Mendelian Disease Associations Reveal Novel Insights into Inflammatory Bowel Disease

Lichy Han, BS,[*] Mateusz Maciejewski, PhD,[†] Christoph Brockel, PhD,[‡] Lovisa Afzelius, PhD, MBA,[†] and Russ B. Altman, MD, PhD[§,¶,]*

Background: Monogenic diseases have been shown to contribute to complex disease risk and may hold new insights into the underlying biological mechanism of Inflammatory Bowel Disease (IBD).

Methods: We analyzed Mendelian disease associations with IBD using over 55 million patients from the Optum's deidentified electronic health records dataset database. Using the significant Mendelian diseases, we performed pathway enrichment analysis and constructed a model using gene expression datasets to differentiate Crohn's disease (CD), ulcerative colitis (UC), and healthy patient samples.

Results: We found 50 Mendelian diseases were significantly associated with IBD, with 40 being significantly associated with both CD and UC. Our results for CD replicated those from previous studies. Pathways that were enriched consisted of mainly immune and metabolic processes with a focus on tolerance and oxidative stress. Our 3-way classifier for UC, CD, and healthy samples yielded an accuracy of 72%.

Conclusions: Mendelian diseases that are significantly associated with IBD may reveal novel insights into the genetic architecture of IBD.

Key Words: Mendelian diseases, claims data, comorbidities

INTRODUCTION

Inflammatory bowel disease (IBD) is a complex, heterogeneous disease that affects over 1 in 300 people in the United States.[1] IBD consists of 2 main diseases, ulcerative colitis (UC) and Crohn's disease (CD), which are very similar and both result in gastrointestinal inflammation. Though there have been numerous attempts to discover implicated genes and identify causal variants,[2–4] much of the genetic architecture of IBD remains unknown.[5]

The advancement of genomics has led to the discovery of the underlying genetic cause for many diseases, and particularly for Mendelian diseases, which are typically monogenic, highly penetrant diseases that are caused by a variant at a single locus. However, attempts to find highly penetrant variants that contribute to the development complex, polygenic diseases, have been limited and can suffer from low reproducibility.[6] At the same time, it has been shown that many Mendelian diseases predispose patients to nonMendelian, complex diseases, such as Friedreich's ataxia with type 2 diabetes.[7] These comorbidities have driven the idea that a combination of mutations in Mendelian genes may contribute to complex disease risk and may be a useful avenue for discovering implicated genes in complex diseases.

In a recent review, Uhlig arbitrarily selected 40 monogenic diseases that are associated with IBD-like gastrointestinal inflammation.[8] Uhlig noted that in children with early-onset IBD, a proportion also suffer from a Mendelian disease, some of which have been studied to gain further insight into IBD pathogenesis. To test the hypothesis that Mendelian genes contribute to complex disease risk at a large scale, Blair, et al analyzed millions of patient records from claims data to discover significant associations between Mendelian and complex diseases.[9] First, they showed that hits from genome- wide association studies (GWAS) for complex diseases are significantly enriched for Mendelian loci, a further indication that genes and pathways implicated in Mendelian disorders may contribute to complex disease risk. Then, they analyzed 95 Mendelian and 65 complex diseases by constructing pairwise contingency tables and mixed-effects models to assess the relative risk. They showed that there is significant comorbidity between the 2 sets of diseases, and they modeled the contributory risk of the Mendelian diseases using additive and combinatorial models.

Recently, Melamed et al replicated this work in cancer, as there also are notable relationships between Mendelian diseases and cancer risk, such as Li-Fraumeni syndrome leading to multiple cancers due to mutations in *TP53*.[10] Melamed

Received for publications August 3, 2017; Editorial Decision October 26, 2017.

*Biomedical Informatics Training Program, Stanford University, Stanford, CA 94305; †Inflammation & Immunology, Pfizer Inc., 1 Portland Street, Cambridge, MA 02139; ‡Hill's Pet Nutrition, 1035 NE 43rd St, Topeka KS 66617, §Department of Genetics and ¶Department of Bioengineering, Stanford University, Stanford, CA 94305.

Conflicts of Interest:The authors have no conflicts of interest to declare.

Supported by: This work is funded by the National Institutes of Health R01 GM102365 and T32 GM007365, F30 AI124553, and by IC2014-1387 from Pfizer, Inc.

Address correspondence to: Shriram Room 209, MC: 4245, 443 Via Ortega Drive, Stanford, CA 94305-4145. Email: russ.altman@stanford.edu.

doi: 10.1093/ibd/izx087
Published online 16 February 2018

et al then used the Mendelian disease comorbidities to identify enriched genes and pathways shared by Mendelian diseases and certain cancers. For example, they found a significant association between Diamond-Blackfan anemia and multiple brain cancers, including glioblastoma. Genes implicated in Diamond-Blackfan anemia include *RPL5*, *RPL11*, and *RPS7*, all of which have a role in repressing *MDM2*. Amplification of *MDM2* is present in 15% of the glioblastoma cases in The Cancer Genome Atlas and, thus, loss of repressor genes due to Diamond-Blackfan anemia could explain the significant association found between these 2 diseases.

In this work, we apply the abovementioned approaches to CD and UC to investigate the potential contribution of Mendelian comorbidities and identify candidate genes that may contribute to IBD risk. CD was included by Blair et al, and our work validates these previously published results. In addition, we investigate the ability of the genes and pathways associated with the Mendelian comorbidities to discover underlying differences between 2 similar diseases with unknown etiologies. We use Optum's longitudinal clinical repository (Optum's deidentified Electronic Health Record dataset 2007–2016), which contains over 55 million patients, to apply the work from Blair et al to IBD. We then demonstrate further utility in Mendelian-complex disease associations by using the Mendelian gene associations unique to CD and UC to build a 3-way classifier to differentiate CD, UC, and healthy tissue samples using transcriptomic data. We then analyze these genes to gain insight into the mechanisms driving CD and UC.

MATERIALS AND METHODS

Clinical Data

We extracted patient information from Optum's database, which includes patient data from January 1, 2007 to March 30, 2017 (Optum deidentified Electronic Health Record dataset 2007–2016, https://www.optum.com/). Optum's longitudinal clinical repository is derived from dozens of healthcare provider organizations in the United States that include more than 650 Hospitals and 6600 Clinics; treating more than 69 million patients receiving care in the United States. The data is certified as deidentified by an independent statistical expert following HIPAA statistical deidentification rules and managed according to Optum's customer data use agreements.[1,2] Clinical, claims, and other medical administrative data are obtained from both Inpatient and Ambulatory electronic health records (EHRs), practice management systems, and numerous other internal systems; and the data are processed, normalized, and standardized across the continuum of care from both acute inpatient stays and outpatient visits. Optum's data elements include demographics, medications prescribed and administered, immunizations, allergies, lab results (including microbiology), vital signs and other observable measurements, clinical and inpatient stay administrative data, and coded diagnoses and procedures. In addition, Optum Analytics uses natural language processing (NLP) computing technology to extract critical facts from physician notes into usable datasets. The NLP data provides detailed information regarding signs and symptoms, family history, disease-related scores (ie, RAPID3 for RA, or CHADS2 for stroke risk), genetic testing, medication changes, and physician rationale behind prescribing decisions that might never be recorded in the EHR. Though this is a different, smaller claims database than was used by Blair et al, patient overlap may exist between these different US databases. Database access was provided by Pfizer, Inc. All data extraction and analyses were performed using R 3.2.1 (R Core Development Team, Vienna, Austria).

