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

Rivas, Manuel A.

2018-05


http://hdl.handle.net/10138/245419
https://doi.org/10.1371/journal.pgen.1007329

Downloaded from Helda, University of Helsinki institutional repository.
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.
Insights into the genetic epidemiology of Crohn’s and rare diseases in the Ashkenazi Jewish population


1 Medical and Population Genetics, Broad Institute, Cambridge, MA, United States of America.
2 Department of Biomedical Data Science, Stanford University, Stanford, CA, United States of America.
3 Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, United States of America.
4 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.
5 Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland.
6 Translational Genomics Unit, F. Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, United States of America.
7 Hadassah-Hebrew University Medical Center, Endocrinology and Metabolism Service Department of Internal Medicine, Jerusalem, Israel.
8 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States of America.
9 Department of Genetics and Medicine, Albert Einstein College of Medicine, Bronx, NY, United States of America.
10 Faculty of Natural Sciences, University of Haifa, Haifa, Israel.
11 Division of Medicine, University College London, London, United Kingdom.
12 UCL Genetics Institute, University College London, London, United Kingdom.
13 Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany.
14 Gastroenterology Department, Saint-Antoine Hospital, AP-HP, UPMC Univ Paris, Paris, France.
15 Department of Internal Medicine, University Hospital Schleswig-Holstein, Kiel, Germany.
16 Gastroenterology Practice, Minden, Germany.
17 IBD Pharmacogenetics, Royal Devon and Exeter NHS Trust, Exeter, United Kingdom.
18 Peninsula College of Medicine and Dentistry, Exeter, United Kingdom.
19 Juliet Keidan Institute of Pediatric Gastroenterology and Nutrition, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel.
20 Research Programs Unit, Immunobiology, and Department of Medical and Clinical Genetics, University of Helsinki, Helsinki, Finland.
21 Department of Medicine, Division of Gastroenterology, Helsinki University Hospital, Helsinki, Finland.
22 Department of Medicine, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland.
23 Department of Neurology, Massachusetts General Hospital, Boston, MA, United States of America.
24 Meyerhoff Inflammatory Bowel Disease Center, Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, United States of America.
25 Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States of America.
26 Division of Gastroenterology, Hepatology and Nutrition, Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States of America.
27 Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, United States of America.
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 Wellcome Trust Centre for Human Genetics, Oxford University, Oxford, United Kingdom.
33 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, United Kingdom.
34 Bonei Olam, Center for Rare Jewish Genetic Disorders, Brooklyn, NY, United States of America.
35 Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease and Center for...
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 (hosted in part at https://ibd.broadinstitute.org, also available in gnomAD at http://gnomad.broadinstitute.org). We estimate that 34% of protein-coding alleles present in the Ashkenazi Jewish population at frequencies greater than 0.2% are significantly more frequent (mean 15-fold) than their maximum frequency observed in other reference populations. Arising via a well-described founder effect approximately 30 generations ago, this catalog of enriched alleles can contribute to differences in genetic risk and overall prevalence of diseases between populations. As validation we document 148 AJ enriched protein-altering alleles that overlap with "pathogenic" ClinVar alleles (table available at https://github.com/macarthur-lab/clinvar/blob/master/output/clinvar.tsv), including those that account for 10–100 fold differences in prevalence between AJ and non-AJ populations of some rare diseases, especially recessive conditions, 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 attempt to evaluate whether strong acting rare alleles, particularly protein-truncating or otherwise large effect-size 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 enriched in AJ (p < 0.005), 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 (sites VCF available via FTP at ftp.broadinstitute.org/pub/ExAC_release/release0.3/) to pinpoint genetic variation that contributes to variable disease predisposition across populations.

Author summary

The Ashkenazim are a people with ancestry in northern-European Jewish groups. A founder effect caused a bottleneck in this population approximately one thousand years ago, leading to a reduced genetic diversity and an increased frequency of certain alleles, particularly those associated with Mendelian diseases. However, the impact of these alleles on common diseases like Crohn's disease is less apparent. Our study provides a resource of enriched alleles and demonstrates the utility of exome sequencing data for understanding genetic risk factors in specific populations.
ago, resulting in a group of enriched alleles in their genetic makeup. A higher documented prevalence of Crohn’s Disease in the Ashkenazim indicates that some enriched alleles may confer risk of having this disease. By studying which genes are enriched, and which of these contribute to Crohn’s Disease risk, we are better able to understand the genetic architecture of the affected population, and of the disease itself. Further, we are able to develop a resource containing tables of significantly enriched alleles that are known or suspected to contribute to other disease.

Introduction

Genetic population isolates like the Ashkenazim, Jews who trace their ancestry to eleventh century central European Jewish groups[1], have previously facilitated the mapping of alleles contributing to human disease predisposition[2–5]. The documented 2–4 fold enrichment of Crohn’s Disease (CD) prevalence in the Ashkenazi Jewish (AJ) population[6,7] motivated the use of exome sequencing and genome-wide array data to evaluate the degree to which bottle-neck-enriched protein-altering alleles and unequivocally implicated common variants contribute an excess CD genetic risk to AJ[6]. Despite the progress in mapping genes and alleles for rare diseases with increased prevalence in the AJ population, precise estimates of the risk-allele frequency and the carrier rate in the AJ population have not yet been determined[8]. Through this study, we provide a frequency resource of protein-coding alleles from over 2,000 non-CD AJ individuals with low admixture that will serve to improve interpretation of rare disease risk alleles in the AJ population and which we employ to discover new Crohn’s risk alleles by comparison to 1855 AJ Crohn’s cases.

Results

We generated a jointly called whole-exome sequence dataset consisting of 18,745 individuals from international Inflammatory Bowel Disease (IBD) and non-IBD cohorts[9,10] (S1 Fig). Given the increased prevalence of Crohn’s disease in the AJ population, our global sequencing efforts had specifically included 5,652 individuals self-reporting as Jewish and, as we aimed to focus on variation observed in the AJ population in comparison to reference populations in ExAC[9,11] (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[12].

