Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease

Hari K. Somineni, Emory University
Sini Nagpal, Georgia Institute of Technology
Suresh Venkateswaran, Emory University
David Cutler, Emory University
David Okou, Emory University
Tanvi Dhere, Emory University
Heba Iskandar, Emory University
Michael Zwick, Emory University
Gregory Gibson, Emory University
Subramaniam Kugathasan, Emory University

Journal Title: AMERICAN JOURNAL OF HUMAN GENETICS
Volume: Volume 108, Number 3
Publisher: Elsevier (Cell Press) | 2021-03-04, Pages 431-445
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1016/j.ajhg.2021.02.001
Permanent URL: https://pid.emory.edu/ark:/25593/vv7dz

Final published version: http://dx.doi.org/10.1016/j.ajhg.2021.02.001

Copyright information:
© 2021 The Author(s)
This is an Open Access work distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Accessed December 3, 2022 2:28 AM EST
Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease

Hari K. Somineni,1,2 Sini Nagpal,3 Suresh Venkateswaran,2 David J. Cutler,4 David T. Okou,2 Talin Haritunians,5 Claire L. Simpson,6 Ferdouse Begum,7 Lisa W. Datta,7 Antonio J. Quiros,8 Jenifer Seminerio,9 Emebet Mengesha,5 Jonathan S. Alexander,10 Robert N. Baldassano,11 Sharon Dudley-Brown,12 Raymond K. Cross,13 Themistocles Dassopoulos,14 Lee A. Denson,15 Tanvi A. Dhere,16 Heba Iskandar,16 Gerald W. Dryden,17 Jason K. Hou,18

Summary

Whether or not populations diverge with respect to the genetic contribution to risk of specific complex diseases is relevant to understanding the evolution of susceptibility and origins of health disparities. Here, we describe a large-scale whole-genome sequencing study of inflammatory bowel disease encompassing 1,774 affected individuals and 1,644 healthy control Americans with African ancestry (African Americans). Although no new loci for inflammatory bowel disease are discovered at genome-wide significance levels, we identify numerous instances of differential effect sizes in combination with divergent allele frequencies. For example, the major effect at PTGER4 fine maps to a single credible interval of 22 SNPs corresponding to one of four independent associations at the locus in European ancestry individuals but with an elevated odds ratio for Crohn disease in African Americans. A rare variant aggregate analysis implicates Ca2+-binding neuro-immunomodulator CALB2 in ulcerative colitis. Highly significant overall overlap of common variant risk for inflammatory bowel disease susceptibility between individuals with African and European ancestries was observed, with 41 of 241 previously known lead variants replicated and overall correlations in effect sizes of 0.68 for combined inflammatory bowel disease. Nevertheless, subtle differences influence the performance of polygenic risk scores, and we show that ancestry-appropriate weights significantly improve polygenic prediction in the highest percentiles of risk. The median amount of variance explained per locus remains the same in African and European cohorts, providing evidence for compensation of effect sizes as allele frequencies diverge, as expected under a highly polygenic model of disease.

Introduction

The inflammatory bowel diseases (IBDs [MIM: 604519, 266600, 191390]), Crohn disease (CD [MIM: 266600]), and ulcerative colitis (UC [MIM: 191390]) arise in the context of inappropriate activation of the intestinal immune system in response to an environmental trigger in individuals who are genetically predisposed. Genome-wide association studies (GWASs) of common and low-frequency variants have so far identified 241 loci that confer significant risk for disease susceptibility.1–3 Although inflammatory bowel disease is one of the most successfully studied polygenic diseases with respect to identifying loci and risk alleles, four major challenges remain: (1) missing heritability—only a small fraction of disease liability is explained by the thus far known genetic risk factors (13% for...
Crohn disease and 8% for ulcerative colitis); (2) uncertainty as to the true causal genetic variants underlying GWAS associations—as most loci span several kb to tens of kb in length, containing credible sets that usually range from several to hundreds of variants in tight linkage disequilibrium (LD) having similar evidence of association; (3) lack of molecular insights into established genetic signals—as most association signals map to non-coding regions with their mechanistic effects largely unknown; and (4) lack of understanding on the different genetic architectures across populations.

Genetic discoveries of inflammatory bowel disease have been made primarily in populations of European ancestry and utilizing genome-wide genotype data. This predominance, combined with a focus on common alleles, has left our understanding of the role of rare variants among non-European populations incomplete. To this end, we have performed the whole-genome sequencing study of inflammatory bowel disease-affected individuals as compared to non-inflammatory bowel disease control individuals from over 3,600 Americans with African ancestry. Although most GWASs are performed with genotyping arrays, whole-genome sequencing offers advantages such as assessment of rare heterozygous effects—for example, type 2 diabetes (T2D [MIM: 125853]) and on blood metabolites—and comprehensive assessment of non-imputed common variants (for example, for COPD [MIM: 606963]), as well as mapping in the presence of high variability and short LD blocks in mixed ancestry populations (for example, plasma lipoproteins and serum peptides).

Our goals were 4-fold: first, we hypothesized that genetic analysis of this understudied population would facilitate new locus discovery of common variants that have not been previously interrogated, including those that are specific to African-ancestry populations or shared across divergent populations. Second, given the high genetic diversity in African populations, we hypothesized that rare and potentially high-risk inflammatory bowel disease variants, within or near protein coding genes, have yet to be identified. Third, we evaluated the potential of our genetically diverse African American cohort to enhance fine-mapping of credible intervals. Fourth, we evaluated the degree to which effect sizes of GWAS variants differ across populations and assessed the impact on polygenic risk assessment based on established inflammatory bowel disease loci.

Subjects and methods

Study samples

This was a multi-center collaborative study involving self-identified African American subjects recruited from five primary sites and their collaborating centers across the US. These sample recruitment centers include Emory University (recruited as part of the GENESIS study and Emory African American Inflammatory Bowel Disease Consortium) and 12 other collaborating centers; Johns Hopkins/Rutgers (recruited as part of the Multicenter African American Inflammatory Bowel Disease Study) and 17 other collaborating centers; Cedars-Sinai Medical Center; Mount Sinai Medical Center; and Washington University (recruited as part of the Centers for Common Disease Genomics network). Sample breakdown,
along with the proportion of affected individuals versus control individuals, per center is shown in Table S1. This study was approved by the institutional review boards at each of the participating sites and informed consent was obtained from all the participants. Deidentified datasets including genetic data are housed at Emory University with the approval of the local ethical board.

