Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population

Manuel A. Rivas[1,2], Jukka Koskela[1,3], Hailiang Huang[1,3], Christine Stevens[1], Brandon E. Avila[1,3], Talin Haritunians[4], Benjamin M. Neale[1,3], Mitja Kurki[1,3], Andrea Ganna[1,3], Daniel Graham[1], Benjamin Glaser[5], Inga Peter[6], Gil Atzmon[7,8], Nir Barzilai[7], Adam P. Levine[9], Elena Schiff[9], Nikolas Pontikos[9,10], Ben Weisburd[1,3], Konrad J. Karczewski[1,3], Eric V. Minikel[1,3], Britt-Sabina Petersen[11], Laurent Beaugerie[12], Philippe Seksik[12], Jacques Cosnes[12], Stefan Schreiber[13], Bernd Bokemeyer[14], Johannes Bethge[13], NIDDK IBD Genetics consortium, T2D-GENES consortium, Graham Heap[15], Tariq Ahmad[16], Vincent Plagnol[10], Anthony W. Segal[9], Stephan Targan[4], Dan Turner[17], Paivi Saavalainen[18], Martti Farkkila[19], Kimmo Kontula[20], Matti Pirinen[21,22], Aarno Palotie[1,21,23], Steven R. Brant[24,25], Richard H. Duerr[26,27], Mark S. Silverberg[28], John D. Rioux[29,30], Rinse K. Weersma[31], Andre Franke[11], Daniel G. MacArthur[1,3], Chaim Jalas[32], Harry Sokol[12], Ramnik J. Xavier[1,33], Ann Pulver[34], Judy H. Cho[35], Dermot P.B. McGovern[4], Mark J. Daly[1,3]

Abstract: As part of a broader collaborative network of exome sequencing studies, we developed a jointly called data set of 5,685 Ashkenazi Jewish exomes. We make publicly available a resource of site and allele frequencies, which should serve as a reference for medical genetics in the Ashkenazim. We estimate that 30% of protein-coding alleles present in the Ashkenazi Jewish population at frequencies greater than 0.2% are significantly more frequent (mean 7.6-fold) than their maximum frequency observed in other reference populations. Arising via a well-described founder effect, this catalog of enriched alleles can contribute to differences in genetic risk and overall prevalence of diseases between populations. As validation we document 151 AJ enriched protein-altering alleles that overlap with "pathogenic" ClinVar alleles, including those that account for 10-100 fold differences in prevalence between AJ and non-AJ populations of some rare diseases including Gaucher disease (*GBA*, p.Asn409Ser, 8-fold enrichment); Canavan disease (*ASPA*, p.Glu285Ala, 12-fold enrichment); and Tay-Sachs disease (*HEXA*, c.1421+1G>C, 27-fold enrichment; p.Tyr427IlefsTer5, 12-fold enrichment). We next sought to use this catalog, of well-established relevance to Mendelian disease, to explore Crohn's disease, a common disease with an estimated two to four-fold excess prevalence in AJ. We specifically evaluate whether strong acting rare alleles, enriched by the same founder-effect, contribute excess genetic risk to Crohn's disease in AJ, and find that ten rare genetic risk factors in *NOD2* and *LRRK2* are strongly enriched in AJ, including several novel contributing alleles, show evidence of association to CD. Independently, we find that genomewide common variant risk defined by GWAS shows a strong difference between AJ and non-AJ European control population samples (0.97 s.d. higher, $p<10^{-16}$). Taken together, the results suggest coordinated selection in AJ population for higher CD risk alleles in general. The results and approach illustrate the value of exome sequencing data in case-control studies along with reference data sets like ExAC to pinpoint genetic variation that contributes to variable disease predisposition across populations.

Correspondence:
mrivas@stanford.edu, Dermot.McGovern@cshs.org,
judy.cho@mssm.edu and mjdaly@atgu.mgh.harvard.edu

Recent advances in genome sequencing technologies are improving our understanding of the etiology of human diseases[1,2]. Similarly, epidemiological studies over the past century have improved our understanding of their global distribution[3–5]. To date, it remains unclear the extent to which genetics may play a role in population-based differences in prevalence and/or incidence. Efforts to increase inclusion of a broader representation of populations in genomic studies will likely improve our interpretation of these observed differences[6]. Here, we present a study on the relative contribution of DNA sequence variants to the risk and prevalence of Crohn's disease (CD) and rare diseases in the Ashkenazi Jewish (AJ) population.

Genetic population isolates, defined as populations that start with a small group of founders and that may experience bottlenecks altering with periods of population growth[7], have facilitated the mapping of alleles contributing to human disease predisposition[8–11]. Tight bottlenecks scatter the relative contributions of genes and as a consequence may make it easier to discover some disease-associated genes, although other genes, where the alleles may have been depleted from the isolated population, will be harder to detect[10].

Genome-wide association studies (GWAS) and follow-up targeted sequencing efforts have been unusually successful in CD and have established a substantial role for low frequency variants across the more than 200 loci[2,12] defined to date. In addition, the documented 2-4 fold enrichment of CD prevalence in the AJ population[13,14], a population with an established founder effect, motivated the use of exome sequencing along with genome-wide array data to evaluate the degree to which bottleneck-enriched protein-altering alleles and unequivocally implicated common variants contribute an excess CD genetic risk to AJ, and as a consequence, to the documented increased prevalence of CD in AJ[13] (Figure S1).

Additionally, founder effects, and possible selection, have made some rare diseases more prevalent in the Ashkenazi Jewish (AJ) population[15,16,17,18], akin to the well-documented 40 rare diseases known as "Finnish heritage diseases", which are much more common in Finland[9] and whose difference in prevalence has largely been attributed to genetics. Despite the remarkable progress in mapping genes and alleles for some of these rare diseases, precise estimates of the risk-allele frequency and the carrier rate in the AJ population isolate have not yet been determined[16]. Through this study we provide a frequency resource of protein-coding alleles from over 2,000 non-CD AJ samples with low admixture fraction that will serve to improve interpretation of the carrier rate of rare disease risk alleles in the AJ population (Figure S1).

Results

We generated a jointly called exome dataset consisting of 18,745 individuals from international IBD and non-IBD cohorts[1,19]. As the present study aimed to focus on variation observed in the AJ population in comparison to reference populations in ExAC[19,20] (including non-Finnish Europeans (NFE), Latino (AMR), and African/African-American (AFR)) populations, we chose a model-based approach to estimate the ancestry of the study population using ADMIXTURE[21].

To identify AJ individuals and estimate admixture fractions we included a set (n=21,066) of LD-pruned common (MAF>1%, see Supplementary Note for additional details) variants after filtering for genotype quality (GQ>20). The 18,745 samples were assigned to four groups (K=4) using ADMIXTURE. In one of the four groups, 3,522 samples had estimated ancestry fraction > 0.9, with the majority of the samples labelled as "AJ" by contributing study sites (Figure S1). Because we were interested in computing an enrichment statistic that would not be affected by possible admixture we obtained alternate (non-reference) allele frequency estimates by restricting the

enrichment analysis to the 2,178 non-IBD Ashkenazi Jewish samples that passed QC and relatedness filtering and had AJ focused ancestry fraction > 0.9 (Figure S1).

To estimate parameters of enrichment in the AJ population, including proportion of enriched alleles and degree of enrichment, we used the observed alternate allele counts and total number of alleles to take into account uncertainty in estimated allele frequencies from AJ and NFE (n=33,370), AFR (n=5,203), and AMR (n=5,789) available from ExAC release 0.3 dataset (n_{total}=60,706). We focused on protein-coding alleles with estimated allele frequency of at least 0.002 in AJ ($n_{alleles}$=103,878), and applied a three group Bayesian mixture model, we refer to as the Population Isolate Enrichment Mixture Model (PIEMM) (see Supplementary Note), to classify the observed alleles into three groups: "depleted", "similar", or "enriched". We estimate that 30% of the analyzed protein-coding alleles have a mean 7.6-fold increased frequency compared to reference populations with markedly different proportion of alleles belonging to the "enriched" group depending on variant annotation: 44% for predicted protein-truncating variants (PTV); 36% for predicted protein-altering variants (PRA); and 27% for synonymous variants (Figure 1, Figure S2, $p < 10^{-16}$ across comparisons of PTV and PRA to synonymous variants, two-proportion test, Supplementary Note).

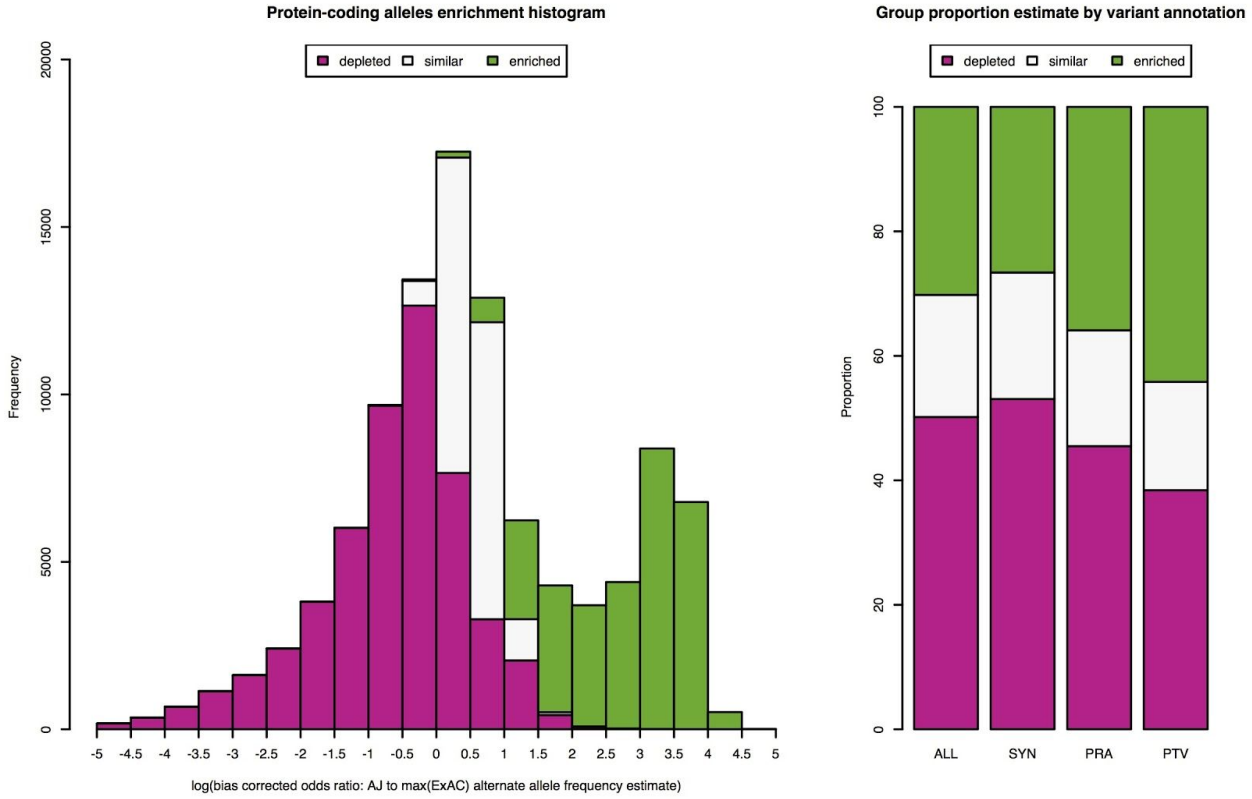


Figure 1. Enrichment of alleles discovered in AJ exome sequencing project. **A)** Density plot of estimated log enrichment statistic, defined as the log of the bias corrected odds ratio comparing the allele frequency in AJ population to the maximum allele frequency estimated from NFE, AFR, and AMR populations in ExAC. For each histogram bin we show a barplot of the expected number of alleles belonging to the three different groups we analyzed: 1) "depleted" (magenta); 2) "similar" (light gray); and 3) "enriched" (green). **B)** Bar plots of estimated proportion of alleles belonging to the three different groups we analyzed for all protein-coding (ALL), synonymous (SYN), protein-altering (PRA), and protein-truncating variants (PTV). An estimate of 30% of protein-coding alleles observed in AJ with a mean shift of 7.6-fold increase in allele frequency compared to other reference populations.

