bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

MIRAGE: a Bayesian rare variant association analysis method incorporating functional information of variants

Shengtong Han*[+1,2], Xiaotong Sun*[2], Laura Sloofman[3,4] F. Kyle Satterstrom[5,6,7], Gao Wang[8], Lifan Liang[2], Nicholas Knoblauch[2], Siming Zhao[9], Yuwen Liu[10], Wenhui Sheng[11], Hoang T. Nguyen[12], Xin He[+2,13]

1 Marquette University, School of Dentistry, 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 Neurology, Columbia University Vagelos College of Physicians and Surgeons, New York City, NY, USA
9 Department of biomedical data science, Dartmouth College, Hanover, NH, USA
10 AGIS, Chinese Academy of Agricultural Sciences, Shenzhen, China
11 Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, WI, USA
12 Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA
13 Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, University of Chicago, Chicago, IL, USA
* The authors contributed equally to this work
[+] Corresponding to Xin He, xinhe@uchicago.edu; Shengtong Han, shengtong.han@marquette.edu

Abstract

Many biobanks and consortiums have generated large whole-exome and whole-genome sequencing datasets, making it possible to perform rare variant association studies. Compared to common variants, rare variants tend to have larger effect sizes and generally have much lower linkage disequilibrium (LD), making it easier to identify causal variants. These potential benefits make association analysis with rare variants a priority for human genetics researchers. Because of the low frequencies of these variants, however, the power of single-variant association is often low. To improve the power, numerous methods have been developed to aggregate information of all variants of a gene to better identify risk genes. These gene-based methods, however, often make unrealistic assumptions, e.g. all rare variants in a risk gene have the same effects. In practice, current gene-based analysis methods often fail to show any advantage over simple single-variant analysis. In this work, we develop a Bayesian method: MIxture model based Rare variant Analysis on GEnes (MIRAGE). MIRAGE captures the

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

heterogeneity of variant effects by treating all variants of a gene as a mixture of risk and non-risk variants, and models the prior probabilities of being risk variants as function of external information of variants, such as allele frequencies and predicted deleterious effects. MIRAGE uses an empirical Bayes approach, combining information across genes to estimate these prior probabilities. 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. Our results highlight several novel Autism genes. MIRAGE is available at https://xinhe-lab.github.io/mirage.

Author summary

Introduction

Genome-wide association studies (GWAS) have successfully identified many loci associated with various complex traits [1–3]. However, in most of these loci, the causal variants and their target genes remain unknown. Additionally, most common variants (with minor allele frequency greater than 5%) discovered by GWAS have small effect sizes [2, 3]. These limitations make it difficult to translate GWAS findings into molecular mechanisms of diseases and to guide therapeutic development. 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], and can only be discovered by sequencing studies. Another benefit of rare variant analysis is that linkage disequilibrium is much weaker, 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 these sequencing data becomes an important area in human genetics [8].

Existing rare variant studies often focus on the protein-coding portions of the genome, using Whole Exome Sequening (WES). WES studies has achieved notable successes in medically important traits such as Alzheimer's disease and Schizophrenia [9, 10]. Nevertheless, these studies typically discovered relatively few variants associated with the traits. This is due to a fundamental challenge that the power of detecting rare variant association is low. A natural strategy to address this challenge is to aggregate information from all rare variants in a genomic region or gene to test if they collectively associate with the phenotype [11]. A number of methods have been developed to perform variant set tests [8]. The most commonly used method is the burden test, which collapse all rare, potentially deleterious variants in a gene, and test the association of the total frequency of these variants (burden) with a phenotype [12–15]. Other commonly used tests include the Sequence Kernel Association Test (SKAT) [16], which tests if the variance of the effect sizes of variants is equal to 0, and Aggregated Cauchy association test (ACAT) [17], which combines the p-values of variants based on the Cauchy distribution.

Despite these efforts, the rare variant set test has limited successes. Often, the power of these tests is lower than the much simpler single-variant association test [18], suggesting some key weakness of existing methods. Indeed, both burden tests and SKAT assume that if a gene is a risk gene, then most variants in that gene should have causal effects. In reality, it is likely that even within a risk gene, causal variants are sparse, leading to the power loss by these methods. ACAT was designed to be a more sensitive test when a small number of variants have causal effects. Nevertheless, ACAT is a way to combine p-values from all variants, but does not explicitly model the heterogeneity of effects across variants. Additionally, ACAT still relies on collapsing

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

variants with frequencies below a threshold. Researchers have attempted to improve the power of rare variant analysis by predicting likely deleterious variants, and then run burden, SKAT or ACAT on the subsets of variants. However, such bioinformatic predictions are far from perfect, so the uncertainty and heterogeneity of variant effects remains largely un-addressed.

We address the challenge of rare variant analysis in the context of family (trio)-sequencing studies. In such studies, variants from parents can be classified as transmitted to children or not. Conceptually, the analysis of such data is statistically equivalent to a case-control study, where the number of times a variant was transmitted is similar to the allele frequency in cases, and the number of non-transmissions is similar to the allele frequency in controls [23]. An imbalanced transmission vs. un-transmission ratio of a variant would suggest association of the variant with the trait. It was well-known that this study design is free of population structure, a common issue of case-control studies [23, 24]. Thus studying inherited variants from trio studies would allow us to focus on the key challenge of rare variant association test, the modeling of heterogeneity of variant effects. Indeed, trio-sequencing studies of autism spectrum disorder (ASD) has revealed an overall tendency of transmission of deleterious variants [25], 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 perform rare variant analysis in a gene or a genomic region, for inherited variants in trio studies. When population structure is not a concern, the method can also be applied to case-control studies. Our key motivation is to better account for the heterogeneity of the variant effects, while taking advantage of functional information of the variants. We model 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 functions. 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, so a simple filter may miss many functional variants; and conversely, many variants passing the filter may have no effects. Importantly, the parameters linking the annotations of variants and their prior probabilities were estimated using an Empirical Bayes strategy, by pooling data from all genes analyzed. To summarize, our approach allows us to effectively account for uncertainty in risk variants, and at the same time to put more emphasis on variants with putative functional effects.

We demonstrated the advantage of the new 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 exome sequencing data of ASD.

Results

Overview of MIRAGE

MIRAGE is designed to analyze rare variant data from trio-sequencing or population-matched case-control studies. For simplicity, we assume our data is from the latter. MIRAGE only 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 missense variants with predicted deleterious effects [35] (Figure 1A). MIRAGE models

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

all input genes as a mixture of risk and non-risk genes, with the proportion of risk genes δ. For a non-risk gene, all its variants are non-risk variants, and their expected frequencies are equal between cases and controls. For a risk gene, any of its variants has a prior probability of being a risk variant, with the probability equal to η_c for the variants in the category c. For a risk variant, its frequencies would be generally different between cases and controls, with the ratio given by the relative risk of the variant (see below). MIRAGE estimates the parameters, including δ and η_c for all categories, over the entire dataset of all genes (Figure 1A). Then for each gene, it assesses its evidence as a risk gene by computing its Bayes factor (BF). BF is similar to likelihood ratio test, comparing the null model (non-risk gene) and the alternative model (risk gene). BF of a gene combines the evidence of all its variants, with larger contributions from more functionally important categories (those with larger values of η_c). 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 a gene being 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 [29].