**Library construction and whole-genome sequencing**

All DNA samples investigated in this study (a total of 3,610 before quality control [QC]) were processed and sequenced at the Broad Institute of Harvard and MIT (Cambridge, MA) following the same protocol. Briefly, genomic DNA (350 ng in 50 μL) extracted from the blood of sampled participants was fragmented to a target size of 385 bp fragments via a Covaris Focused-ultrasonicator. Fragmented DNA was subjected to further size selection via a SPRI cleanup. Libraries were then constructed with kits commercially available from KAPA Biosystems (KAPA Hyper Prep without amplification module, product KK8505) and with palindromic forked adapters with unique 8-base index sequences embedded within the adaptor (purchased from Roche). Completed libraries were quantified with quantitative PCR (kit purchased from KAPA Biosystems), normalized to 2.2 nM, and were pooled into 24-plexes. Sample pools were combined with HiSeqX Cluster Amp Reagents EPX1, EPX2, and EPX3, and cluster generation was performed with the Illumina cbot according to the manufacturer’s protocol (Illumina). DNA libraries were sequenced with the HiSeqX sequencing system utilizing sequencing-by-synthesis kits to produce 151 bp paired-end reads. Because library amplification by PCR introduces substantial bias, we sequenced our samples by using the PCR-free protocol. Output from Illumina software was processed by the Picard data-processing pipeline to yield CRAM files containing demultiplexed, aggregated aligned reads.

**Data processing and variant calling**

The sequence reads from each sample were aligned to human reference genome build hg38 (GRCh38 assembly). Each sample was sequenced to an average depth of 30×. Variants were joint called across 3,610 samples via the Genome Analysis Toolkit (GATK) Best Practices for germline variants and were annotated via our in-house Bystro10 software. Our initial output, VCF file, had data for 3,610 samples and about 145 million variants, both single nucleotide polymorphisms (SNPs) and short insertions (INDELs). In addition, we applied RepeatMasker12 to mask variants from repetitive and low complexity regions of the genome. For some common variant association analysis, we also included the pre-masked dataset in order to obtain summary statistics for some of the lead variants from the previously known disease loci (identified via populations of predominantly European ancestry) that were masked by the RepeatMasker program.

**Variant QC and filtering**

After the removal of samples from above, we then filtered out variants genotyped in <95% of the samples (missingsnness > 5%) and those that showed a significant deviation from Hardy-Weinberg equilibrium in control individuals (p < 1 × 10⁻¹⁰). These procedures resulted in a final dataset of 3,418 samples and 93.4 million variants that include both SNPs and INDELs. In addition, we applied RepeatMasker12 to mask variants from repetitive and low complexity regions of the genome. For some common variant association analysis, we also included the pre-masked dataset in order to obtain summary statistics for some of the lead variants from the previously known disease loci (identified via populations of predominantly European ancestry) that were masked by the RepeatMasker program.

**Principal-component analysis of sequence data**

After excluding samples and variants with low quality, principal-component analysis of the whole-genome sequencing dataset was performed with EIGENSTRAT.13 Principal components were computed on the basis of a pruned version of the dataset consisting of ~1.4 million LD-independent (r² < 0.1), high frequency (MAF > 1%) variants. The first five principal components (based on the inspection of the scree plot) were included as covariates to control for population stratification within the whole-genomes dataset for all analyses (Figure S1).

**Single-variant association testing of sequence data**

We used a logistic regression model to test for association at each individual variant with the first five principal components of the genotype matrix included as covariates. Variants were tested separately for association with Crohn disease, ulcerative colitis, and inflammatory bowel disease (Crohn disease and ulcerative colitis together with inflammatory bowel disease-type unknown [IBDU]). We defined common variants as those that are present in at least 1% of the general African population from gnomAD and have an observed MAF > 1% in this dataset, yielding 14.9 million variants (pre-masked; ~7 million variants in the post-masked dataset). The genomic control (λ_gcr) values for these individual analyses of common variants ranged from 1.02 to 1.04, indicating little or no inflation or deflation due to population
stratification. We defined rare variants as those that are either absent or present at an MAF of $< 0.1\%$ in the general African population from gnomAD. With these criteria, we observed 64.2 million rare variants in our dataset, and the genomic control $\hat{\epsilon}_{GC}$ values for these individual analyses of rare variants ranged from 0.61 to 0.84, indicating deflation due to the limited number of rare alleles in the dataset.

**Known association in ADCY7**

Adenylate cyclase 7 (ADCY7 [MIM: 600385]) has recently been implicated in ulcerative colitis in European populations; using a low-pass whole-genome sequenced case-control cohort in conjunction with well-imputed newly genotyped and pre-existing GWAS case-control datasets of European ancestry and affected individuals and non-inflammatory bowel disease control individuals from the UK Biobank, Luo et al. identified a low-frequency (MAF = 0.006), missense variant, rs78534766, in ADCY7 in association with an increased risk of ulcerative colitis. In fact, rs78534766 is the largest effect allele that was identified to date for ulcerative colitis (odds ratio [OR] = 2.19; $p = 9 \times 10^{-15}$). However, although there is no evidence of association for ulcerative colitis in our study, we noticed an effect trending in the opposite direction (OR = 0.46; $p = 0.47$). The effect allele was seen in seven control individuals (MAF = 0.002) and one affected individual with ulcerative colitis (and four affected individuals with Crohn disease) in our whole-genome sequencing dataset: this could in part be due to the limited sample size of our dataset, fueled by the rarity of this allele, rs78534766, in general African populations (MAF in gnomAD v3 = 0.00098).

**Power analysis**

For power calculations, we assumed that disease prevalence (0.19% for Crohn disease and 0.20% for ulcerative colitis) and effect sizes are homogeneous across ancestries. We converted population frequencies (reported in African populations from gnomAD) and odds ratios (reported in the latest meta-analysis of European descent individuals) to controlled and affected individual frequencies and calculated power of our whole-genomes cohort to detect previously known disease loci at $p < 5 \times 10^{-8}$ or $p < 0.05$ as a function of the MAF and the genotype relative risk of the variant under an additive model.

