Association mapping of inflammatory bowel disease loci to single variant resolution

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

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

*These authors contributed equally to this work.

§These authors contributed equally to this work (corresponding authors). E-mails: hhuang@atgu.mgh.harvard.edu (H.H.); michel.georges@ulg.ac.be (M.G.); mjdaly@atgu.mgh.harvard.edu (M.J.D.) and jb26@sanger.ac.uk (J.C.B.)

†A full list of members and affiliations appears at the end of the paper.

Summary (150 words)

Inflammatory bowel disease (IBD) is a chronic gastrointestinal inflammatory disorder that affects millions worldwide. Genome-wide association studies (GWAS) have identified 200 IBD-associated loci, but few have been conclusively resolved to specific functional variants. Here we report fine-mapping of 94 IBD loci using high-density genotyping in 67,852 individuals. Of the 139 independent associations identified in these regions, 18 were pinpointed to a single causal variant with >95% certainty, and an additional 27 associations to a single variant with >50% certainty. These 45 variants are significantly enriched for protein-coding changes (n=13), direct disruption of transcription factor binding sites (n=3) and tissue specific epigenetic marks (n=10), with the latter category showing enrichment in specific immune cells among associations stronger in CD and gut mucosa among associations stronger in UC. The results of this study suggest that high-resolution, fine-mapping in large samples can convert many GWAS discoveries into statistically convincing causal variants, providing a powerful substrate for experimental elucidation of disease mechanisms.

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

The widespread correlation structure of the human genome (known as linkage disequilibrium, or LD) often results in similar evidence for association among many nearby variants. However, unless LD is perfect ($r^2 = 1$), it is possible, with sufficiently large sample size, to statistically resolve causal variants from neighbors even at high levels of correlation (Extended Data Figure 1 and van de Bunt *et al.*[5]). Novel statistical approaches applied to very large datasets have begun to address this problem[6] but also require that the highly correlated variants are directly genotyped or imputed with certainty. Truly high-resolution mapping data, when combined with increasingly sophisticated and comprehensive public databases annotating the putative protein-coding and regulatory function of DNA variants, are likely to reveal novel insights into disease pathogenesis[7-9] and the mechanistic involvement of disease-associated variants.

92

93 **Genetic architecture of IBD associated loci**

94 As part of a large collaborative effort led by the International IBD Genetics Consortium
95 (IIBDGC), 67,852 study subjects of European ancestry, including 33,595 IBD (18,967
96 CD and 14,628 UC) and 34,257 healthy controls were genotyped using the Illumina™
97 (San Diego, CA, USA) Immunochip. This custom genotyping array was designed to
98 include all known variants from European individuals in the February 2010 release of the
99 1000 Genomes Project[10,11] in 186 high-density regions known to be associated to one or
100 more of 12 immune-mediated diseases[12]. We evaluated ninety-seven of these regions
101 previously associated with IBD[3] and containing one or more associated variants ($p < 10^{-6}$)
102 in this data set. The major histocompatibility complex was excluded from these analyses
103 as fine-mapping has been reported elsewhere[13]. Because fine-mapping uses subtle
104 differences in strength of association between tightly correlated variants to infer which is
105 most likely to be causal, it is particularly sensitive to data quality. We therefore
106 performed stringent quality control (QC) to remove genotyping errors and batch effects,
107 including manual cluster plot inspection for 905 variants (Methods). After QC, we
108 imputed this dataset using the 1000 Genomes reference panel (December 2013,
109 downloaded from IMPUTE2[14,15] website) to fill in missing variants or genotype data
110 dropped in chip design or QC (Figure 1a).

111 We applied three complementary Bayesian fine-mapping methods that used
112 different priors and model selection strategies both to identify independent association
113 signals within a region (Supplementary Methods), and to assign a posterior probability of
114 causality to each variant (Figure 1a). For each independent association signal, we sorted

all variants by the posterior probability of association, and added variants to the 'credible set' of associated variants until the sum of their posterior probability exceeded 95% – that is, the credible set contains the minimum list of DNA variants that are >95% likely to contain the causal variant (Figure 1b). These sets ranged in size from one to > 400 variants. We merged these results (Methods) and subsequently focused (Figure 1a) only on signals where an overlapping credible set of variants was identified by at least two of the three methods and all variants were either directly genotyped or well imputed (Methods). Fluorescent signal intensity cluster plots were manually reviewed for all variants in credible sets with ten or fewer variants, and a second round of imputation and analysis was performed if any genotypes were removed based on this review.

In 3 out of 97 regions, a consistent credible set could not be identified; when multiple independent effects exist in a region with several highly correlated signals, multiple distinct fine-mapping solutions may not be distinguishable (Supplementary Notes). Sixty-eight of the remaining 94 regions contain a single credible set, while 26 harbored two or more independent association signals, for a total of 139 independent associations defined across the 94 regions (Figure 2a). Only *IL23R* and *NOD2* (both previously established to contain multiple associated protein-coding variants[16]), contain more than three independent signals. Consistent with previous reports[3], the vast majority of signals are associated with both CD and UC. However, many of these have significantly stronger association with one subtype than the other. For the purposes of enrichment analyses below, we compare 79 signals that are more strongly associated with CD to 23 signals that are more strongly associated with UC (the remaining 37 are equally associated with both subtypes) ("list of credible sets" sheet, Supplementary Table 1).

Using a restricted maximum likelihood mixed model approach[17], we evaluated the proportion of total variance in disease risk attributed to these 94 regions and how much of that is explained by the 139 specific associations. We estimated that 25% of CD risk was explained by the specific associations described here, out of a total of 28% explained by these loci (the corresponding numbers for UC are 17% out of 22%). This indicates that our credible sets capture most of the IBD genetic risk at these loci. The single strongest signals in each region contribute 76% of this variance explained and the remaining associations contribute 24% (Extended Data Figure 2b), highlighting the importance of secondary and tertiary associations in the articulation of GWAS results[13,18].

Associations mapped to a single variant

For 18 independent signals, the 95% credible set consisted of a single variant (hereafter referred to as 'single variant credible sets') and for 24 others, the credible set consisted of two to five variants (Figure 2b). The single variant credible sets included five previously reported coding variants: three in *NOD2* (fs1007insC, R702W, G908R), a rare protective allele in *IL23R* (V362I) and a splice variant in *CARD9* (c.IVS11+1G>C) [16,19]. The remaining single variant credible sets were comprised of three missense variants (I170V in *SMAD3*, I923V in *IFIH1* and N289S in *NOD2*), four intronic variants (in *IL2RA*, *LRRK2*, *NOD2* and *RTEL1*/*TNFRSF6B*) and six intergenic variants (located 3.7kb downstream of *GPR35*; 3.9kb upstream of *PRDM1*; within a EP300 binding site 39.9 kb upstream of *IKZF1*; 500 bp before the transcription start site of *JAK2*; 9.4kb upstream of *NKX2-3;* and 3.5kb downstream from *HNF4A*) (Table 1). A customizable browser (https://atgu.shinyapps.io/Finemapping) enabling review of the detailed fine-mapping

161 results in each region along with all annotations discussed below has been prepared. Of
162 note, while physical proximity does not guarantee functional relevance, the credible set of
163 variants for 29 associated loci now resides within 50 kb of only a single gene – improved
164 from only 3 so refined using an earlier HapMap-based definition. Using the same
165 definitions, the total number of potential candidate genes was reduced from 669 to 331.
166 Examples of IBD candidate genes clearly prioritized in our data are described in the
167 Supplementary Box.
168
169 **Sequence-level consequences of associated variants – protein coding variation**
170 We first annotated the possible functional consequences of the IBD variants by their
171 effect on the amino acid sequences of proteins. Thirteen out of 45 variants that have
172 >50% posterior probability are non-synonymous (Table 1 and Figure 2c), an 18-fold
173 enrichment (p-value=$2x10^{-13}$, Fisher's exact test) relative to randomly drawn variants in
174 our regions. By contrast, only one variant with >50% probability is synonymous
175 (p=0.42). All common coding variants previously reported to affect IBD risk are
176 included in a 95% credible set including: *IL23R* (R381Q, V362I and G149R); *CARD9*
177 (c.IVS11+1G>C and S12N); *NOD2* (S431L, R702W, V793M, N852S and G908R,
178 fs1007insC); *ATG16L1* (T300A); *PTPN22* (R620W); and *FUT2* (W154X). While this
179 enrichment of coding variation (Figure 3a) provides assurance about the accuracy of our
180 approach, it does not suggest that 30% of all associations are caused by coding variants;
181 rather, it is almost certainly the case that associated coding variants have stronger effect
182 sizes, making them more amenable to fine mapping.
183

Sequence-level consequences of associated variants – non-coding variation

We next examined the best understood non-coding aspect of DNA sequence: conserved nucleotides in high confidence binding site motifs of 84 transcription factor (TF) families[20] (Methods). There was a significant positive correlation between TF motif disruption and IBD association posterior probability (p-value=0.006, binomial regression) (Figure 3a), including three variants with >50% probability (two >95%). In the *RTEL1/TNFRSF6B* region, rs6062496 region is predicted to disrupt a TF binding site (TFBS) for EBF1 and overlaps DNaseI hypersensitivity clusters. EBF1 is a TF involved in the maintenance of B cell identity and prevention of alternative fates in committed cells[21]. The second example, rs74465132, is a low frequency (3.6%) protective variant that creates a binding site for EP300 less than 40kbp upstream of *IKZF1* (zinc-finger DNA binding protein). The third notable example of TFBS disruption, although not in a single variant credible set, is detailed in the Supplementary Box for the association at *SMAD3*.

Recent studies have shown that trait associated variants are enriched for epigenetic marks highlighting cell type specific regulatory regions [22,23]. We compared our credible sets with ChIPseq peaks corresponding to chromatin immunoprecipitation with H3K4me1, H3K4me3 and H3K27ac in 120 adult and fetal tissues, assayed by the NIH Roadmap Epigenomics Mapping Consortium[24] (Figure 3b). Using a threshold of $p=1.3\times10^{-4}$ (0.05 corrected for 360 tests), we observed significant enrichment of H3K4me1 in 6 immune cell types and for H3K27ac in 3 gastrointestinal (GI) samples (sigmoid colon and colonic and rectal mucosa) (Figure 3b and Supplementary Table 2). Furthermore, the subset of signals that are more strongly associated with CD overlap

more with immune cell chromatin peaks, whereas UC signals overlap more with GI chromatin peaks (Supplementary Table 2).

These three chromatin marks are correlated both within tissues (we observe additional signal in other marks in the tissues described above) and across related tissues. We therefore defined a set of "core immune peaks" for H3K4me1 and "core GI peaks" for H3K27ac as the set of overlapping peaks in all enriched immune cell and GI tissue types, respectively. These two tracks (immune-K4me1 and gut-K27ac) are independently significant and capture the observed enrichment compared to "control peaks" made up of the same number of ChIPseq peaks across our 94 regions in non-immune and non-GI tissues (Figure 3c,d). These two tracks summarize our epigenetic-GWAS overlap signal, and the combined excess over the baseline suggests that a substantial number of regions, particularly those not mapped to coding variants, may ultimately be explained by functional variation in recognizable enhancer/promoter elements.

Overlap of IBD credible sets with expression QTLs

Variants that change enhancer or promoter activity might precipitate changes in gene expression, and baseline expression of many genes has been found to be regulated by genetic variation[25-27]. Indeed, these so-called expression quantitative trait loci (eQTLs) have been suggested to underlie a large proportion of GWAS associations [25,28]. We therefore searched for variants that are both in an IBD associated credible set with 50 or fewer variants and the most significantly associated eQTL variant for a gene in the GODOT study[29] of peripheral blood mononuclear cells (PBMC) from 2,752 twins. Sixty-eight of the 76 regions with signals fine-mapped to < 50 variants harbor at least one

significant eQTL (defined as influencing expression of a gene within 1 Mb of the region with a p-value < 10^{-5}). Despite this apparent abundance of eQTLs in fine-mapped regions, only 3 credible sets overlap eQTLs, compared with 3.7 expected by chance (Methods). Data from a more recent independent study (Westra *et al.*)[30] using PBMCs from 8,086 individuals did not yield a substantively different outcome, demonstrating a modest but non-significant enrichment (8 observed overlaps, 4.2 expected by chance, p=0.07). Using a more lenient definition of overlap which requires the lead eQTL variant to be in LD ($R^2 > 0.4$) with an IBD credible set variant increased the number of potential overlaps but again these numbers were not greater than chance expectation (GODOT: observed 14, expected 12.2; Westra *et al.*: observed 11, expected 9.1).

As PBMCs are a heterogeneous collection of immune cell populations, cell type-specific signals, or signals corresponding to genes expressed most prominently in non-immune tissues, may be missed. We therefore tested the enrichment of eQTLs that overlap credible sets in 5 primary T cell populations (CD4+, CD8+, CD19+, CD14+ and CD15+), platelets, and 3 distinct intestinal locations (rectum, colon and ileum) isolated from 350 healthy individuals (ULg dataset, Methods). We observed a significant enrichment of credible SNP/eQTL overlaps in CD4+ cells and ileum (Extended Table 1): 3 and 2 credible sets overlapped eQTLs, respectively, compared to 0.4 and 0.3 expected by chance (p-value=0.007 and 0.025). An enrichment was also observed for the naïve CD14+ cells from another study[31] (Knight dataset, Extended Data Table 1): eight overlaps observed compared to 2.7 expected by chance (p-value=0.005). We did not observe enrichment of overlaps in stimulated (with interferon or lipopolysaccharide) CD14+ cells from the same source (Extended Data Table 1).

