Published in final edited form as:

Curr Protoc Hum Genet. 2019 April ; 101(1): e83. doi:10.1002/cphg.83.

Methods for the analysis and interpretation for rare variants associated with complex traits

J. Dylan Weissenkampen[1], **Yu Jiang**[1], **Scott Eckert**[1], **Bibo Jiang**[1], **Bingshan Li**[2], and **Dajiang J. Liu**[1]

[1]Department of Public Health Sciences, Penn State College of Medicine, Hershey PA,

[2]Department of Molecular Physiology and Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN

Abstract

With the advent of Next Generation Sequencing (NGS) technologies, whole genome and whole exome DNA sequencing has become affordable for routine genetic studies. Coupled with improved genotyping arrays and genotype imputation methodologies, it is increasingly feasible to get rare genetic variant information in large datasets. Such datasets allow researchers to get a more complete understanding of the genetic architecture of complex traits due to rare variants. We review state-of-art statistical methods for the statistical genetics analysis of sequence-based association, including efficient algorithms for association analysis in biobank-scale datasets, gene-association tests, meta-analysis, fine mapping methods that integrate functional genomic dataset and phenome wide association studies (PheWAS). We expect that these methods will be highly useful for the next generation statistical genetics analysis in the era of precision medicine.

Keywords

Rare variant; GWAS; PheWAS; genome sequencing; genetic association; complex traits

INTRODUCTION

Over the last three decades, advances in DNA-sequencing have led to enormous progress in the field of statistical genetics and genomics. The cost effective sequencing of the human genome and exome (i.e. protein coding regions of the genome) has enabled large scale genetic association studies, which have identified many genetic variations associated with disease states, such as hypertension, heart attack, and early-onset Parkinson's disease (Barth & Tomaselli, 2016; Kathiresan et al., 2008; Klein & Westenberger, 2012), and other continuous traits such as height (Lango Allen et al., 2010) and lipid levels (D. J. Liu et al., 2017). Compared to common variants, rare genetic variants are more likely to be functional (Fu et al., 2013), and hence can more easily lead to novel biological and clinical insights. The identified genetic association has led to novel therapeutic targets such as the lipoprotein pathway for lipid levels (Cohen, Boerwinkle, Mosley, & Hobbs, 2006; Tg et al., 2014; J. Wu

Correspondences should be addressed to: Dajiang J. Liu (dxl46@psu.edu) or Bingshan Li (bingshan.li@vanderbilt.edu).

et al., 2007). Despite the paramount successes, for most complex diseases, the known mechanisms are rare, and a large portion of the heritability remain unexplained ("missing heritability"). More in-depth genetic association analysis and functional experiments are necessary to get a more complete understanding on the disease mechanisms.

There is great interest in the field to unveil the "missing heritability" that is attributable to rare variants with minor allelic frequencies (MAF) of less than 1% (Ladouceur, Dastani, Aulchenko, Greenwood, & Richards, 2012). The analysis of rare variants in large sample sizes has been enabled by cost effective sequencing and genotyping technologies and advanced genotype imputation methods. The cost of whole genome sequencing at 30× or higher has fallen below the $1,000 barrier, while genotyping genome-wide single nucleotide polymorphisms (SNPs) using arrays can be as low as ~$50. Coupled with advanced genotype imputation algorithms (Das et al., 2016) and high-quality haplotype reference panels (McCarthy et al., 2016), low frequency variants can be imputed with high accuracy. For example, with the haplotype reference consortium panel, the average imputation quality (as measured by R^2) for low frequency variants with MAF of 0.1% can be as high as 80% (McCarthy et al., 2016). It has now become feasible to collect comprehensive genetic information from cohorts of hundreds of thousands or even millions of individuals.

Even with the large sample sizes, the power for detecting associations with rare variants may still be limited, as each rare allele may still appear only a few times in a given dataset. Thus, it is important to develop sophisticated data analysis methods, to aggregate multiple signals in a gene region, prioritize likely causal variants over non-causal variants and enable the efficient analysis in large datasets (Lee, Abecasis, Boehnke, & Lin, 2014). In addition, as rare variants often arise more recently in history, and may be disproportionally stratified within certain populations (Mathieson & McVean, 2012), refined techniques are necessary to make sure that association analysis results are not spurious and influenced by the presence of population structure or cryptic relatedness.

Here we review contemporary methods and tools for rare variant association analysis. The goal of this review is to introduce techniques which efficiently perform the genetic association analysis of rare variants, and aid in the functional interpretation of rare variant association signals.

KEY CONCEPTS

Rare Variants

The definition of rare variant differs in various contexts. Typically, a rare variant is a genetic variant with minor allele frequency (MAF) <1%. The term "low frequency" variants is often used to refer to genetic variants with MAF between 1% and 5%.

Complex Disease

Complex disease is often defined in contrast to monogenic disorders, which are influenced by one or a few genes. Complex diseases may be influenced by the effects of multiple genes, often on the scale of hundreds or even thousands, in combination with lifestyles and

environmental factors. Most common diseases are complex, which may also be called multifactorial, or polygenic.

Biobanks

Biobanks and biorepositories contain phenotypic data, such as health records, lifestyle variables, and mental health questionnaire data, often alongside genetic data (based upon sequencing or genotyping) in a large sample of participants. Biobanks have been developed based upon hospital patients or from the general population. Biobanks have also been developed and curated for a variety of disorders including neuropsychiatric disorders and cancer. These databases can be utilized to detect biomarkers for traits of interest, to prioritize associated genetic variants, or to determine sub-phenotypes within a trait of interest.

General biobanks such as UK Biobank (Sudlow et al., 2015) ascertain behavioral and medical phenotypes and genetic data from the general population. In addition, many hospital-based biobanks have been developed by various countries and institutions, such as the Vanderbilt BioVU biobank (Roden et al., 2008), the Geisinger biobank (Carey et al., 2016), etc.

Biobanks have also been developed to study specific diseases. For example, a group at Johns Hopkins, The Bioinformoodics group, maintains several neuropsychiatric biobanks and databases (Pirooznia et al., 2014). They have made available to researchers aggregate findings from genetic studies of mood disorders, provides results from association, expression, and linkage studies for hypothesis formation, clinical phenotypes of 5,000 individuals used in genetics studies of bipolar disorder, and gene annotation and prioritization software based on neuronal signaling and synaptic pathways (Askland, Read, & Moore, 2009; Fromer et al., 2014; Fukata & Fukata, 2017; Grover et al., 2007; Henstridge, Pickett, & Spires-Jones, 2016; Pirooznia et al., 2012). The Early Detection Research Network by the National Cancer Institute provides information and data regarding potential biomarkers for many common types of cancer (Sokoll et al., 2010). This information is also made available to researchers to further complement their findings, or to provide researchers with ideas toward hypotheses to test in cancer research (Marks et al., 2015; Williams et al., 2012).

