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

Authors

Shengtong Han, Xiaotong Sun, Laura Sloofman, ..., Autism Sequencing Consortium, Joseph Buxbaum, Xin He

Correspondence

shengtong.han@marquette.edu (S.H.), xinhe@uchicago.edu (X.H.)

In this paper, we proposed a computational Bayesian method—MIRAGE—to perform rare-genetic-variant analysis in population-matched case-control studies or transmitted variants from families. Several plausible autism-risk genes are identified by applying MIRAGE to whole-exome-sequencing data, offering insight into the mechanisms of autism.

Han et al., 2026, The American Journal of Human Genetics *113*, 1–16
January 8, 2026
https://doi.org/10.1016/j.ajhg.2025.11.013

CellPress

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

ARTICLE

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

Shengtong Han,[1,2,13,*] Xiaotong Sun,[2,13] Laura Sloofman,[3,4,13] 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,*]

Summary

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.

Introduction

Genome-wide association studies (GWASs) 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% and 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 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 whether 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 whether 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

[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
[13]These authors contributed equally
*Correspondence: shengtong.han@marquette.edu (S.H.), xinhe@uchicago.edu (X.H.)
https://doi.org/10.1016/j.ajhg.2025.11.013.

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

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, one gene for body mass index, three genes for type 2 diabetes, and two genes for asthma.[27] Often, the power of these tests is lower than the single-variant association test.[28] In a Crohn disease WES, researchers found 45 rare risk variants from single-variant analysis but only two 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 that 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. untransmitted 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 untransmitted 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 to share 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 WES studies.

Material and methods

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

$$X_{ij}\Big|Z_{ij}=0 \sim \mathrm{Pois}\left(q_{ij}N_1\right), X_{ij}^{(0)}\Big|Z_{ij}=0 \sim \mathrm{Pois}\left(q_{ij}N_0\right). \quad \text{(Equation 1)}$$

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 TADA.[33] So we have

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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

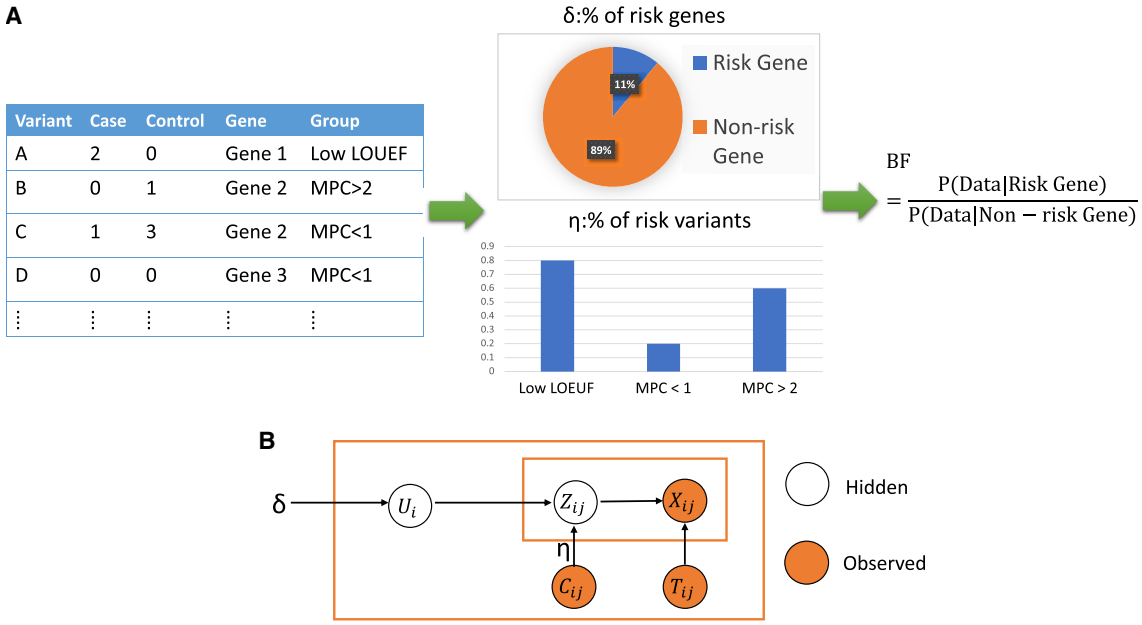

Figure 1. MIRAGE work-flow and model

Upper: 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, LoF 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. Lower: 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 refers to unobserved data.

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 Gamma($\bar{\gamma}$, σ). 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]), although, 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}:

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

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

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

We are now ready to describe the likelihood function and our inference procedure. We denote $\boldsymbol{X}$ as the data of allele counts in cases of all variants in all genes. Similarly, we denote $\boldsymbol{T}$ 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_cs for all variant categories. The likelihood function is given by

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

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

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

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

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

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 LD 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 and Z_{ij}, 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 $\widehat{\delta}$ and $\widehat{\eta}$, we can determine the Bayes factor (BF) of a gene i, B_i, and its posterior probability (PP) of being a risk gene, PP_i, as

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

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

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

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 $logB_i$ can be partitioned as the sum of contributions of each variant, $\log\left(1 - \eta_{C_{ij}} + \eta_{C_{ij}} B_{ij}\right)$. 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 PPs of all genes, we control for multiple testing by performing Bayesian false discovery rate (FDR) control.[33]

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 loss of function (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, $U_i \approx \mathrm{Bernoulli}(\delta)$. 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 $Z_{ij} \approx \mathrm{Bernoulli}(\eta_c)$. 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 $\gamma_{ij} \sim \mathrm{Gamma}(\overline{\gamma}, \sigma)$. Both $\overline{\gamma}$ and σ were set as user-specified parameters. In the simulations, we used $\overline{\gamma} = 5$ 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 Beta (α_0, β_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 Beta (α, β), 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.

