Bayesian model comparison for rare variant association studies of multiple phenotypes

Christopher DeBoever[1], Matthew Aguirre[1], Yosuke Tanigawa[1], Chris C. A. Spencer[2], Timothy Poterba[3], Carlos D. Bustamante [1,4], Mark J. Daly[3,5], Matti Pirinen[6,7,8*], Manuel A. Rivas[1*],

1 Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
2 Genomics plc, Oxford, UK
3 Broad Institute of MIT and Harvard, Cambridge, MA, USA
4 Department of Genetics, Stanford University, Stanford, CA, USA
5 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
6 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
7 Department of Public Health, University of Helsinki, Helsinki, Finland
8 Helsinki Institute for Information Technology HIIT and Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

* matti.pirinen@helsinki.fi
* 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. We consider the use of summary statistic data to apply univariate and multivariate gene-based meta-analysis models for identifying rare variant associations with an emphasis on protective protein-truncating variants that can expedite drug discovery. Through simulation studies, we demonstrate that the proposed model comparison approach can improve ability to detect rare variant association signals. We also apply the model to two groups of phenotypes from the UK Biobank: 1) asthma diagnosis (43,626 cases), eosinophil counts, forced expiratory volume, and forced vital capacity; and 2) glaucoma diagnosis (5,863 cases), intra-ocular pressure, and corneal resistance factor. We are able to recover known associations such as the protective association between rs146597587 in *IL33* and asthma ($\log_{10}$(Bayes Factor) = 29.4). We also find evidence for novel protective associations between rare variants in *ANGPTL7* and glaucoma ($\log_{10}$(Bayes Factor) = 13.1). Overall, we show that the MRP model comparison approach is able to retain and improve upon useful features from widely-used meta-analysis approaches for rare variant association analyses and prioritize protective modifiers of disease risk.

Author summary

Due to the continually decreasing cost of acquiring genetic data, we are now beginning to see large collections of individuals for which we have both genetic information and trait data such as disease status, physical measurements, biomarker levels, and more. These datasets offer new opportunities to find relationships between inherited genetic variation and disease. While it is known that there are relationships between different traits, typical genetic analyses only focus on analyzing one genetic variant and one phenotype at a time. Additionally, it is difficult to identify rare genetic variants that are associated with disease due to their scarcity, even among large sample sizes. In this work, we present a method for identifying associations between genetic variation and disease that considers multiple rare variants and phenotypes at the same time. By sharing information across rare variant and phenotypes, we improve our ability to identify rare variants associated with disease compared to considering a single rare variant and a single phenotype. The method can be used to identify candidate disease genes as well as genes that might represent attractive drug targets.

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.

While many large genetic consortia 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 or severely alter 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.

We applied MRP to summary statistics for two groups of related phenotypes from the UK Biobank. First, we applied 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 (FEV_1, INI3063), and forced vital capacity (FVC, INI3062) and recovered the reported association between a rare PTV in *IL33* and asthma [15, 16]. We also applied MRP to glaucoma (HC276), intra-ocular pressure (INI5263), and corneal resistance factor (INI5265) and find evidence that rare coding variants 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

Description of MRP

In this section, we provide an overview of the MRP model comparison approach. Refer to S1 Appendix for a detailed description. MRP models GWAS summary statistics as being distributed according to one of two models. The null model is that the regression effect sizes obtained across all studies for a group of variants and a group of phenotypes is zero. The alternative model is that summary statistics are distributed according to a multivariate normal distribution with mean zero and covariance matrix described below. MRP compares the evidence for the null and alternative model using a Bayes Factor (BF) that quantifies the amount of evidence for each model as the ratio of the marginal likelihoods of the observed data under two models.

To define the alternative model, we must specify the prior correlation structure, scale, and location (direction) of the effect sizes. Let N be the number of individuals and K the number of phenotype measurements on each individual. 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. When considering multiple studies ($S > 1$), multiple rare variants ($M > 1$), and multiple phenotypes ($K > 1$), we define the prior correlation structure of the effect sizes as an $SMK \times SMK$ 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 $S \times S$ matrix $\mathbf{R}_{\text{study}}$ containing 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 [17];
- an $M \times M$ matrix $\mathbf{R}_{\text{var}}$ containing 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, 18] or prior information obtained through integration of additional data sources, such as functional assay data [5, 19], otherwise zero correlation of genetic effects may be assumed, which is used in dispersion tests like C-alpha [20, 21] and SKAT [14]; and
- a $K \times K$ matrix $\mathbf{R}_{\text{phen}}$ containing the correlations of genetic effects among phenotypes, which may be obtained from common variant data [22–24].

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 estimates (or log odds ratios), and the standard errors of the effect size estimates (S1 Appendix).

MRP allows users to specify priors that reflect knowledge of the variants and phenotypes under study. For instance, we can define an independent effects model where each variant in the model may have different effect sizes. In this case, $\mathbf{R}_{\text{var}}$ is the identity matrix which reflects the assumption that the effect sizes of the variants are not correlated. We can also define a similar effects model by setting every value of $\mathbf{R}_{\text{var}}$ to ~ 1. This model assumes that all variants under consideration have similar effect sizes (with possibly differences in scale). This model may be appropriate for PTVs where each variant completely disrupts the function of the gene, leading to a gene knockout. The prior on the scale of effect sizes can also be used to denote which variants may have larger effect sizes. For instance, emerging empirical genetic studies have shown that within a gene, PTVs may have stronger effects than missense variants [25]. This can be reflected by adjusting the prior spread of effect sizes (σ) for PTVs (S1 Appendix).

Similarly, we can utilize a prior on the location (direction) of effects to specify alternative models where we seek to identify variants with protective effects against disease. Thus far we have assumed that the prior mean, or location, of genetic effects is zero which makes it feasible to analyze a large number of phenotypes without enumerating the prior mean across all phenotypes. To proactively identify genetic variants that have effects that are consistent with a protective profile for a disease, we can include a non-zero vector as a prior mean of genetic effect (S1 Appendix). We can exploit information from Mendelian randomization studies of common variants, such as recent findings where rare truncating loss-of-function variants in *PCSK9* were found to decrease LDL and triglyceride levels and decrease CAD risk [11, 26–28] to identify situations where such a prior is warranted.