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 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 the risk variant status of variant j, denoted as Z_{ij}. For a non-risk gene ($U_i = 0$), then $Z_{ij} = 0$ for all variants. For a risk gene ($U_i = 1$), then 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 any 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:

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

where γ_{ij} is the relative risk of variant j, modelled as a Gamma distribution. We used an Expectation Maximization (EM) algorithm [28] to estimate the proportion of risk genes (δ), and the proportions of risk variants in each variant category, 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 variants counts, X_{ij}'s, under $U_i = 1$ vs. $U_i = 0$ (see Methods).

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

We performed simulations to compare MIRAGE with existing methods in identifying risk genes from rare variants. We fixed sample sizes at 3,000 cases and 3,000 controls. We simulated data of 1000 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, however, the actual number may be smaller because a variant with no count in cases and controls from simulations will be filtered, see Methods). For a risk gene, its variant belongs to one of three functional categories, mimicing benign missense variants, damaging missense variants and LoF variants. The proportions of variants in these categories are 60%, 30%, 10%, respectively. The proportions of risk variants, η_c, are 0.05, 0.2 and 0.5 for the three categories, respectively. 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 [23, 30], we set the mean relative risks, $\bar{\gamma} = 3$ for the first two variant categories and 5 for the last one.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

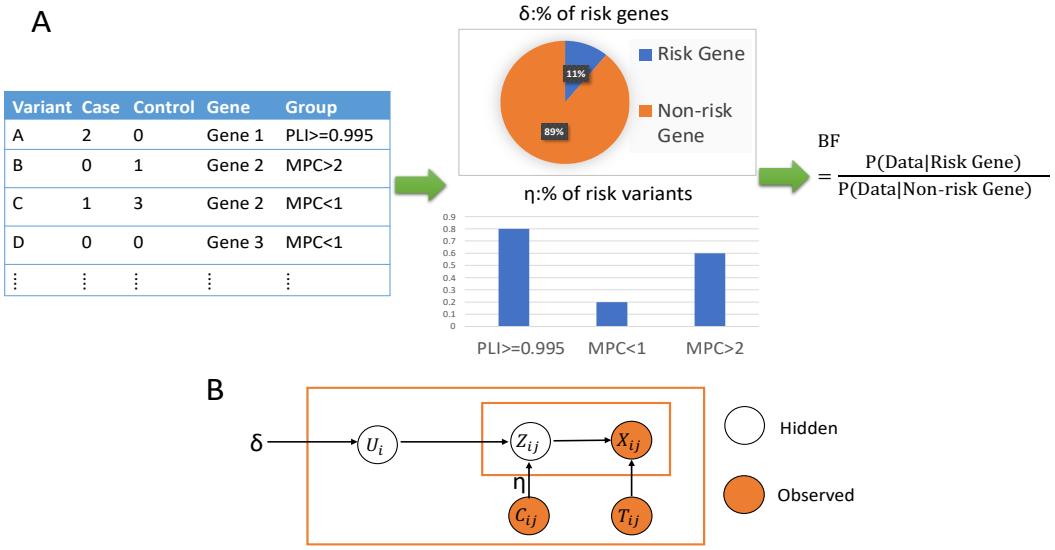

Fig 1. MIRAGE work-flow and model.(A) MIRAGE work-flow. The input has information of all the variants, including case and control counts and function groups. MIRAGE estimates the percent of risk genes and percent of risk variants in each variant class. It uses these values to compute the Bayes factor (BF) of all genes. (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.

When running MIRAGE, we first estimated the model parameters using the EM algorithm, and then computed the BF of each gene. We compared MIRAGE to several variations of burden tests (a baseline Burden test, CMC [12] and ASUM [31]), SKAT and ACAT. In practice, Burden test is often applied to different categories of variants of a gene separately to increase the power. We thus consider two other versions of burden test as well, including "Burden-adj" which tests each of the three categories separately, and returns the minimum p value of three tests; and "Burden-combine" which combines three p values by Fisher's method. The results of these two tests in simulations, however, are very similar to the baseline Burden test (Figure 2AB, Figure S3), so we consider only the baseline version here.

We compare the performance of the methods in distinguishing risk from non-risk genes, using the ROC curves ($\delta = 0.02$ and $\delta = 0.1$ in Figure 2AB, $\delta = 0.05$ and $\delta = 0.2$ in Figure S1). Under all the settings, we found that MIRAGE has Area-under-Receiver Operating of ROC (AUROC) above 80%, outperforming all other methods.

MIRAGE uses a Bayesian approach to controlling false discovery rates (FDR). We performed additional simulations to assess if 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. For each true value of $\bar{\gamma}$, ranging from 3 to 6, we ran analysis with $\bar{\gamma}$ from 3 to 6, in MIRAGE, and computed Bayesian FDR. Our results showed that Bayesian FDR are generally close to the true False Discovery Proportions ($\bar{\gamma} = 3, 5$ in Figure 2CD, $\bar{\gamma} = 4, 6$ in Figure S2).

MIRAGE identifies putative risk genes of ASD

We applied MIRAGE to whole exome sequencing (WES) data of 7291 trios of children affected with ASD and their parents. The dataset combines data from Autism

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

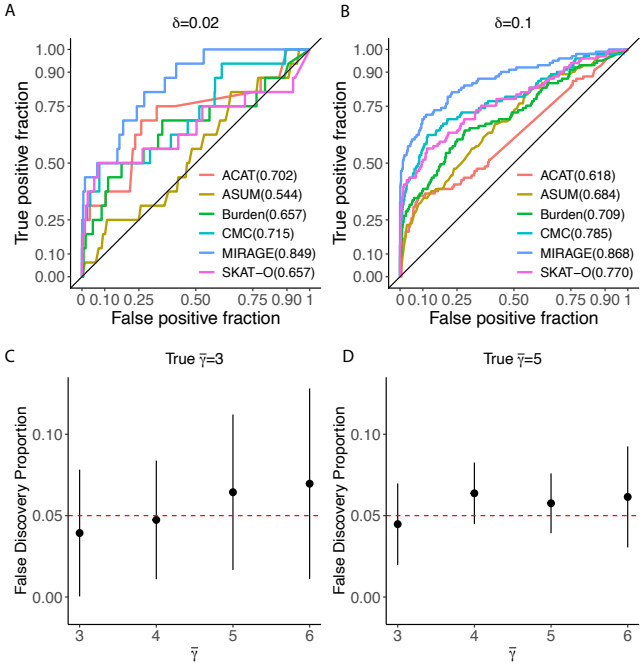

Fig 2. Simulation results. (A)(B) ROC curves of different methods for classifying risk genes. We simulated 1000 genes with varying proportion of risk genes, $\delta = 0.02, 0.1$. AUC values are shown in the bracket. Solid black reference line is in the diagonal. (C)(D) FDR calibration by MIRAGE. We simulated 20 datasets with true $\bar{\gamma} = 3$ or 5. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sequencing Consortium and the SPARK study [39]. We treated transmitted alleles as "cases" and non-transmitted ones as "controls" [23]. We considered only rare variants with allele frequency (AF) below 5%, and filtered all synonymous variants. To annotate variants, we used the same functional categories as in an earlier study [32], based on pLI scores (probability of loss-of-function intolerance) for LoF variants [33, 34], and MPC scores(Missense badness, PolyPhen-2, and Constrain scores) for missense variants [35]. 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).

