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 [1–4].

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 [5]; and third, which statistical model should be used [6]? 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. [7] and Cichonska et al. [8] 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) [9, 10] 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 [11–13]. 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) [14], 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

MRP model comparison for association testing

We consider the multivariate linear regression model

$$\underset{(N\times K)}{\mathrm{Y}} = \underset{(N\times K)}{\boldsymbol{\Psi}} + \underset{(N\times M)(M\times K)}{\mathrm{X}\ \mathrm{B}} + \underset{(N\times K)}{\mathrm{E}},$$

where the matrices $\mathrm{Y} = [y_{ik}]$, $\mathrm{X} = [x_{im}]$, $\mathrm{B} = [\beta_{mk}]$ and $\mathrm{E} = [e_{ik}]$ describe the phenotype values (y_{ik}), copies of minor allele (x_{im}), variant-phenotype effects (β_{mk}), and residual errors (e_{ik}), for individual i, phenotype k, and variant m. We assume that each phenotype has been transformed to a standard normal distribution and that the columns of X have been centered, which means that the estimate for the intercept term $\boldsymbol{\Psi}$ is 0 and independent of the estimate of B. We use vectorized notation where the rows of B form vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)^\mathsf{T}$ of length MK.

We define the MRP model comparison as a Bayes factor (BF) between the alternative model, where at least one variant affects at least one phenotype, and the null model, where all variant-phenotype effects are zero. BF is the ratio of the marginal likelihoods for these two models:

$$\mathrm{BF} = \frac{\int_{\boldsymbol{\beta}} p\left(\mathrm{Data}|\boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}|\mathrm{ALT}\right) d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}} p\left(\mathrm{Data}|\boldsymbol{\beta}\right) p\left(\boldsymbol{\beta}|\mathrm{NULL}\right) d\boldsymbol{\beta}},$$

where Data can correspond either to the effect size estimates $\widehat{\boldsymbol{\beta}}$ and the estimated

variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathrm{V}}_{\boldsymbol{\beta}}$, or to the original phenotypes and genotypes, $\underset{(N\times K)}{\mathrm{Y}}$ and $\underset{(N\times M)}{\mathrm{X}}$, and any other covariates that we want to regress out from the phenotypes.

The prior distribution for the null model, $p(\boldsymbol{\beta}|\mathrm{NULL})$, is simply the point mass at $\boldsymbol{\beta} = 0$. In subsection we show how we approximate the likelihood function for $\boldsymbol{\beta}$, $p(\mathrm{Data}|\boldsymbol{\beta})$, in subsection we define the prior distribution $p(\boldsymbol{\beta}|\mathrm{ALT})$ for the alternative model, and finally, in subsection , we compute the BF.

Likelihood function

A maximum likelihood estimator of B is given by the ordinary least-squares method

$$\widehat{\mathrm{B}} = \left(\mathrm{X}^{\intercal}\mathrm{X}\right)^{-1}\mathrm{X}^{\intercal}\mathrm{Y},$$

that in vectorized form is denoted by $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_M\right)^{\intercal}$. An estimator of the variance-covariance of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\mathrm{V}}_{\boldsymbol{\beta}} = \left(\mathrm{X}^{\intercal}\mathrm{X}\right)^{-1} \otimes \widehat{\mathrm{V}}_{\mathrm{Y}},$$

where $\widehat{\mathrm{V}}_{\mathrm{Y}}$ is the estimated residual variance-covariance matrix of Y given X.

Following Band et al. [15], we approximate the likelihood function of $\boldsymbol{\beta}$ by a multivariate normal distribution with mean $\widehat{\boldsymbol{\beta}}$ and variance-covariance matrix $\widehat{\mathrm{V}}_{\boldsymbol{\beta}}$. Note that by approximating $\widehat{\mathrm{V}}_{Y}$ by the trait correlation matrix, this likelihood approximation does not require access to the individual level data X and Y but only to the summary data of effect sizes $\widehat{\boldsymbol{\beta}}$, LD-matrix $\mathrm{X}^{\intercal}\mathrm{X}$ and a trait correlation estimate.

Prior of $\boldsymbol{\beta}$ in the alternative model

We construct the prior distribution $p(\boldsymbol{\beta}|\mathrm{ALT})$ for the alternative model in three steps allowing user to specify correlations between effects of different variants on different traits across different studies.

