# A Bayesian method for rare variant analysis using functional annotations and its application to Autism

Shengtong Han[1,2], Nicholas Knoblauch[2], Gao Wang[2], Siming Zhao[2], Yuwen Liu[2], Yubin Xie[3], Wenhui Sheng[4], Hoang T. Nguyen[5], Xin He[2,6]

**1** Joseph J. Zilber School of Public Health, University of Wisconsin, Milwaukee, WI, USA
**2** Department of Human Genetics, University of Chicago, Chicago, IL, USA
**3** Weill Cornell Medicine, New York, USA
**4** Department of Mathematics, Statistics and Computer Science, Marquette University, Milwaukee, WI, USA
**5** Division of Psychiatric Genomics, Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
**6** Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA

* xinhe@uchicago.edu

## Abstract

Rare genetic variants make significant contributions to human diseases. Compared to common variants, rare variants have larger effect sizes and are generally free of linkage disequilibrium (LD), which makes it easier to identify causal variants. Numerous methods have been developed to analyze rare variants in a gene or region in association studies, with the goal of finding risk genes by aggregating information of all variants of a gene. These methods, however, often make unrealistic assumptions, e.g. all rare variants in a risk gene would have non-zero effects. In practice, current methods for gene-based analysis often fail to show any advantage over simple single-variant analysis. In this work, we develop a Bayesian method: MIxture model based Rare variant Analysis on GEnes (MIRAGE). MIRAGE captures the heterogeneity of variant effects by treating all variants of a gene as a mixture of risk and non-risk variants, and models the prior probabilities of being risk variants as function of external information of variants, such as allele frequencies and predicted deleterious effects. MIRAGE uses an empirical Bayes approach to estimate these prior probabilities by combining information across genes. We demonstrate in both simulations and analysis of an exome-sequencing dataset of Autism, that MIRAGE significantly outperforms current methods for rare variant analysis. In particular, the top genes identified by MIRAGE are highly enriched with known or plausible Autism risk genes. Our results highlight several novel Autism genes with high Bayesian posterior probabilities and functional connections with Autism. MIRAGE is available at https://xinhe-lab.github.io/mirage.

## Introduction

Genome-wide association studies (GWAS) have successfully identified thousands of loci associated with human complex traits [1–3]. However, in most of these loci, the causal variants and their target genes remain unknown. Additionally, most common variants (with minor allele frequency greater than 5%) discovered by GWAS have small effect sizes, modifying disease risk by less than two fold [2,3]. Sequencing studies focusing on rare variants have the potential to improve our understanding of complex diseases beyond GWAS. Because of purifying selection, deleterious variants with large effects on disease risks tend to be rare in the population, as seen in the cases of many Mendelian diseases [4–7].

Furthermore, linkage disequilibrium is much weaker for rare variants, making it less complicated to fine-map causal variants. Exome sequencing studies have particular advantages because of their relatively low costs, and the ability to directly implicate risk genes [8].

Statistical association tests for individual rare variants are usually under-powered due to their low allele frequency. This poses a significant challenge for rare variant analysis. A natural strategy is to aggregate all rare variants in a genomic region or gene, to test the collective association of the region or gene with phenotype [9]. Over the past decade, a number of methods have been proposed to perform rare variant association tests, see [8] for a review. These methods can broadly be categorized as either Burden tests or variance component tests. Burden tests collapse all rare, potentially deleterious variants in a gene, and test the association of the total frequency of these variants (burden of a gene) with phenotype [10–13]. Burden tests make the implicit assumption that aggregated variants are all risk variants of the same magnitude of effects. The variance component test, exemplified by Sequence Kernel Association Test (SKAT), relaxes the assumption of constant effect by treating variant effects as random following a normal distribution. SKAT tests if the variance of the random effect is equal to 0 [14]. Methods have also been developed as variations of these approaches [15, 16] or to combine burden test and variance component test, including SKAT-O [17].

Despite these research efforts, relatively few exome sequencing studies have identified exome-wide significant genes for complex traits. This is contrary to what researchers had expected: if rare risk variants do have large effect sizes, we ought to be able to find them even with relatively small sample sizes. The fact that this prediction is often not materialized, sometimes in large sequencing studies, suggests that rare disease variants may be less frequent and have more heterogeneous effects than what we had anticipated. Indeed, both burden tests and SKAT assume that if a gene is a risk gene, then most variants in that gene should have some effects. In reality, it is likely that most rare variants will have no effects. One way to address this challenge is to focus on variants with deleterious effects on protein functions [18]. However, most existing methods are not designed to systematically leverage functional information of variants. In practice, researchers may limit burden test or SKAT to variants that are likely to be deleterious as predicted by bioinformatic methods. However, such predictions are far from perfect, so the uncertainty and heterogeneity of variant effects are still not adequately addressed.

These limitations of current methods motivate the development of our new Bayesian statistical method to better account for heterogeneity of variant effects of a gene, and to better prioritize putative risk variants using external information. Our key idea is to model variants in a gene as a mixture of risk and non-risk variants. Each variant has a prior probability of being a risk variant, which depends on the functional annotations of the variant, e.g. conservation score, its effect on protein structure. This prior probability is generally low, reflecting sparsity of risk variants, but also varies considerably across variants based on their likely functions, reflecting heterogeneity of effects. The Bayesian strategy of incorporating functional information as prior has significant advantages over simply filtering variants based on their likely functions. In general, the external annotations have limited accuracy in predicting functional effects, so a simple filter may lose many functional variants; and conversely, many variants passing the filter may have no functional effects. For simplicity, we assume each variant belongs to one of many non-overlapping groups, with the groups defined by functional annotations. Each group has a different proportion of risk variants, with deleterious groups having higher risk proportions. To better estimate these risk proportions, we pool information across all genes being analyzed using a Bayesian hierarchical model. This strategy allows us to effectively account for uncertainty in estimating the effect of risk variants and puts more emphasis on variants with putative functional effects.

We implement our statistical ideas into a method called, MIxture model based Rare variant Analysis on GEnes (MIRAGE). The simpler version of MIRAGE tests if the proportion of risk variants in a given variant set is greater than 0 (we denote it as MIRAGE-VS). This test is a straightforward mixture model based analysis and can be used for assessing, for instance, all rare variants of a gene or a group of genes (pathway). The full version of MIRAGE, would infer risk genes from genome-wide analysis or joint analysis of a large number of genes. Using extensive simulations, we demonstrate that both MIRAGE-VS and full MIRAGE are significantly more powerful than existing methods of gene or pathway association tests. We then applied MIRAGE to a exome sequencing study of autism spectrum disorder (ASD). While standard burden test and SKAT-O identify no signals in this dataset, the top genes of MIRAGE are highly enriched with putative ASD risk genes.

# Results

## Overview of MIRAGE

MIRAGE is designed to analyze rare variant data from case-control studies. It requires only the counts of each variant in cases and controls respectively, assuming that cases and controls are well-matched. An important case is pedigree sequencing studies of parent-child trios, where transmitted and non-transmitted variants from parents to affected children would be free of population structure, and can be viewed as perfect case-control studies [19, 20]. We start by a description of the simpler MIRAGE-VS test. Often, researchers are interested in testing whether a given set of variants, e.g. all rare missense variants in a gene set, are more frequent in cases than in controls [21, 22]. Such variant burden may suggest that at least some of the genes in the gene set are associated with the disease risk. This is especially important when no individual genes pass the threshold in rare variant association tests, which is often the case in exome sequencing studies. Statistical testing of variant sets is often accomplished by so-called Burden test, such as Fisher's exact test, which compares variant counts in cases vs. controls. MIRAGE-VS takes a different approach to variant set analysis. It is motivated by the observation that, if the proportion of risk variants is low, we may not see significant difference in the total variant counts between cases or controls. Thus MIRAGE-VS explicitly models all variants in the input set as a mixture of risk and non-risk variants, and tests if the fraction of risk variants is greater than 0 (Figure 1A).

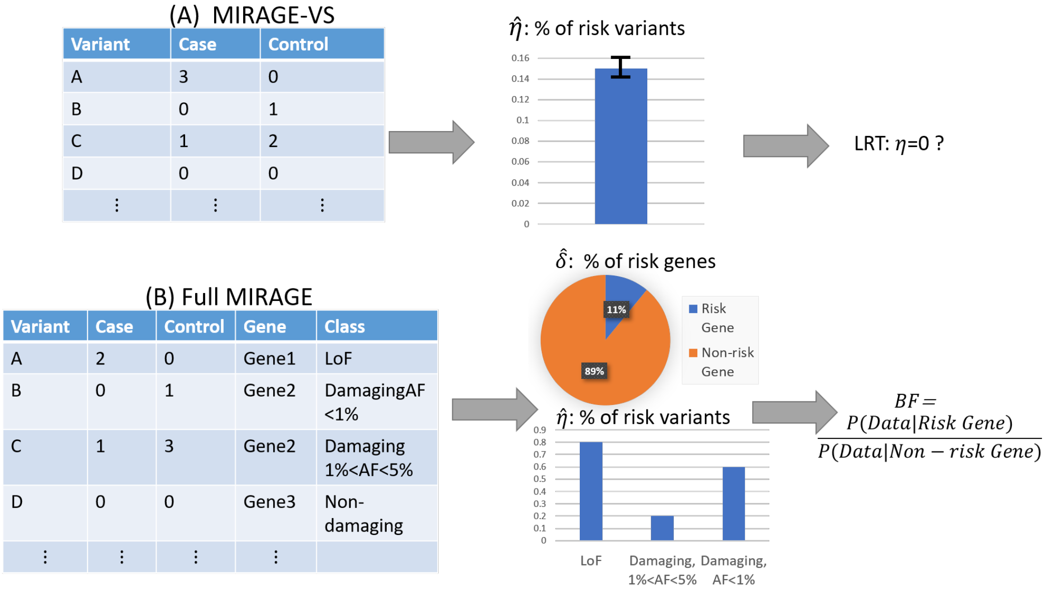
Specifically, for the $j$-th variant, let $X_j$ and $T_j$ be its allele count in cases and total allele counts in cases and controls, respectively. If $j$ is not a risk variant, $X_j$ given $T_j$ follows binomial distribution with probability determined by case and control sample sizes ($N_1$ and $N_0$). If $j$ is a risk variant, $X_j$ conditioned on $T_j$ follows binomial distribution, with the parameter determined by both the effect of the variant and sample sizes. Let $Z_j$ be an indicator of whether $j$ is a risk variant, and we denote the prior probability that $j$ is a risk variant as $P(Z_j) = \eta$. We can write the model as:

$$X_j|T_j, Z_j = 0 \sim \text{Bin}\left(T_j, \frac{N_1}{N_1 + N_0}\right) \qquad X_j|T_j, Z_j = 1 \sim \text{Bin}\left(T_j, \frac{\gamma_j N_1}{\gamma_j N_1 + N_0}\right), \tag{1}$$

where $\gamma_j$ is the relative risk of variant $j$ ($\gamma_j > 1$ for disease predisposing variants), modelled as a Gamma distribution with mean $\bar{\gamma}$. Our model defines a likelihood function of $\eta$, and we use the Expectation Maximization (EM) algorithm [23] to estimate $\eta$ given variant counts. We then test if $\eta = 0$ using likelihood ratio test (Figure 1A). We note that the hyperprior parameter $\bar{\gamma}$ is not estimated, but treated as a user-defined parameter. In our simulation and analysis, we use values between 3 and 7, informed from analysis of exome sequencing studies [19, 24], though simulations show that MIRAGE-VS is quite robust to the exact values of $\bar{\gamma}$.

The full MIRAGE differs from MIRAGE-VS in two ways: first, a gene may consist of variants from multiple functional categories, e.g. rare loss-of-function (LoF) variants or conserved missense variants. The proportions of risk variants in these categories may vary substantially. Second, it is difficult to estimate the values of $\eta$ for all categories using data from a single gene because of sparsity of some of these categories (e.g. a gene may have a single rare LoF variant). This motivates a hierarchical modeling strategy, as we described below.

The input data of MIRAGE consist of case and control counts of all rare variants across all genes being analyzed (Figure 1B), which could be all genes in the genome, or a subset of genes believed to be enriched with disease susceptibility genes. In addition, we have functional features of these variants such as predicted damaging effects, allele frequencies (AFs) in a large reference cohort, and evolutionary conservation. We assume we can define disjoint categories of variants, which may be formed by combining multiple features. MIRAGE models all input genes as a mixture of risk and non-risk genes, with the proportion of risk genes $\delta$. For non-risk genes, all the variants by definition are non-risk variants, regardless of functional features, and their counts follow the binomial model defined above for



| Variant | Case | Control |
|---|---|---|
| A | 3 | 0 |
| B | 0 | 1 |
| C | 1 | 2 |
| D | 0 | 0 |
| ⋮ | ⋮ | ⋮ |

| Variant | Case | Control | Gene | Class |
|---|---|---|---|---|
| A | 2 | 0 | Gene1 | LoF |
| B | 0 | 1 | Gene2 | DamagingAF <1% |
| C | 1 | 3 | Gene2 | Damaging 1%<AF<5% |
| D | 0 | 0 | Gene3 | Non-damaging |
| ⋮ | ⋮ | ⋮ | ⋮ | |

**Fig 1.** Work flow of MIRAGE.(A) MIRAGE-VS for variant set analysis. The method test if the percent of risk variants is equal to 0. (B) MIRAGE for identifying risk genes. It estimates the percent of risk variants in risk genes in each variant class, and use these values to compute the Bayes factor (BF) of all genes.

non-risk variants. For a risk gene, any of its variants has a prior probability of being a risk variant, with the probability equal to $\eta_c$ if the variant belongs to the category $c$. We assume $\eta_c$ are shared among variants in the category $c$ of all risk genes. The full details are described in Methods.

