Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7

Yang Luo1–3,26, Katrina M de Lange1,26, Luke Jostins4,5, Loukas Moutsianas1, Joshua Randall1, Nicholas A Kennedy6,7, Christopher A Lamb8, Shane McCarthy1, Tariq Ahmad6,7, Cathryn Edwards9, Eva Goncalves Serra1, Ailsa Hart10, Chris Hawkey11, John C Mansfield12, Craig Mowat13, William G Newman14,15, Sam Nichols1, Martin Pollard1, Jack Satsangi16, Alison Simmons17,18, Mark Tremelling19, Holm Uhlig17,20, David C Wilson21,22, James C Lee23, Natalie J Prescott24, Charlie W Lees16, Christopher G Mathew24,25, Miles Parkes23, Jeffrey C Barrett1,27 & Carl A Anderson1,27

To further resolve the genetic architecture of the inflammatory bowel diseases ulcerative colitis and Crohn's disease, we sequenced the whole genomes of 4,280 patients at low coverage and compared them to 3,652 previously sequenced population controls across 73.5 million variants. We then imputed from these sequences into new and existing genome-wide association study cohorts and tested for association at ~12 million variants in a total of 16,432 cases and 18,843 controls. We discovered a 0.6% frequency missense variant in ADCY7 that doubles the risk of ulcerative colitis. Despite good statistical power, we did not identify any other new low-frequency risk variants and found that such variants explained little heritability. We detected a burden of very rare, damaging missense variants in known Crohn's disease risk genes, suggesting that more comprehensive sequencing studies will continue to improve understanding of the biology of complex diseases.

Crohn's disease and ulcerative colitis, the two common forms of inflammatory bowel disease (IBD), are chronic and debilitating diseases of the gastrointestinal tract that result from the interaction of environmental factors, including the intestinal microbiota, with the host immune system in genetically susceptible individuals. Genome-wide association studies (GWAS) have identified 215 IBD-associated loci that have substantially expanded understanding of the biology underlying these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases1–8. The correlation between nearby common variants in human populations underpins the success of the GWAS approach, but this also makes it difficult to infer precisely which variant is causal, the molecular consequence of that variant, and often even which gene is perturbed. Rare variants, which plausibly underlie these diseases.
sequencing studies concentrated on the protein-coding sequence in GWAS-implicated loci, which can be naturally extended to the entire exome. However, coding variation explains at most 20% of the common variant associations in IBD GWAS loci, and others have more generally observed that the substantial majority of variants associated with complex disease lie in noncoding, presumed regulatory regions of the genome. Low-coverage whole-genome sequencing has been proposed as an alternative approach that captures this important noncoding variation while being inexpensive enough to enable thousands of individuals to be sequenced. As expected, this approach has proven valuable in exploring rarer variants than those accessible in GWAS, but it is not ideally suited to the analysis of extremely rare variants.

Our aim was to determine whether low-coverage whole-genome sequencing provides an efficient means of interrogating these low-frequency variants and how much these variants contribute to IBD susceptibility. We present an analysis of the whole-genome sequences of 4,280 patients with IBD and 3,652 population controls sequenced as part of the UK10K project, both via direct comparison of sequenced individuals and as the basis for an imputation panel in an expanded UK IBD GWAS cohort. This study allows us to examine, on a genome-wide scale, the role of low-frequency (0.1% ≤ minor allele frequency (MAF) < 5%) and rare (MAF < 0.1%) variants in IBD risk.

RESULTS
Whole-genome sequencing of 7,932 individuals
Following quality control (Supplementary Tables 1 and 2, and Supplementary Note), the whole-genome sequences of 2,513 patients with Crohn’s disease (median coverage of 4×) and 1,767 patients with ulcerative colitis (2×) were jointly analyzed with 3,652 population controls sequenced as part of the UK10K project (Fig. 1). We discovered 87 million autosomal single-nucleotide variants (SNVs) and 7 million short indels (Supplementary Table 3 and Supplementary Note). We then applied support vector machines (SVMs) for SNVs and Genome Analysis Toolkit (GATK) VQSR for indels to distinguish true sites of genetic variation from sequencing artifacts (Fig. 1 and Supplementary Note). We called genotypes jointly across all samples at the remaining sites, followed by genotype refinement using the BEAGLE imputation software. This procedure leverages information across multiple individuals and uses the

![Figure 1](https://example.com/figure1.png)