Applying MRP to variants from a testing unit $\boldsymbol{G}$ yields a BF for that testing unit that describes the evidence that rare variants in that testing unit have a nonzero effect on the traits used in the model. For instance, consider genes as testing units. By running MRP, we obtain a BF for each gene that represents the evidence that rare variants in that gene affect the traits of interest. These BF can be used to identify specific genes that may be linked to disease. Although we see advantages in adopting a

Bayesian perspective for MRP, our approach could be used in a frequentist context by calculating a BF and using it as a test statistic to compute p-values (S1 Appendix, Fig 2).

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 (S1 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 [15]. For asthma, we used the Firth fallback in PLINK, a hybrid algorithm which normally uses the logistic regression code described in [29], 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 and glaucoma 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), forced expiratory volume in 1-second (FEV_1), intra-ocular pressure, and corneal resistance factor. The phenotype codes used throughout (asthma=HC382, eosinophil count=INI30150, FEV_1=INI3063, FVC=INI3062, glaucoma=HC276, intra-ocular pressure=INI5263, and corneal resistance factor=INI5265) correspond to the phenotype codes used on the Global Biobank Engine [`https://biobankengine.stanford.edu`].

Genetic Correlations

We calculated the genetic correlation between the two groups of traits (asthma, eosinophil counts, FVC, FEV_1 and glaucoma, intra-ocular pressure, corneal resistance factor) using the MultiVariate Polygenic Mixture Model (MVPMM) [30]. 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. We observed similar genetic correlations using LD score regression (S2 Fig) [24].

UK Biobank Asthma and Glaucoma Applications

For each group of traits (asthma, eosinophil counts, FVC, FEV_1 and glaucoma, intra-ocular pressure, corneal resistance factor), we applied MRP individually to each phenotype as well as performing a joint analysis using all traits. We also applied a model that prioritizes protective variants where we used non-zero priors for the variant effect size of -0.5 for PTVs and -0.2 for missense alleles. For each analysis, we applied MRP assuming an independent effects model and a similar effects model. We applied Bayesian model averaging to the results of the independent and similar effects models by summing the $\log_{10}$ BF for each gene from each model and dividing by two. The Bayesian model averaging results are reported in the main text while the results for each individual model are included in the Supporting Information.

For the Manhattan plots and tables, we removed any genes with non-unique gene symbols. In cases where genes overlapped such that they shared rare variants and therefore the same BF, we removed one gene. *ANGPTL7* protein expression was assessed using the HIPED protein expression database accessed through genecards.org on 2017/1/29 [31]. We identified the protein 1JC9_A as homologous to the *ANGPTL7* protein using the "3D structure mapping" link from dbSNP (`https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=28991009`). We retrieved the 3D structure image from the iCn3D Structure Viewer (`https://www.ncbi.nlm.nih.gov/Structure/icn3d/icn3d.html`).

Variant Filtering

We used the `variant_filter_table.tsv` file available at `https://github.com/rivas-lab/public-resources` (6f9f726) to filter variants on the UK Biobank array for use with MRP. We first chose variants with minor allele frequency less than 1%. We then filtered out all variants with `all_filters` less than one. This removes variants with missingness greater than 1% (calculated on an array-specific basis for array-specific variants) or Hardy-Weinberg equilibrium $p < 10^{-7}$. This also removes some PTVs for which manual inspection revealed irregular cluster plots [15]. We LD pruned the variants by only using variants with `ld` equal to one. We included missense variants and PTVs indicated by the following annotations: `missense_variant`, `stop_gained`, `frameshift_variant`, `splice_acceptor_variant`, `splice_donor_variant`, `splice_region_variant`, `start_lost`, `stop_lost`. We removed variants whose regression effect size had standard error greater than 0.15.

Results

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 BF as the test statistic and approximating it using distribution properties of quadratic forms (S1 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 [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) [28, 32–34]. 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 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 MRP separately to asthma and three related traits as well as glaucoma and two related traits.

Asthma, eosinophil counts, forced expiratory volume, and forced vital capacity

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 [15, 16]. FEV_1 and FVC are measures of pulmonary function that are used to diagnosis and classify pulmonary disease [35]. To demonstrate the advantage of considering the phenotypes jointly, we applied MRP to rare missense

variants and PTVs (MAF $< 1\%$) for each phenotype separately (Fig 3A-D) as well as to all phenotypes jointly (Fig 3E,F) and obtained $\log_{10}$ BF for each gene. We applied both independent and similar effects models and used Bayesian model averaging to compute a single BF per gene [36]. In agreement with previous studies, we observed evidence that rare missense variants and/or 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 ($\log_{10}$ BF $= 29.3$, S1 Table) [15, 16]. We performed an analysis focused on identifying protective variants which also identified the *IL33* association ($\log_{10}$ BF $= 29.4$, Fig 3F). The results were similar using only either the independent effects (S3 Fig) or similar effects models (S4 Fig). We inspected the effect sizes from the marginal GWAS regressions for the rare variants included in the analysis and found that the association identified by MRP is likely driven by the PTV rs146597587 (Fig 3G).

We also found moderate evidence for association between rare coding variants in *CCR3* and asthma. The $\log_{10}$ BFs for *CCR3* was 3.3 in the joint model compared to only -0.5 in the asthma-only analysis (Fig 3, S1 Table). *CCR3* is a chemokine receptor that is highly expressed on eosinophils and has been a therapeutic focus for asthma [37, 38]. *CCR3* was not reported in a large GWAS for allergic disease including asthma [39] though *CCR3* is near a locus associated with atopy in a previous meta-analysis [40]. These results demonstrate that MRP can identify biologically meaningful therapeutic targets that may be missed by standard GWAS approaches.

Considering multiple phenotypes jointly allows for the efficient prioritization of disease genes. For instance, some genes like *IL18RAP*, *ATP2A3*, and *FLG* had $\log_{10}$ BFs greater than 4 in the asthma-only analysis but much smaller BFs in the joint analyses indicating that rare variants in these genes are less likely to affect this group of traits. Similarly, there were other genes like *RP11-39K24.9* and *IL17RA* that had larger BFs in the eosinophil count-only analysis but small BFs for the joint analyses demonstrating MRP's ability to integrate information across all phenotypes considered.

