The American Journal of Human Genetics, Volume *106*

Supplemental Data

Co-localization between Sequence Constraint and Epigenomic Information Improves Interpretation of Whole-Genome Sequencing Data

Danqing Xu, Chen Wang, Krzysztof Kiryluk, Joseph D. Buxbaum, and Iuliana Ionita-Laza

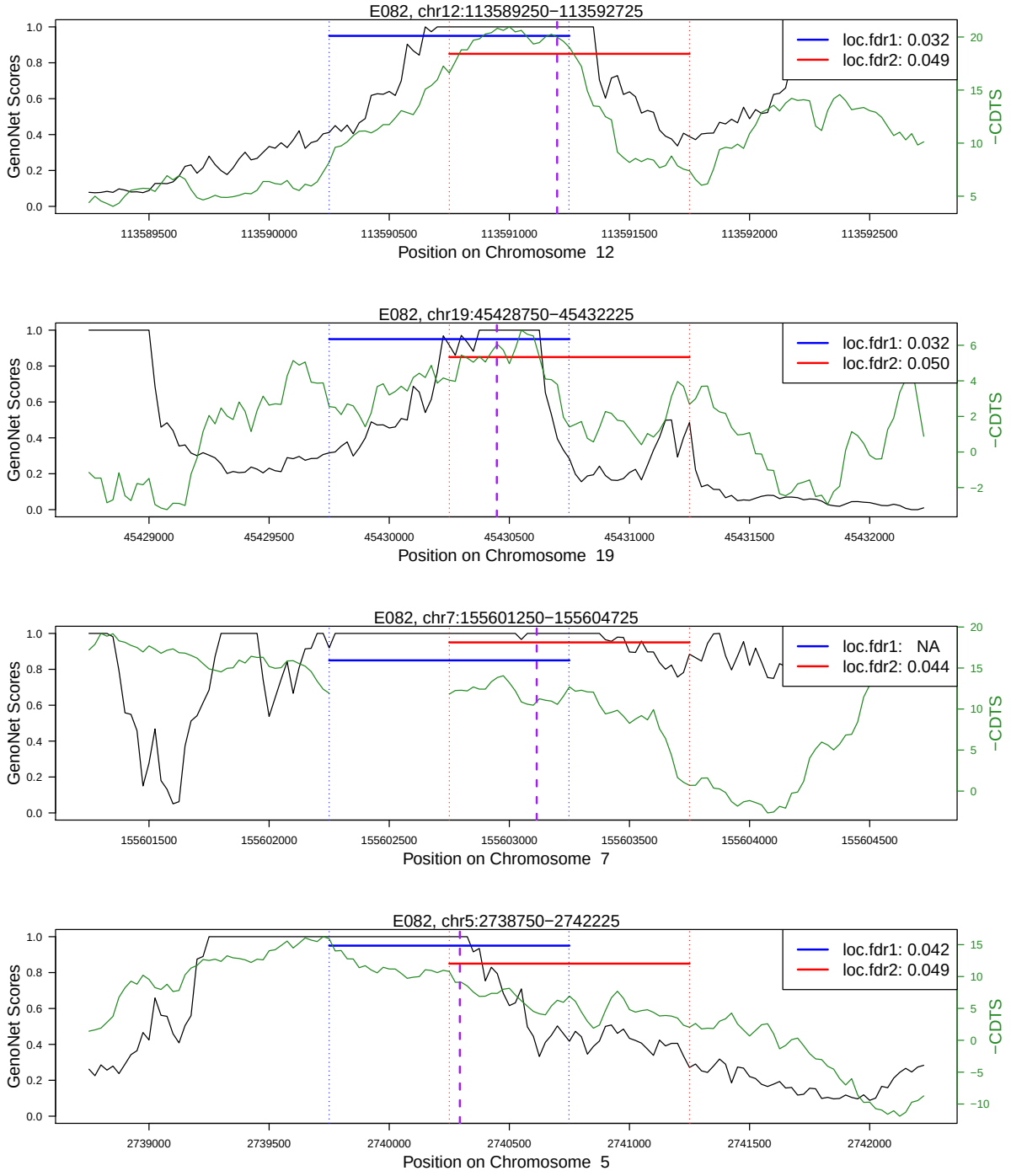

Figure S1: Examples of co-localized regions. GenoNet scores (black solid lines) and -CDTS scores (green solid lines), and two 1 Kb windows (blue and red horizontal solid lines for range, vertical dotted lines for start and end positions) are shown for one random position in each region (indicated by the vertical purple line), along with their co-localization local fdr's.

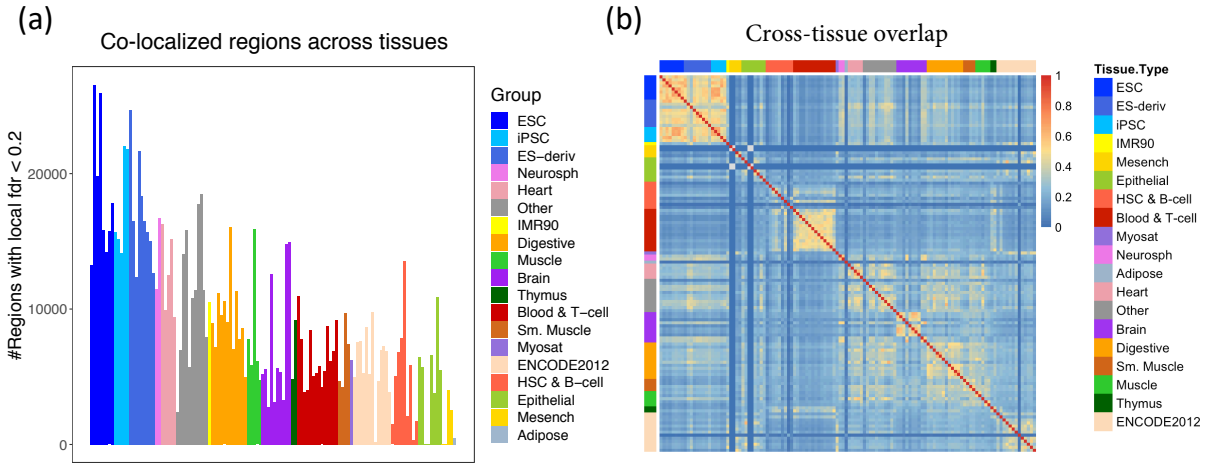

Figure S2: (a) Number of 1 Kb regions with co-localization local fdr ≤ 0.2, for each tissue/cell type in Roadmap. (b) Jaccard index of overlap between different tissues using regions with co-localization local fdr less than 0.2 in each of 127 tissues/cell types in Roadmap.