In a single study, the prior density for $\boldsymbol{\beta}$ incorporates the expected correlation of genetic effects among a group of variants ($\mathbf{R}_{\mathrm{var}}$) and among a group of phenotypes ($\mathbf{R}_{\mathrm{phen}}$). In addition, we incorporate an expected spread of the effect size of each variant by scaling $\mathbf{R}_{\mathrm{var}}$ as

$$\mathbf{S}_{\mathrm{var}} = \Delta(\sigma_m)\,\mathbf{R}_{\mathrm{var}}\Delta(\sigma_m),$$

where $\Delta(\sigma_m)$ is a diagonal matrix with entries σ_m determining the spread of the effect size distribution for each variant $m \leq M$. Thus, we can model settings where, e.g., protein-truncating variants have larger effect sizes ($\sigma = 1$) than missense variants ($\sigma = 0.1$). Note that when $\sigma_m = 1$ for all m then $\mathbf{S}_{\mathrm{var}} = \mathbf{R}_{\mathrm{var}}$.

All in all, our prior density for $\boldsymbol{\beta}$ under alternative model is

$$\boldsymbol{\beta}|\mathrm{ALT} \sim \mathcal{N}(\mathbf{0}, \mathbf{U}), \text{ where } \mathbf{U} = \mathbf{S}_{\mathrm{var}} \otimes \mathbf{R}_{\mathrm{phen}}.$$

When we have data from multiple studies we allow for possible differences in genetic effects across ethnicities or populations extending the Approximate Bayes Factors of Band et al. [15] and the summary statistics approach of RAREMETAL [7] from univariate to multivariate phenotypes. Let $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_{s,m,k}\right) = \left(\widehat{\beta}_{1,1,1}, \widehat{\beta}_{1,1,2}, \ldots, \widehat{\beta}_{1,1,K}, \widehat{\beta}_{1,2,1}, \ldots, \widehat{\beta}_{1,2,K}, \ldots, \widehat{\beta}_{1,M,K}, \widehat{\beta}_{2,1,1}, \ldots, \widehat{\beta}_{S,M,K}\right)$, where S is the number of studies, M is the number of variants, and K is the number of phenotypes. As with a single study, we incorporate the expected correlation of genetic effects between a pair of variants and a single phenotype using the matrix $\mathbf{S}_{\mathrm{var}}$, between a variant and a

pair of phenotypes using the matrix $\mathbf{R}_{\text{phen}}$, and we introduce the matrix $\mathbf{R}_{\text{study}}$ to specify prior on the similarity in effect sizes across the studies. Thus, the prior is

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \mathbf{U}), \text{ where } \mathbf{U} = \mathbf{R}_{\text{study}} \otimes (\mathbf{S}_{\text{var}} \otimes \mathbf{R}_{\text{phen}}).$$

It is straightforward to include a non-zero vector $\boldsymbol{\mu}$ as a prior mean of genetic effects, in which case the prior is

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{U}).$$

We use this, for example, when screening for protective rare variants that have a pre-specified beneficial profile on a set of risk factors.

$\mathbf{BF_{MRP}}$

The Bayes Factor is the ratio of the marginal likelihoods between the alternative and the null model. The marginal likelihood for the alternative model is

$$\int_{\boldsymbol{\beta}} p(\text{Data}|\boldsymbol{\beta})\, p(\boldsymbol{\beta}|\text{ALT})\, d\boldsymbol{\beta} = c \times \mathcal{N}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{\mu}, \widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)$$

and the marginal likelihood for the null model is

$$\int_{\boldsymbol{\beta}} p(\text{Data}|\boldsymbol{\beta})\, p(\boldsymbol{\beta}|\text{NULL})\, d\boldsymbol{\beta} = c \times \mathcal{N}\left(\widehat{\boldsymbol{\beta}}; 0, \widehat{\mathrm{V}}_{\boldsymbol{\beta}}\right).$$

The Bayes Factor (derived is below) is given by

$$\text{BF}_{\text{MRP}} = \frac{\det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)^{\intercal}\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)\right]}{\det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\widehat{\boldsymbol{\beta}}^{\intercal}\,\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\,\widehat{\boldsymbol{\beta}}\right]}.$$

When $\boldsymbol{\mu} = 0$, BF_{MRP} is an increasing function of the following quadratic form

$$Q(\widehat{\boldsymbol{\beta}}; \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U}) = \widehat{\boldsymbol{\beta}}^{\intercal}\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1} - (\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U})^{-1}\right)\widehat{\boldsymbol{\beta}}. \quad (1)$$

Furthermore, this quadratic form is the only part of the BF_{MRP} that depends on $\widehat{\boldsymbol{\beta}}$. Thus, by deriving a distribution of $Q(\widehat{\boldsymbol{\beta}}; \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U})$ under the null model we can compute a p-value when BF_{MRP} is used as a test statistic. According to basic properties of quadratic forms of Gaussian variables, $Q(\widehat{\boldsymbol{\beta}}; \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U}) \sim \sum_{i=1}^{n} d_i \chi_i^2$, where χ_i^2 are an independent sample from χ_1^2 distribution (chi-square with one degree of freedom), and d_i are the eigenvalues of matrix $I - (\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U})^{-1}\,\widehat{\mathrm{V}}_{\boldsymbol{\beta}}$. The distribution function for a mixture of chi-squares can be numerically evaluated by the R-package 'CompQuadForm' [16].

