Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility

Sazonovs, A.; Stevens, C.R.; Venkataraman, G.R.; Yuan, K.; Avila, B.; Abreu, M.T.; ... ; UK IBD Genetics Consortium

nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01156-2

Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility

Genome-wide association studies (GWASs) have identified hundreds of loci associated with Crohn's disease (CD). However, as with all complex diseases, robust identification of the genes dysregulated by noncoding variants typically driving GWAS discoveries has been challenging. Here, to complement GWASs and better define actionable biological targets, we analyzed sequence data from more than 30,000 patients with CD and 80,000 population controls. We directly implicate ten genes in general onset CD for the first time to our knowledge via association to coding variation, four of which lie within established CD GWAS loci. In nine instances, a single coding variant is significantly associated, and in the tenth, *ATG4C*, we see additionally a significantly increased burden of very rare coding variants in CD cases. In addition to reiterating the central role of innate and adaptive immune cells as well as autophagy in CD pathogenesis, these newly associated genes highlight the emerging role of mesenchymal cells in the development and maintenance of intestinal inflammation.

GWASs in CD, and inflammatory bowel disease (IBD) more generally, have successfully identified more than 200 loci contributing to risk of disease[1–4]. While most GWAS hits do not immediately implicate an obvious functional variant or gene, a subset have been directly mapped to coding variants (for example, *NOD2*, *IL23R*, *ATG16L1*, *SLC39A8*, *FUT2*, *TYK2*, *IFIH1*, *SLAMF8*, *PLCG2*)[5], providing more direct clues to pathogenesis. Further, targeted and genome-wide sequencing approaches have revealed additional, lower-frequency, disease-associated coding variants (for example, *CARD9*, *RNF186*, *ADCY7*, *INAVA/C1orf106*, *SLC39A8*, *NOD2*)[6–9] originally undetected by GWASs. Such coding variants, common and rare, have led to functional follow-up experiments demonstrating causal mechanisms for at least ten genes and have provided the most direct biological insights to emerge from genetic studies of IBD[10–13].

Results

To further advance the interpretation of GWAS loci—and to define novel CD-associated genes using variation rarer than that routinely detected by GWASs—we pursued large-scale exome sequencing using CD case and control collections from more than 35 centers in the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). The primary analysis consisted of exome sequencing of 18,816 CD cases across 35 IBD studies and 13,412 non-IBD control samples from the same studies. These samples were all sequenced at the Broad Institute and were supplemented with 22,536 population controls from approved non-IBD studies sequenced contemporaneously at the Broad Institute and accessed from dbGAP (Supplementary Table 1 and Extended Data Fig. 1). Two different exome capture platforms were employed during the course of the study (referred to hereafter as Nextera (Illumina) and Twist (Twist Biosciences)). Details of capture and sequencing of these cohorts (and those subsequently used in follow-up) are provided in the Methods and Supplementary Information.

Calling and quality control (QC) of data from the two exome capture platforms were conducted in parallel (Table 1, Extended Data Fig. 2 and Methods). Sensitivity to detect low-frequency coding variants was evaluated in each callset post-QC by comparison with passing sites in gnomAD v.2.1 that had $0.0001 <$ non-Finnish European (NFE) minor allele frequency (MAF) < 0.1 (Methods). We observed that 84% of all exonic SNPs in this frequency range were detected in both CD datasets with sufficiently high quality to enter meta-analysis. Analysis of each dataset was conducted in SAIGE using a logistic mixed-model[14] and a standard inverse-variance weighted (IVW) meta-analysis conducted across 164,149 nonsynonymous variants with MAF (gnomAD NFE) between 0.0001 and 0.1. A study-wide significance threshold of 3×10^{-7} was applied (Supplementary Table 2). As a control for the entire study, we also analyzed 96,326 synonymous variants. Forty-three sites (Supplementary Table 3) failed a heterogeneity-of-effect test between the Nextera and Twist discovery cohorts (IVW $P_{HET} < 0.0001$) and were eliminated from further analysis. We did not observe an inflation in the exome-wide distribution of test statistics.

Exonic variants significantly associated with IBD. The most significantly associated variants ($P < 10^{-10}$) in the Discovery stage were previously known CD variants (or variants in linkage disequilibrium (LD) with them), indicating that the QC and analysis pipeline removed highly-associated false positives. Twenty-eight variants achieved study-wide significance, including known variants within CD genes established in previous GWASs and sequencing efforts: *NOD2*, *IL23R*, *LRRK2*, *TYK2*, *SLC39A8*, *IRGM* and *CARD9*. Excluding synonymous variants in LD with these known associated nonsynonymous variants, synonymous variants showed little deviation from the null and none reached study-wide significance (Extended Data Fig. 3). Encouraged by this, we then nominated a list of 116 variants (including known variants) with $P < 0.0002$ for further evaluation in three follow-up cohorts (Supplementary Table 2).

Additional exome and genome sequencing was undertaken at the Sanger Institute on an independent cohort of 9,731 CD cases ascertained by the UK IBD Genetics Consortium and IBD BioResource. Genome sequencing with a target depth of 15× was performed on 6,000 patients with CD. Whole-genome sequences from 11,852 individuals from the INTERVAL blood donor cohort were used as population controls. Another 3,731 patients with CD were exome sequenced using the Agilent SureSelect Human All

A full list of affiliations appears at the end of the paper.

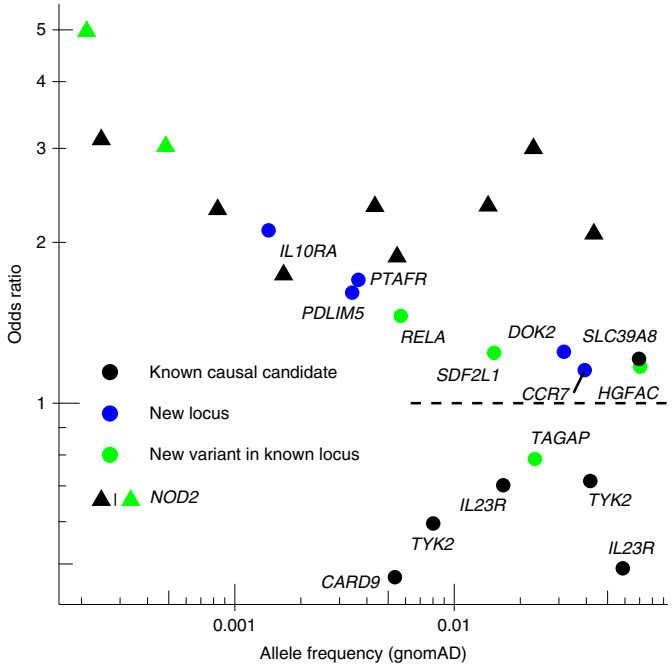


Fig. 1 | Odds ratio and MAF for exome-wide significant findings that are not tagging stronger, established noncoding association signals. Known causal candidate: in a credible set from a fine-mapping study[5] with PIP > 5% or reported in previous studies[6,8] (Methods). New locus: in a locus not yet implicated by GWASs. New variant in known locus: in a known GWAS locus, but represents an association independent from previously reported IBD putative causal variants (Methods).

Table 1 | Sample characteristics

Sequencing center	Data type	Exome capture	Study stage	No. CD cases	No. controls
Broad	WES	Nextera	Discovery	11,125	25,145
Broad	WES	Twist	Discovery	6,109	6,064
Sanger	WGS	N/A	Follow-up	6,000	11,852
Sanger	WES	Agilent	Follow-up	3,731	33,704*
Regeneron	WES	Agilent	Follow-up	4,071	4,223

Numbers listed are post-QC and are of samples entered into the analyses. *WES data derived from the UK Biobank cohort, sequenced by Regeneron using the IDT xGen Exome Research Panel v1.0 including supplemental probes, were used as population controls for the Sanger WES cases (Methods). N/A, not applicable, as no exome capture was used for whole-genome sequencing.

Exon V5 capture. In total, 33,704 individuals without IBD or other related diseases from the UK Biobank were used as controls for the Sanger whole-exome sequencing (WES) cases. These UK Biobank samples were sequenced by Regeneron using the IDT xGen Exome Research Panel v1.0 (including supplemental probes), and thus QC and subsequent analyses were restricted to the intersection of the Agilent and the IDT capture regions. Exome and genome datasets were processed in parallel with similar QC parameters (Methods). Association analyses were performed using a logistic mixed-effects model implemented in the REGENIE v.1.0.6.7 and v.2.0.2 software, correcting for the case–control imbalance using the Firth correction. Of 116 variants, 28 were associated ($P < 4.3 \times 10^{-4}$ (0.05/116)) with CD in the meta-analysis of the two Sanger cohorts and 94 replicated the direction of effect seen in the discovery cohort ($P = 3 \times 10^{-12}$, binomial test). Summary statistics from a German dataset of 4,071 CD cases and 4,223 controls exome sequenced at Regeneron (Methods) were also ascertained and a meta-analysis was carried out across all five cohorts (Table 1). Of the 116 variants, 45 exceeded the study-wide significance threshold, $P < 3 \times 10^{-7}$ (Supplementary Table 4 and Supplementary Data 1). Of note, all 45 study-wide significant sites in the discovery stage exome-wide scan showed stronger evidence of association in the meta-analysis, and none showed significant evidence of heterogeneity of effect across studies. The 'scan' exome-wide and the combined meta-analyses have similar power to detect the same true associations at their respective significance thresholds (Extended Data Fig. 4).

Among the 164,149 low-frequency nonsynonymous variants tested for association in this study, 14 were mapped to a credible set with posterior inclusion probability (PIP) > 5% from a previous IBD fine-mapping study[5]. Eight of the 14 variants reached exome-wide significance (in *NOD2*, *IL23R*, *TYK2*, *PTPN22* and *CARD9*). The remaining six variants (in *GPR65*, *MST1*, *NOD2* and *SMAD3*) have genetic effects consistent with those previously reported, with *P* values ranging from 0.025 to 8×10^{-5} in the WES discovery cohort. Together, these results demonstrate the accuracy of our exome sequencing study in the lowest frequency range covered by previous GWAS approaches.

To identify new loci not yet implicated in CD and independent exonic association signals at known loci, we accounted for the LD between the 45 exome-wide significant variants and previously reported IBD GWAS hits, as well as previous rare-variant discoveries (Methods). We identified five coding variants in genes not previously implicated in IBD, even by their proximity to previous GWAS signals. We also discovered six independently associated novel exonic variants in genes previously known to harbor coding mutations underpinning CD or IBD risk, two of which are in *NOD2* ('New locus' and 'New variant in known locus' in Fig. 1, Supplementary Table 4 and Extended Data Fig. 5). Fourteen significant variants recaptured known IBD causal candidates from fine-mapping, including variants in *CARD9*, *IL23R* and *NOD2*, and the remaining 20 variants either tag the known causal variants through LD or have very small PIP from fine-mapping and thus are highly unlikely to be CD causal variants ('Known causal candidate' and 'Unlikely causal' in Fig. 1, Supplementary Table 4 and Extended Data Fig. 5). A harmonized summary with findings from this study and the fine-mapping study[5] for genes implicated by the 45 exome-wide significant variants is available in Supplementary Table 5. Of note, evidence from earlier GWASs and this study cannot be considered independent and trivially combined since there is considerable sample overlap—while 46% of scan samples come from clinical sites or national cohorts not previously involved in the largest GWAS[4], the remainder are from sites which also contributed earlier recruited samples (generally 10 or more years ago) to previously reported GWASs and have been exome sequenced in this study. Similarly, at least 10% of the Sanger whole-genome sequencing (WGS) samples used for meta-analysis were previously included in past genotyping studies.

Some of the newly implicated CD genes (Table 2 and Box 1) contribute to biological pathways previously implicated via GWASs, such as autophagy (*ATG4C*), or Mendelian forms of IBD, such as the IL-10 signaling pathway[15], or by extensive functional studies of inflammatory response in IBD, such as the NF-κB family of transcriptional regulators[16]. In contrast, many of the newly associated genes appear to be linked to the roles of mesenchymal cells (MCs) in intestinal homeostasis, a pathway not previously implicated by genetic studies. The mesenchyme is composed of non-hematopoietic, nonendothelial and nonepithelial cells such as fibroblasts, myofibroblasts (stromal cells) and pericytes[17]. In the intestine, mucosal MCs act as a second barrier through their interactions with both immune and epithelial cells[18]. Under physiological conditions, MCs regulate immune cell maturation, migration and recruitment of immune cells[19] as well as maintenance of the stem cell niche in the intestinal crypt and mucosal repair through

Table 2 | Novel variants achieving study-wide significance that implicate genes directly in general onset CD

Chr	Pos	A0	A1	MAF	Scan *P*	Meta *P*	OR	Gene	AA conseq.
1	28,150,681	T	C	0.0037	1.68×10^{-07}	2.96×10^{-12}	1.702	*PTAFR*	N114S
4	3,447,925	G	A	0.0704	1.37×10^{-10}	6.92×10^{-15}	1.170	*HGFAC*	R516H
4	94,573,345	C	T	0.0034	1.79×10^{-05}	2.55×10^{-07}	1.610	*PDLIM5*	spl reg
6	159,041,392	C	T	0.0233	7.60×10^{-06}	4.04×10^{-10}	0.786	*TAGAP*	E147K
8	21,909,729	G	A	0.0316	1.34×10^{-04}	2.09×10^{-13}	1.248	*DOK2*	P274L
11	65,658,293	C	T	0.0057	6.23×10^{-05}	2.31×10^{-07}	1.457	*RELA*	D288N
11	117,998,788	C	T	0.0014	1.13×10^{-05}	6.29×10^{-09}	2.107	*IL10RA*	P295L
17	40,558,934	T	C	0.0393	6.16×10^{-06}	4.70×10^{-08}	1.153	*CCR7*	M7V/M1?
22	21,643,991	G	A	0.0152	2.44×10^{-05}	2.21×10^{-07}	1.242	*SDF2L1*	R161H

Four of these variants (in *TAGAP*[1,45], *SDF2L1* (ref. [1]), *RELA*[1] and *HGFAC*[2]) are in regions highlighted in previous GWASs but represent independent associations directly implicating these genes (Methods). Chr, chromosome; Pos, genomic position in hg38; Scan *P*, *P* value from the exome-wide discovery, including subjects exome sequenced at the Broad Institute; Meta *P*, *P* value from the full meta-analysis of the five cohorts shown in Table 1; OR, odds ratio; AA conseq., consequence on the amino acid; A0, reference allele; A1, tested/effect allele; spl reg, splice-site region.

epithelial-to-mesenchymal transition (EMT)[17,18,20]. MCs are highly activated by proinflammatory signals during chronic inflammation, resulting in subepithelial myofibroblasts proliferation and extracellular matrix production, and, not surprisingly, are involved in the development of fibrotic disorders[21].

Among the newly discovered IBD genes, *PDLIM5* is highly expressed in subepithelial myofibroblasts[22]. PDLIM5 is a cytoskeleton-associated protein well known as a regulator of EMT through TGF-β1/SMAD3 signaling[23]. Knocking out its expression also leads to an alteration in extracellular matrix assembly, specifically by decreasing collagen IV network density[23–26] (Fig. 2). Interestingly, *SDF2L1*, also among the newly identified CD genes, has previously been shown to be elevated in plasma cells within the lamina propria of patients with CD failing to achieve durable remission on anti-TNF therapy[27]. *SDF2L1*, which is induced in response to endoplasmic reticulum (ER) stress, is also expressed in Paneth and goblet cells[28]. Impairment of ER stress is closely linked to intestinal inflammation[29]. Therefore, a mutation in this gene could putatively impair epithelial homeostasis in many ways such as preventing goblet cell differentiation, migration and proper production of mucus[30] or perturbing the production of antibodies by plasma cells[31].