We first confirmed that the ASD dataset was challenging for current methods. We applied Burden test 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 binomial test on each variants within each 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 Minor Allele Frequency (MAF) [51], as described previously [17]. The QQ plots of p values showed that none of these three methods is able to detect any significant gene (Figure 3A).

To run MIRAGE, we used an earlier estimate based on de novo mutations that 5% of genes are ASD risk genes [56]. Given that the signals in the transmission data is considerably weaker than de novo mutations [39], 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 categories of PTV (Protein-truncating variants) with high and med pLI showed the highest proportions, with 70% of variants being risk variants. Among three missense variant groups, high MPC had the largest risk variant proportion of 47%, and the last two categories showed much lower proportions. These estimates across six variant groups generally agree with the deleteriousness of these variant annotations.

With the estimated parameters, we calculated Bayes factors and posterior probabilities (PP) for every gene and perform Bayesian FDR control. At PP > 0.7 or FDR ≤ 0.2, MIRAGE was able to identify 7 putative ASK risk genes (Figure 3C). These results thus support higher sensitivity of MIRAGE in detecting risk genes than existing rare variant association methods.

To better understand the MIRAGE results, we evaluated the contribution of each of the variant group to identify what may drive the association of these genes with ASD. Following earlier work, we partitioned the evidence of each gene, in terms of $\log_2$BFLoF variants. In, 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 across genes. For instance, PLXNA1 is predominantly driven by the Low MPC group of missense variants. while the association signals for MYH10 and NAV3 are largely attributed to the High pLI group of LoF variants.In about half of the genes, two or more variant groups make substantial contributions to the signals of the genes. Overall, the high/med pLI groups of LOF variants and the high MPC variant group drive the associations in most of the genes. Together, these results highlight that MIRAGE is able to effectively combine statistical signals across multiple variant groups to discover risk genes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

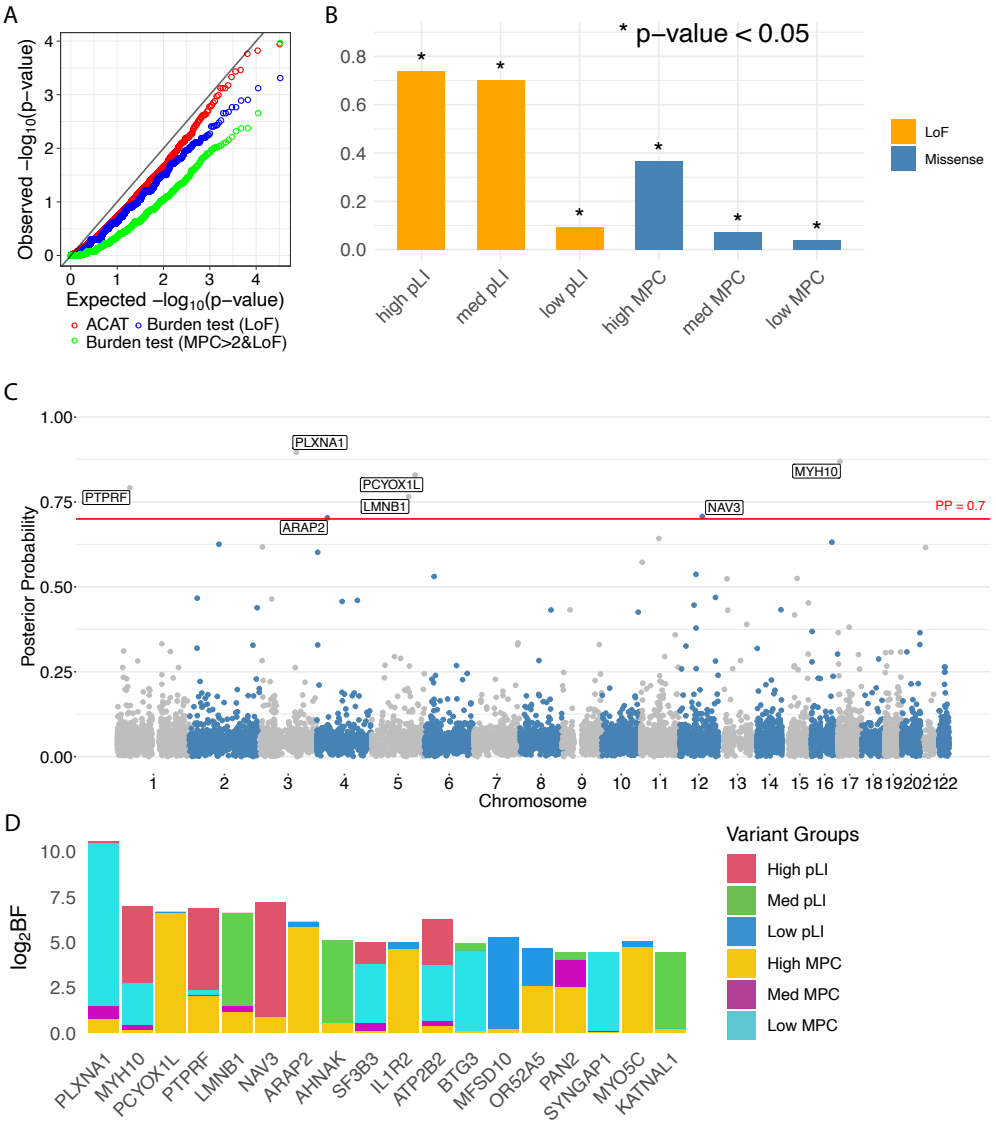

Fig 3. Summary of MIRAGE results from ASD data analysis. (A) QQ plot of all genes in the whole genome by ACAT and burden tests. (B) The proportion of risk variants in 6 variant categories. LoF categories: pLI ≥ 0.995 (high), $0.995 >$ pLI ≥ 0.5 (med), pLID < 0.5(low). Missense categories: MPC≥ 2 (high), $2 >$ MPC ≥ 1 (med), MPC < 1 (low). (C) Manhattan plot of posterior probabilities for all gene analyzed. Genes with PP > 0.7 are labeled. (D) The $\log_2$BF of a gene is partitioned into contributions of variants in each group. Shown are results of 18 genes at PP > 0.5. The negative $\log_2$BF of a variant group is truncated to 0.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

We first evaluated the candidate genes by assessing the enrichment of ASD-related gene sets. We selected top 20 genes identified by MIRAGE, and for comparison, the same number of top genes by burden tests and ACAT.

