Title
Prediction of the effects of rare genetic variation on neurodevelopmental disorders

Author
Chow, Julie Carol

Publication Date
2022

Peer reviewed|Thesis/dissertation

Prediction of the effects of rare genetic variation on neurodevelopmental disorders

By

JULIE CHOW

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Fereydoun Hormozdiari, Chair

C. Titus Brown

John McPherson

Committee in Charge

2022

Acknowledgements

I would like to thank Fereydoun for all his guidance and mentorship throughout the years. This work was possible because of his vision. I could always rely on him to offer new directions to pursue, provide feedback, and search for opportunities for me to grow as a scientist. He has been such a helpful and understanding major professor. Thanks to my past and current lab mates for your help and the fun we had at conferences and lab gatherings. I would like to thank my dissertation committee members, Titus Brown and John McPherson, who also served on my qualifying exam committee and devoted their time to reviewing this work and discussing my research over the years.

I would like to thank my family for their encouragement, support, and confidence in me. They have taken the best care of me my whole life, and as I enter the next stage of my life, I will finally be able to take care of them.

Thanks to my long-time friends of nearly twenty years Megan (Mo.), Megan (Mi.), and Kate, and friends of nearly a decade, Shierly and Fan, for being there for so long.

Thanks to the new friends I made during graduate school, especially our cohort, the 'Callipyges' – Alex, Benny, Dani, Ellen, José, Kelly, Kevin, Kyle, Noemie, Oran, Sara, Sichong, and Thiago. I will miss hosting game nights and our near-monthly birthday parties and hearing about everything you had to say.

Thank you, everyone! I am so grateful to be who I am because of you all.

Julie Chow

Chapter 1
Material from: 'Chow, J., Jensen M., Amini H. *et al*., Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders, Genome Medicine, published 2019, BioMed Central', 'Chow, J., Zhou, R., and Hormozdiari, F., MAGI-MS: multiple seed-centric module discovery, Bioinformatics Advances, published 2022, Oxford University Press'.
Chapter 2
Material from: 'Chow, J., and Hormozdiari, F., Prediction of neurodevelopmental disorders based on *de novo* coding variation, Journal of Autism and Developmental Disorders, published 2022, Springer Nature'.

Abstract

Neurodevelopmental disorders (NDDs), such as autism spectrum disorder (ASD), intellectual disability, and developmental disability are genetically and phenotypically heterogeneous disorders that display high comorbidity and relatively high heritability. Although non-coding and common variation contribute to a substantial proportion of all NDD cases, rare coding genetic variation has proved invaluable to the identification of NDD risk genes. NDD cases possess a significantly larger burden of *de novo* variation, a form of rare genetic variation that is not inherited from either parent, compared to unaffected controls. The enrichment of non-synonymous *de novo* coding variation in cases compared to controls enables the discovery of genetic modules, the early prediction of a subset of affected cases at low false positive rates, and the identification of critical cell-types relevant to specific modules.

Modules are networks of genes that participate in a certain biological function. The module discovery tools MAGI-S and its extension MAGI-MS are introduced in Chapter 1, which identify modules that can dissect specific phenotypes given 'seed' gene(s) that are members of biological pathways of interest. MAGI-S and MAGI-MS provide evidence of the dissection of the epilepsy phenotype from more general NDD phenotypes and the enrichment of non-synonymous *de novo* mutation in cases compared to controls among module genes.

In Chapter 2, a shallow neural network (SNN) with a false positive rate (FPR) minimizing loss function uses non-synonymous *de novo* mutation and features related to genic constraint and conservation to identify a small subset of NDD cases at very low FPR. Compared to traditional machine learning techniques and heuristics derived from genic constraint metrics and known NDD risk genes, the SNN achieves greater true positive rates (TPR) at near-zero FPR and ranks candidate NDD risk genes.

Given modules such as those generated by MAGI-S and MAGI-MS and single-cell expression data, MoToCC identifies groups of cells that selectively express the module genes. MoToCC is a linear programming approach that maximizes the gene co-expression amongst selected cells with consideration of cell-cell similarity and K-nearest neighbor connectivity. By allowing users to vary the number of cells to return as a solution, cell-types relevant to the module and shifting percent composition can be visualized at varied scales, as shown in Chapter 3 for three NDD modules.

The described computational tools seek to use the predictive power of *de novo* coding variation to further characterize the genetic etiology of neurodevelopmental disorders and lead to improvements in the well-being of affected patients.

Table of contents

Introduction

Neurodevelopmental disorders (NDDs) are complex disorders that affect the development of the central nervous system and are characterized by impairment in cognition, memory, language, and motor skills. Examples of NDDs include autism spectrum disorder (ASD), intellectual disability (ID), developmental disability (DD), attention-deficit hyperactivity disorder (ADHD), and motor disorders [1], and may more broadly include conditions such as epilepsy and schizophrenia [2]. NDDs arise from the disruption of typical molecular processes in the developing human brain by genetic and environmental factors [3,4]. In the United States, approximately 17% of children aged 3-17 were diagnosed with developmental disabilities from 2009-2017, indicating a growing prevalence of NDDs among US children [5]. Although many cases of ASD and DD can be detected before age five via behavioral and motor assessments, NDD diagnoses may be delayed by variation in symptom severity, the presence of comorbid conditions, ascertainment bias, and access to healthcare [6–9].

Current methods of diagnoses for NDDs include diagnostic checklists, structured interviews, imaging, and genetic testing. For ASD, the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) are among the most widely used tools of behavioral assessment [10–12]. Additionally, chromosomal microarrays are frequently used to identify chromosomal aberrations that may contribute to ID, DD, and ASD [13–15]. The combined use of electroencephalography (EEG) and magnetic resonance imaging (MRI) can aid in the diagnosis of epilepsy, prediction of seizure recurrence, detection of lesions, abnormal formations, and epileptogenic zones [16–19]. Prenatal testing has consisted of both invasive and non-invasive procedures; amniocentesis and chorionic villus sampling constitute the most common invasive procedures to detect chromosomal abnormalities [20], whereas more recent non-invasive

techniques such the collection of cell-free DNA from maternal blood have become more widespread and have shown promising results towards the screening of genetic disorders via the identification of *de novo* and point mutations in addition to chromosomal rearrangements [21–23].

Both environmental and genetic factors influence NDD susceptibility. It has been established that several environmental risk factors, such as parental age, maternal prenatal medication use, viral infection, and some environmental pollutants can modify gene expression via epigenetic mechanisms [24,25]. Of note is the strong correlation between parental age with the incidence of NDDs, in which advanced paternal age has been associated with increased accumulation of *de novo* mutations in the germline, maternal age with an elevated rate of chromosomal anomalies, and the prevalence of age-related DNA methylation in the germlines of both sexes [26–28]. The alteration of epigenetic regulation can significantly influence the manifestation of NDD phenotypes, particularly in combination with inherited genetic background [29,30].

The study of the genetic causes of NDDs is complicated by its extensive genetic and phenotypic heterogeneity, in addition to high degree of comorbidity. Multiple genetic mechanisms can result in similar NDD phenotypes, and, simultaneously, a single genetic mechanism can result in varied phenotypes [3,31–33]. High comorbidity among NDDs, such as the increased co-occurrence of ASD and ID, schizophrenia and ASD, ASD and ADHD, and epilepsy and ASD suggests shared genetic etiology among NDDs [34–38]. A substantial portion of NDD diagnoses have been attributed to genetic variation; from twin and family studies, estimates of heritability for ASD have ranged from 0.50 to 0.90 [39,40], greater than 0.70 for ADHD [41], and greater than 0.40 for ID [42–44]. Monozygotic twins have displayed a high degree of concordance for ASD and ADHD, whereas the concordance rate among dizygotic twins tends to be less than half of that observed for

monozygotic twins [45,46]. Similarly but to a lesser degree, concordance is lower in dizygotic twins than monozygotic twins with ID [47].

With consideration of the high degree of heritability observed among NDDs, multiple modes of inheritance have been proposed for NDDs. Monogenic inheritance has been observed for certain NDD phenotypes, such as Rett syndrome and Fragile X syndrome [48,49]. Additionally, large copy number variants (CNVs) and structural variants (SVs) affecting the expression of certain risk genes or the copy number of specific genomic regions, such as the 16.p11.2 deletion or duplication, have been identified as primary causes for various NDD phenotypes [50,51]. However, many NDD cases appear to result from the effects of a polygenic mode of inheritance, in which the collective effect of variation in multiple genes, particularly common variation, contributes to the phenotype [31,52]. The proposed 'omnigenic' model may also apply to NDD inheritance [53], in which a smaller set of 'core' genes that directly affect relevant biological pathways are regulated by numerous 'peripheral' genes, resulting in a large, interconnected network of genes that influence the phenotype [54].

Variants within both the protein coding and noncoding regions of the genome contribute to an individual's phenotype. Given that fewer than 2% of the human genome consists of protein coding sequence, vast noncoding regions likely contain functional and regulatory elements relevant to NDDs that are just beginning to become characterized [55,56]. Like common variation, which can be defined as variants with allele frequency greater than 5%, noncoding elements are hypothesized to explain a large proportion of genetic risk in NDDs [52,57,58]. Previously, the identification of causal, common variants has been slowed by small sample sizes, insufficient statistical power, and the relatively small effect size of common variants, but more recent genome-wide association studies have begun to discover common risk variants with increased sample sizes

[59–61]. Although noncoding and common variants together form a substantial component to NDD susceptibility, currently, many successful diagnoses have depended upon the identification of NDD risk genes using rare coding genetic variation.

Rare genetic variation, or variants with a minor allele frequency of less than 1%, have played a crucial role in the characterization of hundreds of NDD risk genes. *De novo* variation is a form of rare genetic variation that arises during gametogenesis or post-zygotically and is thereby absent in the genomes of either parent but present in the genome of their child. Mistakes during DNA replication, failure of DNA repair mechanisms, or DNA lesions resulting from exogenous or endogenous mutagens can cause the formation of *de novo* variants [62,63]. On average, an individual possesses 50-100 *de novo* single-nucleotide variants, the majority of which fall in noncoding regions of the genome [62,64]. The rate at which *de novo* mutations are generated is correlated with parental age, particularly with paternal age due to the continuous cell division of sperm cells during spermatogenesis [65,66]. Because *de novo* mutations are not inherited, *de novo* mutations have not been subjected to purifying selection and may potentially arise in genes critical to typical neurodevelopment [62,67].

Many NDD risk genes have been identified through the study of *de novo* mutation by applying whole exome or whole genome sequencing to simplex families, consisting of unaffected parents, their affected child, and unaffected sibling(s) [68,69]. Because coding *de novo* variants are very rare, a gene with a resulting loss-of-function or likely gene-disruptive (LGD) mutations (and to a lesser extent, missense mutations) observed in multiple, independent NDD cases implicates the gene as an NDD risk gene [70]. In fact, the presence of a coding *de novo* LGD mutation in two independent probands was previously found to provide significant statistical evidence for risk gene status [71]. The transmitted and *de novo* association (TADA) model has been used to find NDD risk

genes using *de novo* or inherited missense and LGD variants, as well as variants from case-control studies [72]. For example, 102 ASD-associated genes were recently identified via a modified TADA model [73], complementing previous discoveries [74–76]. NDD risk genes tend to also be 'constrained', in that they display a statistically lower number of *de novo* LGD mutations than by chance with consideration of factors such as mutation rate and gene length [77]. Genic constraint metrics such as probability of loss-of-function intolerance (pLI) and loss-of-function observed/expected upper bound fraction (LOEUF) were created to quantify the degree to which genes are sensitive to loss-of-function mutation [70,78]. pLI and LOEUF have been found to distinguish high and low risk NDD genes [79,80].

Missense variants in NDD risk genes show significant, but weaker enrichment in NDD cases compared to unaffected controls [81], and although the presence of missense variants in risk genes can be indicative of their deleteriousness [82,83], the significance of many missense variants is uncertain due to position-specific effects. Recent computational methods have emerged to address the challenge of characterizing variants of unknown significance while integrating information pertaining to evolutionary conservation, solvent accessibility, protein structure, genomic sequence context, functional annotations, and genic constraint [84–86]. Aside from in silico predictions, functional assays, such as saturation genome editing, have provided high-throughput assessment of variants of unknown significance even for variants that have not yet been observed [87,88]. The combined efforts of computational prediction methods and functional assays continually improve understanding of complex biological systems and the mechanisms of disease.

By identifying risk genes and characterizing variants of unknown significance, the early prediction of NDDs can be achieved. Accurate early prediction enables early treatment and therapeutic intervention, which has been shown to significantly affect the developmental trajectory

of affected individuals. For example, interventions in children less than three years old with behavioral symptoms of ASD and ID resulted in significant, sustained increases in child attentiveness and communication compared to children who received no intervention [89–92]. Most early treatments of ASD consist of behavioral interventions; medications have been used to lessen associated symptoms, such as sleep disorders and gastrointestinal issues, but there is not yet sufficient evidence of the efficacy of medication that targets the core symptoms of ASD [93]. The positive effects of the early prediction of NDDs are not limited to a patient's improved social outcomes and well-being, but also extend to reduced parenting stress and lifetime costs [94,95]. Given the importance of the early prediction of NDDs to patient and family outcomes, the identification of NDD risk genes via the examination of rare genetic variants holds considerable predictive power that continues to be explored.

The predictive ability of rare genetic variation in module discovery, early phenotypic prediction, and identification of critical cell-types

Genetic module discovery can be used to dissect complex phenotypes. Modules are groups of genes that contribute to a shared biological function and are highly connected in gene co-expression and or protein-protein interaction (PPI) networks. If a module is constructed while limiting the number of severe, protein-truncating *de novo* variation observed in a control population, then modules can identify networks that are enriched in *de novo* mutation in affected cases compared to controls relative to a general phenotype such as NDDs, as was accomplished via the method Merging Affected Genes into Integrated networks (MAGI) [96]. Briefly, MAGI first scores each candidate module gene according to its enrichment of *de novo* non-synonymous mutation and creates seed pathways consisting of high scoring genes, then clusters seed pathways into candidate modules. In Chapter 1, two extensions of MAGI are described, referred to as MAGI-

Seed (MAGI-S) and MAGI-Multiple-Seed (MAGI-MS) [97,98]. MAGI-S differs from MAGI in that module construction is seeded from a single 'seed' gene. Using MAGI-S, genes are scored according to their degree of co-expression with the seed gene, resulting in modules that are highly co-expressed relative to the seed gene [97]. MAGI-MS further extends upon MAGI-S by permitting the selection of multiple seed genes and applying normalization to gene scores to reduce selection of generally highly expressed genes in constructed modules [98].

The enrichment of *de novo* non-synonymous coding mutation that is observed among NDD cases compared to controls permits the identification of a subset of NDD cases at very low false positive rates (FPR). Although most NDD cases are attributed to other types of genetic variation, by focusing specifically on *de novo* coding variation, near-zero FPR can be achieved, which is an important aspect to early prediction methods due to the possible severe, negative effects of erroneous prediction. Previously, the combinatorial framework Oracle for Disorder predictioN (Odin) sought to accurately predict a small subset of NDD cases using *de novo* LGD mutation and gene co-expression via a weighted unicolor clustering with dimensionality reduction [99]. Chapter 2 introduces a shallow neural network (SNN) with a custom FPR-minimizing loss function that incorporates *de novo* LGD and missense variants and features related to genic constraint and evolutionary conservation to identify NDD cases at FPR < 0.01[100] The trained SNN additionally performs gene prioritization to reveal novel NDD risk genes.

Single-cell transcriptomic analyses have uncovered novel cell-types and yielded insights into specific molecular mechanisms relevant to disease at a high resolution. For given a module or a set of genes of interest, it is possible to identify individual cells or tissues that selectively express the supplied genes, as shown with previous methods such as CSEA and TissueEnrich [101,102]. However, previous methods are limited in their flexibility to select variably sized subsets of cells

that correspond to distinct communities of cell-types in arbitrary single-cell datasets. In Chapter 3, the tool MoToCC uses a linear programming approach to select a critical subset of cells that selectively express a given module while deriving measures of cell-cell similarity and neighborhood connectivity from single-cell gene expression values. Repeated iterations of MoToCC with a varied upper bound (k) of number of cells to return as a solution highlight relevant cell-types and shifts in percent composition as a progressively greater number of cells are returned and visualized.

References

1. Psychiatry.org - DSM.
2. Savatt, J.M., and Myers, S.M. (2021). Genetic Testing in Neurodevelopmental Disorders. Front. Pediatr. *9*,.
3. Cardoso, A.R., Lopes-Marques, M., Silva, R.M., Serrano, C., Amorim, A., Prata, M.J., and Azevedo, L. (2019). Essential genetic findings in neurodevelopmental disorders. Hum. Genomics *13*,.
4. Stiles, J., and Jernigan, T.L. (2010). The Basics of Brain Development. Neuropsychol. Rev. *20*, 327–348.
5. Zablotsky, B., Black, L.I., Maenner, M.J., Schieve, L.A., Danielson, M.L., Bitsko, R.H., Blumberg, S.J., Kogan, M.D., and Boyle, C.A. (2019). Prevalence and Trends of Developmental Disabilities among Children in the US: 2009-2017. Pediatrics *144*, e20190811.
6. Micai, M., Fulceri, F., Caruso, A., Guzzetta, A., Gila, L., and Scattoni, M.L. (2020). Early behavioral markers for neurodevelopmental disorders in the first 3 years of life: An overview of systematic reviews. Neurosci. Biobehav. Rev. *116*, 183–201.
7. Leader, G., Hogan, A., Chen, J.L., Maher, L., Naughton, K., O'Rourke, N., Casburn, M., and Mannion, A. (2022). Age of Autism Spectrum Disorder Diagnosis and Comorbidity in Children and Adolescents with Autism Spectrum Disorder. Dev. Neurorehabilitation *25*, 29–37.
8. Kentrou, V., de Veld, D.M., Mataw, K.J., and Begeer, S. (2019). Delayed autism spectrum disorder recognition in children and adolescents previously diagnosed with attention-deficit/hyperactivity disorder. Autism *23*, 1065–1072.
9. Lockwood Estrin, G., Milner, V., Spain, D., Happé, F., and Colvert, E. (2021). Barriers to Autism Spectrum Disorder Diagnosis for Young Women and Girls: a Systematic Review. Rev. J. Autism Dev. Disord. *8*, 454–470.
10. Akshoomoff, N., Corsello, C., and Schmidt, H. (2006). The Role of the Autism Diagnostic Observation Schedule in the Assessment of Autism Spectrum Disorders in School and Community Settings. Calif. Sch. Psychol. CASP Calif. Assoc. Sch. Psychol. *11*, 7–19.
11. Lord, C., Rutter, M., and Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J. Autism Dev. Disord. *24*, 659–685.
12. Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L., DiLavore, P.C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. J. Autism Dev. Disord. *30*, 205–223.

13. Lee, J.S., Hwang, H., Kim, S.Y., Kim, K.J., Choi, J.S., Woo, M.J., Choi, Y.M., Jun, J.K., Lim, B.C., and Chae, J.-H. (2018). Chromosomal Microarray With Clinical Diagnostic Utility in Children With Developmental Delay or Intellectual Disability. Ann. Lab. Med. *38*, 473–480.
14. Battaglia, A., Doccini, V., Bernardini, L., Novelli, A., Loddo, S., Capalbo, A., Filippi, T., and Carey, J.C. (2013). Confirmation of chromosomal microarray as a first-tier clinical diagnostic test for individuals with developmental delay, intellectual disability, autism spectrum disorders and dysmorphic features. Eur. J. Paediatr. Neurol. *17*, 589–599.
15. Shoukier, M., Klein, N., Auber, B., Wickert, J., Schröder, J., Zoll, B., Burfeind, P., Bartels, I., Alsat, E., Lingen, M., et al. (2013). Array CGH in patients with developmental delay or intellectual disability: are there phenotypic clues to pathogenic copy number variants? Clin. Genet. *83*, 53–65.
16. Drenthen, G.S., Jansen, J.F.A., Gommer, E., Gupta, L., Hofman, P.A.M., Kranen-Mastenbroek, V.H. van, Hilkman, D.M., Vlooswijk, M.C.G., Rouhl, R.P.W., and Backes, W.H. (2021). Predictive value of functional MRI and EEG in epilepsy diagnosis after a first seizure. Epilepsy Behav. *115*,.
17. Salmenpera, T.M., and Duncan, J.S. (2005). Imaging in epilepsy. J. Neurol. Neurosurg. Psychiatry *76*, iii2–iii10.
18. Li, A., Chennuri, B., Subramanian, S., Yaffe, R., Gliske, S., Stacey, W., Norton, R., Jordan, A., Zaghloul, K.A., Inati, S.K., et al. (2018). Using network analysis to localize the epileptogenic zone from invasive EEG recordings in intractable focal epilepsy. Netw. Neurosci. *2*, 218–240.
19. Noachtar, S., and Rémi, J. (2009). The role of EEG in epilepsy: a critical review. Epilepsy Behav. EB *15*, 22–33.
20. Akolekar, R., Beta, J., Picciarelli, G., Ogilvie, C., and D'Antonio, F. (2015). Procedure-related risk of miscarriage following amniocentesis and chorionic villus sampling: a systematic review and meta-analysis. Ultrasound Obstet. Gynecol. *45*, 16–26.
21. Bowman-Smart, H., Savulescu, J., Mand, C., Gyngell, C., Pertile, M.D., Lewis, S., and Delatycki, M.B. (2019). 'Is it better not to know certain things?': views of women who have undergone non-invasive prenatal testing on its possible future applications. J. Med. Ethics *45*, 231–238.
22. Chan, K.C.A., Jiang, P., Sun, K., Cheng, Y.K.Y., Tong, Y.K., Cheng, S.H., Wong, A.I.C., Hudecova, I., Leung, T.Y., Chiu, R.W.K., et al. (2016). Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc. Natl. Acad. Sci. *113*, E8159–E8168.
23. Rabinowitz, T., and Shomron, N. (2020). Genome-wide noninvasive prenatal diagnosis of monogenic disorders: Current and future trends. Comput. Struct. Biotechnol. J. *18*, 2463–2470.
24. Karimi, P., Kamali, E., Mousavi, S.M., and Karahmadi, M. (2017). Environmental factors influencing the risk of autism. J. Res. Med. Sci. Off. J. Isfahan Univ. Med. Sci. *22*, 27.
25. Modabbernia, A., Velthorst, E., and Reichenberg, A. (2017). Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. Mol. Autism *8*, 13.
26. Reichenberg, A., Gross, R., Weiser, M., Bresnahan, M., Silverman, J., Harlap, S., Rabinowitz, J., Shulman, C., Malaspina, D., Lubin, G., et al. (2006). Advancing Paternal Age and Autism. Arch. Gen. Psychiatry *63*, 1026–1032.
27. Croen, L.A., Najjar, D.V., Fireman, B., and Grether, J.K. (2007). Maternal and Paternal Age and Risk of Autism Spectrum Disorders. Arch. Pediatr. Adolesc. Med. *161*, 334–340.
28. Adkins, R.M., Thomas, F., Tylavsky, F.A., and Krushkal, J. (2011). Parental ages and levels of DNA methylation in the newborn are correlated. BMC Med. Genet. *12*, 47.
29. Eshraghi, A.A., Liu, G., Kay, S.-I.S., Eshraghi, R.S., Mittal, J., Moshiree, B., and Mittal, R. (2018). Epigenetics and Autism Spectrum Disorder: Is There a Correlation? Front. Cell. Neurosci. *12*, 78.
30. Iwase, S., Bérubé, N.G., Zhou, Z., Kasri, N.N., Battaglioli, E., Scandaglia, M., and Barco, A. (2017). Epigenetic Etiology of Intellectual Disability. J. Neurosci. *37*, 10773–10782.
31. Parenti, I., Rabaneda, L.G., Schoen, H., and Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. Trends Neurosci. *43*, 608–621.
32. Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. Brain Res. *1380*, 42–77.

33. Girirajan, S., Rosenfeld, J.A., Coe, B.P., Parikh, S., Friedman, N., Goldstein, A., Filipink, R.A., McConnell, J.S., Angle, B., Meschino, W.S., et al. (2012). Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N. Engl. J. Med. *367*, 1321–1331.
34. Mpaka, D.M., Okitundu, D.L.E.-A., Ndjukendi, A.O., N'situ, A.M., Kinsala, S.Y., Mukau, J.E., Ngoma, V.M., Kashala-Abotnes, E., Ma-Miezi-Mampunza, S., Vogels, A., et al. (2016). Prevalence and comorbidities of autism among children referred to the outpatient clinics for neurodevelopmental disorders. Pan Afr. Med. J. *25*, 82.
35. Goldin, R.L., Matson, J.L., and Cervantes, P.E. (2014). The effect of intellectual disability on the presence of comorbid symptoms in children and adolescents with autism spectrum disorder. Res. Autism Spectr. Disord. *8*, 1552–1556.
36. Louzolo, A., Gustavsson, P., Tigerström, L., Ingvar, M., Olsson, A., and Petrovic, P. (2017). Delusion-proneness displays comorbidity with traits of autistic-spectrum disorders and ADHD. PLOS ONE *12*, e0177820.
37. Jeste, S.S., and Tuchman, R. (2015). Autism Spectrum Disorder and Epilepsy: Two Sides of the Same Coin? J. Child Neurol. *30*, 1963–1971.
38. Doshi-Velez, F., Ge, Y., and Kohane, I. (2014). Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics *133*, e54–e63.
39. Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The Heritability of Autism Spectrum Disorder. JAMA *318*, 1182–1184.
40. Tick, B., Bolton, P., Happé, F., Rutter, M., and Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta□analysis of twin studies. J. Child Psychol. Psychiatry *57*, 585–595.
41. Faraone, S.V., and Larsson, H. (2019). Genetics of attention deficit hyperactivity disorder. Mol. Psychiatry *24*, 562–575.
42. Lichtenstein, P., Tideman, M., Sullivan, P.F., Serlachius, E., Larsson, H., Kuja-Halkola, R., and Butwicka, A. Familial risk and heritability of intellectual disability: a population-based cohort study in Sweden. J. Child Psychol. Psychiatry *n/a*,.
43. Haworth, C.M.A., Wright, M.J., Luciano, M., Martin, N.G., de Geus, E.J.C., van Beijsterveldt, C.E.M., Bartels, M., Posthuma, D., Boomsma, D.I., Davis, O.S.P., et al. (2010). The heritability of general cognitive ability increases linearly from childhood to young adulthood. Mol. Psychiatry *15*, 1112–1120.
44. Panizzon, M.S., Vuoksimaa, E., Spoon, K.M., Jacobson, K.C., Lyons, M.J., Franz, C.E., Xian, H., Vasilopoulos, T., and Kremen, W.S. (2014). Genetic and Environmental Influences of General Cognitive Ability: Is g a valid latent construct? Intelligence *43*, 65–76.
45. Castelbaum, L., Sylvester, C.M., Zhang, Y., Yu, Q., and Constantino, J.N. (2020). On the Nature of Monozygotic Twin Concordance and Discordance for Autistic Trait Severity: A Quantitative Analysis. Behav. Genet. *50*, 263–272.
46. Langner, I., Garbe, E., Banaschewski, T., and Mikolajczyk, R.T. (2013). Twin and Sibling Studies Using Health Insurance Data: The Example of Attention Deficit/Hyperactivity Disorder (ADHD). PLOS ONE *8*, e62177.
47. Reichenberg, A., Cederlöf, M., McMillan, A., Trzaskowski, M., Kapra, O., Fruchter, E., Ginat, K., Davidson, M., Weiser, M., Larsson, H., et al. (2016). Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. Proc. Natl. Acad. Sci. *113*, 1098–1103.
48. Renieri, A., Meloni, I., Longo, I., Ariani, F., Mari, F., Pescucci, C., and Cambi, F. (2003). Rett syndrome: the complex nature of a monogenic disease. J. Mol. Med. *81*, 346–354.
49. Rajaratnam, A., Shergill, J., Salcedo-Arellano, M., Saldarriaga, W., Duan, X., and Hagerman, R. (2017). Fragile X syndrome and fragile X-associated disorders. F1000Research *6*, 2112.
50. Fetit, R., Price, D.J., Lawrie, S.M., and Johnstone, M. (2020). Understanding the clinical manifestations of 16p11.2 deletion syndrome: a series of developmental case reports in children. Psychiatr. Genet. *30*, 136–140.

51. Vicari, S., Napoli, E., Cordeddu, V., Menghini, D., Alesi, V., Loddo, S., Novelli, A., and Tartaglia, M. (2019). Copy number variants in autism spectrum disorders. Prog. Neuropsychopharmacol. Biol. Psychiatry *92*, 421–427.
52. Gaugler, T., Klei, L., Sanders, S.J., Bodea, C.A., Goldberg, A.P., Lee, A.B., Mahajan, M., Manaa, D., Pawitan, Y., Reichert, J., et al. (2014). Most genetic risk for autism resides with common variation. Nat. Genet. *46*, 881–885.
53. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.
54. Iakoucheva, L.M., Muotri, A.R., and Sebat, J. (2019). Getting to the Cores of Autism. Cell *178*, 1287–1298.
55. Turner, T.N., and Eichler, E.E. (2019). The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. Trends Neurosci. *42*, 115–127.
56. Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat. Genet. *51*, 973–980.
57. Grove, J., Ripke, S., Als, T.D., Mattheisen, M., Walters, R.K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O.A., Anney, R., et al. (2019). Identification of common genetic risk variants for autism spectrum disorder. Nat. Genet. *51*, 431–444.
58. Williams, S.M., An, J.Y., Edson, J., Watts, M., Murigneux, V., Whitehouse, A.J.O., Jackson, C.J., Bellgrove, M.A., Cristino, A.S., and Claudianos, C. (2019). An integrative analysis of non-coding regulatory DNA variations associated with autism spectrum disorder. Mol. Psychiatry *24*, 1707–1719.
59. Matoba, N., Liang, D., Sun, H., Aygün, N., McAfee, J.C., Davis, J.E., Raffield, L.M., Qian, H., Piven, J., Li, Y., et al. (2020). Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism. Transl. Psychiatry *10*, 1–14.
60. Park, J.-H., Gail, M.H., Weinberg, C.R., Carroll, R.J., Chung, C.C., Wang, Z., Chanock, S.J., Fraumeni, J.F., and Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc. Natl. Acad. Sci. *108*, 18026–18031.
61. Anney, R.J.L., Ripke, S., Anttila, V., Grove, J., Holmans, P., Huang, H., Klei, L., Lee, P.H., Medland, S.E., Neale, B., et al. (2017). Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. Mol. Autism *8*, 21.
62. Acuna-Hidalgo, R., Veltman, J.A., and Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. Genome Biol. *17*, 241.
63. Goldmann, J.M., Veltman, J.A., and Gilissen, C. (2019). De Novo Mutations Reflect Development and Aging of the Human Germline. Trends Genet. *35*, 828–839.
64. Veltman, J.A., and Brunner, H.G. (2012). De novo mutations in human genetic disease. Nat. Rev. Genet. *13*, 565–575.
65. Cioppi, F., Casamonti, E., and Krausz, C. (2019). Age-Dependent De Novo Mutations During Spermatogenesis and Their Consequences. In Genetic Damage in Human Spermatozoa, E. Baldi, and M. Muratori, eds. (Cham: Springer International Publishing), pp. 29–46.
66. Wong, W.S.W., Solomon, B.D., Bodian, D.L., Kothiyal, P., Eley, G., Huddleston, K.C., Baker, R., Thach, D.C., Iyer, R.K., Vockley, J.G., et al. (2016). New observations on maternal age effect on germline de novo mutations. Nat. Commun. *7*, 10486.
67. Alonso-Gonzalez, A., Rodriguez-Fontenla, C., and Carracedo, A. (2018). De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis. Front. Genet. *9*, 406.
68. Ní Ghrálaigh, F., Gallagher, L., and Lopez, L.M. (2020). Autism spectrum disorder genomics: The progress and potential of genomic technologies. Genomics *112*, 5136–5142.
69. Wang, W., Corominas, R., and Lin, G.N. (2019). De novo Mutations From Whole Exome Sequencing in Neurodevelopmental and Psychiatric Disorders: From Discovery to Application. Front. Genet. *10*,.

70. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
71. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., DiLullo, N.M., Parikshak, N.N., Stein, J.L., et al. (2012). De novo mutations revealed by whole exome sequencing are strongly associated with autism. Nature *485*, 237–241.
72. He, X., Sanders, S.J., Liu, L., Rubeis, S.D., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., et al. (2013). Integrated Model of De Novo and Inherited Genetic Variants Yields Greater Power to Identify Risk Genes. PLOS Genet. *9*, e1003671.
73. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e23.
74. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233.
75. De Rubeis, S., He, X., Goldberg, A.P., Poultney, C.S., Samocha, K., Cicek, A.E., Kou, Y., Liu, L., Fromer, M., Walker, S., et al. (2014). Synaptic, transcriptional, and chromatin genes disrupted in autism. Nature *515*, 209–215.
76. Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T.N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L.G., Muzny, D.M., et al. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. Npj Genomic Med. *4*, 1–14.
77. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.
78. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
79. Rapaport, F., Boisson, B., Gregor, A., Béziat, V., Boisson-Dupuis, S., Bustamante, J., Jouanguy, E., Puel, A., Rosain, J., Zhang, Q., et al. (2021). Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. Proc. Natl. Acad. Sci. *118*, e2001248118.
80. Coe, B.P., Stessman, H.A.F., Sulovari, A., Geisheker, M.R., Bakken, T.E., Lake, A.M., Dougherty, J.D., Lein, E.S., Hormozdiari, F., Bernier, R.A., et al. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat. Genet. *51*, 106–116.
81. Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
82. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S., and Goldstein, D.B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLOS Genet. *9*, e1003709.
83. Huang, Y.-F. (2020). Unified inference of missense variant effects and gene constraints in the human genome. PLOS Genet. *16*, e1008922.
84. Qi, H., Zhang, H., Zhao, Y., Chen, C., Long, J.J., Chung, W.K., Guan, Y., and Shen, Y. (2021). MVP predicts the pathogenicity of missense variants by deep learning. Nat. Commun. *12*, 510.
85. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.
86. Sundaram, L., Gao, H., Padigepati, S.R., McRae, J.F., Li, Y., Kosmicki, J.A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. Nat. Genet. *50*, 1161–1170.
87. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.

88. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J., and Shendure, J. (2014). Saturation Editing of Genomic Regions by Multiplex Homology-Directed Repair. Nature *513*, 120–123.
89. Green, J., Pickles, A., Pasco, G., Bedford, R., Wan, M.W., Elsabbagh, M., Slonims, V., Gliga, T., Jones, E., Cheung, C., et al. (2017). Randomised trial of a parent□mediated intervention for infants at high risk for autism: longitudinal outcomes to age 3 years. J. Child Psychol. Psychiatry *58*, 1330–1340.
90. Whitehouse, A.J.O., Varcin, K.J., Pillar, S., Billingham, W., Alvares, G.A., Barbaro, J., Bent, C.A., Blenkley, D., Boutrus, M., Chee, A., et al. (2021). Effect of Preemptive Intervention on Developmental Outcomes Among Infants Showing Early Signs of Autism: A Randomized Clinical Trial of Outcomes to Diagnosis. JAMA Pediatr. *175*, e213298.
91. French, L., and Kennedy, E.M.M. (2018). Annual Research Review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review. J. Child Psychol. Psychiatry *59*, 444–456.
92. Guralnick, M.J. (2017). Early Intervention for Children with Intellectual Disabilities: An Update. J. Appl. Res. Intellect. Disabil. *30*, 211–229.
93. Politte, L.C., Howe, Y., Nowinski, L., Palumbo, M., and McDougle, C.J. (2015). Evidence-Based Treatments for Autism Spectrum Disorder. Curr. Treat. Options Psychiatry *2*, 38–56.
94. Keen, D., Couzens, D., Muspratt, S., and Rodger, S. (2010). The effects of a parent-focused intervention for children with a recent diagnosis of autism spectrum disorder on parenting stress and competence. Res. Autism Spectr. Disord. *4*, 229–241.
95. Horlin, C., Falkmer, M., Parsons, R., Albrecht, M.A., and Falkmer, T. (2014). The Cost of Autism Spectrum Disorders. PLoS ONE *9*, e106552.
96. Hormozdiari, F., Penn, O., Borenstein, E., and Eichler, E.E. (2015). The discovery of integrated gene networks for autism and related disorders. Genome Res. *25*, 142–154.
97. Chow, J., Jensen, M., Amini, H., Hormozdiari, F., Penn, O., Shifman, S., Girirajan, S., and Hormozdiari, F. (2019). Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. Genome Med. *11*, 65.
98. Chow, J.C., Zhou, R., and Hormozdiari, F. (2022). MAGI-MS: multiple seed-centric module discovery. Bioinforma. Adv. *2*, vbac025.
99. Huynh, L., and Hormozdiari, F. (2018). Combinatorial Approach for Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders. Genetics *210*, 1483–1495.
100. Chow, J.C., and Hormozdiari, F. (2022). Prediction of Neurodevelopmental Disorders Based on De Novo Coding Variation. J. Autism Dev. Disord.
101. Xu, X., Wells, A.B., O'Brien, D.R., Nehorai, A., and Dougherty, J.D. (2014). Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. J. Neurosci. *34*, 1420–1431.
102. Jain, A., and Tuteja, G. (2019). TissueEnrich: Tissue-specific gene enrichment analysis. Bioinformatics *35*, 1966–1967.

Chapter 1

Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders

Julie Chow, Matthew Jensen, Hajar Amini, Farhad Hormozdiari, Osnat Penn, Sagiv Shifman, Santhosh Girirajan & Fereydoun Hormozdiari

Genome Medicine volume 11, Article number: 65 (2019)

Abstract

Background

Neurodevelopmental disorders (NDDs) such as autism spectrum disorder, intellectual disability, developmental disability, and epilepsy are characterized by abnormal brain development that may affect cognition, learning, behavior, and motor skills. High co-occurrence (comorbidity) of NDDs indicates a shared, underlying biological mechanism. The genetic heterogeneity and overlap observed in NDDs make it difficult to identify the genetic causes of specific clinical symptoms, such as seizures.

Methods

We present a computational method, MAGI-S, to discover modules or groups of highly connected genes that together potentially perform a similar biological function. MAGI-S integrates protein-protein interaction and co-expression networks to form modules centered around the selection of a single "seed" gene, yielding modules consisting of genes that are highly co-expressed with the seed gene. We aim to dissect the epilepsy phenotype from a general NDD phenotype by providing MAGI-S with high confidence NDD seed genes with varying degrees of association with epilepsy, and we assess the enrichment of *de novo* mutation, NDD-associated genes, and relevant biological function of constructed modules.

Results

The newly identified modules account for the increased rate of *de novo* non-synonymous mutations in autism, intellectual disability, developmental disability, and epilepsy, and enrichment of copy number variations (CNVs) in developmental disability. We also observed that modules seeded with genes strongly associated with epilepsy tend to have a higher association with epilepsy phenotypes than modules seeded at other neurodevelopmental disorder genes. Modules seeded with genes strongly associated with epilepsy (e.g., *SCN1A*, *GABRA1*, and *KCNB1*) are significantly associated with synaptic transmission, long-term potentiation, and calcium signaling pathways. On the other hand, modules found with seed genes that are not associated or weakly associated with epilepsy are mostly involved with RNA regulation and chromatin remodeling.

Conclusions

In summary, our method identifies modules enriched with *de novo* non-synonymous mutations and can capture specific networks that underlie the epilepsy phenotype and display distinct enrichment in relevant biological processes. MAGI-S is available at https://github.com/jchow32/magi-s.

Background

Phenotypic heterogeneity in neurodevelopmental disorders (NDDs) has been well documented and includes variability in the severity of symptoms, age of onset, and comorbidity of distinct clinical phenotypes in affected individuals [1]. For example, more than 30% of individuals with autism spectrum disorders are estimated to have epilepsy [2], and individuals with epilepsy have an increased comorbidity of autism and intellectual disability/developmental disability (ID/DD) compared with individuals without epilepsy [3, 4]. The comorbidity of nosologically distinct

phenotypes is reflected in an overlap of causative genes and the involvement of similar molecular processes for these disorders [5, 6]. For example, *SCN2A*, the causative gene for epilepsy-associated Dravet syndrome, is also a primary candidate gene for familial autism [7, 8], while *NRXN1* has been associated with epilepsy as well as autism, schizophrenia, and developmental disability [9, 10]. In fact, nearly all genes with identified *de novo* mutations in epilepsy cases [11, 12] also have identified *de novo* mutations for other NDDs [13, 14].

While indicative of the shared biological pathways of NDDs, the high degree of pleiotropy for candidate NDD genes has made the classification of candidate genes and the discovery of novel genes towards distinct developmental features difficult. To date, several computational approaches have been devised to identify shared pathways of candidate genes for genetic disorders [15,16,17,18,19,20,21,22,23]. These approaches generally combine mutations identified from sequencing data of affected individuals with gene interaction networks and/or co-expression data to group genes with mutations in the same pathway. For example, the previously described tool MAGI was used to identify modules of genes significantly enriched for *de novo* variants in individuals with autism and ID/DD by integrating both protein-protein interaction networks and RNA sequencing data with variant calls [16]. Using this method, we identified distinct gene modules for signaling pathways and synaptic transmission from a set of *de novo* variants, and patients with mutations in these modules were observed to have more severe ID phenotypes than other patients. However, these methods do not allow for isolation of gene modules and pathways that are associated with a specific phenotype, such as epilepsy, compared with those that are more generally associated with multiple NDDs. Network and expression-based integration approaches that can accomplish this task are necessary to further understand the phenotypic heterogeneity of NDD-associated genes [1, 24].

Here, we present MAGI-S, an extension of our method MAGI, that identifies modules consisting of genes with high connectivity in the co-expression and protein-protein interaction networks that are also highly co-expressed with an input "seed gene." We have used MAGI-S to predict potential NDD modules that might help in dissecting the wide phenotypic and genotypic heterogeneity of NDDs. Our approach is based on the assumption that variants in genes that are highly interacting in protein-protein interactions networks and are highly co-expressed during brain development have a higher chance of manifesting similar phenotypes than variants in genes with a low degree of interaction. Using diverse sets of known candidate NDD genes, we utilized MAGI-S to identify modules of genes that are associated with NDD and can dissect the epilepsy phenotypes in NDD. We found that (i) most modules are significantly enriched for *de novo* mutations in affected probands with NDDs versus unaffected siblings, (ii) the union of genes in all modules related to epilepsy contains novel gene candidates for epilepsy, and (iii) these modules can dissect the epilepsy phenotypes for some NDD cases. Based on this analysis, we provide evidence that studying modules of related genes can be useful for better understanding the biomolecular causes of epilepsy phenotypes in NDDs.

Methods

MAGI-S

We previously developed MAGI [16], a tool for predicting pathways and modules significantly enriched for *de novo* variants associated with NDDs in cases compared to controls [16]. MAGI is a randomized algorithm that constructs genetic modules containing a set of related genes that are highly co-expressed during brain development, highly connected in protein-protein interaction networks, have very few severe variants in control populations, and are significantly enriched

among *de novo* variants in affected individuals. Specifically, MAGI first assigns a score s_i to each gene *i* based on the number of *de novo* variants present in the affected cases, while accounting for gene length and distribution of *de novo* non-synonymous mutation [16]. Next, MAGI finds a set of genes, *M*, that maximizes a standardized score of the selected genes (i.e., $S_M = \frac{\sum_{i \epsilon M} s_i}{\sqrt{|M|}}$ while satisfying the connectivity conditions for both protein interaction and co-expression networks.

Here, we developed MAGI-S, a method which differs from MAGI in that MAGI-S uses a *known disease gene* as the input "seed gene" to identify a module that is highly co-expressed with the seed gene, rather than using *de novo* variants observed in affected cases, as in MAGI [16]. The objective of MAGI-S is to discover a set of genes (i.e., module) that share similar biological function with the *seed gene*. MAGI-S utilizes the co-expression network built using the BrainSpan Atlas of the Developing Human Brain [25], high-quality protein interactions from the Human Protein Reference Database and STRING, and loss-of-function (LOF) variants from a set of *normal*/*control* samples (see Additional file 1) (Fig. 1, Additional file 1: Figure S1) [26, 27].

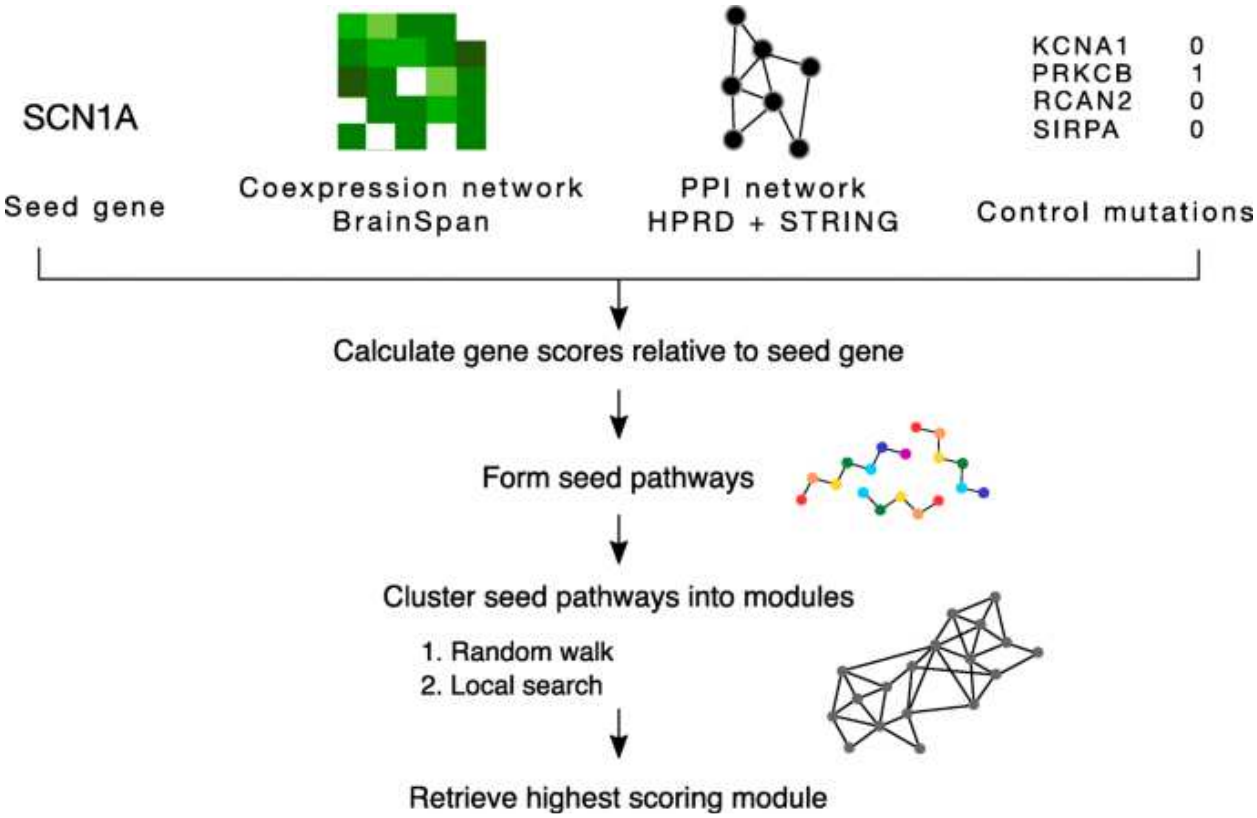


Fig 1. General overview of MAGI-S. A seed gene (e.g., SCN1A), protein-protein interaction (PPI) network, co-expression network, and LOF mutations in control samples are provided to MAGI-S to produce a seed centric module. Each gene in the PPI and co-expression networks is assigned a score based on the gene's degree of co-expression with the seed gene relative to all other genes in the networks. Seed pathways are high-scoring simple paths formed from genes that are highly co-expressed relative to the seed gene, connected in the PPI network, and have a low number of LOF variants in control samples. Seed pathways are clustered into modules via a random walk of a graph created by seed pathways, and the total score of a module is improved by local search (similar to the MAGI algorithm in Hormozdiari et al. [16]). MAGI-S is run with varied parameters related to module size, minimum co-expression, and minimum PPI density, and the highest scoring module is retrieved. We have used the human developmental data from BrainSpan Atlas for the

co-expression network construction. Furthermore, the combination of protein interactions from HPRD and STRING datasets was used as the PPI networks in our analysis

MAGI-S assigns a score to each gene based on the relative ranking of the co-expression of that gene and the *input seed gene*. MAGI-S then finds a set of genes that are highly connected across interaction networks, have a low number of severe variants in control samples, and are highly co-expressed with the input seed gene. The MAGI-S algorithm, similar to MAGI, has two main steps. First, it finds a set of connected paths with a length between 5 and 8 genes in protein interaction networks that have a high summation of gene scores. Second, similar to MAGI, it utilizes a random walk and local search approach to cluster the constructed paths found in the first step into modules while satisfying the connectivity and co-expression constraints (see Additional file 1). This procedure is repeated, and the module with the highest score is selected.

Seed genes

MAGI-S allows any gene to be considered as the seed gene and produces modules centered around that gene. In this study, we consider over 100 well-known neurodevelopmental genes as input seed genes. We applied MAGI-S on a comprehensive set of seed genes known to contribute to NDDs found using different whole-exome and genome sequencing studies. We have considered all the genes reported in the SFARI gene list which were ranked as having the most evidence for contribution to autism by their analysis [28]. More formally, known NDD seed genes were selected from the following main databases: (i) the genes from SFARI Gene database with most evidence of contribution to NDD (i.e., gene scores of either 1 or 2 with a total of 84 genes), (ii) the genes that have been concurrently reported to be associated with epilepsy in OMIM, DDG2P, EpilepsyGene, and a recent review paper of epilepsy genes (total of 41 genes, 4 of which also have

SFARI gene scores of either 1 or 2) [28,29,30,31,32], and (iii) an additional 6 genes moderately associated with epilepsy (*FLNA*, *FMR1*, *GRIN1*, *HNRNPU*, *NECAP1*, *NEDD1L*) (Additional file 1: Table S1). We have mainly focused on epilepsy as the phenotype of interest to investigate in patients with NDDs from discovered modules. In summary, we have considered a total of 127 genes which are known to be significantly associated with NDD phenotypes as input seed genes to MAGI-S. Due to a required minimum average co-expression, 16 potential seed genes failed to produce a module, yielding a total of 111 distinct modules. Note that many of these genes were selected based on the results available through whole-exome sequencing (WES) or whole-genome sequencing (WGS) of NDD cases/probands.

We first assigned the seed genes into three groups according to the known level of association with the epilepsy phenotype based on available disease-phenotype databases and literature [28, 30,31,32,33,34]. The three seed gene groups (classes) were defined based on reported epilepsy annotation from the following well-known sources: OMIM, DDG2P, EpilepsyGene, and Wang et al. [30, 31, 34]. We assigned the seed genes which were concurrently annotated by all four of these resources to be associated with epilepsy as *class 1*. Genes which were annotated to be associated in only a subset of the above resources were assigned to c*lass 2*. Finally, seed genes which were not associated with epilepsy in any of the above resources were assigned to *class 3* (Additional file 1: Table S1).

These three different c*lasses* of seed genes represent the degree of evidence in the literature for their association with epilepsy phenotype. The specified grouping is based on the decreasing degree of known association with seizure of these seed genes as follows:

- *Class*

  1 (*ARHGEF9*, *ALDH7A1*, *ALG13*, *CACNA1H*, *CACNB4*, *CDKL5*, *CHD2*, *CHRNB2*, *DE*

PDC5, *DNM1*, *EEF1A2*, *GABRA1*, *GABRB3*, *GABRG2*, *GNAO1*, *GRIN2A*, *GRIN2B*, *HCN1*, *KCNB1*, *KCNMA1*, *KCNQ2*, *KCNT1*, *KCTD7*, *LGI1*, *PCDH19*, *PRRT2*, *SCN1A*, *SCN1B*, *SCN2A*, *SCN8A*, *SLC25A22*, *SPTAN1*, *STX1B*, *STXBP1*, *TBC1D24*)

- *Class 2* (*ASH1L*, *BCKDK*, *CACNA1D*, *CNTNAP2*, *DIP2C*, *DYRK1A*, *FLNA*, *FMR1*, *GRIN1*, *HNRNPU*, *KMT2A*, *MBOAT7*, *MECP2*, *NECAP1*, *NEDD4L*, *PTEN*, *RANBP17*, *SCN9A*, *SLC6A1*, *SYNGAP1*, *TRIO*)
- *Class 3* (*ADNP*, *ANK2*, *ANKRD11*, *ARID1B*, *ASXL3*, *BAZ2B*, *BCL11A*, *CHD8*, *CIC*, *CTNND2*, *CUL3*, *DDX3X*, *DSCAM*, *ERBIN*, *GIGYF2*, *GRIA1*, *GRIP1*, *ILF2*, *INTS6*, *IRF2BPL*, *KDM5B*, *KDM6A*, *KMT2C*, *KMT5B*, *LEO1*, *MED13*, *MED13L*, *MET*, *MYT1L*, *NAA15*, *NCKAP1*, *NLGN3*, *NRXN1*, *PHF3*, *POGZ*, *RIMS1*, *SETD5*, *SHANK2*, *SHANK3*, *SMARCC2*, *SPAST*, *SRCAP*, *SRSF11*, *TAOK2*, *TBL1XR1*, *TBR1*, *TCF20*, *TNRC6B*, *TRIP12*, *UBN2*, *UPF3B*, *USP15*, *USP7*, *WAC*, *WDFY3*)

Class 1 seed genes include genes which have the most indication of being involved with epilepsy-associated phenotypes based on available databases and literature [28, 30,31,32,33,34]. On the other hand, c*lass 3* are seed genes which have the least/no amount of evidence to be involved with the epilepsy phenotype based on the literature and are more associated with other neurodevelopmental phenotypes.

Enrichment of de novo mutations and CNV from affected cases in modules

To assess the enrichment of *de novo* mutation in NDD cases relative to controls, *de novo* mutations were retrieved from denovo-db (version 1.6) [27]. denovo-db is a database of germline *de novo*

variants that have been identified by next-generation sequencing technology aggregated from 54 different studies with rigorous phenotyping standards (Additional file 1). Variants within denovo-db have been curated to include information such as genomic position, reference and alternate alleles, functional category, associated phenotypes of the individual possessing the variant, and orthogonal validation status. The largest set of *de novo* variants used in our analysis are from the Simons Simplex Collection (SSC), which includes the *de novo* variation of both affected ASD probands and unaffected siblings. The other denovo-db studies used in our analysis have also had the highest quality of *de novo* call sets with a very low false discovery rate and similar rates of *de novo* variation. The complete set of missense (and missense-near-splice) or loss-of-function (frameshift, splice donor, splice acceptor, stop-gained, stop-gained-near-splice, stop-lost) mutations from the denovo-db resource for Simons Simplex Collection set [35,36,37,38,39,40,41], Autism Sequencing Consortium (ASC) [42], MSSNG [43, 44], Deciphering Developmental Disorders (DDD) [29], Epi4K [11], Helbig et al. [45], intellectual disability studies [46,47,48,49], and schizophrenia studies [50,51,52,53,54] were considered. In total, we study 12,199 NDD patients with ASD, ID, DD, or epilepsy and 1933 sibling/control individuals (Additional file 2: Table S2: "denovo-db").

To determine the enrichment of copy number deletions and duplications within NDD cases relative to controls, we retrieve a copy number variant (CNV) morbidity map previously constructed from 29,085 children with developmental delay and 19,584 controls [60]. We assess the intersection of CNVs with genes within each module (Additional file 1).

Dissection of epilepsy phenotype by enrichment of epilepsy genes within modules

To evaluate the enrichment of *de novo* non-synonymous variation specific to different cohorts of NDDs within each module, we use Fisher's exact test to measure the enrichment of *de novo*

variation in probands with either (1) ASD, ID, or DD, or (2) epilepsy relative to controls. Additionally, to quantify the enrichment of modules for genes with phenotypic annotations associated with either (1) NDDs without epilepsy or (2) NDDs with epilepsy, we calculated an enrichment score for each module as $(M_P/M_{P'})/(G_P/(19,986-G_P))$, where M_P is the number of genes annotated as a certain NDD phenotype inside a module and $M_{P'}$ is the complement, and G_P is the total number of genes annotated as a certain phenotype. There is a total of 19,986 protein-coding genes in the human genome (Gencode GRCh38.p12). Phenotypic annotations were retrieved from SFARI, OMIM, DDG2P, EpilepsyGene, or Wang et al. (Additional file 1) [28, 30,31,32].

Pathway and ontology enrichment and expression analyses of modules

To describe pathway, gene ontology, and disease enrichment within a module, we provided a list of the genes within a module and its respective seed gene to Enrichr (http://amp.pharm.mssm.edu/Enrichr/) to produce pathway and GO biological process and Reactome pathway enrichments and OMIM disease annotations [55, 56]. We provided the same gene lists and the union of gene lists belonging to the same *class* to the cell type-specific expression analysis (CSEA), specific expression analyses (SEA), and tissue-specific expression analyses (TSEA) tools to assess the selective expression profiles of modules in the human brain and body [57].

Results

We hypothesized that the phenotypic heterogeneity observed in NDDs can be better understood by dissecting the phenotype based on the pathways and modules disrupted in these disorders.

Given the high comorbidity of NDDs, we tested the ability of MAGI-S to identify modules that can explain the association of specific genes to distinct phenotypes. Focusing on the more common comorbid feature of seizures, we applied MAGI-S to a subset of 111 seed genes strongly associated with NDDs, producing 1 module per seed gene. The size of the modules (i.e., the number of genes in each module) ranged from 25 to 79 genes, with an average size of 54 genes per module (Additional file 1: Figure S2).

