MIRAGE: a Bayesian statistical method for gene-level rare variant analysis incorporating functional annotations

Shengtong Han[1,2,*+], Xiaotong Sun[2*], Laura Sloofman[3,4*], F. Kyle Satterstrom[5,6,7], Xizhi Xu[2], Lifan Liang[2], Nicholas Knoblauch[2], Wenhui Sheng[8], Siming Zhao[9], Tan-Hoang Nguyen[10], Gao Wang[11], Autism Sequencing Consortium[12], Joseph Buxbaum[3,4], and Xin He[2,12,+]

[1]School of Dentistry, Marquette University, Milwaukee, WI, USA
[2]Department of Human Genetics, University of Chicago, Chicago, IL, USA
[3]Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[5]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[7]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
[8]Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, WI, USA
[9]Department of Biomedical Data Science, Dartmouth College, Hanover, NH, USA
[10]Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA
[11]Department of Neurology, Columbia University Vagelos College of Physicians and Surgeons, New York City, NY, USA
[12]Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA
[*] *Contribute equally to this project*
[+] *Corresponding to Xin He, xinhe@uchicago.edu; Shengtong Han, shengtong.han@marquette.edu*

Abstract

Rare variant analysis is commonly used in whole-exome or genome sequencing studies. Compared to common variants, rare variants tend to have larger effect sizes and often directly point out causal genes. These potential benefits make association analysis with rare variants a priority for human genetics researchers. To improve the power of such studies, numerous methods have been developed to aggregate information of all variants of a gene. However, these gene-based methods often make unrealistic assumptions, e.g. the commonly used burden test effectively assumes that all variants chosen in the analysis have the same effects. In practice, current methods are often underpowered. We propose a Bayesian method: MIxture model based Rare variant Analysis on GEnes (MIRAGE). MIRAGE analyzes summary statistics, i.e. variant counts from inherited variants in trio-sequencing or from ancestry-matched case-control studies. MIRAGE captures the heterogeneity of variant effects by treating all variants of a gene as a mixture of risk and non-risk variants, and uses external information of variants to model the prior probabilities of being risk variants. We demonstrate in both simulations and analysis of an exome-sequencing dataset of autism, that MIRAGE significantly outperforms current methods for rare variant analysis. The top genes identified by MIRAGE are highly enriched with known or plausible autism risk genes. MIRAGE is available at https://xinhe-lab.github.io/mirage.

1 Introduction

Genome-wide association studies (GWAS) have identified many loci associated with various complex traits[1–3]. However, identifying causal variants and their target genes is often difficult. Additionally, most common variants discovered by GWAS have small effect sizes[2,3]. These limitations make it difficult to translate GWAS findings into molecular mechanisms of diseases. Sequencing studies focusing on rare variants have the potential to address these challenges. Large effect variants, because of purifying selection, are often rare in the population[4–7]. Indeed, rare variants were estimated to explain between 24% to 50% of heritability of complex traits[8,9]. Another benefit of rare variant analysis is that linkage disequilibrium (LD) is much weaker[10], making it straightforward to identify causal variants. Furthermore, reduced sequencing cost has made it feasible to sequence large numbers of exomes or genomes. Thus, discovering rare variants underlying the risks of complex diseases from sequencing studies is an important area in human genetics[11].

Existing rare variant studies often focus on the protein-coding portions of the genome, using Whole Exome Sequencing (WES). WES studies have achieved notable successes in medically important traits such as Alzheimer's disease and schizophrenia[12–14]. Nevertheless, these studies typically discovered relatively few variants associated with the traits. A natural strategy to improve the power of rare variant studies is to aggregate information from all rare variants in a gene to test if they collectively associate with the phenotype[15]. Many methods have been developed to perform variant set tests[11]. The most commonly used method is the burden test, which collapses all rare, potentially deleterious variants in a gene, and tests the association of the variant "burden" with a phenotype[16–20]. The Sequence Kernel Association Test (SKAT)[21] assesses if the variance of the effect sizes of the variants is equal to 0. Other tests combine the p values of individual variants, through Fisher's or similar methods[19,22], Aggregated Cauchy association test (ACAT)[23], or the Generalized Berk-Jones (GBJ) test[24]. Additionally, rare variant analysis methods have been developed to incorporate functional information of variants and to deal with unbalanced case-control data[25,26].

Despite these efforts, the rare variant set test have limited successes. In a WES study of 450,000 subjects from UK Biobank, the burden test found few risk genes for many common traits, for example, 1 gene for Body Mass Index, 3 genes for Type 2 Diabetes and 2 genes for asthma[27]. Often, the power of these tests is lower than the single-variant association test[28]. In a Crohn's disease WES, researchers found 45 rare risk variants from single-variant analysis, but only 2 risk genes from the burden analysis. These results suggest that causal variants are sparse even within risk genes, leading to a loss of power when collapsing all rare variants in a gene together. Although p-value combination methods such as ACAT were designed to be more sensitive when a small number of variants have causal effects, they do not explicitly model the heterogeneity of effects across variants and continue to collapse variants with frequencies below a threshold[23]. Researchers have attempted to improve the power of rare variant analysis by predicting likely deleterious variants and then running association analysis on these variants, but bioinformatic predictions remain imperfect.

We aim to better model variant effects with a mixture method. Our statistical model can be applied to any settings where we have variant counts from cases, and from population-matched control samples. In other words, it is designed for situations which have properly controlled for population stratification. This simplifies the statistical model and allows us to

focus on the challenge of modeling variant effects. In practice, one can obtain such datasets through population-matched controls[29,30] or through family-based trio studies. In the former, researchers project the genetic data from both cases and potential control subjects into the Principal Component (PC) space, and then select controls that best match the cases in this space. In the latter, variants from parents can be classified as transmitted to children or not. For any variant not associated with the case phenotype, the ratio of transmitted vs un-transmitted alleles should converge to one as sample sizes grow large. Following transmission disequilibrium test (TDT)[31,32], an imbalanced ratio of a variant would suggest association of the variant with the trait[33]. Thus one can treat transmitted variants as the variant count from cases, and the un-transmitted ones as controls (pseudo-control). Indeed, trio-sequencing studies of autism spectrum disorder (ASD) has revealed an overall tendency of transmission of deleterious variants[34], but few individual genes have been identified using inherited data, highlighting the need of better data analysis methods.

We propose a Bayesian statistical method to better account for the heterogeneity of the variant effects in a gene or a genomic region. We model the variants in a gene as a mixture of risk and non-risk variants. The prior probability of a variant being a risk variant depends on the functional annotations of the variant. These prior probabilities are generally low, reflecting the sparsity of risk variants, but also vary considerably across variants based on their likely functional effects. This Bayesian strategy of incorporating functional information as prior has significant advantages over simply filtering variants based on their likely effects. In general, the external annotations have limited accuracy in predicting functional effects; and even if a variant is functional, it does not necessarily have an effect on the particular trait of interest. Importantly, the parameters linking the annotations of variants and their prior probabilities are estimated using an Empirical Bayes strategy, by pooling data from all genes. Our model has the additional advantage that it requires only summary statistics (variant counts). This makes it computationally efficient and eliminates the need of sharing individual-level data.

We demonstrated the advantage of the proposed method, MIxture model based Rare variant Analysis on GEnes (MIRAGE), in detecting putative risk genes over existing methods, in both simulation studies, and the analysis of inherited variants from ASD whole exome sequencing studies.

2 Methods

2.1 MIRAGE model

The input data of MIRAGE include rare variant counts in cases and controls with sample sizes N_1 and N_0, respectively. For variant j of gene i, we denote X_{ij} its allele count in cases, $X_{ij}^{(0)}$ in controls, and T_{ij} its total allele counts in cases and controls. We also have annotations of each variant. We assume each variant belongs to one variant category, denoted as c_{ij} for variant j of gene i. MIRAGE is a probabilistic graphical model that describes a generative process of the variant count data (**Figure 1B**).

Specifically, we denote U_i the indicator of whether gene i is a risk gene. U_i is a Bernoulli random variable with mean δ. We denote Z_{ij} the hidden indicator of whether variant j of gene i is a risk variant. The distribution of Z_{ij} 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 η_c the proportion of risk variants in the variant category c. Then Z_{ij} follows Bernoulli distribution with mean $\eta_{c_{ij}}$.

The allele counts of a variant in cases and controls follow Poisson distributions. Their rates depend on the status of whether a variant is a risk variant. We denote q_{ij} the allele frequency of variant j in the controls. If j is a non-risk variant ($Z_{ij} = 0$), its allele frequency in cases would also be q_{ij}. So we have:

$$X_{ij}|Z_{ij} = 0 \sim \text{Pois}(q_{ij}N_1), \quad X_{ij}^{(0)}|Z_{ij} = 0 \sim \text{Pois}(q_{ij}N_0). \tag{1}$$

If j is a risk variant ($Z_{ij} = 1$), its allele frequency in cases would generally be elevated. Let γ_{ij} be the fold increase of allele frequency. It can be interpreted as the relative risk of variant j, as shown in[33]. So we have:

$$X_{ij}|Z_{ij} = 1 \sim \text{Pois}(\gamma_{ij}q_{ij}N_1), \quad X_{ij}^{(0)}|Z_{ij} = 1 \sim \text{Pois}(q_{ij}N_0). \tag{2}$$

It is generally difficult to estimate γ_{ij} for individual rare variants, so as in TADA[33], we treat γ_{ij} as a random variable following $\text{Gamma}(\bar{\gamma}, \sigma)$. The hyper-parameter $\bar{\gamma}$ is the prior mean of relative risk of risk variants, and σ is the dispersion parameter. These hyperparameters need to be provided by the users, e.g. by estimating them from the data[35], though in practice, we found that the results are relatively robust to the exact values.

We note that q_{ij} 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 allows us to eliminate q_{ij}:

$$X_{ij}|T_{ij}, Z_{ij} = 0 \sim \text{Bin}\left(T_{ij}, \frac{N_1}{N_1 + N_0}\right) \qquad X_{ij}|T_{ij}, Z_{ij} = 1 \sim \text{Bin}\left(T_{ij}, \frac{\gamma_{ij}N_1}{\gamma_j N_1 + N_0}\right). \tag{3}$$

We marginalize γ_{ij} in evaluating the probability of allele counts for risk variants:

$$P(X_{ij}|T_{ij}, Z_{ij} = 1) = \int \text{Bin}\left(X_{ij}; T_{ij}, \frac{\gamma_{ij}N_1}{\gamma_{ij}N_1 + N_0}\right) \text{Gamma}(\gamma_{ij}; \bar{\gamma}, \sigma) d\gamma_j. \tag{4}$$

We are now ready to describe the likelihood function and our inference procedure. We

denote $\mathbf{X}$ as the data of allele counts in cases of all variants in all genes. Similarly, we denote $\mathbf{T}$ the data of allele counts in cases and controls combined of all variants in all genes. We also denote C as the set of variant annotations c_{ij}'s. Our primary parameters of interest are δ, the proportion of risk genes, and η, the vector of η_c's for all variant categories. The likelihood function is given by:

$$P(\mathbf{X}|\mathbf{T}, \mathbf{C}, \delta, \eta) = \prod_i [(1-\delta)P(X_i|T_i, U_i = 0) + \delta \cdot P(X_i|T_i, C_i, U_i = 1, \eta)], \tag{5}$$

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

$$P(X_i|T_i, U_i = 0) = \prod_j \text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right). \tag{6}$$

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

$$P(X_i|T_i, C_i, U_i = 1, \eta) = \prod_j \left[(1-\eta_{C_{ij}})\text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right) + \eta_{C_{ij}} P(X_{ij}|T_{ij}, Z_{ij} = 1)\right], \tag{7}$$

where $P(X_{ij}|T_{ij}, Z_{ij} = 1)$ is given by Equation (4). We note that in computing the likelihood of a single gene, we ignore potential linkage disequilibrium and assume all variants are independent. Since we focus on variants with AF $< 0.1\%$, this assumption is generally valid. In practice, we can also perform LD pruning to create independent set of variants.

Given the likelihood function involving latent variables U_i's and Z_{ij}'s, we derive the Expectation-Maximization (EM) algorithm to estimate the parameters δ and η (see EM Algorithm in the Supplements), with their initial values being randomly chosen. We note that the update rules of the parameters in the M-step have simple, closed forms.

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, PP_i, as:

$$B_i = \frac{P(X_i|T_i, C_i, U_i = 1, \hat{\eta})}{P(X_i|T_i, U_i = 0)}, \quad \text{PP}_i = \frac{\delta B_i}{1 - \delta + \delta B_i}. \tag{8}$$

From Equations 6 and 7, it is easy to show that B_i can be related to the evidence at the single variant level:

$$B_i = \prod_j [(1-\eta_{C_{ij}}) + \eta_{C_{ij}} B_{ij}], \quad B_{ij} = \frac{P(X_{ij}|T_{ij}, Z_{ij} = 1)}{P(X_{ij}|T_{ij}, Z_{ij} = 0)}, \tag{9}$$

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 η_c will contribute more to the gene level evidence. We also note that the log-BF of a gene $\log B_i$ can be partitioned as the sum of contributions of each variant, $\log(1 - \eta_{C_{ij}} + \eta_{C_{ij}} B_{ij})$. This partition is used when we assessed the contribution of individual variants, or variant groups, to the evidence of a gene in real data analysis. Once we determine the BFs and posterior probabilities of all genes, we control for multiple testing by performing Bayesian FDR control[33].

2.2 Simulation procedure

We simulated the variant counts of a set of genes, 1,000 in our simulations, under case-control data with sample sizes $N_1 = N_0 = 3,000$. We assumed each gene has a mixture of variants in different categories, with fixed proportions of variant categories in each gene. We used three categories mimicking LoF, deleterious missense variants and the rest. The proportions of variants in these three categories were 10%, 30% and 60%, respectively. Our simulation started 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 sampled 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 $Z_{ij} \sim \text{Bernoulli}(\eta_c)$. We set η_c 0.5, 0.2 and 0.05 for the three categories, respectively. For a risk variant $Z_{ij} = 1$, we sampled the relative risk $\gamma_{ij} \sim \text{Gamma}(\bar{\gamma}, \sigma)$. Both $\bar{\gamma}$ and σ were set as user-specified parameters. In the simulations, we used $\bar{\gamma} = 5$ for the first category of variants and 3 for the other two categories, and used $\sigma = 1$.

Having sampled the status of risk variants and their relative risk, we sampled the variant counts. First, we sampled the allele frequency q_{ij} from a Beta distribution. If $Z_{ij} = 0$, we sampled from $\text{Beta}(\alpha_0, \beta_0)$. We set $\alpha_0 = 0.1, \beta_0 = 1000$ in our simulations. If $Z_{ij} = 1$, we assumed variants would be even rarer, so we sampled from $\text{Beta}(\alpha, \beta)$, where $\alpha = 0.1, \beta = 2000$ (so mean AF is two times lower than non-risk variants). Now we sample the genotype of each individual in cases and in controls. In controls, the genotype of a variant j in gene i of a subject would follow Bernoulli distribution with mean q_{ij}; and in cases, the genotype would follow Bernoulli distribution with mean $\gamma_{ij} q_{ij}$. We note that we sampled the genotype of each variant independently, assuming no LD between variants. From these genotype data, we can collapse variant counts in cases and in controls, which would be used by MIRAGE and other tests.

In additional simulations (**Figure S2**), we varied the number of variants per gene. We sampled the variant number uniformly from 50 to 500 in every gene. The rest of simulations was the same as before.

2.3 WES data of ASD

Whole exome sequencing (WES) data of autism were obtained from Autism Sequencing Consortium (ASC) and published in Fu et al[36]. Specifically we aggregated inherited variants from ASC B14, ASC B15-16, SPARK Pilot and SPARK main freeze (WES1), resulting in a total of 14,578 ASD probands and 5,391 unaffected siblings. ASC B14 contains ASC samples and SSC (Simons Simplex Collection), contributing 7,291 probands (6,026 male and 1,265 female) and 2,348 silbings (1,158 male and 1,190 female). ASC B15-16 has 279 probands (223 male and 56 female), and 11 siblings (6 male and 5 female). SPARK main freeze has 6,543 probands (5,219 male and 1,324 female) and 3,032 siblings (1,554 male and 1,478 female). SPARK Pilot has 465 probands (376 male and 89 female) only. To evaluate transmitted/untransmitted variants, dataset specific filters were applied, such as different VQSLOD thresholds. Additional filters were applied to ASC B14 to standardize the samples sequenced across the past several years. After applying these filters, transmitted/untransmitted alleles were called and annotations were produced. We note that the transmitted and untransmitted counts reflect the total number of transmission events, rather than the number of individuals or the specific familial configurations. For instance, we do not distinguish between a variant transmitted once each in two separate families, and a variant transmitted from both heterozygous parents to a single

homozygous alternate child in another family. In both cases, the tally would increase by two. More details on the variant calling, variant compilation, quality control can be found in Fu et al[36], and variant annotations in Satterstrom et al[37]. Synonymous variants and variants with allele frequency $> 0.1\%$ are filtered.

2.4 Applying MIRAGE to ASD data

We annotated variants using both loss-of-function (LoF) and missense annotations. For LoF variants, we applied two commonly used metrics: the probability of loss-of-function intolerance (pLI)[38,39] and the loss-of-function observed/expected upper bound fraction (LOEUF)[40], which quantify gene-level intolerance to LoF mutations. For missense variants, we used two approaches: the missense badness, PolyPhen-2, and constraint (MPC) score[41], and AlphaMissense, a deep learning-based pathogenicity predictor[42].

Based on these annotations, we defined variant groups as follows. For LOEUF: decile 1 (low LOEUF, most intolerant), deciles 2–3 (med LOEUF, moderately intolerant), and deciles >3 (high LOEUF, tolerant). For pLI: pLI ≥ 0.995 (high), $0.995 >$ pLI ≥ 0.5 (med), and pLI < 0.5 (low). For MPC: MPC ≥ 2 (high), $2 >$ MPC ≥ 1 (med), and MPC < 1 (low). For AlphaMissense: likely pathogenic, ambiguous, and likely benign.

All genes in the genome were used to run MIRAGE. 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 model parameter δ, we randomly chose initial values, and the algorithm converges if the change of parameter estimates in two iterations is less than 10^{-5}. Once these parameters are estimated, their values are assumed to be known, and are used in calculating BF of each gene.

After running MIRAGE on all genes, we checked the LD pattern of the variants supporting all genes at PP > 0.5. We defined supporting variants as those with logBF > 1. We computed pairwise LD of all the supporting variants in a gene, one gene at a time. From these results, only *LMNB1* and *SV2B* have variants in LD (**Figure S6**). We then chose the variants with the highest logBF, and filtered out the ones in LD. The BFs of the genes would then be updated.

The results of MIRAGE of individual genes were displayed using the lollipop plots. They were generated by trackViewer[43].

2.5 Other programs for rare variant analysis

We used statistical software R (R version 4.4.0) to run other programs, using the same settings in both simulations and real data analyses.

For burden test of a set of variants within a gene, we collapsed the total variant count in cases and in controls, then compared the difference of the burden between the two groups using Fisher's exact test. The R package *AssotesteR* was employed for CMC and ASUM analyses. For CMC, three MAF cutoffs, i.e. $MAF < 1/3000$, $1/3000 < MAF < 5/3000$, $5/3000 < MAF < 20/3000$, $MAF > 20/3000$ were applied to partition the variants. We performed 100 permutations for both CMC and ASUM to obtain p-values, as increasing the number of permutations did not significantly impact the results. R package *SKAT* (version 2.2.5) was used to perform SKAT-O without covariates.

For the ACAT analysis of a gene, we first tested the difference of case and control counts for each variant using Fisher's exact test. Variants with minor allele counts below 10 were collapsed for testing, following the approach outlined by Liu et al[23]. We then aggregated the p values

of all variants within a variant group, using the *ACAT* function from the R package *ACAT* (version 0.91), available at https://github.com/yaowuliu/ACAT. Finally, we applied ACAT to group level p values to get the p value of a gene.

In the analysis of ASD data, we also ran the standard single variant analysis on 1,539,388 variants. This test compares transmitted and non-transmitted variant counts using the binomial test (with the success probability 0.5). As another gene level test, we consider the minimum p value of all variants in a gene. The minimum p values, however, are not calibrated. Under the assumption that all variants are independent, the null distribution of minimum p values follows the Beta distribution, $\text{Beta}(1, n)$ where n is the number of variants in a gene. We thus corrected the minimum p values of all genes using these Beta distributions. We noticed that the resulting empirical p values for both single variant test and minimum-p test are deflated (Supplementary **Figure S5**). This probably reflects that there are a large number of singleton variants (a single total variant count), resulting in p value = 1 from the binomial test. In other words, the p values of most null variants are not uniformly distributed.

2.6 Gene set enrichment analysis

MIRAGE identified 18 genes with PP > 0.5. We thus initially selected the top 18 genes as identified by MIRAGE, burden tests, and ACAT. Additionally, a comprehensive set comprising all 16,469 genes analyzed were used as a baseline for comparative purposes. Our primary objective was to determine whether these genes exhibit enrichment in gene sets associated with ASD. Here, "enrichment" refers to the proportion of overlap between our top-ranked genes and those within various ASD-related gene sets.

The ASD-related gene sets incorporated into our analysis include: (1) known ASD genes from literature: genes from the SFARI database[44], including only categories 1 and 2, totaling 769 genes; (2) genes identified in *de novo* mutation studies, TADA[36] (TADA FDR < 0.05, comprising 146 genes) ; (3) genes associated with Intellectual Disability (ID)[45], encompassing 252 genes in total; (4) Schizophrenia (SCZ) risk genes[46], with a total of 1,796 genes; (5) genes involved in relevant biological processes, such as the Post-synaptic density (PSD)[46], comprising 661 genes, and FMRP target genes[46], totaling 783 genes; (6) evolutionarily constrained genes as identified by RVIS[47], including 846 genes, alongside Haplo-insufficient (HI) genes[47], totaling 1,440 genes; and (7) major depressive disorder (MDD)[48], totalling 296 genes.