Variant	HGVS	Gene	Enrichment	AJ AF	max EXAC AF	Curated Traits	Inheritance
16:3293310_A/G	p.Val726Ala	MEFV	13.31	0.0416	0.0033	Familial Mediterranean fever	AR
1:155205634_T/C	p.Asn409Ser	GBA	8.40	0.0296	0.0036	Susceptibility to Lewy body dementia, Gauch	AR
4:187201412_T/C	p.Phe301Leu	F11	19.29	0.0273	0.0015	Hereditary factor XI deficiency	AR
13:20763553_CA/C	p.Leu56Argfs	GJB2	17.80	0.0199	0.0011	Autosomal recessive deafness	AR
4:187195347_G/T	p.Glu135Ter	F11	15.61	0.0195	0.0013	Factor XI deficiency	AR
12:11421038_G/A	p.Arg49Cys	PRB3	10.10	0.0189	0.0019	Salivary peroxidase	AR
9:111662096_A/G	c.2204+6T>C	IKBKAP	14.37	0.0168	0.0012	Familial dysautonomia	AR
1:215848678_C/T	p.Arg4192His	USH2A	13.06	0.0106	0.0008	Retinitis pigmentosa	AR
10:99371368_TGAG/T	p.Glu315del	HOGA1	14.05	0.0101	0.0007	Primary hyperoxaluria	AR
15:72638920_G/GGATA	p.Tyr427IlefsTer5	HEXA	11.60	0.0105	0.00106	Tay-Sachs disease	AR
7:117282620_G/A	p.Trp1282Ter	CFTR	12.67	0.0085	0.0007	Cystic fibrosis	AR
11:17418602_C/T	c.3992-9G>A	ABCC8	25.24	0.0076	0.0003	Hyperinsulinemic hypoglycemia	AR, AD
17:3402294_A/C	p.Glu285Ala	ASPA	12.10	0.0076	0.0006	Canavan disease	AR
2:98986540_G/A	c.101+1G>A	CNGA3	15.10	0.0074	0.0005	Achromatopsia	AR
7:117250575_G/C	p.Leu997Phe	CFTR	2.22	0.0073	0.0033	Idiopathic pancreatitits	AR
13:32914437_GT/G	p.Ser1982Argfs	BRCA2	14.43	0.0069	0.0005	Hereditary cancer, multiple types	risk factor
9:97934315_T/A	c.456+4A>T	FANCC	15.96	0.0069	0.0004	Fanconi anemia	AR
9:108382330_G/GA	p.Phe390Ilefs	FKTN	13.08	0.0067	0.0005	Limb-girdle muscular dystrophy-dystroglycar	AR
12:40734202_G/A	p.Gly2019Ser	LRRK2	10.34	0.0064	0.0006	Parkinson disease	risk factor
17:41055964_C/T	p.Arg83Cys	G6PC	6.99	0.0062	0.0009	Glycogen storage disease	AR
1:26764719_A/G	p.Lys42Glu	DHDDS	23.91	0.0051	0.0002	Retinitis pigmentosa	AR
3:150690352_A/C	p.Asn48Lys	CLRN1	13.58	0.0051	0.0004	Usher syndrome	AR
12:49312533_GTA/G	p.Ile293Profs	CCDC65	16.14	0.0048	0.0003	Ciliary dyskinesia without situs inversus	AR
6:80878662_G/C	p.Arg183Pro	BCKDHB	13.42	0.0046	0.0003	Maple syrup disease	AR
10:56077174_G/A	p.Arg245Ter	PCDH15	12.37	0.0046	0.0004	Usher syndrome	AR
7:107555951_G/T	p.Gly229Cys	DLD	11.05	0.0046	0.0004	Maple syrup disease	AR
15:72638575_C/G	c.1421+1G>C	HEXA	26.91	0.0044	0.0002	Tay-Sachs disease	AR
5:178699927_G/A	p.Gln225Ter	ADAMTS2	19.19	0.0041	0.0002	Ehlers-Danlos syndrome, dermatosparaxis ty	AR
16:50745656_G/A	p.Ala612Thr	NOD2	5.22	0.0039	0.0008	Early-onset sarcoidosis	risk factor
11:6415434_G/T	p.Arg498Leu	SMPD1	18.50	0.0039	0.0002	Niemann-Pick disease	AR
11:61161437_G/T	p.Arg73Leu	TMEM216	16.30	0.0039	0.0002	Joubert syndrome	AR
1:53676583_CAG/C	p.Lys414ThrfsTer7	CPT2	18.78	0.0037	0.0002	Carnitine palmitoyltransferase II deficiency	AR
1:53676688_T/C	p.Phe448Leu	CPT2	18.78	0.0037	0.0002	Carnitine palmitoyltransferase II deficiency	AR
3:172737276_C/T	p.Arg283Gln	SPATA16	8.31	0.0037	0.0004	Spermatogenic failure	AR
11:86017416_G/C	p.Val54Leu	C11orf73	18.79	0.0037	0.0002	Hypomyelinating leukodystrophy	AR
8:77896060_G/A	p.Arg119Ter	PEX2	15.36	0.0034	0.0002	Peroxisome biogenesis disorder	AR
11:118951899_T/G	p.Cys845Gly	VPS11	18.03	0.0030	0.0002	Hypomyelinating leukodystrophy	AR
6:80203353_G/A	p.Gln279Ter	LCA5	13.19	0.0028	0.0002	Leber congenital amaurosis	AR
19:7591645_A/G	c.406-2A>G	MCOLN1	8.89	0.0028	0.0003	Mucolipidosis	AR
16:56530894_C/G	p.Arg632Pro	BBS2	13.24	0.0028	0.0002	Retinitis pigmentosa	AR
17:41276044_ACT/A	p.Glu23Valfs	BRCA1	6.16	0.0025	0.0004	Hereditary cancer, multiple types	risk factor
4:100543913_G/T	p.Gly865Ter	MTTP	20.78	0.0025	0.0001	Abetalipoproteinaemia	AR
2:99013302_G/A	p.Gly557Arg	CNGA3	16.97	0.0023	0.0001	Achromatopsia	AR
7:107557794_G/A	p.Glu375Lys	DLD	13.90	0.0021	0.0001	Maple syrup disease	AR
17:41209079_T/TG	p.Gln1756Profs	BRCA1	7.48	0.0021	0.0003	Hereditary cancer, multiple types	risk factor
10:99371292_G/T	p.Gly287Val	HOGA1	11.66	0.0021	0.0002	Primary hyperoxaluria	AR
22:29091207_G/A	p.Ser428Phe	CHEK2	18.85	0.0103	0.0006	Hereditary cancer, multiple types	risk factor
8:43054647_G/A	p.Ala615Thr	HGSNAT	2.26	0.0137	0.0062	Mucopolysaccharidosis	AR
15:72105913_G/A	p.Arg311Gln	NR2E3	8.65	0.0042	0.0005	Enhanced s-cone syndrome	AR

Table 1. Forty-nine ClinVar "pathogenic" alleles enriched in AJ. HGVS and Gene is the allele nomenclature in ClinVAR and gene symbol. Enrichment corresponds to the comparison of allele frequency in AJ (AJ AF) to maximum frequency among three population groups (max EXAC AF): 1) NFE; 2) AMR; and 3) AFR. Curated trait is based on the trait description in the Online Mendelian Inheritance in Man (OMIM). Inheritance corresponds to the inheritance description in OMIM (AR: autosomal recessive, AD: autosomal dominant, risk factor: not specified genetic risk factor). Alleles are sorted in decreasing order by AJ AF.

As validation of our approach to identify alleles that contribute to differences in genetic risk to disease, we intersected the list of protein-coding alleles identified in the AJ exome sequencing study with reported pathogenic and non-conflicted alleles (n=42,226) in ClinVar[22] resulting in 151 alleles found both in ClinVar and with posterior probability greater than 0.5 of belonging to the enriched group (Table S1). In OMIM, 49 of the 151 alleles included documentation of a disease subject with AJ ancestry (Table 1). This set of enriched alleles includes 9/14 alleles described in the American College of Medical Genetics and Genomics 2008 screening guideline study for the AJ population[23]. In the setting of autosomal recessive disorders these differences in population allele frequencies may contribute to an order of squared enrichment difference in genetic risk and prevalence between populations (see Supplementary Note). For instance, a 12-fold enriched frameshift indel, p.Tyr427IlefsTer5, in *HEXA*, contributes a 144-fold enrichment in genetic risk in AJ to non-AJ population to

Tay-Sach's disease. This large cohort of adult Ashkenazi exome database further supports recent publications of founder mutations for rare pediatric disorders including: *FKTN* and Walker Warburg syndrome[24]; *CCDC65* and Primary ciliary dyskinesia[25]; *TMEM216* and Joubert syndrome[26]; *C11orf73* and Leukoencephalopathy[27]; *PEX2* and Zellweger syndrome[28]; *VPS11* and Hypomyelination and developmental delay[29]; and *BBS2* and Bardet-Biedl syndrome[30]. This resource will undoubtedly assist in prioritizing variants for gene discovery to further identify founder mutations in AJ (Table S2).

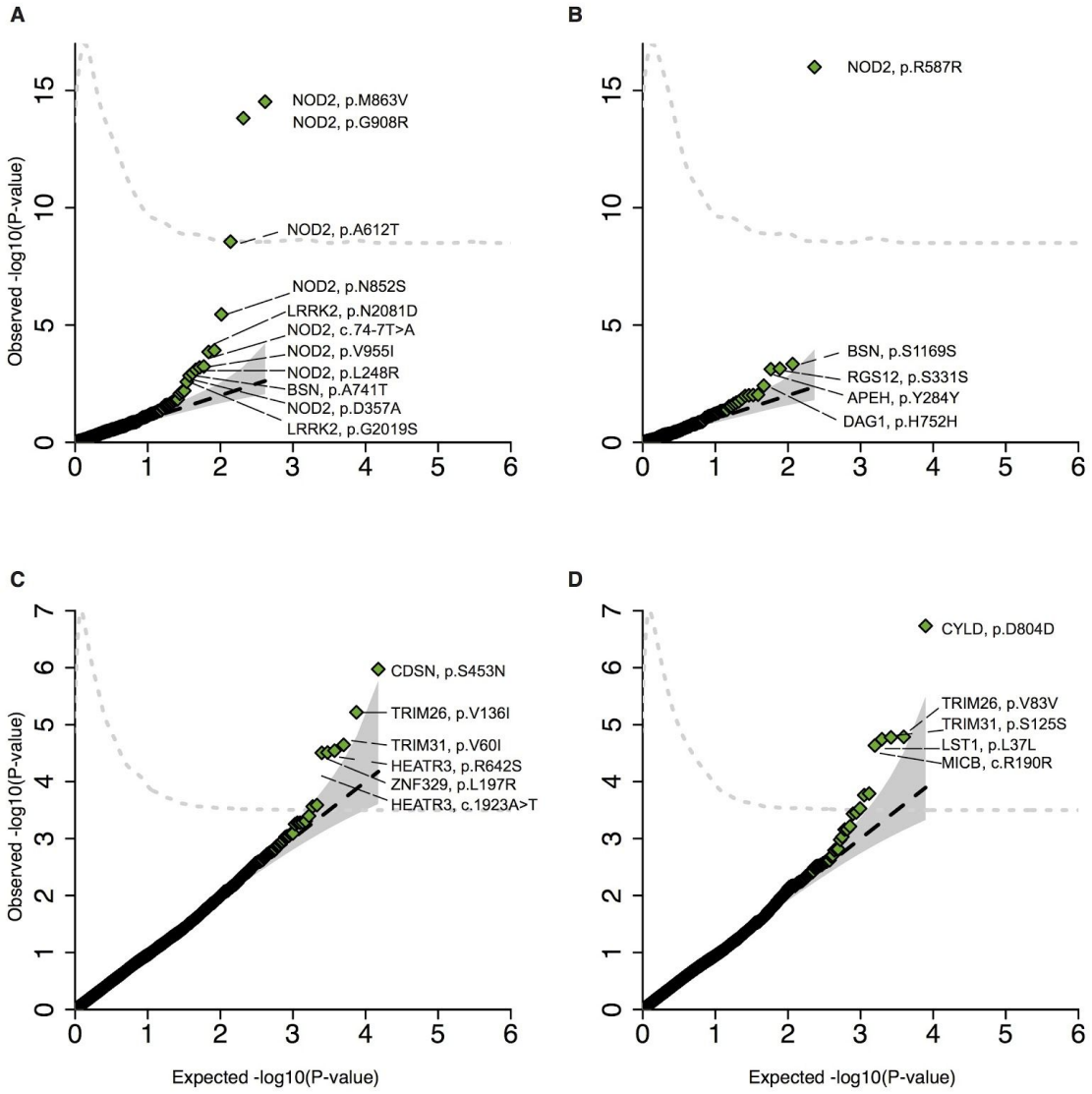


Figure 2. Q-Q plots of enriched alleles. Q-Q plots of Crohn's disease association for: **A**) AJ enriched protein-altering (protein-truncating and missense) and **B**) synonymous alleles in GWAS regions; and AJ enriched **C**) protein-altering and **D**) synonymous alleles outside of GWAS regions. For each Q-Q plot variants with a corresponding p-value less than or equal to a threshold where expected number of false discoveries is equal to one are annotated. The black dashed line is y = x, and the grey shapes show 95% confidence interval under the null. The gray dashed line represents the observed density of -log10 p-values.

To assess whether protein-coding alleles enriched in AJ population contribute to differences in CD genetic risk we performed case-control association analyses. Because individuals with partial AJ ancestry will still carry bottleneck-enriched alleles, here we included all samples with estimated AJ ancestry fraction of at least 0.4 (Figure S2), which resulted in a dataset of 4,899 AJ samples (1,855 Crohn's disease and 3,044 non-IBD). To improve ability to detect association we performed a meta-analysis with CD and non-IBD case-control exome sequencing data from two separate ancestry groups: 1) non-Finnish European (NFE) (2,296 CD and 2,770 non-IBD); and 2) Finnish (FINN) (210 CD and 9,930 non-IBD samples) from a separate callset described in a previous publication[31] for a total of 4,361 CD samples and 15,744 non-IBD samples.

Study-specific association analysis was performed with Firth bias-corrected logistic regression test[32,33] and four principal components as covariates using the software package EPACTS[34] (Figure S4). We combined association statistics in a meta-analysis framework using the Bayesian models in Band et al[35]. We used the correlated effects model, obtained a Bayes factor (BF) by comparing it with the null model where all the prior weight is on an effect size of zero, reported p-value approximation using the BF as a test statistic, and assessed whether heterogeneity of effects exist across studies for downstream QC (see Supplementary Note). We separately assessed CD observed vs. expected associations for enriched protein-altering (pra) and synonymous (syn) alleles in protein-coding genes in CD implicated GWAS loci ($n_{gwas,pra}$=413; $n_{gwas,syn}$=231), and outside implicated GWAS loci ($n_{non\text{-}gwas,pra}$=14,961; $n_{non\text{-}gwas,syn}$=7,858, Figure 2).

We identified ten AJ enriched CD risk alleles ($p<0.005$): the previously published risk haplotypes in *LRRK2* and *NOD2* (*LRRK2*: p.N2081D; *NOD2*: p.N852S, p.G908R, p.M863V+fs1007insC)[36,2], in addition to newly implicated alleles (*NOD2*: p.A612T, $p=2.8\times10^{-9}$; c.74-7T>A, $p=1.4\times10^{-4}$; p.L248R, $p=6.4\times10^{-4}$; p.D357A, p=0.0011; *LRRK2*: p.G2019S, p=0.0014, a Parkinson's disease risk allele[37]). To assess whether the new *NOD2* enriched alleles are conditionally independent of the previously established associated *NOD2* alleles we performed conditional haplotype association analysis in PLINK and Bayesian model averaging[38] for variable selection, both of which suggested independent effects for all alleles (Figure S5, Table S3).