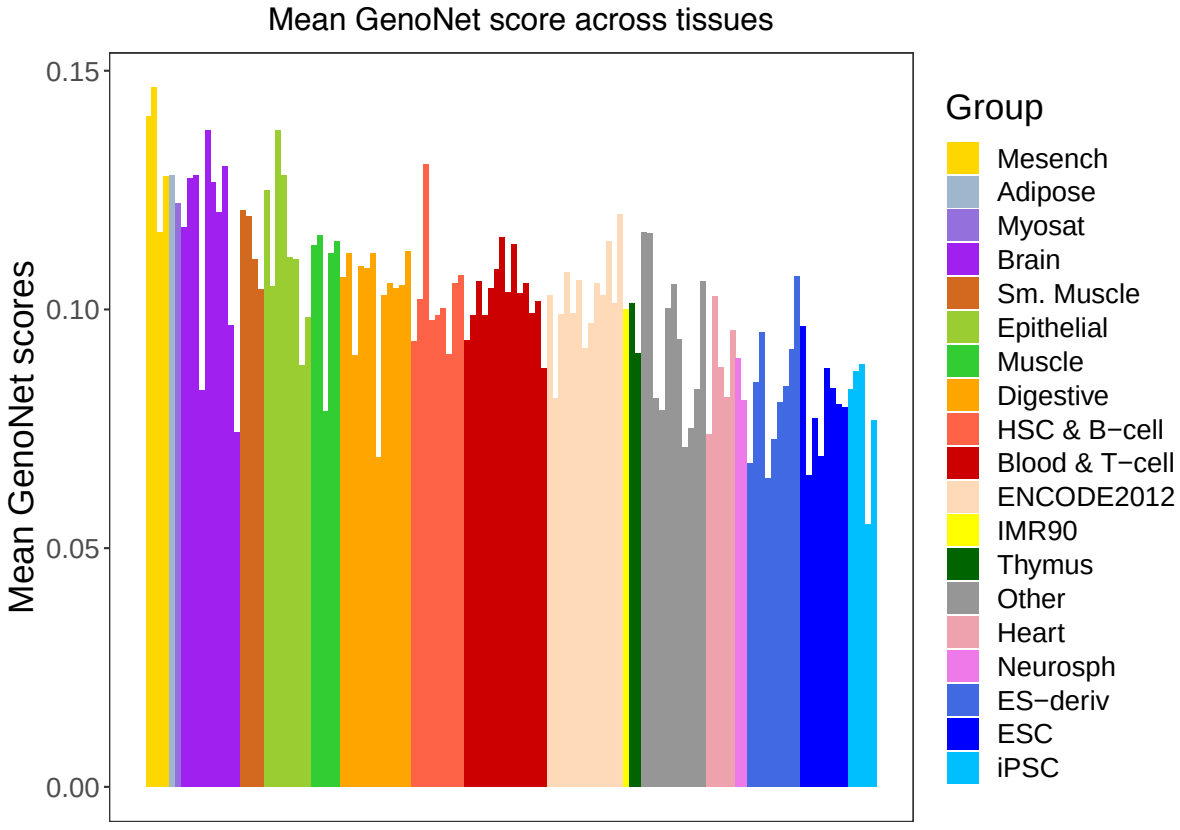

Figure S3: The mean GenoNet scores for each tissue/cell type in Roadmap, ordered by tissues group mean.

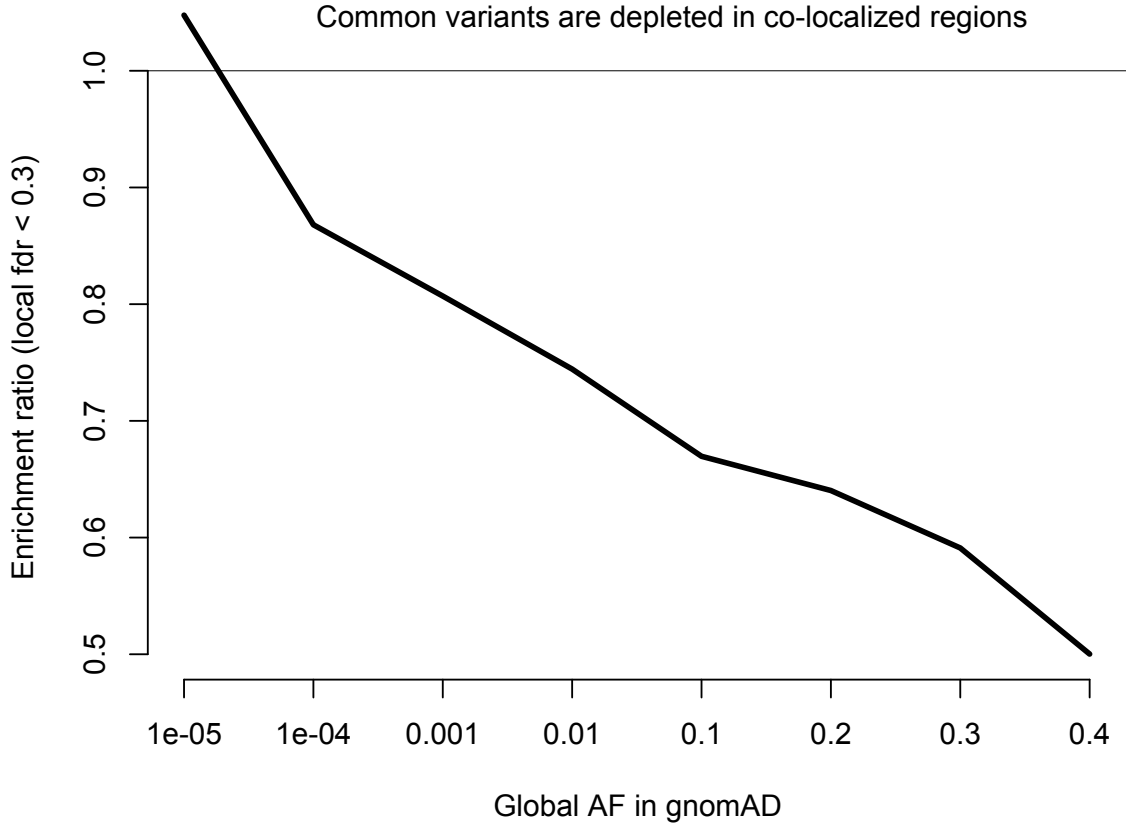

Figure S4: Depletion of common variants in co-localized regions. We considered several frequency bins on the x-axis, and the enrichment ratio is defined as $P(\text{variant is in specific frequency bin} \mid \text{local fdr} < 0.3)/P(\text{variant is in specific frequency bin})$.