To identify AJ individuals and estimate admixture fractions we used a set (n = 21,066) of LD-pruned common variants (MAF>1%, see Supplementary Note for additional details) filtered for genotype quality (GQ>20). The 18,745 individuals were assigned to four groups (K = 4) using ADMIXTURE (further described in Supplementary Note, also see S3 Fig). One group of 5,685 individuals was found consisting mostly (84%) of self-reported AJ individuals, while 3,522 of these individuals were further found with high ancestry fraction (>0.9) mapping to this group (S2 Fig, S1 Table). Thus, many self-reported AJ individuals were not included, as they did not have high enough ascertained AJ ancestry fraction. As 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 individuals that passed QC and relatedness filtering and had AJ ancestry fraction (genotype ancestry grouping closely with other AJ individuals) of >0.9. Our study includes exomes throughout Europe and Israel but the vast majority...
(86%) of these high ancestry fraction AJ individuals were collected in major US cities including
Los Angeles, Boston, Baltimore, and New York (S2 Table).

To explore AJ exome population genetics, including proportion of enriched alleles and
degree of enrichment, we used the observed alternate allele counts and total number of alleles
available from ExAC release 0.3 dataset \( \text{N}_{\text{total}} = 60,706; \text{NFE} (n = 31,902; \text{after excluding AJ}
\text{individuals from ExAC}), \text{AFR} (n = 5,203), \text{and AMR} (n = 5,789) \}. We focused on protein-cod-
ing alleles with estimated allele frequency of at least 0.002 and less than .1 in AJ \( \text{N}_{\text{alleles}} =\)
73,228; practical cutoff of what could be statistically defined as convincing enrichment, see S4
Fig), and applied a one-sided Fisher’s exact test on allele counts (see Supplementary Note), to
classify the observed alleles into two groups: “enriched” or “not enriched”. This analysis identi-
ﬁed 34% of protein-coding alleles as significantly enriched, with mean 15-fold increased odds
of the alternate allele compared to other populations. Different proportions of alleles belong to
the enriched group depending on variant annotation: 36% for predicted protein-truncating
variants (PTV); 38% for predicted protein-altering variants (PRA); and 31% for synonymous
variants. The substantially higher PTV+PRA:synonymous ratio observed in the enriched cate-
gory is consistent with those alleles being drawn randomly from a large pool of much rarer
alleles (where the functional:synonymous ratio is higher[3]) and abruptly boosted in frequency
(Fig 1, \( p < 10^{-16} \)) across comparisons of PTV and PRA to synonymous variants, two propor-
tion test, Supplementary Note). Since much rarer alleles have a higher probability of being
damaging (e.g., they have a higher missense/synonymous ratio), the advantage to gene map-
ing arises from the fact that enriched alleles of a certain frequency are more damaging/delete-
rious on average than non-enriched alleles of the same frequency.

Additionally, we may expect that a “depleted” set of alleles arises from the founder effect,
but in reality, many of these already rare variants are simply eliminated during the bottleneck.
Of course, it is more difficult and less interesting to search for depleted alleles, as their absence
provides no opportunity to obtain significant statistics on population enrichment or disease
association.

We intersected the list of protein-coding alleles identified in the AJ exome sequencing
study with alleles reported to be pathogenic with no conflicting evidence \( n = 42,226 \) in Clin-
Var[14] resulting in 148 alleles found both in ClinVar and with \( p \)-value less than .005 of
belonging to the AJ enriched set (S1 Data File). In OMIM, 48 of the 148 alleles included docu-
tmentation of a disease subject with AJ ancestry (Table 1). This set of enriched alleles includes
all of the major AJ mutations for 8 diseases described in the American College of Medical
Genetics and Genomics 2008 screening guideline study[15]. In the setting of autosomal reces-
sive disorders these differences in population allele frequencies may contribute a factor pro-
portional to the squared enrichment difference to genetic risk and prevalence between
populations (see Supplementary Note). For instance, a 19-fold enriched frameshift indel, p.
Tyr427IlefsTer5, in \( \text{HEXA} \), contributes a 361-fold enrichment in genetic risk in AJ to non-AJ
population to Tay-Sachs disease. Enrichment in this large adult Ashkenazi exome database
reinforces recent publications of founder mutations for rare pediatric disorders including
\( \text{FKTN} \) (Walker Warburg syndrome)[16], \( \text{CCDC65} \) (Primary ciliary dyskinesia)[17], \( \text{TMEM216} \)
(Joubert syndrome)[18], \( \text{C11orf73} \) (Leukoencephalopathy)[19], \( \text{PEX2} \) (Zellweger syndrome)
[20], \( \text{VPS11} \) (Hypomyelination and developmental delay)[21] and \( \text{BBS2} \) (Bardet-Biedl syn-
drome)[22]. While many alleles on this pathogenic list may demonstrate incomplete pene-
trance (as in the case of p.V726A in \( \text{MEFV} \)[23] for Familial Mediterranean fever) and some
may not show recessive inheritance, this resource should provide considerable assistance in
gene discovery and clinical genetic screening in AJ (S2 Data File).

To assess whether AJ-enriched protein-coding alleles also contribute to the established dif-
fERENCE in CD genetic risk we performed case-control association analyses. Since individuals
with only partial AJ ancestry will still carry bottleneck-enriched alleles, here we included samples with estimated AJ ancestry fraction > 0.4 (Supplementary Note, S2 Fig), resulting in a dataset of 4,899 AJ samples (1,855 Crohn’s disease and 3,044 non-IBD). To improve ability to detect a true association, we performed a meta-analysis with CD and non-IBD case-control exome sequencing data from two additional ancestry groups: 1) non-Finnish European (NFE) (2,296 CD and 2,770 non-IBD); and 2) Finnish (FIN) (210 CD and 9,930 non-IBD samples) from a separate callset described in a previous publication[24] for a total of 4,361 CD samples and 15,744 non-IBD samples. By calling additional non-AJ samples, we hoped to discern which of the AJ-enriched alleles contributed a significant risk factor across all populations. The meta-analysis performed across several populations described should mitigate biases by confirming consistency in effect size across these population groups.

