nature portfolio

https://doi.org/10.1038/s41593-021-00887-5

Supplementary information

Coexpression network architecture reveals the brain-wide and multiregional basis of disease susceptibility

In the format provided by the authors and unedited

Co-expression network architecture reveals the brain-wide and multi-regional basis of disease susceptibility

Supplementary Note

Comparison to Latent Factor Correction

Hidden Covariates with a Prior[1] (HCP) was run within each tissue, using all measured covariates as the prior matrix variable within the tissue samples, using 10 factors, and penalization parameters $\lambda_1=1$, $\lambda_2=5$, $\lambda_3=1$. These settings were identified using a grid search within the BRNCBL tissue over the grid (K=5,10,15,20; λ=0.1,1,5,10) to maximize the AUC of GO prediction using a linear SVM (C=1), using the top 30 expression singular vectors as features.

We found that the estimated factor matrices, between tissues, differed substantially in terms of their singular values and their product, and thus were not likely to be near-rotations of one another, suggesting that the latent factors or their effects may be tissue-specific.

We sought to rank the two corrected datasets (linear-model and HCP) on the basis of biological signal-to-noise. To do this, we evaluated co-expression module size, and their preservation in microarray datasets; the extent to which module eigengenes predicted gene ontologies ("GO prediction task"), and by using a bootstrapped version of the integrated correlation coefficient.[2]

Preservation was computed in the same way for both correction methods (see "Module preservation in microarray data"). HCP correction has a profound impact on module size, tending to result in moderately more modules of significantly smaller size than linear-model based correction. As a result, we found that the preservation scores were larger for the lm-corrected rWGCNA networks – and a simple proxy for signal (# of genes in modules with Zsummary > 8) is nearly always larger for lm-corrected networks (depending on the evaluation data). GO prediction was performed using a linear SVM, using the module kME matrix as input predictors. The integrated correlation coefficient was generated using a jackknife estimate, where samples in a tissue were randomly repartitioned 1000 times into two groups, and the integrated correlation coefficient (ICC) computed for each partitioning, and the final ICC given by the mean over all partitions. The results of these analyses appear in **figure S1**.

Forward-backward covariate selection using MARS (earth)

A key step in the treatment of RNA-seq data is identifying what technical or biological covariates are strong drivers of measured expression. RNASeqQC produces a large set of alignment metrics derived from the aligned RNA-seq bams. We combined these with the splicing metrics output by STAR. Separately, each of these data were scaled and the top 5 PCs calculated to summarize the bulk of the technical covariate distribution, producing an additional 10 potential covariates. This final set of technical covariates are combined with the sample-level individual-level information provided by GTEx (ischemic time, age, biological sex, RIN, ethnicity, race).

We then used the `earth` package in R to select covariates that explained a large amount of expression variance across many genes. We set the parameters so that no non-linear splines were used, but that cross terms up to degree 3 were allowed, enabling the model to select tissue-by-covariate or covariate-by-covariate effects.

earth builds a forward model by selecting the covariate (or cross term) which most improves the total R^2 across all genes considered; and when a diminishing-returns threshold is reached (for us, an improvement of 0.01), prunes the terms using a penalized R^2 heuristic.

We ran earth 100 times on a random sample of 1,000 genes; each run producing an estimate of variance explained for all covariates (covariates *not* included in the model are assumed to explain 0% of expression variance). We summarized the impact of each covariate by taking the upper 20% of the variance explained (**figure S1a**). Any covariate whose summary estimate was >5% variance explained was included in our final model for covariate correction. For group variables (such as tissue); if any subgroup exceeded the variance explained threshold, then the entire group variable was selected.

This analysis identified the following main features: sequencing principal components seq_pc1 (13%), seq_pc2 (26.9%), and seq_pc3 (3.4%); RNA integrity RIN (7.6%), Exonic mapping fraction SMEXNCRT (27.6%), and splice alignments Number_of_splices_GT/AG (14.6%). In addition, out of an abundance of caution, we also included ischemic time TRISCHD (1.1%), and the delay between death and tissue extraction DTHCODD_CAT (3.4%). The three low %VE variables were selected by the EARTH-based procedure; and it is therefore likely that they explain large proportion of the residual variance, once the other variables are accounted for.

