Gastro Hep Advances 2022;1:171–179

ORIGINAL RESEARCH—BASIC

A Machine Learning Approach to Identifying Causal Monogenic Variants in Inflammatory Bowel Disease

Daniel J. Mulder,[1,2] Sam Khalouei,[3] Michael Li,[3] Neil Warner,[1,4,5] Claudia Gonzaga-Jauregui,[6] Eric I. Benchimol,[1] Peter C. Church,[1] Thomas D. Walters,[1] Arun K. Ramani,[3] Anne M. Griffiths,[1] Amanda Ricciuto,[1,*] and Aleixo M. Muise[1,4,5,*]

[1]*Division of Gastroenterology, Hepatology and Nutrition, The Hospital for Sick Children, Toronto, Ontario, Canada;* [2]*Departments of Pediatrics; Medicine and Biomedical and Molecular Sciences, Queen's University, Kingston, Ontario, Canada;* [3]*Centre for Computational Medicine; The Hospital for Sick Children, Toronto, Ontario, Canada;* [4]*SickKids Inflammatory Bowel Disease Centre and Cell Biology Program; Research Institute, Hospital for Sick Children, Toronto, Ontario, Canada;* [5]*Department of Pediatrics and Biochemistry; Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada; and* [6]*Regeneron Genetics Center, Regeneron Pharmaceuticals Inc, Tarrytown, New York*

BACKGROUND AND AIMS: Diagnosis of monogenic disease is increasingly important for patient care and personalizing therapy. However, the current process is nonstandardized, expensive, and time consuming. There is currently no accepted strategy to help identify disease-causing variants in monogenic inflammatory bowel disease (IBD). The aim of the study is to develop a prioritization strategy for monogenic IBD variant discovery through detailed analysis of a whole-exome sequencing (WES) data set. **METHODS:** All consenting pediatric patients with IBD presenting to our tertiary care hospital during the study period were enrolled and underwent WES (n = 1005). Available family members also underwent WES. Variants were analyzed en masse using the GEMINI framework and were further annotated using data from dbNSFP, Combined Annotation Dependent Depletion, and gnomAD. Known disease-causing variants (n = 36) were used as positive controls. Machine learning algorithms were optimized and then compared to assist with identifying monogenic IBD case characteristics. **RESULTS:** Initial gene-level analysis identified 11 genes not previously linked to IBD that could potentially harbor IBD-causing variants. Machine learning algorithms identified 4 primary variant characteristics (Combined Annotation Dependent Depletion score, dbNSFP score, relationship with a known immunodeficiency gene, and alternate allele frequency), and optimal threshold values for each were determined to assist with identifying monogenic IBD variants. Based on these characteristics, an automated variant prioritization pipeline was then created that filters and prioritizes variants from >100,000 variants per patient down to a mean of 15. This pipeline is available online for all to use. **CONCLUSION:** Leveraging a large WES data set, we demonstrate a statistically rigorous strategy for prioritization of variants for monogenic IBD diagnosis.

Keywords: Whole-exome Sequencing; Monogenic Disease; Pediatric IBD; Machine Learning

Background

Inflammatory bowel disease (IBD) is a chronic lifelong disorder that affects millions throughout the world[1] and results in substantial long-term health care costs.[2] It has been estimated that 3% of pediatric patients with IBD[3] have monogenic IBD, where pathogenic variants in the genome lead to disease. Monogenic IBD cases are amenable to personalized therapy, including about one-third of whom become eligible for curative hematopoietic stem cell transplant.[3] Thus, next-generation sequencing is currently recommended for very young patients presenting with IBD and those with features of a genetic disease, to improve prognosis and inform treatment decisions.[4]

Although valuable, the process of identifying monogenic forms of IBD through next-generation sequencing is time consuming and challenging because of many factors: the very high number of nonmonogenic variants returned per person (usually >100,000), the rarity of disease-causing variants in most populations, and the complex mechanisms through which variants cause pathogenicity.[5] Dozens of pathogenic variants leading to monogenic forms of IBD have been described.[4,6] Monogenic disease-causing variants are often discovered by an unstructured process after subjective manual analysis on a patient-by-patient and variant-by-variant basis, even in large-scale cohorts.[3,7] There is currently no widely accepted analysis pipeline available to

*Co-senior author.

Abbreviations used in this paper: AIC, Akaike information criterion; AUC, area under the curve; CADD, Combined Annotation Dependent Depletion; CART, classification and regression trees; CDG, Closest Disease-causing Gene; IBD, inflammatory bowel disease; kNN, k nearest neighbors; LOEUF, loss-of-function observed/expected upper bound fraction; PID, primary immunodeficiency; ROC, receiver operator characteristic; SMOTE, synthetic minority over-sampling technique; SVM, support vector machine; WES, whole-exome sequencing.

Most current article

2772-5723
https://doi.org/10.1016/j.gastha.2021.11.002

help identify disease-causing variants in monogenic IBD, limiting diagnosis to expert centers with previous experience. This is a widespread challenge across many fields involved in rare genetic diagnosis, including identification of cancer-driving variants in oncology.[8] A variant prioritization strategy for identifying candidate monogenic IBD variants could assist investigators and clinicians worldwide in this difficult but invaluable process to help reduce diagnostic odysseys.

Statistical analysis of large complex data sets, such as genomic databases, can be challenging to manually analyze.[9] Trial-and-error approaches to testing a hypothesis may not always be able to uncover meaningful patterns in vast data sets. Thus, machine learning algorithms can be used to identify important relationships between specific aspects of a data set that would otherwise be obscured by the size of the data.

In this study, we developed a prioritization strategy for identifying candidate variants that could lead to monogenic IBD by performing in-depth statistical analysis using a previously curated large whole-exome sequencing (WES) data set. We approached this problem by combining our prior experience in variant identification with a statistical approach through machine learning techniques. This open-access variant prioritization pipeline could aid clinicians worldwide in rapidly identifying disease-causing variants in monogenic IBD cases with high sensitivity.