Fine Mapping

Fine mapping refers to the statistical approaches used to narrow down the list of causal variants from association analysis. Genome-wide association studies (GWAS) have been widely used to identify disease-associated variants and loci for traits of interest. However, these GWAS hits are often not causal (MacArthur et al., 2014). Some genetic variants in close proximity may return as the most significant hits, but often are not causal. This is due to variants in proximity to each other inheriting together, also known as linkage Disequilibrium (LD). This manifests as variants in LD with the causal variants may also have highly significant p-values. Through integrating functional genomic information, fine mapping methods can identify functional categories that are most relevant for the diseases of interest, and prioritize causal variants based upon both the strength of the association and the importance of the their functional annotation (Schaid, Chen, & Larson, 2018b).

STRATEGIC APPROACH

Genetic Association Test

The starting point for a statistical genetic analysis is to conduct an association analysis. To analyze low frequency or rare variants, we often perform a combination of single variant and gene-level association tests.

Single Variant Association Test

Statistical evaluation of the association between variants and a trait of interest has traditionally been straightforward. Regression models are often applied to analyze common variants by investigating variants individually and correcting for possible confounders in the data, including genetic principal components, age, sex, etc. (Dudbridge & Gusnanto, 2008). Many software packages can do single variant association analysis with ease, given the standard input of genotypes [e.g. in BGEN format (Band & Marchini, 2018) or in VCF format (Danecek et al., 2011)] and phenotypes [e.g. in PLINK PED (Purcell et al., 2007) file format].

Gene-level Association Test

For low frequency variants, single variant association analysis usually is underpowered. One alternative technique to analyze rare variants is to aggregate rare variants within individuals and compare whether rare variants in a functional unit (e.g. a gene) are associated with a trait of interest (B. Li, Liu, & Leal, 2013). Numerous methods and software packages have been developed to conduct gene-level association test. But most popular among these methods are the burden test (B. Li & Leal, 2008), variable threshold tests (Price et al., 2010), and sequence kernel association tests (SKAT) (Lin & Tang, 2011; M. C. Wu et al., 2011).

The burden test aggregates all rare variants with MAF less than a pre-defined threshold (e.g. 0.01) in a gene region, then tests for association between a phenotype and the total number of rare variants (B. Li & Leal, 2008; Madsen & Browning, 2009). This method implicitly assumes all variants in the same gene have the same direction of effect. When this assumption does not hold, the association signals of different variants may cancel out, and lead to considerable loss of power. This issue can be mitigated by the careful choice of potentially causal variants that will be included in the burden tests. One way to decide is to investigate the allele frequency. The frequency can be a first approximation for functionality, where lower frequency variants are more likely functional. In addition to allele frequencies, selecting variants that modify protein coding can improve the likelihood of selecting casual variants. For example, it has become a standard practice to analyze only nonsynonymous or loss-of-function variants in gene-level association test to reduce noise and potentially improve power. Variable threshold tests conduct burden tests under different MAF thresholds, correcting for the multiple comparison with a “minimal p-value” approach and can increase power over simple burden tests when causal variants are predominately rare. Alternatively, the inverse of the variant MAFs can be used as weights, in order to upweight potentially causal variant.

On the other hand, mixed effects models, such as those used by sequence kernel association test (SKAT) (M. C. Wu et al., 2011), investigate the distribution of the rare variants between cases and controls. SKAT improves power with variants of opposite effects, but it may be less powerful if the data has a large proportion of same-direction effects (B. Li et al., 2013).

To combine the strength of burden tests and SKAT, the SKAT-Optimal test (SKAT-O) (Lee et al., 2012), a hybrid burden and SKAT technique, was developed for situations in which both deleterious and protective variants are within the gene. SKAT-O can be robust against the presence of causal variants with opposite effects and does not lose much power compared to burden tests, when all causal variants in the gene region have unidirectional effects.

In practice, genetic studies often employ a combination of multiple rare variant association tests, such as simple burden tests or SKAT. As gene-level tests may well be driven by one causal variant, their reporting is often focused on the genes that are distant from significant single variant associations and driven by multiple rare variants that do not individually reach genome-wide significance.

As the field moves from whole exome sequencing to whole genome sequencing, gene-level association tests also evolve to region-based association tests, where the analysis unit needs to be redefined. Previously, a protein coding gene could be a natural analysis unit, but for whole genome sequencing, the most straightforward analysis unit may be a consecutive block of variants (e.g. a sliding window (Natarajan et al., 2018)). For whole genome analysis, the determination of causal variants can be more challenging, and new method development will be warranted to make progress (see the section below for more discussions).

Integrating Functional Genomic Data to Prioritize Rare Variants

A key step in genetic association analysis of rare variants is to distinguish causal variants from non-causal ones. The presence of non-causal variants can be highly detrimental to the power of detecting associations (B. Li & Leal, 2008; D. J. Liu & Leal, 2010). Through integrating functional information, higher weights can be assigned to likely causal variants, which can be an effective way to improve the power for gene-level association tests.

Despite the paramount importance of these issues, it is very challenging to determine whether a given variant is functional or causal. To achieve this, the most widely used approach uses variant allele frequency as a proxy. This is based on the idea that lower frequency variants may have a selective disadvantage and remain at low frequency due to purifying selection (Fu et al., 2013). Thus, using a weight that is inversely proportional to the variant allele frequency can be a useful approach to prioritize causal variants (Madsen & Browning, 2009).

Another appealing approach is to prioritize functional variants using integrative approaches based upon functional genomic data. Machine learning techniques allow for an array of techniques to be performed to classify variant functionality and the resulting variant functionality scores can be used to weigh association results to improve power. For example,

techniques such as CADD (Kircher et al., 2014), MetaLR (Dong et al., 2015), and REVEL (Ioannidis et al., 2016), DVAR (H. Yang et al., 2018), LINSIGHT (Huang, Gulko, & Siepel, 2017), EIGEN (Ionita-Laza, McCallum, Xu, & Buxbaum, 2016), GWAVA (Ritchie, Dunham, Zeggini, & Flicek, 2014), have been developed to rank whether a variant is likely to be deleterious or silent by utilizing an ensemble of techniques. The functionality score from these software programs can be used in conjunction with rare variant association tests to improve power. Among these variant functional annotations, some use supervised machine learning, such as CADD, and others use unsupervised methods, such as DVAR and EIGEN. Unlike coding variants, the function of the noncoding genome is largely unknown, and the available noncoding risk variants for complex diseases are often biased towards those with large effect sizes. This can make the training set unrepresentative of typical risk variants (those likely to have weak effects) for complex diseases. Accordingly, unsupervised approaches, such as DVAR and EIGEN, address this challenge without relying on pre-selected disease-causing noncoding risk variants, and are expected to be more effective in prioritizing weak effect variants for complex diseases.