Glaucoma, intra-ocular pressure, and corneal resistance factor

We also applied MRP to missense variants and PTVs for glaucoma, intra-ocular pressure, and corneal resistance factor as well as performing joint analyses. Intra-ocular pressure is a measure of the fluid pressure in the eye, is associated with glaucoma risk, and has been linked to genetic variants associated with glaucoma [41]. Corneal resistance factor is a measure of the cornea's ability to resist mechanical stress and has been associated with glaucoma presence and severity [42–44]. While the individual glaucoma analysis did not yield any associations with $\log_{10}$ BF greater than three, the joint analysis identified rare coding variants in *ANGPTL7* ($\log_{10}$ BF $= 12.2$), *KLHL22* ($\log_{10}$ BF $= 3.7$), and *WNT10A* ($\log_{10}$ BF $= 2.6$) as associated with glaucoma (Fig 4A-D, S2 Table). Applying the protective MRP model also identified the protective association for *ANGPTL7* against glaucoma and added support for associations for *KLHL22* and *WNT10A* (Fig 4E). We obtained similar results using the independent effects (S5 Fig) or similar effects models (S6 Fig).

Expression of *ANGPTL7* is upregulated in glaucoma and has been proposed to regulate intra-ocular pressure and glaucoma risk [45, 46]. The GWAS summary statistics for the rare variants in *ANGPTL7* suggest that the association with glaucoma is driven by the missense variant rs28991009 that changes residue 175 from glutamine to histidine (Fig 4F, G). According to the HIPED protein expression database, *ANGPTL7* protein is expressed at ~ 0.7 parts per million in vitreous humor, the material between the lens and retina of the eyeball; in contrast, the expression of *ANGPTL7* protein is less than 0.01 parts per million in 68 other normal tissues [31]. Such tissue-specific activity may make *ANGPTL7* a useful therapeutic target. *KLHL22* has not been previously

associated with glaucoma though a suggestive association was reported for retinopathy in individuals without diabetes [47]. *WNT10A* also has not been previously associated with glaucoma though an exonic variant rs121908120 in *WNT10A* is associated with central cornea thickness and increased risk of keratoconus, a disease of the cornea, indicating that this gene may play a role in ocular diseases [48].

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 two groups of traits from the UK Biobank: asthma diagnosis, eosinophil counts, FEV_1, and FVC; and glaucoma diagnosis, intra-ocular pressure, and corneal resistance factor. We identified strong evidence for the previously described association between the PTV rs146597587 in *IL33* and asthma [15, 16]. We also found evidence for a link between rare variants in *ANGPTL7* and glaucoma, consistent with previous experiments that suggested a role for *ANGPTL7* in glaucoma [45, 46]. These results demonstrate the ability of the MRP model comparison approach to leverage information across multiple phenotypes and variants to discover rare variant associations.

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.

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.P. is financially supported by the Academy of Finland [288509 and 294050]. C.D.B. and M.A.R. are supported by the GSP Coordinating Center (U24 HG008956). M.A.R., C.D., and C.D.B. are supported by Stanford University and a National Institute of Health center for Multi- and Trans-ethnic Mapping of Mendelian and Complex Diseases grant (U01 HG009080). M.A.R. is a Faculty Fellow at the Stanford Center for Population Health Sciences. 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 (T32 LM012409). 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. M.A.R. and M.P. are paid consultants in Genomics PLC. CDB is a member of the scientific advisory boards for Liberty Biosecurity, Personalis, 23andMe Roots into the Future, Ancestry.com, IdentifyGenomics, and Etalon and is a founder of C.D.B. Consulting. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

MAR and MP designed the method and derived all analytical calculations. MAR, MP, and CD wrote the manuscript. MAR, MP, CCAS, YT, MA and CD provided analysis and designed figures. TP designed HDF5 tables and implementation of loaders. MJD and CDB provided critical feedback on methodology.

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. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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;.

25. 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.

26. 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.

27. 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.

28. 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.

29. Hill A, Loh PR, Bharadwaj RB, Pons P, Shang J, Guinan E, et al. Stepwise Distributed Open Innovation Contests for Software Development: Acceleration of Genome-Wide Association Analysis. GigaScience. 2017;6(5):1–10.

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

31. Fishilevich S, Zimmerman S, Kohn A, Iny Stein T, Olender T, Kolker E, et al. Genic insights from integrated human proteomics in GeneCards. Database : the journal of biological databases and curation. 2016;2016.

32. 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.

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

34. 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.

35. Swanney MP, Ruppel G, Enright PL, Pedersen OF, Crapo RO, Miller MR, et al. Using the lower limit of normal for the FEV1/FVC ratio reduces the misclassification of airway obstruction. Thorax. 2008;63(12):1046–1051. doi:10.1136/thx.2008.098483.

36. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian Model Averaging: A Tutorial. Statistical Science. 1999;14(4):382–401.

37. Neighbour H, Boulet LP, Lemiere C, Sehmi R, Leigh R, Sousa AR, et al. Safety and efficacy of an oral CCR3 antagonist in patients with asthma and eosinophilic bronchitis: a randomized, placebo-controlled clinical trial. Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology. 2014;44(4):508–516.

38. Pease JE, Horuk R. Recent progress in the development of antagonists to the chemokine receptors CCR3 and CCR4. Expert opinion on drug discovery. 2014;9(5):467–483.

39. 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.

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

41. RN W, T A, FA M. The pathophysiology and treatment of glaucoma: A review. JAMA. 2014;311(18):1901–1911. doi:10.1001/jama.2014.3192.

42. Franco S, Lira M. Biomechanical properties of the cornea measured by the Ocular Response Analyzer and their association with intraocular pressure and the central corneal curvature. Clinical and Experimental Optometry. 2009;92(6):469–475. doi:10.1111/j.1444-0938.2009.00414.x.

43. Mansouri K, Leite MT, Weinreb RN, Tafreshi A, Zangwill LM, Medeiros FA. Association Between Corneal Biomechanical Properties and Glaucoma Severity. American Journal of Ophthalmology. 2012;153(3):419 – 427.e1. doi:https://doi.org/10.1016/j.ajo.2011.08.022.

