Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

_____ July 31, 2025

Ranjit Pelia Date

Characterizing the Genetic Landscape of Inflammatory Bowel Disease Across Populations

By

Ranjit Singh Pelia

Genetics and Molecular Biology

Dr. Subramaniam Kugathasan
Advisor

Accepted:

Kimberly Jacob Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

Date

Characterizing the Genetic Landscape of Inflammatory Bowel Disease Across Populations

By

Ranjit Singh Pelia

Associate of Arts, Oxford College of Emory University, 2015
Bachelor of Science, Emory University, 2018
Master of Public Health, Rollins School of Public Health, Emory University, 2022

Thesis Committee Chair: Subramaniam Kugathasan, M.D.

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Genetics and Molecular Biology
2025

Abstract

Characterizing the Genetic Landscape of Inflammatory Bowel Disease Across Populations

By

Ranjit Pelia

Background: Inflammatory bowel disease (**IBD**) is a complex, polygenic, and multi-faceted disease comprised of two main forms, Crohns disease (**CD**) and ulcerative colitis (**UC**). Despite the similar prevalence of IBD among Americans, there are stark differences in severity amongst individuals with European versus African, admixed, ancestries. Over 300 susceptibility loci have been identified in IBD. Detecting statistically significant genetic variants across populations will aid clinicians by optimizing for IBD-subset and patient heterogeneity.

Objective: The goal of this proposal is to characterize novel and assess known IBD-risk associated loci using one of the most diverse genomic datasets to date. Our combined- blended-Genome-Exome (cBGE) approach will elucidate novel IBD related associations, improve genetic risk predictions across all populations, and provide mechanistic insights. Our overarching objective is to characterize IBD-risk associated loci across diverse populations using combined-Blended Genome Exome sequencing (cBGE).

Method: Peripheral blood samples derived from IBD patients and controls was sequenced using cBGE and combined with Whole Genome Sequencing to generate a predominantly African American population dataset. Variants were identified following best practices using GATK, annotated with Bystro, and quality control analyses using PLINK2. Common and rare variant association testing, Polygenic Risk Scores and pathway analyses was performed using IBD, CD, and UC specific variants across populations. Results were compared with previously discovered IBD associated loci and novel findings were reported.

Results: A merged, n=1794 cBGE and n=3608 WGS, dataset was harmonized leading to n=5374 after harmonization. Over 6.5 million variants were observed in IBD patients comprised of SNPs and INDELs. Genomic inflation, λ in IBD, CD, and UC was λ=1.014, λ=1.012, and λ=1.014. We observed *PTGER4*, *CARD9*, and *IL23R* common and rare variants across disease subtypes. Polygenic risk scores were more similar across IBD and CD compared to UC. Pathway analysis highlight cell adhesion in IBD, chromatin remodeling in CD, and T-cell regulation in UC.

Conclusion: The duality of cBGE with WGS increased our power from previous investigations leading to validation of known IBD-associated loci. No significant novel variants were observed showing the limitations of cBGE. Here, we provide the largest known African American population genetic dataset in IBD.

Characterizing the Genetic Landscape of Inflammatory Bowel Disease Across Populations

By

Ranjit Singh Pelia

Associate of Arts, Oxford College of Emory University, 2015
Bachelor of Science, Emory University, 2018
Master of Public Health, Rollins School of Public Health, Emory University, 2022

Thesis Committee Chair: Subramaniam Kugathasan, M.D.

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science
in Genetics and Molecular Biology
2025

Acknowledgements

I am eternally grateful to Dr. Subramaniam Kugathasan for his endless support, advise, and providing me with the best resources to succeed. I am grateful for my lab manager, Anne Dodd, co-mentors Dr. Vasantha Kolachala, Dr. Jason Matthews, and Dr. David J. Cutler. Thank you to the Inflammatory Bowel Disease Genetics Consortium and the Broad Institute at MIT for providing us the sequencing, cBGE platform, and analytical support for this project. I would like to thank Dr. David Katz, Dr. Jie Jiang, and Dr. Roger Deal for their continuous guidance throughout my time at the Genetics and Molecular Biology program. My journey in the program would not be possible without them.

Table of Contents

Introduction

Inflammatory Bowel Disease: Inflammatory bowel disease (**IBD**) is a complex, polygenic, and multi-faceted disease comprised of two main forms, Crohns disease (**CD**) and ulcerative colitis (**UC**). IBD is generally described as chronic inflammation of the gastrointestinal tract (**GI**) with alternating patterns of clinical relapse and remission. IBD is a very heterogenous disease which symptoms differing based upon type of IBD, disease location, extent of inflammation, and other comorbidities. Genesis of IBD can occur at any stage of life, with diagnosis occurring in early childhood or between 20 to 40 years. If untreated, IBD can lead to significant weight loss, diarrhea, intestinal bleeding, blood clots, and fatigue. In severe cases, patients exhibit abscesses, fistulas, obstructions, and perforations across the GI tract. Oncogenesis, specifically cholangiocarcinoma, colorectal cancer, intestinal lymphoma, and small bowel cancer has been associated with IBD[1]. The etiology and pathogenesis of IBD is unknown with genetic and environmental factors being involved.

The cases of IBD have been increasing rapidly across the world, making it a substantial global health burden. Developing regions in Asia and Africa have shown to exhibit the greatest rise in incidence compared to developed, Western, countries. IBD prevalence was shown to affect 5 million individuals, with an incidence of 400,000 new cases according to the 2019 Global Burden of Disease report. In industrialized regions, EpiCom/Epi-IBD showed 15 cases for every 100,000 person-years incidence[2]. The stark rise in prevalence and incidence of IBD amongst developing and industrialized areas suggests environmental factors contributing to the rise of IBD. The overarching hypothesis is that genetic and environmental factors lead to modifications in the lining of the gut, disruption of microbiota, and creating a state of chronic inflammation. There are multiple gaps in our understanding of IBD pathophysiology, role of genetics, environmental, and overall heterogeneity. Somineni et al[3] highlighted four key hurdles: I) missing heritability, II) no known causal variants, III) limited functional understanding of known genetic loci associated with IBD, and IV) paucity in our knowledge of IBD in different populations, i.e. ancestries. Here, I summarize the biological, clinical, and genetic background of IBD and build upon prior works to showcase novel results contributing to the genetic landscape of IBD across populations.

The gut epithelium barrier functions to regulate host-to-environment interactions through immune and microbiota regulation. IBD patients exhibit a “leaky gut” where dysbiosis of immunological, bacterial, and epithelium occur. Intestinal barrier regulation is quintessential and the primary center of focus for cellular dysbiosis in IBD. The main function of the intestinal barrier is to serve as a gate for communication amongst food, microbiota, and the GI tract. Core components of the intestinal barrier include bile acids, epithelial layer, enterocyte layer, luminal enzymes, and water layer. Physically from outermost to inner, the layers of the intestinal barrier are the gut lumen, mucus layers (thick, thin), unstirred water layer, epithelium, and the lamina propria. The gut lumen houses gastric acid, pancreatic enzymes, and commensal bacteria which excrete antimicrobial peptides to deter pathogens. Below this protective layer, the thick-, thin-mucus, unstirred water layer, and glycocalyx deter adhesion of bacteria via immunoglobulin A secretion (**IgA**) and physical properties. Below the mucus layers, the epithelium comprises of tight junctions that function to link epithelial cells and luminal contents. Barrier integrity proteins are dispersed throughout tight junctions, consisting of claudin-1, claudin-4, junctional adhesion molecules, occluding, and zonula occludens-1[4]. At the bottom, the lamina propria houses innate immune cells and aids in communication with the endocrine and enteric nervous system. Poor intestinal barrier function leading to foreign, pathogenic, or disruptive macromolecules

progressing to the lamina propria activates inflammation and multiple immune pathways. IBD patients exhibit poor intestinal gut barrier function which is often reflected by decreased tight junctions and increased permeability[5]. There is a myriad of ways gut epithelium barrier function can be dysregulated, which contributes to the heterogeneity of IBD.

Clinicians use endoscopic evaluations via colonoscopy for diagnosis as the current gold standard. Barrier disfunction is often measured through a combination of histological assessment of mucosal biopsies, confocal endomicroscopy of duodenum, and oral probes measuring urinary excretion of di- and monosaccharides[6]. Alongside, histological, laboratory, radiologic, and serological markers are also used for validating IBD and subtype[7]. Less invasive tools such as stool samples for increased levels of fecal calprotectin (**FCP**) is used for determining levels of inflammation. FCP is often released because of cellular damage by activated immune cells and is found in epithelial, macrophages, and monocytes cells. to serological measurements are also used in diagnostics. In which, increased levels of C-reactive protein (**CRP**) and erythrocyte sedimentation rate (**ESR**) are often indicators of chronic, prolonged, inflammation. Although these levels could be due to inflammation from any part of the body thus testing with other diagnostic tools is required for accurate determination of IBD. Differing patterns of serological markers such as Antineutrophil cytoplasmic antibody (**ANCA**), Anti-*Saccharomyces cerevisiae* (**ASCA**) antibody, and antineutrophil cytoplasmic antibody (**pANCA**) are used to differentiate between UC and CD. Specificity of using pANCA+ and ASCA- to accurately sequester UC from CD was shown to be 94.4%[8]. Follow-up studies of using serological markers as diagnostic tools for IBD have shown that not all patients express these antibodies and their levels can be impacted by treatments.[9] Proper diagnosis, treatment, and patient-specific multi-modal therapies are quintessential to maintain, improve, and provide adequate quality of life for IBD patients.

To limit the role of heterogeneity in IBD when identifying disease subtypes, classification systems are used for both CD and UC to promote consensus amongst clinicians and researchers. The Montreal system of classification was established in 2005 based upon multiple studies that suggest disease location, behavior, and progression can be accurately identified by genetic, endoscopic, and serological markers[10]. For Crohns disease, the system uses three main categories: age at diagnosis, location, and behavior. Age is grouped by A1 (below 16 years), A2 (between 17- and 40 years), and A3 (above 40 years). Location is based on L1 (ileal), L2 (colonic), L3 (ileocolonic), and L4 (isolated or concomitant upper gastrointestinal disease). Behavior categories include B1 (non-stricturing, non-penetrating), B2 (structuring), B3 (penetrating), and p (perianal) modifier. Ulcerative colitis classification is based upon extent and severity of disease. Extent comprises of E1 (ulcerative proctitis where inflammation is distal to rectosigmoid junction), E2 (left-sided, involvement of colorectum distal to splenic flexure), and E3 (pancolitis, proximal to splenic flexure). Severity of UC involved four groups, S0 (remission or asymptomatic), S1 (mild UC, minimal blood in stool with 4 or less passages per day, normal erythrocyte sedimentation rate (**ESR**)), S2 (moderate UC, more than 4 stool passages per day with low systematic toxicity), and S3 (severe UC, more than 6 bloody stool passages per day, high temperature, or ESR above 30 mm/h)[11]. Recently, the Pediatric Ulcerative Colitis Activity Index (**PUCAI**) was introduced to provide a numerical, score, based system to assess phenotype severity of UC. It considers six main components: abdominal pain, rectal bleeding, stool consistency, number of daily stools, nocturnal stools, and activity level. These categories contain a range of scores from 0, 5, to 10 and are summed with a range of 0 (least severity) to 85 (most severe)[12]. In summary, these categorical tools used by clinicians allow researchers to ascertain

disease subtype, severity, and other phenotype specific conclusions by simplifying the complexity or heterogeneity of IBD through classification systems.

Diverse populations in the gut are essential for proper function. The role of microbiota in IBD is hypothesized to effect disease by alterations of mucosal barrier, cellular immune communication, and metabolizing dietary and host-derived compounds[13]. IBD subtypes can also be classified based on gut microbiota populations. Pathogenically, bacterial presence of *Aeromonas*, *Campylobacter*, *Clostridium Difficile*, *E. coli*, *Salmonella*, *Shigella*, tuberculosis, and *Yersinia* are hypothesized to contribute to etiology. Acute gastroenteritis greatly increases risk of IBD which can arise through *Campylobacter* and *Salmonella*[14]. The two main forms CD and UC share clinical characteristics of chronic relapsing abdominal pain and diarrhea. All forms of IBD exhibit some degree of gut epithelium barrier dysregulation. In CD, tight junction proteins show upregulated levels of claudin-2 and myosin-light-chain-kinase, MLCK, activation, often before clinical onset or relapse. Whereas UC patients show decreases in occludin proteins amongst tight junctions and cytoskeleton[15]. CD is primarily observed in a transmural inflammatory pattern amongst any part of the GI tract and has a longer diagnostic delay compared to other subtypes.[7] Other characteristics associated with CD include: terminal ileal involvement, skip-lesions, strictures, fistulas, or perianal disease. UC manifestation in the GI tract is limited to the mucosal layer of the colon with patients often exhibiting rectal bleeding, diffuse mucosal inflammation, back-wase ileitis, or continuous lesions. In general, CD patients have more severe symptoms, require more intensive medical care, and are at higher risk of developing complications. Patients that cannot be accurately diagnosed with CD or UC and do not have irritable bowel syndrome (**IBS**) are classified into another subtype of IBD, indeterminate colitis. Approximately 5-15% of IBD patients have indeterminate colitis[16]. CD and UC are distinct forms of IBD that require different treatments, therapies, and management.