Identifying causal variants in non-coding regions of the genome can be more difficult than in exonic regions. Missense mutations within the exome can alter protein functionality by prematurely terminating transcription or altering amino acid sequence. However, mutations within the non-coding regions, can have significantly more subtle effects by modifying promoter, enhancer, repressor, or binding regions for gene expression or modulating the actions of long non-coding RNA (Kumar et al., 2013; F. Zhang & Lupski, 2015). Although much progress has been made in the area, many regulatory domains remain unknown within the genome. Nonetheless, researchers have developed some successful methods for identifying potentially causal variants within non-coding regions of the DNA.

One major roadblock for using machine learning for rare variant analysis is that there is a much larger number of genetic variants in the non-coding regions (Schubach, Re, Robinson, & Valentini, 2017). This limits the size of training sets available for these techniques. Methods such as hyperSMURF (Schubach et al., 2017), take this into account by utilizing the synthetic minority oversampling technique (SMOTE) method to partition data from unbalanced data sets (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This method works by partitioning probable non-deleterious variants and simulating additional positive variants within similar genomic attributes, allowing for a balanced dataset between non-deleterious variants and potentially deleterious variants associated with a trait of interest (Schubach et al., 2017).

Machine learning techniques have started to be used for complex rare variant analysis, especially within the non-coding regions of the genome (Schubach et al., 2017). Future improvements in these methods may allow for complex decision tree mapping for powerful analyses of rare variant associations and integration of data information.

Genetic Association Test in Large Datasets

Modern genetic datasets have grown to an unprecedented scale, which often includes hundreds of thousands, or even millions of individuals. These datasets quickly render GWAS software packages obsolete that are not scalable up to biobank-scale data.

A new generation of software packages (Table 1) were developed to analyze biobank-scale data for genetic association analysis. The majority of the new software programs focused on a couple of aspects.

First, large biobanks often contain subtle population structures and cryptic relatedness. For example, about 1/3 of individuals in the UK Biobank dataset have 2nd degree (or closer) relatives in the dataset (Bycroft et al., 2017; P.-R. Loh, Kichaev, Gazal, Schoech, & Price, 2018). Linear mixed model-based methods are desired, as they can effectively control for population structure and cryptic relatedness without the need to explicitly model those confounders. In addition, linear mixed models can often improve power for the association analysis. Through modeling the polygenic component of diseases, the model reduces the variance of the association test statistics, and hence improve the non-centrality parameter and power. It has been suggested that linear mixed models are the desirable approach to analyze biobank-scale data and should be considered a default method of choice (J. Yang, Zaitlen, Goddard, Visscher, & Price, 2014).

The linear mixed model-based methods differ by the algorithms to fit the null models, including quadratic time and linear time algorithms.

Algorithms whose running time scales with the number of inputs squared (shown by the equation: $O(N^2)$ where N is the sample size), are referred to quadratic time complexity algorithms. For quadratic time algorithms, for example, the time to analyze 10,000 samples will be 100 times longer than analyzing 1,000 samples. The quadratic algorithm can scale well up to 50,000 individuals. Software packages implementing quadratic time algorithms include EMMAX (Kang et al., 2010), GEMMA (X. Zhou & Stephens, 2012), fastLMM (Lippert et al., 2011), RVTESTS (Zhan, Hu, Li, Abecasis, & Liu, 2016) etc. These methods first calculate the genetic relationship matrix, which models the correlation of the random effects in the linear mixed model. Then the methods factorize the kinship matrix, and rotate the original genotype and phenotype data, so that the calculation of association statistic has the same complexity as a linear regression model. For datasets with half a million or more individuals, storing and factorizing the kinship matrix is extremely memory and time intensive. These algorithms become infeasible for biobank-scale datasets.

More recently, linear time algorithms were adapted to fit linear mixed models. Instead of calculating and storing kinship matrix, they work by repeatedly loading chunks of the genotype matrix into the memory to save the memory for computing. To estimate the variance components parameters, they use either a Monte Carlo approach (e.g. in BOLT-LMM (P. R. Loh et al., 2015)) or method of moment approach (in RVTESTS (Zhan et al., 2016)) to replace maximum likelihood or restricted maximum likelihood. Such methods allow fitting linear mixed model to samples of half a million individuals, greatly facilitating the use of biobank-scale datasets for genetic discoveries. Another bottleneck lies in the evaluation of the variance of the score statistic. Exact calculation requires quadratic time complexity. Approximations were proposed to lower this time complexity to linear ($O(N^2)$), thereby significantly improving the speed of the algorithms (Svishcheva, Axenovich, Belonogova, van Duijn, & Aulchenko, 2012). Association statistics resulting from this approximation are almost identical to the exact statistics.

Gene-level association analysis techniques for biobank-scale datasets are also being developed. Extending the linear time algorithm for calculating single variant association statistics, researchers proposed efficient linear time algorithms for calculating covariance matrices between score statistics. Generally, the approximate covariance can be computed with time complexity $O(wp^2N)$, where p is the number of markers in a gene, N is the sample size, and w is the number of genes. This is a dramatic improvement over calculating the exact covariance which has a quadratic time complexity with respect to the sample size N given by $O(wpN^2)$. In simulations, they showed that the gene-level association tests with the approximate covariance matrix produced nearly identical results as the test that uses exact covariance matrices, but with a much lower computation time. RVTESTS (Zhan et al., 2016) implemented this approximation, substantially improving the computational efficiency.

Meta-Analysis

The identification of low frequency variants requires large sample sizes. One way to increase the sample size is to aggregate association statistics from multiple studies. This technique, known as meta-analysis, is often easier to implement as compared to pooling individual-level data (mega-analysis), more protective to study participant privacy, and more robust against potential heterogeneities between studies (D. J. Liu et al., 2014).

Meta-analysis is a well-established method, which is broadly applied in statistical genetics with user friendly software packages, such as METAL (Willer, Li, & Abecasis, 2010) or META (J. Z. Liu et al., 2010). Meta-analysis methods can be broadly classified into fixed effects and random effects methods. Fixed effect methods assume that the genetic effects are equal across studies. Standard approaches for fixed effect meta-analysis include inverse variance weighted meta-analysis and weighted Z-score statistic method. When the fixed effect is true and the genetic effect for a variant is constant across studies, inverse-variance weighted meta-analysis is provably optimal, maximizing the non-centrality parameter for the chi-square statistic under the alternative hypothesis. One potential limitation in the use of inverse-variance weighted meta-analysis is that included studies must measure the phenotype in the same unit (Willer et al., 2010). For example, lipid levels in some studies are measured in milligrams per deciliter (mg/dL) for some studies and in millimoles per liter (mmol/L) for others. It is essential that the phenotype measurements in different studies are harmonized before applying inverse variance weighted meta-analysis. The results can be very different if this prerequisite is not satisfied. On the other hand, the method that weighs the Z-score statistic by the square-root of the sample size is more robust against the potential measurement heterogeneities. The weighted Z-score method is equivalent to the inverse variance weighted score meta-analysis if the variance of the genetic effect estimates is proportional to the inverse of the sample size – a scenario that holds when the allele frequencies between participating studies are similar.