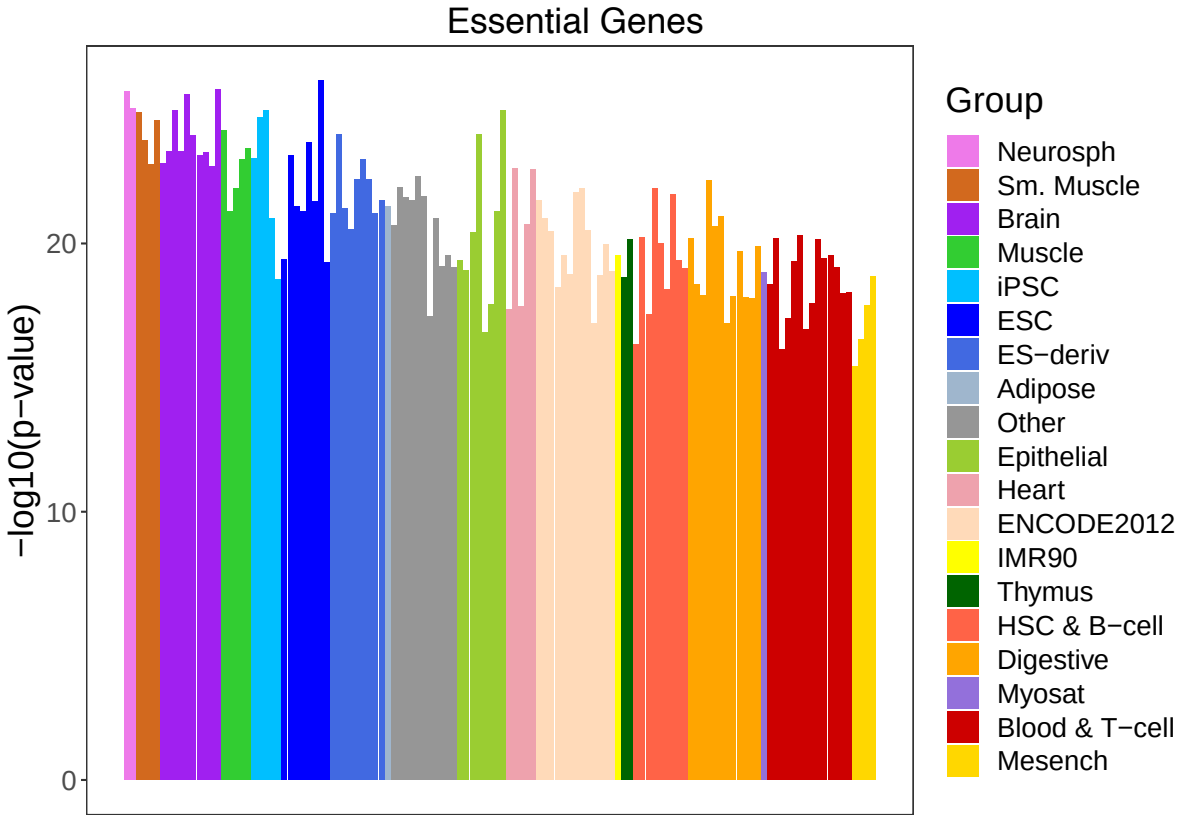

Figure S5: P values from Wilcoxon rank-sum tests comparing local fdr values for 3 kb regions upstream of transcription start sites for human orthologs of mouse essential genes with those for the rest of the genes.