Methods

Patients

We performed WES on a single-center cohort of 1005 ethnically diverse (Figure A1) pediatric patients with IBD and their relatives. Previously, approximately 3% of patients in our cohort were found to have monogenic variants, which was supported by functional studies.[3] From the initial 40 known variants from 31 patients, there were 4 variants excluded from the present study because of having been sequenced using different methods from the remainder of the patients. The full details of the variants, including the patient clinical information, variant location, and impact of the variants, are detailed in Table 2 and further in Table S6 in the study by Crowley et al[3] (excluded from the present study were patients 11, 15, and 26). In the present study, these cases served as the positive controls for the algorithms developed and validated in this study. This study was approved by the Hospital for Sick Children Research Ethics Board (REB 1000024905), and all patients or guardians provided informed consent. Any consenting, eligible patients younger than 18 years old with a diagnosis of IBD who were followed at Hospital for Sick Children were enrolled. All probands and available primary family members underwent WES (total n = 2305). Patients were excluded if they were referred for a second opinion or had known syndromic disease, chromosomal abnormalities, or previously diagnosed primary immunodeficiency.

Whole-exome Sequencing

Peripheral venous blood samples were collected in ethylenediaminetetraacetic acid-containing tubes. Genomic DNA was extracted using a Puregene Blood Kit (Qiagen, Hilden, Germany) as per the manufacturer's instructions. Approximately 2 mg of DNA was used for WES. DNA was fragmented through ultrasonication to a mean size of approximately 150 base pairs. Exomes were captured using the NimbleGen VCRome capture design kit, version 2.1 (Roche, Basel, Switzerland). Samples were sequenced by our collaborators at the Regeneron Genetics Center on the Illumina HiSeq 2500 platform using paired-end 75 bp reads and 2 indexing reads.[10] Further details on alignment, variant calling, and annotation can be found in the Supplemental Methods.

Variant Annotation and Feature Selection

Genetic variant calls from all probands were collected in a central database created using the GEMINI framework (version 0.18)[11] and annotated for basic clinical phenotype (sex and age at diagnosis), population frequency (minor allele frequency, via gnomAD v2.1.1),[12] evolutionary constraint (loss-of-function observed/expected upper bound fraction [LOEUF], via gnomAD v2.1.1), presence of previously reported variants (from ClinVar Nov 2020),[13] and in silico damaging prediction metrics (via dbNSFP v4.1a[14] and Combined Annotation Dependent Depletion (CADD) phred score v1.3[15]).

Lists of genes likely to harbor disease-causing variants were identified from current literature in 4 gene lists (Table A2). For the known monogenic IBD-associated genes, we reviewed current definitive guidelines[4,16] and recent publications.[17,18] The Closest Disease-causing Gene (CDG) list was created from the 99 known monogenic IBD genes, using the online Human Gene Mutation Database CDG server (http://pec630.rockefeller.edu:8080/CDG-OMIM/).[19] The primary immunodeficiency (PID) gene list was created using the latest summaries of known monogenic PID genes.[6,7,20] The IBD genome-wide association list was created using the latest genes identified in the literature.[21–23] Variants were identified as being in proximity to a gene of interest if found within the start or end of the gene coordinates as per current Ensembl coordinates (v102, http://ensembl.org,[24] +/−5000 bp).

Gene interaction data were extracted from the STRING database (https://string-db.org, v11).[25] Interconnections between the 99 previously known monogenic IBD genes were quantified by the following parameters: "full network" type, "confidence" value (for network edges), "experiments, databases, co-expression, neighborhood, and gene fusion" interaction sources, and minimum confidence was set to 0.7 (high).

Variant frequency histograms and curves for the annotations are found in Figure A2. Two ClinVar features were evaluated for each variant. The "ClinSig Simple" value, provided by ClinVar, denotes variants that have at least one reported pathogenic or likely pathogenic variant at that location in the database. Given that the "ClinSig Simple" value can be restrictive and that the aim of this study was to search broadly, we created a "ClinVar Broad" value for each variant calculated using a custom R script where a variant was identified as positive if its ClinVar entry contained at least one label of the following: "affects", "risk factor", "association", "likely pathogenic", "pathogenic", or "uncertain significance", while excluding any variants with labels of "benign" or "likely benign" or "protective".

For dbNSFP annotation, a custom R script was created to summarize only the dbNSFP (version 4.1a) scores that provide

classification of a variant (rather than only numerical values) by the database or by the prediction algorithm authors (eg, 'benign' or 'damaging' rather than only a raw score). Full details of the classification recommendations for each score are available in the dbNSFP readme file. This functionality is included in the final pipeline (https://github.com/DanJMulder/monoibdpriority). Given that each algorithm can give multiple predictions, the summary recommendations were given a value of one if more than 50% of the total predictions for that algorithm were damaging (or "high" or "medium"); otherwise, the algorithm's prediction was given a value of zero. There are 21 prediction algorithms in dbNSFP that have a recommendation that can be summarized in this way. Thus, the total number of algorithms predicting a variant to be damaging by this criterion was summarized as a numeric score out of 21 for each variant in the database.

Machine Learning

Several machine learning models were used in the data analysis pipeline to explore the predictive value of each of the 16 annotated features (the selection process detailed in the Supplemental Methods section) to identify disease-causing variants compared with identification previously carried out through manual filtration (Figure 1). This process then guided the construction of our final filtering strategy. Algorithms (univariate logistic regression, multivariate logistic regression, classification and regression trees [CART], k nearest neighbors [kNN], support vector machine [SVM], and random forests) were selected for their applicability to the problem type (binary classification), performance with an unbalanced data set, and explainability (interpretability of feature contribution to the model).

Data Preprocessing. Overall, a total of >5 million variants were called in the cohort. Sixteen commonly used variant annotation features were evaluated (Table A1), as available, for each variant. The expected distributions of each annotation were evaluated by exploratory data analysis (Figure A2). Variants with missing data (6.2%) were removed for the machine learning data analysis step. Numerical features were normalized using the *caret* R package (v6.0) to a value between 0 and 1.