Currently, one appealing approach is to leverage the latest and largest reference panels to re-impute participating studies and re-conduct genetic meta-analysis, to leverage the better imputed low frequency variants and identify novel associations. For example, a recently conducted meta-analysis using haplotype reference consortium panels imputed genotypes to study the genetics of smoking and drinking addictions (Liu M, 2018). It was noted that a

compromise between weighted Z-score statistic and inverse-variance weighted meta-analysis achieves the optimal balance between robustness and statistical optimality (Liu M, 2018). Specifically, assume that the z-score statistic from K participating studies are given by $Z_k, k=1 \ldots K$, the imputation quality for the K studies (R) are $R_1,\ldots,R_K$ and the allele frequencies are $P_1,\ldots,P_K$. We defined the weight for study k as $w_k = N_k R_k^2 p_k(1-p_k)$. The meta-analysis statistic is given by

$$Z_{meta} = \frac{\sum_k w_k Z_k}{\left(\sum_k w_k^2\right)^{1/2}}$$

This new statistic is imputation quality aware (Zaitlen & Eskin, 2010), but can also accommodate potential differences in the variant allele frequencies between studies. This approach gives the best results among all fixed effects meta-analysis methods in the meta-analyses that we conducted.

In the statistical genetics field, several methods have been developed recently to conduct meta-analysis for rare variant association tests, in particular the gene-level association tests, such as the burden tests, (optimal) sequence kernel association tests, and variable threshold tests. These methods are based upon a similar principle that constructs gene-level association tests from single variant association statistics, and their variance-covariance matrix between them (Lee et al., 2014; Lee, Teslovich, Boehnke, & Lin, 2013; D. J. Liu et al., 2014; Tang & Lin, 2013, 2014, 2015). These methods also differ by the fixed and random effects assumptions. The most widely used approach is fixed effect methods, but random effects meta-analysis methods are also available and implemented.

As genetic research is shifting to include non-European populations, differing allelic frequencies of variants and different mutations may lead to the identification of novel associations. A few reasons may lead to the genetic effect heterogeneity between studies. First, the SNP used in the association analysis may not be the causal variant. Due to the potential linkage disequilibrium (LD) differences across populations, the SNP may have differential LD with the causal variant in different studies, which lead to difference in the measured genetic effect. Second, due to potential gene-by-environment interactions, the marginal effect may differ due to different environmental exposures. Third, if the causal variant is rare, it is likely population specific, which can also lead to the heterogeneity in the genetic effects. Random effects assumption in a trans-ethnic meta-analysis can be very useful. There has been growing emphasis on extending the genetic studies to non-European populations. For example, the Trans-Omics Precision Medicine Sequencing program aims at sequencing >100,000 individuals from diverse human populations (Group, 2015). Such projects, when completed, would allow for the analysis of diverse populations to evaluate reproducibility of results.

Trans-ethnic meta-analysis methods were developed to jointly analyze samples from distinct ancestries. The underlying assumption for these methods is that for genetically closely

related populations, genetic effects are similar and fixed effect meta-analysis can be used to group genetically similar populations. For cohorts that are genetically dissimilar, the genetic effects can be different and random effect meta-analysis can be used to group these dissimilar populations. For single variant analysis, MANTRA (Morris, 2011) and Mr. MEGA (Magi et al., 2017) have been state-of-the-art approaches and widely applied to study a variety of complex human diseases including type II diabetes (Mahajan et al., 2014). More recently, the trans-ethnic meta-analysis has been extended to rare variant gene-level association tests, which combines the use of fixed and random-effect meta-analyses methods. Several methods were available with companion software packages, including MASS (Tang & Lin, 2013) and metaSKAT (Lee et al., 2013).

Joint Analysis of Multiple Phenotypes

There is increasing interest in the field to extend beyond single variant and single trait analysis and jointly study multiple traits. For traits with shared genetic basis, jointly analyzing multiple phenotypes may increase the statistical power for detecting associations.

The multi-trait analysis of GWAS summary statistics (MTAG) method was designed to improve the amount of variance explained as compared to the traditional GWAS statistics (Turley et al., 2018). This technique analyzes multiple summary statistics from GWAS, thus analyzing potentially overlapping samples (Turley et al., 2018). When utilizing multiple GWAS datasets, this method can identify more associated loci than the individual tests themselves. Additionally, numerous methods have been developed for jointly testing multiple phenotypes and perform an omnibus test. Beside omnibus tests, Bayesian methods have been developed to perform model comparison and dissect subset of associated phenotypes in addition to perform omnibus tests (Stephens, 2013).

Phenome-Wide Association Analysis

Another new development in rare variant analysis is the phenome wide association study (PheWAS). Traditionally, PheWAS was developed to follow up the phenotypic consequences of a subset of genetic variants on a variety of phenotypes, e.g. the phenotypes derived from electronic medical record (Denny et al., 2010). However, when the cost of sequencing and genotyping continues to decrease, candidate gene studies have been replaced by genome-wide association studies. PheWAS has started to systematically analyze the associations between all genetic variants and all phenotypes.

PheWAS can be used in any phenotype-rich datasets such as UK Biobank (Sudlow et al., 2015), or BioVU (Roden et al., 2008). Many of these analyses use medical codes such as from the International Classification of Diseases 9th Revision (ICD9) as phenotypic traits for the analysis (Fritsche et al., 2018). Such techniques allow for the evaluation of variant-associated phenomes such that DNA variants can be tested for association across medical phenotypes (Cortes et al., 2017).

As PheWAS phenotypes are often derived from electronic medical records, power can be improved by incorporating the hierarchal structures of ICD9 or ICD10 codes into the analyses. A new method along this direction is TreeWAS(Cortes et al., 2017). This technique utilizes the clustering of phenotypes by related ICD medical codes to employ a hierarchal

"tree" of phenotypes for association with variants (Cortes et al., 2017). This technique allows for a more powerful association analysis with EMR-based biobanks compared to standard methods that ignore the tree structure in the billing codes.

Approaches have been developed to improve the efficiency of PheWAS computation within large-scale databases, such as biobanks. Biobanks often contain extremely unbalanced number of cases and controls, with a far greater number of controls compared to cases. Such unbalanced samples can lead to inflated type I errors when the association statistics are evaluated using a normal distribution. On the other hand, the calculation of p-values can be improved via the use of saddle point approximation. The saddle point approximation is much faster than the approach based upon Firth correction but can achieve comparable performance for controlling the type I error for unbalanced case control studies. Software packages such as SGA(Dey, Schmidt, Abecasis, & Lee, 2017) and SAIGE(W. Zhou et al., 2018) implemented these methods for PheWAS based upon EMR data.

Fine Mapping Causal Variants

One immediate step following association analysis is to identify causal variant using fine mapping. Fine mapping is an active area of research, where numerous methods have been developed and utilized. Most of the approaches are based on a similar Bayesian framework, where genetic variants in the same annotation category are assumed to have a similar genetic effect distribution and similar likelihood of being causal. When jointly modeled with the association statistics (such as the genetic effect size estimate, their standard deviation, etc.), functional categories that are enriched with statistically significant associations will be identified. Often posterior probability of association is calculated, which can be used to prioritize causal variants over non-causal variants.