Significant enrichment of de novo mutations in neurodevelopmental modules

We used a set of well-known neurodevelopmental disorder genes as the input seed genes to MAGI-S for producing relevant modules [28, 30,31,32]. We then investigated if the identified modules as a whole were enriched with *de novo* mutations found in the largest independent NDD studies in denovo-db, including 8426 neurodevelopmental disorder patients from (1) Simons Simplex Collection (SSC), (2) MSSNG, and (3) Deciphering Developmental Disabilities (DDD) 2017 cohorts relative to 1933 sibling/control samples (data from denovo-db version 1.6, Additional file 2: Table S2: "denovo-db") [27, 29, 43, 44, 58].

We compared the average number of loss-of-function (LOF), missense, and synonymous *de novo* mutations among probands and siblings/controls in the following sets: (1) the seed genes (total of 111 genes), (2) the union of all modules excluding seed genes (total of 1215 genes), (3) the union of all the genes in modules excluding the seed genes and 128 genes previously reported as significantly associated with ASD, ID, or DD [28, 34, 37, 42, 59] (Additional file 2: Table S2: "established NDD genes") (a total of 1184 genes), and (4) all other genes possessing *de novo* mutations outside of the union of all constructed modules (total of 17,758 genes).

First, as expected, we observed a significant enrichment of *de novo* variants in probands versus siblings for the seed genes ($p < 9.72e{-}52$, $p < 2.90e{-}12$, and $p < 1.68e{-}57$, for non-synonymous,

missense, and LOF variants, respectively) (Fig. 2). Second, more importantly, significant enrichment was observed for *de novo* variants disrupting the genes within these modules while excluding the seed genes ($p < 1.25e{-}10$, $p < 2.32e{-}6$, and $p < 1.74e{-}8$, for non-synonymous, missense, and LOF variants, respectively). Third, we also observed a significant enrichment of *de novo* mutations disrupting the union of genes in modules after excluding the seed genes and genes recently reported in the literature to be significantly enriched with *de novo* variants in NDDs ($p < 2.67e{-}4$, $p < 3.35e{-}3$, and $p < 5.22e{-}3$, for non-synonymous, missense, and LOF variants, respectively). We note that this indicates the set of genes identified in these modules, even after removing the seed genes and the known neurodevelopmental genes, is still enriched for *de novo* variants in affected probands versus unaffected siblings/controls. Thus, we conclude that the set of genes in these modules should be enriched in novel NDD genes. Finally, for the remaining set of 17,758 genes, we did not observe any significant difference in *de novo* non-synonymous or synonymous variation between affected probands and unaffected siblings/controls (Fig. 2). Due to an unequal ratio of cases (8426) to controls (1933) sampled, we performed bootstrapping for 20,000 iterations per comparison to estimate the accuracy of the reported average number of mutations per individual (Additional file 1), finding the same pattern of increased enrichment of *de novo* non-synonymous variation in cases relative to controls. We found that seed genes contribute to the largest percentage of NDD diagnosis, followed by module genes, indicating that the modules capture a significant proportion of *de novo* mutations that affect NDDs even while excluding identified ASD/ID/DD genes (Additional file 2: Table S2: “enrichment (union)”).

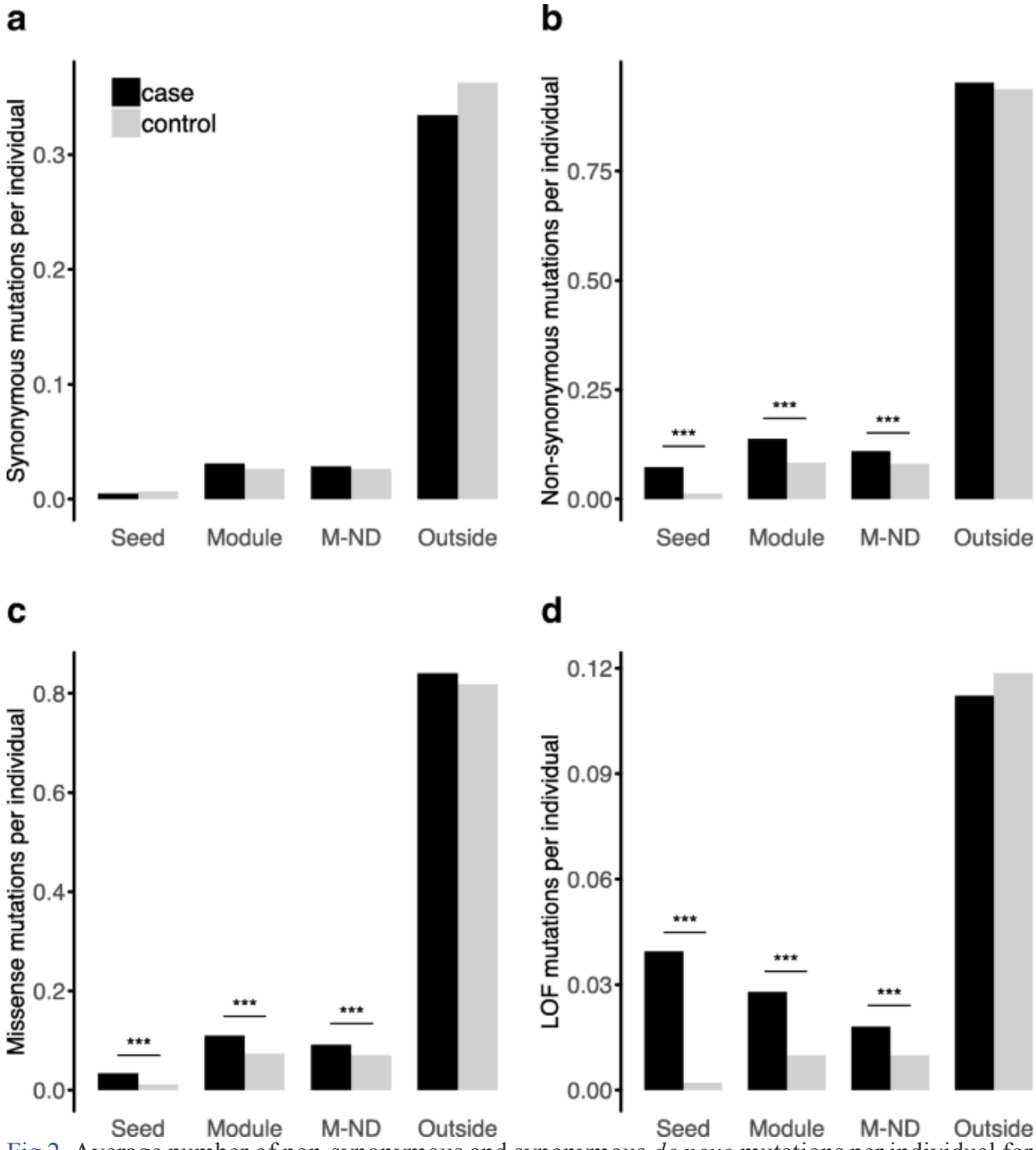


Fig 2. Average number of non-synonymous and synonymous *de novo* mutations per individual for probands and controls in seed genes (Seed), modules excluding seed genes (Module), module genes excluding 128 previously reported neurodevelopmental disorder genes (M-ND) (Additional file 2: Table S2: "established NDD genes"), and genes outside of any module (Outside). **a** No significant difference in the number of synonymous mutations exists between

cases and controls. Cases display significantly more non-synonymous (**b**), including missense (**c**) and loss-of-function (**d**), variants than controls in the Seed, Module, and M-ND groups

We also compared the proportions of *de novo* mutations associated with autism spectrum disorder (ASD) [35,36,37,38,39,40,41,42,43,44], intellectual disability (ID) [46,47,48,49], developmental disability (DD) [29], epilepsy [11, 45], and schizophrenia (SCZ) [50,51,52,53,54] in genes inside and outside of *each of the 111 modules independently*. A total of 12,199 NDD probands and 1933 sibling/control samples were examined (Additional file 2: Table S2: "denovo-db") [58]. For missense and LOF variants annotated with ASD, ID, DD, or epilepsy phenotypes, we evaluated the contingency tables and observed that we have an odds ratio significantly greater than 1 (with $p < 0.05$) for a large fraction of these modules, indicating enrichment of *de novo* mutations in neurodevelopmental probands versus controls in each of these modules (Additional file 2: Table S2). Resampling of contingency tables by 5000 iterations of permutation testing supports an increased enrichment of non-synonymous *de novo* mutation in individual modules (Additional file 2: Table S2: "contingency permutation").

Dissection of epilepsy phenotype in neurodevelopmental disorders using genetic modules

We next investigated the contribution of genetic modules in dissecting the epilepsy phenotype of NDDs. To assess the relevance of each of these 111 modules, we first measured the enrichment of *de novo* variants for ASD, ID, DD, epilepsy, and SCZ cohorts disrupting the modules selected for each of the seed genes (Fig. 3a). When considering any type of non-synonymous *de novo* variant associated with ASD, ID, DD, or epilepsy, we find that 64 of the 111 modules show significant enrichment in *de novo* non-synonymous mutations in these affected probands with neurodevelopmental disorders relative to unaffected siblings/controls (Fig. 3a). This shows that

most of these modules (64/111 > 61%) are indeed significantly enriched in *de novo* mutations observed in the neurodevelopmental disorder probands versus unaffected siblings/controls. Furthermore, we observed that almost all modules (100/111 > 90%) are enriched in coding copy number variations (CNVs) that were detected via array comparative genomic hybridization (aCGH) in probands with developmental disorders relative to controls [60] (Additional file 2: Table S2). Additionally, probands with *de novo* non-synonymous mutations display (1) significantly lower verbal, non-verbal, or full-scale IQ in 30 of 111 modules and (2) an enrichment in macrocephaly in 7 of 111 modules relative to probands with *de novo* non-synonymous mutations outside of the modules (Additional file 2: Table S2). As expected, none of the modules are significantly enriched in synonymous mutations in probands relative to siblings/controls (Additional file 3: Table S3). In addition, enrichment of *de novo* non-synonymous mutation in cases relative to controls for each module was assessed for penetrant missense mutations with CADD score greater than 15 (Additional file 4: Table S2a), revealing increased *de novo* mutation burden in cases relative to controls in the union of all modules and in 64/111 individual modules (Additional file 1: Figure S3S-4).

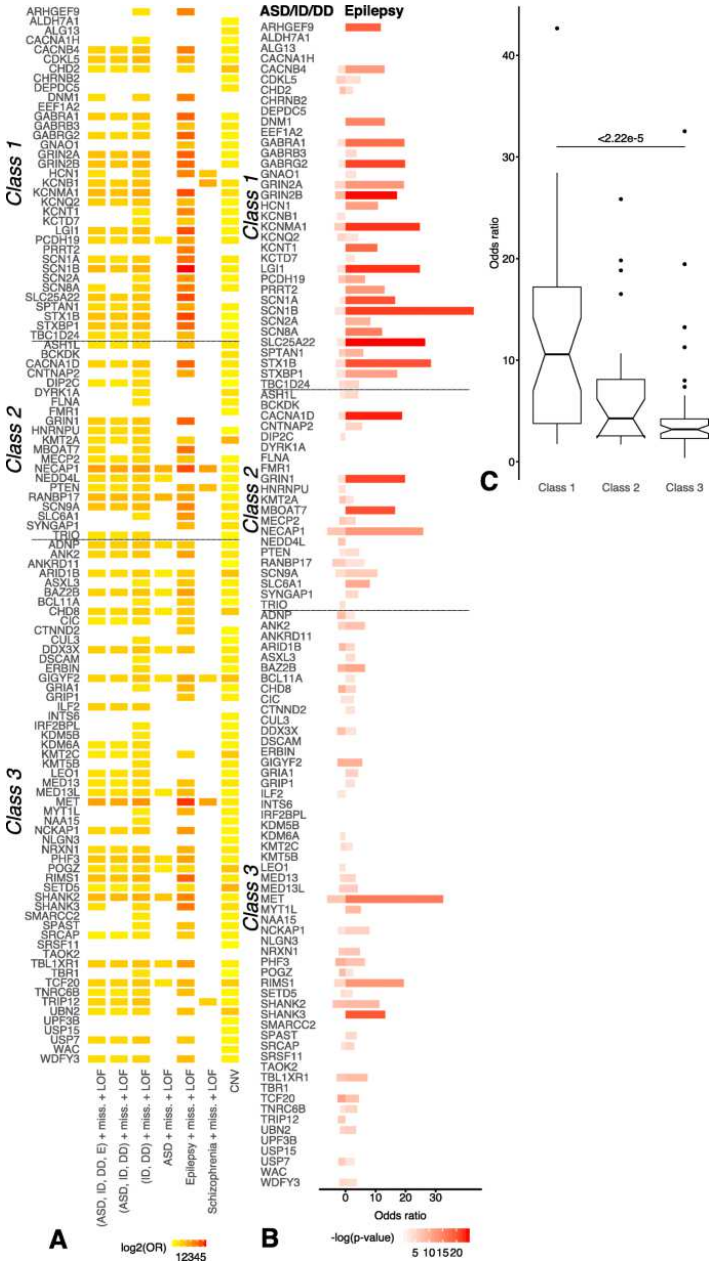


Fig 3. Summary of significant enrichment in *de novo* mutation and copy number variation (CNV) overlap in neurodevelopmental modules. Modules are grouped by class to indicate the degree of association of the seed gene with the epilepsy phenotype. Class 1, class 2, and class 3 modules correspond to the seed genes that have strong, moderate, and weak evidence of association with epilepsy, respectively. **a** Significant enrichment of missense (miss.) and loss-of-function (LOF) mutations for autism spectrum (ASD), intellectual disability (ID), developmental disability (DD),

epilepsy (E), and schizophrenia cohorts within modules. **b** Comparison of log2 of significant ($p < 0.05$) enrichment of *de novo* mutation for variants annotated as ASD/ID/DD (left) or epilepsy (right). **c** Average odds ratio of *de novo* mutations annotated in epilepsy cases relative to controls is significantly greater in class 1 modules compared to class 3 modules

We next studied the capacity of these modules to dissect the epilepsy phenotypes in neurodevelopmental disorders. We investigated the enrichment of non-synonymous *de novo* mutation in probands with either ASD/ID/DD or epilepsy (E) phenotype relative to controls. We first compared the odds ratio of *de novo* variants in ASD/ID/DD cohorts for each module to the odds ratio of *de novo* variants from the epilepsy cohort (Fig. 3). Note that *class 1* modules were constructed using neurodevelopmental seed genes with high evidence of association to epilepsy based on OMIM, DDG2P, and EpilepsyGene databases [28,29,30,31,32], whereas *class 3* modules were constructed using neurodevelopmental seed genes with minimal evidence of association with epilepsy in these databases. We compared the odds ratio of *de novo* variants observed in probands versus controls separately for the (1) ASD/ID/DD cohort and (2) the epilepsy cohort for each of the modules (Fig. 3b). The odds ratio for the ASD/ID/DD cohort is significantly greater than expected ($p < 0.05$) for 62 of 111 modules from all three classes. Similar fraction of modules from *classes 1*, *2*, and *3* had a higher than expected odds ratio for *de novo* mutations in the ASD/ID/DD cohort (19/35 > 54%, 13/21 > 61%, 30/55 > 54%, respectively).

On the other hand, a much larger fraction of modules from *class 1* (31/35 > 89%) had significantly greater than expected odds ratio for the *de novo* mutations in the *epilepsy cohort* (Fig. 3b). In contrast, the fraction of modules significantly enriched for *de novo* mutations is almost the same between ASD/ID/DDD and epilepsy cohorts for *class 3* modules (Fig. 3b).

We also compared the average odds ratio of modules for *de novo* non-synonymous variants in probands from the epilepsy cohort versus siblings/controls for modules in *class 1*, *class 2*, and *class 3* (Fig. 3c). We observed a significantly higher average odds ratio for *de novo* variants in the epilepsy cohort for modules in *class 1* compared with *class 3* ($p < 2.22e{-5}$). These results support the hypothesis that modules predicted using seed genes can help in dissecting the epilepsy phenotype in neurodevelopmental disorders (Fig. 3b, c).

Epilepsy genes enriched in modules built using class 1 seed genes

To investigate the ability of modules to dissect epilepsy phenotypes, we assessed the enrichment of the epilepsy genes in the modules predicted by MAGI-S. Modules seeded with genes from *class 1* gene set contain a significantly higher number of genes previously reported to be associated with epilepsy (Additional file 1) than modules found using seed genes from either *class 2* or *class 3* gene sets (Fig. 4) [28, 30, 32, 33]. Similarly, the average number of genes associated with epilepsy in *class 1* modules was significantly higher than that in *class 2* or *class 3* modules ($p < 8.21e{-3}$ and $p < 4.63e{-8}$, respectively). Genes most frequently shared among *class 1* modules (such as *DLG4*, *GRIN2A*, *PRKCB*, and *SNAP25*) have been associated with epilepsy, synaptic function, and neuronal processes [61,62,63,64].

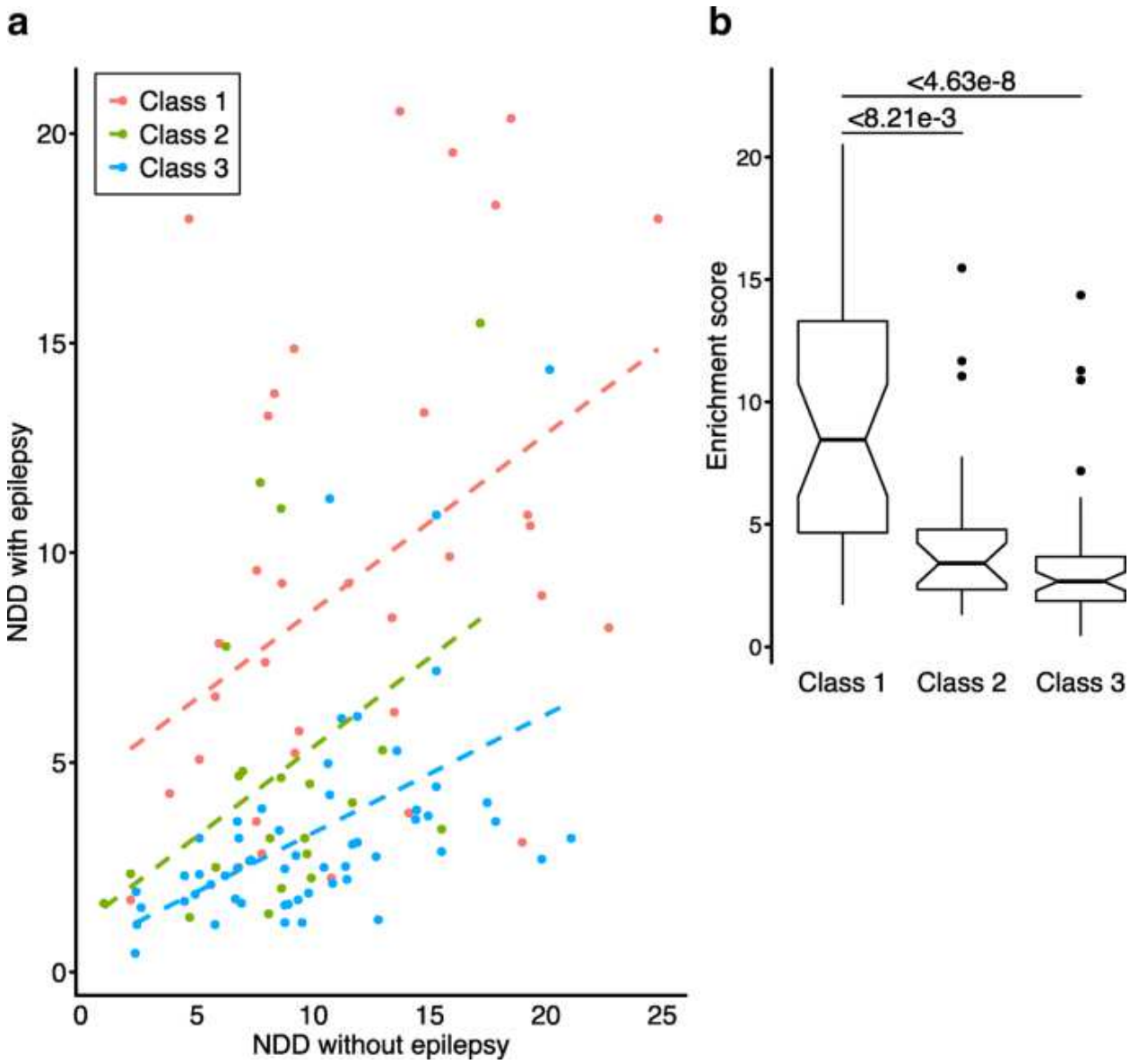


Fig 4. Phenotypic enrichment of genes in modules while including the seed gene. Enrichment is defined $(M_P/M_{P'})/(G_P/(19{,}986 - G_P))$, where M_P is the number of genes annotated as a certain neurodevelopmental disorder (NDD) phenotype inside a module, $M_{P'}$ is the complement of M_P, and G_P is the total number of genes annotated as a certain phenotype. The total number of genes in the human genome is 19,986 (Gencode GRCh38.p12). Increased enrichment of NDD with or without epilepsy for a module corresponds respectively to the presence or absence of epilepsy phenotypes associated with the seed gene. Modules are grouped by evidence of epilepsy association of the seed genes—i.e., class 1 (strong), class 2 (moderate), and class 3 (weak

association). **a** Increased enrichment of NDDs with epilepsy observed in class 1 modules are indicated by an increased *y*-intercept of class 1 regression line relative to class 2 and class 3. **b** Average enrichment of NDD with epilepsy is significantly greater in class 1 modules compared to class 2 or class 3 modules

Modules show enrichment in neuronal and epileptic processes

To assess the biological relevance of the identified modules, we analyzed the Gene Ontology (GO) and Reactome pathway enrichment for genes in each of the modules. The study of genetic modules disrupted in NDDs enables the identification of biological processes and functions that most contribute to these disorders. Enrichments from *class 1* and *class 2* modules indicate processes relevant to epilepsy and seizures, including GABAergic, cholinergic, dopaminergic, glycinergic, noradrenergic, and serotonergic synaptic transmission, and postsynaptic, excitatory, and inhibitory chemical synaptic transmission (Additional file 5: Table S4) [55].

Most modules ($22/27 > 81\%$) that contained a large proportion of genes associated with epilepsy (enrichment score greater than 7.5) were enriched in chemical synaptic transmission pathways (Additional file 1: Figure S5). On the other hand, all remaining modules were enriched for genes related to chromatin regulation and or axon guidance.

Furthermore, many of *class 1* and *class 2* modules ($33/56 > 58\%$) were enriched for interleukin signaling ($p < 0.0001$), which has been previously associated with epilepsy, and the MAPK, Ras, and VEGFR2 signaling pathways [65,66,67,68,69,70] (Additional file 1: Figure S6). Notch and TGF-beta signaling pathways were primarily enriched in *class 2* and *class 3* modules.

Using Enrichr analysis [56], we found that most ($18/35 > 51\%$) *class 1* modules are enriched for genes significantly associated with the OMIM disease terms “epilepsy,” “seizures,” or “ataxia.” Conversely, the genes that occur most commonly in *class 3* modules

(*UBC*, *EP300*, *SMAD2*, *CSNK2A1*, and *ABL1*) are associated with autism and/or intellectual disability [62, 71,72,73], and most (44/55 = 80%) *class 3* modules are enriched for genes associated with the term "autism" (Additional file 5: Table S4). We believe these results support the hypothesis that modules and networks can be utilized to dissect the phenotypic heterogeneity observed in NDDs.

Selective expression of specific cell types and regulation in neurodevelopmental modules

We also sought to use our modules to pinpoint the neuronal critical cell types involved with specific neurodevelopmental disorder phenotypes. Knowing the neuronal cell types involved would help further study of gene expression dysregulation in those cell types in affected patients. We observed that most modules from *class 1* are selectively expressed in layer 5 and 6 cortical neurons and D1+ and D2+ spiny neurons in the striatum. This is also true for the union of the genes in all *class 1* modules (Fig. 5), according to the cell type-specific expression analyses (CSEA) tool that uses RNAseq data from BrainSpan [57]. Furthermore, genes in *class 1* modules show expression in early infancy to young adulthood in the developing brain, whereas genes in *class 2* and *class 3* modules separately are expressed primarily during fetal stages of development. Additionally, *class 1* modules are enriched specifically in the brain relative to other tissues, which complements the enrichment of pathways involved in growth and development in *class 3* modules (Additional file 1: Figure S7, Additional file 6: Table S5).

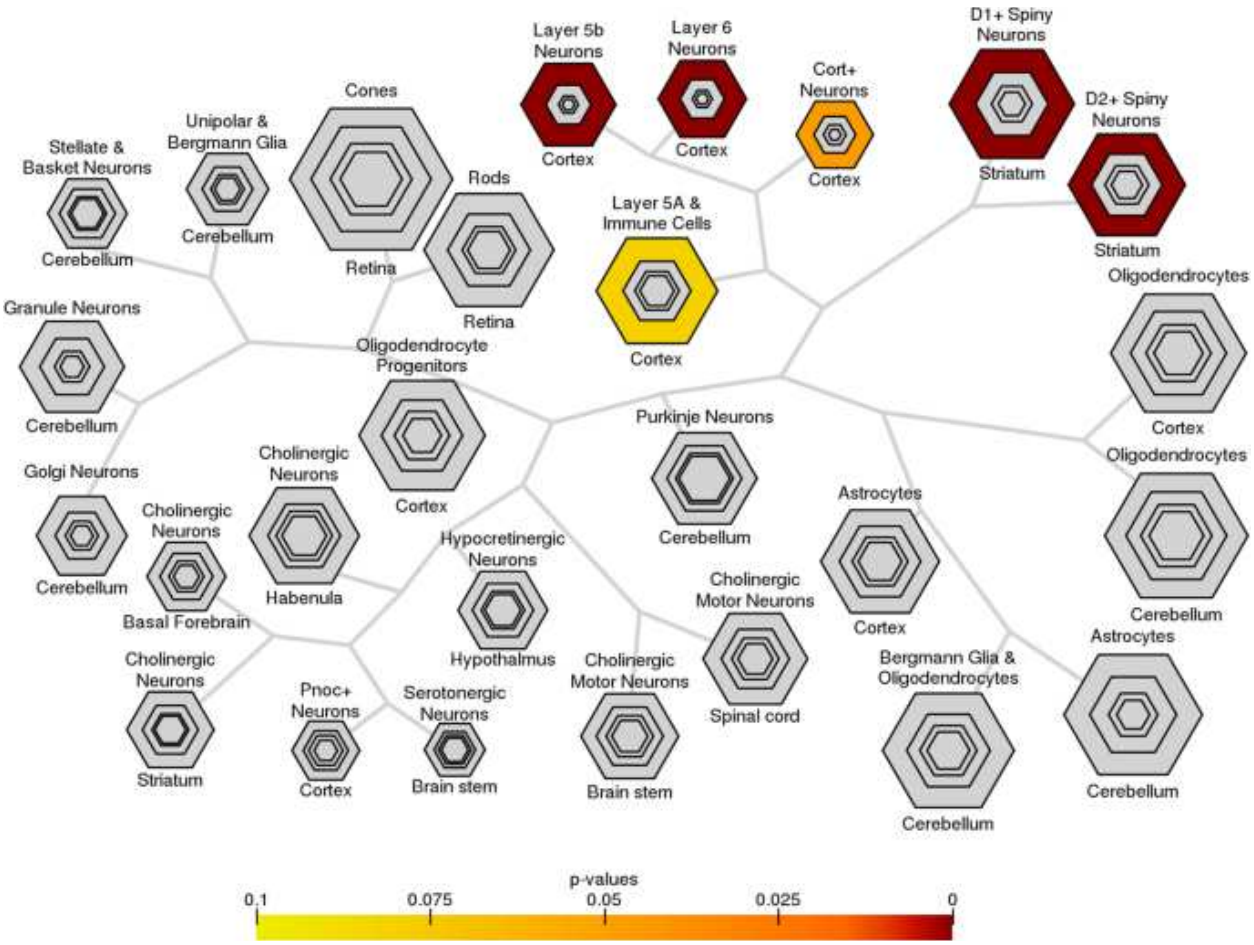


Fig 5. Cell type-specific expression analyses (CSEA) profile for the union of class 1 modules. Transcripts from provided gene lists that overlap significantly in specific cell types are indicated by intensity of color. Modules with seed genes strongly associated with epilepsy (class 1) show selective expression in the cortical neurons and spiny neurons in the striatum

Discussion

We have applied MAGI-S to construct modules from seed genes relevant to NDDs with and without association with epileptic phenotypes. The high degree of pleiotropy that exists among NDD genes complicates the understanding of the role of candidate genes in neurodevelopment. However, the ability to dissect the specific epilepsy phenotype from more general, heterogenous NDD phenotypes enables the improved characterization of candidate NDD genes in relation to

specific NDD subtypes. To dissect specific phenotypes from a more general phenotype, the choice of seed genes with varied degrees of association with the specific phenotype of interest is critical to module discovery. MAGI-S produces modules that are highly co-expressed with the seed gene. Thus, we investigated the hypothesis that the selection of seed genes that are strongly associated with epilepsy would produce modules that participate in pathways that underlie the epilepsy phenotype. Seed genes were selected from the aggregation of different, large-scale studies, including whole-genome and whole-exome sequencing studies. It is important to note that the selected seed genes have a significant impact on the modules found by MAGI-S. To construct modules in a spatio-temporal context, we chose to use the BrainSpan Atlas that describes the gene co-expression in the developing human brain at a range of life stages [25]. Critical processes that underlie typical neurodevelopment are performed by co-expressed genes that may be vulnerable to deleterious mutation and are thus relevant to NDDs [74, 75]. As co-expression resources for affected probands at varying developmental stages are developed, modules constructed by MAGI-S will be able to more accurately reflect pathway dysregulation over time.

Genes that occur frequently and exclusively in *class 1* or *class 2* module groups point to potential novel candidate genes, such as *DLG4*, *PRKCB*, *STX1A*, and *YWHAH*, that may play important roles in NDDs with epileptic phenotypes. Among the genes that have not previously been defined as epilepsy genes, *DLG4* is the most commonly shared gene among *class 1* modules and has been implicated in autism, intellectual disability, and synaptic function [76, 77]. *PRKCB* is a candidate gene for partial epilepsies and possibly involved in microRNA dysregulation in patients with mesial temporal lobe epilepsy [63, 78]. *STX1A* is a presynaptic gene to which its paralog *STXBP1* binds to regulate the SNARE complex, associated with epilepsy [79], and *STX1A* knockout mice experience reduced dense-core vesicle exocytosis and abnormal

monoaminergic transmission [80]. *YWHAx* genes, including *YWHAH*, have been hypothesized to be involved in neurological disorders including familial partial epilepsy [81, 82]. Genes that occur frequently in *class 3* modules but are absent in *class 1* modules such as *MYC* and *SIRT1* are implicated in tumorigenesis and metabolism [83, 84]; *SIRT1* is involved in learning and memory [85].

An accumulation of *de novo* missense and LOF mutations contribute to the manifestation of several NDDs [35, 86]. We found that most modules have significantly more *de novo* mutations in NDD probands than in controls (Fig. 2, Additional file 2: Table S2), and 88% of *class 1* modules are significantly enriched in epilepsy cohort-specific variants relative to controls. However, the potentially high degree of comorbidity among NDD probands and pleiotropy in NDDs suggests that particular *de novo* variants may impart the risk on several NDD phenotypes, although full comorbid phenotype information is not available from denovo-db. Thus, the enrichment of cohort-specific variants may not capture all genetic variation associated with a specific NDD phenotype, such as epilepsy. Analyses which concern all NDD-associated variants, such as the enrichment of *de novo* mutation within modules, are not dependent on phenotypic annotation and reflect the diversity of NDD phenotypes that may associate with seed genes and module genes. Seed genes and the union of all modules excluding seed genes capture a large proportion (~ 46%) of the *de novo* mutation signal that contributes to NDDs (Fig. 2, Additional file 2: Table S2: "enrichment (union)"). We observed that the union of genes identified in the modules is significantly enriched in *de novo* variants in NDD probands versus siblings/controls. This enrichment was still true after removing genes that were previously reported to be significantly enriched in *de novo* variants in these disorders (Additional file 2: Table S2: "established NDD genes"). However, we did not observe a significant difference between the genes not found in any module (17,758 genes) for *de*

novo variants in NDDs versus siblings/controls. We believe this supports a polygenic model for *de novo* variation in NDDs, in which mutations accumulate in genes that modulate pathways that underlie complex disease, in comparison with an omnigenic model, in which disease-associated signals are widespread across the genome. The penetrance of rare genetic variation may also be affected by common variation to result in a wide phenotypic heterogeneity among NDDs with typically monogenic forms [87]. Additionally, the overlap of coding CNVs with individual modules, confirmed via permutation tests, indicates that there is a significantly greater proportion of CNVs that overlap genes inside modules in probands than in controls, suggesting that copy number variation of genes within modules may also disrupt normal neurodevelopmental function. We assessed the relevance of modules by comparing enrichments in biological processes, signaling pathways, and selective expression in the human brain during different developmental stages. Modules seeded with genes that are strongly associated with epilepsy tend to cluster more distinctly than other module groups in relation to GO biological processes and Reactome pathways [55] (Additional file 1: Figure S5, Additional file 1: Figure S6). Most modules seeded with epilepsy genes are strongly related to chemical synaptic transmission, while modules produced with seed genes associated with other NDDs without epileptic phenotypes are related to chromatin organization and regulation, suggesting that the biological processes of a module correspond to its respective seed gene. Indeed, genome-wide analyses have previously associated autism genes with chromatin regulation [42, 59, 88, 89]. *Class 1* and *class 2* modules that have NDD with epilepsy enrichment scores consistently greater than 7.5 are enriched in similar biological processes involving chemical synaptic transmission and are selectively expressed in deep cortical neurons and spiny neurons in the striatum, which may indicate a stronger role of certain *class 2* seed genes in epilepsy than previously suggested. The selective expression of *class 1* modules in layer 5 and

6 cortical neurons is consistent with the epilepsy phenotype. Loss of excitatory neurons and the initiation of epileptic discharge have been observed in deep cortical layers, including layers 5 and 6, in individuals with epilepsy [90,91,92]. Additionally, in the striatum, direct and indirect neural pathways respectively modulate motor function via dopamine receptors D1 and D2 [93,94,95].

Conclusion

We have constructed modules of high connectivity relevant to NDDs. The choice of gene used for seed module construction is critical to module formation. To minimize bias in selecting seed genes, we selected all high confidence and strong candidate NDD genes curated from multiple, high-quality whole-exome and genome sequencing studies as per the SFARI Gene database and all genes that have been reported to be concurrently associated with epilepsy according to OMIM, DDG2P, EpilepsyGene, and a recent review [28, 30,31,32]. From our choices of seed genes, we describe three general classes of modules by the strength of evidence of epilepsy association. We find that the majority of modules are significantly enriched in *de novo* mutations, and modules constructed with seed genes that are strongly associated with epilepsy tend to be (1) significantly enriched in *de novo* mutation from individuals affected by epilepsy relative to unaffected controls, (2) enriched in epilepsy-associated genes, and (3) enriched for biological function relevant to seizure. Genes with *de novo* mutations that have not been traditionally associated with NDDs but are present in modules constructed from relevant seed genes could play an important role in disease. Furthermore, MAGI-S may be applied to dissect the genetic complexity of other diseases characterized by specific clinical features and identify candidate genes in diseases with strong *de novo* mutation components. The seed-centric approach to module discovery integrates interaction

networks and identifies a core set of genes strongly associated with phenotypes attributed to the seed gene and supported by biological evidence.

Availability of data and materials

denovo-db (version 1.6) is available at http://denovo-db.gs.washington.edu/denovo-db/. SFARI gene scores are available at https://gene.sfari.org/database/gene-scoring/. OMIM annotations are available at https://www.omim.org. DDG2P annotations are available at https://decipher.sanger.ac.uk/ddd#ddgenes. Enrichr is hosted at https://amp.pharm.mssm.edu/Enrichr/. CSEA, SEA, and TSEA tools are available at http://genetics.wustl.edu/jdlab/csea-tool-2/. MAGI source code, PPI network, co-expression hash tables and dataset, and control mutations are available at https://eichlerlab.gs.washington.edu/MAGI/. MAGI-S source code is available at https://github.com/jchow32/magi-s.

References

1. Geschwind DH, Levitt P. Autism spectrum disorders: developmental disconnection syndromes. Curr Opin Neurobiol. 2007;17(1):103–11.
2. Tuchman R, Rapin I. Epilepsy in autism. Lancet Neurol. 2002;1(6):352–8.
3. Amiet C, Gourfinkel-An I, Bouzamondo A, Tordjman S, Baulac M, Lechat P, et al. Epilepsy in autism is associated with intellectual disability and gender: evidence from a meta-analysis. Biol Psychiatry. 2008;64(7):577–82.
4. Polyak A, Rosenfeld JA, Girirajan S. An assessment of sex bias in neurodevelopmental disorders. Genome Med. 2015;7:94.
5. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316(5823):445–9.
6. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585–U125.
7. Weiss LA, Escayg A, Kearney JA, Trudeau M, MacDonald BT, Mori M, et al. Sodium channels SCN1A, SCN2A and SCN3A in familial autism. Mol Psychiatry. 2003;8(2):186–94.

8. Shi X, Yasumoto S, Nakagawa E, Fukasawa T, Uchiya S, Hirose S. Missense mutation of the sodium channel gene SCN2A causes Dravet syndrome. Brain and Development. 2009;31(10):758–62.
9. Ching MS, Shen Y, Tan WH, Jeste SS, Morrow EM, Chen X, et al. Deletions of NRXN1 (neurexin-1) predispose to a wide spectrum of developmental disorders. Am J Med Genet B Neuropsychiatr Genet. 2010;153B(4):937–47.
10. Harrison V, Connell L, Hayesmoore J, McParland J, Pike MG, Blair E. Compound heterozygous deletion of NRXN1 causing severe developmental delay with early onset epilepsy in two sisters. Am J Med Genet A. 2011;155A(11):2826–31.
11. Epi KC, Epilepsy Phenome/Genome P, Allen AS, Berkovic SF, Cossette P, Delanty N, et al. De novo mutations in epileptic encephalopathies. Nature. 2013;501(7466):217–21.
12. Mefford HC, Muhle H, Ostertag P, von Spiczak S, Buysse K, Baker C, et al. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet. 2010;6(5):e1000962.
13. Gonzalez-Mantilla AJ, Moreno-De-Luca A, Ledbetter DH, Martin CL. A cross-disorder method to identify novel candidate genes for developmental brain disorders. JAMA Psychiatry. 2016;73(3):275–83.
14. Jensen M, Girirajan S. Mapping a shared genetic basis for neurodevelopmental disorders. Genome Med. 2017;9(1):109.
15. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011;70(5):898–907.
16. Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders. Genome Res. 2015;25(1):142–54.
17. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci. 2016;19(11):1454–62.
18. Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, Cicek AE, et al. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism. 2014;5(1):22.
19. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246–50.
20. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. 2013;155(5):1008–21.
21. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38(Web Server issue):W214–20.
22. Vandin F, Upfal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res. 2012;22(2):375–85.
23. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18(3):507–22.
24. Jeste SS, Geschwind DH. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nat Rev Neurol. 2014;10(2):74–81.

25. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. 2013;41(Database issue):D996–D1008.
26. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database-2009 update. Nucleic Acids Res. 2009;37:D767–D72.
27. Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, Stessman HAF, et al. NAR breakthrough article denovo-db: a compendium of human de novo variants. Nucleic Acids Res. 2017;45(D1):D804–D11.
28. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013;4:36.
29. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433–8.
30. Ran X, Li J, Shao Q, Chen H, Lin Z, Sun ZS, et al. EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. Nucleic Acids Res. 2015;43(Database issue):D893–9.
31. Wang J, Lin ZJ, Liu L, Xu HQ, Shi YW, Yi YH, et al. Epilepsy-associated genes. Seizure. 2017;44:11–20.
32. Wright CF, Fitzgerald TW, Jones WD, Clayton S, Mcrae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305–14.
33. Berkovic S, Cossette P, Delanty N, Dlugos D, Eichler E, Epstein M, et al. Epi4K: gene discovery in 4,000 genomes. Epilepsia. 2012;53(8):1457–67.
34. Mcrae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433.
35. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216–U136.
36. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47(6):582–8.
37. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun. 2014;5:5595.
38. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012;338(6114):1619–22.
39. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic patterns of de novo mutation in simplex autism. Cell. 2017;171(3):710–22 e12.
40. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am J Hum Genet. 2016;98(1):58–74.
41. Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018;50(5):727–36.

42. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209–U119.
43. Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, et al. Genome-wide characteristics of de novo mutations in autism. NPJ Genom Med. 2016;1:160271–1602710.
44. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602.
45. Helbig KL, Farwell Hagman KD, Shinde DN, Mroske C, Powis Z, Li S, et al. Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy. Genet Med. 2016;18(9):898–905.
46. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367(20):1921–9.
47. Halvardson J, Zhao JJ, Zaghlool A, Wentzel C, Georgii-Hemming P, Mansson E, et al. Mutations in HECW2 are associated with intellectual disability and epilepsy. J Med Genet. 2016;53(10):697–704.
48. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380(9854):1674–82.
49. Lelieveld SH, Reijnders MR, Pfundt R, Yntema HG, Kamsteeg EJ, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci. 2016;19(9):1194–6.
50. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014;506(7487):179–84.
51. Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell. 2013;154(3):518–29.
52. Kranz TM, Harroch S, Manor O, Lichtenberg P, Friedlander Y, Seandel M, et al. De novo mutations from sporadic schizophrenia cases highlight important signaling genes in an independent sample. Schizophr Res. 2015;166(1–3):119–24.
53. McCarthy SE, Gillis J, Kramer M, Lihm J, Yoon S, Berstein Y, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. Mol Psychiatry. 2014;19(6):652–8.
54. Smedemark-Margulies N, Brownstein CA, Vargas S, Tembulkar SK, Towne MC, Shi J, et al. A novel de novo mutation in ATP1A3 and childhood-onset schizophrenia. Cold Spring Harb Mol Case Stud. 2016;2(5):a001008.
55. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2016;44(D1):D481–D7.
56. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7.

57. Xu XX, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. J Neurosci. 2014;34(4):1420–31.
58. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192–5.
59. Sanders SJ, Xin H, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87(6):1215–33.
60. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46(10):1063–71.
61. Addis L, Virdee JK, Vidler LR, Collier DA, Pal DK, Ursu D. Epilepsy-associated GRIN2A mutations reduce NMDA receptor trafficking and agonist potency - molecular profiling and functional rescue. Sci Rep-Uk. 2017;7:66.
62. Crawley JN, Heyer WD, LaSalle JM. Autism and cancer share risk genes, pathways, and drug targets. Trends Genet. 2016;32(3):139–46.
63. Danis B, van Rikxoort M, Kretschmann A, Zhang J, Godard P, Andonovic L, et al. Differential expression of miR-184 in temporal lobe epilepsy patients with and without hippocampal sclerosis - influence on microglial function. Sci Rep. 2016;6:33943.
64. Rohena L, Neidich J, Cho MT, Gonzalez KDF, Tang S, Devinsky O, et al. Mutation in SNAP25 as a novel genetic cause of epilepsy and intellectual disability. Rare Dis. 2015;1:e26314.
65. Goines P, Van de Water J. The immune system's role in the biology of autism. Curr Opin Neurol. 2010;23(2):111–7.
66. Griffin WST, Yeralan O, Sheng JG, Boop FA, Mrak RE, Rovnaghi CR, et al. Overexpression of the neurotrophic cytokine S100-beta in human temporal-lobe epilepsy. J Neurochem. 1995;65(1):228–33.
67. Lee C, Agoston DV. Inhibition of VEGF receptor 2 increased cell death of dentate hilar neurons after traumatic brain injury. Exp Neurol. 2009;220(2):400–3.
68. Maroso M, Balosso S, Ravizza T, Liu J, Bianchi ME, Vezzani A. Interleukin-1 type 1 receptor/Toll-like receptor signalling in epilepsy: the importance of IL-1beta and high-mobility group box 1. J Intern Med. 2011;270(4):319–26.
69. Vezzani A, French J, Bartfai T, Baram TZ. The role of inflammation in epilepsy. Nat Rev Neurol. 2011;7(1):31–40.
70. Vezzani A, Maroso M, Balosso S, Sanchez MA, Bartfai T. IL-1 receptor/Toll-like receptor signaling in infection, inflammation, stress and neurodegeneration couples hyperexcitability and seizures. Brain Behav Immun. 2011;25(7):1281–9.
71. Kaufman L, Ayub M, Vincent JB. The genetic basis of non-syndromic intellectual disability: a review. J Neurodev Disord. 2010;2(4):182–209.
72. Okur V, Cho MT, Henderson L, Retterer K, Schneider M, Sattler S, et al. De novo mutations in CSNK2A1 are associated with neurodevelopmental abnormalities and dysmorphic features. Hum Genet. 2016;135(7):699–705.
73. Wincent J, Luthman A, van Belzen M, van der Lans C, Albert J, Nordgren A, et al. CREBBP and EP300 mutational spectrum and clinical presentations in a cohort of Swedish patients with Rubinstein-Taybi syndrome EP300 mutational spectrum and

clinical presentations in a cohort of Swedish patients with Rubinstein-Taybi syndrome (vol 4, pg 39, 2016). Mol Genet Genom Med. 2016;4(3):367
74. Andersen SL. Trajectories of brain development: point of vulnerability or window of opportunity? Neurosci Biobehav Rev. 2003;27(1–2):3–18.
75. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013;155(5):997–1007.
76. Feyder M, Karlsson RM, Mathur P, Lyman M, Bock R, Momenan R, et al. Association of mouse Dlg4 (PSD-95) gene deletion and human DLG4 gene variation with phenotypes relevant to autism spectrum disorders and Williams' syndrome. Am J Psychiat. 2010;167(12):1508–17.
77. Krishnan ML, Van Steenwinckel J, Schang AL, Yan J, Arnadottir J, Le Charpentier T, et al. Integrative genomics of microglia implicates DLG4 (PSD95) in the white matter development of preterm infants. Nat Commun. 2017;8:428.
78. Kasperaviciute D, Catarino CB, Heinzen EL, Depondt C, Cavalleri GL, Caboclo LO, et al. Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study. Brain. 2010;133:2136–47.
79. Gerber SH, Rah JC, Min SW, Liu XR, de Wit H, Dulubova I, et al. Conformational switch of syntaxin-1 controls synaptic vesicle fusion. Science. 2008;321(5895):1507–10.
80. Fujiwara T, Kofuji T, Akagawa K. Dysfunction of the hypothalamic-pituitary-adrenal axis in STX1A knockout mice. J Neuroendocrinol. 2011;23(12):1222–30.
81. Foote M, Zhou Y. 14-3-3 proteins in neurological disorders. Int J Biochem Mol Biol. 2012;3(2):152–64.
82. Morales-Corraliza J, Gomez-Garre P, Sanz R, Diaz-Otero F, Gutierrez-Delicado E, Serratosa JM. Familial partial epilepsy with variable foci: a new family with suggestion of linkage to chromosome 22q12. Epilepsia. 2010;51(9):1910–4.
83. Dang CV. MYC, metabolism, cell growth, and tumorigenesis. Cold Spring Harb Perspect Med. 2013;3(8):a014217.
84. Song NY, Surh YJ. Janus-faced role of SIRT1 in tumorigenesis. Ann N Y Acad Sci. 2012;1271:10–9.
85. Michan S, Li Y, Chou MMH, Parrella E, Ge HY, Long JM, et al. SIRT1 is essential for normal cognitive function and synaptic plasticity. J Neurosci. 2010;30(29):9695–707.
86. Veltman JA, Brunner HG. Applications of next-generation sequencing de novo mutations in human genetic disease. Nat Rev Genet. 2012;13(8):565–75.
87. Niemi MEK, Martin HC, Rice DL, Gallon G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562(7726):268.
88. Lasalle JM. Autism genes keep turning up chromatin. OA Autism. 2013;1(2):14.
89. Cotney J, Muhle RA, Sanders SJ, Liu L, Willsey AJ, Niu W, et al. The autism-associated chromatin modifier CHD8 regulates other autism risk genes during human neurodevelopment. Nat Commun. 2015;6:6404.
90. Chauvette S, Soltani S, Seigneur J, Timofeev I. In vivo models of cortical acquired epilepsy. J Neurosci Meth. 2016;260:185–201.
91. Sloviter RS. Decreased hippocampal inhibition and a selective loss of interneurons in experimental epilepsy. Science. 1987;235(4784):73–6.

92. Swann JW, Al-Noori S, Jiang MH, Lee CL. Spine loss and other dendritic abnormalities in epilepsy. Hippocampus. 2000;10(5):617–25.
93. Gagnon D, Petryszyn S, Sanchez MG, Bories C, Beaulieu JM, De Koninck Y, et al. Striatal neurons expressing D-1 and D-2 receptors are morphologically distinct and differently affected by dopamine denervation in mice. Sci Rep. 2017;7:41432.
94. Gerfen CR, Engber TM, Mahan LC, Susel Z, Chase TN, Monsma FJ, et al. D1 and D2 dopamine receptor regulated gene-expression of striatonigral and striatopallidal neurons. Science. 1990;250(4986):1429–32.
95. Perreault ML, Hasbi A, O'Dowd BF, George SR. The dopamine D1-D2 receptor heteromer in striatal medium spiny neurons: evidence for a third distinct neuronal pathway in basal ganglia. Front Neuroanat. 2011;5:31.

Supplementary information

Supplementary information and Additional files 2-6 are available online at

https://doi.org/10.1186/s13073-019-0678-y.

Additional file 1:

Supplementary methods, Figure S1-S7, and Table S1.

Additional file 2:

Table S2. Summary of analyses performed per module, including determinations of enrichment of *de novo* mutation, overlap with coding copy number variations. Module membership and frequency of occurrence for all genes selected in any module are displayed in the 'modules' tab. Number of cases and controls for ASD, ID, DD, and epilepsy cohorts within denovo-db are displayed in the 'denovo-db' tab. Contingency tables for Fisher's exact test were constructed to assess the enrichment of *de novo* mutation and copy number variations in modules. Contingency table permutation empirical *p-values* are displayed in the 'contingency permutations' tab. Percent contribution to neurodevelopmental disorder diagnoses and comparison of average number of

mutations per individual are displayed in the 'enrichment (union)' tab. Tabs corresponding to a module name show the total number of *de novo* variants, associated phenotype, type of variant, and neurodevelopmental disorder-related descriptions per module.

Additional file 3:

Table S3. Proportions of synonymous mutations in neurodevelopmental cases relative to controls. Tabs correspond to modules and respective total number of synonymous *de novo* variants.

Additional file 4:

Table S2a. Similar to Additional file 2: Table S2, Additional file 4: Table S2a displays a summary of analyses performed per module while requiring a CADD score greater than 15 for missense variants.

Additional file 5:

Table S4. Significant GO terms, KEGG, and Reactome pathway enrichments, and OMIM disease terms per module (*p-value* < 0.05).

Additional file 6:

Contains Table S5. Selective expression profiles for union of modules based on strength of epilepsy association (*Classes 1*, *2*, and *3* as C1, C2, and C3, respectively), including: Cell-type specific Expression Analyses (CSEA), Specific Expression Analyses (SEA) for adult brain regions and development, and Tissue-Specific Expression Analyses (TSEA).

Additional File 1

Supplementary Methods

A total of 111 seed genes (*ADNP*, *ALDH7A1*, *ALG13*, *ANK2*, *ANKRD11*, *ARHGEF9*, *ARID1B*, *ASH1L*, *ASXL3*, *BAZ2B*, *BCKDK*, *BCL11A*, *CACNA1D*, *CACNA1H*, *CACNB4*, *CDKL5*, *CHD2*, *CHD8*, *CHRNB2*, *CIC*, *CNTNAP2*, *CTNND2*, *CUL3*, *DDX3X*, *DEPDC5*, *DIP2C*, *DNM1*, *DSCAM*, *DYRK1A*, *EEF1A2*, *ERBIN*, *FLNA*, *FMR1*, *GABRA1*, *GABRB3*, *GABRG2*, *GIGYF2*, *GNAO1*, *GRIA1*, *GRIN1*, *GRIN2A*, *GRIN2B*, *GRIP1*, *HCN1*, *HNRNPU*, *ILF2*, *INTS6*, *IRF2BPL*, *KCNB1*, *KCNMA1*, *KCNQ2*, *KCNT1*, *KCTD7*, *KDM5B*, *KDM6A*, *KMT2A*, *KMT2C*, *KMT5B*, *LEO1*, *LGI1*, *MBOAT7*, *MECP2*, *MED13L*, *MED13*, *MET*, *MYT1L*, *NAA15*, *NCKAP1*, *NECAP1*, *NEDD4L*, *NLGN3*, *NRXN1*, *PCDH19*, *PHF3*, *POGZ*, *PRRT2*, *PTEN*, *RANBP17*, *RIMS1*, *SCN1A*, *SCN1B*, *SCN2A*, *SCN8A*, *SCN9A*, *SETD5*, *SHANK2*, *SHANK3*, *SLC25A22*, *SLC6A1*, *SMARCC2*, *SPAST*, *SPTAN1*, *SRCAP*, *SRSF11*, *STX1B*, *STXBP1*, *SYNGAP1*, *TAOK2*, *TBC1D24*, *TBL1XR1*, *TBR1*, *TCF20*, *TNRC6B*, *TRIO*, TRIP12*, UBN2*, *UPF3B*, *USP15*, *USP7*, *WAC*, *WDFY3*) associated with neurodevelopmental disorders (NDDs) including autism spectrum disorders (ASD), intellectual disability (ID), developmental disability (DD), or epilepsy were selected to produce modules via MAGI-S. Seed genes were selected from the following databases: (i) all genes from SFARI Gene database with gene scores of either 1 (high confidence ASD gene) or 2 (strong candidate gene for ASD) (total of 84 genes), (ii) the genes that have been **concurrently** reported to be associated with epilepsy in 1) OMIM, 2) DDG2P, 3) EpilepsyGene, and 4) a recent review paper of epilepsy genes (total of 41 genes, 4 of which also have SFARI gene scores of either 1 or 2) (1-5), (iii) and an additional 6 genes associated with NDDs.

Due to few protein-protein interactions (PPIs) or co-expression values (**Figure S1**) associated with certain gene names (*ERBIN*, *IRF2BPL*, *KMT2A*, *KMT2C*, *KMT5B*, *MBOAT7*, *NAA15, SRSF11*),

respective alternate gene names (*ERBB2IP*, *C14orf4*, *MLL*, *SUV420H1*, *MLL3*, *LENG4*, *NARG1*, *SFRS11*) were provided to MAGI-S for module discovery. Parameters related to minimum size (20-35), minimum average co-expression value (0.425-0.52), and minimum PPI density (0.085-0.14) of modules were tested through multiple trials to identify the optimal module producing the highest score. Potential seed genes *CACNA1A*, *CACNA2D3*, *CHRNA2*, *CHRNA4*, *CNTN4*, *DEAF1*, *FOXP1*, *KAT2B*, *KATNAL2*, *MAGEL2*, *MSNP1AS*, *PTCHD1*, *RELN*, *SLC1A2*, *SZT2*, *WWOX*, were omitted from enrichment analysis and failed to produce modules due to average co-expression values below the specified range for minimum average co-expression value. Modules ranged in size from 25 to 79 genes (**Figure S2**). Genes within modules were renamed according to approved gene symbols for enrichment analyses.

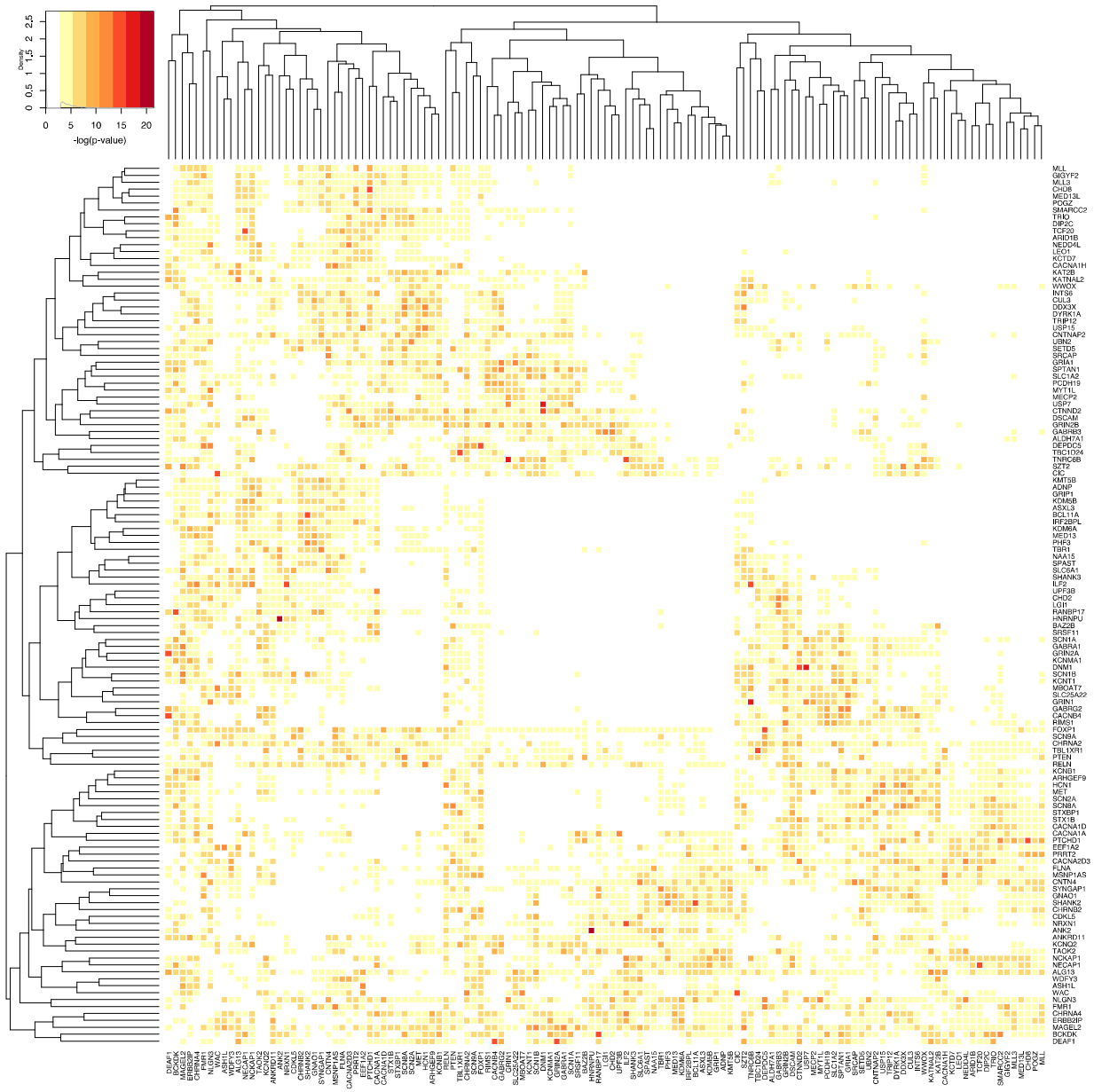


Figure S1. Co-expression between seed genes. Co-expression values were determined by adjacency and Topological Overlap Matrix (TOM) matrices with power of 2 to reveal significant ($p < 0.05$) co-expression among seed genes.