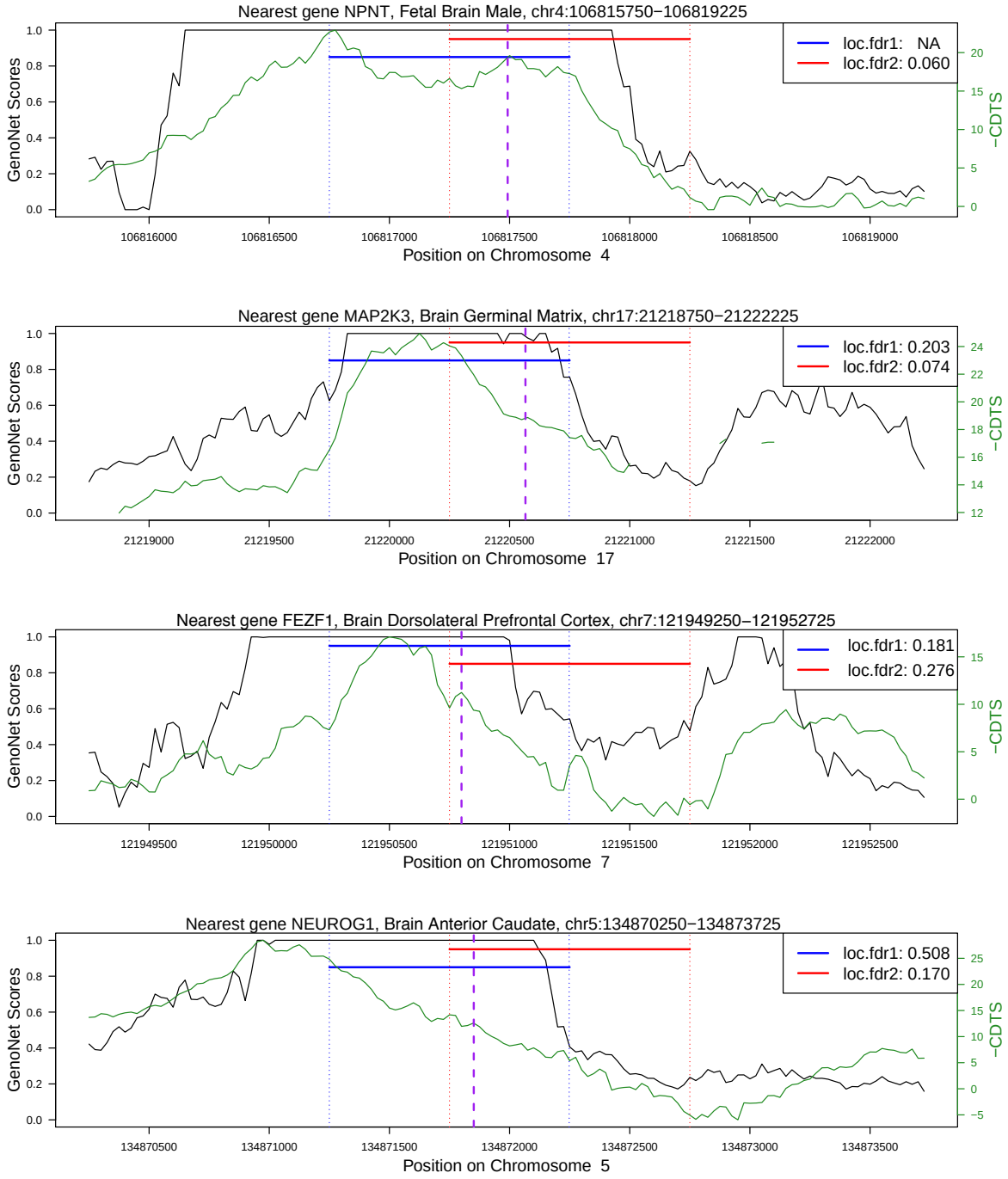

Figure S6: Co-localization results for *de novo* mutations with small local fdr's (the brain tissue with the smallest local fdr is shown for each mutation) and experimental evidence of effects on transcriptional activity. GenoNet scores (black solid lines) and -CDTS scores (green solid lines), and two 1 Kb windows (blue and red horizontal solid lines for range, vertical dotted lines for start and end positions) are shown for each mutation (indicated by the vertical purple line), along with their co-localization local fdr's. The NA values for some of the segments reflect constant values for one of the scores in the window.

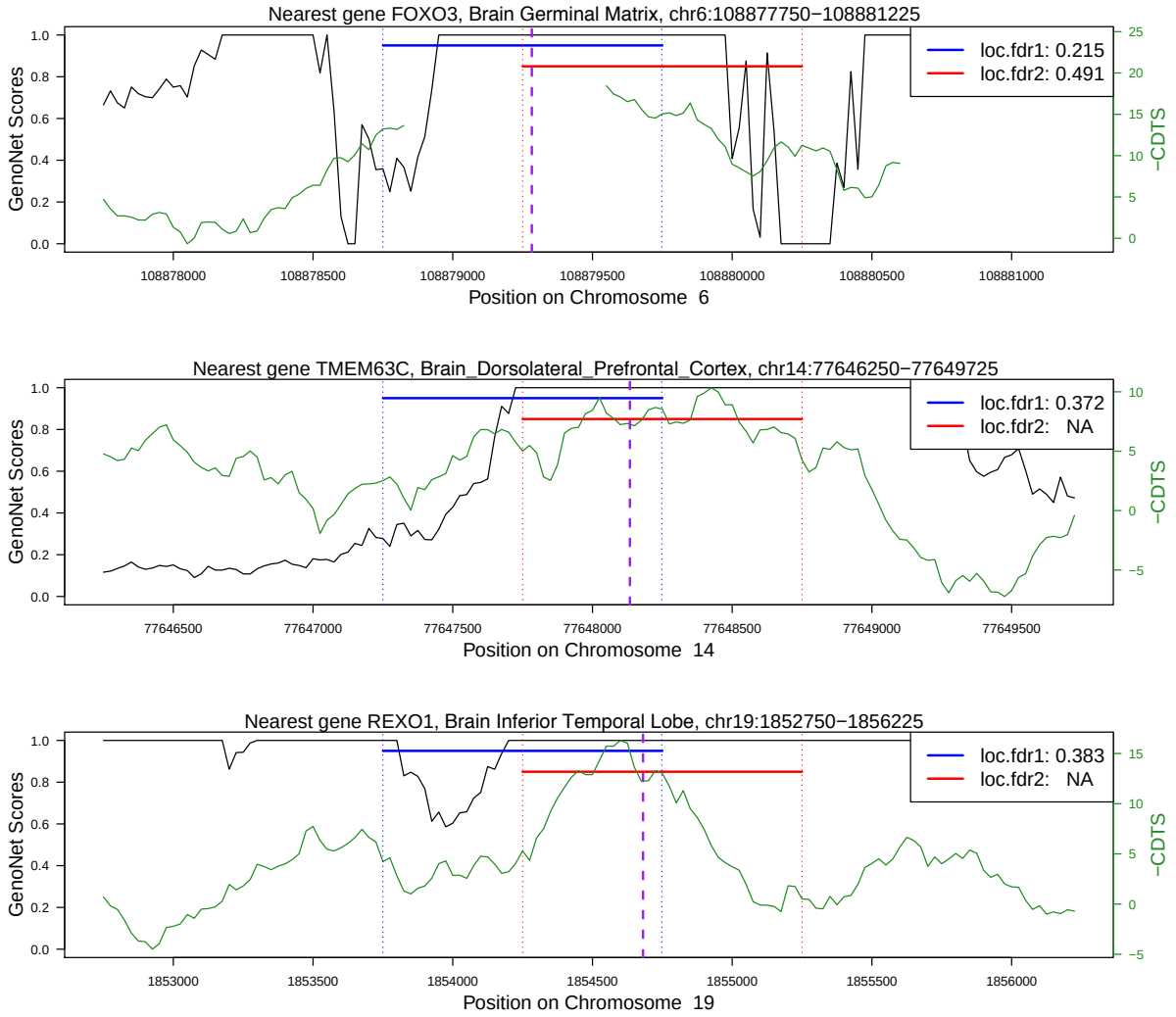

Figure S7: Co-localization results for *de novo* mutations with small local fdr's (the brain tissue with the smallest local fdr is shown for each mutation) and experimental evidence of effects on transcriptional activity. GenoNet scores (black solid lines) and -CDTS scores (green solid lines), and two 1 Kb windows (blue and red horizontal solid lines for range, vertical dotted lines for start and end positions) are shown for each mutation (indicated by the vertical purple line), along with their co-localization local fdr's. The NA values for some of the segments reflect constant values for one of the scores in the window.

