Bayesian model comparison for rare variant association studies of multiple phenotypes

Christopher DeBoever[1], Matthew Aguirre[1], Yosuke Tanigawa[1,2], Chris C. A. Spencer[3], Timothy Poterba[4], Mark J. Daly[4,5], Matti Pirinen[6], Manuel A. Rivas[1*],

1 Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
2 Biomedical Informatics Training Program, Stanford University, Stanford, CA, USA
3 Genomics plc, Oxford, UK
4 Broad Institute of MIT and Harvard, Cambridge, MA, USA
5 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
6 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
7 Oxford Centre for Diabetes, Endocrinology and Metabolism, Oxford, UK
8 Institute for Molecular Medicine Finland, University of Helsinki, Finland

* mrivas@stanford.edu

Abstract

Whole genome sequencing studies applied to large populations or biobanks with extensive phenotyping raise new analytic challenges. The need to consider many variants at a locus or group of genes simultaneously and the potential to study many correlated phenotypes with shared genetic architecture provide opportunities for discovery and inference that are not addressed by the traditional one variant-one phenotype association study. Here we introduce a model comparison approach we refer to as MRP for rare variant association studies that considers correlation, scale, and location of genetic effects across a group of genetic variants, phenotypes, and studies. In so doing we consider the use of summary statistic data to apply 1) standard univariate and multivariate gene-based meta-analysis models, 2) models for assessing heterogeneity of genetic effects, which may be used in practice for downstream quality control, and 3) models for identifying protective protein-truncating variants, which can expedite drug discovery. Through simulation studies, we demonstrate that the proposed model comparison approach may improve ability to detect rare variant association signals. Finally, we demonstrate its use with rare variant data combined with asthma diagnosis and haematological and spirometry measures for individuals in the UK Biobank. We show that we are able to retain useful features from widely-used meta-analysis approaches and prioritize protective modifiers of disease risk.

Author summary

TODO

Introduction

Sequencing technologies are quickly transforming human genetic studies of complex traits: it is increasingly possible to obtain whole genome sequence data on thousands of

samples at manageable costs. As a result, the genome-wide study of rare variants (minor allele frequency [MAF] $< 1\%$) and their contribution to disease susceptibility and phenotype variation is now feasible [**?**, **?**, **?**, **?**].

In genetic studies of diseases or continuous phenotypes, rare variants are hard to assess individually due to the limited number of copies of each rare variant. Hence, to boost the ability to detect a signal, evidence is usually 'aggregated' across variants. When designing an 'aggregation' method, there are three questions that are usually considered. First, across which biological units should variants be combined; second, which variants mapping within those units should be included [**?**]; and third, which statistical model should be used [**?**]? Given the widespread observations of shared genetic risk factors across distinct diseases, there is also considerable motivation to use gene discovery approaches that leverage the information from multiple phenotypes jointly. In other words, rather than only aggregating variants that may have effects on a single phenotype, we can also bring together sets of phenotypes for which a single variant or sets of variants might have effects.

In this paper, we present a Bayesian multiple rare variants and phenotypes (MRP) model comparison approach for identifying rare variant associations as an alternative to current widely-used statistical tests. The MRP framework exploits correlation, scale, or location (direction) of genetic effects in a broad range of rare variant association study designs including: case-control; multiple diseases and shared controls; single continuous phenotype; multiple continuous phenotypes; or a mixture of case-control and multiple continuous phenotypes (Fig 1). MRP makes use of Bayesian model comparison, whereby we compute a Bayes Factor (BF) defined as the ratio of the marginal likelihoods of the observed data under two models: 1) a pre-specified null where all genetic effects are zero; and 2) an alternative model where factors like correlation, scale, or location of genetic effects are considered. The BF is an alternative to p-values from traditional hypothesis testing. For MRP, the BF represents the statistical evidence for a non-zero effect for a particular group of rare variants on the phenotype(s) of interest.

