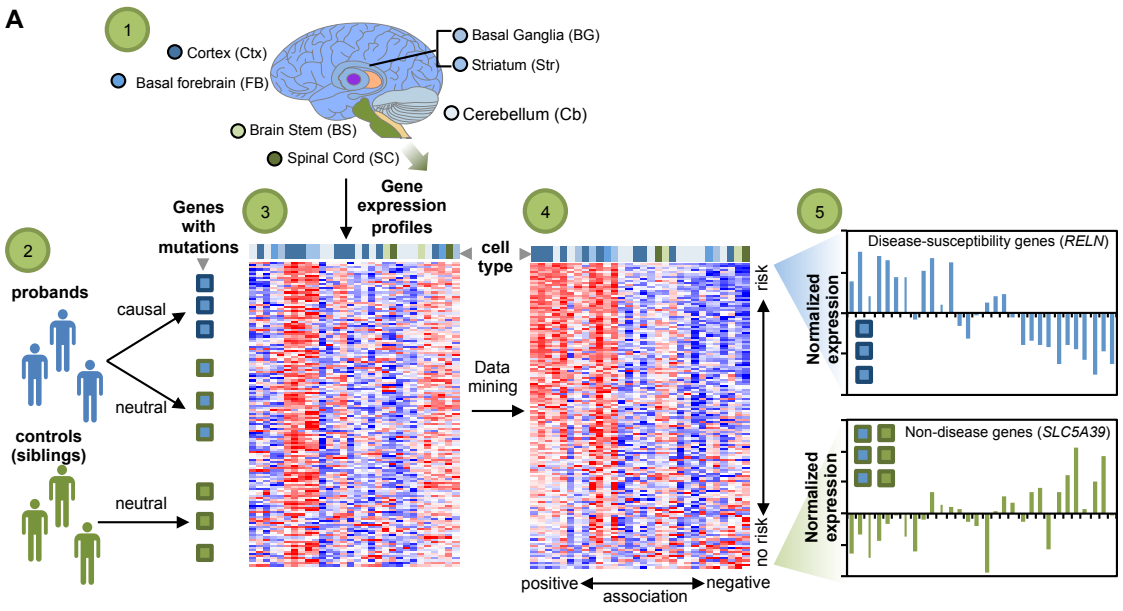
Supplementary Materials

A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes

Chaolin Zhang, Yufeng Shen

cz2294@columbia.edu (C.Z.)

ys2411@cumc.columbia.edu (Y.S.)



B

Data set	Type of data	Purpose
TRAP cell-type specific gene expression	Gene expression	D-score method development
Genes with de novo LGD mutations in pre-2013 papers on autism genetics (123 from cases and 35 from controls)	De novo mutation in autism and controls	D-score method development
Genes with mutations in post-2013 autism papers (468 genes from 3960 cases, and 173 genes from 1911 controls)	De novo mutation in autism and controls	1. Benchmark for prioritizing genes 2. Develop the ensemble method
De novo CNVs in autism cases as reported in (Sanders 2011)	De novo CNVs in autism cases	Evaluate haploinsufficiency of the affected genes
Differential expression of autistic vs. control post-mortem human brains	Gene expression	Evaluate haploinsufficiency of the affected genes
Krishnan et al. candidate genes (Krishnan et al 2016)	Candidate gene list based on expression and other functional annotations	Comparison with D-score
ExAC mutation intolerance metrics (pLI, lof-Z, and mis-Z) (Lek et al 2016)	Gene mutation intolerance score based on population data	1. Evaluate haploinsufficiency of the affected genes 2. Comparison with D-score 3. Input for the ensemble method
De novo mutations from DDD preprint (McRae et al 2016)	De novo mutation in related developmental disorders	Benchmark of candidate genes with singleton LGD mutations (ensemble method vs DAWN)
DAWN-inferred candidate genes (Liu et al 2015)	Candidate gene list based on mutations in autism cases and gene network	Comparison with the ensemble method on candidate genes with singleton LGD mutations

Figure S1: Overview of DAMAGES analysis to prioritize autism-susceptibility genes.

A. A schematic illustration of DAMAGES analysis. To distinguish disease-causing and non-disease-causing genes, this study used microarray expression profiling data of 24 mouse CNS cell types isolated from six brain regions as well as unselected RNA representing all cell types from each of these regions (1). DAMAGES analysis then starts with a list of genes with mutations

detected in patients with ASDs or unaffected siblings used as controls. This study focuses on candidate autism-susceptibility genes which individually have high penetrance. Genes with likely gene disrupting (LGD) mutations in controls are regarded as non-disease causing, while genes with LGD mutations in ASD probands represent a mixture of disease-causing and non-disease causing genes (2). Microarray expression data for mouse orthologs of genes with LGD mutations in ASD probands and siblings are extracted and shown as a heat map. Red and blue colors represent high and low expression, respectively (3). Data mining approaches such as principal component analysis (PCA) and regularized regression are used to find molecular signatures that can differentiate autism-causing genes and non-disease genes as well as the association of each cell type with autism (4). Expression pattern of two demonstrative genes predicted to confer high (*RELN*) and low (*SLC5A39*) risk of autism (5).