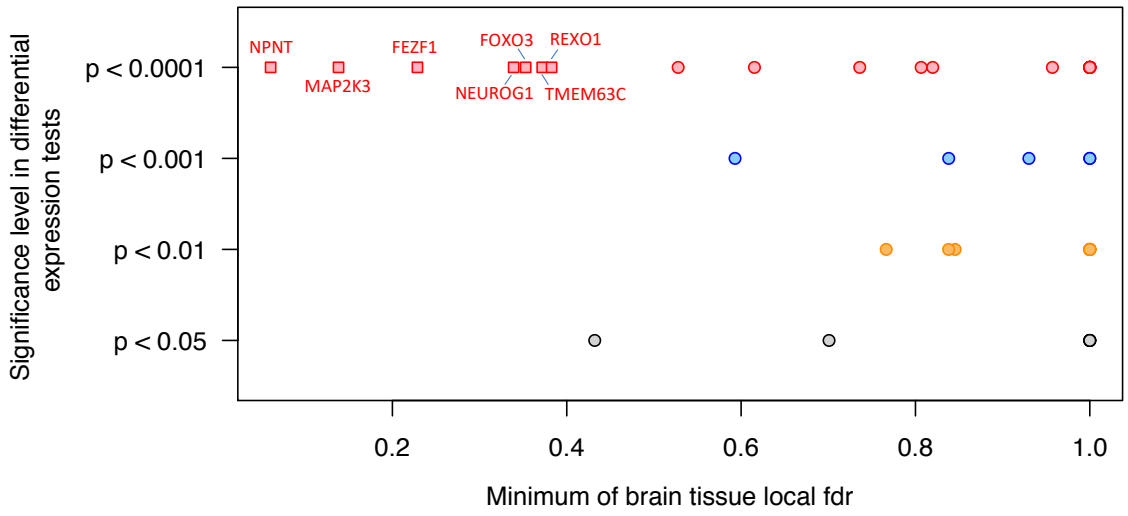

Figure S8: Significance level for testing differential expression for proband and sibling allele in a dual-luciferase assay, versus minimum co-localization local fdr of ten brain tissues for 51 *de novo* mutations in ASD probands; mutations with minimum co-localization local fdr below 0.4 are represented as squares, and labeled with their nearest gene; all other mutations are shown as circles. Significance levels for testing differential expression were computed on the basis of a t test and Fisher's combined probability test (two sided; gray for $p < 0.05$, orange for $p < 0.01$, blue for $p < 0.001$, red for $p < 0.0001$).

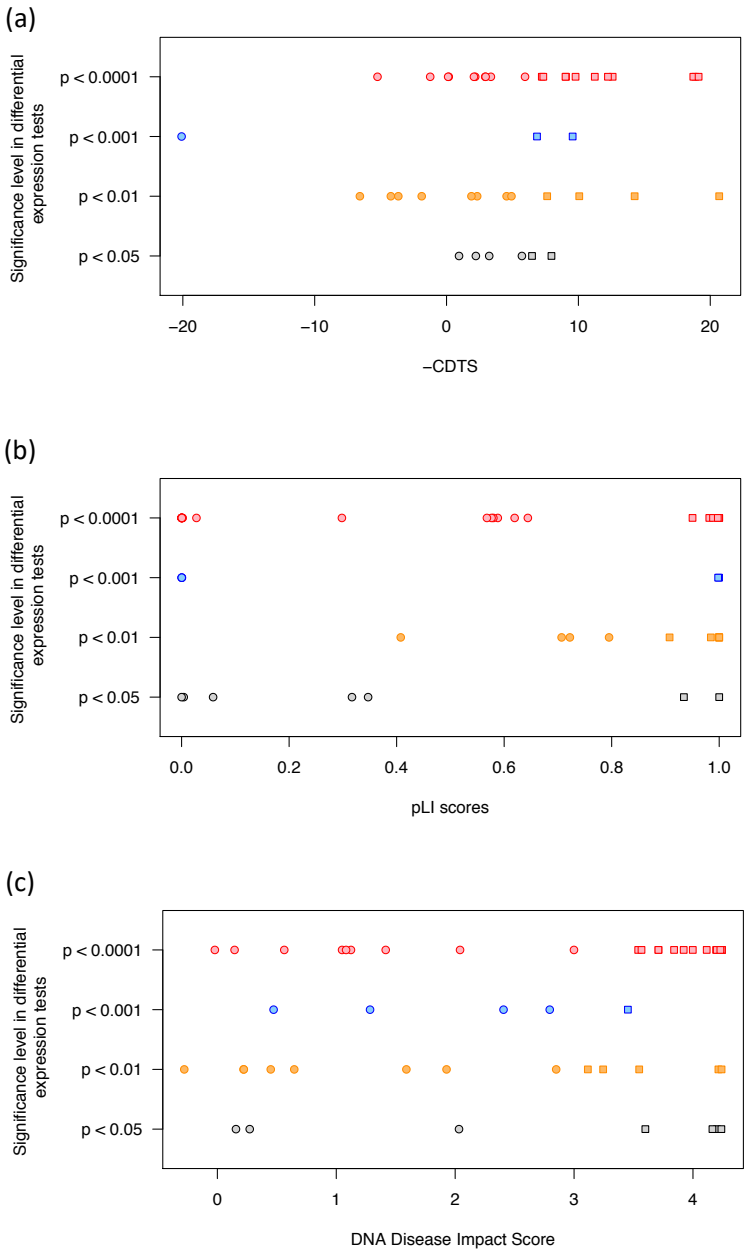

Figure S9: Significance level for testing differential expression for proband and sibling allele in a dual-luciferase assay, versus (a) -CDTS scores for 42 *de novo* mutations in ASD probands; mutations with -CDTS score above 6 are represented as squares; all other mutations are shown as circles. (b) pLI score for genes closest to 48 *de novo* mutations in ASD probands; mutations that correspond to genes with pLI above 0.8 are represented as squares; all other mutations are shown as circles. (c) DNA disease impact scores for 51 *de novo* mutations in ASD probands; mutations with scores above 3 are represented as squares; all other mutations are shown as circles. Significance levels for testing differential expression were computed on the basis of a t test and Fisher's combined probability test (two sided; gray for $p < 0.05$, orange for $p < 0.01$, blue for $p < 0.001$, red for $p < 0.0001$).