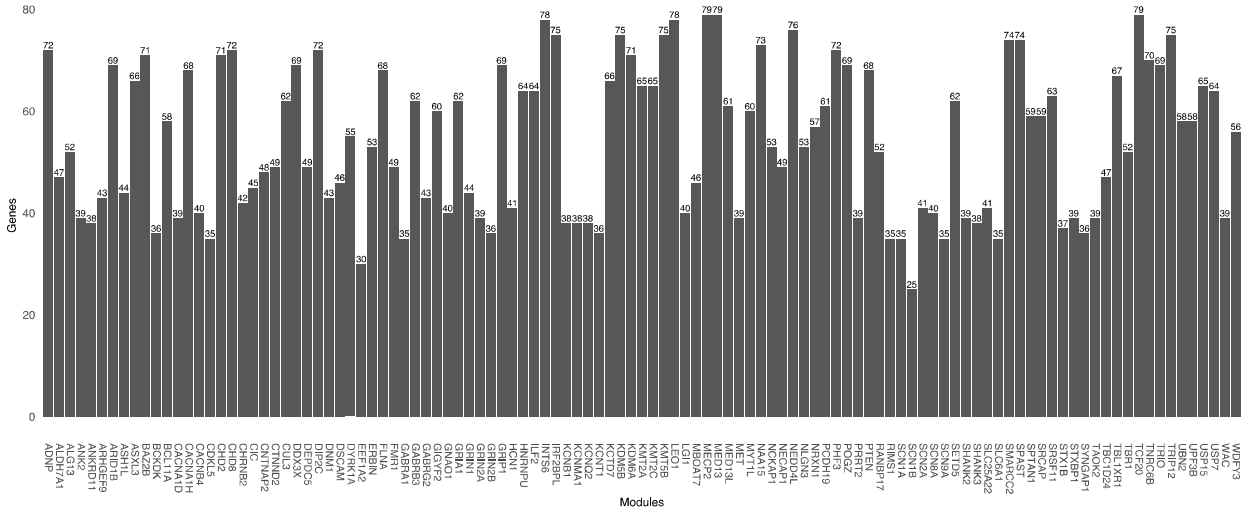


Figure S2. Number of genes within each module excluding seed gene.

Module groups (*Classes*) were defined by **concurrent** epilepsy annotations from the following sources (**Table S1**): *Class 1* (OMIM, DDG2P, EpilepsyGene, and Wang et al. 2017)*, Class 2* (a **subset** of *Class 1* sources), *Class 3* (**none** of *Class 1* sources) (3-5).

Determining enrichment of de novo mutations within modules

De novo mutations were retrieved from denovo-db (version 1.6) (6). The total number of missense (or missense-near-splice) or loss of function (frameshift, frameshift-near-splice, splice donor, splice acceptor, stop-gained, stop-gained-near-splice, stop-lost) mutations from the denovo-db Simons Simplex Collection (SSC) set (7-13), Autism Sequencing Consortium (ASC) (14), MSSNG (15, 16), Deciphering Developmental Disorders (DDD) (2), Epi4K (17), Helbig et al. 2016, and selected intellectual disability (18-21) and schizophrenia studies (22-26) were recorded. Rigorous phenotyping standards were applied in contributing studies. The Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS), among other measures (https://www.sfari.org/resources/ssc-instruments/), were recorded for probands with autism. For the SSC cohort, phenotyping was uniform across 12 university-affiliated clinics serving children with autism (27). For the Epi4K cohort, epilepsy phenotyping was accomplished by magnetic resonance imaging (MRI), electroencephalogram (EEG) findings, collection of medical records, and structured interviews (28). For intellectual disability cohorts, individuals with intellectual disability who were referred to a tertiary referral center for clinical genetics were further evaluated by a clinical geneticist (18), patients with intellectual disability were recruited by the Genetic Diagnostics Unit at Uppsala University Hospital (19), and patients with severe non-syndromic intellectual disability were selected from the German Mental Retardation Network (20). For the developmental disability cohort, individuals with severe undiagnosed developmental disability were recruited, and phenotypes were described using the Human Phenotype Ontology

(2). Patients in the schizophrenia cohort were recruited from psychiatric treatment settings (22-26).

We retrieved the total number of non-synonymous and synonymous mutations in genes in probands and controls and normalized the number of mutations by number of SSC, MSSNG, and DDD probands (8,426) and controls (1,933) considered (**Additional file 2: Table S2**: **'denovo-db'**). To compare the average number of *de novo* mutations per individual among probands and controls in 1) seed genes, 2) the union of all modules excluding seed genes, 3) the union of all modules excluding seed genes and 128 previously identified ASD/DD genes from the sources: de Rubeis et al. 2014, Mcrae et al. 2017 (DDD), O'Roak et al. 2014, Sanders et al. 2015, SFARI (score of 1) (1, 9, 14, 29, 30), and 4) outside of modules and seeds, we applied a one-tailed two-sample t-test on normalized counts of mutations per individual. To assess the accuracy of the t-test to measure true difference in normalized average number of mutations per individual, we applied 20,000 iterations of bootstrapping per comparison to calculate an empirical *p-value*. To determine an empirical *p-value*, we created bootstrap samples with replacement of cases (8,426) and controls (1,933) and calculated the t-test statistic for the bootstrapped sample and its respective *p-value*. If this *p-value* from the bootstrap sample was less than the *p-value* calculated prior to bootstrapping, then a 'total score' was incremented by one. The empirical *p*-value was then calculated as the total score divided by the number of iterations (20,000) plus 1.

We additionally constructed contingency tables of the raw counts of *de novo* mutations and evaluated Fisher's exact test to compare proportions of non-synonymous mutation among probands and controls within the seed genes, the union of all modules excluding seed genes, and outside of modules. Percent contribution to the neurodevelopmental phenotypes was calculated by dividing the difference between the normalized number of mutations in probands and controls by the

normalized number of mutations in probands (**Additional file Table S2: 'enrichment (union)'**). We also assessed the average number of *de novo* mutations among probands and controls while requiring a CADD score greater than 15 for missense variants to examine likely penetrant non-synonymous mutations (**Figure S3**).

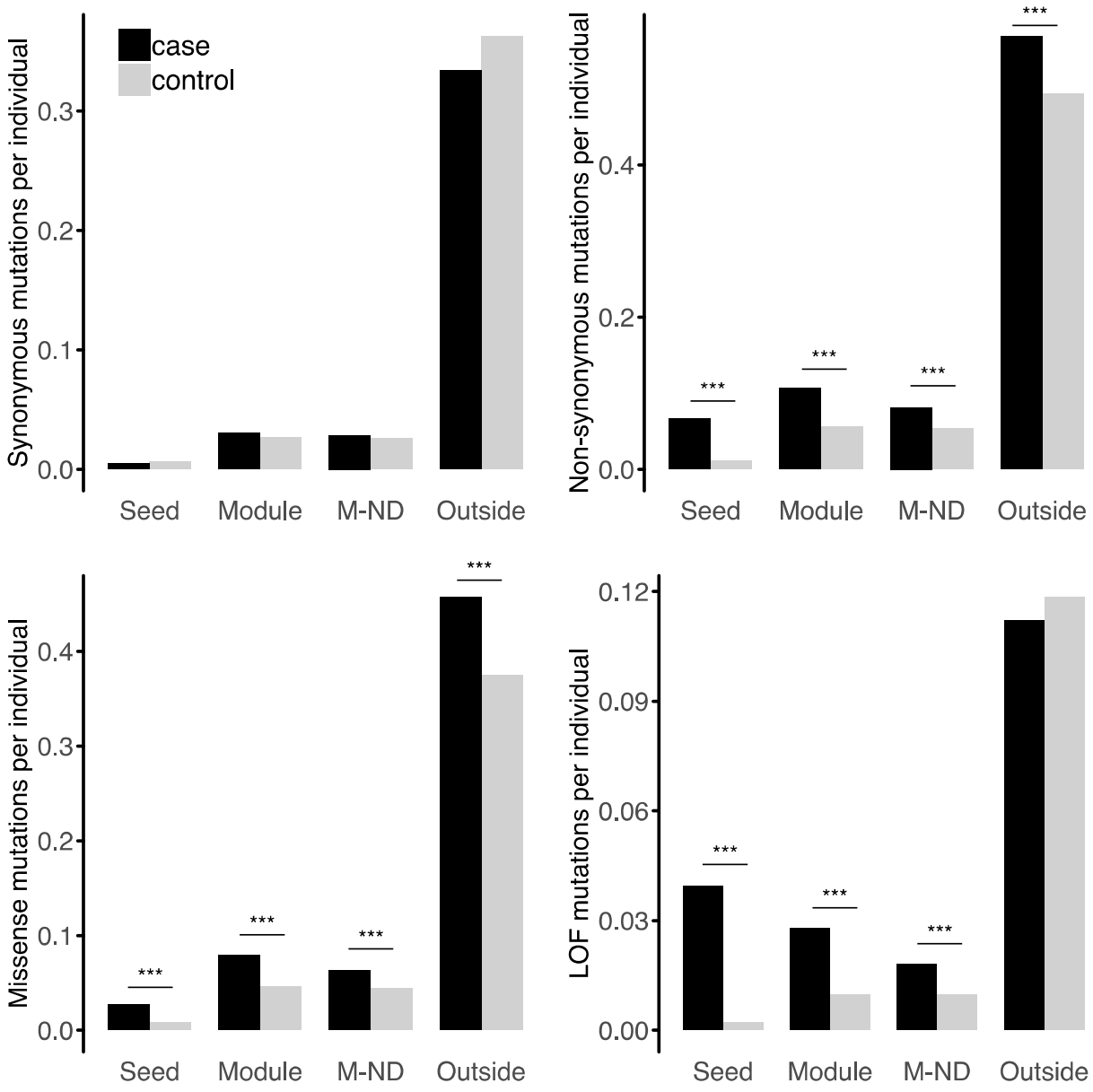


Figure S3. Average number of non-synonymous and synonymous *de novo* mutations per individual for probands and controls in seed genes ('Seed'), modules excluding seed genes ('Module'), Module genes excluding 128 previously reported neurodevelopmental disorder genes ('M-ND'). Penetrant missense mutations are examined by requiring CADD score to be greater than 15.

To determine if significant enrichment of non-synonymous *de novo* mutations within modules exists in probands with NDDs relative to controls, we compared the number of *de novo* missense and loss of function mutations inside and outside of the module via Fisher's exact test with consideration of a) only autism, developmental disorder, or intellectual disability variants (ASD, DD, ID), b) only ID or DD variants, c) only ASD variants, d) only epilepsy variants, and e) ASD, DD, ID and epilepsy variants (**Additional file 2: Table S2: 'denovo-db'**). Additionally, we further assessed the significance of *de novo* mutation enrichment in probands by considering a) missense or loss of function mutations, b) only missense, or c) only loss of function mutations. We repeated the above analyses while excluding variants attributed to the seed gene. To assess the accuracy of contingency tables applied to test the increased enrichment of non-synonymous mutation in cases relative to controls while excluding the seed gene, we applied resampling via 5,000 iterations of permutation testing per comparison. Cases and controls were randomly sampled indiscriminately to yield two sets of size equal to the number of cases and controls. Fisher's exact test was evaluated for each permuted set, and contingency tables were created to determine significant difference in proportions of non-synonymous mutation in or outside modules. We incremented a 'total score' for every permuted *p-value* less than the *p-value* calculated prior to permutation testing and calculated an empirical *p-value* as the total score divided by the number of iterations (5,000) plus 1.

The absence of any *de novo* mutations in controls in certain modules results in an infinitely large odds ratio. Thus, to better visualize significant enrichment of *de novo* mutation for modules with zero *de novo* mutations in controls, we increased the count of *de novo* mutation to one. We repeated the above analyses requiring missense variants to have a CADD score greater than 15 (**Figure S4**).

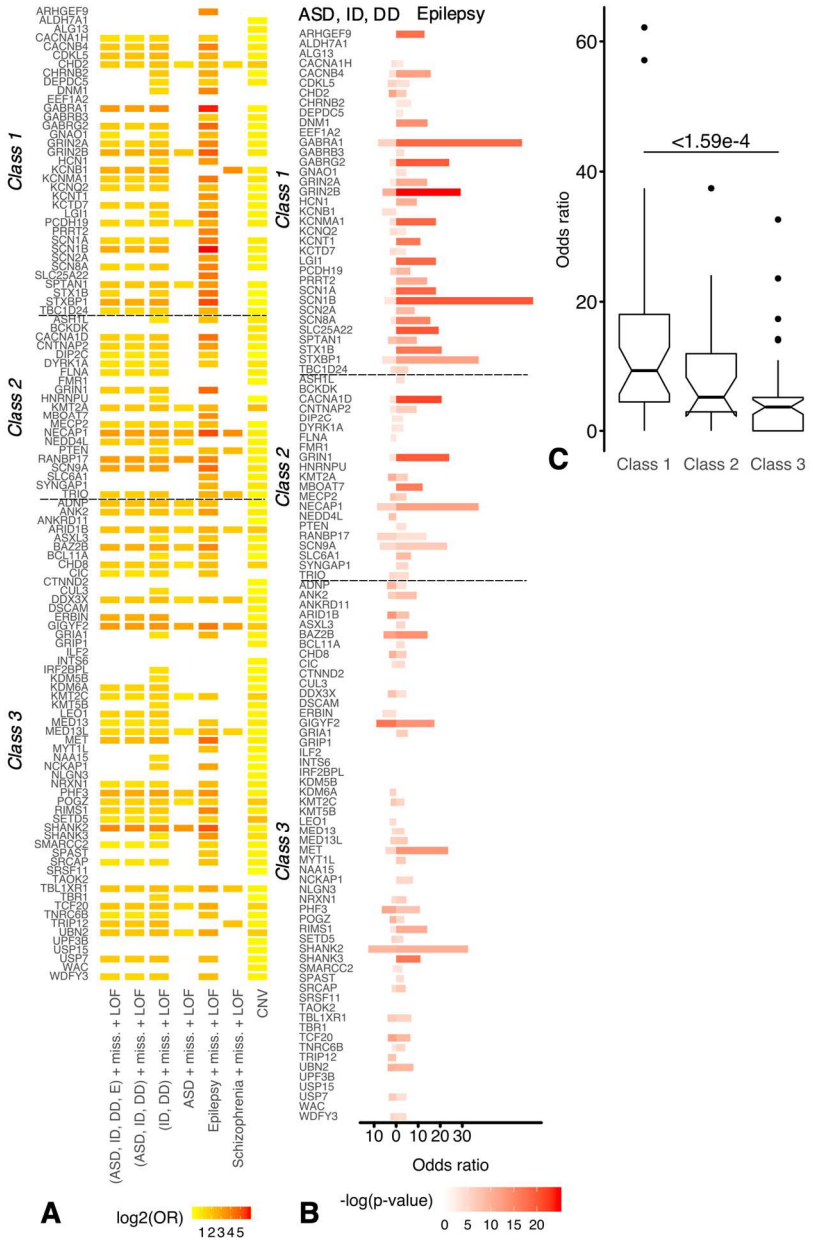


Figure S4. Summary of significant enrichment in *de novo* mutation and copy number variation (CNV) overlap in neurodevelopmental modules for missense variants with CADD score greater than 15. Modules are grouped by Class to indicate degree of association of the seed gene with the epilepsy phenotype. Class 1, Class 2, and Class 3 modules correspond to seed genes that have strong, moderate, and weak evidence of association with epilepsy, respectively. **A**) Enrichment of missense (miss.) and loss of function (LOF)

mutations for autism spectrum (ASD), intellectual disability (ID), developmental disability (DD), epilepsy (E), and schizophrenia cohorts within modules. **B**) Comparison of log2 of significant ($p<0.05$) enrichment of *de novo* mutation for variants annotated as ASD/ID/DD (left) or epilepsy (right). **C**) Average odds ratio of *de novo* mutations annotated in epilepsy cases relative to controls is significantly greater in Class 1 modules compared to Class 3 modules.

Determining overlap of copy number variation morbidity map and modules

From a previously described copy number variant (CNV) morbidity map (31), we retrieve copy number deletions or duplications that overlap any of the genes within a module to determine if significant enrichment of coding copy number deletion and duplication exists in probands with developmental delay relative to copy number deletions in controls. We construct contingency tables to compare the proportion of coding CNVs in probands with CNVs from controls. To account for CNV burden in probands and controls, we conducted 5,000 permutation tests in which coding CNVs containing genes from the module of interest were randomly assigned to two groups of unequal size, with the size of each group corresponding to the number of coding CNVs in probands and in controls. Within a group, we determined how many CNVs contained genes inside or outside the module and constructed a contingency table. If the *p-value* of this contingency table was less than the initial observed *p-value*, then we increment a 'total score'. We calculate an empirical *p-value* by dividing the total score plus by the number of permutations plus 1. A significant empirical *p-value* indicates that an initial assessment of CNV enrichment as significant is indeed significant.

Assessing phenotypic differences in individuals with mutations within and outside modules

To determine if individuals with *de novo* missense or loss of function mutations within a module have lower IQ and higher Social Responsiveness Scale (SRS) T-scores than individuals with *de novo* mutations in genes outside of the module, we intersect Simons Simplex Collection (SSC) individuals with denovo-db and compare average verbal, non-verbal, and full scale IQ and SRS T-scores via a two-sample t-test (6, 27). To determine if the proportion of 1) male and female individuals or 2) individuals with macrocephaly differs within a module, we conducted Fisher's exact tests for individuals with either missense or loss of function mutations and a phenotype of either autism, developmental disability, intellectual disability, or epilepsy. Macrocephaly scores were retrieved for SSC individuals, and scores > 3 were defined as macrocephalic.

Dissection of epilepsy phenotype by enrichment of epilepsy genes within modules

A gene was considered to have an epilepsy annotation if reported by OMIM or DDG2P to have an annotation of 'epilepsy', 'ataxia', 'seizure', or 'Ohtahara', or reported in EpilepsyGene or Wang et al. 2017 to be an epilepsy gene (3, 4). A gene was considered to have an ASD, ID, or DD annotation if the gene has a SFARI gene score of 1 or 2 (1), or is reported by OMIM or DDG2P (5) to be annotated with any of the following case-insensitive terms: autism, Angelman, fragile, intellect, Rett, retardation, Coffin, Bainbridge, CNOT3, Cognitive impairment, Cornelia, CSNK2A1, Developmental, Smith-Kingsmore, Feingold, Floating, GNAI1, Joubert, Kabuki, KBG, KCNQ3, KMT5B, Noonan, Megalencephaly-polymicrogyria-polydactyly-hydrocephalus, Mowat-Wilson, Myhre, Nijmegen, nonspecific severe ID, Opitz-Kaveggia, Phelan, Potocki-Shaffer, Riddle, Rubinstein, Temple-Barraister, Temple Barraister, Weaver, Wiedemann-Steiner, Woodhouse-Sakati, Tatton-Brown-Rahman, Aicardi-Goutieres, Au-Kline, CHOPS, CRASH, Dias-Logan, FG

syndrome, Gabriele-de Vries, Helsmoortel-van der, Lopes-Maciel-Rodan, Kleefstra, Koolen-De Vries, Lujan-Fryns, Nicolaides-Baraitser, Pilarowski-Bjornsson, Pitt-Hopkins, Rubinstein-Taybi, Schuurs-Hoeijmakers, Seckel syndrome, Stankiewicz-Isidor, Takenouchi-Kosaki, White-Sutton, Witteveen-Kolk syndrome, You-Hoover-Fong.

Enrichment of NDDs with or without epilepsy was calculated by counting the number of genes within a module annotated with epilepsy or non-epilepsy associated terms with the formula $(M_P/M_{P'}) \,/\, (G_P \,/\, (19{,}986 - G_P))$, where M_P is the number of genes annotated as a certain NDD phenotype inside a module M_P, is the complement, and (G_P is the total number of genes annotated as a certain phenotype. The total number of genes in the human genome (Gencode GRCh38.p12) is 19,986 genes.

As supplemental phenotypic descriptions, the terms 'epilepsy', 'seizure', 'ataxia', 'convulsion', 'autism', 'macrocephaly', 'intellectual', or 'neurodevelopment' were retained from the Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine (http://www.informatics.jax.org/allele). MGD annotations were not considered in finding NDD phenotypic associations. SFARI gene scores ranging from minimal evidence (4) to high confidence (1) and DDG2P and OMIM descriptions are noted for genes within modules (1, 5).

Pathway and ontology enrichment and expression analyses of modules

Separate lists of genes within a module and respective seed genes were provided to Enrichr (http://amp.pharm.mssm.edu/Enrichr/) to produce pathway and GO biological process and Reactome pathway enrichments and OMIM disease annotations (**Figure S5, Figure S6**) (32, 33). Gene lists and the union of gene lists belonging to the same *Class* were provided to the Cell-type Specific Expression Analysis (CSEA), Specific Expression Analyses (SEA), and Tissue Specific Expression Analyses (TSEA) tools to assess selective expression profiles of modules in the human

brain and body (**Figure S7**) (34). To visualize shared pathway and biological processes, we performed UPGMA hierarchical clustering on selected significant terms (p<0.0001) that occurred in at least ten modules and were related to synapses, neurons, neurodevelopment, neurotransmitters, axons, chromatin, the brain, nervous system, potentiation, or signaling pathways.

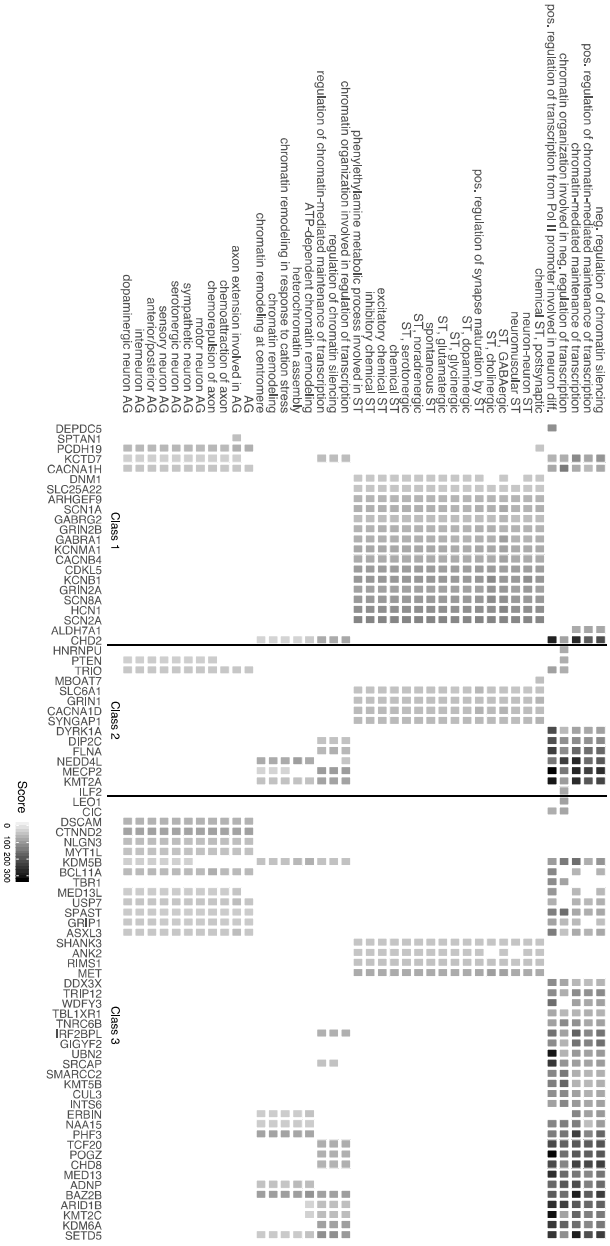


Figure S5. Significant GO Biological Processes. GO terms related to neurodevelopment, synapses, and chromatin organization that are significantly enriched ($p<0.0001$) in at least 10 modules are displayed with combined enrichment scores calculated via Enrichr. Seed genes are grouped as Class 1, Class 2, and Class 3.

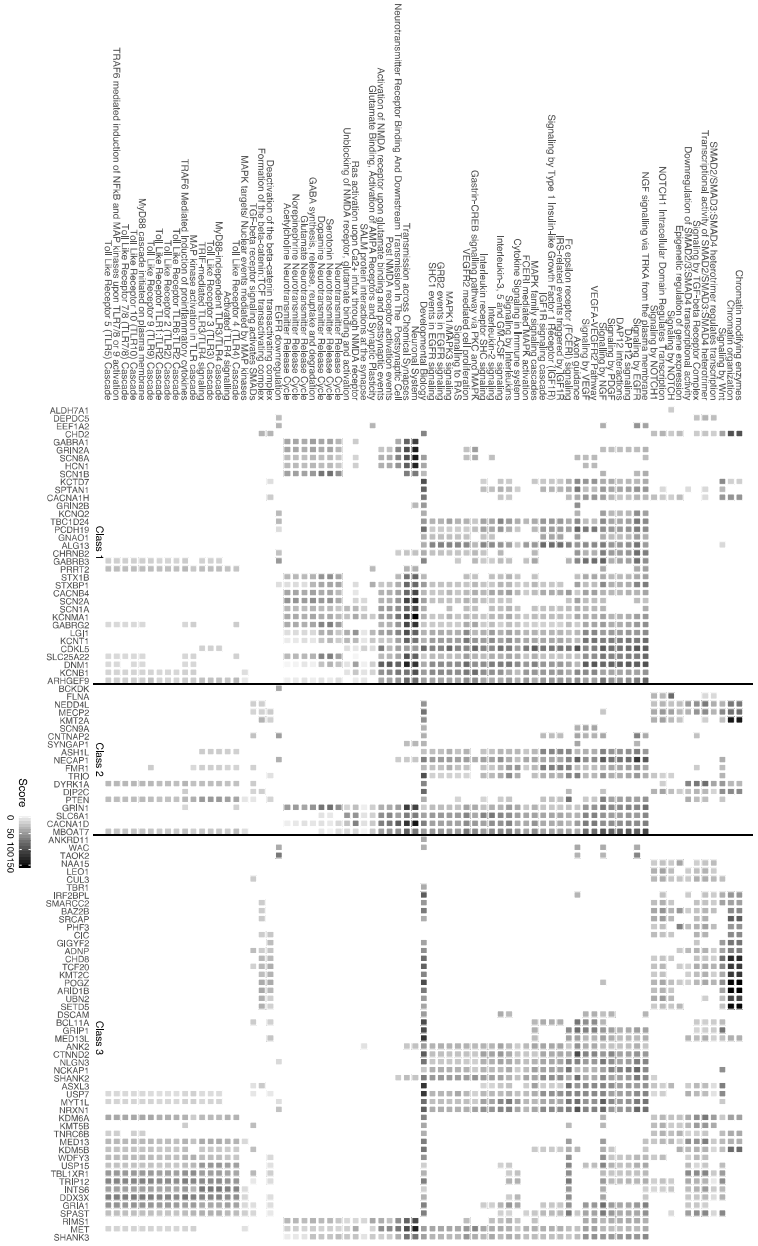
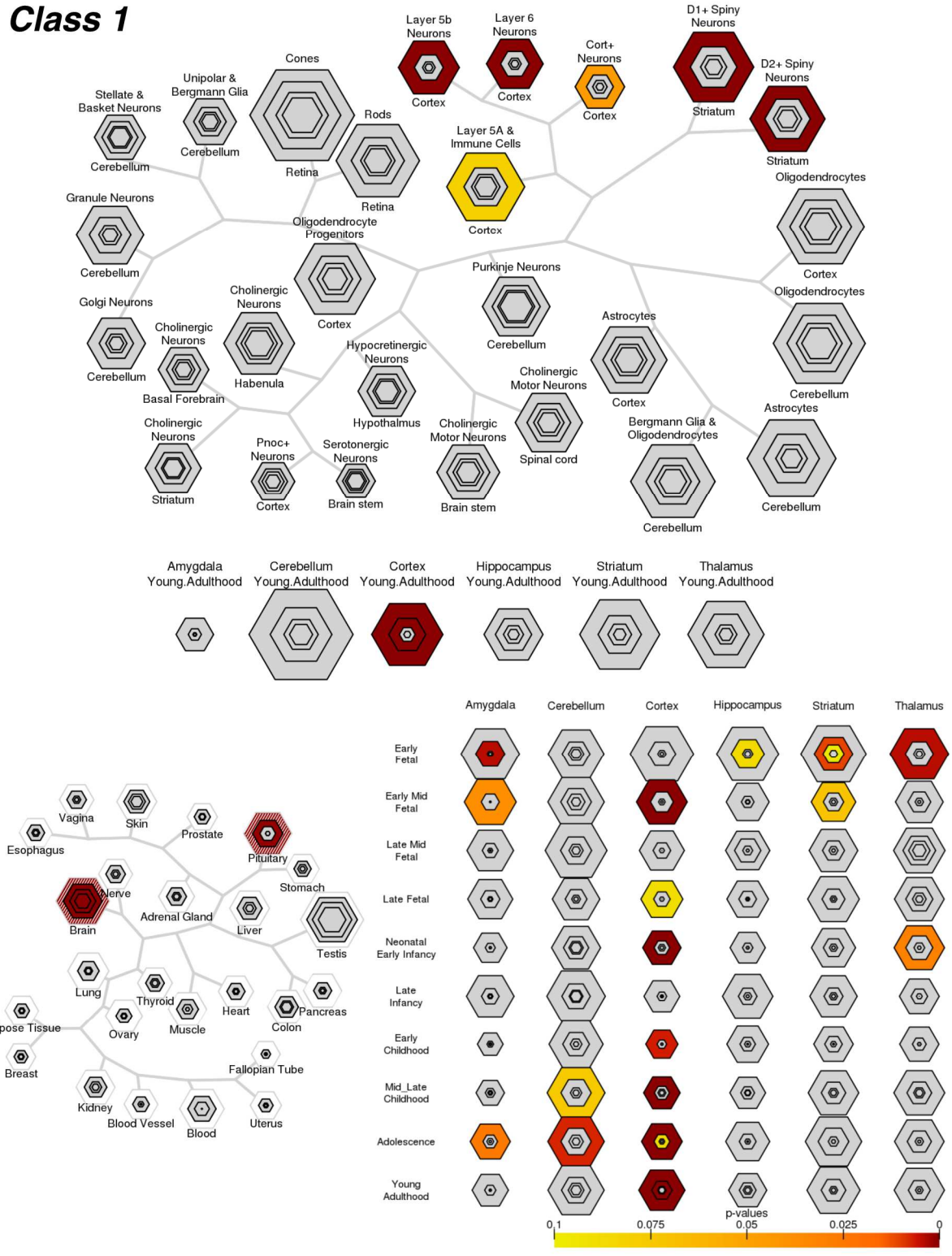
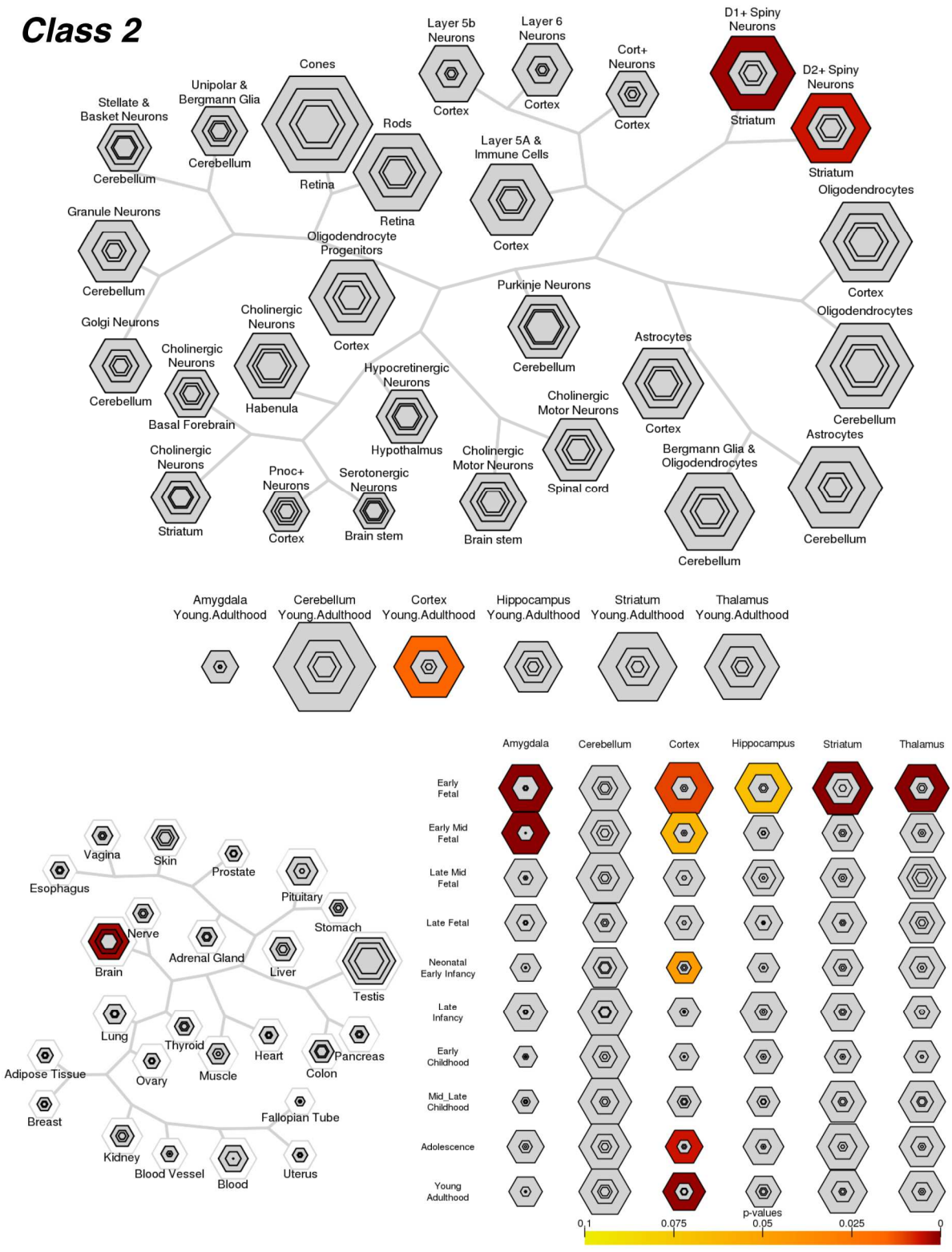


Figure S6. Selected significant ($p<0.0001$) Reactome pathway terms present in at least 10 modules with combined Enrichr score.


Class 1
Layer 5b Neurons
Cortex
Layer 6 Neurons
Cortex
Cort+ Neurons
Cortex
D1+ Spiny Neurons
Striatum
D2+ Spiny Neurons
Striatum
Cones
Retina
Rods
Retina
Unipolar & Bergmann Glia
Cerebellum
Stellate & Basket Neurons
Cerebellum
Layer 5A & Immune Cells
Cortex
Granule Neurons
Cerebellum
Oligodendrocyte Progenitors
Cortex
Oligodendrocytes
Cortex
Purkinje Neurons
Cerebellum
Oligodendrocytes
Cerebellum
Golgi Neurons
Cerebellum
Cholinergic Neurons
Habenula
Cholinergic Neurons
Basal Forebrain
Hypocretinergic Neurons
Hypothalmus
Astrocytes
Cortex
Cholinergic Motor Neurons
Spinal cord
Cholinergic Neurons
Striatum
Pnoc+ Neurons
Cortex
Serotonergic Neurons
Brain stem
Cholinergic Motor Neurons
Brain stem
Bergmann Glia & Oligodendrocytes
Cerebellum
Astrocytes
Cerebellum
Amygdala Young.Adulthood
Cerebellum Young.Adulthood
Cortex Young.Adulthood
Hippocampus Young.Adulthood
Striatum Young.Adulthood
Thalamus Young.Adulthood
Vagina
Skin
Prostate
Esophagus
Pituitary
Stomach
Nerve
Brain
Adrenal Gland
Liver
Testis
Lung
Thyroid
Heart
Pancreas
Adipose Tissue
Ovary
Muscle
Colon
Breast
Fallopian Tube
Kidney
Blood Vessel
Blood
Uterus
Amygdala
Cerebellum
Cortex
Hippocampus
Striatum
Thalamus
Early Fetal
Early Mid Fetal
Late Mid Fetal
Late Fetal
Neonatal Early Infancy
Late Infancy
Early Childhood
Mid_Late Childhood
Adolescence
Young Adulthood
p-values
0.1
0.075
0.05
0.025
0


Class 2
Layer 5b Neurons
Cortex
Layer 6 Neurons
Cortex
Cort+ Neurons
Cortex
D1+ Spiny Neurons
Striatum
D2+ Spiny Neurons
Striatum
Cones
Retina
Rods
Retina
Unipolar & Bergmann Glia
Cerebellum
Stellate & Basket Neurons
Cerebellum
Layer 5A & Immune Cells
Cortex
Oligodendrocytes
Cortex
Granule Neurons
Cerebellum
Oligodendrocyte Progenitors
Cortex
Purkinje Neurons
Cerebellum
Oligodendrocytes
Cerebellum
Cholinergic Neurons
Habenula
Cholinergic Neurons
Basal Forebrain
Golgi Neurons
Cerebellum
Hypocretinergic Neurons
Hypothalmus
Astrocytes
Cortex
Cholinergic Motor Neurons
Spinal cord
Astrocytes
Cerebellum
Cholinergic Neurons
Striatum
Pnoc+ Neurons
Cortex
Serotonergic Neurons
Brain stem
Cholinergic Motor Neurons
Brain stem
Bergmann Glia & Oligodendrocytes
Cerebellum
Amygdala Young.Adulthood
Cerebellum Young.Adulthood
Cortex Young.Adulthood
Hippocampus Young.Adulthood
Striatum Young.Adulthood
Thalamus Young.Adulthood
Vagina
Skin
Prostate
Esophagus
Pituitary
Stomach
Nerve
Brain
Adrenal Gland
Liver
Testis
Lung
Thyroid
Heart
Pancreas
Adipose Tissue
Ovary
Muscle
Colon
Breast
Fallopian Tube
Kidney
Blood Vessel
Blood
Uterus
Amygdala
Cerebellum
Cortex
Hippocampus
Striatum
Thalamus
Early Fetal
Early Mid Fetal
Late Mid Fetal
Late Fetal
Neonatal Early Infancy
Late Infancy
Early Childhood
Mid_Late Childhood
Adolescence
Young Adulthood
p-values
0.1
0.075
0.05
0.025
0

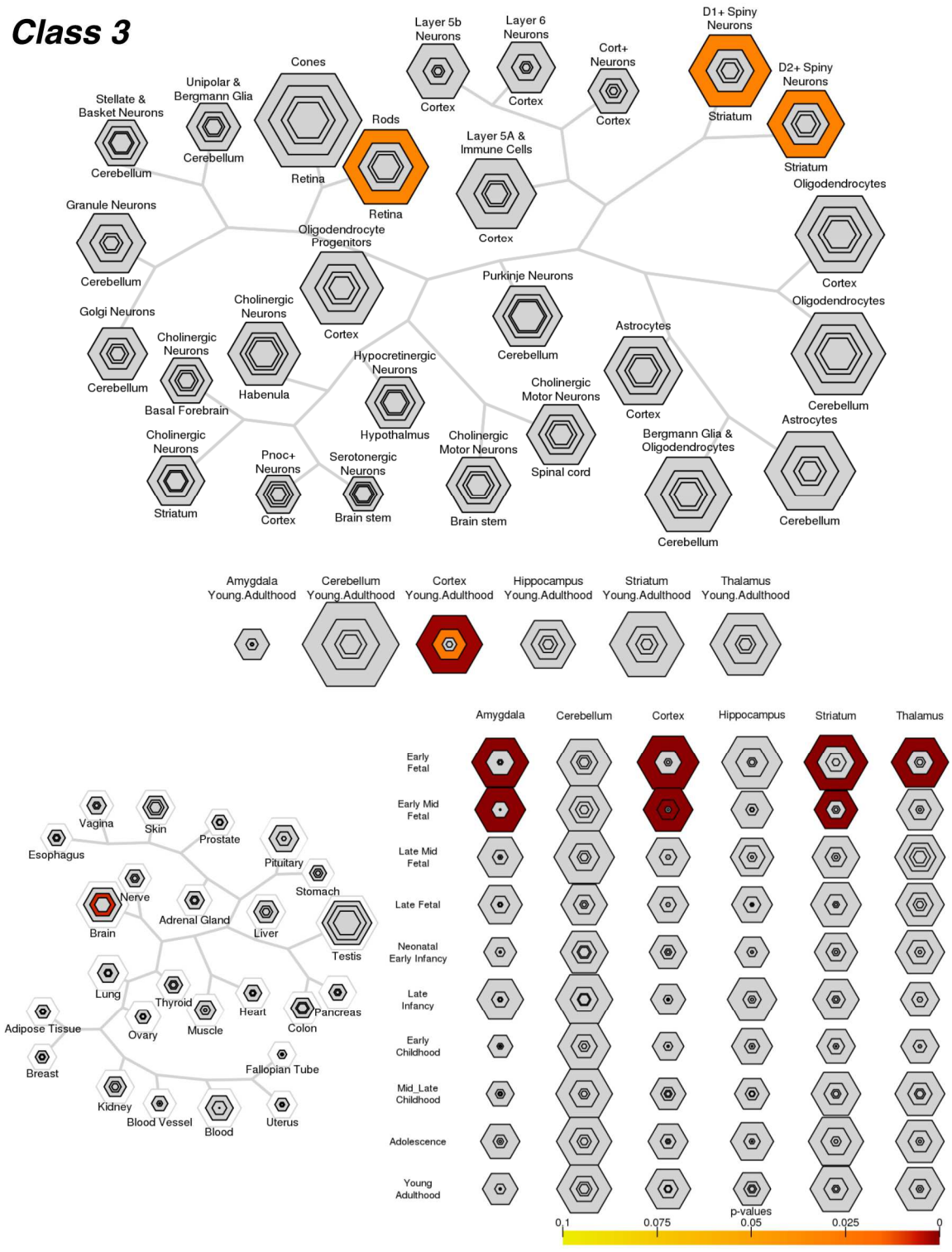

Class 3
Layer 5b Neurons
Cortex
Layer 6 Neurons
Cortex
Cort+ Neurons
Cortex
D1+ Spiny Neurons
Striatum
D2+ Spiny Neurons
Striatum
Cones
Retina
Rods
Retina
Unipolar & Bergmann Glia
Cerebellum
Stellate & Basket Neurons
Cerebellum
Layer 5A & Immune Cells
Cortex
Oligodendrocytes
Cortex
Granule Neurons
Cerebellum
Oligodendrocyte Progenitors
Cortex
Purkinje Neurons
Cerebellum
Oligodendrocytes
Cerebellum
Golgi Neurons
Cerebellum
Cholinergic Neurons
Habenula
Cholinergic Neurons
Basal Forebrain
Hypocretinergic Neurons
Hypothalmus
Astrocytes
Cortex
Cholinergic Motor Neurons
Spinal cord
Cholinergic Neurons
Striatum
Pnoc+ Neurons
Cortex
Serotonergic Neurons
Brain stem
Cholinergic Motor Neurons
Brain stem
Bergmann Glia & Oligodendrocytes
Cerebellum
Astrocytes
Cerebellum
Amygdala Young.Adulthood
Cerebellum Young.Adulthood
Cortex Young.Adulthood
Hippocampus Young.Adulthood
Striatum Young.Adulthood
Thalamus Young.Adulthood
Vagina
Skin
Prostate
Esophagus
Pituitary
Stomach
Nerve
Brain
Adrenal Gland
Liver
Testis
Lung
Thyroid
Heart
Pancreas
Adipose Tissue
Ovary
Muscle
Colon
Breast
Fallopian Tube
Kidney
Blood Vessel
Blood
Uterus
Amygdala
Cerebellum
Cortex
Hippocampus
Striatum
Thalamus
Early Fetal
Early Mid Fetal
Late Mid Fetal
Late Fetal
Neonatal Early Infancy
Late Infancy
Early Childhood
Mid_Late Childhood
Adolescence
Young Adulthood
p-values
0.1
0.075
0.05
0.025
0

Figure S7. Specific expression analyses profiles for Class 1, 2, and 3 modules. Significance of overlap of provided gene lists with transcripts enriched in specific cell-types or tissue types are indicated by intensity of color.

Supplementary Tables

Supplementary Table 1. Neurodevelopmental phenotypes associated with seed genes. Autism (ASD), intellectual disability (ID), and developmental disability (DD) associations are listed according to the SFARI Gene database (gene score of 1 or 2), Online Mendelian Inheritance in Man (OMIM), Developmental Disorders Genotype-Phenotype Database (DDG2P), and literature. Epilepsy phenotypes are retrieved from OMIM, the DDG2P, and literature. Number of genes in modules associated with autism, ID, or DD (G_D), or epilepsy (G_E) and total number of genes in the module (G_T) including seed gene are shown.

	ASD, ID, DD	G_D	Epilepsy	G_E	G_T
Strong epilepsy association: *Class 1*					
ARHGEF9		12	(3-5), OMIM	10	44
ALDH7A1		3	(3-5), OMIM	3	48
ALG13		3	(3-5), OMIM	3	53
CACNA1A			(3-5), OMIM		
CACNA1H	(1)	16	(3-5), OMIM	5	69
CACNB4		12	(3-5), OMIM	13	41
CDKL5	(2)	14	(3-5), OMIM	8	36
CHD2	(1, 2, 9, 30)	25	(3-5), OMIM	7	72
CHRNA2			(3-5), OMIM		
CHRNA4			(3-5), OMIM		
CHRNB2		6	(3-5), OMIM	8	43
DEPDC5		8	(3-5), OMIM	4	50
DNM1	(2)	8	(3-5), OMIM	9	44
EEF1A2	OMIM, (2)	3	(3-5), OMIM	4	32
GABRA1		10	(3-5), OMIM	15	36
GABRB3	(1, 2, 14, 30)	11	(3-5), OMIM	7	63

GABRG2		9	(3-5), OMIM	15	44
GNAO1	OMIM, (2)	10	(3-5), OMIM	10	41
GRIN2A	OMIM	14	(3-5), OMIM	11	40
GRIN2B	OMIM, (1, 2, 5, 9, 14, 30)	13	(3-5), OMIM	10	37
HCN1		13	(3-5), OMIM	17	42
KCNB1		12	(3-5), OMIM	10	39
KCNMA1	OMIM	16	(3-5), OMIM	15	39
KCNQ2	(2)	8	(3-5), OMIM	6	39
KCNT1		7	(3-5), OMIM	12	37
KCTD7		12	(3-5), OMIM	6	67
LGI1		8	(3-5), OMIM	10	41
PCDH19		17	(3-5), OMIM	11	62
PRRT2	(5)	5	(3-5), OMIM	6	40
SCN1A	(2)	12	(3-5), OMIM	14	36
SCN1B		3	(3-5), OMIM	10	26
SCN2A	(1, 2, 5, 14, 30)	15	(3-5), OMIM	10	42
SCN8A	(5)	14	(3-5), OMIM	17	41
SLC1A2			(3-5), OMIM		
SLC25A22		6	(3-5), OMIM	9	42
SPTAN1		17	(3-5), OMIM	7	60
STX1B		7	(3-5), OMIM	12	38
STXBP1	(2)	7	(3-5), OMIM	10	40
SZT2			(3-5), OMIM		
TBC1D24		10	(3-5), OMIM	8	48
WWOX			(3-5), OMIM		
Moderate epilepsy association: *Class 2*					
ASH1L	OMIM, (1, 5, 30)	12	(3)	7	45

BCKDK	(1)	1	(3)	2	37
CACNA1D	(1)	13	(5), OMIM	14	40
CNTNAP2	OMIM, (1)	8	(3, 5), OMIM	7	49
DIP2C	(1)	18	(3)	9	73
DYRK1A	OMIM, (1, 2, 5, 14, 30)	12	(3)	5	56
FLNA	OMIM	8	(3, 5)	3	69
FMR1	OMIM, (5)	7	(5), OMIM	4	50
GRIN1	OMIM	8	(3, 5)	13	45
HNRNPU	(2)	12	(3, 5), OMIM	3	65
KMT2A	(1, 2, 5)	20	(3)	7	66
MBOAT7	OMIM, (1, 5)	7	(5)	10	47
MECP2	(1, 2, 5)	17	(3)	8	80
NECAP1		8	(3, 4), OMIM	7	50
NEDD4L		15	(3)	5	77
PTEN	OMIM, (1, 2, 5, 9, 14, 30)	15	(3)	5	69
RANBP17	(1, 30)	3	(3)	4	53
RELN	(1, 14)		OMIM		
SCN9A	(1)	7	(3, 4), OMIM	5	36
SLC6A1	(1, 2, 30)	7	(4, 5), OMIM	10	36
SYNGAP1	OMIM, (1, 2, 5, 9, 14, 30)	8	(3, 5)	5	37
TRIO	OMIM, (1, 5)	13	(3)	7	70
Weak epilepsy association: *Class 3*					
ADNP	OMIM, (1, 2, 5, 9, 14, 30)	17		5	73
ANK2	(1, 14, 30)	12		8	40
ANKRD11	OMIM, (1, 2, 5)	9		5	39
ARID1B	OMIM, (1, 2, 5, 14, 30)	26		7	70
ASXL3	OMIM, (1, 2, 5, 14)	12		8	67

BAZ2B	(1)	18		7	72
BCL11A	OMIM, (1, 2, 5, 14, 30)	8		4	59
CACNA2D3	(1, 14)				
CHD8	OMIM, (1, 2, 5, 9, 14, 30)	23		9	73
CIC	OMIM, (1)	11		8	36
CNTN4	(1)				
CTNND2	(1)	8		5	50
CUL3	(1, 5, 14, 30)	10		5	63
DDX3X	OMIM, (1, 2, 5)	17		5	70
DEAF1	OMIM, (1, 5)				
DSCAM	(1, 30)	8		4	47
ERBIN	(1)	6		3	54
FOXP1	OMIM, (1, 2, 5, 30)				
GIGYF2	(1, 30)	18		7	61
GRIA1	(1)	10		7	63
GRIP1	(1)	11		4	70
ILF2	(1, 30)	12		3	65
INTS6	(1)	11		3	79
IRF2BPL	(1, 30)	15		6	76
KAT2B	(1, 30)				
KATNAL2	(1, 14, 30)				
KDM5B	(1, 2, 5, 30)	15		3	76
KDM6A	OMIM, (1, 5)	19		3	72
KMT2C	OMIM, (1, 5, 30)	20		6	66
KMT5B	(1, 5)	15		4	76
LEO1	(1)	5		3	79
MAGEL2	(1)				

MED13	(1)	21		7	80
MED13L	OMIM, (1, 2, 5)	15		5	62
MET	(1)	12		11	40
MSNP1AS	(1)				
MYT1L	OMIM, (1, 2, 5, 30)	14		9	61
NAA15	OMIM, (1)	12		4	74
NCKAP1	(1, 30)	6		4	548
NLGN3	OMIM, (1, 5)	8		4	54
NRXN1	OMIM, (1, 5, 30)	16		9	58
PHF3	(1)	18		7	73
POGZ	OMIM, (1, 2, 5, 14, 30)	25		6	70
PTCHD1	OMIM, (1, 5)				
RIMS1	(1)	13		12	36
SETD5	OMIM, (1, 2, 5, 30)	21		7	63
SHANK2	OMIM, (1, 5, 30)	10		7	40
SHANK3	OMIM, (1, 5, 14, 30)	9		11	39
SMARCC2	(1)	17		6	75
SPAST	(1, 30)	15		4	75
SRCAP	OMIM, (1, 5)	18		8	60
SRSF11	(1)	4		4	64
TAOK2	(1)	5		3	40
TBL1XR1	OMIM, (1, 2, 5)	14		6	68
TBR1	(1, 5, 9, 14, 30)	11		3	53
TCF20	(1, 2)	23		9	80
TNRC6B	(1, 30)	12		6	71
TRIP12	OMIM, (1, 5, 9, 30)	16		3	76
UBN2	(1)	17		7	58

UPF3B	OMIM, (1, 5)	4		3	59
USP15	(1)	8		4	66
USP7	(1, 5)	14		4	65
WAC	(1, 2, 5, 30)	5		4	40
WDFY3	(1, 30)	11		6	57

Supplementary Table 2. Summary of analyses performed per module, including determinations of enrichment of *de novo* mutation, overlap with coding copy number variations. Module membership and frequency of occurrence for all genes selected in any module are displayed in the 'modules' tab. Number of cases and controls for ASD, ID, DD, and epilepsy cohorts within denovo-db are displayed in the 'denovo-db' tab. Contingency tables for enrichment of *de novo* mutation and copy number deletions in modules are shown in the 'tables' tab, and contingency table permutation empirical *p-values* are displayed in the 'contingency permutations' tab. Percent contribution to neurodevelopmental disorder diagnoses and comparison of average number of mutations per individual are displayed in the 'enrichment (union)' tab. The tab 'SSC tables' show contingency tables of macrocephaly, IQ, and SRS T-scores values for Simons Simplex Collection individuals. Tabs corresponding to a module name show the total number of *de novo* variants, associated phenotype, type of variant, and neurodevelopmental disorder-related descriptions per module. Similarly, **Supplementary Table 2a** displays a summary of analyses performed per module while requiring a CADD score greater than 15 for missense variants.

Supplementary Table 3. Proportions of synonymous mutations in neurodevelopmental cases relative to controls. Tabs correspond to modules and respective total number of synonymous *de novo* variants.

Supplementary Table 4. Significant GO terms, KEGG and Reactome pathway enrichments, and OMIM disease terms per module (*p-value* < 0.05).

Supplementary Table 5. Selective expression profiles for union of modules based on strength of epilepsy association, including: Cell-type specific Expression Analyses (CSEA), Specific Expression Analyses (SEA) for adult brain regions and development, and Tissue Specific Expression Analyses (TSEA).

References

1. Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Mol Autism. 2013;4.
2. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433-8.
3. Ran X, Li J, Shao Q, Chen H, Lin Z, Sun ZS, et al. EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. Nucleic Acids Res. 2015;43(Database issue):D893-9.
4. Wang J, Lin ZJ, Liu L, Xu HQ, Shi YW, Yi YH, et al. Epilepsy-associated genes. Seizure. 2017;44:11-20.
5. Wright CF, Fitzgerald TW, Jones WD, Clayton S, Mcrae JF, van Kogelenberg M, et al. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet. 2015;385(9975):1305-14.
6. Turner TN, Yi Q, Krumm N, Huddleston J, Hoekzema K, Stessman HAF, et al. NAR Breakthrough Article denovo-db: a compendium of human de novo variants. Nucleic Acids Res. 2017;45(D1):D804-D11.
7. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-U136.
8. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47(6):582-8.
9. O'Roak BJ, Stessman HA, Boyle EA, Witherspoon KT, Martin B, Lee C, et al. Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nat Commun. 2014;5.
10. O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science. 2012;338(6114):1619-22.
11. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017;171(3):710-22 e12.
12. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. Am J Hum Genet. 2016;98(1):58-74.

13. Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet. 2018;50(5):727-36.
14. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209-U119.
15. Yuen RK, Merico D, Cao H, Pellecchia G, Alipanahi B, Thiruvahindrapuram B, et al. Genome-wide characteristics of de novo mutations in autism. NPJ Genom Med. 2016;1:160271-1602710.
16. Yuen RKC, Merico D, Bookman M, Howe JL, Thiruvahindrapuram B, Patel RV, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20(4):602-+.
17. Epi KC, Epilepsy Phenome/Genome P, Allen AS, Berkovic SF, Cossette P, Delanty N, et al. De novo mutations in epileptic encephalopathies. Nature. 2013;501(7466):217-21.
18. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367(20):1921-9.
19. Halvardson J, Zhao JJ, Zaghlool A, Wentzel C, Georgii-Hemming P, Mansson E, et al. Mutations in HECW2 are associated with intellectual disability and epilepsy. J Med Genet. 2016;53(10):697-704.
20. Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet. 2012;380(9854):1674-82.
21. Lelieveld SH, Reijnders MR, Pfundt R, Yntema HG, Kamsteeg EJ, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci. 2016;19(9):1194-6.
22. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, Gormley P, et al. De novo mutations in schizophrenia implicate synaptic networks. Nature. 2014;506(7487):179-84.
23. Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell. 2013;154(3):518-29.
24. Kranz TM, Harroch S, Manor O, Lichtenberg P, Friedlander Y, Seandel M, et al. De novo mutations from sporadic schizophrenia cases highlight important signaling genes in an independent sample. Schizophr Res. 2015;166(1-3):119-24.
25. McCarthy SE, Gillis J, Kramer M, Lihm J, Yoon S, Berstein Y, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. Mol Psychiatr. 2014;19(6):652-8.
26. Smedemark-Margulies N, Brownstein CA, Vargas S, Tembulkar SK, Towne MC, Shi J, et al. A novel de novo mutation in ATP1A3 and childhood-onset schizophrenia. Cold Spring Harb Mol Case Stud. 2016;2(5):a001008.
27. Fischbach GD, Lord C. The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors. Neuron. 2010;68(2):192-5.
28. Collaborative E, Abou-Khalil B, Alldredge B, Bautista J, Berkovic S, Bluvstein J, et al. The epilepsy phenome/genome project. Clin Trials. 2013;10(4):568-86.
29. Mcrae JF, Clayton S, Fitzgerald TW, Kaplanis J, Prigmore E, Rajan D, et al. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433-+.