253 To more deeply investigate eQTL overlaps we applied two colocalization
254 approaches (one based on permutations, one Bayesian, Methods) to eQTL datasets where
255 primary genotype and expression data were available (ULg dataset). We confirmed
256 greater than expected overlap with eQTLs in CD4+ and ileum described above (Figure 4
257 and Extended Data Table 1). The number of colocalizations in other purified cell
258 types/tissues was largely indistinguishable from what we expect under the null using
259 either method, except for moderate enrichment in rectum (4 observed and 1.4 expected,
260 p=0.039) and colon (3 observed and 0.8 expected, p=0.04). Of these robust
261 colocalizations, only two correspond to an IBD variant with causal probability > 50%
262 (Table 1 and Extended Data Figure 3a).
263
264 **Discussion**
265 We have performed fine-mapping of 94 previously reported genetic risk loci for IBD.
266 Rigorous quality control followed by a integration of three novel fine-mapping methods
267 was employed to generate a list of genetic variants accounting for 139 independent
268 associations across these loci. These associations account for more than 80% of the total
269 variance explained by these loci. Our results substantially improve on previous fine-
270 mapping efforts using a preset LD threshold (e.g. $r^2 > 0.6$[32]) (Figure 5) by formally
271 modeling the posterior probability of association of every variant. Much of this
272 resolution derives from the very large sample size we employed, because the number of
273 variants in a credible set significantly decreases with increasing test statistics (p-value =
274 0.0069, Extended Data Figure 4). For example, at 10% allele frequency, 31% of signals

are fine-mapped to ≤ 5 variants – this improves to 53% if the sample size were to double again.

Additionally, the high-density of genotyping also aids in improved resolution. For instance, the primary association at IL2RA has now been mapped to a single variant associated with CD, rs61839660. This variant was not present in the Hapmap 3 reference panel and was therefore not reported in earlier studies[3,33] (nearby tagging variants, rs12722489 and rs12722515, were reported instead). Imputation using the 1000 genomes reference panel and the largest assembled GWAS dataset[3] did not separate rs61839660 from its neighbors (unpublished results), due to the loss of information in imputation using the limited reference. Only direct genotyping, available in the immunochip high-density regions, permitted the conclusive identification of this as the causal variant.

Accurate fine-mapping should, in many instances, ultimately point to the same variant across diseases in shared loci. Among our single-variant credible sets, we fine-mapped a UC association to a rare missense variant (I923V) in *IFIH1*, which is also associated with type 1 diabetes (T1D)[34] with an opposite direction of effect (Supplementary Box). The intronic variant noted above (rs61839660, AF=9%) in *IL2RA* was also similarly associated with T1D, again with a discordant directional effect[35] (Supplementary Box). Simultaneous high-resolution fine-mapping in multiple diseases should therefore better clarify both shared and distinct biology.

High-resolution fine-mapping demonstrates that causal variants are significantly enriched for variants that alter protein coding variants or disrupt transcription factor binding motifs. Enrichment was also observed in H3K4me1 marks in immune related cell types and H3K27ac marks in sigmoid colon and rectal mucosal tissues – with CD

298 loci demonstrating a stronger immune signature and UC loci more enriched for gut
299 tissues. By contrast, overall enrichment of eQTLs is quite modest compared with prior
300 reports and not seen in excess of chance in our well-refined credible sets. This result
301 underscores not only the importance of the high-resolution mapping but also the careful
302 incorporation of the high background rate of eQTLs. It is worth noting that evaluating
303 the overlap between two distinct mapping results is fundamentally different than
304 comparing genetic mapping results to fixed genomic features, and depends on both
305 mappings being well-resolved. While these data strongly challenge the paradigm that
306 easily surveyed baseline eQTLs explain a large proportion of non-coding GWAS signals,
307 the modest excesses observed in smaller but cell-specific data sets suggest that much
308 larger tissue or cell-specific studies (and under the correct stimuli or developmental time
309 points) will resolve the contribution of eQTLs to GWAS hits.

310 Resolving multiple independent associations may often help target the causal gene
311 more precisely. For example, the *SMAD3* locus hosts a non-synonymous variant and a
312 variant disrupting the conserved transcription factor binding site (also overlapping the
313 H3K27ac marker in gut tissues), unambiguously articulating a role in disease and
314 providing an allelic series for further experimental inquiry. Similarly, the *TYK2* locus has
315 been mapped to a non-synonymous variant and a variant disrupting a conserved
316 transcription factor binding site (Extended Data Figure 5).

317 One-hundred and sixteen associations have been fine-mapped to ≤ 50 variants.
318 Among them, 27 associations contain coding variants, 20 contain variants disrupting
319 transcription factor binding motifs, and 45 are within histone H3K4me1 or H3K27ac
320 marked DNA regions. However, 40 non-coding associations were not mapped to any

known function (Extended Data Figure 3b) despite extensive efforts to integrate with all available annotation, epigenetic and eQTL data.

The best-resolved associations - 45 variants having >50% posterior probabilities for being causal (Table 1) – are similarly significantly enriched for variants with known or presumed function from genome annotation. Of these, 13 variants cause non-synonymous change in amino acids, 3 disrupt a conserved TF binding motif, 10 are within histone H3K4me1 or H3K27ac marked DNA regions in disease-relevant tissues, and 2 co-localize with a significant cis-eQTL (Extended Data Figure 3a).

This analysis leaves, however, 21 non-coding variants, all of which have extremely high probabilities to be causal (5 are in the >95% list), that are not located within known motifs, annotated elements, nor in any experimentally determined ChIPseq peaks or eQTL credible sets yet discovered. While we have identified a statistically compelling set of genuine associations (often intronic or within 10 kb of strong candidate genes), we can make little inference about function. For example, the single variant credible set only 500 bp from the transcription start site of JAK2 has no annotation, eQTL or ChIPseq peak of note. This underscores the incompleteness of our knowledge regarding the function of non-coding DNA and its role in disease. That the majority of the best refined non-coding associations have no available annotation is perhaps sobering with respect to how well we may be able to currently interpret non-coding variation in medical sequencing efforts. It does suggest, however, that detailed fine-mapping of GWAS signals down to single variants, combined with emerging high-throughput genome-editing methodology, may be among the most effective ways to advance to a greater understanding of the biology of the non-coding genome.

List of Figures

Figure 1. Procedures in the fine-mapping analysis. a, Flowchart of fine-mapping steps. Dashed line means the imputation has been performed only once after manual inspection (not iteratively). **b, An** example output from fine-mapping. This region has been mapped to two independent signals. For each signal, fine-mapping reports the phenotype it is associated with, the variants it is fine-mapped to and their posterior probabilities.

Figure 2. Summary of fine-mapped associations. a, sixty-eight loci hosting a single association and 26 loci hosting multiple independent associations. **b,** Number of variants in credible sets. 18 associations were fine-mapped to a single variant, and 116 to ≤ 50 variants. Only credible sets having ≤ 50 variants were advanced for set-enrichment analyses (epigenetics and eQTL). **c,** distribution of the posterior probability in credible sets having ≤ 50 variants. 45 variants have posterior probability > 50% and were advanced for variant-based enrichment analyses (coding, TFBS disruption and epigenetics).

Figure 3. Functional annotation of causal variants. a, Proportion of variants that are protein coding, disrupting/creating transcription factor binding motifs or synonymous. **b,** Epigenetic peaks overlapping credible variants in various cell lines. Sample categories were taken from the Roadmap Epigenomics Consortium[36]. Significant cell line-peak pairs have been marked with asterisks. **c,** Proportion of credible variants that overlap H4K4Me1 peaks. **d,** Proportion of credible variants that overlap H3K27ac peaks. In panels **a**, **c** and **d**, the vertical dotted lines mark the 50% probability and the horizontal dashed lines show the background proportions of each functional category.

366 **Figure 4. Number of credible sets that colocalize eQTLs.** The violin plot shows the
367 distribution of the number of colocalizations by chance (background) and the solid points
368 shows the observed number of colocalizations. P-values of the enrichment were shown
369 next to the solid points. Both the background and the observed numbers were calculated
370 using the permutation based approach (Methods).
371 **Figure 5. Fine-mapping improved the resolution of genetic associations.** We
372 compare the numbers of variants that are mapped in each independent signal using the
373 fine-mapping approach (y axis) and the $R^2 > 0.6$ cut-off (x axis). Fine-mapping maps
374 most signals to smaller numbers of variants.
375

List of Tables

Table 1: Summary of variants having posterior probability >50%. Variants were sorted by their posterior probabilities. AF: allele frequency. PROB: posterior probability for being a causal variant. FUNC: functional annotations including coding (C), Epegenetic peaks (E), disrupting transcription factor binding sites (T) and colocalization with eQTL (Q).

VARIANT	CHR	POSITION	TRAIT	AF	PROB	FUNC	ANNOTATION
Signals mapped to a single variant							
rs7307562	12	40724960	CD	0.398	0.999		LRRK2 (intronic)
rs2066844	16	50745926	CD	0.063	0.999	C	NOD2(R702W)
rs2066845	16	50756540	CD	0.022	0.999	C	NOD2(G908R)
rs6017342	20	43065028	UC	0.544	0.999	E	HNF4A (downstream), Gut_H3K27ac
rs61839660	10	6094697	CD	0.094	0.999	E	IL2RA (intronic), Immune_H3K4me1
rs5743293	16	50763781	CD	0.964	0.999	C	fs1007insC
rs6062496	20	62329099	IBD	0.587	0.996	T	RTEL1-TNFRSF6B (ncRNA_intronic), EBF1 TFBS
rs141992399	9	139259592	IBD	0.005	0.995	C	CARD9(1434+1G>C)
rs35667974	2	163124637	UC	0.021	0.994	C	IFIH1(I923V)
rs74465132	7	50304782	IBD	0.034	0.994	T,E	IKZF1 (upstream), EP300 TFBS, Immune_H3K4me1
rs4676408	2	241574401	UC	0.508	0.994		GPR35 (downstream)
rs5743271	16	50744688	CD	0.007	0.993	C	NOD2(N289S)
rs10748781	10	101283330	IBD	0.55	0.990	E	NKX2-3 (upstream), Gut_H3K27ac
rs35874463	15	67457698	IBD	0.054	0.989	C,E	SMAD3(I170V), Gut_H3K27ac
rs72796367	16	50762771	CD	0.023	0.983		NOD2 (intronic)
rs1887428	9	4984530	IBD	0.603	0.974		JAK2 (upstream)
rs41313262	1	67705900	CD	0.014	0.973	C	IL23R(V362I)
rs28701841	6	106530330	CD	0.116	0.971		PRDM1 (upstream)
Signals mapped to ≥ 2 variants but the lead variant have posterior probability > 50%							
rs76418789	1	67648596	CD	0.006	0.937	C	IL23R(G149R)
rs7711427	5	40414886	CD	0.633	0.919		
rs1736137	21	16806695	CD	0.407	0.879		

rs104895444	16	50746199	CD	0.003	0.865	C	NOD2(V793M)
rs56167332	5	158827769	IBD	0.353	0.845		IL12B
rs104895467	16	50750810	CD	0.002	0.833	C	NOD2(N852S)
rs630923	11	118754353	CD	0.153	0.820		
rs3812565	9	139272502	IBD	0.402	0.815	Q	eQTL of INPP5E in CD4 and CD8; CARD9 in CD14, SEC16A in CD15
rs4655215	1	20137714	UC	0.763	0.784	E	Gut_H3K27ac
rs145530718	19	10568883	CD	0.023	0.762		
rs6426833	1	20171860	UC	0.555	0.752		
chr20:43258079	20	43258079	CD	0.041	0.736		
rs17229679	2	199560757	UC	0.028	0.716		
rs4728142	7	128573967	UC	0.448	0.664	E	Immune_H3K4me1
rs2143178	22	39660829	IBD	0.157	0.662	T,E	NFKB TFBS, Gut_H3K27ac
rs34536443	19	10463118	CD	0.038	0.649	C	TYK2(P1104A)
rs138425259	16	50663477	UC	0.009	0.648		
rs146029108	9	139329966	CD	0.036	0.643		
rs12722504	10	6089777	CD	0.26	0.615		
rs60542850	19	10488360	IBD	0.17	0.591		
rs2188962	5	131770805	CD	0.44	0.590	E,Q	Gut_H3K27ac, eQTL of SLC22A5 in CD14, CD15 and IL
rs2019262	1	67679990	IBD	0.4	0.586		
rs3024493	1	206943968	IBD	0.171	0.537	E	Immune_H3K4me1
rs7915475	10	64381668	CD	0.304	0.528		
rs77981966	2	43777964	CD	0.077	0.521		
rs9889296	17	32570547	CD	0.264	0.512		
rs2476601	1	114377568	CD	0.908	0.508	C	PTPN22(W620R)

382

Supplemental Materials

Extended Data Figure 1, Power (y axis) to distinguish which variant in a correlated pair (strength of correlation shown by color) is causal increases with the significance of the association (x axis), and therefore with sample size and effect size. The vertical dashed line flags 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 possible causal variants, and we define 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). Using equation (8) in Supplementary Methods, we have

$$\log\mathrm{BF} = \log\frac{\Pr(\boldsymbol{Y}\,|\mathrm{SNP1})}{\Pr(\boldsymbol{Y}\,|\mathrm{SNP2})} \approx \log\frac{\Pr(\boldsymbol{Y}\,|\mathrm{SNP1},\theta_1^*)}{\Pr(\boldsymbol{Y}\,|\mathrm{SNP2},\theta_2^*)}$$

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

$$\log\mathrm{BF} \sim -\frac{1}{2}(\chi^2_{\mathrm{SNP1}} - \chi^2_{\mathrm{SNP2}}) = -\frac{1}{2}\lambda(1-r^2)$$

in which λ 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, logBF 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 calculated as the probability that logBF > log(10), given by the CDF of the non-central chi-squared distribution.

Extended Data Figure 2, a, Genomic distance that variants in 95% credible set span. **b,** Variance explained normalized to the primary association in each locus.

Extended Data Figure 3. a, Functional annotation for 45 variants having posterior probability > 50%. **b,** Functional annotation for 116 associations that are fine-mapped to ≤ 50 variants.

Extended Data Figure 4, **a,** Number of variants in credible set decreases with the significance of the signal. **b,** Number of variants in credible set increases with the minor allele frequency of the signal. The solid line shows the fitted trend in both panels, and the shaded region shows the variance of the trend.

Extended Data Figure 5, SMAD3 (**a**) and TYK2 (**b**) regions after fine-mapping. The implicated region has been reduced to a smaller number of genes (shown in black). Color ticks are variants mapped to their functions and black ticks are variants not mapped to a function. The width of the tick scales with the posterior probability.