**Figure 1** Overview of our study. Variants were called from raw sequence reads in three groups of samples and jointly filtered using SVMs. The resulting genotypes were refined using BEAGLE and incorporated into the reference panel for a GWAS imputation-based meta-analysis, which discovered a low-frequency association in *ADCY7*. A separate gene-based analysis identified a burden of rare, damaging variants in certain known Crohn’s disease susceptibility genes.
correlation between nearby variants to produce high-quality data from relatively low sequencing depth. We noted that genotype refinement was locally affected by poor-quality sites that failed further quality control analyses, so we ran BEAGLE a second time after these exclusions, yielding a set of 73.5 million high-quality sites (Supplementary Figs. 1–3, Supplementary Table 4, and Supplementary Note). Over 99% of common SNVs (MAF ≥5%) were also found in 1000 Genomes Project Phase III Europeans, indicating high specificity. Among rarer variants, 54.6 million were not seen in the 1000 Genomes Project, demonstrating the value indicating high specificity. Among rarer variants, 54.6 million were also found in 1000 Genomes Project Phase III Europeans, indicating high specificity. Among rarer variants, 54.6 million were not seen in the 1000 Genomes Project, demonstrating the value of directly sequencing the IBD cases and UK population controls (Supplementary Table 5). Over 99% of common SNVs (MAF ≥5%) were also found in 1000 Genomes Project Phase III Europeans, indicating high specificity. Among rarer variants, 54.6 million were not seen in the 1000 Genomes Project, demonstrating the value of directly sequencing the IBD cases and UK population controls (Supplementary Table 5).

We also discovered 180,000 deletions, duplications and multial- lelic copy number variants (CNVs) using GenomeStrip 2.0 (ref. 26), but we noted large differences in sensitivity between the three different sample sets (Supplementary Fig. 4). After quality control (Supplementary Note), including removal of CNVs with length <60 kb, we observed an approximately equal number of variants in cases and controls but retained only 1,475 CNVs. However, we still note a genome-wide excess of rare CNVs in controls (P = 0.002), indicating that even after stringent filtering the data remain too noisy for meaningful conclusions to be drawn. We suggest that high-coverage whole-genome sequencing balanced in cases and controls will be required to evaluate the contribution of rare CNVs to IBD risk.

We individually tested 13 million SNVs and small indels with MAF ≥0.1% for association and observed that we had successfully eliminated systematic differences due to sequence depth (λ1,000_UC = 1.05, λ1,000_CD = 1.04, λ1,000_IBD = 1.06; Supplementary Fig. 5), while still retaining power to detect known associations. While we estimate that this stringent quality control produced well-calibrated association test statistics for more than 99% of sites, this analysis yielded many extremely significant P values at SNPs outside of known loci (for example, ~7,000 SNPs with P < 1 × 10−15), 95% of which had an allele frequency below 5%. In contrast to GWAS, where routine procedures almost completely eliminate false positive associations, the heterogeneity of our sequencing depths makes it challenging to discern true associations from these data alone.

**Imputation into GWAS**

As noted by a previous study of type 2 diabetes27 with a similar design, our whole-genome sequencing data set alone is not well powered to identify new associations, even if all samples were sequenced at the same depth. We therefore built a phased reference panel of 10,971 individuals from our low-coverage whole-genome sequences and 1000 Genomes Project Phase III haplotypes (Supplementary Note), to use imputation to leverage IBD GWAS to increase our power. Previous data have shown that such expanded reference panels significantly improve the imputation accuracy of low-frequency variants28. We next generated a new UK IBD GWAS data set by genotyping 8,860 patients with IBD not having previous GWAS data and combining them with 9,495 UK controls from the Understanding Society project, all genotyped using the Illumina HumanCoreExome v12 chip. We then added previous UK IBD GWAS samples that did not overlap with those in our sequencing data set29,30. Finally, we imputed all of these samples using the PBWT31 software and the reference panel described above, and we combined these imputed genomes with our sequenced genomes to create a final data set of 16,267 IBD cases and 18,843 UK population controls (Supplementary Table 6).

This imputation produced high-quality genotypes at 12 million variants that passed typical GWAS quality control (Supplementary Note), and these variants represented more than 90% of sites with MAF >0.1% that we could directly test in our sequences. In comparison to the most recent meta-analysis by the International IBD Genetics Consortium1, which used a reference panel almost ten times smaller than ours, we tested an additional 2.5 million variants for association with IBD. Because our GWAS cases and controls were genotyped using the same arrays, they should not be differentially affected by the variation in sequencing depths in the reference panel and thus not susceptible to the artifacts observed in the sequence-only analysis. Indeed, in comparison to the thousands of false positive

