Large-Scale Annotation of Genome/Exome Variants for Downstream Analysis: Building prediction models of Inflammatory Bowel Disease (IBD) Phenotypes

Minju Kim

Dr. Yuval Itan (Thesis Advisor)
Icahn School or Medicine at Mount Sinai | Masters in Biomedical Data Sciences

Table of Contents

Abstract

In this study, a robust human genome/exome variant annotation pipeline comprising 19 steps was developed. The pipeline sources information from 65 external databases. The pipeline yielded a total of 612 features at different levels, including Genomic, Gene, Transcript, Protein, Regulatory, Functional Prediction, and Population Frequency. Also, the annotation pipeline demonstrated high coverage and reliability, effectively annotating both small-scale datasets (324 variants) and large-scale datasets (~4 million variants). On average, more than 98% of the variants were annotated across all features, irrespective of the dataset. During the application of the pipeline to various datasets, one notable issue was the runtime efficiency. As the dataset size increased, the annotation process required a significantly longer time. This trend was observed in both the overall runtime and the per variant annotation time. Therefore, improving the runtime efficiency of the pipeline for large-scale datasets should be a focus for future enhancements. To assess the applicability of the annotation results for downstream analysis, I utilized two datasets to build machine learning models for predicting the phenotype of Inflammatory Bowel Disease (IBD). While the classifiers were limited by a small set of labeled data and require further improvements to achieve a desirable performance, their successful development illustrates the utility of the large-scale annotation pipeline.

Introduction

With the rapid advancements in sequencing technology, an increasing number of DNA and RNA sequences from various species, including humans, have been decoded and accumulated. This accumulation includes the sequencing of DNA from numerous individuals and patients, which holds great value for both biological and clinical applications. Therefore, it is crucial to accurately and comprehensively annotate these high-throughput sequencing data in order to conduct downstream analyses such as building machine learning models, elucidating disease pathogenicity, and developing personalized therapies [1].

Addressing this need, this thesis presents a comprehensive annotation approach that involves multiple steps to annotate human genome/exome variants at different levels, including gene, transcript, and protein functions, by utilizing various external sources. Additionally, using the annotated data, this thesis explores the process and results of creating and applying machine learning classifiers to demonstrate the clinical and biological applicability of the annotated data. Specifically, the classifiers were designed to predict the phenotypes of Inflammatory Bowel Disease (IBD), which encompasses a group of chronic inflammatory disorders affecting the gastrointestinal tract. Within IBD, two major phenotypes are known as Crohn's disease and Ulcerative colitis [2].

Methods

1. Data

In this study, three different variant datasets were prepared. The first dataset was obtained from the Human Gene Mutation Database (HGMD). From the professional version (released in 2021) of variants, disease-causing mutations (annotated with 'DM') were selected, resulting in a final subset of 324 variants (314 SNPs and 10 Indels). These variants were categorized into three classes: 'IBD' (253), 'Crohn's Disease' (57), and 'Ulcerative Colitis' (17) [3]. The purpose of collecting this dataset was to exemplify the utility of the annotation pipeline by using its output as features for a machine learning classifier developed to predict two major sub-types of IBD.

The other two datasets were obtained from the Bio*Me* Biobank of the Charles Bronfman Institute for Personalized Medicine at Mount Sinai Hospital. The first dataset consists of biallelic variants, totaling 3,948,623 variants, including 3,811,794 SNPs and 136,829 Indel records. The second dataset comprises multiallelic variants, with a total of 549,303 variants, including 491,931 SNPs and 57,372 Indels across the genome (**Table 1**).

Table 1. Information of the datasets

No	File Name	Number of Variants (SNPs + Indels)	Source	Purpose
1	HGMD_professional_2021.vcf	324 (314 + 10)	HGMD	Annotation, Building models
2	SINAI_BioMe_Biobank.biallelic.vcf	3,948,623 (3,811,794 + 136,829)	Bio*Me* Biobank	Annotation, Validation
3	SINAI_BioMe_Biobank.multiallelic.vcf	549,303 (491,931 + 57,372)	Bio*Me* Biobank	Annotation

2. Annotation

2.1. Annotation Pipeline Scheme

The annotation pipeline consists of a total of 19 steps (excluding Step 0, where the user inputs initial information). The entire pipeline is modularized into 7 stages, each containing a different number of steps: Stage 1 (4), Stage 2 (4), Stage 3 (4), Stage 4 (1), Stage 5 (1), Stage 6 (1), and Stage 7 (4) (**Fig. 1**).

Stage 1 consists of three steps: STEP1, STEP2, and STEP3. STEP1 performs Variant Effect Predictor (VEP) annotation (version.108), taking the initial VCF file as input. It yields 470 features at seven different levels sourced from 61 databases, including Ensembl, UniProtKB, gnomAD, etc [7]. STEP2 converts the output of STEP1 into a tab-separated format file (TSV). STEP3 adds Mutation Significance Cutoff (MSC) and Genetic Damage Index (GDI) annotations to the variants [8-10]. After STEP3, the annotation moves to Stage 2, which consists of four independent steps: STEP4_1, STEP4_2, STEP4_3, and STEP4_4. Each step accepts the output of STEP3 as input and can be executed simultaneously. STEP4_1 performs protein-level annotation by adding six features sourced from two databases. STEP4_2 adds annotation at the level of functional prediction, genomic level, resulting in eight features sourced from two databases. STEP4_3 adds eight new features at three different levels: functional prediction, protein, and transcript, sourced from three databases. STEP4_4 adds annotation across variants on gene expression in 108 tissues using binary figures sourced from a database, GTEx (**Sup. 1**) [13]. Continuing from STEP4_1, the output file is used as input for both Stage 3 and Stage 6. Stage 3, composed of 4 steps: STEP5, STEP6, STEP7, and STEP8, conducts annotations at the level of functional predictions related to protein structures and functions. These annotations are computed using the computational tools, IUPred and NetSurfP [11][12].

The output from STEP4_1 is also utilized as input in Stage 6, which contains only one step: STEP10. This stage is designed to map post-translational modifications based on protein names and positions. It is sourced from two databases: PhosphoSitePlus and UniProt [14][15] (**Fig. 1**). The five output files from Stage 3 (output file of STEP8), Stage 6 (output file of STEP10) and Stage 2 (output files of STEP4_2, STEP4_3, STEP4_4) are merged in the first step of Stage 7, which is STEP11 (**Fig.1**). After merging, the output is annotated to determine whether the genomic position belongs to a blacklist in STEP12. The blacklist refers to a set of genomic regions or coordinates that are considered problematic or should be excluded from downstream analysis [16]. Particularly, in the 'blacklist' column, each variant was annotated as either True or False as reference to a database, 'the blacklist of non-pathogenic variants' [16][17]. By excluding blacklisted regions, researchers can improve the accuracy and reliability of their analyses by avoiding potential biases or artifacts introduced by these problematic genomic regions. After being marked with blacklist results, the output is further annotated with information about gain-of-function (GoF), loss-of-function (LoF), and neutral variants sourced from a database, 'goflof' [18][19]. A GoF variant refers to a genetic alteration that results in an increased or abnormal function of the gene or protein it affects. This can involve enhanced activity, increased expression, or the acquisition of new functions. On the other hand, a LoF variant refers to a genetic alteration that disrupts or impairs the normal function of the gene or protein. LoF variants can lead to a reduction or complete loss of protein function, resulting in diminished or absent protein activity. Neutral variants are genetic alterations that do not have a significant impact on gene function or protein activities. By annotating the variants with these three features, useful information regarding identifying disease-causing variants and underlying molecular mechanisms of disease can be obtained [19]. The variants in the three datasets were annotated using the three features on a position-by-position basis. If a variant position existed in the database, the corresponding probability values were added to the respective row of that position. If a variant position did not have a match in the database, the field was left empty.

