



## King's Research Portal

DOI:

[10.1038/nature22969](https://doi.org/10.1038/nature22969)

*Document Version*

Peer reviewed version

[Link to publication record in King's Research Portal](#)

*Citation for published version (APA):*

International Inflammatory Bowel Disease Genetics Consortium (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. *NATURE*, 547(7662), 173-178. <https://doi.org/10.1038/nature22969>

### **Citing this paper**

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

### **General rights**

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

### **Take down policy**

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

# 1 **Fine-mapping inflammatory bowel disease loci to single variant**

## 2 **resolution**

3 Hailiang Huang<sup>1,2,\*§</sup>, Ming Fang<sup>3,4\*</sup>, Luke Jostins<sup>5,6\*</sup>, Maša Umičević Mirkov<sup>7</sup>, Gabrielle  
4 Boucher<sup>8</sup>, Carl A Anderson<sup>7</sup>, Vibeke Andersen<sup>9,10</sup>, Isabelle Cleynen<sup>11</sup>, Adrian Cortes<sup>5</sup>, François  
5 Crins<sup>3,4</sup>, Mauro D'Amato<sup>12,13</sup>, Valérie Deffontaine<sup>3,4</sup>, Julia Dimitrova<sup>3,4</sup>, Elisa Docampo<sup>3,4</sup>,  
6 Mahmoud Elansary<sup>3,4</sup>, Kyle Kai-How Farh<sup>1,2,14</sup>, Andre Franke<sup>15</sup>, Ann-Stephan Gori<sup>3,4</sup>, Philippe  
7 Goyette<sup>8</sup>, Jonas Halfvarson<sup>16</sup>, Talin Haritunians<sup>17</sup>, Jo Knight<sup>18</sup>, Ian C Lawrance<sup>19,20</sup>, Charlie W  
8 Lees<sup>21</sup>, Edouard Louis<sup>3,22</sup>, Rob Mariman<sup>3,4</sup>, Theo Meuwissen<sup>23</sup>, Myriam Mni<sup>3,4</sup>, Yukihide  
9 Momozawa<sup>3,4,24</sup>, Miles Parkes<sup>25</sup>, Sarah L Spain<sup>26,27</sup>, Emilie Théâtre<sup>3,4</sup>, Gosia Trynka<sup>7</sup>, Jack  
10 Satsangi<sup>21</sup>, Suzanne van Sommeren<sup>28</sup>, Severine Vermeire<sup>11,29</sup>, Ramnik J Xavier<sup>2,30</sup>, International  
11 IBD Genetics Consortium<sup>†</sup>, Rinse K Weersma<sup>28</sup>, Richard H Duerr<sup>31,32</sup>, Christopher G  
12 Mathew<sup>26,33</sup>, John D Rioux<sup>8,34</sup>, Dermot PB McGovern<sup>17</sup>, Judy H Cho<sup>35</sup>, Michel Georges<sup>3,4§</sup>, Mark  
13 J Daly<sup>1,2§</sup>, Jeffrey C Barrett<sup>7§</sup>

14 <sup>1</sup>Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School,  
15 Boston, Massachusetts, USA. <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. <sup>3</sup>Unit  
16 of Medical Genomics, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA-R) Research  
17 Center, University of Liège, Liège, Belgium. <sup>4</sup>Faculty of Veterinary Medicine, University of Liège, Liège,  
18 Belgium. <sup>5</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Headington, UK. <sup>6</sup>Christ  
19 Church, University of Oxford, St Aldates, UK. <sup>7</sup>Wellcome Trust Sanger Institute, Hinxton (Cambridge),  
20 UK. <sup>8</sup>Research Center, Montreal Heart Institute, Montréal, Québec, Canada. <sup>9</sup>Medical Department, Viborg  
21 Regional Hospital, Viborg, Denmark. <sup>10</sup>Organ Center, Hospital of Southern Jutland Aabenraa, Aabenraa,  
22 Denmark. <sup>11</sup>Department of Clinical and experimental medicine, Translational Research in GastroIntestinal  
23 Disorders (TARGID), Katholieke Universiteit (KU) Leuven, Leuven, Belgium. <sup>12</sup>Clinical Epidemiology  
24 Unit, Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden. <sup>13</sup>Department of Liver and  
25 Gastrointestinal Diseases, BioDonostia Health Research Institute San Sebastian, and IKERBASQUE  
26 Basque Foundation for Science Bilbao, Bilbao, Spain. <sup>14</sup>Illumina, San Diego, California, USA. <sup>15</sup>Institute of  
27 Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. <sup>16</sup>Department of  
28 Gastroenterology, Faculty of Medicine and Health, Örebro University, Örebro, Sweden. <sup>17</sup>F. Widjaja  
29 Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los  
30 Angeles, California, USA. <sup>18</sup>Data Science Institute and Lancaster Medical School, Lancaster University,  
31 LA1 4YG, UK. <sup>19</sup>Centre for Inflammatory Bowel Diseases, Saint John of God Hospital, Subiaco, WA,  
32 Australia. <sup>20</sup>Harry Perkins Institute for Medical Research, School of Medicine and Pharmacology,  
33 University of Western Australia, Murdoch, WA, Australia. <sup>21</sup>Gastrointestinal Unit, Wester General Hospital  
34 University of Edinburgh, Edinburgh, UK. <sup>22</sup>Division of Gastroenterology, Centre Hospitalier Universitaire  
35 (CHU) de Liège, Liège, Belgium. <sup>23</sup>Livestock and Aquacultural Sciences, Norwegian University of Life  
36 Sciences, Ås, Norway. <sup>24</sup>Laboratory for Genotyping Development, Center for Integrative Medical Sciences,  
37 RIKEN, Yokohama, Japan. <sup>25</sup>Inflammatory Bowel Disease Research Group, Addenbrooke's Hospital,  
38 Cambridge, UK. <sup>26</sup>Department of Medical and Molecular Genetics, King's College London, London, UK.  
39 <sup>27</sup>Centre for Therapeutic Target Validation, Wellcome Trust Genome Campus, Hinxton (Cambridge), UK.  
40 <sup>28</sup>Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center  
41 Groningen, Groningen, The Netherlands. <sup>29</sup>Division of Gastroenterology, University Hospital Gasthuisberg,  
42 Leuven, Belgium. <sup>30</sup>Gastroenterology Unit, Massachusetts General Hospital, Harvard Medical School,  
43 Boston, Massachusetts, USA. <sup>31</sup>Division of Gastroenterology, Hepatology and Nutrition, Department of  
44 Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA. <sup>32</sup>Department of  
45 Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania,  
46 USA. <sup>33</sup>Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand,  
47 Johannesburg, South Africa. <sup>34</sup>Faculté de Médecine, Université de Montréal, Montréal, Québec, Canada.  
48 <sup>35</sup>Department of Genetics, Yale School of Medicine, New Haven, Connecticut, USA.

49 \*These authors contributed equally to this work.

50 §These authors contributed equally to this work (corresponding authors).

51 †A full list of members and affiliations appears in the Supplementary Information.

52 **Summary**

53 The inflammatory bowel diseases (IBD) are chronic gastrointestinal inflammatory  
54 disorders that affect millions worldwide. Genome-wide association studies have  
55 identified 200 IBD-associated loci, but few have been conclusively resolved to specific  
56 functional variants. Here we report fine-mapping of 94 IBD loci using high-density  
57 genotyping in 67,852 individuals. We pinpointed 18 associations to a single causal  
58 variant with >95% certainty, and an additional 27 associations to a single variant with  
59 >50% certainty. These 45 variants are significantly enriched for protein-coding changes  
60 (n=13), direct disruption of transcription factor binding sites (n=3) and tissue specific  
61 epigenetic marks (n=10), with the latter category showing enrichment in specific immune  
62 cells among associations stronger in CD and in gut mucosa among associations stronger  
63 in UC. The results of this study suggest that high-resolution fine-mapping in large  
64 samples can convert many GWAS discoveries into statistically convincing causal  
65 variants, providing a powerful substrate for experimental elucidation of disease  
66 mechanisms.

67 The inflammatory bowel diseases (IBD) are a group of chronic, debilitating disorders of  
68 the gastrointestinal tract with peak onset in adolescence and early adulthood. More than  
69 1.4 million people are affected in the USA alone<sup>1</sup>, with an estimated direct healthcare  
70 cost of \$6.3 billion/year. IBD affects millions worldwide, and is rising in prevalence,  
71 particularly in pediatric and non-European ancestry populations<sup>2</sup>. IBD has two subtypes,  
72 ulcerative colitis (UC) and Crohn's disease (CD), which have distinct presentations and  
73 treatment courses. To date, 200 genomic loci have been associated with IBD<sup>3,4</sup>, but only a  
74 handful have been conclusively ascribed to a specific causal variant with direct insight  
75 into the underlying disease biology. This scenario is common to all genetically complex  
76 diseases, where the pace of identifying associated loci outstrips that of defining specific  
77 molecular mechanisms and extracting biological insight from each association.

78 The widespread correlation structure of the human genome (known as linkage  
79 disequilibrium, or LD) often results in similar evidence for association among many  
80 neighboring variants. However, unless LD is perfect ( $r^2 = 1$ ), it is possible, with a  
81 sufficiently large sample size, to statistically resolve causal variants from neighbors even  
82 at high levels of correlation (Extended Data Figure 1 and ref 5). Novel statistical  
83 approaches applied to very large datasets that address this problem<sup>6</sup> require that the  
84 highly correlated variants are directly genotyped or imputed with certainty. Truly high-  
85 resolution mapping data, when combined with increasingly sophisticated and  
86 comprehensive public databases annotating the putative regulatory function of DNA  
87 variants, are likely to reveal novel insights into disease pathogenesis<sup>7-9</sup> and the  
88 mechanisms of disease-associated variants.