MIRAGE first estimates the parameters, including $\delta$ and $\eta_c$ for all categories, by maximizing likelihood over the entire dataset of all genes (Figure 1B). Then for each gene, it assesses its evidence of association with the phenotype using all its variants by computing its Bayes factor (BF). BF is similar to likelihood ratio test, comparing the null model (non-risk gene) and the alternative model (risk gene). BF of a gene naturally combines the evidence of all its variants, with larger contributions from more functionally important categories (those with larger values of $\eta_c$). Multiple testing is controlled by a Bayesian False Discovery Rate (FDR) approach [25].

## MIRAGE-VS improves variant set analysis in simulations

We first use simulations to assess the performance of MIRAGE in detecting the presence of risk variants, in a given variant set, mimicking the gene set analysis commonly used in practice. We simulate case-control data of a mixture of risk and non-risk variants. The count of a variant in controls follows Poisson distribution, with the rate depending on sample size and baseline allele frequency. For non-risk variants, their rates in cases follow the same distribution (adjusting for sample size). For a risk variant, its count in cases also follows Poisson distribution, but the rate would be generally higher. In our simulations, the relative risk of a variant is treated as random, and is sampled from a common distribution shared among all risk variants, with mean relative risk $\bar{\gamma} > 1$. Most often a variant increases the risk, but a small percent of variants may be protective. We choose the values of $\bar{\gamma}$, in the range of 3 to 7, that are informed by empirical rare variant studies, in particular exome sequencing studies of ASD [19, 26]. We vary the sample size $N_1 = N_0$, the proportion of risk variants $\eta$, and the mean relative risk $\bar{\gamma}$ in simulations (see Methods for details), and assess the performance of MIRAGE in estimating and testing if $\eta > 0$.

We confirm that MIRAGE-VS is able to accurately estimate $\eta$ under various values of sample sizes, ranging from 1000 to 5000 and $\bar{\gamma}$ (Suppl. Figure 8). We next assess type I error of MIRAGE, by

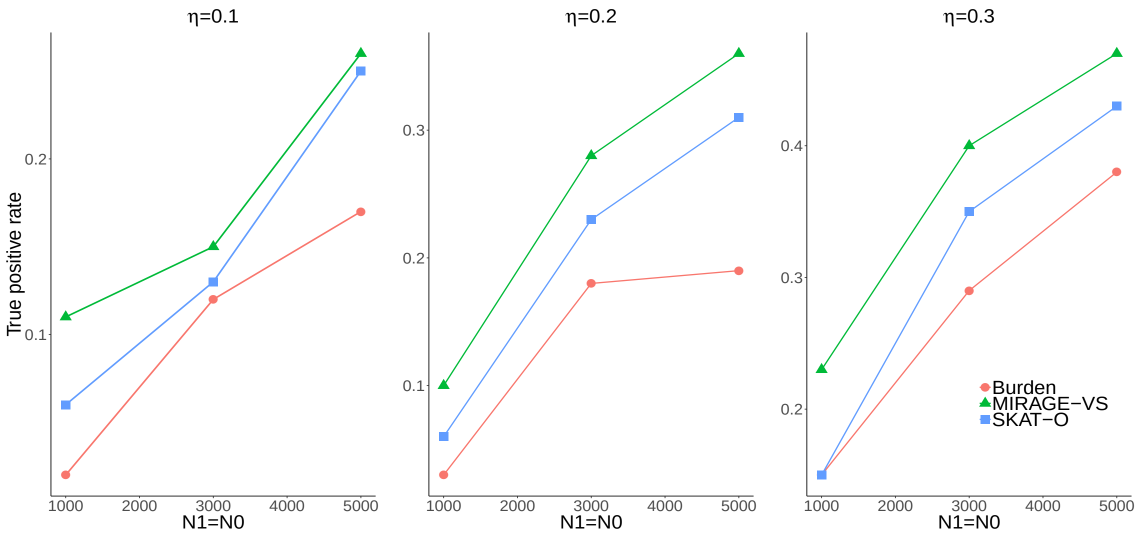


**Fig 2.** Comparison of power at $p < 0.05$, of different methods for variant set analysis. $\eta$ is the proportion of risk variants in simulations.

performing 1000 simulations under $\eta = 0$ (i.e. no risk variants) and varying sample sizes. For comparison, we run MIRAGE-VS, Burden test (Fisher's exact test) and SKAT-O on each simulated variant set. We found that all three methods effectively control type I error at $p < 0.05$, with MIRAGE-VS being the most conservative (Suppl. Figure 9). To compare the power of the three methods, we generated simulated data under $\eta = 0.1, 0.2, 0.3$, with other parameters the same as before. As we expect, the power at $p < 0.05$ increases at larger sample sizes and $\eta$. In all parameter settings, MIRAGE-VS has significantly higher power than burden test and SKAT-O (Figure 2).

In simulations above, we assumed that MIRAGE-VS knows the true value of $\bar{\gamma}$, the prior mean of relative risk, used in generating the simulated data. To evaluate the effect of mis-specified $\bar{\gamma}$, we performed sensitivity analysis in the studies of both type I error and power. We set $\bar{\gamma} = 5$ in simulations, but used $\bar{\gamma} = 3, 4, 5, 6$ in MIRAGE-VS. Type I error of MIRAGE-VS is robust to mis-specified values of $\bar{\gamma}$ (Figure 10). Similarly, the power of MIRAGE-VS does not vary significantly with the value of $\bar{\gamma}$, and remains higher than Burden test and SKAT-O in all settings even if it uses mis-specified value of $\bar{\gamma}$ (Figure 11). The robustness of MIRAGE to $\bar{\gamma}$ is perhaps not surprising, as the value is only used in specifying the prior distributions of variant effects. Indeed, the distributions under different values of $\bar{\gamma}$ overlap significantly.

## MIRAGE is more powerful in identifying risk genes than existing methods in simulations

We next perform simulations that mimic a real exome sequencing study with the goal of identifying specific risk genes. We fix sample sizes at $N_1 = N_0 = 3000$. We simulate data of 1000 genes, with the proportion of risk gene, $\delta$, varying from 0.01, 0.05, 0.1, to 0.2. For simplicity, we assume every gene has the same number of variants (100, however, the actual number may be smaller because a variant with count 0 and 0 in cases and controls from simulations will be filtered). For a risk gene, its variant belongs to one of three categories with the proportions 60%, 30%, 10%, respectively. These categories mimic, roughly, synonymous and benign missense variants, damaging missense variants and LoF variants. The prior mean relative risk, $\bar{\gamma}$, is set at 3 for the first two categories and 5 for the last category. The proportions of risk variants, $\eta_c$, are 0.05, 0.2 and 0.5 for the three categories, respectively.

When running MIRAGE (full version), we first estimate the model parameters including the proportion of risk genes $\delta$ and $\eta_c$ for each of the three categories using the EM algorithm applied to the entire dataset. We then compute the BF of each gene. We compared MIRAGE to SAKT-O and several

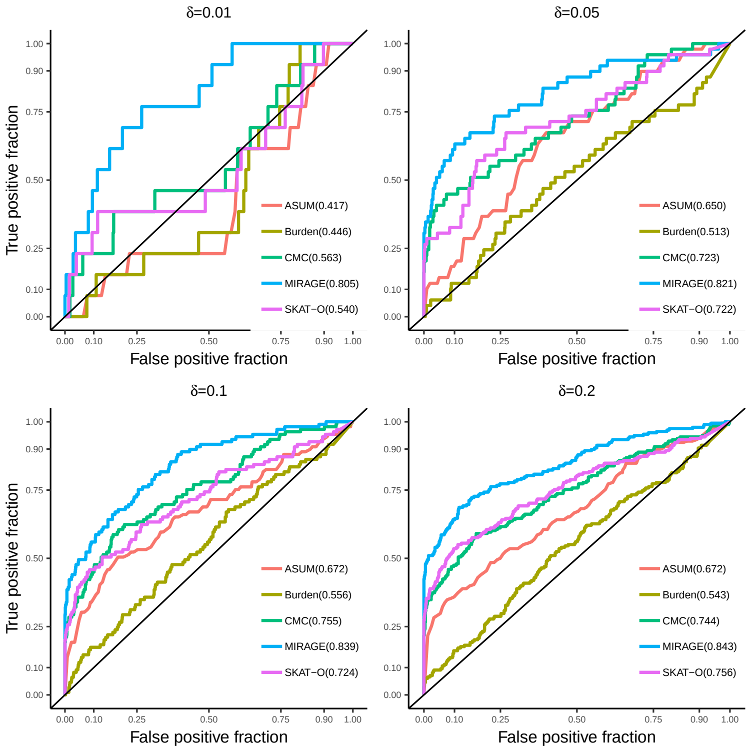


**Fig 3.** ROC curves of different methods for classifying risk genes. We simulate 1000 genes with varying proportion of risk genes, $\delta = 0.01, 0.05, 0.1, 0.2$. AUC values are shown in the bracket. Solid black reference line is in the diagonal.

variations of burden test, including Burden (baseline version), CMC [10] and ASUM [27]. In practice, Burden test is often applied to different categories of variants of a gene separately to increase the power. We thus consider two other versions of burden test as well, including Burden-adj which tests each of the three categories separately, and returns the minimum $p$ value of three tests; and Burden-combine which combines three $p$ values by Fisher's method. The results of these two tests in simulations, however, are very similar to the baseline Burden test (Suppl. Figure 12), so we consider only the baseline version here.

We compare the performance of the methods in distinguishing risk from non-risk genes, using the ROC curves (Figure 3). When $\delta = 0.01$, all methods except MIRAGE behaves close to random guesses, while MIRAGE works well with AUC about 0.8. At larger values of $\delta$, all methods perform better, but MIRAGE still significantly outperforms all other methods. SKAT-O and CMC are ranked next, with similar performance in terms of AUC. These results thus demonstrate the advantages of MIRAGE, in treating variants as mixture of risk and non-risk variants, and in taking into account the functional importance of variants.

In practice, when applying MIRAGE for gene discovery, it would be desirable to control the false discovery rate (FDR). MIRAGE does this by using a Bayesian FDR approach that converts BFs to posterior probabilities. We thus perform additional simulations to assess if the Bayesian FDR is calibrated, and whether the FDR is sensitive to mis-specification of $\bar{\gamma}$. To make simulations simpler, we run similar simulations as before, but use a common value of $\bar{\gamma}$ for all three variant categories. For each true value of $\bar{\gamma}$, ranging from 3 to 6, we used the true value and three mis-specified values of $\bar{\gamma}$ in MIRAGE, and computed Bayesian FDR. Our results show that Bayesian FDR are generally close to true FDR and only slightly inflated when true $\bar{\gamma}$ is large (greater than 5) (Suppl. Figure 13).

## MIRAGE-VS identifies variant sets associated with ASD

We applied MIRAGE to whole exome sequencing (WES) data of 4315 trios of parents and children affected with ASD. Following a method we developed earlier for analyzing trio-sequencing data, we treat transmitted alleles as “case” and non-transmitted ones as “controls” [19]. Risk variants are expected to be transmitted more often than expected (1/2 by chance). We note that the transmission data naturally avoids population structure that may confound case-control comparison. We consider only rare variants with allele frequency (AF) below 5% in our analysis. Additionally, we filter all synonymous variants from analysis except those close to exon-intron boundaries.

We first annotate the functional features of variants using ANNOVAR [28]. We identify loss-of-function (LoF) variants as the union of stop loss, stop gain, frameshift indels and splice site substitutions. For missense variants, we use PolyPhen, CADD and SIFT to define likely deleterious variants (PolyPhen score greater than 0.957, CADD score top 10% or SIFT score $< 0.05$) [29–31]. For comparison, we also include “non-damaging” variants according to PolyPhen (score less than 0.957), as a variant annotation. Since AFs of variants are highly informative of deleteriousness of variant effects [32], we also stratify variants by their AFs in ExAC [33].

We perform variant set analysis using MIRAGE. In addition to the variant level features described above, we have gene-level features for variants. We use 10 gene sets that have been implicated as potentially involved in ASD. Combining variant annotations, MAFs and gene-level features, we define a total of $5 \times 3 \times 10 = 150$ overlapping variant sets (Figure 14).