**Aggregate rare-variant association testing**

Using the optimal sequence kernel association test (SKAT-O), we performed gene-wise aggregation analyses to detect aggregate association of rare, likely deleterious (combined annotation dependent depletion (CADD) > 15) variants with the three traits. For aggregate tests, we selected all rare, likely deleterious (CADD > 15) variants (n = 1.5 million) across the genome and assigned them to the nearest gene. We then assessed the association of each gene with a collection of rare, likely deleterious variants in a SKAT-O model implemented in the R package “SKAT.” To interpret statistical significance, we applied experimental-wise, Bonferroni-corrected significance threshold of $p < 2.2 \times 10^{-6} = 0.05/22,521$.

**GWAS genotype data, QC, imputation, and association testing**

Sample information, genotype data, and the application of QC procedures for the two existing GWAS cohorts considered in the current study were described extensively elsewhere. Briefly, genome-wide genotype data from non-overlapping African American affected individuals and matched control individuals generated with either the Illumina Omni (398 with Crohn disease, 238 with ulcerative colitis, and 1,551 control individuals) or the Affymetrix Axiom Genome-Wide AFR 1 World Array (451 with Crohn disease, 186 with ulcerative colitis, and 3,038 control individuals) SNP chips were considered for replicative evidence. Sample and variant QC, determination of principal components, and removal of outliers was done as described in the original paper. Both datasets were lifted from human reference build hg19 to hg38 via liftOver.

**Imputation**

We phased the whole-genome sequences described above (n = 3,418; after QC) with Eagle v2.4 to create a reference panel. These pre-phased whole-genome sequences with MAF > 0.5% were imputed into each GWAS dataset, separately, via minimac3 software. By design, all the sequenced individuals have African descent and about half of these are inflammatory bowel disease-affected individuals, thereby enriching the reference panel for African-specific alleles that increase or decrease inflammatory bowel disease risk.

**Common variant association testing for replicative evidence**

After removing samples that were directly sequenced in the discovery phase, genotyped and imputed variants with INFO score > 0.6 were tested for association with Crohn disease, ulcerative colitis, and inflammatory bowel disease, separately, within each GWAS case-control dataset via SNPTEST 2.5.2, performing an additive frequentist association test conditioned on the first ten principal components. For sites that were present in both datasets and passed our QC filters, we performed meta-analysis by using META. For a common variant with $p < 5 \times 10^{-8}$ in the discovery cohort to be inferred to be associated with a trait, it had to have a directionally consistent effect and demonstrate at least a nominal evidence of association ($p < 0.05$) in the meta-analysis of the two GWAS datasets. At the 5p13.1 locus near Prostaglandin E Receptor 4 (PTGER4 [MIM: 601586]), the previously defined peak SNP rs7711427 was initially relegated to a “sub-PASS” tranche during GATK’s automated analysis, but closer inspection revealed that the reason for exclusion was an excess of heterozygotes in affected individuals. As this is a well-characterized disease-associated variant and strong disease alleles will give rise to excess heterozygosity in affected individuals, we elected to “rescue” this allele. Because this site is in complete LD ($r^2 = 1$) with several other sites, no conclusions are altered, but we retain rs7711427 to facilitate comparison with published results.

**Genetic risk score calculation**

Assessment of precision (equivalent to predicted prevalence in a percentile of risk) is sensitive to the ratio of affected individuals to control individuals, so it cannot be conducted directly on the whole-genome sequencing cohort. In order to directly contrast accuracy in a population with a maximal inflammatory bowel disease prevalence of ~1%, we thus generated 150,000 pseudo-controls given expected genotype frequencies observed in our 1,644 African American whole-genome sequenced control individuals. Rather than using an external reference database of allele frequencies in African Americans, since these are known to be heterogeneous due to varying degrees of admixture, we used the control estimates from our whole-genome sequencing for internal consistency and simply generated random genotypes according to Hardy-Weinberg expectations. These were computed for 215 of the 241 known loci that were present in both the African American pseudo-controls cohort and UK Biobank. In the pseudo-controls, we combined genotypes between loci independently to minimize LD within the pseudo-control population. Note that
we only use the lead SNPs and, as a result, make no attempt to capture independent secondary signals at each locus or to disentangle the combined effects of multiple interacting sites in high LD. Instead, we simply acknowledge that some of the observed differential effect sizes may be due to the combined influence of secondary sites in LD or epistasis. By ignoring neighboring sites with main effects, we are certainly under-representing the total amount of variance explained per locus but eliminating any influence of these effects on the derivation of pseudo-controls. In the absence of a very large external dataset of African ancestry inflammatory bowel disease genotypes, this is currently the least-biased approach to estimating per-locus effects in polygenic risk score (PRS) estimation, but it should be recognized that there may be other systemic biases to the use of pseudo-controls that contribute to the ancestry-specific assessments. To confirm that long-range LD is also not biasing the analysis, we computed the contribution to the ancestry-specific assessments. To confirm that biased approach to estimating per-locus effects in polygenic risk inflammation bowel disease genotypes, this is currently the least-influence of these effects on the derivation of pseudo-controls.

LD. Instead, we simply acknowledge that some of the observed tangle the combined effects of multiple interacting sites in high genotype correlations ($r$) between $2\%$ and $7\%$, with an overall average $r^2$ of just $0.07\%$. By design, there was no observed significant LD among the variants in the pseudo-controls. Consequently, bias introduced by LD among the lead SNPs appears to be measurably small in the African American pseudo-controls dataset, which has 1,774 inflammatory bowel disease affected individuals and 150,000 pseudo-controls. We similarly generated a UK Biobank dataset of the same size by down-sampling to the same case-control ratio; in this case, actual genotypes were used because UK Biobank is sufficiently large.

Next, to compute the genetic risk score, we divided each of the five sets of African American pseudo-controls and UK Biobank into a training set with 70% samples (1,242 affected individuals, 105,000 pseudo-controls) and test sets with the remaining 30% samples (532 affected individuals, 45,000 pseudo-controls). We used the 70% training datasets to estimate effect sizes both on the lnOR scale as well as the liability scale (these data were also used to estimate proportion of variance explained [PVE]). Using these estimated effect sizes at the most significant variants in the 215 of the 241 known loci that were present in both the African American pseudo-controls cohort and UK Biobank, we derived weighted PRSs in the withheld 30% datasets. Furthermore, for each PRS, the prevalence of inflammatory bowel disease, with standard error determined from the five test sets, was computed at each percentile bin on both the liability scale and lnOR scale. The proportion of variance explained by PRS was computed by determining the Negelkerke’s R-square via the “lm” function from the “rms” package in R.