Multiple medical interventions are available as initial treatment for IBD such as, aminosalicylates (**ASA**), corticosteroids, immunomodulators, and surgical interventions. The main purpose of treatments is to treat symptoms, mitigate disease progression, and provide improved quality of life for IBD patients. UC patients have decrease in symptoms using 5-ASA compared to CD and continuous oral therapy has shown to reduce colorectal cancer risk by 75%[17]. For UC patients that are non-responders for ASA and CD patients with mild-to-moderate disease, corticosteroids have shown to decrease inflammation but are not efficacious for maintaining remission[18]. An increased, chronic, inflammation state in the GI tract cascades into pro-inflammatory cytokines production, release, and filtrations. The most notable are interleukins (**IL-12** and **IL-23**), interferon-gamma **(IF-γ**), and tumor necrosis factor (**TNF**)[19]. Canonical medical interventions for IBD also utilize anti-TNF suppressors, Rx Infliximab, Rx Adalimumab, Rx Golimumab, and Rx Ustekinumab. Alongside TNF, the cytokine cascade pathway of Janus kinase (**JAK**) signal transducer and activator (**STAT**) is over-activated in IBD patients. Inhibitors of JAK-STAT pathway, Rx Filgotinib, Rx Tofacitinib, and Rx Upadacitinib are also used as frontline treatments[20]. Often, first-line and second-line medical interventions often fail overtime to reduce symptoms, maintain remission, and mitigate relapse. Anti-TNF has shown to be non-effective in 40% of patients and in patients where it is efficacious initially, it leads to non-responders status after 1-year in 23-46% of IBD patients[21]. Due to patient specificity, heterogeneity, and poly-factorial nature of IBD, multiple therapies in combination with lifestyle, nutritional, and dietary considerations are quintessential for effective, long-lasting treatment.

Genetics of Inflammatory Bowel Disease: Heritability measurements using concordance rates through twin-studies can shed light on the likelihood of a phenotype or disease being genetic, i.e. passed from one generation to another. Adding to the complexity, subtypes of IBD share different concordance rates implying genetics to play disproportional roles and alluding to other factors contributing to disease. Monozygotic and dizygotic rates were shown to be 50% and 10% in CD compared to much lower 15% and 4% in UC[22, 23]. Similarly, having a positive family history of IBD was associated with increased risk of CD, 1.5%-28%, compared to UC, 1.5%-24%[24]. Although no large-scale studies have examined inheritance of microbiota in IBD, there is large evidence suggesting genetic loci to affect gut microbial variation[25]. Monozygotic siblings share greater microbiome than dizygotic, who share a larger proportion than unrelated individuals[26]. Specifically, increased genetic risk for IBD was associated with decreased levels of *Roseburia*, a converter of acetate to butyrate[27]. There is a need for large-scale, multi-population, and incorporation of environmental (diet, microbiota, lifestyle) genetic studies to ascertain putative IBD risk, protective, and causal variants.

With the advent of sequencing, a shift from familial based to population genetics has led to novel discoveries. Whole-Genome Sequencing, Whole-Exome Sequencing, Genome-Wide Association Studies, Microarrays, and DNA-methylation are the primary forms of genetic tools researchers have utilized in studying IBD. Variants detected at specific loci are denoted single nucleotide polymorphisms (**SNP**) and are often represented through their minor allele frequency (**MAF**), odds ratio (**OR**), or effect size. Over 200 genetic loci have been associated with IBD using Genome-Wide Association Studies (GWAS)[28], yet there are no generalizable targets across populations and disease subtypes. Amongst the known IBD associated loci, *NOD2* and *ATG16L1* are some of the most notable genes connected to variants. The first and one of the strongest risk variants for IBD is *NOD2* (encoding nucleotide-binding oligomerization domain-containing protein 2), with an OR=3.1 in CD[29]. *NOD2* is observed in a wide variety of cell types across the GI tract amongst Paneth, T-cells, macrophages, and monocytes. *NOD2* functions in host-microbial immune response by cytosolic receptor pattern recognition. An intracellular receptor is encoded by *NOD2* upon bacterial signaling on bacterial peptidoglycan muramyl dipeptide. Upon activation, muramyl dipeptides form an active oligomer to recruit adapter protein which initiate either an inflammatory response or *ATG16L1* autophagy[30]. Alongside autophagy, *ATG16L1* counteracts endoplasmic reticulum stress and reduces spontaneous apoptosis in intestinal epithelium. Multiple SNPs have been observed in both *NOD2* and *ATG16L1*. Primarily GWAS of IBD patients with African American ancestry have yielded *NOD2* association with limited replication amongst other ancestries: Japanese, Chinese, and Korean[31]. The involvement of different microbiome, alternative variants, or a combination of environmental with genetic intricacies may explain IBD pathogenesis across populations.

Population Genetics and Inflammatory Bowel Disease: Inflammatory Bowel Disease progression is more severe with increased intestinal resections with almost twice the rate, OR = 2.49, amongst admixed African ancestry (**AA**) compared to European ancestry (**EA**) patients[32]. GWAS loci only account for ~15% of IBD heritability in EAs but the genetic risk in AA populations remains undertermined[33]. Although rare alleles are more likely to be population specific, shared genetic IBD loci may imply generalizable mechanisms of pathogenesis across ancestral populations[34]. Liu et al identified 38 new IBD-associated risk loci using a Bayesian trans-ancestry meta-analysis by combining European- and African-ancestry samples into a European-only cohort of 75,105 samples[34]. Only 13 loci were novel as the other 25 overlapped with known trait-associated loci. Of these loci, population specific variants such as *NOD2* and

IL23 demonstrated population-specific genetic heterogeneity. *NOD2* was observed with a larger effect, OR = 2.13-3.03, in European ancestries, but showed a Risk Allele Frequency (**RAF**) of 0 in East Asians[34]. Liu et al had insufficient power to measure known population-specific IBD risk associated variants in African admixed populations.

Prior IBD genetic studies lacked coverage of variants, inaccurate effect sizes, or inadequate statistical power for generalizable interpretations. In comparing common variants across studies, datasets can be jointly combined and assessed using regression coefficients and standard errors[35]. Previously, large scale meta-analysis studies of IBD have not comprised of sufficient samples of African populations[34, 36, 37]. Vast majority of meta-analysis in IBD have predominantly captured European populations which do not capture all genetic variation. For example, Polygenic risk scores derived from predominantly European ancestry GWAS studies are poorly transferable on non-Europeans thus failing to provide generalizable clinical interpretations[38]. *NOD2* was one of the first CD risk associated gene to be identified with an allele frequency of 13% amongst nine low-frequency IBD causal variants in European populations[39]. Interestingly, *NOD2* across these same nine variants was only 0.06% in a cohort of IBD patients with East Asian ancestry[34]. These population-specific observations may be reflected across other IBD-risk associated variants.

A delicate balance of sample size, sequencing depth, coverage, and type of samples must be considered when performing large-scale population genetic studies. The cost of sequencing is affected by coverage, depth, and number of samples. Coverage is defined as the proportion of sample sequenced to its entirety. Depth is a measurement of the number of reads observed at a given nucleotide[40]. GWAS, Whole Exome, and Whole Genome technologies are mainly limited by their cost. Here, I am proposing to lessen this paucity by using combined-blended genome-exome sequencing (**cBGE**), which provides this duality with a fraction of the cost, ~$44 per sample. cBGE allows for 33% exome and 67% genome libraries to be sequenced with low-whole genome (2-3X) combined with higher-coverage exome (30-40X)[41]. Despite the benefit of increased coverage by performing WES or WGS separately, the cost is ten-fold cheaper by using cBGE. Our first WGS[3] comprised of AAs with IBD consisted of n=2121 cases and n=4922 controls. Despite our sample size, we were unable to reach genome-wide significance for previously observed IBD associated loci[34].

By adding other population groups, we aim to capture variants in non-European ancestries that were in too low frequencies in previous IBD association studies. My overarching objective is to characterize IBD associated loci across diverse populations using combined-Blended Genome Exome sequencing (cBGE) and cross-validate previous discoveries using a fixed-effect meta-analysis. I hypothesize that the much of the higher incidence, worse prognosis, and higher complication rates of IBD in African American compared to European populations are due to previously unrecognized genetic variants. My specific aim I is to identify common and rare variants associated with IBD patients with admixed African American ancestry using cBGE. I hypothesize that rare variants in IBD effectively differentiate participants with primarily European- from primarily African ancestry compared to rare variants. My specific aim II is to calculate a polygenic risk score (**PRS**) of known IBD associated loci by using my cBGE dataset. I hypothesize that IBD associated variants contribute ancestry-specific risk amongst populations. Together, we aim to elucidate novel discoveries and validate previous findings by assessing IBD associated variants across populations.

Methods

Inflammatory Bowel Disease Genetics Consortium (IBDGC): Established in 2002 through the NIDDK, IBDGC's main mission is to increase our current understanding of IBD through furthering research into the genetic structure, pathophysiology, and improve patient outcomes[42]. Other underlying objectives encompass I) identification of underlying genetic risk factors of IBD, CD, UC, and related clinical phenotypes, II) the relationships of non-genetic risk factors with genetic in development of IBD, CD, UC, and related phenotypes, and III) ensuring availability, communication, and sharing of data across the international research and scientific communities[43] IBDGC is a multi-site consortium across the United States. All samples from WGS (n=1794) and cBGE (n=3608) datasets were gathered through BDGC and sequenced at MIT BROAD Institute. Phenotypic data was captured at each site, compiled through IBDGC consortium, and de-identified before becoming available for analytics.

DNA Sample Collection, Extraction, Clinical Phenotypes, and Sequencing: For all cBGE samples, DNA Extraction and Processing for cBGE: DNA (~500 µl saliva) was collected, purified with prepIT.L2P reagent and processed with Oragene OGR-500 Kit from DNA Genotek. Consequentially, DNA was clarified by centrifugation, ethanol precipitated, pelleted, and resuspended in TE buffer. Collected DNA samples were then processed with the Blended Genome Exome (BGE) sequencing reagent. Briefly, a whole genome library is generated with an aliquot PCR amplified for exome regions. Then for each sample, genome and exome libraries are combined for sequencing at 33% exome and 67% genome. This yields low-coverage whole genome (2X-3X) combined with higher coverage exome (30X-40X)[41]. Sequencing was performed on the Illumina NovaSeq X Plus platform. Reads were aligned using Illumina Dragen aligner and HG38 reference genome which results in a single CRAM file per sample. Due to differences in library preparation and aligner used in cBGE, the blended genome- and exome- are not directly combinable thus need to be processed separately for WGS and WES analyses. Post alignment, variant call format (VCF) files were received and stored locally at Emory University Human Genetics Cluster Core (HGCC).

Previous WGS samples from Somineni et al[3] were processed using genomic DNA (350ng per 50 µl) of peripheral blood samples from controls and patients with IBD. Fragmentation to 385 bp was performed using Cavaris Focused-ultra sonification with SPRI size selection. Libraries were constructed with KAPA Hyper Prep without amplification and Roche based adapters with unique 8-base indexes and palindromic forked-adapters. Afterwards, quantified PCR, normalization to 2.2 nM, and pooling into 24-plexes was performed for genomic libraries. These pooled samples were treated with HiSeqX Cluster Amp Reagents EPX1, EPX2, and EPX3 before Illumina cBot cluster generation. Sequencing was conducted with HiSeqX with 151 bp paired-end reads. CRAM files were generated from sequencing output using Picard pipeline from the BROAD Institute to create aligned and demultiplexed reads for analysis. All files are stored locally at HGCC.