WES data of ASD

WES data of autism were obtained from the 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 Simons Simplex Collection (SSC), contributing 7,291 probands (6,026 male and 1,265 female) and 2,348 siblings (1,158 male and 1,190 female). ASC B15–16 has 279 probands (223 male and 56 female), and 11 siblings (six male and five 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, and 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.

Applying MIRAGE to ASD data

We annotated variants using both LoF and missense annotations. For LoF variants, we applied two commonly used metrics: the probability of LoF intolerance (pLI)[38,39] and the LoF 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 (medium LOEUF, moderately intolerant), and deciles >3 (high LOEUF, tolerant). For pLI, pLI $\geq$ 0.995 (high), 0.995 > pLI$\geq$0.5 (medium), and pLI < 0.5 (low). For MPC, MPC $\geq$ 2 (high), 2 > MPC $\geq$ 1 (medium), 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, $\overline{\gamma}$, 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.

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

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 log BF >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 log BF 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]

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 and 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 minor allele frequency (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, 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 (Figure S5). This probably reflects that there are a large number of singleton variants (a single total variant count), resulting in $p = 1$ from the binomial test. In other words, the p values of most null variants are not uniformly distributed.

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 haploinsufficient (HI) genes,[47] totaling 1,440 genes; and (7) major depressive disorder (MDD),[48] totaling 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.

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 PP > 0.5 and for 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 then tested the significance using Poisson test.

Results

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, LoF variants, and likely deleterious missense variants[41] (Figure 1A). For each gene, MIRAGE assesses how likely its data are 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 generally be 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 the BF.

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 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 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 material and methods for the general setting), we have

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

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 EM algorithm[50] to estimate the proportion of risk genes, (δ), and the proportions of risk variants, (δ)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 material and 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 Equations 8 and 9 in material and methods). This partition allows us to assess the contribution of individual variants, or variant groups, to the evidence of a gene.

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, mimicking benign missense variants (60% of variants), damaging missense variants (30%), and 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 as $\bar{\gamma} = 3$ 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 receiver operating characteristic (ROC) curves ($\delta = 0.02$ and $\delta = 0.1$ in Figures 2A and 2B, $\delta = 0.05$ and $\delta = 0.2$ in Figure S2).

Under all the settings, we found that MIRAGE has area under the ROC (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 ~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 whether the Bayesian FDR is calibrated and whether the FDR is sensitive to mis-specification 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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

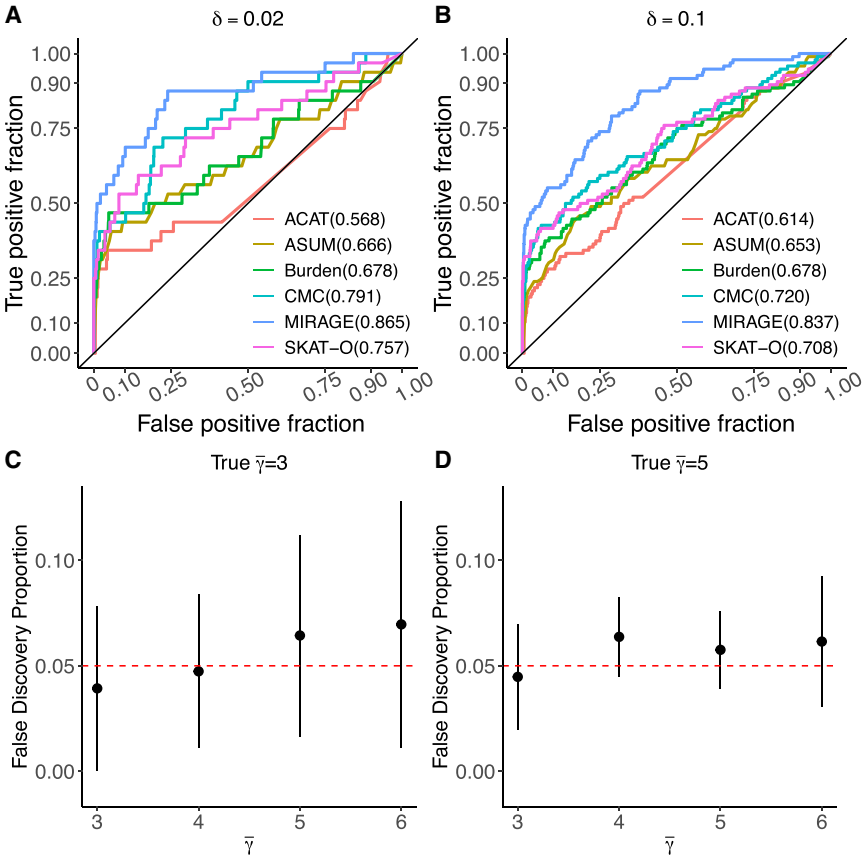

Figure 2. Simulation results

Upper left: ROC curves at $\delta = 0.02$ of different methods for classifying risk genes We simulated 1,000 genes with varying proportion of risk genes. Values for area under the curve (AUC) are shown in the brackets. Solid black reference line is the diagonal. Upper right: ROC curves at $\delta = 0.1$ of different methods for classifying risk genes. Lower left: 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). Lower right: FDR calibration by MIRAGE at $\bar{\gamma} = 5$.