Study-specific association analysis was performed with Firth bias-corrected logistic regression[25,26] and four principal components as covariates using the software package EPACTS [27] (S5 Fig). We combined association statistics in a meta-analysis framework using the Bayesian models in Band et al.[28]. 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

---

**Fig 1. Enrichment of alleles discovered in AJ exome sequencing project.** A) Histogram 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 bar plot of the expected number of alleles belonging to the two groups we analyzed: 1) enriched (green) and 2) not enriched (white). B) Bar plots of estimated percentage of alleles belonging to the two groups we analyzed for all protein-coding (ALL), synonymous (SYN), protein-altering (PRA), and protein-truncating variants (PTV). An estimate of 34% of protein-coding alleles observed in AJ have a mean shift of 15-fold increased odds of the alternate allele compared to other reference populations. This observation is supported by the property that compared to intergenic variants, coding variants tend to be younger for a given frequency and the more pathogenic a variant, the younger it is, therefore tending to be population specific[13].

https://doi.org/10.1371/journal.pgen.1007329.g001
### Table 1. Forty-eight ClinVar "pathogenic" alleles enriched in AJ.

<table>
<thead>
<tr>
<th>Variant</th>
<th>HGVS</th>
<th>Gene</th>
<th>Enrichment Odds Ratio</th>
<th>AJ AF</th>
<th>Max ExAC AF</th>
<th>Curated Traits</th>
<th>Inheritance</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:3293310:A&gt;G</td>
<td>p.Val726Ala</td>
<td>MEFV</td>
<td>26.08</td>
<td>0.0416</td>
<td>0.0017</td>
<td>Familial Mediterranean fever</td>
<td>AR</td>
</tr>
<tr>
<td>5:15072135:A&gt;C</td>
<td>p.Gly87Val</td>
<td>SLC36A2</td>
<td>3.51</td>
<td>0.0414</td>
<td>0.0122</td>
<td>Hyperglycinuria</td>
<td>AD</td>
</tr>
<tr>
<td>1:155205634:T&gt;C</td>
<td>p.Asn409Ser</td>
<td>GBA</td>
<td>11.16</td>
<td>0.0296</td>
<td>0.0027</td>
<td>Susceptibility to Lewy body dementia, Gaucher’s disease, Susceptibility to late onset Parkinson’s disease</td>
<td>AR</td>
</tr>
<tr>
<td>4:187201412:T&gt;C</td>
<td>p.Phe301Leu</td>
<td>F11</td>
<td>47.17</td>
<td>0.0273</td>
<td>0.0006</td>
<td>Hereditary factor XI deficiency</td>
<td>AR</td>
</tr>
<tr>
<td>13:20765353:CA&gt;C</td>
<td>p.Leu56Argfs</td>
<td>GJB2</td>
<td>11.16</td>
<td>0.0296</td>
<td>0.0027</td>
<td></td>
<td>AR</td>
</tr>
<tr>
<td>4:187195347:T&gt;G</td>
<td>p.Glu135Ter</td>
<td>F11</td>
<td>28.20</td>
<td>0.0195</td>
<td>0.0007</td>
<td>Factor XI deficiency</td>
<td>AR</td>
</tr>
<tr>
<td>5:178686540:G&gt;A</td>
<td>p.Arg498Lys</td>
<td>CHEX2</td>
<td>50.06</td>
<td>0.0103</td>
<td>0.0002</td>
<td>Hereditary cancer, multiple types</td>
<td>Risk factor</td>
</tr>
<tr>
<td>10:99371368:G&gt;T</td>
<td>p.Glu315del</td>
<td>HOGA1</td>
<td>29.28</td>
<td>0.0101</td>
<td>0.0003</td>
<td>Primary hyperoxaluria</td>
<td>AR</td>
</tr>
<tr>
<td>7:117282620:G&gt;A</td>
<td>p.Glu2019Ser</td>
<td>LRRK2</td>
<td>23.64</td>
<td>0.0085</td>
<td>0.0004</td>
<td>Cystic fibrosis</td>
<td>AR</td>
</tr>
<tr>
<td>17:3402294:AC</td>
<td>p.Glu285Ala</td>
<td>ASPA</td>
<td>40.36</td>
<td>0.0076</td>
<td>0.0002</td>
<td>Canavan disease</td>
<td>AR</td>
</tr>
<tr>
<td>9:111662096:AG</td>
<td>c.2204+6T&gt;C</td>
<td>IKBKAP</td>
<td>45.22</td>
<td>0.0199</td>
<td>0.0005</td>
<td>Familial dysautonomia</td>
<td>AR</td>
</tr>
<tr>
<td>22:29091207:G&gt;A</td>
<td>p.Ser428Phe</td>
<td>CHEX2</td>
<td>19.14</td>
<td>0.0122</td>
<td>0.00064</td>
<td>Tay-Sachs disease</td>
<td>AR</td>
</tr>
<tr>
<td>12:14421038:G&gt;A</td>
<td>p.Arg49Cys</td>
<td>PRB3</td>
<td>16.12</td>
<td>0.0189</td>
<td>0.0012</td>
<td>Salivary peroxidase</td>
<td>AR</td>