We run MIRAGE-VS on each of the 150 variants sets, testing if the fraction of risk variants is greater than 0 (Figure 4). For comparison, we also perform the Burden test. At the Bonferroni threshold (0.05/150), MIRAGE identifies 7 significant variant sets, while burden test finds only one. At a less stringent threshold of FDR $< 0.05$, MIRAGE has an even larger advantage over Burden test (Figure 4, see Table S? for complete result). Notably, a number of significant sets from MIRAGE are missense variants, which are generally more difficult to study than LoF variants and are completely missed by the burden test (Figure 4). These results thus highlight the substantially higher sensitivity of MIRAGE-VS to identify variant sets associated with diseases, than the standard Burden test.

We observed several broad trends from the variant set results (Figure 4), largely consistent with what we expect. Most of the sets with large $\eta$ (fraction of risk variants) are LoF variants. Significant missense variant sets, in contrast, have very low fractions of risk variants, generally below 5% (Figure 4). This highlights the sparsity of risk variants, even among those deemed deleterious by bioinformatic tools. Additionally, all high confidence missense variant sets have very low AF ($< 0.1\%$), confirming the importance of using AF to prioritize risk variants. Comparing variant sets that differ only in PolyPhen annotation (damaging vs. non-damaging), we notice that the annotation generally improves statistical significance of top missense variant sets (Figure 4).

## MIRAGE identifies putative risk genes of ASD

While variant set analysis above demonstrates that MIRAGE is able to highlight some candidate variant and gene sets associated with ASD, our ultimate goal is to find specific risk genes of ASD. Given the relatively small sample size of the current study, we focus on a set of 1003 most constrained genes (top 10% by pLI scores) that are known to be enriched with ASD risk genes [34]. This allows us to enhance the signal and reduces the burden of multiple testing. For each gene in this set, we divide their variants into eight non-overlapping categories by combining functional effects and AFs. These include two LoF categories ($0.01 < AF < 0.05$ and $AF < 0.01$), and six missense categories combining two functional groups (damaging and non-damaging by PolyPhen) and three AF categories ($0.01 < AF < 0.05$, $0.001 < AF < 0.01$ and $AF < 0.001$).

We first confirm that the ASD data is challenging for current methods, even after we limit to highly constrained genes. We applied Burden test and SKAT-O on each of the 1003 genes. The QQ plots of $p$ values show that neither method is able to detect any significant gene (Figure 5A and B). Indeed, none of the genes would pass any meaningful FDR threshold (say 0.3). We also perform burden and SKAT-O tests on the subsets of variants that are likely deleterious, including LoF and damaging variants (by

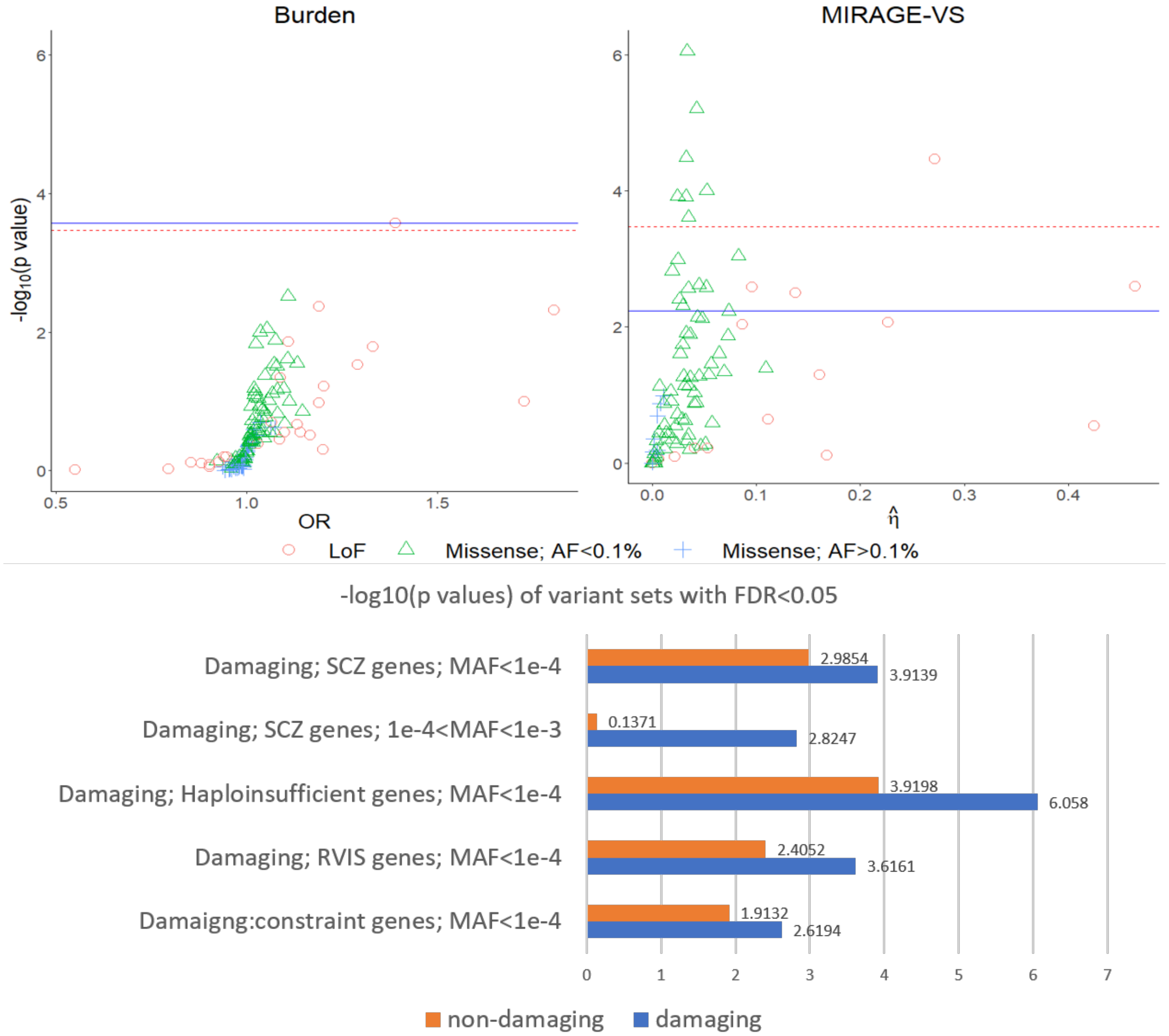


**Fig 4.** Variant set analysis by Burden and MIRAGE-VS. Red and blue horizontal lines correspond to Bonferroni and FDR thresholds at 0.05, respectively. Each variant set is colored and shaped according to its functional effects and AF.

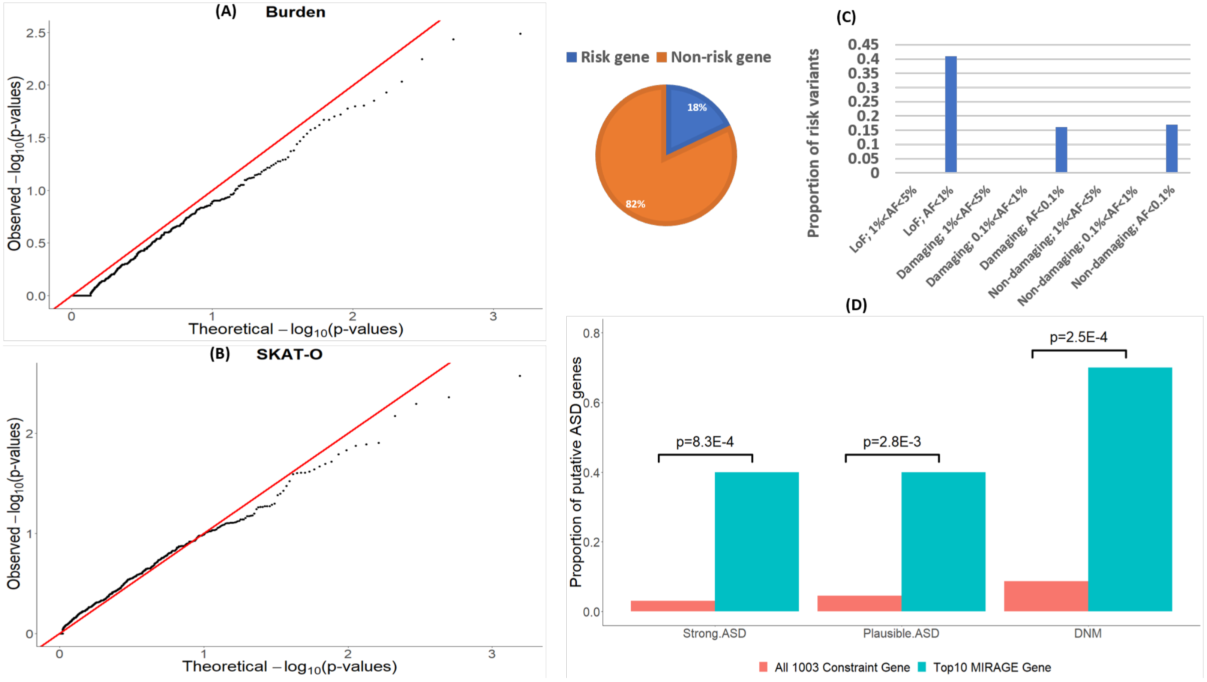


**Fig 5.** (A)QQ plot of 1003 constraint genes by Burden. (B) QQ plot of SKATO. (C) Parameter estimates: the proportion of risk genes $\widehat{\delta}$, and the proportion of risk variants in eight variant categories. (D) Enrichment of putative ASD genes in the top 10 genes found by MIRAGE. Strong ASD: score 1S and 2 by SFARI Gene; plausible ASD: score 3 by SFAIR Gene; DNM: top 1000 genes by TADA based on DNMs.

PolyPhen). The QQ plots show essentially the same pattern (Suppl. Figure 15).

We then applied MIRAGE to this gene set. MIRAGE estimates that about 18% of these genes are ASD risk genes (Figure 5C). MIRAGE also estimates the proportion of risk variants in each of the eight variant categories (Figure 5C). The category of LoF variants with low AF ($< 10^{-2}$) shows the highest proportion at 40%. And very rare missense variants ($AF < 0.001$) also show non-zero proportions at around 15%. Somewhat unexpectedly, the proportions are similar between damaging and non-damaging groups. With the estimated parameters, we calculated Bayes factor for every gene and perform Bayesian FDR control. At $FDR < 0.2$, MIRAGE is able to identify three genes. This number increases to nine at a more relaxed threshold of $FDR < 0.3$. These results thus support much higher sensitivity of MIRAGE in detecting risk genes, comparing with current methods in use.

We evaluate the findings using two sources of ASD risk genes, all independent of the data we used in this study. We focus on the top 10 genes by MIRAGE (posterior probability of risk genes $> 0.5$). The statistical evidence and supporting information are provided in Table 1. None of these genes show any evidence by Bruden and SKAT-O tests. The majority of these 10 genes are involved in ASD, according to SFARI Gene [35]. CHD8 and TRIP12 are known ASD genes (SFARI score 1S). SRCAP and CACNA1D are strong candidates of ASD (score of 2). Four other genes show suggestive evidence (score of 3), including CYFIP1, EP400, FBN1 and DYNC1H1. We found strong evidence of enrichment of SFARI ASD genes in this list, comparing with all 1003 constrained genes (Figure 5D). Even two remaining genes not curated by SFARI Gene show some connection with ASD. DOCK4 is a component of Wnt signaling, a key pathway of neurodevelopment. DOCK4 has been associated with ASD and Schizophrenia [36,37]. ABCA2 has been implicated as a candidate gene of Schizophrenia from multiple lines of evidence [38]. We also assess the connection of these genes with ASD using a list of top 1000 genes ranked by TADA using *de novo* mutations (DNMs [26]). Comparing with all 1003 genes, our list is highly enriched with DNM candidate genes (Figure 5D). These independent evidence thus strongly

**Table 1.** Top 10 genes identified by MIRAGE. BF: Bayes factor. Posterior: Bayesian posterior probability of being a risk gene.

| Gene | BF | Posterior | SKAT-O (p-value) | Burden (p-value) | SFARI score | Evidence [PMID] |
|---|---|---|---|---|---|---|
| CYFIP1 | 27.4 | 0.858 | 0.857 | 1 | 3 | FMR1 (fragile X mental retardation) interacting protein [30784587]) |
| EP400 | 15.3 | 0.771 | 0.857 | 1 | 3 | |
| FBN1 | 11.9 | 0.722 | 0.857 | 1 | 3 | |
| SRCAP | 9.8 | 0.683 | 0.857 | 1 | 2 | retarded speech development and intellectual disability [30425916] |
| DYNC1H1 | 9.5 | 0.675 | 0.863 | 1 | 3 | De novo mutation associated with cortical development [28193117] |
| ABCA2 | 8.9 | 0.662 | 0.857 | 1 | | candidate gene of Schizophrenia from multiple lines of evidence [26666178] |
| DOCK4 | 8.5 | 0.650 | 0.907 | 1 | 4 | Wnt signaling, associated with ASD, Schizophrenia, Developmental Dyslexia [23083465, 26184631] |
| CACNA1D | 6.3 | 0.579 | 0.857 | 1 | 2 | postsynaptic signaling, ASD and epilepsy gene [28472301] |
| CHD8 | 5.2 | 0.532 | 0.863 | 1 | 1S | a top ASD gene from de novo mutation data |
| TRIP12 | 4.8 | 0.513 | 0.857 | 1 | 1S | de novo mutation in this gene may cause ASD [25418537] |

supports the pathological roles of the genes identified by MIRAGE.