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 and 2D, $\bar{\gamma} = 4, 6$ in Figure S4).

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 ASC and SPARK.[36] We treated transmitted parental alleles as cases and non-transmitted ones as controls.[33] We considered only rare variants with 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 for missense variants.[41] LOEUF quantifies how much a gene tolerates LoF 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 (medium LOEUF), and deciles greater than 3 (high LOEUF), MPC $\geq$ 2 (high MPC), 2 > MPC $\geq$ 1 (medium 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 are 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 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 PPs 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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

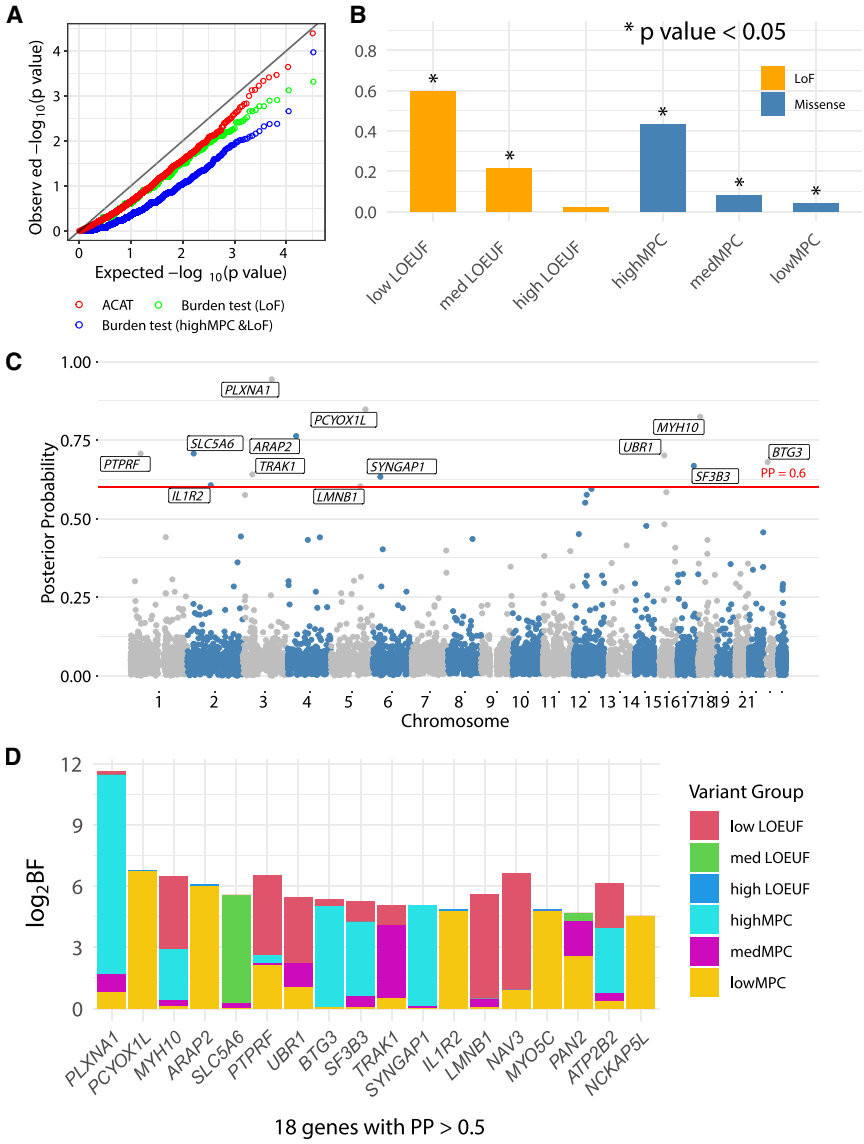

Figure 3. Summary of MIRAGE results from ASD data analysis

(A) QQ plot of all genes in the whole genome by ACAT and burden tests. (B) The proportion of risk variants in six variant categories. LoF categories: LoF observed/expected upper bound fraction (LOEUF) 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 > MPC ≥ 1 (med), MPC < 1 (low). (C) Manhattan plot of posterior probabilities (PPs) 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.

(Figure S6), and pruning of the few remaining variants in LD had small effects on the results (see material and 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* were 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/medium 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, 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).