Logistic Regression. Univariate and multivariate logistic regressions were performed using the base R *glm* function to examine the association between the various annotation features and the confirmed monogenic variants. Once a univariate model was created for each individual feature, multivariate models were created using a stepwise forward addition approach, starting with the features with the lowest Akaike information criterion (AIC) value in the univariate models. The data were not resampled for the logistic regression models.

Resampling. Given that most machine learning algorithms have optimal performance with a 1:1 balanced data set, the highly unbalanced nature of the binary classifier in this study (36 known monogenic variants out of approximately 5 million variants total, a ~1:130,000 ratio) was addressed by resampling using the synthetic minority over-sampling technique (SMOTE) method[26] from the *DMwR* R package (v0.4.1, settings: k = 5, per.over = 1000, perc.under = 2000). After resampling, there were 286 synthetic monogenic variants and 5200 randomly selected nonmonogenic variants. After SMOTE resampling (and after the logistic regression analysis performed previously), the data set was then randomly separated into training and test sets (80% and 20%, respectively).

Four machine learning algorithms were trained and hyperparameters were tuned to optimize recall and then accuracy using a grid search approach. The CART model was created using the *rpart* R package (v4.1), where complexity was optimized. The kNN classification model was created using the *caret* R package (v6.0) by the inflection point on the receiver operator characteristic at area under the curve (ROC AUC) vs k graph (determined to be optimal at k = 9). The single classifier SVM model was created using the caret R package (with a radial basis kernel and sigma = 0.6310028 and cost = 1). The random forest model was created using the *randomForest* R package (v4.6, with the optimized parameters being mtry = 14 and ntree = 500). Once hyperparameter tuning was complete, a 10-fold repeated cross-validation approach was used to finalize the models.

A final variant filtration strategy was created using the features that were consistently most predictive of monogenic disease in both the logistic regression and random forest models. Threshold values for numeric predictors were determined by the value that had the optimal specificity, while maintaining 100% sensitivity (ie, without excluding any of the known monogenic cases).

Software, Dependencies, and Statistical Analysis

All custom scripts were written in R v3.6.1. Custom scripts are available on GitHub (https://github.com/DanJMulder/monoibdpriority). R software packages and versions used for both data processing and statistical analysis can be found in Table A3. Logistic regression models were evaluated by AIC, *P*-value, and area under the ROC curve. CART, kNN, SVM, and random forest models were evaluated primarily by area under the precision-recall curve, but evaluation also included recall, Cohen's kappa, F1 statistic, and area under the ROC curve. Statistics were calculated by the R packages used to create the models. Random forest feature relative importance was calculated by the *inTrees* R package (v1.2).

Results

Overall, custom annotated variants were analyzed and prioritized by 2 processes: manual prioritization of filtered variants (as previously described by Crowley et al[3]) and machine learning. Strongly predictive variant features were incorporated into the final prioritization strategy.

Gene-level Analysis Reveals Novel Potential Monogenic Genes and Relationships

Current guidelines suggest starting with known genes when searching for disease-causing variants.[27] Genes from all 4 lists of potential candidate genes (monogenic IBD genes, IBD genome-wide association study study genes, primary immunodeficiency genes, and the closest disease-causing genes database genes, Table A2) were noted to have substantial overlap (Figure 2A). Notably, of the 4 lists,