</tr>
<tr>
<td>10:99371368:G&gt;T</td>
<td>p.Glu315del</td>
<td>HOGA1</td>
<td>29.28</td>
<td>0.0101</td>
<td>0.0003</td>
<td>Primary hyperoxaluria</td>
<td>AR</td>
</tr>
<tr>
<td>7:10755951:G&gt;T</td>
<td>p.Gly229Cys</td>
<td>DLD</td>
<td>26.55</td>
<td>0.0046</td>
<td>0.0002</td>
<td>Maple syrup disease</td>
<td>AR</td>
</tr>
<tr>
<td>15:72638920:G&gt;GATA</td>
<td>p.Tyr427IlefsTer5</td>
<td>HEXA</td>
<td>19.14</td>
<td>0.0122</td>
<td>0.00064</td>
<td>Tay-Sachs disease</td>
<td>AR</td>
</tr>
<tr>
<td>17:3402294:AC</td>
<td>p.Glu285Ala</td>
<td>ASPA</td>
<td>40.36</td>
<td>0.0076</td>
<td>0.0002</td>
<td>Canavan disease</td>
<td>AR</td>
</tr>
<tr>
<td>9:97934315:T&gt;A</td>
<td>c.456+1G&gt;A</td>
<td>CNGA3</td>
<td>26.11</td>
<td>0.0074</td>
<td>0.0003</td>
<td>Achromatopsia</td>
<td>AR</td>
</tr>
<tr>
<td>13:3291437:GT&gt;G</td>
<td>p.Arg1982Argfs</td>
<td>BRCA2</td>
<td>27.57</td>
<td>0.0069</td>
<td>0.0003</td>
<td>Hereditary cancer, multiple types</td>
<td>Risk factor</td>
</tr>
<tr>
<td>9:108383230:G&gt;GA</td>
<td>p.Phe390Ilefs</td>
<td>FKTN</td>
<td>32.62</td>
<td>0.0067</td>
<td>0.0002</td>
<td>Limb-girdle muscular dystrophy-dystroglycanopathy</td>
<td>AR</td>
</tr>
<tr>
<td>12:40734202:G&gt;A</td>
<td>p.Trp1282Ter</td>
<td>CACNA1A</td>
<td>20.64</td>
<td>0.0064</td>
<td>0.0003</td>
<td>Parkinson’s disease</td>
<td>Risk factor</td>
</tr>
<tr>
<td>15:72638920:G&gt;GATA</td>
<td>p.Tyr427IlefsTer5</td>
<td>HEXA</td>
<td>19.14</td>
<td>0.0122</td>
<td>0.00064</td>
<td>Tay-Sachs disease</td>
<td>AR</td>
</tr>
<tr>
<td>17:410559:G&gt;T</td>
<td>p.Arg83Cys</td>
<td>G6PC</td>
<td>11.04</td>
<td>0.0062</td>
<td>0.0006</td>
<td>Maple syrup disease</td>
<td>AR</td>
</tr>
<tr>
<td>11:641543:G&gt;T</td>
<td>p.Arg498Lys</td>
<td>CHEX2</td>
<td>42.75</td>
<td>0.0069</td>
<td>0.0002</td>
<td>Fanconi anemia</td>
<td>AR</td>
</tr>
<tr>
<td>3:1727372:G&gt;A</td>
<td>p.Arg283Glu</td>
<td>SPATA16</td>
<td>78.35</td>
<td>0.0046</td>
<td>0.0002</td>
<td>Maple syrup disease</td>
<td>AR</td>
</tr>
<tr>
<td>12:4931253:CCDC65</td>
<td>p.Ile293Profs</td>
<td>CCDC65</td>
<td>25.75</td>
<td>0.0048</td>
<td>0.0002</td>
<td>Ciliary dyskinesia without situs inversus</td>
<td>AR</td>
</tr>
<tr>
<td>6:8087862:G&gt;C</td>
<td>p.Arg183Pro</td>
<td>BCKDH</td>
<td>29.42</td>
<td>0.0046</td>
<td>0.0002</td>
<td>Maple syrup disease</td>
<td>AR</td>
</tr>
<tr>
<td>10:5607147:G&gt;A</td>
<td>p.Arg245Ter</td>
<td>PCDH15</td>
<td>26.58</td>
<td>0.0046</td>
<td>0.0002</td>
<td>Usher syndrome</td>
<td>AR</td>
</tr>
<tr>
<td>15:72638920:G&gt;GATA</td>
<td>p.Tyr427IlefsTer5</td>
<td>HEXA</td>
<td>52.65</td>
<td>0.0044</td>
<td>0.0001</td>
<td>Tay-Sachs disease</td>
<td>AR</td>
</tr>
<tr>
<td>11:86017416:G&gt;C</td>
<td>p.Val54Leu</td>
<td>C11orf73</td>
<td>47.03</td>
<td>0.0037</td>
<td>0.0001</td>
<td>Hypomyelinating leukodystrophy</td>
<td>AR</td>
</tr>
</tbody>
</table>

(Continued)
separately assessed CD associations of enriched protein-altering (PRA) and synonymous (SYN) alleles in protein-coding genes in CD implicated GWAS loci ($n_{\text{gwas,pra}} = 351$; $n_{\text{gwas,syn}} = 167$), and outside implicated GWAS loci ($n_{\text{non-gwas,pra}} = 12,529$; $n_{\text{non-gwas,syn}} = 6,202$, Fig 2). See Methods and Materials for a description of these loci.

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+p.fs1007insC)\(^{[29,30]}\), in addition to newly implicated alleles ($NOD2$: p.A612T, $p = 2.8 \times 10^{-9}$; c.74-7T>A, $p = 1.4 \times 10^{-4}$; p.L248R, $p = 6.4 \times 10^{-4}$; p.D357A, $p = 0.0011$; $LRRK2$: p.G2019S, $p = 0.0014$, a Parkinson’s disease risk allele\(^{[31]}\)). 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\(^{[32]}\) for variable selection, both of which suggested independent effects for all alleles (S6 Fig, S3 Table).