Liability scale modeling

Liability effects were measured by assuming an underlying normal distribution (mean 0, variance 1) of liability representing disease status. Affected individuals are assumed to lie above a certain threshold of this continuous distribution. This threshold (Z score on x axis) is determined by computing the inverse of the cumulative normal function of the assumed prevalence of the disease. Assuming an upper limit for the prevalence of all inflammatory bowel diseases as $1\%$, odds ratios determined from logistic regression as the true odds ratios, and using the allele frequencies in control individuals within the study, we determined the additive effects on liability as follows: the displacement of each genotype threshold from the overall population threshold as $2[326-norminv(1-penetrance)]$ where the penetration of the genotype is the overall prevalence ($1\%$) multiplied by the ratio of the genotype between affected individuals and overall. The mean displacement was computed as the sum of the products of the three genotype displacements weighted by the genotype frequencies in the overall sample, which allows computation of the central displacement of each genotype by subtraction of the mean displacement. The liability $\alpha_1$ of the larger effect genotype is then $[p-central displacement_{AA}+q-central displacement_{T_Aa}]$ where $p$ is the frequency of the minor allele A and $q$ is the frequency of the major allele a. Similarly, the liability $\alpha_2$ of the alternate allele is $[p-central displacement_{Aa}+q-central displacement_{AA}]$. The total variance explained by the locus on the liability scale is $2p_\alpha_1^2 + 2q_\alpha_2^2$. The polygenic risk score for an individual is the sum of the liabilities of the relevant genotypes, namely $2\alpha_1 + \alpha_2$, or $2\alpha_2$.

Results

Common variant associations in African Americans

After QC (see subjects and methods) and principal-component analysis (Figure S1) of our deeply sequenced whole-genomes (median coverage of 30×), we present analyses of a total of 3,418 subjects: 1,774 affected individuals (1,335 with Crohn disease, 407 with ulcerative colitis, and 32 with IBDU) and 1,644 non-inflammatory bowel disease control individuals (Table S1) at 93 million variants that comprise both SNPs and short INDELs. These data include 14.9 million common variants (MAF > 1%), 13.9 million low-frequency variants (0.1% < MAF < 1%), and 64.2 million rare variants (MAF < 0.1%). First, we used single-variant analyses to test each variant, regardless of allele frequency, for association with Crohn disease, ulcerative colitis, and inflammatory bowel disease (Crohn disease and ulcerative colitis together with IBDU), separately, in a logistic regression framework conditioned on the first five principal components (see subjects and methods). Following these “discovery” analyses of whole-genome sequence data, we sought replication of the obtained results at common variants in an independent cohort of African American affected individuals and control individuals that were previously genotyped with Axiom or Omni genome-wide SNP arrays (subjects and methods; Table S1). We imputed our whole-genome sequences into these two (Axiom and Omni) existing GWAS datasets, thereby enriching the panel for inflammatory bowel disease risk alleles, and performed case-control association testing by using a logistic regression model, separately, within each dataset. Results from the meta-analysis of these two GWAS datasets served as our replicative evidence for common variation.

With a standard GWAS significance threshold of $p < 5 \times 10^{-8}$ in the discovery cohort and at least nominal (and directionally consistent) evidence of association ($p < 0.05$) in the replication cohort, we identified 22 common variants at a locus proximal to PTGER4 (260 kb) on chromosome 5p, previously discovered in cohorts of European descent,
associated with a decreased risk of Crohn disease (Figure 1 and Table S2). All 22 variants were in complete LD with each other ($r^2 = 1$). Following our previous report of suggestive evidence of association at this locus for Crohn disease in African Americans,\textsuperscript{16} here we present the first evidence of standard (GWAS) genome-wide significance. Consistent with being a Crohn disease-specific locus, the disease-associated alleles demonstrated strong effect sizes for Crohn disease (OR = 0.74; $p < 5 \times 10^{-8}$) but not ulcerative colitis (OR = 0.91; $p = 0.25$; Figure 1) despite being directionally consistent to both. Further, of all the ancestrally divergent populations studied thus far, the strongest effect on Crohn disease at these variants was observed in our African American cohort (Figure S2).

A protective role in Crohn disease of \textit{PTGER4} locus minor alleles was initially discovered in populations of European ancestry,\textsuperscript{32} and 2,819 common variants at the locus had genome-wide significant association. Using conditional stepwise regression to perform fine-mapping, Huang et al. further refined this region to a subset of 189 credible variants representing four independent signals that are more likely to be causal to Crohn disease\textsuperscript{24} (Figure S3). The 22-variant African American signal precisely captures the known primary peak (signal 1) of association in European populations (MAF in African American Crohn disease-affected individuals = 0.32 and control individuals = 0.39; MAF in European Crohn disease-affected individuals = 0.33 and control individuals = 0.39); there is no evidence for an association at signal 2 but nominal evidence at signals 3 and 4. The 22 Crohn disease-associated variants that we detected at this locus in African Americans were in high LD with the strongest signal (signal 1) comprising two potentially causal variants—rs7711427 and rs397897680—from the fine-mapping analysis of European populations\textsuperscript{32} (Figure S4). We note that rs397897680 has since been merged with rs5867512,
yielded deflated summary statistics (Figure S5), we per-
or imputed in previous GWASs of inflammatory bowel dis-
ants that include many alleles that were not genotyped
disease. Our data was comprised of 64.2 million rare vari-
tion of rare variants (MAF < 0.1%) to inflammatory bowel
disease. With the sequencing data, we next assessed the contribu-
tion is affected in diseased epithelial cells.

Experimental manipulation of epithelial cells supports a
role for prostaglandin signaling promoting intestinal
wound healing during inflammatory bowel disease, but
it remains to be established whether PTGER4 is the target
of the primary GWAS signal that lies in a gene desert 250
kb from a cluster of candidate genes and whether expres-
sion is affected in diseased epithelial cells.