To evaluate the statistical significance of the observed differences in gene enrichment between the MIRAGE-identified genes and the baseline set, Fisher's exact test was employed.

2.7 STRING Network analysis of candidate genes.

We used the STRING database (version 12.0, https://string-db.org) to construct gene networks for MIRAGE genes with posterior probability (PP) > 0.5 and for autism spectrum disorder (ASD) genes[36]. The ASD genes were identified by jointly analyzing rare *de novo* and inherited protein-truncating variants, damaging missense variants, and copy number variants from exome sequencing of 63,237 individuals, using an extended Bayesian framework (TADA[33]) to integrate variant types and inheritance patterns.

To generate the gene network shown in **Figure 4C**, we applied the default STRING settings. As a control, we randomly sampled 100 genes from all genes analyzed by MIRAGE, constructed their networks, and calculated the number of links between these random genes and ASD genes.

We then compared the number of links of MIRAGE genes to ASD genes against that of random genes using a Poisson test. Specifically, we observed 42 links between 18 MIRAGE risk genes and ASD genes. In the random set, we observed 1.46 links per gene. We would then test the significance using Poisson test.

3 Results

3.1 Overview of MIRAGE

MIRAGE requires summary statistics as the input, in the form of the counts of all rare variants in cases and in controls (**Figure 1A**). In addition, MIRAGE takes functional information of variants as input, assuming variants are assigned into disjoint categories, based on their likely effects, for example, loss-of-function (LoF) variants, and likely deleterious missense variants[41] (**Figure 1A**). For each gene, MIRAGE assesses how likely its data is generated from a non-risk gene model M_0, vs. a risk gene model M_1. Under M_0, all the variants of a gene are non-risk variants, and their expected frequencies are equal between cases and controls. Under M_1, any variant has a prior probability of being a risk variant, with the probability η_c for the variants in the category c. The frequencies of risk variants would be generally different between cases and controls. Thus each variant would contribute some information: an imbalance between case and control counts would provide some support to M_1, and functionally important variants would make larger contributions. The information across all variants in a gene would then be combined by MIRAGE to form the statistical evidence of M_1 vs. M_0, known as Bayes factor (BF).

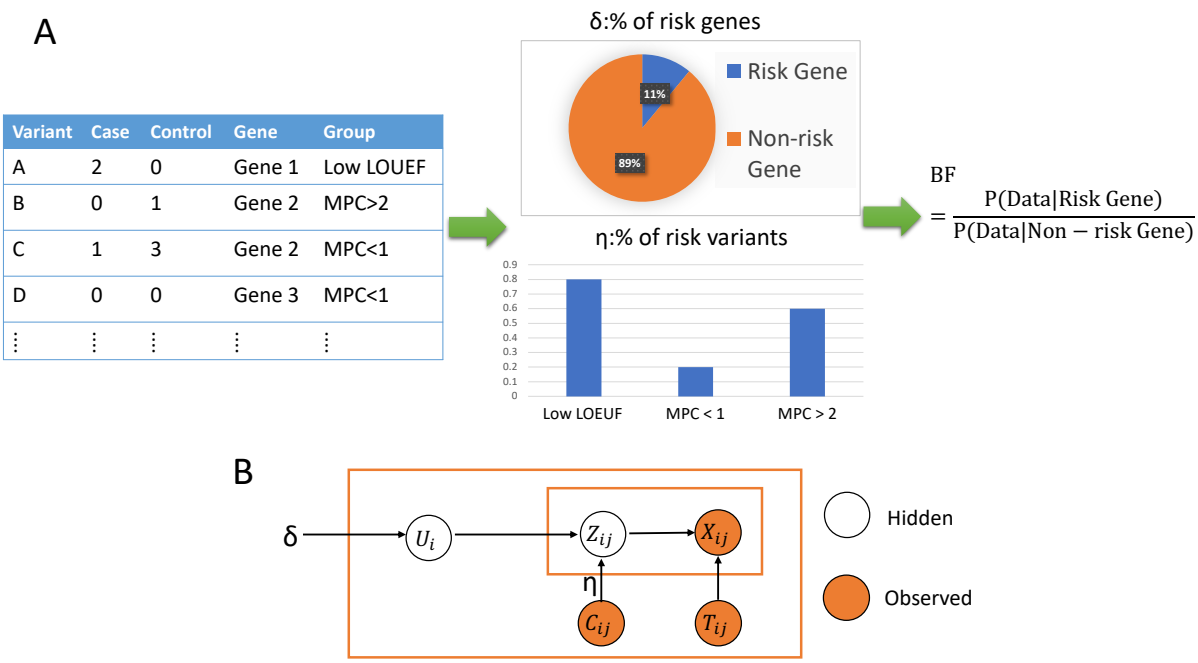

Figure 1: MIRAGE work-flow and model.
(A) MIRAGE work-flow. The input has information of all rare variants, including case and control counts, associated genes, and the functional groups. MIRAGE estimates the parameters and uses these values to compute the Bayes factor (BF) of all genes. LOEUF: Loss-Of-function Observed/Expected Upper bound Fraction, measures how tolerant a gene is to LoF variants; the lower, the more deleterious. MPC: Missense badness, PolyPhen-2, and Constrain scores; the higher, the more deleterious **(B)** MIRAGE model. See text for definitions of variables and parameters. The outer box corresponds to one gene, and the inner box one variant in a gene. Hidden: unobserved data.

MIRAGE estimates all the parameters, including the prior probability of being a risk gene, δ, and the proportion of risk variants, η_c, for all categories, using the entire dataset of all genes (**Figure 1A**). Then for each gene, it assesses its evidence as a risk gene by computing its BF. The BF of a gene can be used to compute its Posterior Probability (PP) of being a risk gene.

Unlike p value, PP has a simple interpretation of the probability of being a risk gene, given all the data we have about the gene. Using the PPs, MIRAGE can also perform multiple testing control, using a Bayesian False Discovery Rate (FDR) approach[49].

MIRAGE can be formulated as a probabilistic graphical model (**Figure 1B**). Let U_i be an indicator variable of whether the gene i is a risk gene (1 if yes and 0 otherwise). Then U_i follows a Bernoulli distribution with probability δ. For the j-th variant of gene i, let X_{ij} and T_{ij} be its allele count in cases and the total allele count, respectively. The distribution of these counts depends on whether the variant j is risk variant or not, denoted as Z_{ij}. For a non-risk gene ($U_i = 0$), $Z_{ij} = 0$ for all variants. For a risk gene ($U_i = 1$), Z_{ij} depends on its variant category, denoted as C_{ij}. If its category is c, then Z_{ij} is a Bernoulli random variable with probability η_c. Given Z_{ij} of a variant, we can model its allele count in cases as a binomial distribution. Assuming equal numbers of cases and controls (see Methods for the general setting), we have:

$$X_{ij}|T_{ij}, Z_{ij} = 0 \sim \text{Bin}\left(T_{ij}, 1/2\right) \qquad X_{ij}|T_{ij}, Z_{ij} = 1 \sim \text{Bin}\left(T_{ij}, \frac{\gamma_{ij}}{\gamma_{ij}+1}\right), \tag{10}$$

where γ_{ij} is the relative risk of variant j, modeled as a Gamma distribution. We assumed all variants are independent. We think this is justified when one analyzes rare or ultra-rare variants, with MAF below 0.1%. In practice, we can also prune variants in LD. With this probabilistic model, we used an Expectation Maximization (EM) algorithm[50] to estimate the proportion of risk genes (δ), and the proportions of risk variants, η_c's. Once the parameters were estimated, we computed the BF of a gene i, as the ratio of the probability of all its variant counts, X_{ij}'s, under $U_i = 1$ vs. $U_i = 0$ (see Methods). We note that the log-BF of a gene only depends on the η_c parameters and can be partitioned as the sum of contributions of each variant (see Equation 8 and 9 in Methods). This partition allows us to assess the contribution of individual variants, or variant groups, to the evidence of a gene.

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

We simulated data by using the genetic architecture information from ASD, to compare MIRAGE with existing methods in identifying risk genes from rare variants. To demonstrate the power of MIRAGE even in small samples, we fixed sample sizes at 3,000 cases and 3,000 controls. We simulated data of 1,000 genes, with the proportion of risk genes, δ, varying from 0.02, 0.05, 0.1, to 0.2. For simplicity, we assumed every gene has the same number of variants (100 variants per gene), however, the results were similar if we vary the number of variants from 50 to 500 (**Figure S2**). For a risk gene, its variants fall into three functional categories, mimicing benign missense variants (60% of variants), damaging missense variants (30%) and Loss-of-function (LoF) variants (10%), respectively. The proportions of risk variants, η_c, are 0.05, 0.2 and 0.5 for the three categories. Once the risk status of a variant is sampled, the allelic counts of the variant in cases and controls would follow Poisson distributions. The ratio of the two Poisson rates is the relative risk of that variant, sampled from Gamma distributions. Based on earlier exome sequencing studies of ASD[33,51], we set the mean relative risks, $\bar{\gamma} = 3$ for the first two variant categories and 5 for the last one.

We compared MIRAGE to several variations of burden tests (a baseline Burden test, CMC[16] and ASUM[52]), SKAT and ACAT. The baseline burden test tested all variants within a gene.

In practice, Burden test is often applied to different categories of variants of a gene separately to increase the power. We thus considered two other versions of burden test as well. The first version tested each of the three variant categories separately, then used the minimum p value of the three, adjusting for three tests. The second combined the three burden p values from the variant categories using Fisher's method. The results of these two tests in simulations were similar to the baseline Burden test (**Figure S1**), so we considered only the baseline version here.

We compared the performance of the methods in distinguishing risk from non-risk genes, using the ROC curves ($\delta = 0.02$ and $\delta = 0.1$ in **Figures 2A–2B**, $\delta = 0.05$ and $\delta = 0.2$ in **Figure S2**).

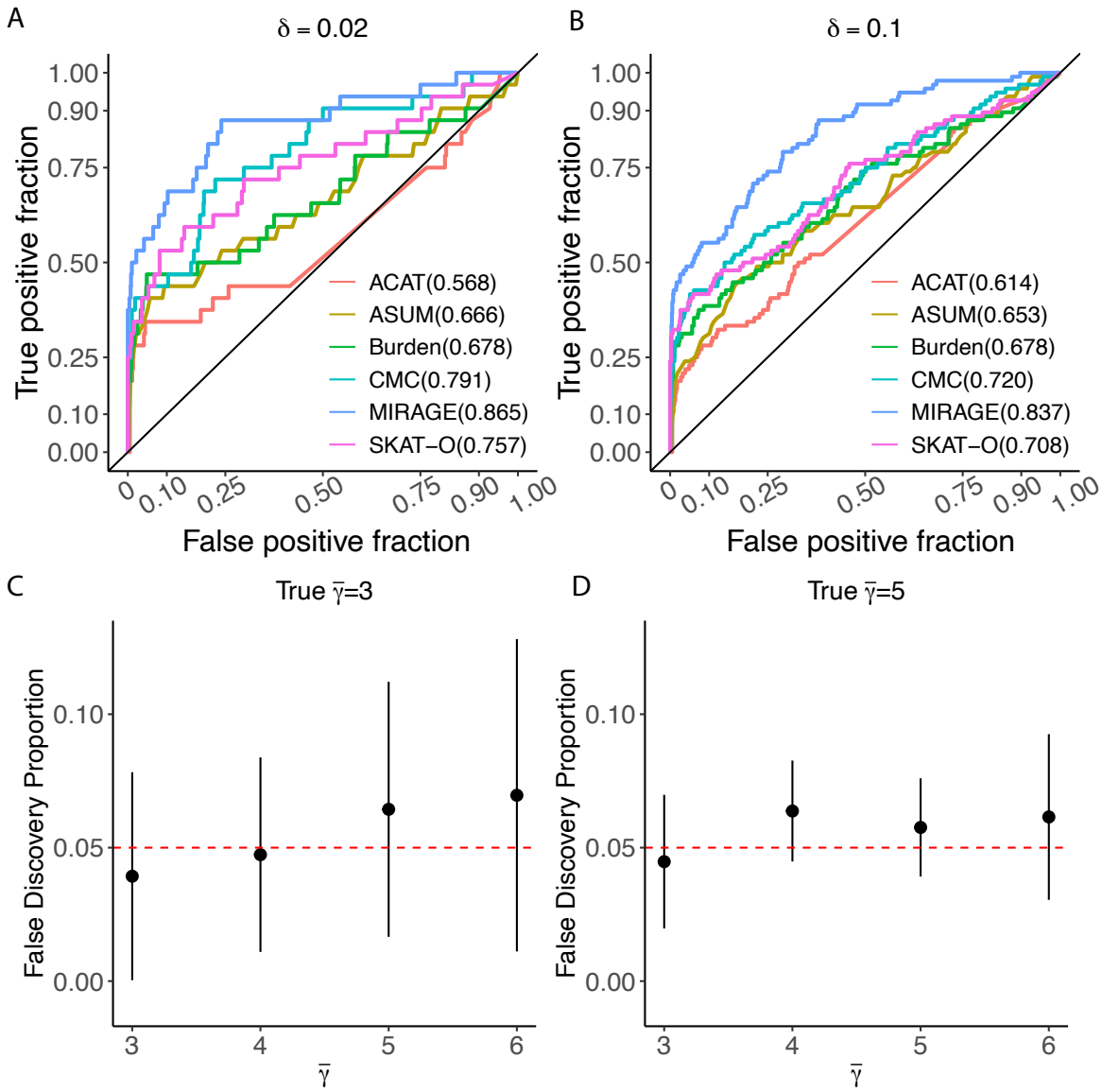

Figure 2: Simulation results.
(A) ROC curves at $\delta = 0.02$ of different methods for classifying risk genes. We simulated 1,000 genes with varying proportion of risk genes. AUC values are shown in the bracket. Solid black reference line is in the diagonal. **(B)** ROC curves at $\delta = 0.1$ of different methods for classifying risk genes. **(C)** False discovery rate(FDR) calibration by MIRAGE. We simulated 20 datasets with true $\bar{\gamma}$(relative risk) = 3. MIRAGE used $\bar{\gamma}$ from 3 to 6. The False Discovery Proportion under each specified value of $\bar{\gamma}$ is shown. Red dashed line is the target FDR level (0.05). **(D)** FDR calibration by MIRAGE at $\bar{\gamma} = 5$

Under all the settings, we found that MIRAGE has Area Under Receiver Operating Characteristic (AUROC) above 80%, substantially outperforming all other methods. To appreciate how much of this difference translates into the difference of power of detecting risk genes, we estimated the number of genes found by each gene at FDR < 0.1 (for MIRAGE, we used Bayesian FDR). We found that the power of MIRAGE is the highest in all settings, and $\sim 20\%$ higher than the second best method (SKAT-O) in the simulations (**Figure S3**).

MIRAGE uses a Bayesian approach to controlling FDR. We performed additional simulations to assess if the Bayesian FDR is calibrated, and whether the FDR is sensitive to misspecification of $\bar{\gamma}$, the average relative risk of risk variants. To make simulations simpler, we ran similar simulations as before, but used a common value of $\bar{\gamma}$ for all three variant categories. In simulations, we used a range of values, from 3 to 6, as $\bar{\gamma}$. When running MIRAGE, we assumed that the true value of $\bar{\gamma}$ was unknown and ran MIRAGE using the value ranging from 3 to 6. Our results were generally robust to the value of this parameter, and the Bayesian FDR was close to the true False Discovery Proportions ($\bar{\gamma} = 3, 5$ in **Figures 2C–2D**, $\bar{\gamma} = 4, 6$ in **Figure S4**).

3.3 MIRAGE identifies putative risk genes of ASD

We applied MIRAGE to WES data of 14,578 trios of children affected with ASD and their parents, combining data from Autism Sequencing Consortium and SPARK[36]. We treated transmitted parental alleles as "cases" and non-transmitted ones as "controls"[33]. We considered only rare variants with minor allele frequency (MAF) (both within cohort MAF and gnomad non-neuro allele frequency) below 0.1%, and filtered all synonymous variants. To annotate variants, we used LOEUF for LoF mutations[40], and MPC scores (Missense badness, PolyPhen-2, and Constrain scores) for missense variants[41]. LOEUF quantifies how much a gene tolerates loss-of-function mutations and is widely used. MPC has been shown to enrich risk variants for ASD in previous studies[37,41]. We included a total of six variant groups defined by the LOEUF decile and MPC score: decile 1 (low LOEUF, the lower, the more deleterious), deciles 2 and 3 (med LOEUF), and deciles greater than 3 (high LOEUF), MPC≥ 2 (high MPC), $2 >$ MPC ≥ 1 (med MPC) and MPC < 1 (low MPC).

We first confirmed that the ASD dataset was challenging for current methods. We applied Burden tests and ACAT to all genes. We used two definitions of burden: 1) burden in all LoF variants; and 2) burden in LoF and missense variants at MPC > 2. In ACAT, we conducted a binomial test on each variant within a gene, comparing the number of transmitted and non-transmitted variants. The statistics of variants within a gene were then combined using the ACAT method, with the variant weights determined by the MAFs[40], as described previously[23]. The QQ plots of p values showed that none of these methods detected any signal (**Figure 3A**). As other baseline methods, we also considered the single-variant test, as well as a gene level test using minimum p value of all variants (adjusting for the number of variants). Neither test showed any significant findings (**Figure S5**).

To run MIRAGE, we used an earlier estimate based on *de novo* mutations that 5% of genes are ASD risk genes[53]. Given that the signals in the transmission data is considerably weaker than *de novo* mutations[36], we think this estimate is more accurate than the value inferred from transmission data alone. Using this proportion of risk genes, MIRAGE estimated the proportions of risk variants across the six variant categories (**Figure 3B**). The LoF variants

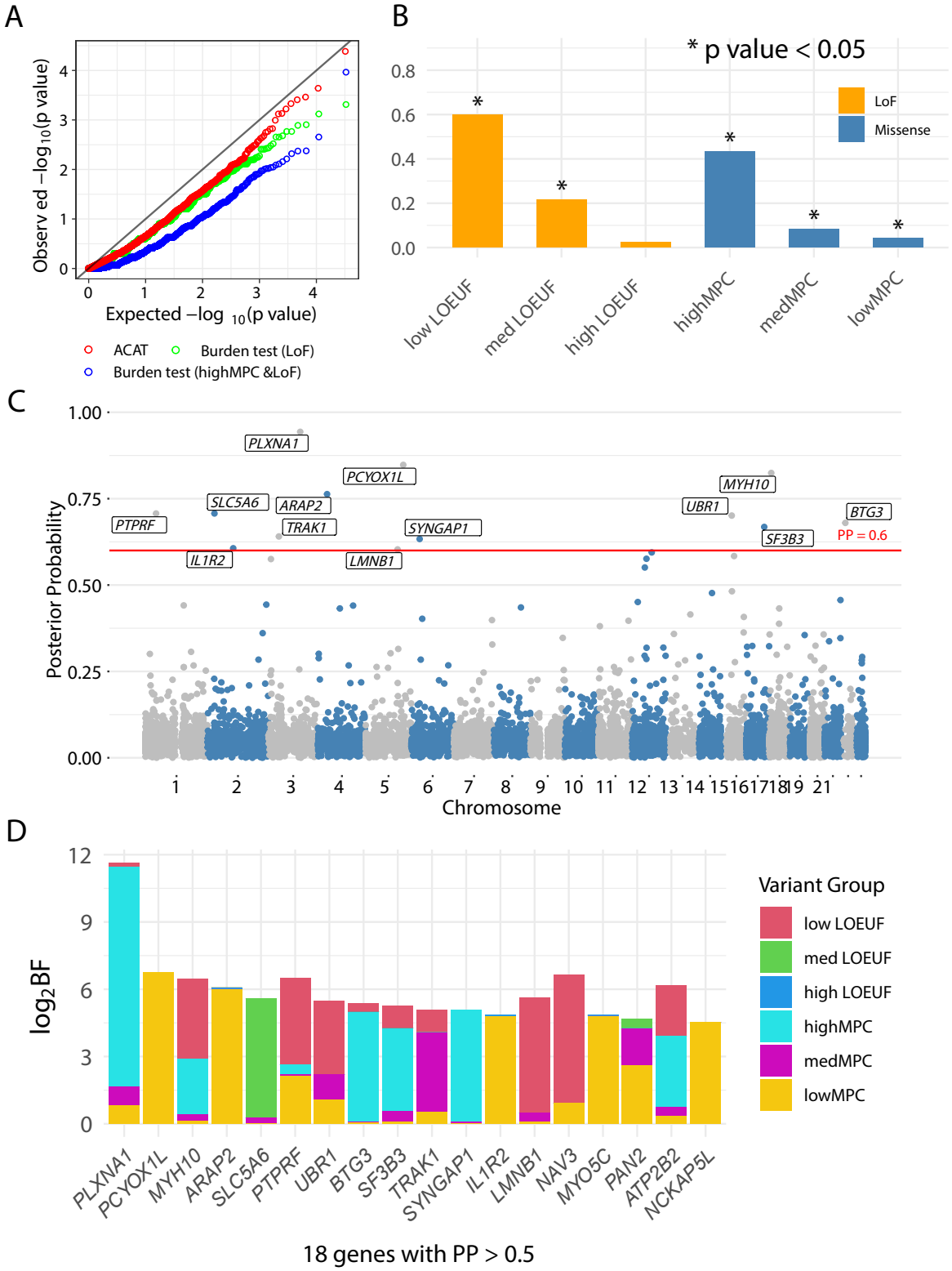

Figure 3: Summary of MIRAGE results from ASD data analysis.
(A) QQ plot of all genes in the whole genome by ACAT and burden tests. **(B)** The proportion of risk variants in 6 variant categories. LoF categories: LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction) decile = 1 (low), LOEUF decile = 2–3 (med), LOEUF decile = 4–10 (high). Missense categories: MPC (Missense badness, PolyPhen-2, and Constrain scores) ≥ 2 (high), $2 > \text{MPC} \geq 1$ (med), MPC < 1 (low). **(C)** Manhattan plot of posterior probabilities (PP) for all gene analyzed. Genes with PP > 0.6 are labeled. **(D)** The $\log_2$BF (BF is defined in **Figure 1A**) of a gene is partitioned into contributions of variants in each group. Shown are results of 18 genes at PP > 0.5, genes are ordered by their PPs. The negative $\log_2$BF of a variant group is truncated to 0.

in genes with high LOEUF score showed the highest proportions, with ∼60% of variants being risk variants. In the missense variant groups, high MPC had the largest risk variant proportion of 43%, and the last two categories showed much lower proportions. These estimates were generally in line with the expected deleteriousness of these variant annotations.