Upon the completion of Stage 7, the annotation progresses to the post-steps, which encompass three individual steps: STEP15, STEP16, and STEP17, dedicated to standard formatting procedures. During these steps, duplicated features and columns are removed, some features are renamed to avoid confusion, and empty columns of the features are dropped. Once the standard formatting is complete, the final results of the annotation are obtained as Comma-Separated Values (CSV) file (**Fig.1**).

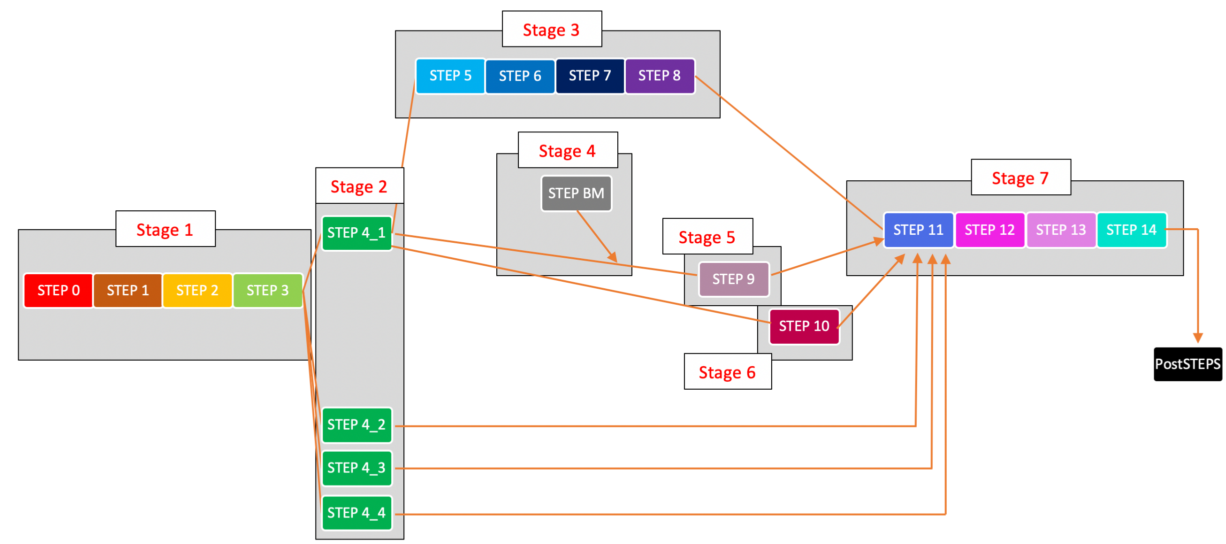


Fig. 1 Annotation Pipeline Scheme

2.2. Automation, Runtime, and Memory Usage

The extensive annotation pipeline was built using two programming languages: Shell scripts and Python (version 3.8.2). The entire workflow was automated using the Load Sharing Facility (LSF) scheduled job submission method. Specifically, each step was scheduled to be executed based on the completion of the previous step, as it requires the output result of the previous step as input. Therefore, the pipeline was designed to proceed to the next step only when the previous step was successfully completed without any errors. To conveniently track the job completion results, the pipeline was designed to generate Standard Error (stderr) and Standard Output (stdout) files within the directory where the output file(s) are generated.

Different from other steps, STEP4_1, STEP4_2, STEP4_3, and STEP4_4, which belong to Stage2, were designed to be executed concurrently by accepting the output of STEP3 as their inputs. This was done to save the overall running time of the pipeline. Stage4 and Stage5, which contain STEP BM (BioMart) [20] and STEP9, were excluded from this annotation work due to their high memory usage and long runtime.

The entire annotation process was performed on Minerva, a high-performance computer (HPC) at the Icahn School of Medicine at Mount Sinai. To execute the automated annotation, the initial information, including the working directory, input VCF file (or VCF.gz file), and project label, were typed in the STEP0.sh script. And then, the 'vep_all_the_way.lsf' script was executed using the 'bsub' command on the HPC, triggering the annotation of the input file through 19 steps **(Sup. 2)**.

Fast and accurate annotation is a crucial consideration. However, due to memory allocation limitations and HPC performance, determining the optimal number of nodes and memory usage was challenging. Instead, these parameters were empirically determined based on observation.

Information regarding the number of nodes, memory usage, and runtime has been recorded for future reference and improvements (**Table 4**).

2.3. Database Use, Data Types, and Levels of Annotations

The annotation pipeline sources information from various external databases, encompassing a total of 612 features derived from 65 different databases (**Sup. 1**). These features consist of diverse data types, including numeric (integer or float), ordinal, categorical, Boolean, and descriptive. Some features contain compound values, necessitating their separation or conversion into multiple components for downstream analyses. The annotation process incorporates multiple levels to annotate large-scale variants. These levels are categorized into seven distinct categories: Genomic Level, Gene Level, Transcript Level, Protein Level, Regulatory Level, Functional Prediction Level, and Population Frequency Level [7].

At the Genomic Level, the annotation provides information about the genomic location and fundamental characteristics of the variant within the genome. Example features of this level include variant classes and genomic location. Gene Level annotation focuses on the genes associated with the variants and their functional characteristics. It provides information such as biotype classification and gene-specific phenotypes. Transcript Level annotation describes the impact of variants on specific transcript isoforms. It includes details such as the precise location and molecular consequence of variants within the transcripts. Protein Level annotation entails features such as effected protein position, domains, and codons. This level describes the effects of variants on protein sequences, structures, and functions. Regulatory Level annotation provides insights into gene regulation and transcriptional control. Functional Prediction Level annotation offers information about the potential functional consequences of variants. Most features in this category provide numeric scores, while a few may provide categorical predictions as string values. Lastly, population Frequency Level annotation focuses on the prevalence of variants in different populations. It provides valuable insights into the frequency of variants and their potential population-specific effects (**Table 2**) [7].

Table 2. The seven categories of the levels of annotations [7]

No	Level	Description	Features
1	Genomic Level	the genomic location, and variant classes	CHROM, POS, REF, ALT, VARIANT_CLASS, etc.
2	Gene Level	the affected genes, and functional characteristics	SYMBOL, Feature_type, MANE, GENE_PHENO, Ensembl_ID_affected_gene, etc.
3	Transcript Level	the specific transcript affected by the variant and its consequences at the transcript level.	CDS_position, cDNA_position, EXON and INTRON HGVSc, etc.
4	Protein Level	the impact of variants on protein sequences, structures, and functions	Protein_position, SWISSPROT, TREMBL Protein_position Amino_acids, etc.
5	Regulatory Level	information on variant impacts on gene regulation and control	Regulatory, MOTIF_NAME, MOTIF_POS, HIGH_INF_POS MOTIF_SCORE_CHANGE, etc.
6	Functional Prediction Level	the potential functional consequences of variants	SIFT, PolyPhen, Condel, REVEL, CADd_PHRED, etc.
7	Population Frequency Level	the prevalence of variants in different populations	AF, gnomAD_AF, AMR_AF, EUR_AF, EAS_AF, etc.

