# Quantifying concordant genetic effects of *de novo* mutations on multiple disorders

Hanmin Guo[1,2], Lin Hou[1,2,3], Yu Shi[4], Sheng Chih Jin[5], Xue Zeng[6,7], Boyang Li[8], Richard P. Lifton[6,7], Martina Brueckner[6,9], Hongyu Zhao[6,8,10,#], Qiongshi Lu[11,#,*]

[1] Center for Statistical Science, Tsinghua University, Beijing, China
[2] Department of Industrial Engineering, Tsinghua University, Beijing, China
[3] MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China
[4] Yale School of Management, Yale University, New Haven, CT, USA
[5] Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[6] Department of Genetics, Yale University, New Haven, CT, USA
[7] Laboratory of Human Genetics and Genomics, Rockefeller University, New York, USA
[8] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
[9] Department of Pediatrics, Yale University School of Medicine, New Haven, CT, USA
[10] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
[11] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

[#] These authors should be considered shared last author

* Correspondence:
Dr. Qiongshi Lu
Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison
425 Henry Mall
Madison, WI, USA 53706
qlu@biostat.wisc.edu

# Abstract

Exome sequencing on tens of thousands of parent-proband trios has identified numerous deleterious *de novo* mutations (DNMs) and implicated risk genes for many disorders. Recent studies have suggested shared genes and pathways are enriched for DNMs across multiple disorders. However, existing analytic strategies only focus on genes that reach statistical significance for multiple disorders and require large trio samples in each study. As a result, these methods are not able to characterize the full landscape of genetic sharing due to polygenicity and incomplete penetrance. In this work, we introduce EncoreDNM, a novel statistical framework to quantify shared genetic effects between two disorders characterized by concordant enrichment of DNMs in the exome. EncoreDNM makes use of exome-wide, summary-level DNM data, including genes that do not reach statistical significance in single-disorder analysis, to evaluate the overall and annotation-partitioned genetic sharing between two disorders. Applying EncoreDNM to DNM data of nine disorders, we identified abundant pairwise enrichment correlations, especially in genes intolerant to pathogenic mutations and genes highly expressed in fetal tissues. These results suggest that EncoreDNM improves current analytic approaches and may have broad applications in DNM studies.

# Introduction

*De novo* mutations (DNMs) can be highly deleterious and provide important insights into the genetic cause for disease[1]. As the cost of sequencing continues to drop, whole-exome sequencing (WES) studies conducted on tens of thousands of family trios have pinpointed numerous risk genes for a variety of disorders[2-4]. In addition, accumulating evidence suggests that risk genes enriched for pathogenic DNMs may be shared by multiple disorders[5-9]. These shared genes could reveal biological pathways that play prominent roles in disease etiology and shed light on clinically heterogeneous yet genetically related diseases[7-9].

Most efforts to identify shared risk genes directly compare genes that are significantly associated with each disorder[10,11]. There have been some successes with this approach in identifying shared genes and pathways (e.g., chromatin modifiers) underlying developmental disorder (DD), autism spectrum disorder (ASD), and congenital heart disease (CHD), thanks to the large trio samples in these studies[3,4,12], whereas findings in smaller studies remain suggestive[13,14]. Even in the largest studies to date, statistical power remains moderate for risk genes with weaker effects[3,15]. It is estimated that more than 1,000 genes associated with DD remain undetected[3]. Therefore, analytic approaches that only account for top significant genes cannot capture the full landscape of genetic sharing in multiple disorders. Recently, a Bayesian framework was proposed to jointly analyze DNM data of two diseases and improve risk gene mapping[9]. Although some parameters in this framework can quantify shared genetics between diseases, the statistical property of these parameter estimates have not been studied. There is a pressing need for powerful, robust, and interpretable methods that quantify concordant DNM association patterns for multiple disorders using exome-wide DNM counts.

Recent advances in estimating genetic correlations using summary data from genome-wide association studies (GWAS) may provide a blueprint for approaching this problem in DNM research[16]. Modeling “omnigenic” associations as independent random effects, linear mixed-effects models leverage genome-wide association profiles to quantify the correlation between additive genetic components of multiple complex traits[17-20]. These methods have identified ubiquitous genetic correlations across many human traits and revealed significant and robust genetic correlations that could not be inferred from significant GWAS associations alone[21-24].

Here, we introduce EncoreDNM (**En**richment **cor**relation **e**stimator for ***De Novo* M**utations), a novel statistical framework that leverages exome-wide DNM counts, including genes that do not reach exome-wide statistical significance in single-disorder analysis, to estimate concordant DNM associations between disorders. EncoreDNM uses a generalized linear mixed-effects model to quantify the occurrence of DNMs while accounting for *de novo* mutability of each gene and technical inconsistencies between studies. We demonstrate the performance of EncoreDNM through extensive simulations and analyses of DNM data of nine disorders.

# Results

## Method overview

DNM counts in the exome deviate from the null (i.e., expected counts based on *de novo* mutability) when mutations play a role in disease etiology. Disease risk genes will show enrichment for deleterious DNMs in probands and non-risk genes may be slightly depleted for DNM counts. Our goal is to estimate the correlation of such deviation between two disorders, which we refer to as the DNM enrichment correlation. More specifically, we use a pair of mixed-effects Poisson regression models to quantify the occurrence of DNMs in two studies.

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \end{bmatrix}\right),$$

$$\log\left(\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \end{bmatrix}\right) = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \log\left(\begin{bmatrix} 2N_1 m_i \\ 2N_2 m_i \end{bmatrix}\right) + \begin{bmatrix} \phi_{i1} \\ \phi_{i2} \end{bmatrix},$$

$$\begin{bmatrix} \phi_{i1} \\ \phi_{i2} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right).$$

Here, $Y_{i1}, Y_{i2}$ are the DNM counts for the $i$-th gene and $N_1, N_2$ are the number of parent-proband trios in two studies, respectively. The log Poisson rates of DNM occurrence are decomposed into three components: the elevation component, the background component, and the deviation component. The elevation component $\beta_k$ $(k = 1,2)$ is a fixed effect term adjusting for systematic, exome-wide bias in DNM counts. One example of such bias is the batch effect caused by different sequencing and variant calling pipelines in two studies. The elevation parameter $\beta_k$ tends to be larger when DNMs are over-called with higher sensitivity and smaller when DNMs are under-called with higher specificity[25]. The background component $\log(2N_k m_i)$ is a gene-specific fixed effect that reflects the expected mutation counts determined by the genomic sequence context under the null[26]. $m_i$ is the *de novo* mutability for the $i$-th gene, and $2N_1 m_i$ and $2N_2 m_i$ are the expected DNM counts in the $i$-th gene under the null in two studies. The deviation component $\phi_{ik}$ is a gene-specific random effect that quantifies the deviation of DNM profile from what is expected under the null. $\phi_{i1}$ and $\phi_{i2}$ follow a multivariate normal distribution with dispersion parameters $\sigma_1$ and $\sigma_2$ and a correlation $\rho$. DNM enrichment correlation is denoted by $\rho$ and is our main parameter of interest. It quantifies the concordance of DNM burden in two disorders.

Parameters in this model can be estimated using a Monte Carlo maximum likelihood estimation (MLE) procedure. Standard errors of the estimates are obtained through inversion of the observed Fisher information matrix. In practice, we use annotated DNM data as input and fit mixed-effects Poisson models for each variant class separately: loss of function (LoF), deleterious missense (Dmis, defined as MetaSVM-deleterious), tolerable missense (Tmis, defined as MetaSVM-tolerable), and synonymous (**Figure 1**). More details about model setup and parameter estimation are discussed in **Methods**.