HGFAC, *PAF-R* and *CCR7* can be linked to IBD-relevant MC functions via their known ligands. Specifically, HGFAC is a serine protease that cleaves hepatocyte growth factor (HGF) to its active form. It has previously been shown that HGF is a paracrine factor secreted by stromal cells (fibroblasts and myofibroblasts) that regulates epithelial homeostasis—in particular, the balance between epithelial proliferation, differentiation and apoptosis—and has been shown to be elevated in the serum of patients with CD[20,32]. In human kidney epithelial cells, HGF has been shown to have antifibrotic properties by upregulating SMAD co-repressor SnoN, resulting in inhibition of EMT, and likely plays a similar role in the intestinal environment[33]. Platelet activating factor receptor (PTAFR - also commonly known as PAF-R) is expressed in epithelial and endothelial cells as well as in pericytes, a population of MCs surrounding blood vessels that regulates angiogenesis[34,35]. PTAFR is a G-protein-coupled receptor for platelet activating factor (PAF), a proinflammatory lipid that is elevated in the mucosa of patients with CD, potentially reflecting disease activity[36]. The PAF-R/PAF axis is known to regulate endothelial and epithelial permeability, which is associated with inflammatory diseases[37,38]. Finally, CCR7$^+$ immune cells have a macrophage-like or dendritic cell morphology. It is known that mesenchymal stromal cells induce CCR7$^+$ dendritic cell migration to mesenteric lymph nodes within inflamed mucosa of patients with IBD[39]. It is also known that CCR7 ligands, CCL19/CCL21, are highly expressed in a recently identified population of proinflammatory stromal cells that appear to prevent the resolution phase that is normally found as part of the wound-healing process[22]. Our identification of CD-associated rare coding variants in these genes suggests that perturbation of these finely balanced cellular processes that are key to intestinal homeostasis causally contributes to CD susceptibility.

The identified coding variants in *RELA*, *TAGAP* and *SDF2L1* are close to, but not in LD ($r^2<0.05$) with, common noncoding variants significantly associated with IBD risk via GWASs (Methods and *TAGAP* as an example in Extended Data Fig. 6a–c). These very likely pinpoint the genes dysregulated by the associated common variant and provide a focus for uncovering the function of those variants, perhaps leading to allelic series of perturbations further informing on the mechanism of their contribution to CD pathogenesis. The associated missense variant in *HGFAC* is in partial LD ($r^2=0.35$ in 1000 Genomes NFE populations) with a common noncoding variant (rs3752440) previously reported as associated with CD[2]. Unfortunately, the missense variant was not included in this previous study—precluding formal assessment of whether this explains the previously observed association signal or represents an independent variant directly implicating *HGFAC*—although we note the missense variant here has a higher odds ratio and greater significance than the variant in the previous report. The two novel *NOD2* associations are not in LD with previously reported putative causal variants; one modestly reduces basal activity and has at least twofold reduction in peptidoglycan-induced NF-κB response[40], while the other is a splice donor variant (Supplementary Table 6). None of the variants described in Table 2 has reached genome-wide significance in previously published GWASs (variant in *HGFAC* has almost reached significance in ref. [4], $P=5\times10^{-8}$ for IBD). The nine new CD-associated variants all had an info score of 1 in the UK Biobank (UKBB) GWAS imputation[41], except for *PTAFR* and *PDLIM5*, which had info scores of 0.72 and 0.9, respectively. Novel variants described in Table 2 together explain around 0.12% of the variance on the liability scale (0.3% on the observed scale). In comparison, the 25 independent coding variants that were included in the meta-analysis together explained 2.1% of variance (5.1% on the observed scale). We performed two gene-based rare-variant (MAF < 0.001) burden tests in the full-exome Nextera and Twist datasets using SAIGE-GENE[42], one restricted to loss-of-function variants and another using all nonsynonymous variants (Supplementary Table 7). The burden test meta-analysis was performed on 15,823 genes for the nonsynonymous variant analysis and 3,953 for the predicted loss-of-function (pLoF) variant only analysis. Correcting for 20,000 genes, associations with $P<2.5\times10^{-6}$ were considered statistically significant. *NOD2*

Box 1 | Genes newly implicated in CD risk

DOK2 (Docking Protein 2, Downstream of Tyrosine Kinase 2) encodes a cytoplasmic signaling protein highly expressed in macrophages and T cells in the terminal ileum. Loss of Dok-2 in mice causes severe Dextran sulfate sodium-induced colitis with reduced *IL-17A* and *IL-22* expression[51], and *DOK2* is known to be differentially methylated in colonic tissue of patients with IBD[52]. *DOK2* regulates both *TLR2*-induced inflammatory signaling and natural killer cell development, and *DOK2* loss-of-function is associated with increased IFN-ɣ production[53,54]. The P274L variant has previously been implicated in atopic eczema where the rare allele was significantly protective for atopic eczema, likely by disturbing the RasGAP activation of DOK2, and transcriptomic analyses also suggest that *DOK2* is a central hub interacting with *CD200R1*, *IL6R* and *STAT3* (ref. [55]).

TAGAP (T-Cell Activation RhoGTPase Activating Protein) has a pivotal role in TH17 development and modulates the risk of autoimmunity through influencing thymocyte migration in thymic selection[56,57]. *TAGAP* expression is upregulated in rectal tissue in patients with IBD, and *TAGAP* is required for Dectin-induced anti-fungal signaling and proinflammatory cytokine production in myeloid cells[58,59].

PTAFR (Platelet Activating Factor Receptor), a hypoxia response gene, has an affinity for bacterial phosphorylcholine (ChoP) moieties[60] and influences development of cigarette-induced chronic obstructive pulmonary disease by inducing neutrophil autophagic death in mice[61]. *PTAFR* regulates colitis-induced pulmonary inflammation through the NLRP3 inflammasome[62].

PDLIM5 (PDZ And LIM Domain 5) encodes a kidney anion exchanger and scaffolding protein. Genetic variation in this gene is associated with prostate cancer, schizophrenia, diverticular disease, diverticulosis, colorectal cancer, testicular cancer and self-reported angina[63]. *PDLIM5* has been reported to be a *STAT3* interaction partner involved in actin binding[64], with *STAT3* previously being identified as an IBD gene[65]. *PDLIM5* is highly expressed in myofibroblast cells, which are important MCs of the intestinal lamina propria[22].

SDF2L1 (Stromal Cell Derived Factor 2 Like 1) has been recently identified to be expressed in the ER stress response in primary intestinal epithelial cells[66]. *SDF2L1* is an ER resident protein that functions within a large protein complex regulating the BiP and Erdj3 chaperone cycle to promote protein folding and secretion[67–69]. Structurally, Sdf2l1 contains an N-terminal signal peptide for entry into the ER lumen and a C-terminal ER retention signal flanking three MIR domains that promote complex assembly. The CD risk variant R161H is located in the third MIR domain. In murine models, deletion of *SDF2L1* in the liver resulted in prolonged ER stress and insulin resistance[70]. In the intestine, single cell transcriptional profiling revealed that *SDF2L1* is predominantly expressed in highly secretory cell lineages, including mucin-secreting goblet cells and immunoglobulin-secreting plasma cells[71]. Moreover, *SDF2L1* expression is dynamically regulated and specifically induced during the acute phase of the unfolded protein response[66]. Together, these observations suggest a critical role for *SDF2L1* in maintaining ER homeostasis and secretory capacity, which may promote barrier function at the level of mucus integrity and/or neutralization of immunoglobulins and antimicrobial peptides that collectively limit interactions between luminal microbes and the host immune system.

CCR7 (chemokine receptor 7) encodes a chemokine receptor. CCR7 and its ligands CCL19/CCL21 promote homing of T cells and dendritic cells to T cell areas of lymphoid tissues where T cell priming occurs. *CCR7* also contributes to adaptive immune functions including thymocyte development, secondary lymphoid organogenesis, high-affinity antibody responses, regulatory and memory T cell function, and lymphocyte egress from tissues. *CCR7* expression is upregulated in an inflamed gut in CD[72], and *CCR7* regulates the intestinal TH1/TH17/Treg balance during Crohn's-like murine ileitis[73]. Genetic variation in *CCR7* is associated with atopy, asthma, chronic obstructive pulmonary disease and IBD in African-Americans[63]. *CCL19* and *CCL21* are highly expressed in a population of stromal cells (designated as S4) that are expanded in IBD inflamed tissues and that continually produce proinflammatory factors, preventing the resolution phase of the wound-healing response[22].

IL10RA (Interleukin 10 receptor A) is a potent regulator of innate and adaptive immune responses, and *IL10RA* genetic variants are associated with very-early onset IBD cases; a subset of very-early onset IBD refractory patients respond to hematopoietic stem cell transplantation[74]. *IL10RA* (also known as *Il10r1)* knockout mice are susceptible to chemical-induced colitis[75].

RELA (Nuclear Factor NF-κB P65 Subunit). NF-κB is a ubiquitous transcription factor, and its most abundant form is NFKB1 complexed together with RELA. *RELA* regulates the Th17 pathway in autoimmune disease models[76], and the FOXO3-NF-κB RelA protein complexes reduce proinflammatory cell signaling and function[77]. *RELA* haploinsufficiency causes autosomal-dominant chronic mucotaneous ulceration[78], and *RELA* is a master transcriptional regulator of EMT in epithelial cells[79]. Genetic variation in *RELA* has been associated with systemic lupus erythematosus, type 2 diabetes, psoriasis, obesity, asthma and atopic dermatitis[63].

ATG4C (Autophagy-Related 4C Cysteine Peptidase) defective autophagy is established as a mechanism contributing to CD risk. This gene encodes one of four Atg4 isoforms (Atg4A, B, C and D) that prime pro-LC3 and GABARAP (orthologues of yeast Atg8), essential proteins required for autophagosome biogenesis[80,81]. These Atg4 proteins, including Atg4C, are involved with proteolytic cleavage of Atg8's C terminus, thus exposing a specific Atg8 glycine residue necessary for phospholipid covalent binding to Atg8. Atg8 lipidation is necessary for autophagosome formation[82].

HGFAC (Hepatocyte Growth Factor Activator) is a serine endopeptidase that converts HGF to its active form in response to thrombin and kallikrein endopeptidases. HGF contributes to neutrophil recruitment. HGF expression is increased in active ulcerative colitis with animal models, suggesting that HGF-MET signaling exacerbates intestinal inflammation[83,84]. Furthermore, HGF promotes colonic epithelial regeneration and mucosal repair[85,86]. *HGFAC* variation is also associated with tuberculosis susceptibility[87].

unsurprisingly stood out far above the expected distribution (P values from gene-based burden tests using pLoF and non-synonymous variants are 7.7×10^{-7} and $<10^{-16}$, respectively). Only one other gene in either analysis exceeded the threshold expected once in the study by chance (*ATG4C*, $\text{NonSyn}_p=3.3\times10^{-6}$). This potentially novel signal in *ATG4C* was driven by three distinct missense variants with individual $P<0.01$ (N75S, R80H and C367Y) (Supplementary Table 8) along with two others with $P<0.05$ (K371R, R389X). The *ATG4C* gene burden signal was examined in the Sanger datasets and replicated, with the meta-analysis reaching exome-wide significance ($P=1.5\times10^{-7}$) driven by several of the same variants. Further examination of results from the single-variant tests

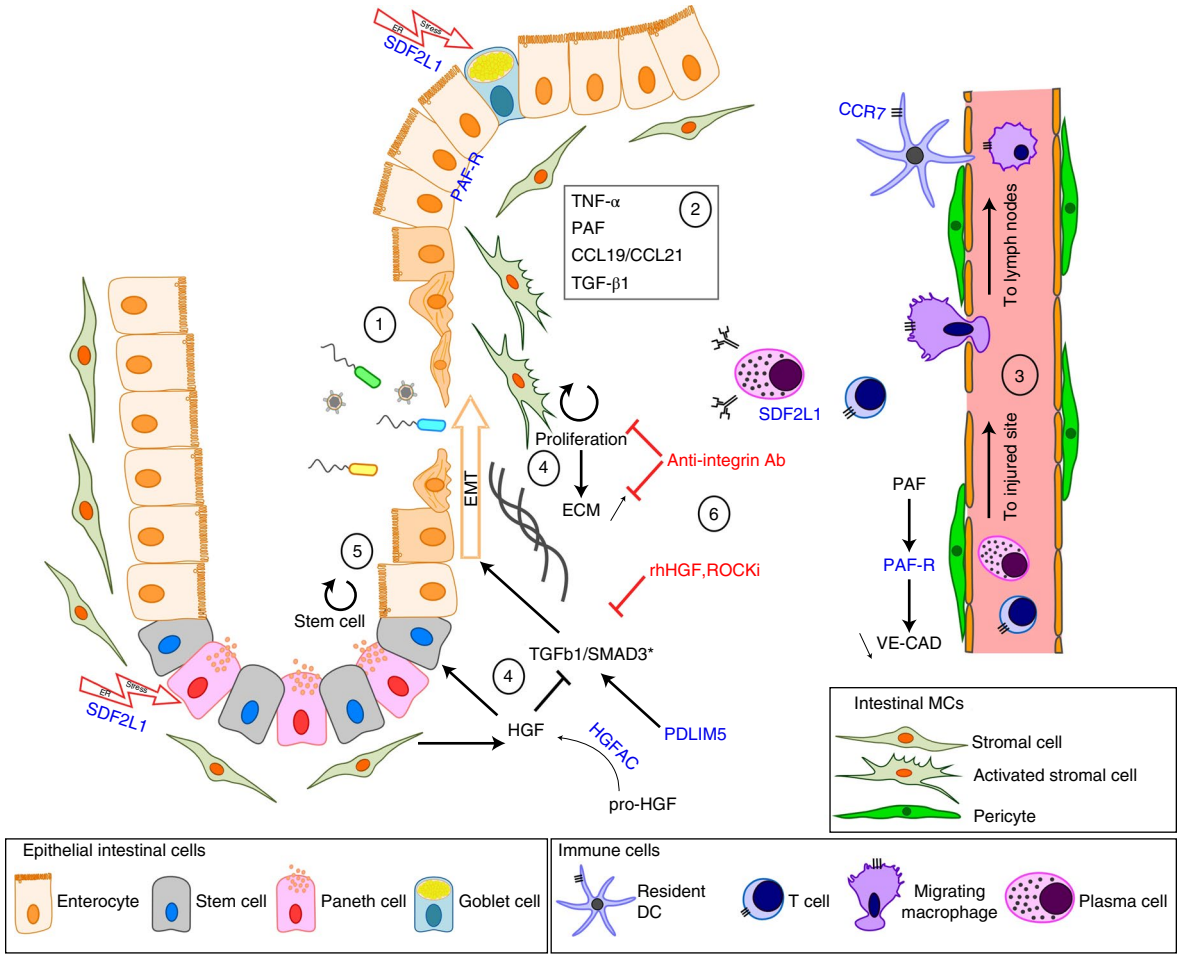


Fig. 2 | Schematic representation of inflamed mucosa showing the mesenchymal-related genes with newly identified mutations. (1) Following mucosal injury, MCs are highly activated by proinflammatory signals such as TNF-α, CCL19/CCL21, PAF and TGF-β1. (2) Among these, TNF-α increases PTAFR expression in intestinal epithelial cells during wound repair[46]. However, prolonged exposure to PAF dissolves cell junctions and increases epithelial permeability[37]. In endothelial cells, the PAF-R/PAF axis has a similar effect on vascular endothelial-cadherin (VE-CAD) assembly[34]. A leaky endothelium can result in an increase in immune cell infiltration and aggravate inflammation at injured sites[47]. (3) Secretion of CCL19/CCL21 by activated stromal fibroblasts in response to epithelial damage or infection attracts dendritic cells (DCs) and other immune cells which then migrate to mesenteric lymph nodes, where the immune response is coordinated. CCR7$^+$ DCs, macrophages and T cells also exacerbate inflammation through proinflammatory mediators such as CCL19/CCL21. Plasma cells express SDF2L1 in response to inflammation and ER stress due to massive antibody production[48]. (4) Mucosal repair mediated by TGF-β1/SMAD3 signaling has a key role in epithelial homeostasis after tissue injury[49]. Importantly, a causal variant in *SMAD3* further supports the importance of this pathway in disease susceptibility[5]. PDLIM5 is a known regulator of SMAD3 stability during EMT[23]. Uncontrolled EMT increases fibroblast proliferation and excessive extracellular matrix (ECM) production leading to fibrosis[43]. Active HGF released by HGFAC antagonizes TGF-β1 resulting in a decrease of EMT. (5) HGF secreted by fibroblasts plays a role in maintaining the stem cell niche in intestinal crypts[50]. SDF2L1 expressed by Paneth cells in response to ER stress may participate in this process. (6) The current genetic findings provide support to the scientific rationale for targeting EMT and fibrosis for the treatment of CD, such as with anti-integrin antibodies (anti-αvβ6), recombinant human HGF (rhHGF) and Rho kinase inhibitor (ROCKi); shown in red[22].

in *ATG4C* identified a frameshift variant with frequency of 0.002 (1:62834058-TTG-T)—too high to be included in our burden test—that just missed our threshold for testing in the follow-up cohorts ($P=0.0003$, Beta = 0.55 in the Broad meta-analysis). This variant also showed evidence of association in the meta-analysis of the Sanger cohorts ($P=1.3\times10^{-5}$), and also exceeded our study-wide significance threshold in the five-way meta-analysis of all cohorts ($P=1.55\times10^{-8}$). Of further note, an additional *ATG4C* frameshift variant specifically enriched in Finland (1:62819215:C:CT) is associated with IBD ($P=6.91\times10^{-8}$, Beta = 1.20) in the publicly released FinnGen resource (r5.finngen.fi). All variants in burden and individual tests increase risk, and the presence of four truncating variants in these analyses suggests that loss-of-function variants in *ATG4C* strongly increase CD risk.