The statistical methods differ by the assumptions that they made. Given the large number of genetic effects in a locus, earlier methods, such as fgwas (Pickrell, 2014) and PAINTOR (Kichaev et al., 2014) often assume that each locus contains only one or a few causal variants. This assumption can be overly restrictive for some loci with extensive allelic heterogeneity, but allows the enumeration of all possible configurations of causal variants in the locus and the calculation of the exact likelihood. Other methods, including RiVIERA (Yue Li, Davila-Velderrain, & Kellis, 2017; Y. Li & Kellis, 2016) and FINEMAP (Benner et al., 2016), instead make use of Markov Chain Monte Carlo methods to approximate the exact likelihood, in order to make the computation feasible.

More recently, the development of fine mapping methods has been extended to combine samples from multiple ancestries. Due to the differential LD patterns between studies, the trans-ethnic fine mapping methods, such trans-ethnic PAINTOR (Kichaev & Pasaniuc, 2015), have the potential to further narrow down the list of causal variants. In addition, methods have been developed to jointly fine map multiple correlated phenotypes in order to further improve the resolution (H. Huang et al., 2017). As sequence data becomes available for large multi-ethnic datasets, efforts have begun to integrate sequence and imputation-based GWAS data for fine mapping. As sequence data and GWAS imputation-based data measure slightly different variant sites, specialized statistical methods were developed to perform joint analysis in the presence of missing data (Jiang et al., 2018). More detailed

information on fine mapping was discussed in the comprehensive review by Schaid et al (Schaid, Chen, & Larson, 2018a).

COMMENTARY

Application

The techniques discussed above are important tools which can assist researchers seeking to analyze the association of rare genetic variants on phenotypes. When looking in the literature, it is fairly common for researchers to utilize several of these techniques when performing their research. Here, two examples are given in which multiple techniques described above are utilized to research phenotypes of interest.

Researchers recently investigated genetic variants associated with smoking and drinking traits, looking to identify and replicate rare variant associations (Brazel et al., 2018). They first aggregated summary statistics from 16 individual studies and combined the data with UK Biobank data for a meta-analysis, using linear mixed models to properly account for relatedness (Brazel et al., 2018). Next, researchers conducted a genome-wide association meta-analysis to identify loci in the genome significantly associated with the phenotypes. Fine mapping was next performed to identify potentially causal variants within each loci using fgwas (Pickrell, 2014) integrating functional data and a Bayesian method (Mahajan et al., 2018) that is based upon association strength only. Additionally, gene-level association tests were conducted using SKAT (M. C. Wu et al., 2011) and simple burden tests using rare variants (Brazel et al., 2018), grouping variants with MAF<1%. These techniques allowed the researchers to identify rare variants which accounted for between 11%−18% of the SNP heritability of these phenotypes (Brazel et al., 2018).

Another research study investigating smoking phenotypes aimed to identify loci that associate with one or more smoking or drinking phenotype and to implicate gene pathways in conferring risk for these phenotypes (Liu M, 2018). They utilized data from a variety of cohorts that were imputed using the Haplotype Consortium Reference Panel. Despite the improved accuracy of imputing rare variants, there can be considerable heterogeneity in the imputation R^2 values between studies, so the researchers adopted an imputation-aware meta-analysis. GWAS summary statistics were obtained using RVTESTS (Zhan et al., 2016), where a linear mixed model was used to analyze the associations controlling for relatedness and genetic architecture. The meta-analysis was performed using rareGWAMA (D. J. Liu et al., 2014), and MTAG (Turley et al., 2018) was used to increase the power for locus discovery. These researchers were able to identify 406 loci associated with multiple stages of smoking, and determined neurotransmitter pathways that may affect susceptibility to smoking (Liu M, 2018).

Future Development

Many techniques have been created within the last decade to analyze genotype association data. Several of these techniques utilize an integration of other types of data in concert with genetic data, allowing for a more powerful analysis of associations between genetics and traits of interest. This enables the identification of rare variants of interest for disease states.

Statistical association analyses will often be followed-up by fine mapping to determine if the associated variant(s) is causal (using either an *in silico* approach or an experimental approach), and if so, through which mechanisms it affects the trait of interest. Cell and animal studies involving knockdown, knockout, silencing, and overexpression of genetic variants can generate data on the effects of variants on cell viability, pathway functionality, or phenotype development.

For the field at large, data continues to be gathered and aggregated into larger and better-annotated databases. This further collection of RNA, proteomic, and medical data in tandem with larger cohorts will greatly increase the power of association tests, refine the set of potential causal variants, and aid in the development of polygenic risk scores. This increase in data available within databases and biobanks will need to be thoroughly vetted through quality control procedures, however, to ensure accuracy of the results within. With increased power comes increased data volume and heightened computational demands for such analysis, generating a need for more sophisticated analytical methods. The need for more processing power may be at least partially overcome by cloud computing methods, which allows researchers to reserve groups of computers for only the time and processing power they need (Langmead & Nellore, 2018). Programs like SEQSpark (D. Zhang et al., 2017) can allow for the analysis of rare variant associations in whole-genome data quickly and efficiently in a highly parallel fashion. This allows for far quicker analysis of large genomic datasets (D. Zhang et al., 2017).

Concluding remarks

Rare variant analysis is of particular interest currently in statistical genetics research, despite the limitations imposed by their low allelic frequency. Increased sensitivity offered by modern statistical analysis methods, large databanks, and heightened computational resources have enabled more research into this field. Studies on such variants have already led to key findings in personalized medicine (Birdwell et al., 2012; MacGregor et al., 2018; Xu et al., 2011) and they hold great promise to further elucidate disease mechanisms. It is imperative that statistical methods are selected carefully to analyze a given dataset to minimize false positive discoveries. We expect that our results will provide a useful resource for the design, analysis, and interpretation of next generation genetic studies.

ACKNOWLEDGMENTS

This research is supported partially by National Institutes of Health grants U01HG009086 (BL) and R01GM126479 and R01HG008983 (DL).

References

Askland K, Read C, & Moore J (2009). Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet, 125(1), 63–79. doi:10.1007/s00439-008-0600-y [PubMed: 19052778]

Band G, & Marchini J (2018). BGEN: a binary file format for imputed genotype and haplotype data. bioRxiv.

Barth AS, & Tomaselli GF (2016). Gene scanning and heart attack risk. Trends Cardiovasc Med, 26(3), 260–265. doi:10.1016/j.tcm.2015.07.003 [PubMed: 26277204]

Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, & Pirinen M (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics, 32(10), 1493–1501. doi:10.1093/bioinformatics/btw018 [PubMed: 26773131]

Birdwell KA, Grady B, Choi L, Xu H, Bian A, Denny JC, Haas DW (2012). The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. Pharmacogenet Genomics, 22(1), 32–42. doi:10.1097/FPC.0b013e32834e1641 [PubMed: 22108237]

Brazel DM, Jiang Y, Hughey JM, Turcot V, Zhan X, Gong J, … Vrieze S (2018). Exome chip meta-analysis fine maps causal variants and elucidates the genetic architecture of rare coding variants in smoking and alcohol use. Biological Psychiatry. doi:10.1016/j.biopsych.2018.11.024

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, … Marchini J (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv.

Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, … Ledbetter DH (2016). The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet Med, 18(9), 906–913. doi:10.1038/gim.2015.187 [PubMed: 26866580]

Chawla NV, Bowyer KW, Hall LO, & Kegelmeyer WP (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Cohen JC, Boerwinkle E, Mosley TH Jr., & Hobbs HH (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med, 354(12), 1264–1272. doi:10.1056/NEJMoa054013 [PubMed: 16554528]

Cortes A, Dendrou CA, Motyer A, Jostins L, Vukcevic D, Dilthey A, McVean G (2017). Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat Genet, 49(9), 1311–1318. doi:10.1038/ng.3926 [PubMed: 28759005]

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, … Genomes Project Analysis, G. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. doi:10.1093/bioinformatics/btr330 [PubMed: 21653522]

Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, … Fuchsberger C (2016). Next-generation genotype imputation service and methods. Nat Genet, 48(10), 1284–1287. doi:10.1038/ng.3656 [PubMed: 27571263]

Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Crawford DC (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics, 26(9), 1205–1210. doi:10.1093/bioinformatics/btq126 [PubMed: 20335276]

Dey R, Schmidt EM, Abecasis GR, & Lee S (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. bioRxiv.

Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, & Liu X (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet, 24(8), 2125–2137. doi:10.1093/hmg/ddu733 [PubMed: 25552646]

Dudbridge F, & Gusnanto A (2008). Estimation of significance thresholds for genomewide association scans. Genet Epidemiol, 32(3), 227–234. doi:10.1002/gepi.20297 [PubMed: 18300295]

Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, Mukherjee B (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet, 102(6), 1048–1061. doi:10.1016/j.ajhg.2018.04.001 [PubMed: 29779563]

Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, O'Donovan MC (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature, 506(7487), 179–184. doi:10.1038/nature12929 [PubMed: 24463507]

Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Akey JM (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature, 493(7431), 216–220. doi:http://www.nature.com/nature/journal/v493/n7431/abs/nature11690.html#supplementary-information [PubMed: 23201682]

Fukata Y, & Fukata M (2017). Epilepsy and synaptic proteins. Curr Opin Neurobiol, 45, 1–8. doi:10.1016/j.conb.2017.02.001 [PubMed: 28219682]

Group PMIW (2015). The Precision Medicine Initiative Cohort Program: Building a Research Foundation for 21st Century Medicine. Retrieved from https://acd.od.nih.gov/documents/reports/DRAFT-PMI-WG-Report-9-11-2015-508.pdf

Grover D, Woodfield AS, Verma R, Zandi PP, Levinson DF, & Potash JB (2007). QuickSNP: an automated web server for selection of tagSNPs. Nucleic Acids Res, 35(Web Server issue), W115–120. doi:10.1093/nar/gkm329 [PubMed: 17517769]

Henstridge CM, Pickett E, & Spires-Jones TL (2016). Synaptic pathology: A shared mechanism in neurological disease. Ageing Res Rev, 28, 72–84. doi:10.1016/j.arr.2016.04.005 [PubMed: 27108053]

Huang H, Fang M, Jostins L, Umicevic Mirkov M., Boucher G, Anderson CA, Barrett JC (2017). Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature, 547(7662), 173–178. doi:10.1038/nature22969 [PubMed: 28658209]

Huang YF, Gulko B, & Siepel A (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet, 49(4), 618–624. doi:10.1038/ng.3810 [PubMed: 28288115]

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Sieh W (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet, 99(4), 877–885. doi:10.1016/j.ajhg.2016.08.016 [PubMed: 27666373]

Ionita-Laza I, McCallum K, Xu B, & Buxbaum JD (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet, 48(2), 214–220. doi:10.1038/ng.3477 [PubMed: 26727659]

Jiang Y, Chen S, McGuire D, Chen F, Liu M, Iacono WG, Liu DJ (2018). Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. PLoS Genet, 14(7), e1007452. doi:10.1371/journal.pgen.1007452 [PubMed: 30016313]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Eskin E (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet, 42(4), 348–354. doi:10.1038/ng.548 [PubMed: 20208533]

Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Orho-Melander M (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet, 40(2), 189–197. doi:10.1038/ng.75 [PubMed: 18193044]

Kichaev G, & Pasaniuc B (2015). Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. Am J Hum Genet, 97(2), 260–271. doi:10.1016/j.ajhg.2015.06.007 [PubMed: 26189819]

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Pasaniuc B (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet, 10(10), e1004722. doi:10.1371/journal.pgen.1004722 [PubMed: 25357204]

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, & Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet, 46(3), 310–315. doi:10.1038/ng.2892 [PubMed: 24487276]

Klein C, & Westenberger A (2012). Genetics of Parkinson's disease. Cold Spring Harb Perspect Med, 2(1), a008888. doi:10.1101/cshperspect.a008888 [PubMed: 22315721]

Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, Wijmenga C (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet, 9(1), e1003201. doi:10.1371/journal.pgen.1003201 [PubMed: 23341781]

Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, & Richards JB (2012). The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet, 8(2), e1002496. doi:10.1371/journal.pgen.1002496 [PubMed: 22319458]

Langmead B, & Nellore A (2018). Cloud computing for genomic data analysis and collaboration. Nat Rev Genet, 19(5), 325. doi:10.1038/nrg.2018.8

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Hirschhorn JN (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature, 467(7317), 832–838. doi:10.1038/nature09410 [PubMed: 20881960]

Lee S, Abecasis GR, Boehnke M, & Lin X (2014). Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet, 95(1), 5–23. doi:10.1016/j.ajhg.2014.06.009 [PubMed: 24995866]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Lin X (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet, 91(2), 224–237. doi:10.1016/j.ajhg.2012.06.007 [PubMed: 22863193]

Lee S, Teslovich TM, Boehnke M, & Lin X (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet, 93(1), 42–53. doi:10.1016/j.ajhg.2013.05.010 [PubMed: 23768515]

Li B, & Leal SM (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet, 83(3), 311–321. doi:10.1016/j.ajhg.2008.06.024 [PubMed: 18691683]

Li B, Liu DJ, & Leal SM (2013). Identifying rare variants associated with complex traits via sequencing. Curr Protoc Hum Genet, Chapter 1, Unit 1.26. doi:10.1002/0471142905.hg0126s78

Li Y, Davila-Velderrain J, & Kellis M (2017). A probabilistic framework to dissect functional cell-type-specific regulatory elements and risk loci underlying the genetics of complex traits. bioRxiv.