Fig 1. Schematic overview of MRP. A: MRP is suitable for a broad range of rare variant association study designs including (from left to right): i) case-control, ii) multiple diseases with shared controls, iii) single quantitative phenotype, and iv) mixture of case-control and quantitative phenotypes. B: Diagram of factors considered in rare variant association analysis including the correlation matrices: $\mathbf{R}_{\text{study}}$ (expected correlation of genetic effects among a group of studies), $\mathbf{R}_{\text{var}}$ (expected correlation of genetic effects among a group of variants), and $\mathbf{R}_{\text{phen}}$ (expected correlation of genetic effects among a group of phenotypes); the scale parameter for genetic variant annotation; and the location of genetic effects, which may be used to prioritize or identify protective modifiers of disease risk.

While many large genetic consortia generally collect both raw genotype and phenotype data, in practice, sharing of individual genotype and phenotype data across groups is difficult to achieve. To address this, MRP can take summary statistics, such as estimates of effect size and the corresponding standard error from typical single variant-single phenotype linear or logistic regressions, as input data. Furthermore, we use insights from Liu et al. [**?**] and Cichonska et al. [**?**] who suggest the use of additional summary statistics, like covariance estimates across variants and studies, respectively, that would enable lossless ability to detect gene-based association signals using summary statistics alone.

Aggregation techniques rely on variant annotations to assign variants to groups for analysis. MRP allows for the inclusion of priors on the spread of effect sizes that can be adjusted depending on what type of variants are included in the analysis. For instance, protein truncating variants (PTVs) [**?**, **?**] are an important class of variants that are

more likely to be functional because they often disrupt the normal function of a gene. This biological knowledge can be reflected in the choices of priors for PTVs in MRP. Since PTVs typically abolish gene function, there is particular interest in identifying protective PTV modifiers of human disease risk that may serve as targets for therapeutics [**?**, **?**, **?**]. We therefore demonstrate how the MRP model comparison approach can improve discovery of such protective signals by modeling the location (direction) of genetic effects which prioritizes variants or genes that are consistent with protecting against disease.

To evaluate the performance of MRP and to study its behavior we use simulations and compare it to other commonly used approaches. Some simple alternatives to MRP include univariate approaches for rare variant association studies including the sequence kernel association test (SKAT) [**?**], and the burden test, which we show are special cases of the MRP model comparison when we assign the prior correlation of genetic effects across different variants to be zero or one. Furthermore, we show that MRP is able to detect the presence of heterogeneity of effects, which in some circumstances, such as those where effects are observed to be heterogeneous across studies, may indicate the presence of technical sources of error.

We also apply MRP to summary statistics for two groups of related phenotypes from the UK Biobank. First, we apply MRP to asthma (HC382: the corresponding phenotype label in Global Biobank Engine [`https://biobankengine.stanford.edu`]), eosinophil count (INI30150), Forced Expiratory Volume in 1-second (FEV1, INI3063), and Forced Vital Capacity (FVC, INI3062). We find that XXX. We next apply MRP to glaucoma (HC276), intra-ocular pressure (INI5263), and corneal resistance factor (INI5265) and find strong evidence that rare PTVs in *ANGPTL7* protect against glaucoma. These analyses show that MRP recovers results from typical single variant-single phenotype analyses while identifying new rare variant associations that include protective modifiers of disease risk.

Materials and methods

A complete description of MRP, along with an analytical derivation, is given in the Online Methods and Appendix. Briefly, we describe the approach.

In Bayesian model comparison, we compare the null and alternative models via the Bayes Factor (BF) defined as the ratio of the marginal likelihoods of the observed data under two models. Here, the null model is that the effect sizes obtained across all studies for a group of variants and a group of phenotypes is zero. To complete the specification of the alternative model we consider the correlation structure, the scale, and the location of the effect sizes (which we find to be informative for alternative models where we seek to identify variants with protective effects).

Let N be the number of individuals and K the number of quantitative phenotype measurements on each individual, for example, here in the applications we focus on asthma, eosinophil count, Forced Expiratory Volume in 1-second, and Forced Vital Capacity, in which case $K = 4$. Let M be the number of variants in a testing unit $\boldsymbol{G}$, where $\boldsymbol{G}$ can be, for example, a gene, pathway, or a network. Let S be the number of studies where data is obtained from - this data may be in the form of raw genotypes and phenotypes or summary statistics including linkage-disequilibrium, effect sizes (or odds ratio), and standard error of the effect size.

