European Journal of Human Genetics (2018) 26:265–274
https://doi.org/10.1038/s41431-017-0041-y

ARTICLE

Bayesian analysis of genome-wide inflammatory bowel disease data sets reveals new risk loci

Yu Zhang[1] · Lifeng Tian[2] · Patrick Sleiman[3] · Soumitra Ghosh[4] · Hakon Hakonarson[5] On behalf of the International IBD Genetics Consortium

Received: 6 May 2017 / Revised: 17 October 2017 / Accepted: 31 October 2017 / Published online: 4 December 2017

Abstract

Genome-wide association studies (GWAS) have identified and validated 200 risk loci for inflammatory bowel disease (IBD) to date, including risk loci for both Crohn's disease and ulcerative colitis. Previously studies mainly used single SNP testing methods and only reported the most significant association within each locus. Advanced methods are needed to detect additional joint effects of multiple SNPs and fine map causal variants in presence of strong linkage disequilibrium. In this study, we applied a powerful Bayesian method to analyze an existing Immunochip data sets for IBD from the international inflammatory bowel disease genetics consortium. The method jointly tested single and set-based SNPs in a unified framework and filtered indirect associations due to linkage disequilibrium, thereby fine-mapping the most likely IBD variants. Using an independent collection of individuals from 11 IBD GWAS as validation, our approach discovered and validated 9 completely new IBD loci and 5 independent signals (excluding the major histocompatibility complex) near known IBD loci reaching genome-wide significance. Several of the replicated new loci implicated functionally more interpretable genes than previous reports. The epigenetic marks at our detected IBD signals demonstrated significant activation signatures in blood cell types and correspondingly substantial repression in stem cells, suggesting regulatory links between genetic variants and IBD. Our analysis of the currently largest IBD datasets therefore added new insights that will be integral to the ongoing efforts in IBD genetics.

Introduction

Inflammatory bowel disease (IBD) in its two major forms, Crohn's disease (CD) and ulcerative colitis (UC), is a complex disease significantly affecting people of European origins and has increasing incidence in other populations recently [1, 2]. Over the past decades, genome-wide association studies (GWAS) on CD and UC have led to major discoveries of genes and loci in the human genome affecting the disease risks [3–8]. There is a substantial overlap between the genetic loci of CD and UC, suggesting that both types of IBD share common biological pathways. Genome-wide meta-analyzes and imputation methods have therefore been used to combine samples from both subtypes of IBD to increase power. Most recently, the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) combined more than 86,640 European individuals and 9846 non-European individuals to unravel a total of 231 genome-wide significant IBD SNPs (including CD and UC) [8]. Merging those SNPs within 100 kb together yielded a total of 200 IBD loci.

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41431-017-0041-y) contains supplementary material, which is available to authorized users.

✉ Yu Zhang
yzz2@psu.edu

[1] Department of Statistics, The Pennsylvania State University, University Park 16802, USA

[2] Bioinformatics Scientist, Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[3] Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[4] Target Sciences, GlaxoSmithKline, King of Prussia, PA 19406, USA

[5] Department of Pediatrics, Center for Applied Genomics, Devision of Human Genetics, The Children's Hospital of Philadelphia, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

We reanalyzed the Immunochip data from IIBGDC. All the samples have European ancestry and were obtained from Jostins et al. [6]. The data we used has been pre-processed, comprising 139,184 SNPs and small indels in 18,458 CD patients, 14,373 UC patients, and 464 patients with undetermined IBD subtypes, and 33,948 controls. The SNPs and small indels were selected primarily based on GWAS analysis of 12 autoimmune and inflammatory diseases, including CD and UC [6]. The Immunochip data therefore could reveal many insights into the shared genetic susceptibility across multiple immune-mediated diseases. We applied a novel Bayesian method, BEAM3 [9, 10], to detect both main effects of individual SNPs and joint effects of multiple SNPs simultaneously. We adjusted for sample stratification captured by the principal-component analysis (PCA), and we obtained estimates of statistical significance of our results by a permutation procedure that preserved the sample stratification information, thereby accounting for any residual inflation missed by PCA. After making discoveries in the Immunochip data set, we then used independent samples from 11 GWAS studies from IIBDGC and additional individuals genotyped by HiSeq to validate the findings.

Results

Detection and replication of new IBD loci

We first mapped IBD variants in the CD and UC combined samples vs. controls and identified 374 lead SNPs reaching Immunochip-wide significance at 0.02 level (Supporting Information). We next analyzed CD and UC samples separately and obtained 196 lead SNPs reaching Immunochip-wide significance at 0.015 level for CD and UC, respectively (Supporting Information). Combining the results, we obtained a total of 504 lead SNPs for IBD, CD, or UC at an overall Immunichip-wide significance of 0.05. These lead SNPs were clustered into 186 loci by grouping within 100 kb via minimum distance. We recaptured 442 lead SNPs in 155 loci that either lied within the most recent compilation of known IBD loci [8] or newly confirmed by a most recent report on chronic inflammatory diseases [11]. The remaining 62 lead SNPs in 31 loci were located outside of any known regions (see Methods and Supporting Information).

We used two independent data sets to replicate the 31 new IBD loci. The first data set included 13,240 individuals genotyped by HiSeq at the Children's Hospital of Philadelphia [12], and the second data set contained 24,350 individuals accumulated from 11 GWAS studies. We used IMPUTE2 [13] to first infer missing SNPs in all replication data sets, including the proxy SNPs within 100 kb and having LD $r^2 > 0.2$ with the lead SNPs. We used EUR samples in the 1000 Genomes release 3 [14] as the reference. At each locus, we ran BEAM3 on the lead SNPs and proxy SNPs in each replication data separately. We used conditional permutation to obtain a locus-wise p-value. Fisher's method was used to combine the p-values from the 11 GWAS data sets to generate a GWAS meta p-value. We then applied Fisher's method again to calculate an overall locus-wise p-value combining the HiSeq and GWAS meta p-values. We finally obtained false discovery rates (FDR) from the overall locus-wise p-values for the 31 IBD loci.