44. Grise-Dulac A, Saad A, Abitbol O, Febbraro JL, Azan E, Moulin-Tyrode C, et al. Assessment of corneal biomechanical properties in normal tension glaucoma and comparison with open-angle glaucoma, ocular hypertension, and normal eyes. Journal of glaucoma. 2012;21(7):486–489.

45. Comes N, Buie LK, Borrás T. Evidence for a role of angiopoietin-like 7 (ANGPTL7) in extracellular matrix formation of the human trabecular meshwork: implications for glaucoma. Genes to cells : devoted to molecular & cellular mechanisms. 2011;16(2):243–259.

46. Kuchtey J, Källberg ME, Gelatt KN, Rinkoski T, Komàromy AM, Kuchtey RW. Angiopoietin-like 7 secretion is induced by glaucoma stimuli and its concentration is elevated in glaucomatous aqueous humor. Investigative ophthalmology & visual science. 2008;49(8):3438–3448.

47. Jensen RA, Sim X, Li X, Cotch MF, Ikram MK, Holliday EG, et al. Genome-Wide Association Study of Retinopathy in Individuals without Diabetes. PLOS ONE. 2013;8(2):1–11. doi:10.1371/journal.pone.0054232.

48. Cuellar-Partida G, Springelkamp H, Lucas SEM, Yazar S, Hewitt AW, Iglesias AI, et al. WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. Human molecular genetics. 2015;24(17):5060–5068.

Figure Legends

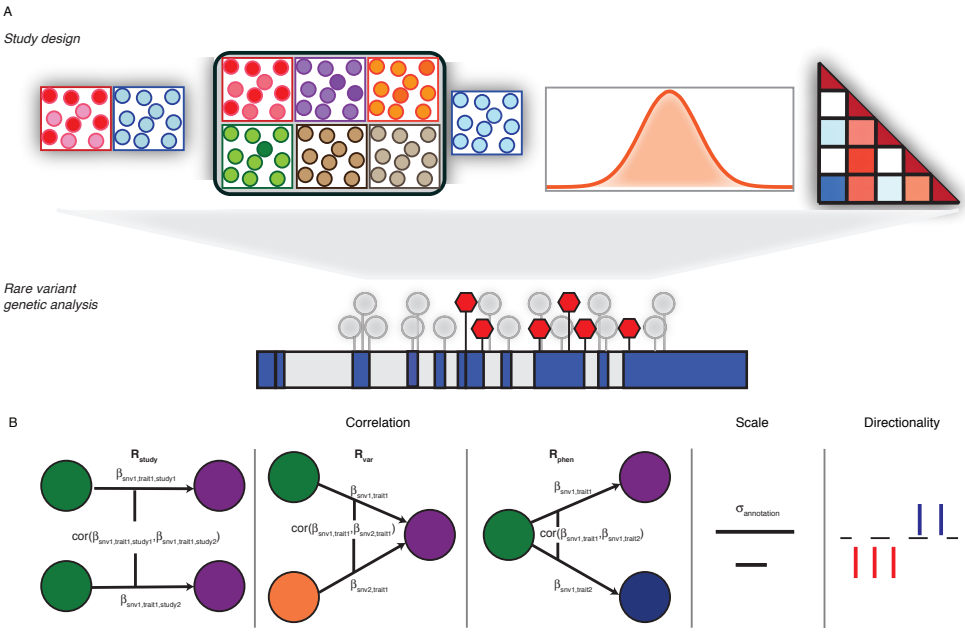


Fig 1. Schematic overview of MRP.
A: MRP is suitable for a broad range of rare variant association study designs including (from left to right): i) case-control, ii) multiple diseases with shared controls, iii) single quantitative phenotype, and iv) mixture of case-control and quantitative phenotypes. 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.

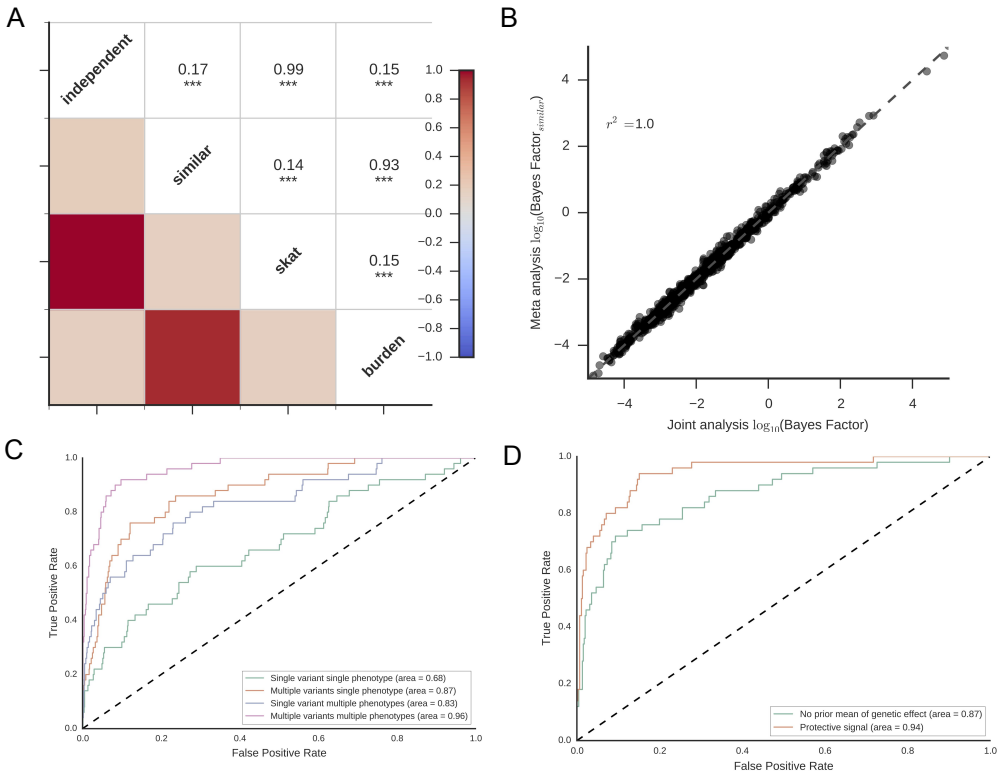


Fig 2. Simulation studies.
A: Comparison of $-\log_{10}$(p-values) from frequentist $\mathrm{BF}_{\mathrm{MRP}}$ approximation for an independent effects and a similar effects model to commonly used gene-based statistical tests (skat and burden). B: Comparison of log10(Bayes Factors) obtained when raw 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 (green) 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.