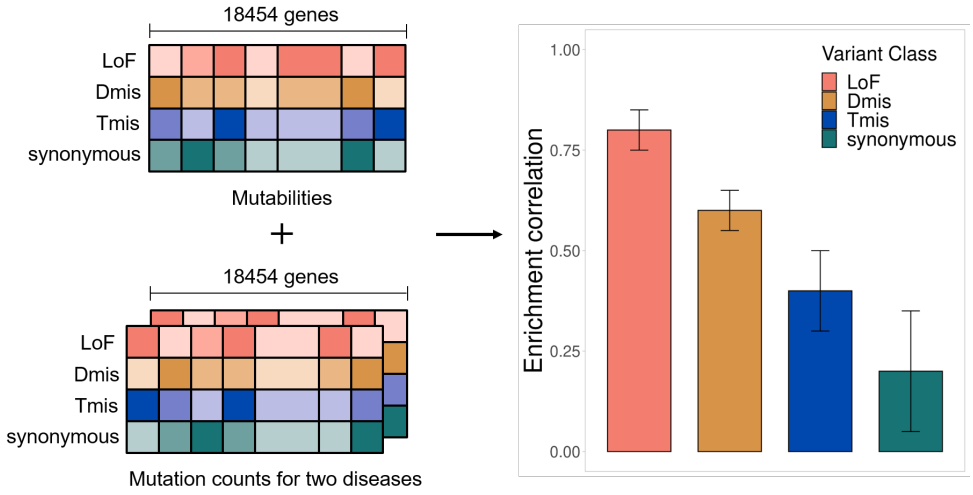


**Figure 1. EncoreDNM workflow.** The inputs of EncoreDNM are *de novo* mutability of each gene and exome-wide, annotated DNM counts from two studies. We fit a mixed-effects Poisson model to estimate the DNM enrichment correlation between two disorders for each variant class.

## Simulation results

We conducted simulations to assess the parameter estimation performance of EncoreDNM in various settings. We focused on two variant classes, i.e., Tmis and LoF variants, since they have the highest and lowest median mutabilities in the exome. We used EncoreDNM to estimate the elevation parameter $\beta$, dispersion parameter $\sigma$, and enrichment correlation $\rho$ (**Methods**). Under various parameter settings, EncoreDNM always provided unbiased estimation of the parameters (**Figure 2** and **Supplementary Figures 1-2**). Furthermore, the 95% Wald confidence intervals achieved coverage rates close to 95% under all simulation settings, demonstrating the effectiveness of EncoreDNM to provide accurate statistical inference.

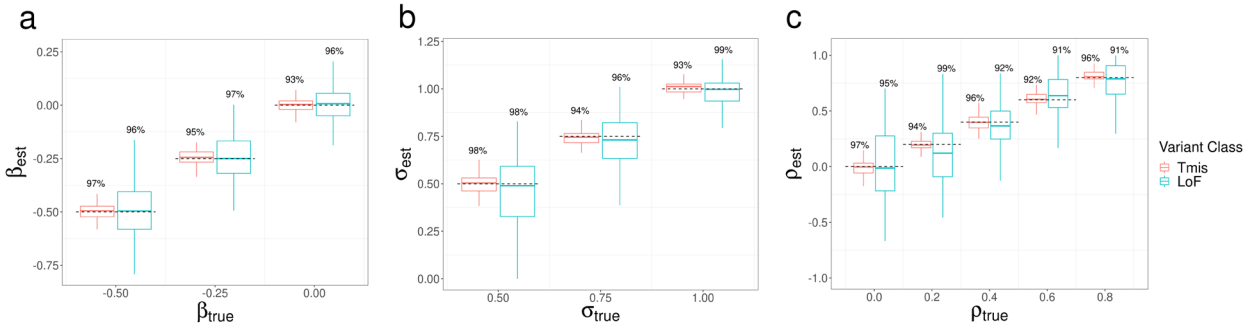


**Figure 2. Parameter estimation results of EncoreDNM.** (a) Boxplot of $\beta$ estimates in single-trait analysis with $\sigma$ fixed at 0.75. (b) Boxplot of $\sigma$ estimates in single-trait analysis with $\beta$ fixed at -0.25. (c) Boxplot of $\rho$ estimates in cross-trait analysis with $\beta$ and $\sigma$ fixed at -0.25 and 0.75. True parameter values are marked by dashed lines. The number above each box represents the coverage rate of 95% Wald confidence intervals. Each simulation setting was repeated 100 times.

Next, we compared the performance of EncoreDNM with mTADA[9], a Bayesian framework that could estimate the proportion of shared risk genes for two disorders. First, we simulated DNM data under the mixed-effects Poisson model. We evaluated two methods across a range of

combinations of elevation parameter, dispersion parameter, and sample size for two disorders. The type-I error rates for our method were well-calibrated in all parameter settings, but mTADA produced false positive findings when the observed DNM counts were relatively small (e.g., due to reduced elevation or dispersion parameters or a lower sample size; **Figure 3a**). We also assessed the statistical power of two approaches under a baseline setting where type-I errors for both methods were controlled. As enrichment correlation increased, EncoreDNM achieved universally greater statistical power compared to mTADA (**Figure 3b**).

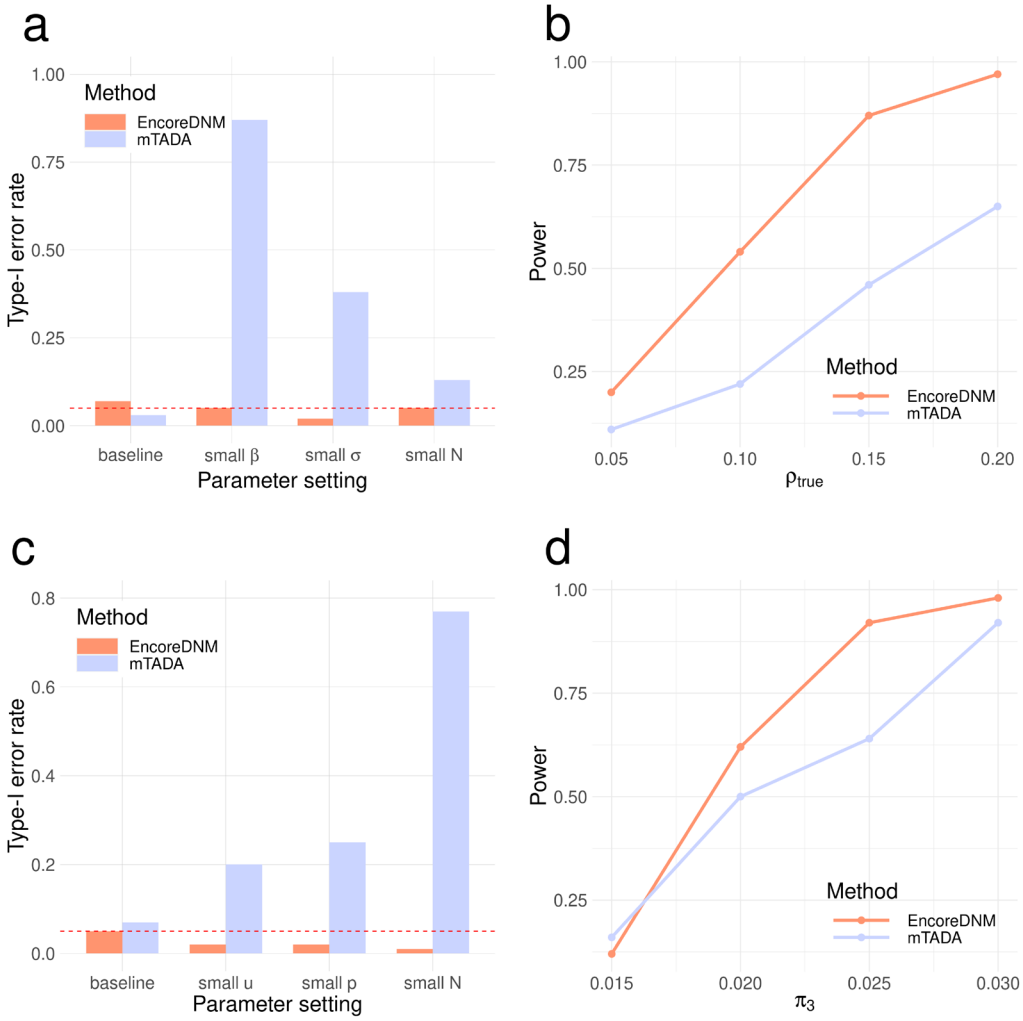


**Figure 3. Comparison of EncoreDNM and mTADA.** (a) Type-I error rates under a mixed-effects Poisson regression model. $(\beta, \sigma, N)$ were fixed at (-0.25, 0.75, 5000) under the baseline setting, (-1, 0.75, 5000) under a setting with small $\beta$, (-0.25, 0.5, 5000) under a setting with small $\sigma$, and (-0.25, 0.75, 1000) under a setting with small N for two disorders. (b) Statistical power of two methods under a mixed-effects Poisson regression model as the enrichment correlation increases. Parameters $(\beta, \sigma, N)$ were fixed at (-0.25, 0.75, 5000) for both disorders. (c) Type-I error rates under a multinomial model. $(u, p, N, \pi^S)$ were fixed at (0.95, 0.25, 5000, 0.1) under the baseline setting, (0.75, 0.25, 5000, 0.1) under a setting with small u (i.e., reduced total DNM counts), (0.95, 0.15, 5000, 0.1) under a setting with small p (i.e., fewer probands explained by DNMs), and (0.95, 0.25, 1000, 0.1) under a setting with lower sample size. (d) Statistical power under a multinomial model with varying proportion of shared causal genes. Parameters $(u, p, N, \pi^S)$ were fixed at (0.95, 0.25, 5000, 0.1) for both disorders. Each simulation setting was repeated 100 times.