At FDR 0.05, 21 out of 31 novel IBD loci were replicated (Table 1). Among the 21 replicated loci, 9 loci implicated entirely new risk regions (Table 1a), and 12 loci suggested potentially new signals near known IBD loci (<1 Mb) showing some levels of LD ($r^2 > 0.2$) with known lead SNPs (Table 1b). These latter loci may be implicating new locations or genes, as they were at least 100 kb away from the previous IBD loci and were indicated as new signals by the BEAM3 method, which uses a joint probabilistic model of all variants to identify direct IBD associations via conditional test. On the other hand, we cannot completely rule out the possibility that our new signals were still just tagging the same (unobserved) risk variants that are also tagged by known variants, for which additional investigation is needed. For example, 7 out of the 12 loci in Table 1b lied in the major histocompatibility complex (MHC), which have extended and complex LD structures in the human population. Interpretation of these MHC signals therefore must be done with caution. Finally, there could be several reasons for why we failed to replicate 10 out of the 31 novel loci in the replication data. First, these loci may have genetic heterogeneity and/or are involved in epistasis, such that when tested by a linear model used in BEAM3, their significance could vary substantially. Secondly, there was substantial genotyping heterogeneity among the data sets we analyzed, for which we used imputation to match SNPs between data sets, and hence lost power. Thirdly, there could be uncontrolled confounding factors in our analysis that have affected the significance differently in different data sets. Finally, we note that not all significant results are meant to be replicated simply because of randomness in the sample. Thus, the 10 unreplicated loci should not be simply taken as all false positives.

We estimated and compared the relative risks of the replicated new IBD signals between the discovery data and the replication data. Figure 1a shows that, for those lead SNPs with nominally significant effects in both discovery data and replication data, their effects directions are always consistent. Of the 52 imputable lead SNPs (in 21 loci), thirty (in 11 loci) had nominally significant risks (p-value < 0.05) in the replication data, six

Table 1 Replicated new IBD loci

ID[a]	IBD Loci (hg19)	Lead SNPs[b]	Rep_FDR	Candidate genes[c]
a) Entirely new loci not reported by previous studies				
1[d]	chr2:181.95–181.96	**rs10176421G > T**, rs6759130G > T	1.94E−04	*AC068196.1;**ITGA4***
2[e]	chr3:57.77	rs7649133A > G	1.41E−03	*SLMAP;**FLNB***
3	chr9:0.24	rs661356A > G	1.66E−03	***DOCK8***
4	chr9:102.37	rs12237953C > T	1.94E−04	*RP11-554F20.1;**NR4A3***
5	chr11:23.28–23.28	**rs10834005A > C, rs1564625A > G**	1.66E−03	;
6	chr13:42.84–42.88	**rs927542A > G**, rs746447A > G, rs9645984C > T, rs1449509A > G, rs61959439C > T, rs59449023A > T, rs17521586C/T	1.45E−03	*AKAP11;**TNFSF11***
7	chr14:35.39–35.4	**rs712303C > T**, rs1712349C > T	1.36E−02	*RP11-85K15.2;Y_RNA;**BAZ1A***
8	chr15:77.38	rs16968665C > G	3.43E−04	*TSPAN3*
9	chr20:39.91	rs6093462A > C	1.13E−04	*ZHX3;**PLCG1***
b) New signals implicating known loci previously reported for CD, UC or IBD				
10	chr1:155.13–155.21	rs4971079A > G, **rs4072037A > G**, rs3768566A > G, rs9628662G > T	1.94E−04	***MUC1**;GBA*
11	chr2:25.5–25.5	rs201014116A > C, **rs2006788G > T**	3.43E−04	*DNMT3A;**POMC***
12	chr2:43.46	rs13402621C > T	3.13E−04	*AC010883.5;THADA*; ***ZFP36L2***;
13	chr6:29.74–29.78	**rs1737041G > T**, rs885940A > G, **rs1633009A > T**, rs1610707G > T	1.01E−05	*HLA-V*
14[f]	chr6:29.93–29.94	**rs2517689A > G, rs2523946A > G**	1.73E-03	*HCG9*
15	chr6:30.38–30.43	rs9261859G > T, **rs11966619G > T**	3.54E-03	;
16	chr6:31.51–31.57	rs9368696A > G, **rs2736191C/G**, rs2515920A > T, rs9267512C > T	2.13E-03	*NCR3;LST1;LTB*
17	chr6:31.92	rs4151651A > G	3.38E-05	*CFB;C2;NELFE*
18[e]	chr6:32.21–32.21	rs439303C > T, rs9267947A > G, **rs411326A > G**	5.94E-06	;
19[e]	chr6:33.05–33.07	**rs1431403C > T**, rs3128927C > T	1.21E-03	*HLA-DPA1;HLA-DPB1*
20	chr16:10.97–11.09	**rs12928665A > G**, rs4781026C > G, rs16957807A > G, rs3813754A > T, rs56363812 A > G, rs4780343A > G, rs3881421C > T, rs8061306A > G	1.03E-03	***CIITA**;CLEC16A*
21[f]	chr17:38.82–38.83	rs7209404G > T, **rs9896791A > C**	1.94E-04	*AC073508.1;KRT222*

Lead SNPs reported by BEAM3 are grouped within 100 kb by minimum distance. Lead SNPs with the maximum posterior probability of association within each loci are marked in bold. Threshold for replication FDR is 0.05

[a]Parenthesis indicate loci where additional SNPs in a nearby known loci were also detected

[b]If a locus involves multiple lead SNPs, the lead SNP with the strongest association is highlighted in bold

[c]Candidate genes are determined as either within 5 kb of the lead SNPs or implicated by GO enrichment analysis (latter shown in bold)

[d]Loci detected in CD only

[e]Loci detected in UC only

[f]Loci detected in both CD and UC, respectively, but not in the combined subtypes (IBD)

(in 3 loci) had nominally significant risks in either CD or UC replication data, but not for IBD (Fig. 1b), and the remaining 16 lead SNPs (in 6 loci) had insignificant relative risks in replication data, which may be due to either epistasis or tagging some causal SNPs ungenotyped in the discovery data.