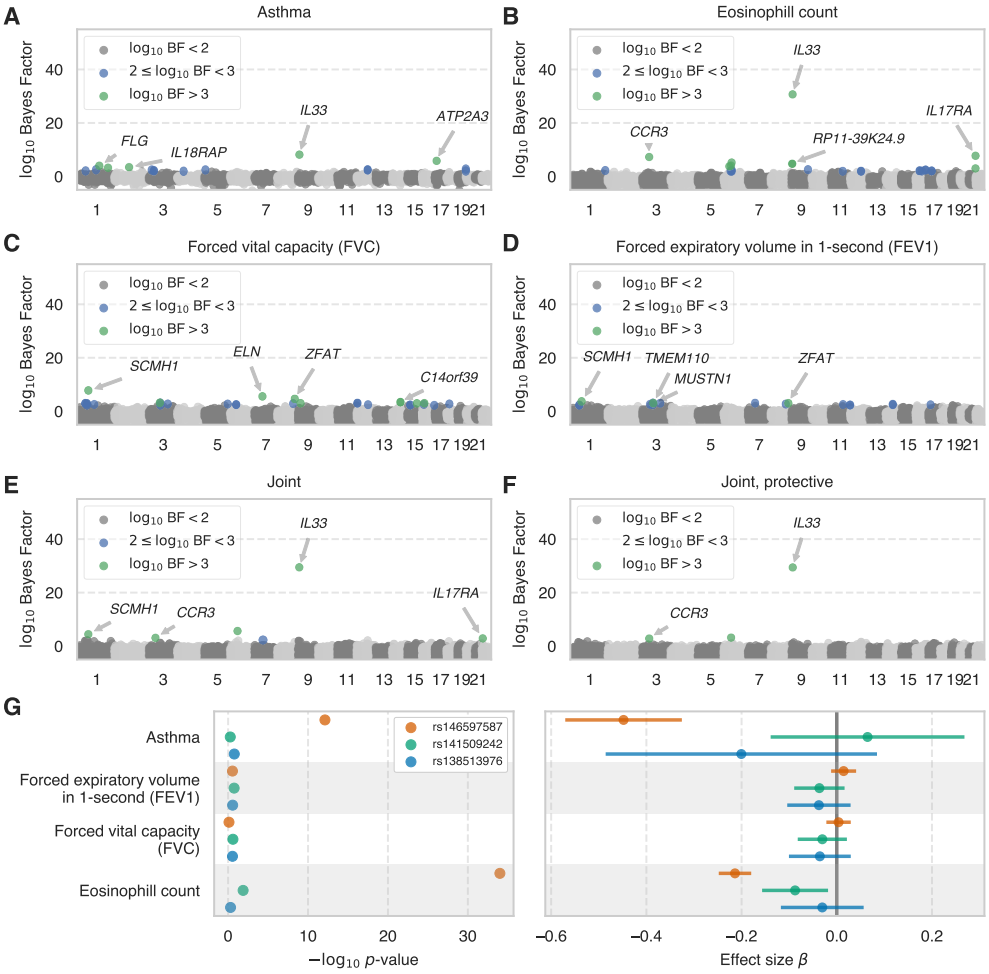


Fig 3. Results for asthma application.
$\log_{10}$ Bayes Factors from applying MRP and Bayesian model averaging to summary statistics for missense and protein-truncating variants from (A) asthma (HC382), (B) eosinophil counts (INI30150), (C) forced vital capacity (FVC, INI3062), (D) forced expiratory volume in 1-second (FEV_1, 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 greater than three 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