Despite the functional relationship between *LRRK2* and *NOD2*[39], we do not observe deviation from additivity between *LRRK2* and *NOD2* (p=0.273). Deviation from additivity has been reported for p.fs1007insC, p.G908R, and p.R702W in *NOD2*[40,41]. We assessed whether any independent evidence of deviation from additivity exists for the newly associated single nucleotide substitutions. In agreement with previous reports, deviation from additivity existed with estimated genotype odds ratios of 1.84 for heterozygous and 7.39 for compound heterozygous/homozygous genotypes (p=0.0038, analysis of deviance, ANOVA), and found significant evidence of deviation from additivity for the newly reported alleles (p=0.00357, odds ratio = 7.53). We found no evidence of deviation from additivity for the associated protein-altering alleles in *LRRK2* (p=0.418).

Given the presence of genetic variants in *NOD2* and *LRRK2* that contribute to differences in genetic risk in AJ population, we next asked whether unequivocally established common variant loci associations may also contribute to differences in genetic risk. We performed polygenic risk score (PRS) analysis using reported effect size estimates from 124 CD alleles including those reported in a previously published study[12] and four variants in *IL23R* from a recent fine-mapping study[42], and excluding variants in *NOD2* and *LRRK2*. We observed an elevated PRS for AJ compared to non-Jewish controls (0.97 s.d. higher, $p<10^{-16}$; Figure 3A; number of non-AJ controls=35,007; number of AJ controls=454). We observed a similar trend for the CD samples (0.54 s.d. higher; $p<10^{-16}$; Figure 3B; number of non-AJ CD cases=20,652; number of AJ CD cases=1,938).

To quantify the relative contribution of CD-implicated alleles to the difference in genetic risk between AJ and non-AJ populations we estimated the expected PRS value of an individual and expected difference in PRS between two populations by simply using summary statistics including the frequency of the minor allele in the two populations and the corresponding odds ratio (Supplementary note, Figures S6-S8).

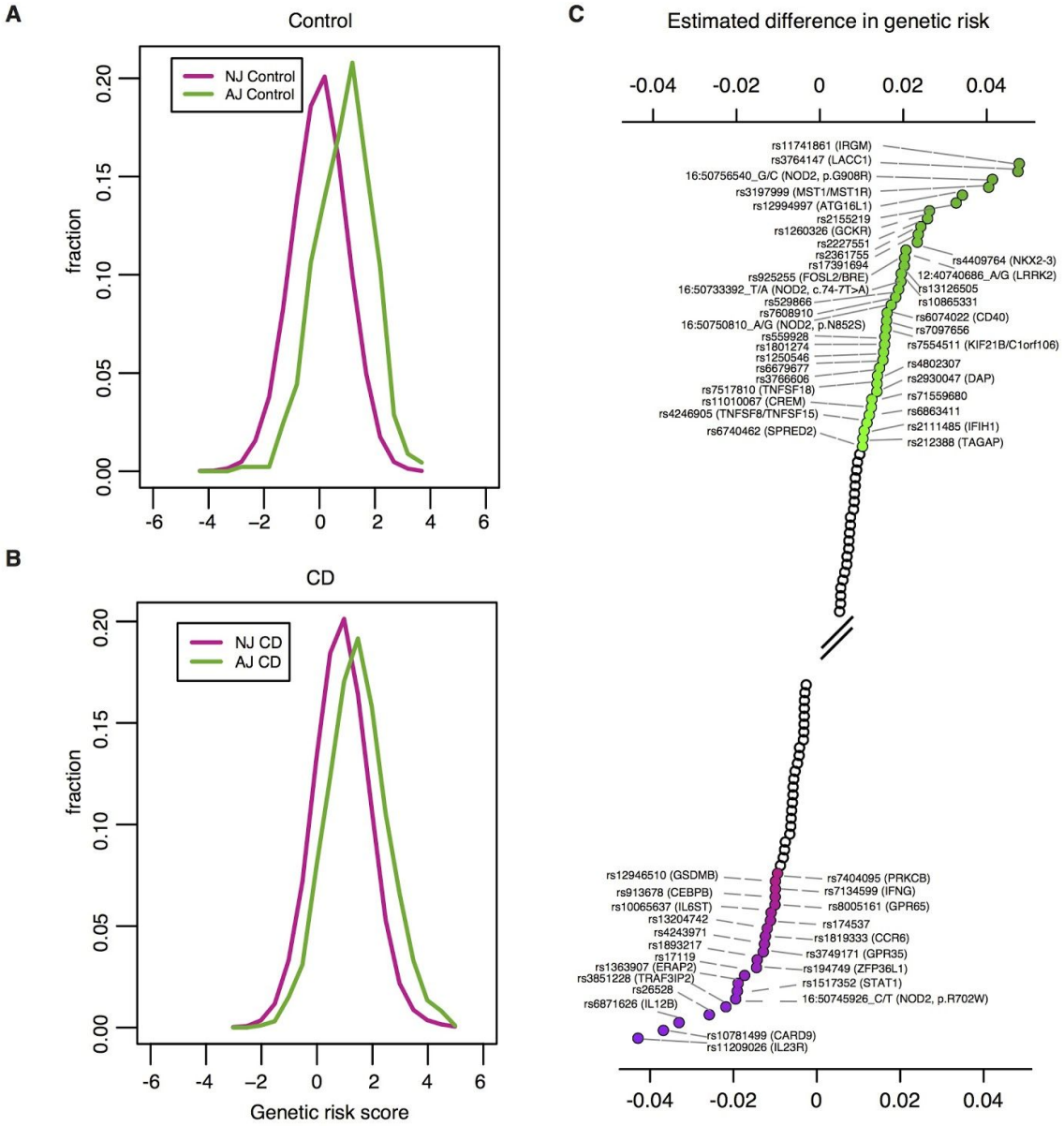


Figure 3. AJ individuals have higher CD polygenic risk score than NJ controls. NJ: non-Jewish; AJ: Ashkenazi Jewish; CD: Crohn's disease; PRS: polygenic risk score. **A**) Density plot of CD polygenic risk scores in 454 AJ (green) and 35,007 NJ (purple) controls. AJ controls have elevated CD polygenic risk score that NJ controls (0.97 s.d. higher, $p<10^{-16}$). **B**) Density plot of CD polygenic risk scores in 1,938 AJ (green) and 20,652 NJ CD (purple) cases (0.54 s.d. higher, $p<10^{-16}$). For both density plots the scores have been scaled to NJ controls, thus resulting in an NJ control PRS density of mean equal to 0 and variance equal to 1 (see Online methods). **C**) Ranked (decreasing order) CD associated variants by estimated contribution to the differences in genetic risk between AJ and NJ. Associated variants with estimated contribution greater than or equal to 0.01, computed as 2*log(odds ratio)*(AJ frequency - NJ frequency), assuming additive effects on the log scale, are highlighted in green. Associated variants with estimated contribution less than or equal to -0.01 are highlighted in purple. Forward slashes represent a break in variants highlighted.

We applied the approach to all CD implicated alleles and observed that variants in GWAS loci annotated as *IRGM*, *LACC1*, *NOD2*, *MST1*, *ATG16L1*, *GCKR, NKX2-3*, and *LRRK2*[12] contribute substantially (>0.01) to the increased genetic risk observed in AJ. It is possibly relevant that variants contributing to increased risk in AJ correspond to autophagy/intracellular defense genes (*IRGM, ATG16L1, LRRK2*) and those contributing to increased risk in non-AJ correspond to anti-fungal/Th17/ILC3 genes[43] (*IL23R*, *IL12B*, *CARD9*, *TRAF3IP2*, *IL6ST*, *CEBPB*; Figure 3C).

Two factors impact our ability to obtain precise estimates of the contribution of genetic risk factors to the documented differences in disease prevalence between populations. First, contemporaneous epidemiological data are largely unattainable. This confounds our ability to obtain any informative estimates due to documented variability in the occurrence of CD over time[3,44]. Second, there has been substantial uncertainty in reported CD prevalence estimates[45,46]. Here, to interpret the impact of shifts in genetic risk score on differences in prevalence, we used the logit risk model[41] and evaluated a new estimate of disease probability, p_{new}, assuming an initial disease probability, p_0, and multiple values for the differences in genetic risk. Assuming log-additive effects, and a log-risk model, we estimate that the observed differences in genetic risk between the AJ and non-AJ populations may contribute an expected 1.5-fold increase in disease prevalence in a population with environmental risk factors corresponding to AJ and baseline genetic risk corresponding to non-AJ populations (Figure S6-S8). To address the extent to which non-additive effects in *NOD2* may impact the observed prevalence we assumed shared heterozygous and compound heterozygous/homozygous odds ratios of 1.84 and 7.39, respectively. We estimate a 1.6% increase in difference in prevalence attributed to the deviation from additivity, suggesting a small effect on differences in population prevalence (Supplementary Note).

Discussion

By drawing on data from 5,685 Ashkenazi Jewish exomes we provide a systematic analysis of AJ enriched protein-coding alleles, which may contribute to differences in genetic risk to CD as well as numerous other rare diseases. We identified protein-altering alleles in *NOD2* and *LRRK2* that are conditionally independent and contribute to the excess burden of CD in AJ. We found evidence that common variant risk defined by GWAS shows a strong elevated difference between AJ and non-AJ European population samples, independent of *NOD2* and *LRRK2*[47], suggesting a coordinated selection in AJ for higher CD risk alleles[48–50].

We made a couple of unique observations by studying CD in the AJ population. First, studying recently bottlenecked populations enables powerful discovery of genetic variation that markedly differ in frequency and as a consequence contributes to differences in genetic risk across population groups. Second, *NOD2* and published common variant associations contribute substantially to the genetic risk of CD, making it more difficult to identify other genes whose ancestral alleles failed to pass through the bottleneck, consistent with predictions from Zuk et al[10].

Finally, we provide an exome frequency resource of protein-coding alleles in AJ along with estimates of enrichment. Our approach and this resource will likely catalyze our understanding of the medical relevance of enriched alleles in population isolates.

References

1. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).
2. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47,** 979–986 (2015).
3. Bernstein, C. N., Blanchard, J. F., Rawsthorne, P. & Wajda, A. Epidemiology of Crohn's Disease and Ulcerative Colitis in a Central Canadian Province: A Population-based Study. *Am. J. Epidemiol.* **149,** 916–924 (1999).
4. Chen, C.-Y. *et al.* Epidemiology and Disease Burden of Ulcerative Colitis in Taiwan: A Nationwide Population-Based Study. *Value in Health Regional Issues* **2,** 127–134 (2013).
5. Mohan, V., Sandeep, S., Deepa, R., Shah, B. & Varghese, C. Epidemiology of type 2 diabetes: Indian scenario. *Indian J. Med. Res.* **125,** 217–230 (2007).
6. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. Genomics for the world. *Nature* **475,** 163–165 (2011).
7. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* **1,** 182–190 (2000).
8. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512,** 190–193 (2014).
9. Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10,** e1004494 (2014).
10. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455–64 (2014).
11. Bahcall, O. & Orli, B. Rare variant association studies. *Nat. Genet.* **46,** 219–219 (2014).

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

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

14. Karban, A., Eliakim, R. & Brant, S. R. Genetics of inflammatory bowel disease. *Isr. Med. Assoc. J.* **4,** 798–802 (2002).

15. Risch, N., Tang, H., Katzenstein, H. & Ekstein, J. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am. J. Hum. Genet.* **72,** 812–822 (2003).

16. Baskovich, B. *et al.* Expanded genetic screening panel for the Ashkenazi Jewish population. *Genet. Med.* **18,** 522–528 (2016).

17. Ostrer, H. A genetic profile of contemporary Jewish populations. *Nat. Rev. Genet.* **2,** 891–898 (2001).

18. Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* **5,** 4835 (2014).

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

20. Karczewski, K. J. *et al. The ExAC Browser: Displaying reference data information from over 60,000 exomes*. (2016). doi:10.1101/070581

21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

22. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2015).

23. Gross, S. J., Pletcher, B. A., Monaghan, K. G. & Professional Practice and Guidelines Committee. Carrier screening in individuals of Ashkenazi Jewish descent. *Genet. Med.* **10,** 54–56 (2008).

24. Chang, W. *et al.* Founder Fukutin mutation causes Walker-Warburg syndrome in four Ashkenazi Jewish families. *Prenat. Diagn.* **29,** 560–569 (2009).

25. Fedick, A. M., Jalas, C., Treff, N. R., Knowles, M. R. & Zariwala, M. A. Carrier frequencies of eleven mutations in eight genes associated with primary ciliary dyskinesia in the Ashkenazi Jewish population. *Mol Genet Genomic Med* **3,** 137–142 (2015).

26. Edvardson, S. *et al.* Joubert syndrome 2 (JBTS2) in Ashkenazi Jews is associated with a TMEM216 mutation. *Am. J. Hum. Genet.* **86,** 93–97 (2010).

27. Edvardson, S. *et al.* Leukoencephalopathy and early death associated with an Ashkenazi-Jewish founder mutation in the Hikeshi gene. *J. Med. Genet.* **53,** 132–137 (2016).

28. Fedick, A., Jalas, C. & Treff, N. R. A deleterious mutation in the PEX2 gene causes Zellweger syndrome in individuals of Ashkenazi Jewish descent. *Clin. Genet.* **85,** 343–346 (2014).

29. Edvardson, S. *et al.* Hypomyelination and developmental delay associated with VPS11 mutation in Ashkenazi-Jewish patients. *J. Med. Genet.* **52,** 749–753 (2015).

30. Fedick, A. *et al.* Carrier frequency of two BBS2 mutations in the Ashkenazi population. *Clin. Genet.* **85,** 578–582 (2014).

31. Rivas, M. A. *et al.* A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. *Nat. Commun.* **7,** 12342 (2016).

32. Firth, D. & D., F. 'Bias reduction of maximum likelihood estimates'. *Biometrika* **82,** 667–667 (1995).

33. Ma, C., Blackwell, T., Boehnke, M., Scott, L. J. & GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37,** 539–550 (2013).