IBD variance better explained by our lead SNPs than by known lead SNPs

Combining the known and replicated new loci, we identified a total of 176 IBD loci containing 493 lead SNPs. Since BEAM3 fine-mapped causative SNPs by removing LD effects (without accounting for LD effects, there were 5438 SNPs with p-value < 1e−8), we were able to estimate the number of independent IBD SNPs within each locus. We estimated that 50.7% IBD lead SNPs (250 out of 493) in our result might directly contribute to the IBD risks (Supporting Information), which we referred to as "direct lead SNPs". The remaining 243 lead SNPs were alternative candidates, but we could not exclude with certainty that their association with IBD may be merely due to LD with the direct lead SNPs.

After removing sample stratification, the direct lead SNPs in our IBD loci (250 lead SNPs in 176 loci) explained 16.40% IBD variance (Fig. 2a), greater than the 15.26% explained by the 231 lead SNPs in 200 IBD loci from Liu

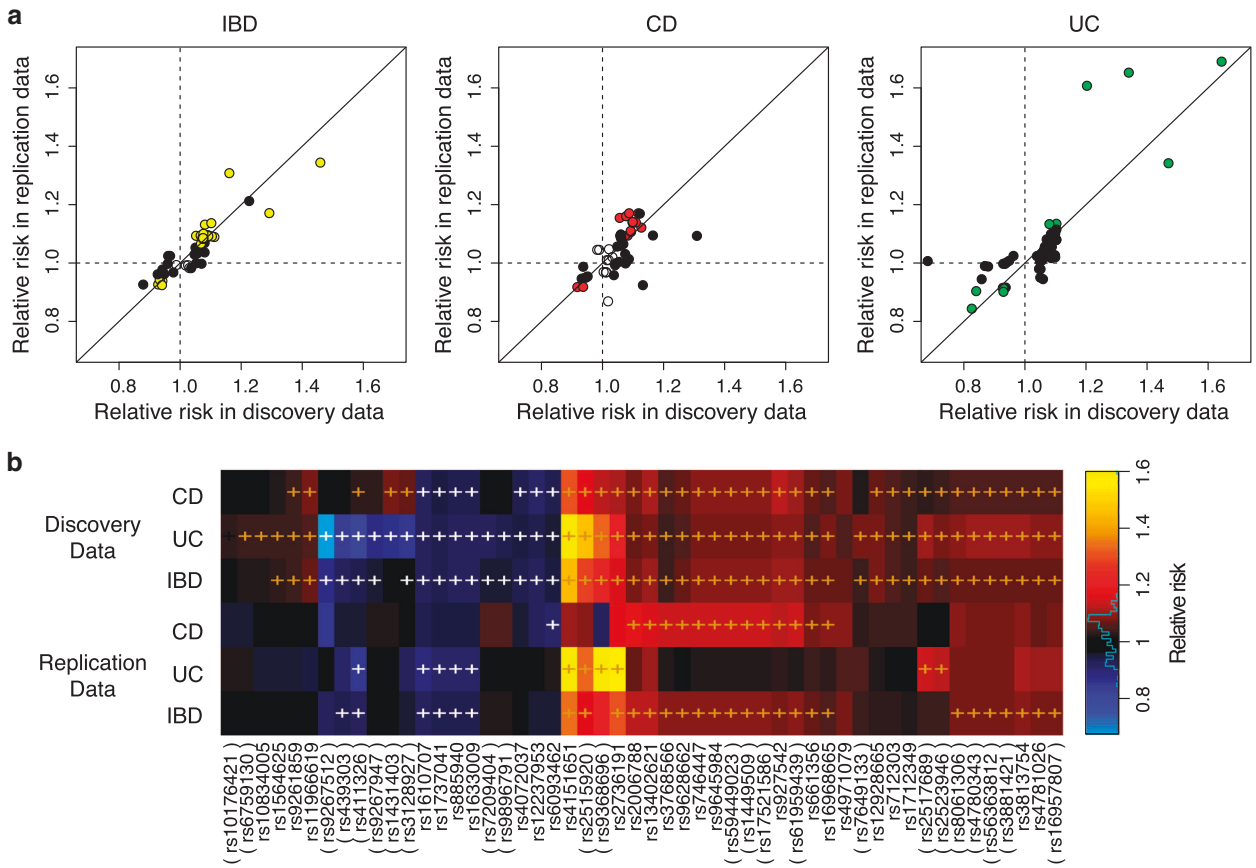

Fig. 1 Replication of relative risks. **a** Scatter plots show the relative risks of minor alleles of each lead SNP in the replicated new loci or new SNPs near known loci. Colored dots denote SNPs with significant marginal effects at 0.05 level in both discovery and replication data; solid black dots denote SNPs with significant marginal effects in discovery data but not in replication data; empty dots denote SNPs with insignificant marginal effects in discovery data. **b** The same relative risks of minor alleles of the lead SNPs in the replicated new loci shown in heatmap, comparing between CD, UC and subtypes combined (IBD). '+' marks the significant marginal effects at 0.05 level. SNP ID in parenthesis indicates lead SNPs for either CD or UC, but not for IBD, in the discovery data

et al. (2015), both on the same logistic scale. Comparing at the same level of explained IBD variance, only top 127 loci (200 direct lead SNPs) in our result were needed to explain the total amount of IBD variance explained by the previous IBD loci (Fig. 2a). In addition, using single best lead SNP per locus only explained 12.86% IBD variance, confirming that there could be multiple IBD variants in a locus contributing to the risks. Also, loci carrying multiple lead SNPs generally ranked higher in their contributions to IBD variance (Fig. 1b). On the other hand, further including the 243 alternative lead SNPs in our result only explained 1.24% more IBD variance (17.64% in total) than using the direct lead SNPs alone. Taken together, the IBD loci inferred by our method in this study better explained the disease variance than the previously reported IBD loci and lead SNPs did.