## 89 **Genetic architecture of associated loci**

90 We genotyped 67,852 individuals of European ancestry, including 33,595 IBD (18,967  
91 CD and 14,628 UC) and 34,257 healthy controls using the Illumina™ ImmunoChip  
92 (Extended Data Table 1). This genotyping array was designed to include all known  
93 variants from European individuals in the February 2010 release of the 1000 Genomes  
94 Project<sup>10,11</sup> in 187 high-density regions known to be associated to one or more immune-  
95 mediated diseases<sup>12</sup>. Because fine-mapping uses subtle differences in strength of  
96 association between tightly correlated variants to infer which is most likely to be causal,  
97 it is particularly sensitive to data quality. We therefore performed stringent quality  
98 control to remove genotyping errors and batch effects (Methods). We imputed into this  
99 dataset from the 1000 Genomes reference panel<sup>13,14</sup> to fill in variants missing from the  
100 ImmunoChip, or filtered out by our quality control (Extended Data Figure 2). We then  
101 evaluated the 97 high-density regions that had previous IBD associations<sup>3</sup> and contained  
102 at least one variant that showed significant association (Methods) in this data set. The  
103 major histocompatibility complex was excluded from these analyses as fine-mapping has  
104 been reported elsewhere<sup>15</sup>.

105 We applied three complementary Bayesian fine-mapping methods that used  
106 different priors and model selection strategies to identify independent association signals  
107 within a region, and to assign a posterior probability of causality to each variant  
108 (Supplementary Methods and Extended Data Figure 2). For each independent signal  
109 detected by each method, we sorted all variants by the posterior probability of  
110 association, and added variants to the ‘credible set’ of associated variants until the sum of  
111 their posterior probability exceeded 95% – that is, the credible set contains the minimum  
112 list of DNA variants that are >95% likely to contain the causal variant (Figure 1). These

113 sets ranged in size from one to > 400 variants. We merged these results and  
114 subsequently focused only on signals where an overlapping credible set of variants was  
115 identified by at least two of the three methods and all variants were either directly  
116 genotyped or imputed with INFO score > 0.4 (Methods and Figure 1).

117 In three out of 97 regions, a consistent credible set could not be identified; when  
118 multiple independent effects exist in a region of very high LD, multiple distinct fine-  
119 mapping solutions may not be distinguishable (Supplementary Note). Sixty-eight of the  
120 remaining 94 regions contain a single association, while 26 harbor two or more  
121 independent signals, for a total of 139 independent associations defined across the 94  
122 regions (Figure 2a). Only *IL23R* and *NOD2* (both previously established to contain  
123 multiple associated protein-coding variants<sup>16</sup>) contain more than three independent  
124 signals. Consistent with previous reports<sup>3</sup>, the vast majority of signals are associated  
125 with both CD and UC, though many of these have a significantly stronger association  
126 with one subtype. For the purposes of enrichment analyses below, we compared 79  
127 signals that are more strongly associated with CD to 23 signals that are more strongly  
128 associated with UC (the remaining 37 were equally associated with both subtypes,  
129 Supplementary Table 1).

130 Using a restricted maximum likelihood mixed model approach<sup>17</sup>, we evaluated the  
131 proportion of total variance in disease risk attributed to these 94 regions and how much of  
132 that is explained by the 139 specific associations. We estimated that 25% of CD risk was  
133 explained by the specific associations described here, out of a total of 28% explained by  
134 these loci (correspondingly for UC: 17% out of 22%). The single strongest signals in  
135 each region contribute 76% of this variance explained and the remaining associations

136 contribute 24% (Extended Data Figure 3), highlighting the importance of secondary and  
137 tertiary associations in GWAS results<sup>15,18</sup>.

### 138 **Associations mapped to a single variant**

139 For 18 signals, the 95% credible set consisted of a single variant ('single variant credible  
140 sets'), and for 24 others the credible set consisted of two to five variants (Figure 2b). The  
141 single variant credible sets included five previously reported coding variants: three in  
142 *NOD2* (fs1007insC, R702W, G908R), a rare protective allele in *IL23R* (V362I) and a  
143 splice variant in *CARD9* (c.IVS11+1G>C)<sup>16,19</sup>. The remaining single variant credible sets  
144 were comprised of three missense variants (I170V in *SMAD3*, I923V in *IFIH1* and  
145 N289S in *NOD2*), four intronic variants (in *IL2RA*, *LRRK2*, *NOD2* and  
146 *RTEL1/TNFRSF6B*) and six intergenic variants (located 3.7kb downstream of *GPR35*;  
147 3.9kb upstream of *PRDMI*; within a EP300 binding site 39.9 kb upstream of *IKZF1*; 500  
148 bp before the transcription start site of *JAK2*; 9.4kb upstream of *NKX2-3*; and 3.5kb  
149 downstream from *HNF4A*) (Table 1). Of note, while physical proximity does not  
150 guarantee functional relevance, the credible set of variants for 30 associated loci now  
151 implicates a specific gene either because it resides within 50 kb of only that gene or has a  
152 coding variant with >50% probability – improved from only 3 so refined using an earlier  
153 HapMap-based definition. Using the same definitions, the total number of potential  
154 candidate genes was reduced from 669 to 233. Examples of IBD candidate genes clearly  
155 prioritized in our data are described in the Supplementary Box, and a customizable  
156 browser (<http://finemapping.broadinstitute.org/>) is available to review the detailed fine-  
157 mapping results.

158

159 **Associated protein coding variants**

160 We first annotated the possible functional consequences of the IBD variants by their  
161 effect on the amino acid sequences of proteins. Thirteen out of 45 variants (Figure 2c)  
162 that have >50% posterior probability are non-synonymous (Table 1), an 18-fold  
163 enrichment (enrichment  $P=2\times 10^{-13}$ , Fisher's exact test) relative to randomly drawn  
164 variants in our regions (Figure 3a). By contrast, only one variant with >50% probability  
165 is synonymous (enrichment  $P=0.42$ ). All common coding variants previously reported to  
166 affect IBD risk are included in a 95% credible set including: *IL23R* (R381Q, V362I and  
167 G149R); *CARD9* (c.IVS11+1G>C and S12N); *NOD2* (S431L, R702W, V793M, N852S  
168 and G908R, fs1007insC); *ATG16L1* (T300A); *PTPN22* (R620W); and *FUT2* (W154X).  
169 While this enrichment of coding variation (Figure 3a) provides assurance about the  
170 accuracy of our approach, it does not suggest that 30% of all associations are caused by  
171 coding variants; rather, it is almost certainly the case that associated coding variants have  
172 stronger effect sizes, making them easier to fine-map.

173

174 **Associated non-coding variants**

175 We next examined conserved nucleotides in high confidence binding site motifs of 84  
176 transcription factor (TF) families<sup>20</sup> (Methods). There was a significant positive  
177 correlation between TF motif disruption and IBD association posterior probability  
178 ( $P=0.006$ , logistic regression) (Figure 3a), including three variants with >50% probability  
179 (two >95%). In the *RTELI/TNFRSF6B* region, rs6062496 is predicted to disrupt a TF  
180 binding site (TFBS) for EBF1, a TF involved in the maintenance of B cell identity and  
181 prevention of alternative fates in committed cells<sup>21</sup>. A low frequency (3.6%) protective



182 allele at rs74465132 creates a binding site for EP300 less than 40kbp upstream of *IKZF1*.  
183 The third notable example of TFBS disruption, although not in a single variant credible  
184 set, is detailed in the Supplementary Box for the association at *SMAD3*.

185         Recent studies have shown that trait associated variants are enriched for  
186 epigenetic marks highlighting cell type specific regulatory regions<sup>9,22,23</sup>. We compared  
187 our credible sets with ChIPseq peaks corresponding to chromatin immunoprecipitation  
188 with H3K4me1, H3K4me3 and H3K27ac (shown previously<sup>22,23</sup> to highlight enhancers,  
189 promoters and active regulatory elements, respectively) in 120 adult and fetal tissues,  
190 assayed by the Roadmap Epigenomics Mapping Consortium<sup>24</sup> (Figure 3b). Using a  
191 threshold of  $P=1.3 \times 10^{-4}$  (0.05 corrected for 360 tests), we observed significant  
192 enrichment of H3K4me1 in 6 immune cell types and for H3K27ac in 2 gastrointestinal  
193 (gut) samples (sigmoid colon and rectal mucosa) (Figure 3b and Supplementary Table 2).  
194 The subset of signals that are more strongly associated with CD overlap more with  
195 immune cell chromatin peaks, whereas UC signals overlap more with gut chromatin  
196 peaks (Supplementary Table 2).

197         These three chromatin marks are correlated both within tissues (we observe  
198 additional signal in other marks in the tissues described above) and across related tissues.  
199 We therefore defined a set of “core immune peaks” for H3K4me1 and “core gut peaks”  
200 for H3K27ac as the set of overlapping peaks in all enriched immune cell and gut tissue  
201 types, respectively. These two sets of peaks are independently significant and capture the  
202 observed enrichment compared to “control peaks” made up of the same number of  
203 ChIPseq peaks across our 94 regions in non-immune and non-gut tissues (Figure 3c,d).  
204 These two tracks summarize our epigenetic-GWAS overlap signal, and the combined

205 excess over the baseline suggests that a substantial number of regions, particularly those  
206 not mapped to coding variants, may ultimately be explained by functional variation in  
207 recognizable enhancer/promoter elements.

208

### 209 **Overlap with expression QTLs**

210 Variants that change enhancer or promoter activity might change gene expression, and  
211 baseline expression of many genes has been found to be regulated by genetic variation<sup>25-</sup>  
212 <sup>27</sup>. Indeed, it has been suggested that these so-called expression quantitative trait loci  
213 (eQTLs) underlie a large proportion of GWAS associations<sup>25,28</sup>. We therefore searched  
214 for variants that are both in an IBD-associated credible set with 50 or fewer variants, and  
215 the most significantly associated eQTL variant for a gene in a study<sup>29</sup> of peripheral blood  
216 mononuclear cells (PBMC) from 2,752 twins. Sixty-eight of the 76 regions with signals  
217 fine-mapped to  $\leq 50$  variants harbor at least one significant eQTL (affecting a gene  
218 within 1 Mb with  $P < 10^{-5}$ ). Despite this abundance of eQTLs in fine-mapped regions,  
219 only 3 credible sets include the most significantly associated eQTL variants, compared  
220 with 3.7 expected by chance (Methods). Data from a more recent study<sup>30</sup> using PBMCs  
221 from 8,086 individuals did not yield a substantively different outcome, demonstrating a  
222 modest but non-significant enrichment (8 observed overlaps, 4.2 expected by chance,  
223  $P=0.06$ ). Using a more lenient definition of overlap which requires the lead eQTL  
224 variant to be in LD ( $R^2 > 0.4$ ) with an IBD credible set variant increased the number of  
225 potential overlaps but again these numbers were not greater than chance expectation.

