UNIVERSITY OF CALIFORNIA

Los Angeles

Developing Machine Learning and Statistical Methods for the Analysis of Genetics and Genomics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Human Genetics

by

Jiajin Li

2021

ABSTRACT OF THE DISSERTATION

Developing machine learning and statistical methods

for the analysis of genetics and genomics

by

Jiajin Li

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2021

Professor Matteo Pellegrini, Chair

With the development of next-generation sequencing technologies, we can detect numerous genetic variants associated with many diseases or complex traits over the past decades. Genome-wide association studies (GWAS) have been one of the most effective methods to identify those variants. It discovers disease-associated variants by comparing the genetic information between controls and cases. This approach is simple and effective and has been used by many studies. Before performing GWAS, we need to detect the genetic variants of the sample population. A subset of these variants, however, may have poor sequencing quality due to limitations in NGS or variant callers. In genetic studies that analyze a large number of sequenced individuals, it is critical to detect and remove those variants with poor quality as they may cause spurious findings. Here, I will present ForestQC, an efficient statistical tool for performing quality control on variants identified from NGS data by combining a traditional filtering approach and a machine learning approach, which outperforms widely used methods by considerably improving the quality of variants to be included in the analysis.

Once this association is identified, the next step is to understand the genetic mechanism of rare variants on how the variants influence diseases, especially whether or how they regulate gene expression as they may affect diseases through gene regulation. However, it is challenging to identify the regulatory effects of rare variants because it often requires large sample sizes and the existing statistical approaches are not optimized for it. To improve statistical power, I will introduce a new approach, LRT-q, based on a likelihood ratio test that combines effects of multiple rare variants in a nonlinear manner and has higher power than previous approaches. I apply LRT-q to the GTEx dataset and find many novel biological insights.

Recent studies have shown that omics data can be used for automatic disease diagnosis with machine learning algorithms. I will introduce an accurate and automated machine learning pipeline for the diagnosis of atopic dermatitis (AD) based on transcriptome and microbiota data. I will demonstrate that this classifier can accurately differentiate subjects with AD and healthy individuals. It also identifies a set of genes and microorganisms that are predictive for AD. I will show that they are directly or indirectly associated with AD.

The dissertation of Jiajin Li is approved.

Jae Hoon Sul

Eleazar Eskin

Sriram Sankararaman

Valerie A. Arboleda

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2021

This work is dedicated to God, my family, friends, and people I love.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

I would like to acknowledge help and support from many people including my family, friends, coworkers, and advisors during my graduate years. First, I would like to thank my mother, Dongmei, for her encouragement and love from China. I also want to thank my father for being so supportive while I was in this degree. I also want to thank my extended family, God Almighty. He has been my strong pillar, my source of strength, confidence, inspiration, wisdom, knowledge and understanding. I had a great time with my friends in our local church, CBCWLA, during these years. They are like my second family here in Los Angeles. Without them, I would not have finished this dissertation.

Second, I would like to thank my advisor Dr. Jae Hoon Sul. He taught me how to do research in statistical genetics, write papers, and everything that I needed to succeed throughout the degree. In Sul's lab, I developed strong working skills in statistics and machine learning. He also gave me valuable advice on my career. I also thank Prof. Matteo Pellegrini for being the chair of my doctoral committee and advising me during my rotation in the first year. I thank Prof. Sriram Sankararaman for teaching me machine learning principles as well as offering me a teaching assistant position. And I thank my committee members, Prof. Eleazar Eskin, Prof. Sriram Sankararaman, and Prof. Valerie A. Arboleda who provided me with valuable feedback and suggestions on this dissertation. I would also like to thank Dr. Giovanni Coppola, my previous committee member, for providing me with opportunities to work with the PSP dataset and giving me feedback on my qualifying exam.

Lastly, I would like to thank my colleagues. I have learned a lot from my labmates throughout the years, including Lingyu Zhan, Sarah Spendlove, Brandon Jew, Ziyuan Jiang, Nahyun Kong,

and Stephen Hwang. I am grateful to friends I have met at UCLA and outside of UCLA for their support and encouragement. I am also grateful to the genetics community at UCLA, especially the Department of Human Genetics and the Bioinformatics program. I would like to thank my internship managers and mentors, Ethan Xu, Alex Fahey, and Juan Borgnino for letting me know more about the science in industires. And I thank my undergraduate advisors, Jianhua Yang and Jinxin Bei, for introducing me to the wonderful world of human genetics and bioinformatics. I would also like to thank Buhm Han for his feedback on my paper.

VITA

EDUCATION

BS in Biotechnology, Sun Yat-sen University	2012-2016
PhD Candidate in Human Genetics, University of California, Los Angeles	2016-2021

PUBLICATIONS

Jiajin Li, Brandon Jew, Lingyu Zhan, Sungoo Hwang, Giovanni Coppola, Nelson B. Freimer, and Jae Hoon Sul. 2019. ForestQC: quality control on genetic variants from next-generation sequencing data using random forest. *PLoS Computational Biology*. 15 (12): e1007556.

Jiajin Li, Nahyun Kong, Buhm Han, and Jae Hoon Sul. 2020. Rare variants regulate expression of nearby individual genes in multiple tissues. *PLoS Genetics*. 17 (6): e1009596.

Ziyuan Jiang*, **Jiajin Li***, and Jae Hoon Sul. 2020. Accurate classification of atopic dermatitis combining RNA sequencing and microbiome data with supervised machine learning. *Allergy, Asthma & Immunology Research (Under review)*. *These authors contributed equally to this work

Brandon Jew*, **Jiajin Li***, and Jae Hoon Sul. 2020. An ultra-fast linear mixed model framework for association studies across multiple contexts. *Workshop on Algorithms in Bioinformatics 2021*. *These authors contributed equally to this work

Lingyu Zhan, **Jiajin Li**, Brandon Jew, Jae Hoon Sul. 2021. Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences. *PLoS Genetics (Accpted)*.

Chapter 1 Introduction

One of the primary goals in genetic and genomic studies is to understand how genetic variations affect phenotypes, especially diseases and complex traits. Studies have found that some genetic variants can increase the risk of getting certain diseases and that individuals carrying some variants can have larger weights or heights. Such information about associations between variants and diseases is essential to understand their genetic basis, which can be fundamental in finding treatments for them.

The genetic basis of diseases or complex traits can be complicated because they can be affected by one or more genetic variants as well as some environmental factors. One approach to identify the genetic mechanisms of complex traits or diseases is association studies. Association studies compare the genotypes of individuals with a disease (cases) with those without a disease (control). Alleles that are overrepresented in cases would be determined to be associated with the disease. With the advent of next-generation sequencing (NGS) technologies, the cost of detecting numerous genetic variants decreases drastically. It enables genetic studies to collect information on tens of millions of single nucleotide polymorphisms (SNP) from a large population. Genome-wide association studies become possible, which perform association studies on a genome-wide level. Over the past decades, GWAS have been playing an important role in identifying genetic variations associated with diseases or complex traits.

However, GWAS have been very successful but still have some limitations. First, several factors can affect the quality of variants detected by NGS, which could, in turn, lead to inaccurate results in the follow-up analysis in GWAS. Errors and biases in NGS, alignment algorithms, and variant calling tools can cause false positives in variant detection. Therefore, it is vital to perform quality control (QC) on genetic variants to remove variants with sequencing errors. Traditionally, genetics studies have two types of QC approaches: “filtering” and “classification” approaches. In the filtering approach, several hard filters are applied to remove problematic variants. One main problem with this type of approach is that the cutoff values are often study-specific and need to be manually fine-tuned for each study. It is also difficult to determine the quality of variants whose metrics are very close to the threshold values. The classification approach attempts to learn variants with low quality using machine learning techniques, which often use public databases to train the models. One of the issues is that the models may be biased to keep known variants in the databases and filter out novel variants. Another issue is that those databases might not always be accurate, contributing to the imprecise classification of variants.

Second, GWAS focus on the associations between common variants and traits, but common variants can explain only a fraction of the heritability. This phenomenon is referred to as “missing heritability”. Rare variants are considered to contribute to the missing heritability. Like common variants, rare variants might affect traits by regulating the expression of nearby genes. However, the effects of rare variants on gene expression remain mostly obscure. There are two major challenges in the analysis of the regulatory effects of rare variants. The first challenge is relatively small sample sizes of datasets with both whole-genome sequencing (WGS) and RNA-seq data. The second challenge is the statistical methods for the analysis. Traditional association methods using a single marker test for each SNP have low statistical power for rare variants. To increase the power, many collapsing approaches that combine the effects of multiple rare variants have been proposed but they are not optimized for analyzing the functional effects of rare variants.

With the accumulation of knowledge about associations between diseases and genomic data, it is possible to make predictions on diseases based on the genetic profiles of patients. Previous studies have found that host gene expression and gut microbiota are associated with atopic dermatitis (AD), a type of inflammatory skin disease. However, there have been few studies on prediction analysis using machine learning based on the gut transcriptome and microbiota in AD. It is desirable to have an accurate and automated diagnosis of AD and an improved set of biomarkers for it because it is challenging to diagnose AD and assess its severity.

My thesis work focuses on developing machine learning applications and statistical methods to improve the analysis of genetics and genomics. Below I will have a brief introduction of the scientific challenges and the methods that I developed to address them.

Chapter 2: Developing machine learning applications for the quality control on genetic variants

Genetic variants of low quality can cause spurious associations in GWAS. Two main types of approaches, filtering and classification approaches have been proposed to perform quality control on the variants detected by NGS, but they have several limitations and fail to improve variant quality for various datasets. The filtering approach requires predefined cutoff values that are often study-specific. It is also difficult to determine the quality of variants whose metrics are very close to the threshold values. The classification approach often uses public databases to train the models, so the models may be biased to keep known variants in the databases and filter out novel variants. Another issue is that those databases might be inaccurate, leading to the poor performance of the models. In Chapter 2, I propose ForestQC, a variant QC method that combines a filtering approach with a classification approach based on a random forest model. It can solve the issues mentioned above and is very scalable. I demonstrate with two high-coverage WGS datasets that ForestQC outperforms existing methods by considerably improving variant quality in both datasets.

Chapter 3: Detecting the regulatory effects of rare variants in multiple tissues

To discover the functional effects of genetic variants, many expression quantitative trait loci (eQTL) studies are interested in identifying genes whose expression levels are influenced by genetic variants (called “eGenes”). We call the genes regulated by rare variants “RV eGenes”. It is challenging to identify RV eGenes because of the limited available datasets with large enough sample sizes and statistical approaches with sufficiently high statistical power. Traditional association test that utilizes a single marker test for each SNP works well for common variants in GWAS but suffers from power loss for rare variants. That is because its power is proportional to the minor allele frequency of SNPs. To boost power, many collapsing approaches that combine the effects of multiple rare variants have been proposed but they are not optimized to find RV eGenes. In Chapter 3, I present a novel powerful approach called LRT-q to detect RV eGenes. It incorporates functional annotations of rare variants and aggregates their statistics in a nonlinear manner to identify a group of potential causal rare variants influencing the expression of a nearby gene. I show by simulated data that LRT-q is more powerful than previous methods. I also apply LRT-q to the Genotype-Tissue Expression (GTEx) v8 dataset to perform the first comprehensive analysis of the regulatory effects of rare variants in multiple tissues. I identify many RV eGenes from this dataset and find several important biological insights.

Chapter 4: Designing machine learning algorithms for the automated diagnosis of atopic dermatitis

Atopic dermatitis (AD) is a type of inflammatory skin disease that can impose a high economic burden and have considerable negative effects on life quality. It is challenging to diagnose AD because of its variable morphology, distribution, and irregularity. And the assessment of disease severity is problematic due to the lack of objective markers. Therefore, an accurate and automated diagnosis of AD and an improved set of biomarkers for it could have a potentially high impact. Studies have shown that AD is related to the expression of some genes and gut dysbiosis. Recent integration and correlation analyses of host gene expression and gut microbiota have emerged as an important opportunity for the diagnosis and prediction of human diseases like AD. However, there have been few studies on AD prediction based on gut transcriptome and microbiota. In Chapter 4, I present a machine learning classifier for an accurate and automated diagnosis pipeline for AD using the transcriptome of gut epithelial colonocytes and gut microbiota. I demonstrate that it can differentiate between subjects with and without AD based on the omics data with high accuracy. And I show that it can also identify a set of predictive genes and microbiota features that may provide novel biological insights and be developed into useful biomarkers for AD.

Chapter 2 Developing machine learning applications for the quality control on genetic variants

2.1 Introduction

Over the past few years, genome-wide association studies (GWAS) have been playing an essential role in identifying genetic variations associated with diseases or complex traits [1,2]. GWAS have found many associations between common variants and human diseases, such as schizophrenia [3], type 2 diabetes [4,5], and Parkinson's Disease [6]. However, these common variants typically explain only a small fraction of heritability for the complex traits [7,8]. Rare variants have been considered as an important risk factor for complex traits and diseases [9–12]. With the next-generation sequencing (NGS) technology, geneticists may now gain insights into the roles of novel or rare variants. For instance, deep targeted sequencing was applied to discover rare variants associated with inflammatory bowel disease [13]. Whole-genome sequencing (WGS) has been used to identify rare variants associated with prostate cancer [14], and with whole-exome sequencing, studies have also detected rare variants associated with LDL cholesterol [15] and autism [16].

However, several factors may adversely influence the quality of variants detected by sequencing. First, NGS is known to have errors or biases [17–21], which might cause inaccuracy in detecting variants. Second, the sequence mappability of different regions may not be uniform but correlated with sequence-specific biological features, leading to alignment biases. For instance, it is shown that introns have significantly lower mappability levels than exons [22]. Third, variant calling algorithms may be the sources of errors as no algorithm is 100% accurate. For example, GATK HaplotypeCaller and GATKUnifiedGenotyper [23], which are the widely used variant callers, have a sensitivity of about 96% and precision of about 98% [24]. Additionally, different variant callers may generate discordant calls [25], and in some instances, different versions of even the same software may generate inconsistent calls. All these factors may generate false-positive variants or incorrect genotypes, which may then lead to false-positive associations in the follow-up association analyses. For example, Alzheimer's Disease Sequencing Project has reported that they found spurious associations in the case-control analysis where one of the causes for the problem could be inconsistent variant discovery pipelines [26].

It is vital to perform quality control (QC) on genetic variants identified from sequencing to remove variants that may contain sequencing errors and hence, are likely to be false-positive calls. Traditionally, genetic studies have utilized two types of QC approaches; we call them "filtering" and "classification" approaches. In the filtering approach, several filters are applied to remove problematic variants such as variants with high genotype missing rate (e.g. > 5%), low Hardy-Weinberg Equilibrium (HWE) p-value (e.g. < 1E-4), or very high or low allele balance of heterozygous calls (ABHet) (e.g. > 0.75 or < 0.25). One main problem with this type of approach is that these thresholds are often study-specific and need to be manually fine-tuned for each study. We may also remove variants whose metrics are very close to the thresholds (e.g., variants with a missing rate of 5.1%). Another type of QC is the classification approach that attempts to learn variants with low quality using machine learning techniques. One example is Variant Quality Score Recalibration (VQSR) of GATK [24,27] that uses a Gaussian mixture model to learn the multidimensional annotation profile of variants with high and low quality. However, one of the issues with VQSR is that one needs training datasets acquired from existing databases on variants such as 1000 Genomes Project [28] and HapMap [29], which may be biased to keep known variants and filter out novel variants. Another issue is that those known databases of genetic variants may not always be accurate, which would lead to inaccurate classification of variants, and they may not even be available for some species. It may also be a challenge to apply VQSR to a variant call set generated by variant callers other than GATK as VQSR needs metrics of variants that are not often calculated by non-GATK variant callers.

In this article, we present ForestQC for performing QC on genetic variants discovered through sequencing. Our method aims to identify whether a variant is of high sequencing quality (high-quality variants) or low quality (low-quality variants) by combining the filtering and classification approaches. We first apply the filtering approach by applying stringent filters to identify truly high-quality or low-quality, while the rest of variants that are neither high-quality nor low-quality are considered to have uncertain quality ("undetermined" variants). Given this set of high-quality and low-quality variants, we train a machine learning model whose goal is to classify whether the undetermined variants are high-quality or low-quality. With an insight that high-quality variants would have higher genotype quality and sequencing depth than do low-quality variants, we use the information of several sequencing quality measures of variants for model training. ForestQC then uses sequencing quality measures of the undetermined variants to predict whether each undetermined variant has high or low sequencing quality. Our approach is different from the filtering strategy in that it only uses filters to identify unambiguously high-quality and low-quality variants and does not attempt to classify undetermined variants with filters. Our method is also different from VQSR as our training strategy allows us to train our model without reference datasets for variants and solves several issues with VQSR mentioned above. Another advantage of our software is that it can be applied to Variant Call Format (VCF) files from most of variant callers that generate standard quality information for genotypes and is very efficient.

To demonstrate the accuracy of ForestQC, we apply it to two high-coverage WGS datasets; 1) large extended pedigrees ascertained for bipolar disorder (BP) from Costa Rica and Colombia [30], and 2) a sequencing study for Progressive Supranuclear Palsy (PSP). The first dataset includes 449 related individuals from families, while the latter dataset consists of 495 unrelated individuals. We show that ForestQC outperforms VQSR and a filtering approach based on ABHet as high-quality variants detected from ForestQC have higher sequencing quality than those from VQSR and the filtering approach in both datasets. This suggests that our tool identifies high-quality variants with higher accuracy than other approaches in both family and unrelated datasets. ForestQC is publicly available at https://github.com/avallonking/ForestQC.

2.2 Results

2.2.1 Overview of ForestQC

ForestQC takes a raw VCF file as input and determines which variants have high or low quality. Our method combines a filtering approach that determines high-quality and low-quality variants by a set of pre-defined filters and a classification approach that uses machine learning to classify whether a variant is high-quality or low-quality. As illustrated in Fig 1, our method first calculates the statistics of each variant for several filters that are commonly used in performing QC in GWAS. These statistics consist of ABHet, HWE p-value, genotype missing rate, Mendelian error rate for family-based datasets, and any user-defined statistics (details described in Materials and Methods). ForestQC then identifies three sets of variants using these statistics as filters: 1) a set of high-quality variants that pass all filters, 2) a set of low-quality variants that fail any filter(s), and 3) a set of undetermined variants that are neither high-quality nor low-quality variants. We use stringent thresholds for filters (S1 and S2 Tables), and hence we are highly confident that high-quality variants have good quality while low-quality variants are indeed false-positives or have unequivocally poor sequencing quality. The next step in ForestQC is to train a random forest machine learning model using the high-quality and low-quality variants we detect in the filtering step. In ForestQC, seven sequencing quality metrics of high-quality and low-quality variants are used as features to train the random forest model, including three related to sequencing depth, three related to genotype quality, and one related to the GC content. Finally, the fitted model predicts whether each undetermined variant is high-quality or low-quality. We combine the predicted high-quality variants from the random forest classifier and the high-quality variants detected in the filtering step, as the complete set of high-quality variants determined by ForestQC. The same procedure is applied to identify low-quality variants.

One major challenge in classifying undetermined variants is to identify a set of sequencing quality metrics that are used as features to train the random forest model. We choose three sets of features based on quality metrics provided by variant callers and prior knowledge in genome sequencing. The first set of features is genotype quality (GQ), where we have three metrics: mean, standard deviation (SD), and outlier ratio. The outlier ratio is the proportion of samples whose GQ scores are lower than a particular threshold, and it measures a fraction of individuals who are poorly sequenced at a mutation site. A high-quality variant is likely to have high mean, low SD, and low outlier ratio of GQ values. The second set of features is sequencing depth (DP), as low depth often introduces sequencing biases and reduces variant calling sensitivity [31]. We also use the same three sets of metrics for DP as those for GQ: mean, SD, and outlier ratio. The last set of features is related to genomic characteristics instead of sequencing quality, which is GC content. Too high or too low GC content may decrease the coverage of certain regions [32,33] and thus may lower the quality of variant calling. Hence, the GC content of the DNA region containing high-quality variants would not be too high or too low. Given these three sets of features, ForestQC learns how those features determine high-quality and low-quality variants, and classifies undetermined variants according to the rules that it learns.

2.2.2 Comparison of different machine learning algorithms

As there are various machine learning algorithms available, we first seek to find the most accurate and efficient algorithm for performing QC on NGS variant call-sets. To ensure the quality of training and prediction, we choose supervised learning algorithms rather than unsupervised algorithms. Several major types of supervised algorithms are selected for comparison: random forest, logistic regression, k nearest neighbors (KNN), Naive Bayes, quadratic discriminant analysis (QDA), AdaBoost, artificial neural network (ANN), and single support vector machine (SVM). We use the BP WGS dataset, which consists of large pedigrees from Costa Rica and Colombia, to compare the performance of different algorithms. We use the three sets of features mentioned above for all these algorithms. We apply the filtering approach (S1 and S2 Tables) to the BP data to identify high-quality, low-quality, and undetermined variants, and we randomly sample 100,000 high-quality and 100,000 low-quality variants for model training. We then randomly choose another 100,000 high-quality and 100,000 low-quality variants from the rest of variants for model testing. Each learning algorithm will be trained with the same training set and tested with the same test set. We use 10-fold cross-validation and calculate area under the receiver operating characteristic curve (AUC) and F1-score to estimate classification accuracy during model testing. F1-score is the harmonic average of precision (positive predictive value) and recall (sensitivity). The closer the F1-score is to 1, the better the performance is. To assess the efficiency of each algorithm, we measure its runtime during training and predicting. We use eight threads for algorithms that support parallelization.

Results show that random forest is the most accurate model in both SNV classification and indel classification with the highest F1-scores, accuracy, and the largest AUC (Table 1, S3 Table, S1 Fig). Its runtime is only 9.85 seconds in model training and prediction (Table 1), which ranks as the fourth fastest algorithm. As random forest randomly divides the entire dataset into several subsets of the same size and constructs decision trees independently in each subset, it is highly scalable, and it has low error rates and high robustness [34]. As for other machine learning algorithms, both SVM and ANN are highly accurate (both with F1-score of 0.97 and AUC > 0.985 in SNV classification), but they are not as efficient as random forest. ANN is the second slowest algorithm that is about 8x slower than random forest because it estimates many parameters. Especially, SVM is the slowest algorithm because of its inability to parallelize, which needs about 125x as much time as random forest (Table 1). This suggests that it may be computationally costly to use SVM in large-scale WGS datasets that have tens of millions of variants. Typically, a real dataset is at least ten times larger than the dataset used here. For example, in the BP dataset, the training set has 2.20 million (M) SNVs, and there are 2.73M undetermined SNVs for prediction. We find that random forest only spends 80.51 seconds in training and predicting, while ANN needs 489.63 seconds, and SVM needs 14.74 hours. Therefore, random forest is much faster than ANN and SVM, although all three algorithms have similar performance in terms of AUC (S1 Fig). Also, there are even a more significant number of variants in large-scale WGS projects such as the NHLBI Trans-Omics for Precision Medicine

(TOPMed) dataset that includes about 463M variants. Hence, it is more practical to use random forest when processing these massive datasets. Logistic regression, Naive Bayes, and QDA are more efficient than random forest, but their predictions are not as accurate as those of random forest. For example, Naive Bayes needs only 0.18 seconds for training and prediction, while its F1-score is the lowest among all algorithms (0.90 and 0.87 in SNV and indel classification, respectively) (Table 1). This result demonstrates that random forest is both accurate and efficient, and hence we use it as the machine learning algorithm in our approach. To further improve the random forest algorithm, we test a different number of trees in the algorithm, and we find that random forest with 50 trees balances efficiency and accuracy (S2 Fig). Also, we consider undetermined variants with the predicted probability of being high-quality variants > 50% as high-quality variants as this probability threshold achieves the highest F1-score (S3 Fig).

2.2.3 Measuring performance of QC methods on WGS data

To evaluate the accuracy of ForestQC and other methods on WGS data, we apply them to two WGS datasets and calculate several metrics. For a family-based dataset, we calculate the Mendelian error rate (ME) of each variant, which measures inconsistency in genotypes between parents and children. Another metric is the genotype discordance rate between microarray and sequencing as individuals in both WGS datasets we analyze are genotyped with both microarray and WGS. These two metrics are important indicators of variant quality because high-quality variants would follow Mendelian inheritance patterns, and their genotypes would be consistent between microarray and sequencing. Additionally, we compute some other metrics that are reported in sequencing studies such as the number of variants (SNVs and indels), transitions/transversions (Ti/Tv) ratio, the number of multi-allelic variants, genotype missing rate. Note that these QC-related metrics are computed separately for SNVs and indels. We use these metrics to compare the performance of ForestQC with that of three approaches. The first is one without performing any QC (no QC). The second method is VQSR, which is a classification approach that requires known truth sets for model training, such as HapMap or 1000 genomes. We use the recommended resources and parameter settings to run VQSR as of 2018-04-04 [35], but we also look at different settings. The third method is the ABHet approach, which is a filtering approach that retains variants according to the allele balance of variants (see Methods).

2.2.4 Performance of ForestQC on family WGS data

We apply ForestQC to the BP WGS dataset that consists of 449 subjects with an average coverage of 36-fold. There are 25.08M SNVs and 3.98M indels [30]. The variant calling is performed with GATK-HaplotypeCaller v3.5. This is an ideal dataset for assessing the performance of different QC methods because this dataset contains individuals from families who are both sequenced and genotyped with microarray. This study design allows us to calculate both ME rate and genotype discordance rate of variants between WGS and microarray. For this dataset, we test ForestQC with two different filter settings, one using ME rate as a filter and the other not using ME as a filter. The results of the former approach would filter out low-quality variants based on ME rate, and hence ME rate of high-quality variants would be very low. However, we observe that both approaches have similar performance in terms of ME rate and other metrics (S4 Table, S4 Fig, S5 Fig), and hence we show results of only ForestQC using ME rate as a filter.

Results show that ForestQC outperforms ABHet and VQSR in terms of the quality of high-quality SNVs while it detects fewer such SNVs than the other approaches (detailed variant-level metrics in Table S5). ForestQC identifies 22.23M (88% of total SNVs) high-quality SNVs, which is fewer than 22.42M (89%) and 24.24M (97%) high-quality SNVs from ABHet and VQSR, respectively (Table 2). However, ABHet has 3.57x and VQSR has 9.99x higher ME rate on high-quality SNVs than ForestQC (Fig 2A), and ABHet has 1.50x (p-value < 2.2e-16) and VQSR has 1.26x higher genotype discordance rate (p-value < 2.2e-16) on high-quality SNVs than ForestQC (Fig 2B). Besides, ABHet and VQSR have 81.48x and 97.72x higher genotype missing rate on high-quality SNVs than ForestQC, respectively (Fig 2C). However, it is important to note that genotype missing rate is used as a filter in ForestQC, so SNVs with high genotype missing rate are filtered out. We observe that VQSR and ABHet have 319 thousand (K) (1.32% of total high-quality SNVs) and 235K (1.05%) high-quality SNVs with very high genotype missing rate (>10%), respectively, and there are also 118K (0.49%, VQSR) and 53K (0.24%, ABHet) high-quality SNVs with very high ME rate (>15%), while ForestQC has none of them due to its filtering approach. We then investigated whether low-quality variants detected by ForestQC are of poor sequencing quality. Our results show that low-quality SNVs detected by our method have higher genotype missing rate, higher ME rates, and higher genotype discordance rate than those of ABHet, and higher genotype missing rate than those of VQSR (S6A, S6B and S6C Fig). The no QC method keeps the greatest number of SNVs (25.08M), but they have the highest ME rate, genotype missing rate, and genotype discordance rate as expected.

Next, we calculate several metrics of high-quality SNVs commonly used in sequencing studies to evaluate the performance of ForestQC. One such metric is the Ti/Tv ratio. It was reported that transitional mutations occurred more frequently than transversional mutations in the human genome [36]. In human WGS datasets, this ratio is expected to be around 2.0 [23]. The lower Ti/Tv ratio is compared to the genome-wide expected value of 2.0, the more false-positive variants are expected in the dataset. We compute the Ti/Tv ratio for each individual across all high-quality SNVs and look at the distribution of those ratios across all individuals (sample-level metrics). We find that the mean Ti/Tv ratio of high-quality known SNVs (present in dbSNP) is around 2.0 for all four methods, which suggests that they have similar accuracy on known SNVs in terms of Ti/Tv ratio (S7A Fig). However, results show that the mean Ti/Tv ratio of high-quality novel SNVs (not in dbSNP) from ForestQC is better than that of those SNVs from other methods; the mean Ti/Tv ratio is 1.68 for ForestQC, which is closest to 2.0 among other methods (1.41 for VQSR, 1.53 for ABHet, and 1.29 for No QC) (Fig 3A). Paired t-tests for the difference in the mean Ti/Tv ratio between ForestQC and other methods are all significant (p-value < 2.2e-16 versus all other methods). This result suggests that novel SNVs predicted to be high-quality by ForestQC are more likely to be true-positives than those novel SNVs from other QC methods. Another metric commonly used in sequencing studies is the percentage of multi-allelic SNVs, which are variants with more than one alternative allele. Given this relatively small sample size (n = 449), true multi-allelic SNVs are not expected to be observed very frequently, so a good portion of them are considered as false-positives. ForestQC has 33.96% and 42.62% smaller fraction of multi-allelic SNVs among high-quality SNVs than do VQSR and no QC methods, while the ABHet approach has the smallest fraction of such SNVs (Table 2). It is important to note that ABHet value is defined as the proportion of reference alleles from heterozygous samples, so ABHet values are not expected to be 0.5 for high-quality multi-allelic mutation sites, but other unknown values. Hence, ABHet does not work properly for multi-allelic variants and may excessively remove multi-allelic SNVs.