Enrichment of biological functions and pathways

To understand the potential biological functions involved in the 9 new IBD loci and the 12 independent IBD signals near known IBD loci, we performed enrichment analysis of gene functions and pathways via GREAT [15]. We removed loci in the MHC region (chr6:29–34 Mb) to avoid bias towards MHC. We identified the terms that were more significant after including our new loci than using the recaptured IBD loci alone. At the level of FDR ≤ 0.05, fold enrichment ≥5, and that including the new loci must reduce the FDR for the enriched term by at least 1 order of magnitude, we identified 102 significantly enriched terms that highlighted the potential function and pathway involvement of our new loci in seven categories of gene ontology. Figure 3 shows the top 15 most significant terms in each gene ontology category. The most significantly enriched terms with gains of significance by including our new loci were cytokine receptor binding ($p = 6.82e{-}08$) in *GO Molecular Functions* and cytokine-mediated signaling pathway ($p = 4.68e{-}13$) in *GO biological process*, which implicated novel genes *PLCG1* and *FLNB* that were not indicated in previous reports. There were also newly added significant terms in *GO biological process* that were not enriched using the

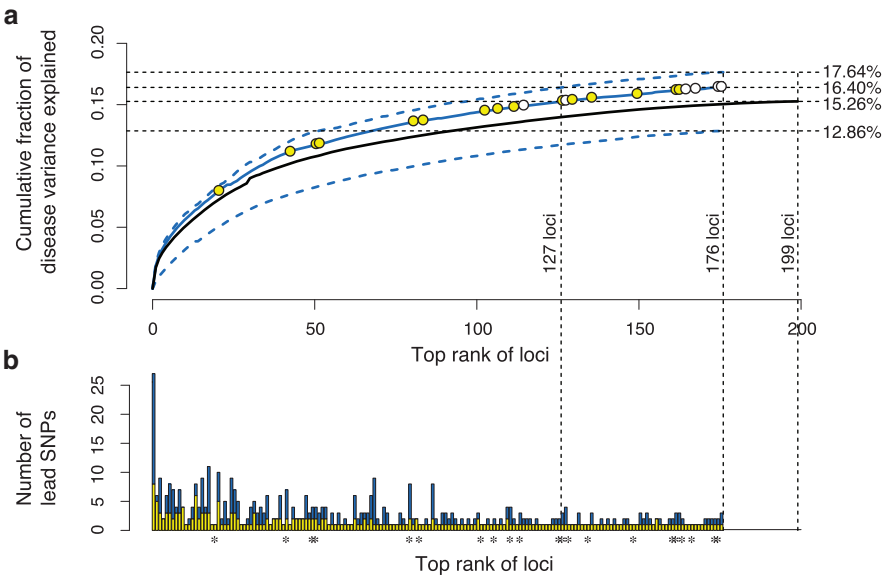

Fig. 2 Proportion of disease variance explained by the lead IBD SNPs. **a** Cumulative fractions (generalized r^2) of IBD variance explained by the lead SNPs after removing sample stratification. Solid blue line shows the fraction explained by the 250 direct lead SNPs in 176 loci detected by BEAM3. Upper dotted blue line shows the fraction explained by including all 504 lead SNPs detected by BEAM3. Lower dotted blue line shows the fraction explained by the single best lead SNPs in 176 loci. Yellow circles mark the replicated new IBD loci or IBD signals near known loci, and white circles mark the replicated new loci or new signals near known loci for either CD or UC, but not combined. Black solid line shows the fraction of IBD variance explained by the 231 previously reported lead SNPs in 200 known IBD loci. The corresponding r^2*s* for each model were shown on the right side. **b** Total number of lead SNPs and the estimated number of direct lead SNPs in each of the 176 loci detected by BEAM3. Stars indicate the replicated new loci or new signals near known loci

recaptured known IBD loci alone, including activation of MAPKK activity ($p = 1.85e{-}02$) and negative regulation of JAK-STAT cascade ($p = 2.79e{-}02$). In *mouse phenotype*, the most significantly enriched terms with gains of significance included abnormal cytokine secretion ($p = 3.31e{-}21$), abnormal interleukin secretion ($p = 1.32e{-}17$), abnormal T cell activation ($p = 4.99e{-}17$), and abnormal T cell proliferation ($p = 2.89e{-}16$), which implicated additional new genes *POMC, DOCK8*. In *disease ontology*, the most significantly enriched terms with grains of significance included autoimmune disease of gastrointestinal tract ($p = 4.65e{-}15$), demyelinating disease of central nervous system ($p = 4.77e{-}12$), multiple sclerosis ($p = 5.00e{-}12$), and demyelinating disease ($p = 8.33e{-}12$), which implicated genes *ITGA4* and *MUC1*. Finally, in *PANTHER pathway, MSigDB pathway and MSigDB immulogic signaturs*, the most significant terms with gains of significance involved JAK/STAT signaling pathway ($p = 7.27e{-}06$), genes involved in regulation of IFNG signaling ($p = 6.64e{-}08$), and genes down−regulated comparing unstimulated vs. stimulated dendritic cells ($p = 1.36e{-}07$), respectively, which implicated new genes *BAZ1A, NR4A3*.