Discussion

Here, we demonstrate that large-scale exome sequencing can complement GWASs by pinpointing specific genes both indirectly implicated by GWASs as well as those not yet observed in GWASs. With high sensitivity to directly test individual variants down to 0.01% MAF, as well as assess burden of ultra-rare mutations, we begin to fill in the low-frequency and rare-variant component of the genetic architecture of CD. This component was not observable by earlier generations of CD GWAS meta-analyses, which have had more limited coverage of low-frequency and rare variation.

Past findings in IBD[5], and most other complex diseases, suggest that while coding variants are vastly outnumbered by noncoding variation, they are highly enriched for associations to common and rare diseases. Furthermore, associated coding variants tend to have

stronger effects than their noncoding counterparts, often keeping them lower in frequency via natural selection. While this alone validates the use of exome sequencing for efficiency's sake, the primary advantage of targeting coding regions for discovery is that coding variants uniquely pinpoint genes, and often pathogenetic mechanisms, in a fashion that is at present far more challenging to achieve routinely for noncoding associations. In the case of several of the new findings (for example, *RELA*, *TAGAP*), the coding variation here provides concrete evidence of genes previously indirectly implicated by independent noncoding GWAS associations. These identify the likely gene underlying these associations and build allelic series of natural perturbations at these genes. Moreover, *IL10RA* and *RELA* are known to harbor mutations causing rare, Mendelian, inflammatory gastrointestinal disorders, and this study extends the phenotypic spectrum resulting from perturbing genetic variation to more complex forms of CD. From a functional perspective, the novel genes identified in the current study reiterate the central roles of innate and adaptive immune cells as well as autophagy in CD pathogenesis. Moreover, the involvement of *PDLIM5*, *SDF2L1*, *HGFAC*, *PAF-R* and *CCR7* pathways, in addition to the previously reported causal variant in *SMAD3* (ref. [5]), highlights the emerging role of MCs in the development and maintenance of intestinal inflammation (Fig. 2)[18]. Also, while previous studies have demonstrated the disruption of MC biology in IBD, the current findings of coding variants in these genes demonstrate that these cells and functions causally contribute to disease susceptibility. Furthermore, the association of these pathways with CD pathogenesis provides an additional rationale for development of therapeutic modalities that can re-establish the balance to the mesenchymal niche, as it is believed that genetic evidence for a drug target has a measurable impact drug development[43,44].

We expect that, in the next year, expanded sequencing efforts underway in ulcerative colitis will come to completion, enabling a more comprehensive survey of low-frequency and rare variation in ulcerative colitis, and IBD in general. Integrated with a much larger GWAS spearheaded in parallel by the IIBDGC, we expect a substantial number of conclusively linked genes and informative allelic series to emerge.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01156-2.

Received: 1 July 2021; Accepted: 12 July 2022;
Published online: 29 August 2022

References

1. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
2. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
3. Luo, Y. et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at *ADCY7*. *Nat. Genet.* **49**, 186–192 (2017).
4. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
5. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
6. 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).
7. Rivas, M. A. et al. A protein-truncating R179X variant in *RNF186* confers protection against ulcerative colitis. *Nat. Commun.* **7**, 12342 (2016).
8. Rivas, M. A. et al. Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population. *PLoS Genet.* **14**, e1007329 (2018).
9. 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).
10. Cao, Z. et al. Ubiquitin ligase TRIM62 regulates CARD9-mediated anti-fungal immunity and intestinal inflammation. *Immunity* **43**, 715–726 (2015).
11. Leshchiner, E. S. et al. Small-molecule inhibitors directly target CARD9 and mimic its protective variant in inflammatory bowel disease. *Proc. Natl Acad. Sci. USA* **114**, 11392–11397 (2017).
12. Sivanesan, D. et al. IL23R (interleukin 23 receptor) variants protective against inflammatory bowel diseases (IBD) display loss of function due to impaired protein stability and intracellular trafficking. *J. Biol. Chem.* **291**, 8673–8685 (2016).
13. Mohanan, V. et al. *C1orf106* is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science* **359**, 1161–1166 (2018).
14. Zhou, W. et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
15. Glocker, E.-O. et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* **361**, 2033–2045 (2009).
16. Liu, T., Zhang, L., Joo, D. & Sun, S.-C. NF-κB signaling in inflammation. *Signal Transduct. Target. Ther.* **2**, 17023 (2017).
17. Koliaraki, V., Prados, A., Armaka, M. & Kollias, G. The mesenchymal context in inflammation, immunity and cancer. *Nat. Immunol.* **21**, 974–982 (2020).
18. Kurashima, Y. et al. Mucosal mesenchymal cells: secondary barrier and peripheral educator for the gut immune system. *Front. Immunol.* **8**, 1787 (2017).
19. Thomson, C. A., Nibbs, R. J., McCoy, K. D. & Mowat, A. M. Immunological roles of intestinal mesenchymal cells. *Immunology* **160**, 313–324 (2020).
20. Koliaraki, V., Pallangyo, C. K., Greten, F. R. & Kollias, G. Mesenchymal cells in colon cancer. *Gastroenterology* **152**, 964–979 (2017).
21. Li, C. & Kuemmerle, J. F. The fate of myofibroblasts during the development of fibrosis in Crohn's disease. *J. Dig. Dis.* **21**, 326–331 (2020).
22. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386.e17 (2018).
23. Shi, Y. et al. PDLIM5 inhibits STUB1-mediated degradation of SMAD3 and promotes the migration and invasion of lung cancer cells. *J. Biol. Chem.* **295**, 13798–13811 (2020).
24. Maier, J. I. et al. EPB41L5 controls podocyte extracellular matrix assembly by adhesome-dependent force transmission. *Cell Rep.* **34**, 108883 (2021).
25. Yuda, A., Lee, W. S., Petrovic, P. & McCulloch, C. A. Novel proteins that regulate cell extension formation in fibroblasts. *Exp. Cell. Res.* **365**, 85–96 (2018).
26. Pompili, S., Latella, G., Gaudio, E., Sferra, R. & Vetuschi, A. The charming world of the extracellular matrix: a dynamic and protective network of the intestinal wall. *Front. Med.* **8**, 610189 (2021).
27. Martin, J. C. et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508.e20 (2019).
28. Treveil, A. et al. Regulatory network analysis of Paneth cell and goblet cell enriched gut organoids using transcriptomics approaches. *Mol. Omics* **16**, 39–58 (2020).
29. Kaser, A. & Blumberg, R. S. Endoplasmic reticulum stress in the intestinal epithelium and inflammatory bowel disease. *Semin. Immunol.* **21**, 156–163 (2009).
30. Zhang, M. & Wu, C. The relationship between intestinal goblet cells and the immune response. *Biosci. Rep.* **40**, BSR20201471 (2020).
31. Wang, X. et al. Function and dysfunction of plasma cells in intestine. *Cell Biosci.* **9**, 26 (2019).
32. Boucher, G. et al. Serum analyte profiles associated with Crohn's disease and disease location. *Inflamm. Bowel Dis.* https://doi.org/10.1093/ibd/izab123 (2021).
33. Yang, J., Dai, C. & Liu, Y. A novel mechanism by which hepatocyte growth factor blocks tubular epithelial to mesenchymal transition. *J. Am. Soc. Nephrol.* **16**, 68–78 (2005).
34. Hudry-Clergeon, H., Stengel, D., Ninio, E. & Vilgrain, I. Platelet-activating factor increases VE-cadherin tyrosine phosphorylation in mouse endothelial cells and its association with the PtdIns3′-kinase. *FASEB J.* **19**, 512–520 (2005).
35. Meran, L., Baulies, A. & Li, V. S. W. Intestinal stem cell niche: the extracellular matrix and cellular components. *Stem Cells Int.* **2017**, 7970385 (2017).
36. Sobhani, I. et al. Raised concentrations of platelet activating factor in colonic mucosa of Crohn's disease patients. *Gut* **33**, 1220–1225 (1992).
37. Chakravarty, V. et al. Prolonged exposure to platelet activating factor transforms breast epithelial cells. *Front. Genet.* **12**, 634938 (2021).

38. Knezevic, I. I. et al. Tiam1 and Rac1 are required for platelet-activating factor-induced endothelial junctional disassembly and increase in vascular permeability. *J. Biol. Chem.* **284**, 5381–5394 (2009).
39. Jang, M. H. et al. CCR7 is critically important for migration of dendritic cells in intestinal lamina propria to mesenteric lymph nodes. *J. Immunol.* **176**, 803–810 (2006).
40. Chamaillard, M. et al. Gene–environment interaction modulated by allelic heterogeneity in inflammatory diseases. *Proc. Natl Acad. Sci. USA* **100**, 3455–3460 (2003).
41. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
42. Zhou, W. et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat. Genet.* **52**, 634–639 (2020).
43. Pariente, B. et al. Treatments for Crohn's disease-associated bowel damage: a systematic review. *Clin. Gastroenterol. Hepatol.* **17**, 847–856 (2019).
44. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
45. Festen, E. A. M. et al. A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for Crohn's disease and celiac disease. *PLoS Genet.* **7**, e1001283 (2011).
46. Birkl, D. et al. TNFα promotes mucosal wound repair through enhanced platelet activating factor receptor signaling in the epithelium. *Mucosal Immunol.* **12**, 909–918 (2019).
47. Cromer, W. E., Mathis, J. M., Granger, D. N., Chaitanya, G. V. & Alexander, J. S. Role of the endothelium in inflammatory bowel diseases. *World J. Gastroenterol.* **17**, 578–593 (2011).
48. Gommerman, J. L., Rojas, O. L. & Fritz, J. H. Re-thinking the functions of IgA$^+$ plasma cells. *Gut Microbes* **5**, 652–662 (2014).
49. Stone, R. C. et al. Epithelial-mesenchymal transition in tissue repair and fibrosis. *Cell Tissue Res.* **365**, 495–506 (2016).
50. Fukushima, T., Uchiyama, S., Tanaka, H. & Kataoka, H. Hepatocyte growth factor activator: a proteinase linking tissue injury with repair. *Int. J. Mol. Sci.* **19**, 3435 (2018).
51. Waseda, M., Arimura, S., Shimura, E., Nakae, S. & Yamanashi, Y. Loss of Dok-1 and Dok-2 in mice causes severe experimental colitis accompanied by reduced expression of IL-17A and IL-22. *Biochem. Biophys. Res. Commun.* **478**, 135–142 (2016).
52. Cooke, J. et al. Mucosal genome-wide methylation changes in inflammatory bowel disease. *Inflamm. Bowel Dis.* **18**, 2128–2137 (2012).
53. Rhodes, J. Erythrocyte rosettes provide an analogue for Schiff base formation in specific T cell activation. *J. Immunol.* **145**, 463–469 (1990).
54. Celis-Gutierrez, J. et al. Dok1 and Dok2 proteins regulate natural killer cell development and function. *EMBO J.* **33**, 1928–1940 (2014).
55. Mucha, S. et al. Protein-coding variants contribute to the risk of atopic dermatitis and skin-specific gene expression. *J. Allergy Clin. Immunol.* **145**, 1208–1218 (2020).
56. Tamehiro, N. et al. T-cell activation RhoGTPase-activating protein plays an important role in TH17-cell differentiation. *Immunol. Cell Biol.* **95**, 729–735 (2017).
57. Duke-Cohan, J. S. et al. Regulation of thymocyte trafficking by Tagap, a GAP domain protein linked to human autoimmunity. *Sci. Signal.* **11**, eaan8799 (2018).
58. Medrano, L. M. et al. Expression patterns common and unique to ulcerative colitis and celiac disease. *Ann. Hum. Genet.* **83**, 86–94 (2019).
59. Chen, J. et al. TAGAP instructs Th17 differentiation by bridging Dectin activation to EPHB2 signaling in innate antifungal response. *Nat. Commun.* **11**, 1913 (2020).
60. Clark, S. E. & Weiser, J. N. Microbial modulation of host immunity with the small molecule phosphorylcholine. *Infect. Immun.* **81**, 392–401 (2013).
61. Lv, X.-X. et al. Cigarette smoke promotes COPD by activating platelet-activating factor receptor and inducing neutrophil autophagic death in mice. *Oncotarget* **8**, 74720–74735 (2017).
62. Liu, G. et al. Platelet activating factor receptor regulates colitis-induced pulmonary inflammation through the NLRP3 inflammasome. *Mucosal Immunol.* **12**, 862–873 (2019).
63. Ochoa, D. et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2020).
64. Blumert, C. et al. Analysis of the STAT3 interactome using in-situ biotinylation and SILAC. *J. Proteomics* **94**, 370–386 (2013).
65. Barrett, J. C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
66. You, K. et al. QRICH1 dictates the outcome of ER stress through transcriptional control of proteostasis. *Science* **371**, eabb6896 (2021).
67. Fujimori, T. et al. Endoplasmic reticulum proteins SDF2 and SDF2L1 act as components of the BiP chaperone cycle to prevent protein aggregation. *Genes Cells* **22**, 684–698 (2017).
68. Meunier, L., Usherwood, Y.-K., Chung, K. T. & Hendershot, L. M. A subset of chaperones and folding enzymes form multiprotein complexes in endoplasmic reticulum to bind nascent proteins. *Mol. Biol. Cell* **13**, 4456–4469 (2002).
69. Hanafusa, K., Wada, I. & Hosokawa, N. SDF2-like protein 1 (SDF2L1) regulates the endoplasmic reticulum localization and chaperone activity of ERdj3 protein. *J. Biol. Chem.* **294**, 19335–19348 (2019).
70. Sasako, T. et al. Hepatic Sdf2l1 controls feeding-induced ER stress and regulates metabolism. *Nat. Commun.* **10**, 947 (2019).
71. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730.e22 (2019).
72. Autschbach, F., Funke, B., Katzenmeier, M. & Gassler, N. Expression of chemokine receptors in normal and inflamed human intestine, tonsil, and liver—an immunohistochemical analysis with new monoclonal antibodies from the 8th international workshop and conference on human leucocyte differentiation antigens. *Cell. Immunol.* **236**, 110–114 (2005).
73. McNamee, E. N. et al. Chemokine receptor CCR7 regulates the intestinal TH1/TH17/Treg balance during Crohn's-like murine ileitis. *J. Leukoc. Biol.* **97**, 1011–1022 (2015).
74. Murugan, D. et al. Very early onset inflammatory bowel disease associated with aberrant trafficking of IL-10R1 and cure by T cell replete haploidentical bone marrow transplantation. *J. Clin. Immunol.* **34**, 331–339 (2014).
75. Pils, M. C. et al. Monocytes/macrophages and/or neutrophils are the target of IL-10 in the LPS endotoxemia model. *Eur. J. Immunol.* **40**, 443–448 (2010).
76. Qu, X. et al. TLR4-RelA-miR-30a signal pathway regulates Th17 differentiation during experimental autoimmune encephalomyelitis development. *J. Neuroinflammation* **16**, 183 (2019).
77. Thompson, M. G. et al. FOXO3-NF-κB RelA protein complexes reduce proinflammatory cell signaling and function. *J. Immunol.* **195**, 5637–5647 (2015).
78. Badran, Y. R. et al. Human RELA haploinsufficiency results in autosomal-dominant chronic mucocutaneous ulceration. *J. Exp. Med.* **214**, 1937–1947 (2017).
79. Tian, B. et al. The NFκB subunit RELA is a master transcriptional regulator of the committed epithelial-mesenchymal transition in airway epithelial cells. *J. Biol. Chem.* **293**, 16528–16545 (2018).
80. Rioux, J. D. et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596–604 (2007).
81. McCarroll, S. A. et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008).
82. Agrotis, A., Pengo, N., Burden, J. J. & Ketteler, R. Redundancy of human ATG4 protease isoforms in autophagy and LC3/GABARAP processing revealed in cells. *Autophagy* **15**, 976–997 (2019).
83. Finisguerra, V. et al. MET is required for the recruitment of anti-tumoural neutrophils. *Nature* **522**, 349–353 (2015).
84. Stakenborg, M. et al. Neutrophilic HGF-MET signaling exacerbates intestinal inflammation. *J. Crohns Colitis* https://doi.org/10.1093/ecco-jcc/jjaa121 (2020).
85. Kanayama, M. et al. Hepatocyte growth factor promotes colonic epithelial regeneration via Akt signaling. *Am. J. Physiol. Gastrointest. Liver Physiol.* **293**, G230–G239 (2007).
86. Tahara, Y. et al. Hepatocyte growth factor facilitates colonic mucosal repair in experimental ulcerative colitis in rats. *J. Pharmacol. Exp. Ther.* **307**, 146–151 (2003).
87. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