Rare-variant associations in African Americans
With the sequencing data, we next assessed the contribu-
tion of rare variants (MAF < 0.1%) to inflammatory bowel
disease. Our data was comprised of 64.2 million rare vari-
ants that include many alleles that were not genotyped
or imputed in previous GWASs of inflammatory bowel dis-
bases. Because our single-variant analyses at rare variants
yielded deflated summary statistics (Figure S5), we per-
formed aggregate analyses by selecting all rare, likely dele-
terious (CADD > 15) variants across the genome and as-
signing them to the nearest gene. In total, 1.5 million
such variants were assigned to 22,521 genes with an
average of 68 variants per gene (range = 1–3,593). Using
the SKAT-O approach,15 we then tested whether any of
these gene sets with a collection of rare, likely deleterious
variants have an aggregate association with inflammatory
bowel disease, Crohn disease, or ulcerative colitis. To inter-
pret statistical significance, we applied a Bonferroni-cor-
corrected significance threshold of $p_{SKAT} < 2.2 \times 10^{-6}$ (0.05
corrected for 22,521 tests).

Using this strategy, we implicate variants in the vicinity
of Calbindin 2 (CALB2 [MIM: 114051]) in ulcerative colitis.
We detected an aggregate association of 35 rare, likely dele-
terious, heterozygous variants within or near CALB2 with
ulcerative colitis ($p_{SKAT} = 1.61 \times 10^{-10}$; Figure 2 and Table
S3). Half of these variants were observed more frequently
in affected individuals with ulcerative colitis compared to
control individuals, while the other half were seen less
frequently, representing a typical SKAT type of signal.
Among the 35 variants that collectively contributed to
this aggregate signal, many are absent in European (non-
Finnish) populations in the gnomAD v3 release34 (15 as
opposed to only 4 variants that were never seen before in
African populations) despite a 0.6-fold larger European
effective sample size, including an African-specific intronic
variant, rs200083611, with a nominal evidence of associa-
tion for increased risk of ulcerative colitis, showing an MAF
of 0.009 in affected individuals and 0.0003 in control
individuals (p = 0.001; OR = 30.5 from single variant asso-
ciation analysis). However, given the high-risk but weak
evidence of association at rs200083611, it appears that
the CALB2 gene-wide signal was driven by multiple addi-
tional rare variants.

This CALB2 signal was approximately 3 Mb away
from, and independent of, the nearby common variant, rs1728785 (intronic region of ZFP90 [MIM: 609451]), that has an established association for ulcerative colitis,2,14,35 indicating that these rare variant associations represent unique effects. CALB2 encodes an intracellular calcium-binding protein, calbindin 2 (also known as calre-
tinin), that plays an important role in neuronal physiology
and the maintenance of Ca^{2+} intracellular homeostasis.
CALB2 has a common expression pattern in both the cen-
tral and peripheral nervous systems. It has high expression
in brain and intermediate expression in sigmoid and trans-
verse colon. The absence of CALB2 in nerve fibers in colon
is a widely used marker for Hirschsprung disease36,37
(HSCR1 [MIM: 142623]), whereas elevated expression of
CALB2 has been reported as a hallmark of rapidly prolif-
erating cancerous cell lines, including in colorectal cancer
cell lines.38 Hirschsprung disease shares many of the clin-
cal features with inflammatory bowel disease, where the
latter is more commonly reported in affected individuals
who had surgical treatment for Hirschsprung disease.39

Conversely, long-standing inflammatory bowel disease is
an established risk factor for colorectal cancer.40,41 Given
the intricate relationship of inflammatory bowel disease
with these companion diseases, our implication of Hirsch-
prung disease- and colorectal cancer-associated CALB2 in
ulcerative colitis makes this signal worthy of experimental
validation.
Genetic landscape at the established disease loci between African Americans and Europeans

With our whole-genomes data, we next assessed whether the genetic landscape at established inflammatory bowel disease risk loci is shared between populations of European and African descent and whether trans-ethnic comparative analysis can be leveraged to further refine established GWAS signals. Of the 241 lead variants from the thus far established loci from the recent meta-analyses of cohorts of European descent (GWAS catalog), in addition to replicating the known PTGER4 locus at $p < 5 \times 10^{-8}$, we replicated 41 lead variants for association with inflammatory bowel disease, Crohn disease, or ulcerative colitis (see Table S4 for all lead variant data). Assuming Crohn disease and ulcerative colitis prevalence of 0.19% and 0.20%, respectively, and no ascertainment bias, our whole-genomes cohort was powered to replicate 42 of the known loci at $p < 0.05$ (see subjects and methods; Figure 3), matching the number of associations actually observed. Despite limited statistical evidence of association at many of the known loci as a result of the size of our African American cohort, we noted a strong concordance in the direction of effects ($73\%$, $p = 2.05 \times 10^{-12}$ for inflammatory bowel disease; $69\%$, $p = 1.26 \times 10^{-8}$ for Crohn disease; and $64\%$, $p = 2.48 \times 10^{-5}$ for ulcerative colitis; sign test) and a strikingly positive correlation in effect sizes between European and African populations ($R = 0.68$, $p < 2.2 \times 10^{-16}$ for inflammatory bowel disease; $R = 0.63$, $p < 2.2 \times 10^{-16}$ for Crohn disease; and $R = 0.43$, $p = 3.0 \times 10^{-9}$ for ulcerative colitis), further supporting the previous notion that the genetic risk of inflammatory bowel diseases conferred by common variants

Figure 3. Comparison of effect sizes across populations
(A) Statistical power of the discovery whole-genome sequencing dataset to replicate previously known variants at $p < 0.05$. The known variants that showed an association ($p < 0.05$) with Crohn disease, ulcerative colitis, or inflammatory bowel disease in African Americans are colored to denote their $p$ value of association.
(B–D) Comparison of effect sizes at known variants in African Americans and Europeans. Each variant is colored to denote the $p$ value of association for Crohn disease (B), ulcerative colitis (C), or inflammatory bowel disease (D) in African Americans from the discovery whole-genome sequencing cohort. The red line shows the linear regression fit to indicate the general trend. Significance and goodness-of-fit are shown. ADCY7 was not included while computing goodness-of-fit in (C) (see subjects and methods).
is, to a great extent, shared across divergent populations (Figure 3).