To ensure a fair comparison, we also considered a mis-specified model setting where we randomly distributed the total DNM counts for each disorder into all genes with an enrichment in causal genes (**Methods**). EncoreDNM showed well-controlled type-I error across all simulation settings, whereas severe type-I error inflation arose for mTADA when the total mutation count, the proportion of probands that can be explained by DNMs, or the sample size were small (**Figure 3c**). Furthermore, we compared the statistical power of two methods under this model in a baseline setting where type-I error was controlled. EncoreDNM showed higher statistical power compared to mTADA as the fraction of shared causal genes increased (**Figure 3d**).

## Pervasive enrichment correlation of damaging DNMs among developmental disorders

We applied EncoreDNM to DNM data of nine disorders (**Supplementary Table 1**; **Methods**): DD (n=31,058; number of trios)[3], ASD (n=6,430)[4], schizophrenia (SCZ; n=2,772)[15], CHD (n=2,645)[12], intellectual disability (ID; n=820)[2], Tourette disorder (TD; n=484)[27], epileptic encephalopathies (EP; n=264)[13], cerebral palsy (CP; n=250)[14], and congenital hydrocephalus (CH; n=232)[28]. In addition, we also included 1,789 trios comprising healthy parents and unaffected siblings of ASD probands as controls[29].

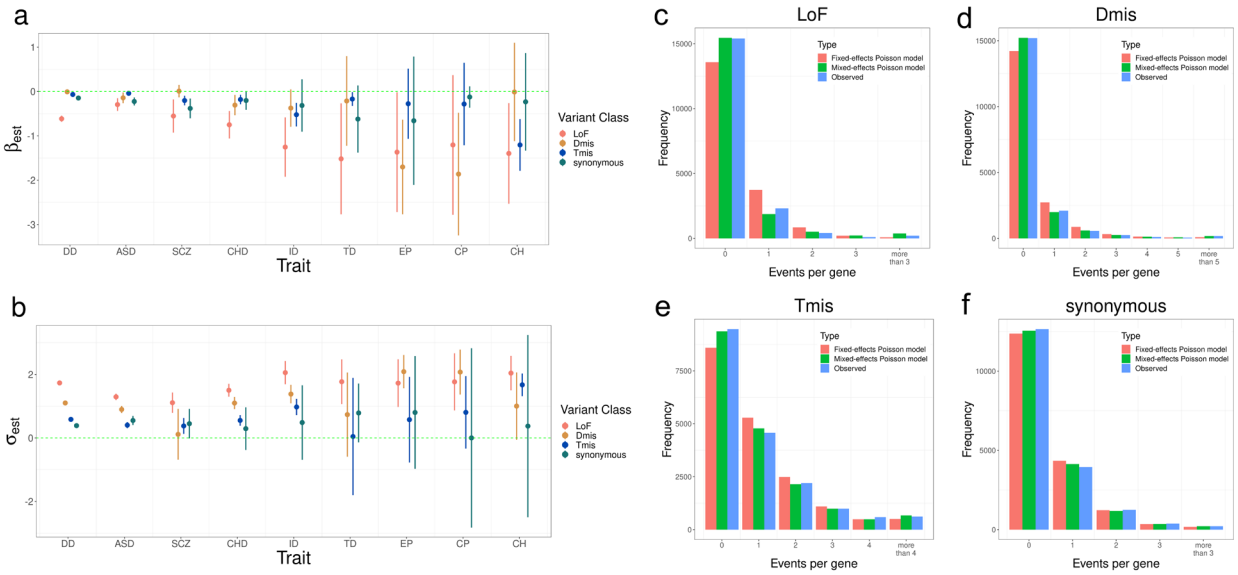


**Figure 4. Model fitting results for nine disorders.** (a, b) Estimation results of $\beta$ and $\sigma$ for nine disorders and four variant classes. Error bars represent 1.96*standard errors. (c-f) Distribution of DNM events per gene in four variant classes for DD. Red and green bars represent the expected frequency of genes under the fixed-effects and mixed-effects Poisson regression models, respectively. Blue bars represent the observed frequency of genes.

We first performed single-trait analysis for each disorder. The estimated elevation parameters (i.e., $\beta$) were negative for almost all disorders and variant classes (**Figure 4a**), with LoF variants showing particularly lower parameter estimates. This may be explained by more stringent quality control in LoF variant calling[12] and potential survival bias[30]. It is also consistent with a depletion of LoF DNMs in healthy control trios[7]. The dispersion parameter estimates (i.e., $\sigma$) were higher

for LoF variants than other variant classes (**Figure 4b**), which is consistent with our expectation that LoF variants have stronger effects on disease risk and should show a larger deviation from the null mutation rate in disease probands. We compared the goodness of fit of our proposed mixed-effects Poisson model to a simpler fixed-effects model without the deviation component (**Methods**). The expected distribution of recurrent DNM counts showed substantial and statistically significant improvement under the mixed-effects Poisson model (**Figures 4c-f** and **Supplementary Figure 3**).

Next, we estimated pairwise DNM enrichment correlations for 9 disorders. In total, we identified 25 pairs of disorders with significant correlations at a false discovery rate (FDR) cutoff of 0.05 (**Figure 5** and **Supplementary Figure 4**), including 12 significant correlations for LoF variants, 7 for Dmis variants, 5 for Tmis variants, and only 1 significant correlation for synonymous variants. Notably, all significant correlations are positive (**Supplementary Table 2**). No significant correlation was identified between any disorder and healthy controls (**Supplementary Figure 5**). The number of identified significant correlations for each disorder was proportional to the sample size in each study (Spearman correlation = 0.70) with controls being a notable outlier (**Supplementary Figure 6**).

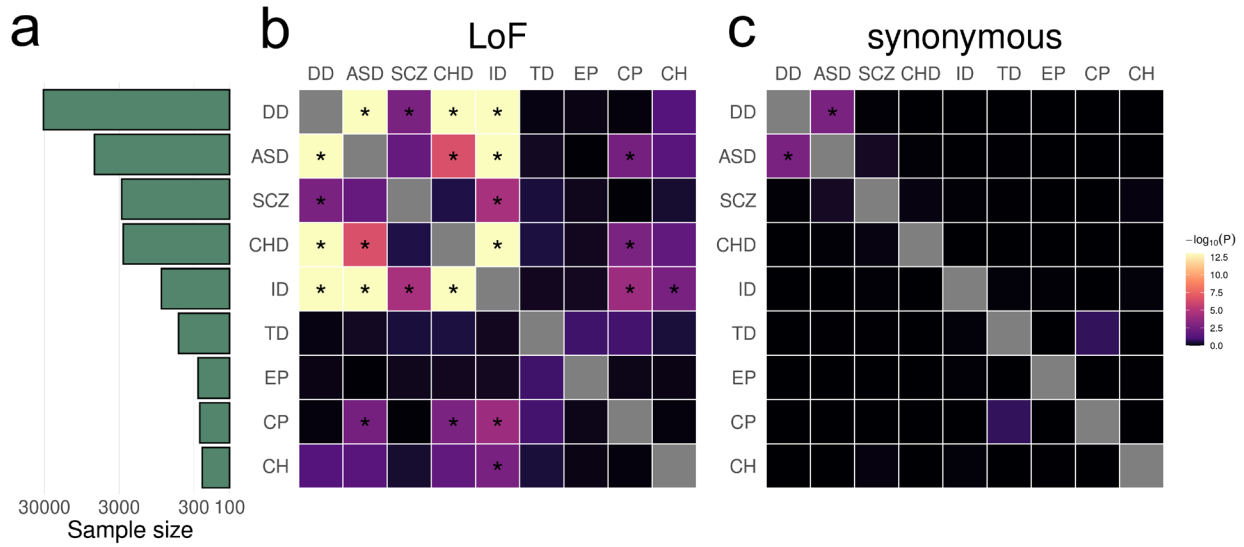


**Figure 5. EncoreDNM identifies pervasive enrichment correlations of damaging DNMs among nine disorders.** (a) shows sample size (i.e., number of trios) for each disease. X-axis denotes sample size on the log scale. (b, c) Heatmap of enrichment correlations for LoF and synonymous DNMs among nine disorders. Significant correlations (FDR<0.05) are marked by asterisks. Results with $-\log_{10} P > 13$ are truncated to 13 for visualization purpose.