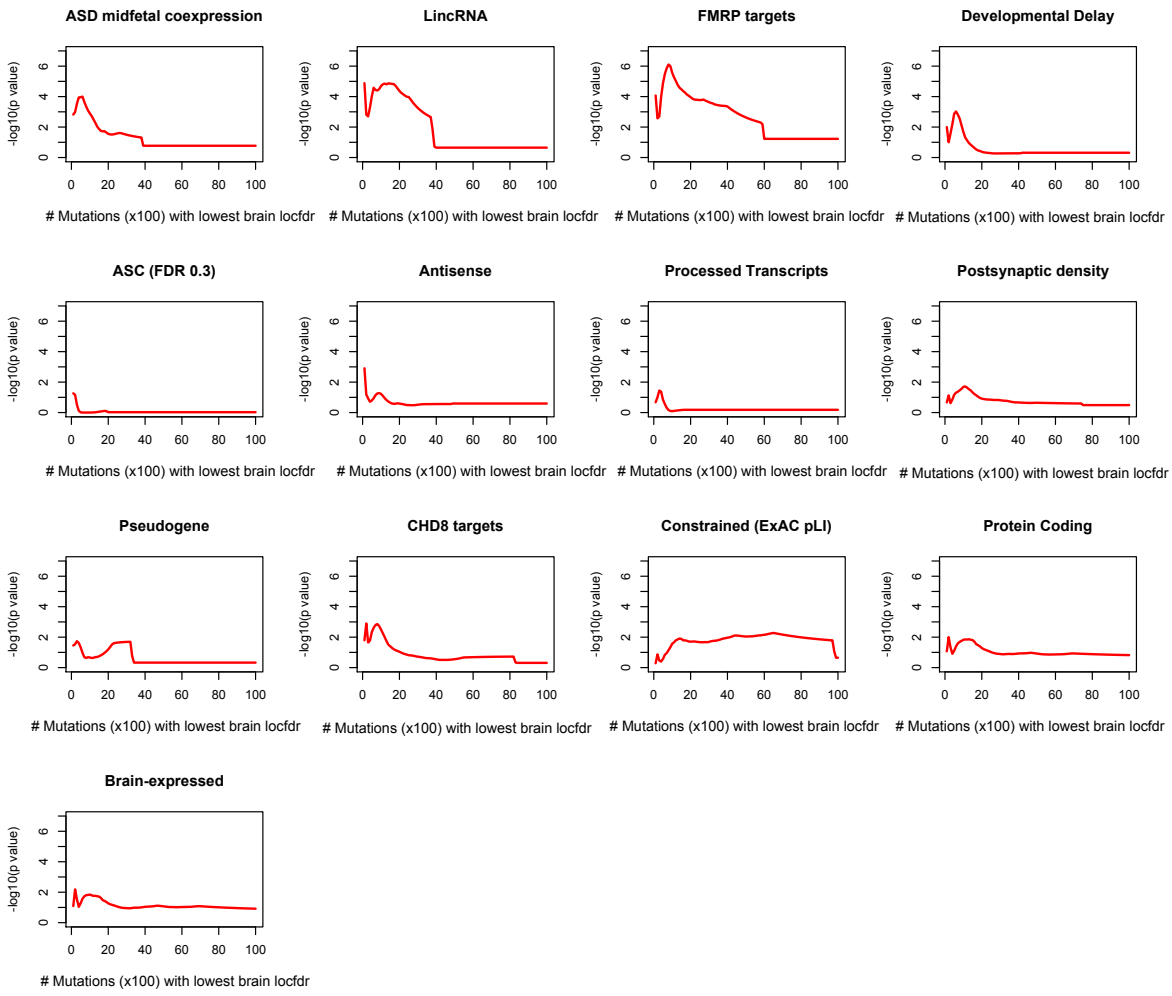

Figure S10: Gene Set Analyses. For each gene set, the -log10(p values) from a Wilcoxon rank sum test of difference between brain local fdr for the mutations residing within 1 Mb of TSS of genes in the set, in ASD probands vs. unaffected siblings are shown. The test is performed on the 100-10000 mutations with the lowest brain local fdr in ASD probands and unaffected siblings, respectively, as described in the Appendix.