Extended Data Figure 6, Tissue and cell line specific expression for genes *SBNO2*, *IL10*, *IL19*, *LRRK2*, *KSR1*, *PRDM1*, *SMAD3*, *SMAD7*, *IFIH1*, *IL2RA*, *RETL1* and *TNFRSF6B*. **Left panels.** Expression levels of selected genes were determined in a panel of human tissues (bone marrow, heart, skeletal muscle (Sk. Muscle), uterus, liver, fetal liver (F. Liver), spleen, thymus, thyroid, prostate, brain, lung, small intestine (Sm. Intestine) and colon) and human cell lines using a custom made Agilent expression array.

The cell lines represent models of human T lymphocytes (Jurkat), monocytes (THP-1), erythroleukemia cells (K562), promyelocytic cells (HL-60), colonic epithelial cells (HCT-15, HT-29, Caco-2), and cells from embryonic kidney (HEK-293). In addition, models of differentiated colonic epithelium (Caco-2 differentiated for 21 days in culture (Caco-2 diff.)), activated T lymphocytes (Jurkat cells stimulated with PMA (40ng/ml) and ionomycin (1ug/ml) for 6 hrs (Jurkat stim.)), and macrophages (derived from THP-1 differentiated for 24 hrs (THP-1 diff.) with IFN-γ (400U/ml) and TNF-α (10ng/ml)) were examined. Intensity values for each tissue/cell line represent the geometric mean with geometric standard deviation of 3 independent measurements; each measurement represents the geometric mean of all probes (one per exon) for each gene followed by a median normalization across all genes on the array. The dotted line indicates the threshold level for detection of basal expression. The reference sample (Ref.) is composed of a mixture RNAs derived from 10 different human tissues. **Right panels.** Expression levels of selected genes were determined in a panel of primary immune cells (neutrophils, monocytes, γδ T cells, B cells, NK cells, $CD4^+$ T cells, $CD8^+$ T cells) isolated from healthy donors, as well as monocyte *in vitro* derived macrophages without and with 24 hours of stimulation using 1 ug/ml of lipopolysaccharide (macrophages+LPS). The results presented in the **left and right panels** were generated and analyzed separately and therefore the expression values are not directly comparable.

Extended Data Table 1, The number IBD credible sets that colocalize with expression QTLs using the naïve, permutation-based and Bayesian-based approaches.

Acknowledgements M.J.D. and R.J.X. acknowledge grant supports from P30DK43351, U01DK062432, R01DK64869, Helmsley grant 2015PG-IBD001 and CCFA. C.A.A. and J.C.B are supported by Wellcome Trust grant 098051. M.G. acknowledges grant support from WELBIO (CAUSIBD), BELSPO (BeMGI), Fédération Wallonie-Bruxelles (ARC IBD@Ulg), and Région Wallonne (CIBLES, FEDER). H.H. acknowledges the ASHG/Charles J. Epstein Trainee Award. J.L. acknowledges Wellcome Trust grant 098759/Z/12/Z. D.M. is supported by the Olle Engkvist Foundation and Swedish Research Council (grants 2010-2976 and 2013-3862). R.K.W. is supported by a VIDI grant (016.136.308) from the Netherlands Organization for Scientific Research (NWO). J.D.R. holds a Canada Research Chair and this work was supported by grants from the U.S. National Institute of Diabetes and Digestive and Kidney Diseases (DK064869; DK062432), a grant (CIHR #GPG-102170) from the Canadian Institutes of Health Research to the "CIHR Emerging Team in Integrative Biology of Inflammatory Diseases", and a *Large-Scale Applied Research Project in Genomics and Personalized Health* grant (GPH-129341) co-funded by the Government of Canada through Genome Canada and the Ministère de l'économie, de l'innovation et de l'exportation du Québec through Génome Québec as well as the Canadian Institutes of Health Research and Crohn's Colitis Canada. J.H.C. is funded by U01 DK62429, U01 DK062422, and the Sanford J. Grossman Charitable Trust. R.H.D. holds the Inflammatory Bowel Disease Genetic Research Chair at the University of Pittsburgh and acknowledges grant support from U01DK062420 and R01CA141743. E.D. benefitted from a Marie-Curie Fellowship and A-S.G a from fellowships from the FNRS and Fonds Léon Fredericq. J.H. is supported by the Örebro University Hospital Research Foundation and the Swedish Research Council (grant no. 521 2011 2764). M.P. acknowledges the NIHR Biomedical Research Centre awards to Guy's & St Thomas' NHS Trust / King's College London and to Addenbrooke's Hospital / University of Cambridge School of Clinical Medicine. D.E.: this work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (SysInflame grant 01ZX1306A). This project received infrastructure support from the DFG Excellence Cluster No. 306 "Inflammation at Interfaces". A.F. receives an endowment professorship by the Foundation for Experimental Medicine (Zuerich, Switzerland).

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

Competing Financial Interests The authors declare no competing financial interests.

Methods

Genotyping and QC

We genotyped 35,197 unaffected and 35,346 affected individuals (20,155 Crohn's disease and 15,191 ulcerative colitis) using the Immunochip array. Genotypes were called using optiCall[37] for 192,402 autosomal variants before QC. We removed variants with missing data rate >2% across the whole dataset, or >10% in any one batch, and variants that failed (FDR < 10^{-5} in either the whole dataset or at least two batches) tests for: a) Hardy-Weinberg equilibrium; b) differential missingness between cases and controls; c) significant heterogeneity in allele frequency across controls from different batches. We also removed noncoding variants that were not in the 1000 Genomes Phase I integrated variant set (March 2012 release), or the HapMap phase 2 or 3 releases, as these mostly represent false positives included on Immunochip from the 1000 Genomes pilot, which often genotype poorly. Where a variant failed in exactly one batch we set all genotypes to missing for that batch (to be reimputed later) and included the site if it passed in the remainder of the batches. We removed individuals that had >2% missing data, had significantly higher or lower (defined as FDR<0.01) inbreeding coefficient (*F*), or were duplicated or related (PI_HAT ≥ 0.4, calculated from the LD pruned dataset described below), by sequentially removing the individual with the largest number of related samples until no related samples remain. After QC, there were 67,852 European-derived samples with valid diagnosis (healthy control, Crohn's disease or ulcerative colitis), and 161,681 genotyped variants available for downstream analyses.

Linkage-disequilibrium pruning and principal components analysis

From the clean dataset we removed variants in long range LD[38] or with MAF < 0.05, and then pruned 3 times using the '--indep' option in PLINK (with window size of 50, step

size of 5 and VIF threshold of 1.25). This pruned dataset (18,123 variants) was used to calculate the relatedness of the individuals and the principal components. Principal component axes were generated within controls using this LD pruned dataset. The axes were then projected to cases to generate the principal components for all samples. The analysis was performed using our in-house C code (https://github.com/hailianghuang/efficientPCA) and LAPACK package[39] for efficiency.

Imputation

Imputation was performed separately in each Immunochip high-density region (184 total) from the 1000 Genomes Phase I integrated haplotype reference panel, downloaded from the IMPUTE2 website (Dec 2013 release). We used SHAPEIT (v2.r769)[40,41] to pre-phase the genotypes, followed by IMPUTE2 (2.3.0)[14,15] to perform the imputation. There were 388,432 variants having good imputation quality (INFO > 0.4) and were used in the fine-mapping analysis.

Manual cluster plot inspection

Variants that had posterior probability greater than 50% or in credible sets mapped to ≤ 10 variants were manually inspected using Evoker v2.2[42]. Each variant was inspected by 3 independent reviewers (10 reviewers participated) and scored as pass, fail or maybe. We remove variants that received one or more fails, or received less than 2 passes. 650 out of 905 inspected variants passed this inspection. A further cluster plot inspection flagged two additional failed variants after removing the failed variants from the first inspection and redoing the imputation and analysis.

Establishing a p-value threshold

We used a multiple testing corrected p-value threshold for associations of 10^{-6}, which was established by permutation. We generated 200 permuted datasets by randomly shuffling phenotypes across samples and carried out association analyses for each permutation across all 161,681 variants in our high-density regions. We stored (i) the ensuing 161,681 x 200 point-wise p-values (α_S), as well as (ii) the 200 "best" p-values (α_B) of each permuted datasets. We then computed the empirical, family-wise p-value (α_M)(corrected for multiple testing) for each of the 161,681 x 200 tests as its rank/200 with respect to the 200 α_B. We then estimated the number of independent tests performed in the studied regions, *n*, as the slope of the regression of log(1-α_M) on log(1-α_S), knowing that $\alpha_M = 1 - (1 - \alpha_S)^n$.

Detecting and fine-mapping association signals

We used three fine-mapping methods to detect independent signals and create credible sets across 103 high-density regions (Supplementary Methods). Signals identified by different methods were merged if their credible sets shared one or more variants. In order to adjudicate differences between methods, we first assigned each candidate signal to the combination of a lead variant and trait (CD, UC or IBD) that maximizes the marginal likelihood from equation (8) in Supplementary Methods. At loci with >1 signal, we fixed the signals reported by all three methods, and then tested all possible combinations of signals reported by one or two methods, selecting whichever combination has the highest joint marginal likelihood. We consider signals to be confidently fine-mapped, and take them forward for subsequent analysis, if they a) are in loci where the lead variant has p < 10^{-6}, b) have a ratio of Bayes factors for the best model and the second best model greater than 10, c) are reported by more than one method and d) passed cluster plot inspection.

Phenotype assignment of signals

We assign each signal as CD-specific, UC-specific or shared, using the Bayesian multinomial model used for fine-mapping method 2 (the method best able to assess evidence of sharing in the presence of potentially correlated effect sizes). For the lead variant for each credible set, we calculate the marginal likelihoods as in equation 13 from Supplementary Methods, restricting either $\beta_{UC} = 0$ (for the CD-only model) or $\beta_{CD} = 0$ (for the UC-only model), as well as using the unconstrained prior (for the associated to both model), and select the model with the highest marginal likelihood. We then calculate the log Bayes factor in favor of sharing, i.e. the log of ratio of marginal likelihoods between the associated-to-both model and the best of the single-phenotype associated models.

Estimating the variance explained by the fine-mapping

We used a mixed model framework to estimate the total risk variance attributable to the IBD risk loci, and to the signals identified in the fine-mapping. We the GCTA software package[43] to compute a gametic relationship matrix (G-matrix) using genotype dosage information for the genotyped variants in the high-density regions (which we will call $\mathbf{G}_{HD}$). We then fit a variety of variance component models by restricted maximum likelihood analysis using an underlying liability threshold model implemented with the DMU package[44]. The first model is a standard heritability mixed-model that includes fixed effects for five principal components (to correct for stratification) and a random effect summarizing the contribution of all variants in the fine-mapping regions, such that the liabilities across all individuals are distributed according to

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

where λ_1 is thus the variance explained by all variants in fine-mapping regions, which we estimate. We then fitted a model that included an additional random effect for the contribution of the lead variants have been specifically identified (with G-matrix $\mathbf{G}_{Signals}$), such that liability is distributed

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

The variance explained by the signals under consideration is then given by the reduction in the variance explained by all variants in the fine-mapping regions between the two models ($\lambda_1 - \lambda'_1$). We used this approach to estimated what fraction of this variance was accounted for by (i) the single strongest signals in each region (as would be typically done prior to fine-mapping), or (ii) the all signals identified in fine-mapping. We used Cox and Snell's method[45] to estimate the variance explained across independent signals (Extended Data Figure 2b) for computational efficiency.

Overlap between transcription factor binding motifs and causal variants

For each motif in the ENCODE TF ChIP-seq data (http://compbio.mit.edu/encode-motifs/, accessed Nov 2014) [20], we calculated the overall information content (IC) as the sum of IC for each position[46], and only considered motifs with overall IC ≥ 14 bits (equivalent to 7 perfectly conserved positions). For every variant in a high-density region we determined if it creates or disrupts a motif at a high-information site (IC ≥ 1.8). For each credible set that contains a motif-affecting variant, we calculated a p-value as $1 - (1 - f)^n$, where *n* is the size of the credible set and *f* is the proportion of all variants in the high-density region that disrupt or a create a motif in that TF family.

Overlap between epigenetic signatures and causal variants

For each combination of 120 tissues and three histone marks (H3K4me1, H3K4me3 and H3K27ac) from the Roadmap Epigenome Project we calculated an overlap score, equal to the sum of fine-mapping posterior probabilities for all variants in peaks of that histone mark in that tissue. We generated a null distribution of this score for each tissue/mark by shifting chromatin marks randomly over the high-density regions (shifting the peaks from their actual position by a random number of bases while keeping inter-peak spacing the same) and calculating the overlap score for each permutation. To summarize these correlated results across many cell and tissue types we defined a set of "core" H3K4me1 immune and H3K27ac gut peaks as sets of overlapping peaks in cells that showed the strongest enrichment ($p<10^{-4}$). Intersects were made using bedtools v2.24.0 default settings[47]. We selected 6 immune cell types for H3K4me1 and 3 gut cell types for H3K27ac (Supplementary Table 2). We also chose controls (Supplementary Table 2) from non-immune and non-gut cell types with similar density of peaks in the fine-mapped regions as compared to immune/gut cell types to confirm the tissue-specificity of the overlap. We used the phenotype assignments (described above) in dissecting the enrichment for the CD and UC signals. Sixty-five CD and 21 UC signals were used in this analysis.

Published eQTL summary statistics

We used eQTL summary statistics from two published studies:

- Peripheral blood eQTLs from the GODOT study[48] of 2,752 twins, reporting loci with MAF>0.5%.
- CD14+ monocyte eQTLs from Table S2 in Fairfax *et al.*[31], comprised of 432 European individuals, measured in a naïve state and after stimulation with interferon-

γ (for 2 or 24 hours) or lipopolysaccharide. Reports loci with MAF>4% and FDR<0.05.