The ASD-related gene sets include known ASD genes from literature (SFARI [37]) and de novo mutation studies (TADA [52]); risk genes of Intellectual Disability (ID) [50] and Schizophrenia (SCZ) [49]; relevant biological processes including Post-synaptic density (PSD) [49] and FMRP target genes [49]; evolutionarily constrained genes from RVIS [53] and Haplo-insufficient (HI) genes [53]. 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 of ASD related gene sets. For example, 35% of MIRAGE candidate genes were likely ASD genes by SFARI ($p = 1.11 * 10^{-5}$, Fisher's Exact Test). The top genes from other methods showed lower or no enrichment. The enrichment of MIRAGE candidate genes in a range of ASD-related gene sets thus strongly supports the likely roles of these genes in ASD.

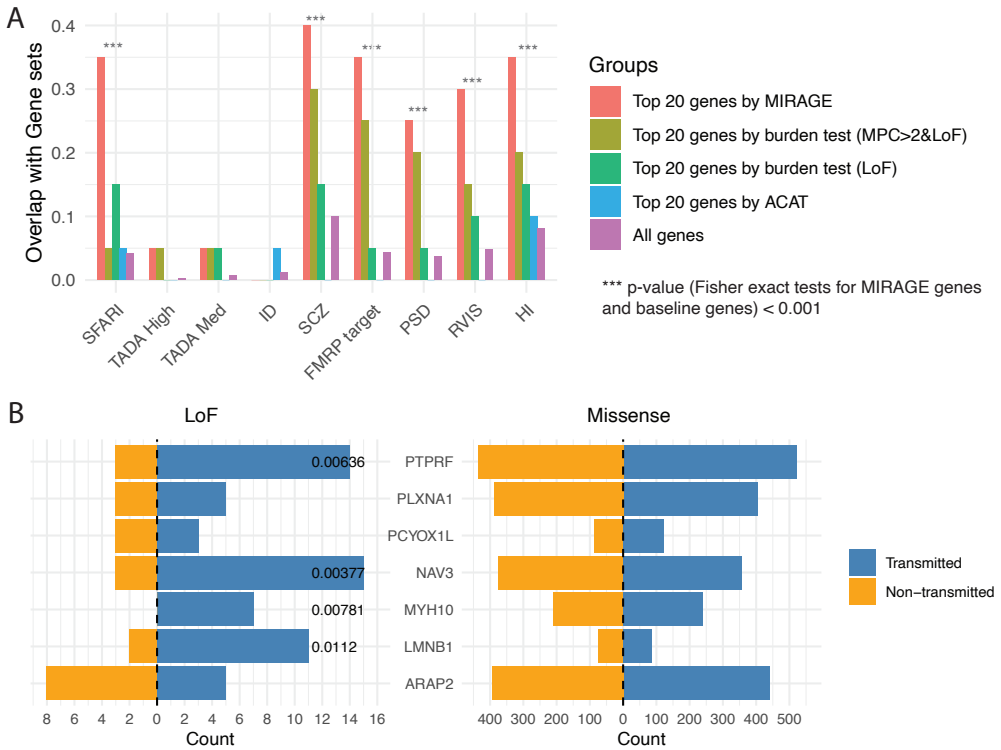

Fig 4. Evaluation of candidate genes from MIRAGE. (A) Enrichment of ASD-related gene sets in the top 20 genes found by MIRAGE, burden tests and ACAT. Enrichment test was based on Fisher's exact test. SFARI genes: Category 1 and 2. TADA high confidence genes: TADA q-value < 0.05, and TADA medium confidence genes: TADA $0.05 <$ q-value < 0.3. (B) 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 at PP > 0.7

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Table 1). Most of the genes were annotated by one or more ASD-related gene sets. Among the seven genes, PLXNA1 and NAV3 were found by transmission disequilibrium analysis in earlier studies [25, 39]. MYH10 was a ASD risk gene according to SFARI (score 2). It was supported by multiple de novo mutation studies [42, 48, 61, 62]. Among the rest, PTPRF, LMNB1 and ARAP2, have known or related functions in neurodevelopment and/or other neuropsychiatric disorders. For example, both PTPRF and ARAP2 have been implicated in SCZ. We discussed the literature support of these genes in details below in Discussion.

We evaluated the statistical support of a few genes in more details. MIRAGE allows us to quantify the statistical signal of association of individual variants in a candidate gene. Briefly, the evidence of a gene, $\log_2$BF, is the sum of $\log_2 BF$ of individual variants. This simple partition allows us to assess if the association of a gene is driven by one or few variants, and whether these variants tend to localize in a certain functional region of the proteins. For this analysis, we plot the transmitted and non-transmitted counts of each variant, and its $\log_2$BF. We highlighted the results here of LMNB1 (Figure 5). The signal is largely driven by two variants, at two locations of the gene. One of them is a frameshift indel, and the other high MPC missense variant. Both variants are transmitted 10 or more times, but non-transmitted only once. Another interesting case is ARAP2, whose association is largely driven by two missense variants that are 4bp away (Figure S4). In other genes, the association signals are more diffused, with many variants together supporting the genes (Figure S5-9).

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 4B, right). While disequilibrium was observed in the LoF group in four genes (at $p < 0.05$) (Figure 4B, left), the p-values would not pass multiple testing threshold in a genomewide setting.

These results together supported the key rationale of MIRAGE. First, by modeling the heterogeneity of variants effects, MIRAGE allows a small number of variants to drive the results, while also leveraging the collective burden that may be present in other variants. This benefit is clear in the case of missense variants, where burden test failed to detect any burden (Figure 4B, right). Secondly, by borrowing information across genes, MIRAGE is able to learn to put more emphasis on deleterious variant groups. This allows MIRAGE to extract statistical signals from relatively modest burden of LOF variants (Figure 4B, left).

Discussion

We proposed a novel method, MIRAGE, for rare variant association test. Despite the importance of rare variants, current methods are not effective at extracting statistical signals from rare variants. In practice, gene-level tests are often not better than much simpler single-variant association tests [18]. 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 framework to assess functional annotations in prioritizing deleterious variants, and to leverage these annotations in identifying risk genes. Simulations under various scenarios confirm 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 are highly enriched with known ASD and risk genes of other neuropsychiatric disorders, and other relevant gene sets. Most of the seven candidate genes, at PP > 0.7, are either reported ASD genes, or have functions in neurodevelopment or related phenotypes, representing novel and plausible ASD risk genes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