30. Sanders SJ, Xin H, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron. 2015;87(6):1215-33.
31. Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet. 2014;46(10):1063-71.
32. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016;44(D1):D481-D7.
33. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90-7.
34. Xu XX, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. J Neurosci. 2014;34(4):1420-31.

MAGI-MS: multiple seed-centric module discovery

Julie C Chow, Ryan Zhou, Fereydoun Hormozdiari

Bioinformatics Advances, Volume 2, Issue 1, 2022, https://doi.org/10.1093/bioadv/vbac025

Abstract

Summary

Complex disorders manifest by the interaction of multiple genetic and environmental factors. Through the construction of genetic modules that consist of highly coexpressed genes, it is possible to identify genes that participate in common biological pathways relevant to specific phenotypes. We have previously developed tools MAGI and MAGI-S for genetic module discovery by incorporating coexpression and protein interaction networks. Here, we introduce an extension to MAGI-S, denoted as Merging Affected Genes into Integrated Networks—Multiple Seeds (MAGI-MS), which permits the user to further specify a disease pathway of interest by selecting multiple seed genes likely to function in the same molecular mechanism. By providing MAGI-MS with seed genes involved in processes underlying certain classes of neurodevelopmental disorders, such as epilepsy, we demonstrate that MAGI-MS can reveal modules enriched in genes relevant to chemical synaptic transmission, glutamatergic synapse and other functions associated with the provided seed genes.

Availability and implementation

MAGI-MS is free and available at https://github.com/jchow32/MAGI-MS.

Supplementary information

Supplementary data are available at Bioinformatics Advances online.

Introduction

The extensive genetic and phenotypic heterogeneity characteristic of complex disorders indicates that the interaction of multiple genes underlies etiology (Parenti *et al.*, 2020). The development of protein–protein interaction (PPI) and coexpression networks has aided in the identification of networks of genes hypothesized to belong to the same functional module and contribute to specific pathways (Chen *et al.*, 2020; Parikshak *et al.*, 2015).

Previously, we described a method called MAGI-S used to dissect complex phenotypes, such as epilepsy, by producing modules seeded from a single gene associated with the phenotype of interest (Chow *et al.*, 2019). We demonstrated that independently providing MAGI-S single seed neurodevelopmental disorder (NDD) genes with varying degrees of association with epilepsy revealed modules enriched in (i) non-synonymous coding *de novo* variation in affected NDD cases relative to controls, (ii) genes associated with epilepsy and (iii) *de novo* mutation specifically retrieved from epilepsy cohorts, suggesting that MAGI-S can uncover networks of genes relevant to a complex disorder.

We introduce an extension to the existing method MAGI-S (Chow *et al.*, 2019), referred to as Merging Affected Genes into Integrated Networks—Multiple Seeds (MAGI-MS). MAGI-MS permits the user to select multiple seed genes from which to construct modules, using either the average or minimum coexpression of other genes relative to the selected seeds during gene score assignment. As a result, modules constructed by MAGI-MS are significantly enriched in specific disease pathways in which the provided seed gene(s) participate. In addition, we have normalized gene scoring prior to seed pathway generation such that seed pathways do not preferentially consist of genes that are generally highly expressed. Furthermore, we have simplified the process of

running the compiled MAGI-MS program by providing example commands, sample input files and suggested parameter combinations for ease of use.

Methods

MAGI-MS uses a PPI network, coexpression network, deleterious mutations within a control population and seed gene(s) to create genetic modules that satisfy constraints related to PPI connectivity and degree of coexpression amongst module genes (Supplementary Data). In the following experiments, we use PPIs retrieved from the HPRD and the STRING databases (Keshava Prasad *et al.*, 2009; Szklarczyk *et al.*, 2011), RNA-seq data from the BrainSpan: Atlas of the Developing Human Brain as the coexpression network (Miller *et al.*, 2014) and control variants from the NHLBI Exome Sequencing Project (ESP; http://evs.gs.washington.edu/EVS/; Supplementary Data). Briefly, MAGI-MS assigns a score (Equation 1) to every gene within the PPI network (Fig. 1, Supplementary Data). High-scoring seed pathways are created by the use of a modified color-coding algorithm to find simple paths that maximize the summation of scores associated with genes (Hormozdiari *et al.*, 2015). Seed pathways are then merged into clusters by a random walk, and clusters are improved incrementally by local search to yield top-scoring modules.

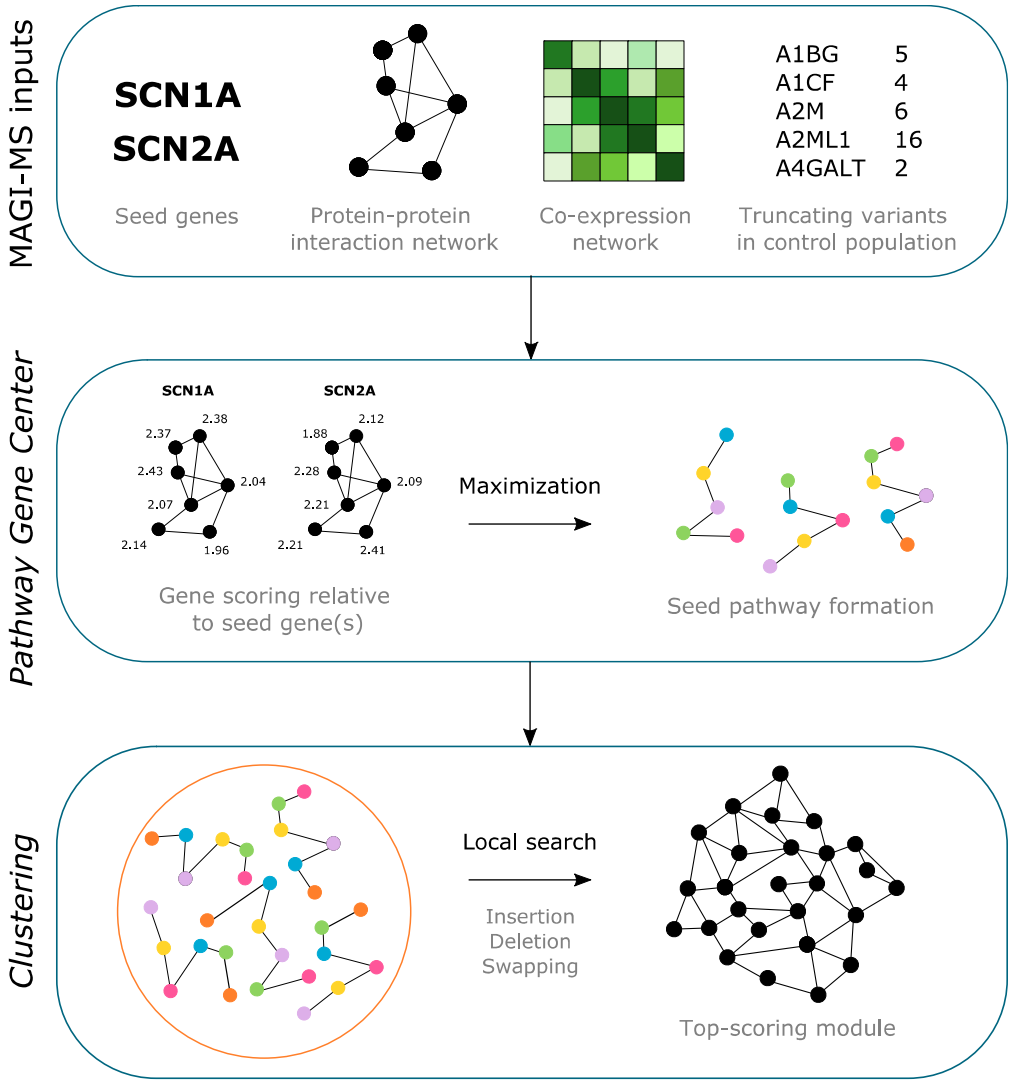


Fig. 1. General methods overview of MAGI-MS. User-selected seed gene(s), a PPI network, a coexpression network and loss-of-function mutations observed in a control population are provided as input to construct modules specific to biological pathways associated with the provided seed genes. During *Pathway Gene Center*, scores are assigned to genes to describe their degree of coexpression with seed gene(s), and seed pathways consisting of high-scoring genes are formed. During *Clustering*, seed pathways are merged and refined to produce candidate modules

To assess the ability of MAGI-MS to dissect a complex phenotype, we provided MAGI-MS with six pairs of seed genes, where each pair consists of genes observed to participate in a similar biological function (Szklarczyk *et al.*, 2021; *CHD8-CREBBP*, *CHD8-CTNNB1*, *GABRA3-GABRB1*, *GRIN2A-GRIN2B*, *SCN1A-SCN2A* and *SHANK2-SHANK3*). We additionally provided MAGI-MS with seed genes that are not hypothesized to participate in the same pathways (*SCN1A-CTNNB1* and *GRIN2A-GRIN2B-ADNP*), randomly selected gene pairs (*BCAS2-SHC1*, *RPL22L1-GEMIN2* and *RPL39L-LRRK2*) and up to 20 genes in the same pathway (long-term potentiation; Supplementary Data). To confirm the presence of relevant functional enrichment and cell-type-specific expression, modules were provided to the tools Enrichr and Cell-type-Specific Expression Analysis (CSEA) and respective enrichment scores were compared (Kuleshov *et al.*, 2016; Xu *et al.*, 2014); we also compared the functional enrichment of MAGI-MS modules with clusters containing seed genes that were generated via PPI clustering algorithms, including MCODE and CytoCluster applications within Cytoscape (version 3.9.0; Bader and Hogue, 2003; Li *et al.*, 2017; Shannon *et al.*, 2003; Supplementary Data).

The number of seeds needed to achieve maximum enrichment may vary depending on the degree of connectivity amongst seed genes and other genes, the extent of shared genes among other related biological pathways and the number of genes in the targeted pathway. It is possible to systematically prioritize candidate seeds by first providing *Pathway Gene Center* with initial seed gene(s) either arbitrarily or based on prior knowledge of importance. *Pathway Gene Center* scores every gene in the PPI network and returns these scores prior to seed pathway construction, where the highest scoring gene displays the highest degree of connectivity with the previously supplied seed(s). Thus, given a list of genes of interest in the same pathway and by retrieving their gene scores, the user can effectively rank candidate seeds and identify a set of seeds to maximize

relevant enrichment. A script to prioritize candidate seeds is provided at https://github.com/jchow32/MAGI-MS.

Results

Given pairs of seed genes involved in the same biological pathway, MAGI-MS produces modules that have significant overlap with modules seeded from either seed gene alone (Supplementary Tables S1 and S2). On average, 49.5% and 61.4% of the genes in paired modules exist, using either minimum or average coexpression values during gene score assignment, respectively, in either of the singly seeded modules. Modules generated by MAGI-MS (Supplementary Table S1) generally display significantly larger enrichment scores (referred to as 'combined scores') compared to singly seeded modules produced by MAGI-S (Supplementary Table S2). For example, most paired-seed modules display significantly greater combined scores or odds ratios in enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways or Gene Ontology (GO) Biological Processes than at least one of the modules produced by a single constituent seed gene.

If MAGI-MS is successively supplied with multiple seeds known to participate in the same pathway, increasingly large enrichment scores can be observed up to a certain point, after which additional seeds do not yield increased enrichment in the targeted pathway. Supplementary Table S3 compares the combined scores of up to 20 seeds involved in the long-term potentiation KEGG pathway. Even while excluding seed genes from the module during functional enrichment analysis, constructed modules using three to five seed genes yield increased enrichment in the long-term potentiation pathway compared to using fewer seeds.

Compared to PPI clustering algorithms such as MCODE and CytoCluster, MAGI-MS produces modules that are seeded from user-selected genes and are specific to pathways in which seed gene(s) participate, whereas modules derived from PPI clustering methods may not necessarily

contain a user's specific seed genes of interest. For PPI clusters containing any seed gene supplied to MAGI-MS, direct comparison of enrichment terms indicate that MAGI-MS shows significantly greater enrichment scores for KEGG pathways and GO Biological Processes compared to MCODE clusters (Supplementary Table S4). Additional modules generated using recent PPI and coexpression data are supplied in Supplementary Table S5.

Modules with paired seeds related to the epilepsy phenotype (*GABRA3-GABRB1*, *GRIN2A-GRIN2B* and *SCN1A-SCN2A*) were enriched in terms such as long-term potentiation, chemical synaptic transmission, among others and showed selective expression in deep cortical neurons (Supplementary Table S1). For seed gene pairs related to more general NDD and autism phenotypes (*CHD8-CREBBP* and *CHD8-CTNNB1*), we observe an enrichment in chromatin organization and regulation of transcription. For modules constructed with seed genes that do not participate in the same biological function (*SCN1A-CTNNB1*, *BCAS2-SHC1*, *RPL22L1-GEMIN2* and *RPL39L-LRRK2*), a module was not formed due to low-scoring seed pathways. For a combination of seeds that do not all participate in the same pathway (*GRIN2A-GRIN2B-ADNP*), a module is produced due to the sufficient degree of connectivity between seeds in the same pathway; however, decreased enrichment in relevant pathways is observed. For example, compared to the *ADNP*(26.76) or *GRIN2A-GRIN2B* (19.61) module, the overall score of the *GRIN2A-GRIN2B-ADNP* module is reduced to 17.65, and functional enrichment of pathways specific to *GRIN2A-GRIN2B*, such as long-term potentiation and glutamatergic synapse, is reduced or absent (Supplementary Table S1). Pathways previously significantly enriched in the *ADNP* module are also reduced, such as ubiquitin-mediated proteolysis, the transforming growth factor-beta signaling pathway and the Wnt signaling pathway. The choice of multiple seed

genes from pathways with similar biological function is critical to form a module that is useful for the dissection of a specific phenotype.

Conclusion

We present an extension to the existing method MAGI-S, denoted as MAGI-MS, which improves upon MAGI-S by (i) permitting the discovery of genetic modules that are specific to certain biological functions by selection of multiple seed genes involved in a pathway of interest, (ii) normalizing gene score assignment to reduce bias during seed pathway formation and (iii) yielding comparable or increased functional enrichment in relevant biological pathways. MAGI-MS is freely available with updated user guides for parameter and input choices.

References

Bader G.D., Hogue C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4, 2.

Chen S. et al. (2020) De novo missense variants disrupting protein–protein interactions affect risk for autism through gene co-expression and protein networks in neuronal cell types. Mol. Autism, 11, 76.

Chow J. et al. (2019) Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. Genome Med., 11, 65.

Hormozdiari F. et al. (2015) The discovery of integrated gene networks for autism and related disorders. Genome Res., 25, 142–154.

Keshava Prasad T.S. et al. (2009) Human protein reference database—2009 update. Nucleic Acids Res., 37, D767–D772.

Kuleshov M.V. et al. (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res., 44, W90–W97.

Li M. et al. (2017) CytoCluster: a cytoscape plugin for cluster analysis and visualization of biological networks. Int. J. Mol. Sci., 18, 1880.

Miller J.A. et al. (2014) Transcriptional landscape of the prenatal human brain. Nature, 508, 199–206.

Parenti I. et al. (2020) Neurodevelopmental disorders: from genetics to functional pathways. Trends Neurosci., 43, 608–621.

Parikshak N.N. et al. (2015) Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. Nat. Rev. Genet., 16, 441–458.

Shannon P. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504.

Szklarczyk D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res., 39, D561–D568.

Szklarczyk D. et al. (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res., 49, D605–D612.
Xu X. et al. (2014) Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. J. Neurosci., 34, 1420–1431.

Supplementary Data

MAGI-MS takes as input a protein-protein interaction (PPI) network, a co-expression network, loss-of-function mutation from control populations, and multiple user-selected seed gene(s). For direct comparison to modules generated from MAGI-S, the PPI network was retrieved from HPRD and STRING (Keshava Prasad *et al.*, 2009; Szklarczyk *et al.*, 2011), using interactions with confidence scores greater than 700 and experimental scores greater than 400. Normalized RPKM values were retrieved from the BrainSpan: Atlas of the Developing Human Brain (Miller *et al.*, 2014) (V6). Truncated variants from the NHLBI Exome Sequencing Project (ESP) (http://evs.gs.washington.edu/EVS/) were used. More recent PPIs from the STRING database (version 11.5) and co-expression data (BrainSpan: Atlas of the Developing Human Brain V10) were used to construct modules displayed in **Supplementary Table 5**.

Genes within modules constructed by MAGI-MS must satisfy constraints related to 1) degree of connectivity in the PPI network, 2) high pairwise co-expression among module genes, and 3) restriction of the number of deleterious loss-of-function mutations in module genes from a control population as described in the Supplementary Material of MAGI (Hormozdiari *et al.*, 2015).

Pathway Gene Center

To construct modules, first, a score is assigned to every gene within the PPI network denoting its degree of co-expression with the seed gene(s). This score ($G_{s,i}$) (**Equation 1**), as in MAGI-S (Chow *et al.*, 2019) and MAGI-MS, is calculated as follows:

$$G_{s,i} = ((H_1)(H_2)) / N^2 \qquad \textbf{Equation 1}$$

For every gene to be scored (*s*) relative to a seed gene (*i*), the score ($G_{s,i}$) is the product of two values which describe the ranking of co-expression between (*s*) and (*i*), referred to as 'coexpression(*s*, *i*)', relative to all other genes in the PPI network. H_1 is the number of pairwise comparisons for which the [co-expression(*s*, *i*) > co-expression(*i*, another gene in the PPI network)]. H_2 is the number of pairwise comparisons for which the [co-expression(*s*, *i*) > co-expression(*s*, another gene in the PPI network)]. *N* is the total number of genes within the PPI network.

Compared to MAGI-S, MAGI-MS differs in that 1) gene scores are calculated for every gene relative to each seed gene rather than a single seed gene, and 2) gene scores are normalized. For example, if two seed genes are provided, then any particular gene will have two scores, where each score is associated with a different seed gene. For each seed gene (*i*), individual gene scores are z-scored (**Equation 2**), such that the scores for any particular seed gene possess a mean score of 0 with standard deviation of 1.

$$z\_score = (G_{s,i} - \mu_i) / \sigma_i \qquad \textbf{Equation 2}$$

Following z-scoring, a final score for every gene in the PPI network is assigned by either taking the average (-avg) or minimum (-min) score among candidate scores from each seed. A larger score indicates a greater degree of co-expression with the seed genes. By calculating an average score, final gene scores will reflect the average degree of co-expression that the gene

possesses with seed genes. By using a minimum score, final gene scores will reflect the largest degree of co-expression observed with any of the seed genes *i* to *j* that were provided.

After final scores have been assigned to every gene in the PPI network, seed pathways are formed to ensure that modules consist of genes that display a high degree of connectivity. Seed pathways consist of *h* genes, where MAGI-MS seeks to maximize the summation of gene scores within the seed pathway by randomly coloring genes with *h* different colors and finding the colorful path via dynamic programming. The use of a modified color coding algorithm permits MAGI-MS to limit the number of deleterious mutations observed from a control population while finding the colorful path (Hormozdiari *et al.*, 2015; Alon *et al.*, 1995). By simultaneously maximizing the summation of gene scores in seed pathways and limiting the number of deleterious mutations observed in a control population, MAGI-MS thus identifies non-random sets of interacting genes. During *Pathway Gene Center*, a total of 16,000 seed pathways are generated using 1,000 iterations of combinations of number of loss-of-function mutations allowed in the control population (0, 1, 2, 3) and number of genes within seed pathways (h = (5, 6, 7, 8)), written to 16 files (*BestPaths* files).

Clustering

During the clustering process, seed pathways generated from *Pathway Gene Center* are merged into high scoring clusters via a random walk. To improve candidate modules, a local search is performed in which individual genes are removed, added, or swapped and the module score is returned. Modules that both satisfy the mentioned constraints and result in an increased module score following local search are produced. The user may independently run several iterations of

Clustering with varied parameters after a single completed execution of *Pathway Gene Center* to compare multiple candidate modules.

Parameter selection

Parameters may be modified during the *Clustering* phase. The minimum (-l) and maximum (-u) size of the constructed module can be specified. We recommend varying the minimum average co-expression of the module (-avgCoExpr, recommended range: 0.425-0.52) and the minimum PPI density of the modules (-avgDensity, recommended range: 0.085-0.14). For seeds with generally low pairwise co-expression values, -avgCoExpr can be further reduced. The parameter (-i) is simply an integer used for the initialization of a random number generator. In practice, the number of deleterious loss of function mutations allowed in genes in the module (-m = 6), the minimum ratio of seed scores allowed (-a = 0.5), and minimum pairwise co-expression value allowed (-minCoExpr = 0.01) are not varied.

Time complexity

MAGI-MS has two main steps: pathway construction and clustering. The pathway construction runs in $O(2^k)$, where k is the maximum length of pathways generated. We bound k to be O(log h), where h is the size of the input graph. The clustering step is linear to the number of pathways constructed (n). Thus, clustering runtime is O(n). On average, for two seed genes, *Pathway Gene Center* completes in 5.30 hours and *Clustering* completes in 1.97 hours running on Ubuntu 16.04.7 LTS via an AMD Opteron(tm) Processor 6380 (Architecture: x86-64, CPUs: 64, CPU MHz: 1396.406).

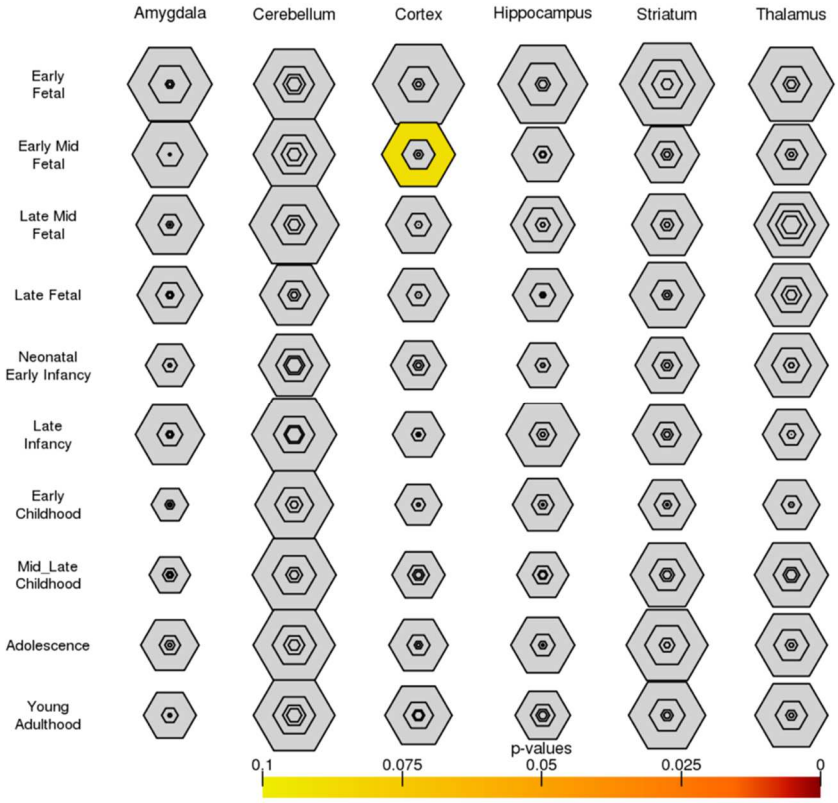
Enrichment analyses among MAGI-MS, MAGI-S, and PPI clustering methods

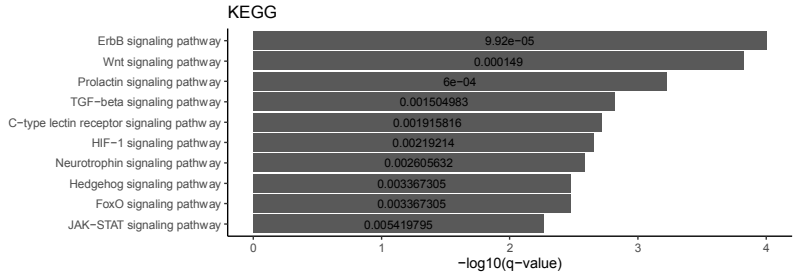
The Cell-type Specific Expression Analysis (CSEA), Specific Expression Analysis (SEA), and Tissue Specific Expression Analysis (TSEA), and Enrichr tools (Xu *et al.*, 2014; Kuleshov *et al.*, 2016) were applied to each of the 6 modules constructed with pairs of seed genes (CHD8-CREBBP, CHD8-CTNNB1, GABRA3-GABRB1, GRIN2A-GRIN2B, SCN1A-SCN2A, and SHANK2-SHANK3). Enriched KEGG, Gene Ontology (GO) Biological Process, and Online Mendelian Inheritance in Man (OMIM) Expanded terms are displayed for each of the 6 modules in **Supplementary Table 1.** The GRIN2A-GRIN2B-ADNP and ADNP modules are also displayed in the GRIN2A-GRIN2B-ADNP tab of **Supplementary Table 1** with corresponding enrichment terms. Modules constructed using single seeds via MAGI-S are shown in **Supplementary Table 2**. In the 'summary' tab of **Supplementary Table 2**, associated p-values resulting from paired t-tests comparing combined scores, odds ratios, and adjusted p-values among shared enrichment terms between a paired seed gene module and corresponding singly-seeded modules are shown. Directional (1-sided) paired t-tests test the hypotheses that 1) the combined score is greater for the paired seed modules than the singly-seeded module, 2) the odds ratio is greater for the paired modules versus the singly-seeded, and 3) the adjusted p-value is smaller for the paired modules versus the singly-seeded. Non-directional (2-sided) t-tests test the hypothesis of equality in shared enrichment terms. In **Supplementary Table 4**, the 'summary' tab similarly indicates significance of directional and non-directional paired t-tests of paired seed modules with clusters resulting from the MCODE and CytoCluster (HC-PIN) PPI clustering methods within the Cytoscape program (version 3.9.0) (Shannon *et al.*, 2003; Bader and Hogue, 2003; Li *et al.*, 2017). Default parameters were used during PPI clustering (MCODE: degree cutoff=2, haircut=enabled, node score cutoff=0.2, k-core=2, max. depth=100; CytoCluster HC-PIN: weak=enabled,

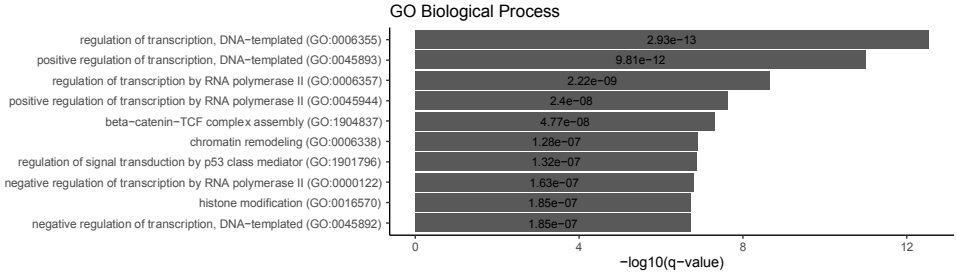
threshold=2.0, complexSize threshold=3). The enrichment scores of a paired module and a PPI cluster were compared if the PPI cluster contained at least one seed gene existing within paired seeds, which include CHD8, CREBBP, CTNNB1, GABRA3, GABRB1, GRIN2A, GRIN2B, SCN1A, SCN2A, and SHANK3. If significant selective expression was detected via CSEA, SEA, and or TSEA for the provided module, respective selective expression plots are displayed in **Supplementary Figure 1**.

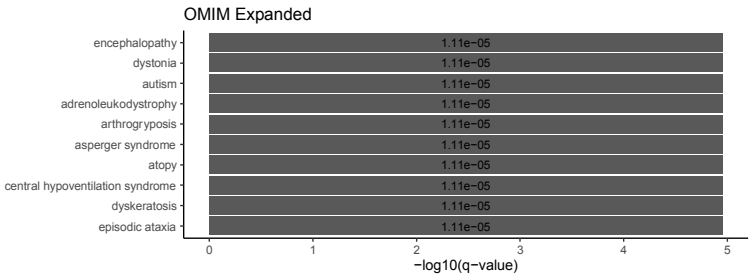
To further compare the enrichment of a specific pathway given multiple seeds that participate in the targeted pathway, up to 20 seeds (**Supplementary Table 3**) in the long-term potentiation KEGG pathway were provided to MAGI-MS. To determine the sequence by which a seed was selected and appended to the list of seeds used as inputs to *Pathway Gene Center*, the following procedure was used. After the user selects one or more seed genes involved in the targeted pathway, additional seeds are prioritized from a list of candidate seeds (all remaining genes in the targeted pathway, or a user-selected list) by calculating the gene scores of candidate seeds as per **Equation 1**. The candidate seed with the largest gene score is then selected as the next seed to append to the previous list of seeds. To select another seed, gene scores are again calculated using the newly appended list of seeds. For example, after providing the seeds GRIN2B-GRIN2A to MAGI-MS, PRKCA possessed the largest gene score among genes in the long-term potentiation pathway and was thus appended as a seed gene (GRIN2B-GRIN2A-PRKCA). Given the seeds GRIN2B-GRIN2A-PRKCA during seed pathway creation, GRIA2 was then identified as the next highest scoring candidate seed. A total of 20 modules seeded via repeated prioritization of genes in the long-term potentiation pathway from the initial seeds GRIN2B-GRIN2A are displayed in **Supplementary Table 3** with associated long-term potentiation pathway enrichment scores.

A) CHD8-CREBBP (-min). SEA: slight enrichment in early mid-fetal cortical tissue.

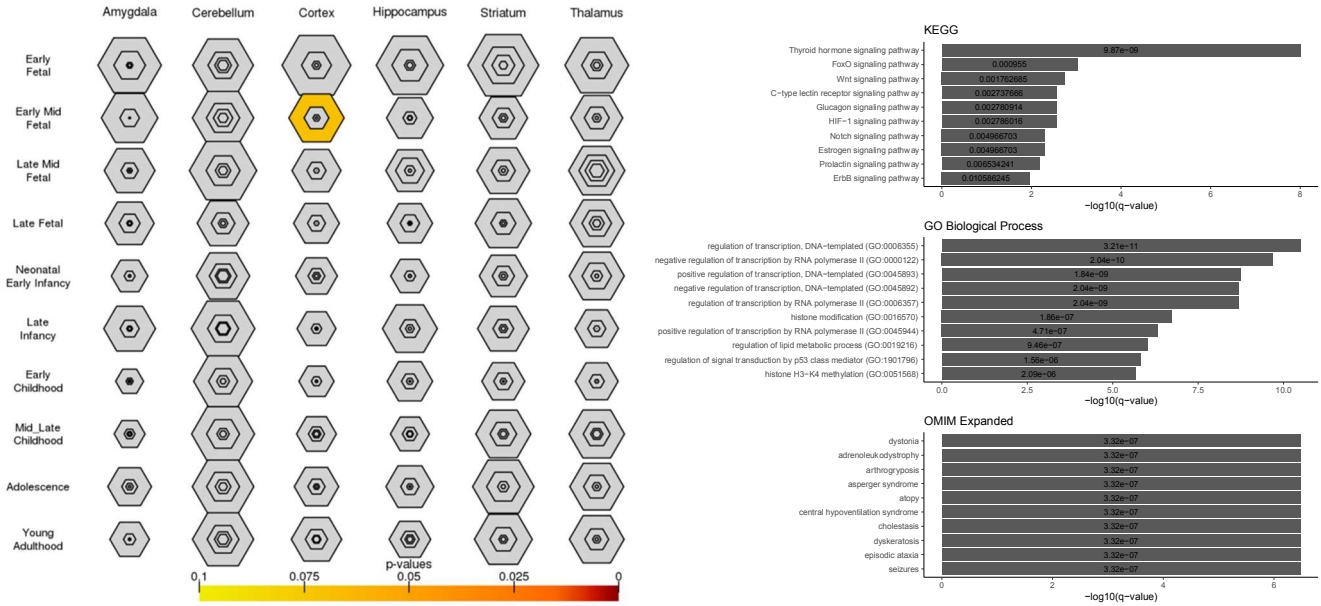


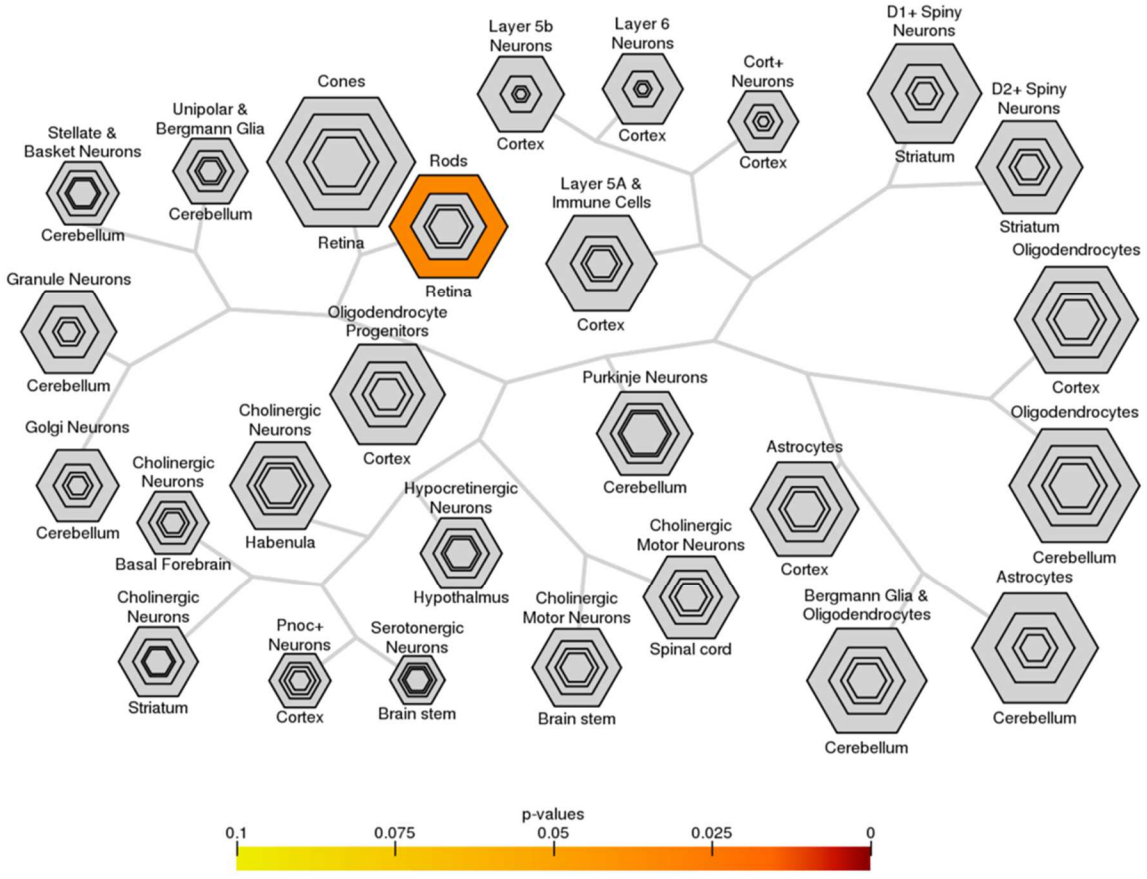




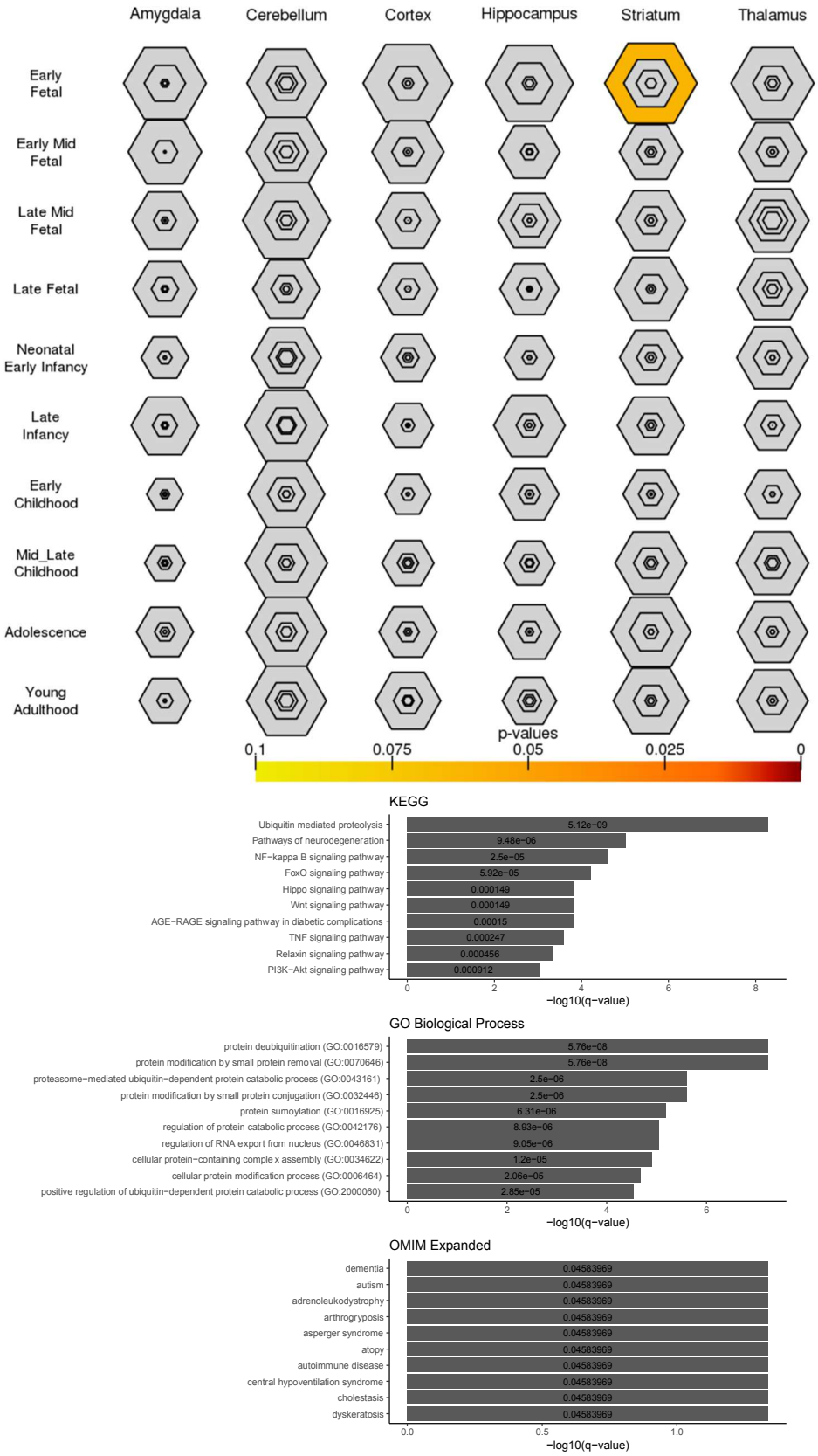


B) CHD8-CREBBP (-avg). SEA: slight enrichment in early mid-fetal cortical tissue. CSEA: enrichment in rods (retina).

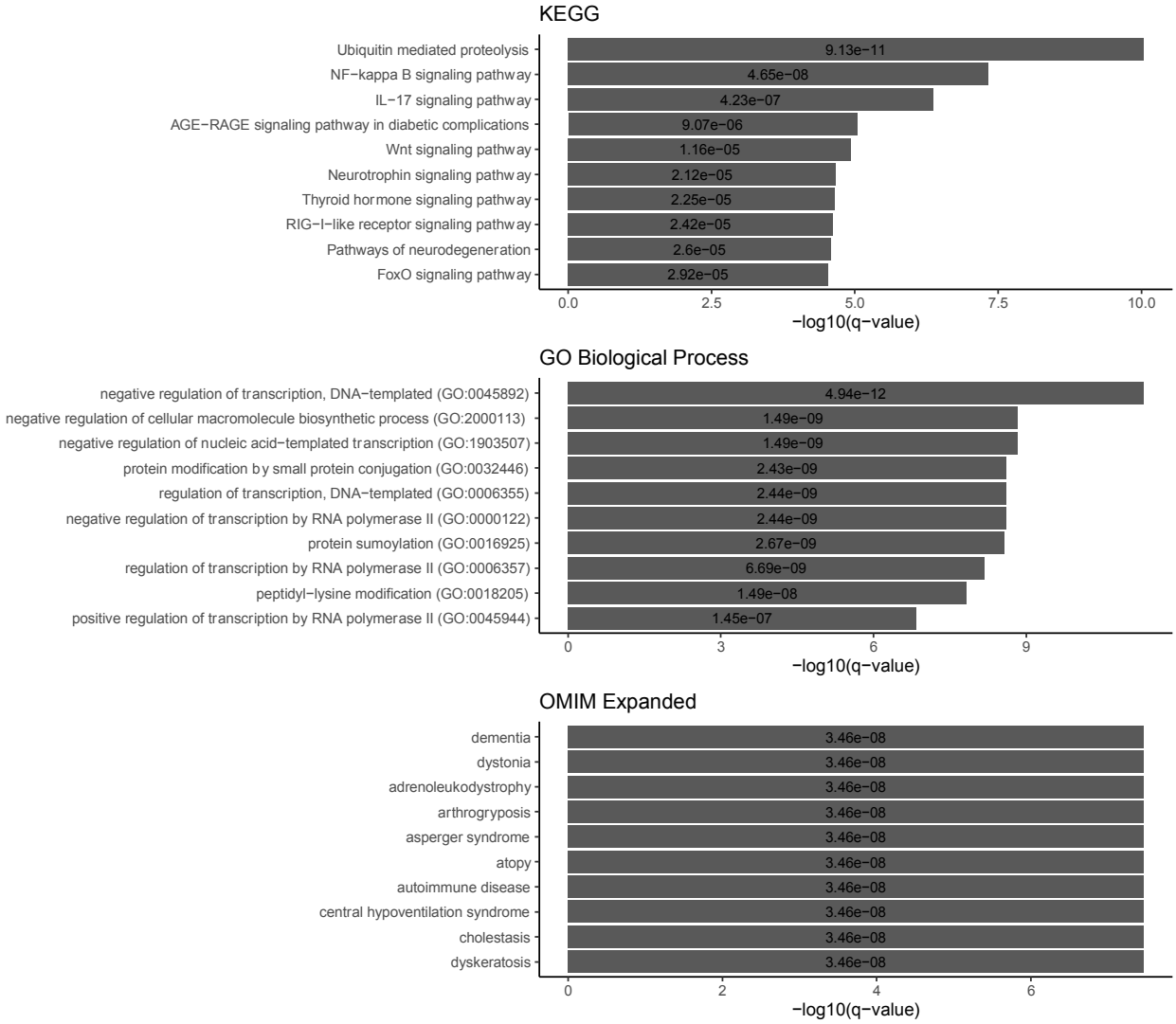




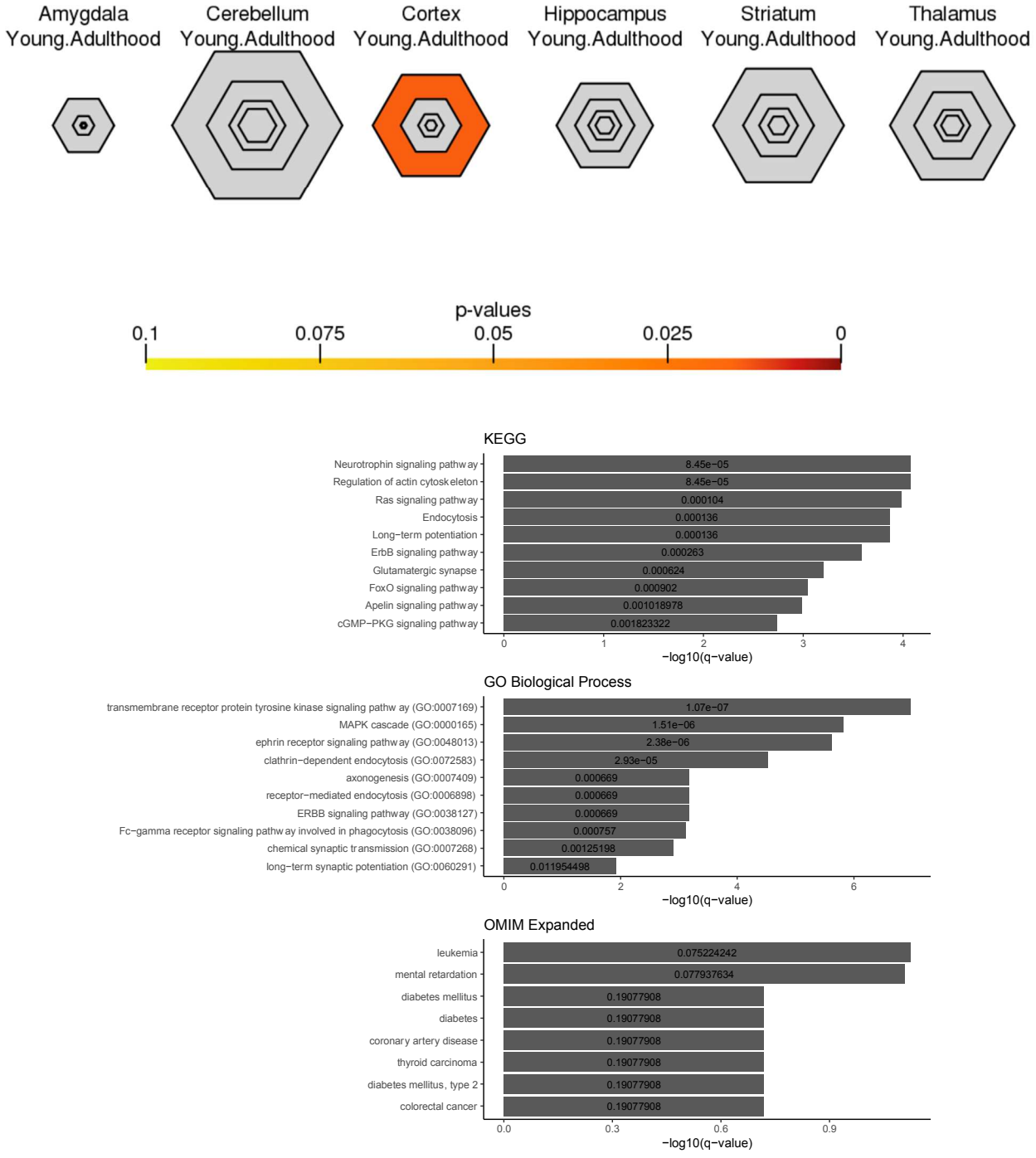
C) CHD8-CTNNB1 (-min). SEA: slight enrichment in early fetal striatum.



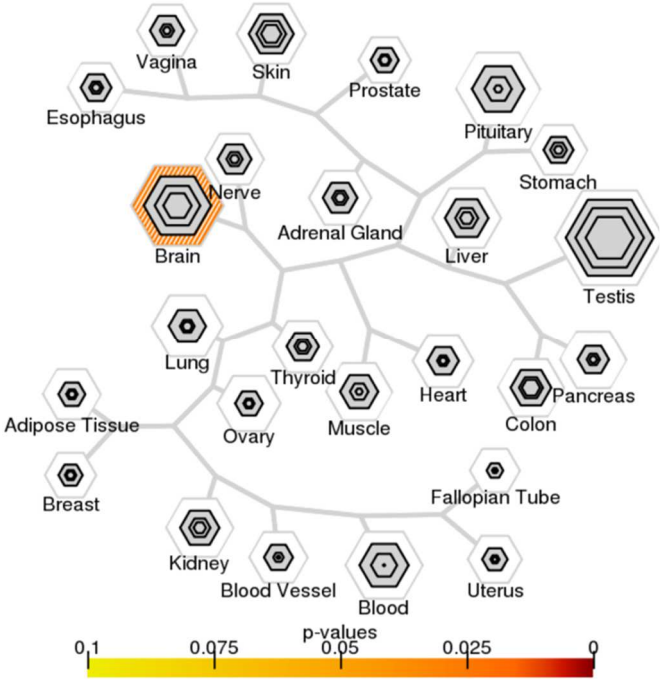
D) CHD8-CTNNB1 (-avg). No significant enrichment via CSEA, SEA, or TSEA tools.

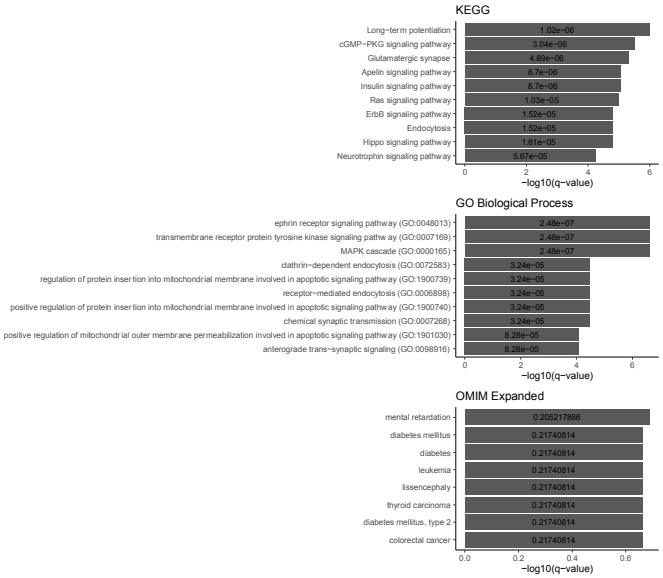


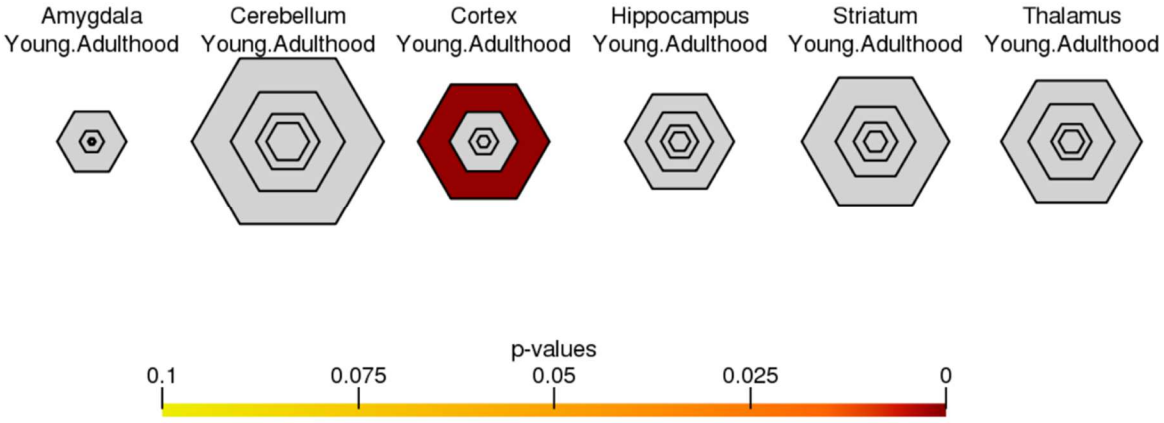
E) GABRA3-GABRB1 (-min). SEA: Enrichment in cortical tissue during young adulthood.



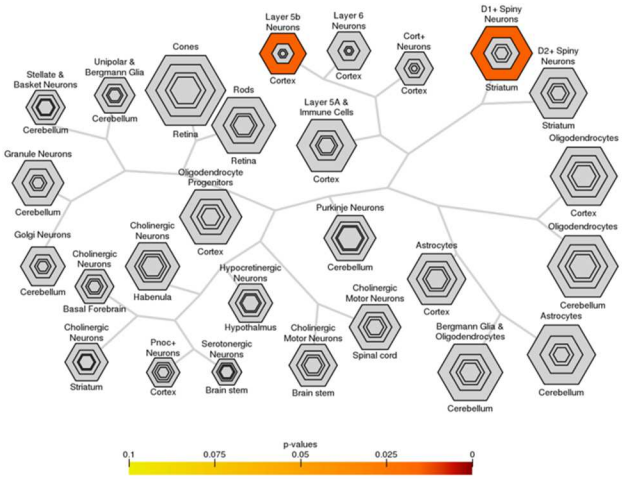
F) GABRA3-GABRB1 (-avg). TSEA: enrichment in brain. SEA: enrichment in cortical tissue during young adulthood.

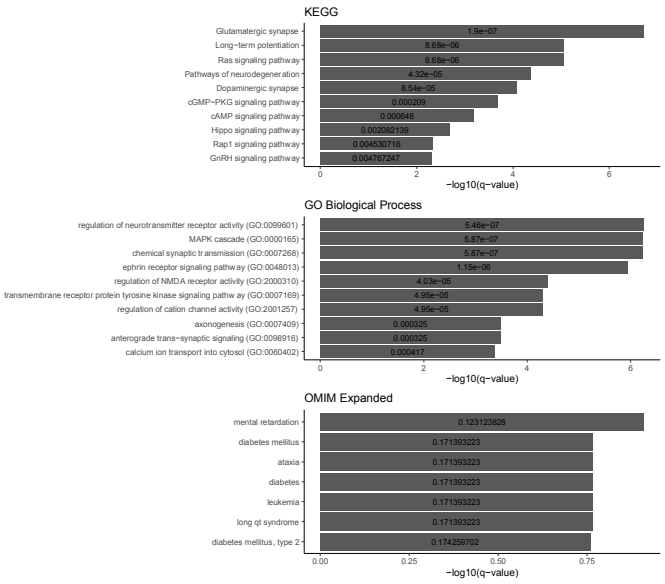


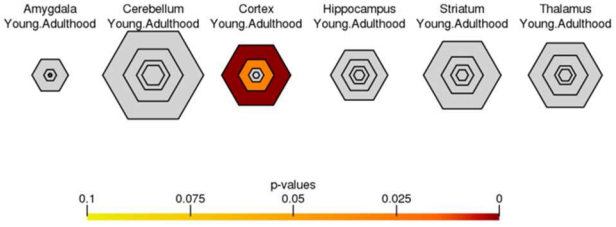


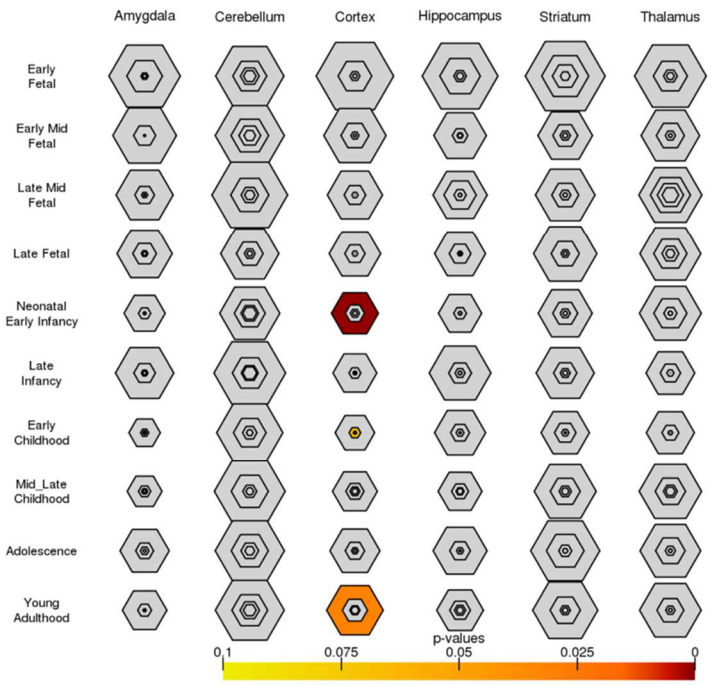


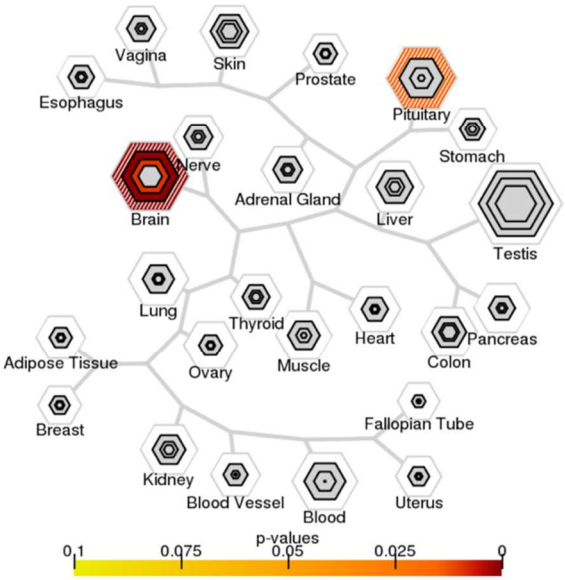
G) GRIN2A-GRIN2B (-min). CSEA: enrichment in layer 5b cortical neurons and D1+ spiny striatal neurons. SEA: increased enrichment in cortical tissues during young adulthood and neonatal early infancy. TSEA: increased enrichment in brain and pituitary gland.