Processing and quality control of new eQTL ULg dataset

A detailed description of the ULg dataset is in preparation (Momozawa et al., in preparation). Briefly, we collected venous blood and intestinal biopsies at three locations (ileum, transverse colon and rectum) from 350 healthy individuals of European descent, average age 54 (range 17-87), 56% female. SNPs were genotyped on Illumina Human OmniExpress v1.0 arrays interrogating 730,525 variants, and SNPs and individuals were subject to standard QC procedures using call rate, Hardy-Weinberg equilibrium, MAF ≥ 0.05, and consistency between declared and genotype-based sex as criteria. We further imputed genotypes at ~7 million variants on the entire cohort using the Impute2 software package[14] and the 1,000 Genomes Project as reference population (Phase 3 integrated variant set, released 12 Oct 2014) [11,15]. From the blood, we purified CD4+, CD8+, CD19+, CD14+ and CD15+ cells by positive selection, and platelets (CD45-negative) by negative selection. RNA from all leucocyte samples and intestinal biopsies was hybridized on Illumina Human HT-12 arrays v4. After standard QC, raw fluorescent intensities were variance stabilized[49] and quantile normalized[50] using the lumi R package[51], and were corrected for sex, age, smoking status, number of probes with expression level significantly above background as fixed effects and array number (sentrix id) as random effect. For each probe with measureable expression (detection p-value < 0.05 in >25% of samples) we tested for cis-eQTLs at all variants within a 500 kilobase window. The nominal p-value of the best SNP within a cis-window was Sidak-corrected for the window-specific number of independent tests, and we estimated false

652 discovery rates (q-values) from the resulting p-values across all probes using the qvalue
653 R package[52]. 480 cis-eQTL with FDR ≤ 0.10 with the lead SNPs within the 97 high-
654 density regions (94 fine-mapped plus 3 unresolved) were retained for further analyses.

655 **Naïve co-localization using lead SNPs**

656 We calculated the proportion of IBD credible sets that contain a lead eQTL variant in a
657 particular tissue. This value is then compared to a background rate:

$$\frac{1}{|S|}\sum_{i\in S}(1-(1-N_i^{-1})^{C_i})$$

658 where N_i is the total number of variants in region i in 1000 Genomes with an allele
659 frequency greater than a certain threshold (equal to the threshold used for the original
660 eQTL study), C_i is the number of these variants that lie in IBD credible sets, and S is a set
661 of regions that have at least one significant eQTL ($|S|$ is the number of regions in this
662 set). P-values can then be calculated assuming a binomial distribution with probability
663 equal to the background rate and the number of trials equal to $|S|$.

664 **Frequentist co-localization using conditional p-values**

665 We next used conditional association to test for evidence of co-localization, as described
666 in Nica et al.[25]. This method compares the p-value of association for the lead SNP of an
667 eQTL before and after conditioning on the SNP with the highest posterior in the credible
668 set, and measures the drop in log(1/p). An empirical p-value for this drop is then
669 calculated by comparing it to the drop for all variants (with MAF ≥ 0.05) in the high-
670 density region. An empirical p-value ≤ 0.05 was considered as evidence that the
671 corresponding credible set is co-localized with the corresponding cis-eQTL. To evaluate
672 whether our 139 credible sets affected cis-eQTL more often than expected by chance we
673 counted the number of credible sets affecting at least one cis-eQTL with p-value ≤ 0.05,

and compared how often this number was matched or exceeded by 1,000 sets of 139 lead variants that were randomly selected yet distributed amongst the 94 loci in accordance with the real credible sets.

Bayesian co-localization using Bayes factors

Finally, we used the Bayesian co-localization methodology described by Giambartolomei et al[53], modified to use the credible sets and posteriors generated by our fine-mapping methods. The method takes as input a pair of IBD and eQTL signals, with corresponding credible sets S^{IBD} and S^{eQTL}, and posteriors for each variant p_i^{IBD} and p_i^{eQTL} (with $p_i^X = 0 \; \forall \; i \notin S^X$). Credible sets and posteriors were generated for eQTL signals using the Bayesian quantitative association mode in SNPTest (with default parameters), with credible sets in regions with multiple independent signals generated conditional on all other signals. Our method calculates a Bayes factor summarizing the evidence in favor of a colocalized model (i.e. a single underlying causal variant between the IBD and eQTL signals) compared to a non-colocalized model (where different causal variants are driving the two signals), given by the ratio of marginal likelihoods

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

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

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

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

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

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

Permutation analysis. To measure enrichment in colocalization Bayes factors compared to the null, we carried out a permutation analysis. In this analysis, we randomly reassigned eQTL signals to new fine-mapping regions to generate a set of simulated null datasets. This is carried out using the following scheme:

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

We can use these permuted Bayes factors to calculate p-values for each IBD credible set, given by the proportion of time the permuted BFs were as large or greater than the one

714 observed in the real dataset. To generate a high-quality set of colocalized eQTL and IBD

715 signals, we take all signals that have BF > 2, p < 0.01 and r^2 between hits of >0.8.

716

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. 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
14. 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).
15. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1,** 457–470 (2011).
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

762 frequency IL23R coding variants protecting against inflammatory bowel disease.
763 *Nat Genet* **43,** 43–47 (2011).
764 20. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of
765 regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42,**
766 2976–2987 (2014).
767 21. Nechanitzky, R. *et al.* Transcription factor EBF1 is essential for the maintenance
768 of B cell identity and prevention of alternative fates in committed cells. *Nat.*
769 *Immunol.* **14,** 867–875 (2013).
770 22. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping
771 complex trait variants. *Nat Genet* **45,** 124–130 (2013).
772 23. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune
773 disease variants. *Nature* **518,** 337–343 (2015).
774 24. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat*
775 *Biotechnol* **28,** 1045–1048 (2010).
776 25. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression
777 QTLs with complex trait genetic associations. *PLoS Genet* **6,** e1000895 (2010).
778 26. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional
779 variation in humans. *Nature* **501,** 506–511 (2013).
780 27. Wallace, C. *et al.* Statistical colocalization of monocyte gene expression and
781 genetic risk variants for type 1 diabetes. *Human Molecular Genetics* **21,** 2815–
782 2824 (2012).
783 28. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and
784 rare variant association signals in celiac disease. *Nat Genet* **43,** 1193–1201 (2011).
785 29. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple
786 tissues in twins. *Nat Genet* **44,** 1084–1089 (2012).
787 30. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of
788 known disease associations. *Nat Genet* **45,** 1238–1243 (2013).
789 31. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory
790 variants upon monocyte gene expression. *Science (New York, NY)* **343,** 1246949–+
791 (2014).
792 32. Schizophrenia Working Group of the Psychiatric Genomics Consortium.
793 Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,**
794 421–427 (2014).
795 33. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of
796 confirmed Crohn's disease susceptibility loci. *Nat Genet* **42,** 1118–1125 (2010).
797 34. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of
798 IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes.
799 *Science (New York, NY)* **324,** 387–389 (2009).
800 35. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based
801 imputation identifies novel and refined associations for the Wellcome Trust Case
802 Control Consortium phase 1 Data. *Eur. J. Hum. Genet.* **20,** 801–805 (2012).
803 36. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes.
804 *Nature* **518,** 317–330 (2015).
805 37. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-
806 frequency and common variants. *Bioinformatics* **28,** 1598–1603 (2012).
807 38. Price, A. L. *et al.* Long-Range LD Can Confound Genome Scans in Admixed

Populations. *Am J Hum Genet* **83,** 132–135 (2008).

39. Anderson, E. *et al*. *LAPACK Users' Guide*. (Society for Industrial and Applied Mathematics, 1999).
40. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9,** 179–181 (2011).
41. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods* **10,** 5–6 (2012).
42. Morris, J. A., Randall, J. C., Maller, J. B. & Barrett, J. C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26,** 1786–1787 (2010).
43. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88,** 76–82 (2011).
44. Madsen, P., Su, G., Labouriau, R. & Christensen, F. DMU-a package for analyzing multivariate mixed models. *Proceedings of the Ninth World Congress on Genetics Applied to Livestock Production* (2010).
45. Cox, D. R. & Snell, E. J. *Analysis of Binary Data, Second Edition*. (CRC Press, 1989).
46. D'haeseleer, P. What are DNA sequence motifs? *Nat Biotechnol* **24,** 423–425 (2006).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
48. Wright, F. A. *et al.* Heritability and genomics of gene expression in peripheral blood. *Nat Genet* **46,** 430–437 (2014).
49. Lin, S. M., Du, P., Huber, W. & Kibbe, W. A. Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic Acids Res* **36,** e11–e11 (2008).
50. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–193 (2003).
51. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (2008).
52. Dabney, A., Storey, J. D. & Warnes, G. Q-value estimation for false discovery rate control. *R package version* **1,** (2006).
53. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10,** e1004383 (2014).

Acknowledgements (original data): We thank all the subjects who contributed samples and the physicians and nursing staff who helped with recruitment globally. UK case collections were supported by the National Association for Colitis and Crohn's disease; Wellcome Trust grant 098051 (L.J., C.A.A., J.C.B.); Medical Research Council UK; the Catherine McEwan Foundation; an NHS Research Scotland career fellowship (R.K.R.); Peninsula College of Medicine and Dentistry, Exeter; the National Institute for Health Research, through the Comprehensive Local Research Network, and through Biomedical Research Centre awards to Guy's & Saint Thomas' National Health Service Trust, King's College London, Addenbrooke's Hospital, University of Cambridge School of Clinical Medicine and to the University of Manchester and Central Manchester Foundation Trust. The British 1958 Birth Cohort DNA collection was funded by Medical Research Council grant G0000934 and Wellcome Trust grant 068545/Z/02, and the UK National Blood Service controls by the Wellcome Trust. The Wellcome Trust Case Control Consortium projects were supported by Wellcome Trust grants 083948/Z/07/ Z, 085475/B/08/Z and 085475/Z/08/Z. North American collections and data processing were supported by funds to the National Institute of Diabetes, Digestive and Kidney diseases (NIDDK) IBD Genetics Consortium, which is funded by the following grants: DK062431 (S.R.B.), DK062422 (J.H.C.), DK062420 (R.H.D.), DK062432 (J.D.R.), DK062423 (M.S.S.), DK062413 (D.P.M.), DK076984 (M.J.D.), DK084554 (M.J.D. and D.P.M.) and DK062429 (J.H.C.). The funding source which enabled our recruitment of New Zealand subjects for the IIBDGC was the New Zealand Ministry of Business, Innovation and Employment.

Members of the International Inflammatory Bowel Disease Genetics Consortium:
Clara Abraham[35], Jean-Paul Achkar[36,37], Tariq Ahmad[38], Leila Amininejad[39,40], Ashwin N Ananthakrishnan[29,41], Vibeke Andersen[9,10,42], Carl A Anderson[7], Jane M Andrews[43], Vito Annese[44,45], Guy Aumais[33,46], Leonard Baidoo[30], Robert N Baldassano[47], Peter A Bampton[48], Murray Barclay[49], Jeffrey C Barrett[7], Theodore M Bayless[50], Johannes Bethge[51], Alain Bitton[52], Gabrielle Boucher[8], Stephan Brand[53], Berenice Brandt[51], Steven R Brant[50], Carsten Büning[54], Angela Chew[20,55], Judy H Cho[34], Isabelle Cleynen[11], Ariella Cohain[56], Anthony Croft[57], Mark J Daly[1,2], Mauro D'Amato[12,13], Silvio Danese[58], Dirk De Jong[59], Martine De Vos[60], Goda Denapiene[61], Lee A Denson[62], Kathy L Devaney[29], Olivier Dewit[63], Renata D'Inca[64], Marla Dubinsky[65], Richard H Duerr[30,31], Cathryn Edwards[66], David Ellinghaus[15], Jonah Essers[67,68], Lynnette R Ferguson[69], Eleonora A Festen[27], Philip Fleshner[17], Tim Florin[70], Denis Franchimont[39,40], Andre Franke[15], Karin Fransen[71], Richard Gearry[49,72], Michel Georges[3,4], Christian Gieger[73], Jürgen Glas[53,74], Philippe Goyette[8], Todd Green[2,67], Anne M Griffiths[75], Stephen L Guthery[76], Hakon Hakonarson[47], Jonas Halfvarson[16], Katherine Hanigan[57], Talin Haritunians[17], Ailsa Hart[77], Chris Hawkey[78], Nicholas K Hayward[79], Matija Hedl[35], Paul Henderson[80,81], Xinli Hu[82], Hailiang Huang[1,2], Jean-Pierre Hugot[83], Ken Y Hui[34], Marcin Imielinski[47], Andrew Ippoliti[17], Laimas Jonaitis[84], Luke Jostins[5,6], Tom H Karlsen[85,86,87], Nicholas A Kennedy[21], Mohammed Azam Khan[88,89], Gediminas Kiudelis[84], Krupa Krishnaprasad[90], Subra Kugathasan[91], Limas Kupcinskas[92], Anna Latiano[44], Debby Laukens[60], Ian C Lawrance[19,20], James C Lee[24], Charlie W Lees[21], Marcis Leja[93], Johan Van Limbergen[75], Paolo Lionetti[94], Jimmy Z Liu[7], Edouard Louis[22], Gillian Mahy[95], John Mansfield[96], Dunecan Massey[24], Christopher G Mathew[25,32], Dermot PB McGovern[17], Raquel