In addition to SNVs, we apply the four QC methods to indels. Similar to the results of SNVs, ForestQC identifies fewer high-quality indels than does VQSR, but the quality of those indels from ForestQC is better than that of high-quality indels from ABHet and VQSR. Out of 3.98M indels in total, ForestQC predicts 2.79M indels (70% of total indels) to have good sequencing quality while VQSR and ABHet find 3.21M (81%) and 2.67M (67%) high-quality indels, respectively (Table 2). High-quality indels from VQSR and ABHet, however, have 8.54x and 3.18x higher ME rate, and 22.25x and 25.28x higher genotype missing rate, than those from ForestQC, respectively (Fig 2D and 2E). Low-quality indels identified by ForestQC have 2.25x and 1.32x higher ME rate, and 1.48x and 2.36x higher genotype missing rate than those from VQSR and ABHet, respectively (S6D and S6E Fig). Besides, we observe that there are 95K (2.97% of total high-quality indels, VQSR) and 86K (3.23%, ABHet) high-quality indels with very high genotype missing rate (>10%) and also 167K (5.21%, VQSR) and 44K (1.66%, ABHet) high-quality indels with very high ME rate (>15%), while there are no such indels in ForestQC. This result suggests that many high-quality indels detected by ABHet or VQSR may be false-positives or indels with poor sequencing quality. One of the reasons why VQSR does not perform well on indels could be the reference database it uses for model training as VQSR considers all indels in the reference database (Mills gold standard call set [37] and 1000G Project [38]) as true-positives. This leads VQSR to have a significantly higher proportion of known indels among its high-quality indels (86% of total high-quality indels), compared with 80% from ForestQC and 82% from ABHet (Table 2). Nevertheless, some indels in the reference database may be false-positives or have poor sequencing quality in the variant call-sets of interest. Hence, the performance of VQSR may be limited by using reference database to identify high-quality variants. It is also important to note that in general, indels have much a higher ME rate (0.41% for no QC) than that of SNVs (0.08% for no QC), which is expected given the greater difficulty in calling indels.

Another significant difference between ForestQC and the other approaches is the allele frequency of variants after QC, as ForestQC keeps a higher number of rare variants in its variant set. Our method has 1.77% and 1.64% higher proportion of rare SNVs, and 5.30% and 15.37% higher proportion of rare indels than ABHet and VQSR do, respectively (S6 Table). We also observe this phenomenon in the variant-level and sample-level metrics for the number of SNVs. The variant-level metrics show that the number of high-quality SNVs detected by ForestQC is similar to those from ABHet (Table 2). However, the sample-level metrics show that each individual on average carries fewer alternative alleles of high-quality SNVs from ForestQC (3.58M total SNVs) than those from VQSR and ABHet (3.99M and 3.77M total SNVs, respectively) (Fig 3B and 3C, S7B Fig). We observe a similar phenomenon for indels between ABHet and ForestQC (Table 2, Fig 3D, S7C and S7D Fig). This phenomenon could be explained by the higher fraction of rare variants among high-quality variants from ForestQC, as individuals would carry fewer variants if there are a higher fraction of rare variants. One main reason why ForestQC has a higher proportion of rare variants is that common variants in the BP dataset have higher ME rate, genotype discordance rate, and genotype missing rate than do rare variants, and therefore, they are more likely to fail the filters of ForestQC (S8 Fig).

ForestQC uses several filters to remove low-quality variants while the other two approaches (VQSR and ABHet) do not use these filters, which might have artificially improved the performance of ForestQC. Hence, to compare ForestQC with other approaches without this potential bias, we measure the performance metrics on only undetermined variants as the filters do not determine their quality in our approach. From 2.73M undetermined SNVs and 1.09M undetermined indels, ForestQC identifies 979K (35.83% of total undetermined SNVs) high-quality SNVs and 532K (48.58% of total undetermined indels) high-quality indels, while ABHet approach detects 620K (22.70%) SNVs and 195K (17.80%) indels, and VQSR selects 2.16M (79.18%) SNVs and 643K (58.76%) indels as high-quality variants, respectively (S7 Table). For high-quality SNVs from undetermined variants, ABHet and VQSR have 2.75x and 22.67x higher ME rate than ForestQC, respectively (S9A Fig), and ABHet and VQSR have 5.15x (p-value = 1.367e-14) and 3.86x (p-value = 1.926e-14) higher genotype discordance rate than ForestQC (S9B Fig). Also, ABHet and VQSR have 15.50x and 7.05x higher genotype missing rate on high-quality SNVs than ForestQC, respectively (S9C Fig). We observed similar results for indels (S9D and S8E Figs). Sample-level metrics also show that ForestQC has better Ti/Tv ratio on known SNVs (mean Ti/Tv: 1.64, 1.85, 1.72, 1.88 for No QC, ABHet, VQSR, ForestQC, respectively), and novel SNVs (mean Ti/Tv: 1.14, 1.04, 1.21, 1.22 for No QC, ABHet, VQSR, ForestQC, respectively) than other methods (S10D and S9E Figs). Paired t-tests for the difference in the mean Ti/Tv ratio of novel SNVs and known SNVs between ForestQC and other methods are all significant (p-value < 0.05 versus all other methods). These results show that ForestQC has better performance than ABHet and VQSR, even on those undetermined variants whose quality is not determined by filtering.

2.2.5 Performance of ForestQC on WGS data with unrelated individuals

To evaluate the performance of ForestQC on WGS datasets that contain only unrelated individuals, we apply it to the PSP dataset that has 495 subjects who are whole-genome sequenced at an average coverage of 29-fold, generating 33.27M SNVs and 5.09M indels. Among the 495 individuals who are sequenced, 381 individuals (77%) of them are also genotyped with microarray, which enables us to check the genotype discordance rate between WGS and microarray data. Because the PSP dataset contains only unrelated individuals, we do not report the ME rate. Similar to the BP WGS dataset, we apply four methods (ForestQC, VQSR, ABHet, and No QC) to the PSP dataset, although the parameter settings of VQSR have slightly changed. As the PSP dataset is called with GATK v3.2, the StrandOddsRatio (SOR) information from the VCF file is missing, which is recommended to use in VQSR. However, we find that SOR information has little impact on the results of VQSR as we test VQSR without SOR information using the BP dataset and obtain similar results with one using SOR information (S11 Fig).

Similar to the results of the BP dataset, high-quality variants identified by ForestQC are fewer but of higher sequencing quality than other approaches (detailed variant-level metrics in Table S8). ForestQC identifies 29.25M (88% of total SNVs) high-quality SNVs, which is slightly fewer than 29.77M (89%) high-quality SNVs from ABHet but about 2 million fewer than 31.28M (94%) high-quality SNVs from VQSR (Table 3). However, high-quality SNVs from ABHet and VQSR have 53.76x and 42.55x higher genotype missing rate than those from ForestQC, respectively (Fig 4A), but it is important to note that missing rate is included as a filter in ForestQC. In addition, there are 311K (0.99% of total high-quality SNVs, VQSR) and 331K (1.13%, ABHet) high-quality SNVs with very high genotype missing rates (>10%), while ForestQC removes all these SNVs. We also observe that low-quality SNVs from ForestQC have a 2.4x higher genotype missing rates than those from ABHet, although low-quality SNVs from GATK have slightly higher missing rates than those from ForestQC (S12A Fig). High-quality SNVs from ABHet and VQSR have 1.28x (p-value < 2.2e-16) and 1.29x higher genotype discordance rate (p-value < 2.2e-16) than those from ForestQC, respectively (Fig 4B). As for the genotype discordance rate of low-quality SNVs, both ABHet and VQSR have higher genotype discordance rate than does ForestQC (S12B Fig), but this may be inaccurate because of the small number of low-quality SNVs genotyped with microarray (10,130, 4,121, and 553 such SNVs for

ForestQC, ABHet, and VQSR, respectively). The variant-level and sample-level metrics also indicate the better quality of high-quality SNVs from ForestQC. Although all methods have mean Ti/Tv ratios of high-quality known SNVs above 2.0, the mean Ti/Tv ratio of high-quality novel SNVs among all sequenced individuals is 1.65 for ForestQC, which is closer to 2.0 than other methods (1.27, 1.54, and 1.24 for VQSR, ABHet, no QC, respectively). (S13A Fig, Fig 5A). Paired t-tests for the difference in the mean Ti/Tv ratio between ForestQC and other methods are all significant (p-value < 2.2e-16 versus all other methods). ForestQC has 16.67% and 33.33% smaller fraction of multi-allelic SNVs among high-quality SNVs than do VQSR and no QC methods, respectively, while the ABHet approach has the smallest proportion of such SNVs (Table 3). ABHet has the smallest number of multi-allelic SNVs because of the reason we discussed in the previous BP dataset analysis. Lastly, consistent with the results of the BP dataset, the sample-level metrics show that each individual on average carries fewer alternative alleles of high-quality SNVs from ForestQC than those from VQSR and ABHet (Fig 5B and 5C, S13B Fig). Rare SNVs in high-quality SNVs from ForestQC account for 1.70% and 1.32% higher proportion, compared with those from ABHet and VQSR (S5 Table). This is because rare SNVs in the PSP dataset have lower genotype missing rate and lower genotype discordance rate, and thus do not fail filters as often as do common SNVs (S14A and S14B Fig).

For indels, ForestQC predicts 3.42M indels (67% of total 5.09M indels) to be high-quality variants, which is slightly more than 3.31M (65%) high-quality indels from ABHet and fewer than 3.68M (72%) high-quality indels from VQSR (Table 3). Because the PSP dataset lacks the ME rate as it contains only unrelated individuals and indels are not detected by microarray, it is difficult to compare the performance of the QC methods on indels. We find that high-quality indels from ABHet and VQSR have 27.02x and 18.77x higher genotype missing rate than those from our method, respectively (Fig 4C). Additionally, VQSR and ABHet have 107K (2.91% of total high-quality indels) and 131K (4.08%) high-quality indels with high genotype missing rate (>10%), respectively, while ForestQC filters out all of these indels. Also, low-quality indels from ForestQC have 2.05x and 1.21x higher genotype missing rate than those from ABHet and VQSR, respectively (S12C Fig). This, however, may be biased comparison as ForestQC removes indels with high genotype missing rate in its filtering step. Consistent with the results of SNVs, the sample-level metrics indicate that each individual has fewer high-quality indels from ForestQC than those from VQSR and ABHet (Fig 5D, S13C, S13D Fig). Among high-quality indels, ForestQC has 6% and 1% more novel indels than VQSR and ABHet, respectively (Table 3). In terms of allele frequency, rare indels detected by ForestQC accounts for 12.35% and 3.49% larger proportions than those identified by VQSR and ABHet, respectively (S9 Table). Similar to the results of the BP dataset, we also observe that the missing rate of rare indels is lower than that of common indels. (S14C Fig).

Similar to the analysis of the BP dataset, we also compare the performance of ForestQC, ABHet approach, and VQSR only on undetermined variants in the PSP dataset. From 3.95M undetermined SNVs and 1.60M undetermined indels, ForestQC identifies 1.71M (43.33% of total undetermined SNVs) high-quality SNVs and 719K (45.01% of total undetermined indels) high-quality indels, while ABHet approach detects 780K (19.74%) SNVs and 248K (15.51%) indels, and VQSR selects 2.75M (69.52%) SNVs and 820K (51.34%) indels as high-quality variants, respectively (S10 Table). For high-quality SNVs from undetermined variants, ABHet and VQSR have 14.84x and 5.38x higher genotype missing rate than ForestQC, respectively (S15A Fig). In addition, ABHet has 2.09x (p-value = 2.183e-11) and VQSR has 2.13x higher genotype discordance rate (p-value = 1.584e-10) on than ForestQC (S15B Fig). For indels, ABHet and VQSR have 9.39x and 3.61x higher genotype missing rate on high-quality indels than ForestQC, respectively (S15C Fig). Sample-level metrics also show that ForestQC has better Ti/Tv ratio on known SNVs (mean Ti/Tv: 1.75, 1.87, 1.82, 1.96 for No QC, ABHet, VQSR and ForestQC, respectively) and novel SNVs (mean Ti/Tv: 1.17, 1.03, 1.20, 1.39 for No QC, ABHet, VQSR and ForestQC, respectively) than other methods (S15D and S15E Fig). Paired t-tests for the difference in the mean Ti/Tv ratio of novel SNVs and known SNVs between ForestQC and other methods are all significant (p-value < 2.2e-16 versus all other methods). Similar to the results of the BP dataset, ForestQC has higher accuracy in identifying high-quality variants from undetermined variants, compared with the ABHet approach and VQSR.

2.2.6 Feature importance in random forest classifier

ForestQC uses several sequencing features in the random forest classifier to predict whether a variant with undermined quality is high-quality or low-quality. To understand how these features determine variant quality, we analyze the feature importance of the fitted random forest classifier. We first find that GC-content has the lowest importance in both BP and PSP datasets and also for both SNVs and indels (S17 Fig). This means that GC-content may not be an informative indicator of the quality of variants as other features related to sequencing quality, such as depth (DP) and genotype quality (GQ). Second, the results show that classification results are not determined by one or two most important features as there is no feature with much higher importance than other features except GC-content. This suggests that all sequencing features except GC-content are essential indicators of the quality of variants and need to be included in our model. We also check correlation among features and find that while specific pairs of features are highly correlated, like outlier GQ and mean GQ, SD DP and mean DP, some features have low correlation to other features, such as GC, suggesting that they may capture different information on quality of genetic variants (S19 Fig). Third, we observe that the same features have different importance between the BP dataset and the PSP dataset. For example, for SNVs, an outlier ratio of the GQ feature has the highest importance for the PSP dataset, while it has the third-lowest importance for the BP dataset (S17A Fig). Also, the importance of features varies between SNVs and indels. For example, SD DP has the highest importance for SNVs in the BP dataset, but it has the third-lowest importance for indels (S17A and S17B Fig). Therefore, these results suggest that each feature may have a different contribution to classification results depending on sequencing datasets and types of genetic variants.

2.2.7 Performance of VQSR with different settings

For SNVs, GATK recommends three SNV call sets for training its VQSR model; 1) SNVs found in HapMap ("HapMap"), 2) SNVs in the omni genotyping array ("Omni"), and 3) SNVs in the 1000 Genomes Project ("1000G"). According to the VQSR parameter recommendation, SNVs in HapMap and Omni call sets are considered to contain only true variants, while SNVs in 1000G consist of both true- and false-positive variants [35]. We call this recommended parameter setting, "original VQSR." We, however, find that considering SNVs in Omni to contain both true- and false-positive variants considerably improves the quality of SNVs from VQSR for the BP dataset. We call this modified parameter setting, "Omni_Modified VQSR". Results show that the mean Ti/Tv on high-quality novel SNVs from Omni_Modified VQSR is 1.76, which is much higher than that from the original VQSR (1.41) and slightly higher than that from ForestQC (1.68) (S19A Fig). We also find that the mean number of total SNVs from Omni_Modified VQSR is 3.68M, which is much smaller than that from the original VQSR (3.99M) but higher than that from ForestQC (3.58M) (S19B Fig). In terms of other metrics, high-quality SNVs from original VQSR have a 3.66x higher ME rate, 7.40x higher genotype missing rate, and 1.16x higher genotype discordance rate (p-value = 0.0001118) than those SNVs from Omni_Modified VQSR (S19C–S19E Fig). Interestingly, we do not observe the improved performance of Omni_Modified VQSR in the PSP dataset as the mean Ti/Tv of high-quality novel SNVs from Omni_Modified VQSR is 1.23, which is slightly smaller than that of original VQSR (1.27) (S19A Fig). Nevertheless, individuals have fewer high-quality SNVs from Omni_Modified VQSR (3.53M) than that from original VQSR (3.75M) (S19B Fig). These results suggest that the performance of VQSR may change significantly depending on whether to consider a reference SNV call-set to contain only true-positive variants or both true- and false-positive variants, and it appears that the difference in performance is more noticeable in certain sequencing datasets than others.

Although Omni_Modified VQSR has slightly better Ti/Tv on high-quality novel SNVs and identifies more high-quality SNVs than does ForestQC, high-quality SNVs from Omni_Modified VQSR have 2.76x higher ME rate, 13.20x higher genotype missing rate, and 1.35x higher genotype discordance rate (p-value < 2.2e-16) than high-quality SNVs from ForestQC (S19C–S19E Fig). Hence, the results show that high-quality SNVs from ForestQC have higher quality than those from VQSR, even with the modified parameter setting.

2.3 Discussion

We developed an accurate and efficient method called ForestQC to identify a set of variants with high sequencing quality from NGS data. ForestQC combines the traditional filtering approach for performing QC in GWAS and the classification approach that uses a machine learning algorithm to classify whether a variant has good quality. ForestQC first uses stringent filters to identify high-quality and low-quality variants that unequivocally have high and low sequencing quality, respectively. ForestQC then trains a random forest classifier using the high-quality and low-quality variants obtained from the filtering step, and predicts whether a variant with ambiguous quality (an undetermined variant) is high-quality or low-quality in an unbiased manner. To evaluate ForestQC, we applied our method to two WGS datasets where one dataset consists of related individuals from families, while the other dataset has unrelated individuals. We demonstrated that high-quality variants identified from ForestQC in both datasets had higher quality than those from other approaches such as VQSR and a filtering approach based on ABHet.

To measure the performance of variant QC methods, one may apply these methods to benchmarking datasets where the true variants with high sequencing quality are verified. A few high-quality benchmarking variant sets have been released, including Genome In A Bottle (GIAB) [39], Platinum Genome (PlatGen) [40], and Syndip [41]. GIAB has seven samples, PlatGen sequenced 17 individuals and derived variant truth sets for two subjects, and Syndip includes only two cell lines, CHM1 and CHM13. The sample sizes of these datasets are very small, while we usually need to perform variant QC on an entire large dataset containing tens of millions of variants from hundreds of subjects or more. In order to apply ForestQC, the variant call-sets should have at least five subjects to calculate the statistics like SD DP and SD GQ accurately. Besides, it is recommended to apply VQSR to variant call-sets with more than 30 samples to achieve reliable results [35]. Thus, these datasets cannot be used as benchmarking datasets for variant QC. Apart, it is not expected to have a new benchmarking dataset with a large sample size soon because it is expensive to construct such a dataset. Hence, in this study, we used real WGS datasets to evaluate different approaches for variant QC. Their large sample sizes allow more accurate calculation of various quality metrics and statistics used by the variant QC methods, and therefore enable more reliable performance evaluation.

To measure the quality of variants, we used 21 sample-level metrics and 20 variant-level metrics, plus genotype missing rate, ME rate, and genotype discordance rate, resulting in a comprehensive evaluation of the performance of different methods. ME rate is found to be nearly linearly correlated with genotype errors [42–44], so it is a useful quality metric for variants with pedigree information. Low genotype missing rate has been considered as an indicator of high-quality variant call set as a variant with high genotype missing rate indicates poor genotyping or sequencing quality [45]. Also, high-quality variants would have the same genotypes generated by different genotyping technologies, such as sequencing and microarray. Thus, variant sequencing quality may be measured with the genotype discordance rate between microarray and sequencing. One challenge with this approach is that genotypes generated by microarray are usually available for only a small proportion of variants in the whole genome, especially for common and known variants, so genotype discordance rate cannot be used to show the quality of the entire variant call-set. Another frequently used variant quality metric is the Ti/Tv ratio [46–49]. It is expected to be around 2.0 for WGS data [23]. That is because transitions occur more frequently according to molecular mechanisms, although the number of transversions is twice as many as transitions. Previous studies found that mitochondrial DNA and some non-human DNA sequences might be biased towards transitions or transversions [50,51]. In this study, we only computed the Ti/Tv ratio for each QC method using the same human variant call set excluding mitochondria, in order to achieve an unbiased evaluation of all methods.

The main advantage of our approach over the traditional filtering approach is that our method does not attempt to classify variants with ambiguous sequencing quality (undetermined variants) using filters. It is difficult to determine the quality of variants using filters if their QC metrics (e.g., genotype missing rate) are close to the thresholds. Hence, ForestQC avoids a limitation of the traditional filtering approaches that determine the quality of every variant using filters, which may exclude some of the high-quality variants from the downstream analysis. We did not compare our approach with the traditional filtering approach used in GWAS that removes variants according to HWE p-values, ME rates, and genotype missing rates. One main reason is that the performance of this approach changes dramatically depending on filters and their thresholds, and there are numerous different thresholds of filters, as well as many combinations of filters that could be tested. Another reason is that its performance could be arbitrarily determined depending on the filters we use. For example, if one filter is to remove any variants having more than zero Mendel errors, the ME rate of high-quality variants would be zero, but we may be removing many other high-quality variants. In this study, we checked the accuracy of a filtering approach based on ABHet as ABHet is often used in performing QC of NGS data and is an important indicator for variant quality [26,52,53]. Also, as this approach is not based on standard QC metrics such as genotype missing rate, its performance is independent of those metrics, unlike the standard filtering approaches. We showed that our approach outperformed the ABHet approach as the high-quality variants from ForestQC have better quality than those from ABHet, regardless of the similar total number of high-quality variants, in terms of ME rate, missing rate, genotype discordance rate, and Ti/Tv ratio in the BP and PSP dataset.

Although our approach is similar to VQSR as both approaches train machine learning classifiers to predict the quality of variants, they have a few differences. First, our approach trains the model using high-quality and low-quality variants detected from sequencing data on which quality control is performed, while VQSR uses variants in existing databases, such as HapMap and 1000 genomes, as its training set. As VQSR uses previously known variants for model training, high-quality variants from VQSR are likely to contain more known (and likely to be common) variants than novel (and rare) variants. We showed in both WGS datasets that VQSR did indeed identify more common and known SNVs and indels as high-quality variants than ForestQC. This may not be a desirable outcome for some sequencing studies if one of their main goals is to identify rare and novel variants not captured in chips. Another difference between ForestQC and VQSR is the set of features used in the classifiers. While both methods use features related to sequencing depth and genotyping quality, VQSR uses some features calculated explicitly by GATK software, while ForestQC uses quality information reported in the standard VCF file. This suggests that our method is more generalizable than VQSR as it can be applied to VCF files generated from variant callers other than GATK. The last difference is the machine learning algorithms that ForestQC and VQSR use. ForestQC trains a random forest classifier while VQSR trains a Gaussian Mixture model. And we found that ForestQC was much faster than VQSR. (S11 Table).

In addition to SNVs, we applied ForestQC to indels in both WGS datasets and found that indels had much lower sequencing quality than do SNVs as the fraction of high-quality indels detected by ForestQC was considerably smaller than that of SNVs. This is somewhat expected because indel or structural variant calling is much more complicated than SNV calling from sequencing data, and some of them are likely to be false-positives [54,55]. It is, however, important to note that VQSR classifies many more indels as high-quality variants than does ForestQC or ABHet, but those high-quality indels from VQSR may not have high sequencing quality. We showed that high-quality indels from VQSR had similar Mendelian error rate to that without performing QC, indicating the poor performance of VQSR on indels. VQSR considers indels from Mills gold standard call set [37] as true-positives. Although those indels might represent true variant sites, it does not necessarily mean that genotyping on those sites is accurate. Therefore, genetic studies need to perform stringent QC on indels to remove those erroneous calls and not to have false-positive findings in their downstream analysis.

We found that the performance of VQSR was improved dramatically in the BP dataset when we considered SNVs in Omni genotyping array to have both true and false-positive sites, compared with when they were assumed to have all true sites. We, however, did not observe this performance enhancement in the PSP dataset. This suggests that users may need to try different parameter settings to obtain optimal results from VQSR for specific sequencing datasets they analyze. Another issue with VQSR and also with ABHet is that some high-quality SNVs or indels have high genotype missing rate and ME rate, which may not be suitable for the downstream analysis such as association analysis. Thus, those variants need to be filtered out separately, which means users may need to perform an additional filtering step in addition to applying VQSR and ABHet to the dataset. As the filtering step is incorporated in ForestQC, our method does not have this issue.

Our approach is an extension of a previous approach that uses a logistic regression model to predict the quality of variants in the BP dataset [30]. While our approach is similar to the previous approach in that they both combine filtering and classification approaches, ForestQC uses a random forest classifier that has higher accuracy than a logistic regression model, according to our simulation results. It includes more low-quality variants for model training, leading to predictions with fewer biases. ForestQC also includes more features than the previous approach as well as more filters to improve the quality of variants. Additionally, compared with the previous approach, ForestQC is more user-friendly and generalizable because users can choose or define different features and filters and tune the parameters according to their research goals.

We want to note that in addition to applying ForestQC, one may do variant calling with high-quality reference genome and accurate variant callers to obtain accurate variant call-sets. As we know, it is crucial to choose a state-of-art variant caller to minimize errors and biases in variant calling. Also, the quality of the reference genome may have an impact on the quality of the resulting variant call-sets [56]. The higher the quality of the reference genome is, the fewer low-quality calls in the variant call-sets are expected. If the quality of the reference genome is expected to be low, we suggest users modify filters or features in ForestQC. For instance, users may want to introduce new features describing the quality of the reference genome, such as an indicator of whether a mutation site is in the high-confidence regions of the reference genome. Then, ForestQC may learn how this information affects variant quality during training, and the performance of the random forest classifier may be improved based on this information.

ForestQC is efficient, modularized, and flexible with the following features. First, users are allowed to change thresholds for filters as needed. This is important because filters that are stringent for one dataset may not be stringent for another dataset. For example, variants from sequence data with a small sample size (e.g., < 100) may not have large enough statistical power to have significant HWE p-values, so higher p-value thresholds should be used, compared with studies with larger sample size. If filters are not stringent enough, there may be many low-quality variants, and ForestQC would train a very stringent classifier, leading to the possible removal of high-quality variants. On the contrary, if the filters are too stringent, there would be too few high-quality variants or low-quality variants, which would lower the accuracy of our random forest classifier. In this study, after the filtering step, 4.39% of SNVs and 15.72% of indels in the BP dataset, and 5.06% of SNVs and 15.66% of indels in the PSP dataset, were determined as low-quality variants. Empirically, we suggest filters for ForestQC such that after the filtering step, a fraction of low-quality variants is about 4–16%. Usually, we recommend the default parameter settings, which are the same sets of filters and features described in this paper. The selection of threshold values for these filters is based on our previous study for WGS data of extended pedigrees for bipolar disorder [30]. Second, users are allowed to use self-defined filters and features provided that they specify values for those new filters and features at each variant site, and our software also allows users to remove existing filters and features. As there may be filters and features that capture the sequencing quality of variants more accurately than the current set of filters and features, this option allows users to improve ForestQC further. For example, users can employ mappability, strand bias, and micro-repeats as features, instead of sequencing depth and genotyping quality used in this study, because DP and GQ might penalize disease-causing variants with low coverage. Also, if users want to obtain more variants after QC, they may lower the standard for high-quality variants, that is, increase the threshold values of ME or missing rate for determining high-quality variants. Third, ForestQC generates the probability of each undetermined variant being a high-quality variant. This probability needs to be higher than a certain threshold for an undetermined variant to be predicted to be high-quality. It can also be used to analyze the sequencing quality of individual variants. If studies find that a particular undetermined variant is associated with a phenotype, they may consider checking whether its probability of being a high-quality variant is high enough. Lastly, ForestQC allows users to change the probability threshold for determining whether each undetermined variant is high-quality or low-quality. Users may lower this threshold if they are interested in obtaining more high-quality variants at the cost of including more low-quality variants.

2.4 Materials and methods

2.4.1 ForestQC

ForestQC consists of two approaches: a filtering approach and a machine learning approach based on a random forest algorithm.

Filtering.

Given a variant call set from next-generation sequencing data, ForestQC first applies several stringent filters to identify high-quality, low-quality, and undetermined variants. High-quality variants are ones that pass all filters, while low-quality variants fail any of them (S1 and S2 Tables). The undetermined variants are variants that neither pass filters for high-quality variants nor fail filters for low-quality variants. We use the following filters in the filtering step.

- Mendelian error (ME) rate. The Mendelian error occurs when a child's genotype is inconsistent with genotypes from parents. ME rate is calculated as the number of ME among all trios divided by the number of trios for a given variant. Note that this statistic is only available for family-based data.
- Genotype missing rate. This is the proportion of missing alleles in each variant.
- Hardy-Weinberg equilibrium (HWE) p-value. This is a p-value for hypothesis testing whether a variant is in Hardy-Weinberg equilibrium. Its null hypothesis is that the variant is in Hardy-Weinberg equilibrium. We use the algorithm from open-source software, VCFtools [57], for the calculation of Hardy-Weinberg equilibrium p-value.
- ABHet. This is the allele balance for heterozygous calls. ABHet is calculated as the number of reference reads from individuals with heterozygous genotypes divided by the

total number of reads from such individuals, which is supposed to be 0.50 for high-quality bi-allelic variants. For variants in chromosome X, we only calculate ABHet for females.

Random forest classifier.

Random forest is a machine learning algorithm that runs efficiently on large datasets with high accuracy [34]. Briefly, random forest builds several randomized decision trees, each of which is trained to classify the input objects. For the classification of a new object, the fitted random forest model passes the input vector down to each of the decision trees in the forest. Each decision tree has its classification result, and then the forest would output the classification that the majority of the decision trees make. To balance efficiency and accuracy, we train a random forest classifier using 50 decision trees (S2 Fig) and a probability threshold of 50% (S3 Fig).