We highlight the result of our top gene CYFIP1 to better understand how it was found by MIRAGE (posterior probability of 0.86), but not Burden or SKAT-O ($p$ value $> 0.8$ for both). This gene has high posterior probability of being a risk gene (0.85), and is supported by multiple line of evidence from literature, including association (copy number) with Autism, Schizophrenia and Intellectual Disability, and its role in regulating synaptic activity and in mediating the function of FMR1, a well-known risk gene of a syndromic form of Autism [39–41]. The signal of CYFIP1 is largely driven by a single very rare (AF $< 0.1\%$) damaging missense variant that occurs 8 times in cases but 0 in controls (Figure 6). Other variants are mostly singletons and do not show clear enrichment in cases vs. controls. Still, MIRAGE is able to derive some evidence from the remaining variants, as the BF remains greater than 1 (2.3) even if we remove the top variant. Indeed, for the remaining variants, 25 are more common in cases than in controls, and only 18 has the opposite pattern. The results of CYFIP1 thus demonstrate two key benefits of MIRAGE: first, by borrowing information across genes via a hierarchical model, MIRAGE is able to learn to focus only on likely deleterious variants (AF $< 0.1\%$ in this case); and second, by modeling the heterogeneity of variants effects, MIRAGE allows a small number of variants to drive the results while also leveraging the collective burden that may be present in other variants.

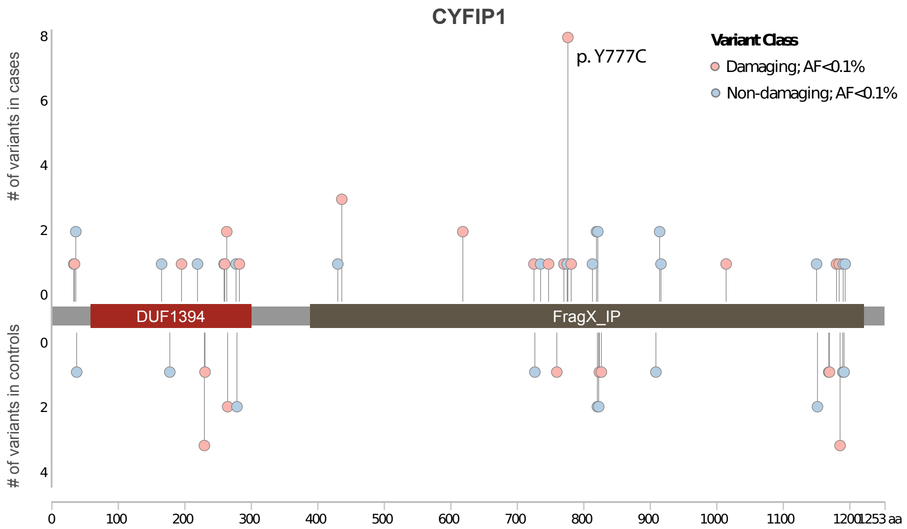


**Fig 6.** .

# Discussion

In this work, we propose a novel method, MIRAGE, for rare variant association test. Despite the importance of rare variants, current methods are not effective at extracting statistical signals from rare variants, and sequencing-based rare variant studies have achieved few successes. MIRAGE addresses two main limitations of current methods. By treating all variants as a mixture of risk and non-risk variants, it better models the heterogeneity of variant effects, particularly the sparsity of risk variants. Furthermore, it provides a framework to assess functional annotations in prioritizing deleterious variants, and to leverage these annotations in identifying risk genes. We provide two implementations of MIRAGE, MIRAGE-VS for detecting burden in variant set analysis and MIRAGE full version for identifying specific risk genes. This makes it flexible for researchers to use MIRAGE in different ways. Simulations under various scenarios confirm the effectiveness of our method. In application to a WES dataset of ASD, we find that MIRAGE-VS is much better at identifying the presence of risk variants in various gene sets than the standard burden analysis. At the level of individual genes, even though current methods fail to find any signal, the results of MIRAGE are highly enriched with ASD risk genes, confirmed by independent evidence.

How to effectively analyze rare variants is a key challenge of the field. The success of MIRAGE in the study of ASD allows us to draw some general lessons that will help address this challenge. First, the effects of rare variants are likely very heterogeneous, and this is better captured by a sparse model where most rare variants have no effects on disease risk. One can see this point clearly from our low estimated fractions of risk variants (Figures 4 and 5C), and from the analysis of individual gene (Figure 6). Our observation is contrary to a common, implicit belief in developers of rare variant association test, that in a risk gene, all rare variants tend to have some effects. Secondly, using external information of variants is critical to improve the signal to noise ratio. In particular, allele frequencies from large population reference are very helpful in separating functional from non-functional variants (Figures 4 and 5C). Variant annotation is an active area of research, and we think some recent methods, e.g. those based on deep learning, may further boost the power of MIRAGE [42]. Finally, to identify specific disease genes, it would be helpful to focus on a set of genes that are enriched with putative risk genes. In our study of ASD, we limit our analysis to 1,003 constrained genes. We suggest a possible strategy for future studies: first use MIRAGE-VS to learn gene sets with burden signals, then take the union of these genes for full MIRAGE analysis.

We find that the ASD genes we identified using inherited rare variants are often supported by independent evidence, particularly from studies using *de novo* mutations (DNMs). The implication is that allelic heterogeneity, i.e. multiple types of variants targeting the same gene, can be exploited to improve detection of risk genes. This idea has been developed in our earlier work, TADA, which combines DNMs with transmitted variants [19]. However, the transmission model of TADA is over-simplified, similar to a typical burden analysis. We think integrated analysis of multiple types of variants is particularly promising in whole-genome sequencing studies, where one can potentially combine coding and non-coding variants, single nucleotide variants (SNVs) and copy number variants (CNVs), affecting the same gene targets. We have demonstrated the feasibility of such an approach in earlier studies with DNMs [26, 43]. We think similar work in the context of whole genome association studies would be important to realize the full potential of whole genome sequencing.

MIRAGE can be further developed in several important directions. First, MIRAGE is designed for analysis of case-control (or transmission) studies. Extending it to association studies of quantitative traits would greatly broaden its applications. One possible strategy is that: instead of modeling mutation counts of individual variants, we can model some form of summary statistics, e.g. Z-scores measuring variant association with quantitative traits. This will allow us to change only the likelihood model of MIRAGE while using the same mixture prior of variant effects. Secondly, MIRAGE does not accommodate sample covariates in analysis, such as age, gender, and population ancestry. Population stratification is of particular concern as it may lead to false positive findings. It is generally difficult to address this issue in generative models such as MIRAGE. One possible strategy is to regress out all the sample covariates, and treat the residuals as quantitative traits so that we can reduce the covariate adjustment problem to the quantitative trait problem just described. Lastly, MIRAGE currently supports only disjoint functional groups as annotations. This simplifies the mathematical model, but restricts the number and types of annotations one may use. A future direction is to extend the prior model of MIRAGE so that the prior probability of a risk variant may depend on a large number of potentially overlapping annotations in a regression model. This type of prior has been successfully used in GWAS and in studies aimed at detecting functional elements in human genome (cite).

# Materials and Methods

## MIRAGE-VS model

The input data of MIRAGE-VS consist of allele counts of a group of variants in well-matched case-control samples, with sample sizes $N_1$ (cases) and $N_0$ (controls). We denote $X_j$ and $X_j^{(0)}$ the number of rare alleles of variant $j$ in cases and in controls, respectively. We also denote $T_j = X_j + X_j^{(0)}$ the total allele count. MIRAGE-VS models these variants as a mixture of risk and non-risk variants. Let $Z_j$ be an indicator of whether the variant $j$ is a risk variant ($Z_j = 1$) or not (0). $Z_j$ follows Bernoulli distribution with mean $\eta$. The goal of MIRAGE-VS is to estimate $\eta$ and test if it is equal to 0.

For a rare variant, its allele count in a set of samples can be described by a Poisson distribution. We denote $q_j$ the allele frequency of variant $j$ in controls. If $j$ is a non-risk variant ($Z_j = 0$), its allele frequency in cases would also be $q_j$. So we have:

$$X_j|Z_j = 0 \sim \text{Pois}(q_j N_1), \quad X_j^{(0)}|Z_j = 0 \sim \text{Pois}(q_j N_0). \tag{2}$$

If $j$ is a risk variant ($Z_j = 1$), its allele frequency in cases would generally be elevated. Let $\gamma_j$ be the fold increase of allele frequency. It can be interpreted as the relative risk of variant $j$, as shown in [19]. So we have:

$$X_j|Z_j = 1 \sim \text{Pois}(\gamma_j q_j N_1), \quad X_j^{(0)}|Z_j = 1 \sim \text{Pois}(q_j N_0). \tag{3}$$

It is generally difficult to estimate $\gamma_j$ for individual rare variants, so following TADA [19], we treat $\gamma_j$ as random, following Gamma$(\bar{\gamma}, \sigma)$. The hyper-parameter $\bar{\gamma}$ is the prior mean of relative risk of risk variants, and $\sigma$ is the dispersion parameter.

We note that $q_j$ is a nuisance parameter of no primary interest. So we take advantage of the property of Poisson distribution that the conditional Poisson random variable follows Binomial distribution. This

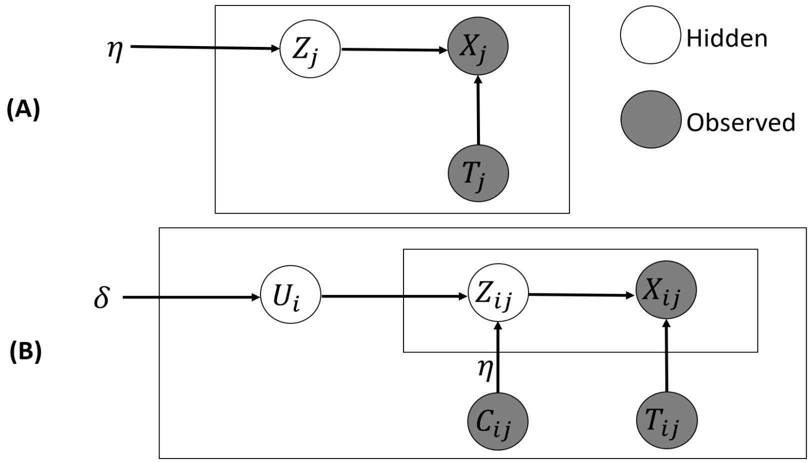


**Fig 7.** Models of MIRAGE-VS (A) and full MIRAGE (B), in the notations of probabilistic graphical models. See text for definitions of variables and parameters. (A) The box corresponds to one variant in a variant set. (B) The outer box corresponds to one gene, and the inner box one variant in a gene.

allows us to eliminate $q_j$:

$$X_j|T_j, Z_j = 0 \sim \text{Bin}\left(T_j, \frac{N_1}{N_1 + N_0}\right) \qquad X_j|T_j, Z_j = 1 \sim \text{Bin}\left(T_j, \frac{\gamma_j N_1}{\gamma_j N_1 + N_0}\right), \tag{4}$$

We marginalize $\gamma_j$ in evaluating the probability of allele counts for risk variants:

$$P(X_j|T_j, Z_j = 1) = \int \text{Bin}\left(X_j; T_j, \frac{\gamma_j N_1}{\gamma_j N_1 + N_0}\right) \text{Gamma}(\gamma_j; \bar{\gamma}, \sigma) d\gamma_j. \tag{5}$$

The full model can be shown as a probabilistic graphical model (Figure 7A). We assume the hyperparameters are given. Let $\mathbf{X}$ be the vector of $X_j$ for all variants and $\mathbf{T}$ be the vector of $T_j$'s. The likelihood function of $\eta$ is given by:

$$P(\mathbf{X}|\mathbf{T}, \eta) = \prod_j [(1-\eta)P(X_j|T_j, Z_j = 0) + \eta P(X_j|T_j, Z_j = 1)], \tag{6}$$

where we assume the variants are independent. Since we focus on variants with AF $< 5\%$, this assumption is generally valid.

The parameter estimation is performed by Expectation Maximization (EM) algorithm [23]. The details are provided in the Supplementary Notes. Once we have maximum likelihood estimation (MLE) of $\eta$, we test if $\eta = 0$ by the likelihood ratio test (LRT). The p-value of the test is determined from the $\chi^2$ distribution.