Aleksejs Sazonovs[1,98], Christine R. Stevens[2,3,4,98], Guhan R. Venkataraman[5,98], Kai Yuan[3,4,98], Brandon Avila[2], Maria T. Abreu[6], Tariq Ahmad[7], Matthieu Allez[8], Ashwin N. Ananthakrishnan[9], Gil Atzmon[10,11], Aris Baras[12], Jeffrey C. Barrett[13], Nir Barzilai[11,14], Laurent Beaugerie[15], Ashley Beecham[16,17], Charles N. Bernstein[18], Alain Bitton[19], Bernd Bokemeyer[20], Andrew Chan[21,22],

Daniel Chung[23], Isabelle Cleynen[24], Jacques Cosnes[25], David J. Cutler[26,27], Allan Daly[28], Oriana M. Damas[29], Lisa W. Datta[30], Noor Dawany[31], Marcella Devoto[31,32,33,34], Sheila Dodge[35], Eva Ellinghaus[36], Laura Fachal[1], Martti Farkkila[37], William Faubion[38], Manuel Ferreira[12], Denis Franchimont[39], Stacey B. Gabriel[35], Tian Ge[3,40,41], Michel Georges[42], Kyle Gettler[43], Mamta Giri[43], Benjamin Glaser[44], Siegfried Goerg[45], Philippe Goyette[46], Daniel Graham[47,48,49], Eija Hämäläinen[50], Talin Haritunians[51], Graham A. Heap[7], Mikko Hiltunen[52], Marc Hoeppner[53], Julie E. Horowitz[12], Peter Irving[54,55], Vivek Iyer[28], Chaim Jalas[56], Judith Kelsen[31], Hamed Khalili[21], Barbara S. Kirschner[57], Kimmo Kontula[58], Jukka T. Koskela[50], Subra Kugathasan[27], Juozas Kupcinskas[59], Christopher A. Lamb[60,61], Matthias Laudes[45], Chloé Lévesque[46], Adam P. Levine[62], James D. Lewis[34,63], Claire Liefferinckx[39], Britt-Sabina Loescher[36], Edouard Louis[42], John Mansfield[60,61], Sandra May[36], Jacob L. McCauley[16,17], Emebet Mengesha[51], Myriam Mni[42], Paul Moayyedi[64], Christopher J. Moran[23], Rodney D. Newberry[65], Sirimon O'Charoen[63], David T. Okou[27,66], Bas Oldenburg[67], Harry Ostrer[68], Aarno Palotie[2,3,4,50,69,70], Jean Paquette[46], Joel Pekow[57], Inga Peter[43], Marieke J. Pierik[71], Cyriel Y. Ponsioen[72], Nikolas Pontikos[62], Natalie Prescott[73], Ann E. Pulver[74], Souad Rahmouni[42], Daniel L. Rice[1], Päivi Saavalainen[75], Bruce Sands[43], R. Balfour Sartor[76], Elena R. Schiff[62], Stefan Schreiber[36], L. Philip Schumm[77], Anthony W. Segal[62], Philippe Seksik[15], Rasha Shawky[78], Shehzad Z. Sheikh[76], Mark S. Silverberg[79], Alison Simmons[80], Jurgita Skeiceviciene[59], Harry Sokol[15], Matthew Solomonson[2], Hari Somineni[81], Dylan Sun[12], Stephan Targan[51], Dan Turner[82], Holm H. Uhlig[83,84], Andrea E. van der Meulen[85], Séverine Vermeire[86,87], Sare Verstockt[87], Michiel D. Voskuil[88], Harland S. Winter[23], Justine Young[63], Belgium IBD Consortium*, Cedars-Sinai IBD*, International IBD Genetics Consortium*, NIDDK IBD Genetics Consortium*, NIHR IBD BioResource*, Regeneron Genetics Center*, SHARE Consortium*, SPARC IBD Network*, UK IBD Genetics Consortium*, Richard H. Duerr[89], Andre Franke[36], Steven R. Brant[30,90], Judy Cho[43], Rinse K. Weersma[88], Miles Parkes[91], Ramnik J. Xavier[47,48,49,92,93,94,95,96], Ma nuel A. Rivas[5], John D. Rioux[46,97], Dermot P. B. McGovern[51], Hailiang Huang[3,4,99] ✉, Carl A. Anderson[1,99] ✉ and Mark J. Daly[2,3,4,50,99] ✉

[1]Genomics of Inflammation and Immunity Group, Human Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [2]Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [5]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [6]Crohn's and Colitis Center, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA. [7]Royal Devon and Exeter Hospital, Exeter, UK. [8]Hopital Saint-Louis, APHP, Universite de Paris, INSERM U1160, Paris, France. [9]Division of Gastroenterology, Crohn's and Colitis Center, Massachusetts General Hospital, Boston, MA, USA. [10]Department for Human Biology, University of Haifa, Haifa, Israel. [11]Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [12]Regeneron Genetics Center, Tarrytown, NY, USA. [13]Human Genetics Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [14]The Institute for Aging Research, The Nathan Shock Center of Excellence in the Basic Biology of Aging and the Paul F. Glenn Center for the Biology of Human Aging Research at Albert Einstein College of Medicine of Yeshiva University, Bronx, NY, USA. [15]Gastroenterology Department, Sorbonne Universite, Saint Antoine Hospital, Paris, France. [16]John P. Hussman Institute for Human Genomics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA. [17]The Dr. John T. Macdonald Foundation Department of Human Genetics, Leonard M. Miller School of Medicine, University of Miami, Miami, FL, USA. [18]University of Manitoba, Winnipeg, Manitoba, Canada. [19]McGill University and McGill University Health Centre, Montreal, Quebec, Canada. [20]Department of Internal Medicine, University Medical Center Schleswig-Holstein, Kiel, Germany. [21]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital, Boston, MA, USA. [22]Channing Division of Network Medicine, Department of Medicine, Brigham and Womens Hospital, Boston, MA, USA. [23]Massachusetts General Hospital, Boston, MA, USA. [24]Department of Human Genetics, KU Leuven, Leuven, Belgium. [25]Professeur Chef de Service chez APHP and Universite Paris-6, Paris, France. [26]Department of Human Genetics, Emory University, Atlanta, GA, USA. [27]Emory University School of Medicine and Children's Healthcare of Atlanta, Atlanta, GA, USA. [28]Human Genetics Informatics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [29]University of Miami, Miami, FL, USA. [30]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [31]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [32]University of Rome Sapienza, Rome, Italy. [33]IRGB - CNR, Cagliari, Italy. [34] Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [35]Genomics Platform, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [36]Christian-Albrechts-University of Kiel and University Medical Center Schleswig-Holstein, Kiel, Germany. [37]Helsinki University Central Hospital, Helsinki, Finland. [38]Mayo Clinic, Rochester, Rochester, MN, USA. [39]Erasme Hospital, ULB, Brussels, Belgium. [40]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [41]Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [42]University of Liège, ULG, Liège,

Belgium. [43]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [44]Department of Endocrinology and Metabolism, Hadassah Medical Center and Faculty of Medicine, Hebrew University of Jerusalem, Jerusalem, Israel. [45]University Medical Center Schleswig-Holstein, Kiel, Germany. [46]Research Center Montreal Heart Institute, Montreal, Quebec, Canada. [47]Infectious Disease and Microbiome Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [48]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [49]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA. [50]Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland. [51]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars Sinai Medical Center, Los Angeles, CA, USA. [52]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [53]Christian-Albrechts-University of Kiel, Kiel, Germany. [54]Department of Gastroenterology, Guys and Saint Thomas Hospital, London, UK. [55]School of Immunology and Microbial Sciences, Kings College London, London, UK. [56]Director of Genetic Resources and Services, Center for Rare Jewish Genetic Disorders, Bonei Olam, Brooklyn, NY, USA. [57]Department of Gastroenterology, University of Chicago Medicine, Chicago, IL, USA. [58]Department of Medicine, Helsinki University Hospital, and Research Program for Clinical and Molecular Metabolism, University of Helsinki, Helsinki, Finland. [59]Department of Gastroenterology and Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania. [60]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK. [61]Department of Gastroenterology, Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK. [62]University College London, London, UK. [63]Crohn's and Colitis Foundation, New York, NY, USA. [64]McMaster University, Hamilton, Ontario, Canada. [65]Washington University School of Medicine, St. Louis, MO, USA. [66]Institut National de Sante Publique (INSP), Abidjan, Côte d'Ivoire. [67]Department of Gastroenterology and Hepatology, University Medical Centre Utrecht, Utrecht, The Netherlands. [68]Albert Einstein College of Medicine, Bronx, NY, USA. [69]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [70]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA. [71]Department of Gastroenterology and Hepatology, Maastricht University Medical Centre, Maastricht, The Netherlands. [72]Department of Gastroenterology and Hepatology, Amsterdam University Medical Centres, Amsterdam, The Netherlands. [73]Department of Medical and Molecular Genetics, Kings College London, London, UK. [74]School of Medicine, Johns Hopkins University, Baltimore, MD, USA. [75]Research Programs Unit, Immunobiology, University of Helsinki, Helsinki, Finland. [76]Center for Gastrointestinal Biology and Disease, University of North Carolina School of Medicine, Chapel Hill, NC, USA. [77]Department of Public Health Sciences, University of Chicago, Chicago, IL, USA. [78]IBD BioResource, NIHR BioResource, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [79]Mount Sinai Hospital, Toronto, Ontario, Canada. [80]MRC Human Immunology Unit, NIHR Biomedical Research Centre, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [81]Department of Pediatrics, Emory University School of Medicine and Children's Healthcare of Atlanta, Atlanta, GA, USA. [82]Shaare Zedek Medical Center, Jerusalem, Israel. [83]Translational Gastroenterology Unit and Biomedical Research Centre, Nuffield Department of Clinical Medicine, Experimental Medicine Division, University of Oxford, Oxford, UK. [84]Department of Pediatrics, John Radcliffe Hospital, Oxford, UK. [85]Department of Gastroenterology and Hepatology, Leiden University Medical Center, Leiden, The Netherlands. [86]University Hospitals Leuven, Leuven, Belgium. [87]Department of Chronic Diseases and Metabolism, KU Leuven, Leuven, Belgium. [88]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [89]University of Pittsburgh, Pittsburgh, PA, USA. [90]Crohn's Colitis Center of New Jersey, Department of Medicine, Rutgers Robert Wood Johnson Medical School and Department of Genetics and the Human Genetics Institute of New Jersey, Rutgers University, New Brunswick and Piscataway, NJ, USA. [91]Department of Gastroenterology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [92]Kurt Isselbacher Professor of Medicine at Harvard Medical School, Cambridge, MA, USA. [93]Core Institute Member, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [94]Klarman Cell Observatory, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [95]Immunology Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [96]Center for Microbiome Informatics and Therapeutics at MIT, Massachusetts Institute of Technology, Cambridge, MA, USA. [97]Faculty of Medicine, Université de Montréal, Montreal, Canada. [98]These authors contributed equally: Aleksejs Sazonovs, Christine R. Stevens, Guhan R. Venkataraman, Kai Yuan. [99]These authors jointly supervised this work: Hailiang Huang, Carl A. Anderson, Mark J. Daly. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: hhuang@broadinstitute.org; carl.anderson@sanger.ac.uk; mjdaly@broadinstitute.org

Methods

Ethics declarations. All relevant ethical guidelines have been followed, and any necessary institutional review board (IRB) and/or ethics committee approvals have been obtained. The details of the IRB/oversight body that provided approval or exemption for the research described are given below:

Study Protocol 2013P002634, The Broad Institute Study of Inflammatory Bowel Disease Genetics, undergoes annual continuing review by the Mass General Brigham Human Research Committee IRB of Mass General Brigham. Ethical approval was given on 27 January 2021 for this study (Mass General Brigham IRB).

All informed consent from participants has been obtained and the appropriate institutional forms have been archived.

DNA samples sequenced at the Sanger Institute were ascertained under the following ethical approvals: 12/EE/0482, 12/YH/0172, 16/YH/0247, 09/H1204/30, 17/EE/0265, 16/WM/0152, 09/H0504/125, 15/EE/0286, 11/YH/0020, 09/H0717/4, REC 22/02, 03/5/012, 03/5/012, 2000/4/192, 05/Q1407/274, 05/Q0502/127, 08/H0802/147, LREC/2002/6/18, GREC/03/0273 and YREC/P12/03.

Broad Institute sequencing pipeline. *Sample processing.* Exome sequencing was performed at the Broad Institute. The sequencing process included sample preparation (Illumina Nextera, Illumina TruSeq and Kapa Hyperprep), hybrid capture (Illumina Rapid Capture Enrichment (Nextera), 37 Mb target, and Twist Custom Capture, 37 Mb target) and sequencing (Illumina HiSeq2000, Illumina HiSeq2500, Illumina HiSeq4000, Illumina HiSeqX, Illumina NovaSeq 6000, 76-base pair (bp) and 150-bp paired reads). Sequencing was performed at a median depth of 85% targeted bases at >20×. Sequencing reads were mapped by BWA-MEM to the hg38 reference using a 'functional equivalence' pipeline. The mapped reads were then marked for duplicates, and base quality scores were recalibrated. They were then converted to CRAMs using Picard 2.16.0-SNAPSHOT and GATK 4.0.11.0. The CRAMs were then further compressed using ref-blocking to generate gVCFs. These CRAMs and gVCFs were then used as inputs for joint calling. To perform joint calling, the single-sample gVCFs were hierarchically merged (separately for samples using Nextera and Twist exome capture).