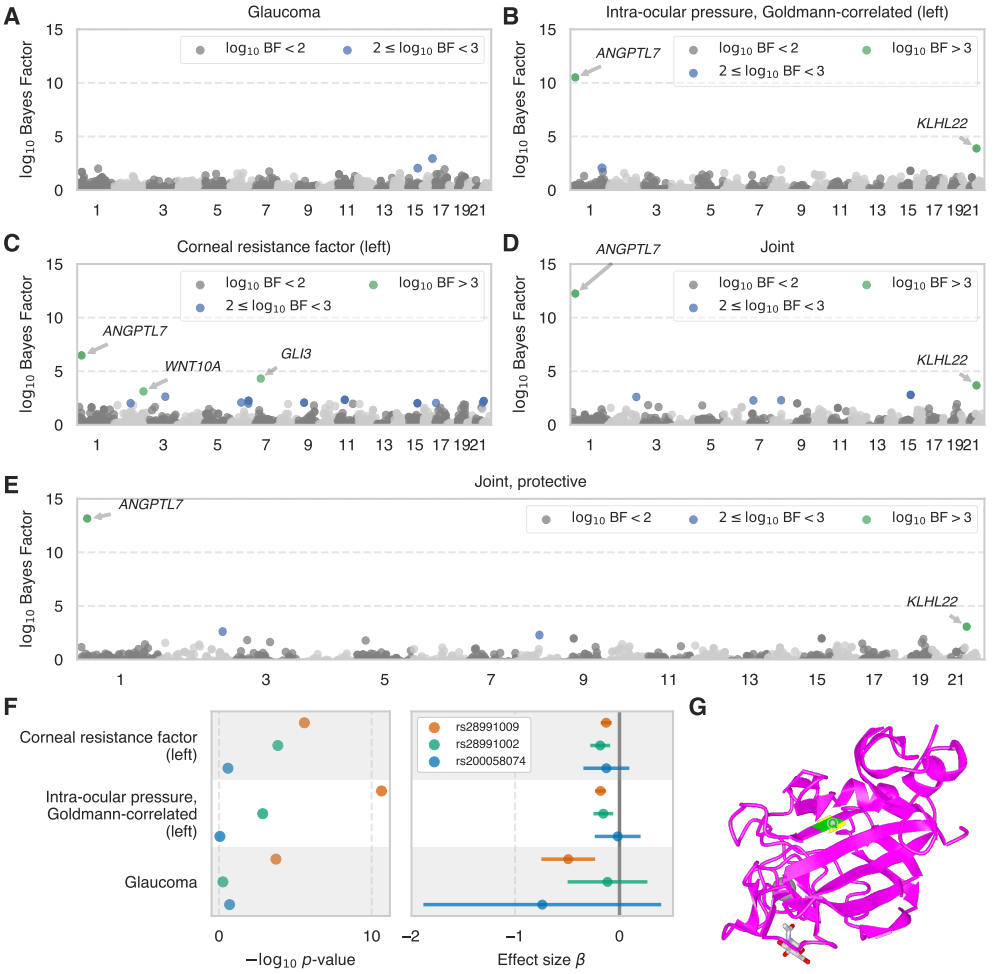
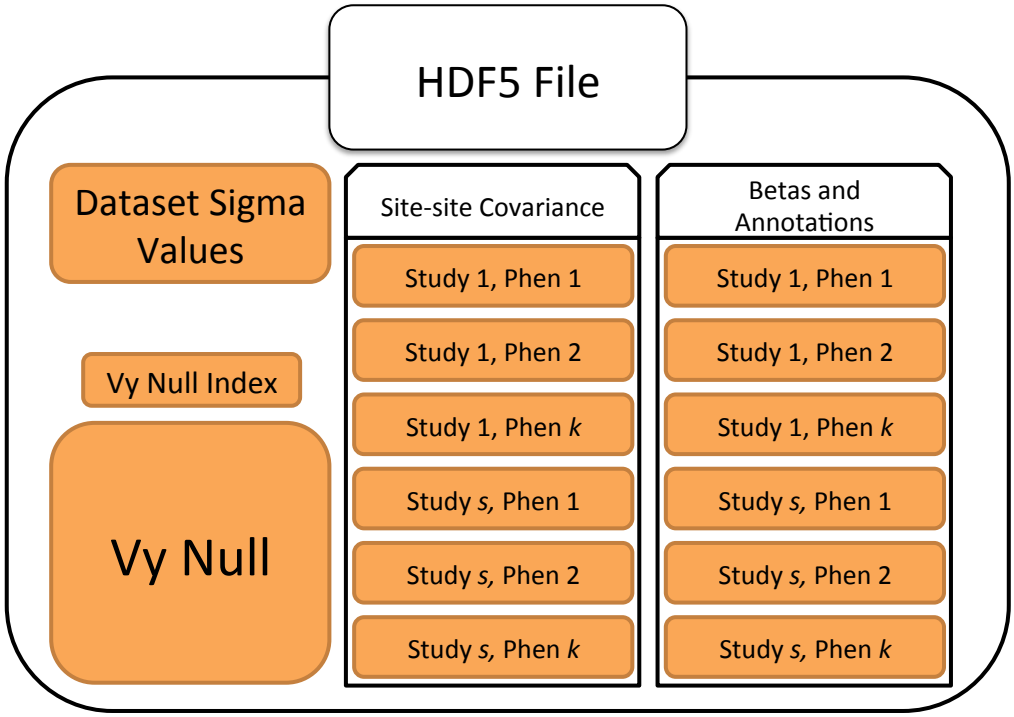


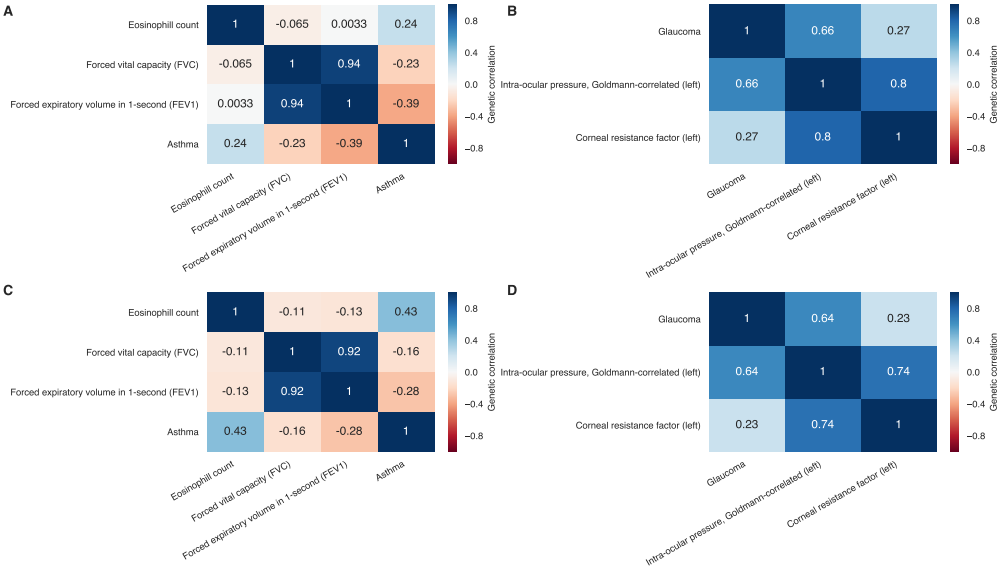
Fig 4. Results for glaucoma application.
$\log_{10}$ Bayes Factors from applying MRP and Bayesian model averaging to summary statistics for missense and protein-truncating variants from (A) glaucoma (HC276), (B) intra-ocular pressure (INI5263), (C) corneal resistance factor (INI5265), and (D) all three 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. G: Location of rs28991009 variant (green, p.Q175H, NM_021146) for the protein 1JC9_A homologous to *ANGPTL7*.