[1]45 CFR 164.514(b)(1).

[2]Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Information Insurance Portability and Accountability Act (HIPAA) Privacy Rule (Dated as September 4, 2012, as first released on November 26, 2012).

Complex-Mendelian Contingency Tables

In concordance with Blair et al,[9] we constructed contingency tables for each complex-Mendelian disease pair. Incidence counts for each complex-Mendelian disease pair were extracted from Optum's deidentified EHR dataset. Specifically, for a given complex-Mendelian disease pair, we extracted the number of patients with both diseases, number of patients with just 1 of the 2 diseases, and number of patients with neither disease. We then calculated the relative risk and applied the Fisher's exact test to each contingency table to assess significance. We accounted for multiple hypothesis testing by using the Bonferroni correction. Mendelian diseases were considered significantly associated with CD or UC if the Bonferroni corrected *P*-value was less than 0.05.

Patients were considered to have a given disease if they received an associated ICD-9 diagnosis code. For the Mendelian diseases, we used the ICD-9 codes curated by Blair et al,[9] which are located in Table S3 of their work. These ICD-9 codes correspond to 95 Mendelian diseases and disease groups, which represent 213 diseases overall. As detailed in the experimental procedures section of Blair et al, the grouping of the 213 diseases into 95 groups was driven by ICD-9 code taxonomy. For CD, which was included in the original work, we used the same ICD-9 codes: 555, 555.0, 555.1, 555.2, and 555.9. For UC, we used the codes 556, 556.0, 556.2, 556.3, 556.4, 556.5, 556.6, 556.8, and 556.9. We did not use code 556.1 (ulcerative ileocolitis), as UC should be restricted to only the colon.

Mixed-effects Poisson Models

To adjust for confounding variables, Blair et al built mixed-effects Poisson models, using the *lme4* package in R.[11] These models are fully detailed in the extended methods in Blair et al. Briefly, they modeled the patient counts as follows:

$$P\left(y_{i,j,k,l}|\lambda_{i,j,k,l}\right)=\frac{\lambda_{i,j,k,l}^{y_{i,j,k,l}}\exp\left[-\lambda_{i,j,k,l}\right]}{y_{i,j,k,l}!}$$

where y is the total number of patients with the complex disease in the subpopulation denoted by the indices *i, j, k*, and *l*. In this model, *i* is the county; *j* is the state; *k* is the patient's age, which is binned into decades; and *l* is a binary variable denoting the presence or absence of the Mendelian disease. For modeling λ, the fixed effects were Mendelian disease status, gender, average per capita income, percent ethnicity, percent insured, percent poor, and percent urban. The random effects were age (binned by decade) and county.

In replicating these models, we made several adjustments to account for differences in the database used in our analysis. Optum's database contains age, gender, Mendelian disease status, average household income, and average percent education, but does not have data pertaining to the county, state, percent ethnicity, percent insured, percent poor, or percent urban. Though county and state level location information were not provided, the average household income and average percent education is based on the 3-digit-zip code the patient resides in. Therefore, we used the 692 unique combinations of average household income and percent education as a proxy for location to group patients in our models. Our modified models for patient counts were thus as follows:

$$P\left(y_{i,k,l}|\lambda_{i,k,l}\right)=\frac{\lambda_{i,k,l}^{y_{i,k,l}}\exp\left[-\lambda_{i,k,l}\right]}{y_{i,k,l}!}$$

where i is the proxy 3-digit-zip code, and all other variables are the same as defined in Blair et al. Our fixed effects were Mendelian disease status, gender, average household income, and percent education, and our random effects were age and our proxy 3-digit-zip code.

Comparison and Validation of Blair et al

We compared our CD results to those presented in Blair et al. Specifically, Table S4 in Blair et al contains the relative risk values for the 44 diseases that were significantly associated with CD in the original work. Using the relative risk scores from their linear model, we performed matched *t* tests with our contingency table relative risk values and those from our own models.

Gene Ontology Analysis of Significantly Associated Mendelian Diseases

We extracted the genes associated with each of the Mendelian diseases from Table S3 of Blair et al. We then updated the gene lists via manual curation from the Online Mendelian Inheritance in Man (OMIM) database. Using these genes, we performed a gene ontology[12] (GO) enrichment analysis. All genes that were associated with a Mendelian disease were annotated with GO terms using the *biomaRt* package.[13] As the enrichment analysis would be dominated by the Mendelian diseases with the most genes, we randomly selected 1 gene to represent each Mendelian disease and assembled 100 such gene sets for CD and for UC using the significantly associated Mendelian diseases for each IBD subtype. By compiling these randomized sets, we mitigate the overrepresentation of clusters of similar genes associated with any 1 disease. We then used the *topGO* package[14] to test for enrichment of biological processes for each of these 100 randomizations and assigned a rank to each biological process per randomization. We take the average rank across all 100 runs to assess the top biological processes associated with CD and UC.

Relating Mendelian genes to Known IBD Genes

We evaluated the candidate genes associated with Mendelian diseases that were significantly associated with CD or UC ("Mendelian IBD genes") by comparing them to IBD genes from the GWAS published by Liu et al ("known IBD genes").[4] We mapped the variants from Liu et al to genes using the Ensembl Variant Effect Predictor.[15] We then used the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) protein-protein interaction database[16] to quantify the relationship between our Mendelian IBD genes and known IBD genes. We mapped genes to STRING gene identifiers and then examined the overlap of the Mendelian IBD genes with the known IBD genes. We then found the length of the shortest path between each Mendelian IBD gene and each known IBD gene. For comparison, we did the same analysis for all genes in STRING.

IBD Models Based on Significant Mendelian Genes

We identified the genes associated with the Mendelian diseases that were uniquely significantly associated with either CD or UC. We then used these candidate genes to distinguish among CD, UC, and healthy patients with gene expression data. We curated 4 publicly available studies: GSE16879[17] (24 UC, 19 CD, and 6 healthy), GSE10616[18] (10 UC, 14 CD, and 11 healthy), GSE9686[19] (5 UC, 11 CD, and 8 healthy), and GSE36807[20] (15 UC, 13 CD, and 7 healthy). For consistency, we used only the baseline CD, UC, and healthy colon biopsy samples from all 4 studies.

All data were processed using robust multiarray average[21] and then ComBat[22] to correct for batch effects, where each study is considered 1 batch. The data were split into a training set consisting of 70%, or 100 samples, and a 30% held- out test set with the remaining 43 samples. Using the *nnet* package,[23] we trained a multinomial logistic regression model on the training set and assessed the accuracy of our classification model on held-out test set.

To assess the significance of the genes used in our model, we constructed additional multinomial logistic regression

models using 1000 randomly selected sets of genes. We then compared the accuracy of our model versus these additional 1000 models.