3. **Downstream Application of the annotated datasets**

To build machine learning models, two fully annotated datasets were selected from the annotation pipeline: HGMD data and the biallelic dataset of Bio*Me* Biobank. In the HGMD data, variants labeled with 'IBD' were dropped as the classifier needed to be designed specifically for classifying IBD subtypes. The remaining 71 variants (57 labeled with 'Crohn's disease' and 17 labeled with 'Ulcerative Colitis') (**Sup. 8**) were used to construct the models using the sklearn library in Python [25]. Initially, the reference dataset was split into a training set and a testing set in an 8:2 ratio. The target variable was set as 'label', which contains the two classes of IBD phenotypes. The training set underwent preprocessing steps. Categorical features were converted to numerical values using one-hot encoding, numeric features were Z-score normalized and scaled between 0 and 1, and missing values were imputed. Additionally, oversampling using the 'SMOTE' module was performed to address the class imbalance, resulting in a balanced ratio of IBD phenotypes [26]. After completing the preprocessing steps, prediction models were built using six different classifiers: Random Forest, Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN). The parameters were determined through the process of hyperparameter tuning (**Sup. 7**).

Results

1. The Structure of Annotation Levels

The annotation pipeline was applied to three datasets of variants: HGMD, biallelic variants of Bio*Me* biobank, and multi-allelic variants from Bio*Me* Biobank. All of the datasets successfully passed the entire steps and resulted in three CSV files containing 612 features (HGMD contains a total of 614 with the features of 'acc_num'(accession number) and 'label' which were inserted before taking the file into the pipeline) (**Sup. 3**). Among the 612 features, almost 80% of them belong to three categories of annotation levels: Population Frequency, Functional Prediction, and Gene. The remaining 20% of features are distributed among the other four categories. Specifically, 34.5% of the annotated features belong to the population frequency category, with a count of 211. Functional prediction level follows with a count of 156, which accounts for 25.5% of all features. The gene level contains 119 features, representing almost 20% (19.4%), followed by genomic level (count: 49, 8%), transcript level (count: 39, 6.4%), protein level (count: 21, 3.4%), and regulatory level (count: 15, 2.5%). The remaining two features were classified as "ID" (**Fig. 2**) (**Sup. 4**).

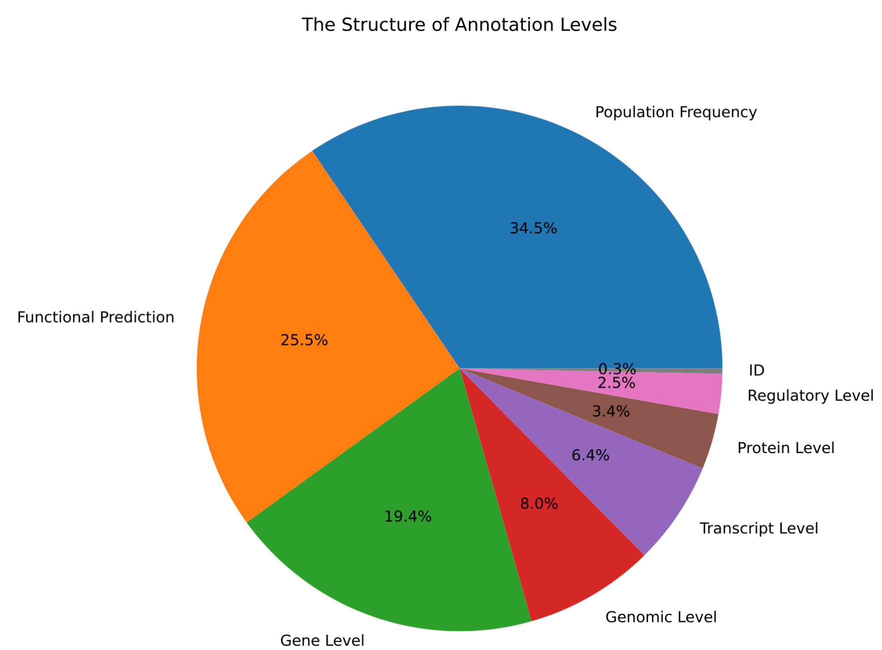


Fig.2 The Structure of Annotation Levels

2. Database Usage Breakdown

Throughout the annotation, a total of 65 databases were used. Among these databases, the Genome Aggregation Database (gnomAD) was the source of 145 features (23.7%). gnomAD provides population-specific allele frequencies for a wide range of genetic variants [21]. The Genotype-Tissue Expression Project (GTEx) was the second-largest source of features, with 111 features (18.1%) (**Fig. 3**). These features represent tissue-specific expression patterns using binary values. The three additional features from GTEx include TSSDistance (Transcription Start Site Distance), GTEx_V8_gene, and GTEx_V8_tissue. TSSDistance provides the distance between the variant and the nearest transcription start site (TSS), indicating the potential impact on gene regulation. GTEx_V8_gene provides information about genes associated with the variant and their expression patterns across different tissues. GTEx_V8_tissue represents tissue-level annotation from the GTEx project, providing information about the expression levels of genes in specific tissues based on the GTEx V8 dataset [13]. Exome Aggregation Consortium (ExAC) served as the source for 50 features (8.2%), providing variant frequencies and functional annotations for protein-coding regions of the genome (**Fig. 3**). Among these, 48 features provide population-specific frequencies, while two features (pLI_gene_value and LoFtool) fall under the functional prediction level. Both pLI_gene_value and LoFtool are used to assess the potential functional consequences of genetic variants, providing information about gene intolerance to loss-of-function mutations and the probability of a variant being a true loss-of-function event [22]. Ensembl, a comprehensive genome annotation database, sourced 41 features (6.7%) (**Fig. 3**). Ensembl provides various types of data (numeric, categorical, and descriptive) across six different annotation levels: Transcript, Gene, Genomic, Protein, and Functional Prediction. Consequence and IMPACT are two representative features sourced from Ensembl. The "Consequence" feature provides information about the predicted impact or effect of a genetic variant on the corresponding transcript or protein, describing the specific type of change introduced by the variant. The "IMPACT" feature indicates the potential impact of a genetic variant on the affected gene or protein, assigning different impact categories based on various criteria [7]. The Database for Non-Synonymous Functional Predictions (dbNSFP) [23] and UCSC Genome Browser [24] ranked fifth and sixth, sourcing 36 (5.9%) and 18 (3%) features, respectively (**Fig. 3**). dbNSFP provides both numeric and categorical data across four different levels: Functional Prediction, Genomic, Protein, and Transcript. The UCSC Genome Browser provides numeric features, including integers and float values, at two different levels: Genomic and Functional Prediction. The remaining 57 databases, although accounting for 34.5% as a whole, individually contribute to less than 3% of the sourced features (**Fig. 3**) (**Sup. 5**).

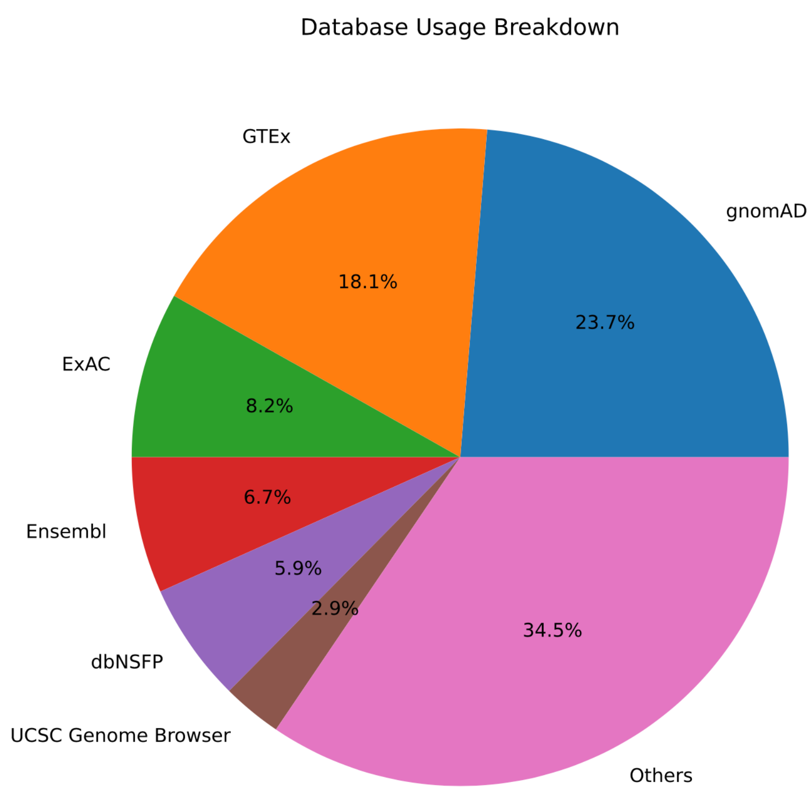


Fig 3. Database Usage Breakdown

3. Annotation assessment

After completing the extensive annotations, the reliability and comprehensiveness of the annotations were assessed by counting the number of variants annotated per feature and computing their percentages. Overall, all three datasets showed very high annotation coverage without missing information.