Deviation from additivity can contribute additionally to individual risk but has been difficult to document in complex disease associations with modest ORs. Despite the functional relationship between $LRRK2$ and $NOD2$\(^{[33]}\), we do not observe deviation from additivity between $LRRK2$ and $NOD2$ ($p = 0.273$); that is, the effect of mutations in both $LRRK2$ and $NOD2$ is no greater than the sum of their individual effects. We assessed whether composite risk carriers (carrier of more than one variant allele) had evidence of deviation from additivity. Deviation from additivity has been reported for p.fs1007insC, p.G908R, and p.R702W in $NOD2$\(^{[34,35]}\). In our AJ exome sequencing data set we estimate a 1-hit effect equal to 1.82 (95% confidence interval [1.59, 2.07]) and a 2-hit effect equal to 8.24 (95% confidence interval [6.06, 11.21]); we found similar evidence for departure from additivity when restricting the analysis to the newly reported alleles only: $p = 0.00357$, odds ratio = 7.53). We confirmed this finding using the larger non-AJ Crohn’s disease ImmunoChip dataset to provide a more precise estimate of the 1-hit effect (OR = 2.17; 95% confidence interval [2.07, 2.27], S4 Table) and the non-additive 2-hit effects in $NOD2$ (OR = 9.93; 95% confidence interval [8.88, 11.13], S5 Table). We found no evidence of deviation from additivity for the associated protein-altering alleles in $LRRK2$ ($p = 0.418$).

Given that enriched genetic variants in $NOD2$ and $LRRK2$ contribute to differences in CD risk in AJ population, we next asked whether unequivocally established common variant associations contribute to differences in CD 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[36] and four variants in IL23R from a recent fine-mapping study [37], 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\(^{-16}\); Fig 3A; number of non-AJ controls = 35,007; number of AJ controls = 454), and as expected when performing the PRS analysis using OR calculated from non-Jewish subset of iCHIP data the signal still remains (p<10\(^{-16}\), S7 Fig). We observed a similar trend for the CD samples (0.54 s.d. higher; p<10\(^{-16}\); Fig 3B; number of non-AJ CD cases = 20,652; number of AJ CD cases = 1,938). We demonstrate this is not a systematic property of common risk alleles in AJ by running the same comparison using instead the comparable set of established schizophrenia associated alleles from the Psychiatric Genomics Consortium[38].
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, S6 Fig).

We applied the approach to all CD implicated alleles and observed that variants in GWAS loci annotated as \textit{IRGM}, \textit{LACC1}, \textit{NOD2}, \textit{MST1}, \textit{ATG16L1}, \textit{GCKR}, \textit{NKX2-3}, and \textit{LRRK2}\cite{36} contribute substantially (\(>0.01\)) to the increased genetic risk observed in AJ. It is possibly...
relevant that variants contributing to increased risk in AJ include many autophagy/intracellular defense genes (IRGM, ATG16L1, LRRK2), while those contributing to increased risk in non-AJ include many anti-fungal/Th17/ILC3 genes[39] (IL23R, IL12B, CARD9, TRAF3IP2, IL6ST, CEBPB; Fig 3C).

Both documented variability in the occurrence of CD over time[40,41] and substantial uncertainty in reported CD prevalence estimates[42,43] impact our ability to precisely estimate the overall contribution of genetics to the established difference in prevalence between populations. To interpret the impact of shifts in genetic risk score on differences in prevalence, we used the logit risk model[35] and evaluated a new estimate of disease probability, $p_{\text{new}}$, assuming an initial disease probability, $p_0$, and multiple values for the differences in genetic risk.

Assuming log-additive effects, and a logit-risk model, we estimate that the observed differences in genetic risk between the AJ and non-AJ populations 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 (S7–S9 Figs). To address the extent to which non-additive effects in NOD2 may impact the observed prevalence we assumed 1-hit and 2-hit odds ratios of 2.17 and 9.93, respectively. We attribute a 6.8% difference in the ratio of estimated disease prevalence in the AJ population to the deviation from additivity, suggesting a small effect on differences in population prevalence (Supplementary Note).

Discussion

Analyzing 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 rare diseases, many of which are transmitted via autosomal recessive inheritance. 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 (0.97 s.d. higher in controls, 0.54 s.d. higher in cases, $p<10^{-16}$ in both), independent of NOD2 and LRRK2[44]. Highly polygenic diseases are unlikely to have substantially altered incidence as a result of a bottleneck alone—for every enriched variant there are those depleted or lost entirely and population genetics simulations[45] suggest no systematic alteration of overall genetic burden as a function of a bottleneck. Thus, the strong (approximately 1.5-fold, see supplementary note) difference in Crohn’s incidence in concert with a systematic enrichment of risk-increasing alleles, unlikely to have arisen by chance, suggests non-random selection in the AJ population for higher CD risk alleles. It seems plausible that, rather than ‘selection for Crohn’s’ per se, this likely suggests a subset of Crohn’s risk alleles may contribute to a common biological process (e.g., a specific immune response) or phenotype that was positively selected for in AJ[46–48]. Such weak, widespread ‘polygenic selection’ has previously been observed with respect to height-associated SNPs in Europe[49], where drift alone could not explain the systematic enrichment of height-increasing alleles in populations of Northern Europe vs. Southern Europe. We found that CD risk alleles that are systematically elevated in AJ are not unusually elevated in another well-established founder population for which we have extensive genotype data (Finland). In Finns, Crohn’s risk alleles were not systematically enriched—they were if anything slightly depleted with 69 risk alleles at higher frequency in Finns than NFE and 79 risk alleles at lower frequency in Finns than NFE. We also demonstrate this is not a systematic property of common risk alleles in AJ by running the same comparison using instead the comparable set of established schizophrenia associated alleles from the Psychiatric Genetics Consortium[50]. We mapped 102 schizophrenia-
associated index SNPs to AJ frequency data and again observed no uneven distribution where risk alleles are systematically more or less common. In total, 52 risk alleles were at higher frequency in AJ than NFE and 50 risk alleles were higher frequency in NFE than AJ.