We identified highly concordant and significant LoF DNM enrichment among DD, ASD, ID, and CHD, which is consistent with previous reports[8-10,31]. SCZ shows highly significant LoF correlations with DD and ID (p=2.0e-3 and 3.7e-5), hints at a correlation with ASD (p=0.012), but does not correlate strongly with CHD. The positive enrichment correlation between ASD and CP in LoF variants ($\rho$=0.81, p=3.3e-3) is consistent with their co-occurrence[32]. The high enrichment correlation between ID and CP in LoF variants ($\rho$=0.68, p=1.0e-4) is consistent with the

associations between ID and motor or non-motor abnormalities caused by CP[33]. A previous study also suggested significant genetic sharing of ID and CP by overlapping genes harboring rare damaging variants[14]. Here, we obtained consistent results after accounting for *de novo* mutabilities and potential confounding bias.

Some significant correlations identified in our analysis are consistent with phenotypic associations in epidemiological studies, but have not been reported using genetic data to the extent of our knowledge. For example, the LoF enrichment correlation between CHD and CP ($\rho$=0.88, p=1.7e-3) is consistent with findings that reduced supply of oxygenated blood in fetal brain due to cardiac malformations may be a risk factor for CP[34]. The enrichment correlation between ID and CH in LoF variants ($\rho$=0.63, p=2.4e-3) is consistent with lower intellectual performance in a proportion of children with CH[35].

Genes showing pathogenic DNMs in multiple disorders may shed light on the mechanisms underlying enrichment correlations (**Supplementary Table 3**). We identified five genes (i.e. *CTNNB1*, *NBEA*, *POGZ*, *SPRED2*, and *KMT2C*) with LoF DNMs in five different disorders and 21 genes had LoF DNMs in four disorders (**Supplementary Table 4**). These 26 genes with LoF variants in at least four disorders were significantly enriched for 63 gene ontology (GO) terms with FDR<0.05 (**Supplementary Table 5**). Chromatin organization (p=7.8e-11), nucleoplasm (p=2.8e-10), chromosome organization (p=6.8e-10), histone methyltransferase complex (p=1.4e-9), and positive regulation of gene expression (p=2.2e-9) were the most significantly enriched GO terms. One notable example consistently included in these gene sets is *CTNNB1* (**Supplementary Figure 7**). It encodes $\beta$-catenin, is one of the only two genes reaching genome-wide significance in a recent WES study for CP[14], and also harbors multiple LoF variants in DD, ID, ASD, and CHD. It is a fundamental component of the canonical Wnt signaling pathway which is known to confer genetic risk for ASD[36]. We also identified 157 recurrent LoF mutations in 45 genes (**Supplementary Table 6**). Most of these recurrent mutations were identified in DD due to its large sample size, but one mutation was identified in joint comparison of other disorders. *FBXO11*, encoding the F-box only protein 31, shows two recurrent p.Ser831fs LoF variants in ASD and CH (**Supplementary Figure 8**; p=1.9e-3; **Methods**). The F-box protein constitutes a substrate-recognition component of the SCF (SKP1-cullin-F-box) complex, an E3-ubiquitin ligase complex responsible for ubiquitination and proteasomal degradation[37]. DNMs in *FBXO11* have been previously implicated in severe ID individuals with autistic behavior problem[38] and neurodevelopmental disorder[39].

For comparison, we also applied mTADA to the same nine disorders and control trios. In total, mTADA identified 117 disorder pairs with significant genetic sharings at an FDR cutoff of 0.05 (**Supplementary Table 7** and **Supplementary Figure 9**). Notably, we identified significant synonymous DNM correlations for all 36 disorder pairs and between all disorders and healthy controls (**Supplementary Figure 10**). These results are consistent with the simulation results and suggest a substantially inflated false positive rate in mTADA.

Further, we applied cross-trait linkage disequilibrium (LD) score regression[18] to five of the nine disorders with publicly available GWAS summary statistics (**Supplementary Table 8**): ASD (n=46,350)[40], SCZ (n=161,405)[41], cognitive performance (used as a proxy for ID; n=257,841)[42], TD (n=14,307)[43], and epilepsy (n=44,889)[44]. In total, we identified 6 trait pairs with significant genetic correlations at an FDR cutoff of 0.05 (**Supplementary Table 9**), suggesting consistent findings made from GWAS and DNM data (Spearman correlation = 0.70; **Supplementary Figure 11**).

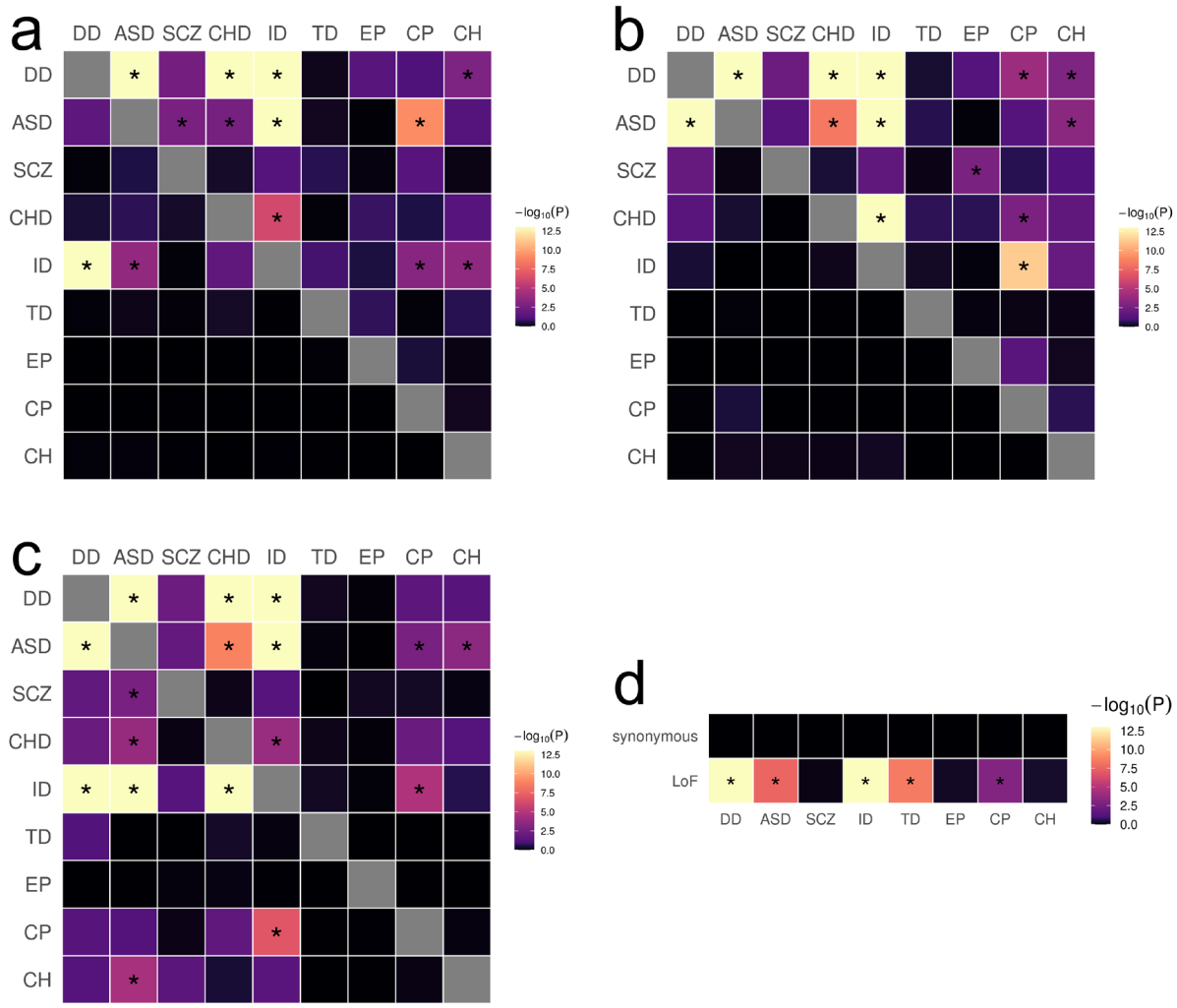


**Figure 6. DNM enrichment correlations in disease-relevant gene sets.** (a) Enrichment correlations in High-pLI genes (upper triangle) and Low-pLI genes (lower triangle) for LoF variants. (b) Enrichment correlations in HBE genes (upper triangle) and LBE genes (lower triangle) for LoF variants. (c) Enrichment correlations in HHE genes (upper triangle) and LHE genes (lower triangle) for LoF variants. (d) Enrichment correlations in CHD-related pathways for LoF and synonymous variants. Significant correlations (FDR<0.05) are marked by asterisks. Results with $-\log_{10} P > 13$ are truncated to 13 for visualization purpose.