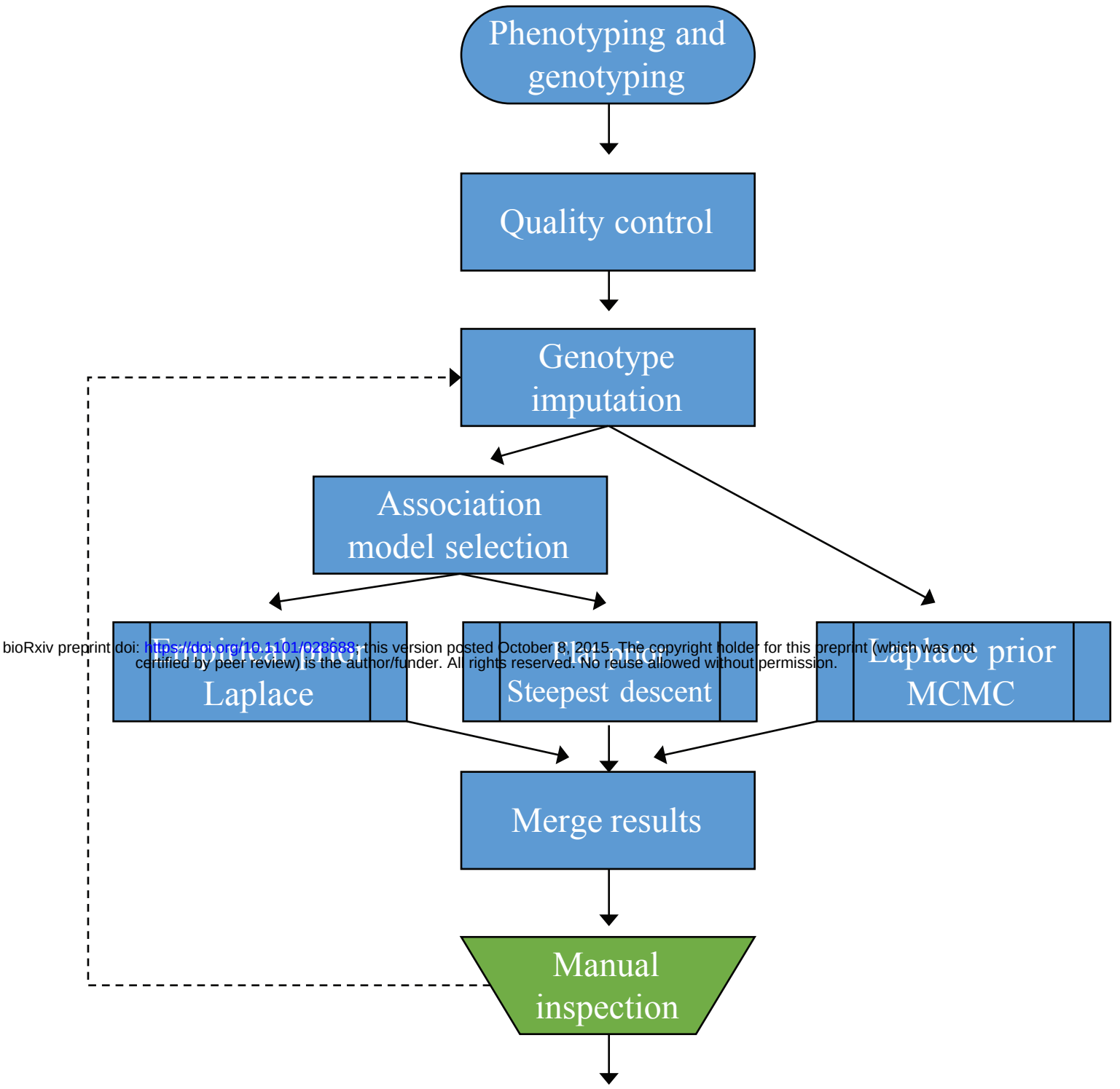
Milgrom[97], Mitja Mitrovic[71,98], Grant W Montgomery[79], Craig Mowat[99], William Newman[88,89], Aylwin Ng[29,100], Siew C Ng[101], Sok Meng Evelyn Ng[35], Susanna Nikolaus[51], Kaida Ning[35], Markus No□then[102], Ioannis Oikonomou[35], Orazio Palmieri[44], Miles Parkes[24], Anne Phillips[99], Cyriel Y Ponsioen[103], Urõs Potocnik[98,104], Natalie J Prescott[25], Deborah D Proctor[35], Graham Radford-Smith[57,105], Jean-Francois Rahier[106], Soumya Raychaudhuri[82], Miguel Regueiro[30], Florian Rieder[36], John D Rioux[8,33], Stephan Ripke[1,2], Rebecca Roberts[49], Richard K Russell[80], Jeremy D Sanderson[107], Miquel Sans[108], Jack Satsangi[21], Eric E Schadt[56], Stefan Schreiber[15,51], Dominik Schulte[51], L Philip Schumm[109], Regan Scott[30], Mark Seielstad[110,111], Yashoda Sharma[35], Mark S Silverberg[97], Lisa A Simms[57], Jurgita Skieceviciene [84], Sarah L Spain[25,26], A. Hillary Steinhart[97], Joanne M Stempak[97], Laura Stronati[112], Jurgita Sventoraityte[92], Stephan R Targan[17], Kirstin M Taylor[107], Anje ter Velde[103], Emilie Theatre[3,4], Leif Torkvist[113], Mark Tremelling[114], Andrea van der Meulen[115], Suzanne van Sommeren[27], Eric Vasiliauskas[17], Severine Vermeire[11,28], Hein W Verspaget[115], Thomas Walters[75,116], Kai Wang [47], Ming-Hsi Wang [36,50], Rinse K Weersma[27], Zhi Wei[117], David Whiteman[79], Cisca Wijmenga[71], David C Wilson[80,81], Juliane Winkelmann[118,119], Ramnik J Xavier[2,29], Bin Zhang[56], Clarence K Zhang[120], Hu Zhang[121,122], Wei Zhang[35], Hongyu Zhao[120], Zhen Z Zhao[79]

[35]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA. [36]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA. [37]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA. [38]Peninsula College of Medicine and Dentistry, Exeter, UK. [39]Department of Gastroenterology, Erasmus Hospital, Brussels, Belgium. [40]Department of Gastroenterology, Free University of Brussels, Brussels, Belgium. [41]Division of Medical Sciences, Harvard Medical School, Boston, Massachusetts, USA. [42]Institute of Regional Health Research , University of Southern Denmark, Odense, Denmark. [43]Inflammatory Bowel Disease Service, Department of Gastroenterology and Hepatology, Royal Adelaide Hospital, Adelaide, Australia. [44]Unit of Gastroenterology, Istituto di Ricovero e Cura a Carattere Scientifico-Casa Sollievo della Sofferenza (IRCCS-CSS) Hospital, San Giovanni Rotondo, Italy. [45]Strutture Organizzative Dipartimentali (SOD) Gastroenterologia 2, Azienda Ospedaliero Universitaria (AOU) Careggi, Florence, Italy. [46]Department of Gastroenterology, Ho□pital Maisonneuve-Rosemont, Montréal, Québec, Canada. [47]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [48]Department of Gastroenterology and Hepatology, Flinders Medical Centre and School of Medicine, Flinders University, Adelaide, Australia. [49]Department of Medicine, University of Otago, Christchurch, New Zealand. [50]Meyerhoff Inflammatory Bowel Disease Center, Department of medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [51]Department for General Internal Medicine, Christian-Albrechts-University, Kiel, Germany. [52]Division of Gastroenterology, Royal Victoria Hospital, Montréal, Québec, Canada. [53]Department of Medicine II, Ludwig-Maximilians-University Hospital Munich-Grosshadern, Munich, Germany. [54]Department of Gastroenterology, Campus Charité Mitte, Universitatsmedizin Berlin, Berlin, Germany. [55]IBD unit , Fremantle Hospital, Fremantle, Australia. [56]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA. [57]Inflammatory Bowel Diseases, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [58]IBD Center, Department of Gastroenterology, Istituto Clinico Humanitas, Milan, Italy. [59]Department of Gastroenterology and Hepatology, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [60]Department of Hepatology and Gastroenterology, Ghent University Hospital, Ghent, Belgium. [61]Center of hepatology, Gastroenterology and Dietetics, Vilnius University, Vilnius, Lithuania. [62]Pediatric Gastroenterology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [63]Department of Gastroenterology, Université Catholique de Louvain (UCL) Cliniques Universitaires Saint-Luc, Brussels, Belgium. [64]Division of Gastroenterology, University Hospital Padua, Padua, Italy. [65]Department of Pediatrics, Cedars Sinai Medical Center, Los Angeles, California, USA. [66]Department of Gastroenterology, Torbay Hospital, Torbay, Devon, UK. [67]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [68]Pediatrics, Harvard Medical School, Boston, Massachusetts, USA. [69]Faculty of

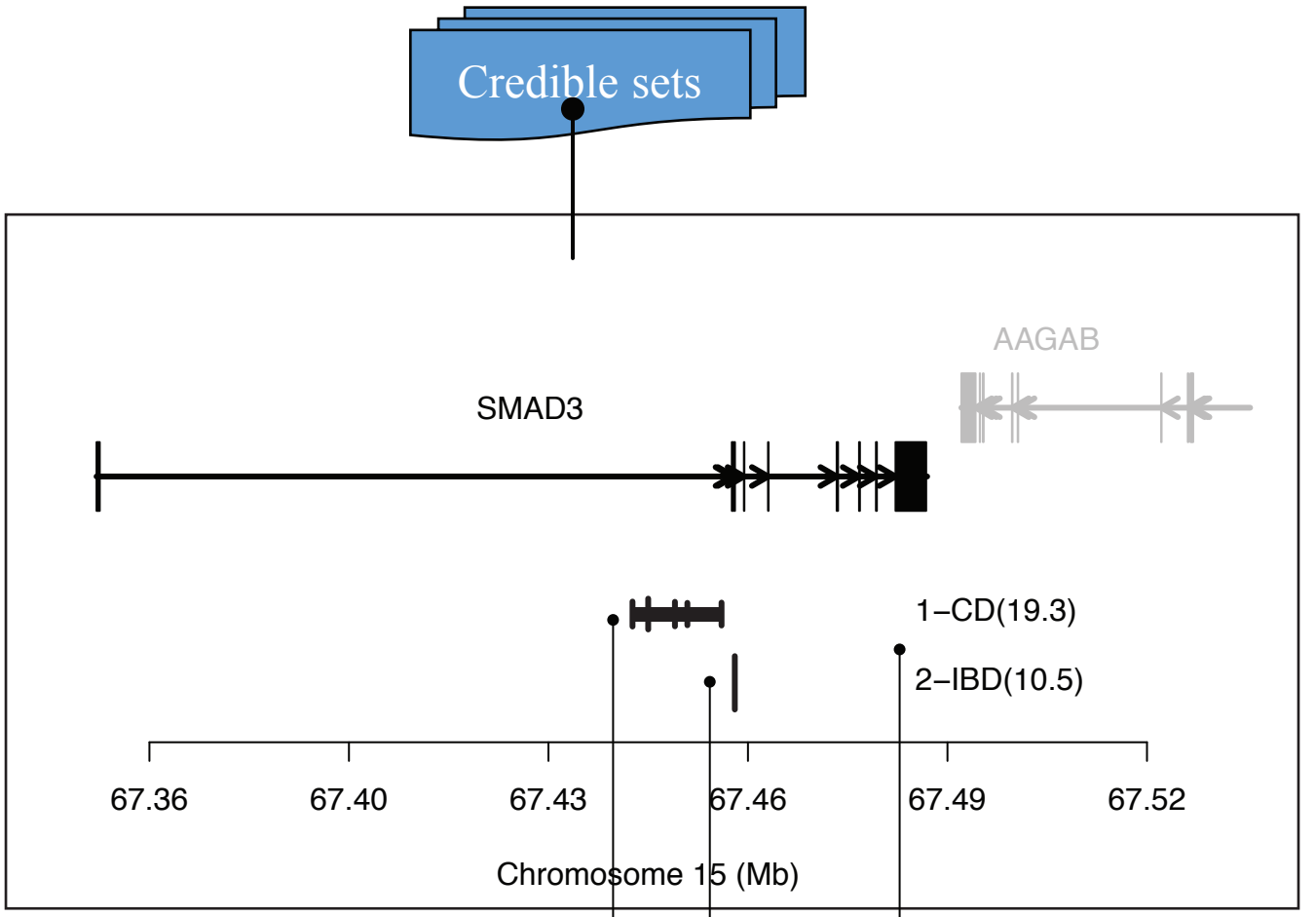
Medical & Health Sciences, School of Medical Sciences, The University of Auckland, Auckland, New Zealand. [70]Department of Gastroenterology, Mater Health Services, Brisbane, Australia. [71]Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands. [72]Department of Gastroenterology, Christchurch Hospital, Christchurch, New Zealand. [73]Institute of Genetic Epidemiology, Helmholtz Zentrum Mu□nchen - German Research Center for Environmental Health, Neuherberg, Germany. [74]Department of Preventive Dentistry and Periodontology, Ludwig-Maximilians-University Hospital Munich-Grosshadern, Munich, Germany. [75]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada. [76]Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah, USA. [77]Department of Medicine, St Mark's Hospital, Harrow, Middlesex, UK. [78]Nottingham Digestive Diseases Centre, Queens Medical Centre, Nottingham, UK. [79]Molecular Epidemiology, Genetics and Computational Biology, Queensland Institute of Medical Research, Brisbane, Australia. [80]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, Edinburgh, UK. [81]Child Life and Health, University of Edinburgh, Edinburgh, Scotland, UK. [82]Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [83]Université Paris Diderot, Sorbonne Paris-Cité, Paris, France. [84]Academy of Medicine, Lithuanian University of Health Sciences, Kaunas, Lithuania. [85]Research Institute of Internal Medicine, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [86]Norwegian PSC Research Center, Department of Transplantation Medicine, Division of Cancer, Surgery and Transplantation, Oslo University Hospital Rikshospitalet, Oslo, Norway. [87]K.G. Jebsen Inflammation Research Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway. [88]Genetic Medicine, Manchester Academic Health Science Centre, Manchester, UK. [89]The Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK. [90]QIMR Berghofer Medical Research Institute, Royal Brisbane Hospital, Brisbane, Australia. [91]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA. [92]Department of Gastroenterology, Kaunas University of Medicine, Kaunas, Lithuania. [93]Faculty of medicine, University of Latvia, Riga, Latvia. [94]Dipartimento di Neuroscienze, Psicologia, Area del Farmaco e Salute del Bambino (NEUROFARBA), Universitˆ di Firenze Strutture Organizzative Dipartimentali (SOD) Gastroenterologia e Nutrizione Ospedale pediatrico Meyer, Firenze, Italy. [95]Department of Gastroenterology, The Townsville Hospital, Townsville, Australia. [96]Institute of Human Genetics, Newcastle University, Newcastle upon Tyne, UK. [97]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada. [98]Center for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Maribor, Slovenia. [99]Department of Medicine, Ninewells Hospital and Medical School, Dundee, UK. [100]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [101]Department of Medicine and Therapeutics, Institute of Digestive Disease, Chinese University of Hong Kong, Hong Kong. [102]Department of Genomics Life & Brain Center, University Hospital Bonn, Bonn, Germany. [103]Department of Gastroenterology, Academic Medical Center, Amsterdam, The Netherlands. [104]Faculty for Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia. [105]Department of Gastroenterology, Royal Brisbane and Womens Hospital, Brisbane, Australia. [106]Department of Gastroenterology, Université Catholique de Louvain (UCL) Centre Hospitalier Universitaire (CHU) Mont-Godinne, Mont-Godinne, Belgium. [107]Department of Gastroenterology, Guy's & St Thomas' NHS Foundation Trust, St-Thomas Hospital, London, UK. [108]Department of Digestive Diseases, Hospital Quiron Teknon, Barcelona, Spain. [109]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA. [110]Human Genetics, Genome Institute of Singapore, Singapore. [111]Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. [112]Department of Biology of Radiations and Human Health, Agenzia nazionale per le nuove tecnologie l'energia e lo sviluppo economico sostenibile (ENEA), Rome, Italy. [113]Department of Clinical Science Intervention and Technology, Karolinska Institutet, Stockholm, Sweden. [114]Gastroenterology & General Medicine, Norfolk and Norwich University Hospital, Norwich, UK. [115]Department of Gastroenterology, Leiden University Medical Center, Leiden, The Netherlands. [116]Faculty of medicine, University of Toronto, Toronto, Ontario, Canada. [117]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA. [118]Institute of Human Genetics, Technische Universita□t Mu□nchen, Munich, Germany. [119]Department of Neurology, Technische Universita□t Mu□nchen, Munich, Germany. [120]Department of Biostatistics, School of Public Health, Yale University, New Haven, Connecticut, USA. [121]Department of Gastroenterology, West China Hospital, Chengdu, Sichuan, China. [122]State Key Laboratory of Biotherapy, Sichuan University West China University of Medical Sciences (WCUMS), Chengdu, Sichuan, China