To train random forest, we use high-quality and low-quality variants identified from the previous filtering step as a training dataset, after balancing their sample size by random sampling. Normally, high-quality variants are much more numerous than low-quality variants, so we randomly sample from high-quality variants with the sample size of low-quality variants. Hence, the sample size of the balanced training set would be twice as large as the sample size of low-quality variants. We also need features in training a random forest, which characterize datasets, and we use the following features.

- Mean and standard deviation of depth (DP) and genotyping quality (GQ). The depth and genotyping quality values are extracted from DP and GQ fields of each sample in VCF files, respectively, and mean and standard deviation are calculated over all samples for each variant.

- Outlier depth and outlier genotype quality. These are the proportions of samples whose DP or GQ is lower than a particular threshold. We choose this threshold as the first quartile value of all DP or GQ values of variants on chromosome 1. We use DP and GQ of variants on only chromosome 1 to reduce the computational costs.
- GC content: We first split a reference genome into windows with a size of 1,000 bp and calculate GC content for each window as (# of G or C alleles) / (# of A, G, C, or T alleles). Then, each variant is assigned a GC content value according to its position in the reference genome.

After training random forest with the training dataset using the above features, we next use the fitted model to make predictions on undetermined variants on being high-quality variants. Undetermined variants with the predicted probability of being high-quality larger than 50% are labeled as predicted high-quality variants. Then the predicted high-quality variants and high-quality variants from the previous filtering step are combined as the final set of high-quality variants. We apply the same procedure to identify low-quality variants.

2.4.2 Comparison of different machine learning algorithms

We compare eight different machine learning algorithms to identify the best algorithm used for ForestQC. They are 1) k-nearest neighbors for supervised two-class classification (eight threads); 2) logistic regression (eight threads); 3) single support vector machine with Gaussian kernel function and penalty parameter C of 1.0 (one thread); 4) random forest with 50 trees (eight threads); 5) naïve Bayes without any prior probabilities of the classes (one thread); 6) artificial neural network with sigmoid function as activation function (eight threads). It has one hidden layer with ten units; 7) AdaBoost with 50 estimators and learning rate of 1.0, which uses SAMME.R real boosting algorithm (one thread); 8) and quadratic discriminant analysis without any prior on classes. Its regularization is 0, and its threshold for rank estimation is 1e-4 (one thread). Other parameters of these machine learning algorithms are the default, as described in the documentation of the Python scikit-learn package [58]. All learning algorithms use the seven features as mentioned earlier: mean and standard deviation of sequencing depth, mean and standard deviation of genotype quality, outlier depth, outlier quality, and GC content.

To test these eight machine learning algorithms, we obtain training and test datasets from the BP dataset, using filters described in S1 and S2 Tables. There are 21,248,103 high-quality SNVs and 2,257,506 high-quality indels while there are 1,100,325 low-quality SNVs and 624,965 low-quality indels. We sample 100,000 variants randomly from high-quality variants and 100,000 variants from low-quality variants to generate a training set. Similarly, 100,000 high-quality variants and 100,000 low-quality variants are randomly chosen from the rest of the variants to form a test set. Each machine learning model shares the same training and test sets. We train the machine learning models and measure training time at a training stage, and then test their accuracy and measure prediction time at a testing stage. We measure the runtime of each algorithm, which is the elapsed clock time between the start and end of each algorithm. To assess the performance of each algorithm, we compute the F1-score for the test set. F1-score is the harmonic average of precision and recall, which is calculated as $2 \cdot precision \cdot \frac{recall}{(precision+recall)}$. The closer F1-score is to 1, the higher classification accuracy is. Recall is the fraction of true-positive results over all samples that should be given a positive prediction. Precision is the number of true-positive results divided by the number of positive results predicted by the classifier. We also measure the model accuracy using 10-fold cross-validation, as well as the area under the receiver operating characteristic curve.

2.4.3 ABHet approach and VQSR

We compare ForestQC with two other approaches for performing QC on genetic variants. One is a filtering approach based on ABHet, and the other is a classification approach called VQSR from GATK software. For the ABHet approach, we consider variants with ABHet > 0.7 or < 0.3 as low-quality variants, and the rest as high-quality variants. We chose this threshold setting of ABHet (> 0.3 and < 0.7) because the ADSP project could not reliably confirm heterozygous calls with ABHet > 0.7 with Sanger sequencing [26]. We also exclude variants with small ABHet values (< 0.3) to ensure high quality. For GATK, we use recommended arguments as of 2018-04-04 [35]. For SNVs, VQSR takes SNVs in HapMap 3 release 3, 1000 Genome Project and Omni genotyping array as training resources, and dbSNP135 as known site resource. HapMap and Omni sites are considered as true sites, meaning that SNVs in these datasets are all true variants, while 1000 Genome Project sites are regarded as false sites, meaning that there could be both true and false-positive variants. The desired level of sensitivity of true sites is set to be 99.5%. In the BP dataset, we run VQSR version 3.5-0-g36282e4 with following annotations; quality by depth (QD), RMS mapping quality (MQ), mapping quality rank sum test (MQRankSum), read position rank sum test (ReadPosRankSum), fisher strand (FS), coverage (DP) and strand odds ratio (SOR) to evaluate the likelihood of true-positive calls. In the PSP dataset, we use VQSR version 3.2-2-gec30cee that uses all annotations above except for SOR and additional inbreeding coefficient (InbreedingCoeff) because variants in PSP dataset do not have the SOR annotation. For indels, VQSR takes indels in Mills gold standard call set [37] as a true training resource and dbSNP135 as a known site resource. The desired level of sensitivity of true sites is set to be 99.0%. We use VQSR version 3.5-0-g36282e4 with QD, DP, FS, SOR, ReadPosRankSum, and MQRankSum annotations to evaluate the likelihood of true-positive calls in the BP dataset, while we run VQSR version 3.2-2-gec30cee with the same annotations except for SOR and additional InbreedingCoeff for the PSP dataset.

2.4.4 BP and PSP WGS datasets

The BP WGS dataset is for studying bipolar disorder whose average coverage is 36-fold. This study recruited individuals from 11 Colombia (CO) and 15 Costa Rica (CR) extended pedigrees in total. 454 subjects from 10 CO and 12 CR families are both whole-genome sequenced and genotyped with microarray. There are 144 individuals diagnosed with BP1 and 310 control samples that are unaffected or have non-BP traits. We use the highly scalable Churchill pipeline [59] to do the variant calling for the BP data set, where GATK-HaplotypeCaller 3.5-0-g36282e4 is used as the variant caller according to the GATK best practices [23], and the reference genome is HG19. After initial QC on individuals, five individuals are removed because of poor sequencing quality and possible sample mix-ups. Finally, 449 individuals are included in an analysis, resulting in 25,081,636 SNVs and 3,976,710 indels. 1,814,326 SNVs in the WGS dataset are also genotyped with microarray, which are used to calculate the genotype discordance rate. In this study, we use the BP dataset before any QC performed on genetic variants. In a previous study [30], genetic variants in the BP WGS dataset are first processed with VQSR and then filtered with a trained logistic regression model to remove variants with low quality.

The PSP WGS dataset is for studying progressive supranuclear palsy with an average coverage of 29-fold. 544 unrelated individuals are whole-genome sequenced, 518 of whom are also genotyped with microarray. Among them, 119 individuals have 547,644 SNPs, and 399 individuals have 1,682,489 SNPs genotyped with microarray, respectively. That 119 individuals would be excluded when calculating the genotype discordance rate in case of biases caused by fewer SNPs. There are 356 individuals diagnosed with PSP and 188 individuals as controls.

Variant calling for the PSP dataset is performed using the Churchill pipeline, where GATK-HaplotypeCaller 3.2-2-gec30cee is used as the variant caller according to the GATK best practices, and the reference genome is HG19. Forty-nine samples are found to have high missing rate, high relatedness with other samples, or are diagnosed with diseases other than PSP, so they are removed. Next, we extract variant data with only 495 individuals with VCFtools. Monomorphic variants are then removed. After preprocessing, the PSP WGS dataset has 33,273,111 SNVs and 5,093,443 indels. There are 1,682,489 SNVs from 381 samples genotyped by both microarray and WGS, which are used for calculating genotype discordance rate.

2.4.5 Performance metrics

Twenty-one sample-level metrics and twenty variant-level metrics are defined to measure the sequencing quality of the variant call-sets after quality control (S12 Table). Note that we do not show all sample-level metrics and variant-level metrics in the main text. Other metrics are available in supplemental materials. Variant-level metrics provide us with a summarized assessment report of the sequencing quality of a variant call set, such as total SNVs of the whole dataset. They are calculated based on the information of all variants in a variant call set. For example, the number and the proportion of multi-allelic SNVs are calculated for the entire dataset. On the other hand, sample-level metrics enable the inspection of the sequencing quality for sequenced individuals in a variant call set. For instance, we check the distribution of novel Ti/Tv or other quality metrics among all individuals in the study. Sample-level metrics are calculated for each sample, using its genotype information on all variants in the dataset. The distribution of those metrics across all individuals is shown as a box plot. For example, the number of SNV singletons on a sample level shows the distribution of the number of SNV singletons across all sequenced individuals. In this study, both sample-level and variant-level metrics are used to evaluate the sequencing quality of WGS variant datasets.

Additionally, we use genotype missing rate, ME rate and genotype discordance rate as variant quality metrics, which are computed using the entire variant call set. The definitions of genotype missing rate and ME rate have been described above. Note that ME rate is only available for family-based datasets, such as the BP dataset, so we do not calculate ME rate for the PSP dataset that only includes unrelated individuals. Genotype discordance rate is the proportion of individuals whose genotypes are inconsistent between next-generation sequencing and microarray. This metric can only be calculated with a subset of variants due to the limited number of variants genotyped by both sequencing and microarray. Note that microarray might also have biases in genotyping, leading to some limitations of genotype discordance rate. For example, microarray usually genotype selected variants, primarily common and known variants, so genotype discordance rate is only available for these selected variants, and it cannot provide quality evaluation for all variants, especially rare variants. Genotype missing rate, ME rate and genotype discordance rate provide us with an accurate evaluation of variant quality because true-positive variants with high quality are very likely to have low values of these three metrics.

Tables

Table 2.1: Performance of eight different machine learning algorithms

Machine learning algorithm	Time cost (sec)	F1-score for indel classification	F1-score for SNV classification
Random Forest	9.85	0.9428	0.9740
ANN	75.34	0.9400	0.9707
SVM	1253.48	0.9381	0.9704
AdaBoost	25.27	0.9270	0.9672
Logistic Regression	2.49	0.9074	0.9668
KNN	24.71	0.9200	0.9486
QDA	0.30	0.9006	0.9241
Naïve Bayes	0.18	0.8716	0.9012

Performance metrics, including F1-scores, total time cost of model fitting and prediction, are ranked by F1-score for SNV classification. Random forest, ANN, logistic regression and KNN are set to run with eight threads. “ANN”: artificial neural network. “SVM”: single support vector machine. “KNN”: K-nearest neighbors classifier. “QDA”: quadratic discriminant analysis

Table 2.2: Variant-level quality metrics of high-quality variants in the BP dataset processed by different methods

Metric	No QC	ABHet	VQSR	ForestQC
Total SNVs	25081636	22415368	24239357	22227503
Known SNVs	21165051	19665276	20675746	19361635
Known SNVs (%)	84.38%	87.73%	85.30%	87.11%
Total indels	3976710	2670647	3212886	2789037
Known indels	3094271	2188996	2758783	2237002
Known indels (%)	77.81%	81.97%	85.87%	80.21%
Multi-allelic SNVs	153836	26549	128894	77693
Multi-allelic SNVs (%)	0.61%	0.12%	0.53%	0.35%

Four methods are compared, including no QC applied, ABHet approach, VQSR and ForestQC. “Known” stands for variants found in dbSNP. The version of dbSNP is 150.

Table 2.3: Variant-level quality metrics of high-quality variants in the PSP dataset processed by four different methods

Metric	No QC	ABHet	VQSR	ForestQC
Total SNVs	33273111	29771182	31281620	29352329
Known SNVs	25960464	24142744	24910728	23514257
Known SNVs (%)	78.02%	81.09%	79.63%	80.11%
Total indels	5093443	3311136	3682319	3418242
Known indels	3679990	2532899	3012662	2567879
Known indels (%)	72.25%	76.50%	81.81%	75.12%
Multi-allelic SNVs	250418	6685	188180	146247
Multi-allelic SNVs (%)	0.75%	0.02%	0.60%	0.50%

Four methods are compared, including no QC applied, ABHet approach, VQSR and ForestQC. “Known” stands for variants found in dbSNP. The version of dbSNP is 150.

Figures

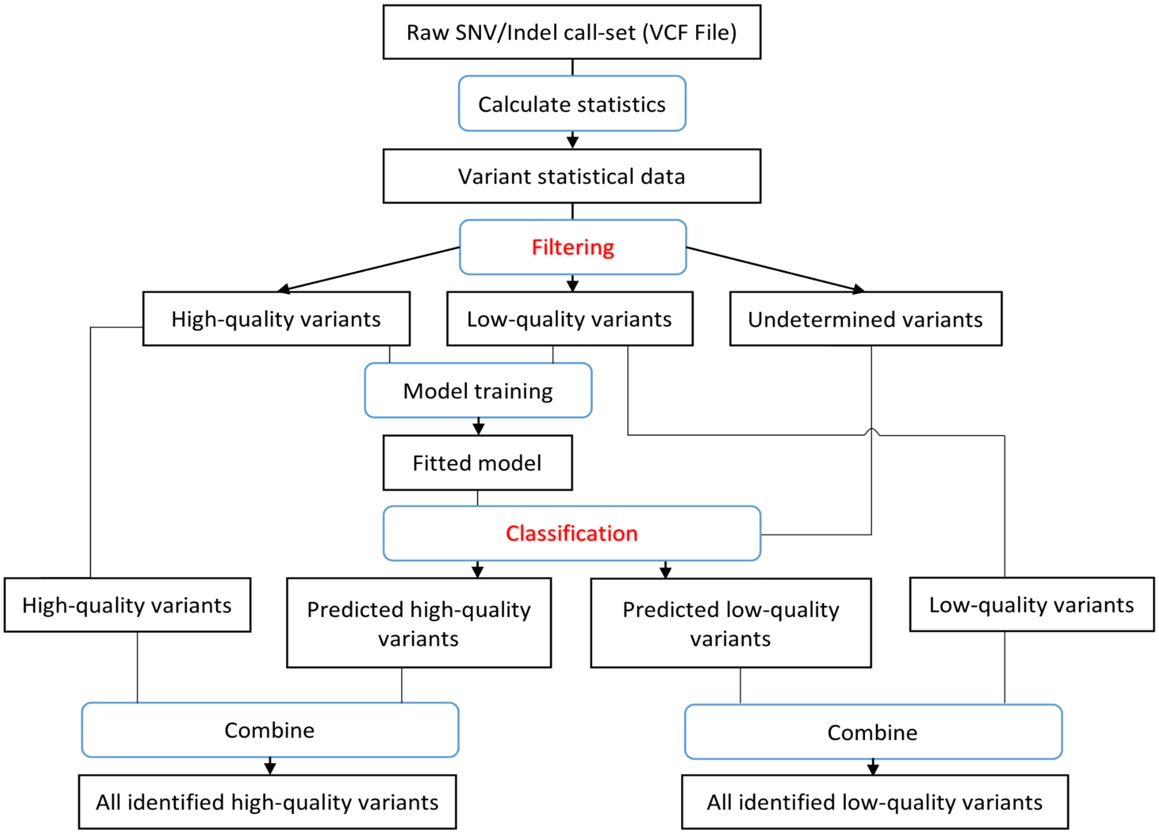

Figure 2.1: Workflow of ForestQC. ForestQC takes a raw variant call set in the VCF format as input. Then it calculates the statistics of each variant, including MAF, mean depth, mean genotyping quality. In the filtering step, it separates the variant call set into high-quality, low-quality, and undetermined variants by applying various hard filters, such as Mendelian error rate and genotype missing rate. In the classification step, high-quality and low-quality variants are used to train a random forest model, which is then applied to assign labels to undetermined variants. Variants predicted to be high-quality among undetermined variants are combined with high-quality variants from the classification step for the final set of high-quality variants. The same procedure applies to find the final set of low-quality variants.

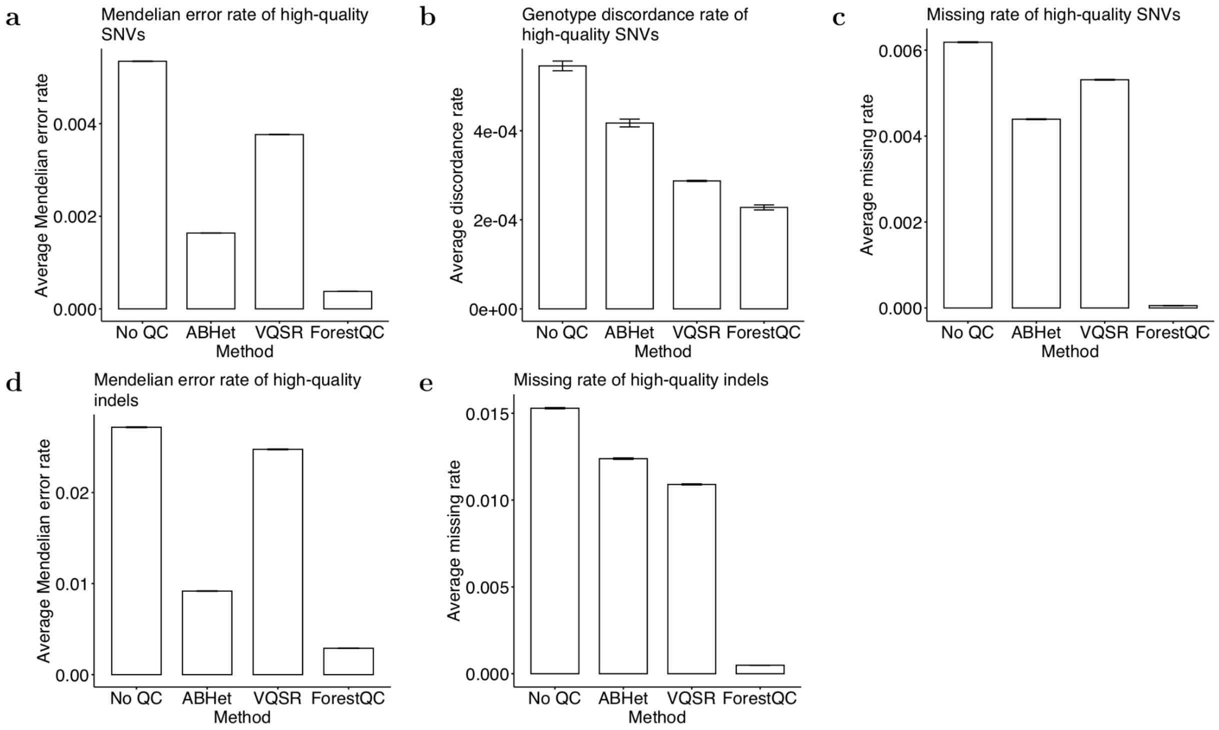

Figure 2.2: Overall quality of high-quality variants in the BP dataset detected by four different methods. (a) The ME rate, (b) the genotype discordance rate, and (c) the missing rate of high-quality SNVs. (d) The ME rate and (e) the missing rate of high-quality indels. Data are represented as the mean ± SEM.

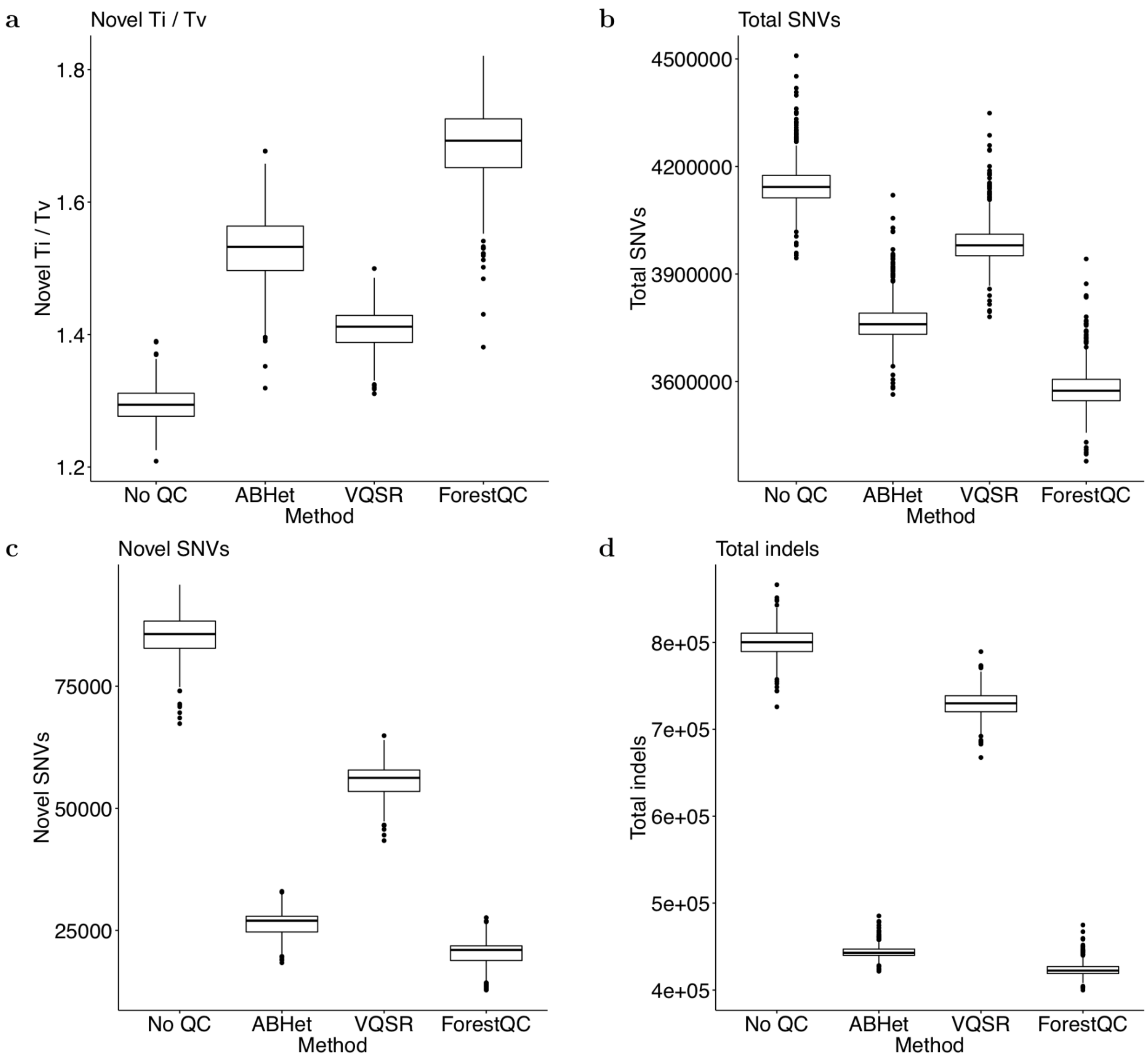

Figure 2.3: Sample-level quality metrics of high-quality variants in the BP dataset identified by four different methods. (a) Ti/Tv ratio of SNVs not found in dbSNP. (b) The total number of SNVs. (c) The number of SNVs not found in dbSNP. (d) The total number of indels. The version of dbSNP is 150.

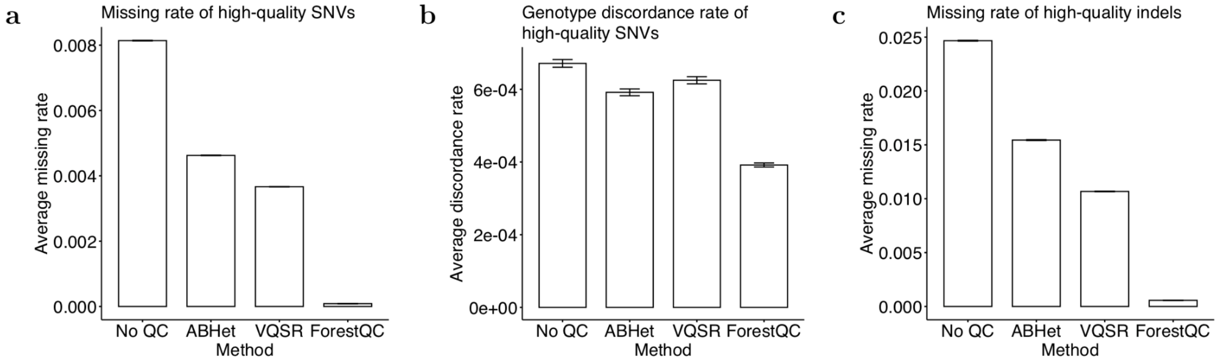

Figure 2.4: Overall quality of high-quality variants in the PSP dataset detected by four different methods. (a) The missing rate and (b) the genotype discordance rate of high-quality SNVs. (c) The missing rate of high-quality indels. Data are represented as the mean ± SEM.

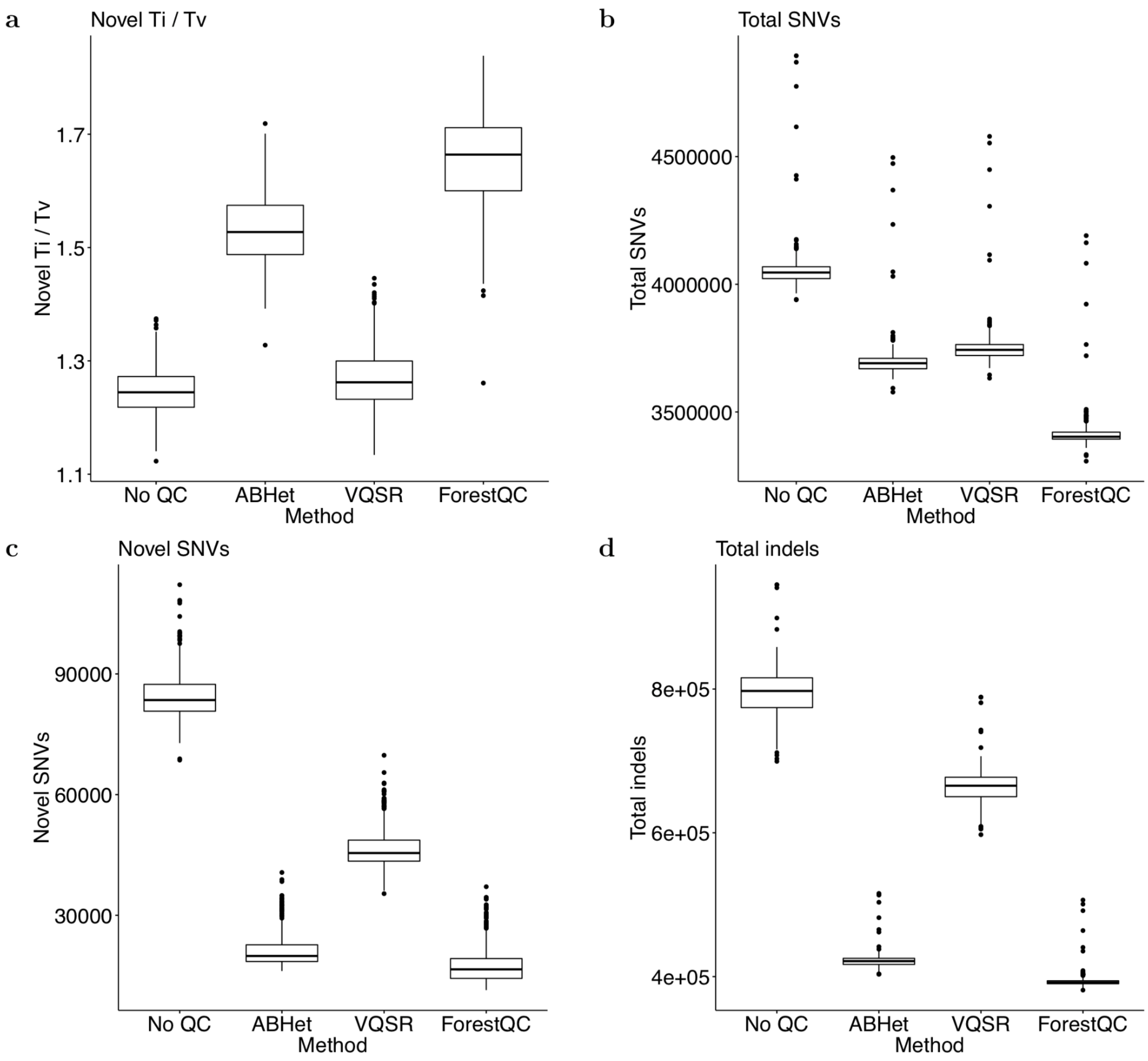

Figure 2.5: Sample-level quality metrics of high-quality variants in the PSP dataset identified by four different methods. (a) Ti/Tv ratio of SNVs not found in dbSNP. (b) The total number of SNVs. (c) The number of SNVs not found in dbSNP. (d) The total number of indels. The version of dbSNP is 150.

Reference to the published article

Jiajin Li, Brandon Jew, Lingyu Zhan, Sungoo Hwang, Giovanni Coppola, Nelson B. Freimer, and Jae Hoon Sul. 2019. ForestQC: quality control on genetic variants from next-generation sequencing data using random forest. *PLoS Computational Biology*. 15 (12): e1007556.