Supporting Information Legends

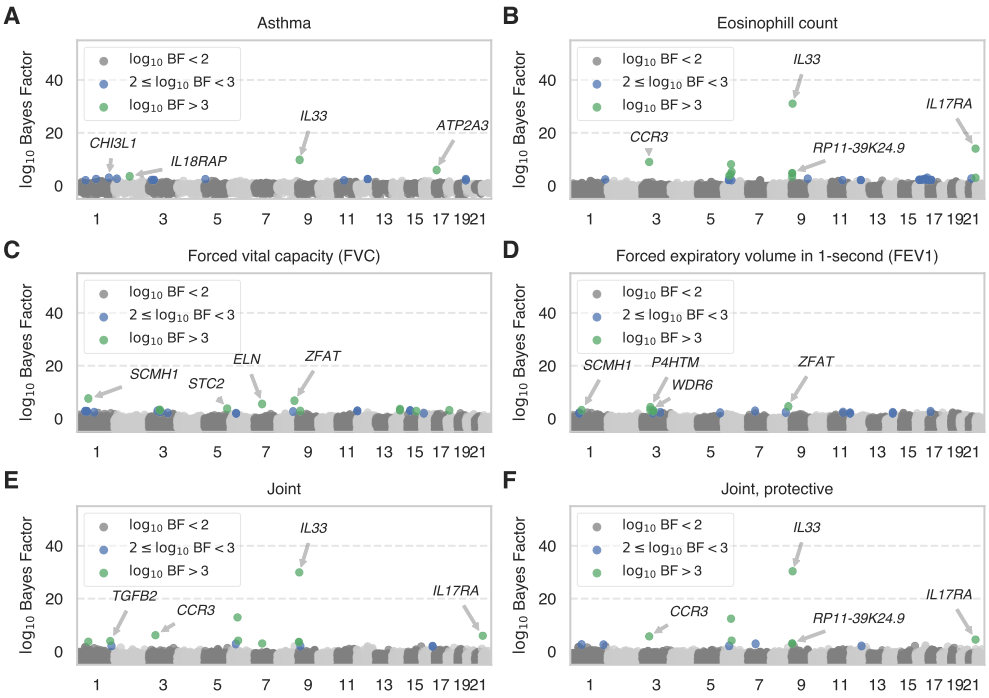
S1 Appendix. MRP model details. Specification of the MRP model including the likelihood function, priors, and Bayes factor calculation.



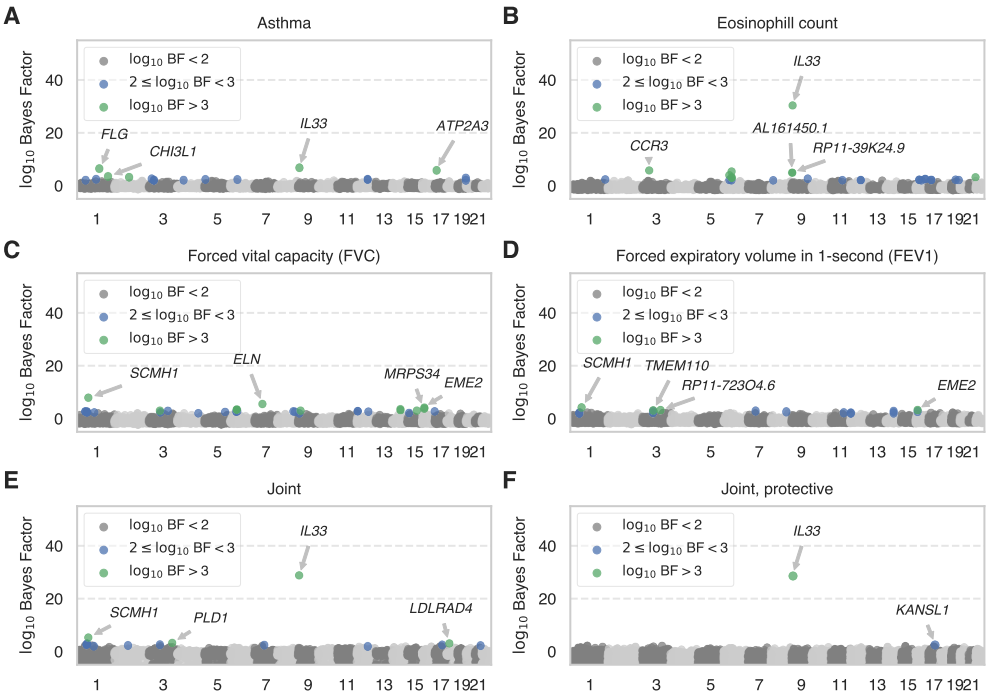
S1 Fig. HDF5 Implementation. Our HDF5 implementation contained the following components: first, a group with one table per annotation file. All effect size (beta) values and study-specific annotations were contained here, and the number of tables is limited by S (the number of studies) $\times$ K (the 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.



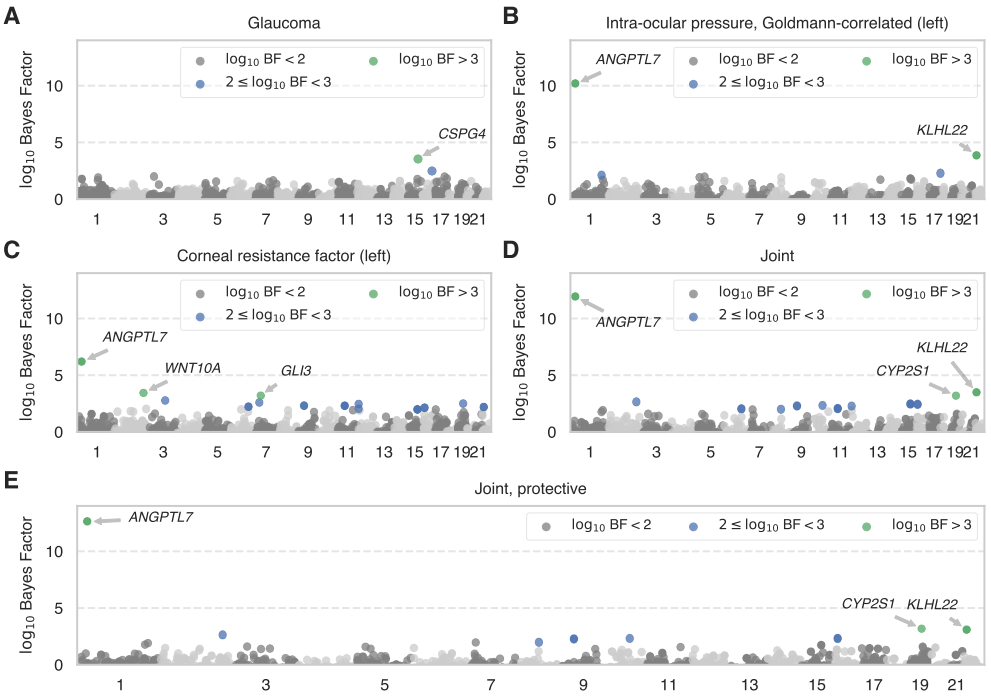
S2 Fig. Genetic correlations. Genetic correlations for (A) asthma and related phenotypes and (B) glaucoma and related phenotypes estimated using MVPMM. Genetic correlations for (C) asthma and related phenotypes and (D) glaucoma and related phenotypes estimated using LD score regression.