Clinical phenotypes reported for all samples (cBGE and WGS) included diagnosis (CD, UC, or control), age, biological sex, family data (.ped), and self-reported ancestry. Due to these samples being recruited from multiple sites across the United States, only samples from our local, Emory University, site contain disease location for CD (Montreal), disease severity for UC (PUCAI), treatment status, and disease complication (surgeries, perianal disease, remission, relapse, or comorbidities). For consistency and simplification, only disease status (IBD, CD, UC, or control) and admixture (AFR, AMR, SAS, EAS, or EUR) were considered for all (cBGE and WGS) samples. An overview of the specific aims and approach is shown in **Figure 1**.

Harmonization, and First-Pass Quality Control: Harmonization is a quintessential process for any multi-dataset analysis where all related datasets are reconciled for maximum compatibility and comparability[44]. Samples from cBGE and Somineni et al[3] will be merged for an increase in sample-size, i.e. power using PLINK2 and Bystro[45] for annotation on Emory University's HGCC. Using patients and participants from IBDGC, I used the merged, cBGE and WGS from Somineni et al[3] to ascertain if combining WGS with cBGE is consistent with expected observations. Due to the novelty of cBGE, we expect low-coverage and high missingness of intergenic regions compared to WGS. Therefore, assessing metrics using a similar platform, Bystro[45], is quintessential for evaluating whether to merge or keep independent by comparing the output summary statistics. To facilitate this, samples were extracted for PLINK2[46] files (.ped, .fam, .bed) for WGS and cBGE and merged using PLINK2[46] and annotated using Bystro[45]. Using a preliminary run of n=1794 cBGE samples and n=3608 WGS from Somineni et al[3].

The merged, cBGE+WGS dataset was aligned with human reference genome build hg38 (GRCh38). Using Genome Analysis Toolkit Best Practices for germline variants (GATK)[47], variants were jointly called and annotated using Bystro[45]. Summary statistics were used to assess if merging preserves the genomic data with minimal effect of sequencing platform, cBGE vs. WGS. Data was harmonized for consistency amongst SNPs or variant identification (ID) numbers. Variant IDs were compared based on chromosome, position, reference allele, alternative allele, and corresponding odds ratios, effect sizes, *p*-values, or standard errors for disease (IBD, CD, or UC), control, and ancestry classifiers.

Using PLINK2, per-individual sample quality control was performed to assess metrics of; I) missingness of SNPs and individuals, II) discordant biological sex, III) Hardy-Weinberg equilibrium (HWE), IV) heterozygosity, V) relatedness, and VI) population stratification. These quality controls steps will be performed for WES-VCF and WGS-VCF files separately. SNP filtering will be performed before any individual or sample filtering. Specifically, variants genotyped in < 95% of samples (missingness > 5%) and those with Hardy-Weinberg equilibrium in control individuals ($p < 1 \times 10^{-9}$) were omitted. After removal of samples that did not meet quality control thresholds, variants with Minor Allele Frequencies (MAF) of <0.05 were defined as common variants and those with MAF <0.01 as rare variants. Variants in repetitive and low-complexity regions were masked using RepeatMasker[48]. Principal components of genetic variants was calculated by EIGENSTRAT[49] for the cBGE-WGS merged dataset.

After alignment and variant calling using GATK, both the merged dataset VCFs were processed through quality control filtering using PLINK2. First, missingness or SNPs that are missing in a large proportion of the samples, at 0.20 threshold, were filtered. After removal of SNPs with low genotype calls, we excluded individuals with high missingness. Sex discrepancy was calculated using X-chromosome heterozygosity/homozygosity rates. Biological females and males denoted by X chromosome homozygosity estimate that is <0.20 and >0.80, respectively. In terms of genotype filtering threshold, these were defined as samples that deviate +/- 3 standard deviations from the mean. Relatedness was determined by calculating identity by state for each pair of individuals, followed by shared ancestry for identity by descent using 2, values ranging from 0 to 1. Duplications or related samples are those with identity by descent values > 0.5 and if detected, one of the two will be discarded based on quality metrics or at-random if both samples share high quality characteristics from any other quality control steps. For first-degree relatives and second-degree relatives, identity by descent values were set as ~ 0.5 and ~0.25, respectively; one individual from each pair was removed with high overall quality being the eliminating criteria.

Common Variant Association Analysis: Our cBGE dataset, in combination with Somineni et al WGS, is the largest IBD dataset of African American individuals to-date. We performed a power calculation using the known IBD associated loci and corresponding odds ratio from Liu et al[34] to infer our ability to detect variants with statistical confidence. Given our sample size of greater than 5000, I am confident in our ability to detect many common and some rare variants in IBD. After quality control of individuals, the remaining samples from cBGE-WGS merged dataset were assessed for common and rare variants. We tested for common variants with MAF >1% and <5% for all three IBD phenotypes, CD, UC, and IBD. SNPTEST v2.4.1 was used for common variant analysis with both Bayesian and Frequentists tests[50]. Using Hardy-Weinberg Equilibrium (**HWE**) via PLINK2, we excluded markers for binary traits which were *p*-value <1e-8 in cases and *p*-value <1e-8 in controls. For quantitative traits, HWE threshold for filtering was *p*-value <1e-8. A logistic regression single-variant association test was used for chromosomes 1-22, excluding X and Y. Any duplicate variants, missing variant IDs, or mismatches were filtered to generate a master-list of variants with corresponding OR, *p*, MAF, and test-statistic. The genomic inflation factor (λ) was calculated using observed chi-squared test statistic values for all detected variants amongst IBD-, CD-, and UC vs controls from PLINK2 output using one degree of freedom. The top 1000 SNPs for each comparison, IBD-, CD-, and UC-vs controls have been provided and further assessed for clinical and biological relevance. Variants were assessed for their allele frequencies across other populations using gnomAD version 4.0[51]. These groups comprised of African/African American (**AFR**), Amish (**AMI**), Admixed American (**AMR**), Ashkenazi Jewish (**ASJ**), East Asian (**EAS**), Finnish (**FIN**), Middle Eastern (**MID**), European non-Finnish (**NFE**), and South Asian **(SAS**).

Rare Variant Association Analysis: Genetic loci with both MAF <0.01 and *p*-value <1e-8 were considered as statistically significant and rare variants. Variants that were likely deleterious based on combined annotation dependent depletion (CADD) greater than 15 were selected. Using the merged cBGE-WGS dataset after quality control, rare variants were aggregated to nearest gene (HG38) and tested for association by optimal sequence kernel association test (SKAT-O) using SKAT package in R. All three phenotypes: IBD vs controls, CD vs controls, and UC vs controls were assessed for potential rare variants that meet CADD, MAF, and genome-wide significance *p*-value <1e-8. Afterwards, the allele frequences of these rare variants were compared to those across populations: AFR, AMI, AMR, ASJ, EAS, FIN, MID, NFE, and SAS using gnomAD version 4.0. Due to the few numbers of rare variants observed across disease subtypes, pathway analysis was not assessed due to lack of minimal hits in query.

Polygenic Risk Scores: Using PLINK2, known IBD associated loci from Liu et al[34] were extracted for corresponding variant IDs, odds ratios, alternative allele, reference allele, and *p* values. The distribution of PRS scores was assessed across IBD subtypes, IBD vs controls, CD vs controls, and UC vs controls, and population groups (AFR, AMR, EUR, SAS, and EAS) for any statistical significance using a standard T-test.

Pathway Analysis: To calculate enrichment scores for Kyto Encyclopedia of Genes and Genomes (**KEGG**) and Gene Ontology biological processes, cellular components, and molecular functions, the top 1000 variants were considered across IBD vs controls, CD vs controls, and UC vs controls. Enrichment testing and evaluation was performed using GO Enrichment Analysis and following best practices in R[52]. All pathways were considered significant if *p* adjusted < 0.01 after correcting for multiple testing using a weighted Fisher test.

Results

Power Calculation and Summary Statistics of Clinical Phenotypes: To generate a foundation for any statistical ascertainments, it is critical to assess for sufficient sample sizes. I used a power calculation based upon known IBD associated loci from Somineni et al[3] to generate theoretical MAFs based upon ORs (**Figure 2**). I assumed 80% power, alpha set to $5x10^{-8}$, and a Chi-square distribution with n=1 degrees of freedom. Using 5000 cases and 5000 controls dataset, I am confident to detect risk allele frequencies of 0.05 with OR=1.45 with statistical significance (**Table 1**). The cBGE sequencing of samples is ongoing at IBDGC with n=1794 samples processed thus far.

To increase the sample size, I harmonized and merged Somineni et al[3] with cBGE dataset. Using the CRAM files from WGS and cBGE, an integrated VCF.gz file was generated. Our initial dataset, pre-quality control consisted of n=5402 samples, of which n=3608 were from WGS and n=1794 from cBGE. To combine the two datasets for consistency, the variables: sample-IDs, disease status (IBD, UC, CD, or control), and type of study (WGS or cBGE) was compiled. Using this list, we first needed to calculate summary statistics of the entire dataset to eliminate any outliers. To do this, the VCF.gz file of all samples was uploaded to BYSTRO[45] to generate sample specific measurements of total variants, theta, and types of SNPs (silent, replacement, transitions, transversions, homozygotes, heterozygotes, exonic, and intronic). Missingness of variants was calculated using PLINK2 and considered for quality control to eliminate any samples with too little data as having too much data, compared to the overall dataset is not a disadvantage in this context.

To generate ancestry profiles for each sample admixture analysis was performed using MANTRA[53] to assess for percentage of population group. Each sample's genotypes were compared to the 1000 Genomes Project and the proportion contributing to each super population group was calculated. These encompass Africans (**AFR**), Admixed Americans (**AMR**), East Asians (**EAS**), Europeans (**EUR**), and South Asians (**SAS**)[54]. For WGS, there were n=3401 AFR, n=35 AMR, n=161 EUR, n=10 SAS, and n=3 EAS. cBGE comprised of n=1174 AFR, n=503 AMR, n=115 EUR, n=2 SAS, and n=0 EAS. In summary, prior to quality control, our merged harmonized dataset contained a total of n=5402 samples with disease status, sequencing platform, and super population group.

Quality Control of Post-Alignment Variant-Called-Files: For quality control thresholds, outliers were defined as any sample that was + or – 4 standard deviations away from the median. We considered theta, theta exonic, ratios of heterozygotes/homozygotes, deletion/insertion, transitions/transversions, and missingness for evaluating putative outlier samples (**Table 2**). By comparing these parameters across cBGE (**Table 3**) with WGS (**Table 4**), there were minimal differences in the overall median, mean, and standard deviations. Theta, a measurement of nucleotide diversity, was the most significant in terms of difference between WGS vs cBGE with $p < 2.22x10^{-16}$ (**Figure 3**). I hypothesized most of this significance was driven by outliers as noted by the trailing sample plots in the box plot. By comparing populations and theta across cBGE vs WGS, only AFR and AMR groups showed statistically significant differences, $p < 2.22x10^{-16}$ and $p = 0.00023$, respectively (**Figure 4**). Much of the differences observed between cBGE and WGS was denoted to sequencing depth and coverage thus we are confident that these two platforms' datasets can be merged to form a cBGE-WGS combined sample set. Therefore, we calculated summary statistics and standard deviations for filtering thresholds based on the merged dataset. Considering Bystro variant metrics, relatedness, and missingness, a total of n=28 samples were removed leading to a total dataset, post-quality control, of n=5374 (**Table 5**).

Common Variant Association Analysis: Using PLINK2, common variant analysis was performed between IBD vs controls, CD vs controls, and UC vs controls. We considered common variants as those with MAF < 0.05 and > 0.01 with $p < 5x10^{-8}$. Using n=2050 controls and n=3324 IBD patients, 6,589,111 variants in the form of SNPs and INDELs were observed of which only 108409 reached genome-wide significance (**Figure 5**). In comparing CD, n=2361, vs controls, n=2050, only 42397 were observed. UC, n=963, vs controls, n=2050, showed n=128413 genome-wide significant common variants. The genomic inflation value, λ, signifies inflation of *p* values with respect to a normal distribution. Using all identified variants, λ was calculated for IBD vs controls (**Figure 6A**), CD vs controls (**Figure 6B**), and UC vs controls (**Figure 6C**). The greatest λ was in IBD and UC with λ=1.014 for both. In contrast, CD showed a λ=1.012 implying lower inflation compared to overall IBD or UC. Using all the variants observed in IBD vs controls, principal component analysis showed variance explained of PC1=37.86% and PC2=5.00% (**Figure 7**). There was minimal clustering observed by population group (**Figure 7A**) or by disease status (**Figure 7B**), suggesting more specific phenotypes are required for further sequestering.