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 SCZ[46]; relevant biological processes including post-synaptic density (PSD)[46] and FMRP target genes[46]; evolutionarily constrained genes from RVIS[47] HI genes[47] and 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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

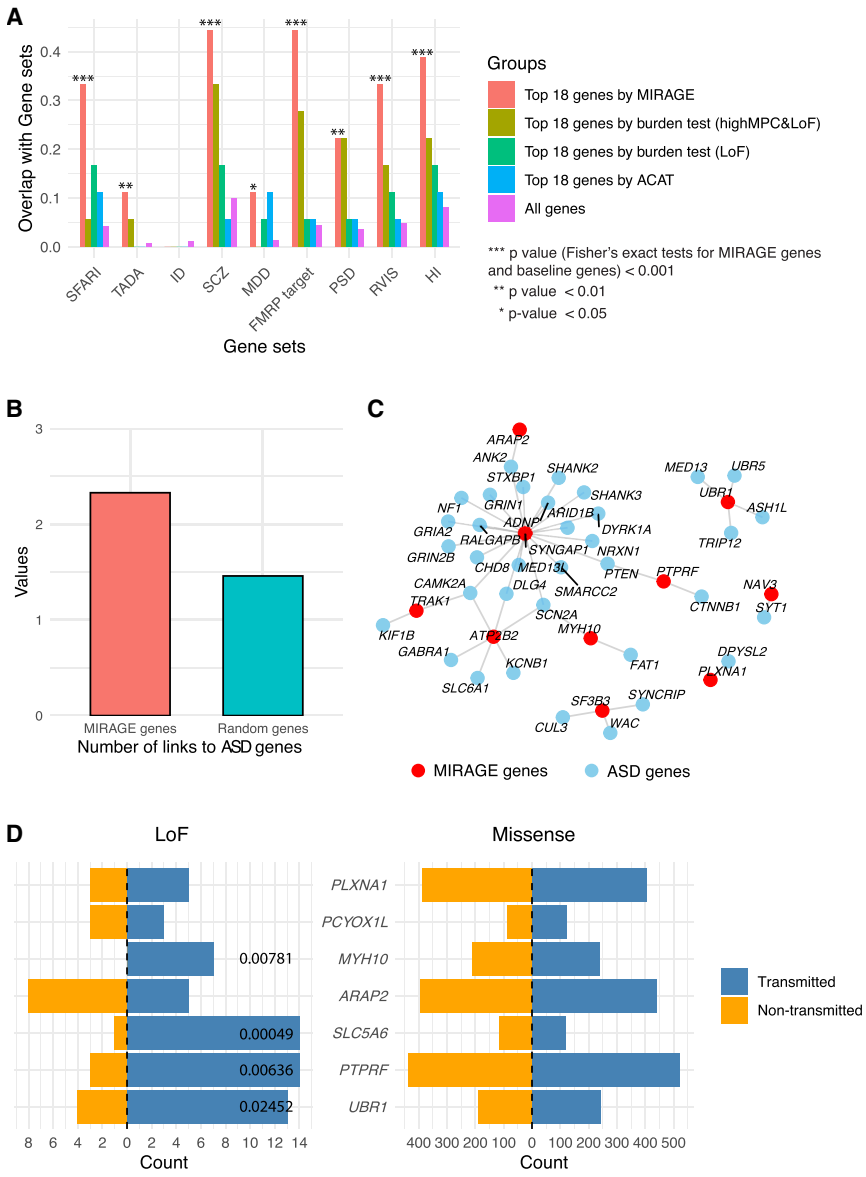

Figure 4. Evaluation of candidate genes from MIRAGE

Upper: 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" refers to the entire set of genes analyzed as a baseline. Middle left: Number of links from MIRAGE genes to TADA genes. Middle right: Gene-interaction network for MIRAGE genes and TADA genes; only genes with connections are shown here. Lower: The transmitted and non-transmitted counts of LoF (left) and missense (right) variants for seven genes with MIRAGE PP > 0.7. Numbers near the bars are p values testing transmission disequilibrium using the binomial test (only shown $p < 0.05$).

were not used previously. We found that the top MIRAGE genes were enriched in most ASD-related gene sets (Figure 4A). For example, ~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.

We next examined the function and plausibility of the top seven genes using the more stringent cutoff, PP > 0.7 (Table 1). Six out of seven genes were annotated by one or more ASD-related gene sets. Among the six genes, *PLXNA1* was found by in earlier ASD studies.[34,36] *MYH10* was an 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 discuss the literature support of these genes in the discussion.

To further validate and characterize the functions of the candidate genes, we used STRING,[59] 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 to 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 phosphatidylinositol 3-kinase (PI3K) signaling and Wnt signaling, respectively. These connections thus suggested that *PTPRF* may act on ASD by modulating PI3K and/or WNT signaling, two pathways important for ASD.[60–63] As another example, *UBR1* (PP = 0.7) is linked to several ASD genes, including *UBR5* and *TRIP12*. All three genes have functions in ubiquitination[64–66] and regulation of protein turnover.[67–69]

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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

Table 1. Genes identified by MIRAGE with PP > 0.7

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[58]	PTP family; synapse formation and differentiation
UBR1	0.70	0.22	HI,[47] SFARI[44]	E3 ubiquitin ligase; intellectual disability

PP, posterior probability; FDR, false discovery rat). The dash (–) indicates no data.

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 4 bp 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*; Figures 3D and 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 genome-wide 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 it 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).