Despite strong effect-size correlations, subtle differences in allele frequency or the magnitude of effects at the established disease loci may reveal differential genetic architectures underlying inflammatory bowel disease between the two populations. Given that our African American dataset is at least one order of magnitude smaller than the existing European datasets, in order to contrast effect estimates directly between the two populations, and assuming a maximal underlying inflammatory bowel disease prevalence of up to ~1%, we generated 150,000 pseudo-controls from the observed genotypes in our whole-genome sequenced control individuals (n = 1,644) to contrast against the 1,774 inflammatory bowel disease-affected individuals. We refer to this from here on as African American pseudo-controls cohort (with 1,774 affected individuals and 150,000 pseudo-controls). For the European counterpart, we took advantage of the UK Biobank (all the observed genotypes were used because the UK Biobank is sufficiently large, 1,774 individuals with inflammatory bowel disease and 150,000 non-inflammatory bowel disease control individuals; see subjects and methods). For both the African American pseudo-controls cohort and UK Biobank, we generated five random subsets retaining 70% of the samples (1,242 affected individuals and 105,000 pseudo-controls) as discovery sets for estimating effect sizes. In a comparative analysis of effects (natural log transformed odds ratios [lnOR]) at the known loci, 69% of the African American estimates are within the 95% confidence interval of UK Biobank estimates (and 75% vice versa), implying general similarity of odds ratios. The distributions of the lnOR are significantly biased toward higher absolute values in African Americans (p = 0.01, two-tailed paired sample t test, n = 215 SNPs) and are highly correlated with those generated directly from the whole-genome sequencing study and shown in Figure 3 (mean r = 0.88). Scores generated directly from the discovery cohort may also bias the estimation (although note that we are only considering lead SNP effects previously identified in predominantly European ancestry studies), but considered together, the two approaches increase the robustness of our findings. The correlation between the odds ratio estimates from the de Lange et al. meta-analysis and the inflammatory bowel disease estimates in five UK Biobank iterations is just 0.77 on average, which is greater than those estimated between the African American pseudo-controls and UK Biobank (0.48) or the direct African American whole-genome sequencing GWAS and European meta-analysis (0.68). Thus, generation of pseudo-controls introduces minimal bias.

Since allele frequencies co-vary with odds ratios, explaining approximately 6.1% (mean R² across five sets) of the difference in odds ratio between the populations (Figure S6), we also computed effect sizes on the liability scale (see subjects and methods), which is not a function of allele frequency. These estimates are also biased slightly upward in African Americans (mean R² = 3.1%; Figure S6). The median percent variance explained per locus is very similar whether on the lnOR or liability scales in the two populations (Figure 4). Nevertheless, we estimate that given either the European or African allele frequencies, typically 40%–50% more variance would be explained with the observed African American effect sizes. Reciprocally, for either distribution of effect sizes, the European allele frequencies lead to 1.35–1.45 times greater variance explained. Collectively, these observations suggest that compensatory differences in both effect sizes and allele frequencies contribute to divergence in inflammatory bowel disease risk between European and African populations. The effect size differences may be due to local ancestry effects of LD with secondary associations at loci and to global interactions with the total genetic background.

**Variance in disease liability in African Americans and Europeans**

For 215 of the 241 known loci that we had data for in both the African American pseudo-controls cohort and UK
Biobank, we estimated the mean PVE by each locus from five discovery subsets in both lnOR and liability scale, and this is illustrated in Figure 5. While the two large effect European Crohn disease-associated alleles at IL23R and NOD2 explained 4.58% and 1.08% PVE on lnOR scale (0.59% and 0.15% on liability scale) in the UK Biobank cohort, they only accounted for 1.02% and 0.24% (0.13% and 0.03% on liability scale) in the African American pseudo-controls cohort, respectively. Conversely, the risk allele at rs3764147 (genes in the region include LACC1 [MIM: 613409] and CCDC122 [MIM: 613408]) that depicted suggestive evidence of association (p = 1.2 × 10⁻⁹) in our discovery African American cohort accounted for 2.71% PVE in the African American pseudo-controls cohort, while it only explained 0.40% in Europeans (0.37% and 0.05% on liability scale, respectively; Figure 5). Similarly, the PTGER4 signal that we detected at p < 5 × 10⁻⁸ in our whole-genome sequencing analysis explained 3.35% of PVE in African Americans versus 0.93% in Europeans (0.45% and 0.12% on liability scale, respectively).

**Transferability of polygenic risk scores across populations**

These subtle differences at each individual locus when combined across a genome-wide feature set, for example while deriving polygenic risk predictors for a common complex disease, lead to significant biases in populations distinct from the discovery samples. In order to directly contrast accuracy in a population with inflammatory bowel disease prevalence of up to ~1%, we utilized the effects estimated from our five discovery sets generated from 70% of 1,774 affected individuals and 150,000 pseudo-controls and used the remaining 30% of the withheld samples from each of the five sets for computing PRSs at 215 of the 241 known loci that we had data for in both the African American pseudo-controls cohort and UK Biobank. Strikingly, after 5-fold cross validation, the African American-estimated betas gave a 7-fold elevation of the top percentile and performed better than the African American beta estimates applied to the UK Biobank. Specifically, for PRSs computed in African Americans, the prevalence in the top percentile was 7.4% (mean PVE by PRS from five test sets: R² = 0.026) with African American summary statistics, 2.5% (mean PVE: R² = 0.011) with UK Biobank summary statistics, and 2.9% (mean PVE: R² = 0.010) with summary statistics obtained from the largest GWAS meta-analysis of European individuals. For PRSs computed in UK Biobank individuals, the prevalence in the top percentile was 2.8% (mean PVE: R² = 0.009) with African American summary statistics, 3.0% (mean PVE: R² = 0.019) with UK Biobank summary statistics, and 3.8% (mean R² = 0.027) with summary statistics from GWAS meta-analysis of European individuals (Figure S7). Although there may be systemic biases in the use of pseudo-controls (see subjects and methods), these results confirm and extend recent observations concerning the propensity for PRS distributions to differ between Europeans and Africans based on discovery ancestry group.