QC. QC analyses were conducted in Hail v.0.2.47 (Extended Data Fig. 2). We first split multiallelic sites and coded genotypes with genotype quality (GQ) < 20 as missing. Variants not annotated as frameshift, inframe deletion, inframe insertion, stop lost, stop gained, start lost, splice acceptor, splice donor, splice region, missense or synonymous were removed from the following analysis. We also removed variants that have known quality issues (have a nonempty QUAL column) in the gnomAD dataset. Sample QC: poor-quality samples that met the following criteria were identified and removed: (1) samples with an extremely large number of singletons (≥500); (2) samples with mean GQ < 40; and (3) samples with missingness rates > 10%. Variant QC: low-quality variants that met the following criteria were identified and removed: (1) variants with missingness rate > 5%; (2) variants with mean read depth (DP) < 10; (3) variants that failed the Hardy–Weinberg equilibrium test for controls with $P < 1 \times 10^{-4}$; and (4) variants with >10% samples that were heterozygous and with an allelic balance ratio <0.3 or >0.7. Variants with different genotypes in WES and WGS in gnomAD were also removed. For Twist exome capture samples, we additionally removed (1) samples that had a significantly high or low inbreeding coefficient (>0.2 or ≤0.2); (2) samples that had a high heterozygosity away from mean (±5 s.d.); and (3) related samples, which were removed sequentially by removing the individual with the largest number of related samples (in PLINK, the individual with PI_HAT > 0.2 when using the '–genome' option) until no related samples remained. For Nextera capture samples, we additionally removed variants showing a significant heterogeneous effect across Ashkenazi Jewish (AJ), Lithuanian (LIT), Finnish (FIN) and NFE samples (see Population assignment below).

Population assignment. We projected all samples onto principal component (PC) axes generated from the 1000 Genomes Project Phase 3 common variants, and classified their ancestry using a random forest method to the European (CEU, TSI, FIN, GBR, IBS), African (YRI, LWK, GWD, MSL, ESN, ASW, ACB), East Asian (CHB, JPT, CHS, CDX, KHV), South Asian (GIH, PJL, BEB, STU, ITU) and American (MXL, PUR, CLM, PEL) samples. We kept samples that were classified as European with prediction probability greater than 80% (Extended Data Fig. 7). For Nextera samples, we used a second random forest classifier to assign EUR samples to AJ, LIT, FIN or NFE, and a third random forest classifier to clean the AJ/NFE split.

Meta-analysis. We used METAL[87] with an IVW fixed-effect model to meta-analyze the SAIGE association statistics from Nextera and Twist samples (Table 1). The heterogeneity test was performed using Cochran's Q with one degree of freedom.

Sanger Institute sequencing pipeline. *Sample processing.* Genome sequencing was performed at the Sanger Institute using the Illumina HiSeqX platform with a combination of PCR ($n = 4{,}751$, controls only) and PCR-free library preparation protocols. Sequencing was performed at a median depth of 18.6×. Exome sequencing of cases was performed at the Sanger Institute using the Illumina NovaSeq 6000 and the Agilent SureSelect Human All Exon V5 capture set. Controls from the UK Biobank were sequenced separately as a part of the UKBB WES50K release using Illumina NovaSeq and the IDT xGen Exome Research Panel v1.0 capture set (including supplemental probes). In total, 33,704 UKBB participants were selected for use as controls, excluding participants with recorded or self-reported CD, ulcerative colitis, unspecified noninfective gastroenteritis or colitis; any other immune-mediated disorders; or a history of being prescribed any drugs used to treat IBD. Exome and genome datasets were analyzed separately but followed a similar analysis protocol.

Reads were mapped to hg38 reference using BWA-MEM v.0.7.12 (WGS) and v.0.7.17 (WES). Variant calls were performed using a GATK Best Practices-like pipeline (v.4.0.10.1 (WGS) and v.4.1.8 (WES)); per-sample intermediate variant calling was followed by joint genotyping across the individual genome and exome cohorts. For the exome cohort, variant calling was limited to Agilent extended target regions. Per-region VCF shards were imported into the Hail software and combined. Multiallelic sites were split. For the exome cohort, we subsetted the calls to the intersection of Agilent and IDT exome captures, further excluding regions recommended for exclusion by the UKBB due to an error in read mapping that results in no variant calls made.

Population assignment. We selected a set of ~14,000 well-genotyped common variants to identify the genetic ancestry of individual participants through the projection of 1000 Genomes Project cohort-derived PCs. For genomes, due to primarily European genetic ancestry of the controls, we excluded samples outside of 4 median absolute deviations from the median point of the European ancestry cluster of 1000 Genomes. For exomes, we implemented a Random Forest technique that classified samples based on PCs into broad genetic ancestry groups (EUR, AFR, SAS, EAS, admixed), with self-reported ancestry as training labels. For these analyses, we only retained the EUR samples, as the number of cases for other groups was too small for robust association analysis.

QC. A combination of hard-cutoff filters and per-ancestry/per-batch outlier filters was used to identify low-quality samples. We applied hard-filters for sample depth (>12× genomes, >15× exomes), call rate (>0.95), chimerism (<0.5) (WGS) and FREEMIX (<0.02) (WGS). We excluded genotype calls with an allelic imbalance (for heterozygous calls, allelic balance ratio < 0.2 or > 0.8), low depth (<2×) and low GQ (<20). We then performed per-ancestry and per-sequencing protocol (AGILENT versus IDT for WES, PCR versus PCR-free for WES) filtering of samples falling outside 4 median absolute deviations from the median per-batch heterozygosity rate, transition/transversion rate, number of called SNPs and INTELs, and insertion and deletion counts/ratio.

An ancestry-aware relatedness calculation (PC-Relate method in Hail[88]) was used to identify related samples. As our association approach (logistic mixed-models) can control for residual relatedness, we only excluded duplicates or MZ twins from within the cohorts and excluded first-, second- and third-degree relatives when the kinship was across the cohorts (for example, parent in WGS, child in WES; kinship metric > 0.1 calculated via PC-Relate method using 10 PCs). In addition, we removed samples that were also present in the Broad Institute's cohorts.

Association testing. Association analysis was performed using a logistic mixed-model implemented in REGENIE software v.1.0.6.7 (singe-variant) and v.2.0.2 (burden). A set of high-confidence variants (>1% MAF, 99% call rate and in Hardy–Weinberg equilibrium) was used for *t*-fitting. To control for case–control imbalance, Firth correction was applied to *P* values < 0.05. To control for residual ancestry and sequencing heterogeneity, we calculated 10 PCs on a set of well-genotyped common SNPs, excluding regions with known long-range LD. These were used as covariates for association analyses. Only variants with call rate above 90% after filtering poor calls were included in the association analysis. For WES, we verified that the >90% call rate condition holds true in both AGILENT and IDT samples. Association analysis was performed on QC-passing calls.

Kiel/Regeneron sequencing pipeline. *Sample preparation and sequencing.* The DNA samples were normalized and 100 ng of genomic DNA was prepared for exome capture with custom reagents from New England Biolabs, Roche/Kapa and IDT using a fully automated approach developed at the Regeneron Genetics Center. Unique, asymmetric 10-bp barcodes were added to each side of the DNA fragment during library preparation to facilitate multiplexed exome capture and sequencing. Equal amounts of sample were pooled before exome capture with a slightly modified version of IDT's xGen v1 probes; supplemental probes were added to capture regions of the genome well covered by a previous capture reagent (NimbleGen VCRome) but poorly covered by the standard xGen probes, the same as the probe library used in UK Biobank exome sequencing. These supplemental probes were included in QC but excluded in the final analysis as we only looked up variants that were in the standard exome captures and reached the nominal significance for replication (Extended Data Fig. 1). Captured fragments were bound to streptavidin-conjugated beads and nonspecific DNA fragments were removed by a series of stringent washes according to the manufacturer's recommended protocol (IDT). The captured DNA was PCR amplified and quantified by quantitative reverse transcription PCR (Kapa Biosystems). The multiplexed samples were pooled and then sequenced using 75-bp paired-end reads with two index reads on the Illumina NovaSeq 6000 platform using S2 flow cells.

Variant calling and QC. Sample read mapping and variant calling, aggregation and QC were performed via the SPB protocol described by Van Hout et al.[89]. Briefly, for each sample, NovaSeq WES reads are mapped with BWA-MEM 0.7.17-r1188 to the hg38 reference genome. Small variants are identified with WeCall v.1.1.2 and reported as per-sample gVCFs. These gVCFs are aggregated with GLnexus into a joint-genotyped, multi-sample VCF (pVCF). SNV genotypes with DP less than 7 and indel genotypes with DP less than 10 are changed to no-call genotypes. After the application of the DP genotype filter, a variant-level allele balance filter is applied, retaining only variants that meet either of the following criteria: (1) at least one homozygous variant carrier or (2) at least one heterozygous variant carrier with an allele balance greater than the cutoff.

Analysis. We combined the gVCF files with bcftools 1.11 using the 'merge' command, then imported the joint VCF into Hail. We then split the multiallelic variants and removed variants with '<NON_REF>' alternative alleles. We applied the QC steps and assigned populations as in the Broad Institute sequencing pipeline.

Statistics and reproducibility. Previous studies show that a large sample size is needed for IBD genetic studies. We have thus included all samples available to us. We excluded samples of non-European ancestries due to their very limited sample size when properly matched between cases and controls (Extended Data Fig. 7). We also excluded data of poor quality from the analysis (Extended Data Fig. 2). These exclusions were necessary to ensure the quality of this study. All criteria were pre-established. We used the logistic mixed-model for the association analysis, followed by meta-analyses to combine multiple cohorts. We have multiple cohorts in the study that serve the purpose of replication. Two large cohorts done at the Broad Institute of different exome capture platforms were used to discover candidate variants. Two independent cohorts done at Sanger and one Kiel/Regeneron cohort were used to replicate the findings (Extended Data Fig. 1). All reported findings have been replicated. No randomization was conducted. No blinding was carried out. Code and pipelines to reproduce our analysis are available on Zenodo[90].

Cross-cohort meta-analysis. We used the Cochran–Mantel–Haenszel test to combine association summary statistics between the Broad Institute, Sanger Institute and Kiel/Regeneron cohorts.

Relation to known IBD causal variants. We assigned the 45 study-wide significant variants to one of the four categories based on their relation with known IBD associations and/or fine-mapping results (Extended Data Fig. 5 and Supplementary Table 4): (1) Known causal candidate: variants in a fine-mapping credible set[5] with PIP > 5%, or reported in the earlier sequencing studies after manual review[6,8]. (2) New locus: variants implicating a genetic locus in general onset CD that have not been previously reported. (3) Unlikely causal: variants with PIP < 5%, or variants tagging the best PIP variants using conditional analysis (see Conditional analysis below, *LRRK2* shown as an example in Extended Data Fig. 6d–g). (4) New variant in known locus: variants in known GWAS loci with MAF < 0.5% (and, thus, no LD to evaluate tagging) remain study-wide significant after conditional analysis using the LD from gnomAD (*TAGAP* shown as an example in Extended Data Figure 6a–c), or after manual review (Exceptions and notes).

Variance explained. Using the Sanger WGS data (6,000 cases, 11,852 controls), we fitted a series of univariate logistic regression (is_case ~ variant_genotype) models and estimated the pseudo-r^2. Pseudo-r^2 estimates were summed to estimate the observed-scale variance explained by a group of variants. To convert the estimate into an estimate of heritability on the liability scale, we assumed that the prevalence of CD is 276 in 100,000 (UK estimate from ref. [91]).

Conditional analysis. For study-wide significant variants not in a previously reported credible set[5], we performed a conditional analysis to test whether they are independent from or tagging the known causal variants[5]. We first classified variants as 'tagging' if they had $r^2 > 0.8$ with any variants in the reported credible sets[5]. For other variants, we performed a conditional analysis using (1) the *P* value estimates from previous fine-mapping studies for credible set variants and (2) the LD calculated from gnomAD. We were unable to directly fit a multivariate model or use the LD from study subjects, because exome sequencing does not cover the noncoding putative causal variants, and the ImmunoChip does not have good quality for rare coding variants. The conditional *z* statistic, z'_{Seq}, for a variant with marginal statistic of z_{Seq} from our study, was calculated as follows:

$$z'_{\mathrm{Seq}} = -\frac{\left|z_{\mathrm{Seq}}\right| - \sum_i^n \left(\left|z_{\mathrm{FM}_i} * r_i * \sqrt{N_{\mathrm{Seq}}/N_{\mathrm{FM}}}\right|\right)}{\prod_i^n \sqrt{1 - r_i^2}}$$

in which z_{FM_i} is the *z* statistic of the variant with the best PIP in the credible set *i*, out of *n* total credible sets, from the fine-mapping study, r_i is the LD between the two variants, and N_{Seq} and N_{FM} are the effective sample sizes for our study and the fine-mapping study, respectively. We used the absolute value in this equation because of the challenges to align the alleles across sequencing, the fine-mapping study and the gnomAD reference panel. Taking the absolute value is a conservative approximation (less likely to declare a variant as novel association) because it assumes that the putative causal variants from fine-mapping have the same direction of effect as the variant being tested when they are in LD. This is very likely to be correct. The effective sample size was calculated as $4/\left(1/N_{\mathrm{case}} + 1/N_{\mathrm{control}}\right)$, in which N_{case} and N_{control} are the sample sizes for cases and controls, respectively. For each variant, we summed the effective sample sizes across all cohorts in which the variant is observed (thus, N_{Seq} can differ from variant to variant). We calculated the conditional *P* value from z'_{Seq} under the standard Gaussian distribution. A variant was classified as 'tagging' if the conditional *P* value failed to reach study-wide significance at 3×10^{-7}.