This study of CD in the AJ population confirms population-genetic expectations. First, recently bottlenecked populations are uniquely powered to discover alleles with markedly increases in frequency, and, as a consequence, contributors to differences in genetic risk across population groups. Second, while NOD2 and published common variant associations contribute substantially to the genetic risk of CD, other genes with causal alleles that failed to pass through the bottleneck are missed, consistent with predictions from Zuk et al [4].

We provide an exome frequency resource of protein-coding alleles in AJ along with estimates of population-specific enrichment. The sets of enriched alleles should be carefully considered when performing case-control analysis. Population structure can easily lead to false positive associations, especially for low frequency and rare variants, if the AJ:nonAJ ratio is slightly different in cases and controls. Our approach and this resource will likely catalyze our understanding of the medical relevance of enriched alleles in population isolates. Most importantly, the frequency reference provides critical guidance in pinpointing or excluding specific risk factors in individuals in clinical and research settings.

**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)[24]. 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[52,53]. We used the same exome sequencing dataset described in Rivas et al. [24]. 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.

**Identifying previously implicated GWAS loci**

CD implicated GWAS loci were those loci defined as reaching genome-wide significance in International IBD Genetics Consortium studies (Jostins, Ripke et al., Nature 2012) and (Liu et al., Nature Genetics 2015)—Credible sets of SNPs around index associations were defined as in (Huang et al., Nature 2017) for fine-mapped loci, and for others credible sets were defined as all SNPs with $r^2 > 0.6$ to the index variant. Genes within 50 kb of the span of credible set SNPs were considered “implicated” by GWAS.
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 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[12]. To identify AJ individuals and estimate admixture proportions we included a set (n = 21,066) of LD-pruned common variants (MAF > 1%) after filtering for genotype quality (GQ > 20) using the PLINK LD-pruning algorithm, whose description is available at http://pngu.mgh.harvard.edu/~purcell/plink/summary.shtml#prune.

For the parameters, we selected a window size of 50 SNPs, a window shift of 5 SNPs at each step, and the variance inflation factor (VIF) threshold equal to 2.

The 18,745 samples were assigned to four groups (K = 4), as ancestry was defined as having a single estimated ancestry fraction ≥ 0.4, and remaining three fractions < 0.4 (S2 Fig). Individuals mostly representing African/African-American and East-Asian ancestry (1,267 and 569 individuals respectively) were discarded from downstream analysis, as well as the 983 admixed individuals with none of the ancestry fractions ≥ 0.4. Thus, a total of 6,093 individuals were considered of Ashkenazi Jewish (AJ) ancestry, while 9,833 were considered to represent Non-Finnish Europeans (NFE). After sample QC and relatedness check, 5,685 individuals of Ashkenazi Jewish and 7,240 of non-Finnish European ancestry were found with valid IBD case/control status (S1 Table). Individuals with Ulcerative Colitis and unspecified and Indeterminate Colitis were further excluded, resulting in 4,899 AJ and 5,066 NFE individuals.

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 (> 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 (S1 Fig). 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 and sample relatedness analysis was done using PLINK[54]. Hail is an open-source software framework for scalably and flexibly analyzing large-scale genetic data sets (https://github.com/broadinstitute/hail). Allele balance was calculated using PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq/).

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

Statistical methods: Fisher’s exact test. To estimate which alleles are enriched in AJ compared to alleles in reference population groups in ExAC we applied Fisher’s exact test one-sided alternative (“greater”).

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 from ExAC we compute a bias corrected log odds ratio estimate, \( \tilde{\beta} \), and its standard error, \( \tilde{SE} \), for odds of the alternate allele as described in the Software DataPlot developed by the National Institute of Standards and Technology (http://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/logoddra.htm,

\[ \hat{\beta}_i = \log(OR_i) = \log \left( \frac{[0.5 + ALT_{AJ}] \cdot [0.5 + REF_{ExAC}]}{[0.5 + REF_{AJ}] \cdot [0.5 + ALT_{ExAC}]} \right), \] 

\[ SE_{\hat{\beta}_i}^2 = \frac{1}{0.5 + REF_{AJ}} + \frac{1}{0.5 + REF_{ExAC}} + \frac{1}{0.5 + ALT_{AJ}} + \frac{1}{0.5 + ALT_{ExAC}}. \]

Precisely, \( \hat{\beta}_i \) is the estimate of the log of the odds ratio of finding the alternate allele in AJ vs in the ExAC population with the highest allele frequency.

We classified a variant as ‘enriched’ if p-value was less than .05/73,228, where 73,228 is the number of variants analyzed with minor allele frequency between .002 and .1.

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 = 31,902; after excluding AJ individuals from ExAC), AFR (n = 5,203), and AMR (n = 5,789) available from ExAC release 0.3 dataset (n\text{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\text{alleles} = 106,377) 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)[14] 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[55] 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 148 belonged to the AJ enriched group with p-value less than .005.

**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[25,26] and four principal components as covariates using the software package EPACTS version 3.2.6[27]. 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[28], which is a normal approximation to the logistic regression likelihood suggested by Wakefield[56]. 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. The final report of results is based on the correlated effects model. To address potential differences in effect sizes for the reported associated variants, we assessed heterogeneity of effects and did not find evidence ($\log_{10}BF > 2$). 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[54] 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[32] 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[37]. We used reported effect size estimates from 124 CD alleles including those reported in a previously published study[36] and four variants in IL23R from a recent fine-mapping study[37], 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[54]. Variants in NOD2 and LRRK2 were excluded from the analysis to assess whether polygenic signal was independent of those genes.

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

$$PRS_i = \sum_{m=1}^{M} \hat{c}_m c_{G_{i,m}}$$

where $\hat{c}_m$ denotes the estimated log odds ratio for variant $m$ and $G_{i,m}$ denotes the genotype dosage of individual $i$ for variant $m$. More specifically, $\hat{c}_m$ is the effect size estimate of variant $m$ on a logit scale in conferring risk of CD in an individual.