## Partitioning DNM enrichment correlation by gene set

To gain biological insights into the shared genetic architecture of nine disorders, we repeated EncoreDNM correlation analysis in several gene sets. First, we defined genes with high/low probability of intolerance to LoF variants using pLI scores[45], and identified genes with high/low brain expression (HBE/LBE)[46] (**Methods**; **Supplementary Table 10**). We identified 11 and 12 disorder pairs showing significant enrichment correlations for LoF DNMs in high-pLI genes and HBE genes, respectively (**Figure 6a-b**). We observed fewer significant correlations for Dmis and Tmis variants in these gene sets (**Supplementary Figures 12-13**). All identified significant correlations were positive (**Supplementary Tables 11-12**). No significant correlations were identified for synonymous variants (**Supplementary Figures 12-13**) or between disorders and controls (**Supplementary Figures 14-15**).

We observed a clear enrichment of significant correlations in disease-relevant gene sets. Overall, high-pLI genes showed substantially stronger correlations across disorders than genes with low pLI (one-sided Kolmogorov-Smirnov test; p=2.3e-6). Similarly, enrichment correlations were stronger in HBE genes than in LBE genes (p=8.8e-7). Among the 11 disorder pairs showing significant enrichment correlations in high-pLI genes, two pairs, i.e., ASD-SCZ ($\rho$=0.68, p=2.4e-3) and DD-CH ($\rho$=0.43, p=1.5e-3), were not identified in the exome-wide analysis. We also identified four novel disorder pairs with significant correlations in HBE genes, including DD-CP ($\rho$=0.80, p=9.5e-5), DD-CH ($\rho$=0.67, p=1.4e-3), ASD-CH ($\rho$=0.82, p=4.7e-4), and SCZ-EP ($\rho$=0.66, p=2.0e-3). These novel enrichment correlations are consistent with known comorbidities between these disorders[47,48] and findings based on significant risk genes[8,28,49,50].

Furthermore, we estimated DNM enrichment correlations in genes with high/low expression in mouse developing heart (HHE/LHE)[7] (**Methods**; **Supplementary Table 10**). We identified 9 significant enrichment correlations for LoF variants in HHE genes (**Figure 6c**). Strength of enrichment correlations did not show a significant difference between HHE and LHE genes (p=0.846), possibly due to a lack of cardiac disorders in our analysis. Finally, we estimated enrichment correlations between CHD and other disorders in known pathways for CHD[51] (**Methods**; **Supplementary Table 10**). We identified 5 significant correlations for LoF variants (**Figure 6d**), including a novel correlation between CHD and TD ($\rho$=0.93, p=3.3e-9). Of note, arrhythmia caused by CHD is a known risk factor for TD[52]. In these analyses, all significant enrichment correlations were positive (**Supplementary Tables 13-14**) and other variant classes showed generally weaker correlations than LoF variants (**Supplementary Figures 16-17**). We did not observe significant correlations in these gene sets between disorders and controls (**Supplementary Figures 18-19**).

# Discussion

In this paper, we introduced EncoreDNM, a novel statistical framework to quantify correlated DNM enrichment between two disorders. Through extensive simulations and analyses of DNM data for nine disorders, we demonstrated that our proposed mixed-effects Poisson regression model provides unbiased parameter estimates, shows well-controlled type-I error, and is robust to exome-wide technical biases. Leveraging exome-wide DNM counts and genomic context-based mutability data, EncoreDNM achieves superior fit for real DNM datasets compared to simpler models and provides statistically powerful and computationally efficient estimation of DNM enrichment correlation. Further, EncoreDNM can quantify concordant genetic effects for user-defined variant classes within pre-specified gene sets, thus is suitable for exploring diverse types of hypotheses and can provide crucial biological insights into the shared genetic etiology in multiple disorders.

Multi-trait analyses of GWAS data have revealed shared genetic architecture among many neuropsychiatric traits[22,53,54]. These findings have led to the identification of pleiotropic variants, genes, and hub genomic regions underlying many traits and have revealed multiple psychopathological factors jointly affecting human neurological phenotypes[55,56]. Although emerging evidence suggests that causal DNMs underlying several disorders with well-powered studies (e.g., CHD and neurodevelopmental disorders[7]) may be shared, our understanding of the extent and the mechanism underlying such sharing remains incomplete. Applied to DNM data for nine disorders, EncoreDNM identified pervasive enrichment correlations of DNMs. We observed particularly strong correlations in pathogenic variant classes (e.g., LoF and Dmis variants) and disease-relevant genes (e.g., genes with high pLI and genes highly expressed in relevant tissues). Genes underlying these correlations were significantly enriched in pathways involved in chromatin organization and modification and gene expression regulation. The DNM correlations were substantially attenuated in genes with lower expression and genes with frequent occurrences of LoF variants in the population. A similar attenuation was observed in less pathogenic variant classes (e.g., synonymous variants). Further, no significant correlations were identified between any disorder and healthy controls. These results lay the groundwork for future investigations of pleiotropic mechanisms of DNMs.

Our study has some limitations. First, a main goal in DNM research is to identify disease risk genes. EncoreDNM leverages exome-wide DNM counts to quantify shared genetic basis in multiple disorders but does not improve the analysis of gene-disease associations. Second, EncoreDNM assumes probands from different input studies to be independent. In rare cases when two studies have overlapping proband samples, enrichment correlation estimates may be inflated and must be interpreted with caution. Finally, genetic correlation methods based on GWAS summary data provided key motivations for the mixed-effects Poisson regression model in our study. Built upon genetic correlations, a plethora of methods have been developed in the

GWAS literature to jointly model more than two GWAS[57], identify and quantify common factors underlying multiple traits[58,59], estimate causal effects among different traits[60], and identify pleiotropic genomic regions through hypothesis-free scans[23]. Future directions of EncoreDNM include using enrichment correlation to improve gene discovery, learning the directional effects and the causal structure underlying multiple disorders, and dynamically searching for gene sets and annotation classes with shared genetic effects without pre-specifying the hypothesis.

Taken together, we provide a new analytic approach to an important problem in DNM studies. We believe EncoreDNM improves the statistical rigor in multi-disorder DNM modeling and opens up many interesting future directions in both method development and follow-up analyses in WES studies. As trio sample size in WES studies continues to grow, EncoreDNM will have broad applications and can greatly benefit DNM research.

# Methods

### Statistical Model

For a single study, we assume that DNM counts in a given variant class (e.g., synonymous variants) follow a mixed-effects Poisson model:

$$Y_i \sim \text{Poisson}(\lambda_i),$$
$$\log(\lambda_i) = \beta + \log(2Nm_i) + \phi_i,$$
$$\phi_i \sim \text{N}(0, \sigma^2), \qquad \text{for } i = 1, \ldots, G,$$

where $Y_i$ is the DNM count in the $i$-th gene, $N$ is the number of trios, $m_i$ is the *de novo* mutabilityfor the $i$-th gene (i.e., mutation rate per chromosome per generation) which is known *a priori*[26] (**Supplementary Table 15**), and $G$ is the total number of genes in the study. The elevation parameter $\beta$ quantifies the global elevation of mutation rate compared to mutability estimates based on genomic sequence alone. Gene-specific deviation from expected DNM rate is quantified by random effect $\phi_i$ with a dispersion parameter $\sigma$. Here, the $\phi_i$ are assumed to be independent across different genes, in which case the observed DNM counts of different genes are independent.