RESULTS

IBD Patients in Optum

We extracted 55,080,118 patients with at least one ICD-9 code from Optum. Of these patients, 177,039 had a UC diagnosis code, and 183,855 had a CD diagnosis code. Of all IBD patients, 81% had at least 2 diagnosis codes, and 9,336 CD patients and 9,264 UC patients have been diagnosed with a Mendelian disorder. The number of IBD patients born after 2000 is enriched for having a Mendelian disorder (2.3% vs 1.3%, $P < 0.0001$), which likely reflects that patients with Mendelian diseases tend to present with IBD at a younger age. Demographic statistics for the UC and CD patients are presented in Table 1. None of the demographic variables were significantly different when comparing CD and UC using the chi-squared test.

IBD Mendelian Signature

Heatmaps showing the relative risk of each Mendelian-IBD pair are shown in Figure 1. Figure 1A depicts the relative risks calculated from the contingency tables and Figure 1B shows the relative risks calculated using the Poisson models. Overall, the relative risk values from the Poisson models were similar to the contingency table relative risk values. However, there were 3 diseases in which the relative risk values increased substantially when using the Poisson models. We note that these diseases are more predominant in a specific portion of the population. For example, sickle cell anemia is more common in Africans and African Americans, and hemophilia and congenital Hirschprung's disease in males, starting at infancy. We believe these larger differences arise when examining subpopulations in our database with fewer IBD patients, as our database contains IBD patients that are mostly Caucasian, more likely to be female, and currently in their 20s or 30s.

TABLE 1: Demographic Statistics for IBD Patients in Optum

	CD	UC	*P*-Value
Number of Patients	183,855	177,039	—
Gender, Male, N(%)	77,992(42.4)	79,151(44.7)	0.20
Birth Year, N(%)			0.23
2011–2016	237(0.13)	222(0.13)	
2001–2010	2835(1.54)	1433(0.81)	
1991–2000	15,137(8.23)	9319(5.26)	
1981–1990	28,778(15.65)	21,808(12.32)	
1971–1980	30,164(16.41)	25,599(14.46)	
1961–1970	32,378(17.61)	30,994(17.51)	
1951–1960	32,206(17.52)	34,888(19.71)	
1941–1950	23,615(12.84)	27,347(15.45)	
1931–1940	12,218(6.65)	16,369(9.25)	
1930 or earlier	6225(3.39)	9007(5.09)	
Race, N(%)			0.21
African American	12,266(6.67)	9510(5.37)	
Asian	1840(1.00)	2482(1.40)	
Caucasian	149,394(81.26)	144,533(81.64)	
Other/Unknown	20,355(11.07)	20,514(11.59)	

Comparison to Previously Published Results

We compared the relative risks from our data and models to those of the 44 Mendelian diseases presented in the work by Blair et al (Fig. 2, S1) for CD. Overall, our relative risk values are not significantly different from the linear model relative risk values in Blair et al, with a *P*-value of 0.08 when comparing against our Poisson model relative risks and 0.31 when comparing against the contingency table relative risk values. The diseases that showed the largest difference between our work and Blair's work were hemophilia and congenital Hirschprung's disease. These differences mirror the main differences between using the contingency table analysis and the linear model in our work. Because of the lack of regional race and additional demographic variables, we proceeded with downstream analysis using the results from the contingency table due to their high congruence with previously published results.[9]

Mendelian Diseases are Associated with IBD Subtypes

The Mendelian diseases that were significantly associated with CD or UC are shown in Table 2. There were 43 diseases significantly associated with CD, and 47 with UC, with 40 in common. These significant Mendelian diseases correspond to 527 candidate genes in total (Table S1), with 490 associated with CD and 503 with UC. The 2 diseases had 466 genes in common, and there were 24 genes uniquely associated with CD and 37 with UC. These 61 genes uniquely associated with CD or UC were selected for downstream classification with gene expression data.

From Figure 1A, the cluster on the left that has the highest relative risk associations with CD and UC consists of 8 metabolic, digestive, and immune-related diseases: Diamond-Blackfan anemia, Bartter's syndrome, congenital Hirschprung's disease, disorders of copper metabolism, disorders of phosphorus metabolism, autoimmune lymphoproliferative syndrome, genetic anomalies of leukocytes, and severe combined immunodeficiency. These diseases correspond to 62 genes, many of which are known to be involved in IBD-related pathways.

The Mendelian diseases that are significantly associated with UC and not CD are Bartter's syndrome, disorders of