In the HGMD dataset, an average of 99.06% (±5.29%) of variants (324) were annotated without missing or empty information across 612 features. From the biallelic variants dataset of the BioMe Biobank, an average of 98.65% (±6.59%) of variants (3,948,623) were annotated for the 612 features. The multi-allelic variants also exhibited a high coverage trend, with 98.37% (±7.74%) of variants annotated across all features. In the annotation results of the HGMD dataset, over 90% of the variants were annotated across 589 features, and 575 of them were fully annotated all 324 variants (100%) (**Fig. 4A**). The features with the lowest coverage percentage of variants were the four features related to the probability of GoF/LoF/Neutral variants (53.70%), but none of the features covered less than 50% of the variants (**Sup. 6A**). In the remaining results of the two Bio*Me* Biobank datasets, 582 features covered more than 90% of the variants (3,948,623 and 549,303 each) (**Fig. 4B and 4C**). Additionally, 572 features in both datasets fully annotated all the variants (100%) (**Sup. 6B and 6C**). From the biallelic variants dataset, the features "Condel_pred" and "Condel_score" had the least coverage, annotating only 48.36% of the variants. Four features, including "PolyPhen_pred" (48.63%), "PolyPhen_score" (48.63%), and the two features aforementioned annotated less than half of the variants in the dataset. Similarly, in the multiallelic dataset, the same two features, "Condel_pred" and "Condel_score," mentioned in the biallelic results, recorded the lowest coverage of variants at 41.93%. Additionally, a total of 8 features annotated less than half of the variants. For example, "PolyPhen_pred" and "PolyPhen_score" covered only 42.44% of the variants, while the four features related to GoF/LoF/Neutral probabilities covered 44.27% of the variants (**Table 3**) (**Sup. 6**).

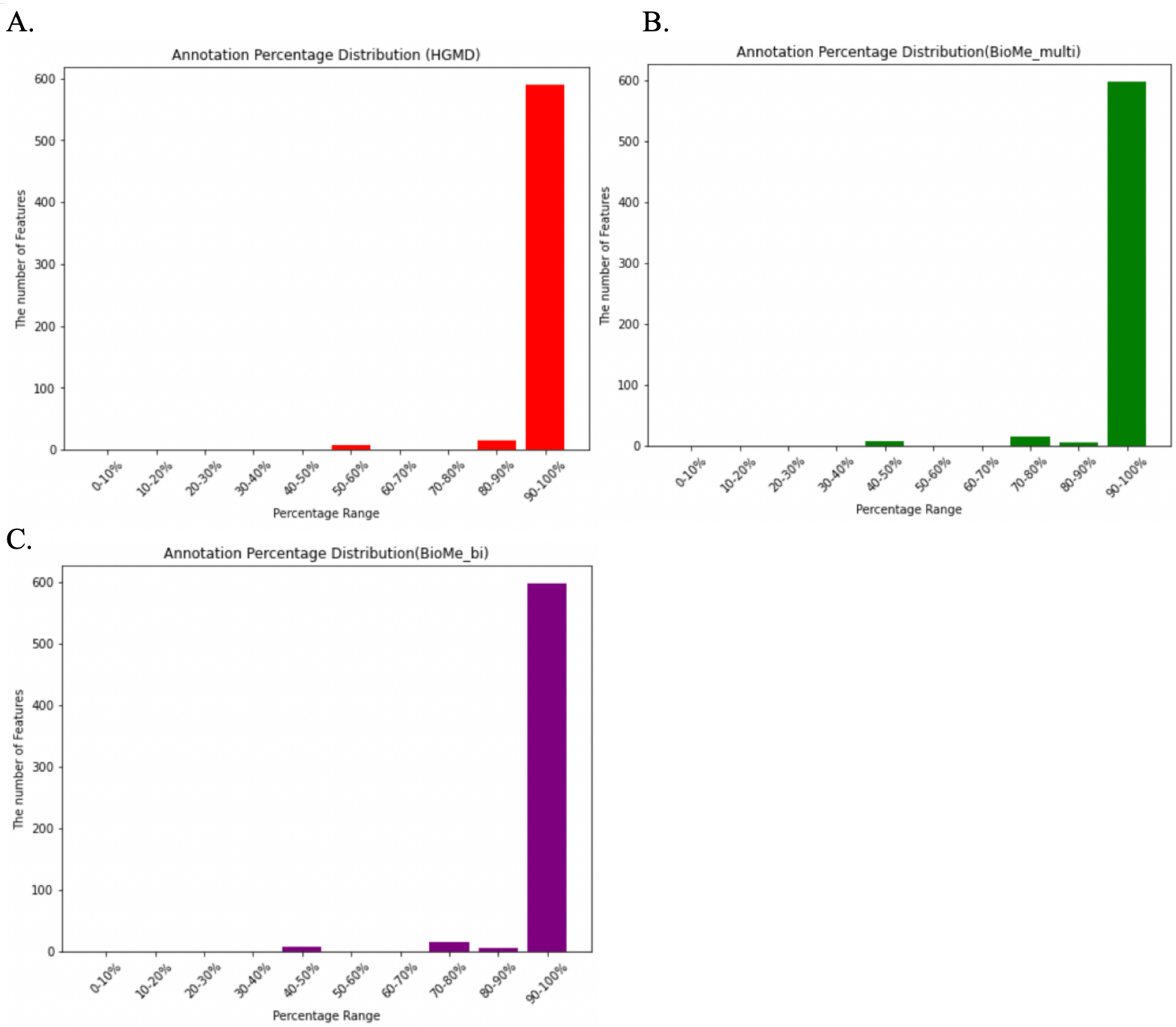


Fig 4. Annotation Percentage Distribution
(A. HGMD, B. Bio*Me* Biobank multi-allelic variants data, C. Bio*Me* Biobank biallelic variants data)

Table 3. The Features with Low Coverage

No	Feature	Description	Source
1	PolyPhen_pred	PolyPhen-2 prediction of the functional impact of the variant	dbNSFP/PolyPhen-2
2	PolyPhen_score	PolyPhen-2 score of the functional impact of the variant	dbNSFP/PolyPhen-2
3	Condel_pred	Combined annotation-dependent depletion prediction	dbNSFP/Condel
4	Condel_score	Combined annotation-dependent depletion score	dbNSFP/Condel
5	LoGoFunc_neutral	Predicted functional impact of the variant on protein structure and function based on a neutral model	goflof
6	LoGoFunc_GOF	Predicted functional impact of the variant on protein structure and function based on a gain-of-function model	goflof
7	LoGoFunc_LOF	Predicted functional impact of the variant on protein structure and function based on a loss-of-function model	goflof
8	GOF/LOF/NEU	Indicates whether the variant is predicted to have a gain-of-function, loss-of-function, or neutral impact on protein function	goflof