226 As PBMCs are a heterogeneous collection of immune cell populations, cell type-  
227 specific signals or signals corresponding to genes expressed most prominently in non-

228 immune tissues may be missed. We therefore tested the enrichment of eQTLs that  
229 overlap credible sets in five primary immune cell populations (CD4+, CD8+, CD19+,  
230 CD14+ and CD15+), platelets, and three distinct intestinal locations (rectum, colon and  
231 ileum) isolated from 350 healthy individuals (Methods). We observed a significant  
232 enrichment of credible SNP/eQTL overlaps in CD4+ cells and ileum (Extended Data  
233 Table 2): three and two credible sets overlapped eQTLs, respectively, compared to 0.4  
234 and 0.3 expected by chance ( $P=0.005$  and  $0.020$ ). An enrichment was also observed for  
235 the naïve CD14+ cells from another study<sup>31</sup>: eight overlaps observed compared to 2.7  
236 expected by chance ( $P=0.001$ ). We did not observe enrichment of overlaps in stimulated  
237 (with interferon or lipopolysaccharide) CD14+ cells from the same source (Extended  
238 Data Table 2).

239 We investigated eQTL overlaps more deeply by applying two colocalization  
240 approaches (one frequentist, one Bayesian, Methods) to the our cell-separated dataset  
241 where primary genotype and expression data were available. We confirmed greater than  
242 expected overlap with eQTLs in CD4+ and ileum described above (Figure 4 and  
243 Extended Data Table 2). These CD4+ colocalized eQTLs also had stronger overlap with  
244 CD4+ ChIPseq peaks than our other credible sets, further supporting a regulatory causal  
245 mechanism. The number of colocalizations in other purified cell types and tissues was  
246 largely indistinguishable from what we expect under the null using either method, except  
247 for moderate enrichment in rectum (4 observed and 1.4 expected,  $P=0.039$ , Frequentist  
248 approach) and colon (3 observed and 0.8 expected,  $P=0.04$ , Bayesian approach). Only  
249 two of these colocalizations correspond to an IBD variant with causal probability  $> 50\%$   
250 (Table 1 and Extended Data Figure 4a).

251

252 **Discussion**

253 We have performed fine-mapping of 94 previously reported genetic risk loci for IBD.  
254 Rigorous quality control followed by an integration of three novel fine-mapping methods  
255 generated lists of genetic variants accounting for 139 independent associations across  
256 these loci. Our methods are concordant with an existing fine-mapping method<sup>6</sup> (67 of 68  
257 credible sets in single signal regions overlap, including exact matches for all single  
258 variant credible sets), and provide extensions to support the phenotype assignment (CD,  
259 UC or IBD) and the conditional estimation of multiple credible sets in loci with multiple  
260 independent signals. The use of multiple methods allowed us to focus our downstream  
261 analyses on loci where the choice of fine-mapping method did not substantially alter  
262 conclusions about the biology of IBD. Our results improve on previous fine-mapping  
263 efforts using a preset LD threshold<sup>32</sup> (e.g.  $r^2 > 0.6$ ) (Extended Data Figure 5) by formally  
264 modeling the posterior probability of association of every variant. Much of this  
265 resolution derives from the very large sample size we employed, because the number of  
266 variants in a credible set decreases with increasing significance ( $P=0.0069$ ).

267         The high-density of genotyping also aids in improved resolution. For instance,  
268 the primary association at *IL2RA* has now been mapped to a single variant associated  
269 with CD, rs61839660. This variant was not present in the Hapmap 3 reference panel and  
270 was therefore not reported in earlier studies<sup>3,33</sup> (nearby tagging variants, rs12722489 and  
271 rs12722515, were reported instead). Imputation using the 1000 genomes reference panel  
272 and the largest assembled GWAS dataset<sup>3</sup> did not separate rs61839660 from its neighbors  
273 (unpublished results), due to the loss of information in imputation using the limited

274 reference. Only direct genotyping, available in the immunochip high-density regions,  
275 permitted the conclusive identification of the causal variant.

276 Accurate fine-mapping should, in many instances, ultimately point to the same  
277 variant across diseases in shared loci. Among our single-variant credible sets, we fine-  
278 mapped a UC association to a rare missense variant (I923V) in *IFIH1*, which is also  
279 associated with type 1 diabetes (T1D)<sup>37</sup> with an opposite direction of effect  
280 (Supplementary Box). The intronic variant noted above (rs61839660, AF=9%) in *IL2RA*  
281 was also similarly associated with T1D, again with a discordant directional effect<sup>38</sup>  
282 (Supplementary Box). Simultaneous high-resolution fine-mapping in multiple diseases  
283 should therefore better clarify both shared and distinct biology.

284 Resolution of fine-mapping can be further improved by leveraging LD from other  
285 ethnicities<sup>34</sup>. However, the sample size from other ethnicities we have collected is small  
286 compared with European samples (9,846 across East-Asian, South-Asian and Middle-  
287 Eastern). Limited access to matched imputation reference panels from all cohorts and the  
288 fact that the smaller non-European sets are not from populations (e.g., African-derived)  
289 with narrower LD also suggest that gains in fine-mapping accuracy would be limited at  
290 this time. Ultimately this effort will be aided by more substantial investment in  
291 genotyping non-European population samples and by developing and applying more  
292 robust trans-ethnic fine-mapping algorithms.

293 A new release of the 1000 genomes (phase 3)<sup>35</sup> and the UK10K<sup>36</sup> project have  
294 introduced new variants that were not present in the reference panel in our study. Our  
295 major findings remain the same using this new reference panel: the 18 single-variant  
296 credible sets are not in high LD ( $r^2 > 0.95$ ) with any new variants in either new dataset,

297 and the 1,426 variants in IBD associations mapped to  $\leq 50$  variants are in high LD with  
298 only 47 new variants (3.3% of the total size of these credible sets, Supplementary Table  
299 1). Given that this release represents a near complete catalogue of variants with minor  
300 allele frequency (MAF)  $> 1\%$  in European populations, we believe our current fine  
301 mapping results are likely to be robust, especially for common variant associations. High-  
302 resolution fine-mapping demonstrates that causal variants are significantly enriched for  
303 variants that alter protein coding variants or disrupt transcription factor binding motifs.  
304 Enrichment was also observed in H3K4me1 marks in immune related cell types and  
305 H3K27ac marks in sigmoid colon and rectal mucosal tissues, with CD loci demonstrating  
306 a stronger immune signature and UC loci more enriched for gut tissues ( $P$  values are  
307 0.014, 0.0005 and 0.0013 respectively for H3K4me1, H3K27ac and H3K4me3; chi-  
308 square test). By contrast, overall enrichment of eQTLs is quite modest compared with  
309 prior reports and not seen strongly in excess of chance in our well-refined credible sets ( $\leq$   
310 50 variants). This result underscores the importance of high-resolution mapping and the  
311 careful incorporation of the high background rate of eQTLs. It is worth noting that  
312 evaluating the overlap between two distinct mapping results is fundamentally different  
313 than comparing genetic mapping results to fixed genomic features, and depends on both  
314 mappings being well resolved. While these data challenge the paradigm that easily  
315 surveyed baseline eQTLs explain a large proportion of non-coding GWAS signals, the  
316 modest excesses observed in smaller but cell-specific data sets suggest that much larger  
317 tissue or cell-specific studies (and under the correct stimuli or developmental time points)  
318 will resolve the contribution of eQTLs to GWAS hits.

319           Resolving multiple independent associations may often help target the causal gene  
320 more precisely. For example, the *SMAD3* locus hosts a non-synonymous variant and a  
321 variant disrupting the conserved transcription factor binding site (also overlapping the  
322 H3K27ac marker in gut tissues), unambiguously articulating a role in disease and  
323 providing an allelic series for further experimental inquiry. Similarly, the *TYK2* locus has  
324 been mapped to a non-synonymous variant and a variant disrupting a conserved  
325 transcription factor binding site (<http://finemapping.broadinstitute.org/>).

326           One-hundred and sixteen associations have been fine-mapped to  $\leq 50$  variants.  
327 Among them, 27 associations contain coding variants, 20 contain variants disrupting  
328 transcription factor binding motifs, and 45 are within histone H3K4me1 or H3K27ac  
329 marked DNA regions. The best-resolved associations - 45 variants having >50% posterior  
330 probabilities for being causal (Table 1) – are similarly significantly enriched for variants  
331 with known or presumed function from genome annotation. Of these, 13 variants cause  
332 non-synonymous change in amino acids, three disrupt a conserved TF binding motif, ten  
333 are within histone H3K4me1 or H3K27ac marked DNA regions in disease-relevant  
334 tissues, and two colocalize with a significant *cis*-eQTL (Extended Data Figure 4a). Risk  
335 alleles of these variants can be found throughout the allele frequency spectrum, with  
336 protein coding variants having somewhat larger effects and more extreme risk allele  
337 frequencies (Extended Data Figure 6a-c).

338           This analysis, however, leaves 21 non-coding variants (Extended Data Figure 4b),  
339 all of which have >50% probabilities to be causal (five have >95%), that are not located  
340 within known motifs, annotated elements, nor in any experimentally determined ChIPseq  
341 peaks or eQTL credible sets yet discovered. While we have identified a statistically

342 compelling set of genuine associations (often intronic or within 10 kb of strong candidate  
343 genes), we can make little inference about function. For example, the intronic single-  
344 variant credible set of *LRRK2* has no annotation, eQTL or ChIPseq peak of note. This  
345 underscores the incompleteness of our knowledge regarding the function of non-coding  
346 DNA and its role in disease, and calls for comprehensive studies on transcriptome and  
347 epigenome in a wide range of cell lines and stimulation conditions. That the majority of  
348 the best-refined non-coding associations have no available annotation is perhaps sobering  
349 with respect to how well we may currently be able to interpret non-coding variation in  
350 medical sequencing efforts. It does suggest, however, that detailed fine-mapping of  
351 GWAS signals down to single variants, combined with emerging high-throughput  
352 genome-editing methodology, may be among the most effective ways to advance to a  
353 greater understanding of the biology of the non-coding genome.