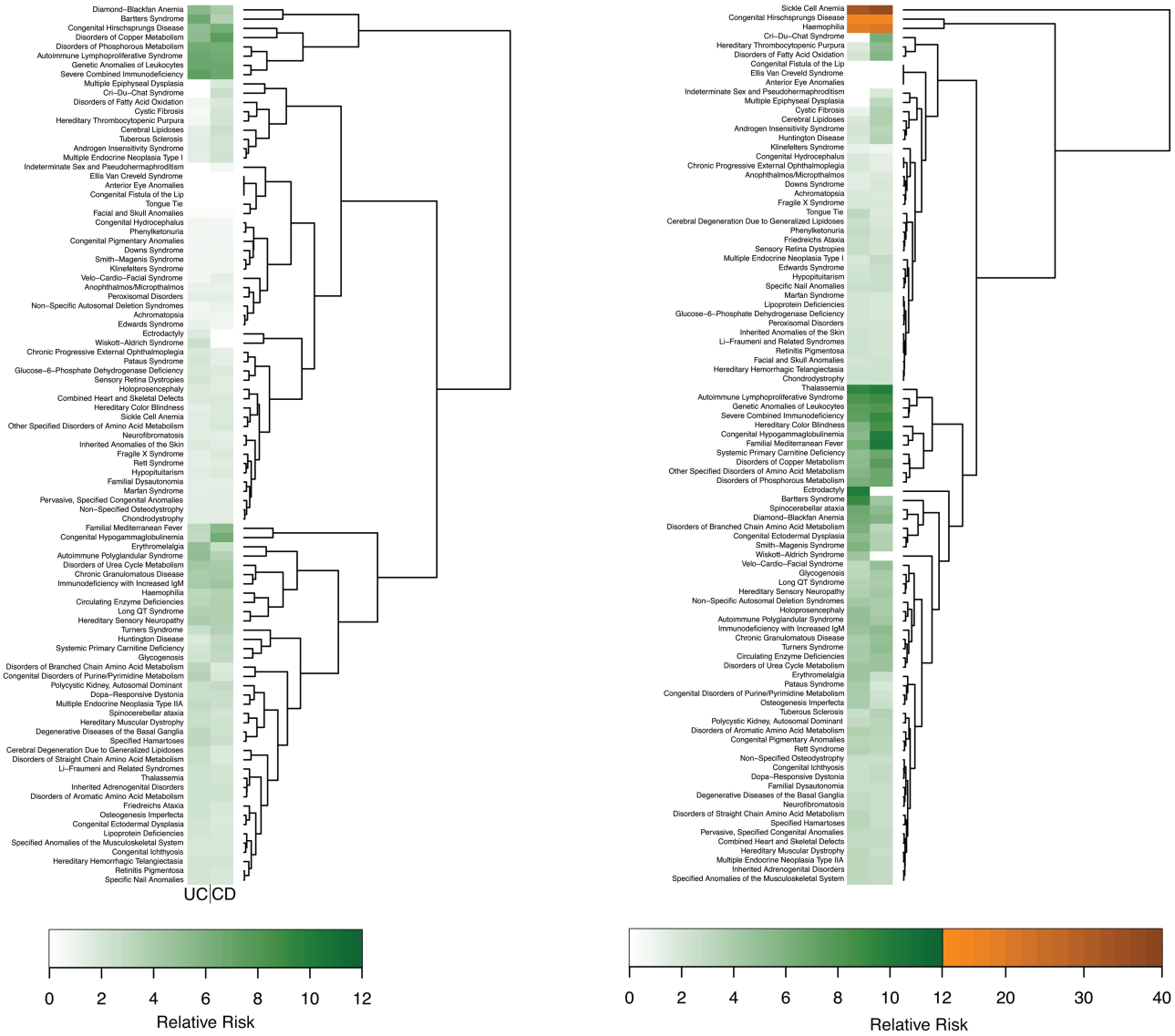

FIGURE 1. Heatmaps depicting relative risk for CD and UC using the contingency table analysis (A) and Poisson mixed-effects models (B) for all 95 Mendelian diseases. Relative risk values are presented using a green gradient color scale, and the larger relative risk values from the Poisson models are in orange. The dendrogram was constructed using Euclidean distance.

straight chain amino acid metabolism, erythromelalgia, glucose-6-phosphate dehydrogenase deficiency, hereditary hemorrhagic telangiectasia, neurofibromatosis, and osteogenesis imperfecta. The Mendelian diseases uniquely associated with CD are congenital hypogammaglobulinemia, Huntington disease, and hypopituitarism.

Enrichment of Biological Processes Associated with CD and UC

The top 15 biological processes are shown in Table 3, ordered by mean rank. There are 10 processes shared by CD and UC, and 5 processes unique to CD or UC. Overall, processes related to the immune system or metabolism were enriched, with those unique to CD more focused on tolerance, and with those unique to UC more focused on response to and regulation of stress and toxic substances.

Proximity of Mendelian IBD Genes to Known IBD Genes

We mapped 501 out of the 527 genes to STRING identifiers. From Liu et al,[4] we mapped 166 genes to STRING. Of our 501 Mendelian candidate IBD genes, 3 (0.6%) were overlapping, 454 (90.6%) were directly connected to one of the known IBD genes from Liu et al, and 44 (8.8%) were 2 links away. In total, there are 19,247 genes in STRING. There are 76.1% directly connected to one of the known IBD genes, 23.0% are 2

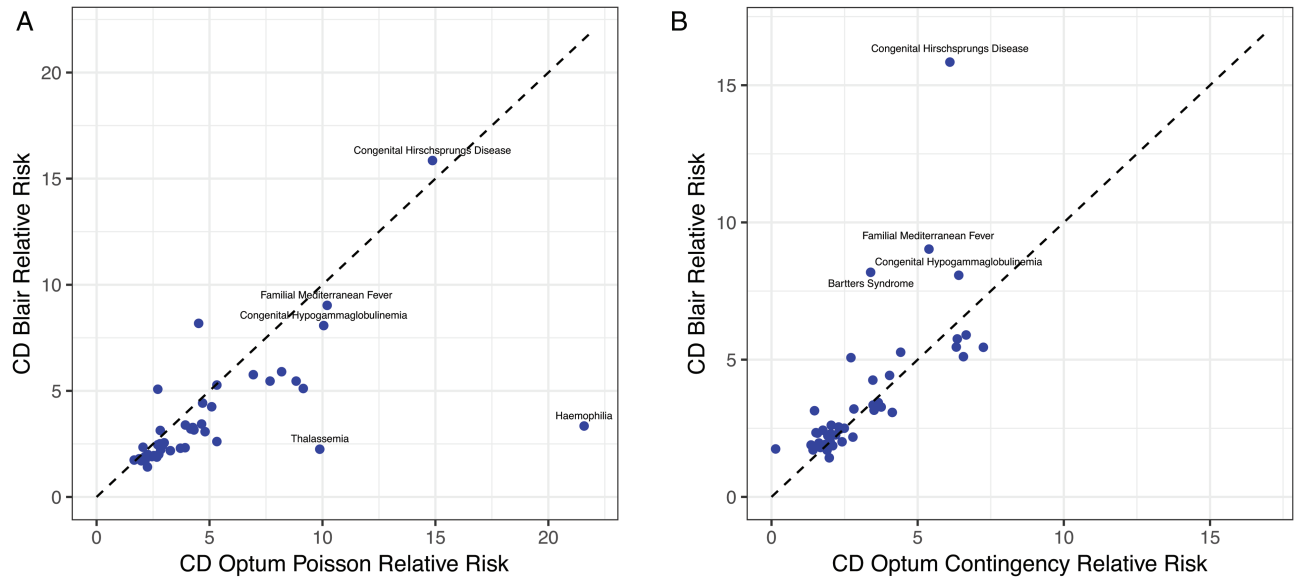

FIGURE 2. Scatterplots showing relative risk results from Optum data as compared to the original results from Blair et al using the Poisson modeling (A) and contingency table (B) approaches. Results using Optum data were not significantly different from the results from Blair et al.

links away, and 0.06% are 3 links away. Thus, though there is not a large direct overlap between Mendelian genes and previously established IBD genes, the implicated Mendelian genes are typically close to known IBD genes than if chosen by random ($P < 0.001$).

Expression of Mendelian Genes Can Distinguish IBD Subtypes

Of the 61 candidate genes uniquely associated with CD or UC, 60 were measured in the IBD expression data. Our 60-gene multinomial logistic regression model attained an accuracy of 72.1% when distinguishing the 3 classes. In the test set of 43 samples, 17 were CD, 18 were UC, and 8 were healthy. Using principal components analysis (PCA), we projected the test set samples into 2 dimensions (Fig. 3). The first and third components were chosen for better visualization. The projection shows the samples grouping by diagnosis, with UC samples on the left, healthy samples on the right, and CD in between. Notably, the samples do not tend to group by study, and that misclassified samples contain samples in all 4 datasets, with 5 from GSE16879, 4 from GSE36807, 2 from GSE9686, and 1 from GSE10616 (Fig. S2). The 12 misclassified samples consist of 6 CD samples predicted as UC, 2 CD samples predicted as healthy, and 4 UC samples as CD.Our 1000 random 60-gene classifiers had a mean accuracy of 57.5% with a standard deviation of 7.8% (Fig. 4). The lowest accuracy obtained was 32.5%, and the maximum was 82.5%. Overall, 44 out of the 1000 random 60-gene classifiers surpassed our 60-gene model. Thus, the chance of picking a random selection of 60 genes that outperforms our model is 0.044.

DISCUSSION

In this work, we investigate the Mendelian comorbidities of IBD to gain new insights into candidate genes contributing to IBD risk. In doing so, we replicated the original work from Blair et al, for CD and extend the original work by performing gene-based classification.