Module comparisons

We considered three alternatives to WGCNA for network building and module identification: ARACNe, GLASSO, and von-Mises-Fisher clustering.

ARACNe was run with default settings (10 permutations, FDR of 0.05); and genes filtered by ARACNe (for having no significant edges) were placed into a background 'grey' module. The resulting network was imported into iGraph[3] and modules identified by Louvain clustering.

As sparse inverse-covariance estimation is computationally intensive, we took an approximate approach. First, we partitioned the genes into initial groups of approximate size 1000 using k-medioids clustering. GLASSO was applied independently to each group to estimate a blockwise precision matrix. Within each block, the penalty parameter was selected using StARS[4], targeting an edge instability of between 0.05 and 0.1. Genes with no partial correlation to any others were grouped into a background 'grey' module. The remaining network was imported into iGraph and modules identified by Louvain clustering.

vMF clustering, unlike the other approaches, does not build a network, but seeks to identify gene clusters directly. Gene expression vectors were pre-processed by transforming their values into ranks (across samples) and normalizing them to unit norm. In this way, an inner product between two gene vectors is effectively their Spearman correlation. The resulting data is modeled as a collection of draws from an *n*-dimensional mixture of *k* von-Mises-Fisher distributions (where *n* is the number of samples). The model was fit using the R package movMF[5] for *k* varying from 8 to 50. The final choice of *k* came from the model that maximized likelihood – 2 * ndim * k; and module assignments were determined from the most likely mixture probability (or 'grey' if that probability was less than 0.8).

We sought to establish the non-inferiority of WGCNA to these methods, first by computing the Jaccard overlaps and their significance, and second by evaluating preservation statistics in orthogonal datasets (**table S10**). We find that most modules from each method methods are preserved; but that it is almost always possible to pick reasonable preservation thresholds (Zsummary, Zdensity, Zconnectivity, size) that favor any particular choice. Because the consensus topological overlap provides a means to hierarchically organize co-expression, the choice of WGCNA is well-motivated and evidently non-inferior to alternative methods.

Whole-brain module comparisons

Beyond comparing modules within each tissue, we sought to compare our hierarchical WGCNA modules with an orthogonal approach for building consensus modules. As consensus modules built from methods already similar to WGCNA would certainly produce similar consensus modules, we considered an alternate approach: tensor decomposition.

First, we built a fully imputed (gene x brain x region) tensor by using probabilistic PCA to impute missing samples within every (brain x region) submatrix for each gene. We then applied CANDECOMP to this tensor to produce 150 feature triplets: {(gene x 1), (brain x 1), (region x 1)}. We treated the gene-level features as a (gene x 150) feature matrix, and ran t-SNE to embed the genes in a 2-dimensional space.

While this embedding did not show distinct visual clusters, it clearly showed regions of high and low density, likely corresponding to modules. Given this intuition, we applied the DBSCAN clustering algorithm, producing a set of 30 whole-brain modules.

We found that the ribosomal, glial, and choroid-plexus modules were in one-to-one correspondence with TD-DBSCAN modules (figure **S1**), and that the neuronal WGCNA modules correspond to multiple TD-DBSCAN modules, with statistically significant overlaps. Visually, the WGCNA modules are localized in the embedded tensor-decomposed space, strongly suggesting that the modules are not driven by the specifics of WGCNA, nor are they induced by the structure of hierarchical merging; but rather that these genes are grouped together by disparate approaches because of an underlying biological signal.