## MIRAGE model

The full MIRAGE model differs from MIRAGE-VS in two ways. First, it analyzes the data of a large number of genes (potentially the whole exome). Only a subset of these genes are risk genes. Secondly, a single risk gene may have multiple distinct variant groups, with different values of $\eta$ (proportion of risk variants). For instance, the LoF variants of a risk gene probably are more enriched with risk variants than its missense variants. We assume each variant belongs to one of multiple, disjoint, categories.

MIRAGE model is shown as a probabilistic graphical model in Figure 7B. We denote $U_i$ the indicator of whether gene $i$ is a risk gene. $U_i$ is a Bernoulli random variable with mean $\delta$. For variant $j$ of gene $i$, we denote $X_{ij}$ its rare allele count in cases and $T_{ij}$ its total allele counts in cases and controls. Each variant belongs to one variant category, denoted as $c_{ij}$ for variant $j$ of gene $i$. Similar to the

MIRAGE-VS model, we denote $Z_{ij}$ the indicator of whether variant $j$ of gene $i$ is a risk variant. The conditional distribution of $X_{ij}$ given $T_{ij}$ and $Z_{ij}$ is exactly the same as above, following Equations (4) and (5). Unlike MIRAGE-VS, the prior distribution of $Z_{ij}$ is not the same for all variants, instead, it depends on the variant category $c_{ij}$ and the gene indicator $U_i$. When $U_i = 0$ (non-risk gene), none of gene $i$'s variants would be risk variant, so $Z_{ij} = 0$ for all $j$. When $U_i = 1$ (risk gene), $Z_{ij}$ would depend on the variant category $c_{ij}$. We denote $\eta_c$ the proportion of risk variants for variant category $c$. Then $Z_{ij}$ follows Bernoulli distribution with mean $\eta_{c_{ij}}$.

Similar to the description of MIRAGE-VS model above, we denote $\mathbf{X}$ and $\mathbf{T}$ as the set of allele counts in cases, and in cases and controls combined. We also denote $C$ as the set of variant annotations $c_{ij}$'s. Our primary parameters of interest are $\delta$, the proportion of risk genes, and $\eta$, the vector of $\eta_c$'s for all variant categories. The likelihood function is given by:

$$P(\mathbf{X}|\mathbf{T}, \mathbf{C}, \delta, \eta) = \prod_i [(1-\delta)P(X_i|T_i, U_i = 0) + \delta P(X_i|T_i, C_i, U_i = 1, \eta)], \tag{7}$$

where $X_i, T_i, C_i$ are the relevant data of all variants in gene $i$. The first probability term in the equation is the likelihood of a non-risk gene, and is simply given by:

$$P(X_i|T_i, U_i = 0) = \prod_j \text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right). \tag{8}$$

The second probability term is the likelihood of a risk gene:

$$P(X_i|T_i, C_i, U_i = 1, \eta) = \prod_j \left[(1-\eta_{C_{ij}})\text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right) + \eta_{C_{ij}} P(X_{ij}|T_{ij}, Z_{ij} = 1)\right], \tag{9}$$

where $P(X_{ij}|T_{ij}, Z_{ij} = 1)$ is given by Equation (5) (adding gene index $i$ in that equation, adding index $i$ only in that single equation could be confusing, unless in the whole section, so do you want to do that?).

The parameters $\delta$ and $\eta$ are estimated by EM algorithm (see EM Algorithm in the Supplements). Given the MLE $\hat{\delta}$ and $\hat{\eta}$, we can determine the Bayes factor of a gene $i$, $B_i$, and its posterior probability of being a risk gene, $\text{PP}_i$, as:

$$B_i = \frac{P(X_i|T_i, C_i, U_i = 1, \hat{\eta})}{P(X_i|T_i, U_i = 0)}, \quad \text{PP}_i = \frac{\delta B_i}{1 - \delta + \delta B_i}. \tag{10}$$

It is easy to show that $B_i$ can be related to the evidence at the single variant level:

$$B_i = \prod_j [(1 - \eta_{C_{ij}}) + \eta_{C_{ij}} B_{ij}], \quad B_{ij} = \frac{P(X_{ij}|T_{ij}, Z_{ij} = 1)}{P(X_{ij}|T_{ij}, Z_{ij} = 0)}, \tag{11}$$

where $B_{ij}$ is the BF of variant $j$ of gene $i$. From this equation, one can see that the more deleterious variant categories with larger values of $\eta_c$ will contribute more to the gene level evidence.

Once we determine BF and posterior probability of all genes, we control for multiple testing by performing Bayesian FDR control [19].

## Simulation procedure

Simulation for MIRAGE-VS analysis: We simulate case-control counts of a variant set, for given sample sizes ($N_1$ and $N_0$ for cases and controls, respectively) and given proportion of risk variants $\eta$. For each variant in the set, we repeat the following steps. (1) We sample the risk variant status $Z_j$ for variant $j$ (1 if it is a risk variant, and 0 otherwise): $Z_j \sim \text{Bernoulli}(\eta)$. And if $Z_j = 1$, we also sample the relative risk $\gamma_j \sim \text{Gamma}(\bar{\gamma}, \sigma)$. Both $\bar{\gamma}$ and $\sigma$ are set as user-specified parameters. We use $\sigma = 1$ in the paper in all simulations. (2) We sample the allele frequency $q_j$ from a Beta distribution. If $Z_j = 0$, we sample from $\text{Beta}(\alpha_0, \beta_0)$. We set $\alpha_0 = 0.1, \beta_0 = 1000$ in our simulations. If $Z_j = 1$, we assume variants would be even rarer, so we sample from Beta($\alpha, \beta$), where $\alpha = 0.1, \beta = 2000$ (so mean AF is two times lower than non-risk variants). (3) We sample the total allele count in the data by $T_j \sim \text{Pois}(N_1 + N_0, q_j)$. It is possible that $T_j$ is 0, and such variants are filtered. (4) We split the total variant count $T_j$ into cases and controls by Binomial distribution. If $Z_j = 0$, we split the count according to sample sizes, so the probability in cases is equal to $N_1/(N_1 + N_0)$. If $Z_j = 1$, the probability in cases is $\frac{N_1 \gamma_j}{N_1 \gamma_j + N_0}$.

Simulation for MIRAGE analysis: We simulate a set of genes under given case-control sample sizes $N_1, N_0$ and the proportion of risk genes $\delta$. We assume each gene has a mixture of variants in different categories, with fixed proportions of variant categories. In the simulations of the paper, we use three categories mimicking LoF, deleterious missense variants and the rest, with fractions 10%, 30% and 60% respectively. Each category is allowed to have different mean relative risks $\bar{\gamma}$ and different proportion of risk variants, $\eta_c$ for category $c$. We use $\bar{\gamma} = 5$ for LoF and 3 for missense categories, and $\eta_c$ 0.5, 0.2 and 0.05 for the three categories, respectively. Our simulation starts with sampling the risk status for gene $i$, $U_i \sim \text{Bernoulli}(\delta)$. When $U_i = 0$, all variants would be non-risk variants. When $U_i = 1$, we sample the risk variant status $Z_{ij}$ for each variant $j$ of gene $i$. For a variant $j$ in a category $c$, its probability of being a risk variant $Z_{ij} \sim \text{Bernoulli}(\eta_c)$. Once we have sampled the risk variant status of all variants, the rest follows the same procedure above for sampling variant counts in cases and controls.

## WES data of ASD families

Transmitted variants from parents to affected children were obtained from 4,315 autism families in De Rubies et al [24], provided by Autism Sequencing Consortium.

## Annotating variants

The software package annovar was used to query the dbNSFP database of functional effect predictions. We annotate with several popular programs including PolyPhen, CADD and SIFT [29–31]. This suite of generic variant annotations was then augmented with 10 gene sets associated with a variety of neuropsychiatric traits (listed in Figure 14). High confidence ASD and moderate confidence ASD genes are defined by the q-values of TADA analysis [19, 26], using $q < 0.1$ and $0.1 \leq q < 0.3$, respectively. The other gene sets are collected from literature (with PMIDs listed in the figure). All the 10 gene lists can be found in Supplementary Table ?

## Applying MIRAGE to ASD data

For MIRAGE-VS analysis of variant sets, we analyze each variant set separately. The hyperprior parameter for relative risk, $\bar{\gamma}$, is set at 6 for LoF and 3 for missense variant sets. In the EM algorithm for estimating the parameter $\eta$, the fraction of risk variants, we randomly choose initial values, and the algorithm converges if the change of parameter estimates in two iterations is less is less than $10^{-5}$.

For MIRAGE analysis of 1003 constrained genes, we create 8 variant categories as described in Results. We use the same hyperprior parameter for LoF and missense variants, as described above. The EM algorithm is used to estimate $\delta$, the proportion of risk genes, and $\eta_c$, the percent of risk variants for each category $c$. Running of EM is similar as above. Once these parameters are estimated, their values are assumed to be known, and are used in calculating BF of each gene.

## Running other programs

We used method SKAT-O in R package *SKAT* (version 1.3.2.1) without covariates, setting method of *SKATO*. We used R package *AssotesteR* for methods CMC and ASUM, and for CMC, we used three MAF cutoffs $5 \times 10^{-6}, 2 \times 10^{-5}, 5 \times 10^{-5}$ to partition the variants. ASUM uses permutation test to get p values, and we set the number of permutations to be 100 as larger number of permutations doesn't make a big difference. For burden test, Fisher's exact test from the R function-*fisher.test* was used.

# Supporting information

**EM Algorithm** We describe the EM algorithm below for Maximum Likelihood parameter estimation of MIRAGE-VS and the full MIRAGE.

**EM for MIRAGE-VS** MIRAGE-VS maximizes the likelihood of $\eta$, as defined in Equation 6 in Methods. For simplicity of notation, we drop $T$ or $T_j$ in the probability or conditional probability terms. We assume the hyperprior parameters $\bar{\gamma}$ and $\sigma$ are given, so we also drop them in the notations. We note though it is possible to estimate the hyperparameters by ML. We denote $\theta = (\eta)$ as parameter of interest and $\theta^{(t)}$ as its value in the $t$-th iteration (this is for generality, as $\theta$ would represent a set of parameters in full MIRAGE).

We denote $B_j$ as the BF of variant $j$:

$$B_j = \frac{P(x_j|Z_j = 1)}{P(x_j|Z_j = 0)}, \tag{12}$$

where the two probabilities are given by Equations (4) and (5).

- E step: calculating the expectation of the log likelihood conditioned on the observed data and $\theta^{(t)}$.

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{Z|X,\theta^{(t)}} \log P(x, Z|\theta) = E_{Z|X,\theta^{(t)}} \log(P(Z|\theta)) + E_{Z|X,\theta^{(t)}} \log(P(x|Z, \theta)) \\ &= \sum_j [E(Z_j|x, \theta^{(t)}) \log(\pi_j(\eta)) + (1 - E(Z_j|x, \theta^{(t)})) \log(1 - \pi_j(\eta))] \\ &+ \sum_j [(1 - E(Z_j|x, \theta^{(t)})) \log(1 - \pi_j(\eta)) + E(Z_j|x, \theta^{(t)})(\log(\pi_j(\eta)) + \log B_j)]. \end{aligned}$$

where $\pi_j(\eta) = P(Z_j = 1|x_j, \eta)$, is the posterior probability that variant $j$ is a risk variant. The derivations are based on:

$$\log P(Z|\theta) = \sum_j [Z_j \log(\pi_j(\eta)) + (1 - Z_j) \log(1 - \pi_j(\eta))]$$

$$\begin{aligned} P(x|Z, \theta) &= \prod_j P(x_j|Z_j, \theta) = \prod_j \sum_{Z_j} P(x_j, Z_j|\theta) = \prod_j [\pi_j(\eta) P(x_j|Z_j = 1) + (1 - \pi_j(\eta)) P(x_j|Z_j = 0)] \\ &\propto \prod_j [I(Z_j = 1)\pi_j(\eta) \times B_j + I(Z_j = 0)(1 - \pi_j(\eta))] = \prod_j [(\pi_j(\eta) B_j^{Z_j} \times (1 - \pi_j(\eta))^{1 - Z_j}] \end{aligned}$$

Thus

$$P(x|Z, \theta) \propto \prod_j [(\pi_j(\eta) B_j)^{Z_j} \times (1 - \pi_j(\eta))^{1 - Z_j}]$$

$$\log(P(x|Z, \theta)) = c + \sum_j [Z_j(\log(\pi_j(\eta)) + \log(B_j)) + (1 - Z_j) \log(1 - \pi_j(\eta))]$$

where $B_j$ is given in (13).

$$\begin{aligned} E(Z_j|x, \theta^{(t)}) = P(Z_j = 1|x, \theta^{(t)}) &= \frac{P(x|Z_j = 1, \theta^{(t)}) P(Z_j = 1|\theta^{(t)})}{P(x|\theta^{(t)})} \\ &= \frac{P(x_j|Z_j = 1, \theta^{(t)}) P(Z_j = 1|\theta^{(t)})}{P(x_j|\theta^{(t)})} = \frac{c_j(\theta^{(t)}) B_j(\theta^t) \pi_j(\eta^{(t)}) \prod_{k \neq j} P(x_k|\theta^{(t)})}{P(x_j|\theta^{(t)}) \times \prod_{k \neq j} P(x_k|\theta^{(t)})} \\ &= \frac{c_j(\theta^{(t)}) B_j(\theta^{(t)}) \pi_j(\eta^{(t)})}{\pi_j(\eta^{(t)}) c_j(\theta^{(t)}) B_j(\theta^{(t)}) + (1 - \pi_j(\beta^{(t)})) c_j(\theta^{(t)})} = \frac{\pi_j(\eta^{(t)}) B_j(\theta^{(t)})}{\pi_j(\eta^{(t)}) B_j(\theta^{(t)}) + (1 - \pi_j(\eta^{(t)}))} \end{aligned}$$

Denote $c_j = P(x_j|Z_j = 0, \theta)$. Then $P(x_j|Z_j = 1, \theta) = c_j B_j$. $B_j$ is the Bayes factor for variant $j$.