4. Runtime and memory use

Due to time limitations and the availability of memory on personal computers or HPC systems, it is crucial to find optimal conditions that satisfy both variables. Although it was not possible to achieve the ideal balance of speed and memory, reasonable setups were determined based on empirical observations. In this section, the number of cores, allocated memory, and the time taken for each step and the entire pipeline are presented. For the complete annotation of the 324 variants in the HGMD dataset across 19 steps, it took a total of 652 seconds (12 minutes), and 0.4969 seconds per variant. The multiallelic dataset, consisting of 549,303 variants, was annotated in 19 steps, taking approximately 58,972 seconds or approximately 16 hours in total. On average, it took 9.3146 seconds to annotate each variant. The biallelic dataset, which included 3,948,623 variants, required a total of 478,262 seconds, approximately 133 hours (5.5days), for full annotation. The average time taken per variant in this dataset was 8.2562 seconds. The same number of cores and memory allocation were applied to all three datasets when submitting the jobs to the HPC (Table 4).

Table 4. The number of cores, Requested Memory, and Runtime of the data annotation per step

STEP	**1**	**2**	**3**	**4_1**	**4_2**	**4_3**	**4_4**	**5**	**6**	**7**	**8**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**Total**
Node (core #)	20	20	20	18	15	15	15	15	15	15	15	15	15	15	15	15	20	1	10	289
Memory (MB)	31.5K	10K	10K	70K	70K	70K	70K	70K	70K	70K	70K	70K	70K	70K	70K	70K	10K	10K	50K	1.031M
Runtime(sec) (HGMD)	192	2	3	3	5	3	9	4	7	2	2	5	2	2	404	1	3	1	2	652
Runtime (sec) (Biobank_multi)	36857	12684	216	95	1765	82	1827	130	3234	87	85	202	561	238	610	12	150	10	127	58,972
Runtime (sec) (Biobank_bi)	373975	51772	1236	697	12781	629	12600	347	2147	659	667	1549	4116	1787	11389	96	1044	85	686	478,262

5. Downstream analysis Results

Despite building the models successfully, their overall performances could be improving. Based on the AUC-ROC curve, only the Decision Tree and KNN models achieved scores greater than 0.5, with both recording an AUC score of 0.54. The other four classifiers performed worse, with scores of 0.29 (Random Forest), 0.33 (Naïve Bayes), 0.44 (Logistic Regression), and 0.19 (SVC). The Decision Tree classifier demonstrated a relatively stronger performance compared to the other five classifiers, achieving an Average Precision (AP) score of 0.36, while the remaining classifiers scored less than 0.2 on the precision-recall curve (Table 5) (Sup. 9).

Table 5. AUC scores and AP values of the six classifiers

Name	**AUC**	**AP**
Random Forest	0.29	0.14
Naïve Bayes	0.33	0.10
Decision Tree	0.54	0.36
Logistic Regression	0.44	0.17
SVC	0.19	0.12
KNN	0.54	0.17

After building the models, they were saved using the 'joblib' library [27] and applied to a large-scale dataset: the biallelic variants of Bio*Me* Biobank. The dataset was not labeled with any IBD phenotypes. Predictions were made using all six models, and the results showed a skewed prediction towards Crohn's disease. Notably, the Random Forest classifier exhibited an extreme bias towards predicting Crohn's disease. The other four classifiers also displayed a bias towards Crohn's disease, except for KNN. Only the KNN algorithm yielded slightly higher predictions for Ulcerative Colitis (**Fig. 5**).

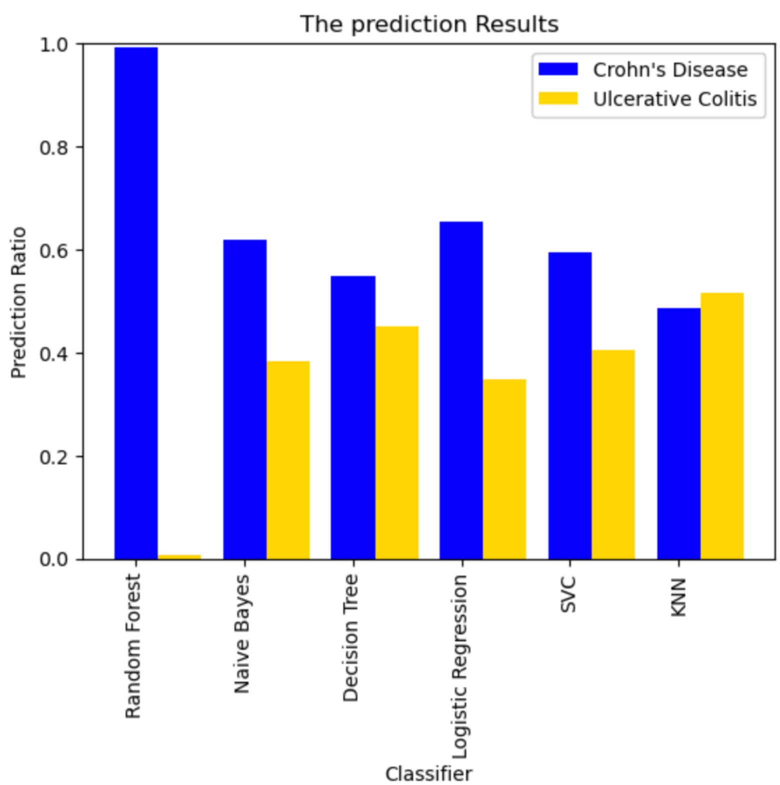


Fig. 5 Prediction Results of the six classifiers

Discussion

The variant annotation process was successfully executed without any failures, demonstrating the reliability of the pipeline. The stepwise approach proved effective even when dealing with large-scale datasets, providing a wealth of information across the seven different levels of annotations. However, it is crucial to address several limitations and issues associated with this approach. Firstly, there is a need to balance the distribution of annotations across the seven levels. Currently, there is an imbalance, with a significant concentration of annotations in three levels: population frequency, functional prediction, and gene, accounting for 80% of the total annotations (**Fig. 2**). The remaining five levels represent only 20% of the annotations. While this distribution can be highly valuable for research focused on population-specific genetic diseases, or population genetics topics, it could render many features irrelevant for other research purposes. Adjusting the number and type of features based on the research objectives, target diseases, or genetic topics is essential to ensure meaningful annotations. Another critical issue is the lengthy annotation time, especially when working with large-scale datasets. Annotating a single variant in the HGMD dataset, which contained 423 variants, took a mere 0.4969 seconds. However, the duration significantly increased to 9.3 and 8.2 seconds per variant for the multiallelic and biallelic datasets, respectively. Notably, annotating approximately 4 million variants from the Bio*Me* Biobank biallelic data required a substantial amount of time, taking over 5 and a half days (133 hours). To enhance the efficiency of annotating large-scale variant datasets, alternative methods to reduce memory usage and improve processing speed should be explored. Lastly, to assess the applicability of the annotation results, further downstream analyses should be conducted. These analyses may include feature importance analysis, and the development of machine learning models based on the annotated data. As an example, I have included a case study involving the prediction of Inflammatory Bowel Disease (IBD) phenotypes through the development of machine learning models. This case study showcases the practical application of the annotation results. Although the classifiers successfully predicted the phenotypes of the unlabeled large-scale dataset, their performances were not deemed good enough and reliable (**Table 5**). This could possibly be attributed to the small size of the reference data used for building the models and its biased label distribution.