Figure 1

a



b



This region has been mapped to two independent signals. Phenotype and -log10(P-val) reported for each signal

Signal 1 fine-mapped to 5 variants spanning 13.88kb

variant	size	position	AF	Probability	TFBS	Epigenetic
rs17293632	5	67442596	0.245	0.400	AP1	
rs72743461	5	67441750	0.246	0.191		
rs56375023	5	67448363	0.245	0.173		
rs56062135	5	67455630	0.245	0.104		
rs17228058	5	67450305	0.245	0.082		Gut_H3K27ac

Posteior probability of the variants. (sum of the probabilities ≥ 95%)

Functional annotation

Signal 2 fine-mapped to a single coding variant with 99% probability: rs35874463 [SMAD3-I170V]

Figure 2

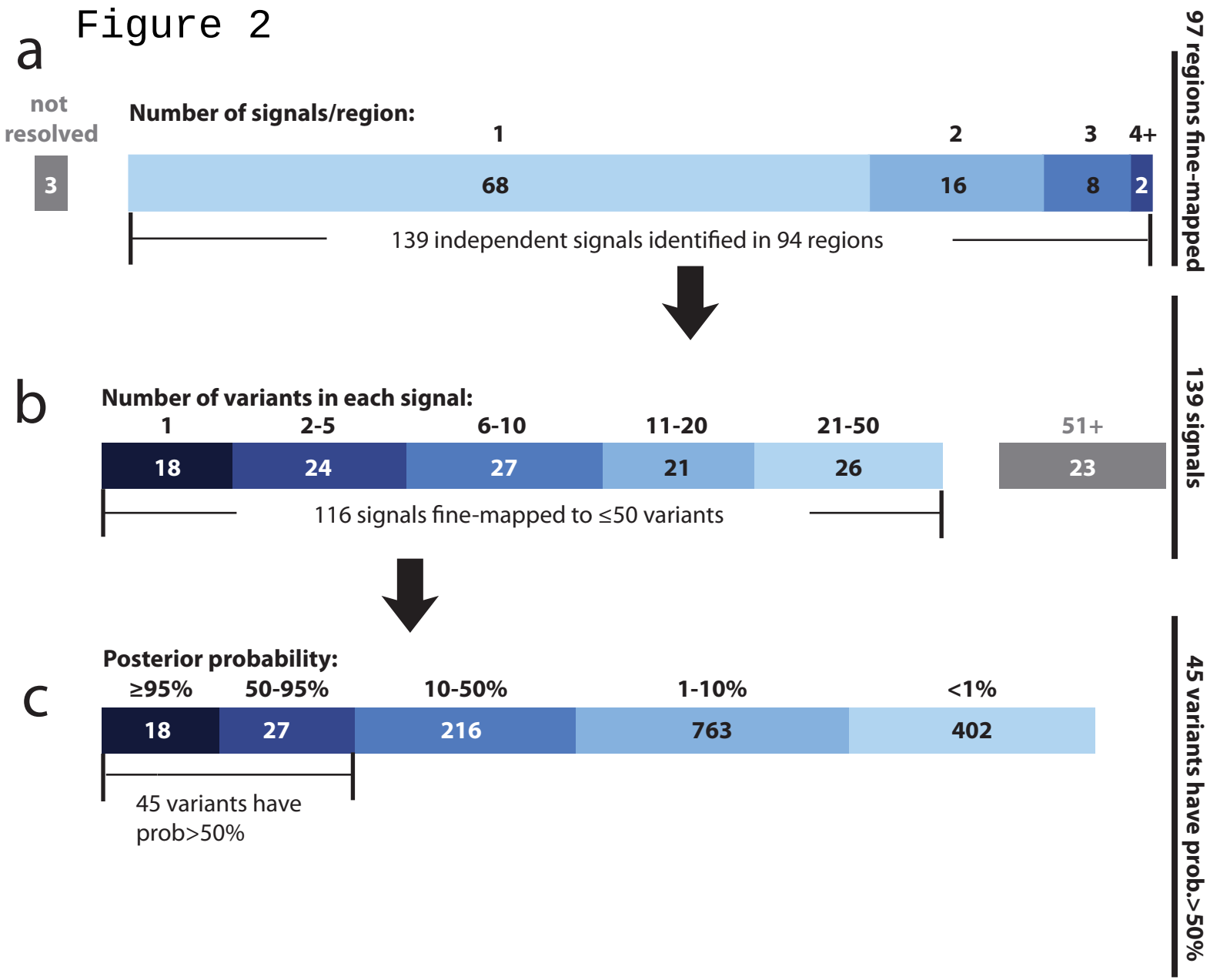
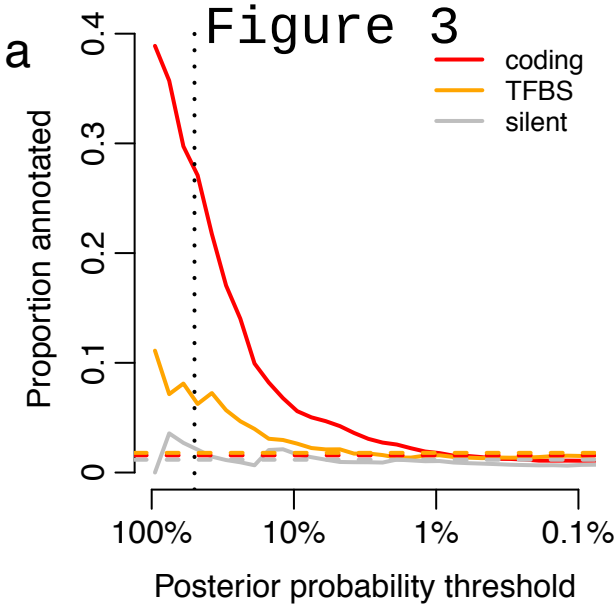
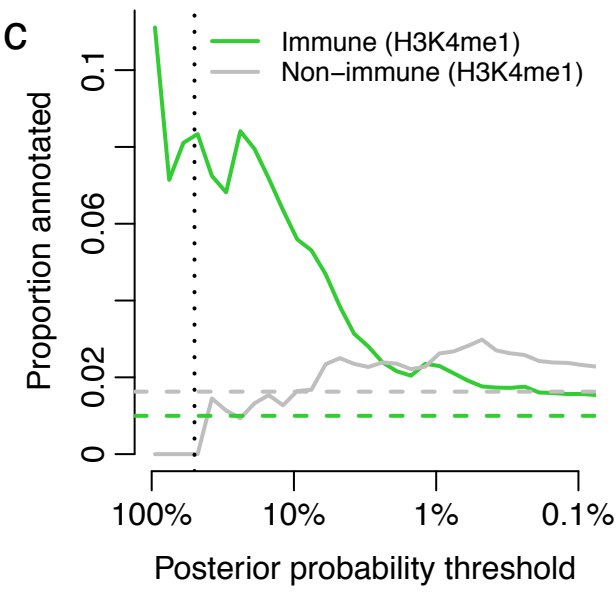
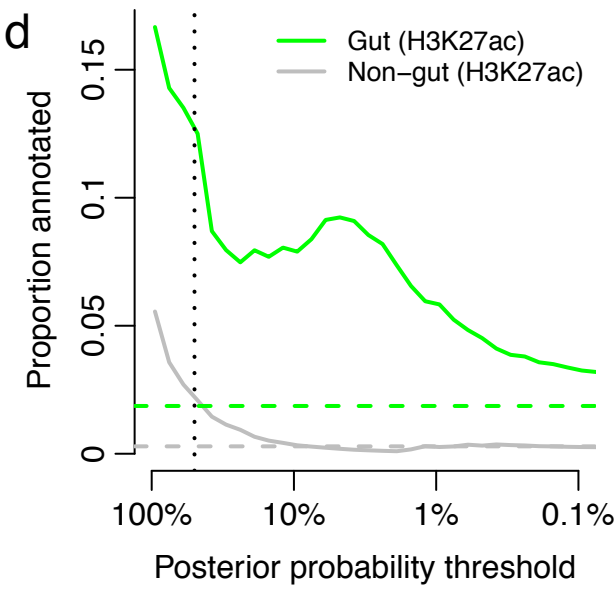

a
not resolved
3
Number of signals/region:
1
2
3
4+
68
16
8
2
139 independent signals identified in 94 regions
97 regions fine-mapped
b
Number of variants in each signal:
1
2-5
6-10
11-20
21-50
51+
18
24
27
21
26
23
116 signals fine-mapped to ≤50 variants
139 signals
c
Posterior probability:
≥95%
50-95%
10-50%
1-10%
<1%
18
27
216
763
402
45 variants have prob>50%
45 variants have prob.>50%

Figure 3


a
Proportion annotated
0
0.1
0.2
0.3
0.4
coding
TFBS
silent
100%
10%
1%
0.1%
Posterior probability threshold


c
Proportion annotated
0
0.02
0.06
0.1
Immune (H3K4me1)
Non−immune (H3K4me1)
100%
10%
1%
0.1%
Posterior probability threshold


d
Proportion annotated
0
0.05
0.1
0.15
Gut (H3K27ac)
Non−gut (H3K27ac)
100%
10%
1%
0.1%
Posterior probability threshold


b
-log10(p-value)
0
1
2
3
4
5
H3K4me1
H3K4me3
H3K27ac
Sample category
Immune
Gut
Others