Epigenetic enrichment patterns at IBD loci

We next evaluated the epigenetic signatures at the new IBD SNPs. We obtained data for 34 epigenetic marks in 127 human cell types from the RoadMap Epigenomics project [16]. Using non-IBD SNPs in the ImmunoChIP data as a reference, we found that the lead SNPs in our new loci tended to be associated with strong signals in most epigenetic marks in non-developmental cell types, particularly in the cluster of differentiation surfaced cells such as B lymphocytes and T lymphocytes. After removing mark effects (averaged over all cell types) and cell type effects (averaged over all marks), we observed a strong enrichment pattern in the residual signals, which represented cell type and mark specific effects (Fig. 4a). One group of cell types (group 1 in Fig. 4a) showed enriched marks for transcription and enhancer activities (RNA-seq, H3K4me1, H3K4me2, and H3K27ac) and depleted marks for repression (H3K27me3, DNA Methylation, and H3K9me3), and the group contained mostly B and T cells. Another group (group 2 in Fig. 4a) showed complementary patterns of depleted active marks and enriched repressive marks, where the group contained mainly primary and derived embryonic stem cells. These patterns were similarly observed at the known IBD loci, suggesting a consistent and unique cell type and mark specific epigenetic signature that distinguishes between IBD and non-IBD loci. The anatomy of cell types further revealed significant enrichment of blood cells, stem cells, and brain tissues (Fig. 4b).

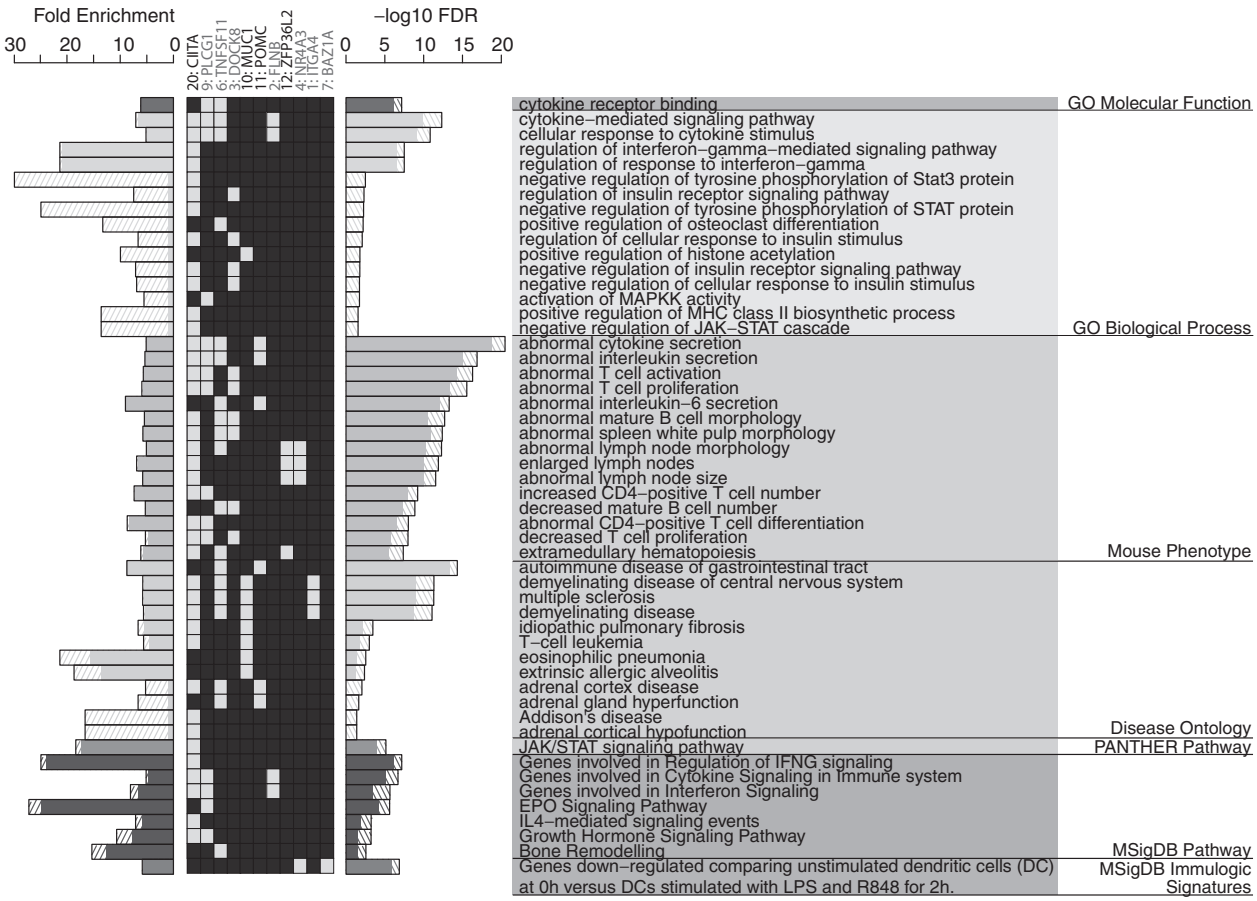

Fig. 3 Gene function and pathway enrichment analysis for the replicated new IBD loci. Solid bars denote enrichments obtained from the recaptured known IBD loci. Bars in shaded lines denote additional enrichments due to inclusion of our replicated new loci. Enriched terms are ranked by FDR within each category, and only the top 15 most significant terms in each category are shown. Yellow and blue boxes in the center show the involvement (yellow boxes) of each locus in the enriched terms. Loci not enriched in any of the shown terms are removed. Best candidate genes implicated in the enriched terms for each locus (new loci or new signals at known loci), and the locus index, are shown on the top. New loci are marked by asteroid. MHC region (chr6:29–34 Mb) is removed from this analysis