Comparing to previously identified genetic loci attributing to IBD, CD, and UC, we observed *PTGER4*, *CARD9*, and *HLA* axis related variants. In IBD vs controls, over 38 genome-wide significant variants near *PTGER4* were observed in the top 1000 variants. For CD vs controls, *PTGER4* was detected 44 times in the top 1000 variants. Also, four SNPs were found with *CARD9* being the nearest gene comprising of OR 1.23-1.24. In UC vs controls, genes belonging to the *HLA* group were observed, *HLA-A*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DQB1-AS1*, and *HLA-DRA*-2, and *HLA-DRB1* with 2, 134, 53, 16, 44, and 52 variants respectively. The top 15 variants for IBD, CD, and UC by *p* value highlight *LIN02511* in IBD and CD, *PTGER4* for CD, and *HLA-DRA* group genes for UC (**Tables 6**, **7**, and **8**). Variant *rs1789907735* on chromosome 6 had the greatest OR = 6.0902 in IBD with no nearby gene based upon the SNP being in an intergenic region. Similarly for CD, *rs1789907735* was also the largest OR = 6.12602 variant. Ulcerative colitis patients showed variant *rs60180187* with OR=14.9477 and *HIVEP2* as the nearest gene. Across IBD, CD, and UC, the variant with the largest OR was found on chromosome 6, implying shared genetic loci potentially playing a role across disease subtypes.

Rare Variant Association Analysis: Rare variants were assigned as SNPs or INDELs with MAF < 0.01 and $p < 5x10^{-8}$. With n=2050 controls and n=3324 IBD patients, 6,589,111 variants were observed with only 78450 meeting the rare variant thresholds. In comparing CD, n=2361, vs controls, n=2050, only 78681 were observed. UC, n=963, vs controls, n=2050, showed n=185663 genome-wide significant rare variants. Using these as input, CADD scores were generated for the variants by uploading to the University of Washington server and following best practices[55]. After filtering, only n=4 variants remained in IBD, n=11 in CD, and n=2 in UC that were statistically significant with CADD > 15 (**Tables 9**, **10**, and **11**).

For IBD vs controls, *rs1165244940* showed lowest MAF of 0.00546 and CADD = 23.8 with *ATP1A4* being the nearest gene. There was no GnomAD population frequencies observed for this SNP implicating its novel nature to our cohort. SNP *rs268065* with CADD=18.5 showed allele frequencies in populations with some proportion of AFR contribution, AFR = 0.0065, AMR = 0.00033, and NFE = 2.94e-05 (**Table 9**). There is no known nearby gene located to *rs268065*. Similarly, CD vs controls also showed *rs1165244940* as the highest CADD = 23.8 and lowest MAF = 0.00546 implicating a potential common role in IBD and CD, but not UC. The OR for *rs1165244940* was higher in CD = 0.14551, compared to IBD = 0.13837. Also in CD, the

SNPs *rs6711382* and *rs1789502* with CADD scores of 18.3 and 15.6, respectively, were the only two out of eleven with AFR population specificity (**Table 10**). For UC, out of the two variants, *rs268065* with CADD = 18.5 exhibited AFR population specificity with only GnomAD AFR and GnomAD AMR showing allele frequencies of 0.0074 and 0.00013 respectively (**Table 11**).

Of the known IBD associated rare variants, *IL23R* was observed in the top 1000 variants by *p* value in both IBD and CD lists. In IBD, *IL23R* had OR =1.244, $p < 5.20E-06$, and MAF=0.279. In CD, *IL23R* showed OR = 1.27, $p < 4.01E-06$, and MAF = 0.279. Although, aggregated SKAT-O testing could not be performed for rare variant association testing, we were able to categorize, highlight, and observe population specific differences across variants with deleterious natures.

Polygenic Risk Scores: Using the known IBD associated genetic loci from Liu et al, polygenic risk scores were calculated for IBD, CD, and UC vs. controls. Due to our cohort being comprised of a large AFR population, we hypothesized population specific differences to be highlighted over disease status. Scores were generated using PLINK2 and showed statistically significant differences in IBD vs controls (**Figure 8A**), CD vs controls (**Figure 8C**), and UC vs controls (**Figure 8E**) with $p < 2.22e-16$. IBD vs controls displayed ancestry specific differences, AFR vs AMR, $p = 2.5e-13$, AFR vs EUR, $p = 2.6e13$, and AMR vs EUR, $p = 0.00031$ (**Figure 8B**). CD vs controls had fewer population based statistically significant differences, AFR vs AMR, $p = 3.3e-12$, AFR vs EUR, $p = 5.7e-12$, and AMR vs EUR, $p = 0.0017$ (**Figure 8D**). UC vs controls had the least number of statistically significant population differences with AFR vs AMR, $p = 1.6e-06$, and AMR vs EUR, $p = 0.00016$ (**Figure 8F**). Across all disease subtypes, EUR vs SAS was not found to be significant and EAS lacked sufficient samples to be tested. Overall, CD and IBD showed a similar distribution of PRS compared to UC (**Table 12**). The observations across IBD, CD, and UC may be due to sample size differences across populations. Nonetheless, our predominantly AFR cohort shows that IBD, CD, and UC are different across ancestorial groups.

Pathway Analysis: Using the top 1000 variants by *p*-value for IBD-, CD-, and UC vs controls, gene ontology pathway analysis was performed for biological processes, cellular components, and molecular functions. In IBD vs controls, a total of 26 pathways that were $p < 0.01$ after multiple-test correcting were observed, n=12 biological processes, n=8 cellular components, and n=6 molecular function (**Table 13**). Whereas in CD vs controls, only 14 pathways were detected, n=4 BP, n=7 CC, and n=3 MF (**Table 14**). In UC vs controls, 21 pathways were observed, comprised of n=17 BP, n=3 CC, and n=1 MF (**Table 15**).

IBD and CD shared multiple pathways pertaining to cell migration (GO:0030335), cell adhesion (GO:0044331, GO:0007156, GO:0016339, and GO:005925), and extracellular regulation (GO:0070062 and GO:0005576). Cell to cell processes were the common theme across all IBD pathways whereas CD showed a variety from cell migration, chromatin remodeling, to protein, ion, and enzymatic binding. The most pathways detected belonged to the lowest sample size group of UC patients. In UC, most of the pathways corresponded to immunological regulation, specifically of T-cells from signaling, binding, to regulation. IBD and UC shared only one pathway, extracellular exosome (GO:0070062) alluding to potential gut-microbiome relationships. UC and CD shared no common pathways thus exemplifying their disease subtype differences. Although these enrichments analysis only utilized the top 1000 variants from IBD, CD, and UC; these highlighted pathways showcase the heterogeneity of IBD and provide potential mechanisms of action to further interrogate.

Discussion: The novelty of cBGE has shown significant advantages and disadvantages, compared to canonically WGS, WES, and GWAS. Being the cheapest solution to obtaining genome and exome sequencing below $100, cBGE can fill the need of sequencing large datasets at reduced costs. Here, we demonstrated that cBGE provides adequate coverage to detect common variants in IBD. Also, cBGE can be combined and harmonized with WGS at 30X coverage to generate a merged dataset thus increasing power. Nonetheless, there are multiple disadvantages to cBGE. The amount of data is significantly reduced in cBGE, compared to WGS, in terms of total variants detected, types of variants, and the quality of the variants. Harmonization of WGS with cBGE required extensive reannotation of variants, re-alignment to HG38, and resulted in minimal improvements in detecting common and rare variants. We hypothesized that this novel technology, in combination with previous WGS dataset, would yield novel variants in AFR populations. The cBGE samples comprised n=1794 and WGS were n=3608. We harmonized the two platforms to increase our sample size by comparing Bystro and PLINK2 output across platforms. The original WGS was of better quality and the merged cBGE-WGS only increased the significance of previous discoveries. The imputing resources and troubleshooting for mapping intronic regions or non-protein coding regions with cBGE are overwhelming compared to developed, well-tested, and published best-practices for GWAS, WES, or WGS. We have provided the top 1000 variants with PLINK2, Bystro, CADD, and GnomAD outputs for IBD (**Supplementary Table 1**), CD (**Supplementary Table 2**), and UC (**Supplementary Table 3**). Due to ongoing data collection, sequencing, and collaborative projects, raw data files are available by request through IBDGC consortium. Future directions of cBGE would entail sequencing in other diseases, across larger cohorts, and comparing to similar coverages (2X-3X) with WES or WGS. In IBD, based on our power calculations (**Figure 2**), gathering n=5000/10000 WGS with n=5000/10000 cBGE to do a matched analysis may highlight more novel common and rare variants across populations.

Efforts to sequence and integrate more cBGE samples are ongoing by IBDGC and Broad Institute. Overall, we report similar odds ratios statistical significance across IBD, CD, and UC of variants near known loci, such as *IL23R*, *PTGER4*, *CARD9*, and *HLA*-group genes. Interestingly, we observed *rs1789007735* on chromosome 6 with OR = 6.0902 in IBD and OR = 6.1260 in CD. Using the vast sample types at the Kugathasan lab biorepository, next directions would entail testing for *rs1789007735* across IBD, CD, and controls using RT-PCR and performing statistical association testing with clinical phenotypes such as treatments, Montreal classifiers, or population group. In UC, the variant with the largest OR=14.9477, *rs60180187*, is also located on chromosome 6. The top variant across IBD subtypes to be on chromosome 6 suggests potential common genetic loci being involved. To further elucidate role of variants on chromosome 6, Hi-C could be performed looking at IBD, CD, UC, and controls to test for variants and potential interactions amongst them.

It will be very beneficial for researchers to test IBD-, CD, or UC vs controls analysis with cBGE vs WGS separately to cross-compare outputs using the same samples. Due to the vast differences in samples sizes for disease subtypes, n=2361 CD and n=963, we did not perform CD vs UC analyses. Future endeavors may be able to use a matched cohort to test for disease subtype specific common and rare variants. In summary, here we provide the largest AFR population IBD dataset to date using merged cBGE-WGS, showed the feasibility of cBGE in detecting common and rare variants, and were able to replicate our previous findings of *PTGER4*, *IL23R*, and the *HLA*-group genes.

Figures

Figure 1: Graphical abstract showing specific aim I of IBD and non-IBD participants collected blood (**A**) of combined-blended-Genome (**B**)-Exome (**C**). Specific aim II with a polygenic risk score (PRS) of IBD associated loci using a merged, cBGE+WGS dataset (**D**). PLINK2 and BYSTRO, software will be used to compare odds ratios and standard errors across known variants and novel variants identified from specific aim I. Summary statistics of variants will be compared across studies, IBD vs controls, subtypes of IBD, and populations (AFR = African, AMR=American, and EUR = European) (**E**).

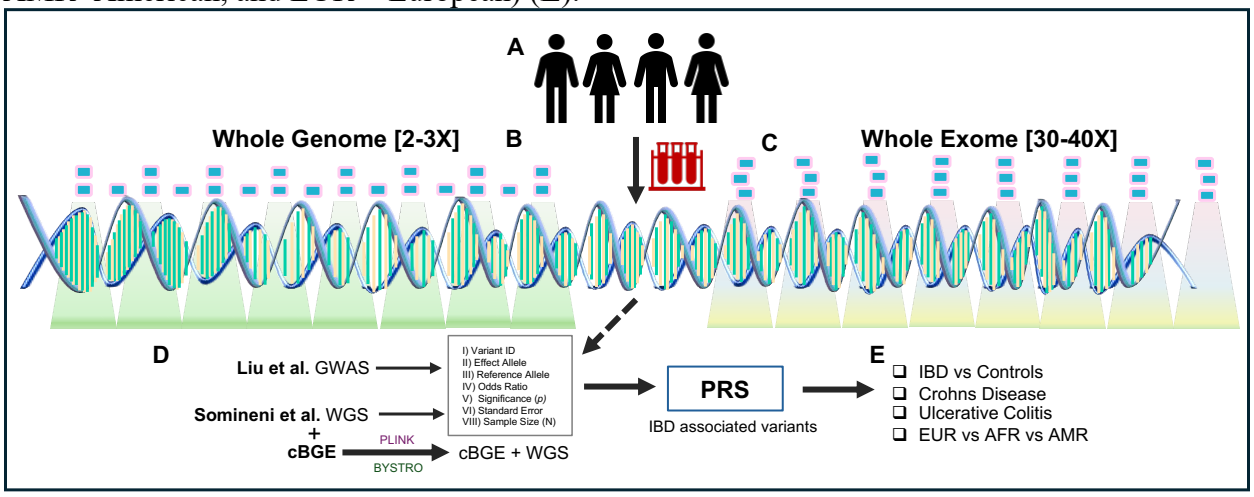


Figure 2: Power calculations with 80% power and alpha set to $5x10^{-8}$ using known IBD risk associated loci (n=271) based upon corresponding minor allele frequencies in IBD patients (x axis) and the absolute value of the log odds ratio of the variants on the y axis. Sample sizes of cases and controls are shown in 1764/1644 (green), 5000/5000 (blue), and 10000/10000 (red). A Chi-square distribution is assumed with n=1 degrees of freedom. Variants that were observed MAF >0.02 in IBD cases by Somineni et al are labeled.