354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399

## References

1. Kappelman, M. D. *et al.* Direct health care costs of Crohn's disease and ulcerative colitis in US children and adults. *Gastroenterology* **135**, 1907–1913 (2008).
2. Molodecky, N. A. *et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology* **142**, 46–54.e42– quiz e30 (2012).
3. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
4. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* (2015). doi:10.1038/ng.3359
5. van de Bunt, M. *et al.* Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS Genet* **11**, e1005535 (2015).
6. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).
7. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).
8. International Multiple Sclerosis Genetics Consortium (IMSGC) *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet* **45**, 1353–1360 (2013).
9. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* **47**, 381–386 (2015).
10. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
11. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
12. Jostins, L. Using Next-Generation Genomic Datasets In Disease Association. (The University of Cambridge, 2012).
13. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
14. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
15. Goyette, P. *et al.* High-density mapping of the MHC identifies a shared role for HLA-DRB1\*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet* (2015). doi:10.1038/ng.3176
16. 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).
17. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* **43**, 519–525 (2011).
18. Huang, H., Chanda, P., Alonso, A., Bader, J. S. & Arking, D. E. Gene-based tests of association. *PLoS Genet* **7**, e1002177 (2011).
19. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease.

- 400 *Nat Genet* **43**, 43–47 (2011).
- 401 20. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of  
402 regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**,  
403 2976–2987 (2014).
- 404 21. Nechanitzky, R. *et al.* Transcription factor EBF1 is essential for the maintenance  
405 of B cell identity and prevention of alternative fates in committed cells. *Nat.*  
406 *Immunol.* **14**, 867–875 (2013).
- 407 22. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping  
408 complex trait variants. *Nat Genet* **45**, 124–130 (2013).
- 409 23. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune  
410 disease variants. *Nature* **518**, 337–343 (2015).
- 411 24. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. **28**,  
412 1045–1048 (2010).
- 413 25. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression  
414 QTLs with complex trait genetic associations. *PLoS Genet* **6**, e1000895 (2010).
- 415 26. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional  
416 variation in humans. *Nature* **501**, 506–511 (2013).
- 417 27. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and  
418 genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21**, 2815–  
419 2824 (2012).
- 420 28. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing  
421 immune gene expression. *Nat Genet* **42**, 295–302 (2010).
- 422 29. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral  
423 blood. *Nat Genet* **46**, 430–437 (2014).
- 424 30. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of  
425 known disease associations. *Nat Genet* **45**, 1238–1243 (2013).
- 426 31. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory  
427 variants upon monocyte gene expression. **343**, 1246949–+ (2014).
- 428 32. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci.  
429 *Nature* **511**, 421–427 (2014).
- 430 33. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of  
431 confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118–1125 (2010).
- 432 34. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Human*  
433 *Molecular Genetics* **24**, R111–9 (2015).
- 434 35. 1000 Genomes Project Consortium *et al.* A global reference for human genetic  
435 variation. *Nature* **526**, 68–74 (2015).
- 436 36. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and  
437 disease. *Nature* (2015). doi:10.1038/nature14962
- 438 37. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of  
439 IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.  
440 **324**, 387–389 (2009).
- 441 38. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based  
442 imputation identifies novel and refined associations for the Wellcome Trust Case  
443 Control Consortium phase 1 Data. *Eur. J. Hum. Genet.* **20**, 801–805 (2012).
- 444 39. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference  
445 human epigenomes. *Nature* **518**, 317–330 (2015).

446 **Acknowledgements** We thank Mariya Khan and Bang Wong for their assistance in  
447 designing illustrations, and Katie de Lange for helpful comments on the Supplementary  
448 Methods. Support from the following grants: M.J.D. and R.J.X.: P30DK43351,  
449 U01DK062432, R01DK64869, Helmsley grant 2015PG-IBD001 and CCFA. C.A.A. and  
450 J.C.B: Wellcome Trust grant 098051. M.G.:WELBIO (CAUSIBD), BELSPO (BeMGI),  
451 Fédération Wallonie-Bruxelles (ARC IBD@Ulg), and Région Wallonne (CIBLES,  
452 FEDER). H.H.: ASHG/Charles J. Epstein Trainee Award. J.L.: Wellcome Trust  
453 098759/Z/12/Z. D.M.: Olle Engkvist Foundation and Swedish Research Council (grants  
454 2010-2976 and 2013-3862). R.K.W.: VIDI grant (016.136.308) from the Netherlands  
455 Organization for Scientific Research. J.D.R.: Canada Research Chair, NIDDK grants  
456 DK064869 and DK062432, CIHR #GPG-102170 from the Canadian Institutes of Health,  
457 GPH-129341 from Genome Canada and Génome Québec, and Crohn's Colitis Canada.  
458 J.H.C.: DK062429, DK062422, DK092235, DK106593, and the Sanford J. Grossman  
459 Charitable Trust. R.H.D.: Inflammatory Bowel Disease Genetic Research Chair at the  
460 University of Pittsburgh, U01DK062420 and R01CA141743. E.D.: Marie-Curie  
461 Fellowship. A-S.G: FNRS and Fonds Léon Fredericq fellowships. J.H.: Örebro  
462 University Hospital Research Foundation and the Swedish Research Council (grant no.  
463 521 2011 2764). C.G.M. and M.P.: NIHR Biomedical Research Centre awards to Guy's  
464 & St Thomas' NHS Trust / King's College London and to Addenbrooke's Hospital /  
465 University of Cambridge School of Clinical Medicine. D.E.: German Federal Ministry of  
466 Education and Research (SysInflame grant 01ZX1306A), DFG Excellence Cluster No.  
467 306 "Inflammation at Interfaces". A.F: professorship of Foundation for Experimental  
468 Medicine (Zuerich, Switzerland). D.P.B.M.: DK062413, AI067068, U54DE023789-01,  
469 305479 from the European Union, and The Leona M. and Harry B. Helmsley Charitable  
470 Trust. Additional acknowledgements for the original data are in the Supplementary  
471 Information.

472

473 **Author Contributions** Overall project supervision and management: M.J.D. J.C.B, M.G.  
474 Fine-mapping algorithms: H.H., M.F., L.J. TFBS analyses: H.H., K.F. Epigenetic  
475 analyses: M.U.M., G.T. eQTL dataset generation: E.L., E.T., J.D., E.D., M.E., R.M.,  
476 M.M., Y.M., V.D., A.G. eQTL analyses: M.F., J.D., L.J., A.C. Variance component  
477 analysis: T.M., M.F. Contribution to overall statistical analyses: G.B. Primary drafting of  
478 the manuscript: M.J.D., J.C.B, M.G., H.H., L.J. Major contribution to drafting of the  
479 manuscript: M.F., M.U.M., J.H.C., D.P.B.M., J.D.R., C.G.M., R.H.D., R.K.W. The  
480 remaining authors contributed to the study conception, design, genotyping QC and/or  
481 writing of the manuscript. All authors saw, had the opportunity to comment on, and  
482 approved the final draft.

483

484 **Author Information** The study protocols were approved by the institutional review  
485 board (IRB) at each center involved with recruitment. Informed consent and permission  
486 to share the data were obtained from all subjects, in compliance with the guidelines  
487 specified by the recruiting center's IRB. The authors declare no competing financial  
488 interests. Correspondence and requests for materials should be addressed to H.H.  
489 (hhuang@atgu.mgh.harvard.edu), M.G. (michel.georges@ulg.ac.be), M.J.D.  
490 (mj Daly@atgu.mgh.harvard.edu) or J.C.B. (barrett@sanger.ac.uk)