MRP Bayes Factor derivation

To compute the Bayes Factor

$$\text{BF}_{\text{MRP}} = \frac{\det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)^{\intercal}\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)\right]}{\det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\widehat{\boldsymbol{\beta}}^{\intercal}\,\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\,\widehat{\boldsymbol{\beta}}\right]},$$

we first consider the term inside the exponential function:

$$\mathcal{E}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\mu}, \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U}\right) = \frac{1}{2}\,\widehat{\boldsymbol{\beta}}^{\intercal}\,\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\,\widehat{\boldsymbol{\beta}} - \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)^{\intercal}\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right)^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right).$$

Since $\widehat{\mathrm{V}}_{\boldsymbol{\beta}}$ and $\mathbf{U}$ are typically defined through Kronecker products of smaller matrices, their inverses are easier to compute than the inverse of their sum. Hence we use Woodbury matrix identity to write

$$\mathcal{E}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\mu}, \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U}\right) = \frac{1}{2}\,\widehat{\boldsymbol{\beta}}^{\intercal}\,\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\,\widehat{\boldsymbol{\beta}} - \frac{1}{2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right)^{\intercal}\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1} - \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\left(\mathbf{U}^{-1} + \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\right)^{-1}\widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\mu}\right).$$

To simplify the determinant calculation we write

$$\det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}} + \mathbf{U}\right) = \det\left(\widehat{\mathrm{V}}_{\boldsymbol{\beta}}\right)\det\left(\mathbf{I} + \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\mathbf{U}\right).$$

The logarithm of the Bayes Factor is then

$$\log\left(\mathrm{BF}_{\mathrm{MRP}}\right) = -\frac{1}{2}\log\left(\det\left(\mathbf{I} + \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{-1}\mathbf{U}\right)\right) + \mathcal{E}\left(\widehat{\boldsymbol{\beta}}, \boldsymbol{\mu}, \widehat{\mathrm{V}}_{\boldsymbol{\beta}}, \mathbf{U}\right).$$

If studies do not share individuals, $\widehat{\mathrm{V}}_{\boldsymbol{\beta}}$ is a block-diagonal matrix

$$\widehat{\mathrm{V}}_{\boldsymbol{\beta}} = \left[\begin{array}{c|c|c|c} \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{1} & 0 & \cdots & 0 \\ \hline 0 & \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{2} & \cdots & 0 \\ \hline \vdots & & \ddots & \vdots \\ \hline 0 & 0 & \cdots & \widehat{\mathrm{V}}_{\boldsymbol{\beta}}^{S} \end{array}\right].$$

If studies share individuals, e.g., controls, we take the approach of Cichonska et al. [8] to use summary level data to estimate the correlation structure of the non-diagonal blocks caused by overlapping individuals.

HDF5 Tables

Although summary statistics are quicker to read and process than raw data, the number of studies meta-analyzed in this work is expected to be sufficiently large to require optimizations in data representation and processing (Fig). Our solution was the use of the HDF5 (Hierarchical Data Format 5) data representation to enable rapid processing of effect size, uncertainty, and cross-trait estimate data. HDF5 is a fast and lightweight file format designed for scientific data. It has bindings for R, Python, C/C++, Java, and nearly every other population programming language. Reading data from a table within a HDF5 file can be an order of magnitude faster than reading text files from a Unix file, and it makes it easier to organize data within an internal structure.