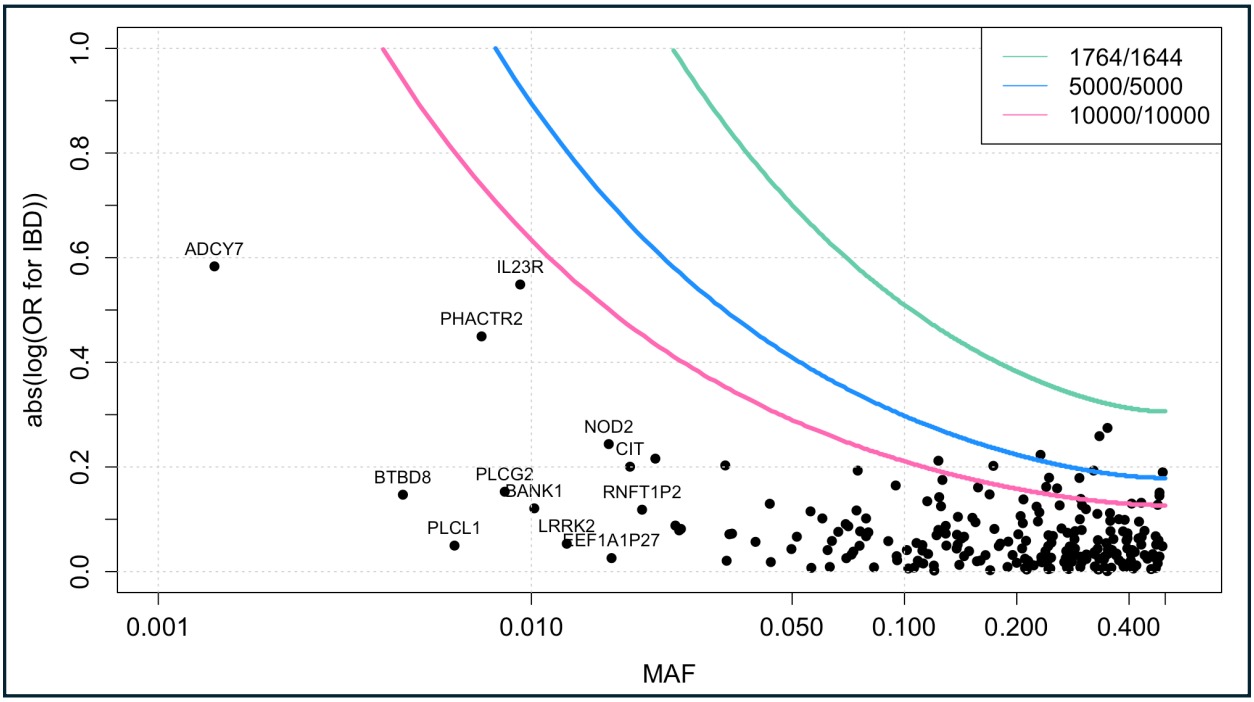


Figure 3: Boxplot showcasing theta values for prior Somineni et al (2019) Whole Genome sequencing and combined-blended Genome Exome sequencing calculated using Bystro. Variants tested totaled n=38008946 with cBGE and WGS samples, n=1794 and n=3608, respectively showing a statistically significant difference using a T-test.

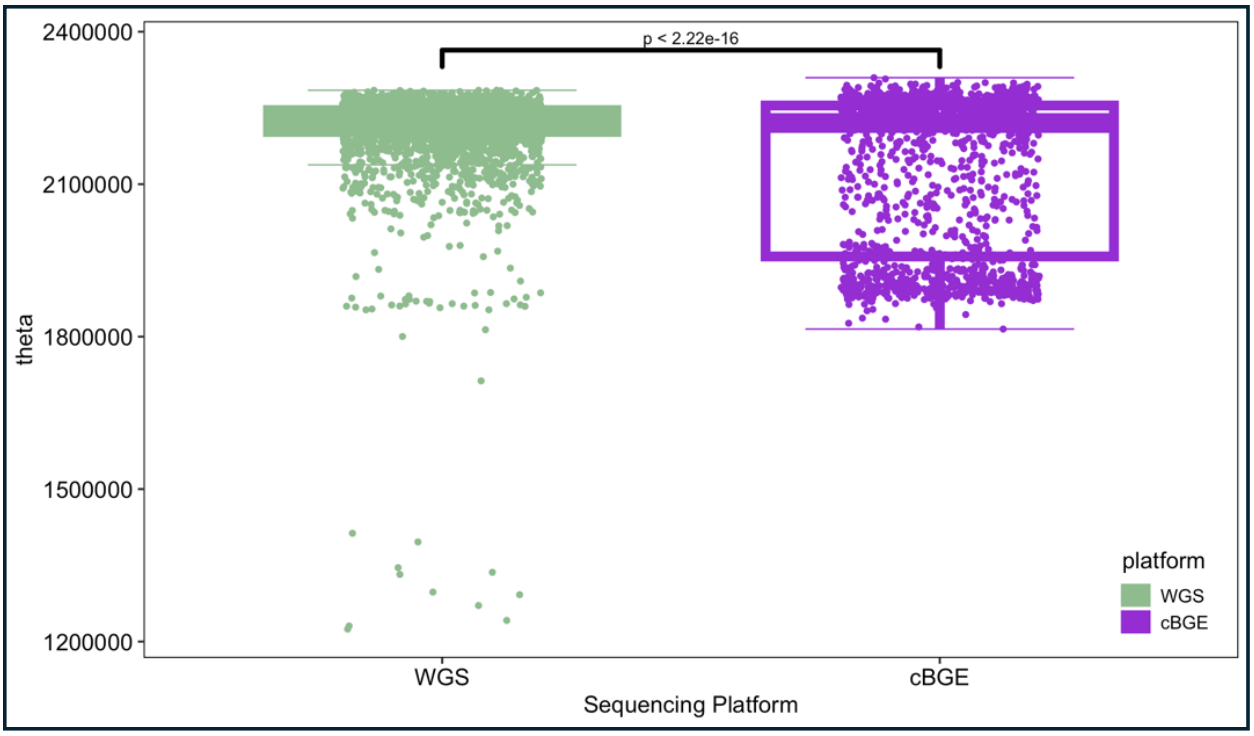


Figure 4: Boxplot showcasing theta values across populations for Whole Genome Sequencing and combined-blended Genome Exome sequencing. The most notable difference was between AFR with ($p < 2.22 \times 10\text{-}16$). Comparing EUR between platforms did not show a statistically significant difference, $p = 0.034$. Whereas AMR showed significance with $p = 0.00023$. There were insufficient samples to compare SAS or EAS populations between platforms.

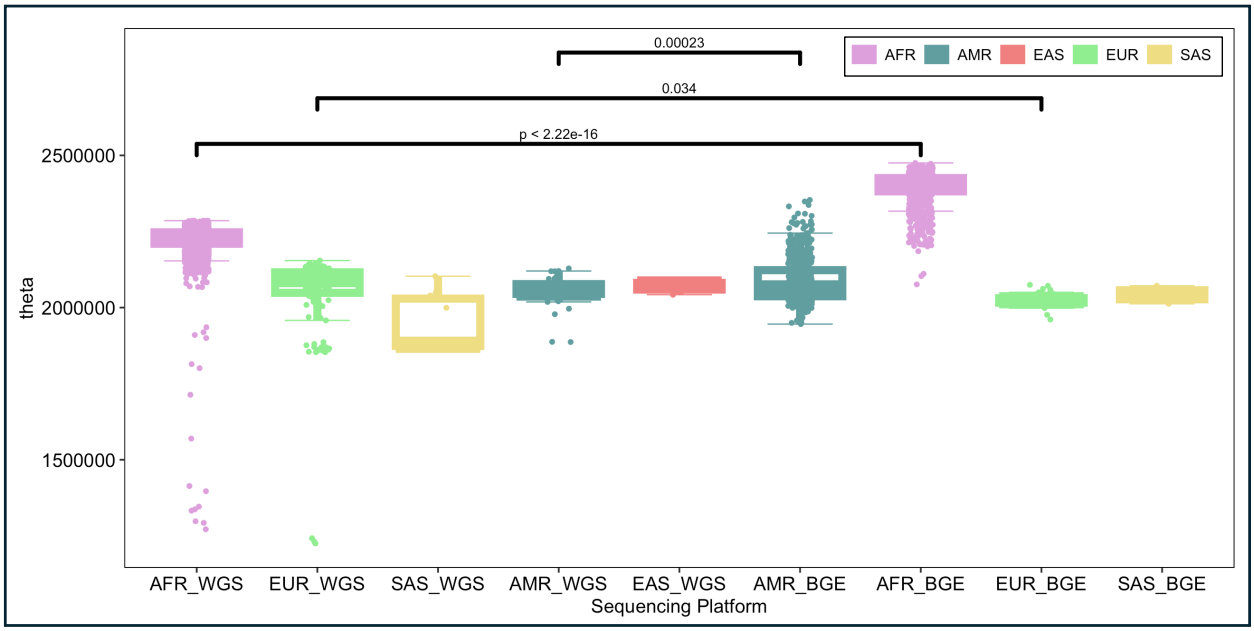


Figure 5: Manhattan plot of common variants observed across autosomes of n=2050 controls vs n=3324 IBD patients. The red line marks genome-wide significance of $p = 5x10^{-8}$ and the blue line denotes standard significance threshold of $p = 1x10^{-5}$. Over 6 million variants were observed with MAF < 0.05, of which 108409 reached genome-wide significance. The top variants by significance have been annotated with the closest nearby gene.

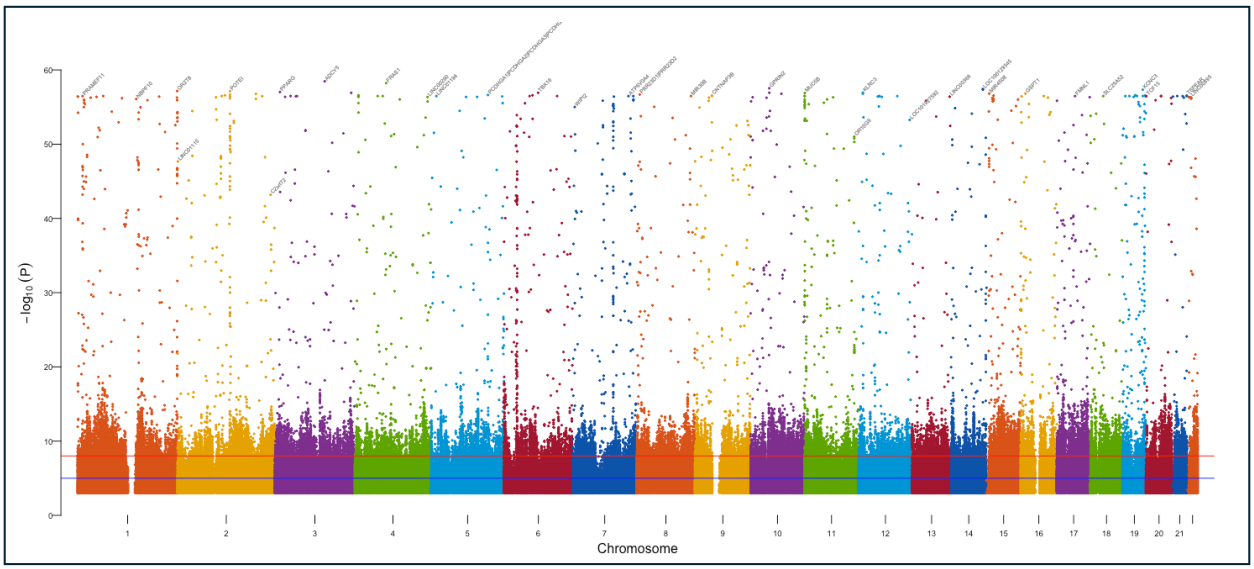


Figure 6: Quantile-Quantile plots of all observed variants in IBD vs controls (A), Crohns disease vs controls (B), and ulcerative colitis patients vs controls (C). Genomic inflation was calculated using a Chi-square distribution with one degree of freedom. The null hypothesis was that any observed *p* values are random and normally distributed.