With the estimated parameters, we calculated BFs and posterior probabilities (PP) for every gene and performed Bayesian FDR control. At PP > 0.5, MIRAGE identified 18 putative ASD risk genes (**Figure 3C**, **Table S1**). We verified that the vast majority of the variants supporting these genes have very low LD (**Figure S6**), and pruning of the few remaining variants in LD have small effects on the results (see Methods). At PP > 0.7, MIRAGE identified seven putative ASD risk genes. These results thus supported higher sensitivity of MIRAGE in detecting risk genes than existing rare variant association methods.

To better understand the MIRAGE results, we evaluated the contributions the variant groups to the associations of these genes with ASD. Following earlier work[36], we partitioned the evidence of each gene, in terms of $\log_2$BF, into contributions of the six variant groups, in all genes at PP > 0.5 (**Figure 3D**). We found that the variant group(s) driving association signals in each gene varied considerably. For instance, *PLXNA1* was predominantly driven by the high MPC group of missense variants, while the association signals for *MYH10* was largely attributed to LoF variants in genes with low LOEUF. In about half of the genes, two or more variant groups made substantial contributions. Overall, the LoF variants from genes with Low/Med LOEUF and the High MPC variant group drove the associations in most of the genes. Together, these results highlighted that MIRAGE was able to effectively combine statistical signals across multiple variant groups to discover risk genes.

Finally, we evaluated the MIRAGE results using different functional annotations, including pLI for LoF variants, and AlphaMissense[42] for missense variants and pLI for LoF variants, and MPC for missense variants. The results from using these annotations broadly agreed with the results here. The estimated parameters reflected the expected severity of variants in the functional categories, and the discovered genes showed enrichment in ASD-related gene sets (**Figures S13**–**S16**)

3.4 Candidate genes found by MIRAGE are supported by multiple lines of evidence

We evaluated the candidate genes by assessing the enrichment of ASD-related gene sets. We selected the 18 genes at PP > 0.5, and for comparison, the same number of top genes by burden tests and ACAT. The ASD-related gene sets include known ASD genes curated by SFARI[44] and from de novo mutation studies using TADA[36]; risk genes of Intellectual Disability (ID)[45] and Schizophrenia (SCZ)[46]; relevant biological processes including Post-synaptic density (PSD)[46] and FMRP target genes[46]; evolutionarily constrained genes from RVIS[47] Haplo-insufficient (HI) genes[47] and major depressive disorder (MDD)[48]. We note that the SFARI and TADA genes were derived from independent datasets. While some samples in our dataset were included in earlier studies, the transmission data were not used previously. We found that the top MIRAGE genes were enriched in most ASD related gene sets (**Figure 4A**). For example, $\sim 35\%$ of MIRAGE candidate genes were likely ASD genes by SFARI ($p = 6.7 \times 10^{-5}$, Fisher's Exact Test). The top genes from other methods showed lower or no enrichment. These enrichment results thus strongly supported the likely roles of the candidate genes in ASD.

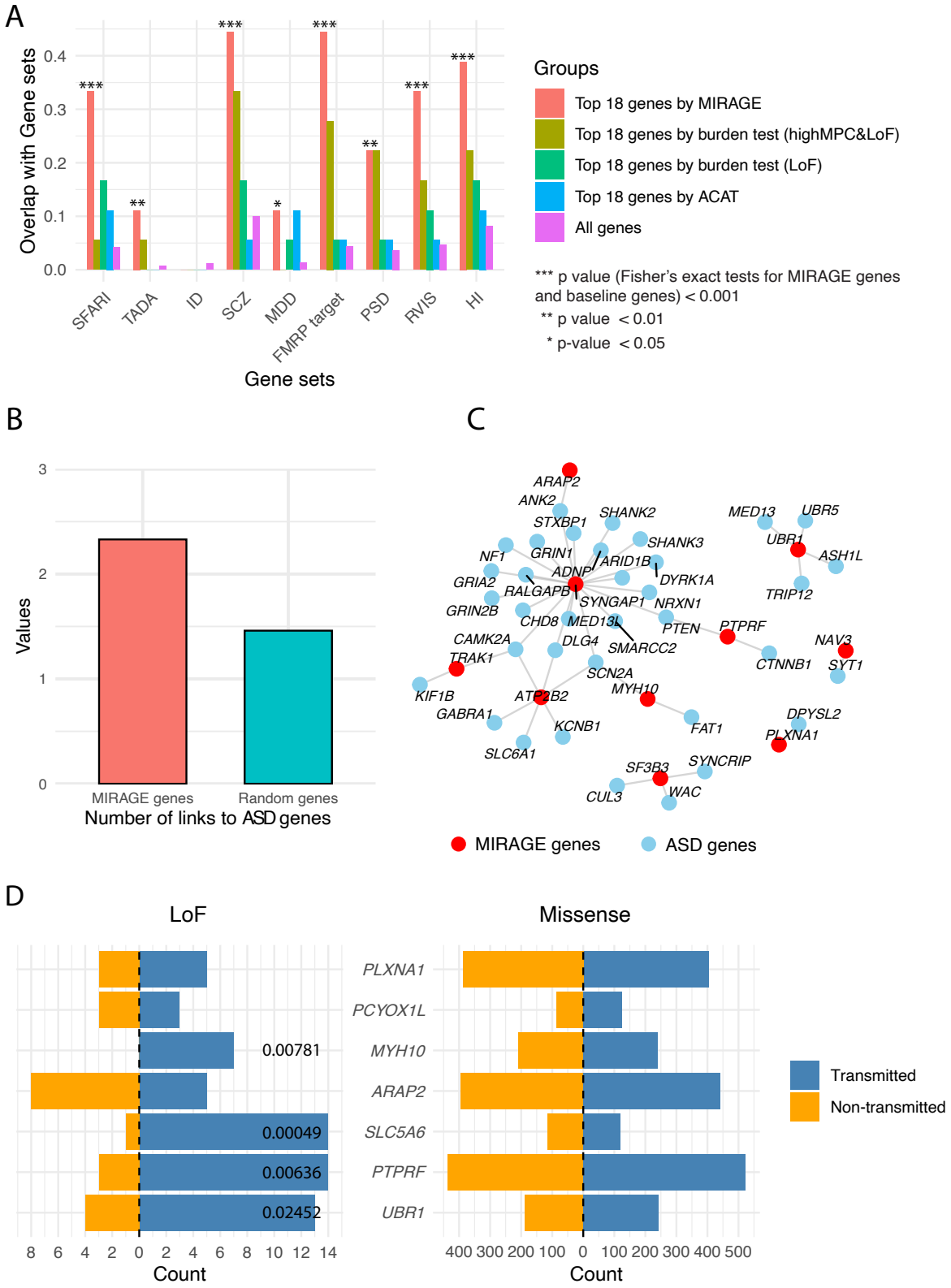

Figure 4: Evaluation of candidate genes from MIRAGE.
(A) Enrichment of ASD-related gene sets in the top 18 genes found by MIRAGE, burden tests and ACAT. Enrichment test was based on Fisher's exact test. SFARI genes: Categories 1 and 2. TADA high confidence genes: FDR < 0.05. 'All genes' refer to the entire set of genes analyzed, as a baseline. **(B)** Number of links from MIRAGE genes to TADA genes. **(C)** Gene interaction network for MIRAGE genes and TADA genes, only genes with connections are shown here. **(D)** The transmitted and non-transmitted counts of LoF (left) and missense (right) variants, for 7 genes with MIRAGE PP > 0.7. Numbers near the bars are p-values testing transmission disequilibrium, using the binomial test (only shown $p < 0.05$).

We next examined the function and plausibility of the top 7 genes using the more stringent cutoff, PP > 0.7 (Table 1). 6 out of 7 genes were annotated by one or more ASD-related gene sets. Among the six genes, *PLXNA1* were found by in earlier ASD studies[34,36]. *MYH10* was a ASD risk gene according to SFARI (score 2, Strong candidate). It was supported by multiple *de novo* mutation studies[54–57]. *PTPRF* and *ARAP2* have known or related functions in neurodevelopment and/or other neuropsychiatric disorders. For example, both of them have been implicated in SCZ. We discussed the literature support of these genes in Discussion.

To further validate and characterize the functions of the candidate genes, we used STRING[58], a widely used resource for gene interactions, to assess the connectivity of our top genes with known ASD genes. Our rationale is that true risk genes would tend to be connected with known risk genes through physical or other forms of functional interactions. We assessed the top 18 genes found by MIRAGE at PP > 0.5, and compared the connectivity with 100 randomly chosen genes. We found that a MIRAGE gene is connected to 2.33 ASD genes[36] on average, significantly higher than random genes (1.46) ($p = 0.004$, Poisson test, **Figure 4B**). These connections provided important clues of how the identified risk genes from MIRAGE may affect the ASD risk (**Figure 4C**). For example, *PTPRF* (PP $= 0.71$) encodes a receptor-type protein tyrosine phosphatase. It is linked to known ASD genes, *PTPN* and *CTNNB1*, two genes important for PI3K signaling and Wnt signaling, respectively. These connections thus suggested that *PTPRF* may act on ASD by modulating the PI3K and/or WNT signaling, two pathways important for ASD[59–62]. As another example, *UBR1* (PP $= 0.7$) is linked to several ASD genes, including *UBR5* and *TRIP12*. All three genes have functions in ubiquitination[63–65] and regulation of protein turnover[66–68].

We evaluated the statistical support of a few genes in more details. In MIRAGE, the evidence of a gene, $\log_2$BF, is the sum of $\log_2$BF of individual variants, allowing us to quantify the contribution of individual variants. For this analysis, we plotted the transmitted and non-transmitted counts of each variant, and its $\log_2$BF. We highlighted here the result of *ARAP2*, a signaling gene. Its association with ASD was largely driven by two missense variants that are 4bp away (**Figure 5**). The two variants are located in the PH3 domain, a domain involved in binding phosphatidylinositol phosphates, an important class of signaling molecules. In other top genes, the association signals tend to be more diffuse, with multiple supporting variants (**Figures S7–S12**).

For comparison, we investigated transmission disequilibrium for the LoF and missense variants, each as a group, of the top seven genes. No apparent burden was observed in the missense variants (**Figure 4D**, right). We note that MIRAGE is able to detect signals in the missense variants even when there is no overall burden between transmitted and non-transmitted variants, e.g. for *PLXNA1* and *ARAP2* (**Figure 3D** and **Figure 4D**). While disequilibrium was observed in the LoF group in four genes (at $p < 0.05$) (**Figure 4D**, left), the p values would not pass the genomewide multiple testing threshold.

These results together supported the key rationales of MIRAGE. First, by modeling the heterogeneity of variant effects, MIRAGE allows a small number of variants to drive the results, and is also able to leverage the collective signal across many variants. This benefit is clear in the case of missense variants, where burden test failed to detect any burden (**Figure 4D**, right). Secondly, by borrowing information across genes, MIRAGE is able to learn to put more emphasis on more deleterious variant groups. This allows MIRAGE to extract statistical signals from relatively modest burden of LoF variants (**Figure 4D**, left).

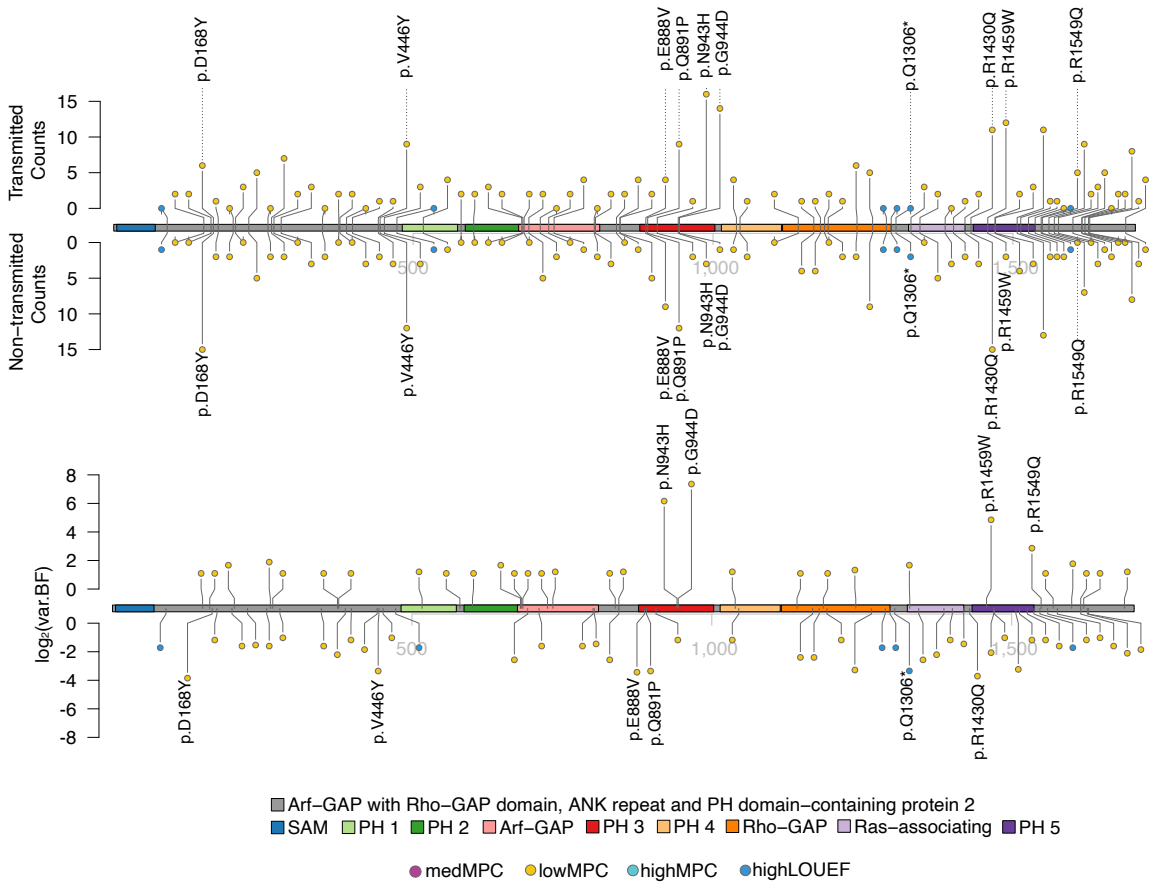

Figure 5: The lollipop plot of *ARAP2* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5 (BF is defined in **Figure 1A**), were plotted. The variants with BF > 3 and BF < 0.1 were labeled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

Table 1: Genes identified by MIRAGE (posterior probability > 0.7). PP (Posterior Probability). FDR (False Discovery Rate). The dash (-) indicates no data.

Gene	PP	FDR	ASD related gene sets	Functions/Roles
PLXNA1	0.94	0.06	PSD[46], FMRP target[46], RVIS[47], HI[47], SCZ[46]	Receptor for semaphorins, signals that guide axons and neurons during development
PCYOX1L	0.85	0.10	MDD[48]	Enzyme involved in polyamine metabolism. Associated with Freemartinism, sexual development
MYH10	0.82	0.13	TADA[36], PSD[46], FMRP target[46], RVIS[47], HI[47], SCZ[46], SFARI[44]	Non-muscle myosin IIB, involved in cell adhesion and migration. Microcephaly, developmental delay, known ASD gene
ARAP2	0.76	0.16	FMRP target[46], HI[47], SCZ[46]	A protein in Rho-GTPase signaling, implicated in ASD
SLC5A6	0.71	0.18	-	Transports essential vitamins and cofactors into cells. Developmental delay, motor neuropathies
PTPRF	0.71	0.20	FMRP target[46], RVIS[47], HI[47], SCZ[46], ADHD[69]	Protein tyrosine phosphatase family. Synapse formation and differentiation.
UBR1	0.70	0.22	HI[47], SFARI[44]	E3 ubiquitin ligase. Intellectual disability.

4 Discussion

We proposed a Bayesian method, MIRAGE, for rare variant association test. MIRAGE addresses two key 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 rigorous framework to leverage functional annotations of variants in identifying risk genes. Simulations confirmed the effectiveness of our method. In application to a WES dataset of ASD, while existing methods failed to detect significant associations, MIRAGE identified a number of candidate genes. The top genes were highly enriched with ASD-related gene sets. Most of the six candidate genes, at PP > 0.7, are either reported ASD genes, or have functions in related phenotypes or neurodevelopment, representing plausible ASD risk genes.

How to effectively analyze rare variants is a key challenge of the field. The success of MIRAGE in the study of ASD offered some general lessons. First, the effects of rare variants are likely heterogeneous, and this is better captured by a sparse model where most rare variants have no effects on disease risks. One can see this point from our estimated fractions of risk variants, especially in missense variants (**Figure 3B**), and from the analysis of individual genes (**Figure 5**, **Figures S7–S12**). Secondly, using external information of variants is critical to improve the signal to noise ratio. Indeed, annotating function of variants is an active area of research, and some recent methods, e.g. those based on deep learning[70–72], may further boost the power of MIRAGE.

MIRAGE is related to some Bayesian statistical methods for analyzing rare variants[73–80]. These methods typically treat the effect sizes of variants as random variables, and some methods explicitly capture the dependency of the effect sizes on variant annotations. For example, in MiST, the prior mean of the effect sizes is a linear function of variant annotations[76]. In a method that generalizes SKAT, the prior variance of effect size is modeled as a function of variant annotations[77]. These methods, however, are not widely used in practice, likely due to several limitations. The effect sizes are often modeled as continuous random variables, similar to SKAT. As we have explained and demonstrated, the risk variants are likely sparse, especially for missense variants. So a mixture approach with the majority of variants as non-risk variants is likely more powerful. Secondly, the existing Bayesian methods analyze one gene at a time. We think it is difficult to learn the parameters about functional annotations from a single gene. For instance, a gene may have only one or few LoF variants, and there is simply not enough information to learn that LoF variants are important. Existing methods may deal with this issue by allowing user-specified weights. We think a better approach is to use Empirical Bayes to borrow information across many genes to estimate the hyperparameters. Lastly, some of these Bayesian methods rely on computationally expensive procedures and are thus difficult to use in practice to analyze thousands of genes. We think these limitations are largely addressed by MIRAGE, through its mixture model of variants, the EM algorithm to estimate hyperparameters from all genes, and the simple binomial distributions for variant counts.

When applying MIRAGE to trio-sequencing data, it is related to TDT on rare variants. Several methods have been proposed to do rare variant analysis on trio- and family-sequencing data, such as RV-TDT[31], PedGen[81], and rare variant Family-based association test (FBAT)[82,83]. These methods are often extensions of burden test, SKAT or other p value aggregation methods, while accounting for family structure. For example, in De et al[82] and in RV-TDT, the TDT or

FBAT statistics were modified by replacing allelic counts of a single variant with genetic burden across multiple variants. In other methods, one may first obtain family-based test statistic (e.g. FBAT) of a single variant, then aggregate over variants using SKAT, or p value aggregation methods[83]. Regardless of the specific approaches, all these methods are closely related to the existing approaches for rare variant analysis, and are thus subject to similar weakness as reviewed earlier.

We discussed in more details the functions of the identified ASD candidate genes. *PTPRF* is a member of protein tyrosine phosphatase (PTP) family. The subfamily containing *PTPRF* plays an important role in neural function such as dendrite and axon growth and guidance, synapse formation and differentiation[84]. A related protein, *PTPRF* interacting protein alpha 1, is a strong ASD risk gene by SFARI[44]. *ARAP2* is an adapter protein in Rho GTPase signaling. A number of studies have linked Rho GTPase signaling to the pathogenesis of ASD[85]. The gene family of *ARAP2* (Arf GAP) are important regulators of the Actin Cytoskeleton, and have been implicated in cell adhesion, motility and neurite outgrowth[86]. *SLC5A6*, which transports biotin and pantothenic acid to the brain, is implicated in brain development and developmental delay[87–89], while *UBR1* is associated with Johanson–Blizzard syndrome, a disorder characterized by cognitive impairment[90]. We presented the supporting evidence from literature of all genes at PP > 0.5 in **Table S2**.

Despite the strong support of our candidate genes from gene set enrichment analysis and literature evidence, it is important to experimentally validate the functions of the identified ASD risk genes in future studies. This may involve CRISPR perturbations of these genes in appropriate cellular models[91], followed by assessment of molecular and cellular phenotypes, including gene expression, neuronal morphology and synaptic activity.

We briefly commented here about how to run MIRAGE. First, we note that the main parameters of MIRAGE are δ, the proportion of risk genes, and η_c, the proportion of risk variants in each functional category. While MIRAGE is able to estimate all the parameters using EM, in practice, it may be better to specify δ, if some approximate value can be provided. The model may have low identifiability when the signal is weak, in other words, different values of δ may give similar likelihood. Next, MIRAGE results depend on functional annotations used. What is the best annotation in a particular dataset may be unclear a priori. For example, while AlphaMissense works well in other contexts, it does not seem to work as well as MPC scores in our ASD dataset (**Figure 4A** vs. (**Figures S13–S14**)). Choosing what annotations to use may be an important consideration in applying MIRAGE.