**Figure 2** Association analysis for the NOD2–ADCY7 region on chromosome 16. Results from the single-variant association analysis are presented in gray, and results after conditioning on seven known NOD2 risk variants are shown in blue. Top, Crohn’s disease (CD); bottom, ulcerative colitis (UC). The dashed red lines indicate genome-wide significance, at α = 5 × 10−8.
associations present in the sequence-only analysis, the imputation-based meta-analysis identified only four previously undescribed genome-wide significant IBD associations. Three of these had MAF >10%, so we carried them forward to a meta-analysis of our data and published IBD GWAS summary statistics32.

p.Asp439Glu in ADCY7 doubles risk of ulcerative colitis

The fourth new association (P = 9 × 10^-12) was a missense variant (p.Asp439Glu, rs78534766) in ADCY7 with a frequency of 0.6% that doubled risk of ulcerative colitis (odds ratio (OR) = 2.19, 95% confidence interval (CI) = 1.75–2.74) and is strongly predicted to alter protein function (SIFT = 0, PolyPhen = 1, MutationTaster = 1). This variant was associated (P = 1 × 10^-6) in a subset of directly genotyped individuals, suggesting that the signal was unlikely to be driven by imputation errors. To further validate it, we genotyped (Online Methods) an additional 450 ulcerative colitis cases and 3,905 controls (P = 0.00169) and looked it up in 982 ulcerative colitis cases and 136,464 controls from the UK Biobank (P = 0.0189). A meta-analysis of all three directly genotyped data sets showed genome-wide significant association (P = 1.6 × 10^-9), no evidence for heterogeneity (P = 0.19) and clean cluster plots (Supplementary Fig. 6 and Supplementary Table 7).

Low-frequency variation makes a minimal contribution to IBD susceptibility

The associated variant in ADCY7 represents precisely the class of variant that our study design was intended to probe: MAF below 1%, OR ~2 and difficult to impute (only one copy of the non-reference allele was observed in the Phase I 1000 Genomes Project and INFO = 0.7 when imputing32 from Phase III), making it notable as our single discovery of this type. We had 66% power to detect this association and reasonable power even for more difficult scenarios (for example, 29% power for variants with MAF of 0.2% and OR = 2, or 11% power for variants with MAF of 0.5% and OR = 1.5). As noted by others39, heritability estimates for low-frequency variants as a class are exquisitely sensitive to potential bias from technical and population differences. We therefore analyzed only the imputed GWAS samples to eliminate the effect of differential sequencing depth and applied a more stringent SNP and sample quality control procedure (Supplementary Fig. 7 and Supplementary Note). We used the restricted maximum likelihood (REML) method implemented in GCTA40 and estimated that autosomal SNPs with MAF >0.1% explain 28.4% (standard error (s.e.) = 0.016) and 21.1% (s.e. = 0.012) of the variation in liability for Crohn’s disease and ulcerative colitis, respectively. Despite SNPs with MAF <1% representing approximately 81% of the variants included in this analysis, they explain just 1.5% of the variation in liability. While these results are underestimates owing to limitations of our data and the REML approach, it seems very unlikely that a large fraction of IBD risk is captured by variants like ADCY7 p.Asp439Glu. Thus, our discovery of ADCY7 actually
serves as an illustrative exception to a series of broader observations\(^\text{41}\) that low-frequency, high-risk variants are unlikely to be important contributors to IBD risk.

The role of rare variation in IBD risk

Our low-coverage sequencing approach does not perfectly capture very rare and private variants because the cross-sample genotype refinement adds little information at sites where nearly all individuals are homozygous for the major allele. Similarly, these variants are difficult to impute from GWAS data: even using a panel of more than 32,000 individuals offers little imputation accuracy below a MAF of 0.1\% (ref. 28). Thus, although our sequence data set was not designed to study rare variants, it is the largest so far in IBD and has sufficient specificity and sensitivity to warrant further investigation (Supplementary Fig. 8). Because enormous sample sizes would be required to implicate any single variant, we used a standard approach from exome sequencing\(^\text{42}\), where variants of a particular functional class are aggregated into a gene-level test. We extended the robust variance score statistic of Derkach \textit{et al.}\(^\text{43}\) to account for our sequencing depth heterogeneity because existing rare variant burden methods gave systematically inflated test statistics.

For each of 18,670 genes, we tested for a differential burden of rare (MAF ≤ 0.5\% in controls, excluding singletons) functional or predicted damaging coding variation in our sequenced cases and controls (Online Methods and Supplementary Tables 8 and 9). We detected a significant burden of damaging rare variants in the well-known Crohn’s disease risk gene NOD2 (\(P_{\text{functional}} = 1 \times 10^{-7}\); Supplementary Fig. 9), which was independent of the known low-frequency NOD2 risk variants (Online Methods). We noted that the additional variants (Fig. 3) that contributed to this signal explain only 0.13\% of the variance in disease liability, as compared to the 1.15\% explained by the previously known variants\(^\text{11}\), underscoring the fact that very rare variants cannot account for much population variability in risk.