491

**Table 1. Variants having posterior probability >50%.**

Variant	Chr	Position	Ns	Phe	AF	Prob	INFO	Func	Annotation
Signals mapped to a single variant									
rs7307562	12	40724960	2	CD	0.398	0.999	1		<i>LRRK2</i> (intronic)
rs2066844	16	50745926	10	CD	0.063	0.999	0.8	C	<i>NOD2</i> (R702W)
rs2066845	16	50756540	10	CD	0.022	0.999	1	C	<i>NOD2</i> (G908R)
rs6017342	20	43065028	2	UC	0.544	0.999	1	E	<i>HNF4A</i> (downstream), Gut_H3K27ac
rs61839660	10	6094697	2	CD	0.094	0.999	1	E	<i>IL2RA</i> (intronic), Immune_H3K4me1
rs5743293	16	50763781	10	CD	0.964	0.999	1	C	<i>NOD2</i> (fs1007insC)
rs6062496	20	62329099	1	IBD	0.587	0.996	1	T	<i>RTEL1</i> - <i>TNFRSF6B</i> (ncRNA_intronic), EBF1 TFBS
rs141992399	9	139259592	3	IBD	0.005	0.995	1	C	<i>CARD9</i> (1434+1G>C)
rs35667974	2	163124637	1	UC	0.021	0.994	1	C	<i>IFIH1</i> (I923V)
rs74465132	7	50304782	3	IBD	0.034	0.994	1	T,E	<i>IKZF1</i> (upstream), EP300 TFBS, Immune_H3K4me1
rs4676408	2	241574401	1	UC	0.508	0.994	0.99		<i>GPR35</i> (downstream)
rs5743271	16	50744688	10	CD	0.007	0.993	1	C	<i>NOD2</i> (N289S)
rs10748781	10	101283330	2	IBD	0.55	0.990	1	E	<i>NKX2-3</i> (upstream), Gut_H3K27ac
rs35874463	15	67457698	2	IBD	0.054	0.989	1	C,E	<i>SMAD3</i> (I170V), Gut_H3K27ac
rs72796367	16	50762771	10	CD	0.023	0.983	1		<i>NOD2</i> (intronic)
rs1887428	9	4984530	1	IBD	0.603	0.974	0.97		<i>JAK2</i> (upstream)
rs41313262	1	67705900	5	CD	0.014	0.973	1	C	<i>IL23R</i> (V362I)
rs28701841	6	106530330	2	CD	0.116	0.971	1		<i>PRDM1</i> (upstream)
Signals mapped to 2-50 variants and the lead variant has posterior probability > 50%									
rs76418789	1	67648596	5	CD	0.006	0.937	0.59	C	<i>IL23R</i> (G149R)
rs7711427	5	40414886	3	CD	0.633	0.919	1		
rs1736137	21	16806695	2	CD	0.407	0.879	1		
rs104895444	16	50746199	10	CD	0.003	0.865	1	C	<i>NOD2</i> (V793M)
rs56167332	5	158827769	2	IBD	0.353	0.845	1		<i>IL12B</i>
rs104895467	16	50750810	10	CD	0.002	0.833	1	C	<i>NOD2</i> (N852S)
rs630923	11	118754353	2	CD	0.153	0.820	0.98		
rs3812565	9	139272502	3	IBD	0.402	0.815	1	Q	eQTL of <i>INPP5E</i> in CD4 and CD8; <i>CARD9</i> in CD14 Gut_H3K27ac
rs4655215	1	20137714	3	UC	0.763	0.784	1	E	
rs145530718	19	10568883	3	CD	0.023	0.762	0.97		
rs6426833	1	20171860	3	UC	0.555	0.752	1		
chr20: 43258079	20	43258079	2	CD	0.041	0.736	0.88		
rs17229679	2	199560757	2	UC	0.028	0.716	1		
rs4728142	7	128573967	1	UC	0.448	0.664	1	E	Immune_H3K4me1
rs2143178	22	39660829	2	IBD	0.157	0.662	1	T,E	NFKB TFBS, Gut_H3K27ac
rs34536443	19	10463118	3	CD	0.038	0.649	1	C	<i>TYK2</i> (P1104A)
rs138425259	16	50663477	10	UC	0.009	0.648	0.92		
rs146029108	9	139329966	3	CD	0.036	0.643	0.92		
rs12722504	10	6089777	2	CD	0.26	0.615	1		
rs60542850	19	10488360	3	IBD	0.17	0.591	0.89		
rs2188962	5	131770805	1	CD	0.44	0.590	1	E,Q	Gut_H3K27ac, eQTL of <i>SLC22A5</i> in CD14, CD15 and IL
rs2019262	1	67679990	5	IBD	0.4	0.586	1		
rs3024493	1	206943968	2	IBD	0.171	0.537	1	E	Immune_H3K4me1
rs7915475	10	64381668	3	CD	0.304	0.528	1		
rs77981966	2	43777964	1	CD	0.077	0.521	1		
rs9889296	17	32570547	1	CD	0.264	0.512	1		
rs2476601	1	114377568	1	CD	0.908	0.508	1	C	<i>PTPN22</i> (W620R)

493

Ns: number of independent signals in the locus. Phe: phenotype. AF: allele frequency.

494

Prob: posterior probability for being a causal variant. INFO: imputation. Func: functional

495

annotations -- coding (C), disrupting transcription factor binding sites (T), overlapping

496

epigenetic peaks (E) and colocalization with eQTL (Q).

497

498 **Figure 1. Fine-mapping procedure and output using the *SMAD3* region as an**  
499 **example. a,** 1) We merge overlapping signals across methods; 2) select a lead variant  
500 (black triangle) and phenotype (color); and 3) choose the best model. Details for each  
501 step are available in Methods. **b,** Example fine-mapping output. This region has been  
502 mapped to two independent signals. For each signal, we report the phenotype it is  
503 associated with (colored), the variants in the credible set, and their posterior probabilities.  
504

505 **Figure 2. Summary of fine-mapped associations. a,** Independent signals. Sixty-eight  
506 loci containing one association and 26 loci containing multiple associations. **b,** Number  
507 of variants in credible sets. 18 associations were fine-mapped to a single variant, and 116  
508 to  $\leq 50$  variants. **c,** Distribution of the posterior probability of the variants in credible sets  
509 having  $\leq 50$  variants.  
510

511 **Figure 3. Functional annotation of causal variants. a,** Proportion of credible variants  
512 that are protein coding, disrupt/create transcription factor binding motif sites (TFBS) or  
513 are synonymous, sorted by posterior probability. **b,** Epigenetic peaks overlapping  
514 credible variants in cell and tissue types from the Roadmap Epigenomics Consortium<sup>39</sup>.  
515 Significant enrichment has been marked with asterisks. Proportion of credible variants  
516 that overlap **(c)** core immune peaks for H4K4me1 or **(d)** core gut peaks for H3K27ac  
517 (Methods). In panels **a, c** and **d**, the vertical dotted lines mark 50% posterior probability  
518 and the horizontal dashed lines show the background proportions of each functional  
519 category.  
520

521 **Figure 4. Number of credible sets that colocalize eQTLs.** Distributions of the number  
522 of colocalizations by chance (violins) and observed number of colocalizations with p-  
523 values (dots). Both the background and the observed numbers were calculated using the  
524 “Frequentist colocalization using conditional *P* values” approach (Methods).

## 525 **Methods**

### 526 **Genotyping and QC**

527 We genotyped 35,197 unaffected and 35,346 affected individuals (20,155 CD and 15,191  
528 UC) using the Immunochip array. Genotypes were called using optiCall<sup>40</sup> for 192,402  
529 autosomal variants before QC. We removed variants with missing data rate >2% across  
530 the whole dataset, or >10% in any one batch, and variants that failed (FDR < 10<sup>-5</sup> in  
531 either the whole dataset or at least two batches) tests for: a) Hardy-Weinberg equilibrium  
532 in controls; b) differential missingness between cases and controls; c) different allele  
533 frequency across different batches in controls, CD or UC. We also removed non-coding  
534 variants that were present in the 1000 Genomes pilot stage but were not in the subsequent  
535 Phase I integrated variant set (March 2012 release) and had not been in releases 2 or 3 of  
536 HapMap as these mostly represent false positives from the 1000 Genomes pilot, which  
537 often genotype poorly. Where a variant failed in exactly one batch we set all genotypes to  
538 missing for that batch (to be reimputed later) and included the site if it passed in the  
539 remainder of the batches. We removed individuals that had >2% missing data, had  
540 significantly higher or lower (defined as FDR<0.01) inbreeding coefficient (*F*), or were  
541 duplicated or related (PI\_HAT ≥ 0.4, calculated from the LD pruned dataset described  
542 below), by sequentially removing the individual with the largest number of related  
543 samples until no related samples remain. We projected all remaining samples onto  
544 principal component axes generated from HapMap 3, and classified their ancestry using a  
545 Gaussian mixture model fitted to the European (CEU+TSI), African  
546 (YRI+LWK+ASW+MKK), East Asian (CHB+JPT) and South Asian (GIH) HapMap  
547 samples. We removed all samples that were classified as non-European, or that lay more

548 than 8 standard deviations from the European cluster. After QC, there were 67,852  
549 European-derived samples with valid diagnosis (healthy control, CD or UC), and 161,681  
550 genotyped variants available for downstream analyses.

### 551 **Linkage-disequilibrium pruning and principal components analysis**

552 From the clean dataset we removed variants in long range LD<sup>41</sup> or with MAF < 0.05, and  
553 then pruned 3 times using the ‘--indep’ option in PLINK (with window size of 50, step  
554 size of 5 and VIF threshold of 1.25). Principal component axes were generated within  
555 controls using this LD pruned dataset (18,123 variants). The axes were then projected to  
556 cases to generate the principal components for all samples. The analysis was performed  
557 using our in-house C code (<https://github.com/hailianghuang/efficientPCA>) and  
558 LAPACK package<sup>42</sup> for efficiency.

### 559 **Controlling for population structure, batch effects and other confounders**

560 We used 2,853 “background SNPs” present on the ImmunoChip but not known to be  
561 associated with immune disorders to calculate the genomic inflation factor  $\lambda_{GC}$ . After  
562 including the first five principal components calculated above as covariates,  $\lambda_{GC} = 1.29$ ,  
563 1.25 and 1.31 for CD, UC and IBD (adding additional principal components did not  
564 further reduce  $\lambda_{GC}$ , Extended Data Table 3a). Because our genotype data were processed  
565 in 15 batches with variable ratios of cases to controls, we conducted two analyses to  
566 ensure possible batch effects were adequately controlled. First, we split the samples into a  
567 “balanced” cohort with studies that have both cases *and* controls and an “imbalanced”  
568 cohort with studies that have exclusively cases *or* controls (Extended Data Table 1). As  
569  $\lambda_{GC}$  under polygenic inheritance scales with the sample size<sup>43</sup>, we randomly down-  
570 sampled the full dataset to match the sample size of the balanced and the imbalanced

571 cohorts respectively. We tested for association in these subsets of our data (and included  
572 batch ID as a covariate in the balanced cohort), and found the  $\lambda_{GC}$  from the balanced and  
573 imbalanced cohorts to be within the 95% confidence interval of size matched values from  
574 our full data, suggesting that batch effects are not systematically inflating our association  
575 statistics (Extended Data Table 3b). We also performed a heterogeneity test for the odds  
576 ratio (OR) of lead variants in each credible set using the balanced and imbalanced  
577 cohorts, and observed no significant heterogeneity after Bonferroni correction  
578 (Supplementary Table 3).

579 We next sought to disentangle the contributions of polygenic inheritance and  
580 uncorrected population structure in our observed  $\lambda_{GC}$ . LD score regression<sup>44</sup> is able to  
581 differentiate these two effects, but requires genome-wide data, so is not possible in our  
582 ImmunoChip dataset. Instead, we compared  $\lambda_{GC}$  and  $\lambda_{1000}$  values calculated using the  
583 same set of background SNPs from the largest IBD meta-analysis with genome-wide  
584 data<sup>45</sup>. For both CD and UC the  $\lambda_{1000}$  values in our ImmunoChip study (1.012 and 1.012)  
585 were equal or less than those in the genome-wide study (1.016 and 1.012). Furthermore,  
586 LD score regression on the genome-wide data shows that the majority of inflation is  
587 caused by polygenic risk (LD score intercept = 1.09 for both CD and UC, compared to  
588  $\lambda_{GC} = 1.23$  and 1.29). Together, these results show that our residual inflation is consistent  
589 with polygenic signal and modest residual confounding. We tested what effect correcting  
590 for the LD score intercept of 1.09 would have on posterior probabilities and credible sets  
591 and found no major differences compared to uncorrected values. The full comparison of  $\lambda$   
592 values is shown in Extended Data Table 3c.

