429 (LPS, JHC), DK063420 (RHD), DK063422 (JDR), DK062423 (MSS), DK067984 and DK084554 (MJD). Additional funds were provided by funding to JHC (DK062429-S1 and Crohn’s & Colitis Foundation of America, Senior Investigator Award [5-2229]), and RHD (CA14743). IBD Research at Cedars-Sinai is supported by grant PO1DK046763 and the Cedars-Sinai F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute Research Funds. DPBM is supported by DK062413, AI067068, and U54DK027389-01, grant HS021747 from the Agency for Healthcare Research and Quality, grant 305497 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 Immunoscop 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 BeMeGI. The University of Liège, Belgium is funded by the Walloon Region (IPSEQ, Crohn & CIBLES projects), by the FEDER, by the Politique Scientifique Fédérale (IAP BeMeGI), by the Fonds National de la Recherche Scientifique (FNRS), and by the Communauté Française de Belgique (ARC IBd@ULg). The University of Liège thanks the GIGA-R gastroentrocsystemic platform for their contribution to the project. 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.12). 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 Grant GR-2008-114485. Activities in Sweden were supported by the Swedish Society of Medicine, Ilre Foundation, Örebro University Hospital Research Foundation, Karolinska Institute, the Swedish National Program for IBD Genetics, the Swedish Organization for IBD, and the Swedish Research Council (JH 521-2011-0665). We acknowledge a grant from Viborg Regional Hospital, Denmark. VAnn 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’s 7th PCRDIT. We acknowledge the following groups that provided biological samples or data for this study: the Inflammatory Bowel disease in South Eastern Norway (IBSEN) study group, including Bjørn Mourn, Morten Vatn, and Marte 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 U01 DK062418.

References
37 Jostins L, Levine AP, Barrett JC. Using genetic prediction from known complex disease loci to guide the design of next-generation sequencing experiments. PloS One 2013; 8: e76128.