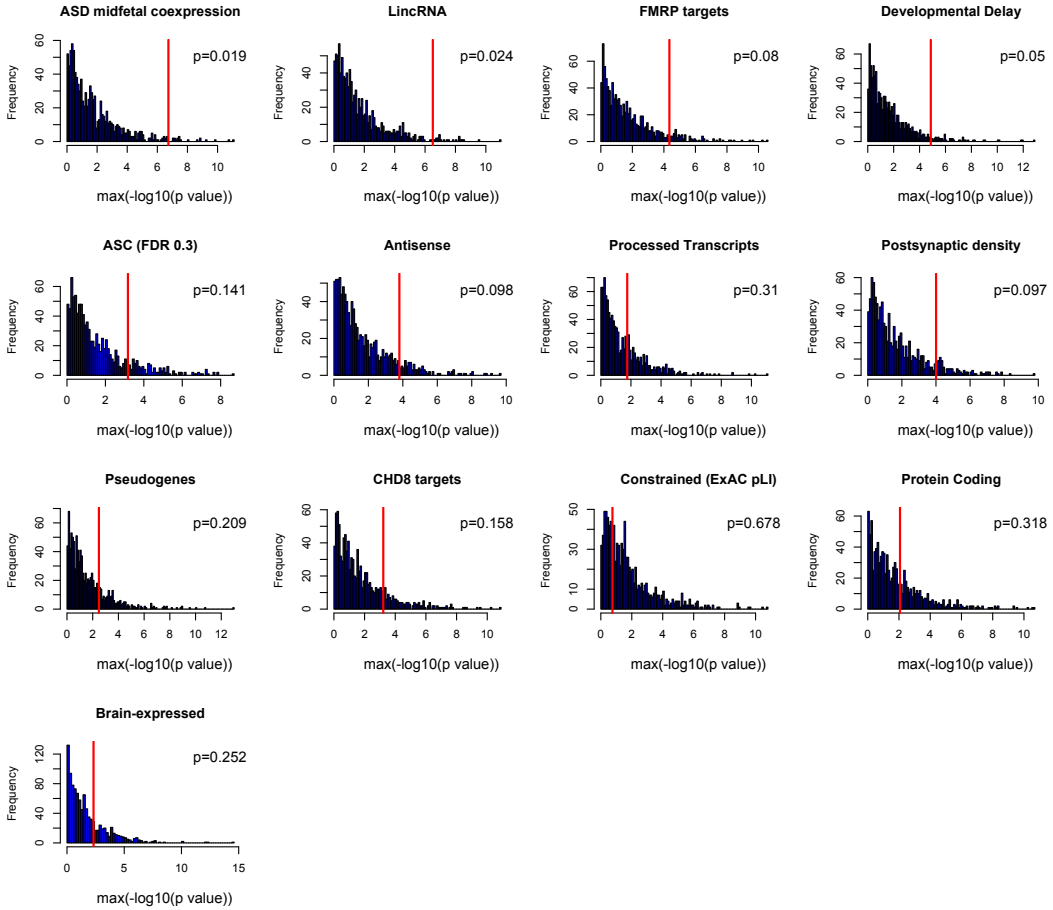

Figure S11: Significance of gene sets results as assessed by permutations (using mutations within ± 2Mb of TSS of genes in each set) . For each gene set, the observed maximum -log 10(p value) is shown using the red vertical line, along with the distribution of the maximum -log 10(p value) in 1000 permutations, where the ASD proband/unaffected sibling status of the *de novo* mutations are permuted as explained in the Appendix. The estimated p value for the observed maximum -log 10(p value) is reported in the topright corner.

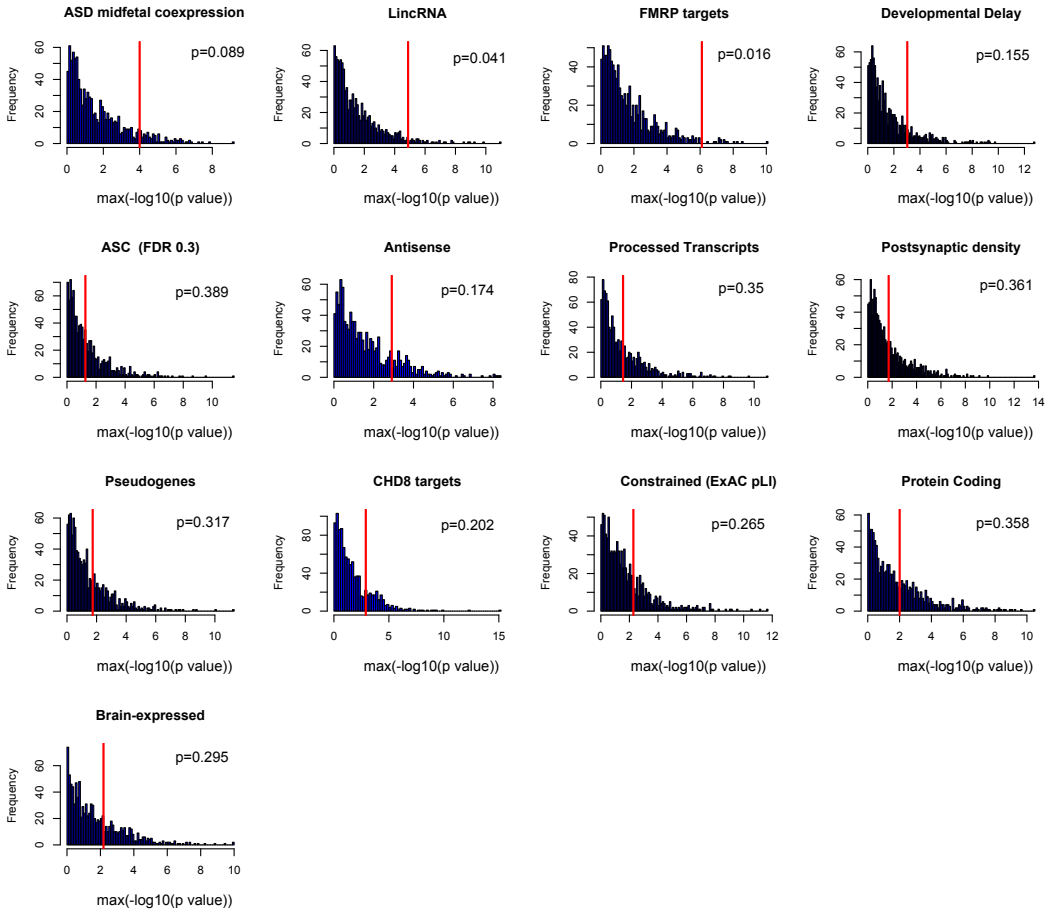

Figure S12: Significance of gene sets results as assessed by permutations (using mutations within ± 1Mb of TSS of genes in each set). For each gene set, the observed maximum -log 10(p value) is shown using the red vertical line, along with the distribution of the maximum -log 10(p value) in 1000 permutations, where the ASD proband/unaffected sibling status of the *de novo* mutations are permuted as explained in the Appendix. The estimated p value for the observed maximum -log 10(p value) is reported in the topright corner.

Table S1: P values for individual genesets (using 1Mb or 2Mb distance from TSS to assign mutations to genesets), along with the number of genes in each set. Cauchy combination p values [**?**], combining the two permutation based p values (for 1Mb and 2Mb), and FDR q-values are also reported. We compute q-values assuming two values for the proportion of null p values: a conservative $\pi_0 = 1$, and a more realistic $\pi_0 = 8/13$.

Geneset	#Genes	p (1Mb)	p (2Mb)	Cauchy p	q ($\pi_0 = 1$)	q ($\pi_0 = 8/13$)
FMRP targets	792	0.016	0.08	0.027	0.134	0.082
ASD midfetal coexpression	429	0.089	0.019	0.031	0.134	0.082
Developmental Delay	521	0.155	0.05	0.076	0.247	0.152
ASC (FDR 0.3)	179	0.389	0.141	0.217	0.352	0.217
Postsynaptic density	1,443	0.361	0.097	0.160	0.330	0.203
CHD8 targets	1,845	0.202	0.158	0.178	0.330	0.203
Constrained (ExAC pLI)	3,230	0.265	0.678	0.455	0.455	0.280
Brain-expressed	14,289	0.295	0.252	0.272	0.353	0.217
LincRNA (GencodeV19)	7,105	0.041	0.024	0.030	0.134	0.082
Antisense (GencodeV19)	5,272	0.174	0.098	0.126	0.327	0.201
Processed Transcripts (GencodeV19)	514	0.35	0.31	0.329	0.365	0.224
Pseudogenes (GencodeV19)	889	0.317	0.209	0.254	0.353	0.217
Protein Coding (GencodeV19)	20,242	0.358	0.318	0.337	0.365	0.224