### 593 **Imputation**



594 Imputation was performed separately in each ImmunoChip autosomal high-density region  
595 (185 total) from the 1000 Genomes Phase I integrated haplotype reference panel. To  
596 prevent the edge effect, we extended each side of the high density regions by 50kbp.  
597 Two imputations were performed sequentially (Extended Data Figure 2) using software  
598 and parameters as described below. The first imputation was performed immediately  
599 after the quality control, from which the major results were manually inspected (Manual  
600 cluster plot inspection, Methods). The second imputation was performed after removing  
601 variants that failed the manual cluster plot inspection. We used SHAPEIT<sup>46,47</sup> (versions:  
602 first imputation: v2.r644, second imputation: v2.r769) to pre-phase the genotypes,  
603 followed by IMPUTE2<sup>13,14</sup> (versions: first: 2.2.2, second: 2.3.0) to perform the  
604 imputation. The reference panels were downloaded from the IMPUTE2 website (first:  
605 Mar 2012 release, second: Dec 2013 release). After the second imputation, there were  
606 388,432 variants with good imputation quality ( $INFO > 0.4$ ). These include 99.9% of  
607 variants with  $MAF \geq 0.05$ , 99.3% of variants with  $0.05 > MAF \geq 0.01$ , and 63.0% of  
608 variants with  $MAF < 0.01$  (Extended Data Figure 6d-f), with similar success rates for  
609 both coding and non-coding variants, making it unlikely that missing variants  
610 substantially affect our fine-mapping conclusions.

### 611 **Manual cluster plot inspection**

612 Variants that had posterior probability greater than 50% or in credible sets mapped to  $\leq$   
613 10 variants were manually inspected using Evoker v2.2<sup>48</sup>. Each variant was inspected by  
614 three independent reviewers (ten reviewers participated) and scored as pass, fail or  
615 maybe. Reviewers were blinded to the posterior probability of these variants. We  
616 removed variants that received one or more fails, or received less than 2 passes. 220 out

617 of 276 inspected variants passed this inspection, and 53 of 56 failed variants were  
618 restored by imputation. There is no difference in MAF between the failed and the passed  
619 variants ( $P=0.66$ ). A further cluster plot inspection flagged two additional failed variants  
620 after removing the failed variants from the first inspection and redoing the imputation and  
621 analysis. Dramatic clustering errors accounted for 27/58 flagged variants, which were  
622 eliminated from final credible sets. The remaining 31 had only minor issues, and the  
623 imputed data for these remained in our final credible sets, with marginally smaller  
624 posteriors (mean of the difference: 9.8%,  $P=0.06$ , paired t test).

### 625 **Establishing a $P$ value threshold**

626 We used a multiple testing corrected  $P$  value threshold for associations of  $1.35 \times 10^{-6}$ ,  
627 which was established by permutation. We generated 200 permuted datasets by randomly  
628 shuffling phenotypes across samples and carried out association analyses for each  
629 permutation across all variants in high-density regions that overlap IBD-associated loci<sup>3</sup>.  
630 We stored (i) all the point-wise  $P$  values ( $\alpha_S$ ), as well as (ii) the “best”  $P$  values ( $\alpha_B$ ) of  
631 each of the 200 permuted datasets. We then computed the empirical, experiment-wide  $P$   
632 value ( $\alpha_M$ )(corrected for multiple testing) for each of the tests as its rank/200 with respect  
633 to the 200  $\alpha_B$ . We then estimated the number of independent tests performed in the  
634 studied regions,  $n$ , as the slope of the regression of  $\log(1-\alpha_M)$  on  $\log(1-\alpha_S)$ , knowing that  
635  $\alpha_M = 1 - (1 - \alpha_S)^n$ , yielding a value of 37,056. The  $P$  value threshold was determined  
636 as  $0.05/n \approx 1.35 \times 10^{-6}$ .

### 637 **Detecting and fine-mapping association signals**

638 We used three fine-mapping methods (Supplementary Methods) to detect independent  
639 signals and create credible sets across 97 ImmunoChip autosomal high-density regions

640 that contained at least one variant with  $p < 1.35 \times 10^{-6}$ . Our process for merging the  
641 results of the three methods is described below and illustrated in Figure 1a.

- 642 1. We merged signals from different methods if their credible sets overlapped. To  
643 ensure a conservative credible set, this new merged credible set included all variants  
644 from all merged signals (the union of constituent credible sets). We assigned each  
645 variant in the merged credible sets a posterior probability equal to the average of the  
646 probabilities from the methods that reported this signal. To filter out technical  
647 artifacts we required genotyped variants in small credible sets to pass manual cluster  
648 plot inspection (see above) and all imputed variants to have  $INFO > 0.4$ . For signals  
649 reported by only one or two methods that contain only imputed variants (i.e. no  
650 directly genotyped variants), we additionally required at least one variant with  
651  $INFO > 0.8$  and  $MAF > 0.01$ .
- 652 2. We next assigned each signal to a provisional combination of lead variant and  
653 phenotype (CD, UC or IBD) that maximized the marginal likelihood of equation 8 in  
654 Supplementary Methods.
- 655 3. At loci with  $>1$  signals, we built a multivariate model with all signals reported by all  
656 three methods, and tested all possible combinations of adding signals reported by one  
657 or two methods, as long as they still had  $p < 1.35 \times 10^{-6}$  when jointly fitted in the  
658 multi-signal model. We selected the combination with the highest joint marginal  
659 likelihood (equation 8 in Supplementary Methods).

#### 660 **Phenotype assignment of signals**

661 The provisional phenotype assignment carried out during the signal merging described  
662 above is merely a point estimate, and does not capture the uncertainty associated with the

663 phenotypic assignment. We therefore recomputed the assignment of each signal as CD-  
664 specific, UC-specific or shared using the Bayesian multinomial model from fine-mapping  
665 method 2, Empirical covariance prior with Laplace approximation<sup>49</sup>, as it is designed to  
666 assess evidence of sharing in the presence of potentially correlated effect sizes. For the  
667 lead variant for each credible set, we calculated the marginal likelihoods as in equation  
668 13 from Supplementary Methods, restricting either  $\beta_{UC} = 0$  (for the CD-only model) or  
669  $\beta_{CD} = 0$  (for the UC-only model), as well as using the unconstrained prior (for the  
670 associated-to-both model). We then calculated the log Bayes factor in favor of sharing,  
671 i.e. the log of the ratio of marginal likelihoods between the associated-to-both model and  
672 the best of the single-phenotype associated models. These sharing log Bayes factors are  
673 given in Supplementary Table 1 (column ‘sharingBF’), and are a probabilistic assessment  
674 of phenotype assignment: for instance, the log Bayes factor of 97.4 for the primary signal  
675 at *IL23R* suggests a very high certainty that this signal is shared across both CD and UC,  
676 whereas the log Bayes factor of 0.4 for the primary signal at *FUT2* is more ambiguous. In  
677 addition to providing the log Bayes factor itself, we also applied a log Bayes factor cut-  
678 off of 10 to select variants with strong evidence of being shared across phenotypes.

### 679 **Final filters**

680 These procedures generated some signals where all three methods largely agreed, and  
681 some where they differed. While the signals where the methods disagree are of interest  
682 for methods development, here we chose to focus on the most concordant signals, as they  
683 are most straightforward to interpret biologically. We therefore discarded all signals  
684 found by only one method (which completely removed one locus), and two loci where  
685 the ratio of marginal likelihoods (equation 8 in Supplementary Methods) for the best

686 model and the second-best model was  $< 10$  (Supplementary Notes). After these filters  
 687 (Extended Data Figure 7) we considered 139 signals from 94 regions (containing a total  
 688 of 181,232 variants) to be confidently fine-mapped, and took them forward for  
 689 subsequent analysis.

690 **Estimating the variance explained by the fine-mapping**

691 We used a mixed model framework to estimate the total risk variance attributable to the  
 692 IBD risk loci, and to the signals identified in the fine-mapping. We used the GCTA  
 693 software package<sup>50</sup> to compute a gametic relationship matrix (G-matrix) using genotype  
 694 dosage information for the genotyped variants in the high-density regions (which we will  
 695 call  $\mathbf{G}_{HD}$ ). We then fit a variety of variance component models by restricted maximum  
 696 likelihood analysis using an underlying liability threshold model implemented with the  
 697 DMU package<sup>51</sup>. The first model is a standard heritability mixed-model that includes  
 698 fixed effects for five principal components (to correct for stratification) and a random  
 699 effect summarizing the contribution of all variants in the fine-mapping regions, such that  
 700 the liabilities across all individuals are distributed according to

701 
$$\mathbf{L} \sim \mathbf{N}(\beta_1 PC_1 + \dots + \beta_5 PC_5, \lambda_1 \mathbf{G}_{HD} + (1 - \lambda_1)I),$$

702 where  $\lambda_1$  is thus the variance explained by all variants in fine-mapping regions, which  
 703 we estimate. We then fitted a model that included an additional random effect for the  
 704 contribution of the lead variants that have been specifically identified (with G-matrix  
 705  $\mathbf{G}_{Signals}$ ), such that the liability is distributed as

706 
$$\mathbf{L} \sim \mathbf{N}(\beta_1 PC_1 + \dots + \beta_5 PC_5, \lambda'_1 \mathbf{G}_{HD} + \lambda_2 \mathbf{G}_{Signals} + (1 - \lambda'_1 - \lambda_2)I).$$

707 The variance explained by the signals under consideration is then given by the reduction  
 708 in the variance explained by all variants in the fine-mapping regions between the two

709 models ( $\lambda_1 - \lambda'_1$ ). We used this approach to estimate what fraction of this variance was  
710 accounted for by (i) the single strongest signals in each region (as would be typically  
711 done prior to fine-mapping), or (ii) all signals identified in fine-mapping. We used Cox  
712 and Snell's method<sup>52</sup> to estimate the variance explained across individual signals  
713 (Extended Data Figure 3b) for computational efficiency.