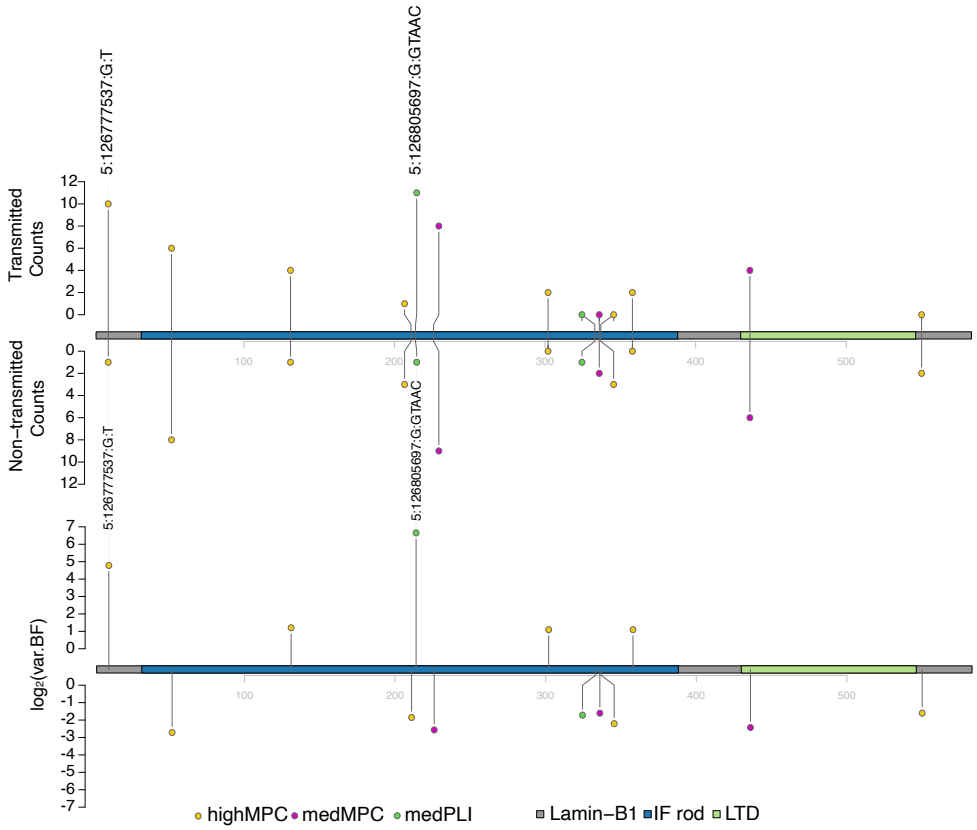

Fig 5. Evidence of LMNB1. Top: the transmission and non-transmission counts of each variant. Bottom: the evidence ($\log_2$BF) of each variant. Only informative variants, defined as BF > 2 or BF< 0.5, were plotted. The variants with BF > 5 were labelled.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Genes identified by MIRAGE (posterior probability > 0.7)

Gene	PP	FDR	ASD related gene sets	Functions/Roles
PLXNA1	0.90	0.10	PSD, FMRP target, RVIS, HI, SCZ	Transmembrane receptor. Essential for axon guidance.
MYH10	0.87	0.12	Moderate TADA, PSD, FMRP target, RVIS, HI, SCZ, SFARI	Myosin heavy chain. Microcephaly, developmental delay, known ASD gene.
PCYOX1L	0.83	0.13	-	Associated with Freemartinism, sexual development.
PTPRF	0.79	0.15	FMRP target, RVIS, HI, SCZ	Protein tyrosine phosphatase family. Synapse formation and differentiation.
LMNB1	0.77	0.17	-	Nuclear lamin B1. Microcephaly, involved in neurodevelopment.
NAV3	0.71	0.19	FMRP target, RVIS, HI, SCZ	Neuron navigator 3. Axon guidance, known ASD gene.
ARAP2	0.70	0.20	FMRP target, HI, SCZ	A protein in Rho-GTPase signaling, implicated in ASD.

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 to address this challenge. First, the effects of rare variants are likely very heterogeneous, and this is better captured by a sparse model where most rare variants have no effects on disease risk. One can see this point clearly from our estimated fractions of risk variants, especially in missense variants (Figures 3B), and from the analysis of individual gene (Figure 5. Our observation is contrary to a common, implicit assumption in development of rare variant association tests, that in a risk gene, all rare variants tend to have some effects. Secondly, using external information of variants is critical to improve the signal to noise ratio. Indeed, annotating function of variants is an active area of research, and we think some recent methods, e.g. those based on deep learning, may further boost the power of MIRAGE [38, 54, 55].

We discussed in more details the functions of the novel 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 [58]. A related protein, PTPRF interacting protein alpha 1, is a strong ASD risk gene by SFARI. The LMNB1 gene encodes Lamin B1, a component of the nuclear lamina. A number of studies have supported important roles for nuclear lamins in the central nervous system, particularly in neuronal migration in the developing brain. LMNB1 mutation in human leads to a slowly progressive disorder of central nervous system. ARAP2 is an adapter protein in Rho GTPase signaling. A number of studies have linked Rho GTPase signaling to the pathogenesis of ASD [45]. 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 [57].

MIRAGE can be further developed in several directions. First, MIRAGE is designed

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for analysis of case-control, or transmission studies. Extending it to association studies of quantitative traits would greatly broaden its applications. One possible strategy is that: instead of modeling mutation counts of individual variants, we can model some form of summary statistics, e.g. Z-scores measuring variant association with quantitative traits. This will allow us to change only the likelihood model of MIRAGE while using the same mixture prior of variant effects. Secondly, MIRAGE does not accommodate sample covariates in analysis, such as age, gender, and population ancestry. Population stratification is of particular concern as it may lead to false positive findings. Incoporating such covariants in the association analsyis is thus a critical 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 extend the prior model of MIRAGE so that the prior probability of a risk variant may depend on potentially overlapping annotations in a regression model. This type of prior has been successfully used in GWAS [59, 60] and in studies of de novo mutations [63].

Materials and Methods

MIRAGE model

MIRAGE model is shown as a probabilistic graphical model in Figure 1B. We denote U_i the indicator of whether gene i is a risk gene. U_i is a Bernoulli random variable with mean δ. Our data include the rare variant counts in cases and in 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 denote Z_{ij} the hidden indicator of whether variant j of gene i is a risk variant. For a rare variant, its allele count in a set of samples can be described by a Poisson distribution. 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:

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

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 [23]. So we have:

$$X_{ij}|Z_{ij} = 1 \sim \text{Pois}(\gamma_{ij}q_{ij}N_1), \quad X_{ij}^{(0)}|Z_{ij} = 1 \sim \text{Pois}(q_{ij}N_0). \tag{3}$$

It is generally difficult to estimate γ_{ij} for individual rare variants, so following TADA [23], we treat γ_{ij} as random, 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, 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}:

$$X_{ij}|T_{ij}, Z_{ij} = 0 \sim \text{Bin}\left(T_{ij}, \frac{N_1}{N_1 + N_0}\right) \qquad X_{ij}|T_{ij}, Z_{ij} = 1 \sim \text{Bin}\left(T_{ij}, \frac{\gamma_{ij}N_1}{\gamma_j N_1 + N_0}\right). \tag{4}$$

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

$$P(X_{ij}|T_{ij}, Z_{ij} = 1) = \int \text{Bin}\left(X_{ij}; T_{ij}, \frac{\gamma_{ij}N_1}{\gamma_{ij}N_1 + N_0}\right) \text{Gamma}(\gamma_{ij}; \bar{\gamma}, \sigma) d\gamma_j. \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The risk variant status Z_{ij}'s are not observed, so we specify the prior distribution of Z_{ij}'s. We assume each variant belongs to one variant category, denoted as c_{ij} for variant j of gene i. The prior 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 $\eta_{c_{ij}}$.