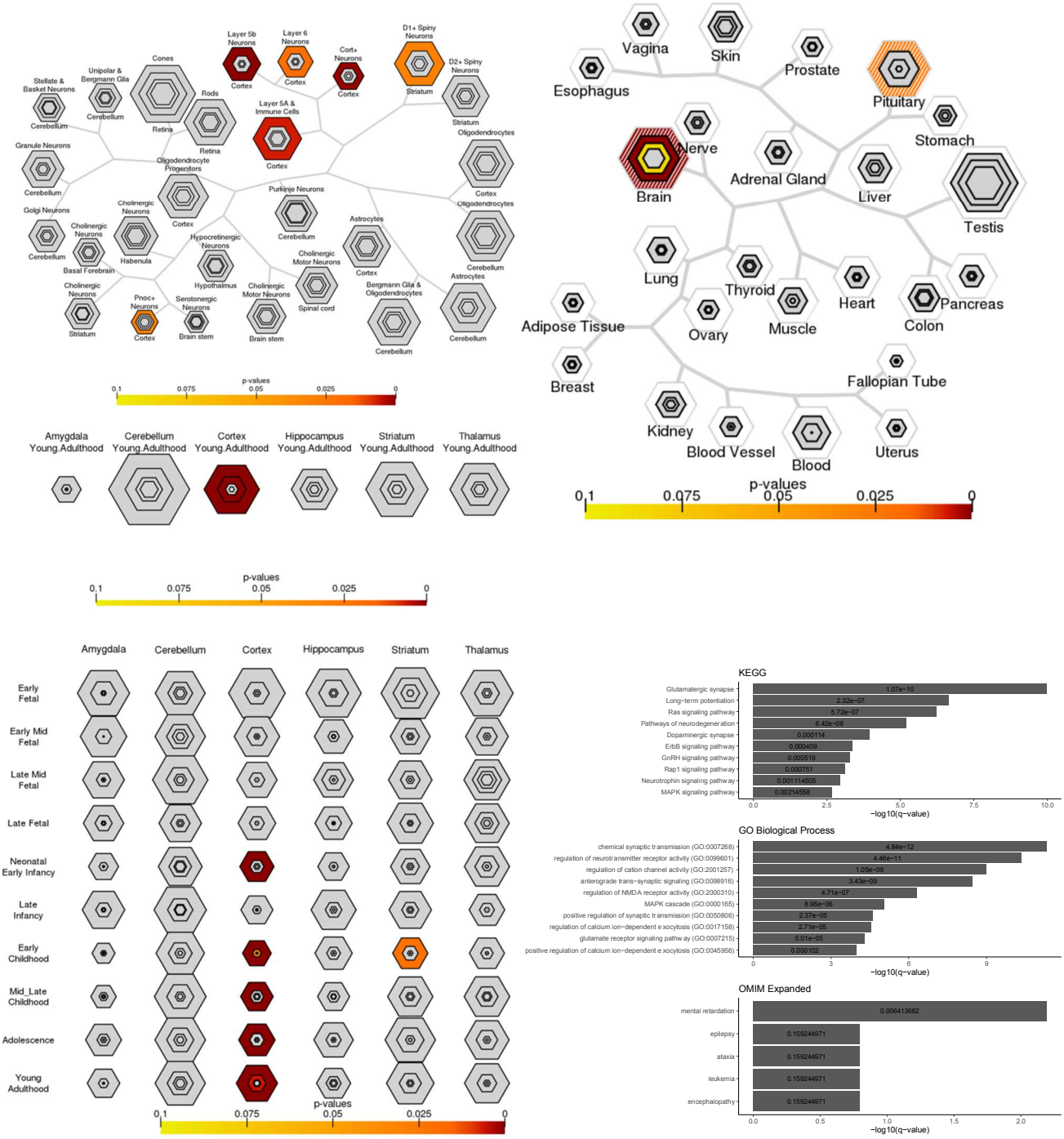




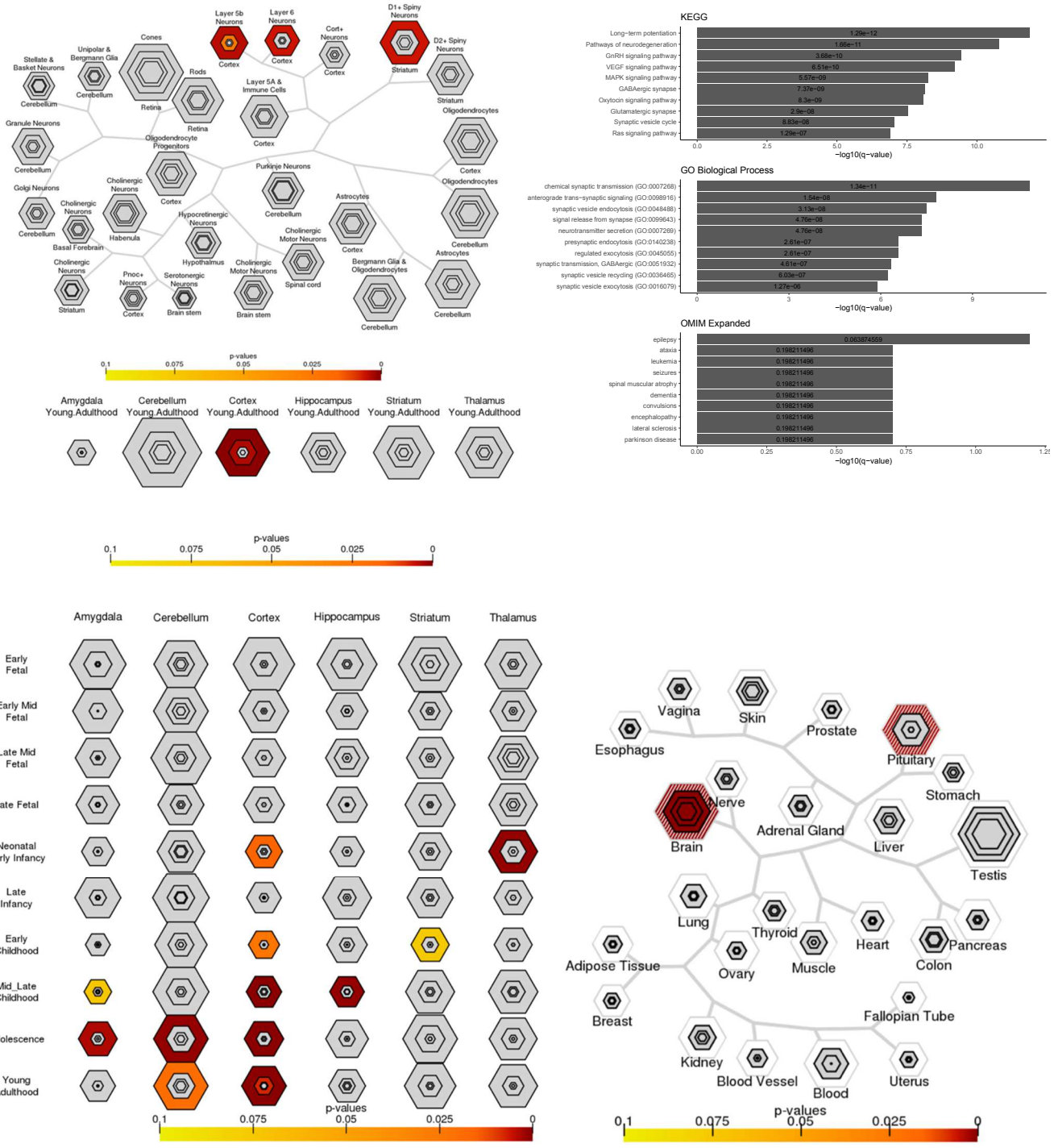




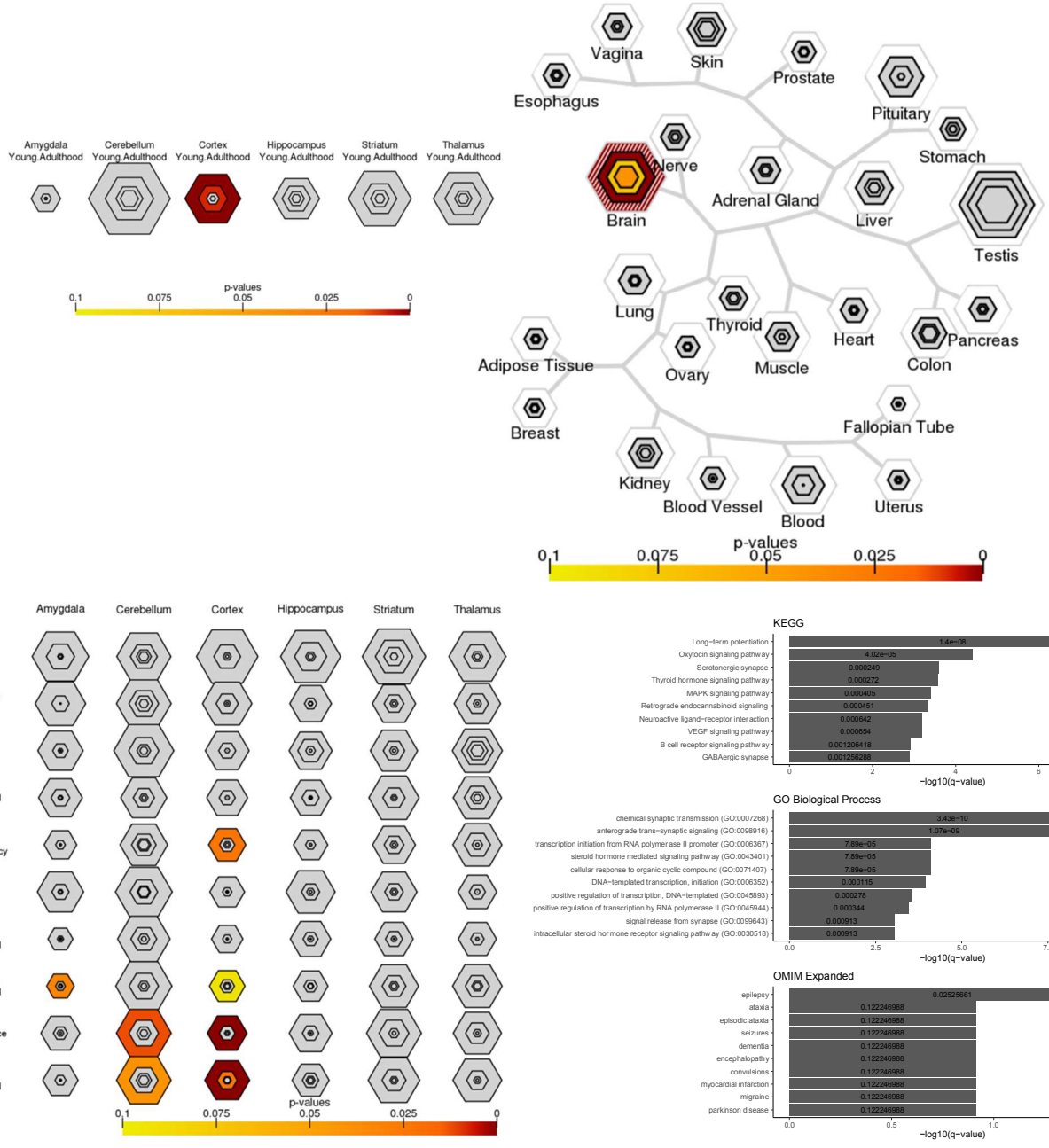
H) GRIN2A-GRIN2B (-avg). CSEA: enrichment in deep cortical neurons and D1+ spiny striatal neurons. SEA: increased enrichment in cortical tissues during neonatal early infancy and from early childhood to young adulthood. Enrichment in striatal tissue during early childhood. TSEA: increased enrichment in brain and pituitary gland.



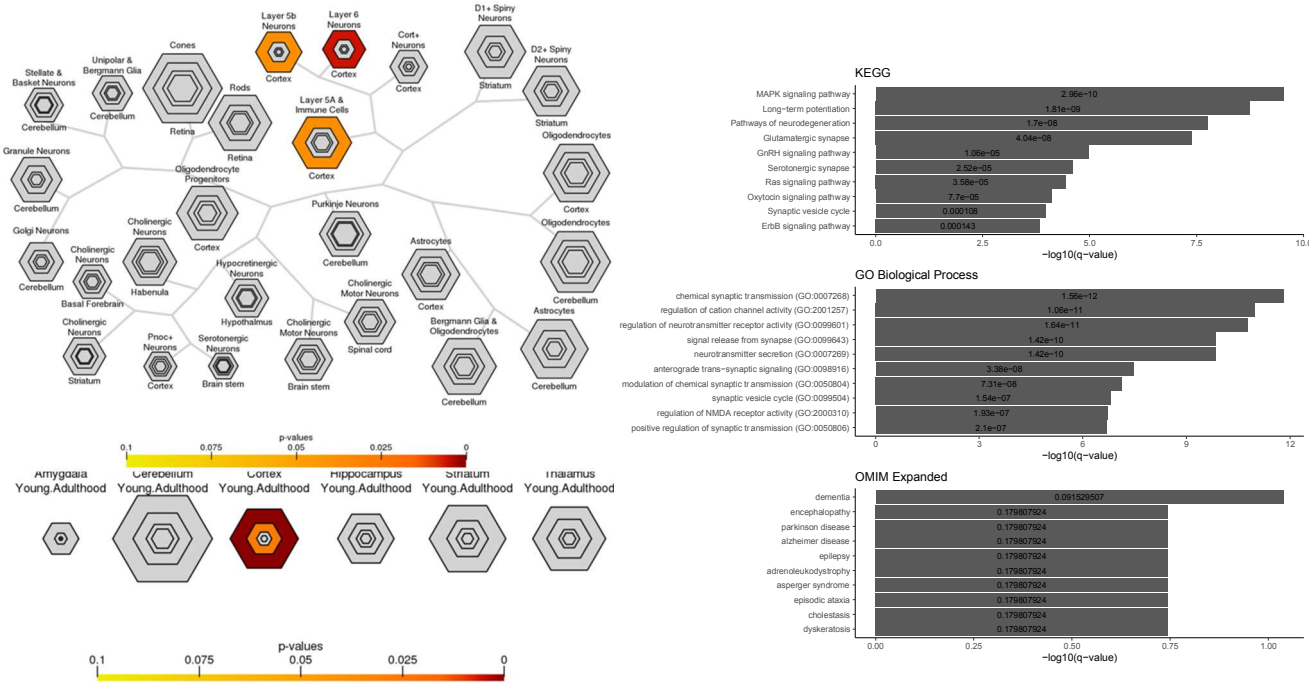
I) **SCN1A-SCN2A (-min).** CSEA: enrichment in layer 5b and 6 cortical neurons and D1+ spiny striatal neurons. SEA: widespread enrichment in the amygdala, cerebellum, cortex, hippocampus, and thalamus from neonatal early infancy to young adulthood. TSEA: increased enrichment in brain and pituitary gland.

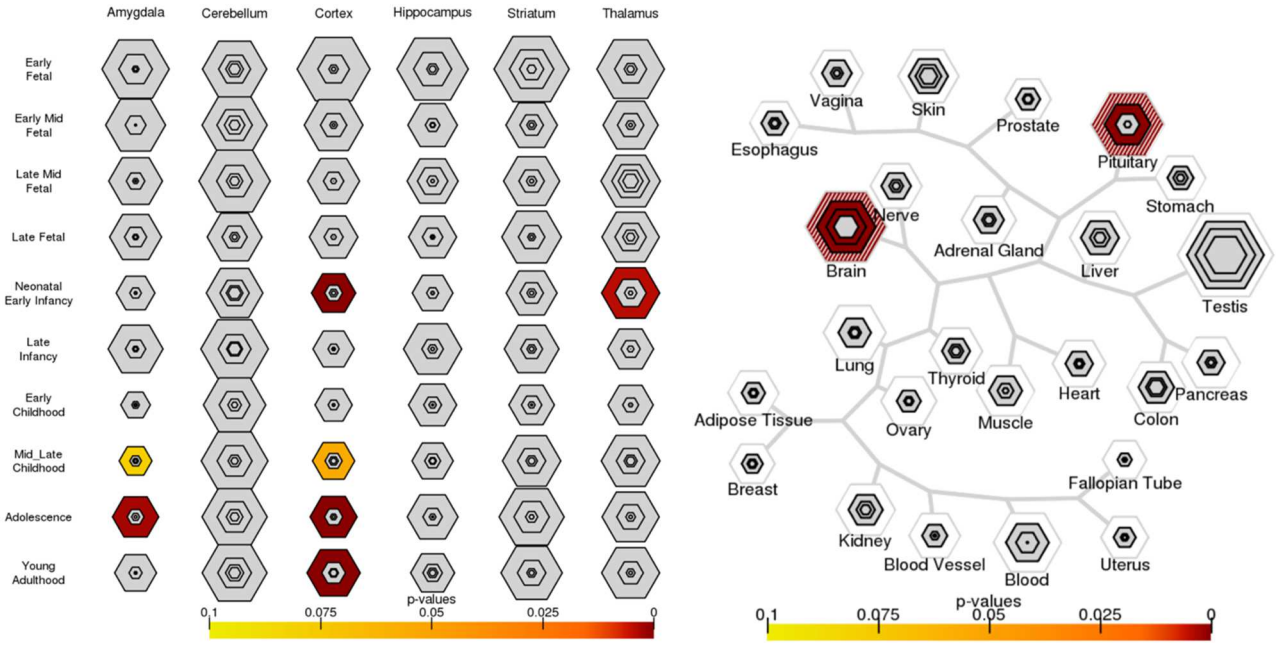


J) SCN1A-SCN2A (-avg). SEA: increased enrichment in cortical tissue during young adulthood, and widespread enrichment in the amygdala, cerebellum, cortex, from mid-late childhood to young adulthood. TSEA: increased enrichment in brain.

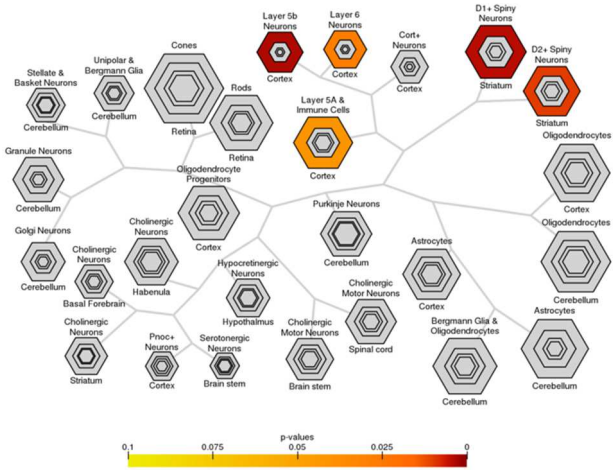


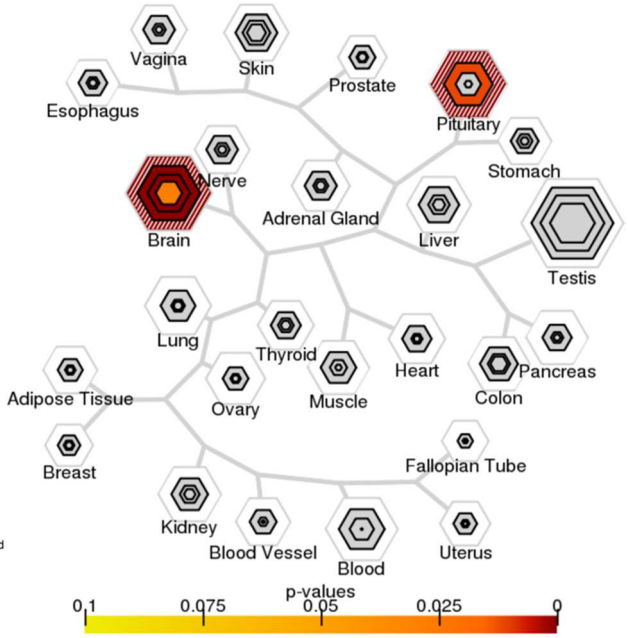
K) SHANK2-SHANK3 (-min). CSEA: enrichment in layer 5b, 5a, and 6 cortical neurons. SEA: enrichment in the amygdala, cortex, and thalamus from neonatal early infancy to young adulthood. TSEA: increased enrichment in brain and pituitary gland.

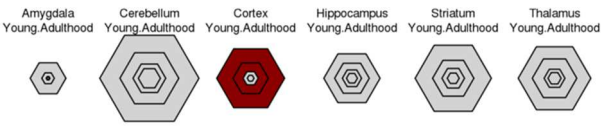


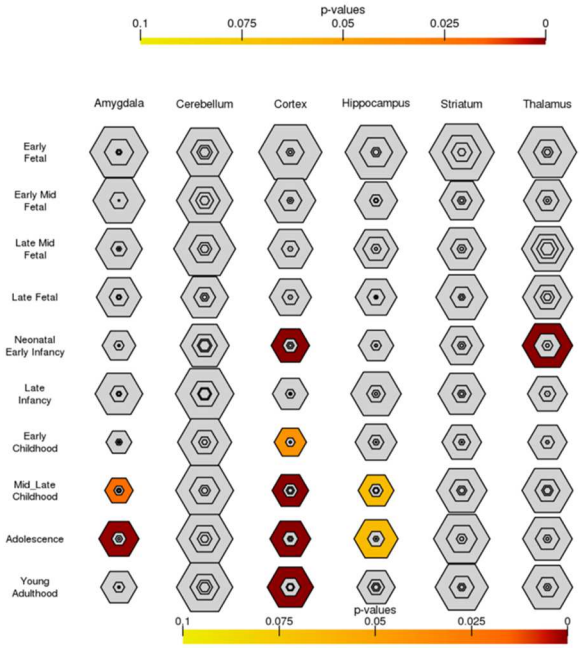


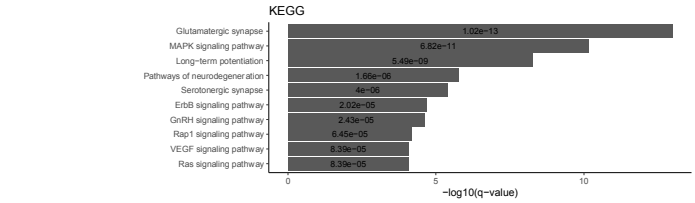
L) **SHANK2-SHANK3 (-avg).** CSEA: enrichment in layer 5b, 5a, and 6 cortical neurons and D1+ and D2+ spiny striatal neurons. SEA: enrichment in the amygdala, cortex, hippocampus, and thalamus from neonatal early infancy and early childhood to young adulthood. TSEA: increased enrichment in brain and pituitary gland.

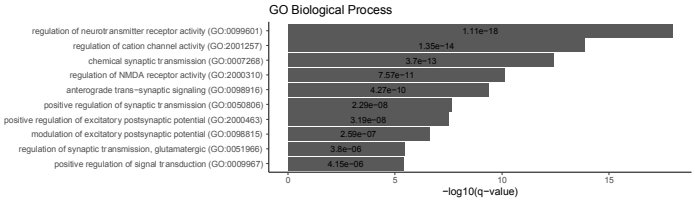


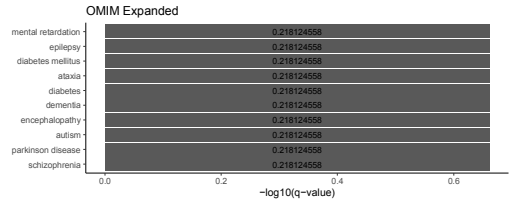












Supplementary Figure 1. Significant cell-type and tissue specific expression analysis for MAGI-MS modules. If significant selective expression is detected via the CSEA, SEA, and TSEA tools for a module generated using either average (-avg) or minimum (-min) co-expression values during score assignment in *Pathway Gene Center*, corresponding figures are shown (-min, -avg): A-B) CHD8-CREBBP, C-D) CHD8-CTNNB1, E-F) GABRA3-GABRB1, G-H) GRIN2A-GRIN2B, I-J) SCN1A-SCN2A, K-L) SHANK2-SHANK3. CSEA and SEA describe the significant enrichment of provided genes across cell-types in the human brain at various developmental time periods. TSEA identifies over-representation of provided (disease) genes with enriched expression in certain human tissues. Significance (p-value) in CSEA, SEA, and TSEA plots is indicated by color intensity, where red indicates p-values close to 0. Top KEGG, GO Biological Processes, and OMIM Expanded enrichment terms are displayed.

Supplementary Table 1. Modules produced via MAGI-MS for selected sets of seed genes. The summary tab displays the number of genes within each paired seed gene module, compared to the number of genes that are shared with respective singly-seed modules created by MAGI-S, using either average (-avg) or minimum (-min) co-expression during gene score assignment. Each paired seed gene tab displays genes within singly-seeded modules and paired modules generated using (-min) or (-avg) parameters, and associated KEGG, GO Biological Processes, and OMIM Expanded enrichment terms retrieved from Enrichr. The GRIN2A-GRIN2B-ADNP tab displays modules seeded using genes that participate in the same pathway (GRIN2A-GRIN2B, ADNP) and associated enrichment terms for the GRIN2A-GRIN2B-ADNP and ADNP modules.

Supplementary Table 2. Modules produced via MAGI-S. The summary tab displays the associated p-values of one-sided and two-sided paired t-test comparisons of Enrichr's Combined Score, odds ratio, and adjusted p-value for KEGG, GO Biological Processes, and OMIM Expanded enrichment terms that are shared between singly-seeded modules created by MAGI-S and paired seed gene modules created by MAGI-MS. Each tab corresponding to a single seed gene displays the associated module and its KEGG, GO Biological Processes, and OMIM Expanded enrichment terms.

Supplementary Table 3. Modules produced via MAGI-MS for up to 20 seeds in the long-term potentiation KEGG pathway. The summary tab displays the seeds in sequential order as they are appended to the list of seeds given to MAGI-MS according to their calculated gene score. Combined scores, odds ratios, and adjusted p-values for modules excluding or including seed genes are shown. Each tab label corresponds to the number of seeds used to generate a module, and each tab displays KEGG enrichment terms for the module while excluding or including seed genes.

Supplementary Table 4. Modules produced via MCODE and CytoCluster that contain seed genes. The summary tab displays associated p-values of one-sided and two-sided paired t-test comparisons of Enrichr's Combined Score, odds ratio, and adjusted p-value for KEGG, GO Biological Processes, and OMIM expanded enrichment terms that are shared between MCODE or CytoCluster clusters that contain seed genes and MAGI-MS modules.

Supplementary Table 5. Modules produced via MAGI-MS for selected pairs of seed genes using recent STRING (version 11.5) and the Atlas of the Developing Human Brain (version 10). The summary tab compares the number of genes, GO Biological Processes, and KEGG terms that are shared or unique among modules generated using older (STRING + HPRD + Allen Brain V6) or more recent (STRING 11.5 + Allen V10) inputs. In general, genes that are shared between modules generated using older and newer inputs constitute the majority of genes in modules generated using newer inputs. Each tab displays associated KEGG, GO Biological Processes, and OMIM Expanded enrichment terms retrieved from Enrichr for every pair of seed genes.

References

Alon,N. *et al.* (1995) Color-coding. *J. ACM*, **42**, 844–856.

Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4, 2.

Chow,J. *et al.* (2019) Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. *Genome Med.*, **11**, 65.

Hormozdiari,F. *et al.* (2015) The discovery of integrated gene networks for autism and related disorders. *Genome Res.*, **25**, 142–154.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kuleshov,M.V. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, **44**, W90–W97.

Li,M. *et al.* (2017) CytoCluster: A Cytoscape Plugin for Cluster Analysis and Visualization of Biological Networks. *Int. J. Mol. Sci.*, **18**, 1880.

Miller,J.A. *et al.* (2014) Transcriptional landscape of the prenatal human brain. *Nature*, **508**, 199–206.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of

proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.
Xu,X. *et al.* (2014) Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. *J. Neurosci.*, **34**, 1420–1431.

Chapter 2

Prediction of neurodevelopmental disorders based on *de novo* coding variation

Julie Chow, Fereydoun Hormozdiari

J Autism Dev Disord (2022). https://doi.org/10.1007/s10803-022-05586-z

Abstract

The early detection of neurodevelopmental disorders (NDDs) can significantly improve patient outcomes. The differential burden of non-synonymous *de novo* mutation among NDD cases and controls indicates that *de novo* coding variation can be used to identify a subset of samples that will likely display an NDD phenotype. Thus, we have developed an approach for the accurate prediction of NDDs with very low false positive rate (FPR) using *de novo* coding variation for a small subset of cases. We use a shallow neural network that integrates *de novo* likely gene-disruptive and missense variants, measures of gene constraint, and conservation information to predict a small subset of NDD cases at very low FPR and prioritizes NDD risk genes for future clinical study.

Introduction

Neurodevelopmental disorders (NDDs), such as autism spectrum disorder (ASD), epilepsy, intellectual disability (ID), and developmental disability (DD) are complex disorders characterized by impairment in cognition, learning, and motor skills. From twin and family studies, it has become apparent that NDDs possess a strong genetic component (Freitag 2007; Gejman et al. 2010). Estimates of heritability for various NDDs have ranged from 0.3 to 0.9, with heritability estimated to be greater than 0.5 for both ASD and ID (Flint 2001; Kaufman et al. 2010; Tick et al. 2016). The evident contribution of genetic factors to NDD diagnoses has provided reason for routine

prenatal whole exome or genome sequencing to identify potentially deleterious genetic variations (Soden et al. 2014; Tărlungeanu and Novarino 2018). In particular, whole exome sequencing has proved a useful tool to identify, at a low cost, coding variants in genes that are highly intolerant to mutation and play important roles in typical neurodevelopment (Srivastava et al. 2019).

The identification and prioritization of NDD risk genes is important for the discovery of underlying biological mechanisms that are perturbed in NDDs (Cardoso et al. 2019). Previous studies have identified many monogenic forms of NDDs and revealed the multifactorial and polygenic nature of most NDD diagnoses (De Felice et al. 2015; Niemi et al. 2018; Sztainberg and Zoghbi 2016). In particular, rare *de novo* mutations that are observed in genes in NDD cases at a significantly higher rate than expected relative to unaffected controls have pinpointed many candidate NDD genes, with more than one thousand genes estimated to be NDD risk genes (De Rubeis et al. 2014; Heyne et al. 2018; Iossifov et al. 2012; Kaplanis et al. 2020; McRae et al. 2017; O'Roak et al. 2012; Sanders et al. 2012; Satterstrom et al. 2020; Wilfert et al. 2017).

De novo mutations are a class of rare genetic variation in which variants, that are not observed in parental genomes, exist in offspring due to mutagenesis in germ cells or errors in replication or recombination (Acuna-Hidalgo et al. 2016). *De novo* mutations may exist as single nucleotide variants, insertions and deletions (indels), and copy number variants. Because *de novo* mutations are not inherited, highly penetrant mutations can arise in genes that are critical to neurodevelopment and likely under purifying selection (Iossifov et al. 2012; Uddin et al. 2014). In fact, individuals affected by NDDs experience a greater burden of non-synonymous *de novo* mutation compared to unaffected controls (Coe et al. 2019; Wilfert et al. 2017). Study of ASD simplex families from the Simons Simplex Collection (SSC) has found that *de novo* likely gene disruptive (LGD) mutations occur at a nearly 2-fold increased rate in affected cases (0.21) relative

to controls (0.12), as well as displaying an increased rate of missense mutation (Iossifov et al. 2014). Furthermore, the study of genetic modules impacted by these *de novo* mutations has pinpointed several biological processes relevant to NDD etiology, such as chromatin remodeling, the Wnt pathway, synaptic transmission, and the long-term potentiation pathway (Chow et al. 2019; Kwan et al. 2016; O'Roak, Vives, Fu, et al. 2012; O'Roak, Vives, Girirajan, et al. 2012; Wilfert et al. 2017).

The benefits associated with successful early prediction of NDDs include the improved ability of parents to make informed decisions about potential early application of treatments (Boivin et al. 2015; Cioni et al. 2016; Corsello 2005). It is important to note that most NDDs cases cannot be predicted using *de novo* coding variation alone; the exome constitutes 1-2% of the human genome and the majority of NDD-associated variants are likely to reside in non-coding regions involved in the regulation of gene expression (Short et al. 2018; Turner and Eichler 2019). Currently, it is estimated that only ~10% of ASD cases and ~20-30% of ID/DD cases have *de novo* LGD variants, and the rate of such variants in the general population is significantly lower (Wang et al. 2021). The genetically and phenotypically heterogeneous nature of NDDs indicates that many factors, including common and non-coding genetic variants and non-genetic factors, account for a large fraction of diagnoses, further complicating our ability for the early prediction of these disorders. However, it is possible to confidently predict a subset of individuals who will likely develop NDDs due to *de novo* coding variation in the form of non-synonymous *de novo* mutations. Despite the polygenic nature of NDDs and the multitude of potential genetic or environmental causes, focusing specifically on un-inherited, *de novo* mutations that disrupt protein coding sequence permits early prediction for a small fraction of cases with very low false positive rates.

The early prediction of NDDs requires a very low false positive rate (FPR) due to potential negative consequences, such as the costs associated with early intervention treatments, that may result from false positive prediction. The minimization of the FPR is clinically most relevant in genetic counseling settings for parents with suspected or confirmed familial risk for NDDs and to aid in the decision to begin early intervention treatments in young children. Early diagnosis of NDDs via a combination of behavioral and motor assessments, imaging, and genetic testing followed by early prediction methods can greatly benefit patient outcomes and lead to timely, appropriate treatment (Hadders-Algra 2021). Previously, a method for the early prediction of complex disorders, Odin, used *de novo* LGD variants observed in cases and controls and co-expression data to identify cases at very low FPR (Huynh and Hormozdiari 2018). The shallow neural net (SNN) with novel objective function introduced here incorporates LGD *de novo* mutation, constraint, and conservation data to achieve a higher (> 0.30129) true positive rate (TPR) at very low FPR (< 0.01) in comparison to traditional classification models such as random forest, support-vector machine (SVM), and logistic regression. Furthermore, the proposed SNN model achieves similar PR-AUC and ROC-AUC to other machine learning approaches. An ensemble model that averages predictions among the SNN, random forest, SVM, and logistic regression models is able to achieve a slightly increased TPR at FPR < 0.01 and comparable PR-AUC. Additionally, the SNN is able to rank genes according to their relative importance in NDDs given LGD or missense *de novo* variation, prioritizing candidate NDD genes.

Methods

Objective*: The main objective is to investigate the potential of using machine learning approaches for early prediction of NDDs using de novo coding genetic variants in a subset of*

cases. More formally, we are interested in utilizing de novo coding variants in maximizing the fraction of affected NDD cases accurately predicted when limiting the false positive rate (FPR) to virtually zero.

Data collection and preprocessing

To distinguish neurodevelopmental disorder (NDD) cases from unaffected controls using *de novo* coding variation, *de novo* likely gene-disruptive (LGD) and missense variants were retrieved from denovo-db (version 1.6.1) (Turner et al. 2017, p.). These data consisted of 9,962 individuals with primary phenotypes of autism spectrum disorder (ASD), intellectual disability, and developmental disability and 2,245 controls, of which 6,509 cases and 1,251 controls possess non-synonymous coding *de novo* mutation (**Supplementary Table 1**). In total, the 7,760 samples possessed 1,974 LGD (cases: 1,715; controls: 259) and 10,777 (cases: 9,073; controls: 1,704) missense *de novo* coding mutations. *PrimateAI* scores were used to quantify the pathogenicity of missense variants, in which position-specific scores were calculated for each missense variant while incorporating conservation, solvent accessibility, and secondary structure data to yield predictions of deleteriousness (Sundaram et al. 2018). Probability of loss-of-function intolerance (pLI) and loss-of-function observed/expected upper bound fraction (LOEUF) scores from the gnomAD browser (v2.1.1), Residual Variation Intolerance (RVIS) scores based on ExAC v2 release 2.0 (March 15, 2017 version), and phastCons element scores were also used as features (Karczewski et al. 2020; Petrovski et al. 2013; Siepel et al. 2005).

LGD-specific and missense-specific feature matrices were generated, in which rows represent individuals with LGD or missense variation from denovo-db and columns represent genes containing *de novo* mutations (**Figure 1A**, **Additional File 1**).

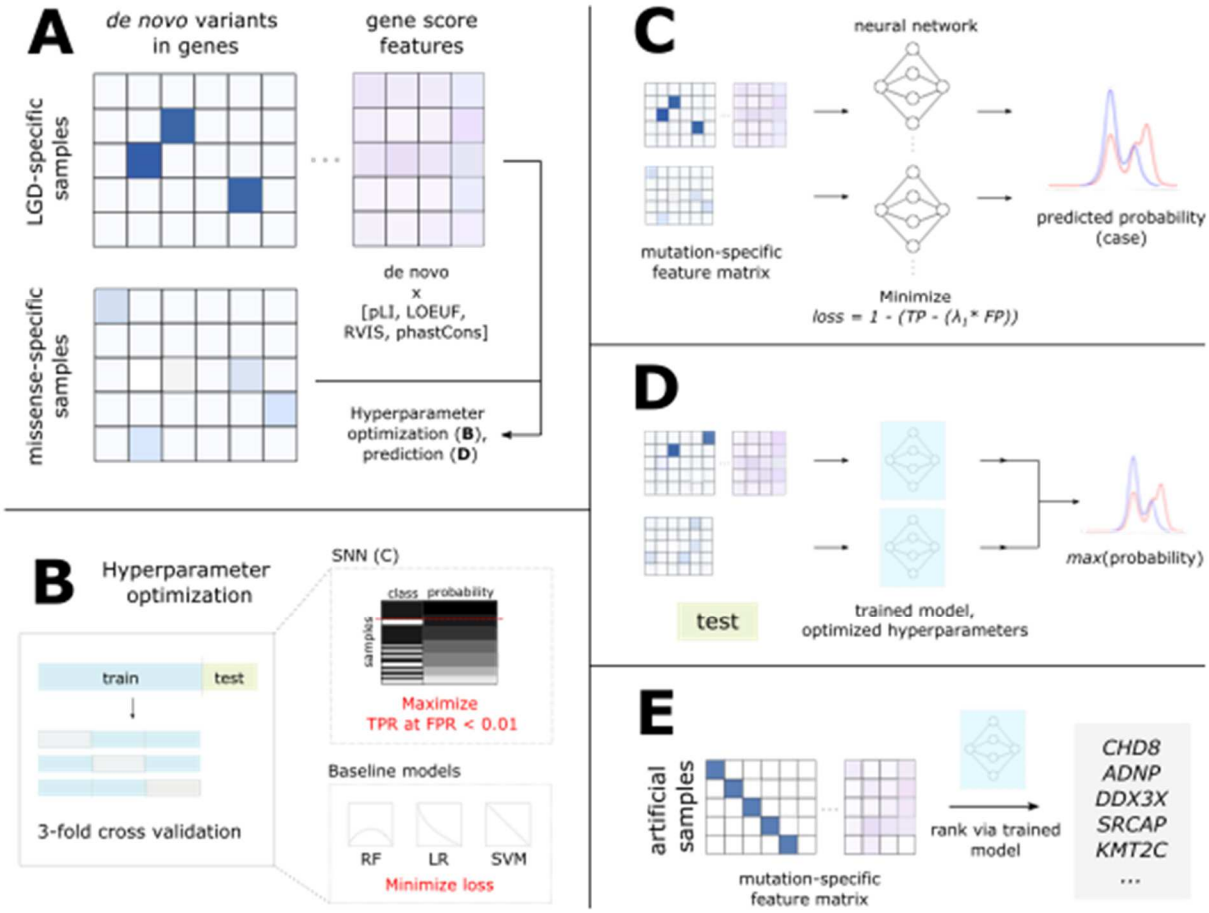


Figure 1. Methods overview. A) *De novo* likely gene-disruptive (LGD) and missense variants from probands with neurodevelopmental disorders and controls were retrieved from denovo-db and arranged into feature matrices. Constraint and conservation information, in the forms of pLI, LOEUF, RVIS, and average phastCons element conservation scores were incorporated as gene score features (Karczewski et al. 2020; Petrovski et al. 2013; Siepel et al. 2005) (**Additional File 1**). B) To perform hyperparameter optimization and model training, samples were divided into training (75%) and testing (25%) sets. Hyperparameter optimization occurs via 3-fold cross validation on the partitioned training set. For the shallow neural net (SNN) model (C), performance is measured as the true positive rate (TPR) at false positive rate (FPR) < 0.01, which is calculated by determining the number of cases (class: black) with predicted probability greater than that of

any control (class: white) in the validation fold. For baseline models, consisting of the random forest (RF), logistic regression (LR), and support-vector machine (SVM) classifiers, respective loss functions are minimized. C) The SNN consists of a single hidden layer and a loss function that seeks to minimize the product of predicted false positives (FP) and a parameter λ_1, subtracted from the true positives (TP). D) During the prediction phase, using the model trained with optimized hyperparameters, a prediction is made on the withheld testing set. For samples that simultaneously have both LGD and missense variation, two separate probabilities are retrieved from LGD- and missense-specific models for a given individual, and the maximum predicted probability is returned per individual. E) For ranking genes based on their importance to NDDs, artificial samples are generated such that each artificial sample has a single *de novo* variant in a unique gene, using either LGD or missense variation, separately. Application of the prediction phase (D) on artificial samples yielded a ranking of the relative importance of a gene to NDDs determined via *de novo* coding variation.

Model architecture development and hyperparameter tuning

Separate models were trained for *de novo* LGD variation and missense variation, referred to as shallow neural net (SNN) LGD-specific and missense-specific models. Each variation-specific SNN consists of two phases, a hyperparameter optimization phase and a prediction phase. After splitting all samples into training (75%) and testing (25%) sets, the hyperparameter optimization phase is applied to the training set, choosing optimal hyperparameters within a selected search space (**Figure 1B**, **Supplementary Table 2**, **Additional File 1**). The purpose of the hyperparameter optimization phase for the SNN is to select a set of hyperparameters that yield the largest true positive rate at very low false positive rates on the training set to use during the prediction phase. Similarly, random forest (sklearn.ensemble.RandomForestClassifier), SVM (sklearn.svm.LinearSVC), and logistic regression (sklearn.linear_model.LogisticRegression) classifiers, hereon referred to as baseline models, are individually subjected to hyperparameter optimization and prediction phases. To allow direct comparison of each model's performance, identical training/testing splits are provided to SNN and baseline models. The performance of SNN and baseline models are additionally compared to the TPR and FPR of the following heuristics, in which an individual is classified as a case if the individual has an LGD mutation in: 1) any gene with a i) SFARI score of 1 (high confidence ASD gene) or ii) SFARI score of 1 or 2 (strong candidate ASD gene) (https://gene.sfari.org/database/gene-scoring/), 2) any gene identified by SPARK as a i) prioritized or ii) risk gene, and 3) any gene with i) pLI >= 0.90 or ii) LOEUF < 0.35 (**Additional File 1**).

In the hyperparameter optimization and prediction phases (**Figure 1C**),

$$loss = 1 - (TP - (\lambda_1 * FP)) \text{ [Equation 1]}$$

is used as a custom loss function (**Equation 1)** for the SNNs, in which the objective is to minimize the product of the number of false positives (FP) and the hyperparameter λ_1 subtracted from the true positives (TP). The value of λ_1 is selected during the hyperparameter optimization phase. The SNN architecture consists of an input layer, a hidden layer with ReLU activation and an optimized number of neurons, and an output layer with sigmoid activation and L2 regularization with an optimized regularization parameter λ_2. The SNN uses the Adam optimization algorithm.

To return a prediction that incorporates both LGD and missense variation for individuals who possess both types of variants simultaneously (referred to as a 'combined' prediction), predictions are retrieved from the separately trained LGD- and missense-specific models for SNN and baseline models. For a given sample with both LGD and missense variation, the maximum predicted probability from the two separately trained variation-specific models is returned as the predicted probability of being a case primarily due to *de novo* coding variation (**Figure 1D**). By using the maximum predicted probability, the model is trained to learn the class of an individual given their *de novo* mutation that is predicted to have the largest deleterious effect.

The average performance of a model over 100 independent training and testing splits is measured by determining the average TPR at FPR < 0.01, ROC-AUC, and PR-AUC for LGD-specific, missense-specific, and combined predictions for the SNN approach using the custom loss function, three baseline models, an ensemble model, and an ensemble model excluding SNN predictions (**Additional File 1**). To demonstrate the importance of gene score features and PrimateAI scores to increased TPR at FPR < 0.01, SNN and baseline models were trivially trained on one-hot encoded feature matrices indicating only the presence or absence (denoted as 1 or 0, respectively) of *de novo* LGD or missense mutation, and performance metrics were returned. To additionally assess the performance of the missense-specific model using only deleterious

missense variation with PrimateAI scores >= 0.803 (as described in Sundaram et al. 2018), the missense-specific model i) was trained using only samples with deleterious missense variation (PrimateAI >= 0.803) without discarding any samples, or ii) was executed while excluding samples without deleterious missense mutation from training and testing sets.

Neurodevelopmental disorder (NDD) gene ranking

To rank genes according to their relative importance to NDDs using *de novo* coding variation in the form of *de novo* LGD mutations or missense mutations, two sets of artificial samples (LGD- and missense-specific) were created. The artificial samples each contain a single LGD (or missense) variant in a unique gene in the human genome (**Figure 1E**). The probability of being a case is predicted for each of these artificial samples using the previously trained SNN LGD- or missense-specific models. For every artificial sample and its corresponding gene containing a *de novo* LGD (or missense) variant, the predicted probability indicates the relative importance of the gene to NDD risk from *de novo* coding variation. Enrichment of *de novo* LGD and missense mutation in NDD cases relative to controls was assessed (**Additional File 1**).

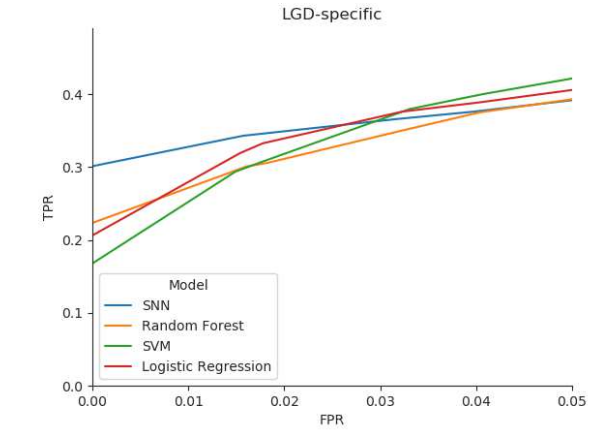
Results

To identify, at very low false positive rates, a subset of affected NDD cases using rare coding variation consisting of *de novo* LGD and missense variation, LGD- and missense-specific feature matrices indicating the presence of *de novo* variation within genes were constructed (**Figure 1A**). Additional features incorporating constraint and conservation data were used to improve classification of NDD cases using LGD variation. The ability of SNNs (**Figure 1C**) to classify NDD cases at very low false positive rates were compared to various classifiers, including

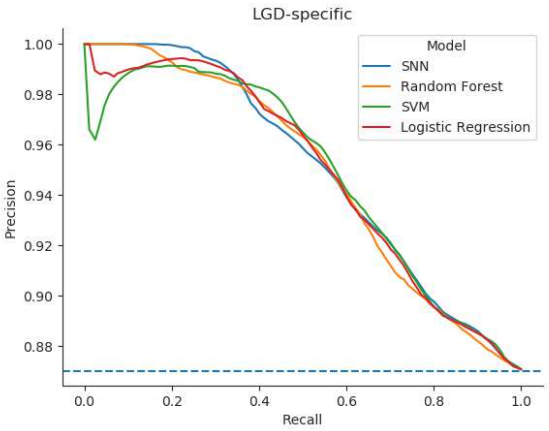
random forest, support-vector machine (SVM), and logistic regression (baseline models), in addition to three heuristics.

De novo likely gene-disruptive (LGD) mutations distinguish a subset of NDD cases from controls with low false positive rate

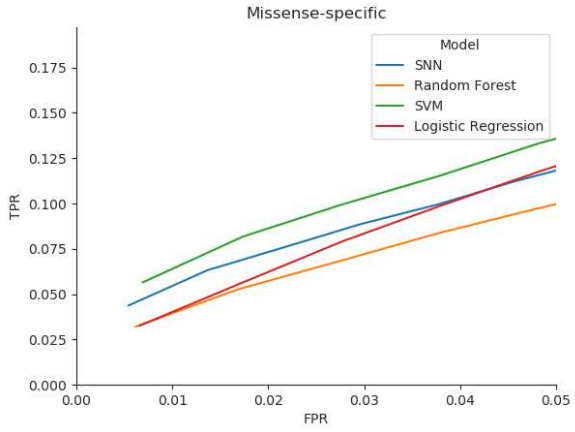
At very low false positive rates (FPR < 0.01), an SNN trained on an LGD-specific feature matrix captures 30.1% of NDD cases possessing any *de novo* LGD coding variation. In comparison to baseline models, the SNN trained on an LGD-specific feature matrix is able to identify 5.29% to 10.25% (95% confidence interval (CI)) more NDD cases at FPR < 0.01 than the random forest classifier, and more than 5.73% to 17.26% (95% CI) NDD cases than SVM or logistic regression models (**Figure 2, Table 1, Supplementary Figure 1**). To measure the extent to which the SNN and other models achieve increased TPR at FPR < 0.01 compared to a randomized model, a z-score was also calculated (**Additional File 1, Table 1**).


LGD-specific
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

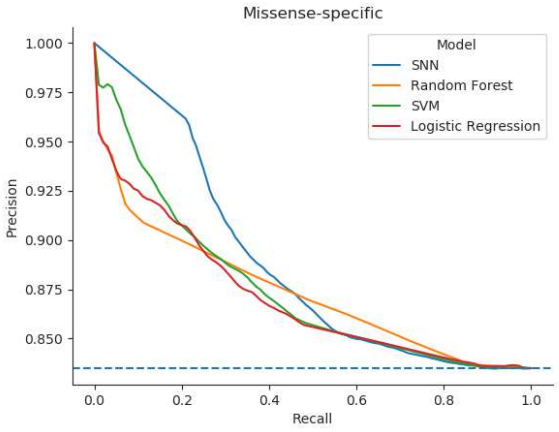
A


LGD-specific
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

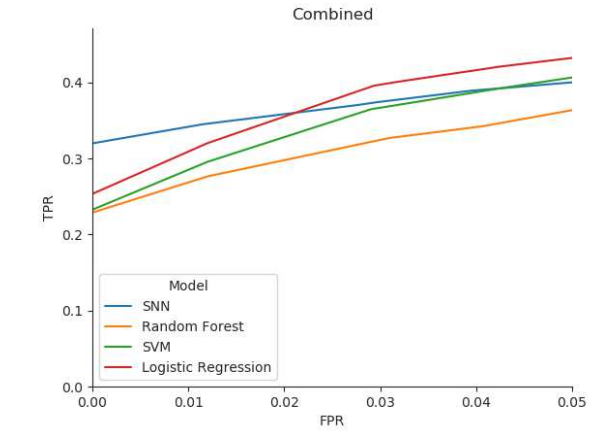
B


Missense-specific
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

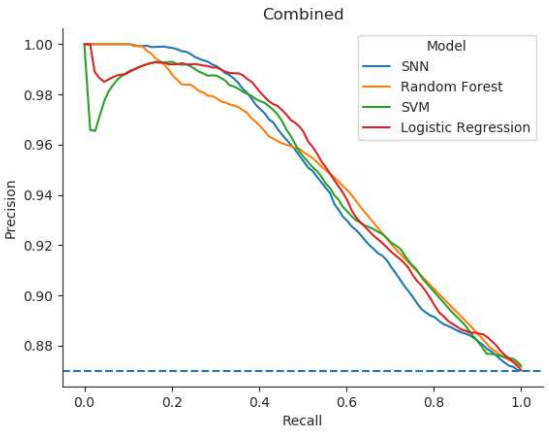
C


Missense-specific
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

D


Combined
TPR
FPR
Model
SNN
Random Forest
SVM
Logistic Regression

E


Combined
Precision
Recall
Model
SNN
Random Forest
SVM
Logistic Regression

F

Figure 2. Receiver operating characteristic (ROC) and precision recall (PR) curves for LGD- and missense-specific and combined predictions for SNN and baseline (random forest, SVM, and logistic regression) models. Random classification is displayed as a dashed blue line in all PR curves. Models trained on LGD-specific variation feature matrices additionally use constraint and conservation gene score information, whereas models provided with missense-specific feature matrices do not use gene score information. For LGD-specific features, the SNN achieves greater TPR at low FPR < 0.01 compared to baseline models, a trend which is evident even at FPR < 0.05 (A), and the SNN achieves comparable precision at lower recall compared to baseline models (B). Models trained on missense-specific variation alone are poor predictors of NDD status; SNN and baseline models show similar TPR at FPR < 0.05, with the SNN achieving slightly higher rates at low FPR (C). The SNN displays comparable precision at low recall thresholds when trained on missense-specific variation (D). E) For combined prediction for samples with both missense and LGD variation, the proportion of cases captured at FPR < 0.01 is largest for the SNN, and similar precision at low recall is observed for the SNN compared to baseline models (F).

For the SNN, ROC-AUC and PR-AUC values of 0.72785 (0.7227 to 0.7326, 95% confidence interval (CI)) and 0.9505 (0.9490 to 0.9519, 95% CI), respectively, were observed (**Table 1**). Observed PR-AUC values were comparable among the SNN and baseline models in their deviance from the randomized model, displaying similar z-scores. Note that due to the large number of cases in proportion to controls in available datasets, PR-AUC values for SNN and baseline models are significantly inflated; the random assignment model has an PR-AUC of over 0.85.

The inclusion of gene score features derived from pLI, LOEUF, RVIS, and phastCons element scores improves upon an SNN trivially trained only on LGD mutations (TPR at FPR < 0.01 = 0.24532, ROC-AUC = 0.66696, PR-AUC = 0.93597) (**Supplementary Table 3, Supplementary Figure 2**). In addition, baseline and SNN models yield similar performance metrics when trivially trained on only LGD mutations, indicating that the inclusion of gene constraint and conservation information is important to accurate classification of NDD cases using *de novo* LGD mutations (**Supplementary Table 3**).

Compared to the TPR and FPR of the three previously described heuristics, in which a sample was classified as a case if the sample possessed an LGD mutation in a set of prioritized genes, decreased TPR at low FPR thresholds in comparison to the SNN was observed for each heuristic (**Supplementary Table 4, Supplementary Figure 3**). No heuristic achieved similar TPR values greater than 0.30 at FPR less than 0.01.

Table 1. Average true positive rate (TPR) at false positive rate (FPR) < 0.01, receiver operating characteristic (ROC) - area under the curve (AUC), and precision recall - area under the curve (PR-AUC) for LGD-specific, missense-specific, and combined shallow neural net (SNN), baseline, ensemble models, and randomized predictions. An ensemble model generated only from the predictions of baseline models while excluding SNN predictions is referred to as 'Ensemble - SNN'. To generate randomized predictions, probabilities drawn from a uniform distribution were randomly assigned to samples. Average performance metrics are measured over 100 independent iterations of randomized training/testing splits on the testing set, in which the same training/testing partition is provided to all models at each iteration. Confidence intervals (95% CI) are indicated in parentheses, followed by a z-score quantifying the deviance from the mean performance metric of a certain model and the randomized model (Additional File 1). The PR-AUC values associated with randomized predictions were calculated by dividing the number of cases in a testing set by the total number of samples within the testing set.

Input features	Model	TPR at FPR < 0.01 (95% CI); z-score	ROC-AUC (95% CI); z-score	PR-AUC (95% CI); z-score
LGD-specific	SNN	0.30129 (0.2906, 0.3124); 4.93244	0.72785 (0.7227, 0.7326); 4.01329	0.95050 (0.949, 0.9519); 5.86600
	Random forest	0.22342 (0.2099, 0.2377); 2.83170	0.71997 (0.7154, 0.7244); 3.95991	0.94866 (0.9472, 0.95); 5.81660
	SVM	0.16790 (0.1398, 0.1962); 1.04685	**0.73199** (0.7278, 0.7365); 4.18017	0.94825 (0.9463, 0.9498); 5.33855
	Logistic regression	0.20632 (0.18, 0.2333); 1.34869	0.72695 (0.7222, 0.7317); 4.06566	0.94877 (0.9471, 0.9504); 5.58760
	Ensemble	**0.30715** (0.2965, 0.3174); 5.08163	0.73037 (0.7261, 0.7347); 4.14049	**0.95176** (0.9504, 0.953); 6.08741

	Ensemble - SNN	0.23347 (0.2213, 0.2453); 3.33032	0.72823 (0.724, 0.7325); 4.10213	0.95023 (0.9488, 0.9515)); 6.00325
	Randomized	0.01660 (0.0135, 0.0202)	0.50627 (0.4963, 0.5164)	0.8698
Missense-specific	SNN	0.02334 (0.0199, 0.0267); 1.09477	0.54378 (0.5391, 0.5483); 1.23832	**0.88139** (0.878, 0.885); 2.40309
	Random forest	0.01279 (0.0109, 0.0151); 0.78867	0.53086 (0.5287, 0.533); 1.17197	0.87220 (0.8705, 0.8738); 3.97519
	SVM	**0.02610** (0.022, 0.0301); 1.09631	0.55910 (0.5564, 0.5618); 2.22556	0.87486 (0.8737, 0.876); 5.88837
	Logistic regression	0.01214 (0.0101, 0.0144); 0.72456	0.55810 (0.5551, 0.5609); 2.13551	0.87071 (0.8694, 0.872); 4.82097
	Ensemble	0.02530 (0.022, 0.0288); 1.18239	**0.56006** (0.5571, 0.5631); 2.18983	0.87374 (0.8726, 0.8749); 5.71154
	Ensemble - SNN	0.02386 (0.0205, 0.0272); 1.13687	0.55915 (0.5564, 0.5619); 2.18614	0.87383 (0.8726, 0.8751); 5.69270
	Randomized	0.00406 (0.0033, 0.0048)	0.50304 (0.4991, 0.5071)	0.8350
Combined	SNN	0.31985 (0.3038, 0.3348); 3.55285	0.71422 (0.7071, 0.7215); 2.93749	0.94685 (0.9445, 0.949); 3.95676
	Random forest	0.22892 (0.2129, 0.2456); 2.39793	0.71830 (0.7121, 0.7246); 3.15223	0.94740 (0.9453, 0.9494); 4.02305
	SVM	0.23267 (0.2058, 0.2598); 1.41386	0.72803 (0.7211, 0.7346); 3.21778	0.94620 (0.9437, 0.9486); 3.67270
	Logistic regression	0.25347 (0.226, 0.2837); 1.48639	0.73280 (0.7269, 0.7389); 3.34153	0.94874 (0.9466, 0.951); 4.05063
	Ensemble	**0.33567** (0.3216, 0.3508); 3.87773	**0.74128** (0.7345, 0.7481); 3.42116	**0.95215** (0.9501, 0.9541); 4.35673
	Ensemble - SNN	0.23961 (0.2249, 0.2549); 2.63914	0.73737 (0.7302, 0.7447); 3.30974	0.94899 (0.9468, 0.951); 4.09443
	Randomized	0.02898 (0.0224, 0.036)	0.53177 (0.5218, 0.5409)	0.8701

Integration of missense and LGD-specific models improves prediction on individuals with both de novo missense and LGD variants

To assess the ability of *de novo* missense mutations to distinguish NDD cases from unaffected controls, *de novo* missense variants from individuals with at least one missense variant were retrieved, consisting of 6,947 samples possessing a total of 10,777 missense mutations. SNN and baseline models trained on missense variation capture less than 2.6% of NDD cases at FPR < 0.01 (**Figure 2, Table 1**), indicating that accurate prediction of NDDs using only missense *de novo* variants is an extremely challenging problem. Slightly increased TPR at FPR < 0.01 is observed when the missense-specific model is trained only on deleterious missense variation without removing any samples from training and testing; excluding samples without deleterious missense variation from training and testing yields 2,242 samples (2,257 cases; 248 controls) with 2,505 deleterious missense variants and increased TPR at FPR < 0.01 (**Supplementary Table 5**).