- M step updates $\theta$ by $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(t)})$. That is

  $$\theta^{(t+1)} = arg\ max_\theta\ Q(\theta|\theta^{(t)})$$

  Then, taking derivatives leads to

  $$\frac{\partial Q}{\partial \eta} = 2\sum_j [\frac{E(Z_j|x,\theta^{(t)})}{\pi_j(\eta)} - \frac{1-E(Z_j|x,\theta^{(t)})}{1-\pi_j(\eta)}]\frac{\partial \pi_j(\eta)}{\partial \eta}$$

  $$\frac{\partial Q}{\partial \bar{\gamma}} = \sum_j \frac{E(Z_j|x,\theta^{(t)})}{B_j}\frac{\partial B_j}{\partial \bar{\gamma}}$$

  $$\frac{\partial Q}{\partial \sigma} = \sum_j \frac{E(Z_j|x,\theta^{(t)})}{B_j}\frac{\partial B_j}{\partial \sigma}$$

  Setting these equations equal to 0 yields

  $$\eta^{(t+1)} = \frac{1}{J}\sum_j E(Z_j|x,\theta^{(t)})$$

  where $J$ is the total number of variants in a variant set. For parameters $\bar{\gamma},\sigma$, there are no closed form solutions for $\bar{\gamma}^{(t+1)},\sigma^{(t+1)}$. Instead, we fix $\bar{\gamma},\sigma$.

**EM for MIRAGE** Let $\theta$ denote all parameters and $\theta^{(t)}$ be the parameter estimate at the $t$th iteration.

- E step: calculates the expectation of the log likelihood of parameter $\theta$ conditional on the observed data and $\theta^{(t)}$.

  $$\begin{aligned}
  &Q(\theta|\theta^{(t)}) = E_{U,Z|X,\theta^{(t)}} \log L(\theta|x,U,Z) = E_{U,Z|X,\theta^{(t)}} \log(P(x,U,Z|\theta))\\
  &= E_{U,Z|X,\theta^{(t)}} \log(P(U,Z|\theta)) + E_{U,Z|X,\theta^{(t)}} \log(P(x|U,Z,\theta))\\
  &= \log(\delta)\sum_i E(U_i|x,\theta^{(t)}) + \log(1-\delta)\sum_i(1-E(U_i|x,\theta^{(t)}))\\
  &+ \sum_i\sum_j[E(U_iZ_{ij}|x,\theta^{(t)})\log(\pi_{ij}(\eta)) + (E(U_i|x,\theta^{(t)}) - E(U_iZ_{ij}|x,\theta^{(t)}))\log(1-\pi_{ij}(\eta))]\\
  &+ \sum_i\sum_j[(E(U_i|x,\theta^{(t)}) - E(U_iZ_{ij}|x,\theta^{(t)}))\log(1-\pi_{ij}(\eta)) + E(U_iZ_{ij}|x,\theta^{(t)})(\log(\pi_{ij}(\eta)) + \log(B_{ij}))]
  \end{aligned}$$

  where $\delta = P(U_i = 1|\theta)$. The last equation holds because

  $$\begin{aligned}
  &\log(P(U,Z|\theta)) = \log(P(U|\theta)\times P(Z|U=1,\theta))\\
  &= \log(\prod_i \delta^{U_i}(1-\delta)^{1-U_i}) + \log[\prod_i\prod_j \pi_{ij}(\eta)^{Z_{ij}}(1-\pi_{ij}(\eta))^{1-Z_{ij}}]^{U_i}\\
  &= \sum_i U_i\log(\delta) + \sum_i(1-U_i)\log(1-\delta) + \sum_i\sum_j[U_iZ_{ij}\log(\pi_{ij}(\eta)) + (U_i - U_iZ_{ij})\log(1-\pi_{ij}(\eta))]
  \end{aligned}$$

  and

  $$P(x|U,Z,\theta) = \prod_i\sum_{U_i} P(x_i|U,\theta) = \prod_i[I(U_i=0)P(x_i|U_i=0,\theta) + I(U_i=1)P(x_i|U_i=1,\theta)]$$

Denote $d_i = P(x_i|U_i = 0, \theta)$. Hence gene $i$ is non-risk gene and $d_i$ should be Binomial probability with $\gamma = 1$, then $P(x_i|U_i = 1, Z_i, \theta) = d_i \times \frac{P(x_i|U_i=1,\theta)}{P(x_i|U_i=0,\theta)} = d_i \times B_i$. Note at gene level, $Z_{ij}$ is useless. $B_i$ is the Bayes factor for gene $i$. Thus

$$P(x|U, Z, \theta) \propto \prod_i [P(x_i|U_i = 1, Z_{ij}, \theta)]^{U_i}$$

When gene $i$ is a causal gene, i.e. $U_i = 1$, the likelihood can be decomposed further into variant level

$$\begin{aligned}
P(x_i|U_i = 1, Z_{ij}, \theta) &= \prod_j P(x_{ij}|U_i = 1, Z_{ij}, \theta) = \prod_j \sum_{Z_{ij}} P(x_{ij}, Z_{ij}|U_i = 1, \theta) \\
&= \prod_j [\pi_{ij}(\eta)P(x_{ij}|U_i = 1, Z_{ij} = 1) + (1 - \pi_{ij}(\eta))P(x_{ij}|U_i = 1, Z_{ij} = 0)] \\
&= \prod_j P(x_{ij}|U_i = 1, Z_{ij} = 0)[I(Z_{ij} = 1)\pi_{ij}(\eta) \times B_{ij} + I(Z_{ij=0})(1 - \pi_{ij}(\eta))] \\
&\propto \prod_j [I(Z_{ij} = 1)\pi_{ij}(\eta) \times B_{ij} + I(Z_{ij} = 0)(1 - \pi_{ij}(\eta))] \\
&= \prod_j [(\pi_{ij}(\eta)B_{ij})^{Z_{ij}} \times (1 - \pi_{ij}(\eta))^{1-Z_{ij}}]
\end{aligned}$$

where $B_{ij}$ is the Bayes factor for $j$-th variant in $i$-th gene and $\pi_{ij}(\eta) = P(Z_{ij} = 1|U_i = 1, \eta)$. Combining together across all the genes yields

$$P(x|U, Z, \theta) \propto \prod_i \Big\{ \prod_j [(\pi_{ij}(\eta)B_{ij})^{Z_{ij}} \times (1 - \pi_{ij}(\eta))^{1-Z_{ij}}] \Big\}^{U_i}$$

$$\begin{aligned}
\log(P(x|U, Z, \theta)) = c + \sum_i \sum_j [&(U_i Z_{ij})(\log(\pi_{ij}(\eta)) + \log(B_{ij})) \\
&+ (U_i - U_i Z_{ij}) \log(1 - \pi_{ij}(\eta))]
\end{aligned}$$

$c$ is a constant free of parameters. Now look at expectation terms,

$$\begin{aligned}
E(U_i|x, \theta^{(t)}) = P(U_i = 1|x, \theta^{(t)}) &= \frac{P(x, U_i = 1|\theta^{(t)})}{P(x|\theta^{(t)})} \\
&= \frac{P(x|U_i = 1, \theta^{(t)})P(U_i = 1|\theta^{(t)})}{P(x|U_i = 1, \theta^{(t)})P(U_i = 0|\theta^{(t)}) + P(x|U_i = 0, \theta^{(t)})P(U_i = 0|\theta^{(t)})} \\
&= \frac{\delta^{(t)} P(x_i|U_i = 1, \theta^{(t)})}{\delta^{(t)} P(x_i|U_i = 1, \theta^{(t)}) + (1 - \delta^{(t)})P(x_i|U_i = 0, \theta^{(t)})} = \frac{\delta^{(t)} B_i(\theta^{(t)})}{\delta^{(t)} B_i(\theta^{(t)}) + (1 - \delta^{(t)})}
\end{aligned}$$

$$
\begin{aligned}
&E(U_i Z_{ij}|x, \theta^{(t)}) = P(U_i = 1, Z_{ij} = 1|x, \theta^{(t)}) \\
&= \frac{P(x|Z_{ij} = 1, U_i = 1, \theta^{(t)})P(Z_{ij} = 1|U_i = 1, \theta^{(t)})P(U_i = 1|\theta^{(t)})}{P(x|\theta^{(t)})} \\
&= \frac{P(x_i|Z_{ij} = 1, U_i = 1, \theta^{(t)})P(Z_{ij} = 1|U_i = 1, \theta^{(t)})P(U_i = 1|\theta^{(t)})}{P(x_i|\theta^{(t)})} \\
&= \frac{c_{ij}(\theta^{(t)})B_{ij}(\theta^{t})\pi_{ij}(\eta^{(t)})\prod_{k\neq j} P(x_{ik}|U_i = 1, \theta^{(t)})\delta^{(t)}}{P(U_i = 1|\theta^{(t)})P(x_i|U_i = 1, \theta^{(t)}) + P(U_i = 0|\theta^{(t)})P(x_i|U_i = 0, \theta^{(t)})} \\
&= \frac{c_{ij}(\theta^{(t)})B_{ij}(\theta^{t})\pi_{ij}(\eta^{(t)})\prod_{k\neq j} P(x_{ik}|U_i = 1, \theta^{(t)})\delta^{(t)}}{\delta^{(t)}P(x_i|U_i = 1, \theta^{(t)}) + (1 - \delta^{(t)})P(x_i|U_i = 0, \theta^{(t)})} \\
&= \frac{c_{ij}(\theta^{(t)})B_{ij}(\theta^{t})\pi_{ij}(\eta^{(t)})\prod_{k\neq j} P(x_{ik}|U_i = 1, \theta^{(t)})\delta^{(t)}}{\delta^{(t)}P(x_{ij}|U_i = 1, \theta^{(t)}) \times \prod_{k\neq j} P(x_{ik}|U_i = 1, \theta^{(t)}) + (1 - \delta^{(t)})P(x_i|U_i = 0, \theta^{(t)})} \\
&= \frac{c_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)})\pi_{ij}(\eta^{(t)})\delta^{(t)}}{\delta^{(t)}[\pi_{ij}(\eta^{(t)})c_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)}) + (1 - \pi_{ij}(\eta^{(t)}))c_{ij}(\theta^{(t)})] + (1 - \delta^{(t)})\frac{P(x_i|U_i=0,\theta^{(t)})}{\prod_{k\neq j} P(x_{ik}|U_i=1,\theta^{(t)})}}
\end{aligned}
$$

Denote $c_{ij} = P(x_{ij}|U_i = 1, Z_{ij} = 0, \theta)$. Then $P(x_{ij}|U_i = 1, Z_{ij} = 1, \theta) = c_{ij}B_{ij}$, $B_{ij}$ is the Bayes factor for variant $j$ in gene $i$,

$$
B_{ij} = \frac{P(x_{ij}|U_i = 1, Z_{ij} = 1)}{P(x_{ij}|U_i = 1, Z_{ij} = 0)}
$$

Note here $x_{ij}$ is the counts of only cases. So we consider the conditional distribution of $x_{1ij}|x_{1ij} + x_{0ij}$, which is

$$
\begin{aligned}
B_{ij} &= \frac{P(x_{1ij}|x_{1ij} + x_{0ij}, U_i = 1, Z_{ij} = 1)}{P(x_{ij}|x_{1ij} + x_{0ij}, U_i = 1, Z_{ij} = 0)} \\
&= \frac{\int Bin(x_{1ij}|x_{1ij} + x_{0ij}, \frac{\gamma N_1}{\gamma N_1 + N_0})\frac{1}{\Gamma(\bar{\gamma}\sigma)\sigma^{\bar{\gamma}\sigma}}\gamma^{\bar{\gamma}\sigma - 1}e^{-\frac{\gamma}{\sigma}}d\gamma}{Bin(x_{1ij}|x_{1ij} + x_{0ij}, \frac{N_1}{N_1 + N_0})}
\end{aligned}
\qquad (13)
$$