Table S2: Several relevant summaries for the ToppFun functional enrichment analyses using *de novo* variants prioritized based on co-localization local fdr

# individuals	ASD	1,902	
	Control	1,902	
# *de novo* variants	ASD	128,627	
	Control	126,117	
# seed genes for *de novo* variants with locfdr < 0.4 in at least one brain tissue	ASD	809	
	Control	815	
# significantly interacting (PPI) genes for *de novo* variants with locfdr < 0.4 in at least one brain tissue	ASD	4	GRIN1, HNRNPL, SHANK3, SYNGAP1
	Control	3	HNRNPL, NF1, TMEM30B
# seed genes for *de novo* variants with locfdr < 0.5 in at least one brain tissue	ASD	1,098	
	Control	1,102	
# significantly interacting (PPI) genes for *de novo* variants with locfdr < 0.5 in at least one brain tissue	ASD	11	DLG4, DLGAP1, GRIN1, HNRNPL, IRF2BP2, IRF2BPL, KALRN, MAPK1, SHANK3, SUMO1, SYNGAP1
	Control	8	FMR1, HNRNPL, NF1, SMAD2, SMAD3, TIAM1, TMEM30B, UBE2I

Table S3: Mouse phenotypes affected by genes orthologous to the interacting genes derived from *de novo* mutations in ASD. *De novo* mutations are prioritized based on having co-localization local fdr < 0.4 in at least one brain tissue. Top 10 phenotypes with Bonferroni adjusted p value < 0.01 as reported by ToppFun analysis are reported. Results are only shown for ASD, since the analogous analysis for the *de novo* mutations derived from the unaffected siblings did not result in any significant results.

Mouse Phenotype	p value	p value Bonferroni	Interacting Genes
social withdrawal	3.649e-08	1.408e-05	GRIN1, SHANK3, SYNGAP1
abnormal discrimination learning	1.021e-07	3.942e-05	GRIN1, SHANK3, SYNGAP1
impaired synaptic plasticity	3.335e-07	1.287e-04	GRIN1, SHANK3, SYNGAP1
abnormal glutamate-mediated receptor currents	1.173e-06	4.530e-04	GRIN1, SHANK3, SYNGAP1
abnormal synaptic plasticity	1.331e-06	5.136e-04	GRIN1, SHANK3, SYNGAP1
increased startle reflex	5.746e-06	2.218e-03	GRIN1, SHANK3, SYNGAP1
abnormal miniature excitatory postsynaptic currents	6.499e-06	2.508e-03	GRIN1, SHANK3, SYNGAP1
reduced NMDA-mediated synaptic currents	8.450e-06	3.262e-03	GRIN1, SHANK3
abnormal medium spiny neuron morphology	8.450e-06	3.262e-03	GRIN1, SHANK3
abnormal social investigation	8.756e-06	3.380e-03	GRIN1, SHANK3, SYNGAP1

Table S4: Several relevant summaries for the ToppFun functional enrichment analyses using *de novo* variants prioritized based on DeepSEA disease impact score

# individuals	ASD	1,902	
	Control	1,902	
# *de novo* variants	ASD	128,627	
	Control	126,117	
# seed genes for *de novo* variants with DeepSEA disease impact score > 4	ASD	883	
	Control	850	
# significantly interacting (PPI) genes for *de novo* variants with DeepSEA disease impact score > 4	ASD	7	CTNNB1, ELAVL1, HDAC1, HNRNPL, RUNX1, SMAD4, YWHAZ
	Control	3	HDAC1, HNRNPL, LEF1

Table S5: Mouse phenotypes affected by genes orthologous to the interacting genes derived from *de novo* mutations in ASD and sibling control. *De novo* mutations are prioritized based on having a DeepSEA disease impact score > 4. Top 10 phenotypes with Bonferroni adjusted p value < 0.01 as reported by ToppFun analysis are reported.

Mouse Phenotype ASD proband	p value	p value Bonferroni	Interacting Genes
abnormal lymph organ size	1.30E-07	2.05E-04	HNRNPL,HDAC1,SMAD4,ELAVL1,CTNNB1,RUNX1,YWHAZ
increased double-negative T cell number	4.63E-07	7.29E-04	HNRNPL,HDAC1,ELAVL1,RUNX1
abnormal thymus physiology	5.68E-07	8.94E-04	HDAC1,SMAD4,ELAVL1,RUNX1
abnormal zigzag hair morphology	8.05E-07	1.27E-03	HDAC1,CTNNB1,RUNX1
abnormal immune system organ morphology	1.01E-06	1.58E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,CTNNB1,RUNX1,YWHAZ
small thymus	1.13E-06	1.77E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,RUNX1
decreased bone marrow cell number	1.46E-06	2.29E-03	HDAC1,ELAVL1,CTNNB1,RUNX1
abnormal thymus size	2.57E-06	4.04E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,RUNX1
increased intestinal adenocarcinoma incidence	2.71E-06	4.26E-03	SMAD4,ELAVL1,CTNNB1
abnormal double-negative T cell morphology	2.92E-06	4.59E-03	HNRNPL,HDAC1,ELAVL1,RUNX1
abnormal immune organ physiology	3.39E-06	5.34E-03	HDAC1,SMAD4,ELAVL1,RUNX1
epithelioid cysts	3.79E-06	5.97E-03	HDAC1,SMAD4
abnormal leukocyte cell number	4.10E-06	6.45E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,CTNNB1,RUNX1,YWHAZ
decreased double-positive T cell number	4.25E-06	6.68E-03	HNRNPL,HDAC1,ELAVL1,RUNX1
abnormal hemopoiesis	4.34E-06	6.83E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,CTNNB1,RUNX1
increased double-positive T cell number	4.41E-06	6.95E-03	HDAC1,ELAVL1,RUNX1
abnormal endocrine gland morphology	5.30E-06	8.34E-03	HNRNPL,HDAC1,SMAD4,ELAVL1,CTNNB1,RUNX1
abnormal epidermis stratum basale morphology	5.48E-06	8.63E-03	HDAC1,SMAD4,CTNNB1
abnormal bone marrow cell number	5.66E-06	8.91E-03	HDAC1,ELAVL1,CTNNB1,RUNX1
Mouse Phenotype Control Sibling	p value	p value Bonferroni	Interacting Genes
decreased CD8-positive, alpha-beta T cell number	1.85E-05	7.52E-03	HNRNPL,LEF1,HDAC1