For samples possessing both *de novo* missense and LGD variants, accurate prediction of NDD cases at low FPR can be improved by taking the maximum predicted probability from two models trained separately on only missense or LGD variation, referred to as a 'combined' prediction (**Figure 2, Table 1)**. Combined prediction on samples with both missense and LGD variation captures an increased fraction of cases. For example, compared to the LGD-specific SNN, an SNN using combined prediction is able to detect at most 4.22% more affected cases at FPR < 0.01 (95% CI). TPR at FPR < 0.01 - associated z-scores for the SNN are greater by 1.41-2.51 than values observed for baseline models using combined predictions.

Ensemble prediction yields increased true positive rates at very low false positive rates compared to separately trained shallow neural net (SNN) and baseline models

An ensemble prediction was generated by returning the average predicted probability from the SNN, random forest, SVM, and logistic regression models for a given sample in the testing set. Compared to SNN and baseline models for LGD-specific, missense-specific, and combined models, the ensemble model consistently yields a larger TPR at low FPR values (**Supplementary Figure 4, Table 1**). The predictive contribution of the SNN to the ensemble model is more substantial than that of the baseline models. For example, the TPR at FPR < 0.01 is greater for LGD-specific and combined prediction SNNs than ensemble models that exclude SNN predictions, referred to as 'Ensemble - SNN' (**Table 1, Supplementary Figure 4**). Additionally, for LGD-specific and combined predictions, there is no overlap of 95% CIs between SNN and Ensemble - SNN models. From the ensemble prediction's constituent models, the SNN performs most similarly to the full ensemble method, differing by 0.586% and 1.582% in TPR at FPR < 0.01 given LGD-specific and combined predictions, respectively. In addition, the ensemble model achieves a slightly higher average PR-AUC, as evidenced by an increased z-score, than any individual SNN or baseline model for corresponding LGD-specific (0.95176) or combined predictions (0.95215) (**Table 1**).

Integration of constraint, conservation, and de novo mutation data permit NDD gene prioritization

Training of SNNs (**Figure 1C**) on variation-specific feature matrices enables NDD risk gene ranking according to the effect of *de novo* missense and LGD mutations within specific genes (**Figure 1E**). For example, using only LGD variants during SNN training reveals genes that are sensitive to LGD mutations and play important roles in typical neurodevelopment. Gene rankings and associated SFARI Gene scores are displayed in **Supplementary Table 6** in descending order according to their relative importance to NDD risk.

For artificial LGD samples (that each possess a single LGD variant in a unique gene), an increased enrichment of LGD variants is observed in NDD cases relative to unaffected controls at increasing predicted probabilities (**Figure 3A**), and a slight increased enrichment of missense variants is also observed in NDD cases for genes ranked according to a trained LGD-specific SNN (**Figure 3B**). The difference in enrichment (E_{diff}) of LGD or missense mutation in cases relative to controls per gene is calculated by **Equation 2** (**Additional File 1**). Significant correlation exists between pLI (*p-value* < 2.25e-79) and LOEUF (*p-value* < 1.09e-63) values with predicted probability ranks for artificial LGD samples (**Figure 3C-D**). For gene rankings produced by a missense-specific SNN, similar trends in enrichment of *de novo* coding variation in NDD cases relative to controls are observed, although the range of probabilities predicted by the missense-specific SNN narrows compared to the LGD-specific SNN, and the strength of correlation amongst pLI and LOEUF values with predicted probabilities is reduced (**Supplementary Figure 5**).

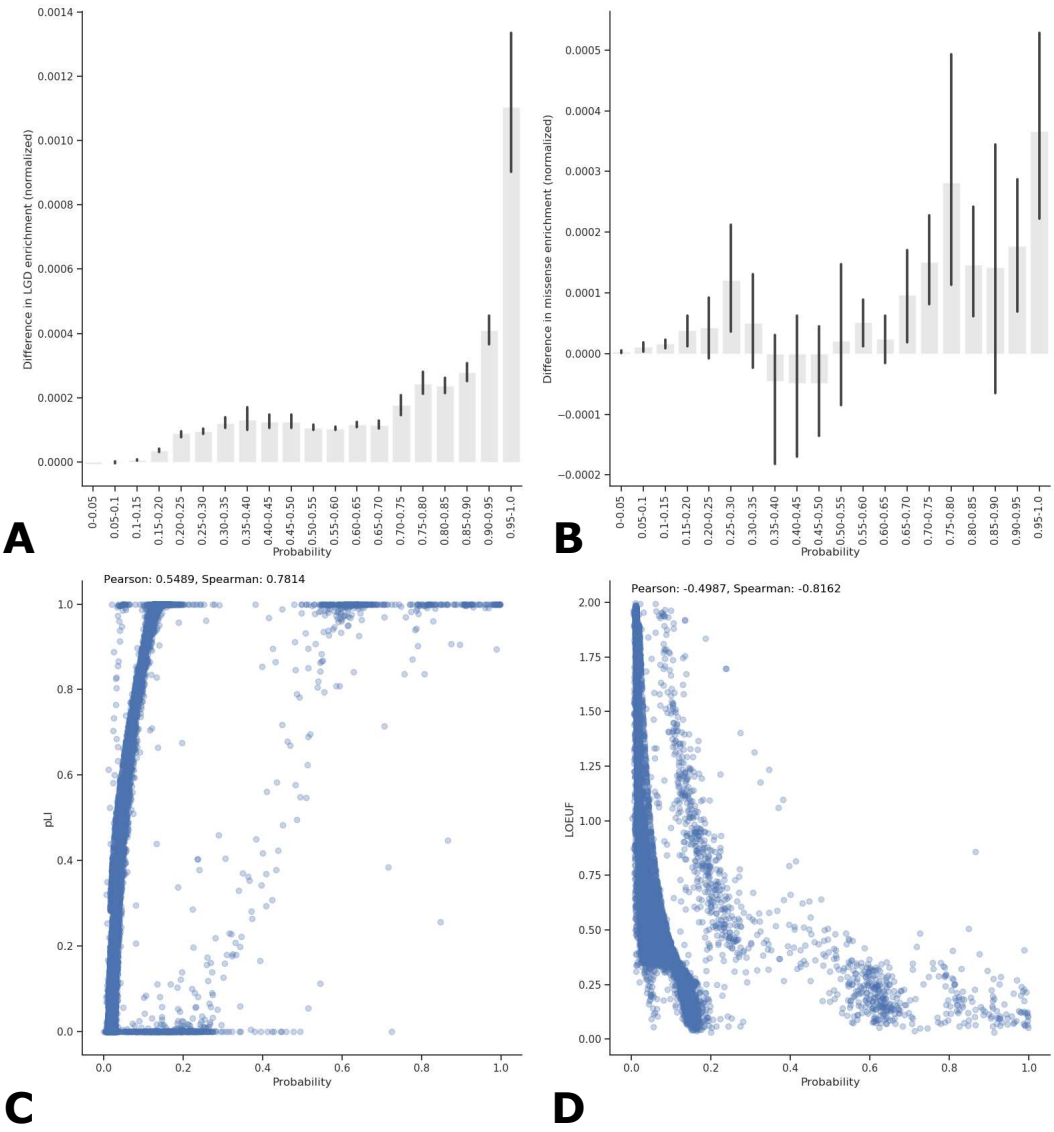


Figure 3. Increased enrichment of *de novo* LGD and missense mutation in NDD cases relative to unaffected controls in highly ranked NDD genes according to an SNN trained on an LGD-specific feature matrix. Applying a trained SNN on artificial samples containing a single unique LGD variant allows the SNN to rank genes according to their relative importance to NDD risk with respect to LGD coding variation. The difference in enrichment in NDD cases versus controls per

ranked gene is calculated by **Equation 2** (**Additional File 1**) and displayed on the y-axes. Increasing probability (x-axes) indicates increasing importance to NDD risk. The average predicted probability was determined for each artificial sample over 100 independent iterations, and 95% confidence intervals are shown. At increasing probabilities for artificial samples with LGD variants, a steady, increased enrichment of LGD in cases (A) is observed, and a slight enrichment of missense variation (B) in cases relative to controls is also observed at increasing probabilities. The probability (ranks) assigned to genes is significantly correlated with both pLI (C) and LOEUF (D) values retrieved from gnomAD (v2.1.1). pLI values range from 0 to 1, where values above 0.9 suggest intolerance to LGD mutation, whereas LOEUF values represent a ratio of observed over expected LGD mutations and values less than 0.35 suggest intolerance to LGD mutation.

For the LGD-specific SNN model, inclusion of gene score features generated from pLI, LOEUF, RVIS, and PhastCons produces rankings with greater enrichment of LGD variation in cases relative to controls at higher probabilities than an LGD-specific SNN model trivially trained on one-hot encoded mutation information that excludes gene score features (**Supplementary Figure 6**).

Discussion

To distinguish neurodevelopmental disorder (NDD) cases from unaffected controls at extremely low false positive rates using *de novo* coding variation and measures of gene constraint and conservation, we developed a shallow neural network (SNN) with a customized objective function to maximize true positives while simultaneously minimizing false positives (**Figure 1**). Although most cases of NDDs arise from a variety of classes of genetic variation, particularly common, non-coding, and structural variants, focusing specifically on *de novo* coding variation of relatively large effect size is a tradeoff to obtain significantly reduced FPR on a small but significant subset of samples. Compared to traditional machine learning techniques, such as random forest, support vector machine (SVM), and logistic regression (referred to as 'baseline' models), the constructed SNN is able to achieve greater true positive rates (TPR) at false positive rates (FPR) less than 0.01 given LGD-specific variation (**Table 1, Figure 2**). The ability of the SNN to capture more than 30% of cases at FPR < 0.01, corresponding to at least 5.29% more cases than any baseline model (**Table 1**), indicates that the use of a SNN with the custom loss function (**Equation 1**) is beneficial in classifying NDD cases at very low FPR. Note that it is estimated that LGD variants have been observed in roughly 10% of ASD cases and up to 30% of DD cases (Wang et al. 2021). Thus, our

results indicate that the proposed SNN should be able to identify >3% of ASD and >10% of all DD cases while having an FPR of virtually zero simply by considering *de novo* LGD variants.

To demonstrate that gene scores related to constraint and conservation, including pLI, LOEUF, RVIS, and phastCons, were useful and necessary for the SNN to yield elevated TPR at FPR < 0.01 compared to baseline models given LGD-specific variation, the performance of trivially trained SNN and baseline models were measured (**Supplementary Table 3, Supplementary Figure 2**). During trivial training, *only* a feature matrix of one-hot encoded values (1 or 0) denoting the presence or absence of a *de novo* coding variation within a gene were provided as input features to models. We note that most *de novo* mutations retrieved from denovo-db were identified via simplex studies that facilitate the identification of *de novo* variants, thus potentially introducing biased prediction in favor of variants identified via simplex rather than multiplex studies. We would also like to note that multiplex NDD cases will have a potentially lower chance of being caused by *de novo* variants and thus reduce the ability of our model's accurate prediction of these cases. Similar TPR at FPR < 0.01 values are reported for trivially trained and trivially trained baseline models, indicating that the inclusion of gene score features greatly contributes to the SNN's improved ability to classify NDD cases at very low FPR.

In addition, a simple ensemble method that uses the average predicted probability from SNN and baseline model predictions is able to identify NDD cases at greater TPR at FPR < 0.01 and slightly increased precision at lowered recall than any of its constituent models (**Table 1, Supplementary Figure 4**). Excluding SNN predictions from the ensemble model reveals that the SNN, compared to baseline models, contributes substantially to the ensemble model's ability to accurately classify NDD cases at low FPR values. In fact, for LGD-specific variation, an ensemble

method that excludes SNN predictions produces decreased TPR at FPR < 0.01 metrics compared to the SNN alone (**Table 1**).

The ability of SNN and baseline models to use only missense variation to identify NDD cases is relatively poor. However, the incorporation of both missense and LGD-specific predictions during 'combined' prediction for samples containing both LGD and missense variation, in which the maximum predicted probability from two separately trained missense- and LGD-specific models are returned, increases average TPR at FPR < 0.01 compared to using only probabilities predicted by an LGD-specific model (**Table 1, Figure 2**). The improved performance of combined predictions indicates that certain samples possessing very deleterious missense variation (in addition to LGD variation) are correctly classified as cases when the predicted probability associated with the missense-specific model, rather than the LGD-specific model, is retrieved.

SNNs trained on LGD- and missense-specific feature matrices containing *de novo* coding variation from NDD cases and controls are able to rank genes according to their relative importance to NDD risk when applied to artificial samples which each contain a single type of *de novo* variant in a single gene (**Supplementary Table 6**). An increased enrichment of *de novo* LGD and missense mutation in NDD cases relative to controls is observed in highly ranked genes (those with higher predicted probability of being a case) using LGD-specific variation (**Figure 3**). Significant, strong correlation exists between predicted probability for artificial samples for both the pLI and LOEUF constraint metrics, showing that the ranking via LGD-specific variation can accurately detect most high risk NDD genes. Among the 50 most highly ranked genes using LGD-specific variation, a total of 47 out of 50 genes are classified as high confidence (39 genes with score 1), strong candidate (6 genes with score 2), and suggestive evidence (2 genes with score 3)

autism spectrum disorder (ASD) risk genes, including genes relevant to syndromes, according to SFARI Gene and OMIM (**Supplementary Table 6**). Among genes with predicted probabilities greater than 0.90 (ranks 1-55), four genes (*WDR45*, *CLTC*, *BRPF1*, and *GATAD2B*) do not possess SFARI annotations, but have been associated with neurodegeneration and intellectual disability according to OMIM annotations. Highly ranked genes lacking both SFARI Gene scores and OMIM annotations suggest candidate NDD genes susceptible to *de novo* LGD variation. Evidence of association with NDDs (*ZFHX3* (Fuller et al. 2018), *CHD5* (Parenti et al. 2021)*, UBR3* (Murcia Pienkowski et al. 2020)) or enrichment of *de novo* LGD mutation in NDD cases (*ANP32A, SKIDA1* (Coe et al. 2019)), neurodegeneration (*ANP32A* (Podvin et al. 2020), *HECTD1* (Schmidt et al. 2021)), gliomas (*LARP4B* (Koso et al. 2016)), synapses and neuronal formation (*LMTK3* (Takahashi et al. 2020), *DOT1L* (Franz et al. 2019)) have been studied in model organisms, cell lines, and families for these candidate NDD genes.

Weaker correlation is observed for missense-specific rankings with pLI and LOEUF values, and enrichment of *de novo* non-synonymous mutation is also present in NDD cases relative to controls, although to a lesser extent compared to LGD-specific rankings (**Supplementary Figure 5**). The missense-specific rankings are distinct from LGD rankings in their ability to identify genes potentially sensitive to missense variation (**Supplementary Table 6**). Among highly ranked genes lacking SFARI Gene scores and OMIM annotations, previous studies suggest association with NDDs and schizophrenia (*OBSCN* (Hashimoto et al. 2016), *PLEC* (Dincer et al. 2015), *RYR2* (Lieve et al. 2019), *ZSWIM8* (Tischfield et al. 2017)), cortical formation and thickness (*LAMA5* (Omar et al. 2017), *GOLGA3* (Kim et al. 2017)), and neurodegenerative diseases (*PKHD1* (Santos-Laso et al. 2020), *DNAH*1 (Thonberg et al. 2017)).

Our results indicate that we can accurately predict a small, yet significant fraction of NDD cases using *de novo* coding variants. Currently, whole-exome or whole-genome sequencing of trios is not common practice. However, to make the early prediction of these disorders a reality, such sequencing should become common practice. Furthermore, our approach only covers a small fraction of affected patients and additional methods that use other types of biomolecular signatures, such as common variants, rare non-coding variants, and epigenomic markers, are needed to increase the reach of early prediction to a larger fraction of cases.

Conclusion

In summary, the described SNN identifies NDD cases at higher TPR while having very low FPR in comparison to traditional machine learning methods. Several factors contribute to the improved performance of the proposed approach, namely: the use of gene constraint and conservation features in LGD-specific prediction and a custom loss function that specifically seeks to maximize the TPR while minimizing the FPR. An ensemble method, aggregated from SNN and baseline model predictions, is able to correctly classify a greater proportion of cases at FPR < 0.01 compared to any individual model. The SNN itself is a major contributor to increased TPR at FPR < 0.01 observed in the ensemble model. Although *de novo* missense mutation alone is a poor predictor of case status relative to LGD mutation, missense-specific predictions are useful during combined prediction for identifying additional cases that possess highly deleterious missense mutation in addition to LGD mutation. Fully trained SNNs on LGD- or missense-specific variation are also useful in NDD risk gene prioritization, revealing candidate NDD genes enriched in *de novo* non-synonymous mutations in NDD cases relative to controls.

References

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. *Genome Biology*, *17*(1), 241. https://doi.org/10.1186/s13059-016-1110-1

Boivin, M. J., Kakooza, A. M., Warf, B. C., Davidson, L. L., & Grigorenko, E. L. (2015). Reducing neurodevelopmental disorders and disability through research and interventions. *Nature*, *527*(7578), S155–S160. https://doi.org/10.1038/nature16029

Cardoso, A. R., Lopes-Marques, M., Silva, R. M., Serrano, C., Amorim, A., Prata, M. J., & Azevedo, L. (2019). Essential genetic findings in neurodevelopmental disorders. *Human Genomics*, *13*. https://doi.org/10.1186/s40246-019-0216-4

Chow, J., Jensen, M., Amini, H., Hormozdiari, F., Penn, O., Shifman, S., et al. (2019). Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. Genome Medicine, 11(1), 65. https://doi.org/10.1186/s13073-019-0678-y

Cioni, G., Inguaggiato, E., & Sgandurra, G. (2016). Early intervention in neurodevelopmental disorders: underlying neural mechanisms. *Developmental Medicine and Child Neurology*, *58 Suppl 4*, 61–66. https://doi.org/10.1111/dmcn.13050

Coe, B. P., Stessman, H. A. F., Sulovari, A., Geisheker, M. R., Bakken, T. E., Lake, A. M., et al. (2019). Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nature Genetics*, *51*(1), 106–116. https://doi.org/10.1038/s41588-018-0288-4

Corsello, C. M. (2005). Early Intervention in Autism. *Infants & Young Children*, *18*(2), 74–85. https://journals.lww.com/iycjournal/fulltext/2005/04000/early_intervention_in_autism.2.aspx. Accessed 17 November 2020

De Felice, A., Ricceri, L., Venerosi, A., Chiarotti, F., & Calamandrei, G. (2015). Multifactorial Origin of Neurodevelopmental Disorders: Approaches to Understanding Complex Etiologies. *Toxics*, *3*(1), 89–129. https://doi.org/10.3390/toxics3010089

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., et al. (2014). Synaptic, transcriptional, and chromatin genes disrupted in autism. *Nature*, *515*(7526), 209–215. https://doi.org/10.1038/nature13772

Dincer, A., Gavin, D. P., Xu, K., Zhang, B., Dudley, J. T., Schadt, E. E., & Akbarian, S. (2015). Deciphering H3K4me3 broad domains associated with gene-regulatory networks and conserved epigenomic landscapes in the human brain. *Translational Psychiatry*, *5*(11), e679–e679. https://doi.org/10.1038/tp.2015.169

Flint, J. (2001). Genetic basis of cognitive disability. *Dialogues in Clinical Neuroscience*, *3*(1), 37–46. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181642/. Accessed 12 November 2020

Franz, H., Villarreal, A., Heidrich, S., Videm, P., Kilpert, F., Mestres, I., et al. (2019). DOT1L promotes progenitor proliferation and primes neuronal layer identity in the developing cerebral cortex. *Nucleic Acids Research*, *47*(1), 168–183. https://doi.org/10.1093/nar/gky953

Freitag, C. M. (2007). The genetics of autistic disorders and its clinical relevance: a review of the literature. *Molecular Psychiatry*, *12*(1), 2–22. https://doi.org/10.1038/sj.mp.4001896

Fuller, T. D., Westfall, T. A., Das, T., Dawson, D. V., & Slusarski, D. C. (2018). High-Throughput Behavioral Assay to Investigate Seizure Sensitivity in Zebrafish Implicates ZFHX3 in Epilepsy. *Journal of neurogenetics*, *32*(2), 92–105. https://doi.org/10.1080/01677063.2018.1445247

Gejman, P., Sanders, A., & Duan, J. (2010). The Role of Genetics in the Etiology of Schizophrenia. *The Psychiatric clinics of North America*, *33*(1), 35–66. https://doi.org/10.1016/j.psc.2009.12.003

Hadders-Algra, M. (2021). Early Diagnostics and Early Intervention in Neurodevelopmental Disorders—Age-Dependent Challenges and Opportunities. Journal of Clinical Medicine, 10(4), 861. https://doi.org/10.3390/jcm10040861

Hashimoto, R., Nakazawa, T., Tsurusaki, Y., Yasuda, Y., Nagayasu, K., Matsumura, K., et al. (2016). Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. *Journal of Human Genetics*, *61*(3), 199–206. https://doi.org/10.1038/jhg.2015.141

Heyne, H. O., Singh, T., Stamberger, H., Abou Jamra, R., Caglayan, H., Craiu, D., et al. (2018). De novo variants in neurodevelopmental disorders with epilepsy. *Nature Genetics*, *50*(7), 1048–1053. https://doi.org/10.1038/s41588-018-0143-7

Huynh, L., & Hormozdiari, F. (2018). Combinatorial Approach for Complex Disorder Prediction: Case Study of Neurodevelopmental Disorders. *Genetics*, *210*(4), 1483–1495. https://doi.org/10.1534/genetics.118.301280

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. *Nature*, *515*(7526), 216–221. https://doi.org/10.1038/nature13908

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron*, *74*(2), 285–299. https://doi.org/10.1016/j.neuron.2012.04.009

Kaplanis, J., Samocha, K. E., Wiel, L., Zhang, Z., Arvai, K. J., Eberhardt, R. Y., et al. (2020). Evidence for 28 genetic disorders discovered by combining healthcare and research data. Nature, 586(7831), 757–762. https://doi.org/10.1038/s41586-020-2832-5

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kaufman, L., Ayub, M., & Vincent, J. B. (2010). The genetic basis of non-syndromic intellectual disability: a review. *Journal of neurodevelopmental disorders*, *2*(4), 182–209. https://doi.org/10.1007/s11689-010-9055-2

Kim, D., Basile, A. O., Bang, L., Horgusluoglu, E., Lee, S., Ritchie, M. D., et al. (2017). Knowledge-driven binning approach for rare variant association analysis: application to neuroimaging biomarkers in Alzheimer's disease. *BMC Medical Informatics and Decision Making*, *17*(Suppl 1), 61. https://doi.org/10.1186/s12911-017-0454-0

Koso, H., Yi, H., Sheridan, P., Miyano, S., Ino, Y., Todo, T., & Watanabe, S. (2016). Identification of RNA-Binding Protein LARP4B as a Tumor Suppressor in Glioma. *Cancer Research*, *76*(8), 2254–2264.

Kwan, V., Unda, B. K., & Singh, K. K. (2016). Wnt signaling networks in autism spectrum disorder and intellectual disability. Journal of Neurodevelopmental Disorders, 8, 45. https://doi.org/10.1186/s11689-016-9176-3

Lieve, K. V. V., Verhagen, J. M. A., Wei, J., Bos, J. M., van der Werf, C., Rosés I Noguer, F., et al. (2019). Linking the heart and the brain: Neurodevelopmental disorders in patients with catecholaminergic polymorphic ventricular tachycardia. *Heart Rhythm*, *16*(2), 220–228. https://doi.org/10.1016/j.hrthm.2018.08.025

McRae, J. F., Clayton, S., Fitzgerald, T. W., Kaplanis, J., Prigmore, E., Rajan, D., et al. (2017).

Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, *542*(7642), 433–438. https://doi.org/10.1038/nature21062

Murcia Pienkowski, V., Kucharczyk, M., Rydzanicz, M., Poszewiecka, B., Pachota, K., Młynek, M., et al. (2020). Breakpoint Mapping of Symptomatic Balanced Translocations Links the EPHA6, KLF13 and UBR3 Genes to Novel Disease Phenotype. *Journal of Clinical Medicine*, *9*(5), 1245. https://doi.org/10.3390/jcm9051245

Niemi, M. E. K., Martin, H. C., Rice, D. L., Gallone, G., Gordon, S., Kelemen, M., et al. (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*, *562*(7726), 268–271. https://doi.org/10.1038/s41586-018-0566-4

Omar, M. H., Campbell, M. K., Xiao, X., Zhong, Q., Brunken, W. J., Miner, J. H., et al. (2017). CNS Neurons Deposit Laminin α5 to Stabilize Synapses. *Cell reports*, *21*(5), 1281–1292. https://doi.org/10.1016/j.celrep.2017.10.028

O'Roak, B. J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (New York, N.Y.), 338(6114), 1619–1622. https://doi.org/10.1126/science.1227764

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, *485*(7397), 246–250. https://doi.org/10.1038/nature10989

Parenti, I., Lehalle, D., Nava, C., Torti, E., Leitão, E., Person, R., et al. (2021). Missense and truncating variants in CHD5 in a dominant neurodevelopmental disorder with intellectual disability, behavioral disturbances, and epilepsy. *Human Genetics*, *140*(7), 1109–1120. https://doi.org/10.1007/s00439-021-02283-2

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics*, *9*(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

Podvin, S., Jones, A., Liu, Q., Aulston, B., Ransom, L., Ames, J., et al. (2020). Dysregulation of Exosome Cargo by Mutant Tau Expressed in Human-induced Pluripotent Stem Cell (iPSC) Neurons Revealed by Proteomics Analyses. *Molecular & Cellular Proteomics : MCP*, *19*(6), 1017–1034. https://doi.org/10.1074/mcp.RA120.002079

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole exome sequencing are strongly associated with autism. *Nature*, *485*(7397), 237–241. https://doi.org/10.1038/nature10945

Santos-Laso, A., Izquierdo-Sanchez, L., Rodrigues, P. M., Huang, B. Q., Azkargorta, M., Lapitz, A., et al. (2020). Proteostasis disturbances and endoplasmic reticulum stress contribute to polycystic liver disease: new therapeutic targets. *Liver international : official journal of the International Association for the Study of the Liver*, *40*(7), 1670–1685. https://doi.org/10.1111/liv.14485

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell, 180(3), 568-584.e23. https://doi.org/10.1016/j.cell.2019.12.036

Schmidt, M. F., Gan, Z. Y., Komander, D., & Dewson, G. (2021). Ubiquitin signalling in neurodegeneration: mechanisms and therapeutic opportunities. *Cell Death & Differentiation*, *28*(2), 570–590. https://doi.org/10.1038/s41418-020-00706-7

Short, P. J., McRae, J. F., Gallone, G., Sifrim, A., Won, H., Geschwind, D. H., et al. (2018). De

novo mutations in regulatory elements in neurodevelopmental disorders. *Nature*, *555*(7698), 611–616. https://doi.org/10.1038/nature25983
Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005
Soden, S. E., Saunders, C. J., Willig, L. K., Farrow, E. G., Smith, L. D., Petrikin, J. E., et al. (2014). Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Science translational medicine*, *6*(265), 265ra168. https://doi.org/10.1126/scitranslmed.3010076
Srivastava, S., Love-Nichols, J. A., Dies, K. A., Ledbetter, D. H., Martin, C. L., Chung, W. K., et al. (2019). Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genetics in Medicine*, *21*(11), 2413–2421. https://doi.org/10.1038/s41436-019-0554-6
Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z
Sztainberg, Y., & Zoghbi, H. Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. *Nature Neuroscience*, *19*(11), 1408–1417. https://doi.org/10.1038/nn.4420
Takahashi, M., Sugiyama, A., Wei, R., Kobayashi, S., Fukuda, K., Nishino, H., et al. (2020). Hyperactive and impulsive behaviors of LMTK1 knockout mice. *Scientific Reports*, *10*(1), 15461. https://doi.org/10.1038/s41598-020-72304-z
Tărlungeanu, D. C., & Novarino, G. (2018). Genomics in neurodevelopmental disorders: an avenue to personalized medicine. *Experimental & Molecular Medicine*, *50*(8), 1–7. https://doi.org/10.1038/s12276-018-0129-7
Thonberg, H., Chiang, H.-H., Lilius, L., Forsell, C., Lindström, A.-K., Johansson, C., et al. (2017). Identification and description of three families with familial Alzheimer disease that segregate variants in the SORL1 gene. *Acta Neuropathologica Communications*, *5*, 43. https://doi.org/10.1186/s40478-017-0441-9
Tick, B., Bolton, P., Happé, F., Rutter, M., & Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta□analysis of twin studies. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(5), 585–595. https://doi.org/10.1111/jcpp.12499
Tischfield, D. J., Saraswat, D. K., Furash, A., Fowler, S. C., Fuccillo, M. V., & Anderson, S. A. (2017). Loss of the neurodevelopmental gene Zswim6 alters striatal morphology and motor regulation. *Neurobiology of disease*, *103*, 174–183. https://doi.org/10.1016/j.nbd.2017.04.013
Turner, T. N., & Eichler, E. E. (2019). The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders. *Trends in Neurosciences*, *42*(2), 115–127. https://doi.org/10.1016/j.tins.2018.11.002
Turner, T. N., Yi, Q., Krumm, N., Huddleston, J., Hoekzema, K., F. Stessman, H. A., et al. (2017). denovo-db: a compendium of human de novo variants. *Nucleic Acids Research*, *45*(Database issue), D804–D811. https://doi.org/10.1093/nar/gkw865
Uddin, M., Tammimies, K., Pellecchia, G., Alipanahi, B., Hu, P., Wang, Z., et al. (2014). Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nature Genetics*, *46*(7), 742–747. https://doi.org/10.1038/ng.2980
Wang, T., Kim, C., Bakken, T. E., Gillentine, M. A., Henning, B., Mao, Y., et al. (2021,

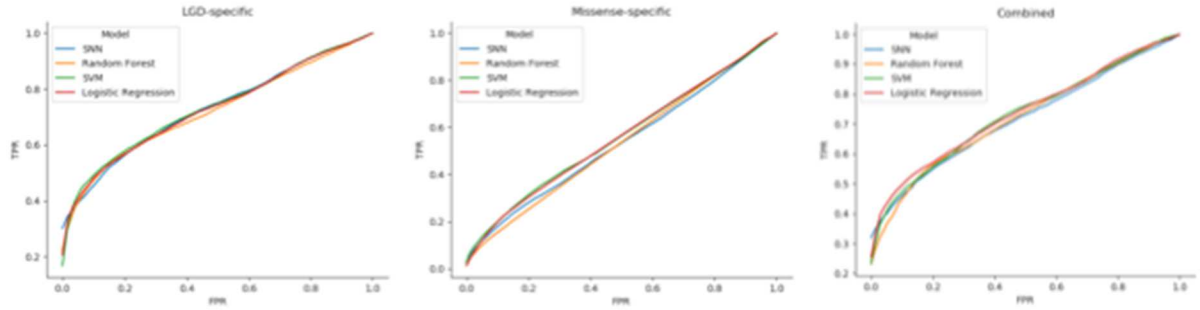
September 16). Integrated gene analyses of de novo mutations from 46,612 trios with autism and developmental disorders. bioRxiv. https://doi.org/10.1101/2021.09.15.460398

Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P., & Eichler, E. E. (2017). Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Medicine*, *9*(1), 101. https://doi.org/10.1186/s13073-017-0498-x

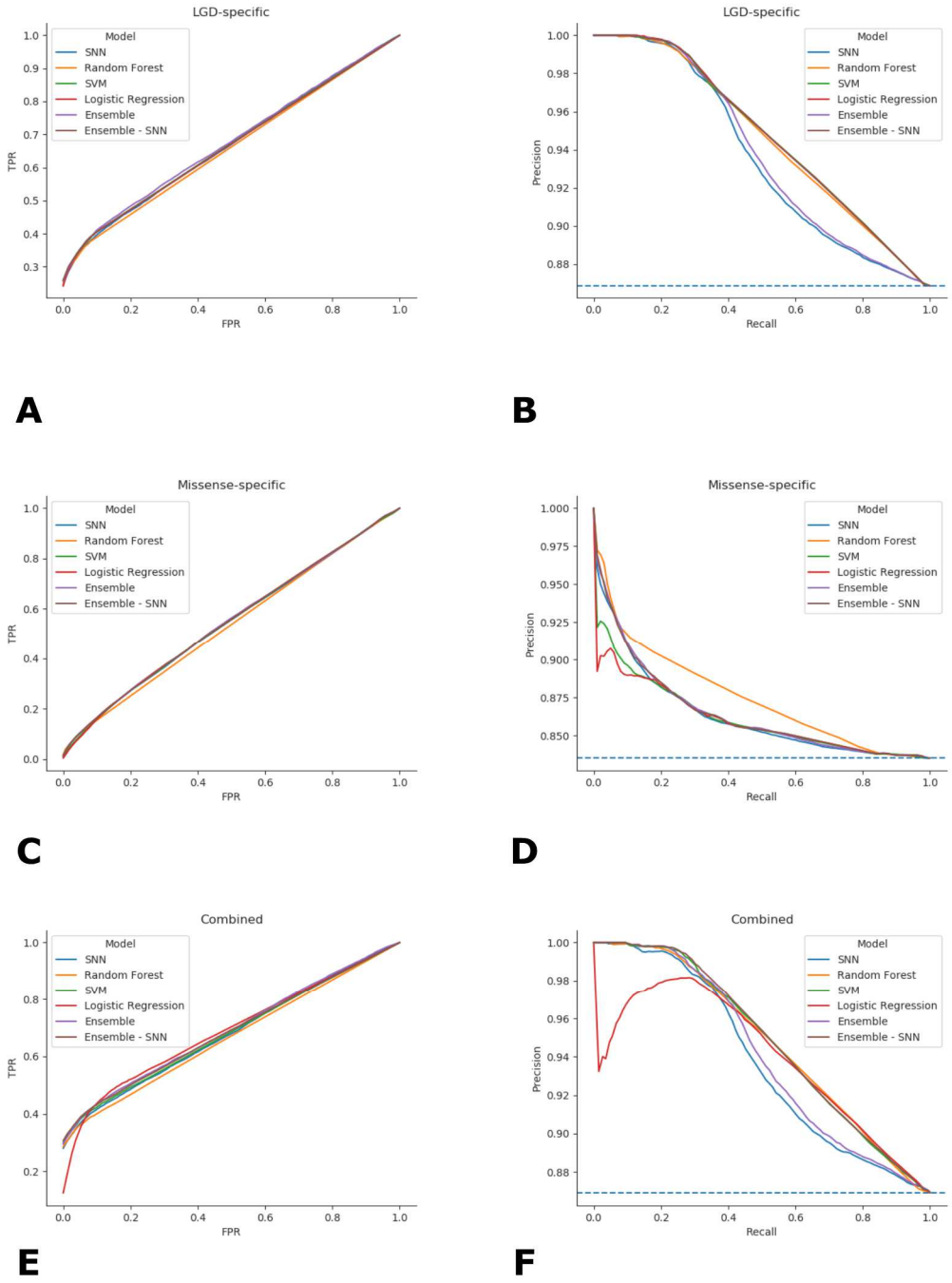
Supplementary information

Additional File 1 (PDF) includes six figures (Supplementary Figures 1-6), five tables (Supplementary Tables 1-5), and supplementary methods.

Supplementary Table 6 (XLS**).** Ranking indicating relative importance of a gene to NDD risk according to rare *de novo* LGD and missense coding variation. Ranks, also known as the predicted probability of being a case for an artificial sample, closer to 1 symbolize greater importance to NDD risk. Multiple rankings are shown based on input features provided to SNN prediction models. Rankings are displayed on separate tabs, in which a tab label beginning with 'LGD' and 'Missense' indicates rankings based on LGD- and missense-specific variation, respectively. Tab labels containing 'Trivial' correspond to rankings created using trivial one-hot encoding of *de novo* mutations in input feature matrices, whereas labels containing 'Final' correspond to non-trivially trained models, in which LGD-specific models use both mutation information and gene score features. SFARI Gene scores ('score') and syndromic status ('syndromic') and OMIM disease associations ('OMIM') are also displayed.

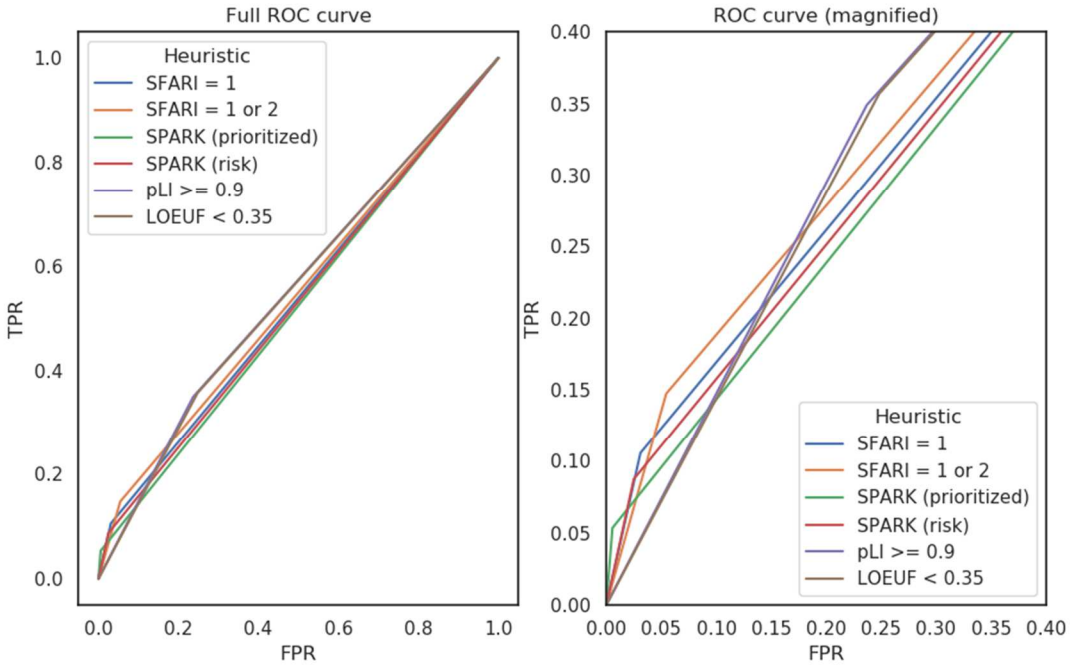


Supplementary Figure 1. Receiver operating characteristic (ROC) curves for LGD- and missense-specific shallow neural net (SNN) and baseline (random forest, SVM, and logistic regression) models at all false positive rate (FPR) thresholds.

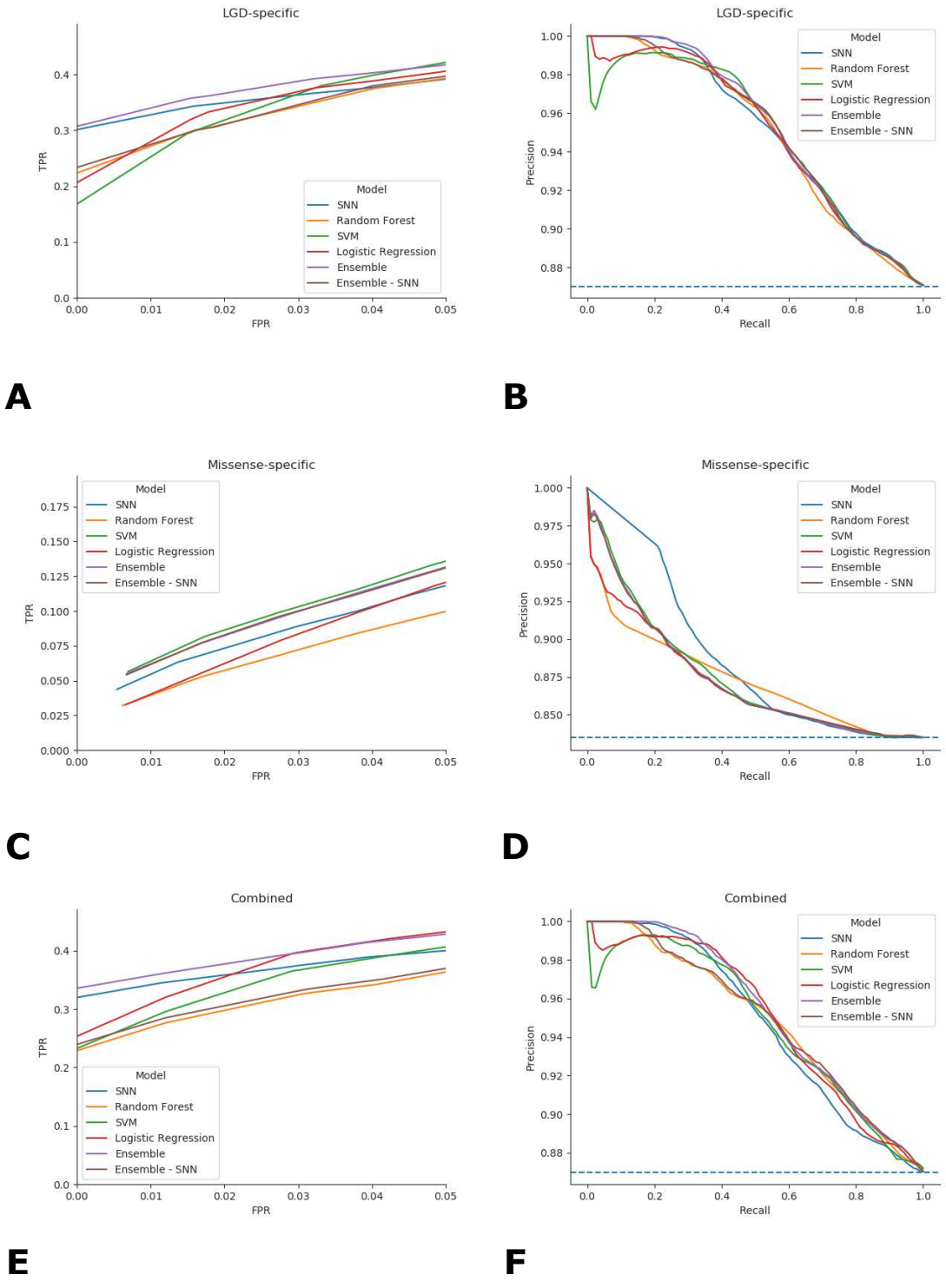


Supplementary Figure 2. Receiver operating characteristic (ROC) and precision recall (PR) curves for LGD- and missense-specific SNN, baseline (random forest, SVM, and logistic regression), and ensemble models trivially trained on only one-hot encoded mutation information. 'Ensemble -

SNN' refers to an ensemble model generated only from the predictions of baseline models. For a given sample, the ensemble model uses the average of the predicted probabilities from SNN and baseline models. SNN, baseline, and ensemble models perform similarly while trained only on LGD-specific (A-B), missense-specific (C-D), and combined mutation information (E-F). Models trained on missense-specific (C-D) variation alone are poor predictors of NDD status.

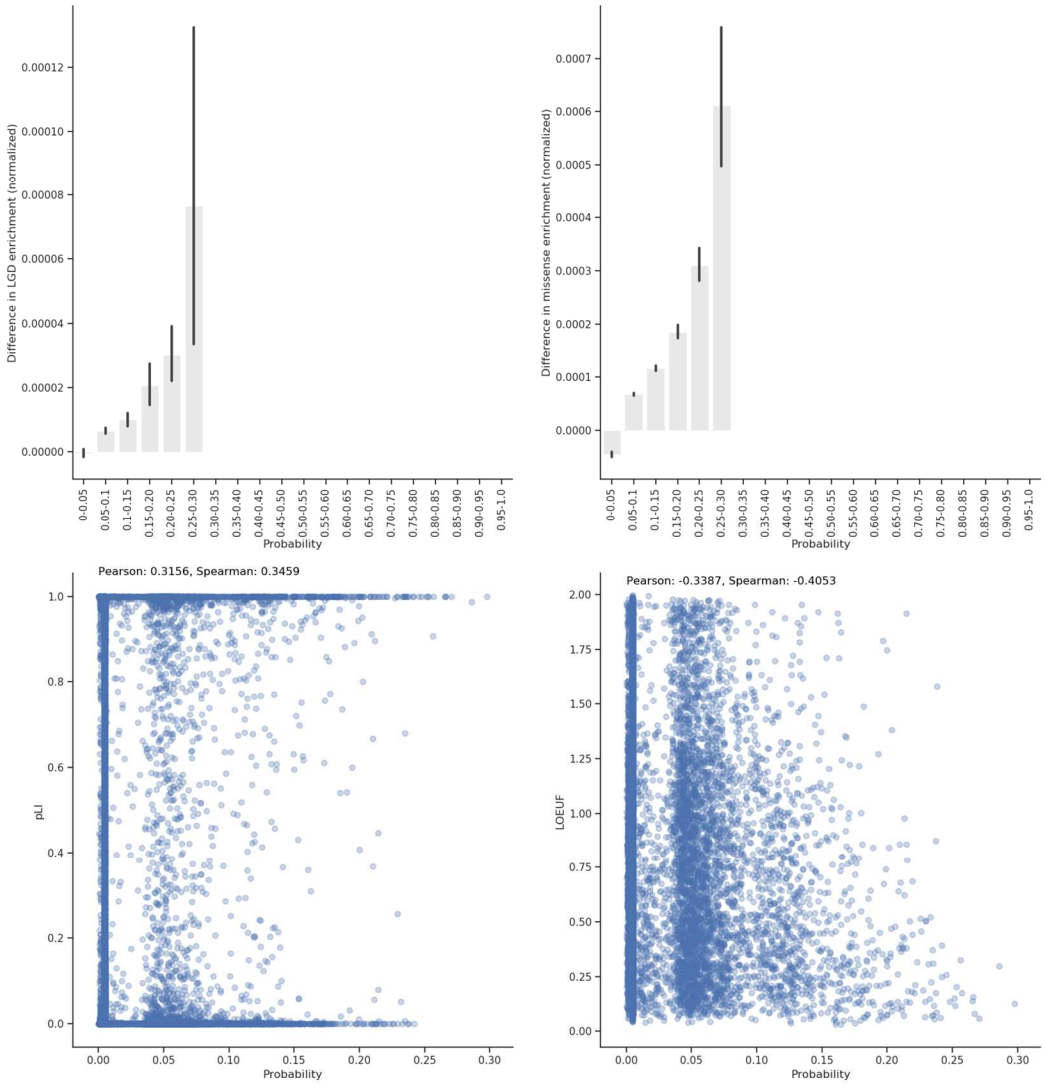


Supplementary Figure 3. Receiver operating characteristic (ROC) curves for three heuristics. For each heuristic, samples with a likely gene-disruptive (LGD) mutation in genes within a particular gene set were classified as cases. The full range of the ROC curve is displayed on the left, and a magnification is displayed on the right.



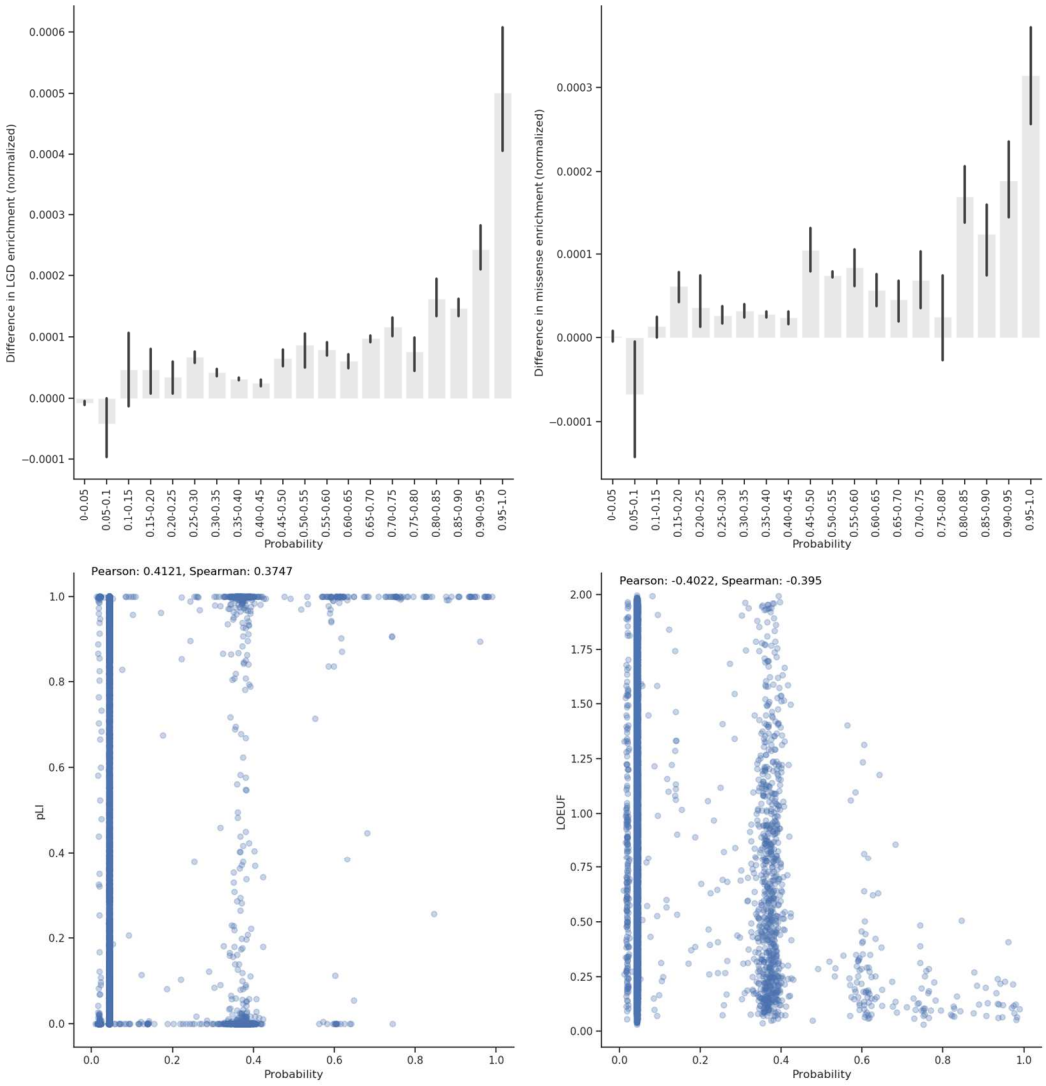
Supplementary Figure 4. Receiver operating characteristic (ROC) curves at low false positive rate (FPR) and precision recall (PR) curves for LGD- and missense-specific SNN, baseline (random forest, SVM, and logistic regression), and ensemble models. Models trained on LGD-specific

variation feature matrices additionally use constraint and conservation gene score information, whereas models provided with missense-specific feature matrices do not use gene score information. For a given sample, the ensemble model uses the average of the predicted probabilities from SNN and baseline models. Ensemble - SNN refers to an ensemble of baseline models while excluding SNN predictions. A-B) For LGD-specific features, the ensemble model and SNN achieve greater TPR at low FPR < 0.01 compared to baseline models, a trend which is evident even at FPR < 0.05. Increased precision at low recall is observed for an ensemble model trained on LGD-specific variation. C-D) Models trained on missense-specific variation alone are poor predictors of NDD status; SNN and baseline models show similar TPR at FPR < 0.05, with the SNN and ensemble models achieving slightly TPR higher rates at low FPR. All models display comparable precision at low recall. E-F) For combined prediction for samples with both missense and LGD variation, the proportion of cases captured at FPR < 0.01 is largest for the ensemble model, followed by the SNN. The ensemble model achieves the largest precision at low recall thresholds.



Supplementary Figure 5. Increased enrichment of *de novo* LGD and missense mutation in NDD cases relative to unaffected controls in more highly ranked NDD genes according to an SNN trained on a missense-specific feature matrix. Applying a trained SNN on artificial samples containing a single unique missense variant allows the SNN to rank genes according to their relative importance to NDD risk with respect to missense coding variation. The difference in

enrichment in NDD cases versus controls is calculated by **Equation 2** (**supplementary methods**) and displayed on the y-axes. Increasing probability (x-axes) indicates increasing importance to NDD risk. The average predicted probability was determined for each artificial sample over 100 independent iterations, and 95% confidence intervals are shown. At increasing probabilities for artificial samples with missense variants, an increased enrichment of LGD in cases (A) and missense in cases (B) is observed. The probability (ranks) assigned to genes is weakly correlated with both pLI (C) and LOEUF (D) values.



Supplementary Figure 6. Increased enrichment of *de novo* LGD and missense mutation in NDD cases relative to unaffected controls in highly ranked NDD genes according to an SNN trivially trained LGD-specific feature matrix consisting only of one-hot encoded mutation information. Applying a trivially trained SNN on artificial samples containing a single unique LGD variant allows the SNN to rank genes according to their relative importance to NDD risk with respect to

LGD coding variation. The difference in enrichment in NDD cases versus controls is calculated by **Equation 2** and displayed on the y-axes. Increasing probability (x-axes) indicates increasing importance to NDD risk. The average predicted probability was determined for each artificial sample over 100 independent iterations, and 95% confidence intervals are shown. At increasing probabilities for artificial samples with LGD variants, a steady, increased enrichment of LGD in cases (A) is observed. A slight enrichment of missense variation (B) in cases relative to controls is also observed at larger probabilities. The probability (ranks) assigned to genes is weakly correlated with both pLI (C) and LOEUF (D) values.

Supplementary Table 1. Neurodevelopmental disorder samples retrieved from denovo-db and associated primary phenotypes.

Study	Primary Phenotype	Cases	Controls
Simons Simplex Collection	Autism	2,508	1,911
ASC	Autism	1,445	
MSSNG	Autism	1,625	
NIMH	Autism	10	
Hashimoto	Autism	30	
GoNL	Control		250
Gulsuner	Control		84
Rauch	Intellectual disability	51	
DDD	Developmental disorder	4,293	

Supplementary Table 2. Search space for optimal hyperparameters. A single parameter is varied while other values are held constant on values most frequently determined to yield the highest average true positive rate (TPR) at false positive rate (FPR) < 0.01 in 100 independent iterations.

Batch size	λ_1	λ_2	Neurons
[8, 16, 32]	100	1e-5	16
32	[70, …, 120]	1e-5	16
32	100	[1e-6, …, 1e-2]	16
32	100	1e-5	[8, 16, 32]

Supplementary Table 3. Average true positive rate (TPR) at false positive rate (FPR) < 0.01, area under the curve (ROC-AUC), and precision recall area under the curve (PR-AUC) for LGD-specific, missense-specific, and combined shallow neural net (SNN), baseline, randomized predictions, and ensemble models trivially trained on feature matrices containing only one-hot encoded mutation information. 'Ensemble - SNN' refers to an ensemble model generated only from the predictions of baseline models. Average performance metrics are measured over 100 independent iterations of randomized training/testing splits on the testing set, in which the same training/testing partition is provided to all models at each iteration. Confidence intervals (95% CI) are indicated in parentheses, followed by a z-score quantifying the deviance from the mean performance metric of a certain model and the randomized model (**supplementary methods**).