34. Kang, H. M. *EPACTS: efficient and parallelizable association container toolbox*. (2012).

35. Band, G. *et al.* Imputation-based meta-analysis of severe malaria in three African populations. *PLoS Genet.* **9,** e1003509 (2013).

36. 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).

37. Ozelius, L. J. *et al.* LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews. *N. Engl. J. Med.* **354,** 424–425 (2006).

38. Raftery, A. E. Bayesian Model Selection in Social Research. *Sociol. Methodol.* **25,** 111 (1995).

39. Zhang, Q. *et al.* Commensal bacteria direct selective cargo sorting to promote symbiosis. *Nat. Immunol.* **16,** 918–926 (2015).

40. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411,** 603–606 (2001).

41. Jostins, L. Using next-generation genomic datasets in disease association. (University of Cambridge, 2013).

42. Huang, H. *et al. Association mapping of inflammatory bowel disease loci to single variant resolution*. (2015). doi:10.1101/028688

43. Richard, M. L., Lamas, B., Liguori, G., Hoffmann, T. W. & Sokol, H. Gut fungal microbiota: the Yin and Yang of inflammatory bowel disease. *Inflamm. Bowel Dis.* **21,** 656–665 (2015).

44. Turunen, P. *et al.* Incidence of inflammatory bowel disease in finnish children, 1987–2003. *Inflamm. Bowel Dis.* **12,** 677–683 (2006).

45. Mayberry, J. F., Rhodes, J. & Newcombe, R. G. Familial prevalence of inflammatory bowel disease in relatives of patients with Crohn's disease. *BMJ* **280,** 84–84 (1980).

46. Sandier, R. S. The epidemiology of inflammatory bowel disease. *Curr. Opin. Gastroenterol.* **6,** 531–535 (1990).

47. Nakagome, S. *et al.* Crohn's disease risk alleles on the NOD2 locus have been maintained by natural selection on standing variation. *Mol. Biol. Evol.* **29,** 1569–1585 (2012).

48. Schurr, E., Erwin, S. & Philippe, G. A Common Genetic Fingerprint in Leprosy and Crohn's Disease? *N. Engl. J. Med.* **361,** 2666–2668 (2009).

49. Liu, H. *et al.* Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. *Nat. Genet.* **47,** 267–271 (2015).

50. Zhang, F.-R. *et al.* Genomewide association study of leprosy. *N. Engl. J. Med.* **361,** 2609–2618 (2009).

51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

52. Rivas, M. A. *et al. A protein truncating R179X variant in RNF186 confers protection against ulcerative colitis*. (2015). doi:10.1101/035105

53. Heliö, T. *et al.* CARD15/NOD2 gene variants are associated with familially occurring and complicated forms of Crohn's disease. *Gut* **52,** 558–562 (2003).

54. Lappalainen, M. *et al.* Novel CARD15/NOD2 mutations in Finnish patients with Crohn's disease and their relation to phenotypic variation in vitro and in vivo. *Inflamm. Bowel Dis.* **14,** 176–185 (2008).

55. *Analysis of protein-coding genetic variation in 60,706 humans*. (2015). doi:10.1101/030338

56. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

57. Tan, A., Abecasis, G. R. & Kang, H. M. Unified representation of genetic variants. *Bioinformatics* **31,** 2202–2204 (2015).

58. Wakefield, J. & Jon, W. A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *Am. J. Hum. Genet.* **81,** 208–227 (2007).

59. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *Lancet* **380,** 1590–1605 (2012).

60. Moonesinghe, R. *et al.* Estimating the contribution of genetic variants to difference in incidence of disease between population groups. *Eur. J. Hum. Genet.* **20,** 831–836 (2012).

Author affiliations

[1]Medical and Population Genetics, Broad Institute, Cambridge, MA, USA
[2]Department of Biomedical Data Science, Stanford, CA, USA
[3]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[4]F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA
[5]Hadassah-Hebrew University Medical Center, Endocrinology and Metabolism Service Department of Internal Medicine, Jerusalem, Israel
[6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[7]Department of Genetics and Medicine, Albert Einstein College of Medicine, Bronx, NY, USA
[8]Faculty of Natural Sciences, University of Haifa, Haifa, Israel
[9]Division of Medicine, University College London, London, UK
[10]UCL Genetics Institute, University College London, London, UK
[11]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany
[12]Gastroenterology Department, Saint-Antoine Hospital, AP-HP, UPMC Univ Paris 06, Paris, France
[13]Department of Internal Medicine, University Hospital Schleswig-Holstein, Kiel, Germany
[14]Gastroenterology Practice, Minden, Germany
[15]IBD Pharmacogenetics, Royal Devon and Exeter NHS Trust, Exeter, UK
[16]Peninsula College of Medicine and Dentistry, Exeter, UK
[17]Juliet Keidan Institute of Pediatric Gastroenterology and Nutrition, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel
[18]Department of Medical Genetics, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland
[19]Department of Medicine, Division of Gastroenterology, Helsinki University Hospital, Helsinki, Finland
[20]Department of Medicine, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
[21]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[22]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
[23]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[24]Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA.
[25]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA.
[26]Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, USA.
[27]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, Pennsylvania, USA.

[28]Inflammatory Bowel Disease Centre, Mount Sinai Hospital, Toronto, Ontario, Canada
[29]Research Center, Montreal Heart Institute, Montréal, Québec, Canada
[30]Department of Medicine, Université de Montréal, Montréal, Québec, Canada
[31]Department of Gastroenterology and Hepatology, University Medical Center Groningen, Groningen, The Netherlands
[32]Bonei Olam, Center for Rare Jewish Genetic Disorders, Brooklyn, NY, USA.
[33]Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease and Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[34]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
[35]Icahn School of Medicine at Mount Sinai, Dr Henry D. Janowitz Division of Gastroenterology, New York, New York, USA

Author contributions

M.A.R, D.P.B.M, M.J.D participated in the study design. M.A.R, J.K., M.K., and M.J.D analyzed whole exome data. M.A.R, H.H, T.H, and B.A analyzed the SNP chip data. B.M.N, A.G, D.G, B.G, I.P, G.A, N.B, A.P.L, E.S, N.P, Ben Weisburd, K.J.K, E.V.M, B.P, L.B, P.S, J.C, Graham Heap, T.A, V.P, A.W.S, S.T, Dan Turner, P.S, M.F, K.K, M.P, Aarno Palotie, S.R.B, D.G.M, R.H.D, Mark S. Silverberg, J.D.R, R.K.W, A.F, H.S, R.J.X, A.P, J.H.C, D.P.B.M provided reagents, methods, and tools for analysis. C.S. managed the project. C.J and J.H.C provided detailed analysis of rare diseases in AJ. All authors commented on the manuscript. J.D.R, Mark J. Daly, S.R.B, R.H.D, Mark S. Silverberg, J.H.C, and D.P.B.M are members of the NIDDK IBD Genetics consortium. Manuel A. Rivas, J.K., B.A, and Mark J Daly wrote the manuscript.

Acknowledgments

M.J.D. is supported by grants from the following: the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) and the National Human Genome Research Institute (NHGRI; DK043351, DK064869 and HG005923); the Crohns and Colitis Foundation (3765); the Leona M. & Harry B. Helmsley Charitable Trust (2015PG-IBD001); the Stanley Center; and Amgen (2013583217). R.J.X. is supported by grants from Amgen (2013583217) and CCFA (3765). J.D.R. is funded by grants from NIDDK (DK064869 and DK062432). G.A. is supported by NIH R01 grant AG042188. N.B. is supported by NIH grants AG618381, AG021654, AG038072, and the Glenn Center for the Biology of Human Aging. A.F. and B.P. are supported by the DFG (Deutsche Forschungsgemeinschaft) Cluster of Excellence "Inflammation at Interfaces" and the DFG grant FR 2821/6-1. I.P. is supported by the Leona M. and Harry B. Helmsley Charitable Trust and New York Crohn's Foundation. IBD Research at Cedars-Sinai is supported by grant PO1DK046763, and the Cedars-Sinai F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute Research Funds. D.P.B.M. is supported by DK062413, AI067068 and U54DE023789-01; grant 305479 from the European Union; and The Leona M. and Harry B. Helmsley Charitable Trust and the Crohn's and Colitis Foundation of America. S.R.B. is support by an NIH U01 grant (DK062431). J.H.C. is supported by grants from NIH (U01 DK062429, U01 DK062422, R01 DK092235, SUCCESS), and the Sanford J. Grossman Charitable Trust. H.S. is supported by Equipe ATIP–Avenir 2012 grant and INSERM–ITMO SP 2013. E.V.M. is supported by an NIH F13 award (AI122592-01A1). Researchers at UCL are supported by the Charles Wolfson Foundation Trust. RKW is

supported by a VIDI grant (016.136.308) from the Netherlands Organization for Scientific Research (NWO). R.H.D. is supported by NIH grant U01 DK062420 and the Inflammatory Bowel Disease Genetic Research Chair at the University of Pittsburgh. We thank Dr. Jonathan Bloom for proposed edits and comments on the manuscript. We thank the Broad IT team for assistance with the IBD exomes browser.

Supplementary Materials:

Materials and Methods

Figures S1-S8

Data Files S1-S2

Table S3

Supplementary Material for: "Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population"

Contents

List of Figures

List of Tables

S1. Materials and methods

Initial variant call set. We generated a jointly called dataset consisting of 18,745 individuals from international IBD and non-IBD cohorts. Sequencing of these samples was done at Broad Institute.

Ethics statement. All patients and control subjects provided informed consent. Recruitment protocols and consent forms were approved by Institutional Review Boards at each participating institutions (Protocol Title: The Broad Institute Study of Inflammatory Bowel Disease Genetics; Protocol Number: 2013P002634). All DNA samples and data in this study were denominalized.

Cohort descriptions. For all cohorts, CD was diagnosed according to accepted clinical, endoscopic, radiological and histological findings.

Target selection. G4L WES is a project specific product. It combines the Human WES (Standard Coverage) product with an Infinium Genome-Wide Association Study (GWAS) array. In addition to the array adding to the genomics data, it also acts as a concordance QC, linking 14 SNPs to the exome data. The processing of the exome includes Sample prep (Illumina Nextera), hybrid capture (Illumina Rapid Capture Enrichment - 37Mb target), sequencing (Illumina, HiSeq machines, 150bp paired reads), Identification QC check, and data storage (5 years). Our hybrid selection libraries typically meet or exceed 85% of targets at 20x, comparable to ~60x mean coverage. The array consists of a 24-sample Infinium array with ~245,000 fixed genome-wide markers, designed by the Broad. On average our genotyping call rates typically exceed 98%.

Pre-processing. The sequence reads are first mapped using BWA MEM[51] to the GRCh37 reference to produce a file in SAM/BAM format sorted by coordinate. Duplicate reads are marked – these reads are not informative and are not used as additional evidence for or against a putative variant. Next, local realignment is performed around indels. This identifies the most consistent placement of the reads relative to potential indels in order to clean up artifacts introduced in the original mapping step. Finally, base quality scores are recalibrated in order to produce more accurate per-base estimates of error emitted by the sequencing machines.

Variant Discovery. Once the data has been pre-processed as described above, it is put through the variant discovery process, i.e. the identification of sites where the data displays variation relative to the reference genome, and calculation of genotypes for each sample at that site. The variant discovery process is decomposed into separate steps: variant calling (performed per-sample), joint genotyping (performed per-cohort) and variant filtering (also performed per-cohort). The first two steps are designed to maximize sensitivity, while the filtering step aims to deliver a level of specificity that can be customized for each project.

Variant calling is done by running Genome Analysis Toolkit's (GATK) HaplotypeCaller in GVCF mode on each sample's BAM file(s) to create single-sample gVCFs. If there are more than a few hundred samples, batches of ~200 gVCFs are merged hierarchically into a single gVCF to make the next step more tractable. Joint genotyping is then performed on the gVCFs of all available samples together in order to create a set of raw SNP and indel calls. Finally, variant recalibration is performed in order to assign a well-calibrated probability to each variant call in a raw call set, and to apply filters that produce a subset of calls with the desired balance of specificity and sensitivity as described in Rivas et al. (2016)[52]. Samples with >= 10% contamination are excluded from call sets. Exome samples with less than 40% of targets at 20X coverage are excluded.

Variant annotation.

Variant annotation was performed using the Variant Effect Predictor (VEP) [cite PMID: 20562413] version 83 with Gencode v19 on GRCh37. Loss-of-function (LoF) variants were annotated using LOFTEE (Loss-Of-Function Transcript Effect Estimator, available at https://github.com/konradjk/loftee), a plugin to VEP. LOFTEE considers all stop-gained, splice-disrupting, and frameshift variants, and filters out many known false-positive modes, such as variants near the end of transcripts and in non-canonical splice sites, as described in the code documentation.

Identification of Finnish samples. Finnish CD patients were recruited from Helsinki University Hospital and described in more detail previously[53,54]. We used the same exome sequencing dataset described in Rivas et al.[31] We applied additional PC correction in the Finnish identified individuals to remove individuals with membership of Finnish sub-isolate (Northern Finland) and excluded based on PC2 >= 0.015 (853 excluded, 826 controls, 27 IBD). We recalculated PCs and included the first four PCs in the association analysis.

Ancestry estimation and quality control. As the present study aimed to focus on variation observed in Ashkenazi Jewish (AJ) population in comparison to reference populations in ExAC[55] including (non-Finnish Europeans (NFE), Latino (AMR), and African/African-American (AFR)) we chose a model-based approach to estimate the ancestry of the study population using ADMIXTURE[21]. To identify AJ individuals and estimate admixture proportions we included a set (n=21,066) of LD-pruned common variants (MAF>1%) variants after filtering for genotype quality (GQ>20). We selected 50 window size in SNPs, 5 SNPs to shift the window at each step, and the variance inflation factor (VIF) threshold equal to 2. The 18,745 samples were assigned to four groups (K=4) using ADMIXTURE. In one of the four groups, 3,522 samples had estimated ancestry fraction > 0.9, with the majority of the samples labelled as "AJ" by contributing study sites, belonging to the group with high probability (Figure S1B).