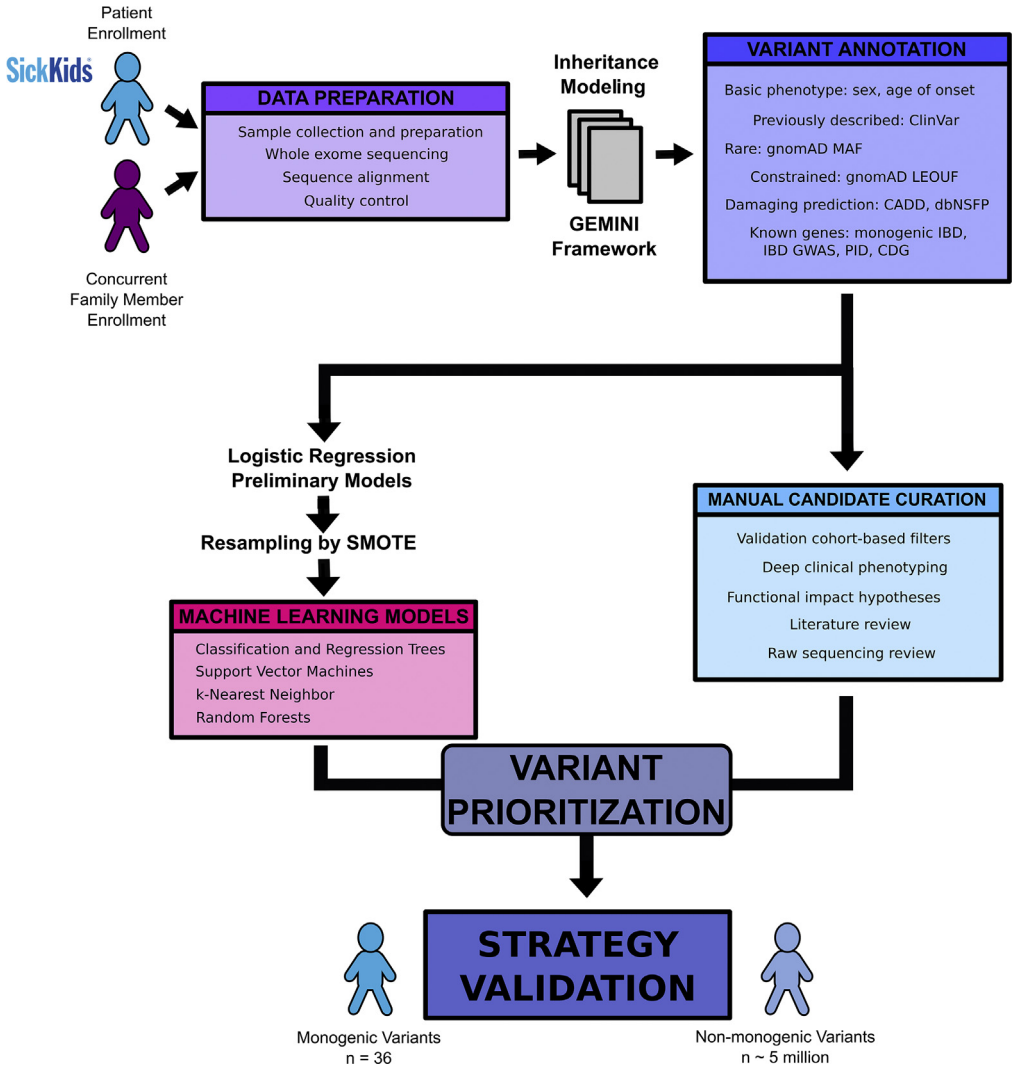

Figure 1. Study flow diagram. Patients were enrolled through local clinical referral and recruitment. Available family members were also enrolled. Whole-exome sequencing was then performed, and variants were identified based on standard alignment and quality control. The GEMINI framework was used to organize variant calls into a central database. Variants were then annotated with 16 separate features, including minor allele frequency (MAF), evolutionary constraint (LOEUF), damaging prediction (from CADD and dbNSFP databases), and occurrence in known genes (from conditions including monogenic IBD, IBD genome-wide association studies, primary immunodeficiency, and the closest disease-causing gene database). Characteristics most likely to be associated with causative variants for monogenic IBD were prioritized using parallel manual curation and machine learning strategies. Manual curation for the local cohort had previously identified 36 variants as monogenic (as described by Crowley et al, Gastroenterology 2020; 158(8):2208–2220), which were used as positive controls for the present study. The data set was resampled using the SMOTE for machine learning. A stepwise filtering strategy was then developed to create an easily implementable automated variant prioritization strategy that was then applied individually to our internal validation cohort (including known positive control cases).

42 genes were present in at least 3 of the 4 lists and 8 genes were present in all 4 lists (Figure A3). Eleven of the genes that appear in at least 3 lists had not previously been associated with monogenic IBD.

Relationships and proximity between known monogenic IBD-causing genes were investigated using network analysis. Figure 2B illustrates a central group of highly interrelated genes associated with monogenic IBD (approximated by the gray circle). The curated gene lists and network analysis were added to the variant annotation pipeline and could also serve as reference for further manual variant curation by phenotype-genotype-mechanism connections.

Statistical Evaluation of Genetic Annotation Features in Identifying Monogenic IBD Variants

Initial univariate logistic regression identified and quantified the contribution of individual variant characteristics to identifying known monogenic variants. The AIC, an estimation of prediction error relative to other features,

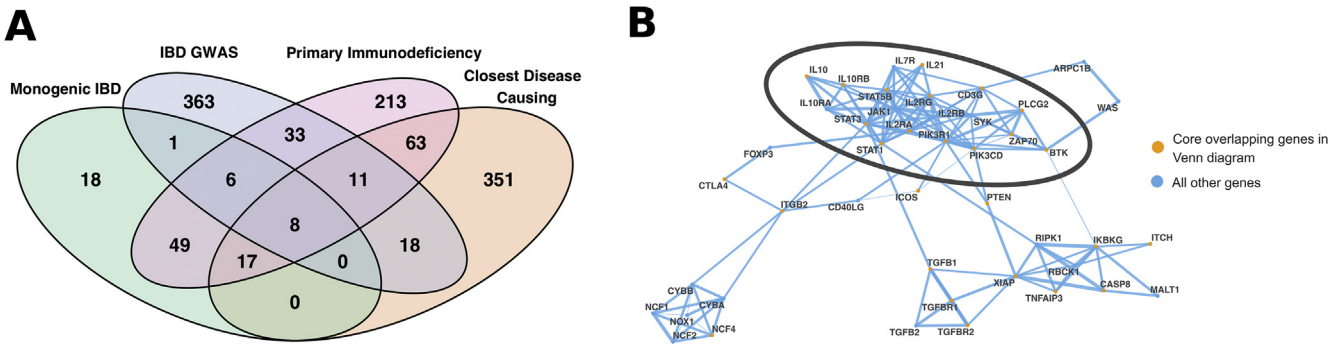

Figure 2. Genes common to IBD and IBD-related conditions have substantial interconnections. (A) A Venn diagram illustrating the substantial number of overlapping genes identified between 4 related groups of important IBD-related genes. Gene lists included the following: known monogenic IBD genes (n = 99), genes identified on IBD genome-wide association study (n = 438), known monogenic primary immunodeficiency genes (n = 400), and genes identified through the Closest Disease-causing Gene tool (input was the 99 known monogenic genes, output n = 468 genes). (B) STRING network analysis of the subset of known monogenic genes that had at least one connection with another known monogenic gene. Edge thickness is scaled based on STRING connection. The central core of genes is circled to highlight the strong interaction in these 19 genes. Genes that appeared in at least 3 of 4 of our gene lists (as illustrated in panel A) are colored orange; all other nodes are colored blue.

(Figure 3A) and ROC AUC (Figure 3B) identified 5 characteristics (out of the initial group of 16 characteristics) that had the highest contribution to prediction of a variant being monogenic relative to the other characteristics (Table A4). Multivariate logistic regression models were then built using combinations of these predictors. A combination of 4 predictors (CADD score, dbNSFP score, PID gene, and allele frequency) was found to have the strongest statistical power in predicting a monogenic variant in the multivariate model (Figure 3C and D). Further feature incorporation beyond these 4 features did not result in meaningful improvement in statistical power (Table A5).