Exceptions and notes. *HGFAC*: despite this locus having been reported in an earlier GWAS[2], the coding variant we identify was not tested for association due to incomplete coverage of this region, and is thus reported in this study as directly implicating this gene ($r^2 = 0.35$ with the previously reported GWAS SNP, rs2073505). We thus assign this variant as 'New variant in known locus'. *RELA*: similar to *HGFAC*, this locus has been reported in an earlier GWAS[2], but the coding variant we identified was not tested for association due to incomplete coverage of this region, and thus is reported in this study as directly implicating this gene ($r^2 = 0.002$ with the previously reported GWAS SNP, rs568617). We thus assign this variant as 'New variant in known locus'. *SLC39A8*: the *SLC39A8* A391T variant was not reported in the fine-mapping paper, as its genetic region was not included in the ImmunoChip design. Because this variant has been published in several papers as an IBD causal variant with genetic and functional evidence[92–94], we assign this variant as 'Known causal candidate'. *TYK2*: the *TYK2* A928V was not reported in the fine-mapping paper[5], likely due to a lack of power. Because this variant has been known to be a causal variant for several autoimmune disorders[95] and in another IBD study[96], we assign this variant as 'Known causal candidate'. *SDF2L1*: this variant has marginal $P = 2 \times 10^{-7}$ and conditional $P = 3.4 \times 10^{-4}$. The r^2 between this variant and the noncoding variant with the best PIP from fine-mapping is 0.045. We manually assigned this variant to 'New variants in known locus', as this is a missense variant. *NOD2*: (1) Previous studies[5–7] have shown evidence that the *NOD2* S431L variant tags the *NOD2* V793M variant, with the latter more likely to be the CD causal variant. In this study, however, S431L reached study-wide significance, but V793M failed to meet the significance cutoff. We therefore retained S431L in Fig. 1 for the purpose of keeping this association signal. (2) Due to the complexity of the *NOD2* locus, we conducted a haplotype analysis using the Twist subjects and additionally classified signed variants that share the same haplotype with known IBD variants as 'tagging'. We found that for the *NOD2* S47L variant, 18 out of 19 copies of the T allele are on the same haplotype as the fs1007insC variant. We therefore classify S47L as 'tagging'. (3) The NOD2 A755V variant is in LD with rs184788345, the best PIP variant from fine-mapping ($r^2 = 0.85$). The marginal *P* value for A755V is one order of magnitude less significant than rs184788345. Considering A755V is a missense variant while none of the variants in the credible set defined by rs184788345 is coding, we assign A755V as a likely 'Known causal candidate'.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We describe all datasets in the manuscript or its Supplementary Information. Genome Reference Consortium Human Build 38 can be accessed at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40/. Sequence data used in this study have been made publicly available in dbGaP Study Accession: phs001642.v1.p1, Center for Common Disease Genomics (CCDG), Autoimmune: Inflammatory Bowel Disease (IBD) Exomes and Genomes (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001642.v1.p1). The summary statistics of Nextera and Twist meta-analysis have been deposited on GitHub (https://github.com/iibdgc/Crohn-s-Disease-WES-meta) (https://doi.org/10.5281/zenodo.6564928). This research has been conducted using the UK Biobank Resource and controls made publicly available by dbGaP (phs001000.v1.p1, phs000806.v1.p1, Myocardial Infarction Genetics Consortium (MIGen); phs000401.v1.p1, NHLBI GO-ESP Project; phs000298.v4.p3, Autism Sequencing Consortium (ASC); phs000572.v8.p4, Alzheimer's Disease Sequencing Project (ADSP); phs001489.v1.p1, Epi25 Consortium; phs001095.v1.p1, T2D-GENES) as well as additional controls from the 1000 Genomes Project, the Epi25 Collaborative, UK-Ireland Collaborators (A. McQuillin, D. Blackwood, A. McIntosh), and collaborators A. Pulver, H. Ostrer, D. Chung, M. Hiltunen and A. Palotie (H2000 and SUPER cohorts) (Supplementary Table 1).

Code availability

The software and code used are described throughout the Methods and can be found at https://github.com/iibdgc/Crohn-s-Disease-WES-meta (https://doi.org/10.5281/zenodo.6564928).

References

88. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
89. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

90. Venkataraman, G. R., Yuan, K. & Huang, H. Crohn's disease WES meta-analysis [Computer software]. *Zenodo* https://doi.org/10.5281/zenodo.6564928 (2022)
91. Pasvol, T. J. et al. Incidence and prevalence of inflammatory bowel disease in UK primary care: a population-based cohort study. *BMJ Open* **10**, e036584 (2020).
92. Nakata, T. et al. A missense variant in SLC39A8 confers risk for Crohn's disease by disrupting manganese homeostasis and intestinal barrier integrity. *Proc. Natl Acad. Sci. USA* **117**, 28930–28938 (2020).
93. Li, D. et al. A pleiotropic missense variant in *SLC39A8* is associated with Crohn's disease and human gut microbiome composition. *Gastroenterology* **151**, 724–732 (2016).
94. Sunuwar, L. et al. Pleiotropic ZIP8 A391T implicates abnormal manganese homeostasis in complex human disease. *JCI Insight* **5**, e140978 (2020).
95. 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).
96. Diogo, D. et al. TYK2 protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLoS ONE* **10**, e0122271 (2015).

Acknowledgements

We thank all of the principal investigators, local staff from individual cohorts and all of the patients who kindly donated samples used in the study for making possible this global collaboration and resource to advance IBD genetics research. This research was funded in whole, or in part, by the US National Institutes of Health grants no. U54HG003067 and no. 5UM1HG008895, the Wellcome Trust grants no. 206194 and no. 108413/A/15/D, and The Leona M. & Harry B. Helmsley Charitable Trust grant no. 2015PG-IBD001. We thank the Broad Institute Genomics Platform for genomic data generation efforts and the Stanley Center for Psychiatric Research at the Broad Institute for supporting control sample aggregation. M.A.R. is in part supported by the NHGRI of the NIH under award no. R01HG010140 and an NIH Center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (no. 5U01HG009080). H.H. acknowledges support from NIDDK grant no. K01DK114379, grant no. P30DK043351 and the Stanley Center for Psychiatric Research. H.S.W. receives philanthropic support from Martin Schlaff, James Brooks and the B. Hasso Family Foundation. H.H.U. and A. Sazonovs. are supported by the NIHR Oxford Biomedical Research Centre and by The Leona M. and Harry B. Helmsley Charitable Trust. A.P. is in part supported by the Academy of Finland Centre of Excellence in Complex Disease Genetics grants no. 312074 and no. 336824. Individual studies contributing to this meta-analysis acknowledge support from NIH grants no. DK062431, no. DK062432, no. DK087694, no. K23DK117054, no. R01DK111843, no. P01DK094779, no. R01HG010140, no. 5U01HG009080 and no. DK062420, and NIDDK grants no. P01DK046763, no. U01DK062413 and no. R01DK104844.

Author contributions

H.H., C.A.A. and M.J.D. designed and supervised the study. C.R.S., H.H., C.A.A. and M.J.D. were responsible for project management. A. Sazonovs, G.R.V., K.Y., B.A., A.D., T.G., D.G., V.I., J.T.K., D.L.R., M. Solomonson, M.A.R., H.H., C.A.A. and M.J.D. performed data analysis. M.T.A., T.A., M.A., A.N.A., G.A., A. Baras, A. Beecham, A. Bitton, J.C.B., N.B., L.B., C.N.B., B.B., A.C., D.C., I.C., J. Cho, J. Cosnes, D.J.C., O.M.D., L.W.D., N.D., M.D., E.E., L.F., M. Farkkila, M. Ferreira, W.F., D.F., M. Georges, M. Giri, K.G., B.G., S.G., P.G., E.H., T.H., G.A.H., M. Hiltunen, M. Hoeppner, J.E.H., P.I., C.J., J. Kelsen, J. Kupcinskas, H.K., B.S.K., K.K., J.T.K., S.K., C.A.L., M.L., C. Lévesque, C. Liefferinckx, A.P.L., J.D.L., B.-S.L., E.L., J.M., S.M., J.L.M., E.M., M.M., P.M., C.J.M., R.D.N., S.O., D.T.O., B.O., H.O., A.P., J. Paquette, J. Pekow, I.P., M.J.P., C.Y.P., N. Pontikos, N. Prescott, A.E.P., S.R., P. Saavalainen, P. Seksik, B.S., R.B.S., E.R.S., S.S., L.P.S., A.W.S., R.S., S.Z.S., M.S.S., A. Simmons, J.S., H. Sokol, H. Somineni, D.S., S.T., D.T., H.H.U., A.E.V., S. Vermeire, S. Verstockt, M.D.V., H.S.W., J.Y., R.H.D., A.F., S.R.B., R.K.W., M.P., R.J.X., J.D.R. and D.P.B.M. were responsible for recruitment, clinical phenotyping, analysis and/or leadership of a contributing study. S.D. and S.B.G. performed sequencing technology development. A. Sazonovs, C.R.S., G.R.V., K.Y., S.R.B., J.D.R., D.P.B.M., H.H., C.A.A. and M.J.D. wrote the manuscript.

Competing interests

A. Baras., M. Ferreira., J.E.H. and D.S are current or former employees and/or stockholders of Regeneron Genetics Center or Regeneron Pharmaceuticals. M.A. is consulting for or part of the advisory board for AbbVie Inc., Bellatrix Pharmaceuticals, Bristol Myers Squibb, Eli Lilly Pharmaceuticals, Gilead, Janssen Ortho, LLC, and Prometheus Biosciences; and teaching, lecturing or speaking at Alimentiv, Arena Pharmaceuticals, Janssen, Prime CME, Takeda Pharmaceuticals. A.B. is an employee of Regeneron and owns stock in Regeneron. O.M.D. has served in the IBD fellowship funding committee for Pfizer and has a funded research project by Pfizer. H.K. receives grant funding from Takeda and Pfizer and has received consulting fees from Takeda. A.P. is a member of Astra Zenecas Genomics Advisory Board. M.A.R. is on the SAB of 54gene and has advised BioMarin, Third Rock Ventures, MazeTx and Related Sciences. G.A.H. is an employee of Takeda, former employee of AbbVie and owns stock in Takeda and AbbVie. C.A.L. reports grants from Genentech, grants and personal fees from Janssen, grants and personal fees from Takeda, grants from AbbVie, personal fees from Ferring, grants from Eli Lilly, grants from Pfizer, grants from Roche, grants from UCB Biopharma, grants from Sanofi Aventis, grants from Biogen IDEC, grants from Orion OYJ, personal fees from Dr Falk Pharma and grants from AstraZeneca, outside the submitted work. H.H.U. reports research collaboration or consultancy with Janssen, Eli Lilly, UCB Pharma, Celgene, MiroBio, OMass and Mestag. D.P.B.M. has consulted for Takeda, Boehringer Ingelheim, Palatin Technologies, Bridge Biotherapeutics, Pfizer and Gilead. M.P. received an unrestricted research grant from Pfizer UK and speaker fees from Janssen. P.I. received lecture fees from AbbVie, BMS, Celgene, Celltrion, Falk Pharma, Ferring, Galapagos, Gilead, MSD, Janssen, Pfizer, Takeda, Tillotts, Sapphire Medical, Sandoz, Shire and Warner Chilcott; financial support for research from Celltrion, MSD, Pfizer and Takeda; advisory fees from AbbVie, Arena, Boehringer Ingelheim, BMS, Celgene, Celltrion, Genentech, Gilead, Hospira, Janssen, Lilly, MSD, Pfizer, Pharmacosmos, Prometheus, Roche, Sandoz, Samsung Bioepis, Takeda, Topivert, VH2, Vifor Pharma and Warner Chilcott. Cedars-Sinai and D.P.B.M. have financial interests in Prometheus Biosciences, a company which has access to the data and specimens in Cedars-Sinais MIRIAD Biobank (including the Cedars-Sinai data and specimens used in this study) and seeks to develop commercial products. H.H. has received consultancy fees from Ono Pharmaceutical and honoraria from Xian Janssen Pharmaceutical. C.A.A. has received consultancy fees from Genomics plc and BridgeBio Inc. and lecture fees from GSK. M.J.D. is a founder of Maze Therapeutics. The remaining authors declare no competing interests.

Additional information

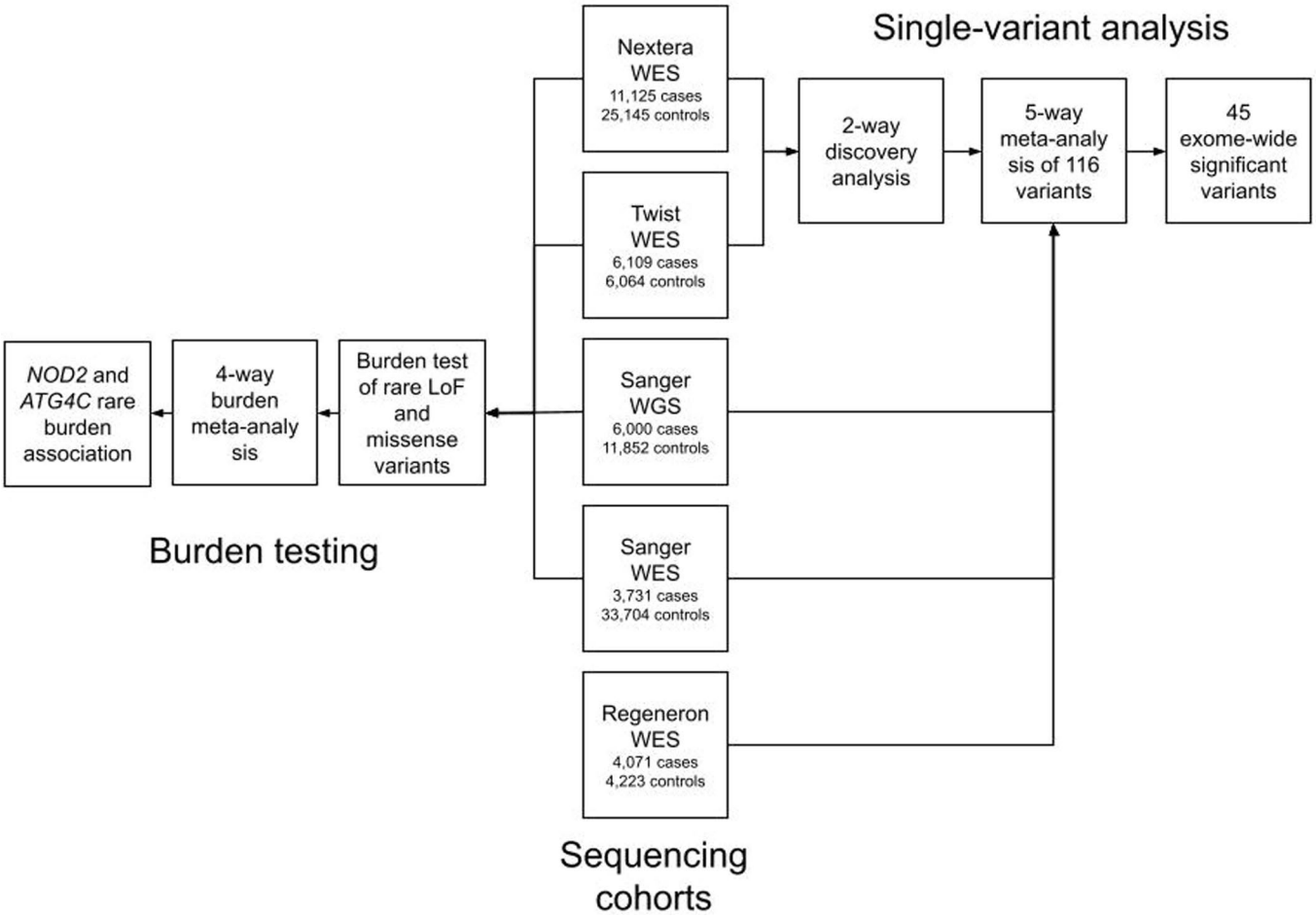
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01156-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01156-2.