Conclusion

A large-scale annotation of human genomic variant data was conducted using a developed annotation pipeline consisting of 19 steps. The annotation process resulted in the comprehensive annotation of all three datasets using 612 features sourced from 65 external databases across 7 different levels. The high coverage of 98% or more indicates the excellent quality of the annotation. However, a notable challenge encountered during the annotation process was the time required to annotate large-scale datasets. Annotating the entire dataset of approximately 4 million variants took more than 5 days to complete. This highlights the need for optimizing the pipeline's efficiency to reduce the annotation duration for such large-scale datasets. To evaluate the applicability of the annotation results, a downstream analysis was conducted by constructing machine learning models. These models were successfully built and used to make predictions on an unlabeled large-scale dataset. However, it is important to note that the performance of these models needs improvement to enhance the reliability of the prediction results. Future efforts should focus on enhancing the performance of the machine learning models by refining the model architecture, optimizing parameters, and considering alternative algorithms. Additionally, the reference dataset should be enriched by including a greater number of variants with a balanced distribution of phenotypes. This enrichment will help the models to be exposed to a more diverse range of examples. By addressing these challenges, it will be possible to further leverage the annotation results for more accurate and reliable downstream analyses.

Reference

[1] Jaravine, Victor, et al. "Annotation of Human Exome Gene Variants with Consensus Pathogenicity." *Genes*, vol. 11, no. 9, Sept. 2020, p. 1076. *Crossref*, https://doi.org/10.3390/genes11091076.

[2] Montero-Meléndez T, Llor X, García-Planella E, Perretti M, Suárez A (2013) Identification of Novel Predictor Classifiers for Inflammatory Bowel Disease by Gene Expression Profiling. PLoS ONE 8(10): e76235. https://doi.org/10.1371/journal.pone.0076235

[3] Stenson, P.D. *et al.* (2017) "The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies," *Human Genetics*, 136(6), pp. 665–677. Available at: https://doi.org/10.1007/s00439-017-1779-6.

[4] The Genome Reference Consortium. "Genome Reference Consortium Human Genome Assembly GRCh38." *National Center for Biotechnology Information*, 2013, www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/

[5]Smith, John. "Sample_Variant_Calls.vcf." VCF. Version 1.2. Database of Genomic Variants, 2022. http://www.example.com/sample_variants.vcf

[6] "Sema4 WES." Mount Sinai Intellectual Property Management. Accessed [date accessed], https://mssm-ipm.atlassian.net/wiki/spaces/IPM/pages/1456439317/Sema4+WES.

[7] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek and Fiona Cunningham. "The Ensembl Variant Effect Predictor."Genome Biology (2016): 17:122.

[8] "Itan Lab" The Charles Bronfman Institute for Personalized Medicine at Mount Sinai http://pec630.rockefeller.edu:8080/MSC/

[9] "Itan Lab" The Charles Bronfman Institute for Personalized Medicine at Mount Sinai https://itanlab.org/wp-content/uploads/2020/06/MSC_v1.6.zip

[10] "Itan Lab" The Charles Bronfman Institute for Personalized Medicine at Mount Sinai http://pec630.rockefeller.edu:8080/GDI/

[11] IUPred: Dosztányi, Zsuzsanna, et al. "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." Bioinformatics, vol. 21, no. 16, 2005, pp. 3433-3434. EBSCOhost, doi:10.1093/bioinformatics/bti541.

[12] NetSurfP: Petersen, Bent, et al. "Prediction of Protein Secondary Structure at High Accuracy Using a Combination of Neural Networks and Structural Templates." BMC Bioinformatics, vol. 11, no. 1, 2010, p. 159. EBSCOhost, doi:10.1186/1471-2105-11-159.

[13] The GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project." *Nature Genetics*, vol. 45, no. 6, 2013, pp. 580-585. *EBSCOhost*, doi:10.1038/ng.2653.

[14] PhosphoSitePlus: Hornbeck, Peter V., et al. "PhosphoSitePlus, 2014: mutations, PTMs and recalibrations." *Nucleic Acids Research*, vol. 43, no. D1, 2015, pp. D512-D520. *EBSCOhost*, doi:10.1093/nar/gku1267.

[15] UniProt: The UniProt Consortium. "UniProt: the universal protein knowledgebase." *Nucleic Acids Research*, vol. 45, no. D1, 2017, pp. D158-D169. *EBSCOhost*, doi:10.1093/nar/gkw1099.

[16] Maffucci, Patrick, et al. "Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 13, 2019, pp. 6223-6231. *EBSCOhost*, doi:10.1073/pnas.1818259116.

[17] "Itan Lab" The Charles Bronfman Institute for Personalized Medicine at Mount Sinai http://pec630.rockefeller.edu:8080/BL/

[18] "Itan Lab" The Charles Bronfman Institute for Personalized Medicine at Mount Sinai https://itanlab.shinyapps.io/goflof/

[19] Bayrak, Cigdem Sevim, et al. "Identification of discriminative gene-level and protein-level features associated with pathogenic gain-of-function and loss-of-function variants." *The American Journal of Human Genetics*, vol. 108, no. 12, 2021, pp. 2301-2318. *EBSCOhost*, doi:10.1016/j.ajhg.2021.10.007.

[20] Durinck S, Spellman P, Birney E, Huber W (2009). "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt." Nature Protocols, 4, 1184–1191.

[21] Chen, S.*, Francioli, L. C.*, Goodrich, J. K., Collins, R. L., Wang, Q., Alföldi, J., Watts, N. A., Vittal, C., Gauthier, L. D., Poterba, T., Wilson, M. W., Tarasova, Y., Phu, W., Yohannes, M. T., Koenig, Z., Farjoun, Y., Banks, E., Donnelly, S., Gabriel, S., Gupta, N., Ferriera, S., Tolonen, C., Novod, S., Bergelson, L., Roazen, D., Ruano-Rubio, V., Covarrubias, M., Llanwarne, C., Petrillo, N., Wade, G., Jeandet, T., Munshi, R., Tibbetts, K., gnomAD Project Consortium, O'Donnell-Luria, A., Solomonson, M., Seed, C., Martin, A. R., Talkowski, M. E., Rehm, H. L., Daly, M. J., Tiao, G., Neale, B. M.†, MacArthur, D. G.† & Karczewski, K. J. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv 2022.03.20.485034 (2022). https://doi.org/10.1101/2022.03.20.485034

[22] Lek, M., Karczewski, K. J.*, Minikel, E. V.*, Samocha, K. E.*, Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., … Daly, M. J., MacArthur, D. G. & Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016). https://doi.org/10.1038/nature19057

[23] Liu X, Li C, Mou C, Dong Y, and Tu Y. 2020. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Medicine. 12:103.

[24] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002 Jun;12(6):996-1006.

[25] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[26] SMOTE. Imbalanced-learn, 2021, imbalanced-learn.org/stable/index.html.

[27] joblib. INRIA, 2021, joblib.readthedocs.io/en/latest/.

Acknowledgements

I would like to acknowledge the following individuals for their enthusiastic support throughout the completion of this thesis: ‘Large-Scale Annotation of Genome/Exome Variants for Downstream Analyses: Building prediction models of Inflammatory Bowel Disease (IBD) Phenotypes’

First and foremost, I express my deepest gratitude to my thesis advisor, Dr. Yuval Itan, for his guidance, expertise, and continuous support. His insightful feedback and mentorship have been instrumental in shaping the direction of this research. I would also like to extend my gratitude to the members of Itan Lab: David Stein and Meltem E. Kars, for their valuable insights, constructive suggestions, computational instructions that have greatly enriched this study.