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

$$PRS_i = \sum_{m=1}^{M} [\hat{c}_m 1_{\text{Het}} + \hat{c}_m 1_{\text{Hom}}].$$

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 differences in contribution of NOD2 alleles to genetic risk in two populations.
The estimated expected PRS value for an individual in population \( j \) is
\[
\mathbb{E}[\text{PRS}_j] = \sum_{i=1}^{N_j} \text{PRS}_i \frac{1}{N_j},
\]
where \( N_j \) is the number of individuals sampled in population \( j \). Substituting equation for \( \text{PRS}_i \) and rearranging terms simplifies the equation as a function of variant frequency:
\[
\mathbb{E}[\text{PRS}_j] = \sum_{i=1}^{N_j} \left( \sum_{m=1}^{M} \hat{\beta}_m \frac{G_{i,m}}{N_j} \right),
\]
\[
\mathbb{E}[\text{PRS}_j] = \sum_{m=1}^{M} \left( \frac{\hat{\beta}_m \sum_{i=1}^{N_j} G_{i,m}}{N_j} \right),
\]
where \( \sum_{i=1}^{N_j} \frac{G_{i,m}}{N_j} = \hat{f}_{m,j} \) and \( \hat{f}_{m,j} \) denotes the frequency of variant \( m \) in population \( j \). Thus, the estimated expected PRS value of an individual in population \( j \) is
\[
\mathbb{E}[\text{PRS}_j] = \sum_{m=1}^{M} (2\hat{\beta}_m \hat{f}_{m,j}).
\]

Assume that we are interested in the expected difference in contribution of the studied variants to the 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 contribution of the studied variants to the PRS as the difference in estimated expected PRS value in two populations:
\[
\mathbb{E}[\text{Difference PRS}] = \mathbb{E}[\text{PRS}_{\text{AJ}}] - \mathbb{E}[\text{PRS}_{\text{NFE}}],
\]
\[
\mathbb{E}[\text{Difference PRS}] = \sum_{m=1}^{M} 2\hat{\beta}_m (\hat{f}_{m,\text{AJ}} - \hat{f}_{m,\text{NFE}}),
\]
which can be used to get an estimated difference in contribution of a variant \( m \) to the polygenic risk score in two populations,
\[
\mathbb{E}[\text{Difference PRS}]_m = 2\hat{\beta}_m (\hat{f}_{m,\text{AJ}} - \hat{f}_{m,\text{NFE}}).
\]

To rank variants according to their relative differences in contribution to genetic risk we included the \textit{NOD2} and \textit{LRRK2} alleles, used the list of estimated effect size from the published studies[36,37], and estimates from this study.

If we substitute \( \text{PRS}' \) for \( \text{PRS} \),
\[
\mathbb{E}[\text{PRS}'_j] = \sum_{m=1}^{M} \left( \sum_{i=1}^{N_j} \left[ \frac{\hat{\beta}_m^{\text{Het}}}{1 - \hat{f}_{m,j}^{\text{Het}}} + \frac{\hat{\beta}_m^{\text{Hom}}}{2 \hat{f}_{m,j}^{\text{Hom}}} \right] \right) \frac{1}{N_j}
\]
\[
= \sum_{m=1}^{M} \left[ 2\hat{f}_{m,j} (1 - \hat{f}_{m,j}) \hat{\beta}_m^{\text{Het}} + \hat{f}_{m,j}^2 \hat{\beta}_m^{\text{Hom}} \right],
\]
Then, the estimated expected difference in PRS* when non-additive effects exist is

\[ \mathbb{E}[\text{Difference PRS}^*] = \sum_{m=1}^{M} \left[ 2 \tilde{\mu}_m^{\text{Het}} \left( \hat{m}_m^{\text{Het}} - \hat{m}_m^{\text{NFE}} \right) - \left( \hat{m}_m^{\text{Het}} - \tilde{\mu}_m^{\text{Het}} \right)^2 \right] + \tilde{\mu}_m^{\text{Het}} \left( \hat{m}_m^{\text{Het}} - \tilde{\mu}_m^{\text{Het}} \right). \]

**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 \( p = \left( 1 + \exp(-\eta) \right)^{-1} \), where \( \eta \) tends towards a normal distribution with parameters \( \mu = \log(p_0/(1-p_0)) + \sum_{m=1}^{M} 2 f_m \mu_m \) and \( \sigma^2 = 2 \sum_{m=1}^{M} (1-f_m) \mu_m^2 \). Here \( p_0 \) refers to a baseline disease probability.

We can see that \( \mu \) 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 this setting where \( \mathbb{E}[\text{PRS}] = 0 \), then

\[ \mathbb{E}[p] = \left( 1 + \exp(-\log(p_0/(1-p_0))) \right)^{-1} = p_0. \]

To evaluate the impact of a shift in the expected value of polygenic risk score to the expected value of \( \mu \) we can express the shift as \( \mu \rightarrow \mu + \beta \). We can compute new values of \( p \) for new values of \( \mu \) 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 \( \beta_0 \). To get an estimate of the absolute prevalence of CD in the AJ population, we must choose a baseline \( \beta_0 \) where \( p_0 \) represents the expected prevalence with zero non-baseline alleles in the population[35], 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 \( \beta_0 \) we can calculate the ratio of expected prevalence as follows. First, use the means (\( \mu_{\text{AJ}} \) and \( \mu_{\text{NAJ}} \)) and variances (\( \sigma^2_{\text{AJ}} \) and \( \sigma^2_{\text{NAJ}} \)) 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

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

where \( \eta \) is the risk score associated with prevalence \( p \), \( g \) is the link function, so \( p = g(\eta) = (1 + e^{-\eta})^{-1} \), and \( \phi \) 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 the same 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 \( \beta_0 = -20.5 \) was chosen in order to obtain a prevalence in the non-AJ population of ~0.5%. At this value of \( \beta_0 \), the ratio of prevalence in the AJ population to that in the non-AJ population was estimated to be 1.5 (\( \mathbb{E}[p_{\text{AJ}}] = 0.82\% \), \( \mathbb{E}[p_{\text{NAJ}}] = 0.55\% \)).