B. Summary of datasets used to derive D-scores and evaluate its performance in comparison with other methods.

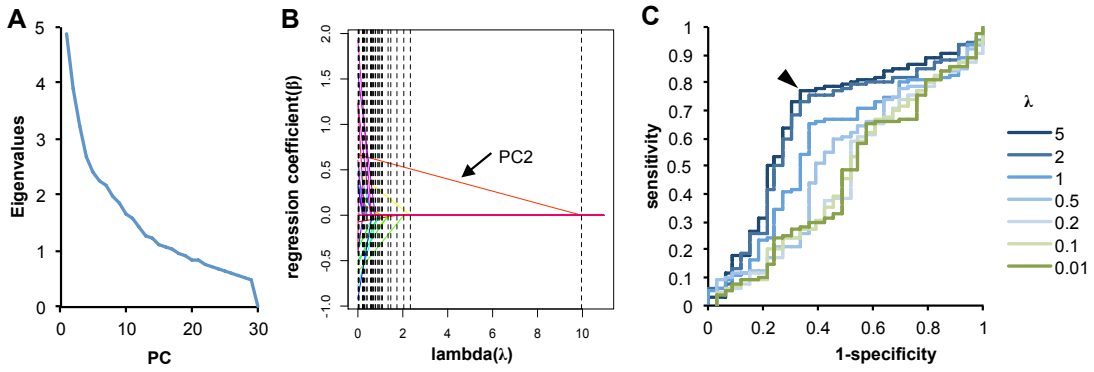


Figure S2: Derivation and characterization of D-score.

A. Scree plot of PCA analysis showing eigenvalues of each PC.

B. A regularized regression analysis is used to evaluate the relevance of each PC in predicting the source of mutations (i.e., probands versus siblings). Parameter λ controls the number of non-zero regression coefficients and therefore the complexity of the models. Each curve represents the trace of one coefficient when varying values of λ are used. The trace for the coefficient of PC2 is indicated. Dotted lines indicate shrinkage of coefficients to zero.

C. Leave-one-out cross-validation (LOOCV) is used to evaluate the specificity and sensitivity in predicting the source of LGD mutations. Prediction is based on genes ranked by the regression score. Each receiving operating characteristic (ROC) curve is derived from one specific value of the regularizing parameter λ. The best performance is achieved when only PC2 is used for prediction (λ=5). The arrowhead indicates the turning point of prediction performance when different thresholds of varying stringency are used for prediction.

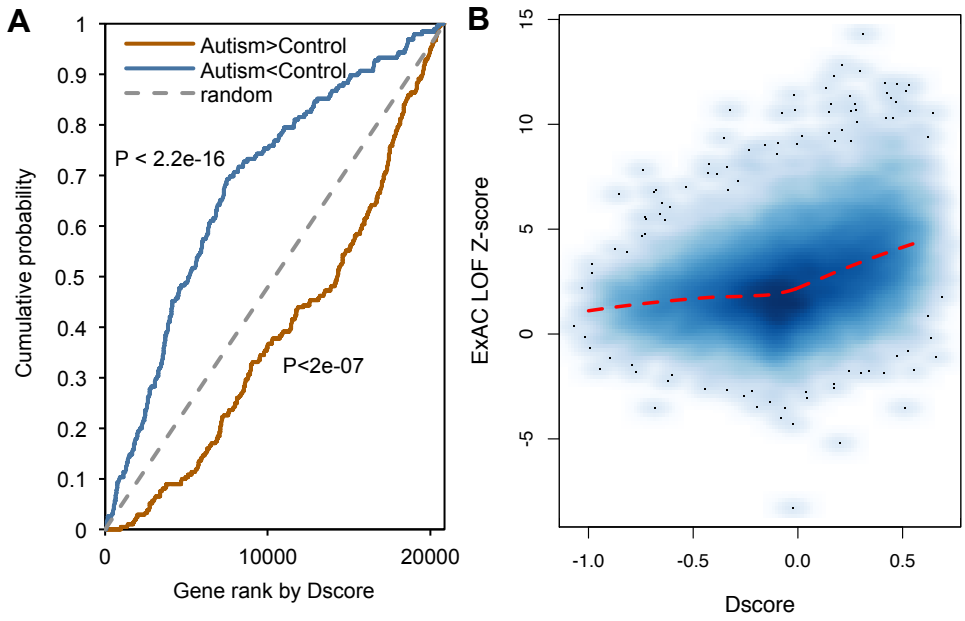


Figure S3: The cell type-specific expression signature reflects loss of function in autism.

A. Genes are ranked by D-score and the cumulative distribution of genes with up- (red) or down-regulation (blue) in autistic vs. control human brains are shown.

B. Correlation of D-score and ExAC LOF Z-score shown as a smooth scatter plot. The lowess regression of the two scores is shown with the red dotted line.