In the multivariate setting we define the prior correlation structure of the effect sizes, denoted by a SMK-by-SMK matrix $\mathbf{U}$. In practice, we define $\mathbf{U}$ as a Kronecker product, an operation of matrices of arbitrary size, of three sub-matrices: an S-by-S matrix $\mathbf{R}_{\text{study}}$ being the correlations of genetic effects among studies where different values can be used to compare different models of association, such as for identifying heterogeneity

of effect sizes between populations [?]; an M-by-M matrix $\mathbf{R}_{\text{var}}$ being the correlations of genetic effects among genetic variants, which may reflect the assumption that all the PTVs in a gene may have the same biological consequence [?, ?, ?] or prior information obtained through integration of additional data sources, such as functional assay data [?, ?], otherwise zero correlation of genetic effects may be assumed, which is used in dispersion tests like C-alpha [?, ?] and SKAT [?]; and a K-by-K matrix $\mathbf{R}_{\text{phen}}$ being the correlations of genetic effects among phenotypes, which may be obtained from common variant data [?, ?, ?]. The variance-covariance matrix of the effect sizes may be obtained from readily available summary statistic data such as in-study LD matrices, effect size estimate (or log odds ratio), and the standard error of the effect size estimate (Online Methods).

For gene discovery, the scale may be used to denote the prior spread of effect sizes. For instance, emerging empirical genetic studies have shown that within a gene protein truncating variants may have stronger effects than missense variants [?].

Thus far we have assumed that the prior mean, or location, of genetic effects is zero. This makes it feasible to analyze a large number of phenotypes without enumerating the prior mean across all phenotypes. However, in practice, we may want to proactively identify genetic variants that have effects that are consistent with a protective profile for a disease, which may exploit information from mendelian randomization studies of common variants, such as recent rare variant findings in *PCSK9* where truncating loss-of-function variants are found to decrease LDL and triglyceride levels, and decrease CAD risk [?, ?, ?, ?]. This is straightforward to consider by including a non-zero vector as a prior mean of genetic effect (Online Methods).

Although we see advantages in adopting a Bayesian perspective, our approach could be used in a frequentist context by calculating a Bayes factor and using it as a test statistic to compute p-values (Methods, Fig 2).

Results

Fig 2. Simulation studies. A: Comparison of $-\log_{10}$(p-values) from frequentist BF_{MRP} approximation for an independent effects and a similar effects model to commonly used gene-based statistical tests (skat and burden). B: Comparison of log10(Bayes Factors) obtained when raw genotype and phenotype data is available to a scenario where summary statistics only was available and similar effects across studies is assumed. C: From single variant and single phenotype to multiple variants and multiple phenotypes gene discovery: ROC curves for detecting gene association to any of the phenotypes using single variant/single phenotype association (blue) to multiple variants and multiple phenotypes association (purple). D: ROC curves for detecting gene association when incorporating prior mean of genetic effects (orange) to identify protective alleles.

Simulation studies

We first verified the analytical derivations and examined the properties of the approach under a simulation framework.

Comparison to frequentist gene tests

For the analysis of multiple rare variants and a single phenotype we compared it to the burden test and the SKAT test, commonly used statistical tests in rare variant

association studies of a single phenotype. We observe concordance between the frequentist methods and the Bayesian models. To compare the Bayesian models we compute p-values by using the Bayes Factor as the test statistic and approximating it using distribution properties of quadratic forms (Appendix). As expected, an independent effects model has high correlation with the gene-based test SKAT ($r^2 = 0.99$), whereas the similar effects model has high correlation with the burden test ($r^2 = 0.93$, Fig 2A).

Summary statistic data

To study the behavior of MRP using summary statistics we simulate two scenarios: first, the scenario where analysts have access to all the raw genotype and phenotype data; and second, the scenario where analysts only have access to summary statistics data [**?**]. We conducted 1000 simulation experiments where we let K (the number of phenotypes) = 3, M (the number of variants) = 10, S (the number of studies) = 2, N_0 (number of individuals in study with access to all the data) = 10000, N_1 (meta-analysis study 1) = 5000, N_2 (meta-analysis study 2) = 5000. We find that, under the scenario where similar effects are assumed across studies, the Bayes Factors obtained using summary statistics alone are strongly correlated ($r^2 = 1$) to Bayes Factors obtained by the full genotype and phenotype data (Fig 2B).