MIRAGE can be further developed in several directions. First, MIRAGE was designed for case-control, or transmission studies. Extending it to quantitative traits would greatly broaden its applications. 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. Incorporating such covariates is thus an important next step. 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 have more flexible models of the prior probabilities of variants. This type of prior has been used successfully in GWAS[92,93] and in studies of *de novo* mutations[51].

Data and code availability

MIRAGE is available at https://xinhe-lab.github.io/mirage/.

Acknowledgements

This work was supported by the National Institutes of Health (NIH) under grants R01MH110531, R01MH106575, and R01HL163523 (to X.H.). We thank other members of He's groups for helpful comments on the work and the manuscript.

Author Contributions

Conceptualization, X.H.; Methodology, S.H., X.H.; Software, G.W., X.S., S.H.; Data Analysis, S.H., X.S., L.S., X.X.; Data Curation, S.H., X.S., L.S., F.K.S., G.W., L.L., N.K., S.Z., Y.L., W.S., H.N., J.B.; Writing, S.H., X.S., X.H.; Supervision, X.H., J.B.; Project Administration, X.H.; Funding Acquisition, X.H..

Declaration of Interests

The authors declare no competing interests.

Web Resources

ASC data: dbGaP (accession ID: phs000298). SPARK data: https://www.sfari.org/resource/spark/.

References

1. 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(7145), 661–78. https://doi.org/10.1038/nature05911.
2. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, and Yang J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics 101(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
3. Shendure J, Findlay GM, and Snyder MW. (2019). Genomic Medicine-Progress, Pitfalls, and Promise. Cell 177(1), 45–57. https://doi.org/10.1016/j.cell.2019.02.003.
4. Gibson G. (2011). Rare and common variants: twenty arguments. Nature Reviews Genetics 13, 135–45. https://doi.org/10.1038/nrg3118.
5. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, and Sunyaev SR. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proceedings of the National Academy of Sciences of the USA 106, 3871–6. https://doi.org/10.1073/pnas.0812824106.
6. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science 335, 823–8. https://doi.org/10.1126/science.1215040.

7. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, and McVean GA1GPC. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65. https://doi.org/10.1038/nature11632.

8. Wainschtein P, Jain D, Zheng Z, Group TAW, Precision Medicine (TOPMed) Consortium NTO for, Cupples LA, Shadyab AH, McKnight B, Shoemaker BM, Mitchell BD, et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nature Genetics 54(3), 263–73. https://doi.org/10.1038/s41588-021-00997-7.

9. Rocheleau G, Clarke SL, Auguste G, Hasbani NR, Morrison AC, Heath AS, Bielak LF, Iyer KR, Young EP, Stitziel NO, et al. (2024). Rare variant contribution to the heritability of coronary artery disease. Nature Communications 15(1), 8741. https://doi.org/10.1038/s41467-024-52939-6.

10. Consortium 1GP, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. (2015). A global reference for human genetic variation. Nature 526(7571), 68–74. https://doi.org/10.1038/nature15393.

11. Lee S, Abecasis GR, Boehnke M, and Lin X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. American Journal of Human Genetics 95(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.

12. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, Salerno WJ, Lancour D, Ma Y, Renton AE, et al. (2020). Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. Molecular Psychiatry 25, 1859–75. https://doi.org/10.1038/s41380-018-0112-7.

13. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, Churchhouse C, Chambert K, Chandler SD, Daly MJ, et al. (2020). Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nature Neuroscience 23(2), 185–93. https://doi.org/10.1038/s41593-019-0564-3.

14. Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD, Bass N, Bigdeli TB, Breen G, Bromet EJ, et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature 604(7906), 509–16. https://doi.org/10.1038/s41586-022-04556-w.

15. Cirulli ET. (2016). The Increasing Importance of Gene-Based Analyses. PLoS Genetics 12(4), e1005852. https://doi.org/10.1371/journal.pgen.1005852.

16. Li B and Leal SM. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. American Journal of Human Genetics 83(3), 311–21. https://doi.org/10.1016/j.ajhg.2008.06.024.

17. Madsen B and Browning S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2), e1000384. https://doi.org/10.1371/journal.pgen.1000384.

18. Price A, Kryukov G, Bakker P de, Purcell S, Staples J, Wei L, and Sunyaev S. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6), 832–8. https://doi.org/10.1016/j.ajhg.2010.04.005.

19. Derkach A, Lawless J, and Sun L. (2013). Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genet Epidemiol 37(1), 110–21. https://doi.org/10.1002/gepi.21689.

20. Ionita-Laza I, Buxbaum J, Laird N, and Lange C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet 7(2), e1001289. https://doi.org/10.1371/journal.pgen.1001289.

21. Wu M, Lee S, Cai T, Li Y, Boehnke M, and Lin X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029.

22. Zhang H and Wu Z. (2023). The generalized Fisher's combination and accurate p-value calculation under dependence. Biometrics 79(2), 1159–72. https://doi.org/10.1111/biom.13634.

23. Liu Y, Chen S, Li Z, Morrison A, Boerwinkle E, and Lin X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet 104(3), 410–21. https://doi.org/10.1016/j.ajhg.2019.01.002.

24. Liu W, Guo Y, and Liu Z. (2021). An Omnibus Test for Detecting Multiple Phenotype Associations Based on GWAS Summary Level Data. Front Genet 12, 644419. https://doi.org/10.3389/fgene.2021.644419.

25. Zhou W, Zhao Z, Nielsen J, Fritsche L, LeFaive J, Gagliano Taliun S, Bi W, Gabrielsen M, Daly M, Neale B, et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat Genet 52(6), 634–9. https://doi.org/10.1038/s41588-020-0621-6.

26. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, Chen H, Sun R, Dey R, Arnett DK, Aslibekyan S, et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat Genet 52(9), 969–83. https://doi.org/10.1038/s41588-020-0676-4.

27. Backman J, Li A, Marcketta A, Sun D, Mbatchou J, Kessler M, Benner C, Liu D, Locke A, Balasubramanian S, et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature 599(7886), 628–34. https://doi.org/10.1038/s41586-021-04103-z.

28. Timpson N, Greenwood C, and Soranzo Nea. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet 19, 110–24. https://doi.org/10.1038/nrg.2017.101.

29. Bodea C, Neale B, Ripke S, Consortium IIG, Daly M, Devlin B, and Roeder K. (2016). A Method to Exploit the Structure of Genetic Ancestry Space to Enhance Case-Control Studies. Am J Hum Genet 98(5), 857–68. https://doi.org/10.1016/j.ajhg.2016.02.025.

30. M. Artomov, A.A. Loboda, M.N. Artyomov, and M.J. Daly. (2024). Public platform with 39,472 exome control samples enables association studies without genotype sharing. Nat Genet 56(2), 327–35. https://doi.org/10.1038/s41588-023-01637-y.

31. Z. He, B.J. O'Roak, J.D. Smith, G. Wang, S. Hooker, R.L. Santos-Cortez, B. Li, M. Kan, N. Krumm, D.A. Nickerson, et al. (2014). Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am J Hum Genet 94(1), 33–46. https://doi.org/10.1016/j.ajhg.2013.11.021.

32. E.A. Ruiz-Narváez and H. Campos. (2004). Transmission disequilibrium test (TDT) for case-control studies. Eur J Hum Genet 12(2), 105–14. https://doi.org/10.1038/sj.ejhg.5201099.

33. He X, Sanders S, Liu L, Rubeis SD, Lim E, Sutcliffe J, Schellenberg G, Gibbs R, Daly M, Buxbaum J, et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet 9(8), e1003671. https://doi.org/10.1371/journal.pgen.1003671.

34. Zhou X, Feliciano P, Shu C, Wang T, Astrovskaya I, Hall JB, Obiajulu JU, Wright JR, Murali SC, Xu SX, et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat Genet 54, 1305–19. https://doi.org/10.1038/s41588-022-01148-2.

35. Nguyen HT, Bryois J, Kim A, Dobbyn A, Huckins LM, Munoz-Manchado AB, Ruderfer DM, Genovese G, Fromer M, Xu X, et al. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. Genome medicine 9(1), 114. https://doi.org/10.1186/s13073-017-0497-y.

36. Fu J, Satterstrom F, Peng M, Brand H, Collins R, Dong S, Wamsley B, Klei L, Wang L, Hao S, et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nature Genetics 54(9), 1320–31. https://doi.org/10.1038/s41588-022-01104-0.

37. Satterstrom F, Kosmicki J, Wang J, Breen M, De Rubeis S, An J, Peng M, Collins R, Grove J, Klei L, et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell 180(3), 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036.

38. Kosmicki J, Samocha K, Howrigan D, Sanders S, Slowikowski K, Lek M, Karczewski K, Cutler D, Devlin B, Roeder K, et al. (2017). Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat Genet 49(4), 504–10. https://doi.org/10.1038/ng.3789.

39. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B, et al. (2016). Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536(7616), 285–91. https://doi.org/10.1038/nature19057.

40. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–43. https://doi.org/10.1038/s41586-020-2308-7.

41. Samocha K, Kosmicki J, Karczewski K, O'Donnell-Luria A, Pierce-Hoffman E, MacArthur D, Neale B, and Daly M. (2017). Regional missense constraint improves variant deleteriousness prediction. bioRxiv. https://doi.org/10.1101/148353.

42. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381(6664), eadg7492. https://doi.org/10.1126/science.adg7492.

43. J. Ou and L.J. Zhu. (2019). trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nat Methods 16, 453–4. https://doi.org/10.1038/s41592-019-0430-y.

44. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, and Packer A. (2013). SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism 4(1), 36. https://doi.org/10.1186/2040-2392-4-36.

45. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. American Journal of Human Genetics 94(5), 677–94. https://doi.org/10.1016/j.ajhg.2014.03.018.

46. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kähler A, et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature 506(7487), 185–90. https://doi.org/10.1038/nature12975.

47. Petrovski S, Wang Q, Heinzen EL, Allen AS, and Goldstein DB. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLOS Genetics 9(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709.

48. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. (2025). Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. Cell 188(3), 640–652.e9. https://doi.org/10.1016/j.cell.2024.12.002.

49. Whittemore A. (2007). A Bayesian false discovery rate for multiple testing. J. Appl. Stat. 34(1), 1–9. https://doi.org/10.1080/02664760600994745.

50. Dempster A, Laird N, and Rubin D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

51. Liu Y, Liang Y, Cicek AE, Li Z, Li J, Muhle RA, Krenzer M, Mei Y, Wang Y, Knoblauch N, et al. (2018). A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies. American Journal of Human Genetics 102(6), 1031–47. https://doi.org/10.1016/j.ajhg.2018.03.023.

52. Han F and Pan W. (2010). A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Human Heredity 70, 42–54. https://doi.org/10.1159/000288704.

53. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature 515(7526), 209–15. https://doi.org/10.1038/nature13772.

54. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, Georgieva L, Rees E, Palta P, Ruderfer DM, et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature 506, 179–84. https://doi.org/10.1038/nature12929.

55. I. Hlushchenko, P. Khanal, A. Abouelezz, V. O. Paavilainen, and P. Hotulainen. (2018). ASD-Associated De Novo Mutations in Five Actin Regulators Show Both Shared and Distinct Defects in Dendritic Spines and Inhibitory Synapses in Cultured Hippocampal Neurons. Front Cell Neurosci 12, 217. https://doi.org/10.3389/fncel.2018.00217.

56. Tuzovic L, Yu L, Zeng W, Li X, Lu H, Lu HM, Gonzalez KD, and Chung WK. (2013). A human de novo mutation in MYH10 phenocopies the loss of function mutation in mice. Rare Diseases 1. https://doi.org/10.4161/rdis.26144.

57. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature 515(7526), 216–21. https://doi.org/10.1038/nature13908.

58. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research 51(D1), D638–D646. https://doi.org/10.1093/nar/gkac1000.

59. Sharma A and Mehan S. (2021). Targeting PI3K-AKT/mTOR signaling in the prevention of autism. Neurochemistry International 147, 105067. https://doi.org/10.1016/j.neuint.2021.105067.

60. Enriquez-Barreto L and Morales M. (2016). The PI3K signaling pathway as a pharmacological target in Autism related disorders and Schizophrenia. Molecular and Cellular Therapies 4(2). https://doi.org/10.1186/s40591-016-0047-9.

61. Bae SM and Hong JY. (2018). The Wnt Signaling Pathway and Related Therapeutic Drugs in Autism Spectrum Disorder. Clinical Psychopharmacology and Neuroscience 16(2), 129–35. https://doi.org/10.9758/cpn.2018.16.2.129.

62. Park G, Jang WE, Kim S, Gonzales EL, Ji J, Choi S, Kim Y, Park JH, Mohammad HB, Bang G, et al. (2023). Dysregulation of the Wnt/-catenin signaling pathway via Rnf146 upregulation in a VPA-induced mouse model of autism spectrum disorder. Experimental Molecular Medicine 55(8), 1783–94. https://doi.org/10.1038/s12276-023-01065-2.

63. Pan M, Zheng Q, Wang T, Liang L, Mao J, Zuo C, Ding R, Ai H, Xie Y, Si D, et al. (2021). Structural insights into Ubr1-mediated N-degron polyubiquitination. Nature 600, 334–8. https://doi.org/10.1038/s41586-021-04099-w.

64. Hodáková Z, Grishkovskaya I, Brunner HL, Bolhuis DL, Belačić K, Schleiffer A, Kotisch H, Brown NG, and Haselbach D. (2023). Cryo-EM structure of the chain-elongating E3 ubiquitin ligase UBR5. The EMBO Journal 42(16), e113348. https://doi.org/10.15252/embj.2022113348.

65. Seo BA, Kim D, Hwang H, Kim MS, Ma SX, Kwon SH, Kweon SH, Wang H, Yoo JM, Choi S, et al. (2021). TRIP12 ubiquitination of glucocerebrosidase contributes to neurodegeneration in Parkinson's disease. Neuron 109(23), 3758–3774.e11. https://doi.org/10.1016/j.neuron.2021.09.031.

66. Rudi O, Hodáková Z, Farias Saad C, Winter N, Grishkovskaya I, Böhm J, Jarck G, Schleiffer A, Haselbach D, and Bachmair A. (2025). The UBR domain of plant Ubr1 homolog PRT6 accommodates basic and hydrophobic amino termini for substrate recognition. Journal of Molecular Biology 437(4), 168939. https://doi.org/10.1016/j.jmb.2025.168939.

67. Jiang H, He X, Feng D, Zhu X, and Zheng Y. (2015). RanGTP aids anaphase entry through Ubr5-mediated protein turnover. Proceedings of the National Academy of Sciences of the United States of America 112(41), E5628–E5637. https://doi.org/10.1073/pnas.1515902112.

68. Kaiho-Soma A, Akizuki Y, Igarashi K, Endo A, Shoda T, Kawase Y, Demizu Y, Naito M, Saeki Y, Tanaka K, et al. (2021). TRIP12 promotes small-molecule-induced degradation through K29/K48-branched ubiquitin chains. Molecular Cell 81(7), 1411–1424.e7. https://doi.org/10.1016/j.molcel.2021.01.023.

69. Demontis D, Walters GB, Athanasiadis G, Walters R, Therrien K, Nielsen TT, Farajzadeh L, Voloudakis G, Bendl J, Zeng B, et al. (2023). Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. Nature Genetics 55(2), 198–208. https://doi.org/10.1038/s41588-022-01285-8.

70. Sundaram L, Gao H, Padigepati S, McRae J, Li Y, Kosmicki J, Fritzilas N, Hakenberg J, Dutta A, Shon J, et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat Genet 50(8), 1161–70. https://doi.org/10.1038/s41588-018-0167-z.

71. K. Minton. (2023). Predicting variant pathogenicity with AlphaMissense. Nat Rev Genet 24, 804. https://doi.org/10.1038/s41576-023-00668-9.

72. N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet 55(9), 1512–22. https://doi.org/10.1038/s41588-023-01465-0.

73. N. Yi and D. Zhi. (2011). Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol 35(1), 57–69. https://doi.org/10.1002/gepi.20554.

74. M.A. Quintana, F.R. Schumacher, G. Casey, J.L. Bernstein, L. Li, and D.V. Conti. (2012). Incorporating prior biologic information for high-dimensional rare variant association studies. Hum Hered 74(3-4), 184–95. https://doi.org/10.1159/000346021.

75. B.A. Logsdon, J.Y. Dai, P.L. Auer, J.M. Johnsen, S.K. Ganesh, N.L. Smith, J.G. Wilson, R.P. Tracy, L.A. Lange, S. Jiao, et al. (2014). A variational Bayes discrete mixture test for rare variant association. Genet Epidemiol 38(1), 21–30. https://doi.org/10.1002/gepi.21772.

76. Sun J, Zheng Y, and Hsu L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol 37(4), 334–44. https://doi.org/10.1002/gepi.21717.

77. He Z, Xu B, Lee S, and Ionita-Laza I. (2017). Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in Metabochip data. Am J Hum Genet 101(3), 340–52. https://doi.org/10.1016/j.ajhg.2017.07.011.

78. G.R. Venkataraman, C. DeBoever, Y. Tanigawa, M. Aguirre, A.G. Ioannidis, H. Mostafavi, C.C.A. Spencer, T. Poterba, C.D. Bustamante, M.J. Daly, et al. (2021). Bayesian model comparison for rare-variant association studies. Am J Hum Genet 108(12), 2354–67. https://doi.org/10.1016/j.ajhg.2021.11.005.

79. H. Susak, L. Serra-Saurina, G. Demidov, R. Rabionet, L. Domènech, M. Bosio, F. Muyas, X. Estivill, G. Escaramís, and S. Ossowski. (2021). Efficient and flexible integration of variant characteristics in rare variant association studies using integrated nested Laplace approximation. PLoS Comput Biol 17(2), e1007784. https://doi.org/10.1371/journal.pcbi.1007784.

80. Y. Yang, S. Basu, and L. Zhang. (2021). A Bayesian hierarchically structured prior for rare-variant association testing. Genet Epidemiol 45(4), 413–24. https://doi.org/10.1002/gepi.22379.

81. D.J. Schaid, S.K. McDonnell, J.P. Sinnwell, and S.N. Thibodeau. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37(5), 409–18. https://doi.org/10.1002/gepi.21727.

82. G. De, W.K. Yip, I. Ionita-Laza, and N. Laird. (2013). Rare variant analysis for family-based design. PLoS One 8(1), e48495. https://doi.org/10.1371/journal.pone.0048495.

83. J. Hecker, F.W. Townes, P. Kachroo, C. Laurie, J. Lasky-Su, J. Ziniti, M.H. Cho, S.T. Weiss, N.M. Laird, and C. Lange. (2021). A unifying framework for rare variant association testing in family-based designs, including higher criticism approaches, SKATs, and burden tests. Bioinformatics 36(22-23), 5432–8. https://doi.org/10.1093/bioinformatics/btaa1055.

84. F. Cornejo, B. I. Cortés, G. M. Findlay, and G. I. Cancino. (2021). LAR Receptor Tyrosine Phosphatase Family in Healthy and Diseased Brain. Front Cell Dev Biol 9, 659951. https://doi.org/10.3389/fcell.2021.659951.

85. Guo D, Yang X, and Shi L. (2020). Rho GTPase Regulators and Effectors in Autism Spectrum Disorders: Animal Models and Insights for Therapeutics. Cells 9(4). https://doi.org/10.3390/cells9040835.

86. C. E. Tanna, L. B. Goss, C. G. Ludwig, and P. W. Chen. (2019). Arf GAPs as Regulators of the Actin Cytoskeleton—An Update. Int J Mol Sci 20(2), 442. https://doi.org/10.3390/ijms20020442.

87. Uchida Y, Ito K, Ohtsuki S, Kubo Y, Suzuki T, and Terasaki T. (2015). Major involvement of Na(+) -dependent multivitamin transporter (SLC5A6/SMVT) in uptake of biotin and pantothenic acid by human brain capillary endothelial cells. Journal of Neurochemistry 134(1), 97–112. https://doi.org/10.1111/jnc.13092.

88. Subramanian VS, Constantinescu AR, Benke PJ, and Said HM. (2017). Mutations in SLC5A6 associated with brain, immune, bone, and intestinal dysfunction in a young child. Case Reports in Human Genetics 136(2), 253–61. https://doi.org/10.1007/s00439-016-1751-x.

89. Holling T, Nampoothiri S, Tarhan B, Schneeberger PE, Vinayan KP, Yesodharan D, Roy AG, Radhakrishnan P, Alawi M, Rhodes L, et al. (2022). Novel biallelic variants expand the SLC5A6-related phenotypic spectrum. European Journal of Human Genetics 30(4), 439–49. https://doi.org/10.1038/s41431-021-01033-2.

90. Noli K, Aleysae N, Alzahrani I, Al-Ghamdi A, Alkazmi M, and Almasoudi A. (2024). Johanson-Blizzard syndrome caused by novel UBR1 mutation in four Saudi patients. JPGN Reports 5(2), 140–7. https://doi.org/10.1002/jpr3.12057.

91. Ensaldo-López A, Veiga-Rúa S, Carracedo Á, Allegue C, and Sánchez L. (2020). Experimental Models to Study Autism Spectrum Disorders: hiPSCs, Rodents and Zebrafish. Genes 11(11), 1376. https://doi.org/10.3390/genes11111376.

92. J. K. Pickrell. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet 94(4), 559–73. https://doi.org/10.1016/j.ajhg.2014.03.004.

93. X. Wen, Y. Lee, F. Luca, and R. Pique-Regi. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am J Hum Genet 98(6), 1114–29. https://doi.org/10.1016/j.ajhg.2016.03.029.