Feature selection using multivariate logistic regression requires manual stepwise model building and iterative decision-making. In an attempt to improve on these limitations, a comparison was made between the ability of 4 supervised machine learning models to maximize the AUC for precision recall (P-R) curves. Positive control data (Figure A4A) was enriched using SMOTE (Figure A4B). Comparison between CART, SVM (Figure 4A), kNN

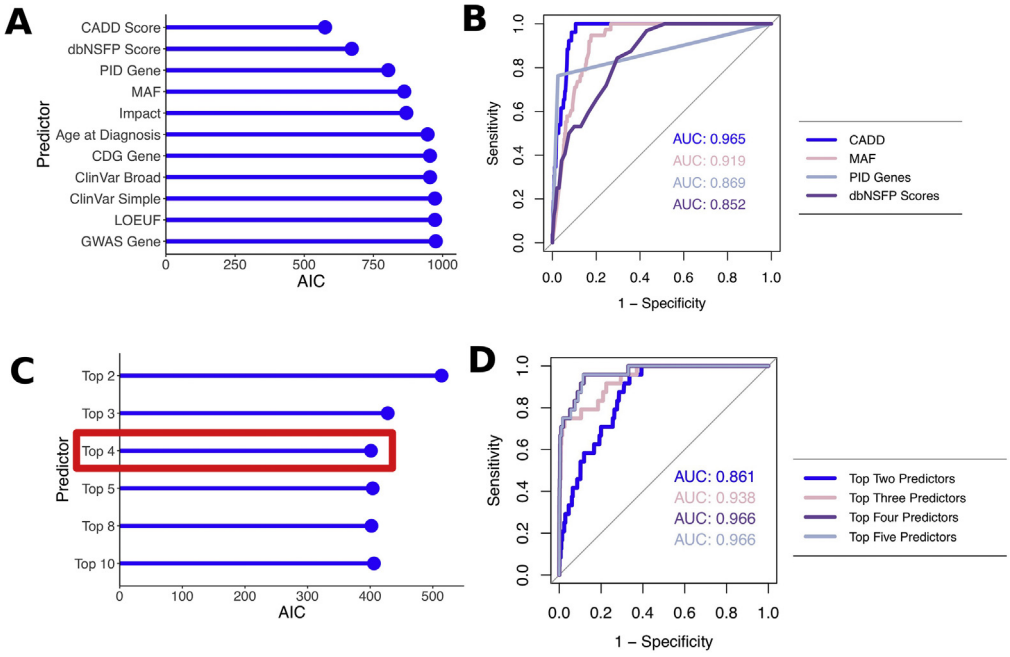

Figure 3. Logistic regression identifies variant characteristics that are associated with monogenic disease. (A) Single predictor (univariate) logistic regression comparison by AIC and (B) ROC AUC, where a lower AIC and higher AUC indicate relative goodness-of-fit for predicting a monogenic variant. Multivariate model comparison by AIC (C) and ROC AUC (D). Overall, the "Top 4" multivariate logistic regression model (highlighted in red), that includes the top 4 univariate predictors in panel A, minimizes the AIC and maximizes the AUC, demonstrating this model as having an optimal goodness-of-fit.

(Figure 4B), and random forests (Figure 4C) demonstrated that the random forest model performed best in terms of P-R AUC and also had the highest F1 statistic of 0.855 (Table A6). Relative feature importance was extracted from the random forest model (Figure 4D) to help inform further variant prioritization.

Stepwise Variant Filtering and Prioritization Strategy

Informed by the abovementioned analysis of the 7 features most predictive of monogenic variants, a variant prioritization strategy was created with the aim that it could be customized for use by other groups looking to find rare variants that could be responsible for monogenic IBD cases. For an individual variant call format file (VCF), after quality control, the following filters are applied by the pipeline: RefSeq "protein coding", CADD score >18, dbNFSP custom score >2, gnomAD allele frequency <0.003, gnomAD LOEUF <1.5, and coordinates within a known monogenic IBD or primary immunodeficiency gene. Thresholds were set by evaluation of the minimum or maximum value in the known monogenic IBD (positive control) cases in the cohort. Notably, LOEUF is a gene-level filter, but at the determined threshold, 2 known monogenic IBD genes (*ORA1* and *CYBA*) are eliminated. Hierarchical prioritization of the selected variants was then performed and sorted stepwise by occurring within a known monogenic IBD gene, RefSeq "exonic" label, allele frequency (descending), and CADD score (ascending).

For internal validation, the known monogenic cases were re-evaluated individually from the raw VCF stage (before splitting data into testing and training sets and before SMOTE resampling) using our custom pipeline script (Figure 5, available online at https://github.com/DanJMulder/monoibdpriority). After initial filtering using the strategy described previously, there remained, on average, 15 variants per proband (an approximate reduction in variants of about 10,000-fold) with a range of 9–26 variants per patient. None of the known monogenic disease-