2.5 References

1. Pray L. Genome-wide association studies and human disease networks. Nat Educ. nature.com; 2008;1(1):220.
2. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005 Feb;6(2):95–108.
3. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul;511(7510):421–7.
4. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature. search.proquest.com; 2007 Feb;445(7130):881–5.
5. Ng MCY, Shriner D, Chen BH, Li J, Chen W-M, Guo X, et al. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. PLoS Genet. journals.plos.org; 2014 Aug;10(8):e1004517.
6. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson’s disease. Nat Genet. nature.com; 2014 Sep;46(9):989–93.

7. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012 Jan;90(1):7–24.

8. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009 Apr;360(17):1696–8.

9. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415–25.

10. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008 Jun;40(6):695–701.

11. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009 Jun;19(3):212–9.

12. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001 Jul;69(1):124–37.

13. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011 Oct;43(11):1066–73.

14. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet. 2012 Dec;44(12):1326–9.

15. Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang Z-Z, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet. 2014 Feb;94(2):233–45.

16. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, et al. Using whole-exome sequencing to identify inherited causes of autism. Neuron. 2013 Jan;77(2):259–73.

17. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008 Sep;36(16):e105.

18. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011 Jul;39(13):e90.

19. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013 May;14(5):R51.

20. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016 Mar;17:125.

21. Manley LJ, Ma D, Levine SS. Monitoring Error Rates In Illumina Sequencing. J Biomol Tech. 2016 Dec;27(4):125–8.

22. Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov M V, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Sci Rep. 2014 Mar;4:4532.

23. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491–8.

24. Highnam G, Wang JJ, Kusler D, Zook J, Vijayan V, Leibovich N, et al. An analytical framework for optimizing variant discovery from personal genomes. Nat Commun. 2015 Feb;6:6275.

25. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. genomemedicine.biomedcentral. …; 2013 Mar;5(3):28.

26. ADSP. Review and Proposed Actions for False-Positive Association Results in ADSP Case-Control Data | ADSP [Internet]. https://www.niagads.org/adsp/content/review-and-proposed-actions-false-positive-association-results-adsp-case-control-data. 2016. Available from: https://www.niagads.org/adsp/content/review-and-proposed-actions-false-positive-association-results-adsp-case-control-data

27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

28. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

29. International HapMap Consortium. The International HapMap Project. Nature. 2003 Dec;426(6968):789–96.

30. Sul JH, Susan K Service, Huang AY, Ramensky V, Hwang S-G, Teshiba TM, et al. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. bioRxiv. 2018 Jul. doi: 10.1101/363267

31. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011 Sep;29(10):908–14.

32. Wang W, Wei Z, Lam T-W, Wang J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci Rep. 2011 Aug;1:55.

33. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011 Feb;12(2):R18.

34. Breiman L. Random Forests. Mach Learn. 2001 Oct;45(1):5–32.

35. GATK Dev Team. Which training sets / arguments should I use for running VQSR? https://software.broadinstitute.org/gatk/documentation/article.php?id=1259. 2017 Sep;

36. Collins DW, Jukes TH. Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics. 1994;20: 386–396.

37. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011 Jun;21(6):830–9.

38. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct;526(7571):75–81.

39. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype

calls. Nat Biotechnol [Internet]. 2014 Mar 16 [cited 2019 Feb 13];32(3):246–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24531798

40. Eberle MA, Fritzilas E, Krusche P, Källberg M, Moore BL, Bekritsky MA, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome Res [Internet]. 2017 Jan [cited 2019 Feb 13];27(1):157–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27903644

41. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods [Internet]. NIH Public Access; 2018 Aug [cited 2019 Feb 13];15(8):595–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30013044

42. Saunders IW, Brohede J, Hannan GN. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference. Genomics [Internet]. Academic Press; 2007 Sep 1 [cited 2018 Dec 14];90(3):291–6. Available from: https://www.sciencedirect.com/science/article/pii/S088875430700136X

43. Sobel E, Papp JC, Lange K. Detection and Integration of Genotyping Errors in Statistical Genetics. Am J Hum Genet [Internet]. Cell Press; 2002 Feb 1 [cited 2018 Dec 14];70(2):496–508. Available from: https://www.sciencedirect.com/science/article/pii/S0002929707639627

44. Hao K, Li C, Rosenow C, Hung Wong W. Estimation of genotype error rate using samples with pedigree information—an application on the GeneChip Mapping 10K array. Genomics [Internet]. Academic Press; 2004 Oct 1 [cited 2018 Dec 14];84(4):623–30. Available from: https://www.sciencedirect.com/science/article/pii/S0888754304001193

45. Hackett CA, Broadfoot LB. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity (Edinb) [Internet]. Nature Publishing Group; 2003 Jan 9 [cited 2018 Dec 14];90(1):33–8. Available from: http://www.nature.com/articles/6800173

46. Durbin RM, Altshuler DL, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature [Internet]. Nature Publishing Group; 2010 Oct 28 [cited 2018 Dec 14];467(7319):1061–73. Available from: http://www.nature.com/doifinder/10.1038/nature09534

47. Guo Y, Zhao S, Sheng Q, Ye F, Li J, Lehmann B, et al. Multi-perspective quality control of Illumina exome sequencing data using QC3. Genomics [Internet]. Academic Press; 2014 May 1 [cited 2019 Jan 12];103(5–6):323–8. Available from: https://www.sciencedirect.com/science/article/pii/S0888754314000354

48. Guo Y, Li J, Li C-I, Long J, Samuels DC, Shyr Y. The effect of strand bias in Illumina short-read sequencing data. BMC Genomics [Internet]. 2012 [cited 2018 Dec 14];13(1):666. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-666

49. Guo Y, Long J, He J, Li C-I, Cai Q, Shu X-O, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics [Internet]. 2012 [cited 2018 Dec 14];13(1):194. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-194

50. Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, et al. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A

[Internet]. National Academy of Sciences; 2008 Jul 8 [cited 2018 Dec 14];105(27):9272–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18583475

51. Lanave C, Tommasi S, Preparata G, Saccone C. Transition and transversion rate in the evolution of animal mitochondrial DNA. Biosystems [Internet]. Elsevier; 1986 Jan 1 [cited 2018 Dec 14];19(4):273–83. Available from: https://www.sciencedirect.com/science/article/pii/0303264786900043

52. Aylward A, Cai Y, Lee A, Blue E, Rabinowitz D, Haddad Jr J, et al. Using Whole Exome Sequencing to Identify Candidate Genes With Rare Variants In Nonsyndromic Cleft Lip and Palate. Genet Epidemiol. 2016 Jul;40(5):432–41.

53. Bellenguez C, Charbonnier C, Grenier-Boley B, Quenez O, Le Guennec K, Nicolas G, et al. Contribution to Alzheimer’s disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. Neurobiol Aging. 2017 Nov;59:220.e1--220.e9.

54. Tattini L, D’Aurizio R, Magi A. Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Front Bioeng Biotechnol [Internet]. 2015 Jun;3:92. Available from: http://journal.frontiersin.org/Article/10.3389/fbioe.2015.00092/abstract

55. Hasan MS, Wu X, Zhang L. Performance evaluation of indel calling tools using real short-read data. Hum Genomics. 2015 Aug;9:20.

56. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8: 61–65.

57. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug;27(15):2156–8.

58. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(Oct):2825–30.

59. Kelly BJ, Fitch JR, Hu Y, Corsmeier DJ, Zhong H, Wetzel AN, et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. Genome Biol. 2015 Jan;16:6.

Chapter 3 Detecting the regulatory effects of rare variants in multiple tissues

3.1 Introduction

Over the past decade, genome-wide association studies (GWAS) have successfully discovered numerous associations between common genetic variants and human complex diseases and traits[1,2]. These studies also found that those common variants typically have small effects and explain a small fraction of heritability[3,4]. Motivated by this finding, many sequencing studies have attempted to identify rare variants associated with complex traits[5,6]. It is hypothesized that rare variants may have larger effect sizes than common variants due to purifying selection and may explain some of the missing heritability[7,8]. Candidate-gene and large-scale sequencing studies have indeed found associations of rare variants with complex diseases and traits[9–11].

An important question after finding the associations of rare variants is to understand their genetic mechanism on how they influence diseases. GWAS have found that common variants associated with diseases are mostly present in non-coding regions of the genome, suggesting that they might affect traits by regulating the expression of nearby genes as recent expression quantitative trait loci (eQTL) studies have identified many common variants with regulatory effects[12,13]. However, the effect of rare variants on gene expression remains mostly obscure, although there have been recent developments in this work. For example, Li et al. discovered that rare variants might result in outlier patterns of over or under expression across multiple human tissues[14] while Zhao et al. found an excess of rare variants was significantly associated with extreme gene expression in human peripheral blood[15]. Hernandez et al. also reported that ultrarare variants make a significant contribution to the heritability of gene expression[16]. These results hint at the possible functional effect of rare variants.

To discover the functional effect of genetic variants, many eQTL studies are interested in identifying genes whose expression levels are influenced by genetic variants (called "eGenes"). The aforementioned studies for rare variants mostly focused on the overall contribution of rare variants to gene expression but did not find individual genes whose expression is associated with rare variants. We call these genes "RV eGenes," and there are two major challenges in finding RV eGenes. The first is relatively small sample sizes of eQTL studies that collected whole-genome sequencing (WGS) as well as RNA-seq data. WGS data instead of whole-exome sequencing data are necessary to discover RV eGenes as many variants regulating gene expression may be present in non-coding regions of the genome. The second challenge is the statistical approach to detect eGenes. While there have been several methods developed to identify common variant eGenes (CV eGenes)[17,18], these methods utilize a single marker test that tests each SNP, which yields low statistical power for rare variants. To increase power to detect association of rare variants, many collapsing approaches that combine the effect of multiple rare variants have been proposed[19,20], but as we will discuss later, they are not optimized to find RV eGenes.

In this paper, we develop a powerful approach called LRT-q to detect RV eGenes and apply this method to WGS and multi-tissue RNA-seq data collected from 681 European individuals in the Genotype-Tissue Expression (GTEx) project (v8). LRT-q incorporates functional annotations of rare variants, observational genotype data, and quantitative phenotype data to identify a group of potential causal rare variants influencing the expression of a nearby gene by aggregating statistics of rare variants in a nonlinear manner. We show using extensive simulations that LRT-q outperforms previous methods for rare variant association testing such as SKAT-O[21] and variable threshold[22]. We also find that LRT-q detects more RV eGenes than previous methods in the GTEx data across all tissues. We investigate the characteristics of those RV eGenes and discover a few important biological insights such as higher tissue specificity of RV eGenes compared to CV eGenes and enrichment of RV eGenes in disease-associated genes. We provide an open-source R package implementing the proposed method, LRTq.

3.2 Overview of LRT-q

An association test between a single rare variant and expression of a gene is likely to result in low statistical power because the power decreases as allele frequency of a variant decreases. To overcome this challenge, many statistical approaches have been developed to aggregate rare variants in a genetic region, like a gene, and to test their cumulative effects on a phenotype. The underlying rationale is that a gene can be regulated by multiple rare variants and thus a larger effect can be observed by grouping them, contributing to increased power. The methods for rare variant testing include burden tests like variable threshold (VT)[22], variance component tests like sequence kernel association test (SKAT)[23], and combined tests like SKAT-O[21,24].

These methods, however, may not be optimal in detecting the effects of rare variants because of the following two reasons. First, they do not attempt to prioritize likely causal variants. As the rare variant methods combine multiple variants, it is important to remove the effects of non-causal variants. Previous methods mostly rely on functional information of variants to prioritize variants[20,25] such as minor allele frequency (MAF) and Combined Annotation Dependent Depletion (CADD) scores[26] as it has been hypothesized that rarer variants may have larger effects than more common variants. However, we may be able to prioritize potential causal variants more accurately by using both functional information and genotype data where the latter may provide additional information on the causal statuses of variants. For example, for gene expression data, individuals with causal rare variants may have significantly different expression patterns from other individuals. The other reason why previous methods may not be optimal is that many of the burden tests combine statistics of multiple variants linearly (e.g. a weighted sum of z-scores). However, it may be desirable to combine the statistics in a nonlinear manner to detect more associations as we show in our results.

To overcome these limitations, we propose a likelihood ratio test for quantitative traits (LRT-q) for detecting rare variants associated with gene expression. This method is an extension of the original LRT[27] that was designed for identifying associations between rare variants and disease status (dichotomous traits). There are two underlying models in LRT-q; 1) the null model that assumes no causal variants among all rare variants, and 2) the alternative model that assumes at least one causal variant. LRT-q calculates a likelihood ratio statistic between the two models and also a p-value using a permutation test. LRT-q calculates the statistic using functional information and observational genotype data that allows LRT-q to prioritize potential causal rare variants. Besides, LRT-q aggregates statistics of multiple rare variants nonlinearly to boost statistical power (see Materials and Methods). Assuming that individuals carrying a rare allele of a variant have different gene expression patterns from those carrying a different allele, we calculate a statistic measuring this difference for each rare variant near a gene. We then combine these statistics from multiple rare variants nonlinearly and generate an aggregated statistic for the gene. LRT-q considers both positive and negative effect sizes of genetic variants on gene expression, and it is very efficient using an adaptive permutation test.

3.3 Verification and comparison

3.3.1 False positive rate of LRT-q

To measure the performance of LRT-q, we first measure the false positive rate using simulated data under the null hypothesis of no causal variants (see Materials and Methods). Each simulation has 1,000 individuals and 33 rare variants on average, and we test several other rare variant association methods such as CMC[28], WSS[29], Burden, VT[22], ACAT-V, ACAT-O[20], and SKAT-O[21] in addition to LRT-q. Here, ACAT-O is an omnibus test constructed by combining p-values of VT, ACAT-V, and SKAT-O. Results show that all methods have well-controlled false positive rates across different significance thresholds such as $\alpha = 0.05$, 0.01, 0.001, and 0.0001 (Table 1).

3.3.2 Power of LRT-q.

Next, we perform power simulation under the alternative hypothesis that there is at least one causal rare variant using several combinations of effect sizes and causal ratio where causal ratio defines the percentage of causal variants among all rare variants (see Materials and Methods). In simulations, half of the causal variants have positive effect sizes, and the rest of causal variants have negative effect sizes as rare variants might increase or decrease expression levels. Additionally, as only a few rare variants might be causal, only 3% to 10% of rare variants are causal in the simulation data. Regarding effect sizes, variants with lower allele frequency have larger effect sizes, which is the assumption often made in simulating the effect of rare variants, and we simulate several different maximum effect sizes of rare variants (from 0.99 to 4.95). For each combination of effect size and causal ratio, we generate 10,000 datasets containing 1,000 subjects. Similar to the false positive rate simulation, we test eight methods, and the power is measured at $\alpha = 0.05$.

Results show higher power of LRT-q over other methods across a variety of simulation settings (Fig 1). Especially, we observe that as the effect size or the proportion of causal variants increases, LRT-q becomes more powerful than other approaches. When 10% of rare variants are causal, LRT-q has the highest statistical power if the maximum effect size is larger than or equal to 1.98. Its power is 147% to 224% as high as the power of SKAT-O, the second most powerful method. Furthermore, the power of LRT-q is slightly smaller or as large as the power of SKAT-O when the effect size of causal rare variants is very small (at most 0.99) (Fig 1A and 1B); in this case, all methods have very low power (<15%). When the effect size is larger, our method is considerably more powerful than SKAT-O (Fig 1C and 1D) where in these settings, the power of the proposed method is 141% to 238% as high as that of SKAT-O. These results demonstrate that prioritizing potential causal variants using the likelihood ratio test boosts statistical power to detect the effects of rare variants across various values of effect size and causal ratio.

Additionally, we perform simulations that add randomness to the effect sizes of causal rare variants by adding noises sampled from a normal distribution $N(1,1)$. We also include simulations that contain much fewer rare variants (19.8 rare variants on average), which is about two-thirds of the number of rare variants in the original simulation (33.1). Note that it is difficult to further decrease the number of rare variants in the new simulations because the proportion of causal rare variants is assumed to be 3–10% and we need to ensure there is at least one causal rare variant. We still observe that LRT-q is much more powerful than other methods in these different settings (S1 Fig).

For the convenience of parameter estimation, the LRT-q model assumes the equal variance of the expression levels of individuals with or without causal rare variants. To examine its robustness when the assumption is violated, we simulate the subjects that carry causal rare variants to have an explicitly different variance from those who do not. We find that our method is robust against the violation of this assumption and can still achieve higher power than other methods (S2 and S3 Figs). Therefore, this assumption simplifies the parameter estimation for the model but does not seem to influence its statistical power and reliability much.

One of the key reasons for the higher statistical power of LRT-q is its nonlinear decision boundary to detect significant associations. A decision boundary of an algorithm determines how it classifies each test (e.g. association test between rare variants and gene expression) into a significant or non-significant association where methods combining effect of rare variants linearly such as CMC have a linear decision boundary while LRT-q that applies nonlinear aggregation of rare variant effects have a nonlinear decision boundary. Using simulations to visualize decision boundaries of multiple rare variant methods (see Materials and Methods), we verify LRT-q has a clear nonlinear decision boundary that separates significant and non-significant associations accurately (S4A Fig). As the nonlinear decision boundary of LRT-q allows it to emphasize contributions from potential causal variants with large effects to its statistic more strongly than those from non-causal variants, LRT-q is more sensitive to causal effects and thus has higher power as demonstrated in the simulation studies above. Decision boundaries of other methods, however, are not as nonlinear as or as obvious as that of LRT-q because some of the significant associations detected by other methods overlap with non-significant associations (S4B–S4D Fig). In other words, the nonlinear decision boundary of LRT-q has higher accuracy in segregating rare variants with causal effects and those without causal effects, which improves statistical power.

3.4 Applications

3.4.1 Identification of RV eGenes across 49 tissues in GTEx.

To demonstrate the utility of our method in real eQTL data, we apply LRT-q to whole-genome sequencing (WGS) and RNA-seq data of 681 individuals with European ancestry from 49 tissues in the GTEx v8 dataset[30] to identify RV eGenes (see Materials and Methods for quality control and data processing); those are genes whose expression is regulated by nearby rare variants. We define rare variants as variants with MAF < 5% among individuals with WGS data, and we combine effects of rare variants present within 20K bp of a transcription start site (TSS) of each gene in each tissue. The GTEx study included common variants with MAF ≥ 1% in their eQTL analysis. This means that there may be some overlaps between rare variants in our analysis and common variants in the GTEx analysis as we used rare variants with MAF < 5%. Hence, we also analyzed rare variants with MAF < 1% to avoid this overlap. The GTEx study analyzed variants within 1 Mb from TSS while we use the 20 Kb window size. The main reason is that the number of rare variants within 1 Mb from TSS is considerably greater than that within 20 kb; we observe 50 times more rare variants within 1 Mb (median of 15,482) than 20 Kb (median of 311) as shown in S5 Fig. Including too many rare variants in a rare variant association test not only will greatly increase the computational cost but also is likely to decrease the power as more non-causal variants are included in the association test. Besides, previous studies[14,31] that analyzed the genetic effects of rare variants on gene expression also considered a smaller window size such as 10 Kb. In our analysis, we choose the window size that is twice as large as that of previous studies to include more variants with potential regulatory effects on gene expression. The sample size varies considerably depending on a tissue type (from 64 to 573) as only subsets of individuals provide RNA-seq data for certain tissues while we have WGS data for the 681 European individuals. To improve power to detect RV eGenes using LRT-q, we utilize a variety of weighting schemes such as assigning them uniform weights and prioritizing them by minor allele frequency (MAF), by their distances to TSS, and by their functional scores such as LINSIGHT[32] and CADD scores[33] as well as different combinations of them (see Materials and Methods). We then compare the performance of our method with that of other methods including ACAT-O, SKAT-O, and VT by applying the same weighting schemes to each method. It is important to note that CMC, Burden, and WSS are not included in this analysis because they have low power as demonstrated in the simulation study. ACAT-V is not under consideration because it uses an aggregated Cauchy association test as ACAT-O does but has lower power than ACAT-O as shown in simulation. We use a false discovery rate (FDR) of 5% to detect RV eGenes in each tissue.

We observe that different weighting schemes of rare variants yield very different numbers of RV eGenes and that our method detects more RV eGenes than other approaches across most of the weighting schemes. Using the whole blood (N = 546) as an example, LRT-q detects more RV eGenes than VT across all eight weighting strategies while we find more RV eGenes than ACAT-O and SKAT-O across four weighting schemes (S1 Table). Regarding the number of eGenes detected using different weighting schemes, the smallest number of RV eGenes LRT-q detects is 211 with TSS distance weighting while we observe about four times as many RV eGenes with a combined weight of MAF and CADD (885). These results show that consistent with the results of our power simulation, our method can detect more RV eGenes in real eQTL data than previous methods and that different weighting schemes could greatly influence the sensitivity of RV eGene detection.

Next, we define the union set of RV eGenes identified with the eight different weighting schemes as the total set of RV eGenes detected by a method for each tissue and compare this number across different methods and tissues. First of all, as expected, the number of RV eGenes detected by LRT-q across tissues is positively correlated with sample sizes of tissues (Pearson's $r = 0.8966$) where this phenomenon is not affected by the number of expressed genes in each tissue (Figs 2A and S6A). When comparing the number of RV eGenes detected by different methods, we find that LRT-q identifies the largest number of RV eGenes in 35 out of 41 tissues where there is at least one RV eGene detected by any method (Fig 2B) while LRT-q detects only one fewer RV eGene than SKAT-O in four tissues (Brain_Putamen_basal_ganglia, Muscle_Skeletal, Ovary, and Vagina). In three of these four tissues, including Brain_Putamen_basal_ganglia, Ovary, and Vagina, have so small sample sizes that LRT-q and VT failed to detect any RV eGenes while SKAT-O and ACAT-O identified at most one RV eGene. In general, LRT-q detects on average 308% more total RV eGenes than SKAT-O (min:1% and max:2,800%), which identifies the second most total RV eGenes in GTEx tissues. Importantly, our method identifies a few RV eGenes in tissues with small sample sizes such as brain—hypothalamus (N = 150) while other methods fail to detect any RV eGenes in these tissues. We find that our method outperforms other methods even when we lower the MAF threshold to 1% to define rare variants although we detect fewer overall RV eGenes with 1% MAF compared to those with 5% MAF, which is expected (S2 and S3 Tables). Results show that our method also discovers more novel RV eGenes than other methods in 38 out of 41 tissues, which are eGenes not reported in the GTEx v8 analysis that only considered the effects of common variants (MAF $\geq$ 1%). LRT-q detects only one fewer RV eGene than SKAT-O in the other three tissues (Brain_Putamen_basal_ganglia, Ovary, and Vagina) (S6B Fig). LRT-q detects on average 204% more novel RV eGenes than SKAT-O that detects the second most novel RV eGenes (min:8% and max:1,000%). These results indicate that our method detects not only more overall RV eGenes but also more novel RV eGenes that have not been discovered using common variants, which may be important in interpreting the functional effects of rare variants.

To examine overlaps among RV eGenes detected by different methods, we look at RV eGenes in four tissues, Muscle_Skeletal, Skin_Sun_Exposed_Lower_leg, Thyroid, and Whole_Blood where we have a good number of RV eGenes. We look at the overlaps of RV eGenes detected by four methods, LRT-q, SKAT-O, ACAT-O, and VT. Using the Venn diagram (S7 Fig), we find that many RV eGenes detected by SKAT-O, ACAT-O, and VT are also detected by LRT-q: on average, LRT-q detects 61.76%, 72.29%, 80.40% of RV eGenes detected by SKAT-O, ACAT-O, and VT, respectively. This result also shows that a majority of RV eGenes detected by ACAT-O and VT are shared with other methods as they detect the smallest numbers of RV eGenes. ACAT-O shares most of RV eGenes with VT and SKAT-O because it is a combination method that uses the results from SKAT-O and VT. SKAT-O also has higher proportions of shared RV eGenes with other methods compared to LRT-q, where it identifies 63.16%, 92.35%, 92.01% of RV eGenes discovered by LRT-q, VT, and ACAT-O, respectively. This result shows that LRT-q detects many RV eGenes detected by other methods and detects additional RV eGenes.

Lastly, we detect RV eGenes using independent rare variants after LD-pruning as one of the assumptions in LRT-q is the independence among variants. We find that LD-pruning increases the number of total RV eGenes and novel RV eGenes detected by LRT-q for the whole-blood tissue by 34.54% and 32.03%, respectively, using the 5% MAF threshold for rare variants. We also observe more RV eGenes for other methods (S4 and S5 Tables). This result shows that the independence assumption may limit the ability of LRT-q to detect RV eGenes, but does not increase FPR as we observe fewer RV eGenes when rare variants are not independent. For the rest of the analysis, we present the results using the 5% MAF threshold and without using LD-pruning because the number of rare variants changes considerably depending on the level of LD-pruning we perform, and it is not obvious which LD-pruning procedure yields the best results.

One important factor that may influence detection of RV eGenes is common SNP eQTLs near rare variants. It is possible that common SNP eQTLs and rare variants may be in weak LD, and LRT-q may detect the common SNP eQTL signal as a rare variant association. Note that this phenomenon does not influence our results on novel RV eGenes since they do not contain common SNP eQTLs. To identify how common SNP eQTLs may affect the detection of non-novel RV eGenes (RV eGenes that have common eQTLs), we regress out the effect of the most significant eQTL from gene expression within 20kb, 50kb, and 100kb from the transcription start sites (TSS) of each non-novel RV gene and perform rare variant association tests with LRT-q to detect RV eGenes in four tissues, including Whole_Blood, Thyroid, Muscle_Skeletal, and Skin_Sun_Exposed_Lower_leg. We select these distance ranges because we consider rare variants within 20kb from TSS while common eQTLs may be up to 1mb from TSS and we only want to regress out common eQTLs that might be in LD with rare variants. We calculate the differences in p-values of non-novel RV eGenes before and after this regression across different weights for rare variants.

The results show that p-values of most non-novel RV eGenes do not change after the regression as the median change in p-value is close to 0.0 (S8 Fig). However, we observe large changes in p-values for some non-novel RV eGenes, and hence, we decide to look at how the number of RV eGenes changes after this regression. For this, we use the fixed p-value threshold (1e-4) to identify RV eGenes instead of FDR. The reason is that we have two groups of genes: 1) genes that have significant eQTLs, and we regress out the effect of these eQTLs from gene expression, and 2) genes that do not have significant eQTLs, and hence we do not apply this regression. We find that by combing these two groups of genes, the p-value distribution changes somewhat significantly, and hence it also changes q-values significantly although the corresponding p-values have not changed much.

We find that both LRT-q and SKAT-O indeed lose some RV eGenes after regressing out the common eQTL effect, which is expected. LRT-q loses about 25.16% of RV eGenes on average where SKAT-O loses a much higher proportion of RV eGenes (36.11% on average) (S6 and S7 Tables). These results suggest that although some of the rare variant associations LRT-q detects may be due to the effect of common SNP eQTLs, they do not seem to appear very frequently. The results also suggest that for these four tissues, although LRT-q detects fewer RV eGenes (FDR < 5%) than SKAT-O before the regression except Thyroid, SKAT-O might have detected more common SNP eQTLs as RV associations as SKAT-O loses a much higher proportion of RV eGenes after the regression.

3.4.2 Patterns of tissue-shared and tissue-specific RV eGenes in GTEx.

We investigate tissue-sharing patterns of RV eGenes in GTEx to determine whether related tissues share more RV eGenes and to compare these patterns to those from CV eGenes, which are eGenes detected from common variants in the previous GTEx analysis. To find a tissue-sharing pattern of RV eGenes between a pair of tissues, we count the number of RV eGenes shared between the two tissues and divide it by the number of RV eGenes in the tissue with fewer RV eGenes. It is important to note that this approach is different from the previous GTEx analysis that used the correlation of effect sizes of common eQTLs between a pair of tissues. As calculating the combined effect size of rare variants is not obvious, we instead calculate the fraction of RV eGenes shared between a pair of tissues and apply the same approach to CV eGenes for comparison. Lastly, as some tissues have very few RV eGenes, we use FDR of 10% to increase the number of RV eGenes in each tissue.

We observe that tissues with related functions share a high fraction of their RV eGenes and are clustered together, such as most brain tissues (11 out of 12 brain tissues) and tissues in the digestion system including stomach, esophagus, colon, and small intestine tissues (Fig 3A). Also, there are a few related tissues that share a high fraction of RV eGenes (S9 Fig). For example, three artery tissues share on average 58.52% of RV eGenes among them, esophagus—muscularis and esophagus—gastroesophageal junction share 53.32% of RV eGenes, and two skin tissues share 48.62% of RV eGenes. The overall tissue sharing patterns of RV eGenes are similar, although attenuated, to patterns of tissue sharing of CV eGenes; we observe stronger tissue sharing patterns of functionally related tissues for CV eGenes (S10A and S11 Figs). Interestingly, we observe two separate clusters of brain tissues in the tissue-sharing matrix (Fig 3A), as the patterns of RV eGenes sharing among brain tissues are not strong where the average fraction of RV eGene sharing is 27.75%, compared to CV eGenes where the average fraction of CV eGene sharing is 58.85%. This may be due to the small numbers of RV eGenes detected in those tissues, where there are 34.08 RV eGenes on average for each brain tissue compared to 6870.15 CV eGenes.