94. Chen S, Xiong J, Chen B, Zhang C, Deng X, He F, Yang L, Chen C, Peng J, and Yin F. (2022). Autism spectrum disorder and comorbid neurodevelopmental disorders (ASD-NDDs): Clinical and genetic profile of a pediatric cohort. Clin Chim Acta 524, 179–86. https://doi.org/10.1016/j.cca.2021.11.014.

95. Matsunami N, Hensel CH, Baird L, Stevens J, Otterud B, Leppert T, Varvil T, Hadley D, Glessner JT, Pellegrino R, et al. (2014). Identification of rare DNA sequence variants in high-risk autism families and their prevalence in a large case/control population. Molecular Autism 5(1), 5. https://doi.org/10.1186/2040-2392-5-5.

96. Wong YH, Lu AC, Wang YC, Cheng HC, Chang C, Chen PH, Yu JY, and Fann MJ. (2010). Protogenin defines a transition stage during embryonic neurogenesis and prevents precocious neuronal differentiation. Journal of Neuroscience 30(12), 4428–39. https://doi.org/10.1523/JNEUROSCI.0473-10.2010.

97. Yoshida F, Nagatomo R, Utsunomiya S, Kimura M, Shun S, Kono R, Kato Y, Nao Y, Maeda K, Koyama R, et al. (2024). Soluble form of Lingo2, an autism spectrum disorder-associated molecule, functions as an excitatory synapse organizer in neurons. Translational Psychiatry 14(1), 448. https://doi.org/10.1038/s41398-024-03167-5.

98. Ji G, Li S, Ye L, and Guan J. (2021). Gene Module Analysis Reveals Cell-Type Specificity and Potential Target Genes in Autism's Pathogenesis. Biomedicines 9(4), 410. https://doi.org/10.3390/biomedicines9040410.

99. Williams SM, An JY, Edson J, Watts M, Murigneux V, Whitehouse AJO, Jackson CJ, Bellgrove MA, Cristino AS, and Claudianos C. (2019). An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder. Molecular Psychiatry 24(11), 1707–19. https://doi.org/10.1038/s41380-018-0049-x.

100. Banks G, Lassi G, Hoerder-Suabedissen A, Tinarelli F, Simon MM, Wilcox A, Lau P, Lawson TN, Johnson S, Rutman A, et al. (2018). A missense mutation in Katnal1 underlies behavioural, neurological and ciliary anomalies. Molecular Psychiatry 23(3), 713–22. https://doi.org/10.1038/mp.2017.54.

Supplementary Information

Supplemental Figures

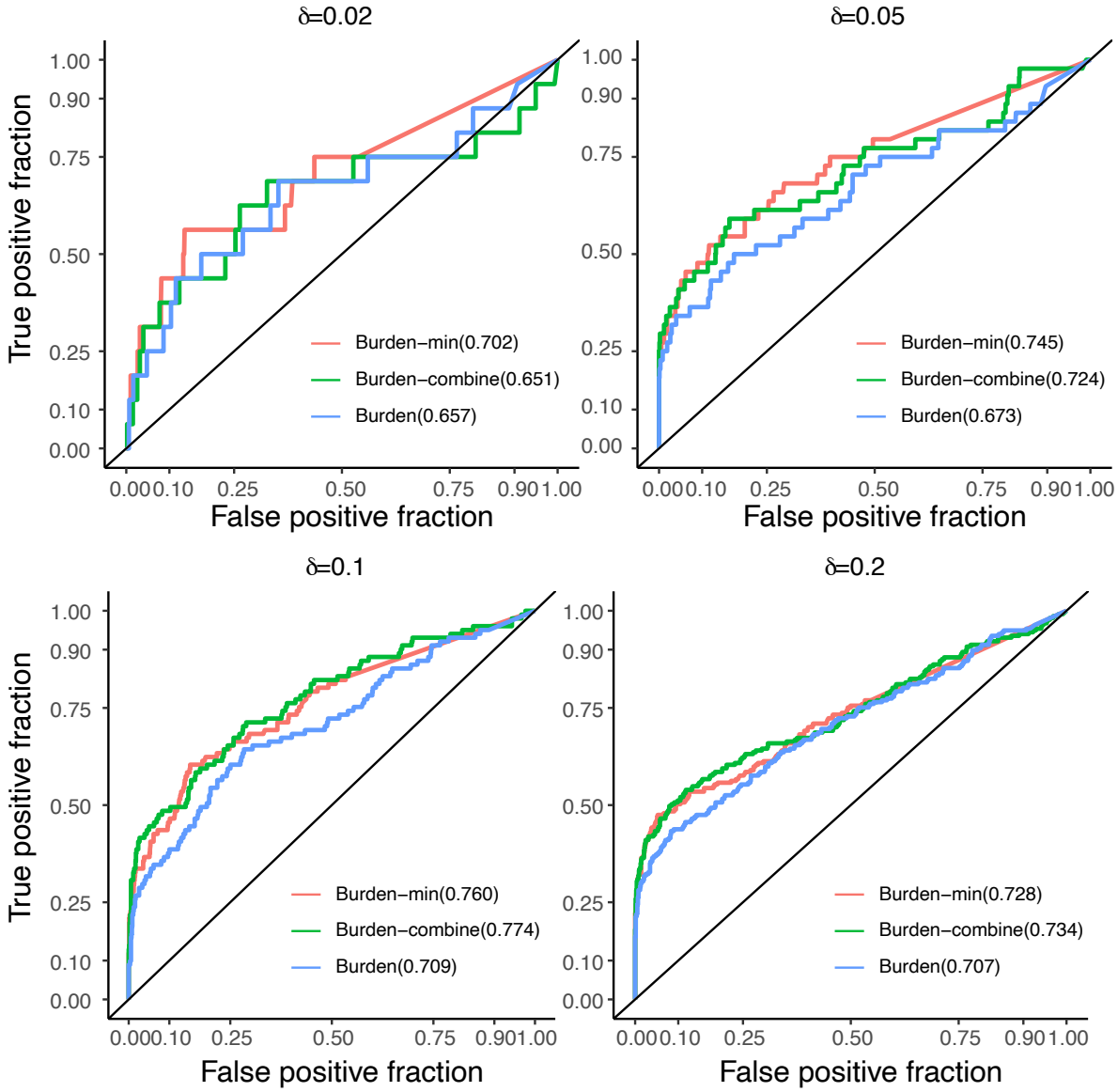

Figure S1: ROCs of several variations of burden test. Burden (basic version): calculates the p value for all variants collectively. Burden-min: calculates the p value for each variant category separately, applies Bonferroni correction to adjust the resulting three p values for each gene, and selects the minimum adjusted p value. Burden-combine: combines the original p values from the burden test across the three categories using Fisher's method.

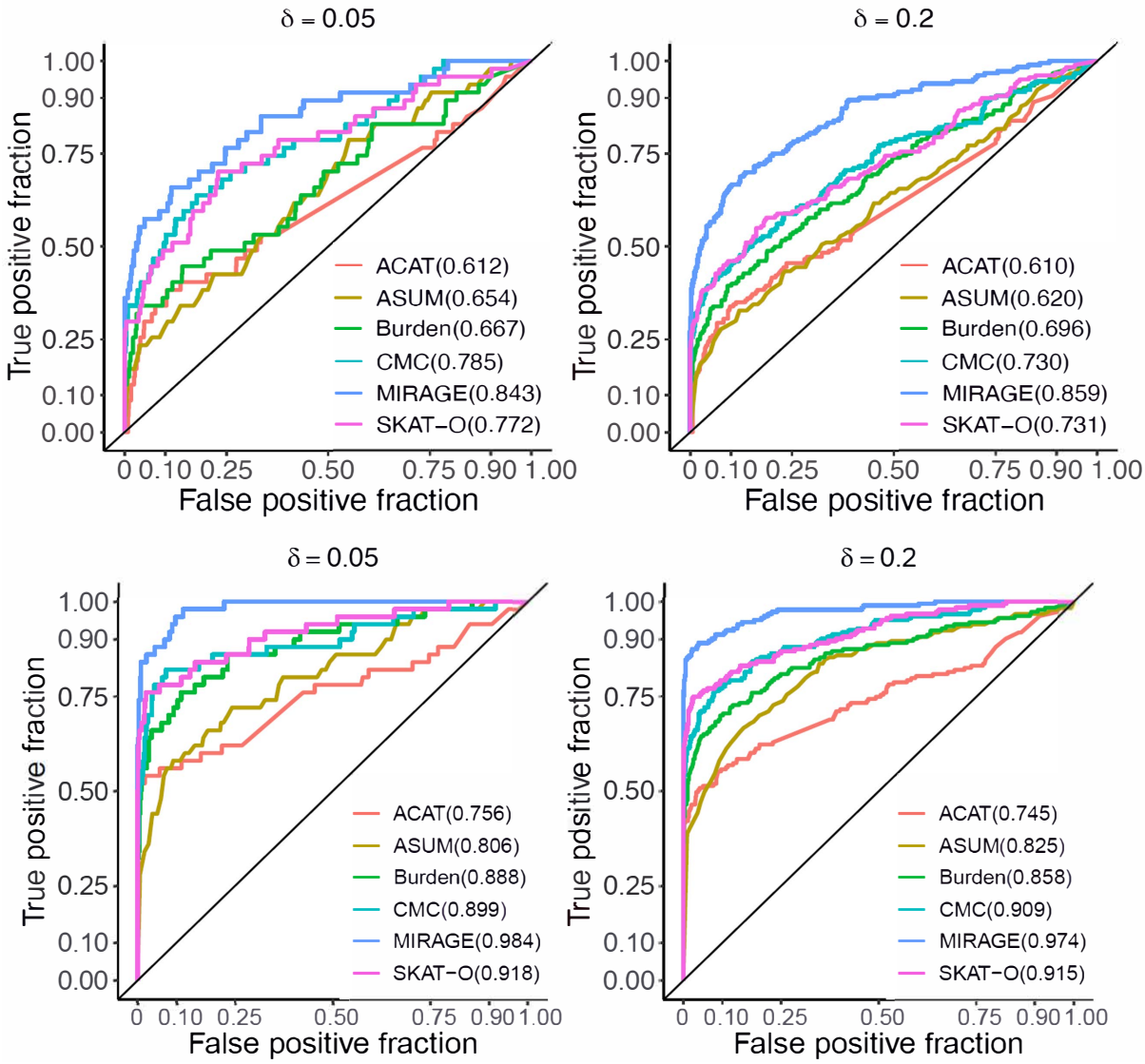

Figure S2: ROC curves of different methods for classifying risk genes. We simulate 1000 genes with varying proportion of risk genes, $\delta = 0.05, 0.2$. AUC values are shown in the bracket. Solid black reference line is in the diagonal. The two figures at the top: results from using the same number of variants, i.e., 100 in a gene, and the bottom two figures: results from sampling random number of variants between 50 and 500 per gene.

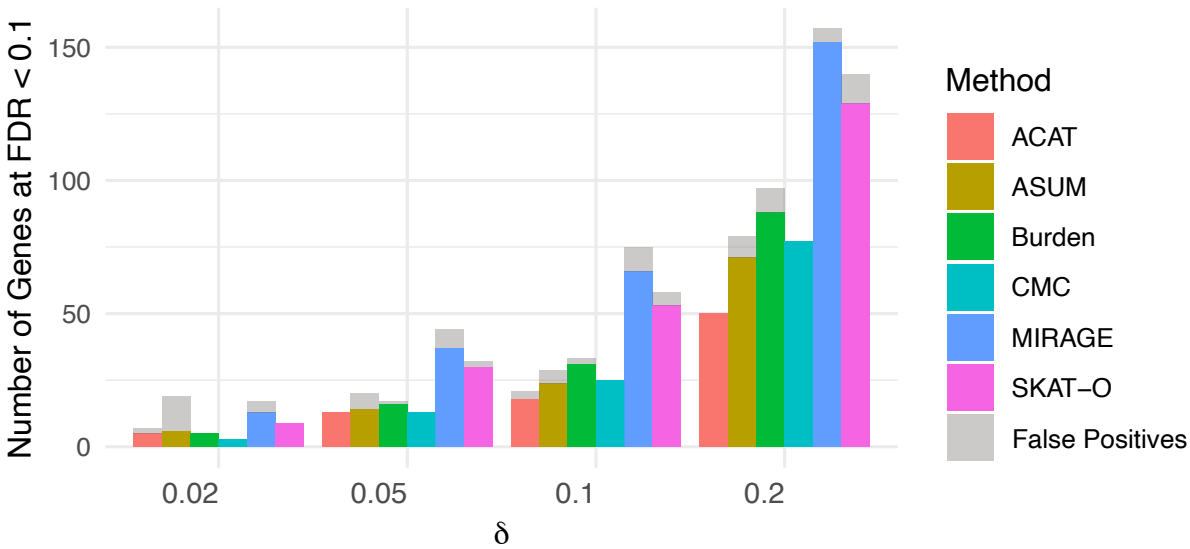

Figure S3: Number of true positive and false positive genes at FDR < 0.1 across different methods. The numbers of true positive and false positive genes for each δ and method are shown in the figure. Colored bars represent true positive genes, while grey bars indicate false positives

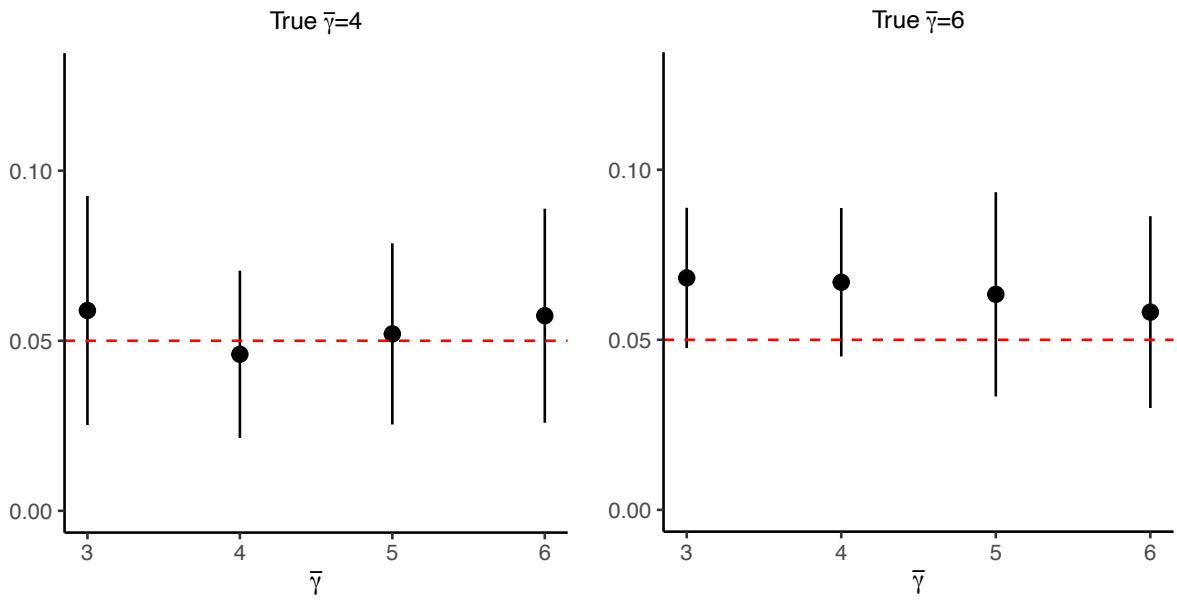

Figure S4: FDR calibration by MIRAGE. We simulated 20 datasets with true $\bar{\gamma} = 4, 6$, use mis-specified $\bar{\gamma}$ to calculate FDR. Red dashed line is the true FDR reference line.

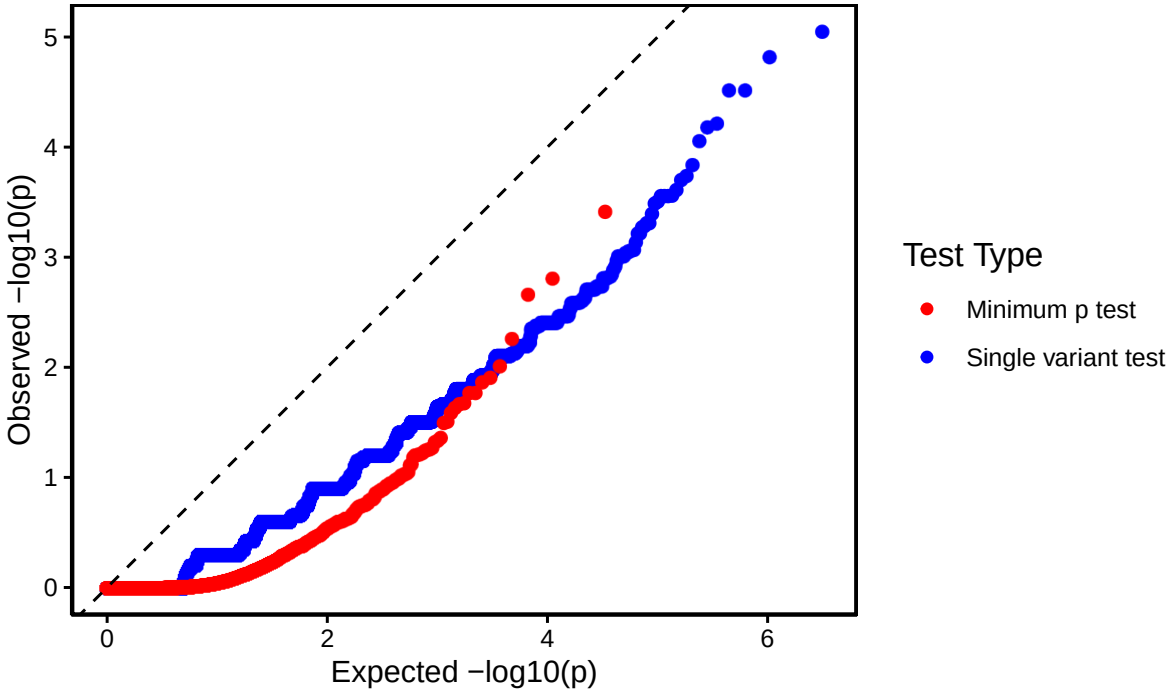

Figure S5: QQ plot of SNPs' p values from the single variant test and genes' p values from minimum-p test in the ASD dataset.

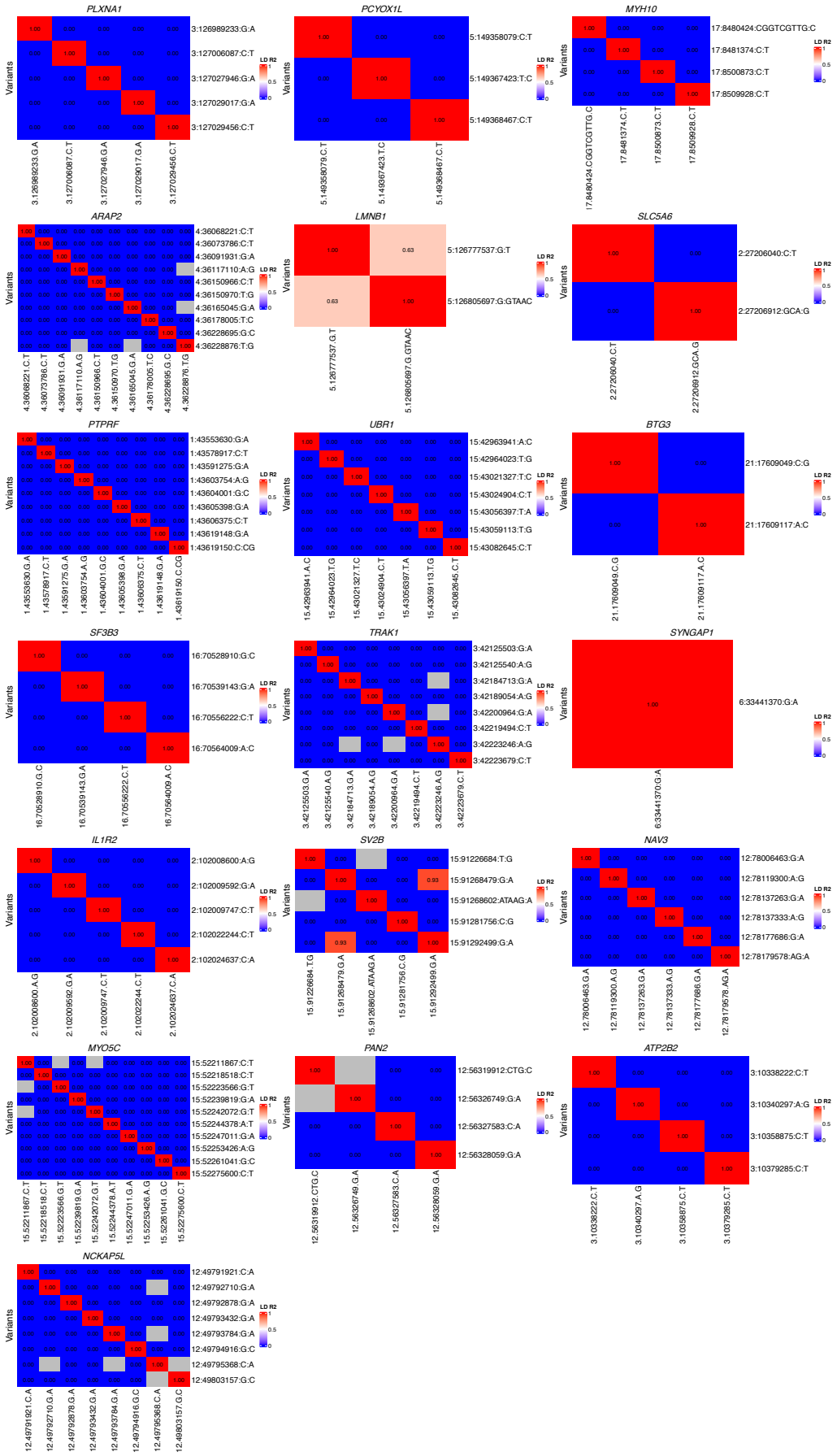

Figure S6: Heatmap showing LD (R^2) of the supporting variants of genes having PP > 0.5 Supporting variants of a gene are defined as the variants whose logBF > 1. Gray elements show the variant pairs whose allele frequencies were too low to compute LD.