From single variant and single phenotype analysis to multiple variants and multiple phenotypes

To validate the flexibility of the approach we conducted a simulation experiment where we assumed an allelic architecture consistent to that discovered for *APOC3* in relation to coronary artery disease (CAD), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), and high-density lipoprotein cholesterol (HDL-C) [**?**, **?**, **?**, **?**]. We simulated three studies and applied the model comparison unit jointly to summary statistic data obtained for each study (Supplementary Note). Overall, we observed that considering the joint effects across multiple studies in a group of variants and phenotypes may improve ability to detect gene-based signals (Fig 2C), and that considering prior mean of genetic effects should aid in efforts to identify protective modifiers of disease risk (Fig 2D).

Applications

We demonstrate its use in rare variant analysis in the UK Biobank.

Model comparison applied to summary statistic from the UK Biobank

We applied the MRP model comparison approach to summary statistic data generated from single variant logistic regression and linear regression analysis for coding variants on the UK Biobank array (Methods). We first applied MRP to the asthma, eosinophil count, forced expiratory volume in 1-second, and forced vital capacity phenotypes. We applied the model to missense variants and PTVs for each phenotype separately (Fig **??**A-D) as well as to all phenotypes jointly (Fig **??**E) and obtained $\log_{10}$ BF for each gene. As expected from recent studies [**?**, **?**], we observed evidence that missense variants and PTVs in *IL33* affect eosinophil counts and offer protection from asthma from the single-phenotype analyses, though the evidence of association was strongest for the joint analysis (Table **??**). We also found evidence for associations between missense variants and PTVs in several other genes and the four phenotypes in the joint model that were not identified when considering asthma alone. For instance, *GPR20*, *CSF2RB*,

PACS2, and *CCR3* all had $\log_{10}$ BF greater than 9 in the eosinophil count-only analysis but $\log_{10}$ BF less than 0 for the asthma-only analysis. In the joint model, the $\log_{10}$ BFs for these genes were all greater than six indicating possible associations with these traits. Besides *IL33*, none of the genes with $\log_{10}$ BF greater than six in the joint model were reported in a large GWAS for allergic disease including asthma [**?**] though *CCR3* is near a locus associated with atopy in a previous meta-analysis [**?**].

We also applied MRP to missense variants and PTVs for glaucoma, intra-ocular pressure, and corneal resistance factor as well as performing a joint analysis and a joint analysis focused on protective associations. While the individual glaucoma analysis did not yield any associations with $\log_{10}$ BF greater than 5, the joint analyses identified missense and PTVs in both *ANGPTL7* and *WNT10A* as associated with protection against glaucoma. These results illustrate the ability of applying model comparison to multiple traits simultaneously to identify protective modifiers of disease.

Fig 3. Results for glaucoma application. $\log_{10}$ Bayes Factors from applying MRP to summary statistics for missense and protein-truncating variants from (A) glaucoma (HC276), (B) intra-ocular pressure (INI5263), (C) corneal resistance factor (INI5265), and (D) all three traits jointly. (E) shows the results of a joint analysis focused on finding rare variants that protect against glaucoma. The genes outside of chromosome 6 with with Bayes Factor greater than three are indicated by arrows. Only $\log_{10}$ Bayes Factors greater than zero are plotted. F: $-\log_{10}$ p-values (left panel) and estimated effect sizes with 95% confidence intervals (right panel) for missense variants and PTVs in *ANGPTL7* for all three phenotypes.

Table 1. $\log_{10}$ **Bayes Factors for genes highlighted in Figure 3.**

Gene	Joint, protective	Joint	Glaucoma	Intra-ocular pressure, Goldmann-correlated	Corneal resistance factor
ANGPTL7	10.1	9.4	1.7	10.1	6.2
KLHL22	3.1	3.6	-0.2	3.9	2.2
GLI3	2.9	2.9	-0.3	-0.7	3.2
WNT10A	2.2	3.0	-0.3	-0.5	3.5
RNASE2	-1.3	-0.3	0.4	-1.0	-1.1
FUK	-6.5	-2.3	2.5	-2.1	-2.1

Discussion

In this study, we developed a Bayesian model comparison approach MRP that shares information across both variants and phenotypes to identify rare variant associations. We used simulations to compare MRP to the widely used burden and SKAT tests for identifying rare variant associations and found that jointly considering both variants and phenotypes can improve the ability to detect associations. We also applied the MRP model comparison framework to summary statistic data from four traits from the UK Biobank: asthma, eosinophil counts, forced expiratory volume in 1-second (FEV1), and forced vital capacity (FVC). We identified strong evidence for the previously described association between *IL33* and asthma [**?**, **?**] as well as evidence of association for several additional genes that have not been previously reported (Table **??**).