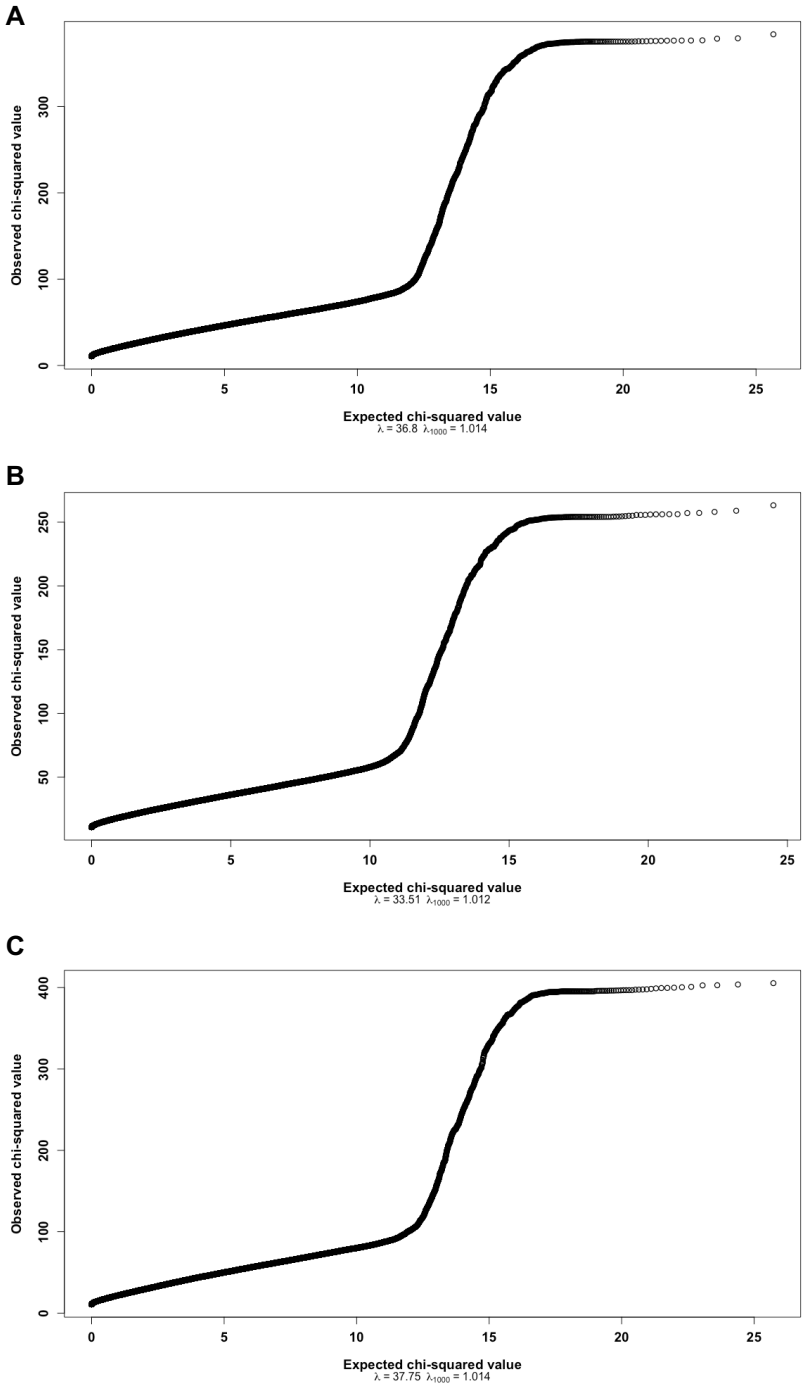


Figure 7: Principal component analysis of all IBD and control samples variants showed a variance explained of PC1=37.86% and PC2=5.00%. Samples are color-coded by ancestry (A) and by disease status (B).

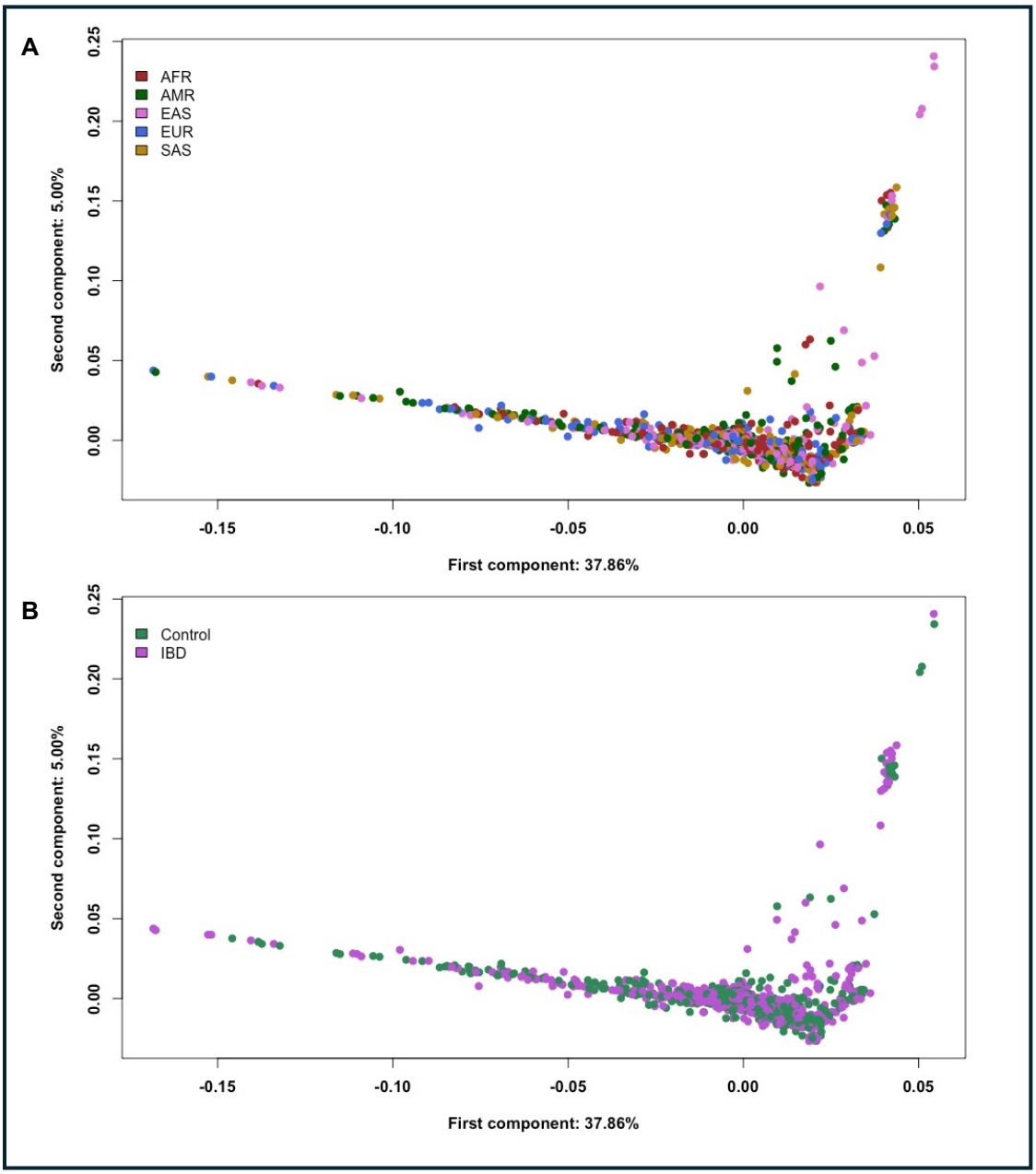
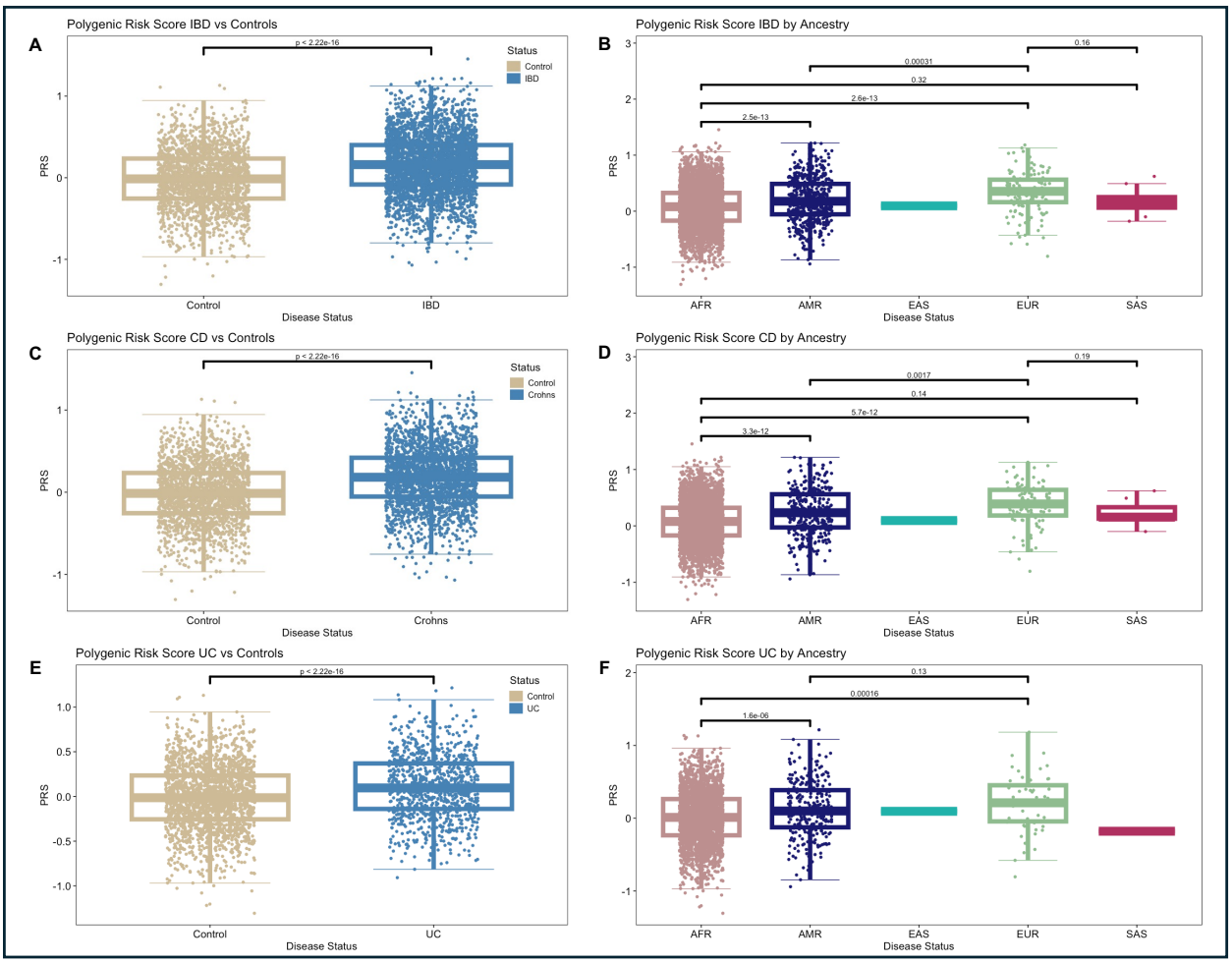


Figure 8: Polygenic risk scores of IBD vs controls (A), Crohns disease vs controls (B), and ulcerative colitis vs controls (C) using known IBD associated loci, n=278. Ancestry-based comparisons for IBD (B), Crohns disease (D), and ulcerative colitis (F) were calculated using T-test and showed AFR vs EUR to be statistically significant across all subtypes of IBD.



Tables

Table 1: Power calculation based on matched cases and controls using Somineni et al observed odds ratios and p-values for IBD associated SNPs, n=278.

Effect Size for 80% Power	Previous WGS Study		With 5000/5000 Case/Controls proposed	
Risk Allele Frequency	Odds Ratio	Variance Explained	Odds Ratio	Variance Explained
0.05	1.84	0.0030	1.45	0.0011
0.1	1.59	0.0032	1.32	0.0011
0.2	1.43	0.0034	1.24	0.0012
0.3	1.38	0.0035	1.21	0.0012
0.5	1.35	0.0037	1.19	0.0013
0.7	1.41	0.0039	1.22	0.0013
0.8	1.49	0.0041	1.26	0.0013
0.9	1.76	0.0045	1.37	0.0014

Table 2: Bystro summary statistics of merged cBGE and WGS dataset before quality control.