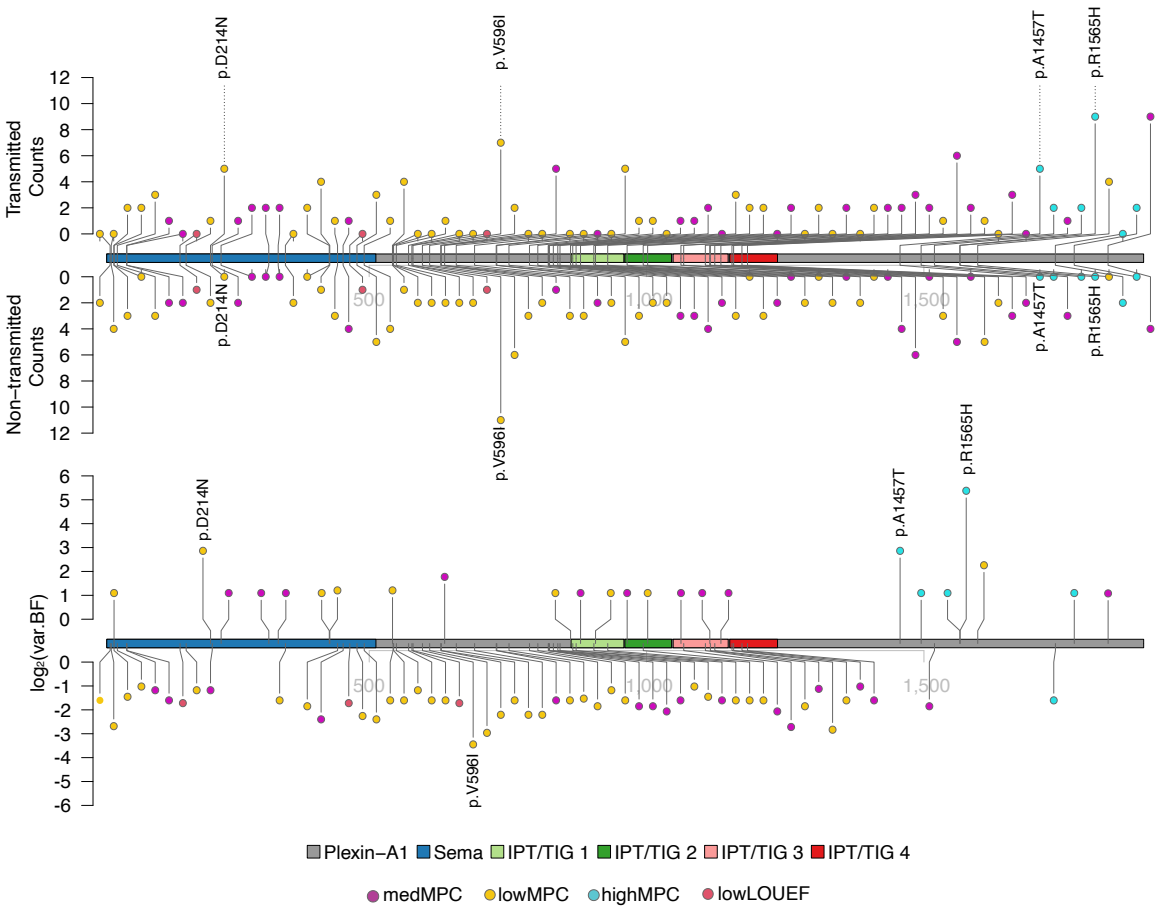

Figure S7: The lollipop plot of *PLXNA1* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 5 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

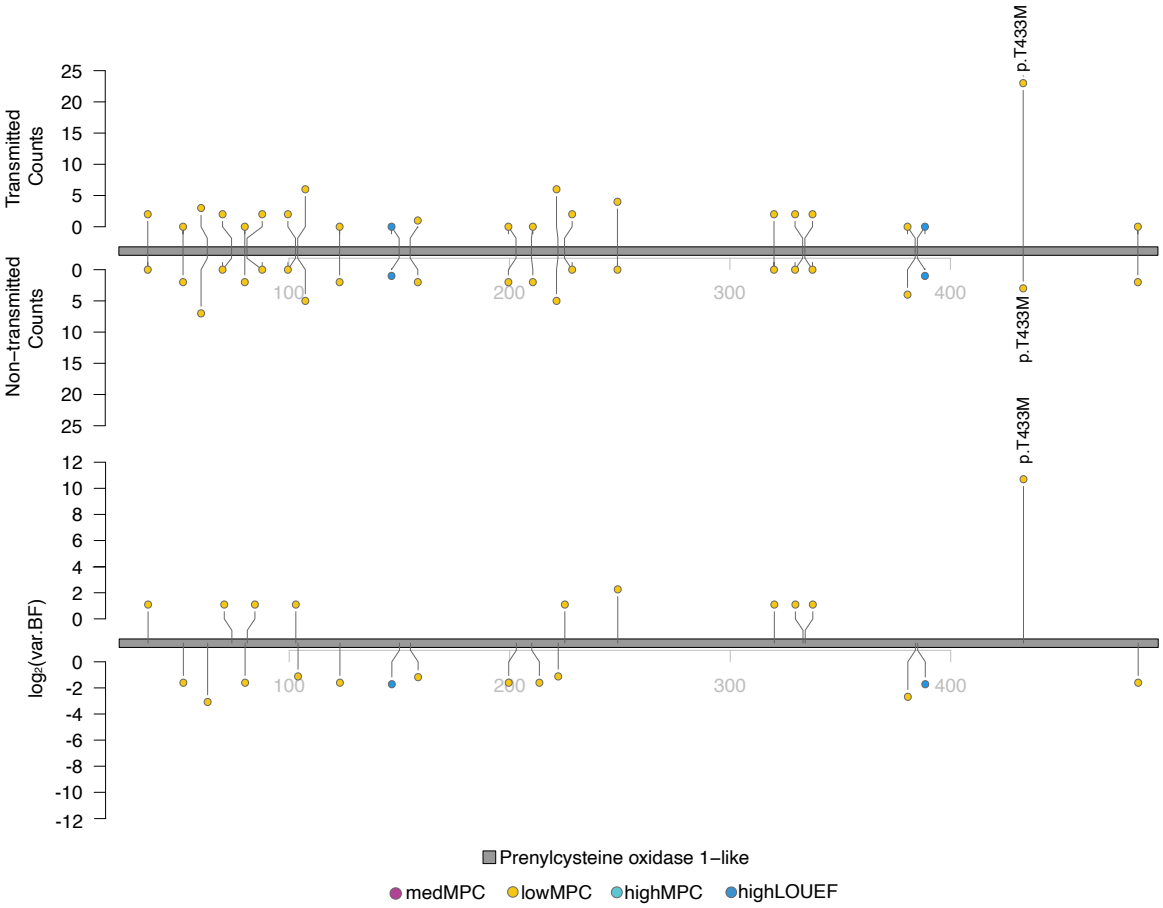

Figure S8: The lollipop plot of *PCYOX1L* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 5 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

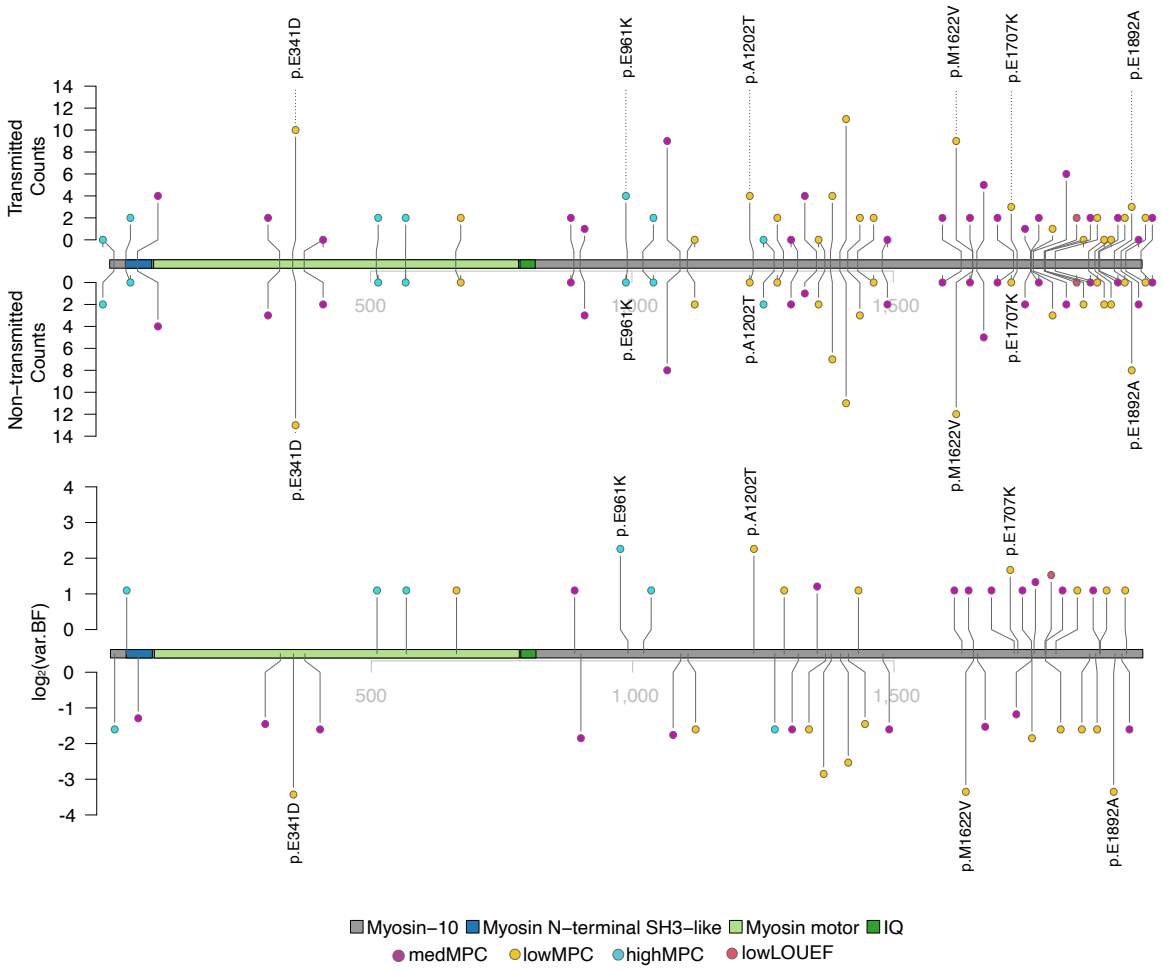

Figure S9: The lollipop plot of *MYH10* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 3 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

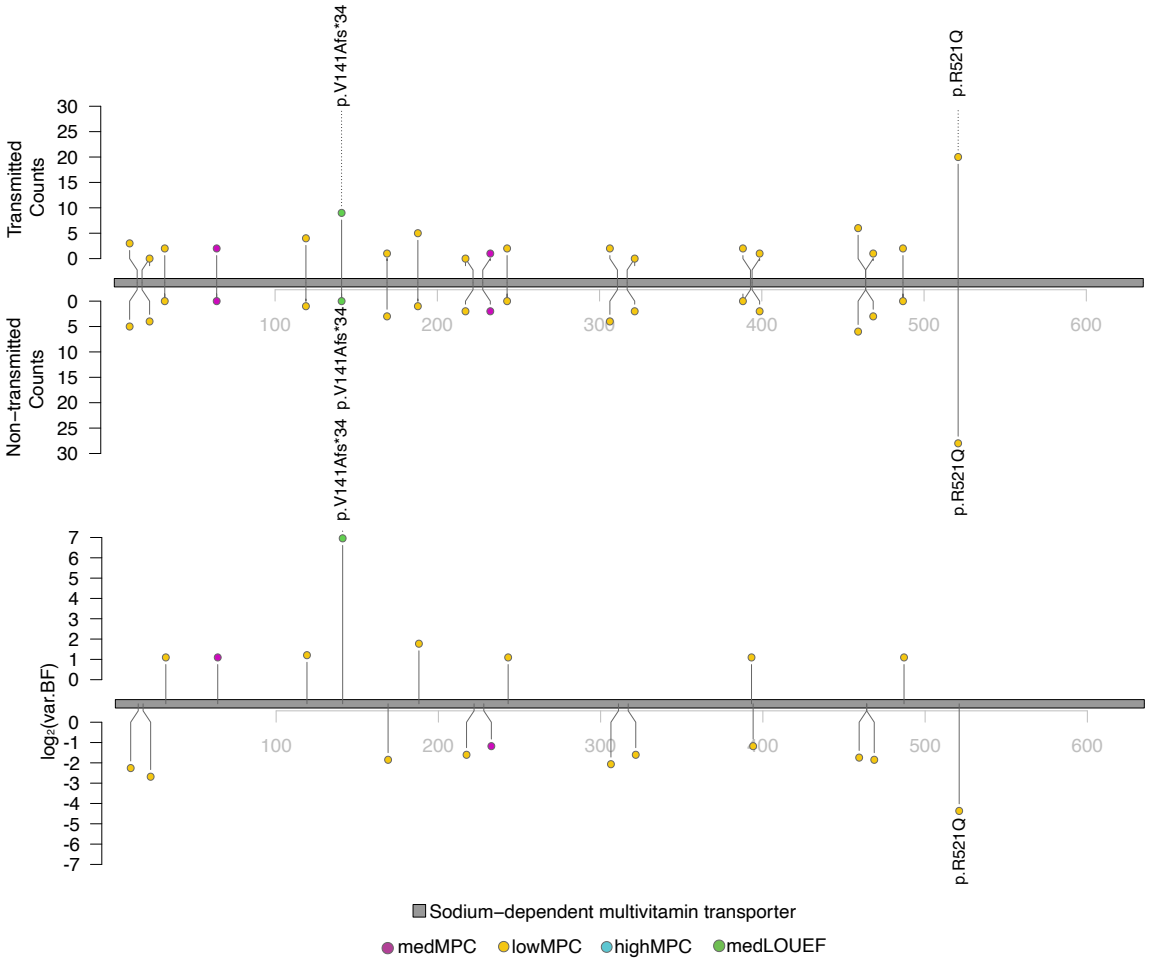

Figure S10: The lollipop plot of *SLC5A6* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 3 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

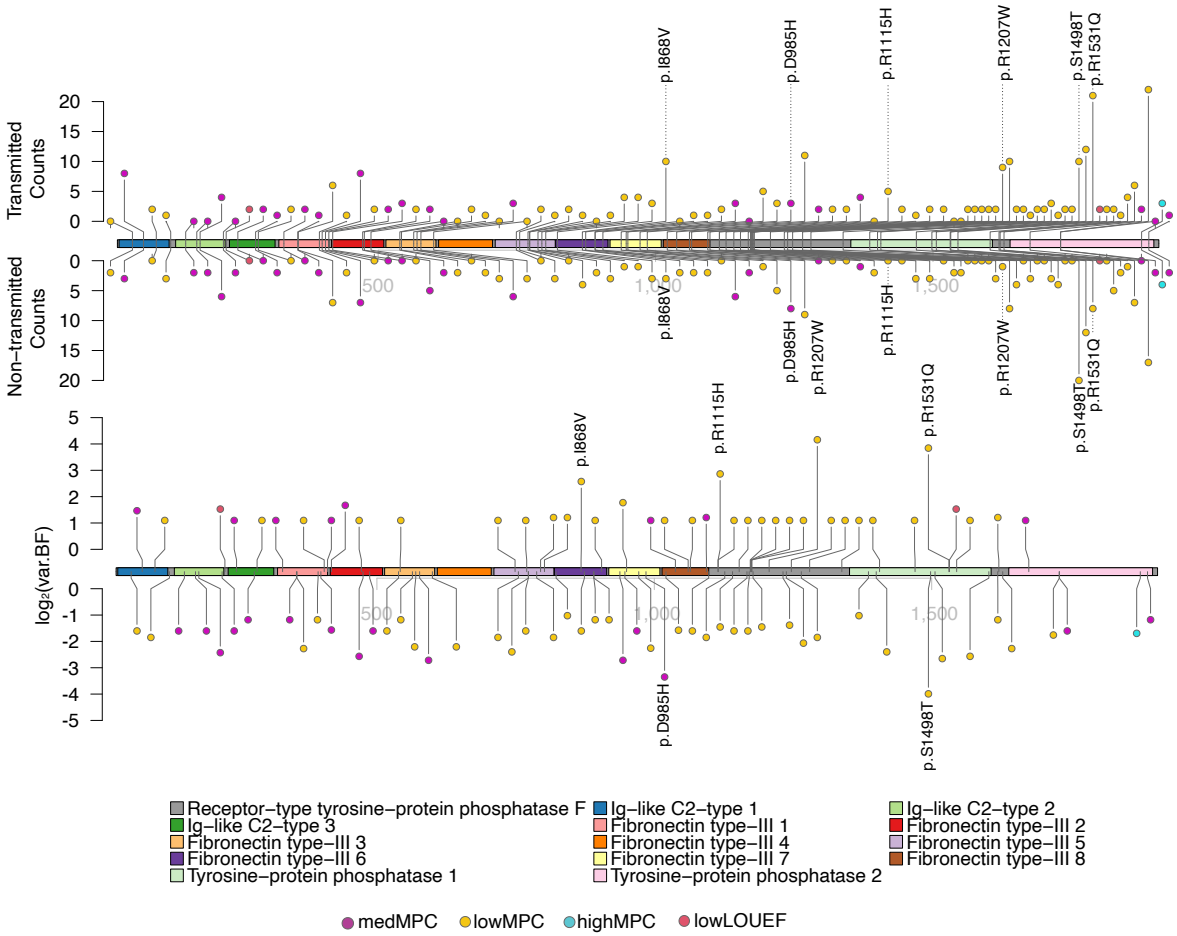

Figure S11: The lollipop plot of *PTPRF* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 5 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

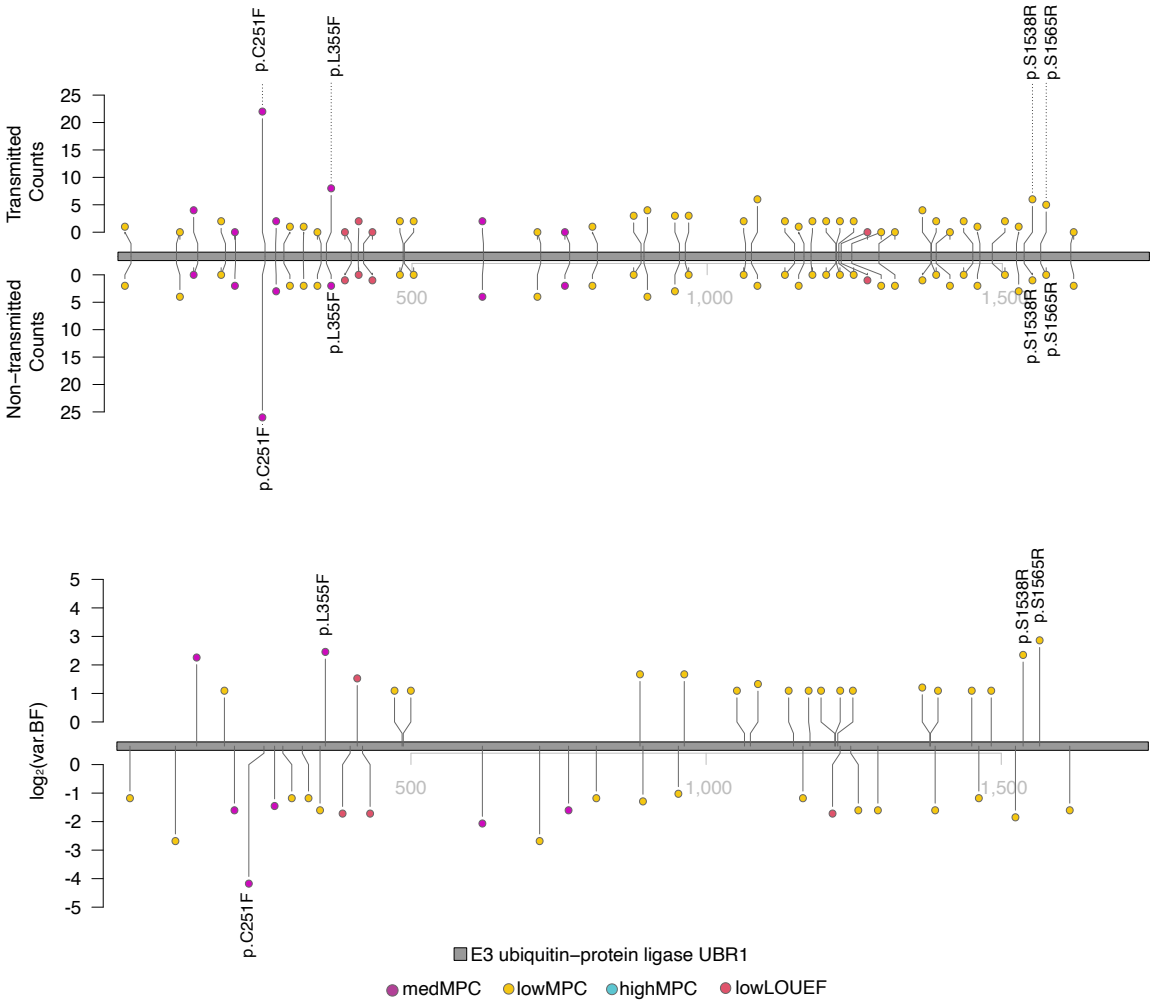

Figure S12: The lollipop plot of *UBR1* Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 5 and BF < 0.1 were labelled. LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