Prior to enrichment and association analysis, 81 samples (of total 18,745) were also filtered due to possible contamination (heterozygous/homozygous ratio < 1), excess of singletons (n>2000), deletion/insertion ratio (>1.5) and mean genotype quality (<40). 275 samples were excluded for relatedness (pi>0.35 cut-off). Genotypes with low genotype quality (<20) were filtered, in addition to variants with low call rate (<80%) and allele balance deviating from 70:30 ratio for greater than 40% of heterozygous samples if at least 7 heterozygous samples were identified.

As we were interested in computing an enrichment statistic that would not be affected by possible admixture, we obtained alternate allele frequency estimates by restricting the enrichment analysis to the 2,178 non-IBD Ashkenazi Jewish samples that passed QC and relatedness filtering and had AJ focused ancestry fraction > 0.9 (Figure S1). Principal Component Analysis (PCA) was done in each ancestry group using the 21,066 variants. Sample QC was done using the Hail software while PCA, differential missingness, allele balance and sample relatedness analysis was done using PLINK[56]. Hail is an open-source software framework for scalably and flexibly analyzing large-scale genetic data sets (https://github.com/broadinstitute/hail).

Estimating fold-enrichment in AJ population compared to reference populations in ExAC

Statistical methods: Population Isolate Enrichment Mixture Model (PIEMM)

To estimate which alleles are enriched in AJ compared to alleles in reference population groups in ExAC we developed a statistical method we refer to as the **P**opulation **I**solate **E**nrichment **M**ixture **M**odel (**PIEMM**). This model estimates the proportion of observed alleles in the population isolate that are enriched or depleted as well as the shift in distribution of those alleles relative to the reference populations.

Using the number of alternate and reference alleles observed in AJ non-IBD samples and in the population (NFE, AFR or AMR) with the highest frequency frequency from ExAC we compute a bias corrected log odds ratio estimate, $\hat{\beta}_i$, and its standard error, $\hat{\mathrm{SE}}_i$, for odds of the alternate allele as

$$\hat{\beta}_i = log(OR_i) = \log\left([(0.5 + \mathrm{ALT}_{\mathrm{AJ}})\,(0.5 + \mathrm{REF}_{\mathrm{ExAC}})]/[(0.5 + \mathrm{REF}_{\mathrm{AJ}})\,(0.5 + \mathrm{ALT}_{\mathrm{ExAC}})]\right), \text{ and}$$

$$\hat{\mathrm{SE}}_i^{\,2} = 1/\left(0.5 + \mathrm{REF}_{\mathrm{AJ}}\right) + 1/\left(0.5 + \mathrm{REF}_{\mathrm{ExAC}}\right) + 1/\left(0.5 + \mathrm{ALT}_{\mathrm{AJ}}\right) + 1/\left(0.5 + \mathrm{ALT}_{\mathrm{ExAC}}\right).$$

These formulas can be understood in a Bayesian framework to be mean posterior estimates of β_i and SE_i based on the ExAC and observed AJ allele frequencies and a prior distribution of Beta(½, ½). This is the Jeffreys prior distribution for the parameter of a binomial distribution.

We use these summary statistics to estimate:

i) which proportion of alleles observed in AJ belong to each of three groups:
- (0) 'similar' - AJ allele has similar frequency to reference populations,
- (1) 'enriched' - AJ allele has shifted increased frequency to reference populations,
- (2) 'depleted' - AJ allele has shifted decreased frequency to reference populations; and

ii) the extent to which enriched and depleted groups are shifted in frequency from the reference populations.

Details: Let $\hat{\beta}_i$ and $\hat{\mathrm{SE}}_i$ be the log odds ratio and standard error estimates of protein-coding allele i in AJ with respect to the reference population with the highest allele frequency. Let γ_i be a variable that is either 0, 1, or 2. Our PIEMM is the following mixture model:

$$
\begin{aligned}
&\pi \sim \text{Dirichlet}(1,1,1)\\
&\gamma_i|\pi \sim \text{Multinomial}(1,\pi)\\
&\mu_0 = 0\\
&\mu_1 \sim \mathcal{N}\left(2, 0.5^2\right)\\
&\mu_2 \sim \mathcal{N}\left(-2, 0.5^2\right)\\
&\sigma_0^2 = 0.2^2\\
&\sigma_1^2 \sim \text{IG}\left(1,1\right)\\
&\sigma_2^2 \sim \text{IG}\left(1,1\right)\\
&\hat{\beta}_i|\gamma_i, \mu_{\gamma_i}, \sigma_{\gamma_i}^2, \hat{\text{SE}}_i{}^2 \sim \mathcal{N}\left(\hat{\beta}_i; \mu_{\gamma_i}, \sigma_{\gamma_i}^2 + \hat{\text{SE}}_i{}^2\right).
\end{aligned}
$$

Motivation of parameters and distribution: The group membership of each AJ protein-coding allele is unknown in advance. As a result the proportion π of the alleles belonging to the similar, enriched, or the depleted group (characterized by an unknown shift in mean μ and variance σ^2) needs to be estimated. Our prior for π is the uniform distribution on the 2d simplex (also known as the $\text{Dirichlet}(1,1,1)$ distribution) so as to not favor *a priori* any particular value of π .

The prior for the variance parameter σ^2 is the inverse gamma distribution with hyperparameter values $\alpha = 1$ and $\beta = 1$ for the enriched and the depleted group. This distribution is relatively flat between 0.3 and 1, and thus covers well the region where we expect the variance parameter to reside. The prior for the variance parameter σ^2 for the 'similar' group is fixed to 0.2^2, thus including alleles in the similar group where small deviation of enrichment is observed. The prior shift hyperparameter values for the 'depleted' and the 'enriched' group are -2 and +2, respectively, indicating separation in shift.

MCMC algorithm: We use a Gibbs sampler, an approximation algorithm, to approximate the posterior distribution of the parameters of the PIEMM. Superscripts for the variables denote their value after the corresponding iteration. Let l be the group index whenever it is not explicitly given.

1. Initialize $\pi^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \left(\sigma_1^2\right)^{(0)}, \left(\sigma_2^2\right)^{(0)}$, and $\gamma_i^{(0)}$ for all i.
2. Repeat for $t = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$

   (a) For $i = 1, 2, \ldots, n_{\text{alleles}}$, generate $\gamma_i^{(t)} \sim \text{Multinomial}\left(1, p_i^{(t)}\right)$ where $p_{il}^{(t)} = \frac{\pi_l^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_l^{(t-1)}, \left(\sigma_l^2\right)^{(t-1)} + \hat{\text{SE}}_i{}^2\right)}{\pi_0^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; 0, \hat{\text{SE}}_i{}^2\right) + \pi_1^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_1^{(t-1)}, \left(\sigma_1^2\right)^{(t-1)} + \hat{\text{SE}}_i{}^2\right) + \pi_2^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_2^{(t-1)}, \left(\sigma_2^2\right)^{(t-1)} + \hat{\text{SE}}_i{}^2\right)}$, for $l = 0, 1, 2$.

   (b) Generate $\pi^{(t)} \sim \text{Dirichlet}\left(1 + n_0, 1 + n_1, 1 + n_2\right)$, where $n_l = \sum_i I\left(\gamma_i^{(t)} = l\right)$, for $l = 0, 1, 2$.

   (c) Update:

   $(\sigma_l^2)^{(t)} \sim \text{IG}\left(1 + 1/2 \sum_i I\left(\gamma_i^{(t)} = l\right), 1 + 1/2 \sum_i I\left(\gamma_i^{(t)} = l\right)\left(\hat{\beta}_i - \mu_l^{(t-1)}\right)^2\right)$, and

   $\mu_l^{(t)} \sim \mathcal{N}\left(a, b\right)$, where

$a = \left[\left(\mu_{l,\text{prior}} \Big/ \sigma^2_{l,\text{prior}}\right) + \left(\sum_i \hat{\beta}_i I\left(\gamma_i^{(t)} = l\right) \Big/ \left(\sigma_l^2\right)^{(t)}\right)\right] \Big/ \left[\left(1 \Big/ \sigma^2_{l,\text{prior}}\right) + \left(\sum_i I\left(\gamma_i^{(t)} = l\right) \Big/ \left(\sigma_l^2\right)^{(t)}\right)\right]$ and
$b = \left[\left(1 \Big/ \sigma^2_{l,\text{prior}}\right) + \left(\sum_i I\left(\gamma_i^{(t)} = l\right) \Big/ \left(\sigma_l^2\right)^{(t)}\right)\right]^{-1}$.

We run the algorithm for $n_{\text{burn}} + n_{\text{iter}} = 1000$ iterations and discard the first $n_{\text{burn}} = 100$ iterations from the analysis. Figure S3 shows a typical example of a plot used to evaluate the performance of the algorithm across all the iterations.

To estimate allele enrichment in AJ compared to reference populations we used 2,178 non-IBD Ashkenazi Jewish samples, after sample and relatedness QC.

We calculated alternate allele frequencies for the Ashkenazi Jewish population and used allele frequency information for NFE (n=33,370), AFR (n=5,203), and AMR (n=5,789) available from ExAC release 0.3 dataset (n_{total}=60,706) and focused on alleles where allele frequency information was available for AJ and the reference populations. For the enrichment plot we focused on alleles with estimated frequency of at least 0.002 in AJ ($n_{alleles}$=103,878) and with alleles observed with an estimated frequency of at least .0001 in the reference populations with depth of coverage of at least 20X in at least 80% of the samples in ExAC.

Overlap of enriched alleles with ClinVar. We harmonized the XML and TXT releases of the ClinVar database (April 11, 2016 data release) [22] into a single tab-delimited text file using scripts that we have released publicly (https://github.com/macarthur-lab/clinvar). Briefly, we normalized variants using a Python implementation of vt normalize[57] and de-duplicated to yield a dataset unique on chromosome, position, reference, and alternate allele. A variant was considered 'pathogenic' if it had at least one assertion of either Pathogenic or Likely Pathogenic for any phenotype. A variant was considered 'conflicted' if it had at least one assertion of Pathogenic or Likely Pathogenic, and at least one assertion of Benign or Likely Benign, each for any phenotype. By these criteria, ClinVar contained n=42,226 identified as pathogenic and non-conflicted. Intersecting with our dataset revealed that 151 belonged to the AJ enriched group with high probability (>.5).

Assessing Crohn's disease association of protein-coding variation that may contribute to difference in disease prevalence in AJ. We focused Crohn's disease association analysis of protein-coding variant to alleles that may account for difference in disease prevalence in AJ population to reference populations. To do so we focused on alleles with high probability of belonging to the enriched group. We included all samples with ADMIXTURE estimated AJ ancestry fraction of at least 0.4 (we excluded any samples that had alternative ancestry fraction of at least .4 in any other group). Samples with Ulcerative Colitis (n=700), unspecified and Indeterminate Colitis (n=86) were excluded from subsequent analysis. This resulted in a dataset of 4,899 AJ samples (1,855 Crohn's disease and 3,044 non-IBD).

Study-specific association analysis was performed with Firth bias-corrected logistic regression test[32,33] and four principal components as covariates using the software package EPACTS version 3.2.6 [34]. Minimum minor allele count (≥1) and variant call rate (≥0.8) filters were used.

For meta-analysis we combined association statistics using the Bayesian models and frequentist properties proposed in Band et al[35], which is a normal approximation to the logistic regression likelihood suggested by Wakefield[58]. As the authors of Band et al. indicate one way of thinking about the approach is that it uses the study-wise estimated log-odds ratio (beta) and its standard error as summary statistics of the data. For each

model of association we assume a prior on the log odds ratio which is normally distributed around zero with a standard deviation of 0.2). By changing the prior on the covariance (or correlation) in effect sizes between studies we can formally compare models where: 1) the effects are independent across studies, and 2) the effects are correlated equally between studies. For each model we can obtain a Bayes factor (BF) for association by comparing it with the null model where all the prior weight is on an effect size of zero. We report p-value approximation using the Bayes factor as a statistic for model 2 where the effects are correlated between studies.

Association statistics were combined based on association analysis across three study groups: 1) AJ (1,855 CD and 3,044 non-IBD samples); 2) NFE (2,296 CD and 2,770 non-IBD); and 3) Finnish (FINN) (210 CD and 9,930 non-IBD samples) for a total of 4,361 CD samples and 15,744 non-IBD samples.

Conditional haplotype based testing and variable selection for *NOD2* alleles. In the conditional haplotype based testing (--chap) analysis we used PLINK v1.08p[56] and set a minimum haplotype frequency of .001 (--mhf). We used PLINKSEQ (https://atgu.mgh.harvard.edu/plinkseq/), an open-source C/C++ library for working with human genetic variation data, and the Python bindings implemented in pyPLINKSEQ to perform Bayesian Model Averaging (BMA). We applied BMA[38] using the R package 'BMA' (https://cran.r-project.org/web/packages/BMA/BMA.pdf).

Polygenic risk scores. The polygenic risk scores were calculated for the international inflammatory bowel diseases consortium European samples. Details of these samples including the QC procedures were described in previous publications [42]. We used reported effect size estimates from 124 CD alleles including those reported in a previously published study[12] and four variants in *IL23R* from a recent fine-mapping study[42], and excluding variants in *NOD2* and *LRRK2*. We used 454 AJ controls; 1,938 AJ CD; 35,007 non-Jewish controls and 20,652 non-Jewish CD samples. Polygenic risk scores were calculated using array genotype data as the sum of the log odds ratio of the variants associated with CD. Scores for missing genotypes were replaced by the imputed expected value using PLINK [56]. Variants in *NOD2* and *LRRK2* were excluded from the analysis to assess whether polygenic signal was independent.

Let PRS_i be the polygenic risk score of individual i, assuming additive effects on the log-odds scale then

$$\mathrm{PRS}_i = \sum_{m=1}^{M} \widehat{\beta_m} G_{i,m},$$

where $\widehat{\beta_m}$ denotes the estimated log odds ratio for variant m and $G_{i,m}$ denotes the genotype dosage of individual i for variant m.