Input features	Model	TPR at FPR < 0.01 (95% CI); z-score	ROC-AUC (95% CI); z-score	PR-AUC (95% CI); z-score
LGD-specific	SNN	0.24532 (0.2364, 0.2544); 3.85542	0.66696 (0.6616, 0.6727); 2.84528	0.93597 (0.9344, 0.9377); 4.39382
	Random forest	0.24593 (0.2358, 0.2559); 3.86630	0.66027 (0.657, 0.6636); 3.13763	0.94557 (0.9440, 0.9469); 5.40379
	SVM	0.25911 (0.2504, 0.2676); 4.43549	0.67015 (0.6668, 0.6734); 3.33195	0.94637 (0.9448, 0.9478); 5.37662
	Logistic regression	0.24141 (0.234, 0.2487); 4.71332	0.66768 (0.6644, 0.6711); 3.28010	0.94670 (0.9452, 0.9482); 5.51134
	Ensemble	0.25526 (0.2463, 0.2645); 4.45173	0.67449 (0.6697, 0.6794); 3.05424	0.93795 (0.9362, 0.9395); 4.59742
	Ensemble - SNN	0.25909 (0.2503, 0.2672); 4.38812	0.66892 (0.6654, 0.6725); 3.30916	0.94658 (0.9451, 0.948); 5.46046
	Randomized	0.01590 (0.0133, 0.019)	0.51564 (0.5086, 0.5233)	0.8684
Missense-specific	SNN	0.01274 (0.011, 0.0146); 0.81483	0.54538 (0.5427, 0.548); 1.81664	0.86321 (0.8621, 0.8643); 4.20065
	Random forest	0.01788 (0.0157, 0.0205); 0.94788	0.53337 (0.5311, 0.5357); 1.39329	0.87390 (0.8723, 0.8757); 3.85907
	SVM	0.00777 (0.0064, 0.0093); 0.41541	0.54479 (0.5424, 0.5472); 1.86765	0.86141 (0.8723, 0.8757); 3.40463

	Logistic regression	0.00442 (0.0037, 0.0052); 0.09893	0.54581 (0.5435, 0.5484); 1.91177	0.86023 (0.8602, 0.8626); 3.57777
	Ensemble	0.01385 (0.0121, 0.0157); 0.87097	0.54842 (0.5458, 0.551); 1.96634	0.86460 (0.8591, 0.8613); 4.51338
	Ensemble - SNN	0.01385 (0.0123, 0.0158); 0.87638	0.54726 (0.545, 0.5496); 1.98008	0.86464 (0.8636, 0.8656); 4.48374
	Randomized	0.00382 (0.0031, 0.0046)	0.49954 (0.496, 0.503)	0.8353
Combined	SNN	0.27952 (0.2671, 0.2929); 2.93042	0.67796 (0.6705, 0.6855); 2.45043	0.93764 (0.9356, 0.94); 3.17344
	Random forest	0.28581 (0.2743, 0.2972); 3.19354	0.66765 (0.6634, 0.6715); 2.74962	0.94663 (0.9449, 0.9484); 3.82637
	SVM	0.29840 (0.2867, 0.309); 3.43782	0.68337 (0.6785, 0.688); 2.89963	0.94854 (0.9466, 0.9506); 3.92888
	Logistic regression	0.12399 (0.1013, 0.1488); 0.63492	0.69015 (0.6841, 0.6963); 2.87051	0.93856 (0.9357, 0.9414); 2.92455
	Ensemble	0.29459 (0.2828, 0.306); 3.34720	0.68981 (0.6825, 0.6968); 2.70058	0.94080 (0.9387, 0.943); 3.42898
	Ensemble - SNN	0.30602 (0.2944, 0.3179); 3.49605	0.68776 (0.6831, 0.6925); 2.99052	0.94837 (0.9466, 0.9503); 3.91210
	Randomized	0.03293 (0.0274, 0.0387)	0.51969 (0.512, 0.5281)	0.8690

Supplementary Table 4. True positive (TPR) and false positive rates (FPR) for three heuristics. A sample was classified as a case if the sample contained a likely gene-disruptive (LGD) mutation in a set of risk genes where the gene 1) i) has a SFARI score of 1 or ii) a SFARI score of 1 or 2, 2) was classified as a SPARK i) prioritized gene or ii) risk gene, and 3) i) pLI >= 0.90 or ii) LOEUF < 0.35.

Heuristic	TPR	FPR
SFARI, score 1	0.1056	0.0311
SFARI, score 1 or 2	0.1474	0.0545
SPARK (prioritized)	0.0535	0.0056
SPARK (risk)	0.0873	0.025
pLI >= 0.9	0.3491	0.2369
LOEUF < 0.35	0.3569	0.2481

Supplementary Table 5. Average true positive rate (TPR) at false positive rate (FPR) < 0.01, area under the curve (ROC-AUC), and precision recall area under the curve (PR-AUC) for missense-specific shallow neural net (SNN), baseline, randomized predictions, and ensemble models using feature matrices containing only one-hot encoded deleterious (PrimateAI score >= 0.803) missense variation i) during training without removing any samples from the dataset or ii) during both training and testing by removing samples without deleterious missense variation from the dataset. 'Ensemble - SNN' refers to an ensemble model generated only from the predictions of baseline models. Average performance metrics are measured over 100 independent iterations of randomized training/testing splits on the testing set, in which the same training/testing partition is provided to all models at each iteration. Confidence intervals (95% CI) are indicated in parentheses, followed by a z-score quantifying the deviance from the mean performance metric of a certain model and the randomized model (**supplementary methods**).

Input features	Model	TPR at FPR < 0.01 (95% CI); z-score	ROC-AUC (95% CI); z-score	PR-AUC (95% CI); z-score
Missense-specific (i)	SNN	0.02829 (0.0211, 0.0359); 1.33101	0.54744 (0.542, 0.5525); 1.97382	0.87008 (0.8681, 0.8722); 6.10794
	Random forest	0.02660 (0.0202, 0.0336); 1.38829	0.54222 (0.5374, 0.5474); 1.76958	0.87564 (0.8715, 0.8797); 3.85456
	SVM	0.02584 (0.0175, 0.0337); 1.04541	0.55408 (0.5489, 0.5595); 2.29522	0.87556 (0.8726, 0.8783); 5.52276
	Logistic regression	0.01033 (0.0062, 0.0151); 0.57026	0.55447 (0.5493, 0.5595); 2.30987	0.87337 (0.8701, 0.8767); 4.51765
	Ensemble	0.02883 (0.0218, 0.0356); 1.43505	0.55427 (0.549, 0.5594); 2.27145	0.87188 (0.8695, 0.8741); 6.24093
	Ensemble - SNN	0.03000 (0.023, 0.037); 1.44297	0.55441 (0.5494, 0.5596); 2.31924	0.87547 (0.8727, 0.8782); 5.59827
	Randomized	0.00323 (0.0022, 0.0044)	0.50044 (0.4931, 0.5083)	0.83492 (0.8342, 0.8357)
Missense-specific (ii)	SNN	0.09378 (0.0719, 0.116); 1.51357	0.63740 (0.626, 0.6487); 3.07228	0.93743 (0.9345, 0.9403); 4.09638

	Random forest	0.08738 (0.0595, 0.1177); 1.07341	0.62847 (0.6178, 0.639); 2.89912	0.94014 (0.9371, 0.9431); 4.26020
	SVM	0.07504 (0.0503, 0.1034); 0.94629	0.63587 (0.6268, 0.6452); 3.23138	0.94094 (0.9377, 0.9442); 4.39978
	Logistic regression	0.05182 (0.0247, 0.0802); 0.56745	0.63210 (0.6209, 0.642); 3.03888	0.93585 (0.9318, 0.9399); 3.39856
	Ensemble	0.08648 (0.0647, 0.1114); 1.37385	0.64388 (0.6332, 0.6529); 3.31811	0.93892 (0.9361, 0.9416); 4.47331
	Ensemble - SNN	0.08155 (0.0578, 0.1037); 1.17750	0.63327 (0.6235, 0.6425); 3.06469	0.94012 (0.9372, 0.9431); 4.47532
	Randomized	0.01051 (0.0062, 0.0154)	0.50026 0.485, 0.5142)	0.89386 (0.8908, 0.8969)

Supplementary methods

Construction of LGD- and missense-specific feature matrices

De novo LGD and missense variants were retrieved from samples with autism spectrum disorder, developmental disability, or intellectual disability directly from denovo-db (version 1.6.1) (C Yuen et al. 2017; De Rubeis et al. 2014; Deciphering Developmental Disorders Study 2017; Genome of the Netherlands Consortium 2014; Gulsuner et al. 2013; Hashimoto et al. 2016; Iossifov et al. 2014; Krumm et al. 2015; Michaelson et al. 2012; B. J. O'Roak et al. 2014; Brian J. O'Roak et al. 2012; Rauch et al. 2012; Turner et al. 2016, 2017; Werling et al. 2018; Yuen et al. 2016).

For a given individual for a particular gene, the presence of an LGD variant was indicated in the feature matrix with a 1, the absence of any *de novo* variants as a 0, and the presence of a missense variant as the associated *PrimateAI* score (Sundaram et al. 2018). For example, for a sample possessing both LGD and missense variation, the presence of missense variation is simply denoted as a 0 in the LGD-specific matrix. In the case of multiple *de novo* variants existing in a single gene in a single sample, the mutation is recorded in the feature matrix as the larger of the scores. For the model trained on an LGD-specific feature matrix (LGD-specific model), gene score features related to pLI, LOEUF, RVIS, and phastCons values were generated by matrix multiplication with the LGD-specific feature matrix (Karczewski et al. 2020; Petrovski et al. 2013; Siepel et al. 2005). The gene scores features were concatenated with the LGD-specific feature matrix to yield a matrix of size (samples by (genes + 4)). The missense-specific feature matrix uses a simplified set of features of size (samples by genes), only indicating the presence of missense mutations in genes.

Following the splitting of all samples into training and testing sets, during model training, min-max scaling (from scikit-learn MinMaxScaler()) is applied to the training set, and the testing

set is transformed with the applied scaler. Class weights were balanced according to sklearn.utils.class_weight.compute_class_weight (version 0.22.1).

Hyperparameter optimization

During hyperparameter optimization (**Figure 1B**), K-fold stratified cross-validation (K=3) is applied to the training set, in which the input set is split into K folds. K-1 folds are used as training folds, and a single fold is used as a validation fold. Over K iterations, a different fold is selected as the validation fold. Optimal hyperparameters for SNNs are selected from the following possible values: batch size={8, 16, 32}, λ_1= {70, 80, 90, 100, 110, 120}, L2 regularization λ_2={1e-2, 1e-3, 1e-4, 1e-5, 1e-6}, and number of neurons in the hidden layer={8, 16, 32}. To decrease compute time and the hyperparameter search space, selected combinations (**Supplementary Table 2**) are evaluated. For the SNNs, for every validation fold and potential set of optimal hyperparameters, the probability of a being a case is retrieved for every individual in the validation fold and the TPR at FPR < 0.01 is determined. 'TPR at FPR < 0.01' is calculated by first identifying the largest predicted probability associated with a control in a validation fold, followed by determining the fraction of cases in the validation fold with predicted probabilities greater than that of the control with the largest predicted probability, which is equivalent to an FPR of 0 and necessarily lower than 0.01To provide a more conservative estimate of the TPR at FPR equal to 0, we refer to this value as 'TPR at FPR < 0.01'. The optimal set of hyperparameters are defined as the set of hyperparameters for which the largest average TPR at FPR < 0.01 is achieved in the validation folds.

During hyperparameter optimization for baseline models, optimal hyperparameters are selected by minimizing the model's corresponding loss function value (random forest: Gini

impurity, SVM: hinge loss, logistic regression: cross entropy). Optimal hyperparameters are selected among the following values for the baselines models: Random Forest: trees={100, 200, 300, 400, 500}, maximum depth={32, 36, 40, 44, 48, 52}; SVM: C={10, 1, 1e-2, 1e-3}; logistic regression: C={10,000, 1,000, 100, 10, 1}. C is inversely proportional to L2 regularization strength in both SVM and logistic regression models.

Assessment of model performance

The average performance of a model is assessed over 100 independent iterations in which the training and testing sets are randomly partitioned and optimal hyperparameters are selected per iteration. For each iteration, the performance metrics TPR at FPR < 0.01, ROC-AUC, and PR-AUC are determined from the predicted probabilities of samples in the testing set. Averages of these performance metrics and bootstrapped 95% confidence intervals are reported for LGD-specific, missense-specific, and combined predictions for our SNN, three baseline models, an ensemble model, and an ensemble model excluding SNN predictions (**Table 1**). The full ensemble model, for every independent iteration, returns the average predicted probability using the predicted probabilities from the SNN and baseline models for every sample in the testing set. The TPR at FPR < 0.01, ROC-AUC, and PR-AUC are reported similarly for the ensemble model on the resultant average probabilities from the SNN and baseline models. For each model, bootstrap confidence intervals (95%) and z-scores were calculated for each performance metric. The z-score was calculated as the difference between the mean performance metric for a certain model and the randomized model divided by the square root of the sum of the variances.

To compare the performance of SNN, baseline, and ensemble models to randomized predictions, probabilities were randomly generated from a uniform distribution and assigned to

samples. Average random PR-AUC values were calculated by dividing the number of cases in a testing set by the total number of samples within the testing set. Models that were 'trivially trained' refer to using one-hot encoded feature matrices indicating only the presence or absence (denoted as 1 or 0, respectively) of *de novo* LGD or missense mutation. TPR and FPR values were retrieved for three heuristics, where a sample was classified as a case if the sample possessed an LGD mutation in a gene that was identified: 1) as a high risk or strong candidate ASD gene according SFARI Gene scores (https://gene.sfari.org/database/gene-scoring/), 2) in the prioritized SPARK gene list (https://simonsfoundation.s3.amazonaws.com/share/SFARI/Prioritized%20SPARK%20Gene%20List_for%20distribution_27Apr21.xlsx), or 3) to have elevated intolerance to mutation (pLI >= 0.9, LOEUF < 0.35) (Karczewski et al. 2020).

Assessing enrichment of de novo mutation in NDD cases for ranked NDD risk genes

To determine if enrichment of LGD (or missense) in NDD cases relative to unaffected controls is observed in highly ranked NDD risk genes, the difference in enrichment among NDD cases and controls (E_{diff}) is calculated per gene by **Equation 2**. The total number of LGD (or missense) mutations observed in cases (M_{cases}) for a certain gene within the test set is divided by the number of NDD cases retrieved from denovo-db ($N_{cases} = 9{,}962$), and the total number of LGD (or missense) mutations observed in controls ($M_{controls}$) for that gene within the test set is divided by the number of controls ($N_{controls} = 2{,}245$).

$$E_{diff} = (M_{cases} / N_{cases}) - (M_{controls} / N_{controls})$$ **[Equation 2]**

Web resources

denovo-db, https://denovo-db.gs.washington.edu

Genome Aggregation Database (gnomAD), https://gnomad.broadinstitute.org/downloads

Online Mendelian Inheritance in Man (OMIM), https://www.omim.org

phastCons, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/

Residual Variation Intolerance Score (RVIS), http://genic-intolerance.org/

Scikit-learn, https://scikit-learn.org/stable/

SFARI Gene, https://gene.sfari.org/

References

C Yuen, R. K., Merico, D., Bookman, M., L Howe, J., Thiruvahindrapuram, B., Patel, R. V., et al. (2017). Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nature Neuroscience, 20(4), 602–611. https://doi.org/10.1038/nn.4524

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., et al. (2014). Synaptic, transcriptional, and chromatin genes disrupted in autism. Nature, 515(7526), 209–215. https://doi.org/10.1038/nature13772

Deciphering Developmental Disorders Study. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature, 542(7642), 433–438. https://doi.org/10.1038/nature21062

Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nature Genetics, 46(8), 818–825. https://doi.org/10.1038/ng.3021

Gulsuner, S., Walsh, T., Watts, A. C., Lee, M. K., Thornton, A. M., Casadei, S., et al. (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell, 154(3), 518–529. https://doi.org/10.1016/j.cell.2013.06.049

Hashimoto, R., Nakazawa, T., Tsurusaki, Y., Yasuda, Y., Nagayasu, K., Matsumura, K., et al. (2016). Whole-exome sequencing and neurite outgrowth analysis in autism spectrum disorder. Journal of Human Genetics, 61(3), 199–206. https://doi.org/10.1038/jhg.2015.141

Iossifov, I., O'Roak, B. J., Sanders, S. J., Ronemus, M., Krumm, N., Levy, D., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature, 515(7526), 216–221. https://doi.org/10.1038/nature13908

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., et al. (2015). Excess of rare, inherited truncating mutations in autism. Nature Genetics, 47(6), 582–588. https://doi.org/10.1038/ng.3303

Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., Jin, X., et al. (2012). Whole Genome Sequencing in Autism Identifies Hotspots for De Novo Germline Mutation. Cell, 151(7), 1431–1442. https://doi.org/10.1016/j.cell.2012.11.019

O'Roak, B. J., Stessman, H. A., Boyle, E. A., Witherspoon, K. T., Martin, B., Lee, C., et al. (2014). Recurrent de novo mutations implicate novel genes underlying simplex autism risk. Nature Communications, 5, 5595. https://doi.org/10.1038/ncomms6595

O'Roak, Brian J., Vives, L., Fu, W., Egertson, J. D., Stanaway, I. B., Phelps, I. G., et al. (2012). Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science (New York, N.Y.), 338(6114), 1619–1622. https://doi.org/10.1126/science.1227764

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., & Goldstein, D. B. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics*, *9*(8), e1003709. https://doi.org/10.1371/journal.pgen.1003709

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet (London, England), 380(9854), 1674–1682. https://doi.org/10.1016/S0140-6736(12)61480-9

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005).

Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Sundaram, L., Gao, H., Padigepati, S. R., McRae, J. F., Li, Y., Kosmicki, J. A., et al. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Turner, T. N., Coe, B. P., Dickel, D. E., Hoekzema, K., Nelson, B. J., Zody, M. C., et al. (2017). Genomic Patterns of De Novo Mutation in Simplex Autism. Cell, 171(3), 710-722.e12. https://doi.org/10.1016/j.cell.2017.08.047

Turner, T. N., Hormozdiari, F., Duyzend, M. H., McClymont, S. A., Hook, P. W., Iossifov, I., et al. (2016). Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. American Journal of Human Genetics, 98(1), 58–74. https://doi.org/10.1016/j.ajhg.2015.11.023

Werling, D. M., Brand, H., An, J.-Y., Stone, M. R., Zhu, L., Glessner, J. T., et al. (2018). An analytical framework for whole genome sequence association studies and its implications for autism spectrum disorder. Nature genetics, 50(5), 727–736. https://doi.org/10.1038/s41588-018-0107-y

Yuen, R. K. C., Merico, D., Cao, H., Pellecchia, G., Alipanahi, B., Thiruvahindrapuram, B., et al. (2016). Genome-wide characteristics of de novo mutations in autism. NPJ genomic medicine, 1, 160271–1602710. https://doi.org/10.1038/npjgenmed.2016.27

Chapter 3

Identification of critical cell-types in neurodevelopmental disorders using genetic modules

Abstract

Single-cell technologies continue to identify novel and rare cell-types that deepen our understanding of the mechanisms of disease at a cellular level. A genetic module is a network of genes with correlated gene expression that contribute to common biological pathways. Cell-types that are critical to a certain biological function are defined as a cluster of similar cells that are most "active" for that function. Given a genetic module indicating a biological function, we propose a method, MoToCC, to find a set of similar cells having the most "locally correlated" gene expression for that module. Application of MoToCC on three neurodevelopmental disorder (NDD) modules and single-cell expression data from the human cortex pinpoints migrating excitatory and excitatory deep layer neurons as critical cell-types for autism spectrum disorder (ASD)- and epilepsy-associated modules, respectively.

Introduction

Single-cell and single-nucleus RNA-sequencing (scRNA-seq and snRNA-seq) technologies can reveal the underlying etiology of complex genetic disorders. ScRNA-seq and snRNA-seq have been used to quantify cellular transcriptomic profiles, the latter with special focus on rare cell-types (1,2). Recent single-cell technologies have enabled exploration of molecular mechanisms in a broad range of biological systems, including tumor cells and their associated microenvironments, as well as previously uncharacterized neuronal subtypes, among others (3,4). Compared to bulk RNA-seq analysis, examining gene expression at a single-cell resolution enables

the dissection of cellular heterogeneity and the identification of specific molecular targets for drug intervention, populations of cells with coordinated expression relevant to common biological functions, and the origins of disease pathogenesis (5–7).

Genetic modules are defined as a group of genes with similar biological function that is distinct from other modules. These modules typically consist of genes that are co-expressed and are highly connected in protein-protein interaction networks. Previous module discovery methods have found modules specific to certain phenotypes and pathways (8–11) and have generated modules that are enriched in deleterious mutations for affected cases compared to unaffected controls (12–14). There are many methods for genetic module discovery that use various biological signals. We have previously developed the methods MAGI and MAGI-S for the discovery of neurodevelopmental disorders (NDDs) using a combination of co-expression and protein-interactions network (12,13). The maturity of genetic module discovery methods and the ability to test these modules *in vitro* and *in vivo* has provided a roadmap to the discovery of a growing list of (disease) genetic modules. To better understand the biological function of these modules and use this knowledge in translational studies, it is key to pinpoint the critical cell-types for which each of these genetic modules are most "active". Previously proposed ideas for predicting the critical cell-types of certain diseases have involved linking the selective expression of modules to specific tissue or cell-types (15,16). In this paper, we propose a formal framework for the discovery of critical cell-types associated with a module.

We have developed software named Module To Critical Cell-types (MoToCC), a novel linear programming approach that returns a subset of cells that selectively express module genes for given gene expression data. Users may vary the desired maximum number of cells to return as a solution, permitting the user to visualize relevant, distinct groups of cells that have similar

expression levels at different scales of resolution. MoToCC source code and associated scripts are freely available at https://github.com/jchow32/MoToCC.

Results

MoToCC identifies a set of critical cells, corresponding to relevant cell-types, whose expression of the genes in the genetic module are correlated. Given normalized single-cell (or single-nucleus) gene expression, measures of cell-cell similarity (from a K-nearest neighbor graph (KNN) and a KNN-derived shared nearest-neighbor (SNN) graph), a genetic module *M*, and an upper bound (*k*) of cells to select, MoToCC selects a subset of cells that maximize the "local correlation" of cell-cell gene expression (15) in module genes *M* while imposing constraints on similarity of cells selected (**Figure 1**).

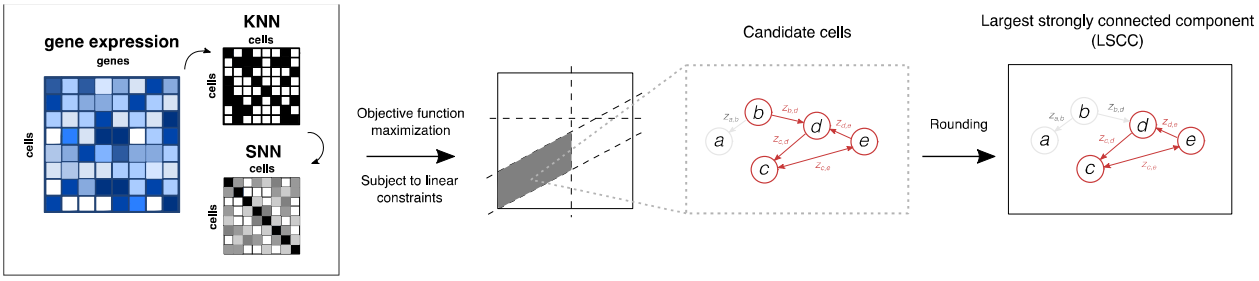


Figure 1. **Overview of** MoToCC. MoToCC uses normalized single-cell gene expression and K nearest-neighbor (KNN) and shared nearest-neighbor (SNN) graphs to select an upper bound of *k* cells that selectively express the genes of a genetic module. Between any two cells *i* and *j*, the edge weight $z_{i,j}$ is calculated as the product of cell-cell similarity from the SNN and the sum of gene expression values for cells *i* and *j* for pairwise combinations of module genes. The objective function maximizes the summation of the edge weight $z_{i,j}$ and the indicator variable $x_{i,j}$ while subject to linear constraints. Following maximization of the objective function, MoToCC returns an initial solution of candidate cells with non-

zero objective function values. A set of cells (a-e) are pictured, where an arrow between two cells represents the associated edge weight with directionality determined by KNN connectivity. Cells and edges with non-zero indicator variable values are highlighted in red. To refine the initial solution, the largest strongly connected component (LSCC) is identified among the candidate cells according to KNN connectivity, and the set of critical cells within the LSCC (red) are returned. Dimensionality techniques may then be used to visualize the set of selected cells to reveal select expression relative to the genetic module.

We evaluated the performance of MoToCC in identifying critical cell-types associated with modules found for NDDs. We used single-cell RNA-seq (scRNA-seq) data from the human cerebral cortex (16), consisting of 33,986 cells previously assigned to 16 cell-types (**Supplementary Data**). Three relevant NDD modules were generated using MAGI and MAGI-S tools (12,14), including the M1 (autism spectrum disorder (ASD), intellectual disability (ID)), M2 (ASD, ID), and *SCN1A* (epilepsy) modules (12,14) (**Supplementary Table 1, Supplementary Data**).

The proposed solution has a user-defined upper bound on the number of cells to return, parameterized as *k.* In our experiments for each module, *k* was varied from 250 to 5,000 cells in intervals of 250 cells, and the silhouette score is calculated after applying K-means clustering to the two-dimensional t-SNE of the normalized gene expression for cells in the LSCC (**Supplementary Data, Supplementary Table 2**).

For all *k*, the total objective function value and each cell's associated indicator variable value associated with the i) initial solution of candidate cells with non-zero indicator variables and the ii) final refined solution of cells in the largest strongly connected component (LSCC) were returned (**Supplementary Tables 2 and 3**). The percent composition of cells in the LSCC for each cell-type was calculated for each *k* (**Supplementary Table 4**).

Neurodevelopmental disorders critical cell-types

We first investigated the critical cell-types impacted in NDDs. We first consider the main NDD modules (M1, M2) found using MAGI (14). We also consider the epilepsy module found using MAGI-S with *SCN1A* as a seed gene, hereon referred to as M_*SCN1A*. The M1 module (80 genes) is significantly functionally enriched in chromatin remodeling and the Wnt pathway, while the M2 module (19 genes) is significantly enriched in chemical synaptic transmission and long-term potentiation (14). The module M_*SCN1A* (36 genes) is significantly enriched in non-synonymous *de novo* mutations from epilepsy cohorts, known epilepsy genes, and pathways such as long-term potentiation, chemical synaptic transmission, and regulation of neurotransmitter activity (12). All NDD modules (M1, M2, M_*SCN1A*) possess a significantly larger initial and final total objective function value at all *k* compared to modules of the same size consisting of random genes (**Supplementary Table 2, Supplementary Data**). The average normalized gene expression values of module genes per previously assigned cell-type are represented via hierarchically clustered heatmaps in **Supplementary Figure 1**.

Neurodevelopmental disorder module - M1

The M1 module is functionally enriched in terms related to regulation of transcription and chromatin remodeling rather than neurotransmitter secretion. CSEA does not identify any significant enrichment of the M1 module in any specific cell-type (17). For the M1 modules at all *k*, the migrating excitatory (ExN) cell-type predominates (**Figure 2**), which is consistent with the role of migrating excitatory neurons in neuronal differentiation and projection development (**Supplementary Table 1**).

Large increases in silhouette scores are indicative of disparate clustering and potential shifts in percent composition. For example, a large, positive difference in silhouette score between two sequential solutions, such as *k* and *k + 250*, indicates that the LSCC at *k* represents a group of cells with a more similar degree of selective expression for the module. For the M1 module, the largest increases in silhouette score are highlighted in red (**Figure 2**). Two-dimensional t-SNE plots for *k* immediately before (750) and after (1,750) the largest increases in silhouette score are also displayed, with cells in the LSCC pictured in red. Full two-dimensional and three-dimensional t-SNE and UMAP plots of cells in the LSCC for all modules at all tested *k* in **Supplementary Figure 3**.

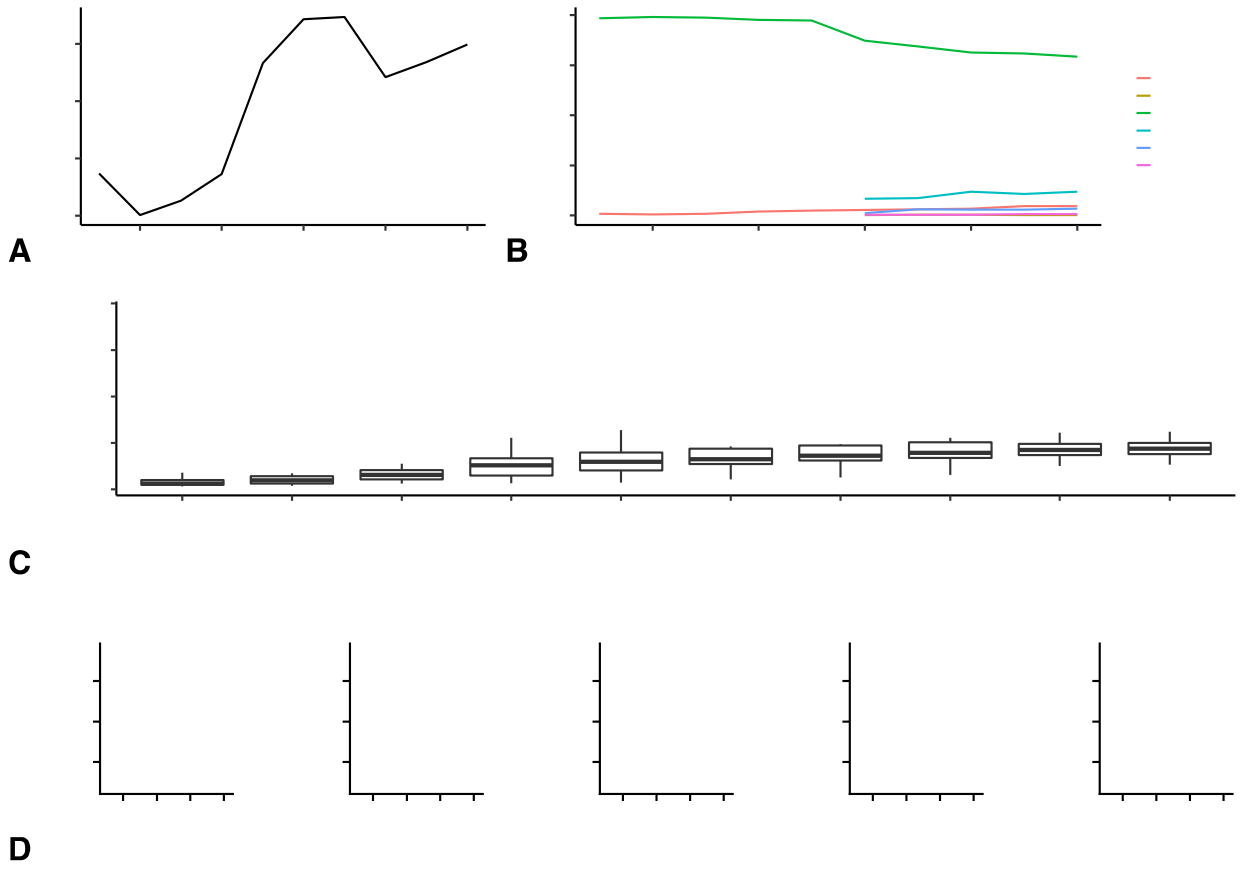


Figure 2. **Silhouette scores, percent composition, final objective function values, and critical cells at varied k for M1**. The upper bound on the final number of cells to return as critical cells, k, is varied from

250 to 5,000 in steps of 250 cells. A) K-means was applied to the two-dimensional t-SNE of critical cells at each k, and silhouette scores were calculated. The largest increases in silhouette scores at k = (1,000; 1,250; 1,500) are highlighted in red. B) Percent composition, defined as the fraction of cells of a certain cell-type among critical cells, are shown. For the M1 module, selected cells are primarily of the migrating excitatory (ExN) cell-type. Cell-type names are abbreviated according to Supplementary Table 4. C) The final objective function of critical cells for the M1 module (red) are compared to values from 20 same-sized modules consisting of random genes (black). At all k, the corresponding final total objective function value of the M1 module is significantly greater than that of the randomized modules (Supplementary Table 2). D) Two-dimensional t-SNE plots including selected critical cells (red) at k corresponding to the greatest increases in silhouette score (red title).

Neurodevelopmental disorder module - M2

At all supplied *k*, critical cells for the M2 module primarily consist of excitatory deep layer 1 and 2 neurons (ExDp1, ExDp2) (**Figure 3**). In fact, more than 63% of all cells belonging to the ExDp2 cell-type are selected when *k* = 500, and ExDp2 percent capture increases to 87% as *k* increases for M2. The elevated percent composition for ExDp1 and ExDp2 in the M2 module is consistent with functional enrichment analyses (**Supplementary Table 1**) which reveal enrichment in synaptic transmission and regulation of neurotransmitter receptor and cation channel activity, and complements the elevated ExDp2 expression levels of pertinent genes such as *GABRB2*, *GRIN2B*, and *STXBP1* (**Supplementary Figure 1**). CSEA (17) also highlights the relevance of M2 genes in deep cortical neurons (**Supplementary Figure 2**).

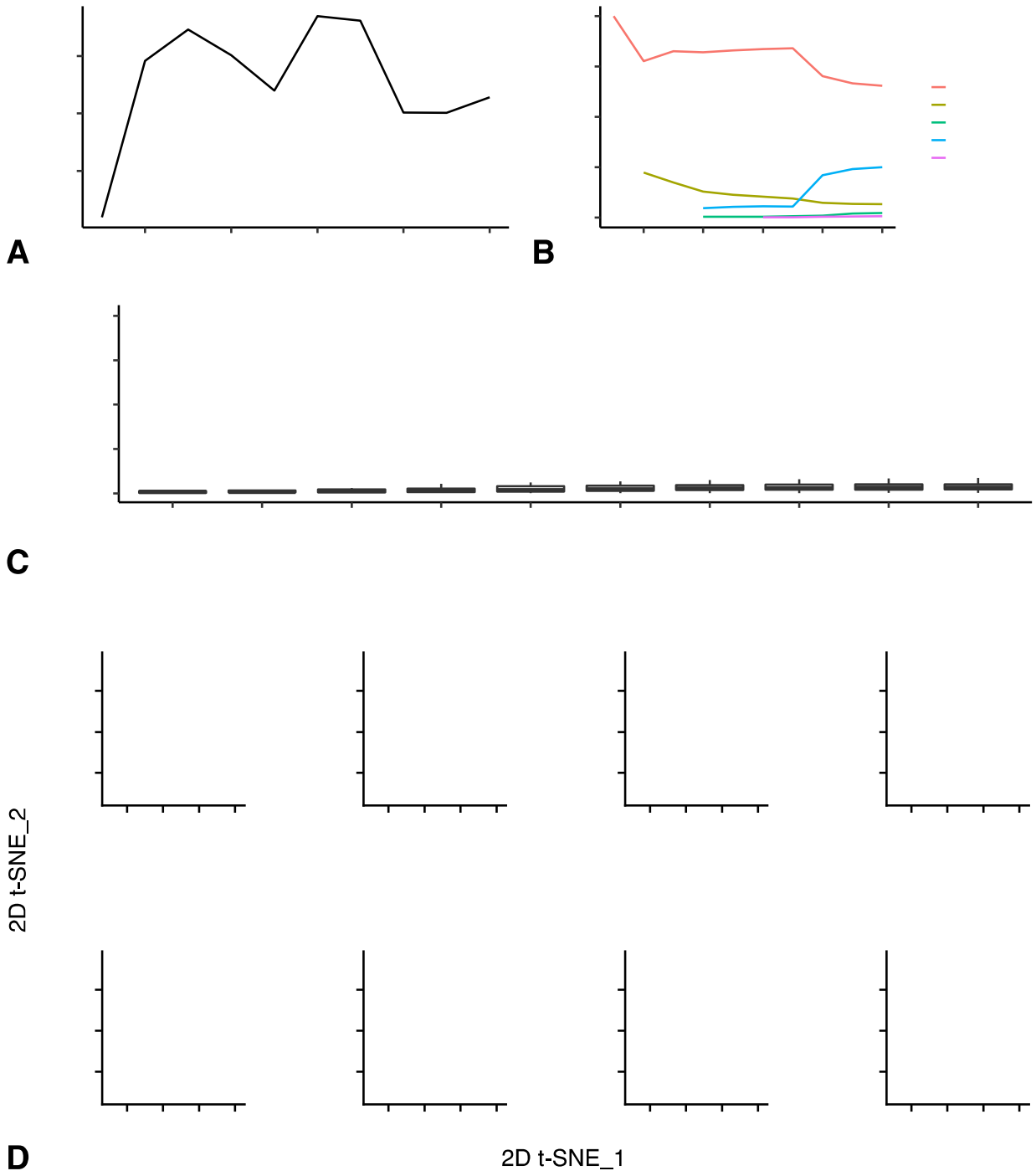


Figure 3. **Silhouette scores, percent composition, final objective function values, and critical cells at varied *k* for M2**. A) The largest increases in silhouette scores at k = (500; 1,500; 4,500) are highlighted in red. B) For the M2 module, selected cells are primarily of the excitatory deep layer 1 and 2 (ExDp1, ExDp2) cell-types. Cell-type names are abbreviated according to **Supplementary Table 4**. C) The final objective

function of critical cells for the M2 module (red) are compared to values from 20 same-sized modules consisting of random genes (black). At all *k*, the corresponding final total objective function value of the M2 module is significantly greater than that of the randomized modules (**Supplementary Table 2**). D) Two-dimensional t-SNE plots including selected critical cells (red) at *k* corresponding to the greatest increases in silhouette score (red title); additional plots are available for all *k* (250 to 5,000) in **Supplementary Figure 3**.

Epilepsy module - M_SCN1A

The M_*SCN1A* and M2 modules display similar functional enrichment due to the large proportion (>30%) of shared genes among modules (**Supplementary Table 1, Supplementary Figure 2**). Like M2, critical cells of the M_*SCN1A* module are primarily labeled as excitatory deep layer neurons. However, at smaller values of *k*, the M_*SCN1A* module's critical cells initially consist of most existing ExDp2 cells (**Figure 4**). Increased ExDp2 percent capture is observed from $k = 1{,}000$ up to a percent capture of 87%. Percent composition of ExDp1 and ExDp2 cells necessarily decreases because of high percentage of capture when *k* is larger than the total number of cells of a certain cell-type (ExDp1: 2,039; ExDp2: 166) (**Supplementary Table 4**). When *k* is increased to $k > 2{,}000$, the maturing excitatory upper enriched (ExM-U) cell-type constitutes approximately 25% of critical cells.

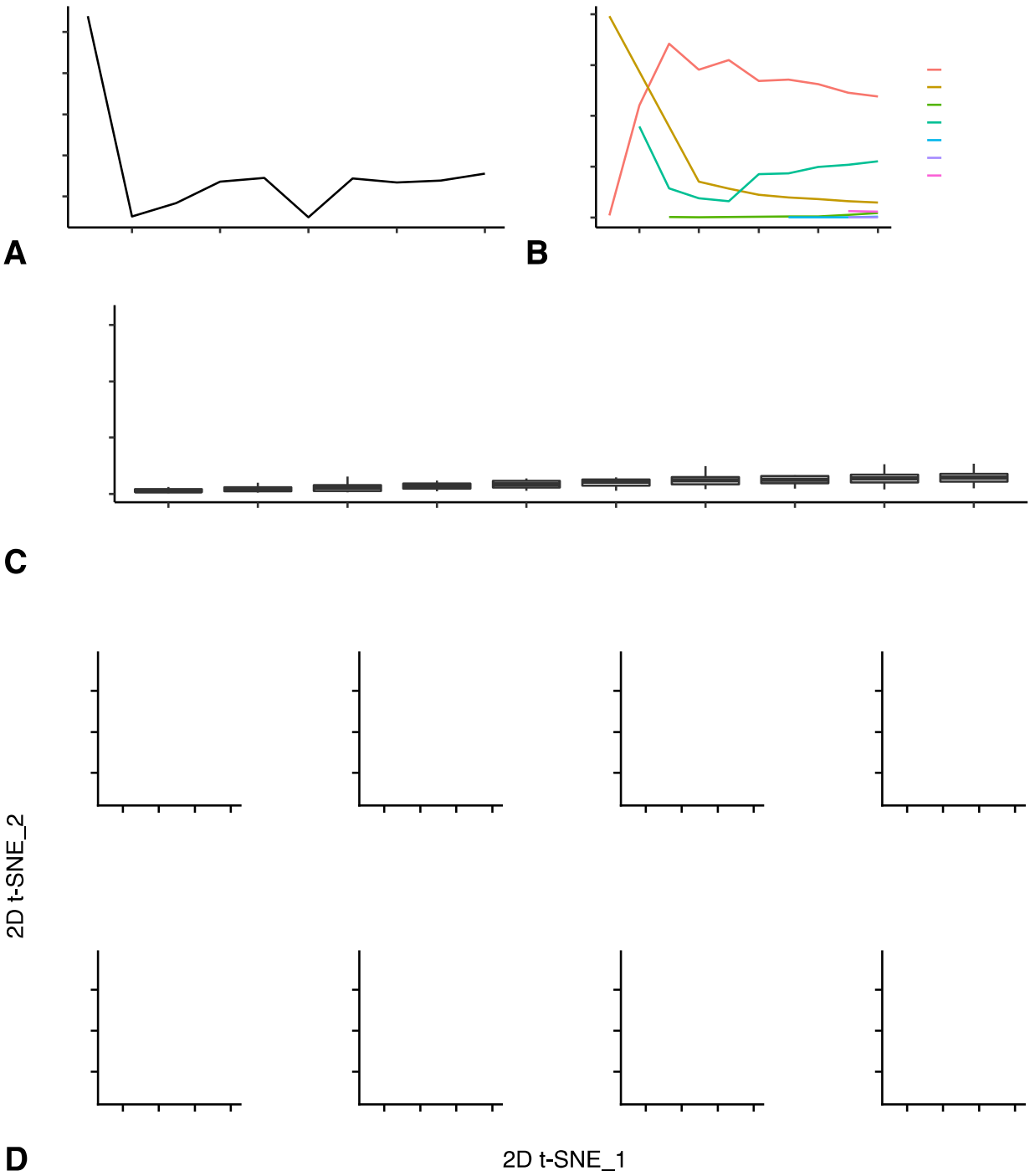


Figure 4. **Silhouette scores, percent composition, final objective function values, and critical cells at varied *k* for M_*SCN1A***. A) The largest increases in silhouette scores at k = (1,000; 1,750; 4,750) are highlighted in red. B) For the M_*SCN1A* module, selected cells are primarily of the excitatory deep layer 1 and 2 (ExDp1, ExDp2) cell-types at lower k, followed by increased selection of the maturing excitatory

upper enriched (ExM-U) cell-type. Cell-type names are abbreviated according to **Supplementary Table 4**. C) The final objective function of critical cells for the M_*SCN1A* module (red) are compared to values from 20 same-sized modules consisting of random genes (black). At all *k*, the corresponding final total objective function value of the M_*SCN1A* module is significantly greater than that of the randomized modules (**Supplementary Table 2**). D) Two-dimensional t-SNE plots including selected critical cells (red) at *k* corresponding to the greatest increases in silhouette score (red title); additional plots are available for all *k* (250 to 5,000) in **Supplementary Figure 3**.

Discussion

MoToCC uses a linear programming approach to identify a set of cells that selectively express genes from a given genetic module. Briefly, MoToCC maximizes an objective function related to the correlated co-expression ("local correlation") among module genes and cell-cell similarity to yield an initial solution. The initial solution is refined using the derived K-nearest neighbor graph to select the largest strongly connected component (LSCC) among cells in the initial solution. Thus, the selected critical cells in the LSCC consist of cells that share a high degree of co-expression in module genes and display similar overall gene expression profiles compared to all other cells in the dataset.

We provided MoToCC with normalized single-cell gene expression values from the adult human cortex and three modules relevant to neurodevelopmental disorder (M1, M2, M_*SCN1A*). For each module, by varying the upper bound (*k*) of the number of cells to return as a solution from 250 to 5,000 cells, we identify breakpoints at which large changes in silhouette score indicate shifts in cellular composition. Percent composition was calculated according to previously assigned cell-type labels (16). Fluctuations in percent composition tend to be preceded by increases in silhouette scores. For example, among the largest increases of silhouette scores for each module

(M1: [1,250; 1,500], M2: [500; 1,500], M_*SCN1A*: [4,750; 1,750]), shifts in percent composition occur at *k* - 250 relative to the *k* associated with the large silhouette score (**Figure 2, Figure 3, Figure 4**). Visualizations of the selected LSCCs associated with each *k* are shown via dimensionality reduction plots, including 2D and 3D t-SNE and UMAP plots (**Supplementary Figure 3**).

We observe that in general, for smaller *k* less than 1,000, a single cell-type tends to predominate except in the case where the total number of cells of a certain cell-type is fewer than *k*. This suggests that at small *k*, the cell-type with the largest percent composition may be the most relevant cell-type to the provided module. At larger *k*, cell-types that are more populous and thus could potentially contain more cells that are members of a strongly connected component are more likely to be returned as solutions. In addition, at all *k*, the total objective function values associated with initial and final solutions are significantly smaller for randomized modules of the same size as true modules (**Supplementary Table 2**), suggesting that cells were non-randomly selected according to their correlated gene expression. In general, we find that MoToCC identifies sets of cells for each module that support existing associations between cell-types and patterns of gene expression, and that breakpoints located via large increases in silhouette scores can emphasize clusters of cells most relevant to the module and the targeted phenotype.

M1 displays widespread expression throughout the human brain during neurodevelopment and shows no evidence of localized enrichment as per the CSEA tool. Yet, the M1 module is significantly enriched in functions related to chromatin remodeling and the Wnt and Notch pathways and in *de novo* loss of function and missense mutation in neurodevelopmental disorder cases relative to unaffected siblings, with significantly increased expression in the fetal brain (14). The migrating excitatory neuron (ExN) forms the majority of cells selected over all *k*, especially

at low $k < 1{,}250$ where percent composition of ExN exceeds 97%. Neuronal migration and genes in the M1 module appear to be highly related; for example, in the M1 module, genes such as *BCL11A, CREB1, CTNNB1, CUL3*, among others, have been shown to regulate cell polarity, migration, and the timing of neurogenesis and differentiation (18–21). The largest silhouette score breakpoints are clustered from $k = 1{,}000$ to 1,500 and correspond to the first decrease in ExN percent composition below 97% for the M1 module. Dimensionality reduction plots corroborate the selection of groups of cells with dissimilar gene expression, particularly between $k = 1{,}250$ to 1,500 (**Supplementary Figure 3**).

Several genes in the M2 and *SCN1A* modules are known to be associated with synaptic transmission or belong to gene families associated with neurotransmitter receptors, such as the DLGAP (*DLGAP1*), GABA-A (*GABRB2*, *GABRG2*), and GRIN (GRIN1, *GRIN2A*, *GRIN2B*) gene families (22–24). Both M2 and M_*SCN1A* modules were most significantly enriched in terms related to the regulation of neurotransmitter secretion and were found to be selectively expressed in deep cortical neurons via CSEA (**Supplementary Table 1, Supplementary Figure 2**), which is complemented by the established importance of synaptic connectivity among deep cortical neurons for functions such as memory formation and perception, among others (25,26). For M2, primarily ExDp1 and ExDp2 cell-types were selected, capturing more than 63% of all ExDp2 cells. The subsets of cells selected by the M_*SCN1A* module resemble those of the M2 module, especially at increased k where cells of type maturing excitatory upper enriched (ExM-U) constitute more than 20% of selected cells. However, the M_*SCN1A* module primarily selects ExDp2 cells at low k and does not provide identical solutions to M2 as is apparent in **Supplementary Figure 3**.

By providing MoToCC with single-cell expression data and genetic modules related to neurodevelopmental disorders, we demonstrate MoToCC ability to select biologically relevant groups of cells that correspond to cell-types that selectively express module genes. The identification of distinct groups of cells depends on the user-defined upper bound (*k*) of the number of cells to select and the dataset's unique cellular composition. Therefore, multiple iterations of MoToCC with varied *k* are recommended to reveal clustering breakpoints that describe cells with dissimilar patterns of gene expression. Example commands, guidelines, and associated scripts for pre-processing and data visualization are freely available at https://github.com/jchow32/MoToCC.

Methods

To identify a subset of critical cells that are selectively expressed in a given genetic module, we propose a linear programming approach, MoToCC. MoToCC takes as input normalized gene expression, a similarity matrix (Shared Nearest-Neighbor, SNN), a K nearest-neighbor (KNN) graph, corresponding cell and gene labels, a user-defined number of cells (*k*), and a genetic module *M*.

1. Data pre-processing

ScRNA-seq gene expression data for the developing human cortex were downloaded and normalized (16) (**Supplementary Data)**. Modules provided to MoToCC were retrieved from MAGI or MAGI-S (12,14) (**Supplementary Data, Supplementary Table 1**). The M1 ('ASD plus ID'; 80 genes) and M2 ('ASD with ID'; 19 genes) modules from MAGI (14) and a module seeded with *SCN1A*, together referred to as 'neurodevelopmental disorder modules', were used to identify three different critical subsets of cells in the human cortex dataset.

2. Formulation and Algorithm

Given a genetic module *M* and a user-defined parameter *k*, the objective of the proposed approach is to select a maximum of *k* cells that (i) are most "active" in genes in module *M* and (ii) cells are similar based on available single-cell expression data over all genes. Our definition of activity is motivated by the local correlation between cells (15). As this problem as defined is intractable, we propose a two step heuristic. The high-level solution we are proposing in the first step finds the most locally correlated set of cells given the module *M,* and then, in the second step, uses the selected cells to find a strongly connected component in the KNN graph (built using all genes).

2.1 Variable definitions

We model this problem of selecting a subset of cells that are most "active" given a gene module *M* as a graph problem. We first represent each cell as a node in a graph and assume for every pair of nodes *i* and *j* there exists an input weight ($w_{i,j}$) indicating the cell-to-cell similarity as retrieved from the SNN graph. Furthermore, given the genes in the input module *G* we pre-calculate the weighted pairwise local correlation score (15) defined as $z_{i,j} = \sum_{n \in G} \sum_{(n \neq p) \in G} w_{i,j}(n_i p_j + p_i n_j)$, where n_i, n_j, p_i, and p_j represent the normalized expression of genes *n* and *p* (from gene input module *G*) in cells *i* and *j* respectively. The objective of the first stage of our approach is to select a maximum of *k* cells such that the summation of $z_{i,j}$ for every pair of cells *i* and *j* selected is maximized. We note that the above problem is indeed *NP*-complete and provide a heuristic using linear programming for solving it.

2.2 Objective function and linear-programming formulation

The objective of the first step of MoToCC is as follows. Given an upper bound *k* to select a set of at most *k* cells, the summation of $z_{i,j}$ for all pairs of selected cells/nodes is maximized. For each cell/node *i* we define variable y_i to indicate that the cell/node is selected. For every pair of cells/nodes (*i, j*) we also define the variable $x_{i,j}$ to indicate that both pair of nodes *i* and *j* are selected (i.e., y_i=1 and y_j=1). We relax the formulation to a linear programming problem:

$$maximize \sum_{(i,j)\in E} z_{i,j}x_{i,j}$$

Subject to:

$$0 \leq x_{i,j} \leq 1 \; \forall (i,j) \in E$$

$$\sum_{i\in C} y_i \leq k$$

$$x_{i,j} \leq y_j$$

$$x_{i,j} \leq y_i$$

$$x_{i,j} \geq y_i + y_j - 1$$

2.3 Rounding and selection of critical cells

The provided real solution must be rounded to an integer value solution. Furthermore, we also need to satisfy that the selected cells must be strongly connected in the KNN graph. Thus, we propose a simple rounding solution that also imposes the connectivity condition for the selected cells in the KNN graph. First, we select the initial solution returned by linear programming of cells with $y_i > 0$ is referred to as potential 'candidate cells'. Next, among the candidate cells, the largest strongly connected component (LSCC) (networkx version 1.11) in the KNN graph is rounded to 1

and returned as the final solution. Finally, the associated objective function values for cells in the LSCC are returned as the final solution.

2.4. Implementation notes

To maximize the objective function (Section 2.2), edge weights ($z_{i,j}$) are calculated between cells that have non-zero similarity ($w_{i,j}$) as per the SNN graph. If pruning (--prune) is enabled, only edge weights outside of one standard deviation from the mean edge weight are retained. For a module, edge weights only need to be calculated once. Thus, if the user wishes to run MoToCC using varied *k*, the quickstart parameter (--quickstart) can be enabled to load edge weights previously calculated by MoToCC using the same module.

2.5. Return and refinement of initial solution

MoToCC returns the silhouette score associated with a 2D t-SNE dimensionality reduction and K-means clustering (K=2) of cells in the LSCC. Given multiple silhouette scores for varied *k*, breakpoints at which distinct groups of cell-types are selected can be viewed. To visualize the selected cells of the refined solution compared to unselected cells (via 2D t-SNE, 3D t-SNE, and UMAP) and to plot silhouette scores versus varied *k*, additional scripts and their usage are described at https://github.com/jchow32/MoToCC.

References

1. Wu H, Kirita Y, Donnelly EL, Humphreys BD. Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis. J Am Soc Nephrol [Internet]. 2019 Jan [cited 2022 Apr 28];30(1):23–32. Available from: https://jasn.asnjournals.org/lookup/doi/10.1681/ASN.2018090912
2. Bakken TE, Hodge RD, Miller JA, Yao Z, Nguyen TN, Aevermann B, et al. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLOS ONE [Internet]. 2018 Dec 26 [cited 2022 Apr 28];13(12):e0209648. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209648

3. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci [Internet]. 2019 Jun 26 [cited 2022 Apr 29];9(1):53. Available from: https://doi.org/10.1186/s13578-019-0314-y
4. Anaparthy N, Ho YJ, Martelotto L, Hammell M, Hicks J. Single-Cell Applications of Next-Generation Sequencing. Cold Spring Harb Perspect Med [Internet]. 2019 Oct [cited 2022 Apr 29];9(10):a026898. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6771363/
5. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet [Internet]. 2019 [cited 2022 Apr 28];10. Available from: https://www.frontiersin.org/article/10.3389/fgene.2019.00317
6. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med [Internet]. 2017 Aug 18 [cited 2022 Apr 28];9(1):75. Available from: https://doi.org/10.1186/s13073-017-0467-4
7. Gawel DR, Serra-Musach J, Lilja S, Aagesen J, Arenas A, Asking B, et al. A validated single-cell-based strategy to identify diagnostic and therapeutic targets in complex diseases. Genome Med [Internet]. 2019 Jul 30 [cited 2022 Apr 28];11:47. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6664760/
8. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell [Internet]. 2013 Nov 21 [cited 2022 Apr 29];155(5):1008–21. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3934107/
9. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform [Internet]. 2018 Jul 20 [cited 2022 Apr 29];19(4):575–92. Available from: https://doi.org/10.1093/bib/bbw139
10. Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet [Internet]. 2013 Oct [cited 2022 Apr 29];14(10):719–32. Available from: http://www.nature.com/articles/nrg3552
11. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun [Internet]. 2018 Mar 15 [cited 2022 Apr 29];9(1):1090. Available from: https://www.nature.com/articles/s41467-018-03424-4
12. Chow J, Jensen M, Amini H, Hormozdiari F, Penn O, Shifman S, et al. Dissecting the genetic basis of comorbid epilepsy phenotypes in neurodevelopmental disorders. Genome Med [Internet]. 2019 Oct 25 [cited 2022 Apr 8];11(1):65. Available from: https://doi.org/10.1186/s13073-019-0678-y
13. Chow JC, Zhou R, Hormozdiari F. MAGI-MS: multiple seed-centric module discovery. Bioinforma Adv [Internet]. 2022 Jan 1 [cited 2022 Apr 29];2(1):vbac025. Available from: https://doi.org/10.1093/bioadv/vbac025
14. Hormozdiari F, Penn O, Borenstein E, Eichler EE. The discovery of integrated gene networks for autism and related disorders. Genome Res. 2015 Jan;25(1):142–54.
15. DeTomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. Functional interpretation of single cell similarity maps. Nat Commun [Internet]. 2019 Sep 26 [cited 2022 Apr 22];10(1):4376. Available from: https://www.nature.com/articles/s41467-019-12235-0
16. Polioudakis D, de la Torre-Ubieta L, Langerman J, Elkins AG, Shi X, Stein JL, et al. A Single-Cell Transcriptomic Atlas of Human Neocortical Development during Mid-gestation. Neuron. 2019 Sep 4;103(5):785-801.e8.
17. Xu X, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell Type-Specific Expression Analysis to Identify Putative Cellular Mechanisms for Neurogenetic Disorders. J Neurosci [Internet]. 2014 Jan 22 [cited 2022 Apr 29];34(4):1420–31. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3898298/
18. Wiegreffe C, Simon R, Peschkes K, Kling C, Strehle M, Cheng J, et al. Bcl11a (Ctip1) Controls Migration of Cortical Projection Neurons through Regulation of Sema3c. Neuron [Internet]. 2015

Jul 15 [cited 2022 Apr 27];87(2):311–25. Available from: https://www.sciencedirect.com/science/article/pii/S0896627315005668
19. Díaz-Ruiz C, Parlato R, Aguado F, Ureña JM, Burgaya F, Martínez A, et al. Regulation of neural migration by the CREB/CREM transcription factors and altered Dab1 levels in CREB/CREM mutants. Mol Cell Neurosci [Internet]. 2008 Nov 1 [cited 2022 Apr 27];39(4):519–28. Available from: https://www.sciencedirect.com/science/article/pii/S1044743108001954
20. Morandell J, Schwarz LA, Basilico B, Tasciyan S, Dimchev G, Nicolas A, et al. Cul3 regulates cytoskeleton protein homeostasis and cell migration during a critical window of brain development. Nat Commun [Internet]. 2021 May 24 [cited 2022 Apr 27];12(1):3058. Available from: https://www.nature.com/articles/s41467-021-23123-x
21. Bocchi R, Egervari K, Carol-Perdiguer L, Viale B, Quairiaux C, De Roo M, et al. Perturbed Wnt signaling leads to neuronal migration delay, altered interhemispheric connections and impaired social behavior. Nat Commun [Internet]. 2017 Oct 27 [cited 2022 Apr 27];8(1):1158. Available from: https://www.nature.com/articles/s41467-017-01046-w
22. Rasmussen AH, Rasmussen HB, Silahtaroglu A. The DLGAP family: neuronal expression, function and role in brain disorders. Mol Brain [Internet]. 2017 Sep 4 [cited 2022 Apr 27];10(1):43. Available from: https://doi.org/10.1186/s13041-017-0324-9
23. Collins AL, Ma D, Whitehead PL, Martin ER, Wright HH, Abramson RK, et al. Investigation of autism and GABA receptor subunit genes in multiple ethnic groups. Neurogenetics [Internet]. 2006 Jul [cited 2022 Apr 27];7(3):167–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1513515/
24. Myers SJ, Yuan H, Kang JQ, Tan FCK, Traynelis SF, Low CM. Distinct roles of GRIN2A and GRIN2B variants in neurological conditions. F1000Research [Internet]. 2019 Nov 20 [cited 2022 Apr 27];8:F1000 Faculty Rev-1940. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6871362/
25. Tritsch NX, Sabatini BL. Dopaminergic modulation of synaptic transmission in cortex and striatum. Neuron [Internet]. 2012 Oct 4 [cited 2022 Apr 27];76(1):33–50. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4386589/
26. Rollenhagen A, Ohana O, Sätzler K, Hilgetag CC, Kuhl D, Lübke JHR. Structural Properties of Synaptic Transmission and Temporal Dynamics at Excitatory Layer 5B Synapses in the Adult Rat Somatosensory Cortex. Front Synaptic Neurosci [Internet]. 2018 [cited 2022 Apr 27];10. Available from: https://www.frontiersin.org/article/10.3389/fnsyn.2018.00024
27. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res [Internet]. 2019 Jan 8 [cited 2022 Apr 8];47(Database issue):D607–13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323986/
28. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res [Internet]. 2016 Jul 8 [cited 2022 Apr 21];44(Web Server issue):W90–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4987924/

Supplementary Data

Supplementary Methods

Data preprocessing

Single-cell RNA-seq (scRNA-seq) data describing gene expression for the developing human cortex (16) (http://geschwindlab.dgsom.ucla.edu/pages/codexviewer) was downloaded. To normalize and scale gene expression data, the Seurat (version 3.2.1) functions NormalizeData() and ScaleData() were used, and the transpose of the resulting data frame of size (cells x genes) was saved. Similarity and K nearest-neighbor matrices were generated following normalization, identification of variable features via the FindVariableFeatures() function and scaling by using the FindNeighbors() function with the 'compute.SNN=TRUE' parameter enabled.