Next, we identify a pattern of tissue-sharing across more than two tissues, and for this analysis, we use 20 tissues that have at least 200 RV eGenes as tissues with only a few RV eGenes would not share many eGenes with other tissues. Among 7,857 unique RV eGenes in those 20 tissues, we find that 60.26% of them are RV eGenes in only one tissue ("tissue-specific"), 28.74% of them are RV eGenes in 2–4 tissues, and only 11.00% of them are eGenes shared in more than 4 tissues (Fig 3B). To compare this result with the tissue-sharing pattern of CV eGenes, we select the top *Nt* of CV eGenes sorted by FDR q-values from tissue *t* where *Nt* is the number of RV eGenes in tissue *t*, so that we compare the same number of CV and RV eGenes from each tissue. This is necessary to make tissue-sharing patterns of CV and RV eGenes comparable as there are many more CV eGenes than RV eGenes in general and many CV eGenes are shared across many tissues without this selection of top CV eGenes. We observe a different pattern of tissue-sharing of top CV eGenes where CV eGenes are less tissue-specific than RV eGenes; 51.04% of CV eGenes are tissue-specific compared to 60.26% of RV eGenes, and 21.15% of CV eGenes are shared in more than 4 tissues, which is about twice higher than the fraction of RV eGenes shared in that number of tissues (Fig 3B). We repeat this experiment selecting 25 tissues with at least 100 RV eGenes and observe similar results (S10B Fig). These results demonstrate that the tissue-sharing patterns of RV eGenes reflect the functional relationship among tissues, and they tend to be more tissue-specific when compared to CV eGenes.

3.4.3 Enrichment of expression outliers, proximal rare variants, and disease-associated genes among RV eGenes in GTEx.

Previous studies have shown that large-effect regulatory rare variants may cause abnormal gene expression, causing individuals carrying those variants to have significantly higher or lower expression for certain genes[14] In this analysis, we investigate whether RV eGenes we detect from LRT-q are enriched with expression outliers who have abnormal gene expression compared to other non-RV eGenes. First, similar to Li et al.[14], we correct gene expression measurements for age, sex, genotyping principal components, and PEER factors, and then generate standardized Z-scores. We define expression outliers as individuals with standardized gene expression |Z-score| > 2 and count the number of outliers in each gene. Because genes may be expressed differently depending on tissues, outliers are defined specific to genes and tissues. In each tissue, we count the number of outliers for each RV eGene and non-RV eGene separately, and then we aggregate these counts across all tissues. We observe that 8,090 unique RV eGenes (FDR < 5%) across all 49 tissues have 19.52 expression outliers on average, which is significantly greater than 13.28 outliers on average in 30,436 non-RV eGenes (t-test p < 2.2e-16). We also look at whether those expression outliers carry rare variants within 20K bp of a TSS of each eGene, and we find that across all tissues, on average, 72.17% of expression outliers carry one or more rare variants (S12 Fig).

Li et al. discovered that expression outliers were enriched for rare variants near the TSS compared to non-outliers, and we investigate whether this enrichment is stronger for RV eGenes compared to non-RV eGenes. This enrichment is defined as the ratio between the proportion of outliers with proximal rare variants (those within 20kb of TSS) and the proportion of non-outliers with the rare variants for each gene, which can be thought of as relative risk of carrying the rare variants in outliers vs. non-outliers. Using FDR of 5% to detect RV eGenes in each tissue, our results show that outliers are significantly enriched for proximal rare variants compared to non-outliers in all tissues except three brain tissues (Brain_Putamen_basal_ganglia, Brain_Hypothalamus, and Brain_Cortex) with limited sample sizes (Fig 4A). For non-RV eGenes, we do not observe this enrichment in all tissues. We observe consistent results when varying Z-score thresholds to define expression outliers; outliers are significantly enriched for adjacent rare variants compared to non-outliers regardless of Z-score thresholds and the enrichment increases as the Z-score thresholds increase (S13 Fig). These results suggest that rare variants with *cis*-regulatory effects may be key factors to explain the large changes in gene expression levels and those rare variants are likely to have significant contributions to RV eGenes.

Lastly, we hypothesize that RV eGenes are more likely to be associated with diseases or traits. For this analysis, we calculate enrichment of RV eGenes among five online disease gene databases (see Materials and Methods) including 6,298 genes from NCBI ClinVar database[34], 2,569 genes from Genotype-to-Phenotype (G2P) database[35], 20,998 reported GWAS genes from NHGRI-EBI catalog[1], 26,352 genes from Online Mendelian Inheritance in Man (OMIM) database[36], and 7,298 genes from OrphaNet database[37]. We choose these databases because they facilitate the development, curation, validation of large-scale datasets for associations between human genetic variants and complex and rare diseases. We also consider genes for non-disease traits as positive controls, which are 212 genes related to body mass index (BMI) and 78 genes related to height that are provided by the GeneRIF database and downloaded from the Harmonizome database[38]. Note that GeneRIF is a public database for the functional annotations of genes based on previous literature. We construct a 2x2 contingency table where an outcome is whether a gene is a disease gene for each database and an exposure is whether a gene is an RV eGene or a non-RV eGene. We use 8,090 RV eGenes we detect from all 49 tissues with FDR of 5%, and p-value is computed with the Fisher's exact test where odds ratio (OR) greater than 1 indicates enrichment of RV eGenes in a disease database while OR less than 1 indicates depletion. Results show that RV eGenes are significantly enriched in five disease databases with OR ranging from 2.00 to 2.97 (p = 0~9.58e-42, Fig 4B), and the largest OR is observed in the OMIM database that contains genes involved in Mendelian disorders. We also observe an odds ratio of around 1.0 for genes related to BMI and height, as expected, because they are two common traits and are not related to any certain diseases. These results suggest that RV eGenes are much more likely to be involved with diseases compared to non-RV eGenes while they are not enriched in non-disease traits.

3.4.4 Analysis of disease-associated RV eGenes.

To discover the clinical importance of RV eGenes we identify, we perform a literature search on RV eGenes using the ClinVar database. Specifically, we attempt to find whether an RV eGene in a specific tissue is associated with a certain disease related to that tissue. First, we find that patients with platelet count disorders carry rare variants in *TUBB1*[39] where *TUBB1* is detected by our method as one of the RV eGenes in the heart left ventricle and skin tissue. The landmark symptom of platelet count disorders is petechiae on the skin[40]. There is one rare variant (rs41303899) in *TUBB1* that was reported as likely pathogenic for this disorder where the gnomAD frequency for this variant is 1.5E-3 in the European population. Interestingly, one individual in GTEx carries this rare variant although the disease status of this individual is not available. We find that the adjusted *TUBB1* expression Z-score of this individual carrying this rare variant is 1.15 in skin tissue, which is relatively high.

Another example of an association between RV eGenes found in a particular tissue and a tissue-specific disease caused by rare variants in those genes is telomere sheltering gene *POT1*, which is one of the RV eGenes found in fibroblasts. Fibroblasts were found to contribute to the growth and drug resistance of melanoma, a potentially lethal form of skin cancer[41]. Previous whole-exome sequencing studies found that rare variants in *POT1* could increase the risk for familial cutaneous malignant melanoma, as one of the rare variants, rs587777477, was discovered to perturb telomere maintenance[42,43]. We also find that *IFIH1* is identified as an RV eGene in skin tissue, and the rare missense variant, rs587777446, in this gene has been shown to be pathogenic for autosomal dominant inflammatory disorder, Aicardi-Goutieres syndrome 7 with the phenotype of skin swelling[44,45]. These examples indicate that some RV eGenes are associated with diseases caused by rare variants in relevant tissues, which demonstrates the clinical importance of RV eGenes.

3.5 Discussion

We have proposed LRT-q as a powerful rare variant association test for identifying the effects of rare variants on gene expression. Our simulation studies showed that the proposed method had a well-controlled false positive rate and higher statistical power compared with other methods.

Through the analysis of gene expression data of 49 tissues from the GTEx dataset, we demonstrated that LRT-q detected more genes whose expression was regulated by nearby rare variants, which we call RV eGenes, compared to other approaches including SKAT-O. More importantly, our method discovered the largest number of novel RV eGenes that were not regulated by common variants reported in GTEx, which might be particular interest to studies analyzing the functional effects of rare noncoding variants. These results show that LRT-q is an effective statistical method for rare variant association analyses for quantitative traits including gene expression.

RV eGenes discovered from 49 tissues in GTEx provided several important biological insights about gene regulation of rare variants. First, we found that as expected, a pair of functionally related tissues shared a high proportion of RV eGenes because their gene expression values were correlated. However, the levels of tissue-sharing patterns of RV eGenes were not as high as those of CV eGenes where one main reason is the limited number of RV eGenes compared to the number of CV eGenes. We detected far fewer RV eGenes than CV eGenes with the same sample size, which is also expected as we have lower power to detect the effects of rare variants than common variants even with the rare variant association methods that combine effects of multiple rare variants[5,46,47]. This suggests that we need larger sample sizes to detect more RV eGenes.

Next, when we checked the tissue sharing patterns of RV and CV eGenes across 20 tissues using the same number of RV and CV eGenes for each tissue, we found that a higher proportion of RV eGenes were detected only in one tissue than CV eGenes where CV eGenes had a much higher proportion of genes shared across more than four tissues. This suggests that the effects of rare variants on gene expression may be more tissue-specific than common variants, which is important in interpreting results of rare variant associations for complex diseases and traits. However, we anticipate that a higher fraction of RV eGenes will be shared across many tissues as more RV eGenes are discovered with a larger sample size as we have observed this phenomenon with CV eGenes[48,49].

Lastly, we explored the characteristics of RV eGenes with a series of enrichment analyses. We found that all RV eGenes had several outliers whose expression levels deviate significantly from the rest of the individuals. These outliers had enrichment of rare variants near the TSS of RV eGenes compared to non-outliers while there was no such enrichment for non-RV eGenes, suggesting that rare variants carried by outliers may play important roles in causing the abnormal expression levels of the outliers for RV eGenes. Additionally, we discovered that RV eGenes were significantly enriched for disease-associated genes across all human disease databases, indicating that genes whose expression is influenced by nearby rare variants have a higher chance of being associated with diseases. Moreover, previous findings provided evidence supporting that rare mutations in our RV eGenes could increase the risk for certain diseases in the same tissues where they were discovered. This further suggests that rare variants with regulatory effects may help identify genes associated with diseases.

There are four main features of LRT-q that make it a highly powerful test as demonstrated in simulation and the GTEx data. First, it prioritizes potential causal rare variants with genotype data and functional information such as CADD scores. With our formulation of the likelihood ratio test, LRT-q attempts to find the most likely scenario of causal statuses of rare variants, which increases the power of detecting potential causal variants. Second, it applies nonlinear aggregation of rare variants, which results in a nonlinear decision boundary in detecting their effects. Using simulations, we show that the nonlinear decision boundary enables LRT-q to emphasize the effects of causal variants in its test statistic, leading to a higher power. Third, as an extension of the original LRT method, LRT-q also computes the likelihoods for all possible scenarios of causal statuses using an efficient decomposition technique, which reduces the computational complexity and enables LRT-q to be applied to large-scale datasets. Fourth, LRT-q considers both directions of rare variant effects as LRT-q statistics are based on the normal distribution that considers the absolute values of effect sizes and not their directions. This is important because some regulatory variants may increase gene expression (positive effect) while other variants may decrease it (negative effect). Results from the real GTEx data appear to suggest that rare variants are likely to have different directions of effect because VT, a method that assumes the same direction of effects of rare variants, detected much fewer RV eGenes than LRT-q; if variants had consistent directions of effect, VT would have detected many more RV eGenes.

The application of LRT-q can be extended to other quantitative traits, such as height and BMI. As the likelihood ratio test is the most powerful test for a particular hypothesis test according to the Neyman-Pearson lemma[50], one is likely to achieve higher power with LRT-q on other quantitative traits than other previous methods. Also, LRT-q can be generalized to a gene-based test or a region-based test as well as analysis of gene sets, pathways, or networks. It is, however, important to find appropriate weights for rare variants because results may change considerably depending on those weights as our results showed that different functional annotations of rare variants affected power to detect RV eGenes. To address this issue, we used a straightforward approach that employs a variety of functional annotations and combines results. Identifying an ideal set of weights for rare variants for gene expression and an optimal approach to combine results remains an important research topic. One of the limitations of LRT-q is that it may be computationally demanding as it needs to perform a permutation test to estimate p-value for each gene. One approach to improve the efficiency of LRT-q is performing an adaptive permutation test that stops the permutation test when observing p-values from a small number of permutations are high (e.g., > 0.05). Assuming that most genes are not RV eGenes, we would only need to perform 1,000 or fewer permutations for the majority of genes. For those genes with small p-values (potentially RV eGenes), we would perform up to 100,000 permutations to obtain more accurate p-values. We find that the adaptive permutation test yields a similar number of RV eGenes compared to the permutation test that uses 100,000 permutations (S14 Fig).

For efficient calculation of LRT-q statistic and the corresponding p-value, we assume the independence between rare variants as previous studies[28,51] have found that there would be very low LD among rare variants. In this study, we found by performing LD-pruning that violation of the assumption of independence between rare variants may reduce the power of LRT-q, but it does not increase FPR. In our analysis of the GTEx data, we did not perform LD-pruning as the optimal LD-pruning approach is not currently known for rare variants. Researchers may want to apply LRT-q to their eQTL data after applying LD-pruning to identify more RV eGenes.

3.6 Materials and methods

3.6.1 LRT-q model

Suppose that we have genotype and gene expression data of a population with size *N*, and perform an association test for a gene with *k* rare variants. For rare variant *i* (1≤*i*≤*k*), there are *mi* individuals not carrying a rare variant (e.g. not carrying a minor allele of a rare variant), whose expression levels are $X_i = \{x_i^1, x_i^2, \cdots, x_i^{m_i}\}$, and *ni* subjects carry a rare variant *i* (e.g. carrying a minor allele), whose expression levels are $Y_i = \{y_i^1, y_i^2, \cdots, y_i^{n_i}\}$. Note that *N* = *mi*+*ni*. There are two assumptions in our model: 1) independence among rare variants (e.g. no linkage disequilibrium (LD) among rare variants) and 2) the normality of gene expression values. Previous studies have suggested that there would be very low linkage disequilibrium (LD) among rare variants because of their low frequencies[28,51]. As for the normality assumption, gene expression values are often quantile normalized in eQTL studies[48,49,52,53], which means that *Xi* and *Yi* can be viewed as random samples from a standard normal distribution. With these two assumptions, *Xi* and *Yi* are independently and normally distributed ($X_i \sim N(\mu_{X_i}, \sigma_{X_i}^2), Y_i \sim N(\mu_{Y_i}, \sigma_{Y_i}^2)$, where μ_{X_i}, μ_{Y_i} stand for the means of *Xi*, *Yi* and $\sigma_{X_i}, \sigma_{Y_i}$ represent the standard deviation of *Xi*, *Yi*, respectively). We are interested in testing the effects of rare variants on gene expression, that is, the difference between *Xi* and *Yi*. Thus, we test the following hypotheses

$$H_0 : \forall 1 \leq i \leq k, \mu_{X_i} = \mu_{Y_i} \; versus \; H_1 : \exists 1 \leq i \leq k, \mu_{X_i} \neq \mu_{Y_i}$$

The null hypothesis (*H*0) asserts that no rare variants have regulatory effects, while the alternative hypothesis (*H*1) states that there is at least one causal rare variant affecting gene expression. Here, σ_{X_i} and σ_{Y_i} are both unknown but assumed to be equal to the pooled variance *σi*.

To boost the statistical power, we want to infer which rare variants are causal. Here, let *vi* be an indicator variable for the causal status of variant *i* (1≤*i*≤*k*); *vi* = 1 if variant *i* is causal and 0 otherwise. Let *V* = {*v*1, *v*2,….,*vk*} be the causal statuses of *k* rare variants. Then there are 2*k* possible values for *V*, because each of the *k* rare variants can have causal effects on gene expression or not. Among them, let $V_q = \{v_1^q, v_2^q, \cdots, v_k^q\}$ be the *qth* vector, representing a specific scenario of causal status. Using the functional information on rare variants, such as CADD scores, we can obtain the probability of variant i being causal $ci = P(vi = 1)$. Using the assumption that rare variants are independent, the probability of each scenario Vq is given by

$$P(V_q) = \prod_{i=1}^{k} c_i^{v_i^q} (1 - c_i)^{1-v_i^q} \tag{1}$$

We calculate the likelihood of the observational data and the inferred causal statuses Vq as follows

$$L(X, Y, V_q) = L(X, Y | V_q) P(V_q) \tag{2}$$

where $X = X1, X2, \cdots, Xk$, $Y = Y1, Y2, \cdots, Yk$ are gene expression levels of individuals without rare variants and with rare variants, respectively. This equation considers both observational data (gene expression and genotype data) and causal statuses of rare variants, and therefore can prioritize causal variants by functional information. We then calculate our statistic as the ratio between the likelihood under the null hypothesis and the likelihood under the alternative hypothesis and use a permutation test to compute the p-value. More detailed information on the derivation of likelihood ratio test, its decomposition for efficient calculation of the test statistic, and parameter estimation is discussed in S1 Text.

3.6.2 Simulation studies

To compare the performance of LRT-q with the widely used existing rare variant association tests, we measure their type I error rates and statistical power in simulation studies. In this study, data are simulated with a similar framework described in Wu et al.'s work[23].

Simulation of genotype data.

The calibration coalescent model[54] (COSI) is used to generate 50,000 haplotypes, assuming that they have the LD structure of individuals of European ancestry. Any pairs of haplotypes could be combined into diplotypes. In each replicate, a 5 kb region is randomly selected to simulate the diplotypes for 1,000 individuals, which contains 33.1 rare variants (MAF < 0.05) on average. We also perform the power simulation with a 3 kb region including 19.8 rare variants on average.

Type I error rate simulation.

Under the null hypothesis of no association between rare variants and gene expression, we simulate the normalized expression levels for individual *j* from the model described as follows.

$$E_j = A_j + \epsilon_j$$

where *Aj*~*N*(0,1) represents the covariates and *ϵj*~*N*(0,1) stands for random errors. Each *Ej* is assumed to be independent. We simulate 100,000 replicates to test the type I error rate at the significance level α = 0.05,0.01,0.001,0.0001. When applying rare variant association methods to the datasets, we use uniform weights for all rare variants (e.g. assuming all rare variants are likely causal). VT and LRT-q are run with 10,000 permutations to measure p-values.

Power simulation.

Under the alternative hypothesis where there is at least one causal rare variant influencing gene expression, we use the following model to simulate the gene expression value for individual *j*.

$$E_j = A_j + \beta^T G_j + \epsilon_j$$

where $Aj \sim N(0,1)$ represents the covariates of individual j and $\epsilon j \sim N(0,1)$ stands for random errors. Here, we randomly sample s variants out of the total k rare variants as causal variants. $Gj = (gj1,\dots,gjs)$ is defined as genotypes of s causal rare variants of individual j where $gji = 0,1,2$ depending on the number of rare alleles for variant i. Their effect sizes are set to be $\beta = a|log10MAF|$, where MAF represents the minor allele frequencies of causal rare variants and a is a constant. In this study, a is set to be a fixed constant 0.3, 0.6, 0.9, 1.2, or 1.5 and we also assume 3%, 5%, 7%, or 10% of rare variants to be causal to simulate different numbers of causal variants with different effect sizes. Thus, in this simulation study, the maximum effect size of a causal rare variant would be 0.99, 1.98, 2.97, 3.96, or 4.95 assuming 1,000 individuals and a MAF cutoff of 5%. Causal rare variants have 50% probability of having negative effect sizes (e.g. decrease gene expression), and 50% probability of having positive effect sizes (e.g. increase gene expression). The statistical power is estimated as the proportion of p-values smaller than $\alpha = 0.05$ in 10,000 simulated datasets. Similar to the type I error simulations, all rare variants are weighted equally. Both LRT-q and VT are run with 1,000 permutations to calculate p-values. We also generate simulations to visualize decision boundaries of LRT-q, SKAT-O, and CMC approaches, and a detailed description of this simulation is discussed in S1 Text.

To examine the robustness of the LRT-q model, we generate simulations using different settings. First, we sample a from $N(1,1)$ to add random noises to effect sizes of rare variants. Second, we perform the power simulation with fewer rare variants in a gene. Third, we simulate Xi (gene expression levels of individuals not carrying the rare variant i) and Yi (gene expression levels of individuals carrying the rare variant i) to have different variances, which violates the assumption of LRT-q model in parameter estimation. Here, we let $Xj = Aj+\epsilon j$, where $Aj \sim N(0,1)$ represents the covariates and $\epsilon j \sim N(0,1)$ stands for random errors. Let $Yj = Bj+\beta TGj+\epsilon j$, where $Bj \sim N(0,2)$

represents the covariates, $\epsilon j \sim N(0,1)$, and βTGj stands for the effects of causal rare variants. Hence, X and Y have different variances.

3.6.3 Analysis of multi-tissue GTEx v8 WGS and RNA-seq data

We download the GTEx dbGaP release v8 RNA-seq data from the GTEx portal and the whole-genome sequencing (WGS) data from dbGaP accession number phs000424.v8.p2. Genotype data and transcriptome data from all 49 GTEx tissues are used in this study. There are 838 subjects with both WGS and RNA-seq data.

Quality control.

We identify 681 individuals of European ancestry using EIGENSTRAT[55]. We consider only Europeans because they are the largest homogenous population in GTEx. We then extract only European samples from each tissue, creating 49 separate genotype datasets for the 49 tissues. We restrict our analysis to autosomal variants. For these 49 genotype datasets, we extract rare single nucleotide variants (SNVs), which are defined as variants with minor allele frequency (MAF) < 5%; we also test a case when rare variants have MAF < 1%. Alleles with genotyping quality (GQ) less than 20 are marked missing. We remove variant sites that have a missing rate larger than 5% or failed variant quality score calibration (VQSR)[56]. Then missing genotypes are imputed as two reference alleles because of the low frequency of rare variants.

RV eGene discovery in the GTEx dataset.

SNVs are functionally annotated with CADD[26] and LINSIGHT[32]. Next, we group variants in a gene and those located within 20kb upstream or downstream of transcription start site (TSS) of a gene. The summary statistics of sample size, the number of genes and rare variants for each tissue after preprocessing is in S8 Table. Covariates of each sample provided by GTEx, which are top 5 genotyping principal components, PEER factors[57] (15 factors for tissues with fewer than 150 samples, 30 factors for those with 150–250 samples, 45 factors for those with 250–350 samples, and 60 factors for those with more than 350 samples), sequencing platform, and sex are used to regress out unwanted confounding effects in gene expression levels for each tissue using a linear model. Then the transformed gene expression levels are normalized with rank-based inverse normal transformation using "RankNorm" in the "RNOmni" R package. When applying rare variant association methods to the GTEx data, different weighting strategies of rare variants are used, including LINSIGHT scores, CADD scores, MAF, distance to TSS, and uniform weights (all rare variants have the same weight). Note that weighting by MAF or TSS distance is to use weights inversely proportional to the values of MAF or TSS distance, so variants with lower frequency or closer to TSS are assigned higher weights while for other weightings, higher scores (e.g. CADD or LINSIGHT scores) mean higher weights for variants. We also combine multiple weights by multiplying two or three weights together for each variant; we create three combined weights, 1) MAF × TSS distance, 2) MAF × CADD scores, and 3) MAF × CADD scores × TSS distance. All weights mentioned above are employed to discover RV eGenes in 49 GTEx tissues. FDR < 5% is applied for multiple testing correction.

Patterns of tissue sharing in RV and CV eGenes.

We first assess tissue-sharing patterns of RV eGenes in a pair of GTEx tissues. We use FDR < 10% to identify RV eGenes in each tissue to increase the RV eGenes. Next, for each pair of tissues, we calculate the fraction of shared RV eGenes as

$$\frac{of\ shared\ RV\ eGenes\ between\ tissues\ 1\ and\ 2}{min(of\ RV\ eGenes\ in\ tissue\ 1, of\ RV\ eGenes\ in\ tissue\ 2)}$$

Similarly, to calculate pairwise tissue-sharing patterns of CV eGenes, we select CV eGenes with FDR < 5% based on the summary statistics in the GTEx v8 dataset and calculate the fraction of shared CV eGenes using the same equation. To assess tissue-sharing patterns of RV eGenes in more than two tissues, we choose 20 tissues with at least 200 RV eGenes (FDR < 5%) and calculate the proportion of RV eGenes shared across different numbers of tissues (i.e. # of RV eGenes present in only one tissue, in 2–4 tissues, or in more than 4 tissues). To find tissue-sharing patterns of CV eGenes in more than two tissues among the same 20 tissues, we choose top *Nt* CV eGenes from tissue *t* where *Nt* is the number of RV eGenes in tissue *t*. We then calculate the proportion of CV eGenes shared across different numbers of tissues. We also repeat this analysis with another group of 25 tissues that have more than 100 RV eGenes (FDR < 5%).

Single-tissue gene expression outlier discovery.

For each individual, we log-transform gene expression value as *log*2(*TPM*+1) for each gene and each tissue, where TPM is the number of transcripts per million RNA molecules. We then standardize gene expression value for each gene in each tissue into Z-score to avoid the shrinkage of outlier gene expression caused by rank-based quantile normalization, using the following equation:

$$Z_{gj}^{t} = \frac{x_{gj}^{t} - \mu_{g}^{t}}{\sigma_{g}^{t}}$$

where x_{gj}^{t} and Z_{gj}^{t} represent unstandardized log-transformed gene expression value and the standardized Z-score of individual *j* for gene *g* in tissue *t*, respectively. μ_{g}^{t} and σ_{g}^{t} are the mean and standard deviation of the unstandardized values across all individuals (x_{gj}^{t}), for gene g in tissue t, respectively. Next, for each gene in each tissue, we regress out the covariates, including top 5 genotyping principal components, PEER factors, sequencing platform, and sex from the transformed and standardized gene expression values using a linear model. The resulting regression residuals are standardized again using the equation above and the resulting Z-scores are used to determine outliers.

Single-tissue gene expression outliers in a gene are defined as the individuals with extreme gene expression levels who have |Z-score| > 2, while the remaining individuals are defined as non-outlier for this gene. Other Z-score thresholds are also tested, including 1, 3, 4, 5, 6, 7, 8, 9, and 10. Under this definition, an outlier is specific to a gene in a certain tissue. Therefore, each gene may have different sets of outliers across tissues, and an individual may be an outlier for multiple genes in one or more tissues. We analyze all outliers in non-RV eGenes and RV eGenes identified by LRT-q in 41 out of all 49 tissues with at least one RV eGene (FDR < 5%).

Enrichment analysis of RV eGenes.

To calculate enrichment of proximal rare variants near RV eGenes in gene expression outliers compared to non-outliers, we consider rare variants (MAF ≤ 5%) within 20 kb of the TSS of a gene. Similar to the analysis conducted by Li et al.[14], enrichment is defined as the ratio of the proportion of outliers carrying rare variants to the corresponding proportion of non-outliers. It is equivalent to the relative risk of having proximal rare SNVs as an outlier. The 95% Wald confidence intervals are calculated with the asymptotic distribution of the log relative risk. We also assess this enrichment by varying Z-score thresholds to define expression outliers (from 1 to 10). Enrichment is similarly calculated for non-RV eGenes.

We also examine the enrichment of RV eGenes for disease- or trait-associated genes in five public databases, including 6,298 genes from NCBI ClinVar database[34], 2,569 genes from Genotype-to-Phenotype (G2P) database[35], 20,998 reported GWAS genes from NHGRI-EBI catalog[1], 26,352 genes from Online Mendelian Inheritance in Man (OMIM) database[36], and 7,298 genes from OrphaNet database[37], We also consider genes related to two non-disease traits, 212 genes related to BMI and 78 genes related to height that are provided by the GeneRIF database and downloaded from the Harmonizome database[38]. We construct a 2x2 contingency table where an outcome is whether a gene is a disease gene for each database and an exposure is whether a gene is an RV eGene or a non-RV eGene. Odds ratios and 95% confidence intervals are computed by applying Fisher's exact test to compare non-RV eGenes and RV eGenes to each of the five lists of disease- or trait-associated genes.

Analysis of disease-associated RV eGenes.

To find evidence supporting the clinical importance of the identified RV eGenes, we do literature research in the ClinVar database. We search the database for the information about known relationships between rare variants in RV eGenes and observed health status. The information includes diseases, tissues, clinical significance, variants and their frequencies, and supporting literature.