Li Y, & Kellis M (2016). Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. Nucleic Acids Res, 44(18), e144. doi:10.1093/nar/gkw627 [PubMed: 27407109]

Lin DY, & Tang ZZ (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet, 89(3), 354–367. doi:10.1016/j.ajhg.2011.07.015. 10.1016/j.ajhg.2011.07.015 [PubMed: 21885029]

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, & Heckerman D (2011). FaST linear mixed models for genome-wide association studies. Nat Methods, 8(10), 833–835. doi:10.1038/nmeth.1681 [PubMed: 21892150]

Liu DJ, & Leal SM (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet, 6(10), e1001156. doi:10.1371/journal.pgen.1001156 [PubMed: 20976247]

Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, Kathiresan S (2017). Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet, 49(12), 1758–1766. doi:10.1038/ng.3977 [PubMed: 29083408]

Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, Abecasis GR (2014). Meta-analysis of gene-level tests for rare variant association. Nat Genet, 46(2), 200–204. doi:10.1038/ng.2852 [PubMed: 24336170]

Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Marchini J (2010). Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet, 42(5), 436–440. doi:10.1038/ng.572 [PubMed: 20418889]

Liu MJY, Wedow R, Li Y, Brazel DM, Chen F^, Datta G, Davila-Velderrain J, McGuire D, Tian C, Zhan X, Choquet H, Docherty AR, Faul JD, Foerster JR, Gabrielsen ME, Gordon SD, Haessler J, Hottenga J, Huang H, Jansen PR, Ling YY, Mägi R, Matoba N, McMahon G, Mulas A, Orrù V, Palviainen T, Pandit A, Reginsson GW, Skogholt AH, Smith JA, Taylor AE, Turman C, Willemsen G, Young H, Young KA, Zajac GJM, Zhao W, Zhou W, Bjornsdottir G, Boardman JD, Boehnke M, Boomsma DI, Chen C, Cucca F, Davies GE, Eaton CB, Ehringer MA, Esko T, Fiorillo E, Gillespie NA, Gudbjartsson DF, Haller T, Harris KM, Heath AC, Hewitt JK, Hickie IB, Hokanson JE, Hopfer CJ, Hunter DJ, Iacono WG, Johnson EO, Kamatani Y, Kardia SLR, Keller MC, Kellis M, Kooperberg C, Kraft P, Krauter KS, Laakso M, Lind PA, Loukola A, Lutz SM, Madden PAF, Martin NG, McGue M, McQueen MB, Medland SE, Metspalu A, Mohlke KL, Nielsen JB, Okada Y, Peters U, Polderman TJC, Posthuma D, Reiner AP, Rice JP, Rimm E, Rose RJ, Runarsdottir V, Stallings MC, Stančáková A, Stefansson H, Thai KK, Tindle HA, Tyrfingsson T, Wall TL, Weir D, Weisner C, Whitfield JB, Winsvold BS, Yin J, Zuccolo L, Bierut LJ, Hveem K, Lee JJ, Munafo MR,NA, Saccone NL, Willer CJ, Cornelis MC, David SP, Hinds D, Jorgenson E, Kaprio J, Stitzel JA, Stefansson K, Thorgeirsson TE, Abecasis G, Liu DJ, Vrieze S (2018). Association studies of

1.2 million individuals yields new insights in the genetic etiology of tobacco and alcohol use. Nat Genet, In press.

Loh P-R, Kichaev G, Gazal S, Schoech AP, & Price AL (2018). Mixed model association for biobank-scale data sets. bioRxiv.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Price AL (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet, 47(3), 284–290. doi:10.1038/ng.3190 [PubMed: 25642633]

MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Gunter C (2014). Guidelines for investigating causality of sequence variants in human disease. Nature, 508(7497), 469–476. doi:10.1038/nature13127 [PubMed: 24759409]

MacGregor S, Ong JS, An J, Han X, Zhou T, Siggs OM, Hewitt AW (2018). Genome-wide association study of intraocular pressure uncovers new pathways to glaucoma. Nat Genet, 50(8), 1067–1071. doi:10.1038/s41588-018-0176-y [PubMed: 30054594]

Madsen BE, & Browning SR (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet, 5(2), e1000384. doi:10.1371/journal.pgen.1000384 [PubMed: 19214210]

Magi R, Horikoshi M, Sofer T, Mahajan A, Kitajima H, Franceschini N, Morris AP (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. Hum Mol Genet, 26(18), 3639–3650. doi:10.1093/hmg/ddx280 [PubMed: 28911207]

Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, Consortium, T. D. G. E. b. N.-g. s. i. m.-E. S. T. D.-G. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet, 46(3), 234–244. doi:10.1038/ng.2897 [PubMed: 24509480]

Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, McCarthy MI (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet, 50(11), 1505–1513. doi:10.1038/s41588-018-0241-6 [PubMed: 30297969]

Marks JR, Anderson KS, Engstrom P, Godwin AK, Esserman LJ, Longton G, Pepe MS (2015). Construction and analysis of the NCI-EDRN breast cancer reference set for circulating markers of disease. Cancer Epidemiol Biomarkers Prev, 24(2), 435–441. doi:10.1158/1055-9965.EPI-14-1178 [PubMed: 25471344]

Mathieson I, & McVean G (2012). Differential confounding of rare and common variants in spatially structured populations. Nat Genet, 44(3), 243–246. doi:10.1038/ng.1074 [PubMed: 22306651]

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Haplotype Reference, C. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet, 48(10), 1279–1283. doi:10.1038/ng.3643 [PubMed: 27548312]

Morris AP (2011). Transethnic meta-analysis of genomewide association studies. Genet Epidemiol, 35(8), 809–822. doi:10.1002/gepi.20630 [PubMed: 22125221]

Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, Group NTLW (2018). Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat Commun, 9(1), 3391. doi:10.1038/s41467-018-05747-8 [PubMed: 30140000]

Pickrell JK (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet, 94(4), 559–573. doi:10.1016/j.ajhg.2014.03.004 [PubMed: 24702953]

Pirooznia M, Seifuddin F, Judy J, Goes FS, Potash JB, & Zandi PP (2014). Metamoodics: meta-analysis and bioinformatics resource for mood disorders. Mol Psychiatry, 19(7), 748–749. doi:10.1038/mp.2013.118 [PubMed: 24018898]

Pirooznia M, Wang T, Avramopoulos D, Valle D, Thomas G, Huganir RL, Zandi PP (2012). SynaptomeDB: an ontology-based knowledgebase for synaptic genes. Bioinformatics, 28(6), 897–899. doi:10.1093/bioinformatics/bts040 [PubMed: 22285564]

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, & Sunyaev SR (2010). Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet, 86(6), 832–838. doi:S0002–9297(10)00207–7 [pii]10.1016/j.ajhg.2010.04.005 [PubMed: 20471002]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Sham PC (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81(3), 559–575. doi:10.1086/519795 [PubMed: 17701901]

Ritchie GR, Dunham I, Zeggini E, & Flicek P (2014). Functional annotation of noncoding sequence variants. Nat Methods, 11(3), 294–296. doi:10.1038/nmeth.2832 [PubMed: 24487584]

Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, & Masys DR (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther, 84(3), 362–369. doi:10.1038/clpt.2008.89 [PubMed: 18500243]

Schaid DJ, Chen W, & Larson NB (2018a). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. doi:10.1038/s41576-018-0016-z

Schaid DJ, Chen W, & Larson NB (2018b). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet, 19(8), 491–504. doi:10.1038/s41576-018-0016-z [PubMed: 29844615]

Schubach M, Re M, Robinson PN, & Valentini G (2017). Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. Sci Rep, 7(1), 2959. doi:10.1038/s41598-017-03011-5 [PubMed: 28592878]

Sokoll LJ, Sanda MG, Feng Z, Kagan J, Mizrahi IA, Broyles DL, Chan DW (2010). A prospective, multicenter, National Cancer Institute Early Detection Research Network study of [−2]proPSA: improving prostate cancer detection and correlating with cancer aggressiveness. Cancer Epidemiol Biomarkers Prev, 19(5), 1193–1200. doi:10.1158/1055-9965.EPI-10-0007 [PubMed: 20447916]

Stephens M (2013). A unified framework for association analysis with multiple related phenotypes. PLoS ONE, 8(7), e65245. doi:10.1371/journal.pone.0065245 [PubMed: 23861737]

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Collins R (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med, 12(3), e1001779. doi:10.1371/journal.pmed.1001779 [PubMed: 25826379]

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, & Aulchenko YS (2012). Rapid variance components-based method for whole-genome association analysis. Nat Genet, 44(10), 1166–1170. doi:10.1038/ng.2410 [PubMed: 22983301]

Tang ZZ, & Lin DY (2013). MASS: meta-analysis of score statistics for sequencing studies. Bioinformatics, 29(14), 1803–1805. doi:10.1093/bioinformatics/btt280 [PubMed: 23698861]

Tang ZZ, & Lin DY (2014). Meta-analysis of sequencing studies with heterogeneous genetic associations. Genet Epidemiol, 38(5), 389–401. doi:10.1002/gepi.21798 [PubMed: 24799183]

Tang ZZ, & Lin DY (2015). Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. Am J Hum Genet, 97(1), 35–53. doi:10.1016/j.ajhg.2015.05.001 [PubMed: 26094574]

Tg, Hdl Working Group of the Exome Sequencing Project, N. H. L., Blood I, Crosby J., Peloso GM, Auer PL, Kathiresan S (2014). Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N Engl J Med, 371(1), 22–31. doi:10.1056/NEJMoa1307095 [PubMed: 24941081]

Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Consortium SSGA (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet, 50(2), 229–237. doi:10.1038/s41588-017-0009-4 [PubMed: 29292387]

Willer CJ, Li Y, & Abecasis GR (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 26(17), 2190–2191. doi:10.1093/bioinformatics/btq340 [PubMed: 20616382]

Williams SB, Salami S, Regan MM, Ankerst DP, Wei JT, Rubin MA, Sanda MG (2012). Selective detection of histologically aggressive prostate cancer: an Early Detection Research Network Prediction model to reduce unnecessary prostate biopsies with validation in the Prostate Cancer Prevention Trial. Cancer, 118(10), 2651–2658. doi:10.1002/cncr.26396 [PubMed: 22006057]

Wu J, Province MA, Coon H, Hunt SC, Eckfeldt JH, Arnett DK, Kraja AT (2007). An investigation of the effects of lipid-lowering medications: genome-wide linkage analysis of lipids in the HyperGEN study. BMC Genet, 8, 60. doi:10.1186/1471-2156-8-60 [PubMed: 17845730]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, & Lin X (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet, 89(1), 82–93. doi:S0002–9297(11)00222–9 [pii]10.1016/j.ajhg.2011.05.029 [PubMed: 21737059]

Xu H, Jiang M, Oetjens M, Bowton EA, Ramirez AH, Jeff JM, Denny JC (2011). Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. J Am Med Inform Assoc, 18(4), 387–391. doi:10.1136/amiajnl-2011-000208 [PubMed: 21672908]

Yang H, Chen R, Wang Q, Wei Q, Ji Y, Zheng G, Li B (2018). De Novo pattern discovery enables robust assessment of functional consequences of noncoding variants. Bioinformatics. doi:10.1093/bioinformatics/bty826

Yang J, Zaitlen NA, Goddard ME, Visscher PM, & Price AL (2014). Advantages and pitfalls in the application of mixed-model association methods. Nat Genet, 46(2), 100–106. doi:10.1038/ng.2876 [PubMed: 24473328]

Zaitlen N, & Eskin E (2010). Imputation aware meta-analysis of genome-wide association studies. Genet Epidemiol, 34(6), 537–542. doi:10.1002/gepi.20507 [PubMed: 20717975]

Zhan X, Hu Y, Li B, Abecasis GR, & Liu DJ (2016). RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics, 32(9), 1423–1426. doi:10.1093/bioinformatics/btw079 [PubMed: 27153000]

Zhang D, Zhao L, Li B, He Z, Wang GT, Liu DJ, & Leal SM (2017). SEQSpark: A Complete Analysis Tool for Large-Scale Rare Variant Association Studies Using Whole-Genome and Exome Sequence Data. Am J Hum Genet, 101(1), 115–122. doi:10.1016/j.ajhg.2017.05.017 [PubMed: 28669402]

Zhang F, & Lupski JR (2015). Non-coding genetic variants in human disease. Hum Mol Genet, 24(R1), R102–110. doi:10.1093/hmg/ddv259 [PubMed: 26152199]

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, Lee S (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet, 50(9), 1335–1341. doi:10.1038/s41588-018-0184-y [PubMed: 30104761]

Zhou X, & Stephens M (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet, 44(7), 821–824. doi:10.1038/ng.2310 [PubMed: 22706312]

Table 1:

Software packages for Biobank-Scale Association Analysis. We compared the features of several widely used software packages that can analyze biobank scale datasets for associations. We considered the input file format supported, the type of phenotypes (binary or quantitative) they can analyze, the gene-level tests they can perform, as well as whether they can handle samples with relatedness.

Software for Biobank-Scale Data Analysis	**BGENIE**	**BOLT-LMM**	**PLINK2**	**RVTESTS**	**SAIGE**
Input File Format	**BGEN**	**BGEN/PLINK/dosage/VCF**	**BGEN/PLINK/dosage/VCF**	**BGEN/VCF/PLINK**	**BGEN/VCF/dosage**
Single Variant Association Analysis					
Quantitative Trait	✓	✓	✓	✓	✓
Binary trait	✓	✓	X (only Wald test)	✓ (include correction for unbalanced case/ control samples)	✓ (include correction for unbalanced case/ control samples)
Gene-level Association Test					
Simple burden; SKAT and Variable Threshold Test	X	X	X	✓	X
Support for Related Individuals and Linear Mixed Model Analysis	X	✓	X	✓	✓
Generation of Summary Statistics	X	X	X	✓	X