In the setting where effects are non-additive, i.e. a genotype-specific effect model, then

$$\mathrm{PRS}_i^* = \sum_{m=1}^{M} \left[\widehat{\beta_m^{\mathrm{Het}}} 1_{[\mathrm{Het}]} + \widehat{\beta_m^{\mathrm{Hom}}} 1_{[\mathrm{Hom}]} \right].$$

For now, we consider the additive scenario, and later we return to the setting where non-additive effects exist, which is relevant for quantifying the contribution of *NOD2* alleles to differences in genetic risk between two populations.

The estimated expected PRS value for an individual in population j is

$\widehat{\mathbb{E}[\mathrm{PRS}]}_j = \sum_{i=1}^{N_j} \frac{\mathrm{PRS}_i}{N_j}$,

where N_j is the number of individuals sampled in population j. Substituting equation for PRS_i and rearranging terms simplifies the equation as a function of variant frequency:

$\widehat{\mathbb{E}[\mathrm{PRS}]}_j = \sum_{i=1}^{N_j} \sum_{m=1}^{M} \frac{\widehat{\beta_m} G_{i,m}}{N_j}$,

$\widehat{\mathbb{E}[\mathrm{PRS}]}_j = \sum_{m=1}^{M} \left(\sum_{i=1}^{N_j} \frac{\widehat{\beta_m} G_{i,m}}{N_j} \right)$,

$\widehat{\mathbb{E}[\mathrm{PRS}]}_j = \sum_{m=1}^{M} \left(\widehat{\beta_m} \sum_{i=1}^{N_j} \frac{G_{i,m}}{N_j} \right)$, where $\sum_{i=1}^{N_j} \frac{G_{i,m}}{N_j} = 2\widehat{f_{m,j}}$, and $\widehat{f_{m,j}}$ denotes the frequency of variant m in population j. Thus, the estimated expected PRS value of an individual in population j is $\widehat{\mathbb{E}[\mathrm{PRS}]}_j = \sum_{m=1}^{M} \left(2\widehat{\beta_m} \widehat{f_{m,j}} \right)$.

Assume that we are interested in the expected difference in PRS between two individuals, say from population 1 being AJ and population 2 being NFE. Also, assume that the effect size of variant m is shared across both populations. Then, using the estimated expected PRS value we define estimated expected difference in PRS as the difference in estimated expected PRS value between two populations:

$\widehat{\mathbb{E}[\text{Difference PRS}]} = \widehat{\mathbb{E}[\mathrm{PRS}]}_{\mathrm{AJ}} - \widehat{\mathbb{E}[\mathrm{PRS}]}_{\mathrm{NFE}}$,

$\widehat{\mathbb{E}[\text{Difference PRS}]} = \sum_{m=1}^{M} 2\widehat{\beta_m} \left(\widehat{f_{m,\mathrm{AJ}}} - \widehat{f_{m,\mathrm{NFE}}} \right)$,

which can be used to get an estimated contribution of a variant m to the difference in polygenic risk score between two populations,

$\widehat{\mathbb{E}[\text{Difference PRS}]}_m = 2\widehat{\beta_m} \left(\widehat{f_{m,\mathrm{AJ}}} - \widehat{f_{m,\mathrm{NFE}}} \right)$.

To rank variants according to their relative contribution to differences in genetic risk we included the *NOD2* and *LRRK2* alleles, used the list of estimated effect size from the published studies [12,42], and estimates from this study (Supplementary Table 3).

If we replace PRS* for PRS

$\widehat{\mathbb{E}[\mathrm{PRS}^*]}_j = \sum_{m=1}^{M} \frac{\left(\sum_{i=1}^{N_j} \left[\widehat{\beta_m^{\mathrm{Het}}} 1_{[\mathrm{Het}]} + \widehat{\beta_m^{\mathrm{Hom}}} 1_{[\mathrm{Hom}]} \right] \right)}{N_j}$,

$= \sum_{m=1}^{M} \left[2\widehat{f_m} \left(1 - \widehat{f_m} \right) \widehat{\beta_m^{\mathrm{Het}}} + \widehat{f_m}^2 \widehat{\beta_m^{\mathrm{Hom}}} \right]$.

Then, the estimated expected difference in PRS* when non-additive effects exist is

$$\widehat{\mathbb{E}[\text{Difference PRS}^*]} = \sum_{m=1}^{M}\left[2\widehat{\beta_m^{\text{Het}}}\left(\left[\widehat{f_m^{\text{AJ}}} - \widehat{f_m^{\text{NFE}}}\right] - \left[\widehat{f_m^{\text{AJ}}}^2 - \widehat{f_m^{\text{NFE}}}^2\right]\right) + \widehat{\beta_m^{\text{Hom}}}\left(\widehat{f_m^{\text{AJ}}}^2 - \widehat{f_m^{\text{NFE}}}^2\right)\right].$$

Estimating fold difference in prevalence for a population with shift in expected genetic risk.

Assuming log-additive effects in the logit risk model the disease probability for an individual is given as $p = (1+\exp(-\eta))^{-1}$, where η tends towards a normal distribution with parameters $\mu = \log(p_0/(1-p_0)) + \sum_{m=1}^{M} 2f_m\beta_m$ and $\sigma^2 = 2\sum_{m=1}^{M} f_m(1-f_m)\beta_m^2$[41]. Here p_0 refers to a baseline disease probability.

We can see that μ may be expressed in terms of the expected polygenic risk score, i.e. $\mu = \log(p_0/(1-p_0)) + \mathbb{E}[\text{PRS}]$. In the setting where $\mathbb{E}[PRS] = 0$ then $\mathbb{E}[p] = (1 + exp(-\log(p_0/(1-p_0))) = p_0$.
To evaluate the impact of a shift in the expected value of polygenic risk score to the expected value of μ we can express the shift as:
$\mathbb{E}[\text{Difference } \mu] = \mathbb{E}[\text{Difference PRS}]$. We can compute new values of p for new values of μ to obtain a fold-increase in prevalence for a population that has undergone such a shift.

We see that this requires a value to be chosen for p_0 and that $\log(p_0/(1-p_0))$ can be represented as a baseline risk score value β_0. To get an estimate of the absolute prevalence of CD in the AJ population, we must choose a baseline β_0, where p_0 represents the expected prevalence with zero non-baseline alleles in the population[41], to which we add a contribution from multiple non-baseline alleles to calculate: 1) an individual's probability of disease, or 2) the expected prevalence of the disease in the population.

Once we have chosen a value for β_0, we can calculate the ratio of expected prevalence as follows. First, use the means (μ_{AJ} and μ_{NAJ}) and variances (σ^2_{AJ} and σ^2_{NAJ}) of risk scores as calculated above to calculate the probability density function of the disease prevalence. In the case of the AJ population, we have

$$f(p) = \frac{d\eta}{dg}\frac{1}{\sigma_{\text{AJ}}}\phi\left(\frac{\eta - \mu_{\text{AJ}}}{\sigma_{\text{AJ}}}\right) = \frac{1}{\sigma_{\text{AJ}}p(1-p)}\phi\left(\frac{1}{\sigma_{\text{AJ}}}log\left(\frac{p}{1-p}\right) - \frac{\mu_{\text{AJ}}}{\sigma_{\text{AJ}}}\right)$$

where η is the risk score associated with prevalence p, g is the link function, so $p = g(\eta) = \left(1+e^{-\eta}\right)^{-1}$, and ϕ is the standard normal density function.

Next, we integrate to get $\int_0^1 p \cdot f(p)dp = \mathbb{E}[p_{\text{AJ}}]$. Finally, we can calculate $\mathbb{E}[p_{\text{NAJ}}]$ in a similar way, and divide the expected prevalence in the AJ population by that in the non-AJ population to get the prevalence ratio, $\mathbb{E}[p_{\text{AJ}}]/\mathbb{E}[p_{\text{NAJ}}]$.

The value of β_0 = -20.5 was chosen in order to obtain a prevalence in the non-AJ population of ~0.5%. At this value of β_0, the ratio of prevalence in the AJ population to that in the non-AJ population was estimated to be 1.5 ($\mathbb{E}[p_{\mathrm{AJ}}] = 0.82\%$, $\mathbb{E}[p_{\mathrm{NAJ}}] = 0.55\%$).

For different choices of β_0, however, this ratio may vary, as the relationship between probability of disease and risk score is non-linear. Figure S6 shows how the values of the disease prevalence and their ratio vary as β_0 is changed. We see that the ratio values range from 1.46 to 1.52 for different values of β_0 with a range of baseline prevalence of .001 to .01 - the range of prevalence estimates for Crohn's disease[3,46,59].

To further understand the effect that choosing a logit based model had on the results, a comparison of the standard logit and probit models was done using the values inferred from the logit model. No full scale probit modelling was done in this analysis, so the values found with the probit model represent only a close approximation of the expected results.

In the logit model for population analysis, we may assume that individual risk scores are chosen from a normal distribution $\mathcal{N}(\mu_{\mathrm{logit}}, \sigma^2_{\mathrm{logit}})$ where μ_{logit} and σ_{logit} represent the mean and standard deviation of the risk scores as defined above. From here, we may calculate the probability density function of probit model risk scores η_{probit} based on that of logit model risk scores η_{logit} as

$$f\left(\eta_{\mathrm{probit}}|\mu_{\mathrm{logit}}, \sigma_{\mathrm{logit}}\right) = f(\eta_{\mathrm{logit}}|\mu_{\mathrm{logit}}, \sigma_{\mathrm{logit}}) d\eta_{\mathrm{logit}}/d\eta_{\mathrm{probit}}$$

and use this to calculate μ_{probit} and σ_{probit}, the estimated mean and standard deviation of the risk scores in the probit model. Using these values, we obtain a probability distribution for the frequency of disease in the populations using the probit model.

While the logit model yielded a prevalence ratio of 1.506, the probit estimation yielded a prevalence ratio of 1.5136, with similar expected prevalence values ($\mathbb{E}[p_{\mathrm{AJ}}] = 0.823\%$, $\mathbb{E}[p_{\mathrm{NAJ}}] = 0.544\%$). These values demonstrate that individual logit and probit analyses would likely give extremely similar results for values of interest. The complete probability densities under the logit and probit models can be seen in Figure S7.

Further, it is interesting to compare the relationship between values of risk scores in the two models. For values of risk scores between -1 and 1 in the logit model, the relationship to those in the probit model is highly linear, with a formula of $\eta_{\mathrm{probit}} = 0.6223 \cdot \eta_{\mathrm{logit}}$, with $r^2 = 1.0000$. This formula may be used to impute single values in one model or the other assuming that the estimated total risk score is otherwise close to zero, and the imputed value is low. It is worth noting, however, that this does not work for all values of η_{logit}, as the relationship between risk score in the logit and probit models deviates from this simple linear model when the risk score values are large.

Difference in prevalence between AJ and NFE attributed to implicated variants. The difference in prevalence due to multiple alleles can be computed as

$$\text{Prevalence difference} = \frac{((p_2 - p_1) - (i_2 - i_1))}{(p_2 - p_1)},$$

where p_j denotes the disease prevalence in population j and i_j denotes the disease prevalence without the risk factors in population j, which according to Moonesinghe et al. [60] is

$$i_j = \frac{p_j}{\prod_{m=1}^{M}(1+f_{m,j}(\mathrm{GRR}_m-1))^2}$$

where GRR_m denotes the genotype relative risk for variant m.

We model the CD prevalence accounted for by CD associated enriched protein-altering alleles separately in both AJ and non-AJ European and determine the amount that CD prevalence would be reduced if this variant were absent from each population. Two-to-four fold difference in prevalence has been documented to exist between these groups[13].We use a standard log-additive effect model and assume an odds ratio and uncertainty as estimated in the meta-analysis for each variant in both populations.

To estimate the difference in prevalence between two populations attributed to genetic risk factors when non-additive effects exist

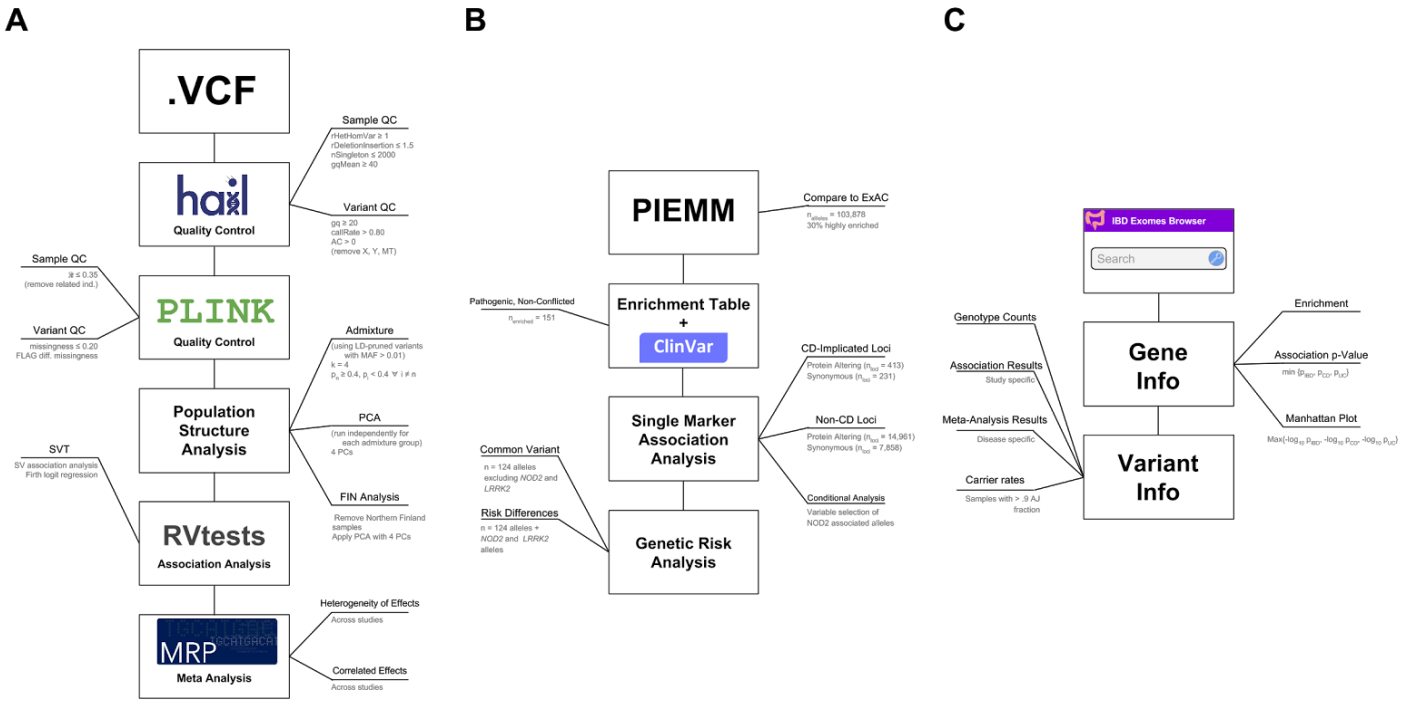
$$i_j = \frac{p_j}{\prod_{m=1}^{M}\left(1+2f_{m,j}\left(\mathrm{GRR}_m^{\mathrm{Het}}-1\right)+f_{m,j}^2\left(\mathrm{GRR}_m^{\mathrm{Hom}}-1\right)\right)}.$$