Some genes implicated by IBD GWAS had suggestive \(P\) values but did not reach exome-wide significance (\(P = 5 \times 10^{-7}\); Supplementary Table 10), so we combined individual gene results into two sets: (i) 20 genes that had been confidently implicated in IBD risk by fine-mapping or functional data and (ii) 63 additional genes highlighted by less precise GWAS annotations (Supplementary Table 11 and Supplementary Note). We tested these two sets (after excluding NOD2, which otherwise dominates the test) using an enrichment procedure\(^\text{44}\) that allows for differing direction of effect between the constituent genes (Supplementary Table 12 and Supplementary Note). We found a burden in the 12 confidently implicated Crohn’s disease genes that contained at least one damaging missense variant (\(P_{\text{damaging}} = 0.0045\)). By contrast, we saw no signal in the second, more generic set of genes (\(P = 0.94\); Fig. 4 and Table 1).

We extended this approach to evaluate rare regulatory variation, using enhancer regions described by the FANTOM5 project (Supplementary Table 13). Within each robustly defined enhancer\(^\text{44}\), we tested all observed rare variants, as well as the subset predicted to disrupt or create a transcription factor binding motif\(^\text{18}\). We combined groups of enhancers with cell-type- and/or tissue-type specific expression, to improve power in an analogous fashion to the gene set tests above. However, none of these tissue- or cell-type-specific enhancer sets had a significant burden of rare variation after correction for multiple testing (Supplementary Table 14).

DISCUSSION

We investigated the role of low-frequency variants of intermediate effect in IBD risk through a combination of low-coverage whole-genome sequencing and imputation into GWAS data (Fig. 5).

![Figure 5](attachment:image.png)
We discovered an association with a low-frequency missense variant in *ADCY7*, which represents one of the strongest ulcerative colitis risk alleles outside of the major histocompatibility complex. The most straightforward mechanistic interpretation of this association is that loss of function of *ADCY7* reduces production of cAMP, leading to an excessive inflammatory response that predisposes to IBD. Previous evidence suggested that general cAMP-elevating agents that act on multiple adenylyl cyclases might, in fact, worsen IBD. While members of the adenylyl cyclase family have been considered potential targets in other contexts, specific upregulation of *ADCY7* has not yet been attempted, raising the intriguing possibility that altering cAMP signaling in a leukocyte-specific way might offer therapeutic benefit in IBD.

To maximize the number of patients with IBD we could sequence, and thus our power to detect association, we sequenced our cases at lower depth than the controls available to us via managed access. While joint and careful analysis largely overcame the bias this difference in coverage introduced, this is just one example of the complexities associated with combining sequencing data from different studies. Such challenges are not just restricted to low-coverage whole-genome sequencing study designs; variable pulldown technology and sequencing depth in the 60,000 exomes in the Exome Aggregation Consortium necessitated a simultaneous analysis of such analytical complexity and computational intensity that it would be prohibitive at all but a handful of research centers. Therefore, if rare variant association studies are to be as successful as those for common variants, computationally efficient methods and accepted standards for combining sequence data sets need to be developed.

We have participated in one such joint analysis by contributing to the Haploype Reference Consortium (HRC), which has collected whole-genome sequencing data from more than 32,000 individuals into a reference panel that allows accurate imputation of low-frequency and common variants. Indeed, imputation into GWAS from the HRC is as accurate as low-coverage sequencing at allele frequencies as low as 0.05% (ref. 28), so by far the most effective way to discover complex disease associations with variants in this range is to reanalyze the huge quantities of existing GWAS data with improved imputation. Although projects like ours have provided wider public benefit through the HRC, there is little need for future low-coverage whole-genome sequencing projects in complex disease.

Despite our study being specifically designed to interrogate both coding and noncoding variation, our sole newly discovered association was with a missense variant. This is perhaps unsurprising, as the only previously identified IBD risk variants with similar frequencies and odds ratios are protein-altering changes to NOD2, IL23R and CARD9. More generally, the alleles with largest effect sizes at any given frequency tend to be coding and are therefore the first to be discovered when new technologies expand the frequency spectrum of genetic association studies. This pattern is further reinforced by the contrast between the tantalizing evidence we found for a burden of very rare coding variants in previously implicated IBD genes and the absence of any signal across the enhancer regions we tested. This distinction emphasizes how dramatically better we can distinguish likely functional from neutral variants in coding as compared to noncoding sequence. For example, if we include all rare coding variants (MAF ≤ 0.5% in controls, n = 136) in IBD genes, the P value is 0.2291, as compared to P = 0.0045 when using the subset of 54 coding variants with CADD ≥ 21. Therefore, the identification of rare variant burdens in the non-coding genome will require not only tens of thousands of samples to be sequenced, but also much better discrimination between functional and neutral variants in regulatory regions.