Tables

Table 3.1: False positive rate of eight rare variant test methods in simulation

α	CMC	WSS	Burden	VT	SKAT-O	ACAT-V	ACAT-O	LRT-q
0.05	0.04935	0.04765	0.04904	0.05134	0.05036	0.04971	0.04916	0.04975
0.01	0.00958	0.00929	0.00988	0.01021	0.01055	0.00973	0.01006	0.01015
0.001	1.17×10^{-3}	1.04×10^{-3}	1.11×10^{-3}	0.99×10^{-3}	1.32×10^{-3}	1.00×10^{-3}	1.00×10^{-3}	0.89×10^{-3}
0.0001	1.00×10^{-4}	1.10×10^{-4}	1.20×10^{-4}	1.10×10^{-4}	1.10×10^{-4}	1.20×10^{-4}	0.50×10^{-4}	0.90×10^{-4}

Figures

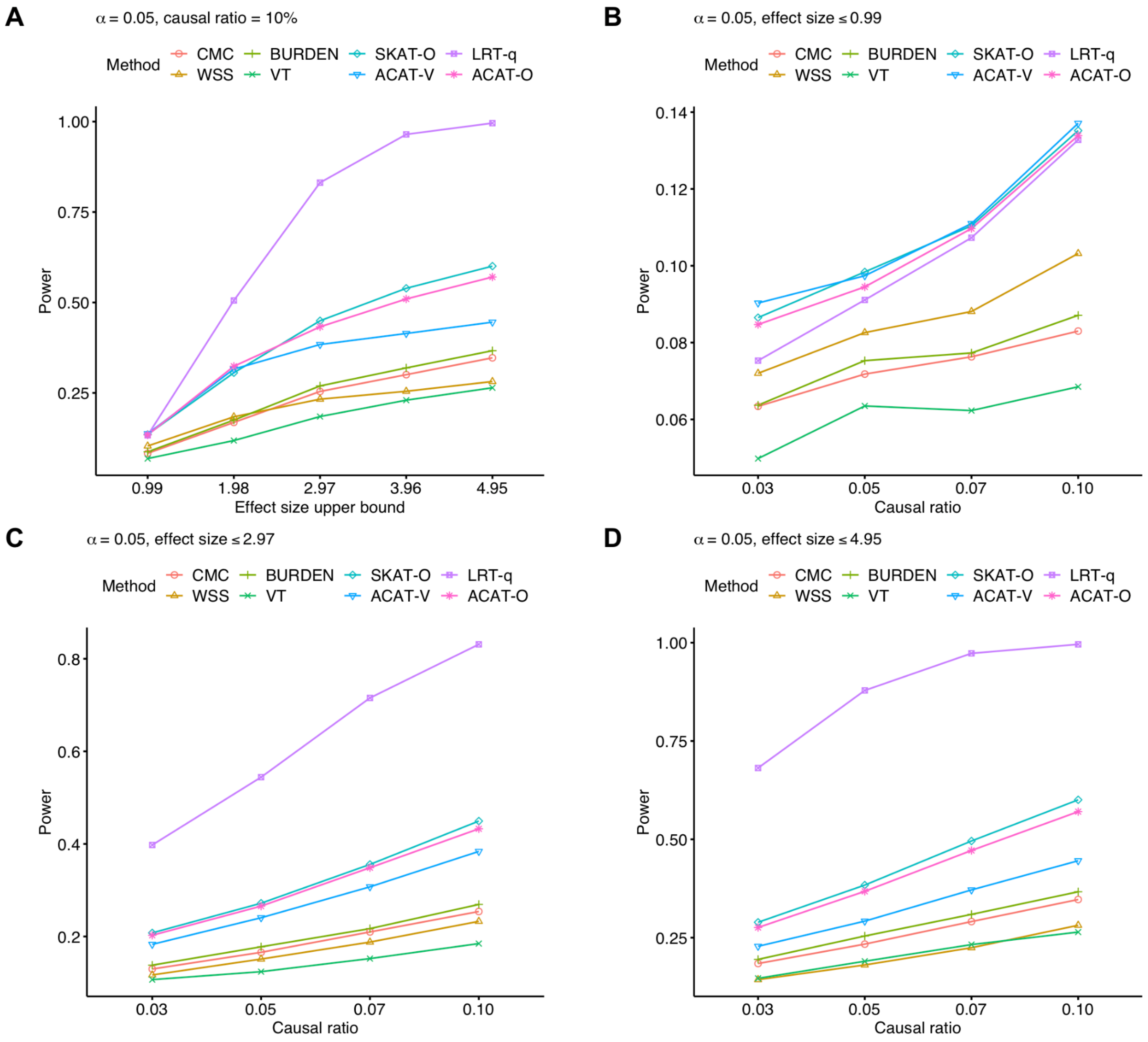

Figure 3.1: Power comparison between LRT-q and seven existing methods on simulated data. A. for different effect sizes and fixed causal ratio (10%), and for fixed effect sizes (B. ≤ 0.99, C. ≤ 2.97, D. ≤ 4.95) and various causal ratios. Significance level α = 0.05.

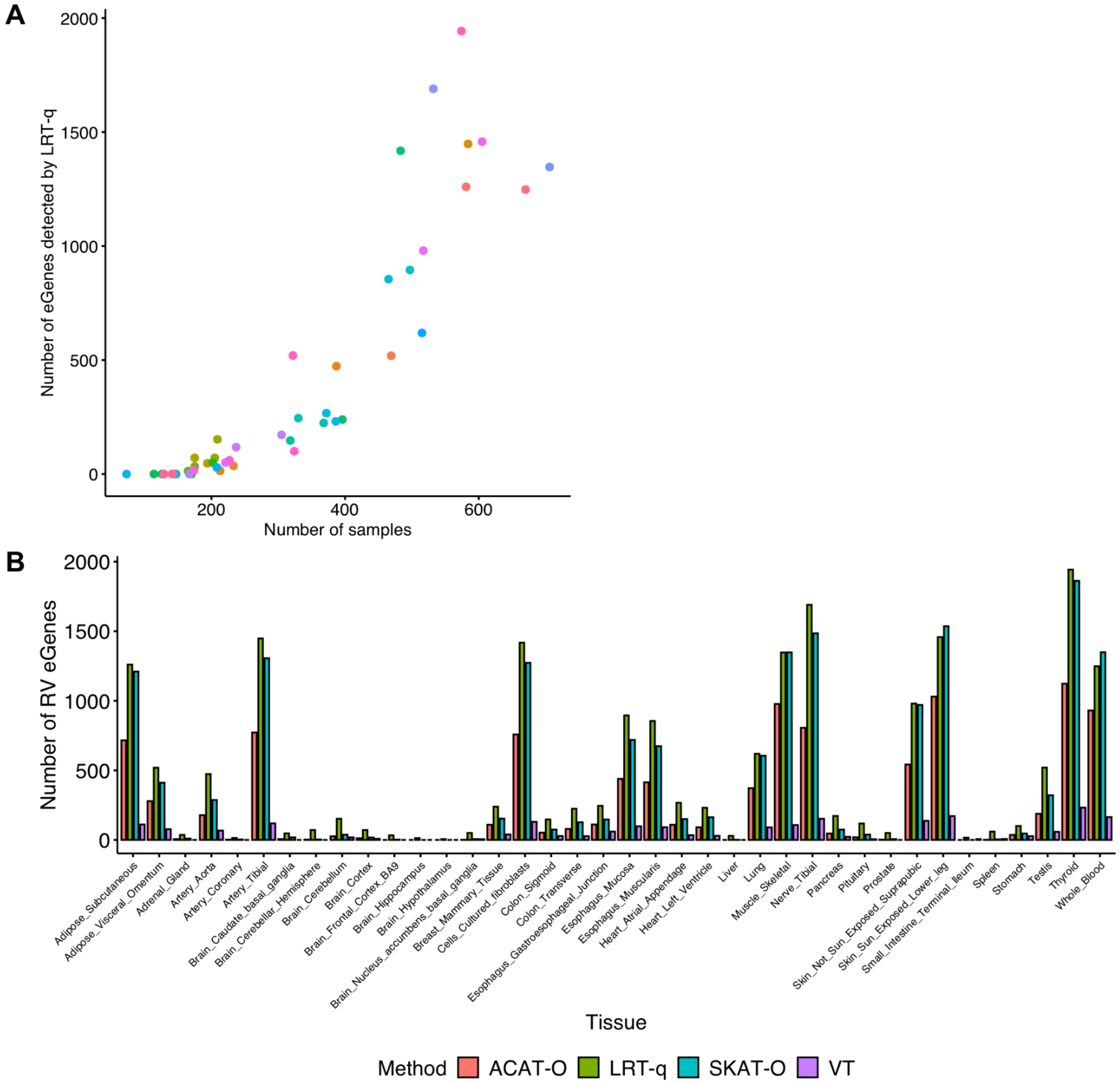

Figure 3.2: RV eGenes detected from 49 tissues in the GTEx v8 dataset. A. The relationship between the number of total RV eGenes detected by LRT-q in each tissue and the sample size of each tissue. The colors of the data points are randomly assigned. Each tissue has its own color. B. The number of total RV eGenes detected by each method. In panel B, only tissues with more than one RV eGene detected by any methods are included.

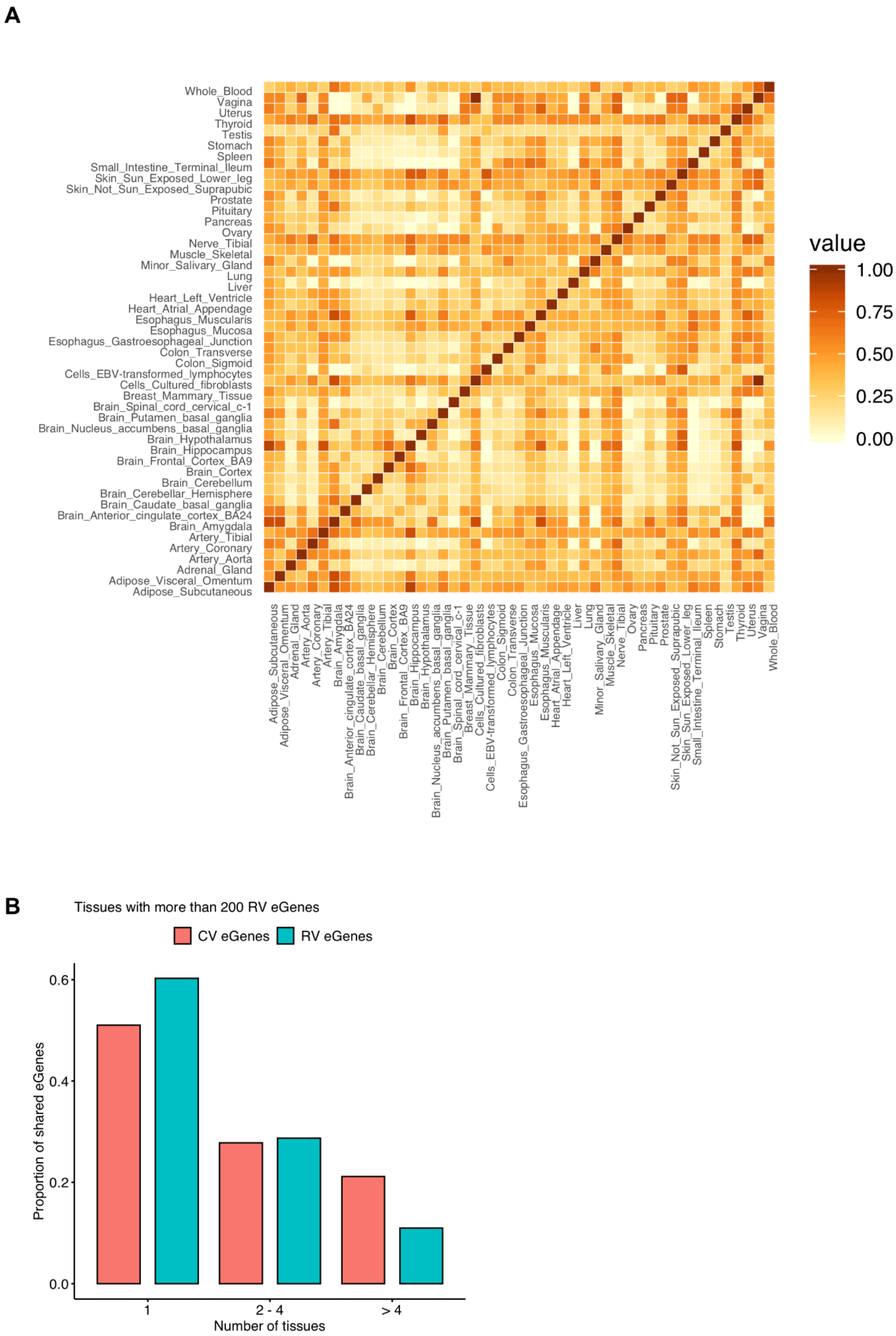
A
Whole_Blood
Vagina
Uterus
Thyroid
Testis
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Skin_Sun_Exposed_Lower_leg
Skin_Not_Sun_Exposed_Suprapubic
Prostate
Pituitary
Pancreas
Ovary
Nerve_Tibial
Muscle_Skeletal
Minor_Salivary_Gland
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV-transformed_lymphocytes
Cells_Cultured_fibroblasts
Breast_Mammary_Tissue
Brain_Spinal_cord_cervical_c-1
Brain_Putamen_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Brain_Hypothalamus
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Brain_Amygdala
Artery_Tibial
Artery_Coronary
Artery_Aorta
Adrenal_Gland
Adipose_Visceral_Omentum
Adipose_Subcutaneous
value
1.00
0.75
0.50
0.25
0.00
B
Tissues with more than 200 RV eGenes
CV eGenes
RV eGenes
Proportion of shared eGenes
0.6
0.4
0.2
0.0
1
2 - 4
> 4
Number of tissues

Figure 3.3: Tissue-sharing patterns of RV eGenes in the GTEx v8 dataset. A. Pairwise tissue-sharing matrix of RV eGenes. It shows the fraction of shared RV eGenes in each pair of tissues. Here we use FDR < 10% to increase the number of RV eGenes. Tissues are sorted by clustering. Only tissues with more than one RV eGenes are included. B. The proportion of RV eGenes and CV eGenes shared among different numbers of tissues. Only tissues with more than 200 RV eGenes are selected. It shows the proportion of eGenes that are only detected in one tissue, in 2-4 tissues, and in more than 4 tissues.

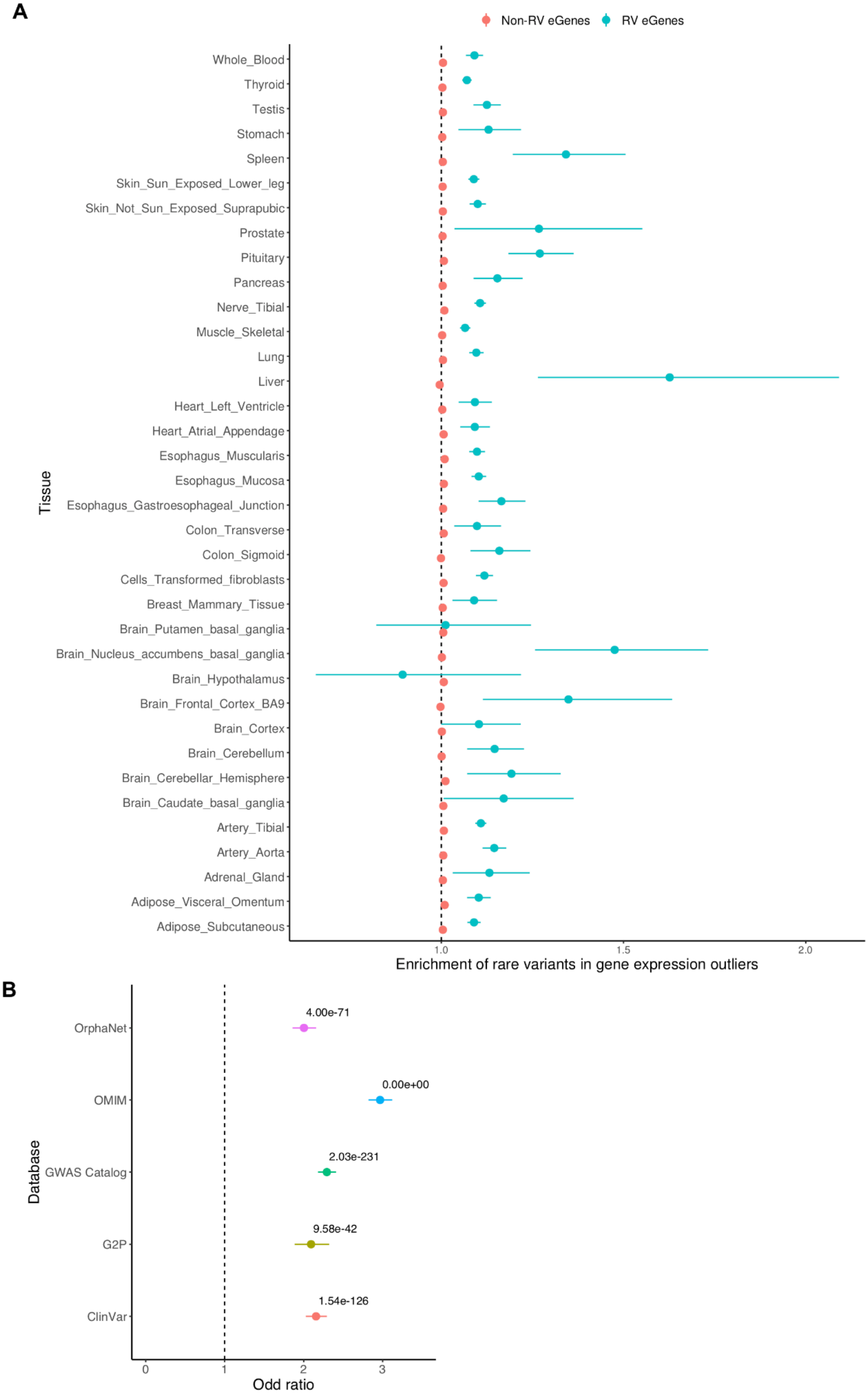
A
Non-RV eGenes
RV eGenes
Tissue
Whole_Blood
Thyroid
Testis
Stomach
Spleen
Skin_Sun_Exposed_Lower_leg
Skin_Not_Sun_Exposed_Suprapubic
Prostate
Pituitary
Pancreas
Nerve_Tibial
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_Transformed_fibroblasts
Breast_Mammary_Tissue
Brain_Putamen_basal_ganglia
Brain_Nucleus_accumbens_basal_ganglia
Brain_Hypothalamus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
Brain_Caudate_basal_ganglia
Artery_Tibial
Artery_Aorta
Adrenal_Gland
Adipose_Visceral_Omentum
Adipose_Subcutaneous
1.0
1.5
2.0
Enrichment of rare variants in gene expression outliers
B
Database
OrphaNet
OMIM
GWAS Catalog
G2P
ClinVar
4.00e-71
0.00e+00
2.03e-231
9.58e-42
1.54e-126
0
1
2
3
Odd ratio

Figure 3.4: Outlier analysis of RV eGenes detected by LRT-q in GTEx v8. A. Enrichment of proximal rare variants in outliers compared to non-outliers for RV eGenes in each tissue. Tissues without RV eGenes are excluded. B. Enrichment of RV eGenes in disease-associated genes and genes related to common traits (BMI and Height) from public databases. The numbers represent p-values. In both panels, we show the mean values as dots and 95% confidence intervals as error bars.

Reference to the published article

Jiajin Li, Nahyun Kong, Buhm Han, and Jae Hoon Sul. 2020. Rare variants regulate expression of nearby individual genes in multiple tissues. *PLoS Genetics*. 17 (6): e1009596.

3.7 References

1. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47: D1005–D1012.

2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106: 9362–9367.

3. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456: 18–21.

4. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461: 747–753.

5. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111: E455–64.

6. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11: 446.

7. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40: 695–701.

8. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13: 135–145.

9. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43: 1066–1073.

10. Beaudoin M, Goyette P, Boucher G, Lo KS, Rivas MA, Stevens C, et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. PLoS Genet. 2013;9: e1003723.

11. Johansen CT, Wang J, Lanktree MB, Cao H, McIntyre AD, Ban MR, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010;42: 684–687.

12. Li Q, Seo JH, Stranger B, McKenna A, Pe'er I. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell. 2013. Available: https://www.sciencedirect.com/science/article/pii/S0092867412015565

13. West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics. 2007;175: 1441–1450.

14. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. Nature. 2017;550: 239–243.

15. Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, et al. A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am J Hum Genet. 2016;98: 299–309.

16. Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. Ultrarare variants drive substantial cis heritability of human gene expression. Nat Genet. 2019;51: 1349–1355.

17. Sul JH, Raj T, de Jong S, de Bakker PIW, Raychaudhuri S, Ophoff RA, et al. Accurate and fast multiple-testing correction in eQTL studies. Am J Hum Genet. 2015;96: 857–868.

18. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics. 2016;32: 1479–1485.

19. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95: 5–23.

20. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019;104: 410–421.

21. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91: 224–237.

22. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Variable Thresholds in Plink/Seq: Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010;86: 832–838.

23. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89: 82–93.

24. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13: 762–775.

25. He Z, Xu B, Lee S, Ionita-Laza I. Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. Am J Hum Genet. 2017;101: 340–352.

26. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315.

27. Sul JH, Han B, Eskin E. Increasing power of groupwise association test with likelihood ratio test. J Comput Biol. 2011;18: 1611–1624.

28. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008;83: 311–321.

29. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5: e1000384.

30. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369: 1318–1330.

31. Ferraro NM, Strober BJ, Einson J, Abell NS, Aguet F, Barbeira AN, et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. Science. 2020;369. doi:10.1126/science.aaz5900

32. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49: 618–624.

33. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47: D886–D894.

34. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42: D980–5.

35. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519: 223–228.

36. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM. [cited 22 Nov 2019]. Available: https://www.omim.org/

37. INSERM. Orphanet: an online database of rare diseases and orphan drugs. [cited 22 Nov 2019]. Available: http://www.orpha.net/

38. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database . 2016;2016. doi:10.1093/database/baw100

39. Downes K, Megy K, Duarte D, Vries M, Gebhart J, Hofer S, et al. Diagnostic high-throughput sequencing of 2396 patients with bleeding, thrombotic, and platelet disorders. Blood. 2019;134: 2082–2091.

40. Nachman RL, Rafii S. Platelets, petechiae, and preservation of the vascular wall. N Engl J Med. 2008;359: 1261–1270.

41. Flach EH, Rebecca VW, Herlyn M, Smalley KSM, Anderson ARA. Fibroblasts contribute to melanoma tumor growth and drug resistance. Mol Pharm. 2011;8: 2039–2049.

42. Shenenberger DW. Cutaneous malignant melanoma: a primary care perspective. Am Fam Physician. 2012;85: 161–168.

43. Shi J, Yang XR, Ballew B, Rotunno M, Calista D, Fargnoli MC, et al. Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. Nat Genet. 2014;46: 482–486.

44. Popp B, Ekici AB, Thiel CT, Hoyer J, Wiesener A, Kraus C, et al. Exome Pool-Seq in neurodevelopmental disorders. Eur J Hum Genet. 2017;25: 1364–1376.

45. Bale R, Putzer D, Schullian P. Local Treatment of Breast Cancer Liver Metastasis. Cancers . 2019;11. doi:10.3390/cancers11091341

46. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012;44: 623–630.

47. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7: 16.

48. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550: 204–213.

49. Aguet F, Barbeira AN, Bonazzola R, Brown A, Castel SE, Jo B, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv. 2019. p. 787903. doi:10.1101/787903

50. Neyman Jerzy, Pearson Egon Sharpe, Pearson Karl. IX. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London Series A, Containing Papers of a Mathematical or Physical Character. 1933;231: 289–337.

51. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease–common variant… or not? Hum Mol Genet. 2002;11: 2417–2423.

52. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13: 204–216.

53. Jansen R, Hottenga J-J, Nivard MG, Abdellaoui A, Laport B, de Geus EJ, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. Hum Mol Genet. 2017;26: 1444–1451.

54. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005;15: 1576–1583.

55. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38: 904–909.

56. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498.

57. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7: 500–507.

Chapter 4 Designing machine learning algorithms for the automated diagnosis of atopic dermatitis

4.1 Introduction

Atopic dermatitis (AD) is a type of inflammatory skin disease resulting in red, itchy, swollen, cracked, and irritated skin, which is a severe form of eczema [1]. It usually begins in childhood where about 70% of cases start in children younger than five years old while only 10% of cases start in adults[2]. Studies found that about 15-20% of children under 13 years of age are affected by AD in the United States[3,4]. In addition to the discomfort in the skin, children with AD may develop inhalant allergic diseases such as asthma and allergic rhinitis[5,6] as well as mental disorders such as anxiety and depression[7]. Hence, AD may impose a high economic burden and have considerable negative effects on life quality[8,9], which is a significant cost to society. However, there is no cure for this disease except a few treatments to relieve the symptoms[10] because its causes are complicated[11].

Recently, the important role of colonic epithelial cells (colonocytes) has been implicated in the host-microbial interactions, and these gut epithelial cells contribute to the microbiota composition and activities following gut dysbiosis, affecting many chronic human diseases[12]. In addition, integration and correlation analyses of host genes expression and gut microbiota have emerged as an important opportunity for diagnosis and prediction of human diseases including AD[13,14]; for instance, associations between enzyme commission genes and microbiota in inflammatory bowel diseases[15], and also between *IL-17* and *Streptococcus* in AD[16]. However, there have been few studies on prediction analysis using machine learning based on the gut transcriptome and microbiota in AD.

It is challenging to diagnose AD because of its variable morphology, distribution, and irregularity. Based on its main clinical features, diagnostic criteria have been developed and used worldwide[17]. Besides, the assessment of disease severity is problematic due to the lack of objective markers[18]. This is concerning as physicians need to make decisions about treatment based on the diagnosis of AD and its severity, which might be related to the prognosis. Therefore, an accurate and automated diagnosis of AD and an improved set of biomarkers for it could have a potentially high impact.

In this paper, we develop a machine learning classifier for an accurate and automated diagnosis pipeline for AD using the transcriptome of gut epithelial colonocytes and gut microbiota data. A classifier is an algorithm that implements classification, which maps the input data to specific classes. In our study, an AD classifier takes transcriptome data and/or microbiota data as input data and output the predicted status of AD. Specifically, we use transcriptome and metagenome data to achieve the comprehensive gene expression and microbiota profiles of individuals with moderate to severe AD and controls. We develop a robust machine learning pipeline including feature selection, model selection, cross-validation, classification, and follow-up statistical analyses, which can differentiate between subjects with and without AD based on the omics data with high accuracy.

4.2 Materials and methods

4.2.1 Sample collection and disease diagnosis

In this study, we collected the transcriptome of gut epithelial colonocytes and gut microbiota data of 161 subjects including 84 cases (patients with AD), 77 controls (healthy individuals). AD subjects were recruited from the Childhood Asthma Atopy Center of Asan Medical Center, Seoul, Korea, and were diagnosed in accordance with the criteria of Hanifin and Rajka[19]. All individuals are children aged from 6 months to 72 months. The SCOring AD (SCORAD) value, as an important AD assessment index for the extent and severity of AD, was assessed by pediatric allergists based on the guidelines for the SCORAD index[20]. Total serum immunoglobulin (IgE) levels in the peripheral blood were measured using the ImmunoCAP system (Phadia AB, Uppsala, Sweden). The parents and guardians of all children provided written informed consent for their participation, and this study protocol was approved by the human ethics committee at Asan Medical Center (Institutional Review Board No. 2008-0616, 2015-1031, and 2017-0923).

4.2.2 Transcriptome and microbiota data

Transcriptome data was obtained from mRNAs extracted the exfoliated colonocytes of each fecal specimen using the GeneChip Human Gene 2.0 ST Array (Affymetrix, Santa Clara, CA) under the manufacturer's protocol. Microbiota data was obtained from the fecal samples using the Power Microbial RNA/DNA Isolation kit (MO BIO/Qiagen, Carlsbad, CA, USA), polymerase chain reaction (PCR) amplification based on primers targeting the V1-V3 variable region of 16S rRNA gene, and sequencing the Roche/454 FLX Titanium system (Roche, Mannheim, Germany) and MiSeq (Illumina, San Diego, CA) under the manufacturer's instructions. Since there was the requirement of actual read counts for quality control and the difficulty in a direct comparison between these two sequencing platforms, we focused on the common phylum and genus. More detailed information on the sequencing method is provided in our previous studies [21,22].

4.2.3 AD machine learning classifier

We built the supervised machine learning pipeline that predicts atopic dermatitis status using transcriptome and microbiota data. This pipeline includes prepossessing[23], feature selection[24], model selection and improvements[25], integration of microbiota data, and performance evaluation. The pipeline is implemented with Python 3.7 and the scikit-learn package[26].

Prepossessing

Initially, there were 161 samples in the transcriptome dataset and the microbiota dataset, respectively. 160 samples were overlapped between the two datasets. For one individual with only transcriptome data, we imputed its microbiota data using a mean values imputation approach that assumes missing values are missing completely at random (MCAR). For one individual with only microbiota data, we removed this sample as it is difficult to impute its transcriptome data due to a large number of genes to impute (one hundred times more genes in the transcriptome data than microorganisms in the microbiota data). At the end of this process, we have the set of identical 161 samples in both transcriptome and microbiota datasets.

Among the 161 samples, there were 84 AD patients and 77 controls, and we split them into the training set ($n = 131$) and the test set ($n = 30$). As the numbers of cases and controls were different, we used a stratified split to guarantee that the balance of cases and controls is consistent across training and test sets. We then used min-max normalization to scale the transcriptome and microbiota data so that the data range from 0 to 1, calculated as:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

where x is the values of a feature. We use this normalization method because we want to ensure that the scaled data are positive. We normalize the training and test sets together.

Feature selection

As it is unlikely that a disease is strongly associated with more than 40,000 genes, most of the genes would be unrelated to the disease or have negligible effects. Therefore, feature selection on the training dataset is necessary to identify a subset of predictive genes, whose expression data could predict atopic dermatitis as accurately as possible. The two main aspects we considered were: (i) the optimal number of features to be selected in the entire dataset, and (ii) the exact features chosen from the original training set.

Typically, there are three types of widely used feature selection methods. They are Filter, Wrapper, and Embedded methods[27]. We selected three methods from each type: Chi-squared Test, Recursive Feature Elimination (RFE), and Random Forest Classifier (RFC), because they are efficient and applied in previous research [PMID: A Nasal Brush-based Classifier of Asthma Identified by Machine Learning Analysis of Nasal RNA Sequence Data]. Specifically, RFE requires the results from the models, and hence we chose Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest Classifier (RFC) as the models in conjunction with RFE. These combinations are referred to as LR-RFE, SVM-RFE, and RFC-RFE respectively.