S2. Online resources
IBD Exomes Browser: http://ibd.broadinstitute.org
Clinvar table: https://github.com/macarthur-lab/clinvar/blob/master/output/clinvar.tsv
ExAC sites VCF: ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/

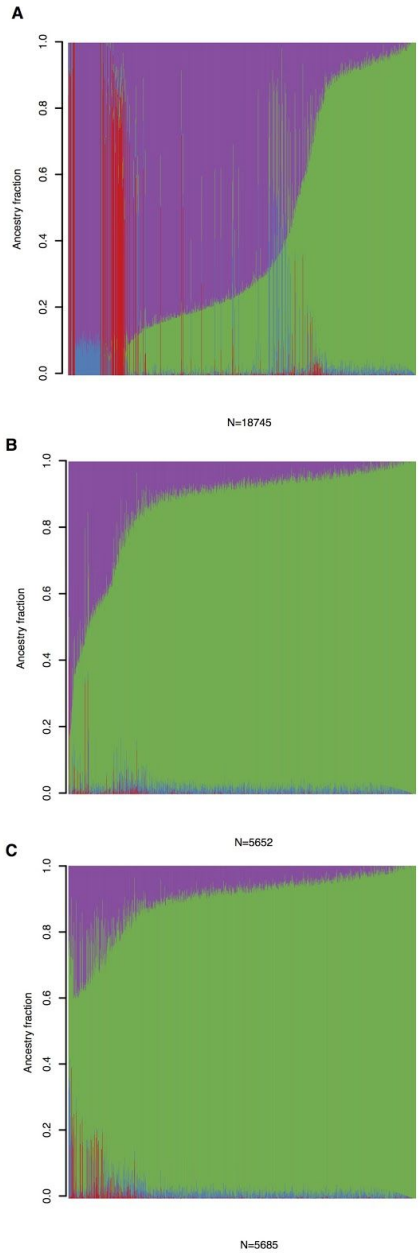
S3. Consortium members
NIDDK IBD Genetics Consortium

Abraham C[34], Achkar JP[35,36], Bitton A[37], Boucher G[5], Croitoru K[38], Fleshner P[39], Kugathasan S[40], Limbergen JV[41], Milgrom R[38], Proctor D[34], Regueiro M[24], Schumm PL[42], Sharma Y[43], Stempak JM[38], Targan SR[4], Wang MH[22]

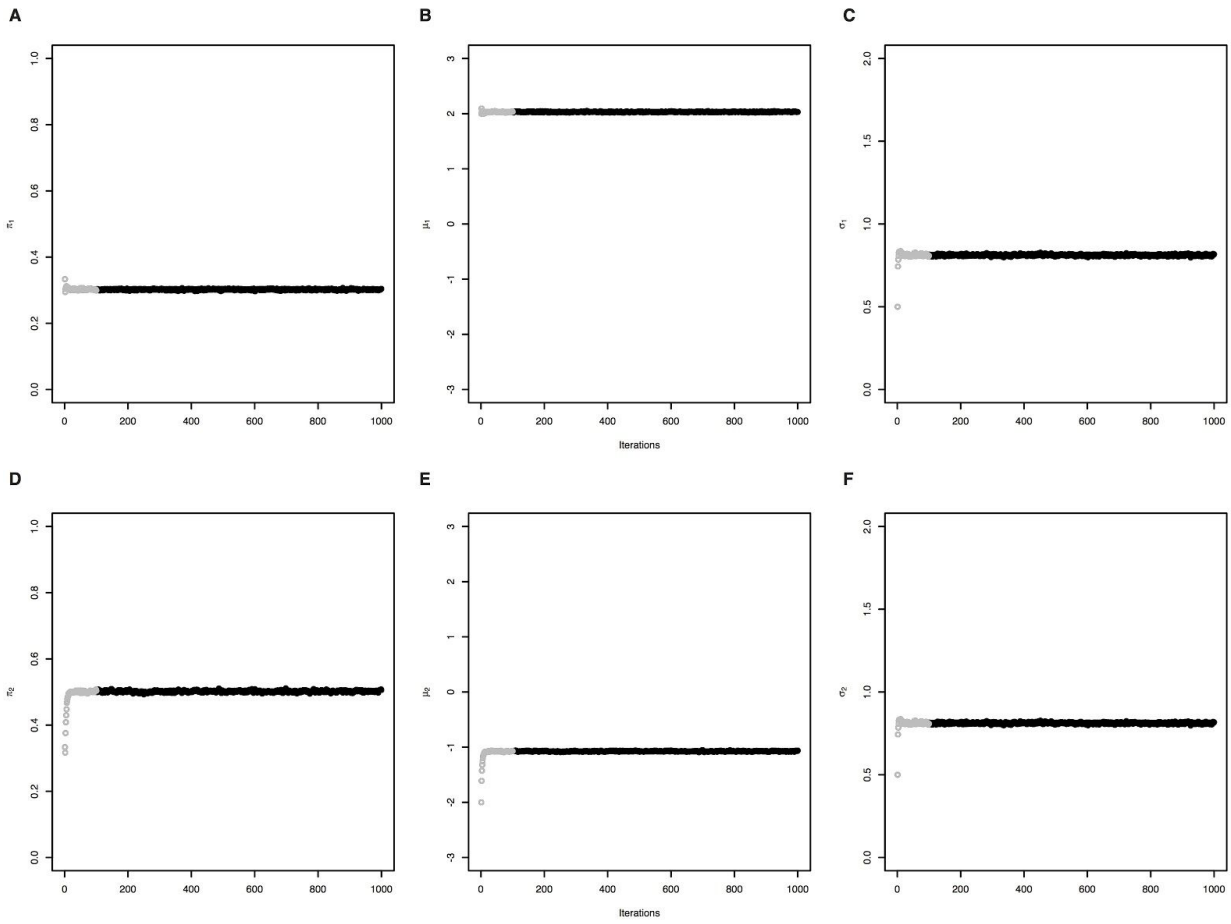
[34]Section of Digestive Diseases, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, USA, [35]Department of Gastroenterology and Hepatology, Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio, USA, [36]Department of Pathobiology, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA, [37]Division of Gastroenterology, Royal Victoria Hospital, Montréal, Québec, Canada, [38]Inflammatory Bowel Disease Group, Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada, [39]Department of Transplantation and Liver Surgery, University of Helsinki, Helsinki, Finland, [40]Department of Pediatrics, Emory University School of Medicine, Atlanta, Georgia, USA, [41]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, Toronto, Ontario, Canada, [42]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA, [43]Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY



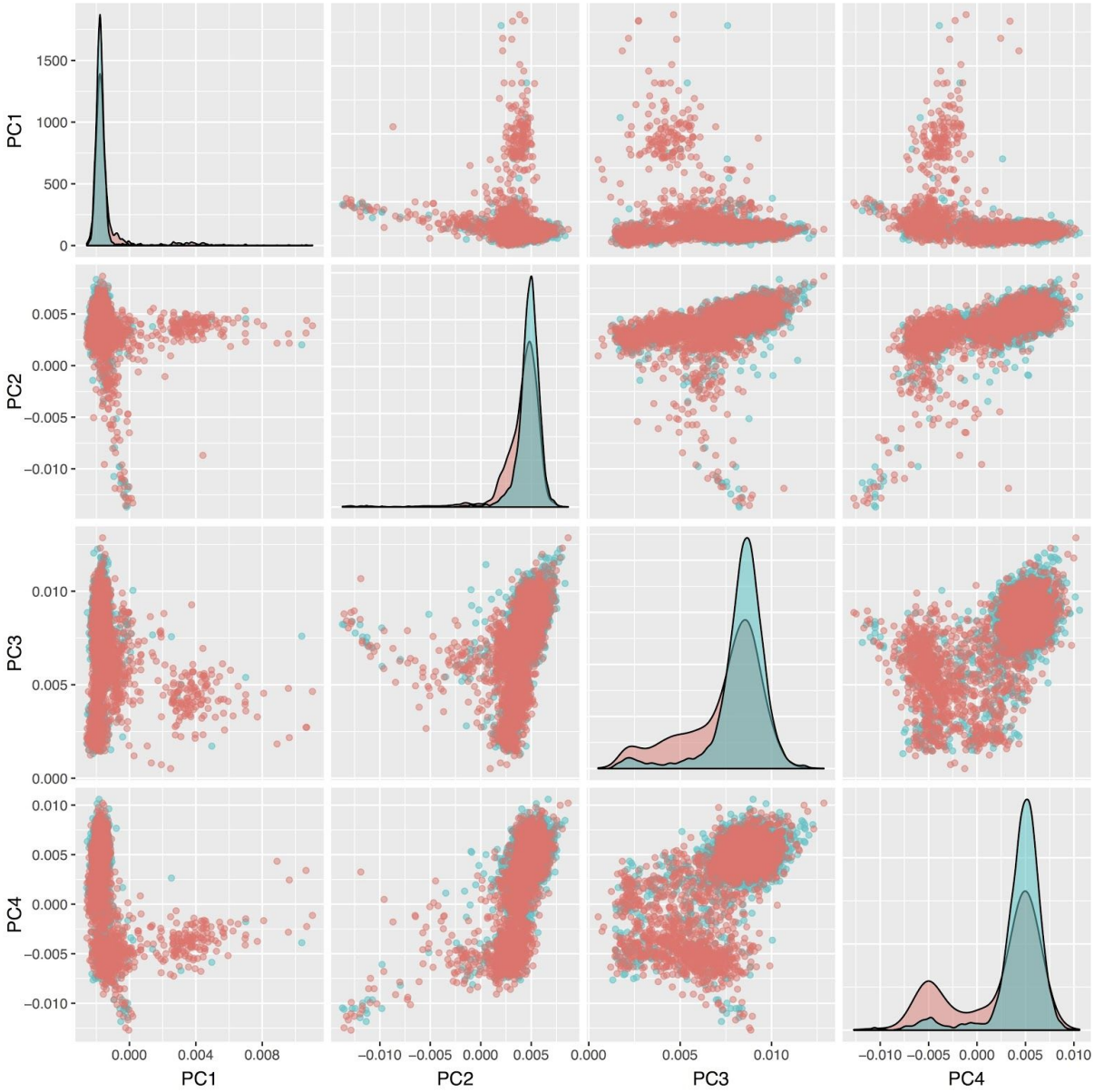
Supplementary Figure 1. Analysis workflow diagram. A) Quality control, population structure, and association analysis workflow. **B)** AJ specific analysis workflow. **C)** Results and summary statistics are uploaded to the IBD browser hosted at: http://ibd.broadinstitute.org - a website that contains a gene and variant search engine, a "Gene Info" landing page that contains a manhattan plot and additional summary statistics, and a detailed "Variant Info" page that contains additional information about the alleles identified in our exome sequencing studies.



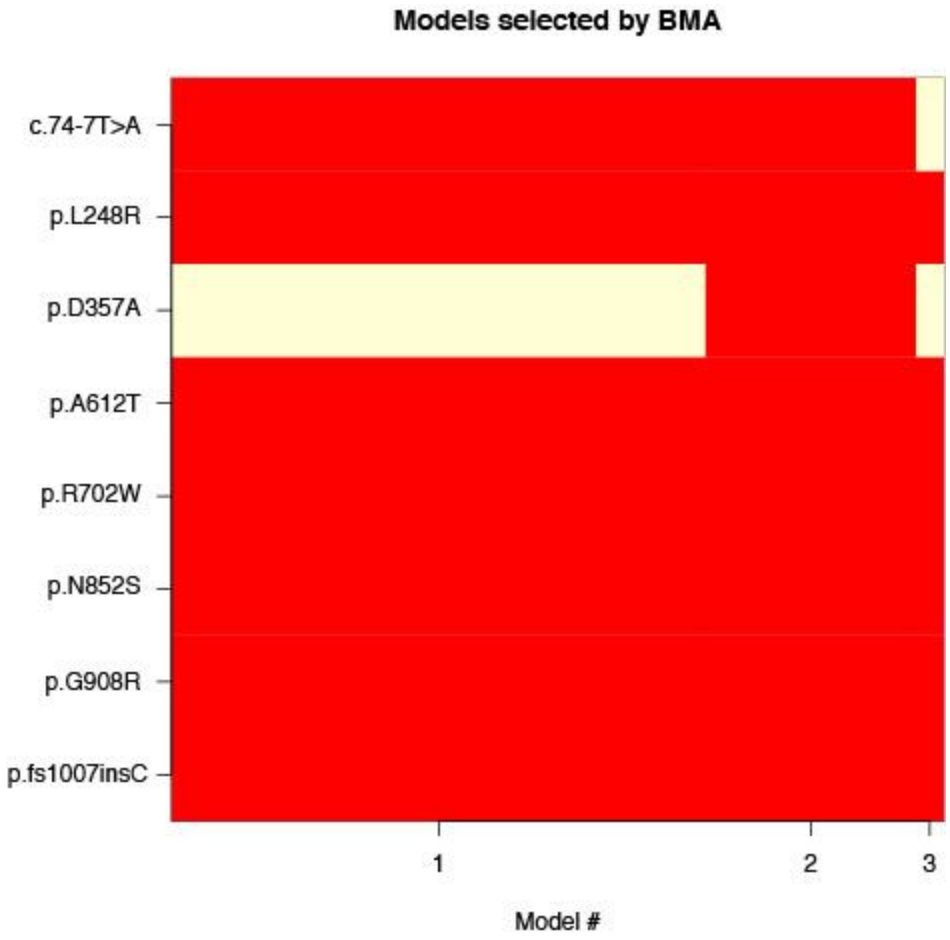
Supplementary Figure 2. Admixture plots. A) Admixture plot, with 4 groups (K=4), for all samples exome sequenced. On the y-axis shown are ancestry mixture fractions for each of the samples (on x-axis) exome sequenced. **B)** Admixture plot for samples with self-reported Ashkenazi Jewish ancestry. **C)** Admixture plot for final samples used in the AJ analysis. All plots are ordered by ancestry fraction mostly loading for AJ samples (green). Group mostly loading for NFE samples is highlighted in magenta, East-Asian in red and African-American in blue.