Nonetheless, it is likely that rare variants have an important role in IBD risk and that many such alleles are regulatory, as is the case for common risk variants. The *ADCY7* association offers a direct window on a new IBD mechanism, but it would probably eventually have been discovered through HRC imputation in existing GWAS samples and is a relatively meager return in comparison to the number of loci discovered more simply by increasing GWAS sample size. Making real progress on rare variant association studies will require much larger numbers of deeply sequenced exomes or whole genomes, especially if ‘ultra-rare’ variants are as important in IBD as they are in, for example, schizophrenia. Extrapolating for IL23R, the IBD-associated gene with the most significant coding burden (P = 0.0005) after NOD2, we would require roughly 20,000 cases to reach genome-wide significance; as we noted above, the challenge is even greater for noncoding regions, where functional variants cannot currently be distinguished from neutral. Together, our discoveries suggest that a combination of continued GWAS, coupled to new imputation reference panels, and large-scale deep sequencing studies will be needed to complete understanding of the genetic basis of complex diseases.

URLs. Understanding Society project, http://www.understandingsociety.ac.uk/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all individuals who contributed samples to the study. This work was co-funded by the Wellcome Trust (098051) and the Medical Research Council, UK (MR/J00314X/1). Case collections were supported by Crohn’s and Colitis UK. K.M.d.L., L.M., Y.L., C.A.L., C.A.A. and J.C.B. are supported by the Wellcome Trust (098051; 093885/2/10/Z). K.M.d.L. is supported by a Woolf Fisher Trust scholarship. C.A.L. is a clinical lecturer funded by the NIHR. H.U. is supported by the Crohn’s and Colitis Foundation of America (CCFA) and the Leona M. and Harry B. Helmsley Charitable Trust. We acknowledge support from the UK Department of Health via NIHR comprehensive Biomedical Research Centre awards to Guy’s and St Thomas’ NHS Foundation Trust in partnership with King’s College London and to Addenbrooke’s Hospital, Cambridge, in partnership with the University of Cambridge, and the BRC to the Oxford IBD cohort study, University of Oxford. This research was also supported by the NIHR Newcastle Biomedical Research Centre. The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The survey was conducted by NatCen, and the genome-wide scan data were analyzed and deposited by the Wellcome Trust Sanger Institute. Information on how to access the data can be found on the Understanding Society website. We are grateful for genotyping data from the British Society for Surgery of the Hand Genetics of Dupuytren’s Disease consortium and L. Southam for assistance with genotype intensities. This research has been conducted using the UK Biobank Resource.

AUTHOR CONTRIBUTIONS

Y.L., K.M.d.L., L.J., L.M., J.C.B. and C.A.A. performed statistical analysis. Y.L., K.M.d.L., L.J., L.M., J.C.L., C.A.L., E.G.S., J.R., M. Pollard, S.N. and S.M. processed the data. T.A., C.E., N.A.K., A.H., C.H., J.C.M., J.C.L., C.M., W.G.N., J.S., A.S., M.T., H.U., D.C.W., N.J.P., C.W.L., M. Parkes and C.G.M. contributed samples and/or materials. Y.L., K.M.d.L., L.M., J.C.L., M. Parkes, C.A.L., N.A.K., J.C.B. and C.A.A. wrote the manuscript. All authors read and approved the final version of the manuscript. J.C.M., M. Parkes, C.W.L., T.A., N.J.P., J.C.B. and C.A.A. conceived and designed experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
CONVERGE Consortium. Sparse whole-genome sequencing identifies two loci for Crohn’s disease in a Japanese population. Gastroenterology 144, 781–788 (2013).

Anderson, C.A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat. Genet. 43, 246–252 (2011).

Kenny, E.E. et al. A genome-wide scan of Ashkenazi Jewish Crohn’s disease suggests novel susceptibility loci. PLoS Genet. 8, e1002559 (2012).

Julia, A. et al. A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis. Hum. Mol. Genet. 23, 6927–6934 (2014).

Yang, S.-K. et al. Genome-wide association study of Crohn’s disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut 63, 80–87 (2014).

Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat. Genet. 48, 510–518 (2016).

Manolio, T.A. et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009).

Zuk, O. et al. Searching for missing heritability: designing realistic null variants association studies. Proc. Natl. Acad. Sci. USA 111, E455–E464 (2014).

Rivas, M.A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43, 1066–1073 (2011).

Beaudoin, M. et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLoS Genet. 9, e1003723 (2013).

Hunt, K.A. et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498, 232–235 (2013).

Prescott, N.J. et al. Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTN2L2 and implicates other immune related genes. PLoS Genet. 11, e1004955 (2015).

Do, R. et al. Exome sequencing identifies rare LDLR and APD5A alleles conferring risk for myocardial infarction. Nature 518, 102–106 (2015).

De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515, 209–215 (2014).

Singh, T. et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat. Neurosci. 19, 571–577 (2016).

Huang, H. et al. Association mapping of inflammatory bowel disease loci to single variant resolution. Preprint at bioRxiv http://dx.doi.org/10.1101/0286685 (2015).

Farh, K.K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).

Li, Y. et al. Effects of forskolin on Kupffer cell production of interleukin-10 and tumor necrosis factor α differ from those of endogenous adenylyl cyclase activators: possible role for adenylyl cyclase 9. Infect. Immun. 73, 7290–7296 (2005).

Duan, B. et al. Distinct roles of adenylyl cyclase VII in regulating the immune responses in mice. J. Immunol. 185, 335–344 (2010).

Jiang, L.-J., Sternweis, P.C. & Wang, J.E. Zymosan activates protein kinase A via adenylyl cyclase VII to modulate innate immune responses during inflammation. Mol. Immunol. 54, 14–22 (2013).

Risase, P.K. et al. Higher TNFα responses in young males compared to females are associated with attenuation of monocyte adenylyl cyclase expression. Hum. Immunol. 76, 427–430 (2015).

Pierre, S., Eschenhagen, T., Geisslinger, G. & Scholich, K. Capturing adenylyl cyclases as potential drug targets. Nat. Rev. Drug Discov. 8, 321–335 (2009).

Bhatia, G. et al. Subtle stratification confounds estimates of heritability from rare variants. Preprint at bioRxiv https://dx.doi.org/10.1101/048181 (2016).

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).

Chen, G.-B. et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and Immunochip data. Hum. Mol. Genet. 23, 4710–4720 (2014).

Purcell, S.M. et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506, 185–190 (2014).

Derkach, A. et al. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. Bioinformatics 30, 2179–2188 (2014).

Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).

Zimmerman, N.P., Kumar, S.N., Turner, J.R. & Dwinell, M.B. Cyclic AMP dysregulates intestinal epithelial cell restitution through PKA and RhoA. Inflamm. Bowel Dis. 18, 1081–1091 (2012).

Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

Genovese, G. et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat. Neurosci. 19, 1433–1441 (2016).

Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124 (2012).
Online Methods

Preparation of genome-wide genetic data. Sample ascertainment and sequencing. UK IBD cases, diagnosed using accepted endoscopic, histopathological and radiological criteria, were sequenced to low depth (2–4×) using Illumina HiSeq paired-end sequencing. Population controls, also sequenced to low depth (7×) using the same protocol, were obtained from the UK10K project. Supplementary Table 2 provides details on sample numbers and quality control filters. Case sequence data were aligned to the human reference used in Phase II of the 1000 Genomes Project (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase1_reference_assembly_sequence/hs37d5.fa.gz). Control data were aligned to an earlier human reference (1000 Genomes Project Phase I) (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_gatk_v37.fasta.gz) and then updated to the same reference as the cases using BridgeBuilder, a tool we developed (Supplementary Note).

Genotype calling and quality control. Variants were jointly called across 8,424 samples, using SAMTools and BCFTools for SNVs and indels and GenomeSTRIP for CNVs. CNVs were filtered using standard GenomeSTRIP quality metrics as described in the Supplementary Note. SNVs were filtered using SVMs trained on variant quality statistics output from SAMTools. Each variant was required to pass with a minimum score of 0.01 in at least two of five independent SVM models. Indels were filtered using gATK VQSR, with a truth sensitivity threshold of 97% (VQSLOD score of 1.0659).