Supplementary Materials

Supplementary Material 1 (Sup. 1). Feature Information about source database example value, and Link

1. Database_Sources

2. Feature explanation

Supplementary Material 2 (Sup. 2). Shell and Python scripts for building a pipeline

Variant Annotation Pipeline scripts

Supplementary Material 3 (Sup. 3). Annotation Results (csv files)

1. HGMD_professional_2021

2. BioMe_Biobank_biallelic

3. BioMe_Biobank_multi-allelic

Supplementary Material 4 (Sup. 4). Annotation Level Count

Level	Count
Population Frequency	211
Functional Prediction	156
Gene Level	119
Genomic Level	49
Transcript Level	39
Protein Level	21
Regulatory Level	15
ID	2
Total	612

Supplementary Material 5 (Sup. 5). Database Usage Breakdown (Count)

Database	Features
gnomAD	145
GTEx	111
ExAC	50
Ensembl	41

dbNSFP	36
UCSC Genome Browser	18
1000 Genomes	13
ClinVar	10
FATHMM	10
CADD	8
UniProtKB	8
PolyPhen-2	7
BayesDel	6
Eigen	6
Aloft	6
MutationAssessor	5
ENCODE	5
MutPred	5
MutationTaster	5
PhosphoSitePlus, UniProt	5
LRT	4
snpEff	4
ESP6500	4
UK10K	4
Variant Call Format (VCF)	4
goflof	4
SIFT	4
ANNOVAR	4
REVEL	3
M-CAP	3
LIST-S2	3
HUVEC	3
H1-hESC	3
PROVEAN	3
GM12878	3
GERP++	3
PrimateAI	3
SIFT4G	3
SiPhy	3
MaxEntScan	3
DEOGEN2	3
ClinPred	3
VEST4	2

Various databases (e.g., NCBI Gene)	2
dbNSFP/PolyPhen-2	2
MVP	2
dbNSFP/Condel	2
MPC	2
ALSPAC	2
GenoCanyon	2
GENCODE	2
Mastermind	2
DANN	2
APPRIS	1
Neanderthal Genome Project	1
Blacklist	1
UniProt	1
DisGeNET	1
Altai Neandertal	1
dbNSFP/dbSNP	1
GeneSplicer	1
InterPro	1
COSMIC	1
HPA	1
CAGI	1
NHGRI-EBI GWAS Catalog	1
PubMed	1
JASPAR	1
Denisova	1
Total	612

Supplementary Material 6 (Sup. 6). Feature Coverage

Sup 6A. HGMD

Sup 6B. Biallelic

Sup 6C. Multi-allelic

Supplementary Material 7 (Sup. 7). The parameters used

1. Random Forest: n_estimators=100, max_depth=10, random_state=42

2. Naïve Baeysian: Not applicable
3. Decision Tree: max_depth=5, random_state=42
4. Logistic Regression: C=1.0, solver='liblinear', random_state=42
5. Support Vector Machine: probability=True, C=1.0, kernel='rbf', gamma='scale', random_state=42
6. K-Nearest Neighbors: n_neighbors=5, p=2, algorithm='auto', leaf_size=30

Supplementary Material 8 (Sup. 8)

❶ 71 Variants from HGMD (2021)

CHROM	POS	REF	ALT	SYMBOL	Ensembl_ID_affected_Gene	rs_dbSNP151	acc_num	label
1	183563302	G	A	NCF2	ENSG00000116701	rs13306575	CM992363	Crohns
1	234607721	T	G	IRF2BP2	ENSG00000168264	rs138385624	CM2211702	Crohns
2	47046329	A	G	TTC7A	ENSG00000068724	rs139010200	CM137039	UC
2	47050043	T	C	TTC7A	ENSG00000068724	rs149602485	CM137040	UC
2	97725241	C	A	ZAP70	ENSG00000115085	.	CM2023215	Crohns
2	203870655	A	G	CTLA4	ENSG00000163599	.	CM153325	Crohns
2	241801858	C	T	GAL3ST2	ENSG00000154252	rs372108744	CM1412990	Crohns
3	30674229	G	A	TGFBR2	ENSG00000163513	rs104893816	CM052921	UC
3	30691474	G	A	TGFBR2	ENSG00000163513	.	CM139810	Crohns
3	30691478	G	A	TGFBR2	ENSG00000163513	rs104893815	CM050762	UC
3	48574261	C	G	COL7A1	ENSG00000114270	.	CS993006	Crohns
6	25661628	G	A	SCGN	ENSG00000079689	rs376721140	CM1915957	UC
6	31810299	T	A	HSPA1L	ENSG00000204390	rs566393477	CM170943	Crohns
6	31811143	G	A	HSPA1L	ENSG00000204390	rs1460318497	CM170942	UC
6	31811171	C	T	HSPA1L	ENSG00000204390	rs34620296	CM170945	Crohns
6	31811173	G	A	HSPA1L	ENSG00000204390	rs139868987	CM170946	Crohns
6	31811744	C	T	HSPA1L	ENSG00000204390	rs368138379	CM170944	UC
6	137877178	TAAAAG	T	TNFAIP3	ENSG00000118503	.	CD2012689	Crohns
7	74777318	C	T	NCF1	ENSG00000158517	rs782800778	CM1612167	Crohns
7	74783529	G	A	NCF1	ENSG00000158517	rs145360423	CM021647	Crohns
9	99144816	G	T	TGFBR1	ENSG00000106799	.	CM064323	Crohns
10	23440184	C	T	OTUD1	ENSG00000165312	.	CM1812286	UC
11	36575160	C	T	RAG1	ENSG00000166349	rs755059628	CM2019041	Crohns
11	117989458	T	C	IL10RA	ENSG00000110324	rs1343534194	CM1415113	Crohns
11	117989504	C	T	IL10RA	ENSG00000110324	rs137853580	CM098175	Crohns
11	117989525	A	G	IL10RA	ENSG00000110324	.	CM137114	Crohns
11	117989554	C	T	IL10RA	ENSG00000110324	rs368287711	CM127465	Crohns
11	117989603	G	A	IL10RA	ENSG00000110324	rs199989396	CM134627	Crohns
11	117993343	A	G	IL10RA	ENSG00000110324	rs1027503096	CM175770	Crohns