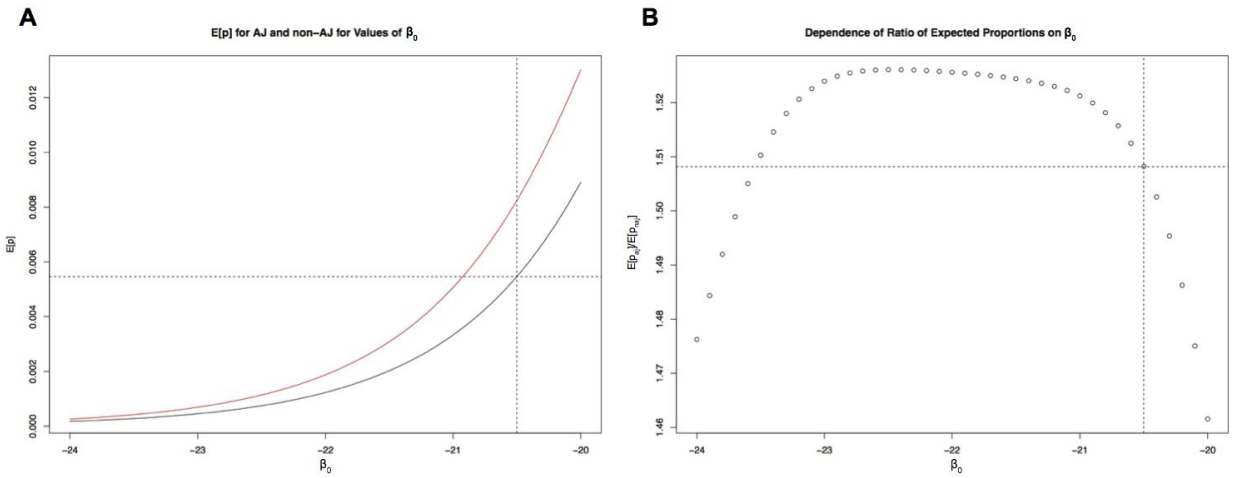
Supplementary Figure 3. Parameter estimates for MCMC chain. When applying MCMC algorithms to a data set it is customary to show the performance of the algorithm across all stages of the experiment (including the burn-in). We show that the PIEMM algorithm generates stable proportion estimates for the two groups: "enriched" **(A-C)** and "depleted" **(D-F)** for parameters: π **(A, D)**, the proportion of alleles belonging to the group; μ **(B, E)**, the shift of the distribution belonging to the group; and σ **(C, F)**, the standard deviation of the distribution belonging to the corresponding group. For each group we demonstrate the parameter estimate during the burn-in (100 iterations, gray circles) and non burn-in (900 iterations, black circles) stage of the experiment used to obtain the parameter estimates reported in the manuscript.



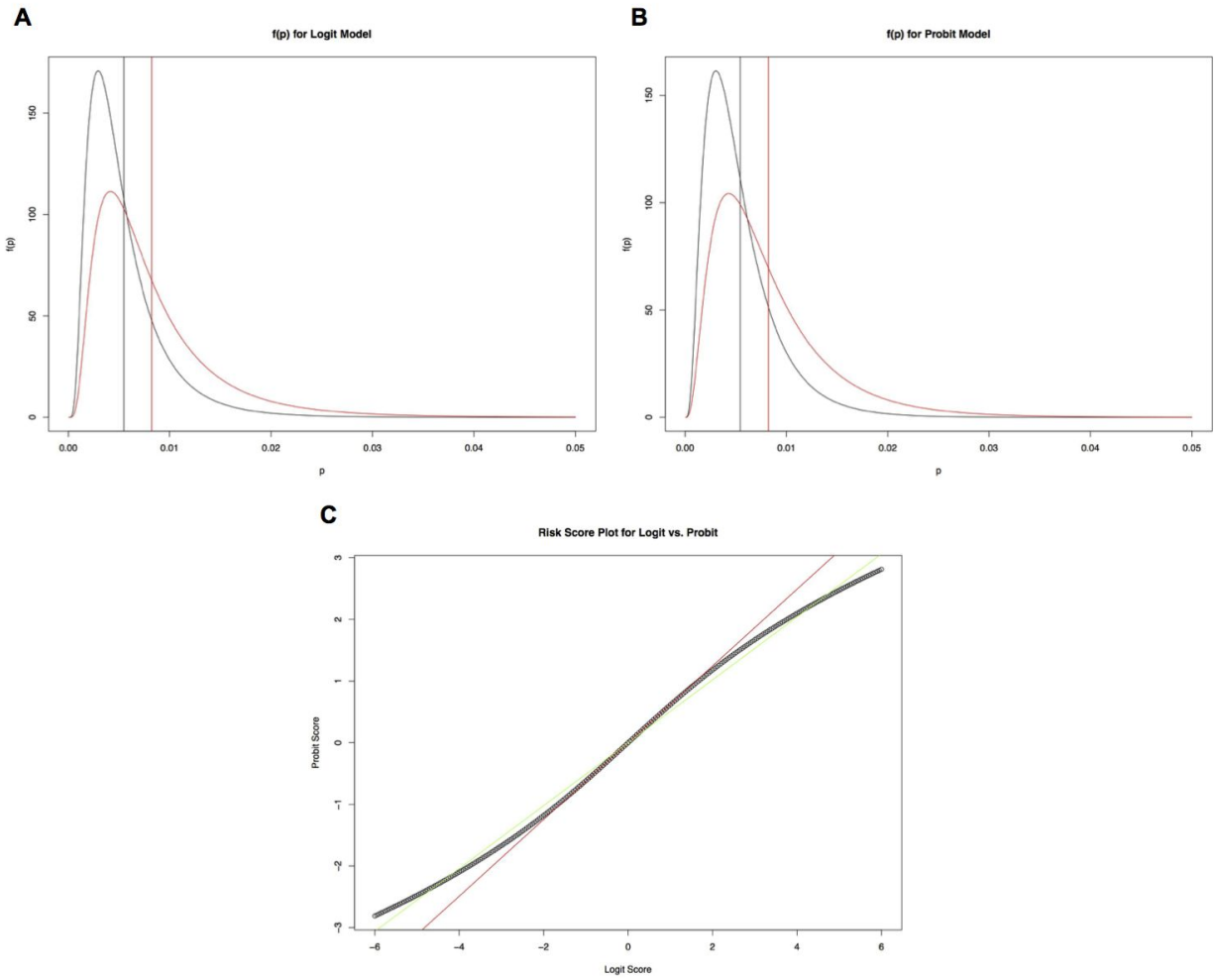
Supplementary Figure 4. Principal components plot for 5,685 AJ individuals. Density plots for each of the PCs is shown (diagonal) separately for IBD cases (red) and controls (blue). Pairwise scatter plots are shown for PC1-PC4 separately for IBD cases and controls.



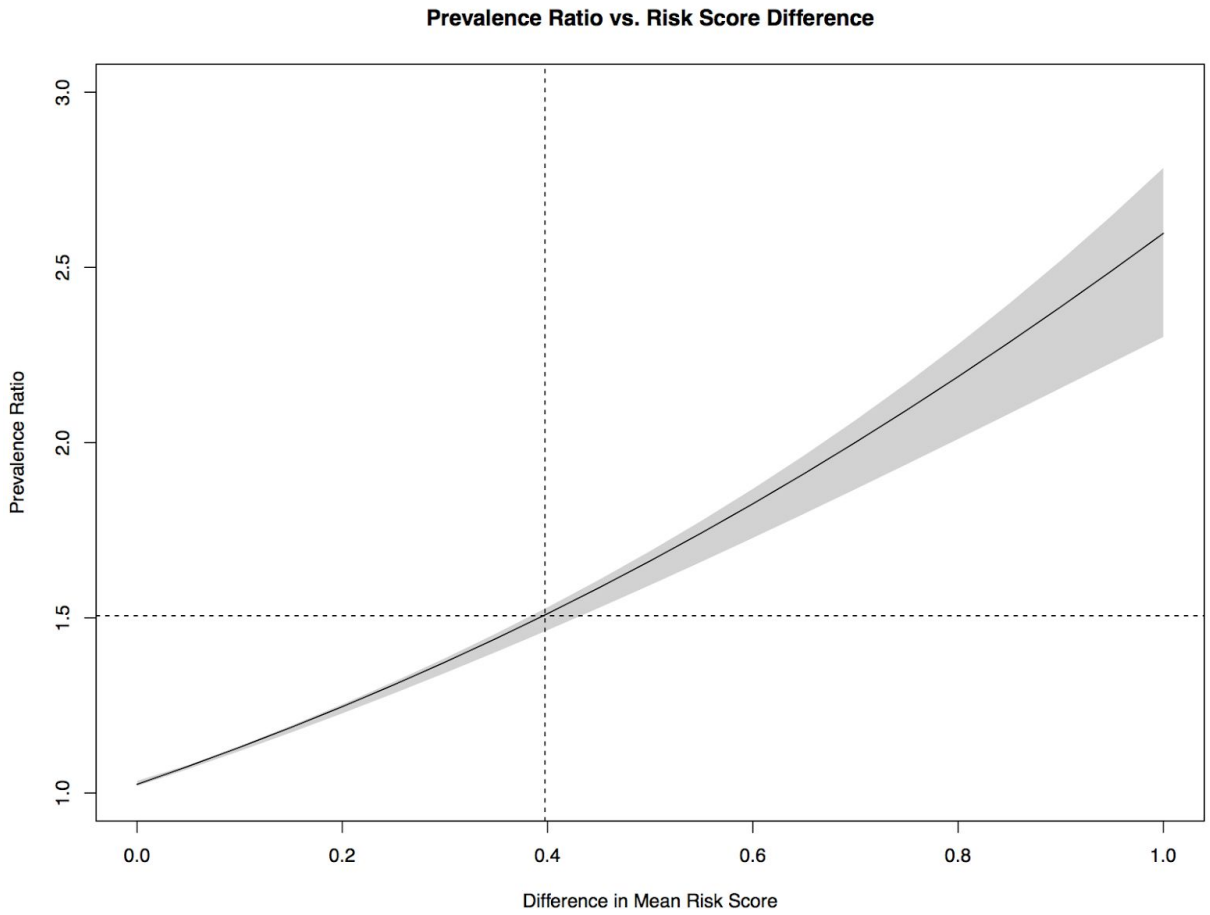
Supplementary Figure 5. **Variable selection using Bayesian model averaging (BMA).** Eight protein-altering alleles with evidence of association and their corresponding membership for the models with high probability (> 0.01) after applying BMA, which accounts for the model uncertainty inherent in the variable selection problem by averaging over the best models in the model class according to approximate posterior model probability. Model #1 corresponds to the model where 7/8 alleles have a non-zero effect (p.D357A is not included) with approximate posterior model probability of .692. Model #2 corresponds to the model where 8/8 alleles have a non-zero effect with approximate posterior model probability of .273. Model # 3 corresponds to the model where 6/8 alleles have a non-zero effect (p.D357A and c.74-7T>A are not included) with approximate posterior model probability of .035.



Supplementary Figure 6. Dependence of population prevalence on β_0. A) Expected values of disease prevalence in non-AJ (black) and AJ (red) populations. The value of β_0 = -20.5 was chosen to approximate the prevalence of disease in the non-AJ population at around 0.005. **B)** Relationship between expected prevalence ratio and choice of β_0. The chosen value of β_0 = -20.5 corresponds to a ratio of 1.506.



Supplementary Figure 7. Probit and logit model analysis. A) Probability distribution of prevalence in non-AJ (black) and AJ (red) populations given the calculated mean and variance of risk score given by the logit model. **B)** Equivalent probability distributions given the calculated mean and variance of the risk score in the probit model. **C**) A comparison of risk score values in logit and probit models. The green line corresponds to a linear fit for the entire shown region, and the red line corresponds to a linear fit for the linear range around zero.



Supplementary Figure 8. The relationship between expected differences in genetic risk score and expected fold difference in disease prevalence. Vertical line represents the estimated expected difference in genetic risk between AJ and non-AJ European population (0.397). For varying levels of expected difference in genetic risk we compute the expected fold-difference in prevalence. The shaded region marks the range of prevalence ratios obtained by varying β_0 in a region of reasonable estimates (-24 to -20).

Variant	HGVS	P-value
16:50763778:G:GC	p.fs1007insC	6.50E-24
16:50745926:C:T	p.R702W	2.18E-05
16:50756540:G:C	p.G908R	1.52E-20
16:50750810:A:G	p.N852S	1.78E-05
16:50745656:G:A	p.A612T	3.98E-08
16:50733392:T:A	c.74-7T>A	2.07E-05
16:50744565:T:G	p.L248R	0.00198
16:50744892:A:C	p.D357A	0.0016

Supplementary Table 3. Conditional haplotype-based testing in *NOD2*. HGVS nomenclature for each allele (HGVS) and corresponding p-values are shown for independent effects given the haplotype formed by residual variants.