Genotype refinement and further quality control. After initial call, SNV and indel quality control, genotypes at all sites that passed were refined via BEAGLE23. Variants were then filtered again to remove those showing significant evidence of deviation from Hardy–Weinberg equilibrium in controls ($P_{\text{HWE}} < 1 \times 10^{-7}$), a significant frequency difference ($P < 1 \times 10^{-3}$) in samples sequenced at the Wellcome Trust Sanger Institute versus the Beijing Genomics Institute or $>10$% missing genotypes after refinement (posterior probability $< 0.9$) and those within 3 bp of an indel. We also filtered to allow only one indel to pass when clusters of indels were separated by two or fewer base pairs. After these exclusions, a second round of genotype refinement was performed. Sample quality control was then applied to remove samples with an excessive heterozygosity rate ($\mu \pm 3.5\sigma$), duplicated or related individuals, and individuals of non-European ancestry (Supplementary Fig. 10 and Supplementary Note).

New GWAS samples. A further 11,768 UK IBD cases and 10,484 population control samples were genotyped on the Human Core Exome v12 chip. Detailed information on ascertainment, genotyping and quality control are provided elsewhere23.

Existing GWAS cohorts. 1,748 Crohn’s disease cases and 2,936 population controls genotyped on the Affymetrix 500k chip, together with 2,361 ulcerative colitis cases and 5,417 population controls genotyped on the Affymetrix 6.0 array, were obtained from the Wellcome Trust Case Control Consortium (WTCCC)29,30. Both data sets were converted to Build 37 using liftOver49.

Imputation. The whole-genome sequences described above were combined with 2,504 samples from the Phase III v5 release of the 1000 Genomes Project (2013-05-02 sequence freeze) to create a phased imputation reference panel enriched in IBD-associated variants. We used PBWT23 to impute from this reference panel (114.2 million total variants) into the three GWAS panels described above, after removing overlapping samples. This resulted in imputed whole-genome sequences for 11,987 cases and 15,189 controls (Supplementary Table 6).

Common and low-frequency variation association testing. Association testing and meta-analysis. We tested for association with ulcerative colitis, Crohn’s disease and IBD separately within the sequenced samples and three imputed GWAS panels using SNPTTEST v2.5, performing an additive frequency-trend association test conditioned on the first ten principal components for each cohort (calculated after exclusion of the MHC region). We filtered out variants with MAF $<0.1\%$, INFO $<0.4$ or strong evidence of deviations from Hardy–Weinberg equilibrium in controls ($P_{\text{HWE}} < 1 \times 10^{-7}$), and we then used METAL (release 2011-03-05)50 to perform a standard-error-weighted meta-analysis of all four cohorts. Only sites for which all cohorts passed our quality control filters were included in the meta-analysis.

Quality control. The output of the fixed-effects meta-analysis was further filtered, and sites with strong evidence of heterogeneity ($P > 0.90$) were discarded. In addition, we discarded all genome-wide significant variants for which the meta-analysis $P$ value was not lower than all of the cohort-specific $P$ values. Finally, to minimize the number of false positive associations due to misimputation, sites that did not have an INFO score $>20.8$ in at least three of the four data sets (two of the three for Crohn’s disease and ulcerative colitis) were removed.

Locus definition. A linkage disequilibrium (LD) window was calculated for every genome-wide significant variant in any of the three traits (Crohn’s disease, ulcerative colitis and IBD), defined by the leftmost and rightmost variants that were correlated with the main variant with $r^2 > 0.6$ or greater. LD was calculated in the GBR (British) and CEU (European) samples from 1000 Genomes Project Phase III, release v5 (using the 20130502 sequence freeze and alignments). Loci with overlapping LD windows, as well as loci whose lead variants were separated by 500 kb or less, were subsequently merged, and the variant with the strongest evidence of association was kept as the lead variant for each merged locus. This process was conducted separately for each trait. A locus was annotated as known when there was at least one variant in it that was previously reported (Supplementary Table 15) to be of genome-wide significance (irrespective of the LD between that variant and the most associated variants in the locus) and as novel otherwise.

Conditional analysis. Conditional analyses were conducted using SNPTEST v2.5 (ref. 51), as for the single-variant association analysis. $P$ values were derived using the score test (default in SNPTEST v2.5). To fully capture the NOD2 signal when investigating the remaining signal in the region, we conditioned on seven variants that are known to be associated (rs2066844, rs2066845, rs2066847, rs2797637, rs2357623, rs18478383 and rs104985944).