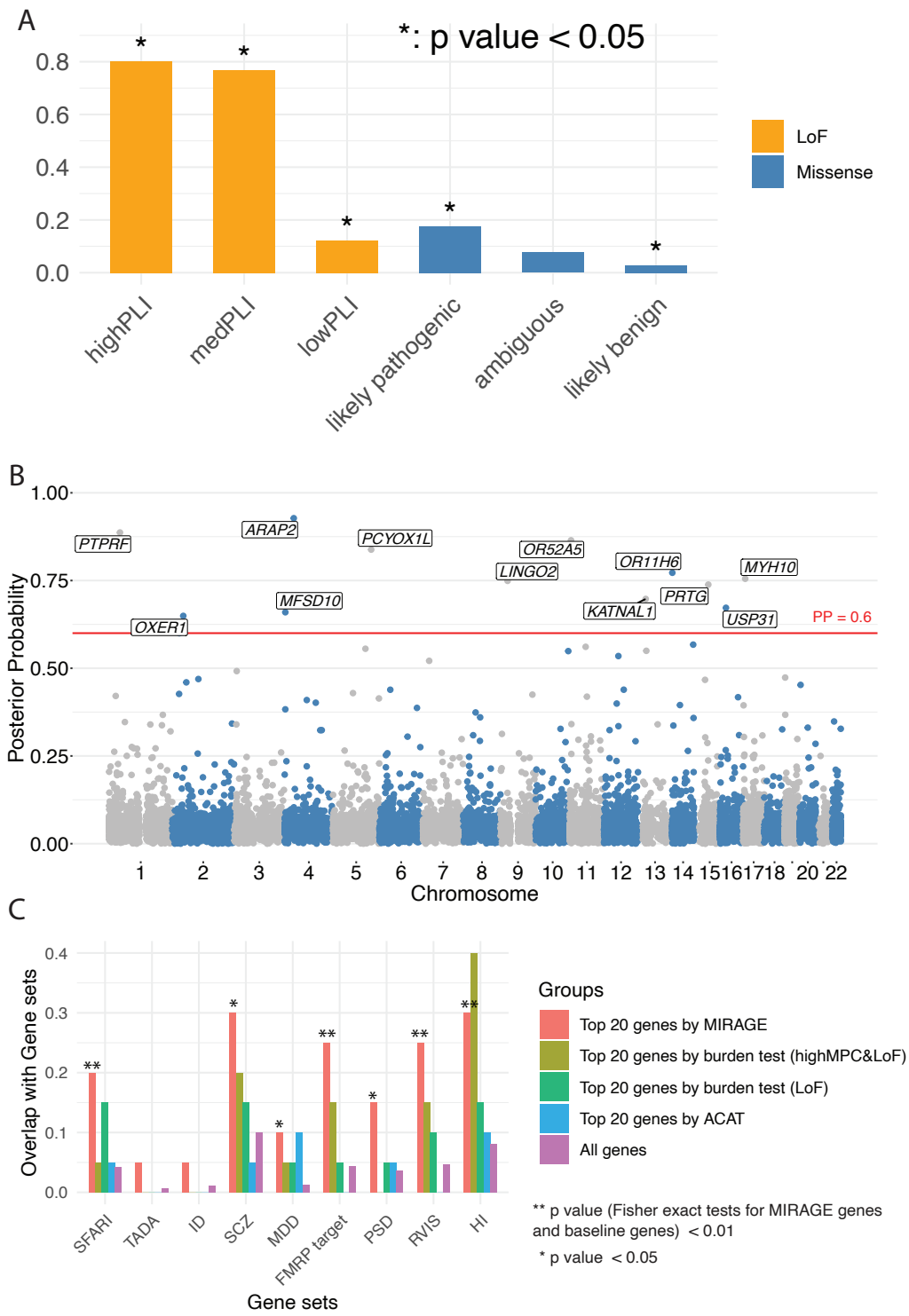

Figure S13: Results from using pLI as LoF annotation and AlphaMissense as missense variant annotation.
(A) The proportion of risk variants in 6 variant categories. LoF categories: pLI (the probability of loss-of-function intolerance) ≥ 0.995 (high), $0.995 > \text{pLI} \geq 0.5$ (med), $\text{pLI} < 0.5$ (low). Missense categories: likely pathogenic, ambiguous, likely benign **(B)** Manhattan plot of posterior probabilities (PP) for all gene analyzed. Genes with $\text{PP} > 0.6$ are labeled. **(C)** Enrichment of ASD-related gene sets in the top 20 genes found by MIRAGE, burden tests and ACAT. Enrichment test was based on Fisher's exact test. SFARI genes: Categories 1 and 2. TADA high confidence genes: $\text{FDR} < 0.05$. "All genes" refer to the entire set of genes analyzed, as a baseline.

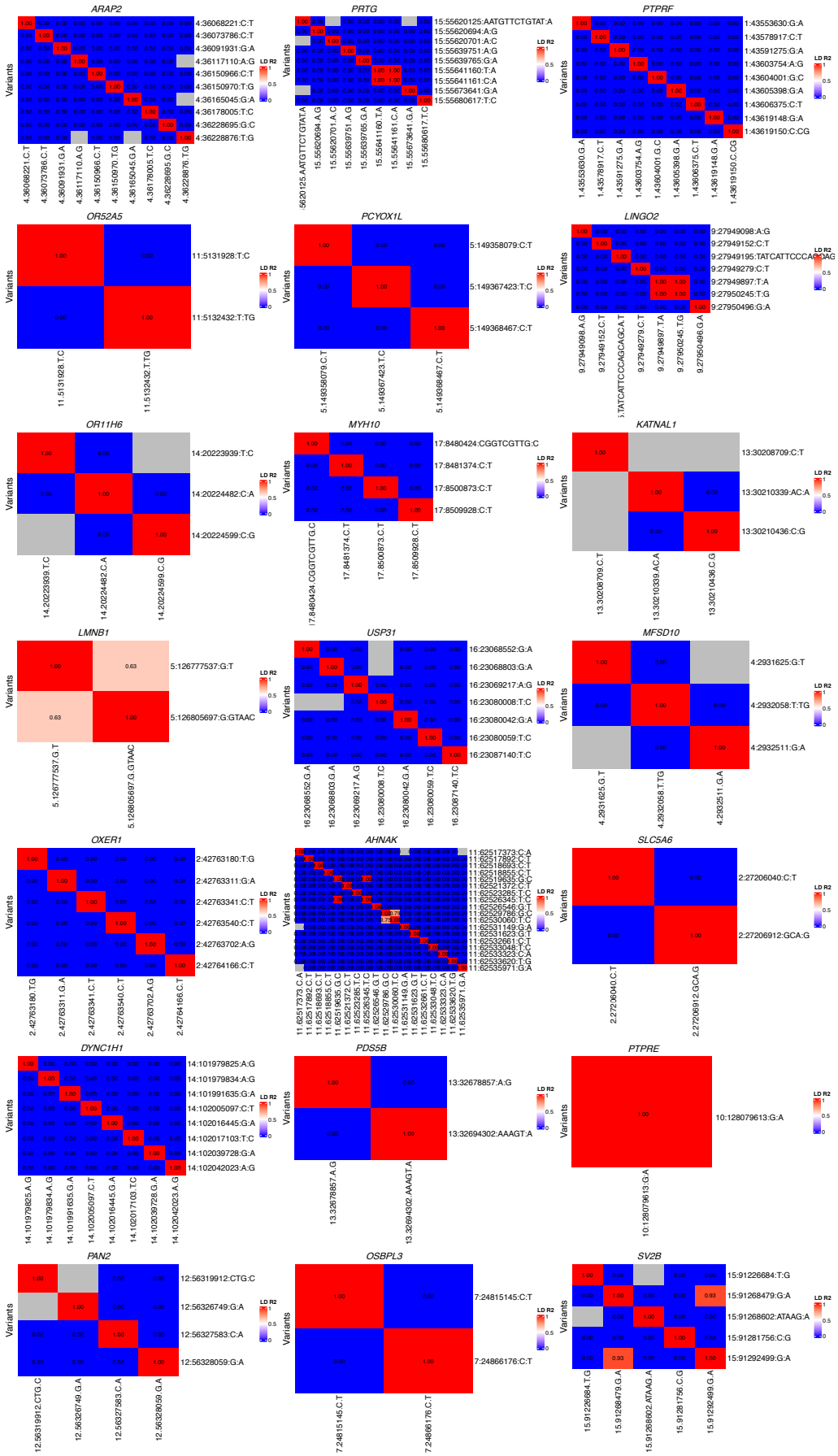

Figure S14: Heatmap showing LD (R^2) of the supporting variants of genes having PP > 0.5. Variant annotations are pLI for LoF variants and AlphaMissense for missense variants. Supporting variants of a gene are defined as the variants whose logBF > 1. Gray elements show the variant pairs whose allele frequencies were too low to compute LD. pLI (the probability of loss-of-function intolerance).

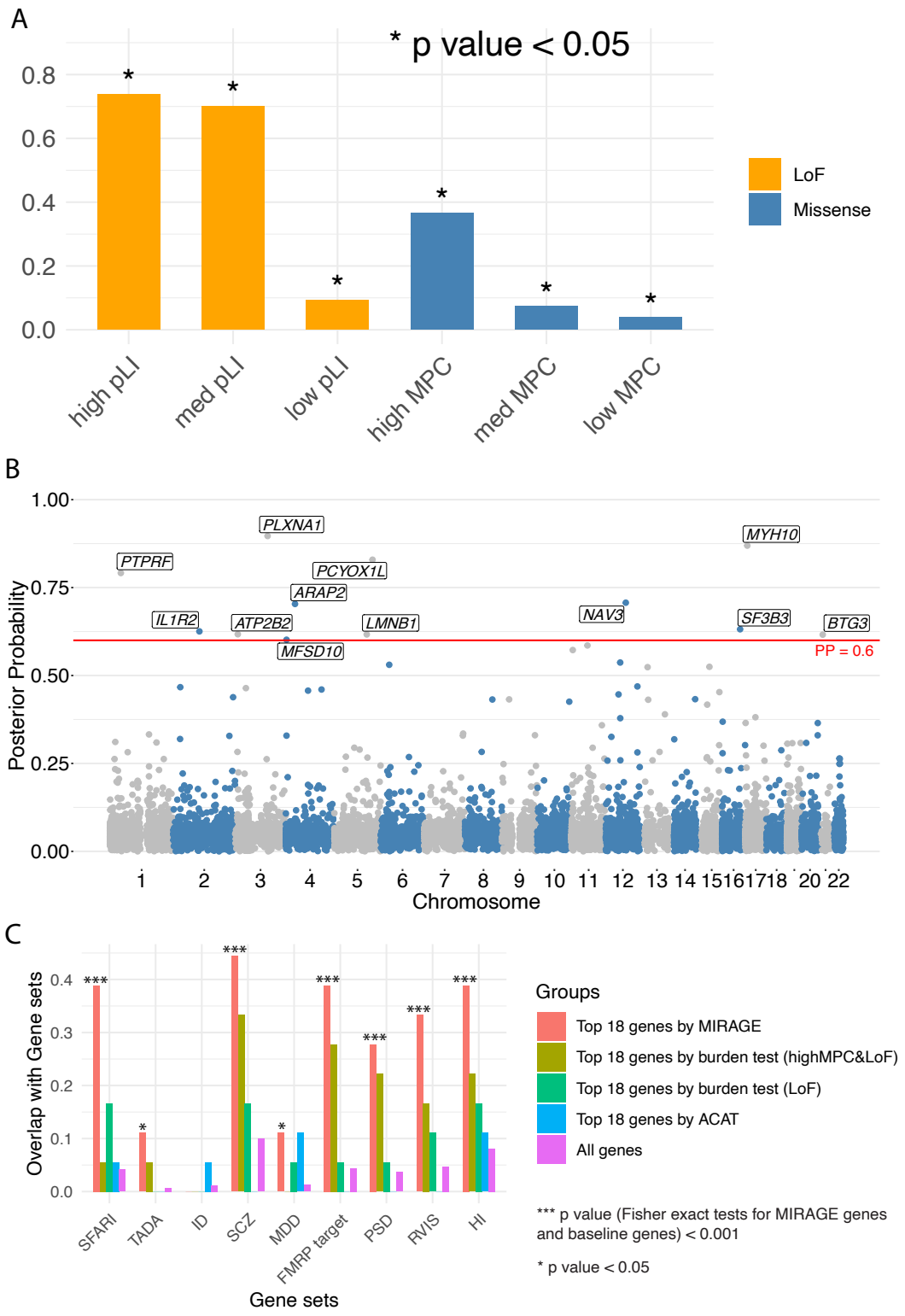

Figure S15: Results from using pLI as LoF annotation and MPC as missense variant annotation.
(A) The proportion of risk variants in 6 variant categories. LoF categories: pLI $\geq$ 0.995 (high), 0.995 > pLI $\geq$ 0.5 (med), pLI < 0.5 (low). Missense categories: MPC$\geq$ 2 (high), 2 > MPC $\geq$ 1 (med), MPC < 1 (low). **(B)** Manhattan plot of posterior probabilities (PP) for all gene analyzed. Genes with PP > 0.6 are labeled. **(C)** Enrichment of ASD-related gene sets in the top 18 genes found by MIRAGE, burden tests and ACAT. Enrichment test was based on Fisher's exact test. SFARI genes: Categories 1 and 2. TADA high confidence genes: FDR < 0.05. "All genes" refer to the entire set of genes analyzed, as a baseline. pLI (the probability of loss-of-function intolerance). MPC (Missense badness **(D)** PolyPhen-2 **(E)** and Constrain scores).

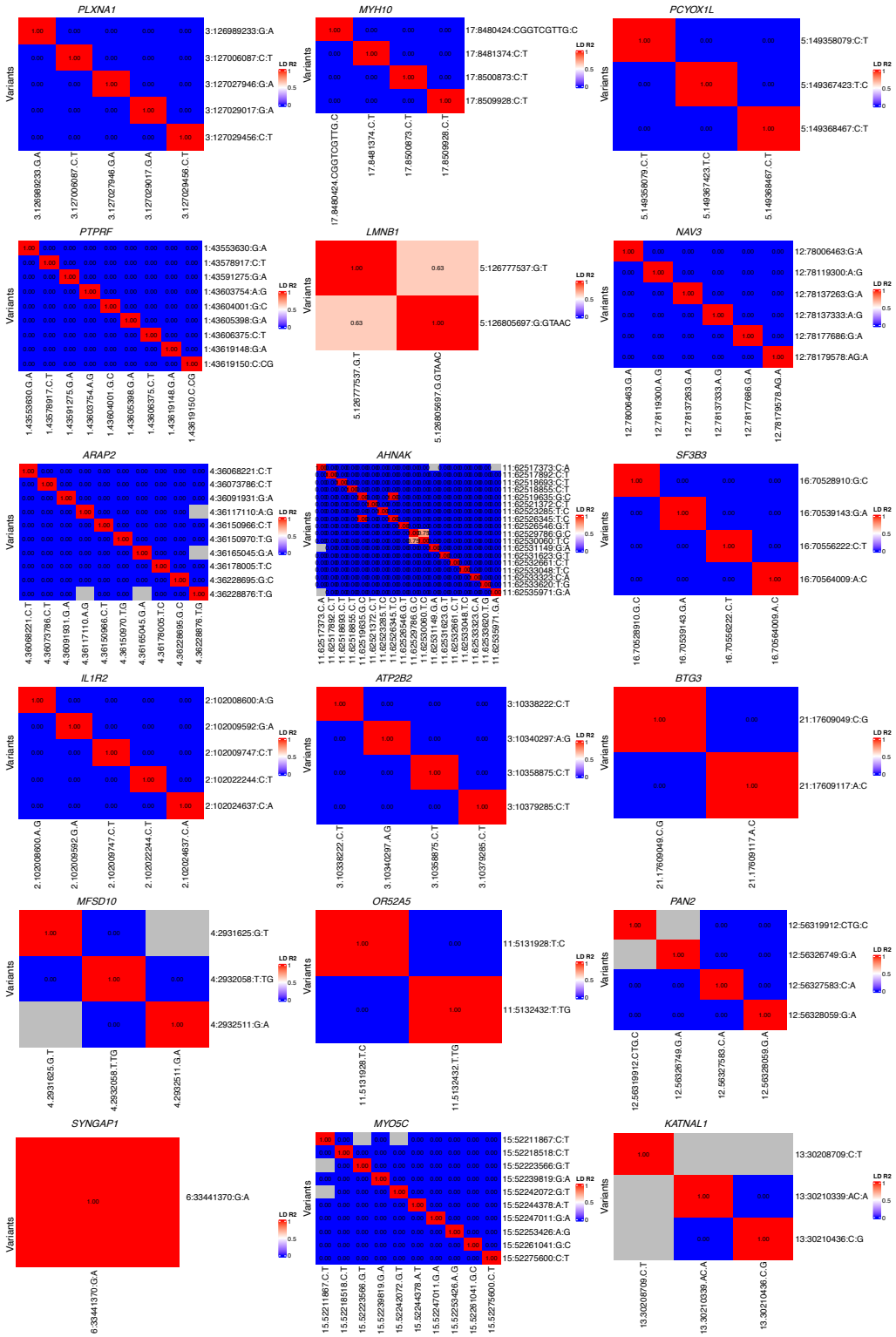

Figure S16: Heatmap showing LD (R^2) of the supporting variants of genes having PP > 0.5. Variant annotations are pLI for LoF variants and MPC for missense variants. Supporting variants of a gene are defined as the variants whose logBF > 1. Gray elements show the variant pairs whose allele frequencies were too low to compute LD. pLI (the probability of loss-of-function intolerance).

Supplemental Tables

Gene	BF	PP	total_count	lowLOEUF_count	lowLOEUF_BF	medLOEUF_count	medLOEUF_BF	highLOEUF_count	highLOEUF_BF	highMPC_count	highMPC_BF	medMPC_count	medMPC_BF	lowMPC_count	lowMPC_BF
PLXNA1	316.6647	0.9434	409 vs 391	5 vs 3	1.1337	-	-	-	-	44 vs 13	888.3763	171 vs 160	0.5599	189 vs 215	0.5616
PCYOX1L	105.7960	0.8478	126 vs 88	-	-	-	-	3 vs 3	0.9880	-	-	-	-	123 vs 85	107.0731
MYH10	89.1755	0.8244	246 vs 209	7 vs 0	12.1017	-	-	-	-	32 vs 20	5.5193	98 vs 81	1.2355	109 vs 108	1.0806
ARAP2	61.2013	0.7631	445 vs 402	-	-	-	-	5 vs 8	0.9510	-	-	-	-	440 vs 394	64.3518
SLC5A6	45.8839	0.7072	133 vs 116	-	-	14 vs 1	40.1181	-	-	-	-	23 vs 15	1.1659	96 vs 100	0.9810
PTPRF	45.8515	0.7070	535 vs 439	14 vs 3	14.8271	-	-	-	-	31 vs 22	0.7518	112 vs 107	0.9421	378 vs 307	4.3657
UBR1	44.5556	0.7010	256 vs 192	13 vs 4	9.6965	-	-	-	-	-	-	67 vs 51	2.1928	176 vs 137	2.0954
BTG3	40.3589	0.6799	52 vs 26	3 vs 1	1.2941	-	-	-	-	22 vs 5	30.2137	8 vs 8	0.9902	19 vs 12	1.0425
SF3B3	38.2542	0.6681	105 vs 72	2 vs 0	2.0052	-	-	-	-	24 vs 9	12.7308	36 vs 28	1.4202	43 vs 35	1.0551
TRAK1	33.8669	0.6406	311 vs 263	2 vs 0	2.0052	-	-	-	-	-	-	46 vs 22	11.8287	263 vs 241	1.4278
SYNGAP1	32.8103	0.6333	95 vs 64	-	-	-	-	-	-	39 vs 15	31.4086	51 vs 45	1.0577	5 vs 4	0.9877
IL1R2	29.2764	0.6064	207 vs 158	-	-	-	-	11 vs 6	1.0702	-	-	-	-	196 vs 152	27.3545
LMNB1	28.8671	0.6031	88 vs 74	11 vs 2	35.3299	-	-	-	-	-	-	21 vs 27	0.7635	56 vs 45	1.0701
NAV3	27.8493	0.5944	371 vs 378	15 vs 3	52.9195	-	-	-	-	-	-	-	-	356 vs 375	0.5263
MYO5C	26.6631	0.5839	478 vs 445	-	-	-	-	21 vs 23	0.9570	-	-	-	-	457 vs 422	27.8587
PAN2	25.8167	0.5761	237 vs 194	-	-	4 vs 2	1.3440	-	-	-	-	92 vs 87	3.1836	141 vs 105	6.0337
ATP2B2	25.7228	0.5752	148 vs 143	4 vs 0	4.7253	-	-	-	-	46 vs 30	9.0907	49 vs 51	0.7691	49 vs 62	0.7786
NCKAP5L	23.2976	0.5508	324 vs 258	-	-	3 vs 2	1.0052	-	-	-	-	-	-	321 vs 256	23.1759

Table S1: 18 putative ASD risk genes identified by MIRAGE. BF (Bayesian Factor). PP (Posterior Probability). LOEUF (Loss-Of-function Observed/Expected Upper bound Fraction). MPC (Missense badness, PolyPhen-2, and Constrain scores).