#### 714 **Overlap between transcription factor binding motifs and causal variants**

715 For each motif in the ENCODE TF ChIP-seq data ([http://compbio.mit.edu/encode-](http://compbio.mit.edu/encode-motifs/)  
716 [motifs/](http://compbio.mit.edu/encode-motifs/), accessed Nov 2014)<sup>20</sup>, we calculated the overall information content (IC) as the  
717 sum of IC for each position<sup>53</sup>, and only considered motifs with overall IC  $\geq 14$  bits  
718 (equivalent to 7 perfectly conserved positions). For every variant in a high-density region  
719 we determined whether it creates or disrupts a motif at a high-information site (IC  $\geq 1.8$ ).

#### 720 **Overlap between epigenetic signatures and causal variants**

721 For each combination of 120 tissues and three histone marks (H3K4me1, H3K4me3 and  
722 H3K27ac) from the Roadmap Epigenome Project we calculated an overlap score, equal  
723 to the sum of fine-mapping posterior probabilities for all variants in peaks of that histone  
724 mark in that tissue. We generated a null distribution of this score for each tissue/mark by  
725 shifting chromatin marks randomly (between 0bp and 44.53Mbp, the length of all high-  
726 density regions) and circularly (peaks at the end of the region shifted to the beginning of  
727 the region) over the high-density regions while keeping the same inter-peak distances.  
728 To summarize these correlated results across many cell and tissue types we defined a set  
729 of "core" H3K4me1 immune and H3K27ac gut peaks as sets of overlapping peaks in  
730 cells that showed the strongest enrichment. Intersects were made using bedtools v2.24.0  
731 default settings<sup>54</sup>. We selected 6 immune cell types for H3K4me1 and 3 gut cell types for

732 H3K27ac (Supplementary Table 2). We also chose controls (Supplementary Table 2)  
733 from non-immune and non-gut cell types with similar density of peaks in the fine-  
734 mapped regions as compared to immune/gut cell types to confirm the tissue-specificity of  
735 the overlap. We used the phenotype assignments (described above) in dissecting the  
736 enrichment for the CD and UC signals. Sixty-five CD and 21 UC signals that were  
737 mapped to  $\leq 50$  variants were used in this analysis.

### 738 **Published eQTL summary statistics**

739 We used eQTL summary statistics from three published studies:

- 740 • Peripheral blood eQTLs from the GODOT study<sup>29</sup> of 2,752 twins, reporting loci with  
741 MAF>0.5%. Imputation was performed using the 1000 genomes reference panel<sup>11</sup>.
- 742 • Peripheral blood eQTLs from the Westra *et al.* study<sup>30</sup> of 8,086 individuals, including  
743 variants with MAF>5%. Imputation was performed using the HapMap 2 CEU  
744 population reference panel<sup>55</sup>.
- 745 • CD14+ monocyte eQTLs from Table S2 in Fairfax *et al.*<sup>31</sup>, comprised of 432  
746 European individuals, measured in a naïve state and after stimulation with interferon-  
747  $\gamma$  (for 2 or 24 hours) or lipopolysaccharide, reporting loci with MAF>4% and  
748 FDR<0.05. Imputation was performed using the 1000 genomes reference panel<sup>10</sup>.

### 749 **Processing and quality control of new eQTL ULg dataset**

750 A detailed description of the ULg dataset is in preparation (Momozawa *et al.*, in  
751 preparation). Briefly, we collected venous blood and intestinal biopsies at three locations  
752 (ileum, transverse colon and rectum) from 350 healthy individuals of European descent,  
753 average age 54 (range 17-87), 56% female. SNPs were genotyped on Illumina Human  
754 OmniExpress v1.0 arrays interrogating 730,525 variants, and SNPs and individuals were

755 subject to standard QC procedures using call rate, Hardy-Weinberg equilibrium, MAF  $\geq$   
756 0.05, and consistency between declared and genotype-based sex as criteria. We further  
757 imputed genotypes at  $\sim 7$  million variants on the entire cohort using the Impute2 software  
758 package<sup>13</sup> and the 1,000 Genomes Project as reference population (Phase 3 integrated  
759 variant set, released 12 Oct 2014)<sup>11,14</sup>. From the blood, we purified CD4+, CD8+,  
760 CD19+, CD14+ and CD15+ cells by positive selection, and platelets (CD45-negative) by  
761 negative selection. RNA from all leucocyte samples and intestinal biopsies was  
762 hybridized on Illumina Human HT-12 arrays v4. After standard QC, raw fluorescent  
763 intensities were variance stabilized<sup>56</sup> and quantile normalized<sup>57</sup> using the lumi R  
764 package<sup>58</sup>, and were corrected for sex, age, smoking status, number of probes with  
765 expression level significantly above background as fixed effects and array number  
766 (sentrix id) as random effect. For each probe with measureable expression (detection  $P$   
767 value  $< 0.05$  in  $>25\%$  of samples) we tested for *cis*-eQTLs at all variants within a 500  
768 kbp window. The nominal  $P$  value of the best SNP within a *cis*-window was Sidak-  
769 corrected for the window-specific number of independent tests. The number of  
770 independent test in each window was estimated exactly in the same manner as for the  
771 number of independent test for fine-mapping methods (Establishing a  $P$  value threshold,  
772 Methods). We estimated false discovery rates (q-values) from the resulting  $P$  values  
773 across all probes using the qvalue R package<sup>59</sup>. 480 *cis*-eQTL with FDR  $\leq 0.10$  for  
774 which the lead SNPs (i.e. the SNP yielding the best  $P$  value for the *cis*-eQTL) mapped  
775 within the 97 high-density regions (94 fine-mapped plus 3 unresolved) were retained for  
776 further analyses.

777 **Naïve colocalization using lead SNPs**



778 We calculated the number of IBD credible sets that contain a lead eQTL variant in a  
779 particular tissue (“observed”). This number is then compared to the background number  
780 of overlaps (“expected”):

$$781 \quad \sum_{i \in S} (1 - (1 - N_i^{-1})^{C_i})$$

782 where  $N_i$  is the total number of variants in region  $i$  in 1000 genomes with an allele  
783 frequency greater than a certain threshold (equal to the threshold used for the original  
784 eQTL study),  $C_i$  is the number of these variants that lie in IBD credible sets, and  $S$  is a set  
785 of regions that have at least one significant eQTL. We simulated 1,000 trials per region  
786 with binomial probability equal to the regional background overlap rate:  $1 -$   
787  $(1 - N_i^{-1})^{C_i}$ . Empirical  $P$  values were estimated by comparing the observed number of  
788 overlaps with the simulated number of the overlaps. More specifically,  $P$  value is defined  
789 as the proportion of trails that have equal or more overlaps in the simulations than the  
790 observed.

### 791 **Frequentist colocalization using conditional $P$ values**

792 We next used conditional association to test for evidence of colocalization, as described  
793 in Nica *et al.*<sup>25</sup>. This method compares the  $P$  value of association for the lead SNP of an  
794 eQTL before and after conditioning on the SNP with the highest posterior in the credible  
795 set, and measures the drop in  $-\log(P)$ . An empirical  $P$  value for this drop is then  
796 calculated by comparing it to the drop for all variants in the high-density region. Because  
797 this method requires full genotypes we could only apply it to the ULg dataset (MAF >  
798 5%). An empirical  $P$  value  $\leq 0.05$  was considered as evidence that the corresponding  
799 credible set is colocalized with the corresponding *cis*-eQTL. To evaluate whether our  
800 fine-mapping associations colocalized with *cis*-eQTL more often than expected by

801 chance we counted the number of credible sets affecting at least one *cis*-eQTL with  
 802  $P \leq 0.05$ , and compared how often this number was matched or exceeded by 1,000 sets of  
 803 variants that were randomly selected yet distributed amongst the loci in accordance with  
 804 the real credible sets. The number of variants per set is same as the number of credible  
 805 sets in this eQTL analysis (MAF matched, size  $\leq 50$ ), shown in Extended Data Table 2.

806 **Bayesian colocalization using Bayes factors**

807 Finally, we used the Bayesian colocalization methodology described by Giambartolomei  
 808 *et al.*<sup>60</sup>, modified to use the credible sets and posteriors generated by our fine-mapping  
 809 methods (similarly only applicable to the ULg full genotype data). The method takes as  
 810 input a pair of IBD and eQTL signals, with corresponding credible sets  $S^{IBD}$  and  $S^{eQTL}$ ,  
 811 and posteriors for each variant  $p_i^{IBD}$  and  $p_i^{eQTL}$  (with  $p_i^X = 0 \forall i \notin S^X$ ). Credible sets  
 812 and posteriors were generated for eQTL signals using the Bayesian quantitative  
 813 association mode in SNPTest (with default parameters), with credible sets in regions with  
 814 multiple independent signals generated conditional on all other signals. Our method  
 815 calculates a Bayes factor (BF) summarizing the evidence in favor of a colocalized model  
 816 (i.e. a single underlying causal variant between the IBD and eQTL signals) compared to a  
 817 non-colocalized model (where different causal variants are driving the two signals), given  
 818 by the ratio of marginal likelihoods

819 
$$BF = \frac{L(\text{Colocalized})}{L(\text{Not colocalized})}$$

820 The marginal likelihood for the colocalized model (i.e. hypothesis  $H_4$  in Giambartolomei  
 821 *et al.*) is given by

822 
$$L(\text{Colocalized}) \propto \frac{1}{N} \sum_{i \in S^{IBD} \cup S^{eQTL}} p_i^{IBD} p_i^{eQTL}$$

823 and the marginal likelihood for the model where the signals are not colocalized (i.e.,  
 824 hypothesis  $H_3$ ) is given by:

$$825 \quad L(\text{Not colocalized}) \propto \frac{1}{N^2 - N} \sum_{i,j \in S^{IBD} \cap S^{eQTL}, i \neq j} p_i^{IBD} p_j^{eQTL}$$

826 In both cases,  $N$  is the total number of variants in the region. We only count towards  $N$   
 827 variants that have  $r^2 > 0.2$  with either the lead eQTL variant or the lead IBD variant.