Next, we describe how we expand this model to quantify the shared genetics of two disorders. We assume DNM counts in a given variant class for two diseases follow:

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \end{bmatrix} \sim \text{Poisson}\left(\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \end{bmatrix}\right),$$
$$\log\left(\begin{bmatrix} \lambda_{i1} \\ \lambda_{i2} \end{bmatrix}\right) = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \log\left(\begin{bmatrix} 2N_1 m_i \\ 2N_2 m_i \end{bmatrix}\right) + \begin{bmatrix} \phi_{i1} \\ \phi_{i2} \end{bmatrix},$$
$$\begin{bmatrix} \phi_{i1} \\ \phi_{i2} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right),$$

where $Y_{i1}, Y_{i2}$ are the DNM counts for the $i$-th gene and $N_1, N_2$ are the trio sizes in two studies, respectively. Similar to the single-trait model, $m_i$ is the mutability for the $i$-th gene. $\beta_1, \beta_2$ are the elevation parameters, and $\phi_{i1}, \phi_{i2}$ are the gene-specific random effects with dispersion parameters $\sigma_1, \sigma_2$, for two disorders respectively. $\rho$ is the enrichment correlation which quantifies the concordance of the gene-specific DNM burden between two disorders. Here, $\beta_1, \beta_2, \sigma_1, \sigma_2, \rho$ are unknown parameters. The gene specific effects for two disorders are assumed to be independent for different genes. We also assume that there is no shared sample for two disorders, in which case $Y_{i1}$ is independent with $Y_{i2}$ given $\begin{bmatrix}\lambda_{i1}\\ \lambda_{i2}\end{bmatrix}$.

## Parameter estimation

We implement an MLE procedure to estimate unknown parameters. For single-trait analysis, the log-likelihood function can be expressed as follows:

$$l(\beta, \sigma|\boldsymbol{Y}) = \sum_{i=1}^{G} \log\left[\int \exp(-\lambda_i)\, \lambda_i^{Y_i} * f(\phi_i) d\phi_i\right] + C,$$

where $\boldsymbol{Y} = [Y_1, \dots, Y_G]^T$, $\lambda_i = 2Nm_i \exp(\beta + \phi_i)$, $C = -\sum_{i=1}^{G} \log(Y_i!)$, and $f(\phi_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\phi_i^2}{2\sigma^2}\right)$. Note that there is no closed form for the integral in the log-likelihood function. Therefore, we use Monte Carlo integration to evaluate the log-likelihood function. Let $\phi_{ij} = \sigma\xi_{ij}$, where the $\xi_{ij}$ are independently and identically distributed random variables following a standard normal distribution. We have

$$l(\beta, \sigma|\boldsymbol{Y}) \approx l'(\beta, \sigma|\boldsymbol{Y}) = \sum_{i=1}^{G} \log\left[\sum_{j=1}^{M} \exp\left(-\lambda_{ij}\right) \lambda_{ij}^{Y_i}\right] + C,$$

where $\lambda_{ij} = 2Nm_i \exp(\beta + \sigma\xi_{ij})$, and $M$ is the Monte Carlo sample size which is set to be 1000. Then, we could obtain the MLE of $\beta, \sigma$ through maximization of $l'(\beta, \sigma|\boldsymbol{Y})$. We obtain the standard error of the MLE through inversion of the observed Fisher information matrix.

The estimation procedure can be generalized to multi-trait analysis. Log-likelihood function can be expressed as follows:

$$l(\beta_1, \beta_2, \sigma_1, \sigma_2, \rho|\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \sum_{i=1}^{G} \log\left[\int \exp(-\lambda_{i1} - \lambda_{i2})\, \lambda_{i1}^{Y_{i1}} \lambda_{i2}^{Y_{i2}} * f(\phi_{i1}, \phi_{i2}) d\phi_{i1} d\phi_{i2}\right] + C,$$

where $\boldsymbol{Y}_1 = [Y_{11}, \dots, Y_{G1}]^T$, $\boldsymbol{Y}_2 = [Y_{12}, \dots, Y_{G2}]^T$, $\lambda_{i1} = 2N_1 m_i \exp(\beta_1 + \phi_{i1})$, $\lambda_{i2} = 2N_2 m_i \exp(\beta_2 + \phi_{i2})$, $C = -\sum_{i=1}^{G} [\log(Y_{i1}!) + \log(Y_{i2}!)]$, and $f(\phi_{i1}, \phi_{i2}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sqrt{1-\rho^2}}\left(\frac{\phi_{i1}^2}{\sigma_1^2} + \frac{\phi_{i2}^2}{\sigma_2^2} - \frac{2\rho\phi_{i1}\phi_{i2}}{\sigma_1\sigma_2}\right)\right]$. We use Monte Carlo integration to evaluate the log-likelihood function. Let $\phi_{i1j} =$

$\sigma_1\xi_{i1j}$ and $\phi_{i2j} = \sigma_2\left(\rho\xi_{i1j} + \sqrt{1-\rho^2}\xi_{i2j}\right)$, where the $\xi_{i1j}$ and $\xi_{i2j}$ are independently and identically distributed random variables following a standard normal distribution. We have

$$l(\beta_1, \beta_2, \sigma_1, \sigma_2, \rho | \boldsymbol{Y}_1, \boldsymbol{Y}_2) \approx l'(\beta_1, \beta_2, \sigma_1, \sigma_2, \rho | \boldsymbol{Y}_1, \boldsymbol{Y}_2) = \sum_{i=1}^{G} \log\left[\sum_{j=1}^{M} \exp\left(-\lambda_{i1j} - \lambda_{i2j}\right) \lambda_{i1j}^{Y_{i1}} \lambda_{i2j}^{Y_{i2}}\right] + C,$$

where $\lambda_{i1j} = 2N_1 m_i \exp(\beta_1 + \sigma_1\xi_{i1j})$ and $\lambda_{i2j} = 2N_2 m_i \exp\left[\beta_2 + \sigma_2\left(\rho\xi_{i1j} + \sqrt{1-\rho^2}\xi_{i2j}\right)\right]$. Then, we obtain the MLE of $\beta_1, \beta_2, \sigma_1, \sigma_2, \rho$ through maximization of $l'(\beta_1, \beta_2, \sigma_1, \sigma_2, \rho | \boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Standard error of MLE can be obtained through inversion of the observed Fisher information matrix.

## DNM data and variant annotation

We obtained DNM data from published studies (**Supplementary Table 1**). DNM data for EP from the original release[13] were not in an editable format and were instead collected from denovo-db[61]. We used ANNOVAR[62] to annotate all DNMs. Synonymous variants were determined based on the 'synonymous SNV' annotation in ANNOVAR; Variants with 'startloss', 'stopgain', 'stoploss', 'splicing', 'frameshift insertion', 'frameshift deletion', or 'frameshift substitution' annotations were classified as LoF; Dmis variants were defined as nonsynonymous SNVs predicted to be deleterious by MetaSVM[63]; nonsynonymous SNVs predicted to be tolerable by MetaSVM were classified as Tmis. Other DNMs which did not fall into these categories were removed from the analysis. For each variant class, we estimated the mutability of each gene using a sequence-based mutation model[26] while adjusting for the sequencing coverage factor based on control trios as previously described[12] (**Supplementary Table 15**). We included 18,454 autosomal protein-coding genes in our analysis. *TTN* was removed due to its substantially larger size.

## Implementation of mTADA

The software mTADA requires the following parameters as inputs: proportion of risk genes ($\pi_1^S, \pi_2^S$), mean relative risks ($\bar{\gamma}_1^S, \bar{\gamma}_2^S$), and dispersion parameters ($\bar{\beta}_1^S, \bar{\beta}_2^S$) for both disorders. We used extTADA[10] to estimate these parameters as suggested by the mTADA paper[9]. mTADA reported the estimated proportion of shared risk genes $\pi_3$ (posterior mode of $\pi_3$) and its corresponding 95% credible interval $[\mathrm{LB}, \mathrm{UB}]$. We considered $\pi_1^S * \pi_2^S$ as the expected proportion of shared risk genes, and there is significant genetic sharing between two disorders when $\mathrm{LB} > \pi_1^S * \pi_2^S$. P-value for $\pi_3$ was calculated by comparing $\pi_1^S * \pi_2^S$ to the posterior distribution of $\pi_3$. Number of MCMC chain was set as 2 and number of iterations was set as 10,000.