Figure 4

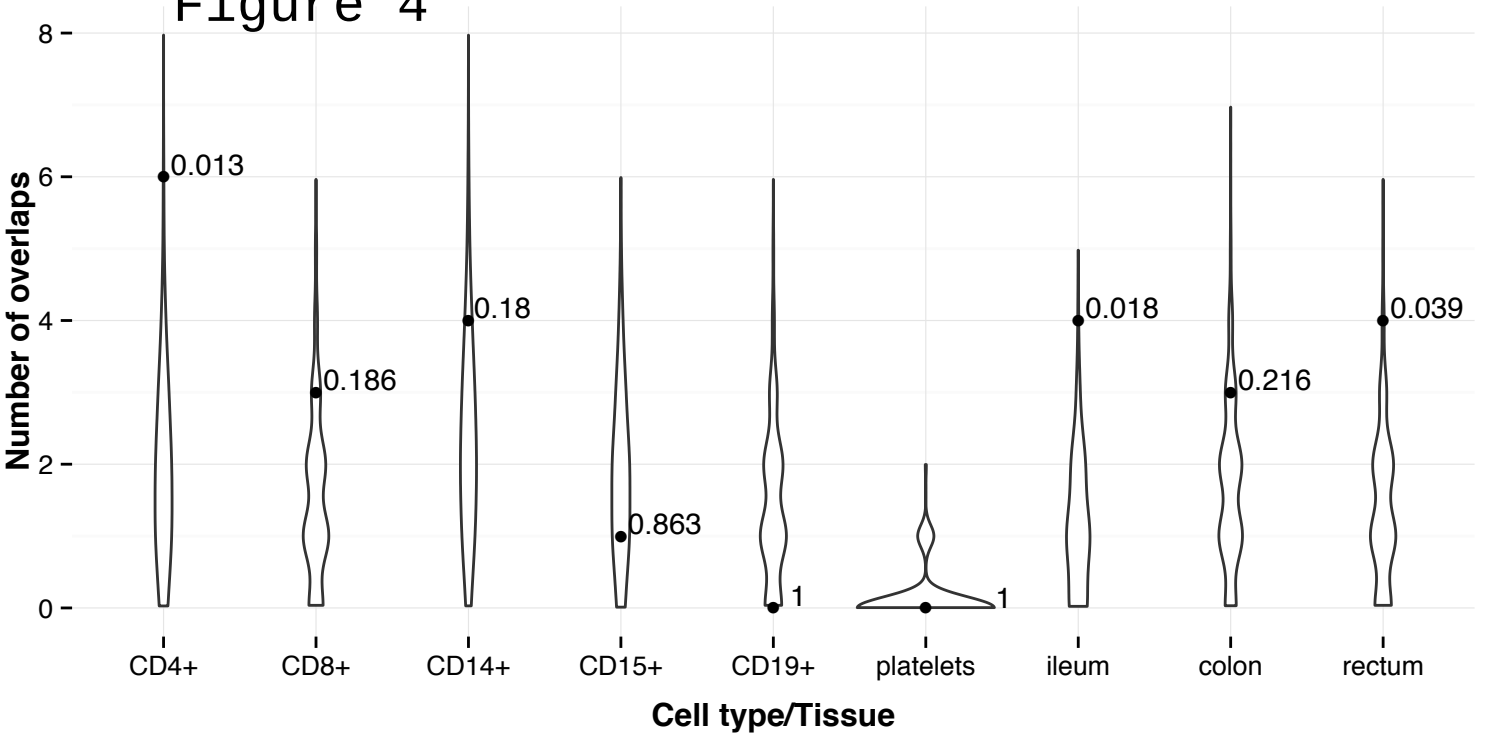
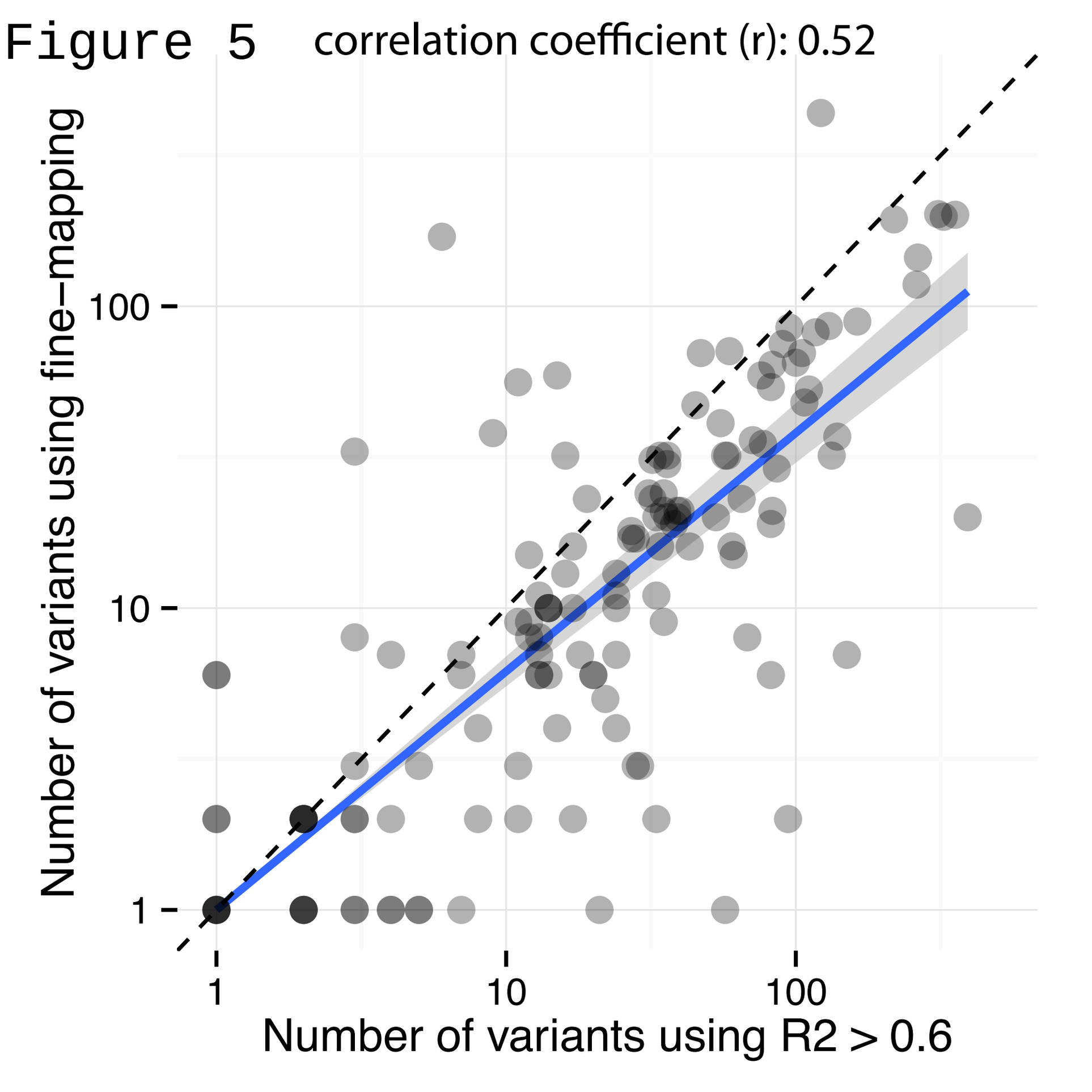
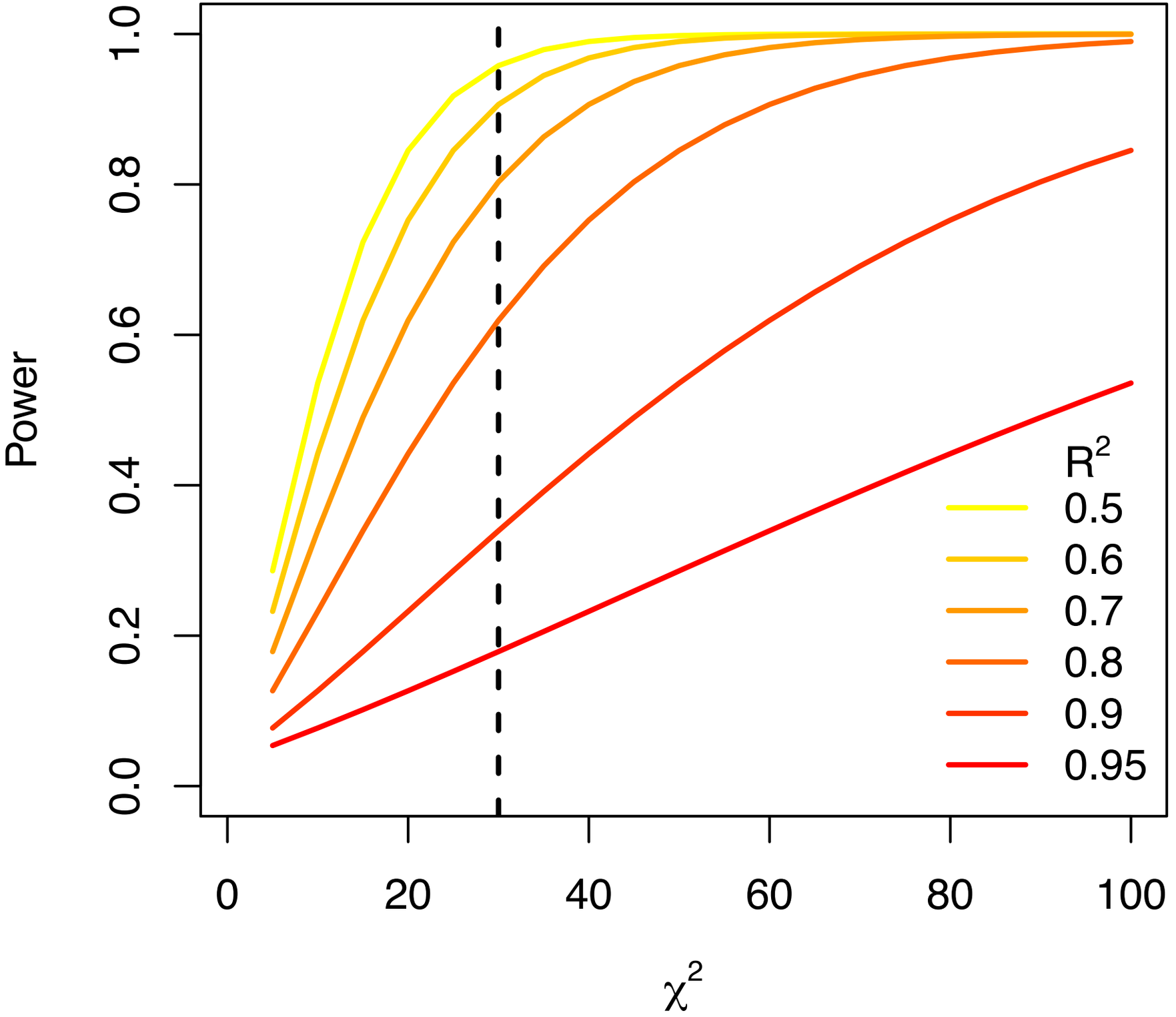

Number of overlaps
8
6
4
2
0
0.013
0.186
0.18
0.863
1
1
0.018
0.216
0.039
CD4+
CD8+
CD14+
CD15+
CD19+
platelets
ileum
colon
rectum
Cell type/Tissue

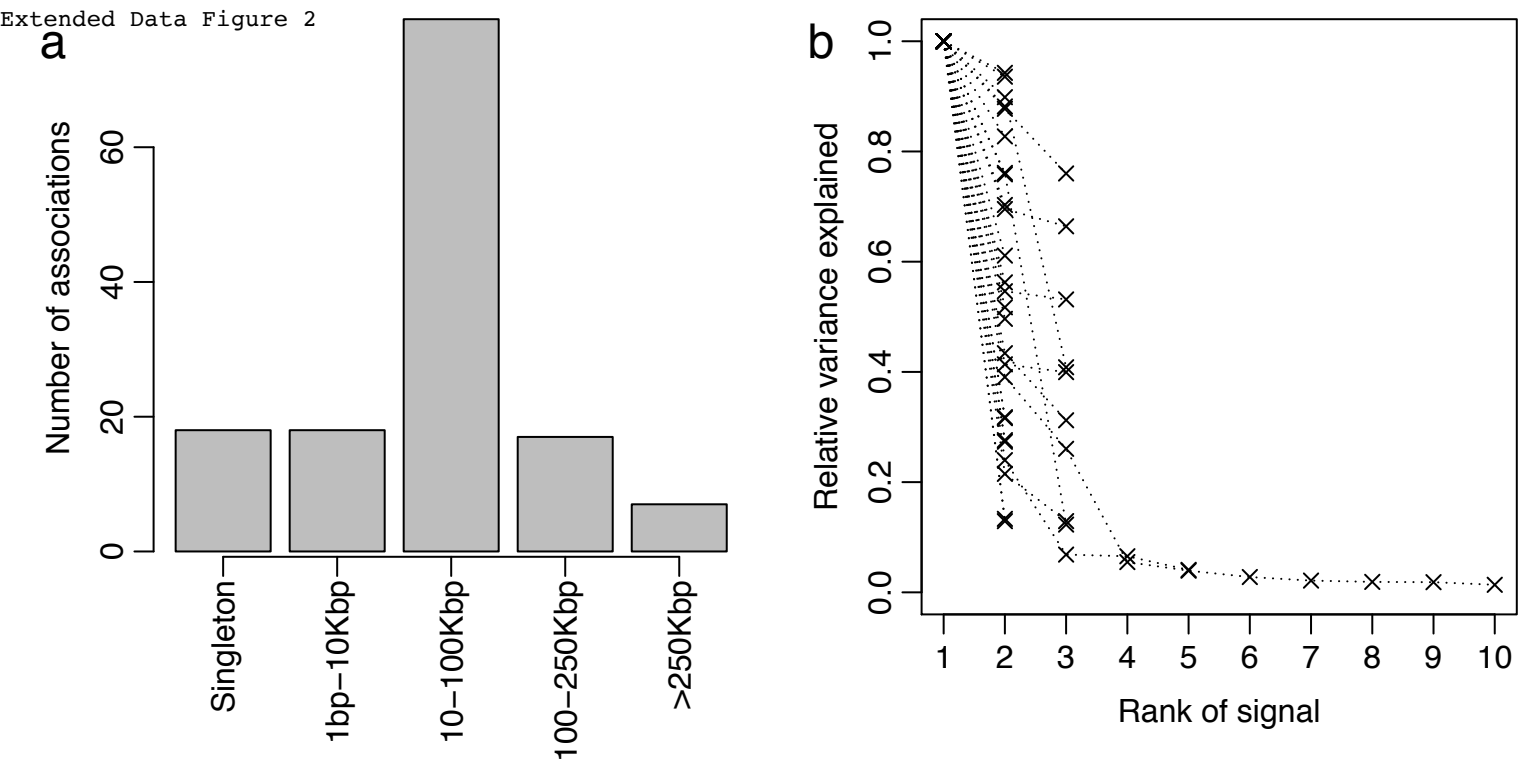
Figure 5


correlation coefficient (r): 0.52
Number of variants using fine−mapping
100
10
1
1
10
100
Number of variants using R2 > 0.6