Discussion

While we recaptured a large portion (77.4%) of known IBD loci in this study, 55 (12.5%) lead SNPs that overlapped with the known IBD loci were mapped to locations at least 100 kb away from the previously reported SNPs, and some of them implicated different genes. For instance, Liu et al. [8] reported rs12946510C > T at chr17:37.91 Mb in hg19 as a lead SNP, whereas our method reported rs4795397A > G at chr17:38.02 Mb in hg19, which was 110 kb away from rs12946510. While rs12946510 is located near the transcription end site of gene *IKZF3*, rs4795397 inferred by our method is located at the transcription start site of *IKZF3*. Liu et al. reported rs3197999C > T at chr3:49.72 Mb in hg19 as a lead SNP near genes *MST1* and *APEH*, whereas our method reported rs35261698C > G at chr3:49.54 Mb in hg19 near gene *DAG1*. GO enrichment analysis by GREAT [15] showed that *DAG1* appeared 93 times in the significantly enriched terms at FDR 0.05. In contrast, both *MST1* and *APEH* did not appear in any significantly enriched terms.

We further performed GO enrichment analysis using our 176 IBD loci and compared with the 200 known IBD loci. At FDR 0.05 and fold enrichment ≥5, we identified 337 significant terms that had at least one magnitude stronger significance at our loci than the known loci. Among these, 183 terms were completely new (i.e., not enriched in the known IBD loci). The most significant new term was *regulation of tyrosine phosphorylation of Stat3 protein* in GO Biological Process, with FDR 1.40e−05. In comparison, we obtained 210 significant terms that had at least one magnitude stronger significance at the known IBD loci than ours. Among these, 128 new terms were completely new, and the most significant term was *decreased IgG2a level* in Mouse Phenotype, with FDR 5.46e−08. While we found the distinctly enriched terms were relevant

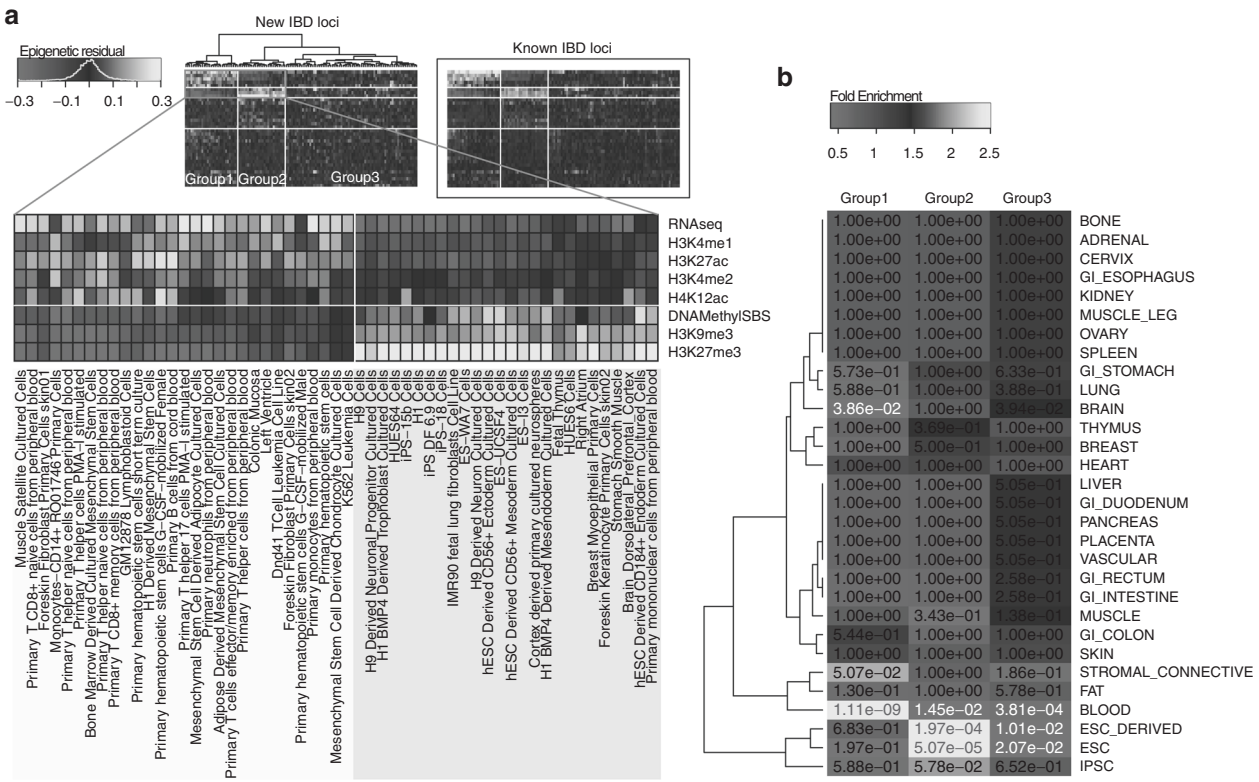

Fig. 4 Enrichment pattern of epigenetic marks. Using 34 epigenetic marks in 127 epigenomes from RoadMap Epigenomics, we calculated mean signals (in log scale) of each mark in each cell type at our detected IBD lead SNPs. We adjusted for SNP density and subtracted mean signals of corresponding marks and cell types at all non-IBD SNPs (see Methods). We further removed mark effects and cell type effects to obtain residual signals. **a** Heatmap of the residual signals at the lead SNPs in the 22 novel IBD loci or new signals near known loci. We observed a clear enrichment pattern that separated the 127 epigenomes into three groups, with the first two groups and a subset of epigenetic marks showing in the zoomed view. Heatmap of the residual signals at known IBD loci is shown in the boxed area for comparison. **b** Fold enrichment of cell type anatomy in the three epigenome groups marked in (**a**), with fisher's exact *p*-values for enrichment showing in the heatmap. Significant *p*-values for enrichment and depletion of cell type anatomy are shown in red and white colors, respectively

to IBD in both sets of loci, our loci implicated more interesting terms than those by the known loci.