In order to further validate these conclusions, we also stress-tested the PRS estimation by a series of perturbations summarized in Figure S8, all of which retain the core result that using African American-estimated weights improves PRS discrimination. Substituting lnOR estimated directly from the whole-genome sequencing (Figures S8A–S8D), the overall PVE actually increases to 4.5% in the African ancestry samples, although the prevalence in the upper percentile is reduced a percentage point, while there is little impact on the European ancestry evaluation. Substitution of European meta-analysis effect-size estimates for ones generated from the UK Biobank also improves the PVE in Europeans (but not Africans), most likely because the down-sampling procedure—like the pseudo-control generation—uses a reduced sample size in the estimation. Furthermore, we removed various large-effect SNPs to evaluate whether they may be driving the improved performance. Figures S8E and S8F show that removing four...
variants with MAF < 0.02, three with OR > 1.2, has just a small impact on the prevalence-risk score relationship, as does removing the five variants with the largest absolute value of the OR, three of which have protective minor alleles.

**Potential of population diversity to improve fine-mapping**

Analysis of genetic data from divergent populations is thought to improve fine-mapping resolution of causal variants by leveraging trans-ethnic differences in LD, effect sizes, and allele frequencies. The fixed-effects meta-analysis of our African American whole-genome sequence data with summary statistics from European individuals\(^2\) at the previously fine-mapped (to 95% credible sets) loci\(^3\) showed marked improvement in fine-mapping resolution. Of the 3,529 credible variants from 94 loci that we meta-analyzed, 552 variants (16%) from 40 signals showed a change in p value of association with the previously assigned phenotype (inflammatory bowel disease, Crohn disease, or ulcerative colitis) by a factor of at least 10 (up to 100,000; Table S5). Of these, we noted an improvement in the strength of the association for 320 variants, while weaker association was observed for 232 variants. For instance, of the three credible variants that constituted a 95% credible set near Interleukin 23 Receptor (\(IL23R\) [MIM: 607562]; HD 7, signal 1 in Libioulle et al.\(^3\)), we noted an improvement in association at the variant with the highest posterior probability (PP) of being causal (PP = 0.49). Conversely, a weaker association was observed for one (PP = 0.26) of the three 95% credible set-variants near \(SBNO2\) (MIM: 615729). Similarly, 98 of the analyzed 120 credible variants near \(IRGM\) (MIM: 608212) (HD 67, signal 1, 95% credible size = 145 variants) and 102 (including a variant with PP = 0.58) of the analyzed 355 credible variants near \(DOCK3\) (MIM: 603123) (HD 50, signal 1, 95% credible size = 437 variants) demonstrated weaker evidence of association, supporting the potential of trans-ethnic meta-analysis as a means to further improve the resolution of credible sets that had been constructed with a homogeneous population.

**Discussion**

To further resolve the genetic architecture of inflammatory bowel disease and better define the differential genetic structure of the disease across divergent ancestries, we have performed the whole-genome sequencing analyses that include many alleles that were not previously examined in a population that remains very significantly understudied. Similar to the findings of a whole-genome sequencing association study of inflammatory bowel disease in 4,280 European affected individuals,\(^14\) we find little evidence for large effect rare variants explaining much of the heritability. We did observe an aggregate association of rare, likely deleterious variants at \(CALB2\) with ulcerative colitis, and many variants were specific to African populations (absent in European [non-Finnish] populations in gnomAD). However, though it exceeded exome-wide significance, this finding needs to be interpreted with caution because of the lack of replication in an independent cohort or functional validation. Our study assesses rare variant contributions to inflammatory bowel diseases in African Americans and will need evidence from future studies to further support the association at \(CALB2\) with ulcerative colitis. Nevertheless, our study highlights that multiple rare variants with small to moderate effects exist, and at least, when clustered in a small number of sets (genes, windows etc.), are likely to account for some of the missing heritability. However, much larger-scale, deep-sequencing studies will be needed to precisely estimate the variance in disease liability explained by such variants.
Besides providing further evidence for the emerging notion that the overall genetic risk of inflammatory bowel disease conferred by common alleles is shared across populations, our data highlight the impact that subtle differences in the effect size and/or allele frequency can have on the phenotype and the likelihood that rare variant contributions exert population-specific effects. Our effect size estimates were based on univariate associations at lead SNPs, an approach that most likely under-represents the variance explained at each locus because it does not capture the effect of secondary associations that are modeled by methods such as LD score regression. Since the apparent proportion of the variance explained by a particular SNP is the sum of the effects of that SNP plus those in LD (weighted by the extent of LD captured by the squared genotype correlation), local differences in LD and rare variant contributions will contribute to observed effect size differences. However, interactions with the genetic background as well as environmental/cultural differences, including access to healthcare and the age distribution of disease, all may also subtly contribute to variable allelic effects across populations. From the point of view of using PRSs in practice, whether the cause of the difference is due to main effects at the lead SNP or combined effects of LD at neighboring SNPs is largely immaterial. It remains to be seen whether such population-specific genetic contributions may provide insights into differences in diseases incidence and progression across populations.

Our analyses also provide an example of how polygenic analysis needs to be adjusted for ancestry when considering ethnic disparities in health care. It is well established that frequency distributions of PRSs can differ markedly across populations, mostly because of deviations in allele frequencies, although ascertainment biases in discovery of common variant associations are also a concern. Fine-scale mapping of diverse clinical, behavioral, and hematological traits in large multi-ethnic cohorts such as PAGE and BioME has been shown to identify new loci, resolve secondary signals, and quantify divergent allelic effects despite high overall levels of repeatability, consistent with our observations. Admixture mapping has also been used to refine PRS estimation, for example producing better discrimination in high risk percentiles for African Americans with multiple myeloma. New methods for incorporating such findings into polygenic prediction algorithms are rapidly emerging, notably those incorporating both local and global ancestry adjustments. Our findings similarly illustrate how observed risk distributions that differ across populations and as a function of the discovery ancestry group can be used to improve risk assessment. Optimal risk scores should combine genetic and clinical features and will require better estimates based on expanded sample collection within each ethnic background and in admixed populations.

It is not known to what extent the rank order of risk is conserved in trans-ethnic risk assessment: if it is, then a simple rescaling of the PRS distribution may suffice. However, we show, similar to a multi-ethnic comparison of diabetes genetic risk, that incorporation of ancestry-specific weights significantly increases the inferred risk at the upper tail of the polygenic risk distribution. This occurs despite overall conservation of the average variance explained per locus accounting for compensation of changes in allele frequency and allelic effect size. It is striking that most of the gains in risk assessment occur at the extreme tails of the PRS distribution. Our intuition about this is that for most people, increases in effects at some loci are offset by decreases at others, since the magnitude and sign of the effects are generally uncorrelated. However, in the top and bottom percentiles, there is an excess of alleles with the same sign of effect and necessarily those individuals share more genetic variations, which establishes a large enough correlation of ancestry-specific estimates to enhance the PRS discrimination in the correct population and add to the error in the incorrect one. Further research into the mechanisms responsible for ancestry-specific effects is warranted, including evaluation of the influences of linked rare variants, cumulative genetic background modifiers, and genotype-by-environment interactions.