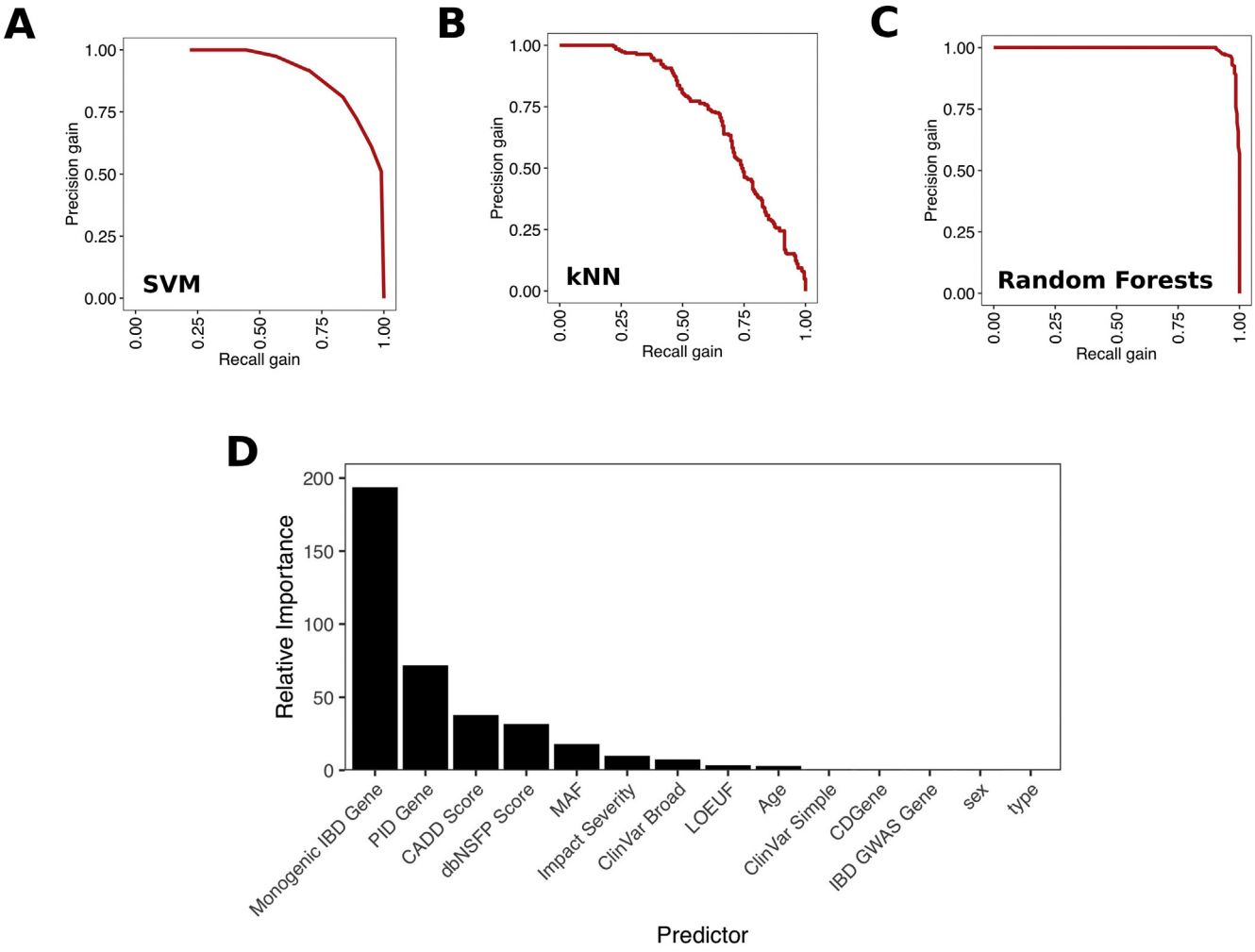

Figure 4. Machine learning algorithms improve feature selection for monogenic variant identification. Precision-recall curves compared the ability of machine learning algorithms to identify the known monogenic cases in the data set. Algorithms evaluated were (A) SVM, (B) kNN, and (C) random forests. The random forest model had the highest AUC on the precision-recall curves. (D) Relative importance of features contributing to the random forest model.

causing variants were filtered out in this process. The prioritization strategy resulted in a monogenic variant being prioritized first in 18 of the 27 cases. For the cases with 2 monogenic variants (ie, either autosomal recessive or compound heterozygous cases), both monogenic variants were ranked in the top 3 variants in 8 of 9 cases. Only 2 variants were ranked lower than third (fourth and sixth). Inheritance patterns for every variant are computed in this pipeline if family member sequencing is available. However, inheritance models were not necessary to enable the high prioritization of the disease-causing variants in our positive control cases. Of note, the open-source nature of this filtration and prioritization code can be easily customized to accommodate different thresholds by the end user to assist with both more and less stringent filtration, as desired.

The final pipeline is open access and requires only the raw patient sequence (VCF) file and local instances of 4 open-source tools (dbNSFP, Annovar, CADD, and gnomAD) to run. Once set up, the pipeline runs entirely offline and is fully customizable.

Discussion

In this study, we used detailed statistical analysis to create a powerful variant prioritization tool for identifying disease-causing monogenic IBD variants. On an individual case basis, our open-source pipeline is able to use a small number of variant characteristics to rapidly filter the possible disease-causing variants from tens of thousands per patient, to approximately 15 per patient. The pipeline is also able to sort variants effectively, ranking the disease-causing variants above all others in most cases, and perform inheritance modeling.

This study addresses the current major success-limiting step for diagnosis of monogenic IBD: disease causing variant prioritization. Using rigorously defined, statistically determined parameters, the present pipeline provides a framework for variant prioritization. Disease-causing monogenic IBD variant discovery has previously been dependent on manual trial-and-error curation and lacks any standardized approach. Genetic sequencing is increasingly touted as cost-effective, but this claim assumes the genetic variant is discovered in the process.[28] By providing a starting-point filtering and prioritization strategy, analysts will have a small and manageable list of high-confidence/probability variants to more likely be able to identify candidate monogenic variants for functional validation and diagnosis.

Despite beginning with 16 variant characteristics across >5 million variants, this study was able to use a machine learning–based approach to narrow down the strategy to 6 variant characteristics that were highly sensitivity for monogenic variant discovery. This strength is dependent on this unique data set which features hundreds of genomes of patients with IBD analyzed over many years. Variant discovery is plagued in many instances by the "needle in the haystack" problem, where 1 or 2 diagnostic variants lie