Module preservation in microarray data

Figures 1c and 1d utilize preservation statistics from multiple microarray-based studies of human brain tissue. For computing module preservation, we used several microarray datasets for validation, matched to the regions under study. These datasets were: Kang 2011 (GSE25219; AMY, BA9, BA24, PFC, CDT, CBH, CBL, HIP, PUT), UKBEC (GSE46706; BA9, BA24, PFC, CBH, CBL, NAcc, SNA), NABEC (GSE15745; BA9, BA24, PFC, CBH, CBL, HIP), Oldham 2008 (GSE1572, GSE3790, GSE5392, GSE7540, GSE12649, GSE12654; BA9, BA24, PFC), and the GTEx pilot microarray dataset (AMY, BA9, BA24, PFC, CDT, CBH, CBL, HIP, HYP, NAcc, PUT, SNA). Because the dissections for these various studies could be broad; we matched areas broadly for preservation; for instance cortical areas of GTEx (BA9, BA24, and PFC) are examined for preservation (separately) in FC, vFC, oFC, dFC, mFC, and m1C (Kang); and the average preservation statistics were retained as a summary, while sub-striatal areas (CDT, PUT, NAcc) could only be compared to bulk striatum (Kang). These datasets were pre-processed to remove by linear regression the effects of: RIN (Kang, UKBEC, GTEx), age (all), sex (all), PMI (Kank, UKBEC, NABEC).

For our WGCNA modules, we find that brain-wide modules show evidence of preservation across all brain regions, with the notable exception of the neuronal module, BW-M4, which is present in all brain areas except the cerebellum and striatum. In contrast, region-specific modules show evidence of preservation only within that region (32/63) or adjacent regions (50/63).

We also obtained RNA-sequencing data from GTEx-v8, and computed module preservation within samples only present in GTEx-v8 (and not used to construct the original modules). Only 4 modules failed to strongly replicate (Zpreservation > 8) in GTEx-v8 samples (BRNACC-M1, BRNCDT-M4, BRNHYP-M2, BRNPUT-M6), and none failed to show moderate evidence of preservation (Zpreservation < 5).

Single-cell data

Quantified single-cell data was downloaded from http://mousebrain.org[6] (mouse) and subset to only cells from the CNS (without spinal chord); and GEO GSE97942[7] was downloaded for human. These data were log-transformed log(1 + x) for counts and log(0.005 + x) for TPM; and the cell type labels from the respective publications were used for all subtype analyses ("In" versus "Ex" for interneuron, "Purkinje" versus all other neuronal cells, and "MSN1" and "MSN2" versus all other cells for medium spiny neurons). Absolute expression values were taken as the mean expression of a cluster; and relative expression was obtained via

Relative = absolute – background

Where the background expression is the average expression of a gene over all cells. To incorporate gene variance information into relative expression, the *relative expression rank* is defined as the lower end of a small confidence-interval for the difference in means:

$$\text{rank} = (\mu_a - \mu_b) - \ 0.5 * \sqrt{\frac{v_a}{n_a} + \frac{v_b}{n_b}}$$

kME enrichments are based on the correlation between module kME and the relative expression rank within a given cell type; and enrichment trends (as fitted by a generalized additive model in the R package mgcv) are plotted in figure 1(c).

Cell-type enrichment and single-cell data

For kME-based enrichments (such as those in figure 2), the shaded region of the figure represents the standard error around the estimated functional relationship between kME and relative expression rank. In all cases it is visually apparent that these lines deviate from 0 by a factor far exceeding 2.5 times their standard error ($p \sim 0.006$).

For gene-set based enrichments such those presented in the text, and those in figure 3, cell type markers were obtained from several sources[8,9,10,11,12,13,14,15] representing various studies performed both in mouse and in human. We also obtained gene lists corresponding to neural progenitor development[16], neuronal migration[17], and neuronal differentiation[18]. The statistical test is a logistic regression using the model:

is.cell.marker ~ 1 + is.in.module + gene.length + gene.gc

adjusting for gene length and GC. We test that the coefficient for module presence is significantly different and greater than zero, implying an enrichment (as opposed to depletion) of cell-type related genes.

This test is performed independently on cell type markers from the various studies, and FDR adjusted across all tests.