Our Mendelian disease association analysis revealed a group of 8 diseases with high relative risk for IBD (Fig. 1). These diseases consist mainly of metabolic, immunological, and hematological disorders and are associated with genes that modulate the immune system in IBD. For example, JAK3, IL2RG, and IL7R, which are associated with severe combined immunodeficiency and are involved in cytokine signaling,[24] which have all been implicated in IBD.[25, 26]

We found 3 Mendelian diseases significantly associated with CD only (congenital hypogammaglobulinemia, Huntington's disease, and hypopituitarism), and 8 significantly associated with UC only (Bartter's syndrome, disorders of straight chain amino acid metabolism, erythromelalgia, glucose-6-phosphate dehydrogenase deficiency, hereditary hemorrhagic telangiectasia, neurofibromatosis, and osteogenesis imperfecta). These associated diseases reveal insights into the pathophysiology of CD and UC. For example, there has been multiple case reports of patients with UC and neurofibromatosis, and it has been postulated that that the 2 diseases may have similar perturbed pathways in common involving mast cells.[27–29] Specifically, the presence of these mast cells in neuromas and in the colon have been correlated with disease progression in both diseases.[30, 31]

As these diseases may share similar implicated pathways with IBD, existing treatments for these diseases have

TABLE 2: Relative Risk and Bonferroni Corrected *P*-values for Significant ($P < 0.05$) Mendelian Diseases Associated with CD and UC[a]

Mendelian Disease	Cases No.	CD RR	CD *P*-Value	UC RR	UC *P*-Value
Disorders of Phosphorous Metabolism	176,087	6.36	0.00E+00	6.42	0.00E+00
Long QT Syndrome	37,916	3.62	4.08E-112	3.68	1.97E-112
Haemophilia	53,855	3.47	6.48E-145	3.03	7.27E-100
Disorders of Urea Cycle Metabolism	18,042	4.04	3.36E-68	4.26	9.18E-74
Genetic Anomalies of Leukocytes	3,917	6.66	1.47E-39	6.83	6.20E-40
Tongue Tie	59,368	0.14	1.98E-49	0.2	4.04E-39
Lipoprotein Deficiencies	83,695	1.9	1.44E-38	2.03	2.93E-47
Disorders of Copper Metabolism	5911	7.25	6.53E-70	5.21	3.17E-36
Thalassemia	46,956	2.19	1.25E-35	2.29	1.24E-39
Chronic Granulomatous Disease	8058	4.13	9.31E-32	3.94	1.47E-27
Hereditary Sensory Neuropathy	8066	3.75	1.08E-25	3.86	2.49E-26
Severe Combined Immunodeficiency	2510	6.57	1.55E-24	7.07	5.11E-27
Inherited Anomalies of the Skin	204,379	1.42	8.84E-23	1.63	1.70E-47
Facial and Skull Anomalies	52,501	0.36	4.24E-20	0.27	1.22E-26
Congenital Hirschsprung's Disease	3537	6.10	2.55E-30	5.02	7.49E-20
Degenerative Diseases of the Basal Ganglia	15,766	2.34	2.53E-14	3.04	2.14E-29
Li-Fraumeni and Related Syndromes	21,635	2.09	8.07E-14	2.36	8.50E-20
Specified Hamartoses	12,036	2.42	4.49E-12	3.03	4.75E-22
Polycystic Kidney, Autosomal Dominant	8605	2.79	5.34E-13	2.75	5.01E-12
Circulating Enzyme Deficiencies	5008	3.65	9.42E-15	3.36	1.49E-11
Inherited Adrenogenital Disorders	9030	2.29	2.41E-07	2.31	2.29E-07
Dopa-Responsive Dystonia	5899	2.49	4.01E-06	2.58	1.11E-06
Turner's Syndrome	6223	3.47	2.40E-16	2.45	5.98E-06
Combined Heart and Skeletal Defects	22,220	1.75	8.62E-07	1.72	5.30E-06
Familial Mediterranean Fever	3283	5.39	4.53E-22	3.03	1.44E-05
Diamond-Blackfan Anemia	1374	4.14	7.07E-05	5.43	1.19E-08
Autoimmune Lymphoproliferative Syndrome	616	6.32	4.90E-05	6.57	3.21E-05
Congenital Pigmentary Anomalies	42,160	0.60	1.09E-04	0.63	7.48E-04
Cerebral Degeneration Due to Generalized Lipidoses	15,353	1.68	1.70E-03	2.53	2.60E-17
Spinocerebellar ataxia	6734	2.05	2.05E-03	2.73	4.66E-09
Congenital Ichthyosis	8224	1.93	2.47E-03	1.93	4.11E-03
Retinitis Pigmentosa	9149	1.83	7.07E-03	2.14	1.02E-05
Autoimmune Polyglandular Syndrome	1196	3.51	1.37E-02	4.94	4.86E-06
Disorders of Aromatic Amino Acid Metabolism	4171	2.23	1.25E-02	2.39	2.34E-03
Congenital Disorders of Purine/Pyrimidine Metabolism	5755	1.98	2.43E-02	3.51	5.62E-15
Hereditary Muscular Dystrophy	3935	2.21	2.47E-02	2.69	8.92E-05
Specific Nail Anomalies	7296	1.85	2.79E-02	2.13	2.58E-04
Pervasive, Specified Congenital Anomalies	30,683	1.46	1.87E-03	1.4	3.45E-02
Glycogenosis	3193	2.82	1.63E-04	2.34	3.88E-02
Sickle Cell Anemia	23,210	1.76	3.23E-07	1.45	4.71E-02
Erythromelalgia	1763	—	—	4.77	1.46E-08
Disorders of Straight Chain Amino Acid Metabolism	5265	—	—	2.78	2.53E-07
Bartters Syndrome	706	—	—	6.61	3.42E-06
Glucose-6-Phosphate Dehydrogenase Deficiency	9501	—	—	1.97	3.52E-04
Neurofibromatosis	18,455	—	—	1.58	5.44E-03
Hereditary Hemorrhagic Telangiectasia	4,319	—	—	2.16	2.43E-02
Osteogenesis Imperfecta	3946	—	—	2.21	3.34E-02
Huntington Disease	4196	2.86	1.91E-06	—	—
Congenital Hypogammaglobulinemia	655	6.40	1.48E-05	—	—
Hypopituitarism	12,418	1.62	4.44E-02	—	—

[a]Rows are Ordered by Average P-value for Mendelian Diseases Significantly Associated with Both CD and UC, UC only, and CD Only. Abbreviation: RR, Relative Risk.