## Simulation settings

We assessed the performance of EncoreDNM under the mixed-effects Poisson model. We performed simulations for two variant classes: Tmis and LoF variants, which have the largest and

the smallest median mutability values across all genes. First, we performed single-trait simulations to assess estimation precision of elevation parameter $\beta$ and dispersion parameter $\sigma$. We set the true values of $\beta$ to be -0.5, -0.25, and 0, and the true values of $\sigma$ to be 0.5, 0.75, and 1. These values were chosen based on the estimated parameters in real DNM data analyses and ensured simulation settings to be realistic. Next, we performed simulations for cross-trait analysis to assess estimation precision of enrichment correlation $\rho$, whose true values were set to be 0, 0.2, 0.4, 0.6, and 0.8. Sample size for each disorder was set to be 5,000. Coverage rate was calculated as the percentage of simulations that the 95% Wald confidence interval covered the true parameter value. Each parameter setting was repeated 100 times.

We also carried out simulations to compare the performance of EncoreDNM and mTADA. Type I error and statistical power for EncoreDNM were calculated as the proportion of simulation repeats that p-value for enrichment correlation $\rho$ was smaller than 0.05. and the proportion of simulation repeats that p-value for estimated proportion of shared risk genes $\pi_3$ was smaller than 0.05 was used for mTADA. We aggregated all variant classes together, so mutability for each gene was determined as the sum of mutabilities across four variant classes (i.e. LoF, Dmis, Tmis, and synonymous).

First, we simulated DNM data under the mixed-effects Poisson model. To see whether two methods would produce false positive findings, we performed simulations under the null hypothesis that the enrichment correlation $\rho$ is zero. We compared two methods under a range of parameter combinations of $(\beta, \sigma, N)$ for both disorders: (-0.25, 0.75, 5000) for the baseline setting, (-1, 0.75, 5000) for a setting with small $\beta$, (-0.25, 0.5, 5000) for a setting with small $\sigma$, and (-0.25, 0.75, 1000) for a setting with small sample size. We also assessed the statistical power of two methods under the alternative hypothesis. True value of enrichment correlation $\rho$ was set to be 0.05, 0.1, 0.15, and 0.2. In the power analysis, parameters $(\beta, \sigma, N)$ were fixed at (-0.25, 0.75, 5000) as in the baseline setting when both methods had well-controlled type-I error.

To ensure a fair comparison, we also compared EncoreDNM and mTADA under a multinomial model, which is different from the data generation processes for the two approaches. For each disorder ($k = 1,2$), we randomly selected causal genes of proportion $\pi_k^S$. A proportion (i.e., $\pi_3$) of causal genes overlap between two disorders. We assumed that the total DNM count to follow a Poisson distribution: $C_k \sim \text{Poisson}(u_k * 2N_k \sum_{i=1}^{G} m_i)$, where $u_k$ represents an elevation factor to represent systematic bias in the data. Let $\mathbf{Y}_k$ denote the vector of DNMs counts in the exome, $\boldsymbol{m}$ denote the vector of mutability values for all genes, and $\boldsymbol{m}_{causal,k}$ denote the vector of mutability with values set to be 0 for non-causal genes of disorder $k$. We assumed that a proportion $p_k$ of the probands could be attributed to DNMs burden in causal genes, and $1 - p_k$ of the probands obtained DNMs by chance:

$$\mathbf{Y}_k = \mathbf{Y}_{causal,k} + \mathbf{Y}_{background,k},$$
$$\mathbf{Y}_{causal,k} \sim \text{Multinomial}\left(p_k C_k, \boldsymbol{m}_{causal,k}\right),$$
$$\mathbf{Y}_{background,k} \sim \text{Multinomial}\left((1 - p_k) C_k, \boldsymbol{m}\right).$$

To check whether false positive findings could arise, we performed simulations under the null hypothesis that $\pi_3 = \pi_1^S * \pi_2^S$ across a range of parameter combinations of $(u, p, N, \pi^S)$ for both disorders: (0.95, 0.25, 5000, 0.1) for the baseline setting, (0.75, 0.25, 5000, 0.1) for a setting with small $u$ (i.e., reduced total mutation count), (0.95, 0.15, 5000, 0.1) for a setting with small $p$ (fewer probands explained by DNMs), and (0.95, 0.25, 1000, 0.1) for a setting with smaller sample size. We also assessed the statistical power of two methods under the alternative hypothesis that $\pi_3 > \pi_1^S * \pi_2^S$. In power analysis, $(u, p, N, \pi^S)$ were fixed at (0.95, 0.25, 5000, 0.1) as in the baseline setting when type-I error for both methods were well-calibrated.

## Comparison to the fixed-effects Poisson model

For single-trait analysis, the fixed-effects Poisson model assumes that

$$Y_i \sim Poisson(\lambda_i),$$
$$log(\lambda_i) = \beta + log(2Nm_i), \quad \text{for } i = 1, \ldots, G.$$

Note that the fixed-effects Poisson model is a special case of our proposed mixed-effects Poisson model when $\sigma = 0$. We compared the two models using likelihood ratio test. Under the null hypothesis that $\sigma = 0$, $2(l_{alt} - l_{null}) \sim \frac{1}{2}\chi_1^2$ asymptotically, where $l_{alt}$ and $l_{null}$ represent the log likelihood of the fitted mixed-effects and fixed-effects Poisson models respectively.

## Recurrent genes and DNMs

We used FUMA[64] to perform GO enrichment analysis for genes harboring LoF DNMs in multiple disorders. Due to potential sample overlap between the studies of DD[3] and ID[2], we excluded ID from the analysis of recurrent DNMs. We calculated the probability of observing two identical DNMs in two disorders using a Monte Carlo simulation method. For each disorder, we simulated exome-wide DNMs profile from a multinomial distribution, where the size was fixed at the observed DNM count and the per-base mutation probability was determined by the tri-nucleotide base context. We repeated the simulation procedure 100,000 times to evaluate the significance of recurrent DNMs. Lollipop plots for recurrent mutations were generated using MutationMapper on the cBio Cancer Genomics Portal[65].

## Implementation of cross-trait LD score regression

We used cross-trait LDSC[18] to estimate genetic correlations between disorders. LD scores were computed using European samples from the 1000 Genomes Project Phase 3 data[66]. Only HapMap 3 SNPs were used as observations in the explanatory variable with the --merge-alleles flag. Intercepts were not constrained in the analyses.

## Estimating enrichment correlation in gene sets

Genes with a high/low probability of intolerance to LoF variants (high-pLI/low-pLI) were defined as the 4,614 genes in the upper/lower quartiles of pLI scores[45]. Genes with high/low brain expression (HBE/LBE) were defined as the 4,614 genes in the upper/lower quartiles of expression in the human fetal brain[46]. Genes with high/low heart expression (HHE/LHE) were defined as the 4,614 genes in the upper/lower quartiles of expression in the developing heart of embryonic mouse[67]. Five biological pathways have been reported to be involved in CHD: chromatin remodeling, Notch signaling, cilia function, sarcomere structure and function, and RAS signaling[51]. We extracted 1730 unique genes that belong to these five pathways from the gene ontology database[68] and referred to the union set as CHD-related genes. We repeated EncoreDNM enrichment correlation analysis in these gene sets. One-sided Kolmogorov-Smirnov test was used to assess the statistical difference between enrichment correlation signal strength in different gene sets.

## URLs

GWAS summary statistics data of ASD, SCZ, and TD were downloaded on the PGC website, https://www.med.unc.edu/pgc/download-results/; Summary statistics of cognitive performance were downloaded on the SSGAC website, https://www.thessgac.org/data; Summary statistics of epilepsy were downloaded on the epiGAD website, http://www.epigad.org/; pLI scores were downloaded from gnomAD v3.1 repository https://gnomad.broadinstitute.org/downloads; mTADA, https://github.com/hoangtn/mTADA; denovo-db, https://denovo-db.gs.washington.edu/denovo-db/; MutationMapper on cBioPortal, https://www.cbioportal.org/mutation_mapper; LDSC, https://github.com/bulik/ldsc.

## Code availability

EncoreDNM software is available at https://github.com/ghm17/EncoreDNM.