Gene	PP	ASD-related gene sets	Gene functions	Ref
PLXNA1	0.94	PSD, FMRP, RVIS, HI, SCZ	Receptor for semaphorins, signals that guide axons and neurons during development.	PMID 31690636, 39756719
PCYOX1L	0.85	MDD	Enzyme involved in polyamine metabolism. Associated with Freemartinism, sexual development.	https://www.sinobiological.com/cdna-clone/human-pcyox1l-hg27312
MYH10	0.82	PSD, FMRP, RVIS, HI, SCZ, SFARI	Non-muscle myosin IIB, involved in cell adhesion and migration. Microcephaly, developmental delay, known ASD gene.	PMID 24463507, 30123108, 25003005
ARAP2	0.76	FMRP, HI, SCZ	A protein in Rho-GTPase signaling. Implicated in ASD.	Genecard https://www.genecards.org/cgi-bin/carddisp.pl?gene=ARAP2 & PMID 32244264
PTPRF	0.71	FMRP, RVIS, HI, SCZ, ADHD	Protein tyrosine phosphatase family. Synapse formation and differentiation.	Genecard https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTPRF PMID 20139422, 19252495
SLC5A6	0.71		Transports essential vitamins and cofactors into cells. Developmental delay, motor neuropathies	PMID: 25809983, 27904971, 35013551
UBR1	0.7	HI, SFARI	E3 ubiquitin ligase. Intellectual disability.	PMID 38756130
BTG3	0.68		Cell cycle regulation. Contributes to neurogenesis in the central nervous system	https://www.genecards.org/cgi-bin/carddisp.pl?gene=BTG3
SF3B3	0.67	Constraint	Pre-mRNA splicing factor. Global loss in zebrafish causes severe developmental defects	PMID 37774458 & https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=hu
TRAK1	0.64	FMRP, RVIS, SCZ	Links mitochondrial to motor protein and microtubule. Variants cause fatal encephalopathy	PMID 28364549, 21454691, 32561740
SYNGAP1	0.63	TADA, PSD, FMRP, SCZ, DD	Regulate synaptic plasticity. Mutations cause ID, ASD.	PMID: 26912996, 32075947
IL1R2	0.61	SFARI	Decoy IL-1 receptor regulating inflammation	PMID: 23195532
LMNB1	0.6		Nuclear lamin B1. Microcephaly, involved in neurodevelopment.	PMID: 37720548, 26510501
NAV3	0.59	FMRP, RVIS, HI, SCZ, SFARI, MDD	Interact with microtubules. Guides axons, migration, and neurite growth.	PMID: 39708122, 38977784, 39038237
ATP2B2	0.58	TADA, PSD, FMRP, RVIS, SCZ, SFARI	Regulate Ca2+ signaling. Implicated in neurodevelopmental disorders with dystonia, ataxia, and seizures	PMID: 37675773 & https://www.creative-biolabs.com/gene-therapy/atp2b2-and-associated-diseases.htm
MYO5C	0.58	HI, SFARI	Myosin family, secretory granule trafficking.	https://www.ncbi.nlm.nih.gov/gene/55930
PAN2	0.58		mRNA deadenylation, control RNA stability.	PMID: 35304602
NCKAP5L	0.55		Microtubule stabilization. Neurodevelopmental disorders, including ASD.	Genecard https://www.genecards.org/cgi-bin/carddisp.pl?gene=NCKAP5L

Table S2: Genes identified by MIRAGE (posterior probability > 0.5, plausible genes are bolded) PP (Posterior Probability).

EM Algorithm We used Maximum Likelihood to estimate the parameters, denoted as $\theta = (\delta, \eta)$, where δ is the proportion of risk genes and η (vector) the proportions of risk variants in each variant category. We note that the model has hyperparameters $\bar{\gamma}$ and σ. We assume the values are provided by the users, and found that in general, the results were relatively robust to the exact values (**Figures 2C–2D**, **Figure S4**). To estimate θ, we note that the model marginalizes latent variables $\mathbf{U}$ - the risk gene status, and $\mathbf{Z}$ - the risk variant status. So we use Expectation-Maximization algorithm to estimate the parameters. We denote $\theta^{(t)}$ the parameter estimate at the t-th iteration.

- E step: we compute the $Q(\cdot)$ function

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_{\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}} \log(P(\mathbf{X},\mathbf{U},\mathbf{Z}|\theta)) \\ &= E_{\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}} \log(P(\mathbf{U},\mathbf{Z}|\theta)) + E_{\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}} \log(P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta)) \end{aligned} \tag{11}$$

First we derive $\log(P(\mathbf{U},\mathbf{Z}|\theta))$ and its expectations.

$$\begin{aligned} &\log(P(\mathbf{U},\mathbf{Z}|\theta)) = \log\Big\{P(\mathbf{U}|\theta) \times P(\mathbf{Z}|\mathbf{U},\theta)\Big\} \\ &= \log\left\{\prod_i \delta^{U_i}(1-\delta)^{1-U_i}\right\} + \log\left\{\prod_i \prod_j \eta_{C_{ij}}^{Z_{ij}}\Big(1-\eta_{C_{ij}}\Big)^{1-Z_{ij}}\right\}^{U_i} \\ &= \sum_i U_i \log(\delta) + \sum_i (1-U_i)\log(1-\delta) \\ &+ \sum_i \sum_j \left\{U_i Z_{ij} \log(\eta_{C_{ij}}) + (U_i - U_i Z_{ij}) \log\Big(1-\eta_{C_{ij}}\Big)\right\}. \end{aligned}$$

Taking expectation over $\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}$:

$$\begin{aligned} E_{\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}} \log(P(\mathbf{U},\mathbf{Z}|\theta)) &= \log(\delta)\sum_i E(U_i|\mathbf{X},\theta^{(t)}) + \log(1-\delta)\sum_i \Big[1 - E(U_i|\mathbf{X},\theta^{(t)})\Big] \\ &+ \sum_i \sum_j \left\{E(U_i Z_{ij}|\mathbf{X},\theta^{(t)})\log(\eta_{C_{ij}}) + \Big[E(U_i|\mathbf{X},\theta^{(t)}) - E(U_i Z_{ij}|\mathbf{X},\theta^{(t)})\Big]\right. \\ &\left.\times \log(1-\eta_{C_{ij}})\right\}. \end{aligned}$$

We will derive below the expectation terms $E(U_i|\mathbf{X},\theta^{(t)})$ and $E(U_i Z_{ij}|\mathbf{X},\theta^{(t)})$.

Next we derive $\log(P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta))$ and its expectation. Note that when gene i is a non-risk gene, i.e. $U_i = 0$, $P(X_i|U_i = 0,\theta)$ is independent of θ, so

$$P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta) = \prod_i \left\{ \sum_{U_i} P(X_i|\mathbf{U},\theta) \right\} = \prod_i \left\{ I(U_i=0)P(X_i|U_i=0,\theta) + I(U_i=1)P(X_i|U_i=1,\theta) \right\}$$
$$\propto \prod_i \left\{ P(X_i|U_i=1,Z_{ij},\theta) \right\}^{U_i}.$$

We expand $P(X_i|U_i=1,Z_{ij},\theta)$ into the variant level probabilities. Using the definition of B_{ij} in Equation 9, we have:

$$P(X_i|U_i=1,Z_{ij},\theta) = \prod_j \left\{ P(X_{ij}|U_i=1,Z_{ij},\theta) \right\} = \prod_j \left\{ \sum_{Z_{ij}} P(X_{ij},Z_{ij}|U_i=1,\theta) \right\}$$
$$= \prod_j \left\{ \eta_{C_{ij}} P(X_{ij}|U_i=1,Z_{ij}=1) + (1-\eta_{C_{ij}})P(X_{ij}|U_i=1,Z_{ij}=0) \right\}$$
$$= \prod_j \left\{ P(X_{ij}|U_i=1,Z_{ij}=0)\Big[I(Z_{ij}=1)\eta_{C_{ij}} \times B_{ij} + I(Z_{ij}=0)(1-\eta_{C_{ij}}) \Big] \right\}$$
$$\propto \prod_j \left\{ I(Z_{ij}=1)\eta_{C_{ij}} \times B_{ij} + I(Z_{ij}=0)(1-\eta_{C_{ij}}) \right\}$$
$$= \prod_j \left\{ \Big(\eta_{C_{ij}} B_{ij}\Big)^{Z_{ij}} \times \Big(1-\eta_{C_{ij}}\Big)^{1-Z_{ij}} \right\},$$

therefore,

$$P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta) \propto \prod_i \left\{ \prod_j \left[\Big(\eta_{C_{ij}} B_{ij}\Big)^{Z_{ij}} \times \Big(1-\eta_{C_{ij}}\Big)^{1-Z_{ij}} \right] \right\}^{U_i},$$

$$\log\Big(P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta)\Big) = c + \sum_i \sum_j \left\{ (U_i Z_{ij}) \left[\log\Big(\eta_{C_{ij}}\Big) + \log(B_{ij}) \right] \right.$$
$$\left. + (U_i - U_i Z_{ij}) \log\Big(1-\eta_{C_{ij}}\Big) \right\},$$

where c is a constant. Now we take the expectation over $\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}$:

$$E_{\mathbf{U},\mathbf{Z}|\mathbf{X},\theta^{(t)}} \log(P(\mathbf{X}|\mathbf{U},\mathbf{Z},\theta)) = c + \sum_i \sum_j \left\{ \Big[E(U_i|\mathbf{X},\theta^{(t)}) - E(U_i Z_{ij}|\mathbf{X},\theta^{(t)}) \Big] \log(1-\eta_{C_{ij}}) \right.$$
$$\left. + E(U_i Z_{ij}|\mathbf{X},\theta^{(t)}) \Big[\log(\eta_{C_{ij}}) + \log(B_{ij}) \Big] \right\}$$

Putting them together, Equation 11 becomes:

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) = \log(\delta)\sum_i E(U_i|\mathbf{X},\theta^{(t)}) + \log(1-\delta)\sum_i \Big[1 - E(U_i|\mathbf{X},\theta^{(t)})\Big] \\
+ \sum_i\sum_j \Big\{ E(U_iZ_{ij}|\mathbf{X},\theta^{(t)})\log(\eta_{C_{ij}}) + \Big[E(U_i|\mathbf{X},\theta^{(t)}) - E(U_iZ_{ij}|\mathbf{X},\theta^{(t)})\Big] \\
\times \log(1-\eta_{C_{ij}})\Big\} + \sum_i\sum_j \Big\{ \Big[E(U_i|\mathbf{X},\theta^{(t)}) - E(U_iZ_{ij}|\mathbf{X},\theta^{(t)})\Big]\log(1-\eta_{C_{ij}}) \\
+ E(U_iZ_{ij}|\mathbf{X},\theta^{(t)})\Big[\log(\eta_{C_{ij}}) + \log(B_{ij})\Big]\Big\}.
\end{aligned}
\tag{12}
$$

In the Q function, two expectations need to be calculated, $E(U_i|\mathbf{X},\theta^{(t)}), E(U_iZ_{ij}|\mathbf{X},\theta^{(t)})$. First,

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

where B_i is the Bayes factor for gene i, as in Equation 8. Note that the BF of a gene depends on the parameters at the t-th iteration, namely $\eta^{(t)}$. This term has a simple interpretation: the posterior probability of the gene i being a risk gene given the current parameters. The second expectation term also has a simple interpretation: the posterior probability that gene i is a risk gene and its variant j is a risk variant, given the current parameters. We denote $\lambda_{ij} = P(X_{ij}|U_i = 1, Z_{ij} = 0, \theta)$, then $P(X_{ij}|U_i = 1, Z_{ij} = 1, \theta) = \lambda_{ij}B_{ij}$,

$$
\begin{aligned}
&E(U_iZ_{ij}|\mathbf{X},\theta^{(t)}) = P(U_i=1,Z_{ij}=1|\mathbf{X},\theta^{(t)})\\
&= \frac{P(\mathbf{X}|Z_{ij}=1,U_i=1,\theta^{(t)})P(Z_{ij}=1|U_i=1,\theta^{(t)})P(U_i=1|\theta^{(t)})}{P(\mathbf{X}|\theta^{(t)})}\\
&= \frac{P(\mathbf{X}_i|Z_{ij}=1,U_i=1,\theta^{(t)})P(Z_{ij}=1|U_i=1,\theta^{(t)})P(U_i=1|\theta^{(t)})}{P(\mathbf{X}_i|\theta^{(t)})}\\
&= \frac{\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^t)\eta^{(t)}_{C_{ij}}\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})\delta^{(t)}}{P(U_i=1|\theta^{(t)})P(\mathbf{X}_i|U_i=1,\theta^{(t)})+P(U_i=0|\theta^{(t)})P(\mathbf{X}_i|U_i=0,\theta^{(t)})}\\
&= \frac{\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^t)\eta^{(t)}_{C_{ij}}\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})\delta^{(t)}}{\delta^{(t)}P(\mathbf{X}_i|U_i=1,\theta^{(t)})+(1-\delta^{(t)})P(\mathbf{X}_i|U_i=0,\theta^{(t)})}\\
&= \frac{\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^t)\eta^{(t)}_{C_{ij}}\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})\delta^{(t)}}{\delta^{(t)}P(X_{ij}|U_i=1,\theta^{(t)})\times\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})+(1-\delta^{(t)})P(\mathbf{X}_i|U_i=0,\theta^{(t)})}\\
&= \frac{\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)})\eta^{(t)}_{C_{ij}}\delta^{(t)}}{\delta^{(t)}\left[\eta^{(t)}_{C_{ij}}\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)})+(1-\eta^{(t)}_{C_{ij}})\lambda_{ij}(\theta^{(t)})\right]+(1-\delta^{(t)})\frac{P(\mathbf{X}_i|U_i=0,\theta^{(t)})}{\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})}}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\frac{P(\mathbf{X}_i|U_i=0,\theta^{(t)})}{\prod_{k\neq j}P(X_{ik}|U_i=1,\theta^{(t)})} = \frac{P(X_{ij}|U_i=1,\theta^{(t)})P(\mathbf{X}_i|U_i=0,\theta^{(t)})}{\prod_j P(X_{ij}|U_i=1,\theta^{(t)})}\\
&= P(X_{ij}|U_i=1,\theta^{(t)})\times\frac{1}{B_i(\theta^{(t)})}\\
&= \left[\eta^{(t)}_{C_{ij}}\lambda_{ij}(\theta^{(t)})B_{ij}(\theta^{(t)})+\left(1-\eta^{(t)}_{C_{ij}}\right)\lambda_{ij}(\theta^{(t)})\right]\times\frac{1}{B_i(\theta^{(t)})},
\end{aligned}
$$

thus,

$$
E(U_iZ_{ij}|\mathbf{X},\theta^{(t)}) = \frac{B_{ij}(\theta^{(t)})\eta^{(t)}_{C_{ij}}\delta^{(t)}}{\left[\delta^{(t)}+\frac{1-\delta^{(t)}}{B_i(\theta^{(t)})}\right]\times\left[\eta^{(t)}_{C_{ij}}B_{ij}(\theta^{(t)})+\left(1-\eta^{(t)}_{C_{ij}}\right)\right]}.
$$

With these expectation terms, we completed our calculation of the Q function.

- M step: we update θ by $\theta^{(t+1)}$ that maximizes $Q(\theta|\theta^{(t)})$ as,

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

We first take derivative with respect to δ. Note that in Equation 12, δ only occurs in the first two terms, so we have:

$$
\frac{\partial Q}{\partial\delta} = \frac{\sum_i E(U_i|\mathbf{X},\theta^{(t)})}{\delta} - \frac{\sum_i\left[1-E(U_i|\mathbf{X},\theta^{(t)})\right]}{1-\delta}.
$$

Setting this equations to 0 leads to:

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

where I is the total number of genes.

Next we derive the update rule of η. Assume there are C annotation groups, then $\eta = (\eta_1, \eta_2, \cdots, \eta_C)$ is a C-dimensional vector. The variant (i, j) in group c has the prior probability of being causal variant η_c. We use $v(i, j)$ to denote the group this variant belongs to. Taking the derivative of Q function with respect to η_c, we have:

$$\frac{\partial Q}{\partial \eta_c} = \sum_i \sum_{j:v(i,j)=c} \left[\frac{E(U_i Z_{ij}|\mathbf{X}, \theta^{(t)})}{\eta_c} - \frac{E(U_i|\mathbf{X}, \theta^{(t)}) - E(U_i Z_{ij}|\mathbf{X}, \theta^{(t)})}{1 - \eta_c} \right] = 0.$$

Solving this equation leads to the update rule:

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

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.

Results using pLI as LoF annotation, AlphaMissense as missense annotation In this section, when annotating variants, we used probability of loss-of-function intolerance (pLI scores) for LoF variants[38,39], and AlphaMissense pathogenicity[42] for missense variants. We included a total of six variant groups: pLI ≥ 0.995 (high pLI), $0.995 >$ pLI ≥ 0.5 (med pLI), pLI < 0.5 (low pLI), in addition, AlphaMissense classification: likely pathogenic, ambiguous, and likely benign.

To run MIRAGE, we used the same settings as the main text did. MIRAGE estimated the proportions of risk variants across the six variant categories (**Figure S13A**). The LoF variants in genes with high and medium pLI showed the highest proportions, with $\sim 80\%$ of variants being risk variants. In the missense variant groups, high MPC had the largest risk variant proportion of $\sim 20\%$, and the last two categories showed much lower proportions. These estimates were generally in line with the expected deleteriousness of these variant annotations. With the estimated parameters, we calculated BFs and PP for every gene and performed Bayesian FDR control. At PP > 0.5, MIRAGE identified 20 putative ASD risk genes (**Figure S13B**). We verified that the vast majority of the variants supporting these genes have very low LD (**Figure S14**), and pruning of the few remaining variants in LD have small effects on the results. At PP > 0.7 , MIRAGE identified nine putative ASD risk genes. These results thus supported higher sensitivity of MIRAGE in detecting risk genes than existing rare variant association methods.

We evaluated the candidate genes by assessing the enrichment of ASD-related gene sets. We selected the 20 genes at PP > 0.5, and for comparison, the same number of top genes by burden tests and ACAT. The ASD-related gene sets include known ASD genes curated by SFARI[44] and from de novo mutation studies using TADA[36]; risk genes of Intellectual Disability (ID)[45] and Schizophrenia (SCZ)[46]; relevant biological processes including Post-synaptic density (PSD)[46]

and FMRP target genes[46]; evolutionarily constrained genes from RVIS[47], Haplo-insufficient (HI) genes[47] and major depressive disorder[48]. We note that the SFARI and TADA genes were derived from independent datasets. While some samples in our dataset were included in earlier studies, the transmission data were not used previously. We found that the top MIRAGE genes were enriched in most ASD related gene sets. The top genes from other methods showed lower or no enrichment. These enrichment results thus strongly supported the likely roles of the candidate genes in ASD.

We next examined the function and plausibility of the top 9 genes using the more stringent cutoff, PP > 0.7). Among the nine genes, *PTPRF* and *ARAP2* have known or related functions in neurodevelopment and/or other neuropsychiatric disorders. For example, both of them have been implicated in SCZ. *MYH10* was a ASD risk gene according to SFARI (score 2, Strong candidate). It was supported by multiple *de novo* mutation studies[54–57]. *PRTG* has been reported as a candidate risk gene for ASD[94,95]. In mouse and chick models, it marks early neural progenitors and suppresses premature neuronal differentiation during neural development[96]. The soluble form of *LINGO2* promotes excitatory synapse formation and has been implicated in autism spectrum disorder (ASD) by multiple studies[97–99]. *KATNAL1* regulates neuronal development, migration, and behavior, with loss-of-function linked to brain abnormalities and ciliary defects in mice[100].

Results using pLI as LoF annotation, MPC as missense annotation In this section, we used the same functional categories as in an earlier study[37] to annotate variants, based on probability of loss-of-function intolerance (pLI scores) for LoF variants[38,39], and MPC scores (Missense badness, PolyPhen-2, and Constrain scores) for missense variants[41]. We included a total of six variant groups: pLI ≥ 0.995 (high pLI), $0.995 >$ pLI ≥ 0.5 (med pLI), pLI < 0.5 (low pLI), MPC≥ 2 (high MPC), $2 >$ MPC ≥ 1 (med MPC) and MPC < 1 (low MPC).

To run MIRAGE, we used the same settings as the main text did. MIRAGE estimated the proportions of risk variants across the six variant categories (**Figure S15A**). The LoF variants in genes with high and medium pLI showed the highest proportions, with $\sim 70\%$ of variants being risk variants. In the missense variant groups, high MPC had the largest risk variant proportion of $\sim 37\%$, and the last two categories showed much lower proportions. These estimates were generally in line with the expected deleteriousness of these variant annotations.

We evaluated the candidate genes by assessing the enrichment of ASD-related gene sets (**Figure S15C**). We selected the 18 genes at PP > 0.5 (**Figure S15B**), and for comparison, the same number of top genes by burden tests and ACAT. The ASD-related gene sets include known ASD genes curated by SFARI[44] and from de novo mutation studies using TADA[36]; risk genes of Intellectual Disability (ID)[45] and Schizophrenia (SCZ)[46]; relevant biological processes including Post-synaptic density (PSD)[46] and FMRP target genes[46]; evolutionarily constrained genes from RVIS[47], Haplo-insufficient (HI) genes[47] and major depressive disorder[48]. We note that the SFARI and TADA genes were derived from independent datasets. While some samples in our dataset were included in earlier studies, the transmission data were not used previously. We found that the top MIRAGE genes were enriched in most ASD related gene sets. The top genes from other methods showed lower or no enrichment. These enrichment results thus strongly supported the likely roles of the candidate genes in ASD.

We next examined the function and plausibility of the top 6 genes using the more stringent cutoff, PP > 0.7). Among the six genes, *PLXNA1* and *NAV3* were found by in earlier ASD

studies[34,36]. *MYH10* was a ASD risk gene according to SFARI (score 2, Strong candidate). It was supported by multiple *de novo* mutation studies *PTPRF* and *ARAP2* have known or related functions in neurodevelopment and/or other neuropsychiatric disorders. For example, both of them have been implicated in SCZ.