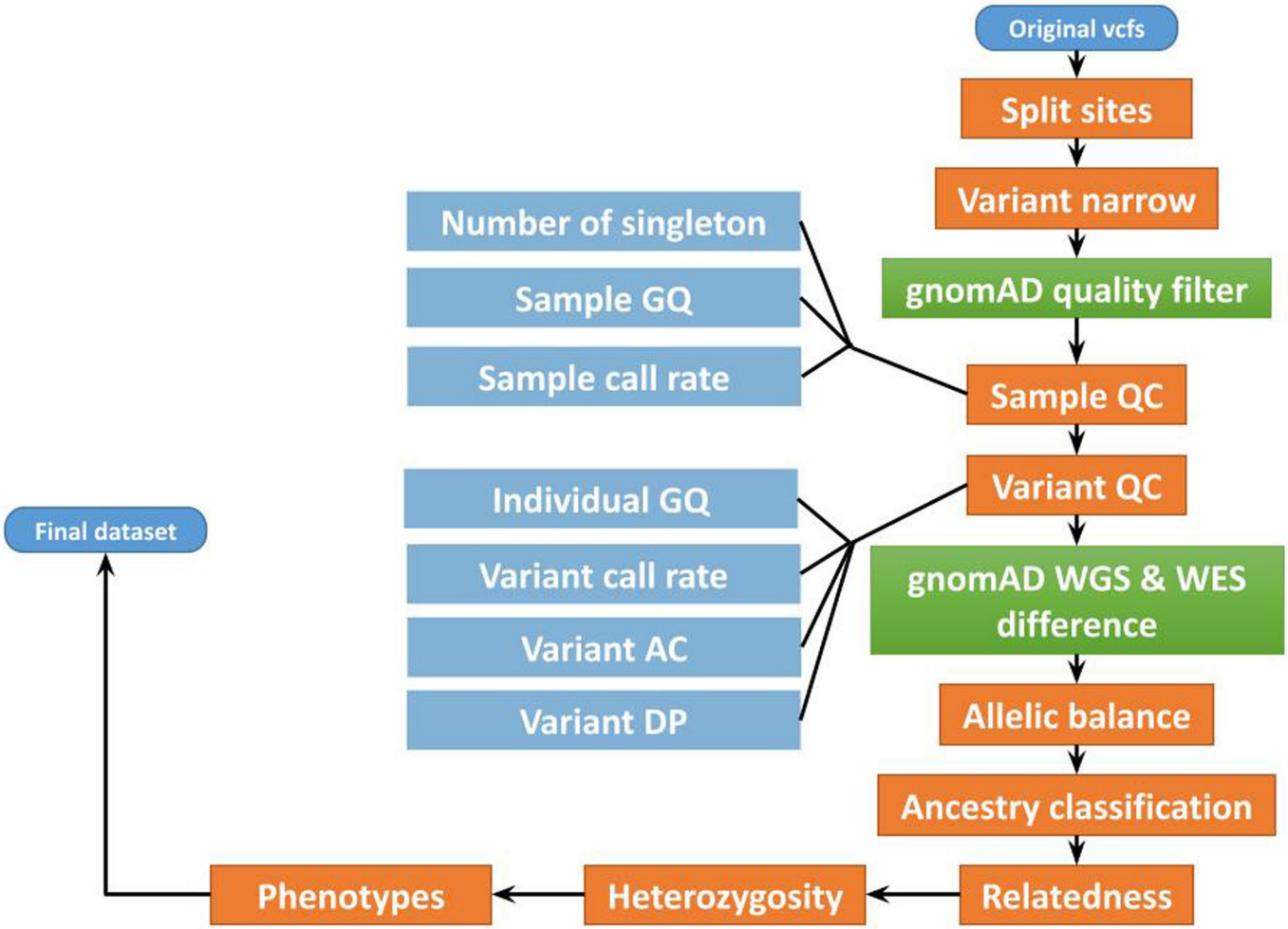
Correspondence and requests for materials should be addressed to Hailiang Huang, Carl A. Anderson or Mark J. Daly.

Peer review information *Nature Genetics* thanks Reedik Mägi, Yukihide Momozawa and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

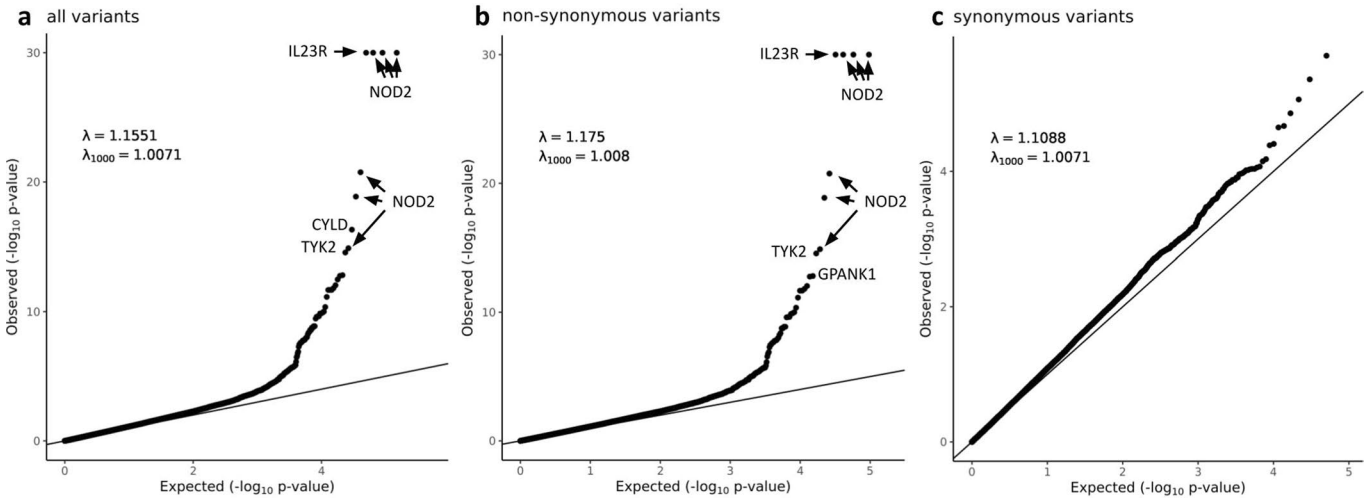
Reprints and permissions information is available at www.nature.com/reprints.



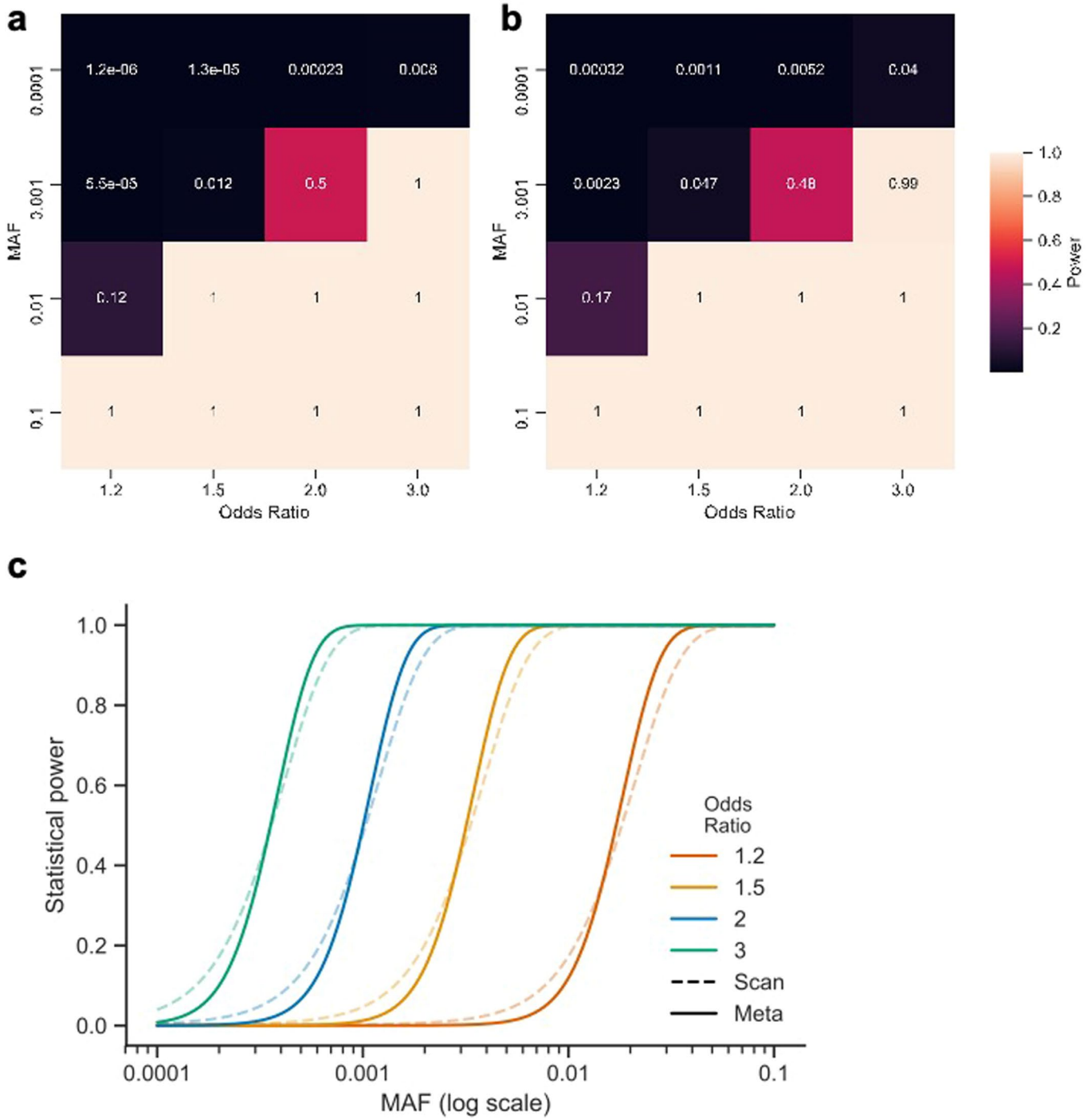
Extended Data Fig. 1 | Overview of the study design. We utilized a logistic mixed-model for the association analysis, followed by meta-analyses to combine multiple cohorts. Multiple cohorts serve the purpose of replication. Two large cohorts at Broad Institute of different exome capture platforms were used to discover candidate variants (Nextera WES and Twist WES). Two independent cohorts at Sanger (Sanger WGS and Sanger WES) and one Kiel/Regeneron cohort (Regeneron WES) were used to replicate the findings.



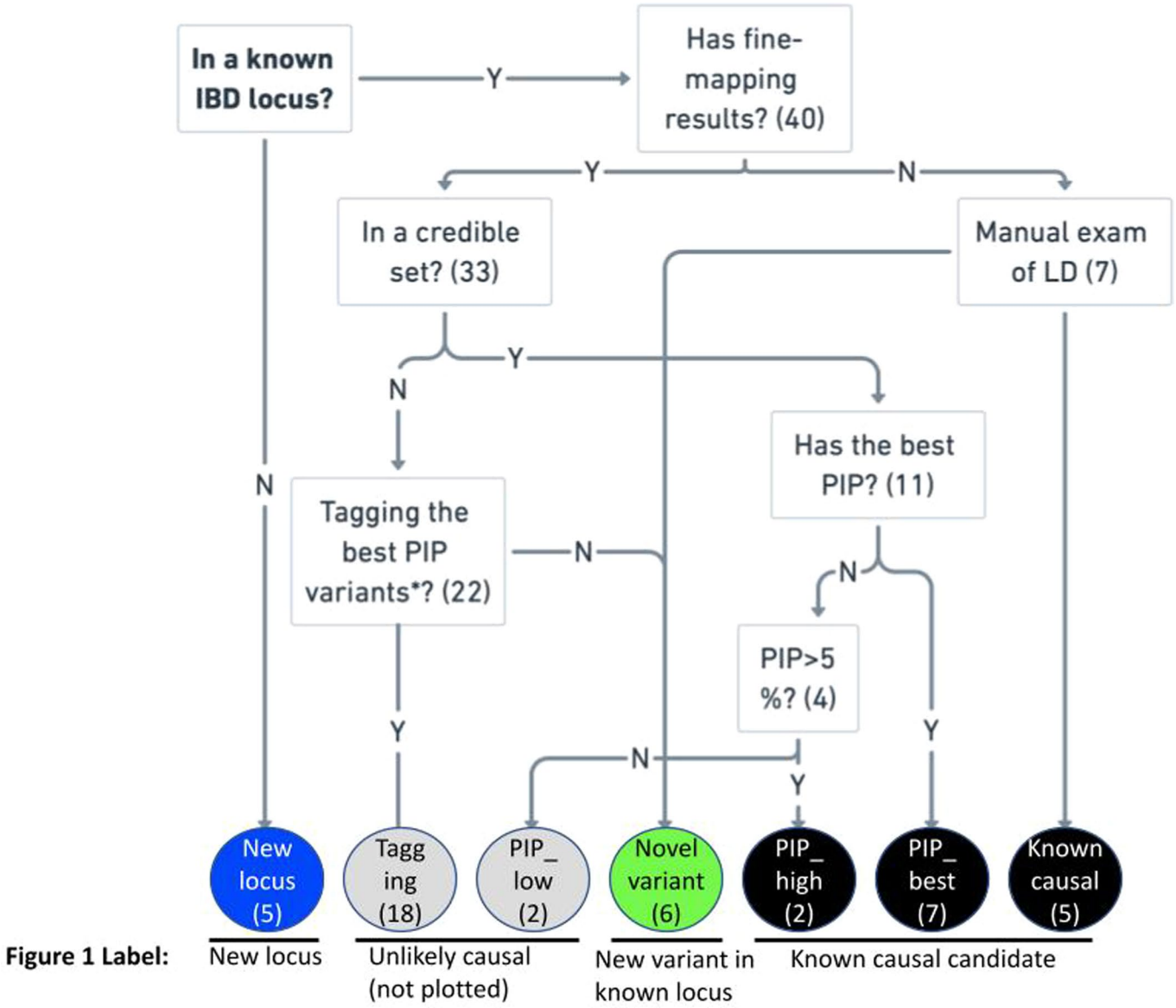
Extended Data Fig. 2 | Quality control procedures applied in the Broad sequencing pipeline. We show as an example the quality control steps performed on variants and subjects from the Broad sequencing platform. Quality controls performed on data from other platforms follow a similar plan and are described in Methods. Quality control steps using external information from gnomAD were colored green. Thresholds and details can be found in Methods.



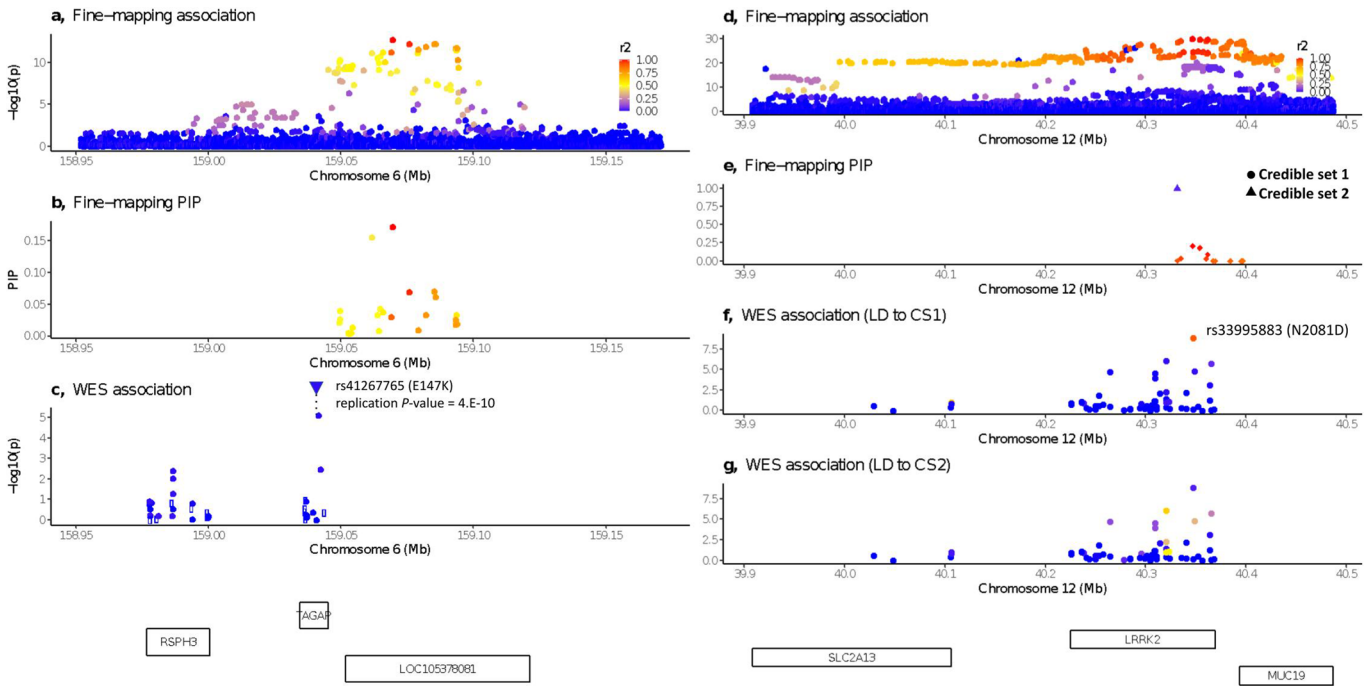
Extended Data Fig. 3 | QQ plots for Nextera and Twist discovery cohorts. Only QC passed variants with minor allele frequency in NFE between 0.0001 and 0.10 were included. **a**, all variants. **b**, non-synonymous variants. **c**, synonymous variants. In **a** and **b**, the y axis is capped at $-\log_{10} p = 30$ while the top four variants (three in *NOD2* and one in *IL23R*) have $-\log_{10} p > 100$. In **c**, to remove the synonymous variants that tag causal non-synonymous variants and artifacts through LD, we removed loci hosting large-effect coding variants (*IL23R*, *NOD2*, *LRRK2*, *TYK2*, *ATG16L1*, *SLC39A8*, *PTGER4*, *IRGM*, *CARD9*), implicated by variants removed in the heterogeneous test (*AHNAK2*, *LILRA*), and with long range LD (MHC).