Data and code availability

Individual-level whole-genome sequencing data relating to this study are available at dbGaP: phs001642.v1.p1. All other data are contained in the paper and its supplemental information or are available upon request.

Acknowledgments

The National Institutes of Health (NIH) grants DK062431 (S.R.B.); DK087694 (S.K.); DK062413 (D.P.B.M. and K.T.); DK046763-19, AI067068, and U54DE023789-01 (D.P.B.M.); DK062429 (J.H.C. and P.S.); DK062422 (J.H.C.); DK062420 (R.H.D.); DK062432 (J.D.R.); and DK062423 (M.S.S.) supported this study. Sequencing at the Broad Institute was supported by NHGRI grants U54HG003067-13 and UM1HG008895. S.R.B. was also supported in part by funding from Rutgers Crohns and Colitis Center of New Jersey. We acknowledge the clinicians and organizations that contributed to samples used in this study. Finally, we are grateful to the many families whose participation made this study possible.

Declaration of interests

The authors declare no competing interests.

Received: September 21, 2020
Accepted: February 1, 2021
Published: February 17, 2021

442 The American Journal of Human Genetics 108, 431–445, March 4, 2021
References

1. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124.

2. Liu, J.Z., van Someren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. 47, 979–986.

3. de Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., Jostins, L., Rice, D.L., Gutierrez-Achury, J., Ji, S.G., et al. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. 49, 256–261.

4. Fuchsberger, C., Flannick, J., Teslovich, T.M., Mahajan, A., Agarwala, V., Gaulton, K.J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D.J., et al. (2016). The genetic architecture of type 2 diabetes. Nature 536, 41–47.

5. Long, T., Hicks, M., Yu, H.C., Biggs, W.H., Kirkness, E.F., Menni, C., Zierer, J., Small, K.S., Mangino, M., Messier, H., et al. (2017). Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nat. Genet. 49, 568–578.

6. Prokopenko, D., Sakornsakolpat, P., Fries, R.H., Qiao, D., Parker, M.M., McDonald, M.N., Manichaikul, A., Rich, S.S., Barr, R.G., Williams, C.J., et al. (2018). Whole-genome sequencing in severe chronic obstructive pulmonary disease. Am. J. Respir. Cell Mol. Biol. 59, 614–622.

7. Zekavat, S.M., Ruotsalainen, S., Handsaker, R.E., Alver, M., Bloom, J., Poterba, T., Seed, C., Ernst, J., Engreitz, J., et al.; NHLBI TOPMed Lipids Working Group (2018). Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. Nat. Commun. 9, 2606.

8. de Vries, P.S., Yu, B., Feofanova, E.V., Metcalf, G.A., Brown, M.R., Zeighami, A.L., Liu, X., Muzny, D.M., Gibbs, R.A., Boerwinkle, E., Nils, A., and Morrison, A.C. (2017). Whole-genome sequencing study of serum peptide levels: the Atherosclerosis Risk in Communities study. Hum. Mol. Genet. 26, 3442–3450.

9. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33.

10. Kotlar, A.V., Trevino, C.E., Zwick, M.E., Cutler, D.J., and Wingo, T.S. (2018). Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol. 19, 14.

11. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

12. Smit, A.F.A., Hubley, R., and Green, P. (2013–2015). RepeatMasker Open-4.0. RepeatMasker Open-4.0.

13. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909.

14. Luo, Y., de Lange, K.M., Jostins, L., Moutsianas, L., Randall, J., Kennedy, N.A., Lamb, C.A., McCarthy, S., Ahmad, T., Edwards, C., et al. (2017). Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat. Genet. 49, 186–192.

15. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiansi, D.C., Wurfel, M.M., Lin, X.; and NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91, 224–237.

16. Brant, S.R., Okou, D.T., Simpson, C.L., Cutler, D.J., Haritunians, T., Bradford, J.P., Chopra, P., Prince, J., Beugm, E., Kumar, A., et al. (2017). Genome-wide association study identifies African-specific susceptibility loci in African Americans with inflammatory bowel disease. Gastroenterology 152, 206–217.e2.

17. Bioconductor Package Maintainer (2020). liftOver: Changing genomic coordinate systems with tracklayer:liftOver. (R package version 1.12.0).

18. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Refshøj, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the HaploTreff Reference Consortium panel. Nat. Genet. 48, 1443–1448.

19. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.L., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. 48, 1284–1287.

20. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913.

21. de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. 17 (R2), R122–R128.

22. Liu, J.Z., Tozzi, F., Waterworth, D.M., Pillai, S.G., Muglia, P., Middleton, L., Berrettini, W., Knuff, C.W., Yuan, X., Waaber, G., et al.; Wellcome Trust Case Control Consortium (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. 42, 436–440.
studies of multiple myeloma among men and women of African ancestry. Blood Adv. 4, 181–190.

49. Martin, E.R., Tunc, I., Liu, Z., Slifer, S.H., Beecham, A.H., and Beecham, G.W. (2018). Properties of global- and local-ancestry adjustments in genetic association tests in admixed populations. Genet. Epidemiol. 42, 214–229.

50. Zhong, Y., Perera, M.A., and Gamazon, E.R. (2019). On using local ancestry to characterize the genetic architecture of human traits: genetic regulation of gene expression in multiethnic or admixed populations. Am. J. Hum. Genet. 104, 1097–1115.

51. Shi, H., Burch, K.S., Johnson, R., Freund, M.K., Kichaev, G., Mancuso, N., Manuel, A.M., Dong, N., and Pasaniuc, B. (2020). Localizing components of shared tranethnic genetic architecture of complex traits from GWAS summary data. Am. J. Hum. Genet. 106, 805–817.

52. Márquez-Luna, C., Loh, P.R., Price, A.L.; South Asian Type 2 Diabetes (SAT2D) Consortium; and SIGMA Type 2 Diabetes Consortium (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. 41, 811–823.