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 (Figures 5 and 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

Please cite this article in press as: Han et al., MIRAGE: A Bayesian statistical method for gene-level rare-variant analysis incorporating functional annotations, The American Journal of Human Genetics (2026), https://doi.org/10.1016/j.ajhg.2025.11.013

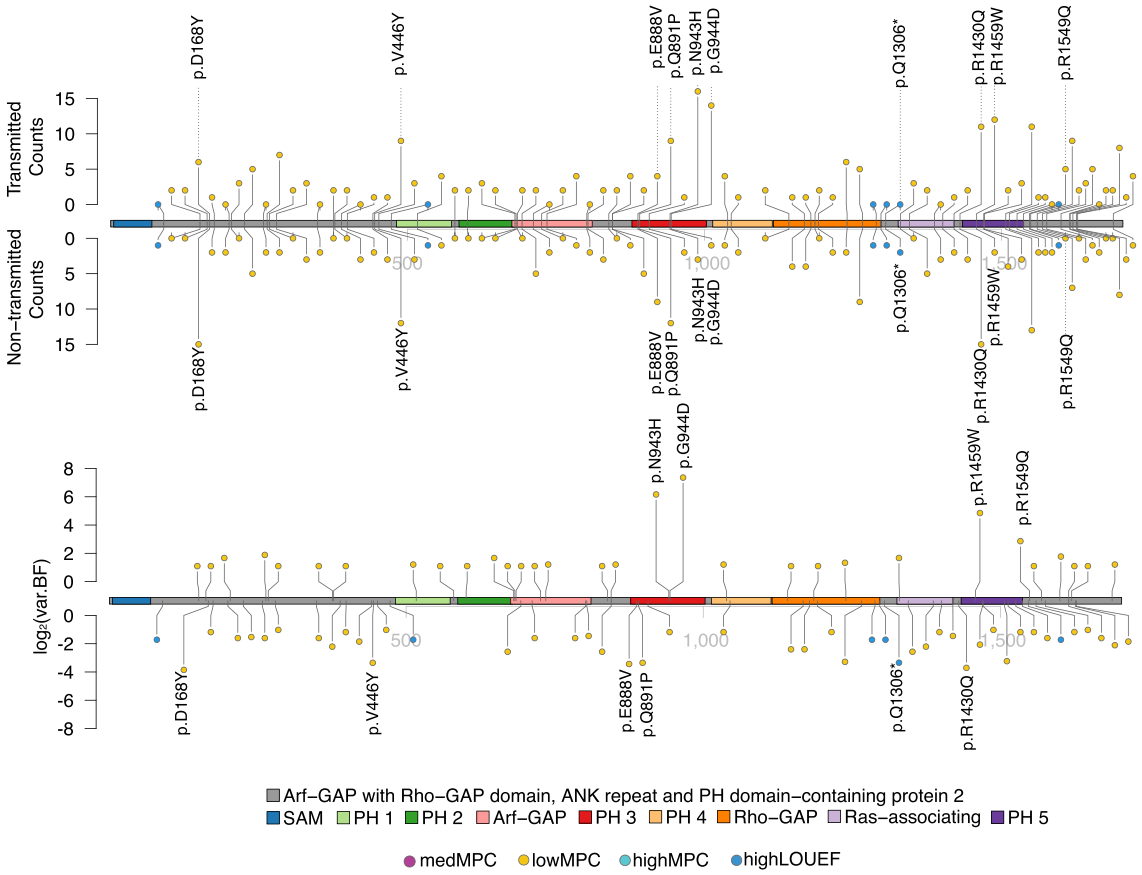

Figure 5. The lollipop plot of *ARAP2*

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), are plotted. The variants with BF > 3 and BF < 0.1 are labeled. LOEUF, LoF observed/expected upper bound fraction; MPC, missense badness, PolyPhen-2, and Constrain scores.

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 a 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 and 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 present the supporting evidence from the 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 the best annotation is 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 and 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 disjointed 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 GWASs[92,93] and in studies of *de novo* mutations.[51]

Data and code availability

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

Acknowledgments

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. and X.H.; software, G.W., X.S., and S.H.; data analysis, S.H., X.S., L.S., and X.X.; data curation, S.H., X.S., L.S., F.K.S., G.W., L.L., N.K., S.Z., W.S., T.-H.N., and J.B.; writing, S.H., X.S., and X.H.; supervision, X.H. and J.B.; project administration, X.H.; funding acquisition, X.H.

Declaration of interests

The authors declare no competing interests.

Web resources

ASC data (dbGaP: phs000298), https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v4.p3
SPARK data, https://www.sfari.org/resource/spark/

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2025.11.013.

Received: March 26, 2025
Accepted: November 21, 2025

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*, 661–678. https://doi.org/10.1038/nature05911.
2. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.
3. Shendure, J., Findlay, G.M., and Snyder, M.W. (2019). Genomic Medicine-Progress, Pitfalls, and Promise. Cell *177*, 45–57. https://doi.org/10.1016/j.cell.2019.02.003.
4. Gibson, G. (2012). Rare and common variants: twenty arguments. Nat. Rev. Genet. *13*, 135–145. https://doi.org/10.1038/nrg3118.
5. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A., and Sunyaev, S.R. (2009). Power of deep, all-exon resequencing for discovery of human trait genes. Proc. Natl. Acad. Sci. USA *106*, 3871–3876. https://doi.org/10.1073/pnas.0812824106.
6. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828. https://doi.org/10.1126/science.1215040.
7. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (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., , et al.; TOPMed Anthropometry Working Group; and NHLBI Trans-Omics for Precision Medicine TOPMed Consortium, Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D. (2022).

Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. *54*, 263–273. https://doi.org/10.1038/s41588-021-00997-7.

9. Rocheleau, G., Clarke, S.L., Auguste, G., Hasbani, N.R., Morrison, A.C., Heath, A.S., Bielak, L.F., Iyer, K.R., Young, E.P., Stitziel, N.O., et al. (2024). Rare variant contribution to the heritability of coronary artery disease. Nat. Commun. *15*, 8741. https://doi.org/10.1038/s41467-024-52939-6.
10. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.
11. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am. J. Hum. Genet. *95*, 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009.
12. Bis, J.C., Jian, X., Kunkle, B.W., Chen, Y., Hamilton-Nelson, K.L., Bush, W.S., Salerno, W.J., Lancour, D., Ma, Y., Renton, A.E., et al. (2020). Whole exome sequencing study identifies novel rare and common Alzheimer's-associated variants involved in immune response and transcriptional regulation. Mol. Psychiatry *25*, 1859–1875. https://doi.org/10.1038/s41380-018-0112-7.
13. Howrigan, D.P., Rose, S.A., Samocha, K.E., Fromer, M., Cerrato, F., Chen, W.J., Churchhouse, C., Chambert, K., Chandler, S.D., Daly, M.J., et al. (2020). Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nat. Neurosci. *23*, 185–193. https://doi.org/10.1038/s41593-019-0564-3.
14. Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature *604*, 509–516. https://doi.org/10.1038/s41586-022-04556-w.
15. Cirulli, E.T. (2016). The Increasing Importance of Gene-Based Analyses. PLoS Genet. *12*, e1005852. https://doi.org/10.1371/journal.pgen.1005852.
16. Li, B., and Leal, S.M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. Am. J. Hum. Genet. *83*, 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024.
17. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384. https://doi.org/10.1371/journal.pgen.1000384.
18. Price, A.L., Kryukov, G.V., de Bakker, P.I.W., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838. https://doi.org/10.1016/j.ajhg.2010.04.005.
19. Derkach, A., Lawless, J.F., 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*, 110–121. https://doi.org/10.1002/gepi.21689.
20. Ionita-Laza, I., Buxbaum, J.D., Laird, N.M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. *7*, e1001289. https://doi.org/10.1371/journal.pgen.1001289.
21. Wu, M.C., 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*, 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*, 1159–1172. https://doi.org/10.1111/biom.13634.
23. Liu, Y., Chen, S., Li, Z., Morrison, A.C., 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*, 410–421. 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.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun, S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat. Genet. *52*, 634–639. https://doi.org/10.1038/s41588-020-0621-6.
26. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., 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*, 969–983. https://doi.org/10.1038/s41588-020-0676-4.
27. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.
28. Timpson, N.J., Greenwood, C.M.T., Soranzo, N., Lawson, D.J., and Richards, J.B. (2018). Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat. Rev. Genet. *19*, 110–124. https://doi.org/10.1038/nrg.2017.101.
29. Bodea, C.A., Neale, B.M., Ripke, S., International IBD Genetics Consortium, Daly, M.J., 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*, 857–868. https://doi.org/10.1016/j.ajhg.2016.02.025.
30. Artomov, M., Loboda, A.A., Artyomov, M.N., and Daly, M.J. (2024). Public platform with 39,472 exome control samples enables association studies without genotype sharing. Nat. Genet. *56*, 327–335. https://doi.org/10.1038/s41588-023-01637-y.
31. He, Z., O'Roak, B.J., Smith, J.D., Wang, G., Hooker, S., Santos-Cortez, R.L.P., Li, B., Kan, M., Krumm, N., Nickerson, D.A., et al. (2014). Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am. J. Hum. Genet. *94*, 33–46. https://doi.org/10.1016/j.ajhg.2013.11.021.
32. Ruiz-Narváez, E.A., and Campos, H. (2004). Transmission disequilibrium test (TDT) for case-control studies. Eur. J. Hum. Genet. *12*, 105–114. https://doi.org/10.1038/sj.ejhg.5201099.
33. He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum,

J.D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. *9*, e1003671. https://doi.org/10.1371/journal.pgen.1003671.

34. Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J.B., Obiajulu, J.U., Wright, J.R., Murali, S.C., Xu, S.X., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat. Genet. *54*, 1305–1319. https://doi.org/10.1038/s41588-022-01148-2.
35. Nguyen, H.T., Bryois, J., Kim, A., Dobbyn, A., Huckins, L.M., Munoz-Manchado, A.B., Ruderfer, D.M., 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 Med. *9*, 114. https://doi.org/10.1186/s13073-017-0497-y.
36. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331. https://doi.org/10.1038/s41588-022-01104-0.
37. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., 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*, 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036.
38. Kosmicki, J.A., Samocha, K.E., Howrigan, D.P., Sanders, S.J., Slowikowski, K., Lek, M., Karczewski, K.J., Cutler, D.J., 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*, 504–510. https://doi.org/10.1038/ng.3789.
39. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.
40. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.
41. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. Preprint at bioRxiv. https://doi.org/10.1101/148353.
42. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science *381*, eadg7492. https://doi.org/10.1126/science.adg7492.
43. Ou, J., and Zhu, L.J. (2019). trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. Nat. Methods *16*, 453–454. https://doi.org/10.1038/s41592-019-0430-y.
44. Abrahams, B.S., Arking, D.E., Campbell, D.B., Mefford, H.C., Morrow, E.M., Weiss, L.A., 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*, 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. Am. J. Hum. Genet. *94*, 677–694. https://doi.org/10.1016/j.ajhg.2014.03.018.
46. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kähler, A., et al. (2014). A polygenic burden of rare disruptive mutations in schizophrenia. Nature *506*, 185–190. https://doi.org/10.1038/nature12975.
47. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLoS Genet. *9*, e1003709. https://doi.org/10.1371/journal.pgen.1003709.
48. Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium Electronic address andrew mcintosh@ed ac uk; and 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*, 640–652.e9. https://doi.org/10.1016/j.cell.2024.12.002.
49. Whittemore, A.S. (2007). A Bayesian false discovery rate for multiple testing. J. Appl. Stat. *34*, 1–9. https://doi.org/10.1080/02664760600994745.
50. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. J. Roy. Stat. Soc. B *39*, 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.
51. Liu, Y., Liang, Y., Cicek, A.E., Li, Z., Li, J., Muhle, R.A., 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. Am. J. Hum. Genet. *102*, 1031–1047. 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. Hum. Hered. *70*, 42–54. https://doi.org/10.1159/000288704.
53. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. Nature *515*, 209–215. https://doi.org/10.1038/nature13772.
54. Fromer, M., Pocklington, A.J., Kavanagh, D.H., Williams, H.J., Dwyer, S., Gormley, P., Georgieva, L., Rees, E., Palta, P., Ruderfer, D.M., et al. (2014). De novo mutations in schizophrenia implicate synaptic networks. Nature *506*, 179–184. https://doi.org/10.1038/nature12929.
55. Hlushchenko, I., Khanal, P., Abouelezz, A., Paavilainen, V.O., and Hotulainen, P. (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, H.M., Gonzalez, K.D., and Chung, W.K. (2013). A human de novo

mutation in MYH10 phenocopies the loss of function mutation in mice. Rare Dis. *1*, e26144. https://doi.org/10.4161/rdis.26144.

57. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221. https://doi.org/10.1038/nature13908.
58. Demontis, D., Walters, G.B., Athanasiadis, G., Walters, R., Therrien, K., Nielsen, T.T., 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. Nat. Genet. *55*, 198–208. https://doi.org/10.1038/s41588-022-01285-8.
59. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., Gable, A.L., Fang, T., Doncheva, N.T., 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 Res. *51*, D638–D646. https://doi.org/10.1093/nar/gkac1000.
60. Sharma, A., and Mehan, S. (2021). Targeting PI3K-AKT/mTOR signaling in the prevention of autism. Neurochem. Int. *147*, 105067. https://doi.org/10.1016/j.neuint.2021.105067.
61. Enriquez-Barreto, L., and Morales, M. (2016). The PI3K signaling pathway as a pharmacological target in Autism related disorders and Schizophrenia. Mol. Cell. Ther. *4*, 2. https://doi.org/10.1186/s40591-016-0047-9.
62. Bae, S.M., and Hong, J.Y. (2018). The Wnt Signaling Pathway and Related Therapeutic Drugs in Autism Spectrum Disorder. Clin. Psychopharmacol. Neurosci. *16*, 129–135. https://doi.org/10.9758/cpn.2018.16.2.129.
63. Park, G., Jang, W.E., Kim, S., Gonzales, E.L., Ji, J., Choi, S., Kim, Y., Park, J.H., Mohammad, H.B., Bang, G., et al. (2023). Dysregulation of the Wnt/β -catenin signaling pathway via Rnf146 upregulation in a VPA-induced mouse model of autism spectrum disorder. Exp. Mol. Med. *55*, 1783–1794. https://doi.org/10.1038/s12276-023-01065-2.
64. 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*, 334–338. https://doi.org/10.1038/s41586-021-04099-w.
65. Hodáková, Z., Grishkovskaya, I., Brunner, H.L., Bolhuis, D.L., Belačić, K., Schleiffer, A., Kotisch, H., Brown, N.G., and Haselbach, D. (2023). Cryo-EM structure of the chain-elongating E3 ubiquitin ligase UBR5. EMBO J. *42*, e113348. https://doi.org/10.15252/embj.2022113348.
66. Seo, B.A., Kim, D., Hwang, H., Kim, M.S., Ma, S.X., Kwon, S.H., Kweon, S.H., Wang, H., Yoo, J.M., Choi, S., et al. (2021). TRIP12 ubiquitination of glucocerebrosidase contributes to neurodegeneration in Parkinson's disease. Neuron *109*, 3758–3774.e11. https://doi.org/10.1016/j.neuron.2021.09.031.
67. 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. J. Mol. Biol. *437*, 168939. https://doi.org/10.1016/j.jmb.2025.168939.
68. Jiang, H., He, X., Feng, D., Zhu, X., and Zheng, Y. (2015). RanGTP aids anaphase entry through Ubr5-mediated protein turnover. Proc. Natl. Acad. Sci. USA *112*, E5628–E5637. https://doi.org/10.1073/pnas.1515902112.
69. Kaiho-Soma, A., Akizuki, Y., Igarashi, K., Endo, A., Shoda, T., Kawase, Y., Demizu, Y., Naito, M., Saeki, Y., Tanaka, K., and Ohtake, F. (2021). TRIP12 promotes small-molecule-induced degradation through K29/K48-branched ubiquitin chains. Mol. Cell *81*, 1411–1424.e7. https://doi.org/10.1016/j.molcel.2021.01.023.
70. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., 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*, 1161–1170. https://doi.org/10.1038/s41588-018-0167-z.
71. Minton, K. (2023). Predicting variant pathogenicity with AlphaMissense. Nat. Rev. Genet. *24*, 804. https://doi.org/10.1038/s41576-023-00668-9.
72. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J., and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. Nat. Genet. *55*, 1512–1522. https://doi.org/10.1038/s41588-023-01465-0.
73. Yi, N., and Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. Genet. Epidemiol. *35*, 57–69. https://doi.org/10.1002/gepi.20554.
74. Quintana, M.A., Schumacher, F.R., Casey, G., Bernstein, J.L., Li, L., and Conti, D.V. (2012). Incorporating prior biologic information for high-dimensional rare variant association studies. Hum. Hered. *74*, 184–195. https://doi.org/10.1159/000346021.
75. Logsdon, B.A., Dai, J.Y., Auer, P.L., Johnsen, J.M., Ganesh, S.K., Smith, N.L., Wilson, J.G., Tracy, R.P., Lange, L.A., Jiao, S., et al. (2014). A variational Bayes discrete mixture test for rare variant association. Genet. Epidemiol. *38*, 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*, 334–344. 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*, 340–352. https://doi.org/10.1016/j.ajhg.2017.07.011.
78. Venkataraman, G.R., DeBoever, C., Tanigawa, Y., Aguirre, M., Ioannidis, A.G., Mostafavi, H., Spencer, C.C.A., Poterba, T., Bustamante, C.D., Daly, M.J., et al. (2021). Bayesian model comparison for rare-variant association studies. Am. J. Hum. Genet. *108*, 2354–2367. https://doi.org/10.1016/j.ajhg.2021.11.005.
79. Susak, H., Serra-Saurina, L., Demidov, G., Rabionet, R., Domènech, L., Bosio, M., Muyas, F., Estivill, X., Escaramís, G., and Ossowski, S. (2021). Efficient and flexible integration of variant characteristics in rare variant association studies using integrated nested Laplace approximation. PLoS Comput. Biol. *17*, e1007784. https://doi.org/10.1371/journal.pcbi.1007784.
80. Yang, Y., Basu, S., and Zhang, L. (2021). A Bayesian hierarchically structured prior for rare-variant association testing. Genet. Epidemiol. *45*, 413–424. https://doi.org/10.1002/gepi.22379.
81. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P., and Thibodeau, S.N. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet. Epidemiol. *37*, 409–418. https://doi.org/10.1002/gepi.21727.

82. De, G., Yip, W.K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. PLoS One *8*, e48495. https://doi.org/10.1371/journal.pone.0048495.
83. Hecker, J., Townes, F.W., Kachroo, P., Laurie, C., Lasky-Su, J., Ziniti, J., Cho, M.H., Weiss, S.T., Laird, N.M., and Lange, C. (2021). A unifying framework for rare variant association testing in family-based designs, including higher criticism approaches, SKATs, and burden tests. Bioinformatics *36*, 5432–5438. https://doi.org/10.1093/bioinformatics/btaa1055.
84. Cornejo, F., Cortés, B.I., Findlay, G.M., and Cancino, G.I. (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*, 835. https://doi.org/10.3390/cells9040835.
86. Tanna, C.E., Goss, L.B., Ludwig, C.G., and Chen, P.W. (2019). Arf GAPs as Regulators of the Actin Cytoskeleton—An Update. Int. J. Mol. Sci. *20*, 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. J. Neurochem. *134*, 97–112. https://doi.org/10.1111/jnc.13092.
88. Subramanian, V.S., Constantinescu, A.R., Benke, P.J., and Said, H.M. (2017). Mutations in SLC5A6 associated with brain, immune, bone, and intestinal dysfunction in a young child. Hum. Genet. *136*, 253–261. https://doi.org/10.1007/s00439-016-1751-x.
89. Holling, T., Nampoothiri, S., Tarhan, B., Schneeberger, P.E., Vinayan, K.P., Yesodharan, D., Roy, A.G., Radhakrishnan, P., Alawi, M., Rhodes, L., et al. (2022). Novel biallelic variants expand the SLC5A6-related phenotypic spectrum. Eur. J. Hum. Genet. *30*, 439–449. 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 Rep. *5*, 140–147. 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*, 1376. https://doi.org/10.3390/genes11111376.
92. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573. https://doi.org/10.1016/j.ajhg.2014.03.004.
93. Wen, X., Lee, Y., Luca, F., and Pique-Regi, R. (2016). Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. Am. J. Hum. Genet. *98*, 1114–1129. https://doi.org/10.1016/j.ajhg.2016.03.029.