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:

$$P(\mathbf{X}|\mathbf{T}, \mathbf{C}, \delta, \eta) = \prod_i [(1-\delta)P(X_i|T_i, U_i = 0) + \delta \cdot P(X_i|T_i, C_i, U_i = 1, \eta)], \quad (6)$$

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:

$$P(X_i|T_i, U_i = 0) = \prod_j \text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right). \quad (7)$$

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

$$P(X_i|T_i, C_i, U_i = 1, \eta) = \prod_j \left[(1-\eta_{C_{ij}})\text{Bin}\left(X_{ij}; T_{ij}, \frac{N_1}{N_1 + N_0}\right) + \eta_{C_{ij}} P(X_{ij}|T_{ij}, Z_{ij} = 1)\right], \quad (8)$$

where $P(X_{ij}|T_{ij}, Z_{ij} = 1)$ is given by Equation (5). 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 $< 5\%$, this assumption is generally valid.

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). 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:

$$B_i = \frac{P(X_i|T_i, C_i, U_i = 1, \hat{\eta})}{P(X_i|T_i, U_i = 0)}, \quad \text{PP}_i = \frac{\delta B_i}{1 - \delta + \delta B_i}. \quad (9)$$

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

$$B_i = \prod_j [(1 - \eta_{C_{ij}}) + \eta_{C_{ij}} B_{ij}], \quad B_{ij} = \frac{P(X_{ij}|T_{ij}, Z_{ij} = 1)}{P(X_{ij}|T_{ij}, Z_{ij} = 0)}, \quad (10)$$

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 [23].

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Simulation procedure

We simulated a set of genes under given case-control sample sizes N_1, N_0 and the proportion of risk genes δ. We assumed each gene has a mixture of variants in different categories, with fixed proportions of variant categories in each gene. In the simulations, 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. Each category is allowed to have different mean relative risks $\bar{\gamma}$ and different proportion of risk variants, η_c for category c. We used $\bar{\gamma} = 5$ for LoF and 3 for missense categories, and η_c 0.5, 0.2 and 0.05 for the three categories, 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 $Z_{ij} \sim \text{Bernoulli}(\eta_c)$.

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. We used $\sigma = 1$ in the paper in all simulations. We sampled the allele frequency q_j 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). We sampled the total allele count in the data by $T_{ij} \sim \text{Pois}(N_1 + N_0, q_{ij})$. It is possible that T_{ij} is 0, and such variants are filtered. We split the total variant count T_{ij} into cases and controls by Binomial distribution. If $Z_{ij} = 0$, we split the count according to sample sizes, so the probability in cases is equal to $N_1/(N_1 + N_0)$. If $Z_{ij} = 1$, the probability in cases is $N_1\gamma_{ij}/(N_1\gamma_{ij} + N_0)$.

WES data of ASD

We apply MIRAGE to whole exome sequencing (WES) of autism from autism sequencing consortium, with 6026 male probands, 1265 female probands, 1158 male siblings, and 1190 female siblings.

Applying MIRAGE to ASD data

We Apply MIRAGE to WES of autism. Synonymous variants and variants with allele frequency $> 5\%$ are filtered. The annotation features utilized in this analysis were consistent with those described by Satterstrom et al. [32]. All genes in the genome-wide are used to run MIRAGE with 6 variant groups [32]. The hyperprior parameter for relative risk, $\bar{\gamma}$, is set at 6 for PTV and 3 for MPC variant sets. In the EM algorithm for estimating the parameter δ, the proportion of risk genes, and η_c, the fraction of risk variants for each category c, we randomly choose initial values, and the algorithm converges if the change of parameter estimates in two iterations is less is less than 10^{-5}. Once these parameters are estimated, their values are assumed to be known, and are used in calculating BF of each gene.

Running other programs

We used statistical software R to run other programs. For burden test, fisher exact test in R was used. R package *SKAT* (version 1.3.2.1) was used to perform SKAT-O without covariates, setting method of *SKATO*. We used R package *AssotesteR* for methods CMC and ASUM, and for CMC, three MAF cutoffs $5 \times 10^{-6}, 2 \times 10^{-5}, 5 \times 10^{-5}$ were used to partition the variants. 100 permutations were used for both CMC and ASUM as larger number of permutations doesn't make a big difference. In the ACAT analysis, the

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ACAT function from the R package-ACAT (version 0.91) was utilized, which was obtained from https://github.com/yaowuliu/ACAT. In every variant group, run ACAT on burden p values of every single variant, and variants are collapsed when its minor allele count less than 10, following Liu et al. [17], to obtain p value per group, then run ACAT on group level p values to get p value of a gene. The lollipop plots were generated by trackViewer [64].

Gene set enrichment analysis

In our analysis, we initially selected the top 20 genes as identified by MIRAGE, burden tests, and ACAT. Additionally, a comprehensive set comprising all 16,554 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 [37], including only categories 1 and 2, totaling 769 genes; (2) genes identified in de novo mutation studies, TADA [52], which are subdivided into high confidence genes (TADA q-value < 0.05, comprising 46 genes) and medium confidence genes (TADA 0.05 ¡ q-value < 0.3, comprising 133 genes); (3) genes associated with Intellectual Disability (ID) [50], encompassing 252 genes in total; (4) Schizophrenia (SCZ) risk genes [49], with a total of 1,796 genes; (5) genes involved in relevant biological processes, such as the Post-synaptic density (PSD) [49], comprising 661 genes, and FMRP target genes [49], totaling 783 genes; and (6) evolutionarily constrained genes as identified by RVIS [53], including 846 genes, alongside Haplo-insufficient (HI) genes [53], totaling 1,440 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 tests were employed.

Acknowledgements

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

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

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

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

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

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

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

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

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

9. Bis, J.C., Jian, X., Kunkle, B.W. et al. 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 (2020). https://doi.org/10.1038/s41380-018-0112-7

10. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, Churchhouse C, Chambert K, Chandler SD, Daly MJ, Dumont A, Genovese G, Hwu HG, Laird N, Kosmicki JA, Moran JL, Roe C, Singh T, Wang SH, Faraone SV, Glatt SJ, McCarroll SA, Tsuang M, Neale BM. Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nat Neurosci. 2020 Feb;23(2):185-193. doi: 10.1038/s41593-019-0564-3. Epub 2020 Jan 13. PMID: 31932770; PMCID: PMC7007385.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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

17. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019 Mar 7;104(3):410-421. doi: 10.1016/j.ajhg.2019.01.002. PMID: 30849328; PMCID: PMC6407498.

18. Timpson, N., Greenwood, C., Soranzo, N. et al. Genetic architecture: the shape of the genetic contribution to human traits and disease. Nat Rev Genet 19, 110–124 (2018). https://doi.org/10.1038/nrg.2017.101

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

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

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

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

23. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, State MW, Devlin B, Roeder K. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8):e1003671. doi: 10.1371/journal.pgen.1003671. Epub 2013 Aug 15. PMID: 23966865; PMCID: PMC3744441.