Our replication study was locus based, that is, using the implicated lead SNPs and its proxy SNPs within each locus. Given that current statistical methods cannot accurately pinpoint causal variants, and the potential confounding effects from epistasis, genetic heterogeneity and sample heterogeneity, our locus-based replication is more appropriate and powerful than replicating the exact SNPs. Particularly for epistasis, while they may be detected in the discovery data, direct replication of their effects in genetic data alone remains a challenging problem, which may require validation at the functional level [17]. These difficulties highlight the importance and usefulness of combining additional information, such as epigenetic signals and functional annotations of genetic variants, to improve the power for mapping causal variants.

In summary, we have presented evidence of new IBD loci following a powerful reanalysis of the Immunochip IBD data from IIBDGC. The new loci consolidate evidence for the cytokine-based pathways that were also prevalent from previous reports. Our analysis implicated new loci such as a DNA demethyl transferase that could have profound effects at a number of distal genes and reveal new pathways to target for IBD. The statistical methods used in this study have played a pivotal role for detecting the new IBD loci and locating the novel lead SNPs within known IBD loci. Without using these advanced methods, only the SNPs showing the strongest signal would be selected, while missing potential epistasis and additional risk variants buried in the same locus.

Materials and methods

Processing of the immunochip data

We obtained the Immunochip genotype data set from the IIBDGC. The data set has already been cleaned (version 5) by the IIBDGC including 68,427 individuals, of which 1184 individuals had no case control status and thus were

removed from the study. After removing non-polymorphic SNPs, SNPs with ≥2% missing genotypes, and SNPs violating Hardy-Weinberg equilibrium (HWE) at p-value < 1e −10, the data set had 139,184 SNPs. We imputed missing genotypes (<2% missingness) by randomly sampling from the observed genotypes at the same SNPs. We used the first four PCs of the Immunochip samples from the IIBDGC to adjust for sample stratification.

Processing of the HiSeq replication data

The HiSeq data set contained 535,931 SNPs genotyped in 2846 cases and 11,104 controls that were independent of the Immunochip samples. We used PLINK [18] to identify closely related individuals in this data set, as measured by the proportion of identity-by-descent between pairs of individuals. We removed 710 individuals whose proportion of identity-by-descent was >0.4, which is the threshold used by Jostins et al. [6]. We also performed PCA using all SNPs and selected the first two PCs as covariates to remove sample stratification.

Processing of the GWAS data

We obtained the GWAS data sets from 13 different studies from the IIBDGC, including 6 studies for CD and 7 studies for UC, genotyped at 11 genotyping centers. We removed individuals who were related to the Immunochip samples with estimated proportions of identity-be-descent >0.4 by PLINK. We then performed PCA using all genotyped SNPs and selected the first PC in each GWAS data set to adjust for sample stratification.

Bayesian multi-locus association mapping

The BEAM3 method [9, 10] is a Bayesian partition method that detects both single-locus and multi-locus association for both common and rare variants. The BEAM3 method is an extension of the original BEAM method [19] that simultaneously achieves three goals. First, it detects multi-locus joint association. Secondly, it removes LD effects such that a set of SNPs will be identified only if it is directly associated with the disease given all other disease SNPs. Thirdly, it is computationally efficient for genome-wide studies because it implicitly models the non-disease SNPs, which includes most of the SNPs in a GWAS. BEAM3 can adjust for sample stratification by including PCs of samples as covariates.

Analysis of the immunochip data

We ran BEAM3 on the Immunochip data by first assigning SNPs into SNP set. We assigned all SNPs whose minor allele frequency (MAF) > 0.05% into its own set (i.e., a single SNP set). For SNPs with MAF < 1%, we then created multi-SNP sets containing five consecutive rare variants per set. We further created multi-SNP sets containing 30 consecutive rare variants per set. As a result, the SNPs with MAF between 0.05 and <1% were tested for both single SNP effects and group effects jointly with other variants. We ran BEAM3 for 100 iterations. The prior probability for each SNP set to be associated with the disease was set at 1e −4, and we identified significant IBD SNPs at a threshold of posterior probability ≥0.1, which corresponded to Immunochip-wide significance of 0.02, estimated by 1000 conditional permutations. The runtime of BEAM3 algorithm is proportional to the product of the number of detectable SNP sets associated with the disease and the total number of SNP sets. In the Immunochip data, there were a few hundreds of IBD associated SNP sets, for which BEAM3 could take a long time to complete. To reduce computing time, we ran BEAM3 on the SNP sets in each chromosome separately, that is, not considering trans-epistasis associations. As a result, BEAM3 finished the analysis in 1 h on a single 2.4 GHz CPU.

Analysis of HiSeq data

For each locus to be replicated in the HiSeq data, we first identified the proxy SNPs of all lead SNPs within the locus. We used SNAP [20] to identify proxy SNPs within 100 kb and in LD ($r^2 > 0.2$) with each lead SNP. We extracted all lead SNPs and their proxy SNPs in the locus, and we imputed missing genotypes by IMPUTE2 [13] using EUR individuals in 1000 Genomes release 3 [14] as reference. We ran BEAM3 on the extracted and imputed data and included the first two PCs for HiSeq samples as covariates. We specified the prior probability of disease-association at 0.03. We calculated the sum of posterior probabilities over all lead and proxy SNPs in each locus and obtained a locus-wise p-value by 1000 conditional permutations.