Because

$$
\begin{aligned}
\frac{P(x_i|U_i = 0, \theta^{(t)})}{\prod_{k\neq j} P(x_{ik}|U_i = 1, \theta^{(t)})} &= \frac{P(x_{ij}|U_i = 1, \theta^{(t)})P(x_i|U_i = 0, \theta^{(t)})}{\prod_j P(x_{ij}|U_i = 1, \theta^{(t)})} = P(x_{ij}|U_i = 1, \theta^{(t)}) \times \frac{1}{B_i(\theta^{(t)})} \\
&= [\pi_{ij}(\eta^{(t)})c_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)}) + (1 - \pi_{ij}(\eta^{(t)}))c_{ij}(\theta^{(t)})] \times \frac{1}{B_i(\theta^{(t)})}
\end{aligned}
$$

thus we have

$$
\begin{aligned}
&E(U_i Z_{ij}|x, \theta^{(t)}) \\
&= \frac{B_{ij}(\theta^{(t)})\pi_{ij}(\eta^{(t)})\delta^{(t)}}{\delta^{(t)}[\pi_{ij}(\eta^{(t)})B_{ij}(\theta^{(t)}) + (1 - \pi_{ij}(\eta^{(t)}))] + (1 - \delta^{(t)})[\pi_{ij}(\eta^{(t)})B_{ij}(\theta^{(t)}) + (1 - \pi_{ij}(\eta^{(t)}))] \times \frac{1}{B_i(\theta^{(t)})}} \\
&= \frac{B_{ij}(\theta^{(t)})\pi_{ij}(\eta^{(t)})\delta^{(t)}}{[\delta^{(t)} + \frac{1-\delta^{(t)}}{B_i(\theta^{(t)})}] \times [\pi_{ij}(\eta^{(t)})B_{ij}(\theta^{(t)}) + (1 - \pi_{ij}(\eta^{(t)}))]}
\end{aligned}
$$

- M step updates $\theta$ by $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(t)})$. That is

$$
\theta^{(t+1)} = arg\ max_{\theta}\ Q(\theta|\theta^{(t)})
$$

Taking derivatives with respect to $\theta$ leads to

$$\frac{\partial Q}{\partial \delta} = \frac{\sum_i E(U_i|x,\theta^{(t)})}{\delta} - \frac{\sum_i (1 - E(U_i|x,\theta^{(t)}))}{1-\delta}$$

Suppose there are $C$ annotation groups. $\eta = (\eta_1, \eta_2, \cdots, \eta_C)$ is a vector. The variant $(i,j)$ belonging to group $c$ has the prior probability of being causal $\pi_{ij}(\eta) = \eta_c$ with group prior. Use $v(i,j) = c$ to denote variant $(i,j)$ in group $c$.

$$\frac{\partial Q}{\partial \eta_c} = \sum_i \sum_{j:v(i,j)=c} [\frac{E(U_i Z_{ij}|x,\theta^{(t)})}{\eta_c} - \frac{E(U_i|x,\theta^{(t)}) - E(U_i Z_{ij}|x,\theta^{(t)})}{1-\eta_c}]$$

$$\frac{\partial Q}{\partial \bar{\gamma}} = \sum_i \sum_j \frac{E(U_i Z_{ij}|x,\theta^{(t)})}{B_{ij}} \frac{\partial B_{ij}}{\partial \bar{\gamma}}$$

$$\frac{\partial Q}{\partial \sigma} = \sum_i \sum_j \frac{E(U_i Z_{ij}|x,\theta^{(t)})}{B_{ij}} \frac{\partial B_{ij}}{\partial \sigma}$$

Setting these equations equal to 0 yields

$$\delta^{(t+1)} = \frac{1}{I} \sum_i E(U_i|x,\theta^{(t)})$$

$I$ is the total number of genes.

$$\eta_c^{(t+1)} = \sum_i \sum_{j:v(i,j)=c} \frac{E(U_i Z_{ij}|x,\theta^{(t)})}{\sum_i \sum_{j:v(i,j)=c} E(U_i|x,\theta^{(t)})}$$

For parameters $\bar{\gamma}, \sigma$, there are no closed solutions for $\bar{\gamma}^{(t+1)}, \sigma^{(t+1)}$. Instead, we fix $\bar{\gamma}, \sigma$ based on empirical evidence.

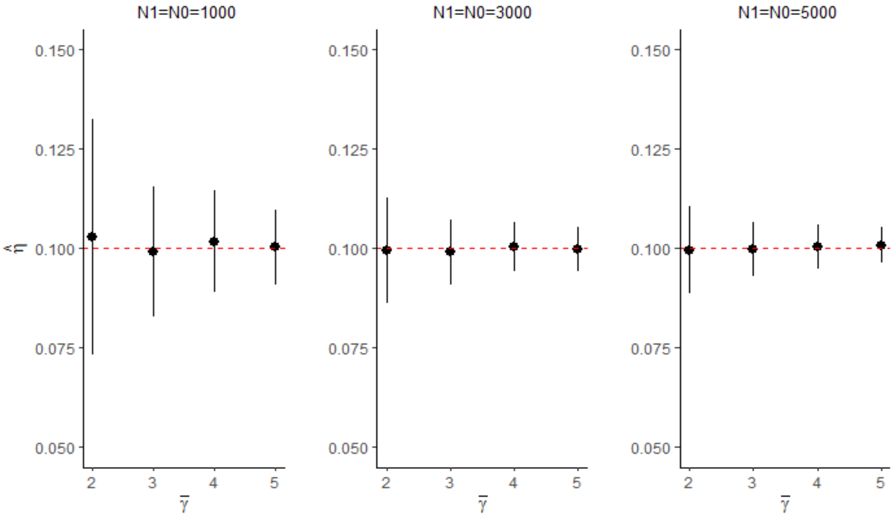


**Fig 8.** Parameter estimates by MIRAGE-VS across 100 simulations with varying sample size $N1 = N0 = 1000, 3000, 5000$ and varying $\bar{\gamma} = 2, 3, 4, 5$. Every data set has 1000 variants with the proportion of risk variants, $\eta = 0.1$.

## Acknowledgements

## References

1. Wellcome Trust Case Control, C.; and Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 447, 661-678.

2. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5-22

3. Shendure J, Findlay GM, Snyder MW Genomic Medicine-Progress, Pitfalls, and Promise. Cell. 2019 Mar 21;177(1):45-57.

4. Gibson, G. (2011). Rare and common variants: twenty arguments. Nat. Rev. Genet. 13, 135-145.

5. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA 106, 3871-3876.

6. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science. 335, 823-828.

7. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature. 491, 56-65.

8. Seunggeung Lee, Goncalo R. Abecasis, Michael Boehnke, and Xihong Lin Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014 Jul 3;95(1):5-23.

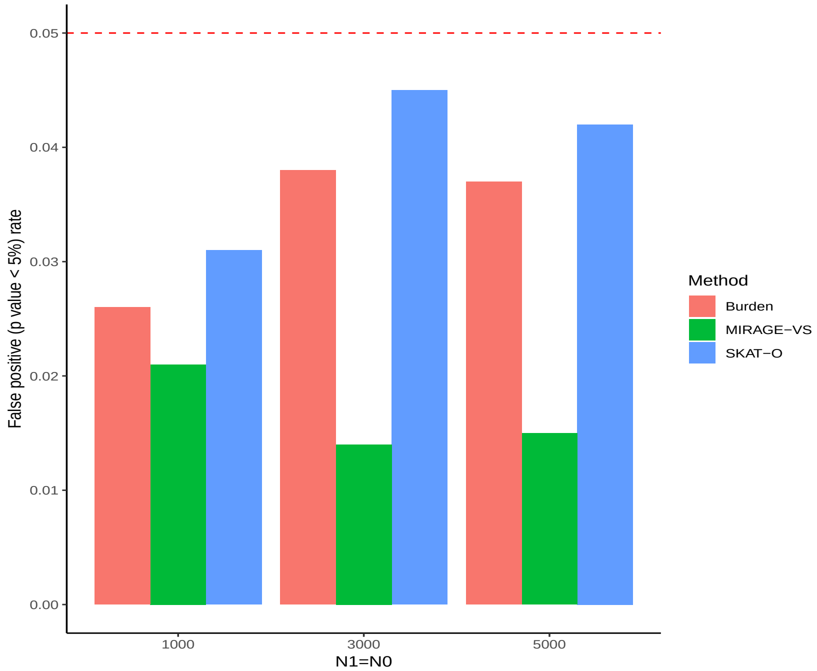


**Fig 9.** False positive rates at $p < 0.05$, across 1000 simulations generated under null hypothesis. Every simulated data set has 100 variants with $\eta = 0$. True $\bar{\gamma} = 5$.

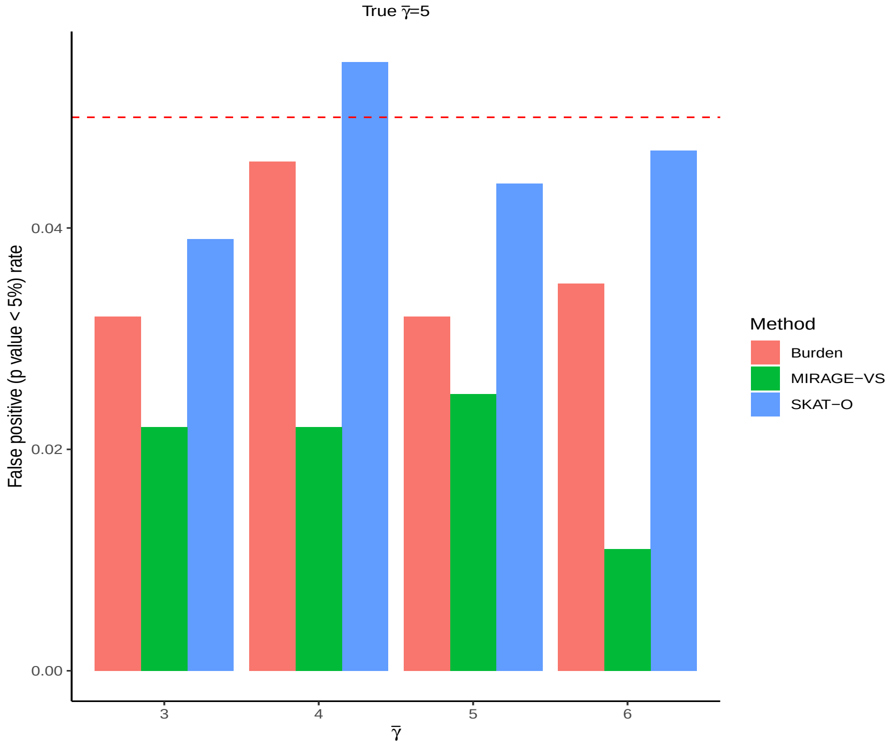


**Fig 10.** False positive rate of MIRAGE-VS with mis-specified $\bar{\gamma}$: Fix N1=N0=3000. 1000 variant sets are randomly generated with $\eta = 0$, each having 100 variants and true $\bar{\gamma} = 5$. Different values of $\bar{\gamma}$ are used to calculate $p$ values by MIRAGE-VS. Y-axis shows the false positive rate at p value less than 5%.

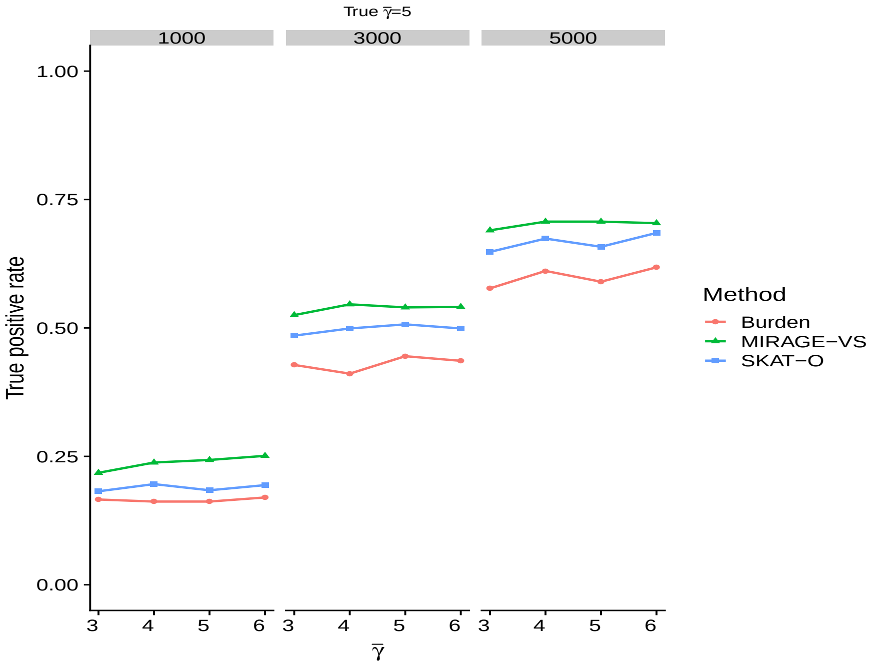


**Fig 11.** Sensitivity of power of MIRAGE-VS to the value $\bar{\gamma}$. We simulated data sets with true $\bar{\gamma} = 5$ and calculate power with $\bar{\gamma} = 3, 4, 5, 6$. We perform this analysis for 1000 random data sets with different sample size $N1 = N0 = 1000, 3000, 5000$.