For different choices of \( \beta_0 \), however, this ratio may vary, as the relationship between probability of disease and risk score is non-linear. S10 Fig shows how the values of the disease prevalence and their ratio vary as \( \beta_0 \) is changed. We see that the ratio values range from 1.46 to 1.52 for different values of \( \beta_0 \) with a range of baseline prevalence of .001 to .01—the range of prevalence estimates for Crohn’s disease[41,43,57].

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 \( N(\mu_{\text{logit}}, \sigma_{\text{logit}}^2) \) where \( \mu_{\text{logit}} \) and \( \sigma_{\text{logit}} \) 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 \( \mu_{\text{probit}} \) based on that of logit model risk scores \( \mu_{\text{logit}} \) as

\[
f(\eta_{\text{probit}}|\mu_{\text{logit}}, \sigma_{\text{logit}}) = f(\eta_{\text{logit}}|\mu_{\text{logit}}, \sigma_{\text{logit}}) d\eta_{\text{logit}} / d\eta_{\text{probit}}
\]

and use this to calculate \( \mu_{\text{probit}} \) and \( \sigma_{\text{probit}} \), 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 (\( E[p_{\text{AJ}}] = 0.823\% \), \( E[p_{\text{NAJ}}] = 0.544\% \)). These values demonstrate that individual logit and probit analyses would likely give similar results for values of interest. The complete probability densities under the logit and probit models can be seen in S8 Fig.

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 \( \eta_{\text{probit}} = 0.6223 \cdot \eta_{\text{logit}} \), 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 \( \eta_{\text{logit}} \), 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

\[
\text{Prevalence difference} = \frac{(p_j - p_i) - (i_j - i_i)}{(p_j - p_i)}
\]

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.[58] is

\[
i_j = \frac{p_j}{\prod_{m=1}^{M} (1 + f_{mj}(GRR_m - 1))^2}
\]

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.

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

\[
i_j = \frac{p_j}{\prod_{m=1}^{M} (1 + 2f_{mj}(GRR_m^m - 1) + f_{mj}^2(GRR_m^m - 1))}
\]
Enrichment testing sensitivity

When modeling enrichment, we chose a standard significance cutoff of $p < 0.05/N$ for classifying variants as enriched. We noted that the number of variants classified as enriched does not change significantly when the p-value threshold changes. See S11 Fig for more information.

Supporting information

S1 Fig. Analysis workflow diagram.
(PNG)

S2 Fig. Admixture plots.
(PNG)

S3 Fig. Cross-validation errors for number of clusters in ADMIXTURE.
(PNG)

S4 Fig. MAF thresholds chosen for enrichment testing.
(PNG)

S5 Fig. Principal components plot for 5,685 AJ individuals.
(PNG)

S6 Fig. Variable selection using Bayesian model averaging (BMA).
(PNG)

S7 Fig. AJ individuals have higher CD polygenic risk score than NJ controls.
(PNG)

S8 Fig. Dependence of population prevalence on $\beta_0$.
(PNG)

S9 Fig. Probit and logit model analysis.
(PNG)

S10 Fig. The relationship between expected differences in genetic risk score and expected fold differences in disease prevalence.
(PNG)

S11 Fig. Sensitivity test for p-value significance cutoff.
(PNG)

S1 Data File. ClinVar pathogenic alleles enriched in AJ.
(XLSX)

S2 Data File. AJ enrichment data for all analyzed alleles.
(TXT)

S1 Table. Origins of moderate ancestry fraction Ashkenazi samples.
(PNG)

S2 Table. Origins of high ancestry fraction Ashkenazi samples.
(PNG)

S3 Table. Conditional haplotype-based testing in NOD2.
(PNG)
S4 Table. Assessing association of a one-hit and two-hit model of NOD2 in the AJ exome sequencing data.

(PNG)

S5 Table. Assessing association of a one-hit and two-hit model of NOD2 in the non-AJ immunoChip data.

(JPG)

Acknowledgments

Manuel A. Rivas is a Faculty Fellow at the Stanford Center for Population Health Sciences. We thank the Broad IT team for assistance with the IBD exomes browser.

NIDDK IBD Genetics Consortium


International IBD Genetics Consortium

Weersma, Zhi Wei, David Whittman, Cisca Wijmenga, David C Wilson, Juliane Winkelmann, Ramnik J Xavier, Bin Zhang, Clarence K Zhang, Hu Zhang, Wei Zhang, Hongyu Zhao, Zhen Z Zhao.

T2D-GENES Consortium

Author Contributions

Conceptualization: Manuel A. Rivas, Judy H. Cho, Dermot P. B. McGovern, Mark J. Daly.


Formal analysis: Manuel A. Rivas, Brandon E. Avila, Jukka Koskela, Hailiang Huang, Matti Pirinen, Monkol Lek, Mark J. Daly.

Funding acquisition: Mark J. Daly.

Investigation: Manuel A. Rivas, Judy H. Cho, Dermot P. B. McGovern, Mark J. Daly.

Methodology: Manuel A. Rivas, Brandon E. Avila, Hailiang Huang, Matti Pirinen, Mark J. Daly.

Project administration: Christine Stevens, Mark J. Daly.


Supervision: Mark J. Daly.

Validation: Brandon E. Avila, Jukka Koskela.

Visualization: Manuel A. Rivas, Brandon E. Avila, Jukka Koskela, Hailiang Huang, Mark J. Daly.

Writing – original draft: Manuel A. Rivas.
Writing – review & editing: Brandon E. Avila, Jukka Koskela, Judy H. Cho, Dermot P. B. McGovern, Mark J. Daly.

References