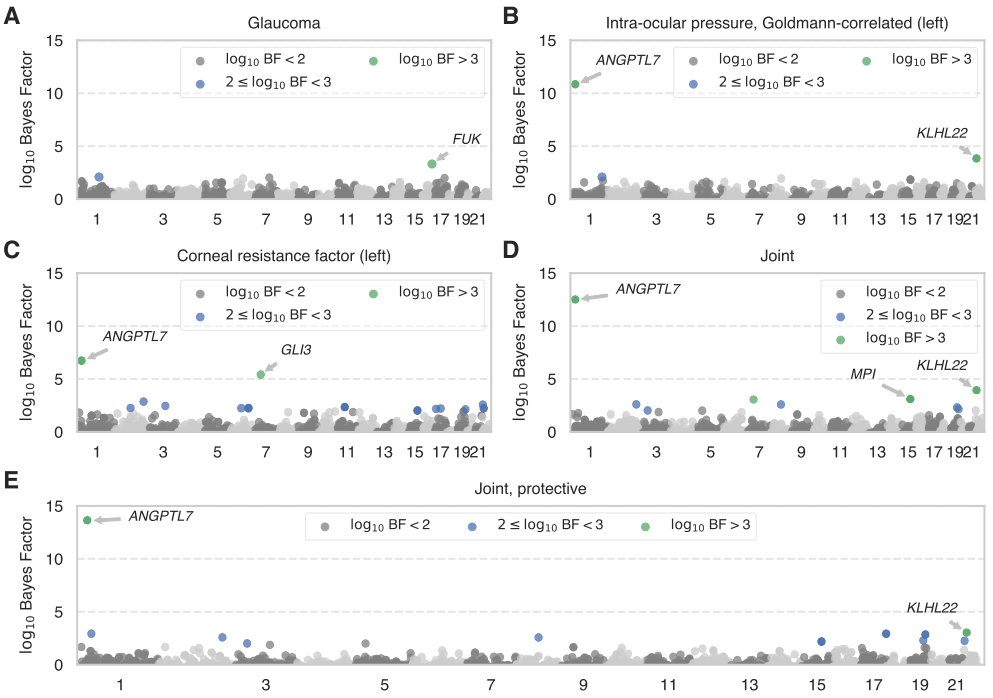
S3 Fig. Results for independent effects model applied to asthma, eosinophil counts, FEV_1, and FVC. $\log_{10}$ Bayes Factors from applying MRP independent effects model to summary statistics for missense and protein-truncating variants from (A) asthma (HC382), (B) eosinophil counts (INI30150), (C) forced vital capacity (FVC, INI3062), (D) forced expiratory volume in 1-second (FEV_1, 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 greater than three are labeled in each plot. Only $\log_{10}$ Bayes Factors greater than -5 are plotted.



S4 Fig. Results for similar effects model applied to asthma, eosinophil counts, FEV_1, and FVC. $\log_{10}$ Bayes Factors from applying MRP similar effects model 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 (FEV_1, 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 greater than three are labeled in each plot. Only $\log_{10}$ Bayes Factors greater than -5 are plotted.



S5 Fig. Results for independent effects model applied to glaucoma intra-ocular pressure, and corneal resistance factor. $\log_{10}$ Bayes Factors from applying MRP independent effects model 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.



S6 Fig. Results for similar effects model applied to glaucoma intra-ocular pressure, and corneal resistance factor. $\log_{10}$ Bayes Factors from applying MRP similar effects model 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.

Tables

Gene	Joint, protective	Joint	Eosinophil count	FVC	FEV_1	Asthma
IL33	29.4	29.3	30.6	-2.3	-2.2	8.1
CCR3	3.1	3.3	7.4	-1.4	-1.6	-0.5
RP11-39K24.9	0.8	1.8	4.9	-0.1	-0.4	0.3
SCMH1	0.5	4.7	-1.5	7.7	3.8	-0.7
MUSTN1	0.4	1.1	-1.2	2.9	2.9	-0.6
ZFAT	0.3	1.3	-2.0	4.7	3.1	-0.4
ELN	0.2	2.5	-1.0	5.6	2.9	-0.6
C14orf39	-0.7	-0.0	-1.1	3.5	2.5	0.0
TMEM110	-0.9	1.1	-1.0	3.3	3.1	-0.6
IL17RA	-4.4	3.1	7.9	-2.7	-2.5	-1.1
IL18RAP	-9.6	-0.9	-1.0	-1.6	-1.7	3.5
ATP2A3	-11.9	-0.8	-1.2	-2.1	-2.2	5.8
FLG	-20.1	-17.2	-6.0	-7.4	-8.5	4.0

S1 Table. Highlighted genes from asthma analysis. $\log_{10}$ Bayes Factors for genes highlighted in Figure 3.

Gene	Joint, protective	Joint	Glaucoma	Intra-ocular pressure, Goldmann-correlated	Corneal resistance factor
ANGPTL7	13.1	12.2	1.7	10.5	6.4
KLHL22	3.1	3.7	-0.2	3.9	2.2
WNT10A	2.6	2.6	-0.2	-0.6	3.1
GLI3	0.5	2.3	-0.4	-0.1	4.3

S2 Table. Highlighted genes from glaucoma analysis. $\log_{10}$ Bayes Factors for genes highlighted in Figure 4.