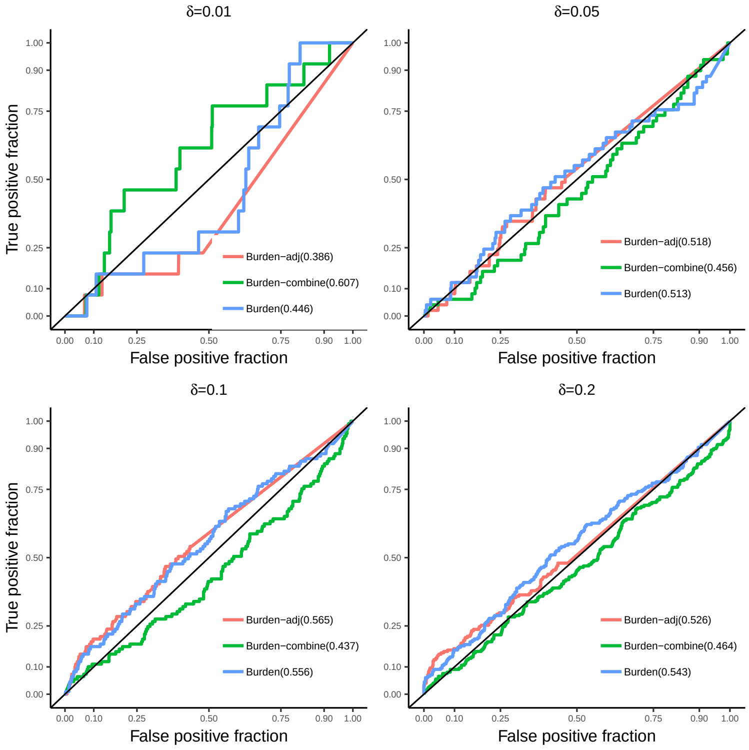


**Fig 12.** Power of several variations of burden test: burden (basic version), burden-min (we calculate p value for each variant category separately, then take the minimum $p$ value) and burden-combine (combine category-specific $p$ values by Fisher method).

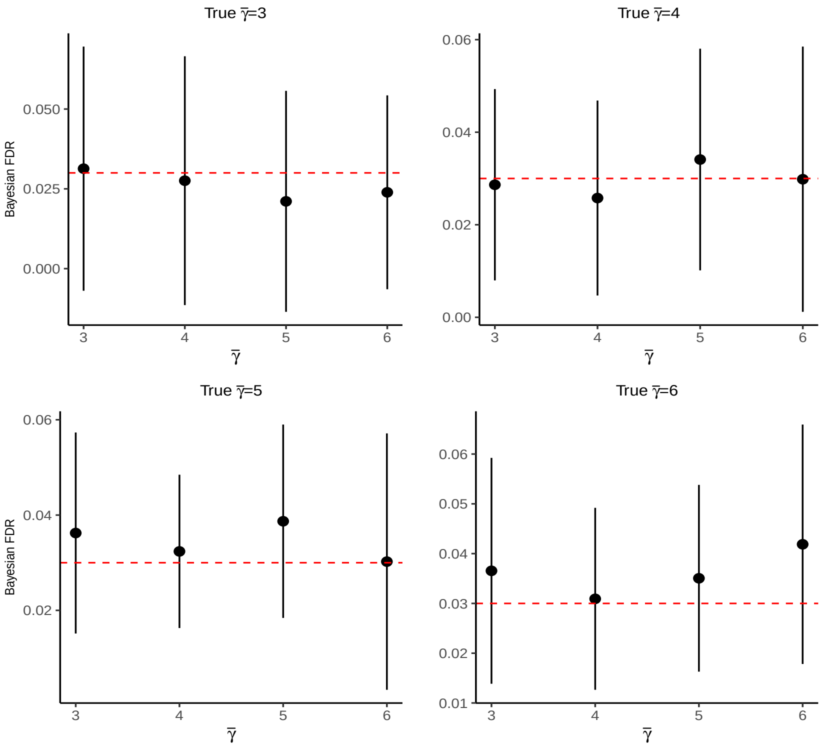
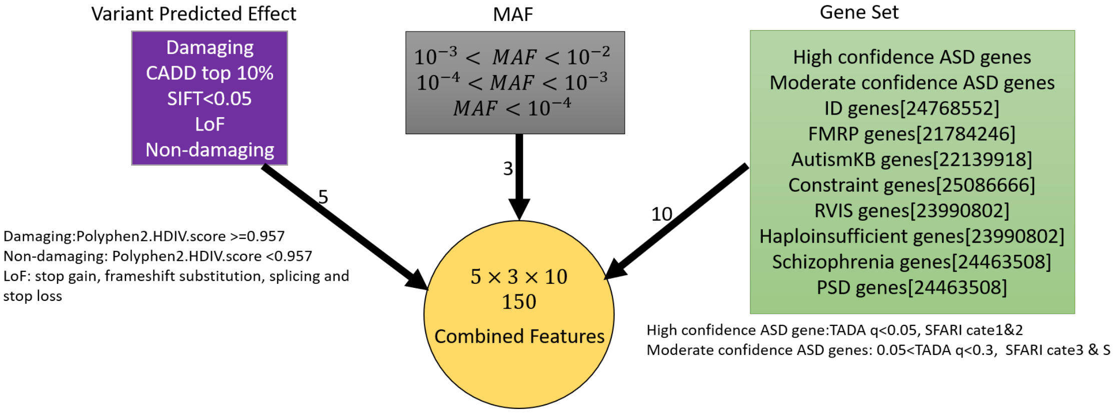


**Fig 13.** Sensitivity of false discovery rate (FDR) to mis-specified $\bar{\gamma}$. For each true $\bar{\gamma} = 3, 4, 5, 6$, we use 3, 4, 5, 6 in the analysis. In each scenario, the red horizontal dashed line is Bayesian FDR of 0.05, the vertical line is the mean of actual FDR with standard deviation across 20 replications. $N1 = N0 = 3000$, $10\%$ of 1000 genes are risk genes. Every gene has 100 variants in three variant groups with proportion of $30\%, 40\%, 30\%$ and the corresponding $\eta$ are $0.05, 0.2, 0.5$.



Damaging:Polyphen2.HDIV.score >=0.957
Non-damaging: Polyphen2.HDIV.score <0.957
LoF: stop gain, frameshift substitution, splicing and stop loss

High confidence ASD gene:TADA q<0.05, SFARI cate1&2
Moderate confidence ASD genes: 0.05<TADA q<0.3, SFARI cate3 & S

**Fig 14.** Definition of variant sets in MIRAGE-VS analysis of ASD. We combine features at two levels, variant and gene set. For variant features, we combine predicted effects and MAF. ID: intellectual disability. PSD: post-synaptic density.

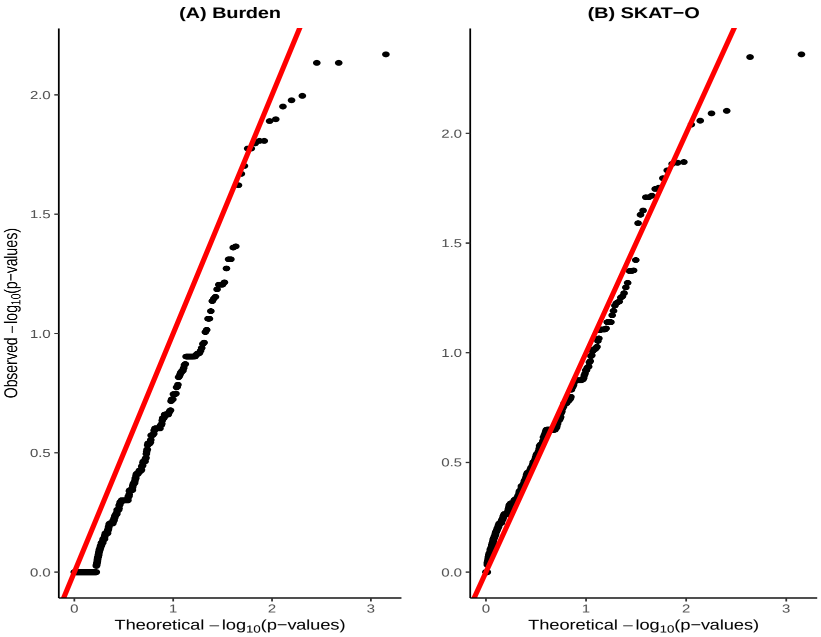


**Fig 15.** QQ plot of burden (A) and SKAT-O (B) in ASD data on LoF and damaging variants from top constrained genes.

9. Cirulli ET The Increasing Importance of Gene-Based Analyses. PLoS Genet. 2016 Apr 7;12(4):e1005852.

10. Li Bingshan, Leal Suzanne M Methods for Detecting Associations with Rare Variants for Common Diseases : Application to Analysis of Sequence Data. American Journal of Human Genetics. 2008, 83(3), 311-321.

11. Madsen BE1, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb;5(2):e1000384.

12. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010 Jun 11;86(6):832-8.

13. Derkach A1, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol. 2013 Jan;37(1):110-21.

14. Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15; 89(1): 82-93.

15. Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet. 2011 Sep 9;89(3):354-67.

16. Chen LS, Hsu L, Gamazon ER, Cox NJ, Nicolae DL. An exponential combination procedure for set-based association tests in sequencing studies. Am J Hum Genet. 2012 Dec 7;91(6):977-86.

17. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–237.

18. Byrnes AE, Wu MC, Wright FA, Li M, Li Y. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. Genet Epidemiol. 2013 Nov;37(7):666-74.

19. Xin He, et al. Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLoS Genet. 9(8): e1003671. doi:10.1371/journal.pgen.1003671.

20. Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 1995 Aug;57(2):455-64.

21. Purcell SM, et al (2014) A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 506(7487):185-90.

22. Fuchsberger C, et al 2016. The genetic architecture of type 2 diabetes. Nature. 2016 Aug 4;536(7614):41-47.

23. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B. (1977), 39 (1): 1-38.

24. Silvia De Rubeis, et al (2014) Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 515, 209-215.

25. Whittemore, A. S. (2007) A Bayesian false discovery rate for multiple testing. J. Appl. Stat. 34(1), 1–9.

26. Yuwen Liu, et al. A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies. American Journal of Human Genetics. 2018, 102(6), 1031-1047.

27. Han Fang, Pan Wei A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Human Heredity. 2010, 70, 42-54.

28. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Research. 38:e164, 2010.

29. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 46(3):310-5

30. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 7(4):248-249 (2010).

31. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.

32. Genovese, Giulio, Fromer, Menachem, et al Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nature Neuroscience. 19, 1433-1441 (2016).

33. Monkol Lek, Konrad J. Karczewski, et al, Exome Aggregation Consortium Analysis of protein-coding genetic variation in 60,706 humans. Nature. 536, pages 285–291.

34. Kaitlin E Samocha, Elise B Robinson, et al. A framework for the interpretation of de novo mutation in human disease. Nature Genetics. 46, 944-950 (2014).

35. Abrahams BS, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013 Oct 3;4(1):36.

36. Kalkman HO. A review of the evidence for the canonical Wnt pathway in autism spectrum disorders. Mol Autism. 2012 Oct 19;3(1):10.

37. Shao S, et al. The Roles of Genes in the Neuronal Migration and Neurite Outgrowth Network in Developmental Dyslexia: Single- and Multiple-Risk Genetic Variants. Mol Neurobiol. 2016 Aug;53(6):3967-3975.

38. Wang Q, et al. Increased co-expression of genes harboring the damaging de novo mutations in Chinese schizophrenic patients during prenatal development. Sci Rep. 2015 Dec 15;5:18209.

39. Hsiao K, Harony-Nicolas H, Buxbaum JD, Bozdagi-Gunal O, Benson DL. Cyfip1 Regulates Presynaptic Activity during Development. J Neurosci. 2016 Feb 3;36(5):1564-76.

40. Abekhoukh S,et al. New insights into the regulatory function of CYFIP1 in the context of WAVE- and FMRP-containing complexes. Dis Model Mech. 2017 Apr 1;10(4):463-474.

41. Davenport EC, Szulc BR, Drew J, Taylor J, Morgan T, Higgs NF, López-Doménech G, Kittler JT. Autism and Schizophrenia-Associated CYFIP1 Regulates the Balance of Synaptic Excitation and Inhibition. Cell Rep. 2019 Feb 19;26(8):2037-2051.e6.

42. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA1, Fritzilas N, Hakenberg J, Dutta A, Shon J, Xu J, Batzoglou S, Li X, Farh KK. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018 Aug;50(8):1161-1170.

43. Sanders SJ, et al, (2015) Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 87(6):1215-1233.