Meta-analysis of 11 GWAS data sets

We excluded two GWAS for UC (nid3 and norw) from the meta-analysis as they had less than 30 cases. For the remaining 11 GWAS data sets, we identified the proxy SNPs of all the lead SNPs within each locus using SNAP. We extracted the lead SNPs and proxy SNPs and imputed missing SNPs by IMPUTE2 using EUR samples in 1000 Genomes release 3 as reference. We ran BEAM3 in each GWAS data set separately and used the first PC of each GWAS as covariate. We obtained locus-wise p-values in the same way as we did in HiSeq data, and we combined the

locus-wise p-values of the 11 GWAS together using the Fisher's method.

Conditional permutation

To account for sample stratification in a permutation study, we used logistic regression to predict and simulate disease status from the PCs of samples. This created randomized disease labels that were correlated with the PCs in the same way as that of the original disease status. Conditioning on the PCS, the randomized disease status was independent of the genotypes.

Explaining IBD variance

We fitted logistic regression models for the IBD status using both the lead SNPs and the PCs of samples as predictors. We calculated generalized r^2 by $1-(L_1/L_0)^{2/n}$, where L_1 and L_0 denote the likelihoods of the alternative and the null model, respectively, and n denotes the sample size.

Estimating the number of independent IBD variants

We estimated the number of independent IBD variants in a locus by the sum of posterior probabilities of association over all SNPs within a locus and we rounded the number up to its nearest integer if its decimal point was ≥0.1.

Estimating relative risks

The relative risks of SNPs were estimated by logistic regression while using PCs as covariates. The relative risk for a SNP is given by the exponential of the SNP coefficient minus 1. We merged the 11 GWAS data sets and the HiSeq data together using the "merge" function in PLINK. When fitting logistic regression models on the merged data, we added an indicator vector for each data set as covariates to adjust for potential inconsistency between data sets.

Enrichment analysis of gene functions and pathways

We generated three sets of loci: (1) our new and replicated IBD loci; (2) the known IBD loci recaptured in this study; and (3) all other loci in ImmunoChIP data that were at least 100 kb away from the known and novel IBD loci. A locus was defined by clustering SNPs within 100 kb using the minimum distance method. We removed loci in MHC region (chr6:29–34 Mb). We ran GREAT in default setting (version 3.0.0) twice: (1) we used the set2 loci against all three sets of loci combined to calculate GO enrichment in the recaptured known IBD loci relative to all loci in the Immunochip regions; and (2) we used set1 and set2 loci combined against all three sets of loci to calculate the GO enrichment in both new and known IBD loci identified in this study, relative to all loci in the Immunochip regions.

Analysis of RoadMap epigenomics data

We downloaded the uniformly processed chromatin marks from the RoadMap Epigenomics project [16]. We used the UCSC utility tool to obtain epigenetic signals at each SNP position. We then took $\log(x+0.01)$ transformation to reduce data skewness. To remove estimation bias due to SNP density, we grouped SNPs in each set by the 100 kb minimum distance method. We modeled the mean signal of each mark i in each cell type j in each SNP set k (=1, 2, 3 defined above), denoted by Y_{ijk}, as $Y_{ijk}=\alpha+\beta+\gamma_j+\varepsilon_{ijk}$. That is, Y_{ijk} is the sum of an overall effect (α_k) of the SNP set, plus epigenetic mark effect (β_{ik}, with mean 0), plus cell type effect (γ_{jk}, with mean 0), and a residual term (ε_{ijk}, with mean 0), where ε_{ijk} were the residual epigenetic signals shown in Fig. 4. We obtained the anatomy information of each cell type from the RoadMap Epigenomics website, and we performed Fisher's exact test to evaluate the significance of each anatomy enriched in each group of cell type cluster.

Data availability

The immunochip data and the GWAS data that support the findings of this study are obtained as described in Jostins et al. [6]. The HiSeq data analyzed in this study are obtained from Imielinski et al. [12], but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publically available. The HiSeq data are however available from its original authors upon reasonable request and with permission of IIBDGC. The epigenetics data analyzed in this study are available in the Roadmap Epigenomics data portal, http://egg2.wustl.edu/roadmap/web_portal/.

Acknowledgements Y.Z. is supported by National Institutes of Health grant R01-HG004718.

Author contributions Y.Z., S.G., and H.H. conceived and designed the study. Y.Z. carried out the analysis. T.L. and P.S. helped in the data processing and analysis. Y.Z., S.G., and H.H. wrote the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

1. Molodecky NA, Soon IS, Rabi DM, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012;142:46–54.
2. Ng SC, Tang W, Ching JY, et al. Incidence and phenotype of inflammatory bowel disease based on results from the Asia-Pacific Crohn's and Colitis Epidemiology Study. Gastroenterology 2013;145:158–65.
3. Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–62.
4. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 2010;42:1118–25.
5. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 2011;43:246–52.
6. Jostins L, Ripke S, Weersma RK, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119–24.
7. Goyette P, Boucher G, Mallon D, 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;47:172–79.
8. Liu JZ, van Sommeren S, Huang H, et al. Association analyzes identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 2015;47:979–86.
9. Zhang Y. A novel bayesian graphical model for genome-wide multi-SNP association mapping. Genet Epidemiol 2012;36:36–47.
10. Zhang Y, Ghosh S, Hakonarson H. Dynamic Bayesian testing of sets of variants in complex diseases. Genetics 2014;198:867–78.
11. Ellinghaus D, Jostins L, Spain SL, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 2016;48:510–18.
12. Imielinski M, Baldassano RN, Griffiths A, et al. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 2009;41:1335–40.
13. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44:955–59.
14. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
15. McLean CY, Bristor D, Hiller M, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 2010;28:495–501.
16. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.
17. Hemani G, Shakhbazov K, Westra HJ, et al. Detection and replication of epistasis influencing transcription in humans. Nature 2014;508:249–53.
18. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 2007;81:559–75.
19. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case–control studies. Nat Genet 2007;39:1167–73.
20. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938–39.