UK Biobank Data

GWAS Summary Statistics

We performed genome-wide association analysis using PLINK v2.00a(17 July 2017) as previously described [17]. For asthma, we used the Firth-fallback in PLINK, a hybrid algorithm which normally uses the logistic regression code described in (Hill 2017), but switches to a port of logistf() (https://cran.r-project.org/web/packages/logistf/index.html) in two cases: (1) one of the cells in the 2x2 allele count by case/control status contingency table is empty (2) logistic regression was attempted since all the contingency table cells were nonzero, but it failed to converge within the usual number of steps. We used the following covariates in our analysis: age, sex, array type, and the first four principal components, where array type is a binary variable that represents whether an individual was genotyped

with UK Biobank Axiom Array or UK BiLEVE Axiom Array. For variants that were specific to one array, we did not use array as a covariate. Asthma cases were defined using both Hospital Episode Statistics and verbal questionnaire responses. We used the provided values from the UK Biobank for eosinophil counts, forced vital capacity (FVC), and forced expiratory volume in 1-second (FEV1). The phenotype codes used throughout (asthma=HC382, eosinophil count=INI30150, Forced Expiratory Volume in 1-second (FEV1)=INI3063, and Forced Vital Capacity (FVC)=INI3062) correspond to the phenotype codes used in on the Global Biobank Engine [`https://biobankengine.stanford.edu`].

Genetic Correlations

We calculated the genetic between asthma, eosinophil counts, forced vital capacity (FVC), and forced expiratory volume in 1-second (FEV1) using the MultiVariate Polygenic Mixture Model (MVPMM) [18]. Briefly, MVPMM estimates genetic correlation given GWAS summary statistics (effect size and standard error of effect size estimate) by modeling GWAS summary statistics as generated from one of two mixture components. Summary statistics from variants in the null component are modeled as being drawn from a multivariate normal distribution with zero mean and covariance matrix that captures correlation in the summary statistics due to the use of shared subjects or other sources of correlation. Summary statistics from variants in the non-null component are modeled as being drawn from a multivariate normal distribution with zero mean, but the covariance matrix for the non-null component combines the covariance matrix from the null component with another covariance matrix that captures the genetic correlation between the phenotypes being considered.

UK Biobank Asthma Application

For the Manhattan plots and Table **??**, we removed any genes with non-unique gene symbols. We also removed the sense overlapping gene *CTD-3064M3.3* and and the antisense gene *ZFAT-AS1* since these genes overlap *GPR20* and *ZFAT* respectively and therefore have the same $\log_{10}$ Bayes Factors as those genes.

Results

Method overview

A complete description of MRP, along with an analytical derivation, is given in the Materials and Methods. 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 **U**. In practice, we define **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 [15]; 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 [9, 10, 19] or prior information obtained through integration of additional data sources, such as functional assay data [5, 20], otherwise zero correlation of genetic effects may be assumed, which is used in dispersion tests like C-alpha [21, 22] and SKAT [14]; 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 [23–25]. 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 [26].

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 [11, 27–29]. 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).

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 (Materials and Methods). 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 [7]. 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) [29–32]. 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 also 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 applied the MRP model comparison approach to asthma and three related traits and glaucoma and two related traits.

Model comparison applied to asthma and related phenotypes

We first applied MRP to GWAS summary statistics for asthma, eosinophil count, forced expiratory volume in 1-second (FEV_1), and forced vital capacity (FVC) phenotypes. Recent work has identified associations between the PTV rs146597587 in *IL33* and asthma and eosinophil counts [17, 33]. FEV_1 and FVC are measures of pulmonary function that are used to diagnosis and classify pulmonary disease [**?**]. To demonstrate the advantage of considering the phenotypes jointly, we applied the model to missense

variants and PTVs for each phenotype separately (Fig **??**A-D) as well as to all phenotypes jointly (Fig 3E) and obtained $\log_{10}$ BF for each gene. In agreement with previous 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 **??**) [17, 33]. We also performed an analysis focused on identifying protective variants which also identified the *IL33* association (Fig 3).

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 [34] though *CCR3* is near a locus associated with atopy in a previous meta-analysis [35].

Fig 3. Results for asthma application. $\log_{10}$ Bayes Factors from applying MRP to summary statistics for missense and protein-truncating variants from (A) asthma (HC382), (B) eosinophil counts (INI30150), (C) forced vital capacity (FVC, INI3062), (D) forced expiratory volume in 1-second (FEV1, INI3063), (E) all four traits jointly, and (F) all four traits jointly with focus on protective effects. The four genes outside of chromosome 6 with the largest Bayes Factors are labeled in each plot. Only $\log_{10}$ Bayes Factors greater than -5 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 *IL33* for each phenotype

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

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

Model comparison applied to glaucoma and related phenotypes

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 4. 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 2. $\log_{10}$ **Bayes Factors for genes highlighted in Figure 4.**

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 [17,33] as well as evidence of association for several additional genes that have not been previously reported (Table 2).