Replication of the ADCYT association. After quality control32, an additional 450 UK ulcerative colitis cases and 3,905 population controls (Dupuytren’s contracture cases), genotyped using Illumina Human Core Exome array v12, were available for replication. An additional 982 ulcerative colitis cases and 136,464 controls from the UK Biobank, genotyped on either the UK Biobank Axiom or UK BiLEVE array, formed a second replication cohort. Quality control for the UK Biobank data was performed as previously described (http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf), and individuals who were not British or Irish were excluded from further analysis. Cases were defined as those with self-reported ulcerative colitis or an ICD-10 code of K51 in their Hospital Episode Statistics (HES) record. Controls were defined as individuals without self-diagnosis or a hospital record of ulcerative colitis or Crohn’s disease (HES code K50). Logistic regression conditional on ten principal components was carried out in both replication cohorts. We used METAL (release 2011-03-05)50 to perform a standard-error-weighted meta-analysis of all three directly genotyped cohorts.

Heritability explained. SNP heritability analysis was performed on the dichotomous case–control phenotype using constrained REML in GCTA40 with prevalence of 0.005 and 0.0025 for Crohn’s disease and ulcerative colitis, respectively. Hence, all reported values of $h^2$ are on the underlying liability scale. To further eliminate spurious associations, we computed genetic relationship matrices (GRMs) restricted to all variants with MAF $\geq 0.1\%$, imputation $r^2 \geq 20.6$, missing rate $\leq 1\%$ and Hardy–Weinberg equilibrium $P \leq 1 \times 10^{-7}$ in controls for each GWAS cohort. We further checked the reliability and robustness of our estimates by performing a joint analysis across all autosomes, a joint analysis between common (MAF $\geq 1\%$) and rare (0.1% $\leq$ MAF $< 1\%$) variants, and LD-adjusted analysis using LDAR28 (Supplementary Fig. 7, Supplementary Table 16 and Supplementary Note).

Rare variation association testing. Additional variant quality control. Additional site filtering was undertaken, as rare variant association studies are more susceptible to differences in read depth between cases and controls (Supplementary Fig. 11). This filtering included removing singlets, as well as sites with (i) missingness rate $>9.9$, when the rate was calculated using genotype probabilities estimated from the SAMtools genotype quality (GQ) field; (ii) low-confidence observations comprising $>1\%$ of non-missing data; or (iii) INFO $<0.6$ in the appropriate cohorts.

Association testing. Individual gene and enhancer burden tests were performed using an extension of the robust variance score statistic35 (Supplementary Note), to adjust for the systematic bias in coverage between cases and controls. This required the estimation of genotype probabilities directly from SAMTools (using the GQ score) as genotype refinement using
imputation results in poorly calibrated probabilities at rare sites. Burden
tests were performed across sites with MAF ≤0.5% in controls and within
genes defined by Ensembl or enhancers selected on the basis of inclusion in
the FANTOM5 ‘robustly defined’ enhancer set. For each gene, two sets
of burden tests were performed: (i) a set with all functional coding variants
and (ii) a set with all functional coding variants predicted to be damaging
(CADD ≥ 21) (Supplementary Table 8). For each enhancer, burden tests were
repeated to include all variants falling within the region or just the subset of
variants predicted to disrupt or create a transcription factor binding motif
(Supplementary Note).

NOD2 independence testing. We evaluated the independence of the rare
NOD2 signal from that for the known common coding variants in this gene
(rs2066844, rs2066845 and rs2066847). Individuals with a minor allele at any
of these sites were assigned to one group, and those with reference genotypes
were assigned to another. Burden testing was performed for this new pheno-
type in both variant sets that contained a significant signal in Crohn’s disease
versus controls.

Set definition. The individual burden test statistic was extended to test across
sets of genes and enhancers using an approach based on the SMP method, whereby the test statistic for a given set is evaluated against the statistics from the complete set (for example, all genes), to account for residual bias
in case–control coverage. The sets of genes confidently associated with IBD
risk were defined on the basis of implication of specific genes in ulcerative
colitis, Crohn’s disease or IBD risk through fine-mapping, eQTL and targeted
sequencing studies (Supplementary Table 11). The broader set of IBD suscep-
tibility genes was defined as any remaining genes implicated by two or more
candidate gene approaches in Jostins et al. Enhancer sets were defined as those
showing positive differential expression in each of 69 cell types and 41
tissues, according to Andersson et al. (Supplementary Table 17).

Data availability. Whole-genome sequence data that supports this study have
been deposited in the European Genome-phenome Archive (EGA) under accessions EGAD00001000409 and EGAD00001000401. Genotype data are
available under accession EGAS00001000924.

49. Hinrichs, A.S. et al. The UCSC Genome Browser Database: update 2006. Nucl. Acids Res. 34, D590–D598 (2006).
50. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of
genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
51. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499–511 (2010).
52. Speed, D. & Balding, D.J. MultiBLUP: improved SNP-based prediction for complex
traits. Genome Res. 24, 1550–1557 (2014).