For Figure 1(c), the genes taken as “Interneuron” markers are those differentially expressed between “Ex” and “In” classes from Lake et al. (2016).

Defining mouse orthologs to human genes

The ensembl API was used, through biomaRt, to query human genes with associated mouse orthologs and the type of orthology; and visa versa. These queries enabled defining genes as one-to-one orthologs, one-to-many orthologs, many-to-many orthologs, or non-orthologous. The ensembl API was also used to obtain human-mouse dN and dS values; and the ratio dN/dS calculated, with 0/0 treated as 0.

GO enrichment

Gene ontology enrichment is performed competitively, with covariate correction, using logistic regression. Briefly, each GO or KEGG[19] category is treated as a binary variable (1 for genes in the category, 0 for genes not in the category – only genes ascertained in our gene expression matrix are part for the regression). Modules are also treated as binary. We include as covariates the average gene expression across all tissues in the brain, the gene GC content, the log gene length, and the gene expression reproducibility (see below). The GO enrichment model is then

GO ~ module.1 + … + module.k + mean.expr + GC + log.gene.length

And is fit using logistic regression. If we detect that convergence fails, an L2-regularized logistic regression is instead applied (using `brglm`). The enrichment p-values are taken to be the statistics that reject ($\beta_i \leq 0$) for all β_i corresponding to a module indicator.

The enrichment p-values are adjusted for all ontologies.

In one instance (TGF-beta signaling in BW-M1), the FDR reported comes from STRING[20] and is annotated as such in the main text.

Meta-GSEA

To aggregate enrichment results (such as GO) from the module level to the module set level, the GO p-values are treated as independent p-values, and Fisher's method is applied: For a given ontology category, a χ^2 value is calculated as -2 * log(p_1*p_2*…*p_k), where the product is taken across modules in the set. In the case of independence, this statistic has 2**k* degrees of freedom; allowing a p-value to be calculated. Because the modules in a set overlap by construction, the resulting statistics are not calibrated probabilities, and are referred to as "scores" or "rankings," and should not be interpreted as reflecting significance. In nearly all cases, the highly-ranked consensus ontology had been significant in one or more of the modules within the set.

The meta-GSEA applied to generate supplemental figures 5b,c was to identify the genes within the regional BW-M4 modules (e.g. PFC-BW-M4) with MAGMA Z-scores > 3.0 (SCZ) or 2.5 (ASD). This generated an indicator variable which was then used to perform gene ontology, using the BW-M4 genes as a background; generating p-values for each ontology. Meta-GSEA was applied to these p-values, generating a score for each ontology, plotted in figure S5.

pLI enrichment

Gene pLI scores were downloaded from the ExAC consortium release[21], and a gene was considered likely to be LoF-intolerant if its pLI score was 0.9 or higher. Enrichment for "hard" module membership (i.e. comparing two gene lists) is performed via Fisher's exact test on the contingency table between module membership and LoF-tolerance/intolerance. "Soft" module enrichment (i.e. based on kME) is computed via a Brownian Bridge statistic.

The genes are ranked by their module membership (kME); and the proportion of all genes which are likely LoF-intolerant (the pLI rate, $r=P/M$) is computed. At a given quantile q of genes, we tabulate how many of the first $q * M$ genes are LoF-intolerant; and denote this cumulative sum by $Cs(q)$. The expected number of LoF-intolerant genes is $\mathrm{Ne}(q) = q * P = q * r * M$. For large M, this cumulative sum converges to a scaled Brownian motion with drift r; and has variance $\mathrm{V}(q) = q * (1 - q) * M * r * (1 - r)$. Z-scores for this cumulative sum at each q are given by $\mathrm{Z}(q) = (\mathrm{Cs}(q) - \mathrm{Ne}(q))/\sqrt{\mathrm{V}(q)}$. An excess of LoF-intolerant genes occurs when $min\_q\ \Phi(Z(q)) < 0.05$. For clearer visualization, we plot $(\mathrm{Cs}(q) - \mathrm{Ne}(q))$ and $2.17 * \sqrt{\mathrm{V}(q)}$ as functions of q.