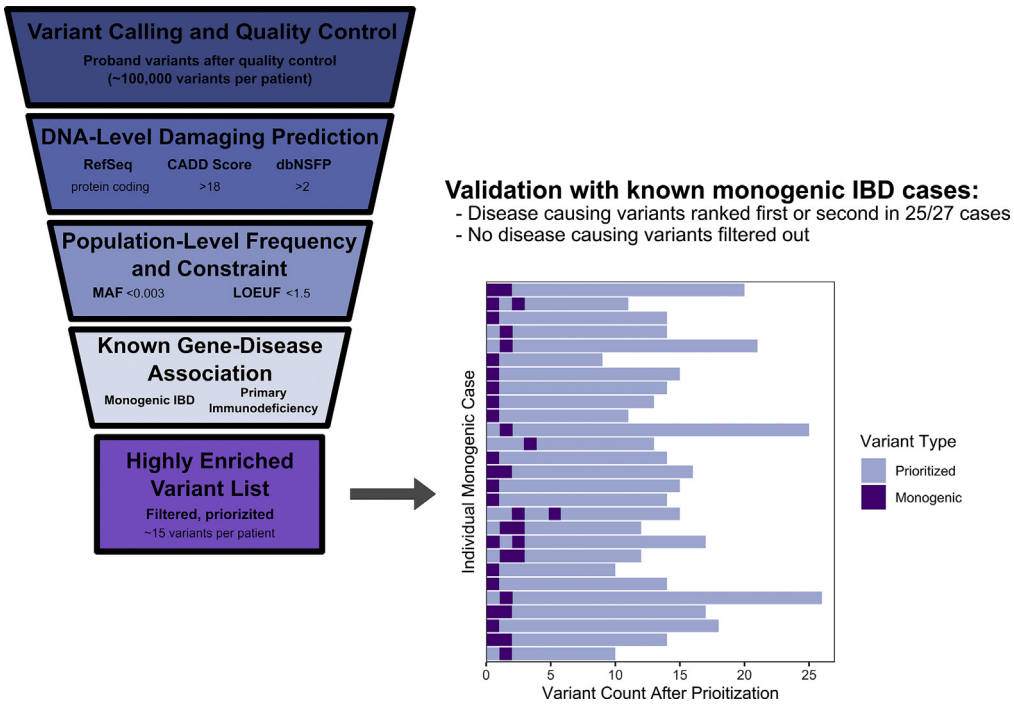

Figure 5. Variant prioritization pipeline and positive control filtering results. The pipeline illustrates the variant characteristics and thresholds used for initial filtering. Variants remaining after filtration are then prioritized (purple box) stepwise by occurring within a known monogenic IBD gene, RefSeq "exonic" label, allele frequency (descending), and CADD score (ascending). The graph illustrates internal validation of the filtering strategy by applying our pipeline via the custom R script for filtering and prioritization to known monogenic IBD cases, where a disease-causing variant was ranked first or second by the pipeline in 25 of the 27 cases.

within a data set of more than tens of thousands of nonpathogenic variants. Our multivariate regression analysis, in particular, clearly demonstrated the minimal improvement that further variant annotations add to disease-causing variant identification. Another strength of this proposed prioritization approach is that providing sequenced family members is not mandatory and is only used to support variant prioritization if available, thus making this strategy more amenable in low-resource settings.

An important limitation of the present study is that the filtration thresholds identified from this data set may be different if different methodology is used. For example, if moving from whole-exome to whole-genome sequencing, the consideration of an intronic variant would require new data characteristics that do not include scores that rely on codon changes or amino acid biochemistry. In addition, it is likely that the gene lists in the present study will continue to expand. Given these considerations, we have provided the open-source code for the filtration pipeline, which can be customized by end users for the many unique aspects of other data sets. Validation on external data sets will be an important future direction for this work. We also will aim to improve the generalizability of this tool ourselves for intronic variants as more monogenic IBD cases are identified through whole-genome sequencing.

Despite the computational power of this pipeline, the role for humans in the analysis process is still necessary. Importantly, a statistical approach such as ours is prone to overfitting, especially when high-dimensional data are used to identify a small number of monogenic IBD cases. This should be taken into consideration when interpreting and adapting the filtration and prioritization steps and applying them to external patient exomes. As with other genetic diseases,[5] monogenic IBD diagnosis must still rely on expert decision-making in parallel to predefined pipelines to ensure appropriate genotype-phenotype correlation and appropriate functional validation of mechanistic hypotheses.[4] Indeed, the strategy used in this study follows the current guidelines for identifying variants implicated in causing disease.[27]

This prioritization strategy is technically limited by the use of WES data, which does not allow for interrogation of structural variants and intronic variants. In addition, despite the demonstrated broad ethnic diversity of this cohort, this was a single-center study. It is still likely that the filtering parameters will need to be altered for cohorts where genetic diversity is limited, leading to higher allele frequencies and complicating inheritance patterns. Notably, the open-source pipeline scripts linked previously can be easily customized to local specifications. Currently, linking the clinical phenotype to monogenic variants suffers from the dual challenge of lack of standardization in the medical record and the clinical heterogeneity of diseases, such as monogenic IBD.[29] Previous work has linked earlier onset and extra-intestinal manifestations with monogenic disease.[3] In our machine learning models, neither sex nor age at diagnosis was statistically associated strongly enough with monogenic cases to warrant inclusion in our prioritization pipeline. As electronic medical records and phenotype entry become more standardized (ie, through projects like the Human Phenotype Ontology), it seems likely that clinical phenotypes will soon be easier to incorporate into the pipeline.

In summary, we present an open-source, novel, generalizable strategy for genetic diagnosis of monogenic IBD. The filtration and prioritization pipeline is based on a large cohort experience with monogenic IBD diagnosis. We envision that this approach could empower multidisciplinary teams worldwide to rapidly identify monogenic IBD-causing variants and could also be adapted to other monogenic diseases. Genetic diagnosis can greatly improve prognosis through previously established personalized therapy.[30]

Supplementary Materials

Material associated with this article can be found in the online version at https://doi.org/10.1016/j.gastha.2021.11.002.