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, 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. HDF5 Implementation. Our HDF5 implementation contained the following components: first, a group with one table per annotation file. All effect size (beta) values and study-specific annotations were contained here, and the number of tables is limited by S (the number of studies) $\times$ K (the number of traits). Second, a group with site-site covariance data. While these covariance matrices may have dimension M (the number of variants) $\times$ M, we store the data as tables, each row specifiying the covariance between two variants. The number of tables should be the same as the previous set, capped by
S (the number of studies) $\times$ K (the number of traits). Third, we store one table with sigma values for each study/phenotype combination. In the event that the traits were rank-normal transformation was performed these sigma values are equal to 1. These are used to compute correlation between two datasets. Finally, we store a matrix/table pair for Vy null and its index. The Vy null matrix has dimensions $(S \times K) \times (S \times K)$ each entry specifying the estimated correlation of effect sizes between two datasets. The index table encodes row/column position of each dataset.

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. TODO: Add C. Spencer contributions.

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.

References

1. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009;324(5925):387–389.

2. 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. Nature. 2010;467(7319):1061–1073.

3. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature Genetics. 2011;43(11):1066–1073.

4. The 1000 Genomes Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.

5. Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, Fontanillas P, et al. Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. Proceedings of the National Academy of Sciences. 2014;111(36):13127–13132.

6. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics. 2014;95(1):5–23.

7. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. Nature Genetics. 2014;46(2):200–204.

8. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016;32(13):1981–1989.

9. Rivas MA, Pirinen M, Neville MJ, Gaulton KJ, Moutsianas L, Lindgren CM, et al. Assessing association between protein truncating variants and quantitative traits. Bioinformatics. 2013;29(19):2419–2426.

10. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015;348(6235):666–669.

11. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. Nature Genetics. 2005;37(2):161–5.

12. Cohen JC, Boerwinkle E, Mosley Jr TH, Hobbs HH. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. New England Journal of Medicine. 2006;354(12):1264–1272.

13. Sullivan D, Olsson AG, Scott R, Kim JB, Xue A, Gebski V, et al. Effect of a monoclonal antibody to PCSK9 on low-density lipoprotein cholesterol levels in statin-intolerant patients: the GAUSS randomized trial. JAMA. 2012;308(23):2497–2506.

14. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics. 2011;89(1):82–93.

15. Band G, Le QS, Jostins L, Pirinen M, Kivinen K, Jallow M, et al. Imputation-based meta-analysis of severe malaria in three African populations. PLoS genetics. 2013;9(5):e1003509.

16. Duchesne P, de Micheaux PL. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. Computational Statistics and Data Analysis. 2010;54:858–862.

17. DeBoever C, Tanigawa Y, McInnes G, Lavertu A, Chang C, Bustamante CD, et al. Medical relevance of protein-truncating variants across 337,208 individuals in the UK Biobank study. bioRxiv. 2017;doi:10.1101/179762.

18. DeBoever C, Rivas MA. Harnessing digital phenotyping to enhance genetic studies of human diseases. Submitted. 2017;.

19. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335(6070):823–8.

20. Findlay GM, Boyle EA, Hause RJ, Klein JC, Shendure J. Saturation editing of genomic regions by multiplex homology-directed repair. Nature. 2014;513(7516):120–123.

21. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genetics. 2011;7(3):e1001322.

22. Clarke GM, Rivas MA, Morris AP. A Flexible Approach for the Analysis of Rare Variants Allowing for a Mixture of Effects on Binary or Quantitative Traits. PLoS Genetics. 2013;9(8):e1003694.

23. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive sharing of genetic effects in autoimmune disease. PLoS genetics. 2011;7(8):e1002254.

24. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013;14(7):483–495.

25. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature genetics. 2015;.

26. Do R, Stitziel NO, Won HH, Jørgensen AB, Duga S, Merlini PA, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015;518(7537):102–106.

27. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. Nature Genetics. 2005;37(2):161–165.

28. Do R, Willer CJ, Schmidt EM, Sengupta S, Gao C, Peloso GM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature Genetics. 2013;45(11):1345–1352.

29. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. New England Journal of Medicine. 2014;371(1):22–31.

30. Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, Horenstein RB, et al. A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection. Science. 2008;322(5908):1702–5.

31. Hofker MH. APOC3 null mutation affects lipoprotein profile APOC3 deficiency: from mice to man. European Journal of Human Genetics. 2010;18(1):1–2.

32. Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. New England Journal of Medicine. 2014;371(1):32–41.

33. Smith D, Helgason H, Sulem P, Bjornsdottir US, Lim AC, Sveinbjornsson G, et al. A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. PLOS Genetics. 2017;13(3):1–24. doi:10.1371/journal.pgen.1006659.

34. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nature genetics. 2017;49(12):1752–1757.

35. Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunological reviews. 2011;242(1):10–30.