TABLE 3: Top 15 Biological Process GO Terms Associated with CD and UC Using the Genes from Significantly Associated Mendelian Diseases

Crohn's Disease	Ulcerative Colitis
Immune response	Reactive oxygen species metabolic process
Blood coagulation, intrinsic pathway	Response to stress
Blood coagulation, fibrin clot formation	***Regulation of cellular protein metabolic process***
Immune system process	***Regulation of protein metabolic process***
Reactive oxygen species metabolic process	Blood coagulation, intrinsic pathway
Vesicle-mediated transport	Blood coagulation, fibrin clot formation
Peripheral T cell tolerance induction	***Response to toxic substance***
Tolerance induction dependent upon immune response	Immune system process
Peripheral tolerance induction	***Regulation of reactive oxygen species metabolism***
Tolerance induction	Peripheral T cell tolerance induction
Central tolerance induction	Tolerance induction dependent upon immune response
Central tolerance induction to self antigens	Peripheral tolerance induction
Tolerance induction to self antigen	***Inflammatory response***
T cell tolerance induction	Tolerance induction
Response to stress	Central tolerance induction

[a]Terms that are bold italicized are unique to the disease column, and the remaining terms are common to both CD and UC

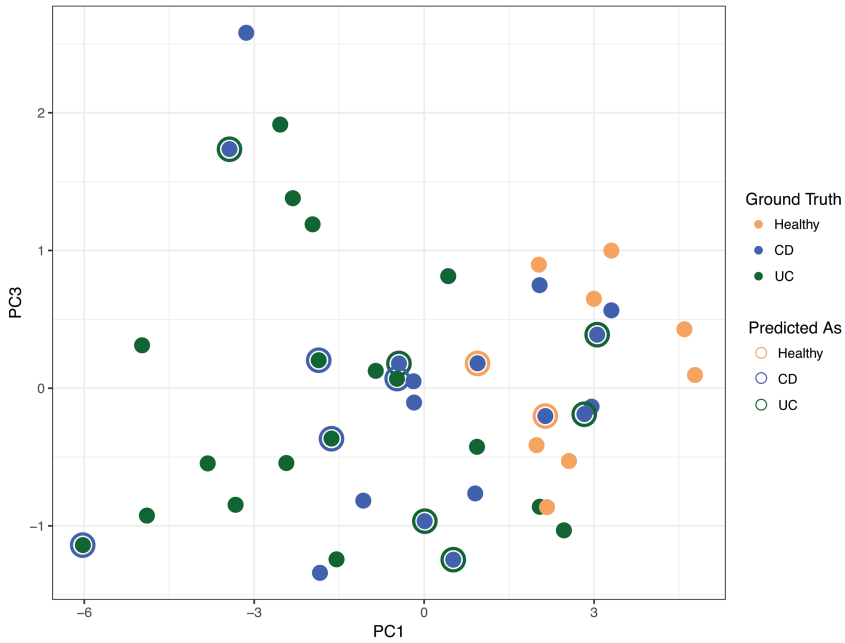

FIGURE 3. Our classification results projected using PCA. The 12 misclassified samples are encircled with the color corresponding to the predicted label. UC samples in green tend to be on the left with a low PC1 value, whereas healthy samples in yellow tend to be on the right with a high PC1 value.

repurposing potential in IBD. For example, laquinimod, an immunomodulator that is being investigated as a potential treatment for Hungtinton's disease, is also being tested for use in CD.[32] Although the exact mechanism of action is unknown, experiments have shown that laquinimod inhibits antigen presenting cells and modulates the release of inflammatory cytokines.[33–35]

Amiloride is a drug that inhibits Na+/H+ exchangers (NHEs) on the surface of epithelial cells and is used to treat Bartter's disease, which is significantly associated with UC. There is an increase in NHE expression in induced colitis models,[36] and it has been shown that lithium, which stimulates NHEs, can induce colitis.[37] Experiments have shown that blockage of these exchangers with amiloride suppresses the inflammatory

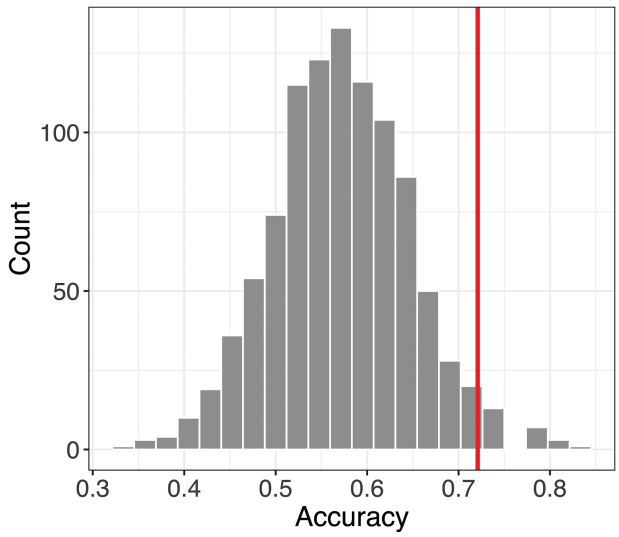

FIGURE 4. A histogram of all 1000 random 60-gene model accuracies. Our 60-gene model is shown using the vertical line at 0.721. 44 out of the 1000 random classifiers outperform our model.

response, leading to a decrease in IL-1β production, IL-8 production, and NF-κB activation.[38, 39] Amiloride and other NHE blockers may therefore have potential to be repurposed for use in IBD. Thus, these Mendelian disease associations may provide another avenue for discovering CD- or UC-specific treatments, and further study is warranted.

The candidate genes associated with these Mendelian diseases may serve as biomarkers for IBD onset and progression or as drug targets for IBD treatment. For example, *SMAD4*, one of the genes associated with hereditary hemorrhagic telangiectasia (HHT), has been implicated in UC. There have been case reports of HHT, caused by *SMAD4* haploinsufficiency, with colonic inflammation resembling IBD,[40] and mouse models have shown that mice with only 1 functional copy of *SMAD4* are more susceptible to induced colitis.[40] Furthermore, mutations in *SMAD4* and other *SMAD* proteins has been implicated in several gastric cancers[41, 42] and is associated with increased risk for colon cancer in UC.[43]

PROK2 is a gene implicated in hypopituitarism, which is significantly associated with CD. *PROK2* and its receptor are involved in nerve cell migration, and loss of *PROK2* hinders migration of neurons in the pituitary and olfactory bulb, resulting in Kallman syndrome.[44] *PROK2* also has been found to have a role in the immune system as a chemoattractant for monocytes and macrophages, and experiments have shown that incubation of monocytes with *PROK1* or *PROK2* stimulates TNF-α transcription, an inflammatory cytokine[45–47] with a key role in IBD.[48] Furthermore, biopsy samples from IBD patients and inflammatory mouse models have shown an increase in *PROK2* gene expression.[49] It has been postulated that blocking the *PROK2* receptor may have a beneficial role in IBD management.[49]

As Mendelian diseases are thought to confer risk for complex diseases by implicating similar pathways, we examined the proximity of genes implicated by associated Mendelian to known IBD genes from Liu et al. Nearly all the Mendelian IBD genes are directly connected to a known IBD gene, suggesting that they may interact with known IBD genes to contribute to the IBD phenotype. We then examined biological processes that are enriched in CD and UC using the candidate genes associated with the significant Mendelian disease comorbidities. Interestingly, though the significant Mendelian diseases associated with IBD are not primarily associated with the immune system or metabolism, such as long QT syndrome or hereditary muscular dystrophy, the top GO terms enriched by these diseases are focused on immune and metabolic processes. For example, reactive oxygen species (ROS) metabolic process was highly ranked in both CD and UC, and thought to play a role in both diseases.[50] Specifically, this was the top ranked process in UC, and regulation of ROS metabolism was unique to UC, suggesting that ROS metabolism may be more dysregulated in UC than in CD. Aminosalicylates, such as sulfasalazine and mesalazine, are one of the primary treatments for UC that has not shown to be as effective in CD.[51] The antioxidant properties of these compounds and their ability to decrease ROS concentration have been postulated as a potential mechanism for their efficacy in UC.[52, 53]

Vesicle-mediated transport is the top ranked process in CD that is not ranked for UC. Whereas most of the other implicated processes involve immune or metabolic regulation, this finding is likely tied to the importance of the autophagy pathway in CD. Multiple autophagy genes, such as *ATG16L1*, *IRGM*, and *LRRK2*, have been consistently shown by multiple GWAS and further experimental studies to be implicated specifically in CD.[54, 55] Modulating the autophagy pathway is being investigated as a potential therapeutic avenue for CD,[56] and further investigation into Mendelian genes and their pathways associated with IBD may reveal additional insights into disease pathogenesis and treatment.