Metric	Mean	Median	Standard Deviation
Heterozygotes/Homozygotes Ratio	2.14	2.19	0.175
Deletion/Insertion Ratio	1.14	1.14	0.0157
Theta	2180000	2220000	115000
Theta Exonic	14400	14700	781
Transitions/Transversions Ratio	2.28	2.28	0.00314

Table 3: Bystro summary statistics of cBGE dataset before quality control filtering.

Metric	Mean	Median	Standard Deviation
Heterozygotes/Homozygotes Ratio	2.05	2.17	0.217
Deletion/Insertion Ratio	1.12	1.12	0.0129
Theta	2280000	2380000	162000
Theta Exonic	16300	16900	1170
Transitions/Transversions Ratio	2.11	2.11	0.00257

Table 4: Bystro summary statistics of WGS dataset before quality control filtering.

Metric	Mean	Median	Standard Deviation
Heterozygotes/Homozygotes Ratio	2.18	2.19	0.131
Deletion/Insertion Ratio	1.15	1.15	0.0134
Theta	2210000	2230000	76200
Theta Exonic	16100	16300	580
Transitions/Transversions Ratio	2.27	2.27	0.00288

Table 5: Total number of samples per disease status and super population group after quality control filtering.

	Inflammatory Bowel Disease		Crohns Disease		Ulcerative Colitis	
Status	Control	Case	Control	Case	Control	Case
AFR	1974	2683	1974	1974	1974	709
AMR	55	517	55	302	55	215
EAS	1	0	1	0	1	0
EUR	20	116	20	78	20	38
SAS	0	8	0	7	0	1
Total	2050	3324	2050	2361	2050	963
Total	5374		4411		3013	

Table 6: Top 15 variants by *p* value amongst Inflammatory Bowel Disease patients vs controls denoted by variant ID, chromosome location, odds ratio (OR), *p* value, and nearest gene if available.

rsID	Chromosome	OR	P	Nearest Gene
rs201803404	4	0.108067	3.49E-16	LINC02511
rs1553450141	2	0.055718	9.24E-16	NA
rs1408990865	16	0.299874	3.01E-13	
rs1305058916	12	0.13615	3.73E-13	
rs71383930	10	0.181469	1.34E-12	
chr16_34574641	16	0.332558	2.82E-12	
rs1474733259	13	0.251291	3.21E-12	
rs36086189	16	3.29121	5.14E-12	
rs920821247	19	2.25141	8.95E-12	
rs201006837	12	3.16657	3.34E-11	TCP11L2
rs145214322	3	0.2781	2.39E-10	
rs112869181	5	0.137725	4.00E-10	NIM1K
rs386673910	4	2.17416	4.42E-10	TMEM33
rs7651040	3	0.719409	5.81E-10	
rs1213444357	2	2.88504	9.03E-10	

Table 7: Top 15 variants by *p* value amongst Crohns disease patients vs controls denoted by variant ID, chromosome location, odds ratio (OR), *p* value, and nearest gene if available.

rsID	Chromosome	OR	P	Nearest Gene
rs201803404	4	0.100331	1.26E-12	LINC02511
rs1553450141	2	0.0672429	1.29E-12	RPRM
rs920821247	19	2.35259	5.47E-12	
rs1408990865	16	0.280806	1.95E-11	
rs201006837	12	3.3279	2.00E-11	TCP11L2
rs36086189	16	3.24755	4.69E-11	
rs1305058916	12	0.115536	1.35E-10	
chr16_34574641	16	0.320357	1.62E-10	
rs71383930	10	0.180228	3.97E-10	
rs1474733259	13	0.251895	6.37E-10	
rs1213444357	2	3.00803	7.58E-10	
rs746451076	7	2.50156	4.47E-09	
rs957100	5	0.756292	6.45E-09	PTGER4
rs375394643	3	2.04255	9.52E-09	LINC01266
rs7730591	5	0.760264	1.34E-08	PTGER4

Table 8: Top 15 variants by *p* value amongst Ulcerative colitis patients vs controls denoted by variant ID, chromosome location, odds ratio (OR), *p* value, and nearest gene if available.

rsID	Chromosome	OR	P	Nearest Gene
rs3891173	6	0.640315	3.81E-09	HLA-DQB1-AS1
rs4321864	6	0.614284	3.85E-09	HLA-DRA
rs7651040	3	0.603998	4.42E-09	NAALADL2
rs386673910	4	2.5435	5.28E-09	TMEM33
rs7745002	6	0.672761	8.90E-09	HLA-DQA1
rs7744613	6	0.666807	1.61E-08	HLA-DQA1
rs36086189	16	3.34286	1.94E-08	
rs35302873	5	0.548893	2.17E-08	
rs9271418	6	1.45781	2.53E-08	HLA-DQA1
rs9279966	6	0.654878	3.45E-08	
rs201621006	6	1.60343	3.81E-08	
rs9271176	6	1.45374	3.95E-08	HLA-DRB1
rs148411365	6	0.614906	4.27E-08	HLA-DQB1
rs796439239	6	1.42342	4.46E-08	HLA-DRB1
rs9270915	6	1.46374	4.66E-08	

Table 9: Statistically significant variants from IBD patients vs controls with CADD scores greater than 15. Amongst all SNPs, only 4 passed CADD threshold of which n=2 rare with MAF <0.01.

Chr	rsID	OR	P	MAF	Gene	CADD	gnomad_AFR	gnomad_AMI	gnomad_AMR	gnomad_ASJ	gnomad_EAS	gnomad_FIN	gnomad_MID	gnomad_NFE
1	rs1165244940	0.13837	6.01E-08	0.00546	ATP1A4	23.8	NA	NA	NA	NA	NA	NA	NA	NA
2	rs268065	0.823385	6.37E-06	0.553	NA	18.5	0.00651474	0	0.000326755	0	0	0	0	2.94E-05
3	rs77655087	0.77924	2.85E-05	0.139	KCNH8	17	0.520758986	0.604395986	0.741672993	0.791930974	0.277756006	0.744047999	0.734694004	0.754576027
8	rs143653444	0.337016	1.79E-05	0.0075	NA	21.6	0.155683994	0	0.073850997	0.032834101	0.263882995	0.017423199	0.071428597	0.019464901

Table 10: Statistically significant variants from Crohns disease vs controls with CADD scores greater than 15. Amongst all SNPs, only 11 passed CADD threshold of which n=2 were common with MAF <0.05 and n=2 rare with MAF <0.01.

Chr	rsID	OR	P	MAF	Gene	CADD	gnomad_AFR	gnomad_AMI	gnomad_AMR	gnomad_ASJ	gnomad_EAS	gnomad_FIN	gnomad_MID	gnomad_NFE
1	rs1165244940	0.14551	2.11E-06	0.00546	ATP1A4	23.8	NA	NA	NA	NA	NA	NA	NA	NA
2	rs77861282	3.0821	2.20E-05	0.0129	KCNH7	17.4	0.199369	0.0131579	0.0695732	0.0326645	0.109986	0.00578107	0.146851	0.0201815
2	rs6711382	0.818898	4.13E-05	0.673	NEB	18.3	0.00651474	0	0.00032676	0	0	0	0	2.94E-05
6	rs805267	1.26077	4.47E-05	0.194	LY6G5B	16.8	0.64199901	0.85197401	0.90261	0.84360403	0.55260098	0.88372701	0.86506099	0.89595997
6	rs2736182	1.26298	4.22E-05	0.189	AIF1	22.5	0.00908302	0.137363	0.0239438	0.0449309	0	0.066182	0.0442177	0.0600842
7	rs141891591	0.602859	3.90E-05	0.0335	NRCAM	15.1	0.204641	0.0164835	0.0736519	0.0518814	0.109853	0.00509831	0.14087801	0.031293
8	rs143653444	0.236153	7.81E-06	0.0075	NA	21.6	0.0539832	0	0.00450863	0	0	0	0.00340136	0.00010288
13	rs386770784	1.59471	1.42E-05	0.0437	NA	15.4	0.205732	0.300439	0.37282801	0.39381999	0.190512	0.35622501	0.32102999	0.36945799
13	rs192826605	1.59471	1.42E-05	0.0437	NA	15.7	0.203959	0.280702	0.38260099	0.427551	0.189583	0.30830899	0.40812099	0.34237099
18	rs1789504	1.26303	4.18E-05	0.231	SLC39A6	15.6	0.0540092	0	0.00450745	0	0	0	0.00340136	0.0001029
18	rs1632169	1.2851	9.92E-06	0.233	NA	15.7	0.0385431	0	0.00405865	0.00547235	0	0	0.00340136	0.0002647

Table 11: Statistically significant variants from Ulcerative colitis patients vs controls with CADD scores greater than 15. Amongst all SNPs, only 2 passed CADD threshold of which only one was rare with MAF <0.01.

Chr	rsID	OR	P	MAF	Gene	CADD	gnomad_AFR	gnomad_AMI	gnomad_AMR	gnomad_ASJ	gnomad_EAS	gnomad_FIN	gnomad_MID	gnomad_NFE
2	rs268065	0.73225	1.32E-06	0.553	LINC01793	18.5	0.0073894	0	0.00013101	0	0	0	0	0
18	rs147810566	5.73711	1.23E-06	0.00602	PIK3C3	21.5	0.52075899	0.60439599	0.74167299	0.79193097	0.27775601	0.744048	0.734694	0.75457603

Table 12: Polygenic Risk Score summary statistics for IBD patients, n=3324 and controls, n=2050. Of the IBD patients, n=2361 had Crohns disease and n=963 ulcerative colitis.

	Inflammatory Bowel Disease	Crohns Disease	Ulcerative Colitis
Minimum	-1.30640	-1.30640	-1.30640
1st Quartile	-0.15588	-0.15915	-0.21730
Median	0.09585	0.09610	0.2400
Mean	0.09461	0.09216	0.02389
3rd Quartile	0.34723	0.34115	0.27970
Maximum	1.45440	1.45440	1.21430

Table 13: Gene ontology analysis using IBD vs controls top 1000 variants as input for detecting biological processes (BP), cellular components (CC), and molecular functions (MF). Fisher's Exact test was performed and pathways with values less than 0.01 were considered. The number contributing is denoted by annotated and of which were deemed statistically qualified was termed significant. A total of 26 pathways were detected, n=12 BP, n=8 CC, and n=6 MF.

Ontology	GO:ID	Term	Annotated	Significant	Expected	Fisher Weight
BP	GO:0016477	cell migration	33	25	17.68	0.00083
BP	GO:0007165	signal transduction	95	61	50.89	0.00176
BP	GO:0045892	negative regulation of DNA-templated transcription	13	11	6.96	0.00243
BP	GO:0007156	homophilic cell adhesion via plasma membrane	5	5	2.68	0.0028
BP	GO:0009617	response to bacterium	16	13	8.57	0.00772
BP	GO:0008150	biological_process	198	131	106.07	0.00893
BP	GO:0007189	adenylate cyclase-activating G protein-c...	5	5	2.68	0.00897
BP	GO:0044331	cell-cell adhesion mediated by cadherin	5	5	2.68	0.00897
BP	GO:0034332	adherens junction organization	4	4	2.14	0.00917
BP	GO:0046330	positive regulation of JNK cascade	4	4	2.14	0.00917
BP	GO:0016339	calcium-dependent cell-cell adhesion via...	4	4	2.14	0.00917
BP	GO:0000902	cell morphogenesis	25	18	13.39	0.0095
CC	GO:0070062	extracellular exosome	28	18	15.46	0.00042
CC	GO:0098978	glutamatergic synapse	14	11	7.73	0.00048
CC	GO:0048471	perinuclear region of cytoplasm	13	10	7.18	0.00122
CC	GO:0005912	adherens junction	10	8	5.52	0.00274
CC	GO:0015629	actin cytoskeleton	11	8	6.07	0.00349
CC	GO:0016323	basolateral plasma membrane	5	5	2.76	0.00363
CC	GO:0005925	focal adhesion	7	6	3.86	0.00591
CC	GO:0009986	cell surface	26	15	14.35	0.0065
MF	GO:0008270	zinc ion binding	30	19	16.09	0.00041
MF	GO:0005178	integrin binding	6	6	3.22	0.00121
MF	GO:0008013	beta-catenin binding	8	7	4.29	0.00223
MF	GO:0051015	actin filament binding	5	5	2.68	0.00374
MF	GO:0045296	cadherin binding	10	8	5.36	0.00703
MF	GO:0005515	protein binding	185	131	99.22	< 1e-30

Table 14: Gene ontology analysis using Crohns disease patients vs controls top 1000 variants as input for detecting biological processes (BP), cellular components (CC), and molecular functions (MF). Fisher's Exact test was performed and pathways with values less than 0.01 were considered. The number contributing is denoted by annotated and of which were deemed statistically qualified was termed significant. A total of 14 pathways were detected, n=4 BP, n=7 CC, and n=3 MF.