828 To measure enrichment in colocalization BFs compared to the null, we carried out  
 829 a permutation analysis. In this analysis, we randomly reassigned eQTL signals to new  
 830 fine-mapping regions to generate a set of simulated null datasets. This is carried out using  
 831 the following scheme on variants and credible sets with the same MAF cut-off as the  
 832 eQTL dataset (ULg, MAF > 5%):

- 833 1. Estimate the standardized effect size  $\beta_g$  for each eQTL signal  $g$ , equal to standard  
 834 deviation increase in gene expression for each dose of the minor allele.
- 835 2. Randomly reassign each eQTL signal to a new fine-mapping region, and then select a  
 836 new causal variant with a MAF within 1 percentage point of the lead variant from the  
 837 real signal. If multiple such variants exist, select one at random. If no such variants  
 838 exist, pick the variant with the closest MAF.
- 839 3. Generate new simulated gene expression signals for each individual from  
 840  $\text{Normal}(\beta_g x_j, 1 - \beta_g^2 2f(1 - f))$  where  $x_j$  is the individual's minor allele dosage at  
 841 the new causal variant and  $f$  is the minor allele frequency.
- 842 4. Carry out fine-mapping and calculate colocalization BFs for each pair of (real) IBD  
 843 signal and (simulated) eQTL signal.
- 844 5. Repeat stages 2-4 1000 times for each tissue type

845 We can use these permuted BFs to calculate  $P$  values for each IBD credible set, given by  
846 the proportion of time the permuted BFs were as large or greater than the one observed in  
847 the real dataset. To generate a high-quality set of colocalized eQTL and IBD signals, we  
848 take all IBD signals that have the colocalization  $BF > 2$ ,  $P < 0.01$  and  $r^2$  (with the eQTL  
849 variant)  $> 0.8$ .

#### 850 **Code availability**

851 Computer code used in this study is provided in the ‘Software availability’ sections in  
852 Supplementary Methods.

#### 853 **Data availability**

854 The data that support the findings of this study are available from the international IBD  
855 Genetics Consortium but restrictions apply to the availability of these data, which were  
856 used under license for the current study, and so are not publicly available. Data are  
857 however available from the authors upon reasonable request and with permission of the  
858 international IBD Genetics Consortium.

859

860

## References

- 861 40. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-  
862 frequency and common variants. **28**, 1598–1603 (2012).
- 863 41. Price, A. L. *et al.* Long-Range LD Can Confound Genome Scans in Admixed  
864 Populations. *Am J Hum Genet* **83**, 132–135 (2008).
- 865 42. Anderson, E. *et al.* *LAPACK Users' Guide*. (Society for Industrial and Applied  
866 Mathematics, 1999).
- 867 43. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European*  
868 *Journal of Human Genetics* **19**, 807–812 (2011).
- 869 44. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from  
870 polygenicity in genome-wide association studies. *Nat Genet* (2015).  
871 doi:10.1038/ng.3211
- 872 45. de Lange, K. M. *et al.* Genome-wide association study implicates immune  
873 activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*  
874 (2017). doi:10.1038/ng.3760
- 875 46. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for  
876 thousands of genomes. *Nature Methods* **9**, 179–181 (2011).
- 877 47. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing  
878 for disease and population genetic studies. *Nature Methods* **10**, 5–6 (2012).
- 879 48. Morris, J. A., Randall, J. C., Maller, J. B. & Barrett, J. C. Evoker: a visualization  
880 tool for genotype intensity data. **26**, 1786–1787 (2010).
- 881 49. Jostins, L. & McVean, G. Trinculo: Bayesian and frequentist multinomial logistic  
882 regression for genome-wide association studies of multi-category phenotypes.  
883 *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw075
- 884 50. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for  
885 Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*  
886 **88**, 76–82 (2011).
- 887 51. Madsen, P., Su, G., Labouriau, R. & Christensen, F. DMU-a package for analyzing  
888 multivariate mixed models. *Proceedings of the Ninth World Congress on Genetics*  
889 *Applied to Livestock Production* (2010).
- 890 52. Cox, D. R. & Snell, E. J. *Analysis of Binary Data, Second Edition*. (CRC Press,  
891 1989).
- 892 53. D'haeseleer, P. What are DNA sequence motifs? **24**, 423–425 (2006).
- 893 54. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing  
894 genomic features. **26**, 841–842 (2010).
- 895 55. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796  
896 (2003).
- 897 56. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing  
898 transformation for Illumina microarray data. *Nucleic Acids Res* **36**, e11–e11  
899 (2008).
- 900 57. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of  
901 normalization methods for high density oligonucleotide array data based on  
902 variance and bias. **19**, 185–193 (2003).
- 903 58. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina  
904 microarray. **24**, 1547–1548 (2008).
- 905 59. Dabney, A., Storey, J. D. & Warnes, G. Q-value estimation for false discovery rate

906 control. *R package version 1*, (2006).  
907 60. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic  
908 association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

**Extended Data Figure 1. Power of the fine-mapping analysis.** Power (y axis) to identify the causal variant in a correlated pair (strength of correlation shown by color) increases with the significance of the association (x axis), and therefore with sample size and effect size. The vertical dashed line shows the genome-wide significance level. To estimate the relationship between the strength of association and our ability to fine-map it, we assumed that the association has only two causal variant candidates, and we defined the signal as successfully fine-mapped if the ratio of Bayes factors between the true causal variant and the non-causal variant is greater than 10 (a 91% posterior, assuming equal priors for the two candidate variants). Using equation (8) in Supplementary Methods, we have

$$\log\text{BF} = \log \frac{\Pr(\mathbf{Y} \mid \text{SNP1})}{\Pr(\mathbf{Y} \mid \text{SNP2})} \approx \log \frac{\Pr(\mathbf{Y} \mid \text{SNP1}, \theta_1^*)}{\Pr(\mathbf{Y} \mid \text{SNP2}, \theta_2^*)}$$

in which  $\theta^*$  is maximum likelihood estimate of the parameter values. The log-likelihood ratio follows a chi-square distribution:

$$\log\text{BF} \sim -\frac{1}{2}(\chi_{\text{SNP1}}^2 - \chi_{\text{SNP2}}^2) = -\frac{1}{2}\lambda(1 - r^2)$$

in which  $\lambda$  is the chi-square statistic of the lead variant and  $r$  is the correlation coefficient between the two variants. Because of the additive property of the chi-square distribution,  $\log\text{BF}$  follows a non-central chi-square distribution with 1 degree of freedom and non-centrality parameter  $\lambda(1 - r^2)/2$ . Therefore, the power can be calculated as the probability that  $\log\text{BF} > \log(10)$ , given by the cumulative distribution function of the non-central chi-squared distribution.

**Extended Data Figure 2. Procedures in the fine-mapping analysis.** Details for each stage are described in Methods. The dashed line means the imputation was performed only once after the manual inspection (not iteratively).

**Extended Data Figure 3. Variance explained.** Variance explained by secondary, tertiary, ... variants as a fraction of the primary signal at each locus.

**Extended Data Figure 4. Functional annotations. a,** Functional annotation for 45 variants having posterior probability  $> 50\%$ . **b,** Functional annotation for 116 association signals that are fine-mapped to  $\leq 50$  variants. Annotations are defined in Methods. We additionally grouped eQTLs into “Immune/Blood” (CD4+, CD8+, CD19+, CD14+ CD15+, platelets) and “Gut” (ileum, transverse colon and rectum). The eQTLs were generated from the ULg dataset using the “frequentist colocalization using conditional  $P$  values” approach (Methods).

**Extended Data Figure 5. Size of credible sets.** Comparison of credible set sizes for primary signals using each of our fine-mapping methods (methods 1, 2 and 3), the combined approach (as adopted in final results) and the approach described in Maller *et al.*<sup>6</sup> (y axis) and the  $R^2 > 0.6$  cut-off (x axis). Fine-mapping maps most signals to smaller numbers of variants.

**Extended Data Figure 6. Distributions of the allele frequency and the imputation quality.** Panels **a-c**: distribution of the risk allele frequency for 45 variants having  $> 50\%$  posterior probability plotted against **(a)** posterior probability, **(b)** significance of the association as  $-\log_{10}(P)$ , and **(c)** odds ratio of the association. Variants are color coded according to their

functions. Odds ratio for IBD associations was the larger of odds ratios for CD and UC. Panels **d-f**: distribution of imputation quality (INFO measure from the IMPUTE2 program) for variants having MAF  $\geq 5\%$  (**d**), between 5% and 1% (**e**) and  $< 1\%$  (**f**).

**Extended Data Figure 7. Merging and adjudicating signals across methods.** The number of signals for each method is shown in the brackets, and for each method a black bar indicates a signal with  $p < 1.35 \times 10^{-6}$ , and a grey bar a signal that does not reach that threshold. The colored bar shows the final status of each signal after merging and model selection (Methods). “Low info” corresponds to INFO  $< 0.8$  (the threshold used for signals reported by 1 or 2 methods) and “rare and imputed” to MAF  $< 0.01$  and no genotyped variants in the credible set, regardless of INFO (Methods).

**Extended Data Table 1. Study samples.** Genotyped samples in each batch for healthy controls (Control), Crohn’s disease (CD) and ulcerative colitis (UC). Batches were grouped into cohorts for further analysis (Controlling for population structure, batch effects and other confounders, Methods).

**Extended Data Table 2. Colocalization with eQTL.** The number of IBD credible sets that colocalize with eQTLs using the naïve, frequentist and Bayesian approaches. Significant observations are boldfaced. ‘Number of credible sets’ reports the number of credible sets that have MAF above the cut-off.

**Extended Data Table 3. Genomic inflation.** Genomic inflation factors and LD score regression intercept for Crohn’s disease (CD), ulcerative colitis (UC) and both (IBD). **a**, Genomic inflation factors using the first four, five and six principal components. The factors were calculated using 2,853 background variants from the ImmunoChip. **b**, Genomic inflation factors for subsets of the data (using five principal components for the same 2,853 background variants). Balanced, imbalanced and down-sampled cohorts are defined in Methods. Numbers in brackets indicate the 95% confidence interval for the inflation factors (only estimated for the down-sampled cohorts). **c**, LD score regression intercept and genomic inflation factors ( $\lambda_{GC}$  and  $\lambda_{1000}$ ) from the largest IBD meta-analyses with genome-wide data (CD:GWAS and UC:GWAS).