In addition to linking the Mendelian diseases to IBD, we investigated the ability of these candidate Mendelian genes to differentiate CD, UC, and healthy patients using expression data from colon biopsies. Our classifier achieves 72.1% accuracy, which is a significantly better performance than expected from a set of randomly selected 60 genes. In constructing our classifier, we used 4 independent datasets so that our results are less sensitive to experiment conditions at any particular institution. Importantly, the samples tend to group by phenotype and not by study, and that misclassified samples come from all 4 studies. Out of the 4 studies, GSE16879 had the most misclassified samples. This may be because patients in GSE16879 are refractory to steroids or immunosuppression, which is not a criterion in the other studies.

Our results depend on claims data, which are often used to study epidemiological trends and disease associations. For

example, data from Optum has been used to study disease outcomes,[57] drug adherence,[58] and health care utilization.[59] However, claims data can be noisy and contain misdiagnoses and coding errors. These limitations from Blair et al apply to our work as well. As the database is limited to a decade of data, we cannot determine the exact onset of disease, and some patients that are currently healthy may develop IBD in the future. As a result, we cannot determine whether the Mendelian disease onset preceded IBD or vice versa. Furthermore, Mendelian diseases that result in early mortality that share underlying genetics with IBD may go undetected due to death before IBD onset.

Our work relies on ICD-9 codes to represent a curated group of diseases. We chose to use the same diseases and ICD-9 codes as Blair et al so our results could be compared for further insight. However, the structure of ICD-9 codes limits our ability to differentiate between disorders that may have both Mendelian and polygenic causes. For example, one third of congential Hirschsprung's disease cases are due to a Mendelian genetic syndrome, such as multiple endocrine neoplasia type 2. In the remaining cases, Hirschsprung's is thought to have a complex genetic architecture.[60] Though we establish these diseases are comorbid with IBD, any shared genetic architecture is challenging to discern and may include genes that are not related to those that result in Mendelian inheritance of the disorder.

Claims data also are subject to miscoding errors. Typically, diseases that are similar are more likely to be miscoded, and Blair et al noted that it is challenging to differentiate between miscoding errors and underlying genetic similarity. Blair et al constructed a second Poisson model to account for these errors, and they noted that the effects of this model were minimal and that the biological signal was unlikely to be attributed to miscoding. Though our methods are based on existing sources of data, we believe this work still makes a contribution toward the understanding of the genetic architecture of IBD.

Aside from revealing insights into the underlying biology of IBD, this work also replicates the previous work by Blair et al, using a different, smaller dataset. Our ability to reproduce the existing work is important as reproducibility of the primary literature is often quoted at 10%–25%.[61] Though the Optum database likely has patients in common with the Truven MarketScan database used by Blair et al, our data still has notable differences. First, we include data from 2013–2016, which is after the publication of the original work. We also have different covariates, and thus we have modified the original approach in our implementation. The concordance of these results across multiple studies and databases also lends further evidence to the existence of underlying genetic similarities between complex and Mendelian diseases.

SUPPLEMENTARY DATA

Supplementary data are available at *Inflammatory Bowel Diseases* online.

REFERENCES

1. Kappelman MD, Moore KR, Allen JK, et al. Recent trends in the prevalence of crohn's disease and ulcerative colitis in a commercially insured US population. *Dig Dis Sci*. 2013;58:519–25.
2. McGovern DP, Kugathasan S, Cho JH. Genetics of inflammatory bowel diseases. *Gastroenterology*. 2015;149:1163–1176.e2.
3. Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411:599–603.
4. Liu JZ, Sommeren S van, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 2015;47:979–986.
5. Loddo I, Romano C. Inflammatory bowel disease: genetics, epigenetics, and pathogenesis. *Front Immunol*. 2015;6:551.
6. Gambaro G, Anglani F, D'Angelo A. Association studies of genetic polymorphisms and complex disease. *Lancet*. 2000;355:308–311.
7. Cnop M, Mulder H, Igoillo-Esteve M. Diabetes in Friedreich ataxia. *J. Neurochem.* 2013;126:94–102.
8. Uhlig HH. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut*. 2013;62:1795–1805.
9. Blair DR, Lyttle CS, Mortensen JM, et al. A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk. *Cell*. 2013;155:70–80.
10. Melamed RD, Emmett KJ, Madubata C, et al. Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes. *Nat Commun*. 2015;6:7033.
11. Bates D, Mächler M, Bolker B, et al. Fitting Linear Mixed-Effects Models using lme4. *eprint arXiv:1406.5823*. 2014;67:51.
12. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: Tool for The Unification of Biology. *Nat. Genet.* 2000;25:25–29.
13. Durinck S, Spellman PT, Birney E, et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*. 2009;4:1184–1191.
14. Alexa A, Rahnenfuhrer J. *topGO: Enrichment Analysis for Gene Ontology. R package version 2.26.0*. 2016.
15. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17:122.
16. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015;43:D447–D452.
17. Arijs I, De Hertogh G, Lemaire K, et al. Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *Plos One*. 2009;4:e7984.
18. Kugathasan S, Baldassano RN, Bradfield JP, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet*. 2008;40:1211–1215.
19. Carey R, Jurickova I, Ballard E, et al. Activation of an IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm Bowel Dis*. 2008;14:446–457.
20. Montero-Meléndez T, Llor X, García-Planella E, et al. Identification of Novel Predictor Classifiers for Inflammatory Bowel Disease by Gene Expression Profiling Calogero RA, ed. *PLoS One*. 2013;8:e76235.
21. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4:249–264.
22. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*. 2007;8:118–127.
23. Venables WN, Ripley BD. *Modern Applied Statistics with S*. Springer;2002.
24. Kalman L, Lindegren ML, Kobrynski L, et al. Mutations in genes required for T-cell development: IL7R, CD45, IL2RG, JAK3, RAG1, RAG2, ARTEMIS, and ADA and severe combined immunodeficiency: huge review. *Genet Med*. 2004;6:16–26.
25. Pedersen J, Coskun M, Soendergaard C, et al. Inflammatory pathways of importance for management of inflammatory bowel disease. *World J Gastroenterol*. 2014;20:64–77.
26. Lange KM de, Barrett JC. Understanding inflammatory bowel disease via immunogenetics. *J. Autoimmun.* 2015;64:91–100.
27. Tavakkoli H, Asadi M, Mahzouni P, et al. Ulcerative colitis and neurofibromatosis type 1 with bilateral psoas muscle neurofibromas: a case report. *J. Res. Med. Sci.* 2009;14:261–265.
28. Adams W, Mitchell L, Candelaria-Santiago R, et al. Concurrent ulcerative colitis and neurofibromatosis type 1: the question of a common pathway. *Pediatrics*. 2016;137:e20150973.
29. Baratelli F, Le M, Gershman GB, et al. Do mast cells play a pathogenetic role in neurofibromatosis type 1 and ulcerative colitis? *Exp. Mol. Pathol.* 2014;96:230–234.
30. Stoyanova II, Gulubova MV. Mast cells and inflammatory mediators in chronic ulcerative colitis. *Acta Histochem*. 2002;104:185–192.
31. Staser K, Yang FC, Clapp DW. Mast cells and the neurofibroma microenvironment. *Blood*. 2010;116:157–164.