After successful neighbor finding via Seurat (version 3.2.1), normalized scRNA-seq gene expression values and the associated similarity matrix (Shared Nearest-Neighbor, SNN) and K Nearest-Neighbor (KNN) graph were compressed using pandas (version 0.25.0) and scipy (version 1.5.2) for Python 3.6. The SNN and KNN were stored as sparse matrices. The assigned cell label (identity) of each cell was downloaded (http://geschwindlab.dgsom.ucla.edu/pages/codexviewer) (16).

After clustering with FindClusters(), three dimensionality reduction techniques were used, including principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) to visualize the clustering of cells based on their gene expression profiles and to highlight selected cells following termination of MoToCC. Preprocessing scripts are available for the downloaded cortex dataset at https://github.com/jchow32/MoToCC.

Genetic module functional enrichment and randomized modules

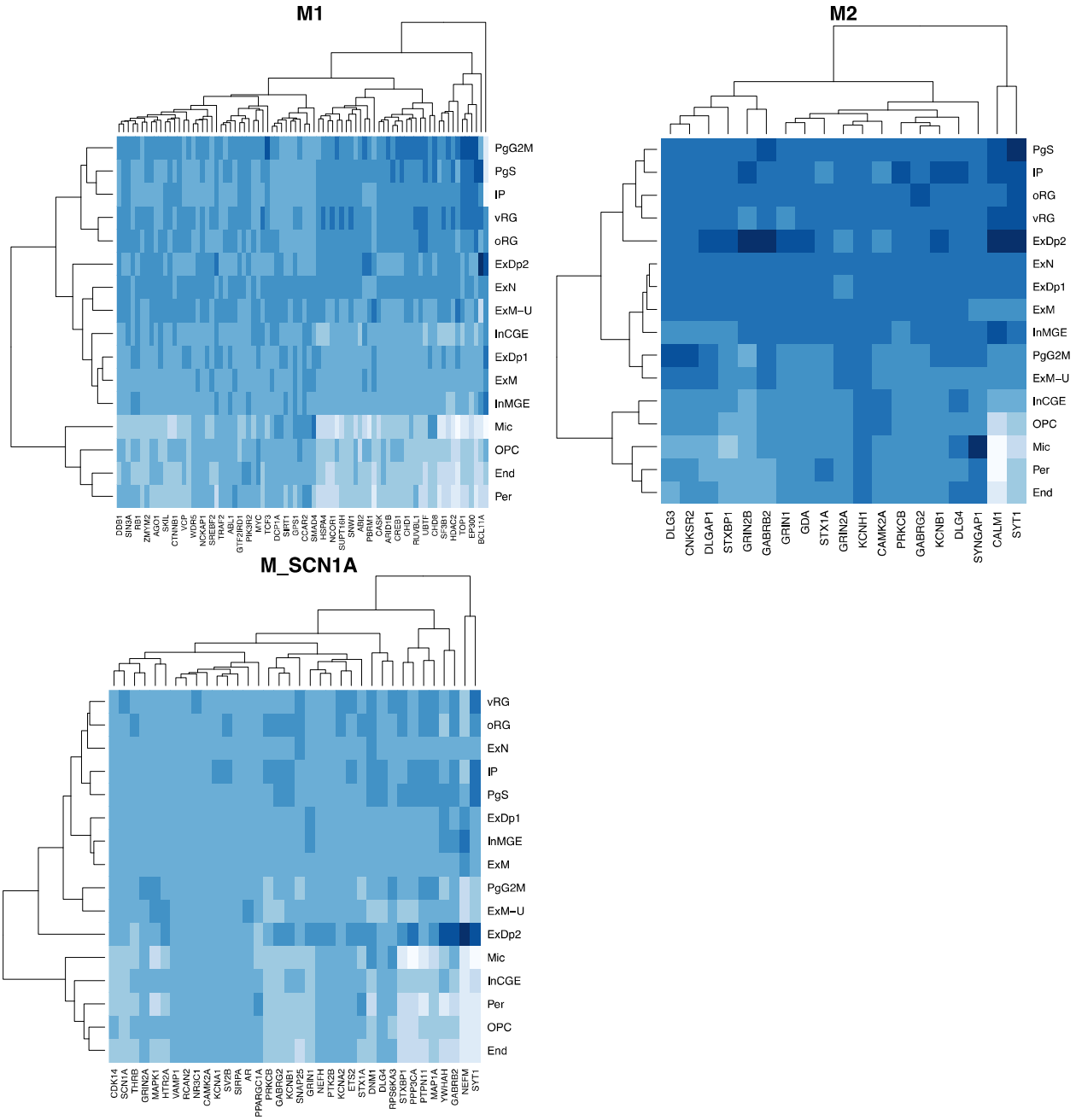
Enrichment terms, including Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) Biological Processes, and Online Mendelian Inheritance in Man (OMIM) Expanded terms, were retrieved from the tool Enrichr (28) for each module.

Twenty randomized modules containing the same corresponding number of genes were created for each module (M1, M2, M_*SCN1A*), each subjected to variable *k* ranging from 250 to 5,000 in steps of 250. Directional one-sample t-tests were used to determine if the average total initial and final objective function values of the corresponding randomized modules were significantly smaller than those of the true modules. Associated p-values and total average objective function values are displayed in **Supplementary Table 2** for each true module.

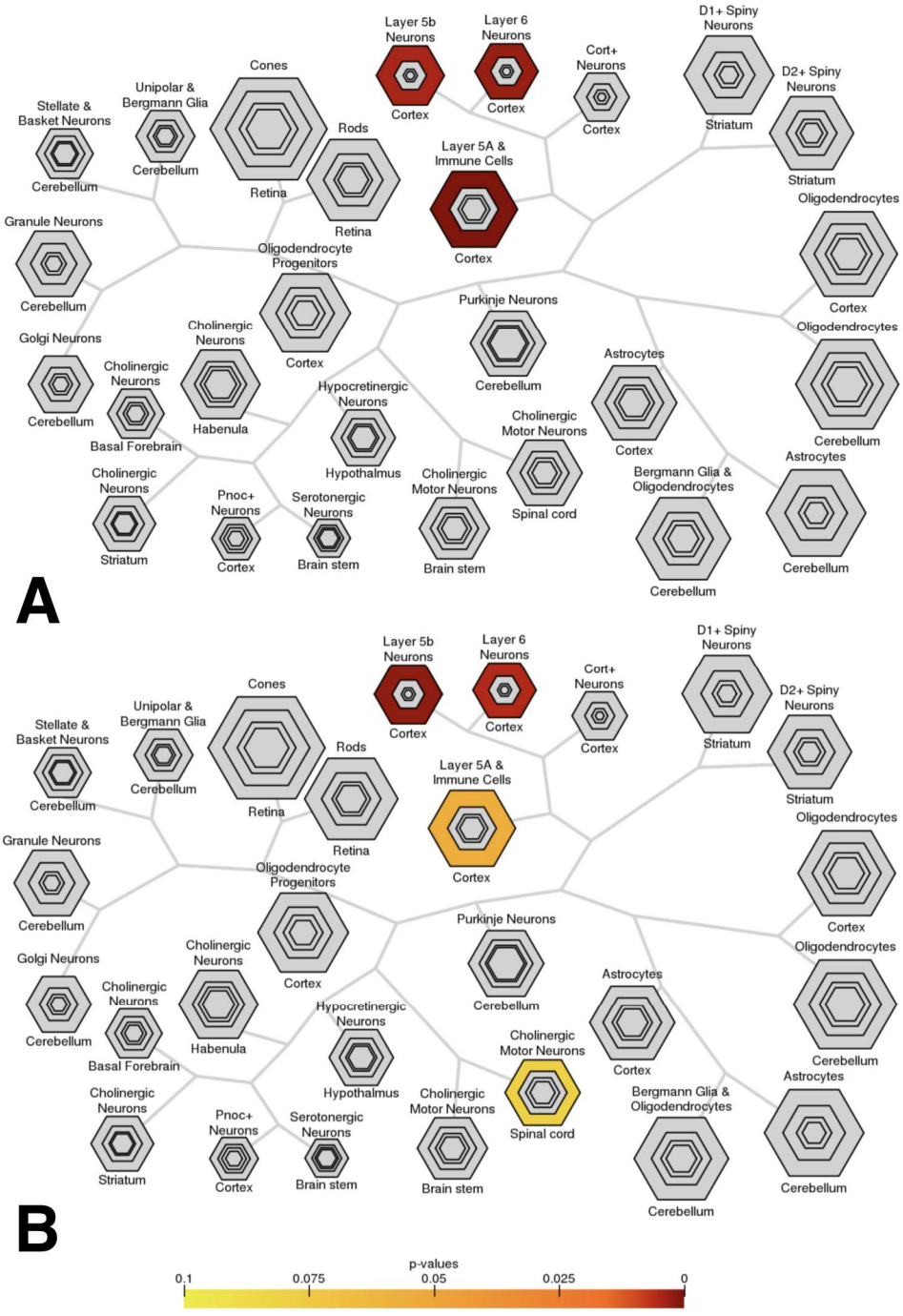
Specifications and runtime

MoToCC was developed and tested on Linux 4.15.0-142-generic x86_64 (model: AMD Opteron(tm) Processor 6380, CPU Mhz: 1396.336). Packages installed via Anaconda (4.10.3) for Python (3.4.5) are listed at https://github.com/jchow32/MoToCC.

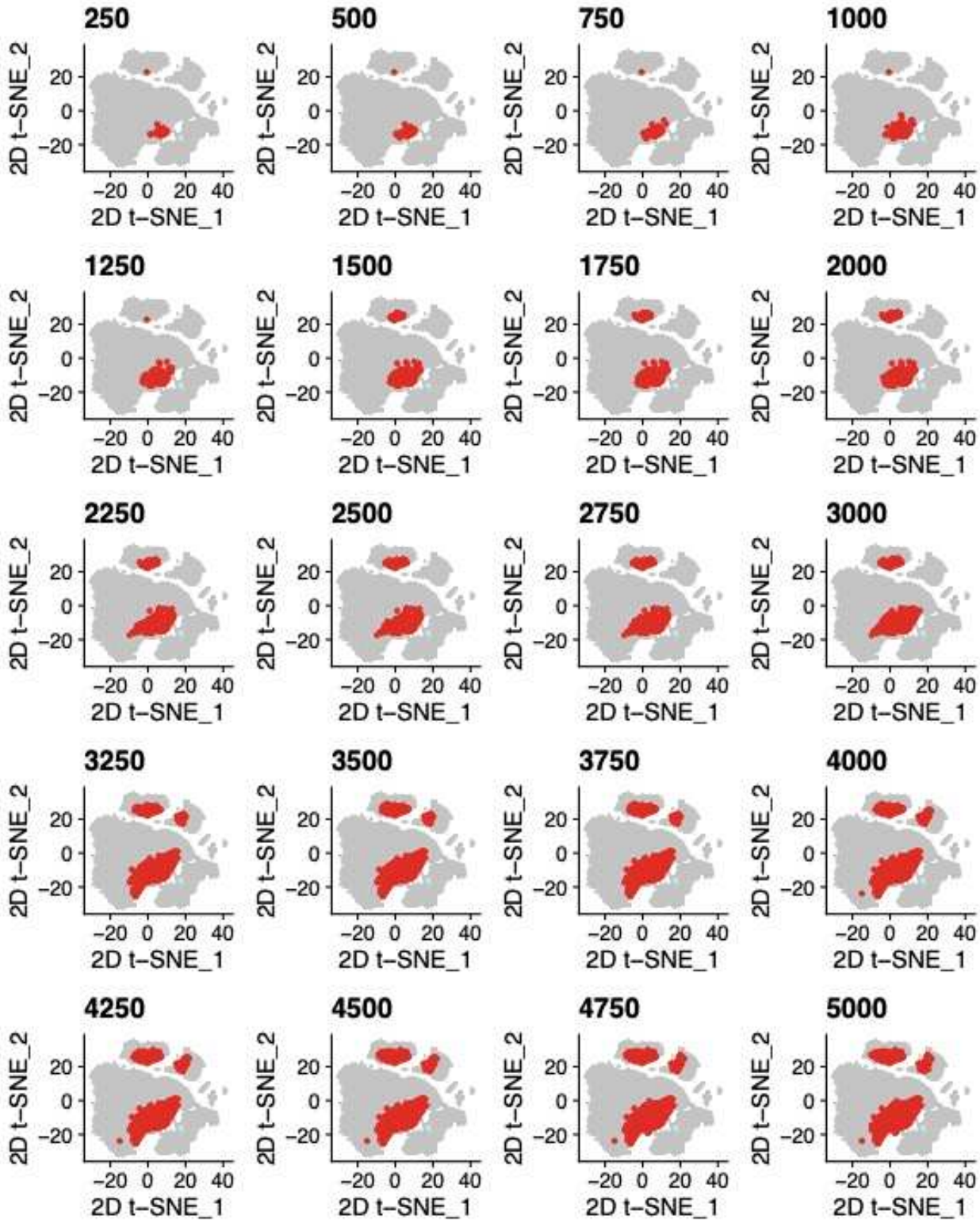
Runtime depends on the number of cells and the size of the module provided to the model. Average runtimes with and without using the quickstart option for loading calculated edge weights are displayed in **Supplementary Table 5**.



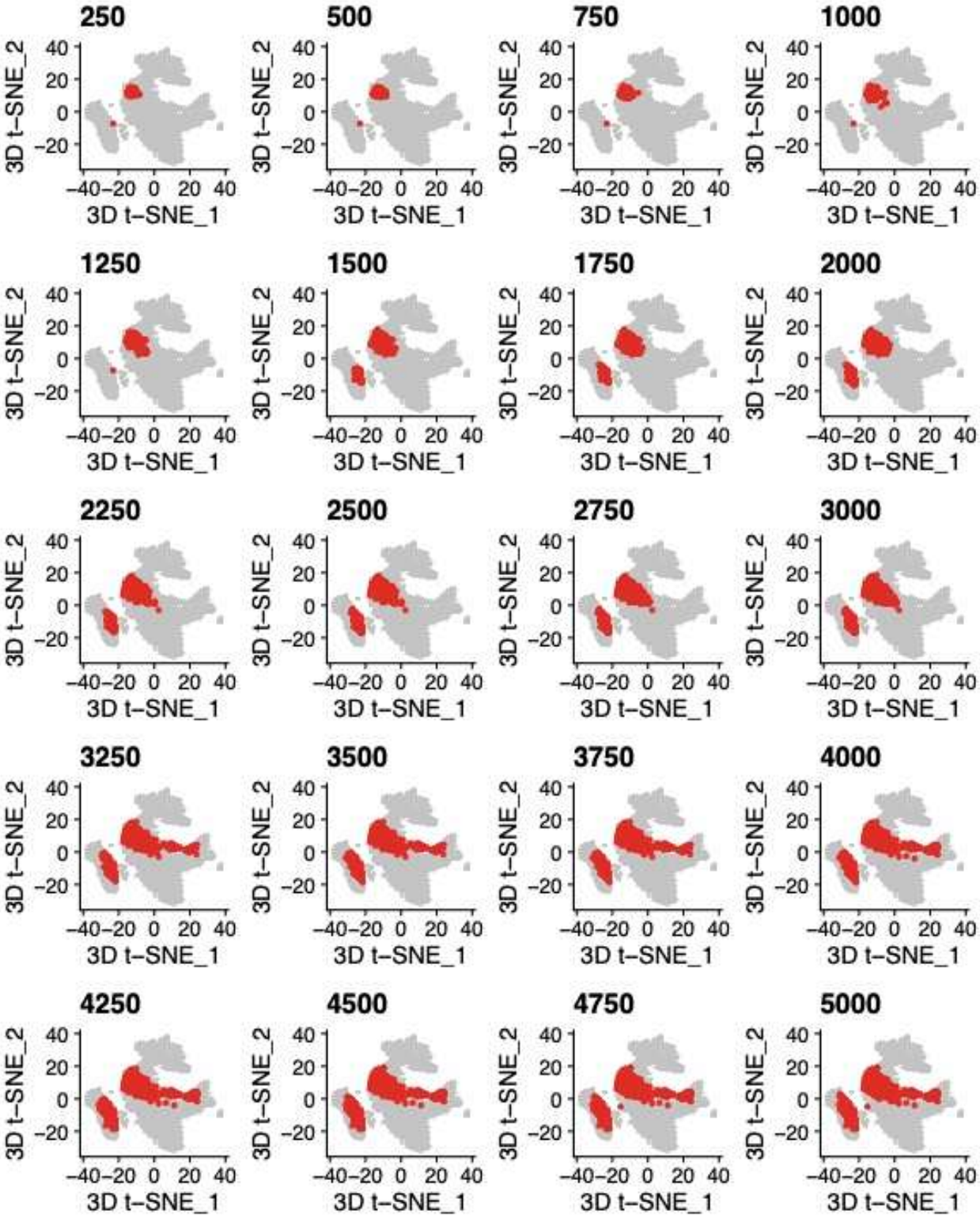
Supplementary Figure 1. Hierarchically clustered heatmaps of average single-cell expression per cell-type for modules. The dendrogram describes similarity among genes and cell-types using Euclidean distance. Darker colors indicate a greater degree of average gene expression for that cell-type.



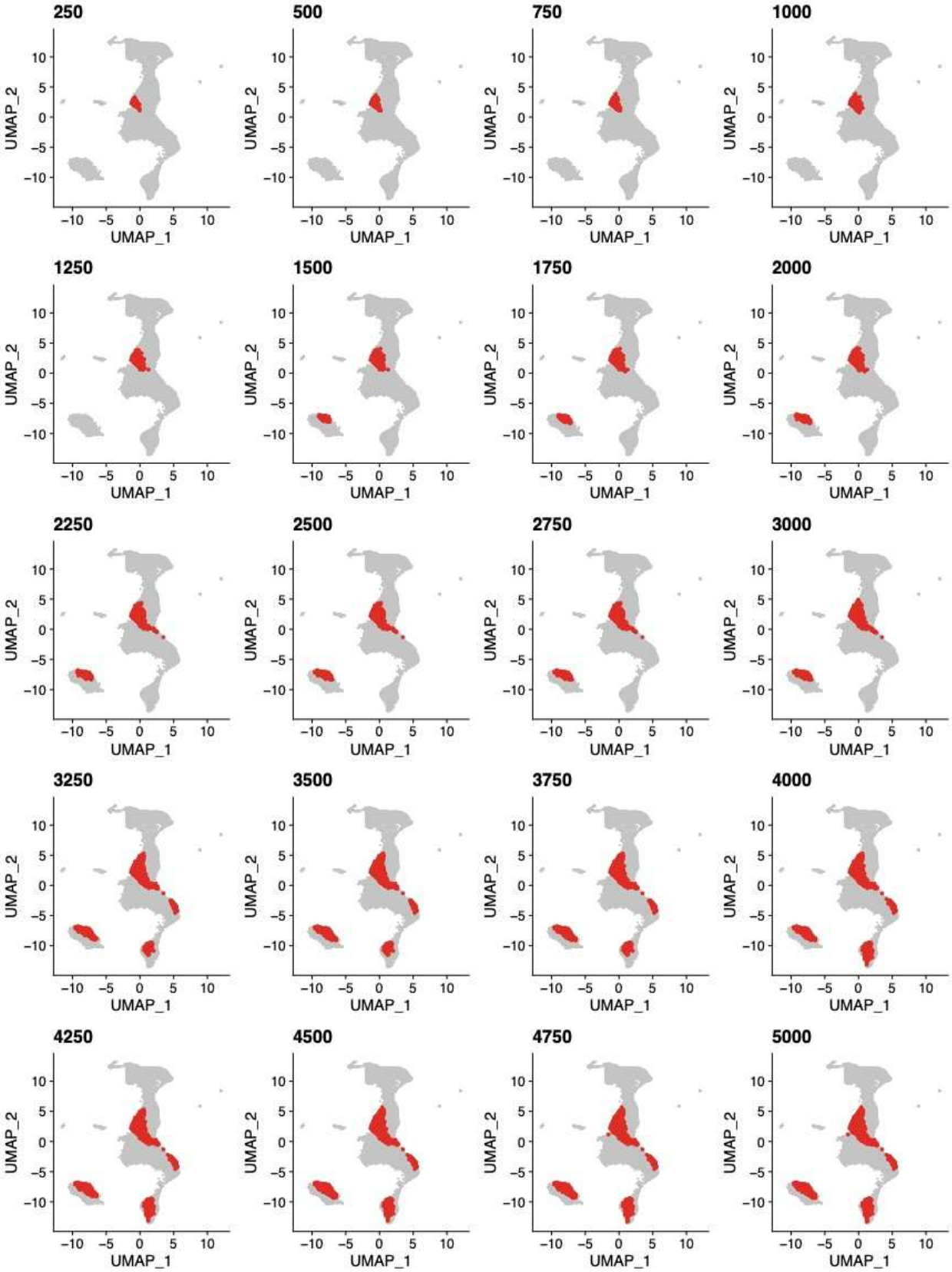
Supplementary Figure 2. Cell-type Specific Expression Analysis (CSEA) tool plots for M2 and M_*SCN1A* modules for the adult human brain. The M1 module does not show significant enrichment in any specific cell-type via the CSEA tool. Enrichment of module genes in cell-types is indicated by intensity of color, corresponding to adjusted p-values.


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
2D t−SNE_1
2D t−SNE_2
−20 0 20 40
20
0
−20

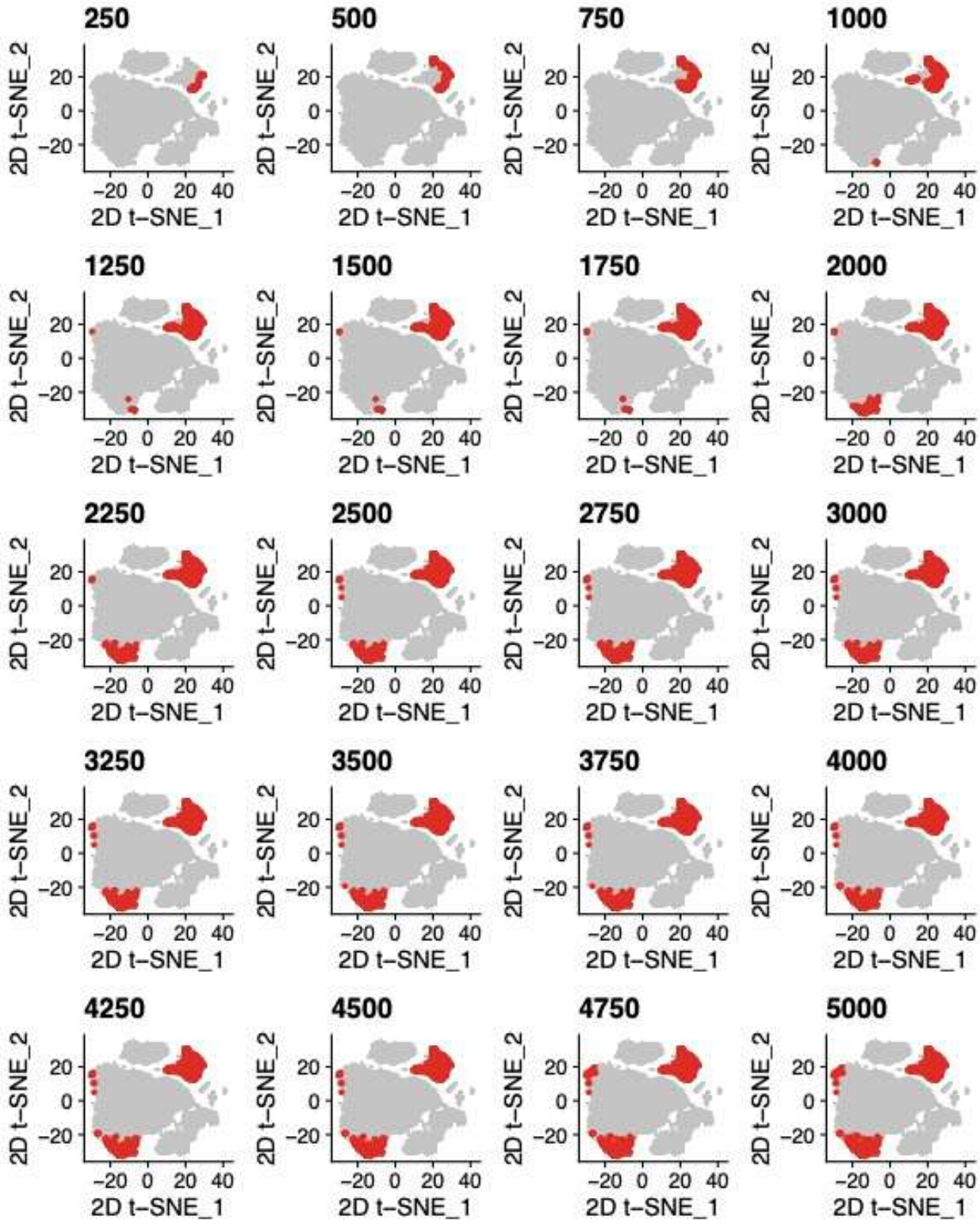
1. M1 2D t-SNE


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
3D t−SNE_2
3D t−SNE_1
40
20
0
−20
−40
−20
0
20
40

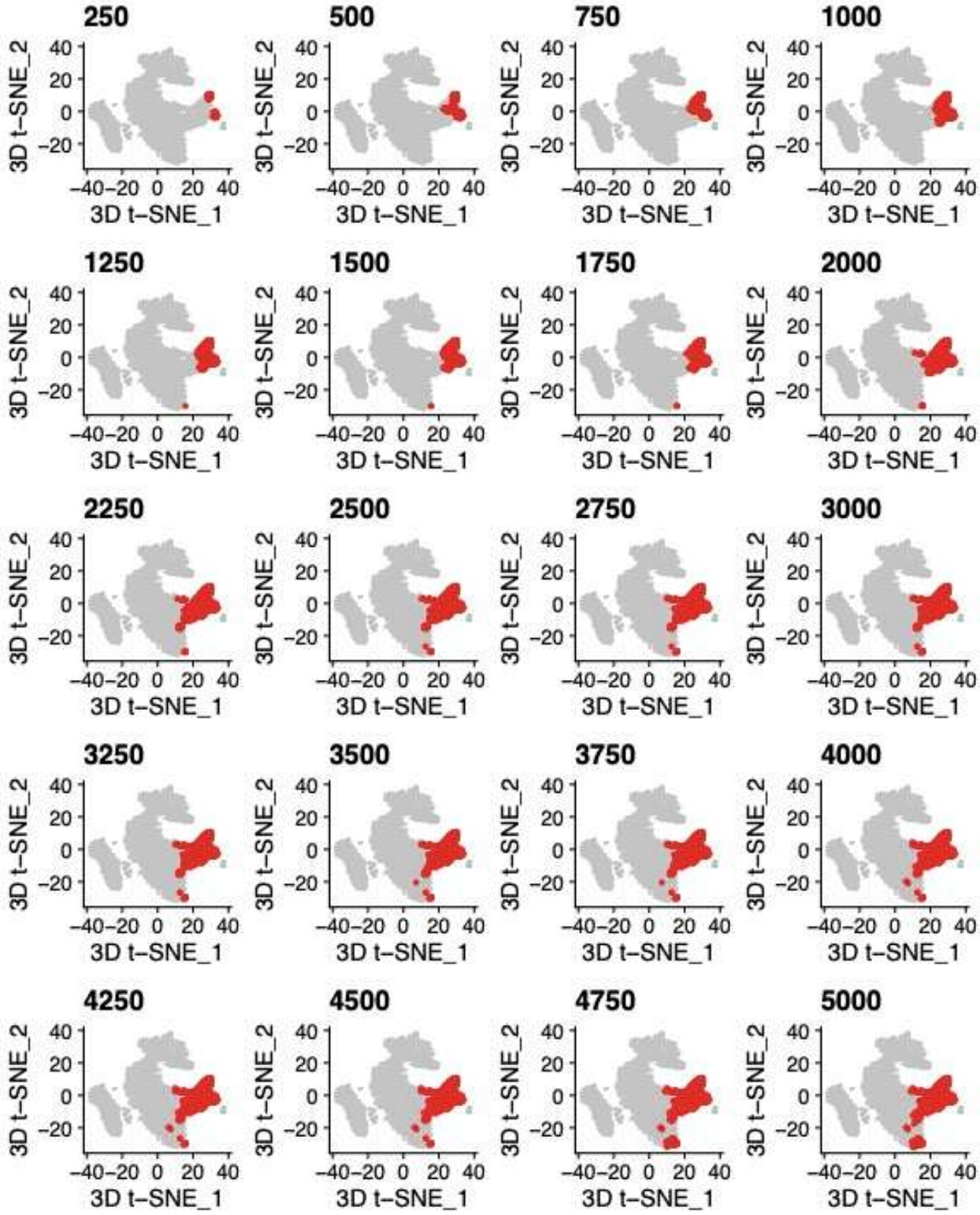
2. M1 3D t-SNE


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
UMAP_1
UMAP_2

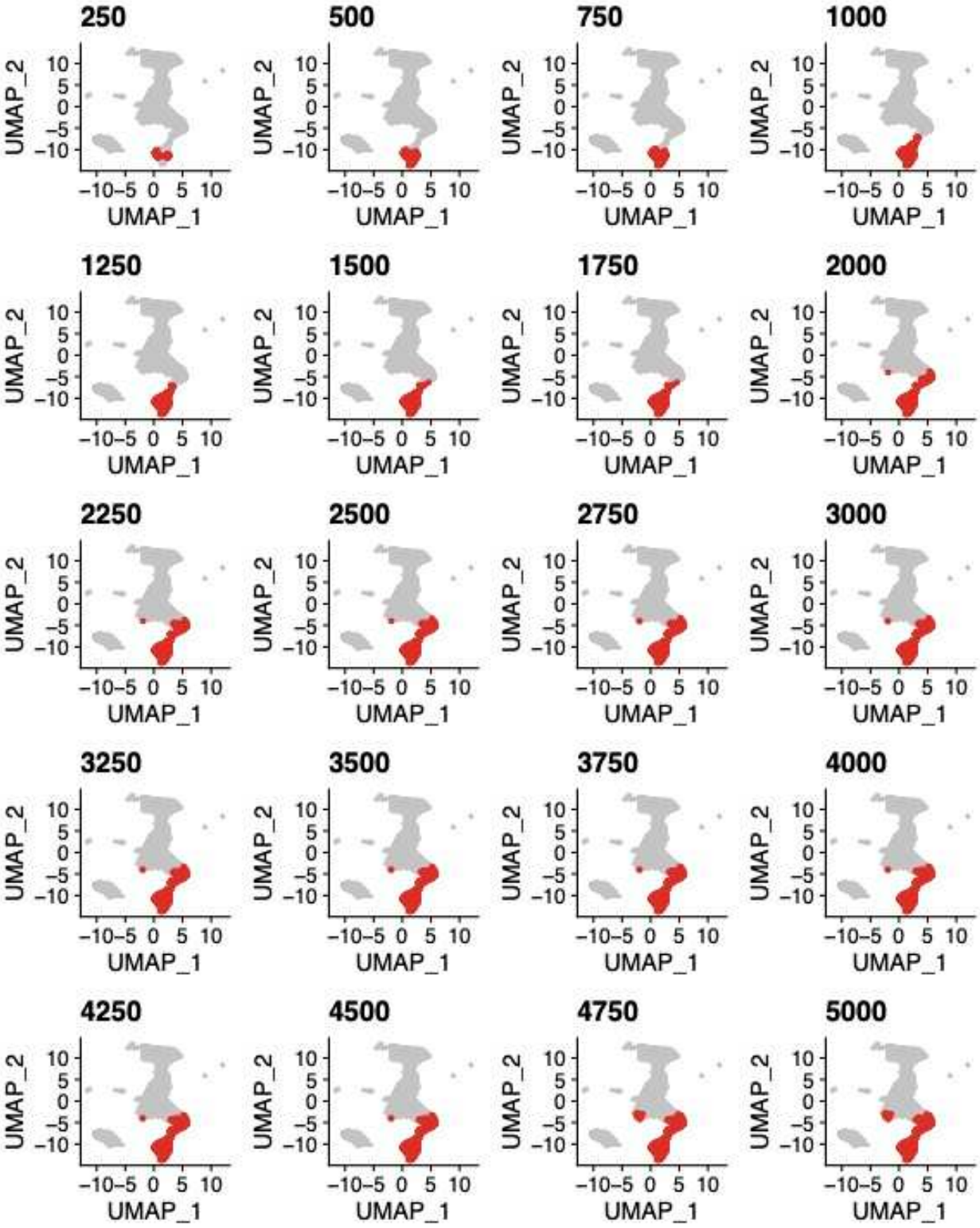
3. M1 UMAP


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
2D t−SNE_1
2D t−SNE_2
−20
0
20
40

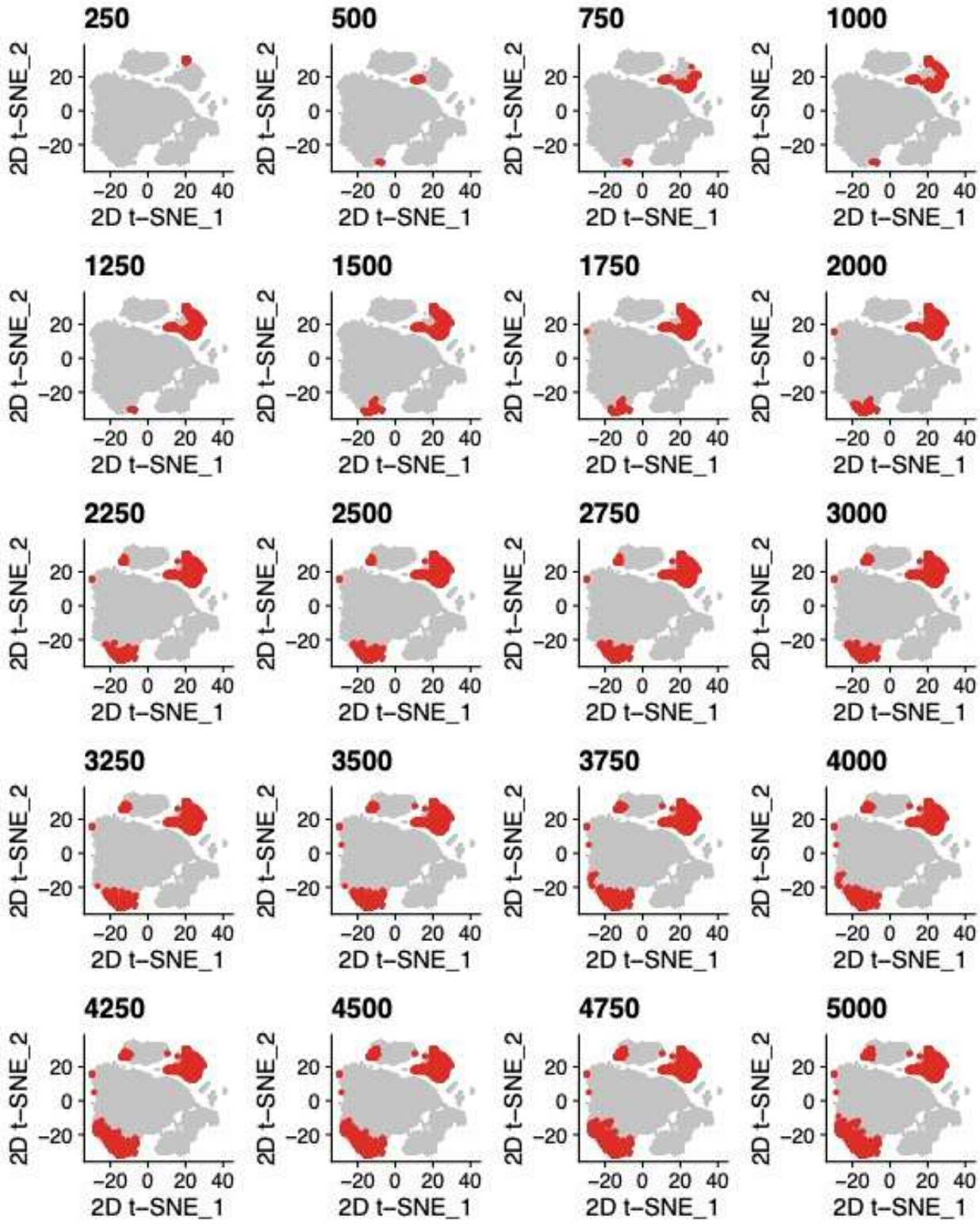
4. M2 2D t-SNE


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
3D t−SNE_1
3D t−SNE_2

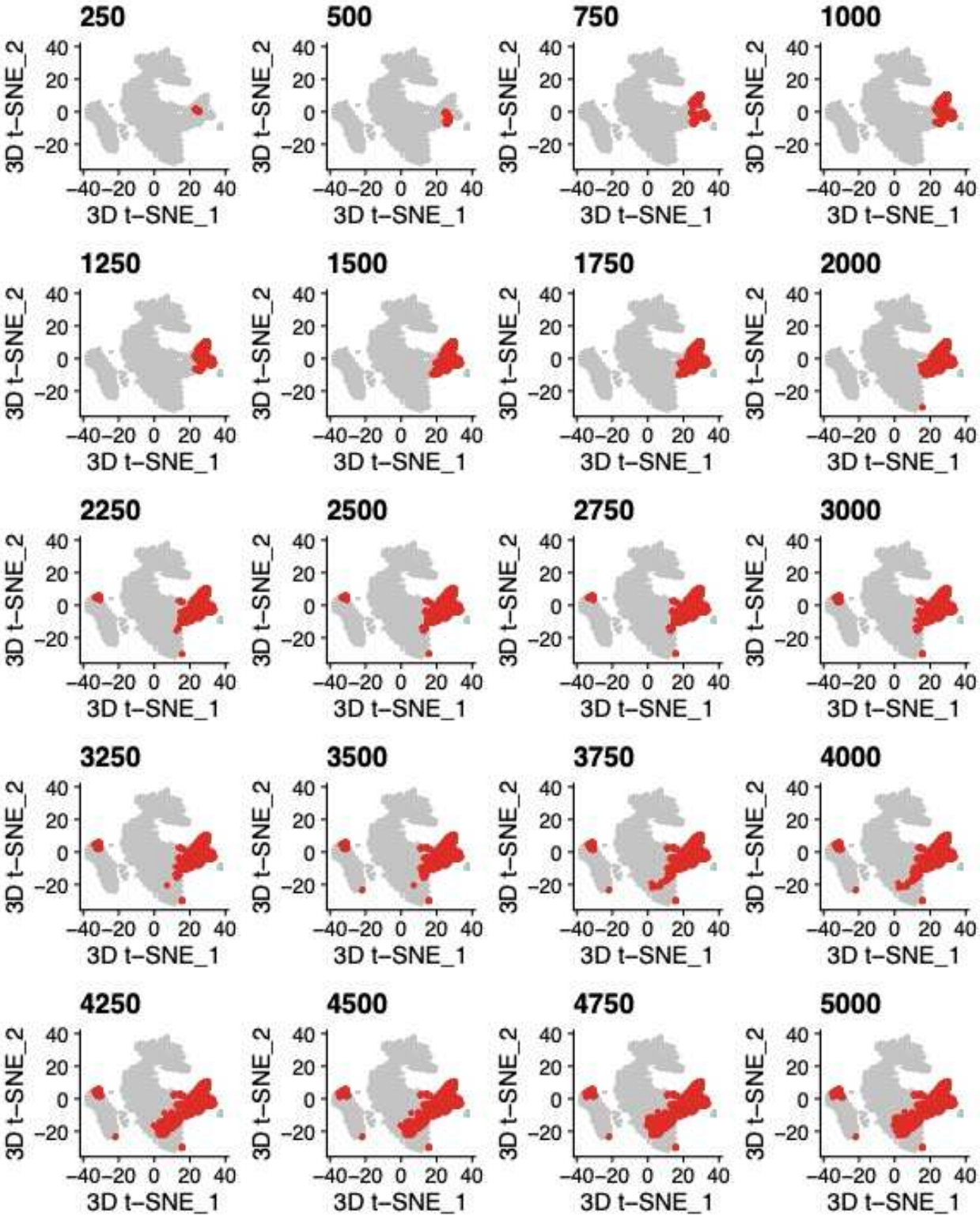
5. M2 3D t-SNE


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
UMAP_1
UMAP_2
−10 −5 0 5 10

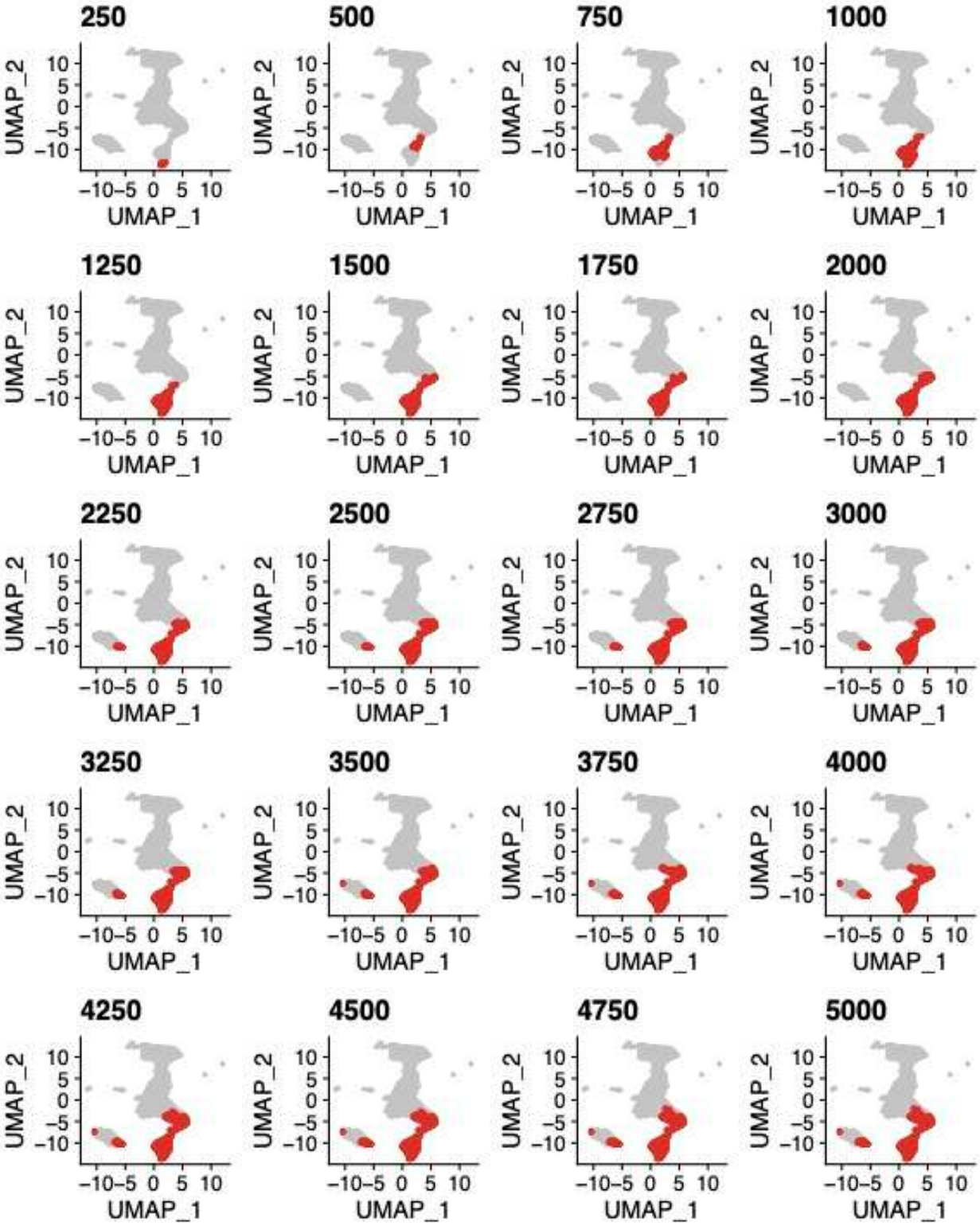
6. M2 UMAP



7. SCN1A 2D t-SNE



8. SCN1A 3D t-SNE


250
500
750
1000
1250
1500
1750
2000
2250
2500
2750
3000
3250
3500
3750
4000
4250
4500
4750
5000
UMAP_1
UMAP_2

9. SCN1A UMAP

Supplementary Figure 3. Visualization of single-cell gene expression via dimensionality reduction techniques at varied *k* for modules (M1, M2, M_*SCN1A*). Two-dimensional (2D) and three-dimensional (3D) t-SNE plots and UMAP plots are shown with cells in the largest strongly connected component (LSCC) highlighted in red. Bold lettering above each plot indicates *k* supplied (*k* = 250 to 5,000, in steps of 250 cells).

Supplementary Tables

Supplementary tables are available at https://github.com/jchow32/MoToCC.

Supplementary Table 1. Modules supplied to MoToCC and associated biology pathway enrichment. Modules (M1, M2, M_*SCN1A*) are displayed separately within each tab. Enrichment terms were retrieved from Enrichr.

Supplementary Table 2. Total objective function values and silhouette scores at varied *k* for modules. The total objective function value of the initial solution ('InitialObjF') and the refined solution ('ObjF') are shown at each *k* from *k* = 250 to 5,000. Silhouette scores calculated from two-dimensional t-SNE reduction of cell gene expression followed by K-means (K=2) are shown in the 'Silhouette' column. 'Silhouette difference' is the difference between the silhouette score at *k* and *k* - 250. 'Selected' represents the number of cells that were selected in the refined solution consisting of cells in the largest strongly connected component (LSCC). Each tab corresponds to a separate module that was supplied to MoToCC. P-values from directional one-sample t-tests and average total initial and final objective function values of twenty randomized modules are also displayed for each true module.

Supplementary Table 3. Indicator variable values at varied *k* for modules. For each value *k* varied from 250 to 5,000 in steps of 250, the initial solution of approximately *k* cells are shown in the 'Cell' column, as are the associated indicator variable values of the initial solution ('Initial value'). The 'LSCC value' column indicates whether a cell was present (1) in the largest strongly connected component (LSCC) that was returned in the refined solution. Each tab corresponds to a separate module that was supplied to MoToCC.

Supplementary Table 4. Percent composition of the largest strongly connected component (LSCC) at varied *k* for modules. The 'identity' tab displays the full cell-type name for each

abbreviated cell-type name. Percent composition is defined as the number of cells of a given cell-type divided by the total number of cells selected within the LSCC. Cell-types with percent composition greater than 0 are displayed for each k from k = 250 to 5,000 in steps of 250 for modules. 'Cell-type total' is the total number of cells of a certain cell-type that were defined in the dataset. 'Percent captured' is the proportion of cells selected in the refined solution ('num. cells selected') divided by 'cell-type total'. Each tab corresponds to a separate module that was supplied to MoToCC.

Supplementary Table 5. Runtime per module in hours. For a given module, edge weights only need to be calculated once and saved. Edge weight calculation runtimes are shown for k = 250. The 'quickstart' option may be enabled to load previously saved edge weights. Average runtime for k = 500 to 5,000 in steps of 250 cells are listed.

	M1	M2	SCN1A
Edge weight calculation (k = 250)	4.22	0.83	1.46
Total runtime (k = 250)	5.47	1.85	1.8
Quickstart runtime			
k = 500	1.32	1.48	1.35
k = 750	1.46	1.42	1.38
k = 1,000	1.35	1.46	1.49
k = 1,250	1.58	1.66	1.29
k = 1,500	1.43	1.59	1.5
k = 1,750	1.32	1.7	1.84
k = 2,000	1.25	1.56	1.82
k = 2,250	1.63	1.72	1.97
k = 2,500	1.89	2.01	1.84

k = 2,750	1.6	2.16	2.09
k = 3,000	1.44	2.0	1.78
k = 3,250	1.66	2.35	2.26
k = 3,500	1.55	2.36	2.19
k = 3,750	2.31	2.6	2.03
k = 4,000	1.69	2.38	2.63
k = 4,250	1.75	2.84	2.62
k = 4,500	1.55	3.35	2.55
k = 4,750	2.05	3.2	3.13
k = 5,000	1.63	2.87	3.77

Conclusion

Rare genetic variation in the form of *de novo* variation within coding regions of the human genome enables phenotypic prediction in neurodevelopmental disorders (NDDs), in large part due to the severe deleteriousness of likely gene-disruptive (LGD) mutation in NDD risk genes. Many NDD risk genes are 'constrained', or intolerant to protein truncating variation and possess a significantly lower frequency of LGD and missense variation than would be expected given mutation rates and genomic context. Additionally, *de novo* variation is not subject to purifying selection as inherited variants are and can potentially disrupt the function of genes necessary for typical neurodevelopment. Although *de novo* coding variation represents only a subset of genetic variation responsible for most cases of NDDs, the predictive power of *de novo* variation has permitted the identification of most known NDD risk genes, further leading to the identification of molecular mechanisms that underlie NDD phenotypes.

In Chapter 1, the module discovery methods MAGI-S and MAGI-MS were introduced. MAGI-S and MAGI-MS construct modules using gene co-expression data, protein-protein interactions, the presence of LGD variation in a control population, and user-provided seed gene(s). MAGI-S and MAGI-MS use the same general procedure to generate modules. Candidate module genes are scored according to their degree of co-expression with the seed gene(s) relative to all other genes, and seed pathways consisting of high scoring genes are formed, then clustered. High scoring modules thus contain genes that are highly co-expressed with respect to the seed gene(s) and connected via protein-protein interactions and contain genes with limited LGD variation in control populations.

MAGI-S demonstrated that it is possible to dissect the epilepsy phenotype from more general NDD phenotypes by providing single NDD seed genes with varying degrees of association

with the epilepsy phenotype. Three classes of seed gene with strong, moderate, and weak association with epilepsy were defined based on concurrent epilepsy annotations from databases and literature. The modules formed by MAGI-S, even while excluding seed genes or well-characterized NDD genes, were significantly enriched in non-synonymous *de novo* mutation in affected NDD cases compared to controls. The remaining set of genes in the human genome showed no significant difference in enrichment among NDD cases and controls, which indicates that the modules generated by MAGI-S captured a substantial portion of genes that contribute to NDD phenotypes. As evidence of the dissection of the epilepsy phenotype, modules that were seeded with seed genes strongly associated with epilepsy were significantly enriched in *de novo* non-synonymous mutation from epilepsy cohorts, contained a significantly greater proportion of genes with annotations of epilepsy association, and showed significant functional enrichment related to seizure compared to modules seeded with genes moderately or weakly associated with epilepsy.

Unlike MAGI-S, MAGI-MS permits users to select one or more seed genes, performs gene score normalization, and allows users to choose if minimum or average co-expression among seed genes is used during gene score calculations. By providing multiple seed genes to MAGI-MS, users can target module discovery towards specific biological pathways. As an example, up to 20 seed genes in the long-term potentiation pathway were provided to MAGI-MS, yielding increased enrichment in the targeted pathway as the number of seeds increased to a certain point, compared to the use of a single seed. Increased enrichment in relevant functional terms was also observed for various modules seeded with multiple seeds compared to the union of modules seeded with single seeds via MAGI-S.

By allowing users to supply gene co-expression data, protein-protein interactions, control population LGD variants, and seed gene(s) of their choice, MAGI-S and MAGI-MS provide users with the means to discover modules specific to their phenotype or biological pathway of interest. Documentation and example usage, including descriptions of all parameters, are available in MAGI-S (https://github.com/jchow32/magi-s) and MAGI-MS repositories (https://github.com/jchow32/MAGI-MS).

In Chapter 2, a shallow neural net (SNN) with a custom false positive rate (FPR) minimizing loss function used to identify a subset of NDD cases from controls was described. Using *de novo* LGD variation from NDD cases and controls and features related to genic constraint and conservation, the SNN achieved increased true positive rate (TPR) at FPR < 0.01 compared to traditional machine learning techniques including random forest, logistic regression, and support-vector machine and a randomized model. When genic constraint and conservation features were excluded from the LGD-specific model, referred to as a 'trivial' model, increased TPR at FPR < 0.01 was not observed, thus indicating the importance of genic constraint and conservation in distinguishing NDD cases from controls at low FPR. For missense-specific models, TPR at FPR < 0.01 for the SNN was not larger than corresponding values from baseline models. However, the simultaneous use of LGD and missense variation in 'combined' prediction for those individuals with both missense and LGD variation yielded greater TPR at FPR < 0.01 due to the existence of particularly deleterious missense variation.

An ensemble model consisting of the average predictions from the SNN and baseline models returned increased TPR at FPR < 0.01 compared to any individual LGD-specific or combined prediction model. Excluding SNN predictions from the ensemble model (Ensemble - SNN) decreased TPR at FPR < 0.01, even to the extent that the TPR at FPR < 0.01 for the SNN

itself was greater than that of Ensemble – SNN, suggesting that SNN predictions were a major contributor to the improved TPR at FPR < 0.01 of the ensemble model.

Although SNN predictions were strongly correlated with measures of gene constraint such as LOEUF and pLI, the SNN still achieved greater TPR at FPR < 0.01 compared to heuristics derived from gene constraint and prior classification of known NDD risk genes. Candidate NDD risk genes were prioritized using an SNN trained on LGD variation, identifying novel risk genes previously associated with related phenotypes, such as neurological disorders. The SNN and supporting documentation are freely available at https://github.com/jchow32/EarlyPredictionSNN.

In Chapter 3, MoToCC, a linear programming approach to identify critical cell-types that selectively express the genes of a module was described. MoToCC, subject to linear constraints, maximizes an objective function in the form of local correlation that is the product of cell-cell similarity and correlated gene expression of module genes among single cells and produces an initial solution of candidate cells. Among candidate cells, the associated K-nearest neighbor graph is used to identify the largest strongly connected component, which is then returned as the solution. Unlike other programs used to identify cells that selectively express certain genes, the maximum number of cells to return as a solution (k) can be varied by the user. By varying the parameter k, distinct groups of cells and cell-types relevant to module genes at varying degrees of resolution can be identified.

Three modules related to NDDs were generated by MAGI and MAGI-S and given to MoToCC with corresponding single-cell expression data from the developing human brain (M1, M2, M_*SCN1A*). Given the NDD modules M2 and M_*SCN1A*, MoToCC found enrichment for excitatory deep layer neurons, as expected due to the functional enrichment of synaptic

transmission in M2 and M_*SCN1A* modules. For the M1 module functionally enriched in chromatin modification, primarily cells of the migratory excitatory neuron cell-type were selected. For each *k* from 250 to 5,000 in steps of 250 cells, silhouette scores were calculated to describe dissimilarity of gene expression for each solution in reduced t-SNE space. Large increases in silhouette score coincided with changes in percent composition and clustering in dimensionality reduction plots among selected cells. MoToCC is available online at https://github.com/jchow32/MoToCC.

The predictive power of rare coding genetic variation in the form of *de novo* variation extends past the identification of neurodevelopmental disorder risk genes. The module discovery tools MAGI-S and MAGI-MS found genetic modules that are enriched in non-synonymous *de novo* mutation in affected cases relative to controls and underlie specific functions relevant to NDDs, as evidenced by the dissection of the epilepsy phenotype. The early prediction of NDDs greatly improves the well-being of affected patients and permits parents to make informed decisions about their reproduction. To act as a proof of principle to the importance of early genetic screening, a shallow neural network (SNN) architecture using LGD *de novo* variation in combination with genic constraint and conservation features can detect more than 30% of NDD cases at near-zero false positive rates, improving upon predictions from traditional machine learning models. As single-cell genomics grows increasingly common, MoToCC was developed as a flexible method to identify distinct groups of cells that selectively express given modules at varying degrees of resolution, leading to the further characterization of molecular mechanisms relevant to disease phenotypes. The computational methods described herein advance our understanding of the role of *de novo* variation in the genetic underpinnings of neurodevelopmental disorders.