As genetic data linked to high-dimensional phenotype data continues to be made available through biobanks, health systems, and research programs, there is a large need for statistical approaches that can leverage information across different genetic variants,

Table 2. $\log_{10}$ **Bayes Factors for genes highlighted in Figure ??.**

Gene	Joint, protective	Joint	Eosinophil count	FVC	FEV1	Asthma
IL33	30.1	29.8	30.9	-3.1	-2.7	9.5
CCR3	5.8	6.3	9.1	-1.7	-2.3	-0.5
IL17RA	4.6	6.1	14.1	-4.1	-3.9	-1.4
AL161450.1	3.2	3.8	4.9	-0.1	-0.4	0.3
RP11-39K24.9	3.2	3.8	4.9	-0.1	-0.4	0.3
ELN	2.9	3.2	-1.0	5.6	2.9	-0.6
SCMH1	2.7	3.9	-2.1	7.5	3.3	-0.7
TGFB2	2.6	4.1	-2.8	1.7	-2.4	0.1
ZFAT	1.3	1.9	-2.8	6.7	4.6	-0.4
STC2	0.2	1.3	-0.8	3.9	2.3	-0.7
P4HTM	-1.3	-0.7	-1.6	2.9	4.3	-1.0
IL18RAP	-2.2	-1.0	-0.8	-2.0	-2.2	3.6
WDR6	-4.1	-2.9	-2.3	1.9	3.4	-0.9
ATP2A3	-4.2	-1.5	-1.8	-3.1	-3.1	5.9
CHI3L1	-4.4	-2.0	-2.0	-1.9	-2.0	3.0

phenotypes, and studies to make strong inferences about disease-associated genes. The approach presented here relies only on summary statistics from marginal association analyses which can be shared with less privacy concerns compared to raw genotype and phenotype data. Combining joint analysis of variants and phenotypes with meta-analysis across studies offers new opportunities to identify gene-disease associations.

Supporting information

S1 Fig. Bold the title sentence. Add descriptive text after the title of the item (optional).

S2 Fig. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 File. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Video. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Appendix. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

S1 Table. Lorem ipsum. Maecenas convallis mauris sit amet sem ultrices gravida. Etiam eget sapien nibh. Sed ac ipsum eget enim egestas ullamcorper nec euismod ligula. Curabitur fringilla pulvinar lectus consectetur pellentesque.

Author Contributions

M.A.R. and M.P. designed the method and derived all analytical calculations. M.A.R., M.P., and C.D. wrote the manuscript. Y.T., M.A., and C.D. provided analysis and designed figures. T.P. designed hDF5 tables and implementation of loaders. M.J.D. provided critical feedback on methodology.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 24983. We thank all the participants in the UK Biobank study. M.A.R. and C.D. are supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (5U01 HG009080). C.D. is supported by a postdoctoral fellowship from the Stanford Center for Computational, Evolutionary, and Human Genomics and the Stanford ChEM-H Institute. Y .T. is supported by Funai Overseas Scholarship from Funai Foundation for Information Technology and the Stanford University Biomedical Informatics Training Program. The primary and processed data used to generate the analyses presented here are available in the UK Biobank access management system (https://amsportal.ukbiobank.ac.uk/) for application 24983, "Generating effective therapeutic hypotheses from genomic and hospital linkage data" (http://www.ukbiobank.ac.uk/wp-content/uploads/2017/06/24983-Dr-Manuel-Rivas.pdf), and the results are displayed in the Global Biobank Engine (https://biobankengine.stanford.edu). We would like to thank the Customer Solutions Team from Paradigm4 who helped us implement efficient databases for queries and application of inference methods to the data, and also implemented optimized versions of truncated singular value decomposition. M.A.R. is a paid consultant in Genomics PLC.