We also used a generalized additive models (“GAM”) and a generalized linear models (“GLM”) to verify findings of constraint. In these cases we applied the (logistic) model:

is.constrained ~ rank(kME)+ gene.length + gene.GC

and found that, for the whole-brain modules, these enrichments were so strong that the three methods were in 100% concordance. The results of the linear models did not change substantively when using competitive as opposed to marginal enrichments.

For supplemental figure 4 (enrichment in pLI and o/e bins), the odds ratio and p-values were computed using a Fisher Exact Test between module membership, and bin membership.

PPI enrichment

We use the InWeb PPI database[22] (brain tissue) for a source of defined protein-protein interactions, with a confidence threshold of 0.2 used as a cutoff for a particular interaction. PPI prediction is treated as edge-related data, where the response variable is binary (presence/absence of PPI), and the predictors the following collection of data relevant to that edge: the (PPI) connectivity of its first vertex, the (PPI) connectivity of its second vertex, the product of kMEs of its vertices (for each module), the product of the GCs of its vertices, and the product of the reproducibilities of its vertices. Or:

$$E_{ij} \sim C_i + C_j + kME\_M1_i * kME\_M2_j + \ldots + kME\_Mk_i * kME\_Mk_j + GC_i*GC_j$$

This equation encodes the model that gene pairs which are mutually close to a given module are more likely to physically interact. The logistic model is fit using `statsmodels` in python, and the hypotheses $\beta i \leq 0$ is assessed for each βi corresponding to a module.

Module Imputation

For our lncRNA analysis, we imputed whole-brain modules into an independent RNA-seq dataset[23] by i) splitting the data into BA9 and BA41-42-22 regions, ii) Calculating module kMEs within each region, and iii) Averaging across the two regions. This generates a set of 11 features (average within-region kME to each module) for each gene. The overlapping genes between the GTEx modules and control brain expression were used as labels to fit a gradient boosted trees classifier[24] (using the R package xgboost with 2000 trees and a learning rate of 0.025). Non-overlapping genes (which contain most lncRNA and a set of held-out, length and GC matched protein-coding genes) are assigned to modules via the prediction of the fitted classifier. Using cross-validation on the matched protein-coding genes, we estimate that the sensitivity and specificity of this approach are 0.63 and 0.53 for BW-M6, with sensitivity ranging from 0.25-0.7 and specificity from 0.2-0.8 across other modules. The most common misclassification (>60%) results from assigning a ‘grey’ gene as in the module, or a BW-M6 gene as ‘grey’. We examined the predicted cell-type lncRNA in published single-cell data[25], and found that the lncRNA predicted to be in cell-type modules are up-regulated in those corresponding cells.

Isoform specificity from sorted cell data

RNA-sequencing data was obtained from GSE73721 (SRA project SRP064454) and quantified at the isoform level with Kallisto (mouse gencode release M16). These data included sorted populations of astrocytes, oligodendrocytes, endothelial cells, a single neuronal population, and a whole-tissue background. Relative isoform expression were obtained as described in “Single-cell data,” with the background set to be the average expression across the whole-tissue background samples.

Isoform switching and validation

Isoform-level TPM values (produced by RSEM) were corrected using a linear model with the same covariates used for correcting gene expression TPMs. Subsequently, each isoform expression (within tissue) was correlated to brain-wide module eigengenes computed within the tissue, and the mean correlation across tissues taken as an estimate of module membership for the isoform.

To determine an appropriate kME threshold, we evaluated the impact of thresholding on cell type enrichments. Each threshold produces a set of isoforms within a module; and each isoform can be annotated with the cell type marker status of its parent gene. Fisher's Exact Test produces an odds ratio and p-value for cell-type enrichment at each threshold. We found that a threshold of 0.45 produced a 15-fold enrichment for both astrocyte and oligodendrocyte markers when looking at kME to their respective modules (M6 and M7); but that when increasing this threshold the odds ratio for oligodendrocytes did not substantially change, while the astrocyte odds ratio increased **(figure S7)**. Based on this we defined the threshold for isoform module membership at 0.45 kME. In the case where an isoform has >0.45 kME to multiple modules, module with highest kME is selected.