After introducing the specific methods for feature selection, we should then consider the problem of overfitting. It is hard to extract correct features from high-dimensional datasets with small sample sizes.

Cross-validation (CV) is important in preventing overfitting[28]. In our task, we designed two plans (Plans A and B) for feature selection using cross-validation. Note that we do cross-validation and feature selection on the training set only so it will not cause data leakage. Plan A: we performed a 5 by 5 nested cross-validation for feature selection, which consisted of a 5-fold inner CV round and a 5-fold outer CV round. We used the outer CV on the entire training set to evaluate the model performance, and the inner CV is applied to the outer CV training split to select the set of predictive features (Supplementary Fig. S1A). In other words, supposing that the outer CV training split named *D* is used for feature selection, we executed a 5-fold CV on *D* (i.e., inner CV) and determined the optimal number of features to select in *D*, which could achieve best average performance in the inner CV. Then we calculated the overlapping features across all training splits of the outer CV. Denote the number of features in the final set as n_A. In a given inner CV training split, all the features are ranked by their weights (feature importance) assigned by the classification model trained with the inner CV training split. Then we selected top *k* features with *k* starting from 44608 (all features) and being reduced by 10% in each iteration until $k = 1$, and trained models with the inner CV training split and evaluated them with the inner CV test split (the validation set). The optimal value of *k* was selected to generate the model with the best performance. In the outer CV training split, all the features are ranked using the same method as applied to the inner CV training split. Then we selected *k* top features to identify the set of predictive features for this outer CV training split, where *k* is the optimal number of features determined in the inner CV. The outer CV test set will be used for model selection and hyper-parameter tuning with the *k* selected features in the follow-up analysis instead of feature selection. We then repeated this process over all the five outer CV training splits and yielded five sets of predictive features. Finally, we selected the intersection set of them as the final set of predictive features for the entire original training set. Plan B: instead of using a nested cross-validation, we only performed one 5-fold cross-validation on the entire training set and directly selected n_B top features on its training splits, *D* (Supplementary Fig. S1B, S2), where the features are ranked by test scores such as the p-values of χ^2 test between the features and the disease statuses. n_B is determined as the value that produces the model with the best average performance in the outer CV test sets (the validation set). The final set of predictive features are the top n_B features chosen in the entire original training set.

As mentioned before, to decide the two main aspects regarding feature selection, we considered different feature selection methods and many possible numbers of features based on a set of criteria. We used 5-fold cross-validation to evaluate the performance. In detail, we employed the average F1-scores from outer CV test sets as metrics. We compared plan A and plan B, combined with five candidate feature selection methods: RFE-LR, RFE-SVM, RFE-RFC, Chi-square Test, and RFC. Based on the comparison results (see supplementary methods), we took the following steps to select the best feature number: we first calculated feature importance in each training split in the outer CV, ranked the features by their average feature importance, and chose the top n features from the training set whose feature importance was greater than a threshold (Plan B). Finally, we chose 35 features using the Chi-squared test in the entire original training set.

Model selection and improvements

We trained four different machine learning models, 1) Logistic Regression (LR), 2) Support Vector Machine (SVM) with linear and rbf kernels, 3) Random Forest Classifier (RFC)[29], and 4) XGBoost[30] with the outer CV training splits. XGBoost is a tree boosting method, demonstrated to perform extremely well in multiple classification tasks. We choose the best model among the four models by comparing the average F1-scores on outer CV test splits.

Jittering

Jittering is a useful tool to mitigate overfitting[31]. We added random noises to the training set of the original data before normalization and feature selection. The noise followed a normal distribution of

$$s \sim N(0, \sigma^2 I)$$

where I is the intensity of noises and the variance $\boldsymbol{\sigma^2 = 1}$. Although jittering might reduce the classification accuracy of the model on training sets, proper noises could increase the robustness of the algorithm, narrowing the performance gap between training and test sets, and therefore reducing the possibility of overfitting. Note that jittering is only performed on the transcriptome data in the training set. It is because the microbiota data contain many zero values and thus adding noises to it will distort the data.

Thresholding

Moreover, we consider the effect of changing the probability threshold (p_t) in prediction. A sample is predicted to be a case by the model if the predicted probability is greater than a certain threshold (p_t) where the default value is 0.5. Different probability thresholds should be examined to see whether they could further improve the model performance. For this improvement, we first selected the model with the best performance using the default probability threshold (0.5). Then we changed the machine learning model such that it generates the probability (p_t) as output. We tested different thresholds and chose the best one based on outer CV test split evaluation.

Feature importance

After comparison, we selected the Chi-squared test as our feature selection method. We used the “SelectedKBest" function in the scikit-learn package[26] to implement the Chi-squared test. After identifying k features, we also want to rank the features based on a certain criterion. This function has an attribute named “*scores_*", and it returns the scores of features. The Chi-squared test is used to test the independence of two events.

In our binary classification problem, we have *X* as the input data with the size of (n_sample, n_feature), which are the number of samples and features respectively, and also *y* as the label of each sample with the size of n_sample. For calculation, we expand the size of *y* to (n_sample, 2). The first column of y represents the first class and the second stands for the second class. For each row, the elements will either be (1, 0) or (0, 1), which indicates that the sample of this row belongs to the first class or the second class.

After that, we calculated the observed result f_{obs},

$$f_{obs} = y^T X$$

Next, we calculated the expected result f_{exp}. To do this, first, we acquired feature_count, which is a (1, n_feature) matrix, and each column is the sum of this feature in each sample. Secondly, we obtained class_prob, which is a (1, 2) matrix, and each column is the mean of this class. Now we can get f_{exp},

$$f_{exp} = class\_prob^T \cdot feature\_count$$

Finally, we calculated the χ^2 value by the following equation,

$$\chi^2 = \frac{\left(f_{obs} - f_{exp}\right)^2}{f_{exp}}$$

where χ^2 is a (2, n_feature) matrix. We summed up the result along the column and calculated the *scores_* vector of size n_feature. It represents the scores of the features, where column *i* of *scores_* is the score of the *i*-th feature.

In our task, higher values suggest higher importance of features. Therefore, we could compare each feature relatively from their feature importance.

Integration of microbiota data

In addition to the transcriptome data, we integrated the microbiota data to improve the performance of the AD classifier. We tested four methods when incorporating the microbiota data and evaluated their performance using the outer CV test set:

1) a method that uses microbiota data only, 2) a method that uses transcriptome data only, 3) a method that combines transcriptome and microbiota data first, then performs feature selection and trains the model, and 4) a method that performs feature selection on transcriptome and microbiota data separately, then combines the two types of data and trains the model. We used similar feature selection methods as described above for microbiota data. The comparison of these four methods is in Supplementary Table S1.

Performance evaluation

We evaluated the prediction accuracy of the AD classifier using the test set, which is not used in training. We calculated several performance metrics including accuracy, precision, recall, and F1-score. In binary classification problems, we calculated those metrics as follows:

	Predicted Class		
Actual Class		True	False
	True	*a* (True Positive)	*b* (False Negative)
	False	*c* (False Positive)	*d* (True Negative)

Precision= $= \frac{a}{a+c}$, Recall= $r = \frac{a}{a+b}$, Accuracy= $\frac{a+d}{a+b+c+d}$, F1-score= $\frac{2pr}{p+r} = \frac{2a}{2a+b+c}$

We used F1-score as the main evaluation metric in this paper as it is a harmonic mean of precision and recall, leading to a more general and reliable assessment of the model performance, especially when classes are imbalanced. F1-score ranges from 0 to 1 where the performance is better when the F1-score is closer to 1. In addition to these metrics, we plotted the Receiver Operating Characteristic (ROC) curve by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings. We then calculated the area under the ROC curve (AUC), which also ranges from 0 to 1 where 1 represents the perfect performance.

Assumptions in experimental settings

In building and testing the AD classifier pipeline, we have several hyperparameters such as feature numbers, feature selection methods, and training models. As examining all combinations of hyperparameters is exponential in the number of hyperparameters, our experiments are based on the assumption that every variable or hyperparameter is weakly correlated to each other. It means that optimizing a hyper-parameter one at a time yields a similar model when optimizing all hyperparameters at the same time.

4.3 Results

4.3.1 Data description

We acquired the transcriptome profiles and the microbiota data of 161 subjects, who were recruited from the Cohort for Childhood Origin of Asthma and Allergic Diseases birth cohort and the Asthma Atopy Center of Asan Medical Center, Seoul, Korea. After preprocessing, there are 84 cases (patients with AD) and 77 controls (healthy individuals). The mean age was higher in the AD patient group than the controls (17.37±3.48 month vs. 10.81±2.15 month, $P = 0.001$), and the serum total IgE levels were significantly higher in the AD group (243.06±160.21 IU/ml vs. 22.83±9.46 IU/ml, $P = 0.004$) as summarized in Table 1. On average, subjects without AD are 3.16 months younger than individuals with the disease. After pre-processing and normalizing the raw gene expression and the microbiota data, there are 44,608 gene expression probes and 366 taxa of microorganisms used for developing a machine learning pipeline.

4.3.2 Developing atopic dermatitis classifiers

To accurately predict AD incidence, we designed two machine learning classifiers: one using only transcriptome data (Fig.1a) and the other using both transcriptome and microbiota data (Fig.1b). Both classifiers consist of several computational steps, and we describe each step briefly here (see Methods for details). First, we preprocess the data such as removing duplicates, imputing missing values, splitting the data into training and test sets, and performing normalization. Second, we perform feature selection using the training set with the cross-validation to identify the best set of features for prediction (e.g., expression of specific genes or specific taxa) as well as to choose the best machine learning model. To improve the performance of the classifier, we consider changing a few hyperparameters such as adding certain levels of noise to expression data and changing the probability threshold to classify cases and controls. Lastly, we apply the trained machine learning model and selected features to the test set for classification and evaluate the performance of the machine learning model. As our dataset consists of a high-dimensional feature set from a limited sample size, we primarily focus on developing a machine learning classifier that is robust with the small sample size, prevents overfitting, and prioritizes genes or features for prediction.

4.3.3 Evaluation of transcriptome only classifier

The transcriptome data are available from 161 individuals whose gene expression profile is measured at 44,608 probes ("features"). As a large number of features have negative effects on classification performance such as causing overfitting, we perform feature selection to identify a subset of informative features. We use a 5-fold cross-validation approach using a training set (n=131) and test several feature selection methods such as recursive feature elimination (RFE), support vector machine (SVM), and chi-squared test. We measure the performance of feature selection methods using F1 score and find that the chi-squared test approach selecting about 35 features from the training set has the best performance. So we decide to select 35 features with feature importance ≥ 0.95 which can achieve the highest performance (Supplementary Fig. S3). Once we identify the set of best features or genes for prediction, we next seek to identify the best machine learning model for prediction. We test several machine learning (ML) models such as Logistic Regression (LR), Support Vector Machine (SVM), Random Forest Classifier (RFC), and XGBoost. We use a 5-fold cross-validation using a training set and 35 features to evaluate the performance of each ML model and find that SVM with the rbf kernel has the highest F1 score (Supplementary Table S2). To improve the performance of our ML classifier, we vary the probability threshold (p_t) when making predictions on the case-control status; an individual is predicted to be a case if the predicted probability is greater than p_t and by default, p_t is 0.5. Results show that p_t of 0.3 generates the best F1 score using a 5-fold cross-validation (Supplementary Table S3). Lastly, another improvement we make to the ML classifier is jittering, which is adding random noises to transcriptome data. With jittering, it may be difficult for the machine learning models to fit the data, and therefore it may enhance the generalization ability and reduce the overfitting. We add different levels of noise and observe the highest F1 score with a noise level of $I = 0.001$ using SVM on a 5-fold cross-validation (Supplementary Table S4).

After we identify the best ML model and features as well as improvements based on the performance using the training set, we evaluate the AD classifier on the test set (n=30). We use a variety of metrics such as F1 score, accuracy, precision, recall, and the area under the curve (AUC) under the receiver operating characteristic (ROC) curve. In addition to the best AD classifier we identified, we also test classifiers without feature selection and the improvements to observe their impact on the performance. Specifically, we test four models: 1) SVM with all features, 2) SVM with the best 35 features, 3) SVM with all features and with jittering and best p_t threshold, and 4) SVM with the best 35 features and with jittering and best p_t threshold.

Results show that feature selection improves the performance as expected; both SVM models with feature selection have higher F1 scores (0.76) than models without feature selection (0.71 and 0.73, Table 2). However, the impact of feature selection is not dramatic as the F1 score improves by at most 0.05, similar to the improvement in the training set (from an F1 score of 0.7397 without feature selection to an F1 score of 0.7809 with feature selection). Also, improvements that include jittering and the best p_t threshold do not improve the performance as the F1 score of the SVM model with those improvements is identical to that without the improvements in the test set, although we observe higher F1 scores with the improvements in the training set where we observed F1 score of 0.80 with the best p_t threshold and F1 score of 0.78 without the improvement. In terms of AUC under the ROC curve, the best AUC is observed when using all features (AUC of 0.75) while the SVM models with feature selection have slightly lower AUC (0.72, Fig. 2). The modest improvements in performance by feature selection and other improvements in the test set may be due to the small sample size of the test set. We also examined the performance of our AD classifier with only 19 of the 35 selected probes, which explicitly represents expressed genes with gene symbols (Supplementary Table S5). We observed the greatly increased performance of our AD classifier, which achieved an F1 score of

0.84 (Supplementary Table S6). Interestingly, applying jittering and thresholding did not further improve its performance. It was probably due to the smaller number of selected features that were more representative and informative. So the overfitting issue might be mitigated and thus it is unnecessary to use jittering and thresholding.

4.3.4 Evaluation of the classifier with microbiota data

In addition to the transcriptome data, we have microbiota data from 161 individuals with 366 phylum and genus features, and we build the ML classifier that uses both transcriptome and microbiota data (Fig. 1). Similar to the transcriptome-only classifier, we first perform feature selection on the microbiota features using a training set (n=131) with the same cross-validation approach and feature selection methods. If using microbiota data only, we observe the best performance in terms of F1 score (0.73) with the SVM approach using 25 microbiota features (Supplementary Table S1). Additionally, we perform feature selection after combining microbiota and transcriptome data and find that 50 microbiota and 35 transcriptome features generate higher F1 scores (Supplementary Table S1). As for the other improvements in the ML model, we use the same probability threshold ($p_t = 0.3$) and noise level ($I = 0.001$) as ones we used for the transcriptome-only classifier; these thresholds and noise levels also generate the best performance (Supplementary Table S7, S8).

Next, we evaluate the prediction ability of the microbiota data on AD using six different classifiers with a test set (n=30): 1) SVM using all microbiota features, 2) SVM using the best 25 microbiota features, 3) SVM using the best 50 microbiota and 35 transcriptome features, 4) the first, 5) the second, and 6) the third models with p_t and jittering improvements. Results show that the classifiers that combine microbiota and transcriptome data (the third and sixth models) are most accurate in predicting AD, achieving an F1 score of 0.78. (Table 3). The classifiers that use only microbiota data generally have lower accuracy (F1 scores between 0.69 and 0.74) than ones that use both microbiota and transcriptome data. Compared to the previous transcriptome-only classifiers that have the best F1 score of 0.76, the microbiota data marginally improves the classifier performance to an F1 score of 0.78. In terms of area under the ROC curve (AUC), the microbiota data does not improve the performance compared to the transcriptome-only classifiers as the best AUC is identical (0.75) between the classifier that combines microbiota and transcriptome data and the transcriptome-only classifier (Fig. 3). Additionally, we selected 19 transcriptome features with gene names and 50 microbiota features to train our AD classifier. We found that it did not perform better than using all 35 selected transcriptome features and 50 microbiota features, where it achieved an F1 score of 0.7273 initially and 0.7778 after applying jittering and thresholding (Supplemental Table S9).

4.3.5 Top genes selected in the AD classifier

Our feature selection algorithm using the transcriptome data identifies 35 features or probes (Fig. 4) that span over 19 unique genes. These genes are selected as they are most informative in predicting AD, which suggests they may be implicated in AD. Hence, we perform a literature search for these 19 genes to examine whether they are known to be related to AD and find a few cases. First, we find that *GRP1* (Probe ID: 16907572) is associated with a type of scaffold protein (Grasp) that potentially influences p53-mediated apoptotic responses in skin[32]. It is known that apoptosis is a crucial process in the development of AD[33,34]. Second, another gene called *CCL22* (Probe ID: 16819478) is known to play an important role in AD pathogenesis. It encodes chemokine (C-C motif) ligand 22, which is involved in the immunoregulatory and inflammatory processes of T cells. Additionally, *CCL22* is found to be one of the important biomarkers of severity in infantile AD according to a study involving 34 patients[35]. This gene has also been reported to be significantly up-regulated with AD in a high-throughput proteomic assay[36] and a transcriptomic analysis[37]. Association studies and functional studies further suggest that the mutations in *CCL22* affect the susceptibility to AD in a gain-of-function manner[38]. Lastly, according to a genome-wide association study, four SNPs associated with Alopecia are mapped to our selected gene, *TTC27* (Probe ID: 16878890)[39]. A previous study found that patients with Alopecia have a higher risk for AD[40]. Overall, these examples demonstrate the clinical importance of our selected genes.

4.3.6 Top microorganisms in microbiota selected in the AD classifier

Our feature selection algorithm using the microbiota data and transcriptome data identifies 50 microbiota features (Fig. 5). These microorganisms are chosen to be top predictors for AD, so they may be involved in AD. To validate our findings, we perform a literature search for these 50 kinds of bacteria to look for related studies and supporting evidence. Here are some examples. First, *Akkermanisia* has the highest feature importance in our AD classifier, indicating that the amounts of *Akkermanisia* can affect our prediction the most. A recent study found that *Akkermanisia* is high in transient AD cases but low in children with persistent AD[41]. So *Akkermanisia* can be a crucial microbiota indicator for AD. Second, a metagenomic analysis of microbe-derived extracellular vesicles discovered that *Verrucomicrobia*, the bacteria with the second highest feature importance in our AD classifier, had significantly different relative abundances between the AD and control groups and could be used as a novel biomarker for AD diagnosis[42]. Lastly, *Propionibacterium* is ranked the sixth most important microorganism in our AD classifier. It was reported that the relative abundance of *Propionibacterium* is usually reduced and the abundance of *Staphylococcus aureus* is elevated in the skin of AD patients[43], leading to dysbiosis. Another study observed a dysbiotic status characterized by a reduction of

Propionibacterium in the gut microbiota of AD patients[44]. And dysbiosis is considered to be an essential driving factor of AD[45,46]. Hence, the selected 50 microbiota features demonstrate the close relationship between gut microbiota and the pathogenesis of AD[47].

4.4 Discussion

AD is a paradigmatic chronic inflammatory skin disease characterized by complex pathophysiology and a wide spectrum of clinical phenotypes. In particular, the phenotype of AD in early childhood may be influenced by genetic factors and gut microbiota. The purpose of this study was to predict the phenotype of atopic dermatitis in early childhood with transcriptome and microbiota data. Therefore, in order to understand this diversity, efforts to find new AD endotypes by ML technique using these multi-omics are needed. In this study, we integrated and took the advantage of one of the largest transcriptomic and microbial profiles for AD patients to the best of our knowledge. We developed an AD classifier solely based on transcriptome and microbiota data, which accurately distinguished subjects with AD from healthy individuals. The most accurate classifier selected 35 genes and 50 microbial features (4 phyla and 46 genera) interpreted via a support vector machine classifier, which can automatically classify AD with high precision (0.70) and recall (0.88). Also, among the selected genes/probes used in the AD classifier, we discovered that at least three genes are reported to be directly or indirectly associated with AD. In summary, our classifier represents the first step toward a precise, automated diagnosis of AD and provides important biological insights into the development of the biomarkers of this disease.

Recently, our colleagues have developed an estimated prediction model by multi-omics analyses and realized the importance of transcriptome data[48]; therefore, this study performed the extended analyses using a larger sample size and a different machine learning model for a more precise prediction. Our AD classifier is the first machine learning classifier for this disease based on the transcriptomic and microbial profiles of patients. To diagnose AD, clinicians typically rely on the clinical features of AD[17,49]. However, the lack of robust objective measures might have negative effects on the assessment of AD[17,18,50,51]. To overcome these challenges, previous studies developed machine learning classifiers for AD diagnosis or severity evaluation based on electronic health records (EHR)[52], camera photos[53], or multiphoton tomography[54]. While these approaches may provide an unbiased diagnosis of AD, they are either not highly accurate, achieving F1 scores of 0.67 using EHR[52] and F1 scores of 0.69 using camera photos[53], or it may be more inconvenient or expensive to obtain these kinds of data for the ML classifier. With the development of high-throughput microarray and sequencing technologies, it may be and is likely to be cheaper in near future to obtain transcriptome and microbiota data. Another advantage of our ML classifier is that it does not require patients to be present in the testing sites or hospitals as they can send their samples to the labs to generate transcriptome and microbiota data and our classifier can predict the risk of AD based on the data. Thus, our classifier enables the convenient, efficient, and cost-effective diagnosis of AD as well as improving the accessibility of medical resources for patients.

In further enrichment analysis using Enrichr (https:// https://maayanlab.cloud/Enrichr/)[55] based on the featured genes, interleukin-7 (IL-7) interactions in the immune response pathway (P = 0.032, Supplementary Table S10) was enriched. IL-7 is a critical cytokine for the development of the group 2 innate lymphoid cells (ILC2s), which are involved in allergic diseases including AD[56]. In addition, several inflammation-related processes (for instance, lymphocyte, chemokine, neutrophil, P < 0.05) were enriched in gene ontology observation. Inflammatory responses associated with lymphocyte, chemokine, and neutrophil are important in AD mediated by CD4+ T cells[57]. These results suggest that featured genes in this study might be potentially valuable for AD diagnosis.

There are a few study limitations. First, the sample size of our dataset is relatively small. As we only used 161 samples recruited from the birth cohort follow-up group and outpatients group, it could cause overfitting and biases when training the ML classifier. To address this issue, we applied nested cross-validation to perform feature selection and model training. We also introduced jittering to add a small amount of artificial noise into the data to reduce overfitting. We showed that we successfully controlled the biases and overfitting as our classifier performed well on the independent test set. Our study also had the limited ability to assess the benefits of adding microbiota data to the ML classifier as we observed the marginal improvement in prediction accuracy, possibly due to the small sample size of the test set when we combined transcriptome and microbiota data. Second, since our subjects from the birth cohort follow-up group are general population and usually considered to have mild severity of AD, there is a possibility that the results may differ from those in the severe patient group. Therefore, in order to validate and improve the ML classifier and to more accurately assess its performance, further studies in a larger sample size and in an independent cohort are required. Third, age should be considered as a confounding factor to affect the gene expression and gut microbiota in infants through developmental stages. The strengths of our study could be an application of non-invasive gut epithelial cells from fecal specimens and a possibility to apply to early prediction for AD patients with mild severity and the general population. In addition, to address the issue of no validation, we created an independent test set from the original dataset and demonstrated the accuracy of our classifier, which could serve as the independent cohort.

In this study, we developed an accurate and automated machine learning pipeline for atopic dermatitis classification. This pipeline can not only be used to predict this skin disease but also be generalized to classify other diseases based on transcriptome and microbiota data. It could assist clinicians in diagnosing and assessing diseases and providing timely treatment to patients and provide new endotypes with performing further research. In addition, we demonstrated the utility of combining genomics and cutting-edge artificial intelligence (AI) technologies like machine learning to detect diseases or identify biomarkers. We expect that the increasing availability of genomics and AI technologies would improve the effectiveness and efficiency of medical diagnosis.

Tables

Table 4.1: Baseline characteristics of the subjects in this study

	All	Cases (AD)	Control (No AD)	Cases (AD) vs Control (No AD) t-test p-value
Average Age: months	14.21±2.14	17.37±3.48	10.81±2.15	0.001
Sex: female	72	32	40	-
*SCORAD	-	32.86±5.49	-	-
Total IgE (IU/ml)	135.191±83.53	243.06±160.21	22.83±9.46	0.004

*SCORAD: SCOring AD value, an AD assessment index that is only available for patients

Table 4.2: The results on different methods with transcriptome data only

Feature selection method (number of features) + Classification method	F1 score	Accuracy	Precision	Recall
All features (44608) + SVM (rbf)	0.7272	0.6000	0.5714	**1.0000**
chi-squared test (35) + SVM (rbf)	**0.7647**	**0.7333**	**0.7222**	0.8125
All features (44608) + SVM (rbf), with noise (*I* = 0.001) and probability threshold = 0.3	0.7111	0.5667	0.5517	**1.0000**
chi-squared test (35) + SVM (rbf), with noise (*I* = 0.001) and probability threshold = 0.3	**0.7647**	**0.7333**	**0.7222**	0.8125

The first method trained the model on the original training set without feature selection. The second method performed feature selection by chi-squared test and selected 35 features. For the last two methods, they are similar with the first two methods respectively while the only difference was that they added the noise and changed the probability threshold. The random seed of the noise was 21, which was the best result on this intensity (*I* = 0.001).

Table 4.3: The first and second methods used microbiota data only

Feature selection method (number of features) + Classification method	F1 score	Accuracy	Precision	Recall
All features (366) + SVM (rbf)	0.7111	0.5667	0.5517	**1.0000**
chi-squared test (25) + SVM (rbf)	0.7442	0.6333	0.5926	**1.0000**
chi-squared test (85) + SVM (rbf)	**0.7778**	**0.7333**	**0.7000**	0.8750
All features (366) + SVM (rbf), with probability threshold = 0.3	0.6957	0.5333	0.5333	**1.0000**
chi-squared test (25) + SVM (rbf), with probability threshold = 0.3	0.7111	0.5667	0.5517	**1.0000**
chi-squared test (85) + SVM (rbf), with noise (*I* = 0.001) and probability threshold = 0.3	**0.7778**	**0.7333**	**0.7000**	0.8750

The first method trained the model on the original training set without feature selection. The second method did feature selection by chi-squared test and selected 25 features, while the third method used both transcriptome and microbiota data, and integrated the data using the fourth plan mentioned above, and selected 85 features (35 for transcriptome and 50 for microbiota). For the last three methods, they are similar with the first three methods respectively. The only difference was that they changed the threshold and added noises.

Figures

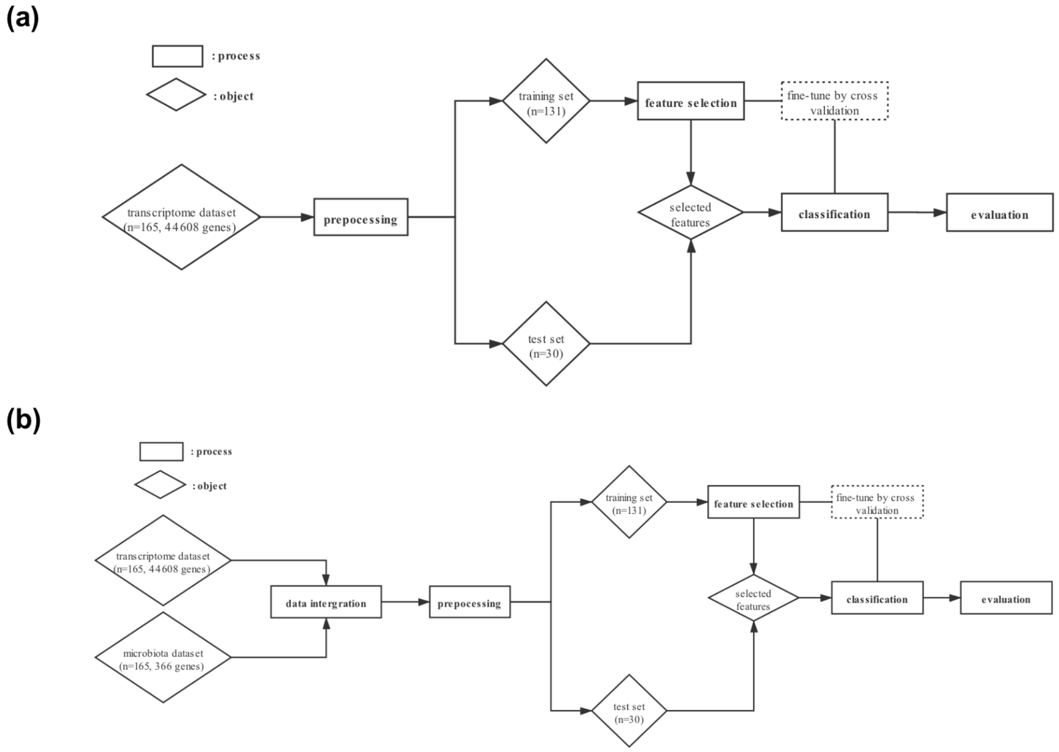

Figure 4.1: The overview of atopic dermatitis classification pipelines in two settings. (a) Transcriptome dataset only, and (b) Transcriptome and microbiota data.