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

25. Zhou, X., Feliciano, P., Shu, C. et al. Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat Genet 54, 1305–1319 (2022). https://doi.org/10.1038/s41588-022-01148-2

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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

32. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, Peng M, Collins R, Grove J, Klei L, Stevens C, Reichert J, Mulhern MS, Artomov M, Gerges S, Sheppard B, Xu X, Bhaduri A, Norman U, Brand H, Schwartz G, Nguyen R, Guerrero EE, Dias C; Autism Sequencing Consortium; iPSYCH-Broad Consortium; Betancur C, Cook EH, Gallagher L, Gill M, Sutcliffe JS, Thurm A, Zwick ME, Børglum AD, State MW, Cicek AE, Talkowski ME, Cutler DJ, Devlin B, Sanders SJ, Roeder K, Daly MJ, Buxbaum JD. Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell. 2020 Feb 6;180(3):568-584.e23. doi: 10.1016/j.cell.2019.12.036. Epub 2020 Jan 23. PMID: 31981491; PMCID: PMC7250485.

33. Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, Lek M, Karczewski KJ, Cutler DJ, Devlin B, Roeder K, Buxbaum JD, Neale BM, MacArthur DG, Wall DP, Robinson EB, Daly MJ. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat Genet. 2017 Apr;49(4):504-510. doi: 10.1038/ng.3789. Epub 2017 Feb 13. PMID: 28191890; PMCID: PMC5496244.

34. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won HH, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG; Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91. doi: 10.1038/nature19057. PMID: 27535533; PMCID: PMC5018207.

35. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, Neale BM, and Daly MJ (2017). Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 10.1101/148353.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

39. Fu JM, Satterstrom FK, Peng M, Brand H, Collins RL, Dong S, Wamsley B, Klei L, Wang L, Hao SP, Stevens CR, Cusick C, Babadi M, Banks E, Collins B, Dodge S, Gabriel SB, Gauthier L, Lee SK, Liang L, Ljungdahl A, Mahjani B, Sloofman L, Smirnov AN, Barbosa M, Betancur C, Brusco A, Chung BHY, Cook EH, Cuccaro ML, Domenici E, Ferrero GB, Gargus JJ, Herman GE, Hertz-Picciotto I, Maciel P, Manoach DS, Passos-Bueno MR, Persico AM, Renieri A, Sutcliffe JS, Tassone F, Trabetti E, Campos G, Cardaropoli S, Carli D, Chan MCY, Fallerini C, Giorgio E, Girardi AC, Hansen-Kiss E, Lee SL, Lintas C, Ludena Y, Nguyen R, Pavinato L, Pericak-Vance M, Pessah IN, Schmidt RJ, Smith M, Costa CIS, Trajkova S, Wang JYT, Yu MHC; Autism Sequencing Consortium (ASC); Broad Institute Center for Common Disease Genomics (Broad-CCDG); iPSYCH-BROAD Consortium; Cutler DJ, De Rubeis S, Buxbaum JD, Daly MJ, Devlin B, Roeder K, Sanders SJ, Talkowski ME. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet. 2022 Sep;54(9):1320-1331. doi: 10.1038/s41588-022-01104-0. Epub 2022 Aug 18. PMID: 35982160; PMCID: PMC9653013.

40. Márquez-Luna, C., Gazal, S., Loh, PR. et al. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nat Commun 12, 6052 (2021). https://doi.org/10.1038/s41467-021-25171-9

41. Dworschak, G.C., Punetha, J., Kalanithy, J.C. et al. Biallelic and monoallelic variants in *PLXNA1* are implicated in a novel neurodevelopmental disorder with variable cerebral and eye anomalies. *Genet Med* 23, 1715–1725 (2021). `https://doi.org/10.1038/s41436-021-01196-9`

42. Fromer M., Pocklington A. J., Kavanagh D. H., Williams H. J., Dwyer S., Gormley P., et al.. De novo mutations in schizophrenia implicate synaptic networks. *Nature* 506, 179–184. *Nature* (2014). `https://doi.org/10.1038/nature12929`

43. Padiath QS. Autosomal Dominant Leukodystrophy: A Disease of the Nuclear Lamina. *Front Cell Dev Biol*. 2019 Mar 20;7:41. doi: 10.3389/fcell.2019.00041. PMID: 30949481; PMCID: PMC6435485.

44. Alcalá-Vida R, Garcia-Forn M, Castany-Pladevall C, Creus-Muncunill J, Ito Y, Blanco E, Golbano A, Crespí-Vázquez K, Parry A, Slater G, Samarajiwa S, Peiró S, Di Croce L, Narita M, Pérez-Navarro E. Neuron type-specific increase in lamin B1 contributes to nuclear dysfunction in Huntington's disease. *EMBO Mol Med*. 2021 Feb 5;13(2):e12105. doi: 10.15252/emmm.202012105. Epub 2020 Dec 28. PMID: 33369245; PMCID: PMC7863407.

45. Guo D, Yang X, Shi L. Rho GTPase Regulators and Effectors in Autism Spectrum Disorders: Animal Models and Insights for Therapeutics. Cells. 2020 Mar 31;9(4):835. doi: 10.3390/cells9040835. PMID: 32244264; PMCID: PMC7226772.

46. Belyk M, Kraft SJ, Brown S; Pediatric Imaging, Neurocognition and Genetics Study. PlexinA polymorphisms mediate the developmental trajectory of human corpus callosum microstructure. J Hum Genet. 2015 Mar;60(3):147-50. doi:

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10.1038/jhg.2014.107. Epub 2014 Dec 18. PMID: 25518740; PMCID: PMC5292032.

47. Holtz AM, VanCoillie R, Vansickle EA, Carere DA, Withrow K, Torti E, Juusola J, Millan F, Person R, Guillen Sacoto MJ, Si Y, Wentzensen IM, Pugh J, Vasileiou G, Rieger M, Reis A, Argilli E, Sherr EH, Aldinger KA, Dobyns WB, Brunet T, Hoefele J, Wagner M, Haber B, Kotzaeridou U, Keren B, Heron D, Mignot C, Heide S, Courtin T, Buratti J, Murugasen S, Donald KA, O'Heir E, Moody S, Kim KH, Burton BK, Yoon G, Campo MD, Masser-Frye D, Kozenko M, Parkinson C, Sell SL, Gordon PL, Prokop JW, Karaa A, Bupp C, Raby BA. Heterozygous variants in MYH10 associated with neurodevelopmental disorders and congenital anomalies with evidence for primary cilia-dependent defects in Hedgehog signaling. Genet Med. 2022 Oct;24(10):2065-2078. doi: 10.1016/j.gim.2022.07.005. Epub 2022 Aug 18. PMID: 35980381; PMCID: PMC10765599.

48. Tuzovic L, Yu L, Zeng W, Li X, Lu H, Lu HM, Gonzalez KD, Chung WK. A human de novo mutation in MYH10 phenocopies the loss of function mutation in mice. Rare Dis. 2013 Aug 14;1:e26144. doi: 10.4161/rdis.26144. PMID: 25003005; PMCID: PMC3927488.

49. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kähler A, Duncan L, Stahl E, Genovese G, Fernández E, Collins MO, Komiyama NH, Choudhary JS, Magnusson PK, Banks E, Shakir K, Garimella K, Fennell T, DePristo M, Grant SG, Haggarty SJ, Gabriel S, Scolnick EM, Lander ES, Hultman CM, Sullivan PF, McCarroll SA, Sklar P. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014 Feb 13;506(7487):185-90. doi: 10.1038/nature12975. Epub 2014 Jan 22. PMID: 24463508; PMCID: PMC4136494.

50. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, Thiruvahindrapuram B, Xu X, Ziman R, Wang Z, Vorstman JA, Thompson A, Regan R, Pilorge M, Pellecchia G, Pagnamenta AT, Oliveira B, Marshall CR, Magalhaes TR, Lowe JK, Howe JL, Griswold AJ, Gilbert J, Duketis E, Dombroski BA, De Jonge MV, Cuccaro M, Crawford EL, Correia CT, Conroy J, Conceição IC, Chiocchetti AG, Casey JP, Cai G, Cabrol C, Bolshakova N, Bacchelli E, Anney R, Gallinger S, Cotterchio M, Casey G, Zwaigenbaum L, Wittemeyer K, Wing K, Wallace S, van Engeland H, Tryfon A, Thomson S, Soorya L, Rogé B, Roberts W, Poustka F, Mouga S, Minshew N, McInnes LA, McGrew SG, Lord C, Leboyer M, Le Couteur AS, Kolevzon A, Jiménez González P, Jacob S, Holt R, Guter S, Green J, Green A, Gillberg C, Fernandez BA, Duque F, Delorme R, Dawson G, Chaste P, Café C, Brennan S, Bourgeron T, Bolton PF, Bölte S, Bernier R, Baird G, Bailey AJ, Anagnostou E, Almeida J, Wijsman EM, Vieland VJ, Vicente AM, Schellenberg GD, Pericak-Vance M, Paterson AD, Parr JR, Oliveira G, Nurnberger JI, Monaco AP, Maestrini E, Klauck SM, Hakonarson H, Haines JL, Geschwind DH, Freitag CM, Folstein SE, Ennis S, Coon H, Battaglia A, Szatmari P, Sutcliffe JS, Hallmayer J, Gill M, Cook EH, Buxbaum JD, Devlin B, Gallagher L, Betancur C, Scherer SW. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014 May 1;94(5):677-94. doi: 10.1016/j.ajhg.2014.03.018. Epub 2014 Apr 24. PMID: 24768552; PMCID: PMC4067558.

51. Karczewski, K.J., Francioli, L.C., Tiao, G., et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581:434–443, 2020. `https://doi.org/10.1038/s41586-020-2308-7`.

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

52. Stephan J. Sanders, X. He, A. Jeremy Willsey, A. Gulhan Ercan-Sencicek, Kaitlin E. Samocha, A. Ercument Cicek, Michael T. Murtha, Vanessa Hus Bal, Sarah L. Bishop, Shan Dong, Amanda P. Goldberg, Jinlu C, John F. Keaney 3rd, Lambertus Klei, Judith D. Mandell, Daniel Moreno-De-Luca, Christopher S. Poultney, Elise B. Robinson, Lana Smith, T. Nowlan Solli, M. Y. Su, N. A. Teran, M. F. Walker, Donna M. Werling, A. L. Beaudet, R. M. Cantor, E. Fombonne, D. H. Geschwind, D. E. Grice, C. Lord, J. K. Lowe, S. M. Mane, D. M. Martin, E. M. Morrow, M. E. Talkowski, J. S. Sutcliffe, C. A. Walsh, T. W. Yu; Autism Sequencing Consortium; D. H. Ledbetter, C. L. Martin, E. H. Cook, J. D. Buxbaum, M. J. Daly, B. Devlin, K. Roeder, M. W. State. "Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci." Neuron, 87(6):1215–1233, 2015. DOI: 10.1016/j.neuron.2015.09.016. PMID: 26402605; PMCID: PMC4624267.

53. Sergey Petrovski, Q. Wang, Erin L. Heinzen, Andrew S. Allen, David B. Goldstein. "Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes." PLOS Genetics, 9(8): e1003709, 2013. DOI: https://doi.org/10.1371/journal.pgen.1003709.

54. K. Minton, "Predicting variant pathogenicity with AlphaMissense," Nat Rev Genet, vol. 24, p. 804, 2023. DOI: `https://doi.org/10.1038/s41576-023-00668-9`

55. N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, V. Ntranos, "Genome-wide prediction of disease variant effects with a deep protein language model," Nat Genet, vol. 55, no. 9, pp. 1512–1522, 2023. DOI: `https://doi.org/10.1038/s41588-023-01465-0`

56. S. De Rubeis, X. He, A. P. Goldberg, et al., "Synaptic, transcriptional and chromatin genes disrupted in autism," Nature, vol. 515, no. 7526, pp. 209–215, 2014. DOI: `https://doi.org/10.1038/nature13772`

57. C. E. Tanna, L. B. Goss, C. G. Ludwig, P. W. Chen, "Arf GAPs as Regulators of the Actin Cytoskeleton—An Update," Int J Mol Sci, vol. 20, no. 2, art. 442, 2019. Published 2019 Jan 21. DOI: `https://doi.org/10.3390/ijms20020442`

58. F. Cornejo, B. I. Cortés, G. M. Findlay, G. I. Cancino, "LAR Receptor Tyrosine Phosphatase Family in Healthy and Diseased Brain," Front Cell Dev Biol, vol. 9, art. 659951, 2021. Published 2021 Dec 13. DOI: `https://doi.org/10.3389/fcell.2021.659951`

59. J. K. Pickrell, "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits," Am J Hum Genet, vol. 94, no. 4, pp. 559–573, 2014. [Published correction appears in Am J Hum Genet. 2014 Jul 3;95(1):126]. DOI: `https://doi.org/10.1016/j.ajhg.2014.03.004`

60. X. Wen, Y. Lee, F. Luca, R. Pique-Regi, "Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors," Am J Hum Genet, vol. 98, no. 6, pp. 1114–1129, 2016. DOI: `https://doi.org/10.1016/j.ajhg.2016.03.029`

61. I. Hlushchenko, P. Khanal, A. Abouelezz, V. O. Paavilainen, P. Hotulainen, "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, vol. 12, art. 217, 2018. Published 2018 Aug 3. DOI: `https://doi.org/10.3389/fncel.2018.00217`

bioRxiv preprint doi: https://doi.org/10.1101/2024.10.24.620131; this version posted October 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

62. I. Iossifov, B. J. O'Roak, S. J. Sanders, et al., "The contribution of de novo coding mutations to autism spectrum disorder," Nature, vol. 515, no. 7526, pp. 216–221, 2014. DOI: https://doi.org/10.1038/nature13908

63. Y. Liu, Y. Liang, A. E. Cicek, et al., "A Statistical Framework for Mapping Risk Genes from De Novo Mutations in Whole-Genome-Sequencing Studies," Am J Hum Genet, vol. 102, no. 6, pp. 1031–1047, 2018. DOI: https://doi.org/10.1016/j.ajhg.2018.03.023

64. J. Ou, L.J. Zhu, "trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data," *Nat Methods*, vol. 16, pp. 453–454, 2019.DOI: https://doi.org/10.1038/s41592-019-0430-y