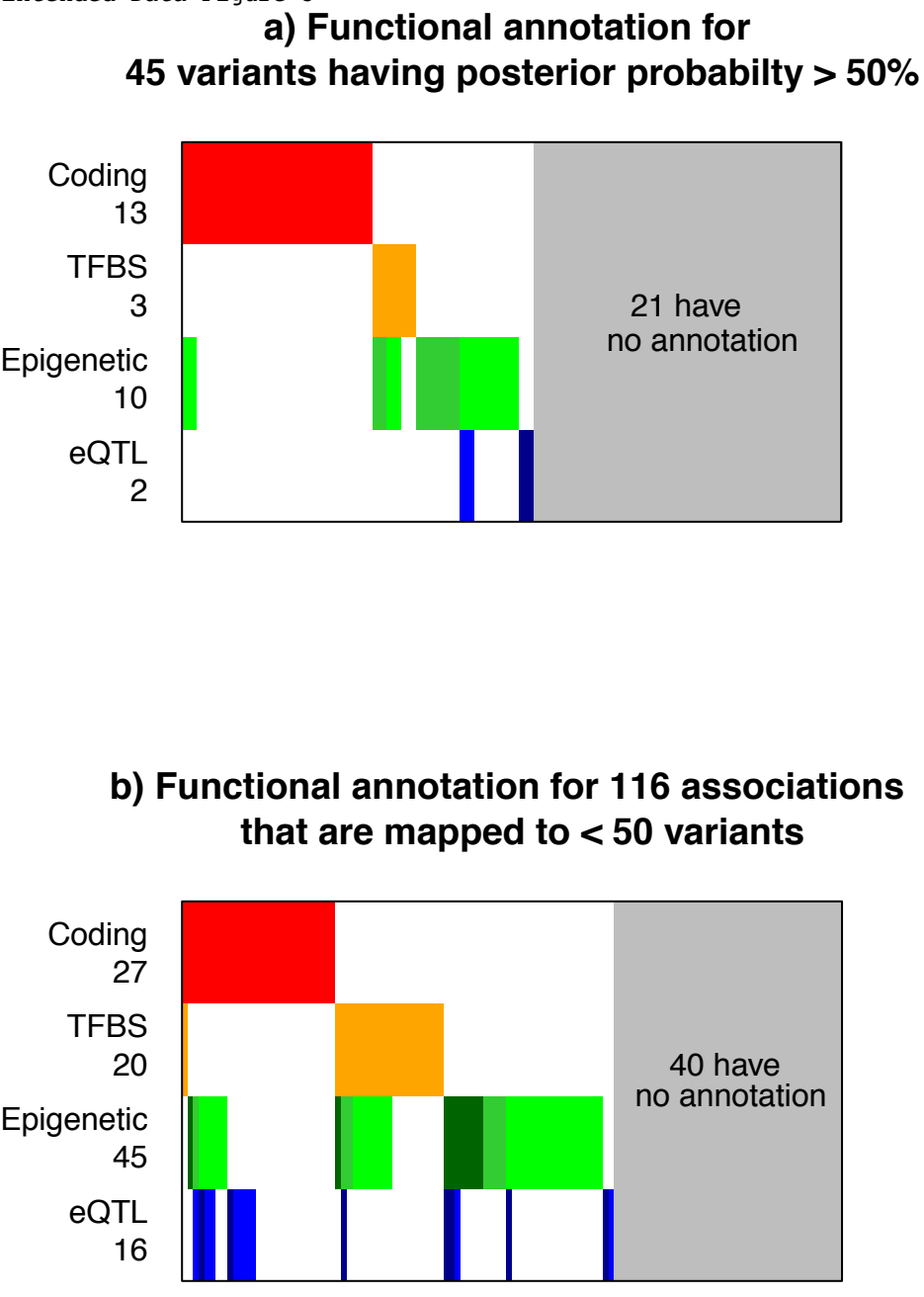
Extended Data Figure 1



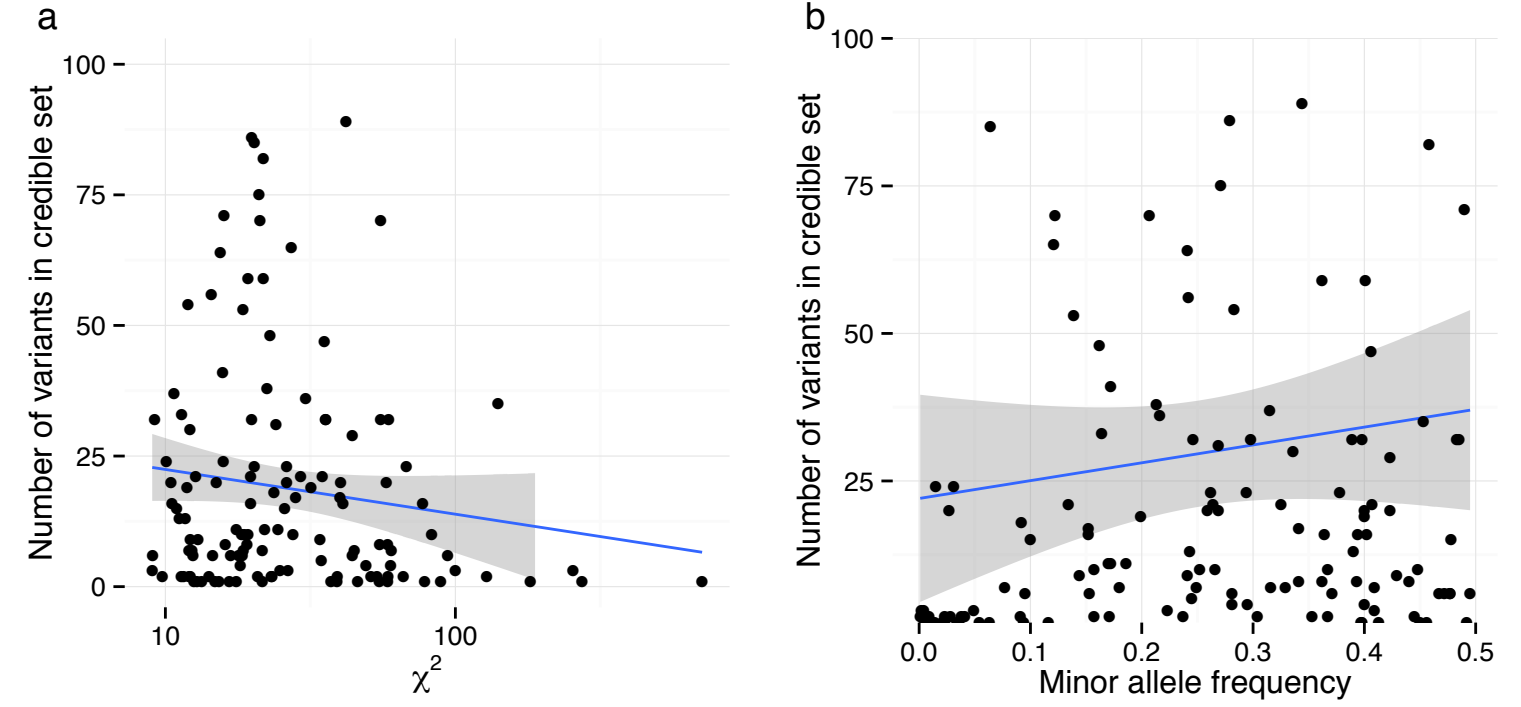
Extended Data Figure 2

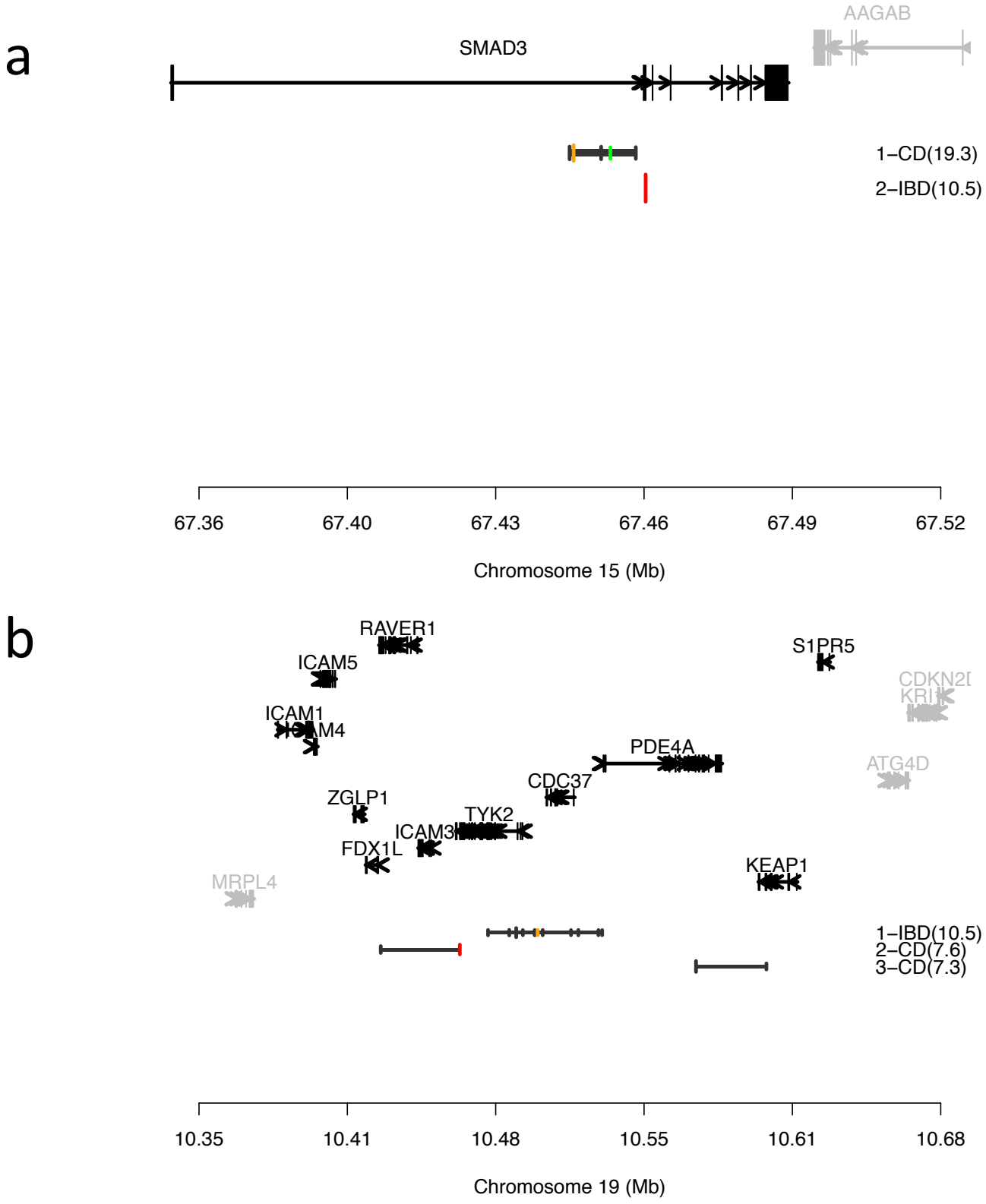


Extended Data Figure 3

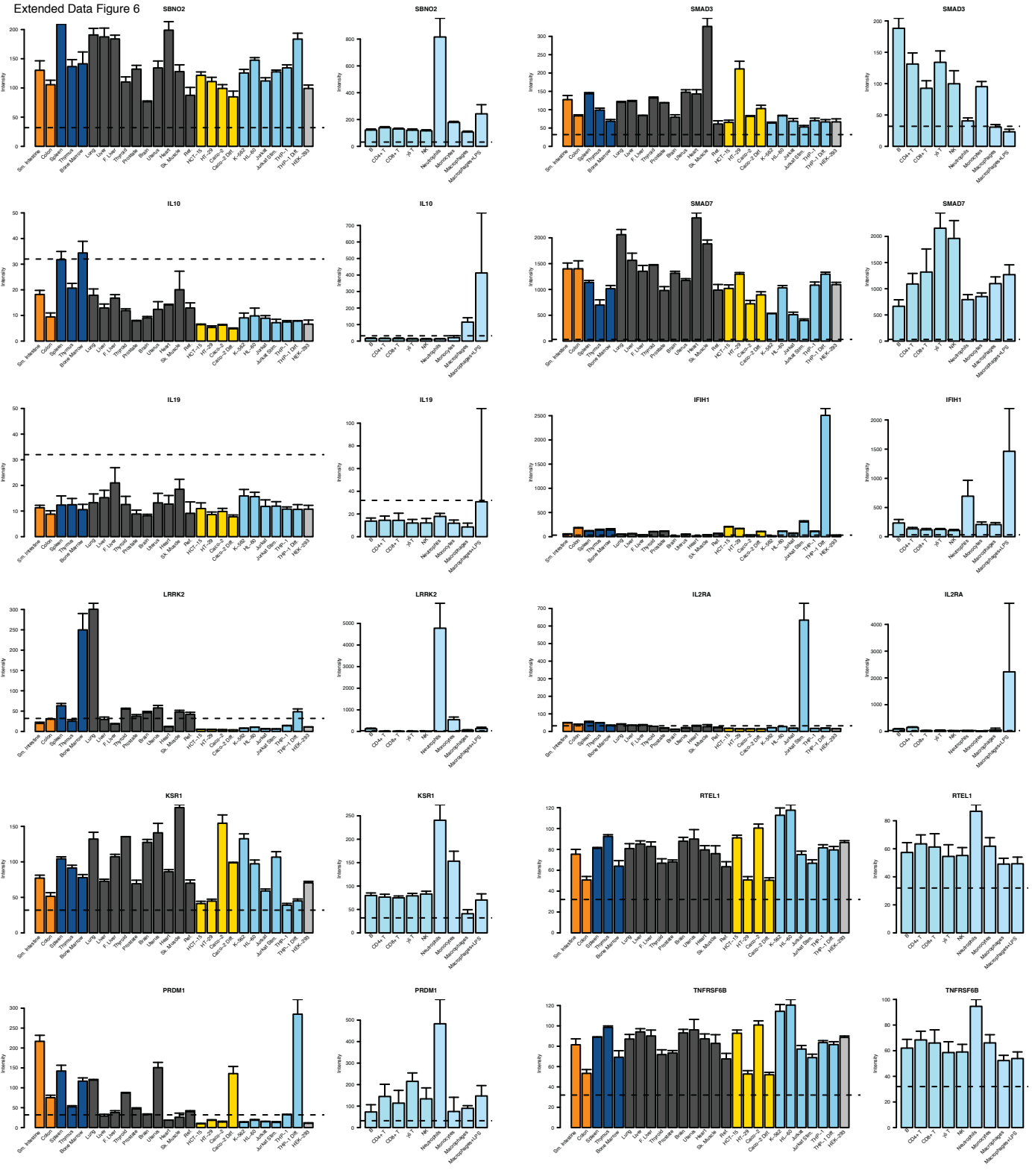


Extended Data Figure 4





Extended Data Figure 6



dataset	method	observed	Expected	p-value
CD4	**Naïve**	**3**	**0.4**	**0.007**
CD8	Naïve	1	0.3	0.296
CD14	Naïve	0	0.2	1.000
CD15	Naïve	1	0.2	0.199
CD19	Naïve	0	0.1	1.000
platelets	Naïve	0	0.0	1.000
Ileum	**Naïve**	**2**	**0.3**	**0.025**
colon	Naïve	1	0.2	0.206
rectum	Naïve	1	0.2	0.187
CD14 naïve	**Naïve**	**8**	**2.7**	**0.005**
CD14 IFN stimulated	Naïve	4	3.2	0.559
CD14 LPS 2h stimulated	Naïve	1	2.1	0.726
CD14 LPS 24h stimulated	Naïve	5	2.5	0.107
CD4	**Bayesian**	**4**	**1.0**	**0.010**
CD8	Bayesian	1	0.8	0.566
CD14	Bayesian	1	0.9	0.595
CD15	Bayesian	0	0.7	1.000
CD19	Bayesian	0	0.6	1.000
platelets	Bayesian	0	0.1	1.000
ileum	Bayesian	2	0.4	0.069
colon	**Bayesian**	**3**	**0.8**	**0.040**
rectum	Bayesian	2	0.6	0.124
CD4	**Permutation**	**6**	**1.9**	**0.013**
CD8	Permutation	3	1.5	0.186
CD14	Permutation	4	2.3	0.180
CD15	Permutation	1	1.8	0.863
CD19	Permutation	0	1.4	1.000
platelets	Permutation	0	0.1	1.000
ileum	**Permutation**	**4**	**1.1**	**0.018**
colon	Permutation	3	1.7	0.216
rectum	**Permutation**	**4**	**1.4**	**0.039**