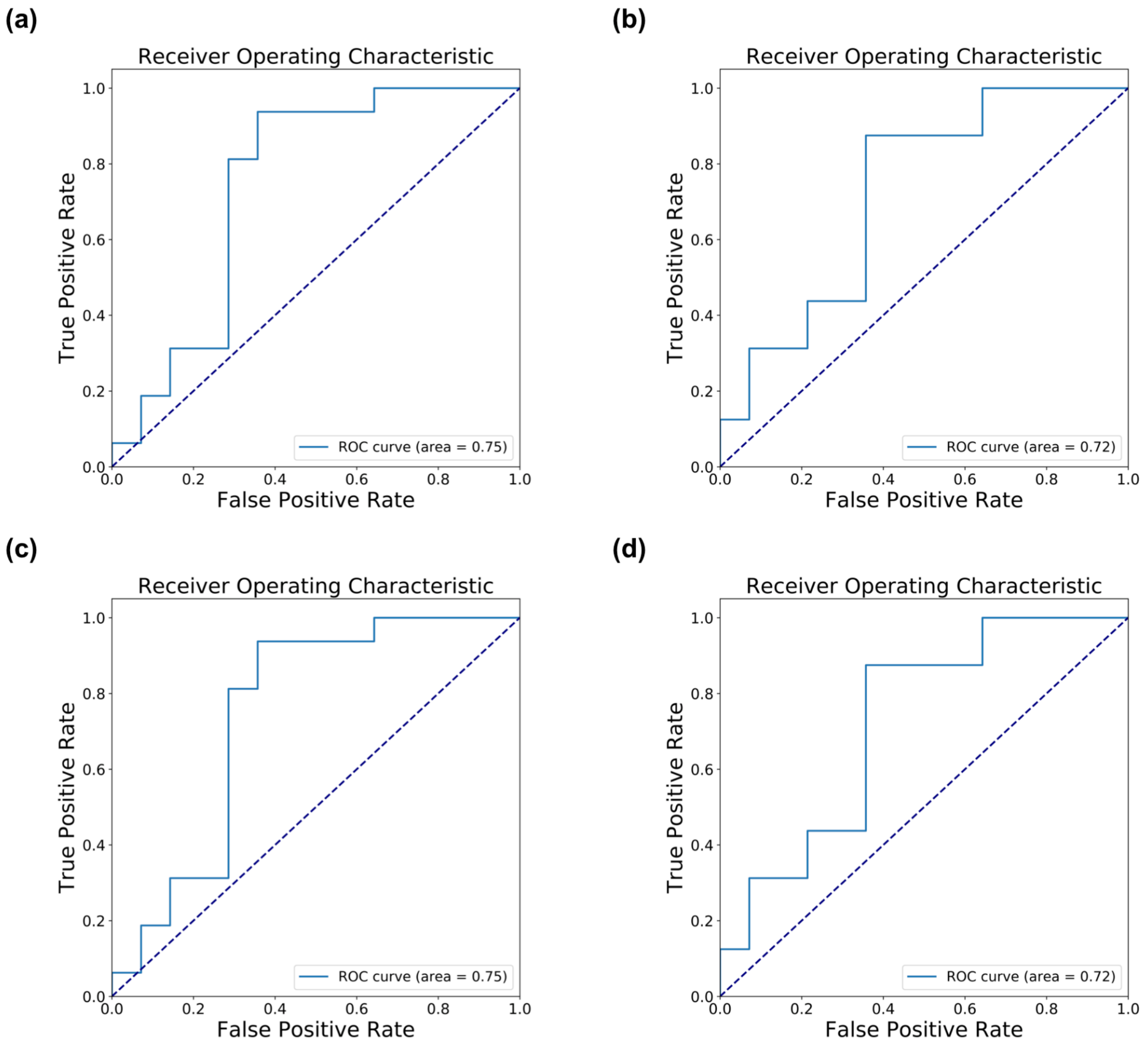

Figure 4.2: The ROC curve of the test set with transcriptome data only. (a) All features (44608) + SVM (rbf). (b) chi-squared test (35) + SVM (rbf). (c) All features (44608) + SVM (rbf), with noise (I = 0.001) and probability threshold = 0.3. (d) chi-squared test (35) + SVM (rbf), with noise (I = 0.001) and probability threshold = 0.3.

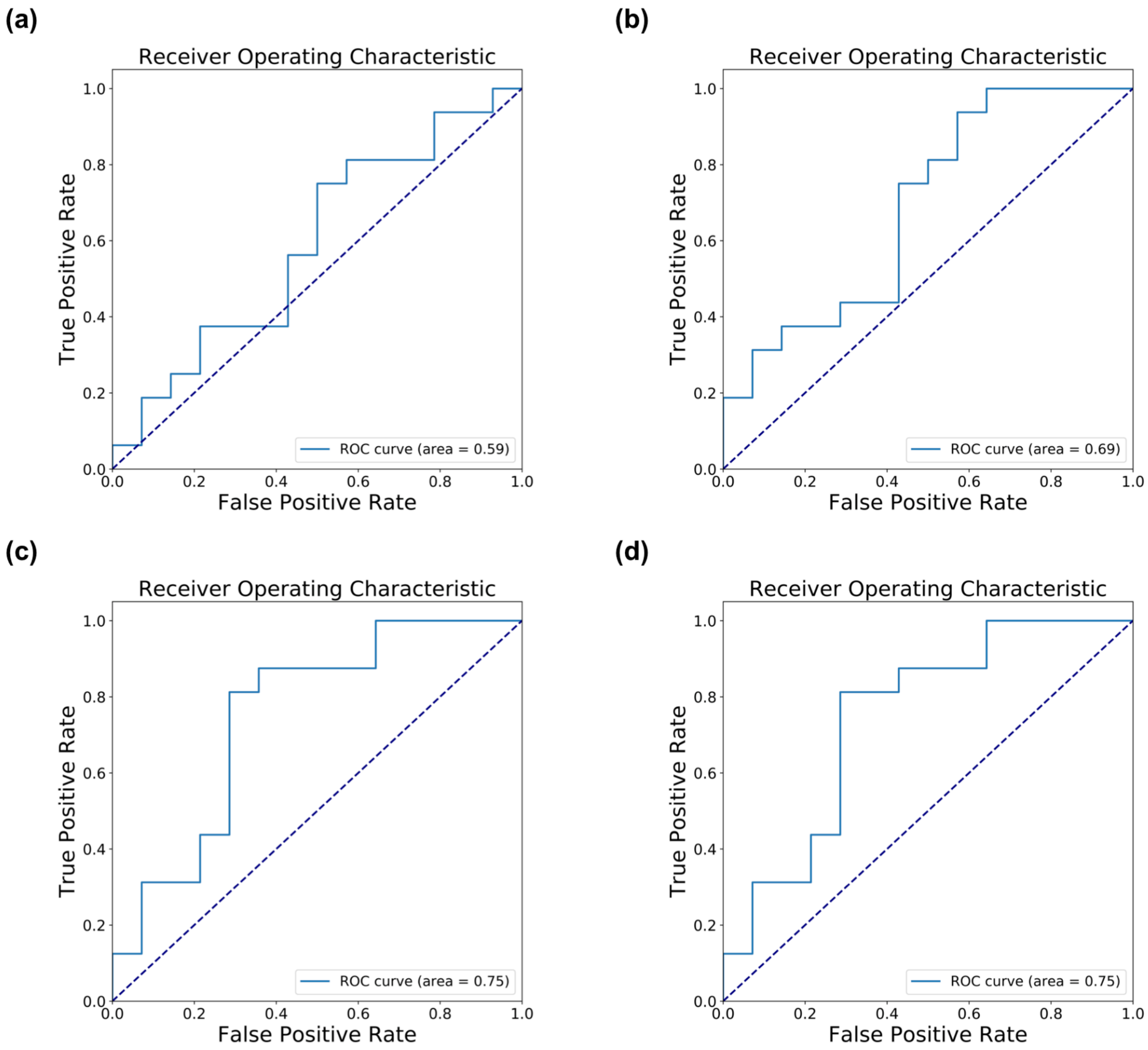

Figure 4.3: The ROC curve of the test set with microbiota data. (a) All features (366) + SVM (rbf). (b) chi-squared test (25) + SVM (rbf). (c) chi-squared test (85) + SVM (rbf). (d) chi-squared test (85) + SVM (rbf), with noise (I = 0.001) and probability threshold = 0.3. For panel (a) and (b), we only use microbiota data, while for (c) and (d) we also include transcriptome data.

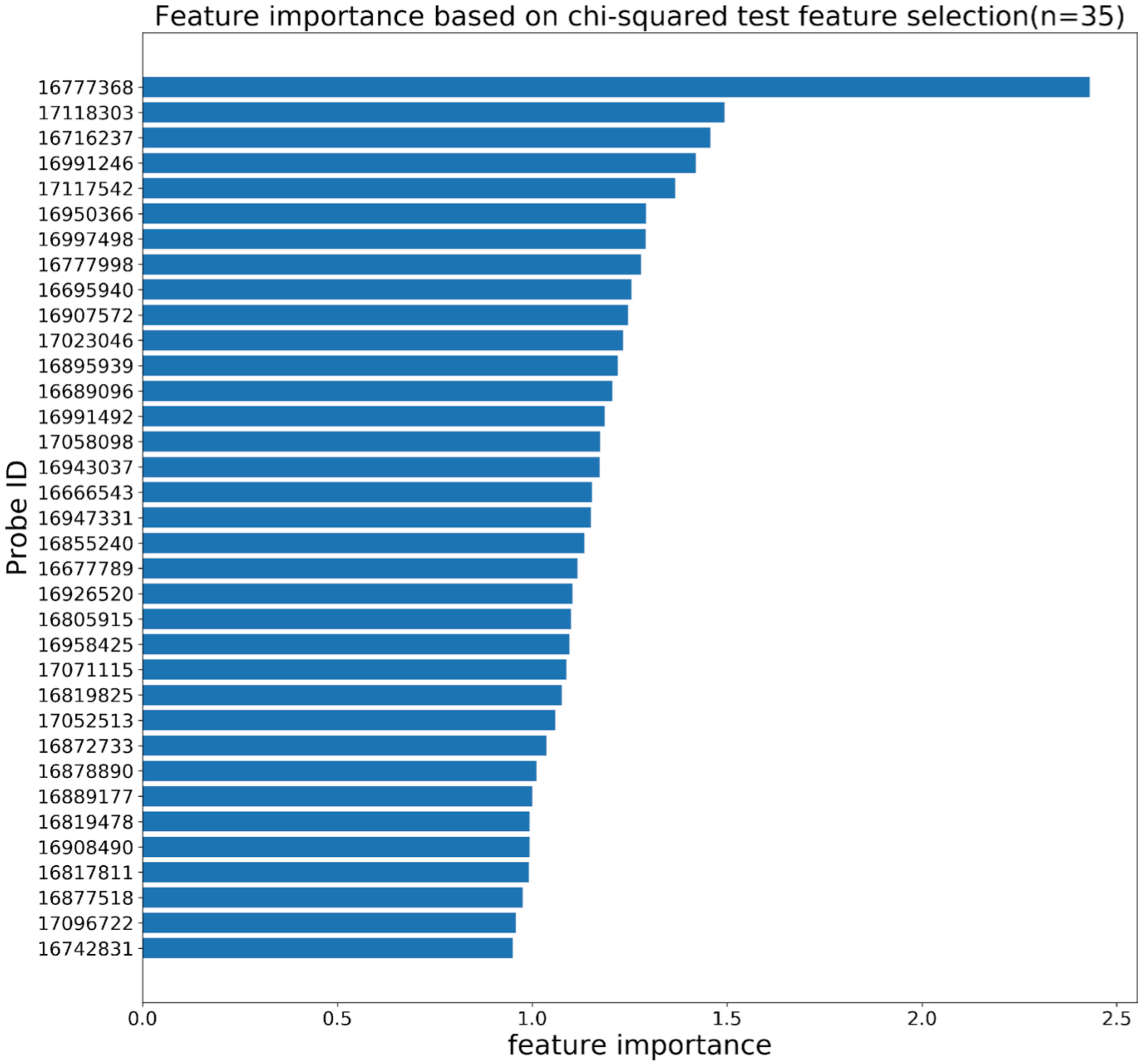

Figure 4.4: The average feature importance of the top 35 selected probes/genes. See more detailed annotation information in Supplementary Table S5.

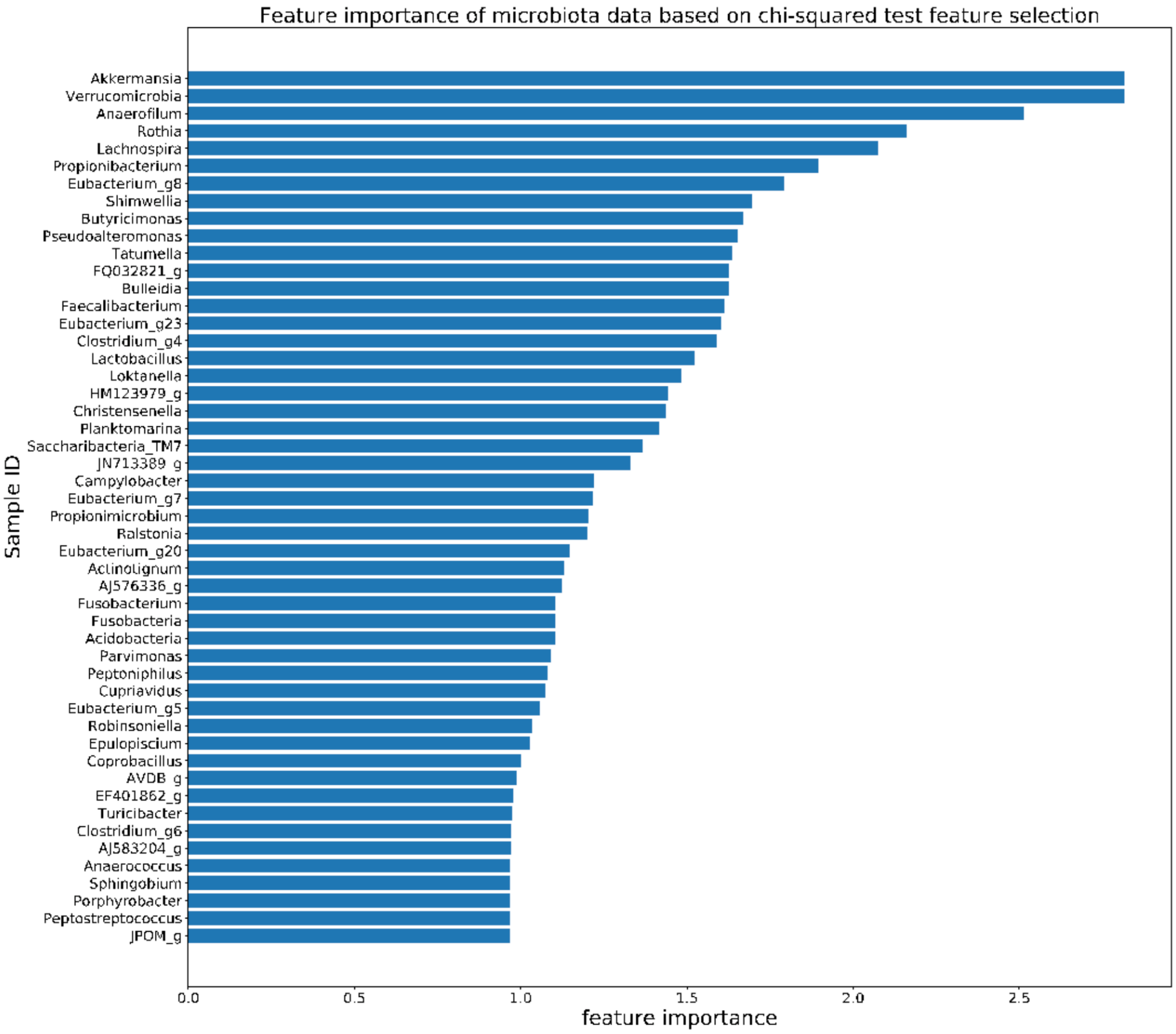

Figure 4.5: The average feature importance of the top 50 selected microorganisms from the microbiota dataset.

4.5 References

1. Aoki T, Fukuzumi T, Adachi J, Endo K, Kojima M. Re-evaluation of skin lesion distribution in atopic dermatitis. Analysis of cases 0 to 9 years of age. *Acta Derm Venereol Suppl* 1992;**176**:19–23.

2. Williams HC. Epidemiology of atopic dermatitis. *Clin Exp Dermatol* 2000;**25**:522–529.

3. Hong S, Son DK, Lim WR, Kim SH, Kim H. The prevalence of atopic dermatitis, asthma, and allergic rhinitis and the comorbidity of allergic diseases in children. *health and toxicology* Published Online First: 2012.https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3282234/

4. Mortz CG, Lauritsen JM, Bindslev-Jensen C, Andersen KE. Prevalence of atopic dermatitis, asthma, allergic rhinitis, and hand and contact dermatitis in adolescents. The Odense Adolescence Cohort Study on Atopic Diseases and Dermatitis. *Br J Dermatol* 2001;**144**:523–532.

5. Spergel JM. From atopic dermatitis to asthma: the atopic march. *Ann Allergy Asthma Immunol* 2010;**105**:99–106; quiz 107–109, 117.

6. Luoma R, Koivikko A, Viander M. Development of asthma, allergic rhinitis and atopic dermatitis by the age of five years. A prospective study of 543 newborns. *Allergy* 1983;**38**:339–346.

7. Silverberg JI, Gelfand JM, Margolis DJ, Boguniewicz M, Fonacier L, Grayson MH et al. Symptoms and diagnosis of anxiety and depression in atopic dermatitis in U.S. adults. *Br J Dermatol* 2019;**181**:554–565.

8. Kemp AS. Cost of illness of atopic dermatitis in children. *Pharmacoeconomics* 2003;**21**:105–113.

9. Jenner N, Campbell J, Marks R. Morbidity and cost of atopic eczema in Australia. *Australas J Dermatol* 2004;**45**:16–22.

10. Reed B, Blaiss MS. The burden of atopic dermatitis. *Allergy Asthma Proc* 2018;**39**:406–410.

11. Olesen AB, Juul S, Thestrup-Pedersen K. Atopic dermatitis is increased following vaccination for measles, mumps and rubella or measles infection. *Acta Derm Venereol* 2003;**83**:445–450.

12. Litvak Y, Byndloss MX, Bäumler AJ. Colonocyte metabolism shapes the gut microbiota. *Science* 2018;**362**. doi:10.1126/science.aat9076

13. Ghosh D, Bernstein JA, Khurana Hershey GK, Rothenberg ME, Mersha TB. Leveraging Multilayered 'Omics' Data for Atopic Dermatitis: A Road Map to Precision Medicine. *Front Immunol* 2018;**9**:2727.

14. Sacco KA, Milner JD. Gene-environment interactions in primary atopic disorders. *Curr Opin Immunol* 2019;**60**:148–155.

15. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;**569**:655–662.

16. Kang MJ, Lee SY, Park YM, Kim BS, Lee MJ, Kim JH et al. Interactions Between IL-17 Variants and Streptococcus in the Gut Contribute to the Development of Atopic Dermatitis in Infancy. *Allergy Asthma Immunol Res* 2021;**13**:404–419.

17. Williams HC, Jburney PG, Pembroke AC, Hay RJ, Party ADDCW. The UK Working Party's diagnostic criteria for atopic dermatitis. III. Independent hospital validation. *Br J Dermatol* 1994;**131**:406–416.

18. Williams H. Objective Measures of Atopic Dermatitis Severity: In Search of the Holy Grail.

Arch Dermatol 2003;**139**:1490–1492.

19. HANIFIN, J. M. Diagnostic features of atopic dermatitis. *Acta Derm Venereol* 1980;**92**:44–47.

20. Kunz B, Oranje AP, Labrèze L, Stalder JF, Ring J, Taïeb A. Clinical validation and guidelines for the SCORAD index: consensus report of the European Task Force on Atopic Dermatitis. *Dermatology* 1997;**195**:10–19.

21. Lee M-J, Kang M-J, Lee S-Y, Lee E, Kim K, Won S et al. Perturbations of gut microbiome genes in infants with atopic dermatitis according to feeding type. *J Allergy Clin Immunol* 2018;**141**:1310–1319.

22. Park J-U, Oh B, Lee JP, Choi M-H, Lee M-J, Kim B-S. Influence of Microbiota on Diabetic Foot Wound in Comparison with Adjacent Normal Skin Based on the Clinical Features. *Biomed Res Int* 2019;**2019**:7459236.

23. Shopov V, Markova V. Impact of Data Preprocessing on Machine Learning Performance. In: *Proceedings of the International Conference on Information Technologies,(InfoTech-2013)*. researchgate.net 2013: 187–192.

24. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–2517.

25. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* 2007;**160**:3–24.

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn:

Machine Learning in Python. *J Mach Learn Res* 2011;**12**:2825–2830.

27. Reunanen J. *Overfitting in feature selection: Pitfalls and solutions*. Aalto University 2012

28. Liu H, Dougherty ER, Dy JG, Torkkola K, Tuv E, Peng H et al. Evolving feature selection. *IEEE Intell Syst* 2005;**20**:64–76.

29. Breiman L. Random Forests. *Mach Learn* 2001;**45**:5–32.

30. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery 2016: 785–794.

31. Loughrey J, Cunningham P. Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets. In: *Research and Development in Intelligent Systems XXI*. Springer London 2005: 33–43.

32. Venkataraman A, Coleman DJ, Nevrivy DJ, Long T, Kioussi C, Indra AK et al. Grp1-associated scaffold protein regulates skin homeostasis after ultraviolet irradiation. *Photochem Photobiol Sci* 2014;**13**:531–540.

33. Trautmann A, Akdis M, Blaser K, Akdis CA. Role of dysregulated apoptosis in atopic dermatitis. *Apoptosis* 2000;**5**:425–429.

34. Trautmann A, Akdis M, Schmid-Grendelmeier P, Disch R, Bröcker EB, Blaser K et al. Targeting keratinocyte apoptosis in the treatment of atopic dermatitis and allergic contact dermatitis. *J Allergy Clin Immunol* 2001;**108**:839–846.

35. Nakazato J, Kishida M, Kuroiwa R, Fujiwara J, Shimoda M, Shinomiya N. Serum levels of

Th2 chemokines, CCL17, CCL22, and CCL27, were the important markers of severity in infantile atopic dermatitis. *Pediatr Allergy Immunol* 2008;**19**:605–613.

36. Brunner PM, Suárez-Fariñas M, He H, Malik K, Wen H-C, Gonzalez J et al. The atopic dermatitis blood signature is characterized by increases in inflammatory and cardiovascular risk proteins. *Sci Rep* 2017;**7**:8707.

37. Brunner PM, Israel A, Leonard A, Pavel AB, Kim HJ, Zhang N et al. Distinct transcriptomic profiles of early-onset atopic dermatitis in blood and skin of pediatric patients. *Ann Allergy Asthma Immunol* 2019;**122**:318–330.e3.

38. Hirota T, Saeki H, Tomita K, Tanaka S, Ebe K, Sakashita M et al. Variants of C-C motif chemokine 22 (CCL22) are associated with susceptibility to atopic dermatitis: case-control studies. *PLoS One* 2011;**6**:e26987.

39. Hagenaars SP, Hill WD, Harris SE, Ritchie SJ, Davies G, Liewald DC et al. Genetic prediction of male pattern baldness. *PLoS Genet* 2017;**13**:e1006594.

40. Mohan GC, Silverberg JI. Association of Vitiligo and Alopecia Areata With Atopic Dermatitis. JAMA Dermatology. 2015;**151**:522.

41. Park YM, Lee SY, Kang MJ, Kim BS, Lee MJ, Jung SS et al. Imbalance of Gut Streptococcus, Clostridium, and Akkermansia Determines the Natural Course of Atopic Dermatitis in Infant. *Allergy Asthma Immunol Res* 2020;**12**:322–337.

42. Yang J, McDowell A, Seo H, Kim S, Min TK, Jee YK et al. Diagnostic Models for Atopic Dermatitis Based on Serum Microbial Extracellular Vesicle Metagenomic Analysis: A Pilot Study. *Allergy Asthma Immunol Res* 2020;**12**:792–805.

43. Bjerre RD, Bandier J, Skov L, Engstrand L, Johansen JD. The role of the skin microbiome in atopic dermatitis: a systematic review. *Br J Dermatol* 2017;**177**:1272–1278.

44. Reddel S, Del Chierico F, Quagliariello A, Giancristoforo S, Vernocchi P, Russo A et al. Gut microbiota profile in children affected by atopic dermatitis and evaluation of intestinal persistence of a probiotic mixture. *Sci Rep* 2019;**9**:4996.

45. Kobayashi T, Glatz M, Horiuchi K, Kawasaki H, Akiyama H, Kaplan DH et al. Dysbiosis and Staphylococcus aureus Colonization Drives Inflammation in Atopic Dermatitis. *Immunity* 2015;**42**:756–766.

46. Dainichi T, Kitoh A, Otsuka A, Nakajima S, Nomura T, Kaplan DH et al. The epithelial immune microenvironment (EIME) in atopic dermatitis and psoriasis. *Nat Immunol* 2018;**19**:1286–1298.

47. Edslev SM, Agner T, Andersen PS. Skin Microbiome in Atopic Dermatitis. *Acta Derm Venereol* 2020;**100**:adv00164.

48. Park J, Lutz SM, Choi S, Lee S, Park S, Kim K et al. Multi-omics analyses implicate EARS2 in the pathogenesis of atopic dermatitis. *Allergy* Published Online First: 31 March 2021. doi:10.1111/all.14837

49. Eichenfield LF, Tom WL, Chamlin SL, Feldman SR, Hanifin JM, Simpson EL et al. Guidelines of care for the management of atopic dermatitis: section 1. Diagnosis and assessment of atopic dermatitis. *J Am Acad Dermatol* 2014;**70**:338–351.

50. Charman C, Chambers C, Williams H. Measuring atopic dermatitis severity in randomized controlled clinical trials: what exactly are we measuring? *J Invest Dermatol* 2003;**120**:932–

941.

51. Brenninkmeijer EEA, Schram ME, Leeflang MMG, Bos JD, Spuls PI. Diagnostic criteria for atopic dermatitis: a systematic review. *Br J Dermatol* 2008;**158**:754–765.

52. Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A Machine Learning Algorithm for Identifying Atopic Dermatitis in Adults from Electronic Health Records. In: *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 2017: 83–90.

53. Pan K, Hurault G, Arulkumaran K, Williams HC, Tanaka RJ. EczemaNet: Automating Detection and Severity Assessment of Atopic Dermatitis. In: *Machine Learning in Medical Imaging*. Springer International Publishing 2020: 220–230.

54. Guimarães P, Batista A, Zieger M, Kaatz M, Koenig K. Artificial Intelligence in Multiphoton Tomography: Atopic Dermatitis Diagnosis. *Sci Rep* 2020;**10**:7968.

55. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–W97.

56. Kabata H, Moro K, Koyasu S. The group 2 innate lymphoid cell (ILC2) regulatory network and its underlying mechanisms. *Immunol Rev* 2018;**286**:37–52.

57. Bień K, Żmigrodzka M, Orłowski P, Fruba A, Szymański Ł, Stankiewicz W et al. Involvement of Fas/FasL pathway in the murine model of atopic dermatitis. *Inflamm Res* 2017;**66**:679–690.

Chapter 5 Conclusions

The development of machine learning applications and statistical methods has become very important in the field of human genetics and genomics because a large amount of omics data has been generated to uncover the genetic basis of diseases and complex traits. As the sequencing cost decreases quickly, GWAS have been widely used to detect associations between genetic variants and diseases or complex traits. Traditional approaches used in GWAS have limitations to achieve other research goals beyond simple associations. Therefore, new methods have to be developed to address the challenges in GWAS. Current proceedings of machine learning and statistics allow us to design more accurate or powerful statistical approaches for the analysis of genetic data.

The first problem I tackled in my thesis was to develop a statistical approach to perform quality control on genetic variants. Variant quality control is an important step before GWAS. Previous filtering and classification approaches have many limitations. My method was one of the first works that combined filtering and classification approaches to perform variant quality control. Our method could improve the quality of genetic variants more than other methods. And it is very scalable. I believe that my approach will be very useful for other researchers. It also provides the science community with new ideas to develop methods for variant quality control with machine learning techniques.

Next, I worked on developing a novel powerful statistical approach to detect the functional effects of rare variants. Numerous statistical methods have been developed for rare variant association tests. However, to the best of my knowledge, they neither are designed specifically for analyzing the regulatory effects of rare variants nor attempt to incorporate the causal statuses of rare variants in the association. Identifying causal variants and utilizing this information can greatly improve the power of tests. My method, LRT-q, employs functional annotations of rare variants, observational genotype data, and quantitative phenotype data to identify potential causal rare variants by aggregating statistics of rare variants in a nonlinear manner. I believe that it is more powerful than current methods, and will be of significant interest to those who want to discover the effects of rare variants on quantitative traits.

Lastly, I developed an accurate and automated machine learning classifier for the diagnosis of atopic dermatitis based on transcriptome and microbiota data. The main challenge is the small sample size of the dataset. To overcome this challenge, I designed an approach to select features with nested cross-validation and reduce overfitting. Our classifier enables the identification of atopic dermatitis with high accuracy and low cost. To the best of our knowledge, it is the first machine learning classifier for the diagnosis of atopic dermatitis solely based on transcriptome and microbiota data. I believe that this pipeline will be of significant interest to physicians. And it can help researchers who want to identify novel biomarkers or potential drug targets for atopic dermatitis as well as those who want to detect the risk of different diseases in using omics data in general.

In addition to the problems above, new machine learning and statistical approaches can address many other problems in genetics and genomics, such as variant calling, fine mapping, and imputation. There are several methods proposed for each of these problems and active research is in progress. I believe that more machine learning and statistical approaches with higher accuracy and power can improve the solutions to these problems and that the work presented in this dissertation will be useful for the scientific community to develop such approaches.