Ontology	GO:ID	Term	Annotated	Significant	Expected	Fisher Weight
BP	GO:0030335	positive regulation of cell migration	11	7	3.55	0.0041
BP	GO:0016601	Rac protein signal transduction	3	3	0.97	0.0045
BP	GO:0051092	positive regulation of NF-kappaB transcr...	3	3	0.97	0.0045
BP	GO:0006338	chromatin remodeling	13	7	4.19	0.0078
CC	GO:0005576	extracellular region	44	19	14.64	0.00016
CC	GO:0009897	external side of plasma membrane	9	6	2.99	0.00121
CC	GO:0000139	Golgi membrane	9	6	2.99	0.00121
CC	GO:0005789	endoplasmic reticulum membrane	18	11	5.99	0.00201
CC	GO:0048471	perinuclear region of cytoplasm	7	5	2.33	0.0035
CC	GO:0005788	endoplasmic reticulum lumen	3	3	1	0.00497
CC	GO:0005783	endoplasmic reticulum	29	18	9.65	0.00557
MF	GO:0042802	identical protein binding	28	13	9.08	0.0017
MF	GO:0005509	calcium ion binding	17	8	5.51	0.0027
MF	GO:0031267	small GTPase binding	5	4	1.62	0.003

Table 15: Gene ontology analysis using ulcerative colitis patients vs controls top 1000 variants as input for detecting biological processes (BP), cellular components (CC), and molecular functions (MF). Fisher's Exact test was performed and pathways with values less than 0.01 were considered. The number contributing is denoted by annotated and of which were deemed statistically qualified was termed significant. A total of 21 pathways were detected, n=17 BP, n=3 CC, and n=1 MF.

Ontology	GO:ID	Term	Annotated	Significant	Expected	Fisher Weight
BP	GO:0001916	positive regulation of T cell mediated c...	3	2	0.03	0.00017
BP	GO:0006959	humoral immune response	4	2	0.04	0.00035
BP	GO:0050852	T cell receptor signaling pathway	4	2	0.04	0.00035
BP	GO:0050778	positive regulation of immune response	9	4	0.1	0.00224
BP	GO:0051262	protein tetramerization	1	1	0.01	0.00881
BP	GO:0042088	T-helper 1 type immune response	1	1	0.01	0.00881
BP	GO:0046598	positive regulation of viral entry into ...	1	1	0.01	0.00881
BP	GO:0035774	positive regulation of insulin secretion...	1	1	0.01	0.00881
BP	GO:0002862	negative regulation of inflammatory resp...	1	1	0.01	0.00881
BP	GO:0042130	negative regulation of T cell proliferat...	1	1	0.01	0.00881
BP	GO:0032653	regulation of interleukin-10 production	1	1	0.01	0.00881
BP	GO:0002842	positive regulation of T cell mediated i...	1	1	0.01	0.00881
BP	GO:0032689	negative regulation of type II interfero...	1	1	0.01	0.00881
BP	GO:0032673	regulation of interleukin-4 production	1	1	0.01	0.00881
BP	GO:0070374	positive regulation of ERK1 and ERK2 cas...	1	1	0.01	0.00881
BP	GO:0033674	positive regulation of kinase activity	1	1	0.01	0.00881
BP	GO:0045657	positive regulation of monocyte differen...	1	1	0.01	0.00881
CC	GO:0001772	immunological synapse	3	2	0.03	0.00017
CC	GO:0070062	extracellular exosome	10	2	0.11	0.00246
CC	GO:0005882	intermediate filament	1	1	0.01	0.00862
MF	GO:0042608	T cell receptor binding	3	2	0.03	0.00017

Bibliography

1. Saez, A., Herrero-Fernandez, B., Gomez-Bris, R., Sanchez-Martinez, H. & Gonzalez-Granado, J.M. Pathophysiology of Inflammatory Bowel Disease: Innate Immune System. *Int J Mol Sci* **24** (2023).
2. Burisch, J. et al. East-West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort. *Gut* **63**, 588-597 (2014).
3. Somineni, H.K. et al. Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease. *Am J Hum Genet* **108**, 431-445 (2021).
4. Camilleri, M. et al. Role for diet in normal gut barrier function: developing guidance within the framework of food-labeling regulations. *Am J Physiol Gastrointest Liver Physiol* **317**, G17-G39 (2019).
5. Dunleavy, K.A., Raffals, L.E. & Camilleri, M. Intestinal Barrier Dysfunction in Inflammatory Bowel Disease: Underpinning Pathogenesis and Therapeutics. *Dig Dis Sci* **68**, 4306-4320 (2023).
6. Camilleri, M. et al. Understanding measurements of intestinal permeability in healthy humans with urine lactulose and mannitol excretion. *Neurogastroenterol Motil* **22**, e15-26 (2010).
7. Hong, S.M. & Baek, D.H. Diagnostic Procedures for Inflammatory Bowel Disease: Laboratory, Endoscopy, Pathology, Imaging, and Beyond. *Diagnostics (Basel)* **14** (2024).
8. Wang, Z.Z., Shi, K. & Peng, J. Serologic testing of a panel of five antibodies in inflammatory bowel diseases: Diagnostic value and correlation with disease phenotype. *Biomed Rep* **6**, 401-410 (2017).
9. Lee, W.I., Subramaniam, K., Hawkins, C.A. & Randall, K.L. The significance of ANCA positivity in patients with inflammatory bowel disease. *Pathology* **51**, 634-639 (2019).
10. Satsangi, J., Silverberg, M.S., Vermeire, S. & Colombel, J.F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. *Gut* **55**, 749-753 (2006).
11. Shrestha, S. et al. The use of ICD codes to identify IBD subtypes and phenotypes of the Montreal classification in the Swedish National Patient Register. *Scand J Gastroenterol* **55**, 430-435 (2020).
12. Dotson, J.L. et al. Feasibility and validity of the pediatric ulcerative colitis activity index in routine clinical practice. *J Pediatr Gastroenterol Nutr* **60**, 200-204 (2015).
13. Lee, M. & Chang, E.B. Inflammatory Bowel Diseases (IBD) and the Microbiome-Searching the Crime Scene for Clues. *Gastroenterology* **160**, 524-537 (2021).
14. Gradel, K.O. et al. Increased short- and long-term risk of inflammatory bowel disease after salmonella or campylobacter gastroenteritis. *Gastroenterology* **137**, 495-501 (2009).
15. Turner, J.R. Intestinal mucosal barrier function in health and disease. *Nat Rev Immunol* **9**, 799-809 (2009).
16. Seyedian, S.S., Nokhostin, F. & Malamir, M.D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J Med Life* **12**, 113-122 (2019).
17. Murray, A., Nguyen, T.M., Parker, C.E., Feagan, B.G. & MacDonald, J.K. Oral 5-aminosalicylic acid for induction of remission in ulcerative colitis. *Cochrane Database Syst Rev* **8**, CD000543 (2020).

18. Lichtenstein, G.R., Sbreu, M.T., Cohen, R. & Tremaine, W. [American Gastroenterological Association Institute technical review on corticosteroids, immunomodulators, and infliximab in inflammatory bowel disease]. *Rev Gastroenterol Mex* **71**, 351-401 (2006).
19. Stallmach, A. et al. Treatment Strategies in Inflammatory Bowel Diseases. *Dtsch Arztebl Int* **120**, 768-778 (2023).
20. Cai, Z., Wang, S. & Li, J. Treatment of Inflammatory Bowel Disease: A Comprehensive Review. *Front Med (Lausanne)* **8**, 765474 (2021).
21. Nakase, H. et al. Significance of measurement of serum trough level and anti-drug antibody of adalimumab as personalised pharmacokinetics in patients with Crohn's disease: a subanalysis of the DIAMOND trial. *Aliment Pharmacol Ther* **46**, 873-882 (2017).
22. Orholm, M., Binder, V., Sorensen, T.I., Rasmussen, L.P. & Kyvik, K.O. Concordance of inflammatory bowel disease among Danish twins. Results of a nationwide study. *Scand J Gastroenterol* **35**, 1075-1081 (2000).
23. Thompson, N.P., Driscoll, R., Pounder, R.E. & Wakefield, A.J. Genetics versus environment in inflammatory bowel disease: results of a British twin study. *BMJ* **312**, 95-96 (1996).
24. Torres, J. et al. Risk Factors for Developing Inflammatory Bowel Disease Within and Across Families with a Family History of IBD. *J Crohns Colitis* **17**, 30-36 (2023).
25. Qin, Y. et al. Author Correction: Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat Genet* **56**, 554 (2024).
26. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227 (2012).
27. Singh, V. et al. Interplay between enterobactin, myeloperoxidase and lipocalin 2 regulates E. coli survival in the inflamed gut. *Nat Commun* **6**, 7113 (2015).
28. Jung, S. et al. Identification of Three Novel Susceptibility Loci for Inflammatory Bowel Disease in Koreans in an Extended Genome-Wide Association Study. *J Crohns Colitis* **15**, 1898-1907 (2021).
29. Turpin, W., Goethel, A., Bedrani, L. & Croitoru Mdcm, K. Determinants of IBD Heritability: Genes, Bugs, and More. *Inflamm Bowel Dis* **24**, 1133-1148 (2018).
30. Bonen, D.K. et al. Crohn's disease-associated NOD2 variants share a signaling defect in response to lipopolysaccharide and peptidoglycan. *Gastroenterology* **124**, 140-146 (2003).
31. Inoue, N. et al. Lack of common NOD2 variants in Japanese patients with Crohn's disease. *Gastroenterology* **123**, 86-91 (2002).
32. Khalessi, A. et al. Differential Manifestations of Inflammatory Bowel Disease Based on Race and Immigration Status. *Gastro Hep Adv* **3**, 326-332 (2024).
33. Adeyanju, O. et al. Common NOD2 risk variants in African Americans with Crohn's disease are due exclusively to recent Caucasian admixture. *Inflamm Bowel Dis* **18**, 2357-2359 (2012).
34. Liu, J.Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).

35. Lee, S., Teslovich, T.M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* **93**, 42-53 (2013).
36. Liu, Z. et al. Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nat Genet* **55**, 796-806 (2023).
37. de Lange, K.M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**, 256-261 (2017).
38. Yap, C.F. & Morris, A.P. Methods for multiancestry genome-wide association study meta-analysis. *Ann Hum Genet* (2024).
39. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173-178 (2017).
40. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121-132 (2014).
41. DeFelice, M. et al. Blended Genome Exome (BGE) as a Cost Efficient Alternative to Deep Whole Genomes or Arrays. *bioRxiv* (2024).
42. Gibson, G. et al. Eleven Grand Challenges for Inflammatory Bowel Disease Genetics and Genomics. *Inflamm Bowel Dis* **31**, 272-284 (2025).
43. Dassopoulos, T. et al. Assessment of reliability and validity of IBD phenotyping within the National Institutes of Diabetes and Digestive and Kidney Diseases (NIDDK) IBD Genetics Consortium (IBDGC). *Inflamm Bowel Dis* **13**, 975-983 (2007).
44. Cheng, C. et al. A General Primer for Data Harmonization. *Sci Data* **11**, 152 (2024).
45. Kotlar, A.V., Trevino, C.E., Zwick, M.E., Cutler, D.J. & Wingo, T.S. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. *Genome Biol* **19**, 14 (2018).
46. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
47. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
48. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10 (2004).
49. Price, A.L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909 (2006).
50. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499-511 (2010).
51. Karczewski, K.J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
52. Thomas, P.D. et al. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci* **31**, 8-22 (2022).
53. Cordero, R.Y. et al. Trans-ancestry, Bayesian meta-analysis discovers 20 novel risk loci for inflammatory bowel disease in an African American, East Asian and European cohort. *Hum Mol Genet* **32**, 873-882 (2023).
54. Gaspar, H.A. & Breen, G. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC Bioinformatics* **20**, 116 (2019).
55. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).