32. D'Haens G, Sandborn WJ, Colombel JF, et al. A phase II study of laquinimod in Crohn's disease. *Gut*. 2014:1–9.
33. Jolivel V, Luessi F, Masri J, et al. Modulation of dendritic cell properties by laquinimod as a mechanism for modulating multiple sclerosis. *Brain*. 2013;136:1048–1066.
34. Gurevich M, Gritzman T, Orbach R, et al. Laquinimod suppress antigen presentation in relapsing-remitting multiple sclerosis: in-vitro high-throughput gene expression study. *J Neuroimmunol*. 2010;221:87–94.
35. Zou LP, Abbas N, Volkmann I, et al. Suppression of experimental autoimmune neuritis by ABR-215062 is associated with altered Th1/Th2 balance and inhibited migration of inflammatory cells into the peripheral nerve tissue. *Neuropharmacology*. 2002;42:731–739.
36. Khan I, al-Awadi FM, Abul H. Colitis-induced changes in the expression of the Na+/H+ exchanger isoform NHE-1. *J Pharmacol Exp Ther*. 1998;285:869–875.
37. Németh ZH, Deitch EA, Szabó C, et al. Lithium induces NF-kappa B activation and interleukin-8 production in human intestinal epithelial cells. *J Biol Chem*. 2002;277:7713–7719.
38. Németh ZH, Deitch EA, Szabó C, et al. Na+/H+ exchanger blockade inhibits enterocyte inflammatory response and protects against colitis. *Am J Physiol Gastrointest Liver Physiol*. 2002;283:G122–G132.
39. Khan I, Oriowo MA, Anim JT. Amelioration of experimental colitis by Na-H exchanger-1 inhibitor amiloride is associated with reversal of IL-1ss and ERK mitogen-activated protein kinase. *Scand J Gastroenterol*. 2005;40:578–585.
40. Szigeti R, Pangas SA, Nagy-Szakal D, et al. SMAD4 haploinsufficiency associates with augmented colonic inflammation in select humans and mice. *Ann Clin Lab Sci*. 2012;42:401–408.
41. Powell SM, Harper JC, Hamilton SR, et al. Inactivation of Smad4 in gastric carcinomas. *Cancer Res*. 1997;57:4221–4224.
42. Duff EK, Clarke AR. Smad4 (DPC4)–a potent tumour suppressor? *Br J Cancer*. 1998;78:1615–1619.
43. Terdiman JP, Aust DE, Chang CG, et al. High resolution analysis of chromosome 18 alterations in ulcerative colitis-related colorectal cancer. *Cancer Genet Cytogenet*. 2002;136:129–137.
44. Dodé C, Rondard P. PROK2/PROKR2 signaling and Kallmann syndrome. *Front Endocrinol (Lausanne)*. 2013;4:19.
45. Martucci C, Franchi S, Giannini E, et al. Bv8, the amphibian homologue of the mammalian prokineticins, induces a proinflammatory phenotype of mouse macrophages. *Br J Pharmacol*. 2006;147:225–234.
46. Monnier J, Samson M. Cytokine properties of prokineticins. *Febs J*. 2008;275:4014–4021.
47. Kisliouk T, Friedman A, Klipper E, et al. Expression pattern of prokineticin 1 and its receptors in bovine ovaries during the estrous cycle: involvement in corpus luteum regression and follicular atresia. *Biol Reprod*. 2007;76:749–758.
48. Rogler G, Andus T. Cytokines in inflammatory bowel disease. *World J Surg*. 1998;22:382–389.
49. Watson RP, Lilley E, Panesar M, et al. Increased prokineticin 2 expression in gut inflammation: role in visceral pain and intestinal ion transport. *Neurogastroenterol Motil*. 2012;24:65–75.
50. Piechota-Polanczyk A, Fichna J. Review article: the role of oxidative stress in pathogenesis and treatment of inflammatory bowel diseases. *Naunyn Schmiedebergs Arch Pharmacol*. 2014;387:605–620.
51. Williams C, Panaccione R, Ghosh S, et al. Optimizing clinical use of mesalazine (5-aminosalicylic acid) in inflammatory bowel disease. *Therap Adv Gastroenterol*. 2011;4:237–248.
52. Couto D, Ribeiro D, Freitas M, et al. Scavenging of reactive oxygen and nitrogen species by the prodrug sulfasalazine and its metabolites 5-aminosalicylic acid and sulfapyridine. *Redox Rep*. 2010;15:259–267.
53. Joshi R, Kumar S, Unnikrishnan M, et al. Free radical scavenging reactions of sulfasalazine, 5-aminosalicylic acid and sulfapyridine: mechanistic aspects and antioxidant activity. *Free Radic. Res.* 2005;39:1163–1172.
54. Cho JH, Brant SR. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology*. 2011;140:1704–1712.e2.
55. Brest P, Corcelle EA, Cesaro A, et al. Autophagy and Crohn's disease: at the crossroads of infection, inflammation, immunity, and cancer. *Curr Mol Med*. 2010;10:486–502.
56. Nys K, Agostinis P, Vermeire S. Autophagy: a new target or an old strategy for the treatment of Crohn's disease? *Nat Rev Gastroenterol Hepatol*. 2013;10:395–401.
57. Buysman EK, Chow W, Henk HJ, et al. Characteristics and outcomes of patients with type 2 diabetes mellitus treated with canagliflozin: a real-world analysis. *BMC Endocr. Disord.* 2015;15:67.
58. Byfield SA, McPheeters JT, Burton TM, et al. Persistence and compliance among U.S. patients receiving pazopanib or sunitinib as first-line therapy for advanced renal cell carcinoma: a retrospective claims analysis. *J Manag Care Spec Pharm*. 2015;21:515–522.
59. Ney JP, Johnson B, Knabel T, et al. Neurologist ambulatory care, health care utilization, and costs in a large commercial dataset. *Neurology*. 2016;86:367–374.
60. Sergi CM, Caluseriu O, McColl H, et al. Hirschsprung's disease: clinical dysmorphology, genes, micro-RNAs, and future perspectives. *Pediatr. Res.* 2017;81:177–191.
61. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res*. 2015;116:116–126.