11	117993410	G	A	IL10RA	ENSG00000110324	.	CS156892	Crohns
11	117994151	T	C	IL10RA	ENSG00000110324	.	CS135619	UC
12	52056030	C	T	NR4A1	ENSG00000123358	rs1255104785	CM2128603	Crohns
12	52898860	G	A	KRT8	ENSG00000170421	rs62636489	CM077755	UC
13	43883879	T	C	LACC1	ENSG00000179630	rs730880295	CM1410012	Crohns
13	108209987	C	G	LIG4	ENSG00000174405	.	CM2023216	Crohns
14	34996744	C	T	SRP54	ENSG00000100883	.	CM2023214	Crohns
14	94378610	C	T	SERPINA1	ENSG00000197249	rs28929474	CM830003	Crohns
15	77037132	G	C	TSPAN3	ENSG00000140391	.	CM154606	UC
16	50699517	C	T	NOD2	ENSG00000167207	.	CM2129175	UC
16	50699527	G	T	NOD2	ENSG00000167207	rs104895487	CM082979	Crohns
16	50699655	C	G	NOD2	ENSG00000167207	rs34936594	CM1926598	Crohns
16	50710976	G	A	NOD2	ENSG00000167207	rs104895488	CM082983	Crohns
16	50711057	C	G	NOD2	ENSG00000167207	rs104895476	CM050186	Crohns
16	50711101	C	T	NOD2	ENSG00000167207	rs150078153	CM2129173	UC
16	50711467	A	C	NOD2	ENSG00000167207	rs368316739	CM2129174	Crohns
16	50712091	C	T	NOD2	ENSG00000167207	rs104895489	CM082980	Crohns
16	50712162	G	A	NOD2	ENSG00000167207	.	CM2129171	Crohns
16	50712168	C	T	NOD2	ENSG00000167207	rs749720540	CM2129172	Crohns
16	50722626	T	C	NOD2	ENSG00000167207	rs104895490	CM082982	Crohns
16	50729881	T	G	NOD2	ENSG00000167207	.	CM198083	Crohns
16	50731751	C	T	NOD2	ENSG00000167207	rs104895491	CM082981	Crohns
16	88646753	A	G	CYBA	ENSG00000051523	.	CS101778	Crohns
17	44075031	C	T	G6PC3	ENSG00000141349	rs911423195	CM2023217	Crohns
19	53810605	G	A	NLRP12	ENSG00000142405	rs199881207	CM112823	Crohns
20	63693247	C	T	TNFRSF6B	ENSG00000243509	.	CM132873	UC
21	33288194	G	A	IL10RB	ENSG00000243646	.	CM175775	Crohns
22	36937930	GC	G	CSF2RB	ENSG00000100368	.	CD1613104	Crohns
X	123885663	A	G	XIAP	ENSG00000101966	.	CM171029	Crohns
X	123885777	G	T	XIAP	ENSG00000101966	rs775237858	CM1412043	Crohns
X	123885925	CA	C	XIAP	ENSG00000101966	.	CD207065	Crohns
X	123885957	G	T	XIAP	ENSG00000101966	.	CM138815	Crohns
X	123885993	C	T	XIAP	ENSG00000101966	.	CM2130297	Crohns
X	123886030	G	A	XIAP	ENSG00000101966	rs368343771	CM187544	Crohns
X	123886162	TA	T	XIAP	ENSG00000101966	.	CD2130298	Crohns
X	123886269	TGTG	T	XIAP	ENSG00000101966	.	CD2015739	Crohns
X	123886326	C	T	XIAP	ENSG00000101966	.	CM111971	Crohns
X	123886360	G	A	XIAP	ENSG00000101966	rs368511826	CM187541	Crohns
X	123888631	A	C	XIAP	ENSG00000101966	.	CM1412044	Crohns
X	123888709	G	A	XIAP	ENSG00000101966	.	CM1412045	Crohns
X	123888710	G	A	XIAP	ENSG00000101966	.	CM1618307	Crohns

X	123891304	TGAG	T	XIAP	ENSG00000101966	.	CD120268	Crohns

❷ Bio*Me* Biobank data
The dataset was provided from Sema4 Data in The Bio*Me* Biobank

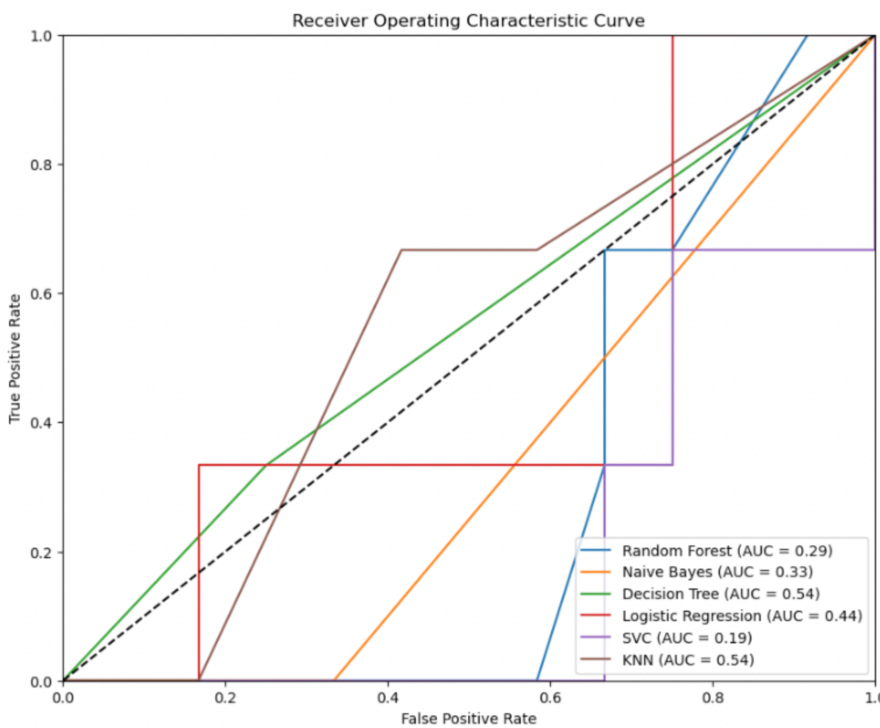
Exome Dataset Directory on the HPC (Minerva):

/sc/private/regen/data/Regeneron/SINAI_Freeze_Two_pVCF/data/pVCF/QC_passed/freeze2-ontarget/biallelic

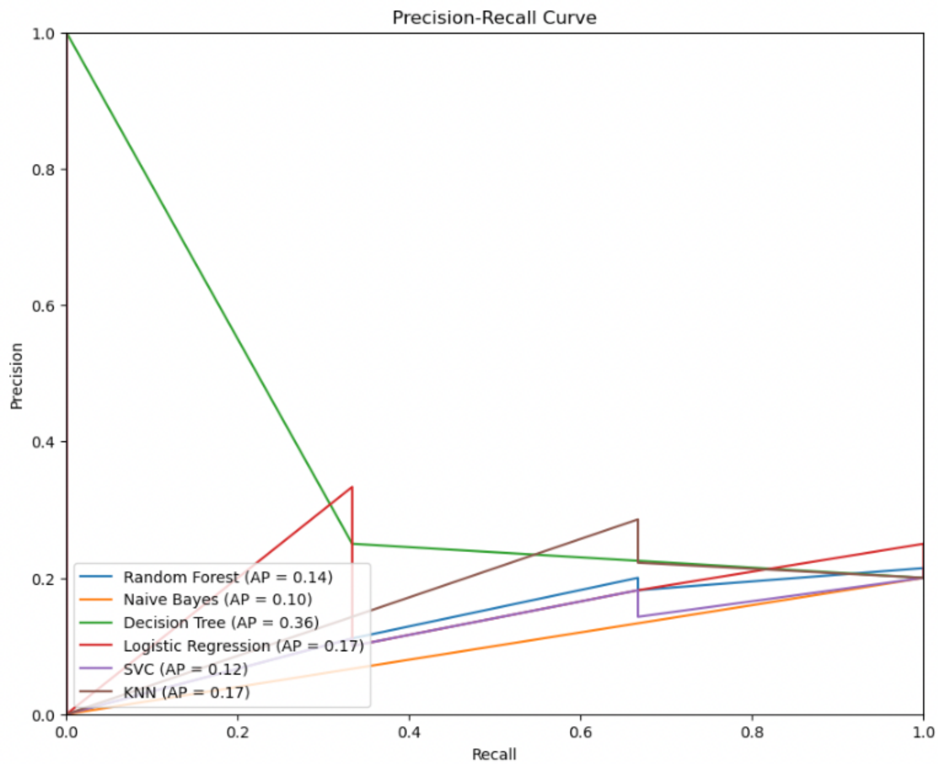
/sc/private/regen/data/Regeneron/SINAI_Freeze_Two_pVCF/data/pVCF/QC_passed/freeze2-ontarget/multiallelic.normalized

Supplementary Material 9 (Sup. 9)

A.



B.



Link:

Thesis_Supplementary (Data, Images, Statistics, and Python and Shell Scripts)