References

1. Kaplan GG. The global burden of IBD: from 2015 to 2025. Nat Rev Gastroenterol Hepatol 2015;12:720–727.
2. Kuenzig ME, Benchimol EI, Lee L, et al. The impact of inflammatory bowel disease in Canada 2018: direct costs and health services utilization. J Can Assoc Gastroenterol 2019;2(Suppl 1):S17–S33.
3. [dataset] Crowley E, Warner N, Pan J, et al. Prevalence and clinical features of inflammatory bowel diseases associated with monogenic variants, identified by whole-exome sequencing in 1000 children at a single center. Gastroenterology 2020;158:2208–2220.
4. Uhlig HH, Charbit-Henrion F, Kotlarz D, et al. Clinical genomics for the diagnosis of monogenic forms of inflammatory bowel disease: a position paper from the Paediatric IBD Porto Group of European Society of Paediatric Gastroenterology, Hepatology and Nutrition. J Pediatr Gastroenterol Nutr 2021;72:456–473.
5. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet 2017;18:599–612.
6. Kelsen JR, Dawany N, Moran CJ, et al. Exome sequencing analysis reveals variants in primary immunodeficiency genes in patients with very early onset inflammatory bowel disease. Gastroenterology 2015;149:1415–1424.
7. Thaventhiran JED, Lango Allen H, Burren OS, et al. Whole-genome sequencing of a sporadic primary immunodeficiency cohort. Nature 2020;583:90–95.
8. Patel AP, Wang M, Fahed AC, et al. Association of rare pathogenic DNA variants for familial hypercholesterolemia, hereditary breast and ovarian cancer syndrome, and Lynch syndrome with disease risk in adults according to family history. JAMA Netw Open 2020;3:e203959.
9. Umlai UI, Bangarusamy DK, Estivill X, et al. Genome sequencing data analysis for rare disease gene discovery. Brief Bioinform 2021 [Epub ahead of print].
10. Dewey FE, Murray MF, Overton JD, et al. Distribution and clinical impact of functional variants in 50,726 whole-

exome sequences from the DiscovEHR study. Science 2016;354:aaf6814.
11. Paila U, Chapman BA, Kirchner R, et al. GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol 2013;9:e1003153.
12. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020;581:434–443.
13. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018;46:D1062–D1067.
14. Liu X, Li C, Mou C, et al. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020;12:103.
15. Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019; 47:D886–D894.
16. Kelsen JR, Sullivan KE, Rabizadeh S, et al. North American Society for Pediatric Gastroenterology, Hepatology, and Nutrition position paper on the evaluation and management for patients with very early-onset inflammatory bowel disease. J Pediatr Gastroenterol Nutr 2020;70:389–403.
17. Dhingani N, Guo C, Pan J, et al. The E3 ubiquitin ligase UBR5 interacts with TTC7A and may be associated with very early onset inflammatory bowel disease. Sci Rep 2020;10:18648.
18. Wang L, Aschenbrenner D, Zeng Z, et al. Gain-of-function variants in SYK cause immune dysregulation and systemic inflammation in humans and mice. Nat Genet 2021; 53:500–510.
19. Requena D, Maffucci P, Bigio B, et al. CDG: an online server for detecting biologically closest disease-causing genes and its application to primary immunodeficiency. Front Immunol 2018;9:1340.
20. Tangye SG, Al-Herz W, Bousfiha A, et al. Human inborn errors of immunity: 2019 update on the classification from the International Union of Immunological Societies Expert Committee. J Clin Immunol 2020;40:24–64.
21. de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet 2017;49:256–261.
22. Liu JZ, van Sommeren S, 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.
23. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature 2020;578:527–539.
24. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. Nucleic Acids Res 2020;48:D682–D688.
25. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019; 47:D607–D613.
26. Taft LM, Evans RS, Shyu CR, et al. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. J Biomed Inform 2009; 42:356–364.
27. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. Nature 2014;508:469–476.
28. Giollo M, Jones DT, Carraro M, et al. Crohn disease risk prediction-best practices and pitfalls with exome data. Hum Mutat 2017;38:1193–1200.
29. Nambu R, Muise AM. Advanced understanding of monogenic inflammatory bowel disease. Front Pediatr 2020;8:618918.
30. Uhlig HH, Muise AM. Clinical genomics in inflammatory bowel disease. Trends Genet 2017;33:629–641.

Received August 10, 2021. Accepted November 8, 2021.

Correspondence:
Address correspondence to: Aleixo M. Muise, MD, PhD, The Hospital for Sick Children, 555 University Ave, Toronto, Ontario M5G 1X8, Canada. e-mail: aleixo.muise@sickkids.ca. Daniel J. Mulder, MD, PhD, Queen's University, 76 Stuart St, Kingston, Ontario K7L 2V7, Canada. e-mail: daniel.mulder@queensu.ca.

Acknowledgments:
The authors thank the patients and their families as well as healthy individuals for participating in this study. The authors thank Karoline Fiedler for assistance with patient-related materials. Bioinformatics analyses were supported in part by the Canadian Centre for Computational Genomics (C3G), part of the Genome Technology Platform (GTP) funded by Genome Canada through Genome Quebec and Ontario Genomics.

Authors' Contributions:
Daniel J. Mulder, Neil Warner, Sam Khalouei, and Aleixo M. Muise contributed to study concept and design. Claudia Gonzaga-Jauregui, Peter C. Church, Thomas D. Walters, Anne M. Griffiths, and Aleixo M. Muise contributed to acquisition of data. All the authors contributed to analysis and interpretation of data. Daniel J. Mulder, Sam Khalouei, Neil Warner, Amanda Ricciuto, and Aleixo M. Muise contributed to drafting of the manuscript. All the authors contributed to critical revision of the manuscript for important intellectual content.

Conflicts of Interest:
C.G.J. is a full-time employee of the Regeneron Genetics Center from Regeneron Pharmaceuticals, Inc and receives stock options as part of compensation. No other competing interests from any authors.

Funding:
A.M.M. is funded by the Leona M. and Harry B. Helmsley Charitable Trust, a Canada Research Chair (Tier 1) in Pediatric IBD, Canada Institutes of Health Research (CIHR) Foundation Grant, and NIDDK (RC2DK118640) Grant. A.M.G. holds the Northbridge Financial Corporation Chair in IBD at SickKids Hospital, University of Toronto. D.J.M. is funded by a CIHR-Canadian Association of Gastroenterology (CAG) Fellowship.

Ethical Statement:
The corresponding author, on behalf of all authors, jointly and severally, certifies that their institution has approved the protocol for any investigation involving humans or animals and that all experimentation was conducted in conformity with ethical and humane principles of research.

Data Transparency Statement:
To protect genetic information from identifying individual patients, individual participant data will not be shared. However, analytic methods and R code will be made available to other researchers on request. R code is available open access currently at https://github.com/DanJMulder/monoibdpriority.

Writing Assistance:
None.