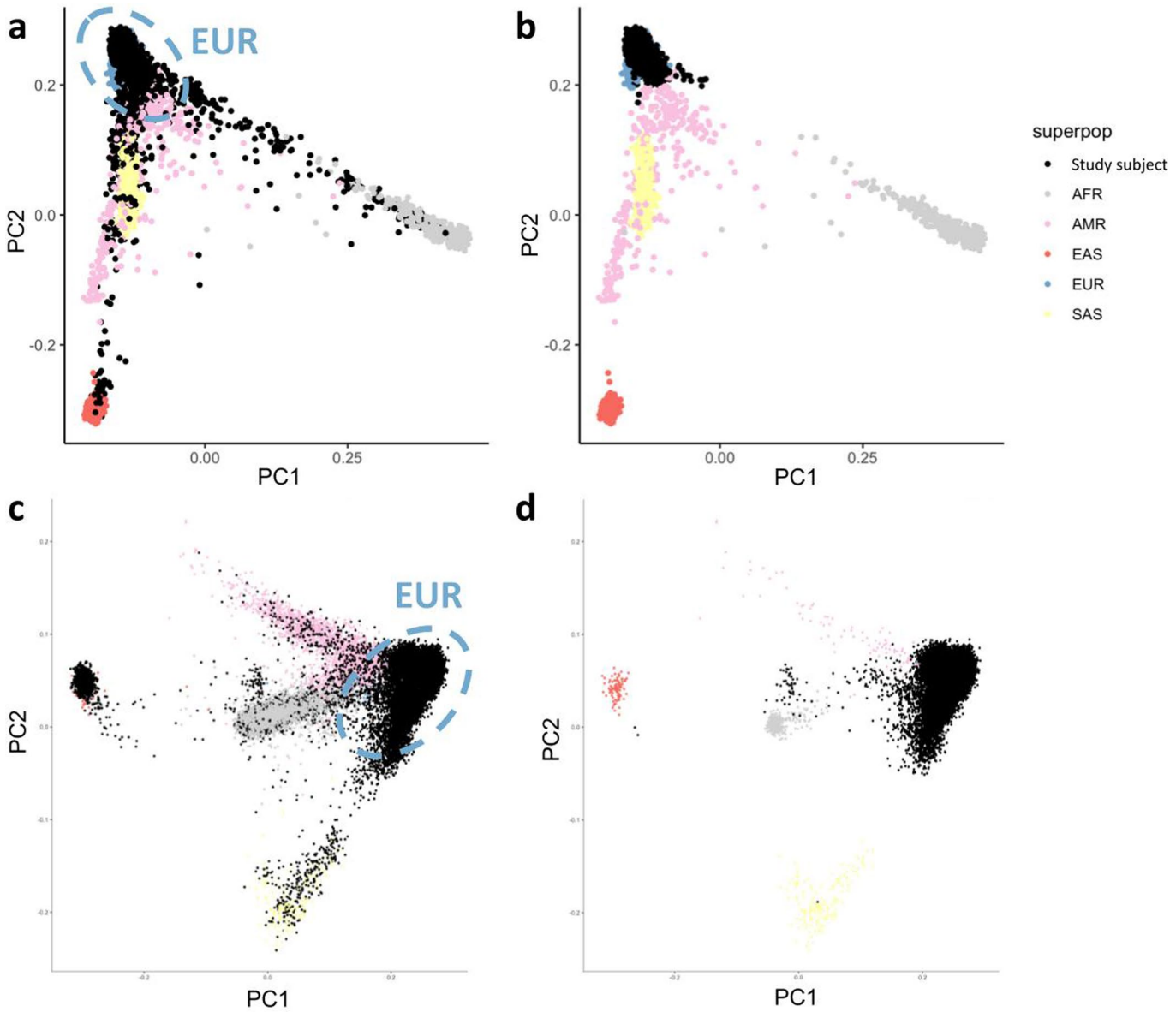
Extended Data Fig. 4 | Power to detect single variant associations. We performed a series of power calculations using the methodology described by Johnson and Abecasis (2017). Our initial 'exome-wide scan' (two cohorts) had fewer samples and a more lenient significance threshold than subsequent meta-analysis (five cohorts). However, both analyses had similar power to detect true associations at their respective significance levels. Our single-variant association analyses did not have the power to uncover association to variants with a MAF = 0.0001 and below (unless the variant has a very strong effect, for example 0.76 power at OR = 8). Similarly, the exome-wide scan had limited power to detect association to variants with a MAF = 0.001 and OR < 2, but was well-powered above these thresholds. **a**, Power of the exome-wide scan analysis **b**, Power of the meta-analysis. **c**, Power to detect single-variant associations at different minor allele frequencies at $\alpha = 0.0002$ ('scan'; dashed lines) and 3 ×10-7 ('meta'; solid lines) and assuming Crohn's disease population prevalence of 276 in 100,000, and an additive effect model.



Extended Data Fig. 5 | Relation to known IBD associations. Numbers in brackets are the number of variants assigned to the categories out of the 45 exome-wide significant variants.



Extended Data Fig. 6 | WES variants from this study implicating known IBD loci. a-c: a novel CD variant implicating *TAGAP*. **d-g:** CD variants tagging fine-mapped IBD associations in *LRRK2*. **a** and **d**, P-value for variants from the fine-mapping study[5]. **b** and **e**, PIP from fine-mapping. **c**, **f** and **g**, P-value for variants from this study. Open circle indicating LD information is missing. LD calculated between the plotted variant and the best variant in **b** for panel **c**, and variants with best PIP in credible sets 1 and 2 (panel **e**) respectively for panels **f** and **g**.



Extended Data Fig. 7 | Nextera and Twist callset population assignment. Principal components for **a, c**, before removing non-European samples for Twist and Nextera respectively. **b, d**, after removing non-European samples for Twist and Nextera respectively. Principal components generated from the 1000 Genome Project Phase III data and different colors stand for different continental / superpopulations. Study subjects (black dots) were projected onto principal components.

nature portfolio

Corresponding author(s): Mark J. Daly

Last updated by author(s): April 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No other direct software used for data collection. GATKv4.1.2.0 used for extracting Plink binary genotype files.
Data analysis	Computer code used in this manuscript: Hail v0.2 (quality control, variants effect annotation; https://hail.is) SAIGE v0.39.5 (variant-based and gene-based burden test for Twist and Nextera samples; https://github.com/weizhouUMICH/SAIGE) METAL v2011-03-25 (meta-analysis; https://genome.sph.umich.edu/wiki/METAL_Documentation) Plink v1.9 (calculating principal components and identity-by-descent relatedness; https://www.cog-genomics.org/plink/) PBK QC pipeline March 30 2020 (ancestry assignment and quality control including the sample heterozygosity; https://github.com/Annefeng/PBK-QC-pipeline) REGENIE v1.0.6.7 (Single-variant test for Sanger samples) REGENIE v2.0.2 (Gene-based burden test for Sanger samples)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genome Reference Consortium Human Build 38 can be accessed at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40/. Sequence data used in this study has been made publicly available in dbGaP Study Accession: phs001642.v1.p1 - Center for Common Disease Genomics [CCDG] - Autoimmune: Inflammatory Bowel Disease (IBD) Exomes and Genomes (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001642.v1.p1). The summary statistics of Nextera and Twist meta-analysis have been deposited on GitHub (https://github.com/yorkklause/Crohn-s-Disease-WES-meta) (doi:10.5281/zenodo.6363106). This research has been conducted using the UK Biobank Resource and controls made publicly available by dbGaP (phs001000.v1.p1, phs000806.v1.p1 - Myocardial Infarction Genetics Consortium (MIGen), phs000401.v1.p1 - NHLBI GO-ESP Project, phs000298.v4.p3 - Autism Sequencing Consortium (ASC), phs000572.v8.p4 - Alzheimer's Disease Sequencing Project (ADSP), phs001489.v1.p1 - Epi25 Consortium, phs001095.v1.p1 - T2D-GENES) as well as additional controls from the 1000 Genomes Project, the Epi25 Collaborative, UK-Ireland Collaborators (A. McQuillin, D. Blackwood, A. McIntosh), and collaborators A. Pulver, H. Ostrer, D. Chung, M. Hiltunen, and A. Palotie. We describe all datasets in the manuscript or its Supplementary Information (Supplementary Table 1).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Previous studies show that larger sample size boost the analysis of IBD [1]. Several recent studies of biobank scale suggest a sample size of over 100,000 individuals will be needed to properly power this study [2-4]. All available samples from collaborators throughout the IIBDGC were used and all replication samples available were included in downstream follow-up.

1. Wei et al. Am J Hum Genet. 2013
2. Wojcik et al. Nature. 2019
3. Feng et al. MedRxiv 2021
4. Kurki et al. MedRxiv 2022

Data exclusions

We first excluded data of poor quality from the analysis (Quality control). After PCA-based ancestry analysis (Population Assignment), we then excluded samples from ancestry groupings with fewer than 250 cases or controls in order to permit robust association analysis. These exclusions were necessary to ensure the quality of the final association statistics from this study. All criteria were established prior to analysis. Here are the details:

Population assignment

Broad Institute cohort and Kiel/Regeneron cohort: We projected all samples onto principal component (PC) axes generated from the 1000 Genomes Project Phase 3 common variants, and classified their ancestry using a random forest method to the European (CEU, TSI, FIN, GBR, IBS), African (YRI, LWK, GWD, MSL, ESN, ASW, ACB), East Asian (CHB, JPT, CHS, CDX, KHV), South Asian (GIH, PJL, BEB, STU, ITU) and American (MXL, PUR, CLM, PEL) samples. We kept samples that were classified as European with prediction probability greater than 80% .
For Nextera samples, we used a second random forest classifier to assign EUR samples to AJ, LIT, FIN, or NFE, and a third random forest classifier to clean the AJ/NFE split.

Sanger Institute cohort: We selected a set of ~14,000 well-genotyped common variants to identify the genetic ancestry of individual participants through the projection of 1000 Genomes Project cohort-derived principal components. For genomes, due to primarily European genetic ancestry of the controls, we excluded samples outside of four median absolute deviations from the median point of the European ancestry cluster of 1000G. For exomes, we implemented a Random Forest technique that classified samples based on principal components into broad genetic ancestry groups (EUR, AFR, SAS, EAS, admixed), with self-reported ancestry as training labels. For these analyses, we only retained the EUR samples, as the number of cases for other groups was too small for robust association analysis.

Quality control

Broad Institute cohort: Poor-quality samples that met the following criteria were identified and removed: 1) samples with an extremely large number of singletons (≥ 500); 2) samples with mean GQ < 40; and 3) samples with missingness rates > 10%. For Twist exome capture samples, we additionally removed 1) samples that had a significantly high or low inbreeding coefficient (> 0.2 or < -0.2); 2) samples that had a high heterozygosity away from mean (± 5 standard deviations); and 3) related samples, which were removed sequentially by removing the individual with the largest number of related samples (in PLINK, the individual with PI_HAT > 0.2 when using the "--genome" option) until no related samples remained. For Nextera capture samples, we additionally removed variants showing a significant heterogeneous effect across Ashkenazi Jewish (AJ), Lithuanian (LIT), Finnish (FIN), and non-Finnish European (NFE) samples

Sanger Institute cohort: A combination of hard-cutoff filters and per-ancestry/per-batch outlier filters were used to identify low-quality samples. We applied hard-filters for sample depth (> 12x genomes, > 15x exomes), call rate (> 0.95), chimerism < 0.5 (WGS) and FREEMIX < 0.02 (WGS). We then performed per-ancestry and per-sequencing protocol (AGILENT vs IDT for WES, PCR vs PCR-free for WES) filtering of samples falling outside 4 MAD from the median per-batch heterozygosity rate, Ti/Tv rate, number of called SNPs and INTELs, and insertion and deletion counts/ratio.

An ancestry-aware relatedness calculation (pc-relate method in Hail) was used to identify related samples. As our association approach (logistic mixed-models) can control for residual relatedness, we only excluded duplicates or MZ twins from within the cohorts and excluded second and third degree relatives when the kinship was across the cohorts (e.g., parent in WGS, child in WES; kinship metric > 0.1 calculated via PC-Relate method using 10 principal components). In addition, we removed samples that were also present in the Broad Institute's cohorts.

Replication

We have multiple cohorts in the study that serve the purpose of replication. Two large cohorts done at Broad Institute of different exome capture platform were used to discover candidate variants. Two independent cohorts at Sanger and one Kiel/Regeneron cohort were used to replicate the findings.
All our findings have been successfully replicated. We nominated an initial list of 116 variants with $p < 0.0002$ in the two Broad cohorts for further evaluation. 28 of 116 variants were associated ($p < 4.3 \times 10\text{-}4$ [0.05/116]) with CD in the meta-analysis of the two Sanger cohorts and 94 replicated the direction of effect seen in the discovery cohort ($p = 3 \times 10\text{-}12$, binomial test). Summary statistics of Kiel/Regeneron were also ascertained and an inverse-variance weighted meta-analysis carried out across all five cohorts. 45 of the 116 variants exceeded the study-wide significance threshold, $p < 3 \times 10\text{-}7$.

Randomization

No randomization was conducted. The current study is a large-scale genetics study. Randomness is achieved from the nature of how alleles are distributed in the population. Alleles are randomly passed from parent to offspring during meiosis, hence the nature of these types of studies tend to be shielded from extraneous confounds of standard epidemiological studies.

Blinding

No blinding was carried out. The nature of the study is in of itself blind to study recruiters. No one would beforehand know the genotype make up of the cases and controls recruited in the current study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Sequencing data was obtained from subjects from North America and Europe. All cases were clinically diagnosed to have IBD. Controls were either unaffected healthy controls from the IBD studies or population based controls from various public resources. 47% of subjects in discovery samples are males and the remaining are females (biological sex). We do not have age, treatment or past diagnosis information, which usually have minimal impact on the quality (e.g., false-positive rates) of large-scale disease genetic studies such as this one.

Recruitment

Please see Supplementary Table 1 for recruitment information.

Ethics oversight

All relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained. The details of the IRB/oversight body that provided approval or exemption for the research described are given below:

Study Protocol 2013P002634, The Broad Institute Study of Inflammatory Bowel Disease Genetics, undergoes annual continuing review by the Mass General Brigham Human Research Committee (MGBHRC) Institutional Review Board (IRB) of Mass General Brigham. Ethical approval was given on January 27, 2021, for this study. Mass General Brigham IRB, Mass General Brigham, 399 Revolution Drive, Suite 710, Somerville, MA 02145

All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

Note that full information on the approval of the study protocol must also be provided in the manuscript.