## Acknowledgements

LH acknowledges research support from the National Science Foundation of China (Grant No. 12071243) and Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01). QL acknowledges research support from the University of Wisconsin-Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation and the Waisman Center pilot grant program at University of Wisconsin-Madison. HZ acknowledges research support from the National Institutes of Health (Grant No. R03HD100883)

## Author contribution

H.G., L.H., and Q.L. designed the study.
H.G. performed data analysis and implemented the software.
Y.S. implemented an early version of the method.
S.C.J., X.Z., and B.L. assisted DNM and mutability data preparation.
R.P.L and M.B. advised on disease biology, data interpretation, and genetic issues.
H.Z. and Q.L. advised on statistical issues.
H.G., L.H., and Q.L. wrote the manuscript.
All authors contributed in manuscript editing and approved the manuscript.

## Competing financial interests

The authors declare no competing financial interests.

# References


1. Veltman, J.A. & Brunner, H.G. De novo mutations in human genetic disease. *Nature Reviews Genetics* **13**, 565-575 (2012).
2. Lelieveld, S.H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature neuroscience* **19**, 1194-1196 (2016).
3. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757-762 (2020).
4. Satterstrom, F.K. *et al.* Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568-584. e23 (2020).
5. Hoischen, A., Krumm, N. & Eichler, E.E. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nature neuroscience* **17**, 764 (2014).
6. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179-184 (2014).
7. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-1266 (2015).
8. Li, J. *et al.* Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from NPdenovo database. *Molecular psychiatry* **21**, 290-297 (2016).
9. Nguyen, T.-H. *et al.* mTADA is a framework for identifying risk genes from de novo mutations in multiple traits. *Nature Communications* **11**, 2929 (2020).
10. Nguyen, H.T. *et al.* Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome medicine* **9**, 114 (2017).
11. Willsey, A.J. *et al.* The psychiatric cell map initiative: a convergent systems biological approach to illuminating key molecular pathways in neuropsychiatric disorders. *Cell* **174**, 505-520 (2018).
12. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nature genetics* **49**, 1593 (2017).
13. Allen, A.S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221 (2013).
14. Jin, S.C. *et al.* Mutations disrupting neuritogenesis genes confer risk for cerebral palsy. *Nature Genetics* **52**, 1046-1056 (2020).
15. Howrigan, D.P. *et al.* Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations. *Nature Neuroscience* **23**, 185-193 (2020).
16. Zhang, Y. *et al.* Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. *Briefings in bioinformatics* (2021).
17. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540-2542 (2012).
18. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature genetics* **47**, 1236 (2015).
19. Lu, Q. *et al.* A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics. *Am J Hum Genet* **101**, 939-964 (2017).
20. Ning, Z., Pawitan, Y. & Shen, X. High-definition likelihood inference of genetic correlations across

human complex traits. *Nature genetics* **52**, 859-864 (2020).

21. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *The American Journal of Human Genetics* **101**, 737-751 (2017).
22. Brainstorm, C. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**(2018).
23. Guo, H., Li, J.J., Lu, Q. & Hou, L. Detecting local genetic correlations with scan statistics. *Nature Communications* **12**, 2033 (2021).
24. Zhang, Y. *et al.* SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome biology* **22**, 1-30 (2021).
25. Wei, Q. *et al.* A Bayesian framework for de novo mutation calling in parents-offspring trios. *Bioinformatics* **31**, 1375-1381 (2015).
26. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944-950 (2014).
27. Willsey, A.J. *et al.* De novo coding variants are strongly associated with Tourette disorder. *Neuron* **94**, 486-499. e9 (2017).
28. Jin, S.C. *et al.* Exome sequencing implicates genetic disruption of prenatal neuro-gliogenesis in sporadic congenital hydrocephalus. *Nature medicine* **26**, 1754-1765 (2020).
29. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nature genetics* **47**, 582-588 (2015).
30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291 (2016).
31. Hormozdiari, F., Penn, O., Borenstein, E. & Eichler, E.E. The discovery of integrated gene networks for autism and related disorders. *Genome research* **25**, 142-154 (2015).
32. Christensen, D. *et al.* Prevalence of cerebral palsy, co-occurring autism spectrum disorders, and motor functioning–A utism and D evelopmental D isabilities M onitoring N etwork, USA, 2008. *Developmental Medicine & Child Neurology* **56**, 59-65 (2014).
33. Reid, S.M., Meehan, E.M., Arnup, S.J. & Reddihough, D.S. Intellectual disability in cerebral palsy: a population-based retrospective study. *Developmental Medicine & Child Neurology* **60**, 687-694 (2018).
34. Garne, E. *et al.* Cerebral palsy and congenital malformations. *European Journal of Paediatric Neurology* **12**, 82-88 (2008).
35. Lumenta, C.B. & Skotarczak, U. Long-term follow-up in 233 patients with congenital hydrocephalus. *Child's Nervous System* **11**, 173-175 (1995).
36. O'Roak, B.J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246-250 (2012).
37. Cardozo, T. & Pagano, M. The SCF ubiquitin ligase: insights into a molecular machine. *Nature reviews Molecular cell biology* **5**, 739-751 (2004).
38. Jansen, S. *et al.* De novo variants in FBXO11 cause a syndromic form of intellectual disability with

behavioral problems and dysmorphisms. *European Journal of Human Genetics* **27**, 738-746 (2019).

39. Gregor, A. *et al.* De novo variants in the F-box protein FBXO11 in 20 individuals with a variable neurodevelopmental disorder. *The American Journal of Human Genetics* **103**, 305-316 (2018).
40. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics* **51**, 431 (2019).
41. Ripke, S., Walters, J.T., O'Donovan, M.C. & Consortium, S.W.G.o.t.P.G. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *MedRxiv* (2020).
42. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* **50**, 1112-1121 (2018).
43. Yu, D. *et al.* Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies. *American Journal of Psychiatry* **176**, 217-227 (2019).
44. Abou-Khalil, B. *et al.* Genome-wide mega-analysis identifies 16 loci and highlights diverse biological mechanisms in the common epilepsies. *Nature Communications* **9**, 5269 (2018).
45. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
46. Werling, D.M. *et al.* Whole-genome and RNA sequencing reveal variation and transcriptomic coordination in the developing human prefrontal cortex. *Cell reports* **31**, 107489 (2020).
47. Kielinen, M., Rantala, H., Timonen, E., Linna, S.-L. & Moilanen, I. Associated medical disorders and disabilities in children with autistic disorder: a population-based study. *Autism* **8**, 49-60 (2004).
48. Kilincaslan, A. & Mukaddes, N.M. Pervasive developmental disorders in individuals with cerebral palsy. *Developmental Medicine & Child Neurology* **51**, 289-294 (2009).
49. Kume, T. *et al.* The forkhead/winged helix gene Mf1 is disrupted in the pleiotropic mouse mutation congenital hydrocephalus. *Cell* **93**, 985-996 (1998).
50. Cao, M. & Wu, J.I. Camk2a-Cre-mediated conditional deletion of chromatin remodeler Brg1 causes perinatal hydrocephalus. *Neuroscience letters* **597**, 71-76 (2015).
51. Zaidi, S. & Brueckner, M. Genetics and genomics of congenital heart disease. *Circulation research* **120**, 923-940 (2017).
52. Gulisano, M. *et al.* Cardiovascular safety of aripiprazole and pimozide in young patients with Tourette syndrome. *Neurological Sciences* **32**, 1213-1217 (2011).
53. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics* **45**, 984 (2013).
54. Gratten, J., Wray, N.R., Keller, M.C. & Visscher, P.M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature neuroscience* **17**, 782 (2014).
55. Lee, P.H. *et al.* Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469-1482. e11 (2019).
56. Wang, Q., Yang, C., Gelernter, J. & Zhao, H. Pervasive pleiotropy between psychiatric disorders and immune disorders revealed by integrative analysis of multiple GWAS. *Human genetics* **134**, 1195-1209 (2015).
57. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics* **50**, 229 (2018).
58. Grotzinger, A.D. *et al.* Genomic structural equation modelling provides insights into the multivariate

genetic architecture of complex traits. *Nature human behaviour* **3**, 513 (2019).

59. Grotzinger, A.D. *et al.* Genetic Architecture of 11 Major Psychiatric Disorders at Biobehavioral, Functional Genomic, and Molecular Genetic Levels of Analysis. *medRxiv* (2020).
60. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nature genetics* **48**, 709 (2016).
61. Turner, T.N. *et al.* denovo-db: A compendium of human de novo variants. *Nucleic acids research* **45**, D804-D811 (2017).
62. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164-e164 (2010).
63. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137 (2015).
64. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nature communications* **8**, 1-11 (2017).
65. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. (AACR, 2012).
66. Consortium, G.P. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
67. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-223 (2013).
68. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25-29 (2000).