To validate these findings, we used RNA-seq data from sorted cells,[26] quantified at the isoform level. Correlation between isoform kME (to a cell-type module) and the rank of the gene expression within the corresponding cell type was moderate (rho=0.286 oligo, 0.258 astrocyte), but significant ($p<10^{-15}$ for both) using Spearman's rho.

An “isoform switch” is defined as two sister isoforms having membership to different modules. Genes linked to Autism (either via known mutations or other genetic evidence) were obtained from AutDB,[27] and the likelihood of observing the four ASD genes among all isoform switching genes was obtained using Fisher's exact test.

Overlap with disease-implicated co-expression networks

Data sets were obtained for normal brain[28,29] , autism[30] , schizophrenia[31,32], cross-psychiatric[33], Alzheimer's disease[34], epilepsy[35], and developing brain[36,37,38] from the main or supplementary tables of the corresponding publications. We computed the Jaccard overlap and its significance (Fisher exact test) between our whole-brain and regional modules, and the disease-relevant modules.

Differential preservation analysis

Modules defined in the GTEx tissue samples were assessed for their preservation in ASD case samples and (separately) in normal samples. These produced a pair of preservation Z-scores per module. We defined differential preservation to be cases where the control Z-score is preserved (>3) while the ASD Z-score is not preserved (<3).

Hub gene co-expression

Gene expression data was obtained for adult human brains from the Allen Brain Atlas. Gene expression values were averaged across individuals, and Z-scored within region. Figure 6g shows that the CTX-M3 hub genes have nearly identical patterns of expression across all cortical regions, while the patterns are more variable in non-cortical regions.

Activity-dependent gene enrichment

Lists of activity-dependent genes were obtained from Schanzenbacher et al.,[39] and gene set enrichment was performed identically to other ontologies (see above).

Network construction and computation of d(G)

Transcription Factor Binding Networks

Bipartite transcription factor binding graphs were obtained from regulatorycircuits.org, and converted to a similarity network as in Marbach2016. Briefly, the probability weights are taken as edge weights, and the random-walk kernel K=(I+W)^4 with W the symmetrically-normalized Laplacian $D^{-1/2}AD^{-1/2}$ of the adjacency matrix; and converted to a dissimilarity via $D_K = 1 - \frac{(K - \min(K))}{(\max(K) - \min(K))}$. A natural set of "core" genes on this network are the most highly-connected genes of K; of which the top 25 are taken. Distances are either the mean or minimum path distance under D_K.

Protein-protein interaction networks

InWeb[40] was used for the protein-protein interaction network. The refined brain-PPI network was obtained from the resource, and a confidence of 0.05 required for an edge to be defined; and the interactions were converted into a binary matrix. Distances were defined as either the minimum or mean path distance in this network.

As with TFBNs, the natural set of hub genes are the most connected genes, of which the top 25 are taken.

Partial Correlation Networks

To compute (approximate) partial correlation networks, covariate-corrected expression data for whole-blood, prefrontal cortex, and prenatal cortex were processed as follows:

1. Genes are partitioned into blocks, of target size 3000, using a number of centers equal to ⌊n.genes/target.block.size⌋*3, using projective k-means
2. GLASSO is run on each block, using the glasso function with `approx=T`. The penalization parameter rho (=1/Lambda) is selected by taking Lambda to be initially high (5000), and repeatedly shrinking by 20% until the total proportion of edges are between 3% and 8% of all possible edges.
3. The final network is unsigned, using the absolute value of estimated partial correlations as the edge weights

As with PPI networks, these edges are used to define the minimum and mean path distances in the network. As the partial correlation is estimated blockwise, it generates a graph of several disconnected components; and rather than using infinity as the maximum distance; the distance between two non-connected genes is set to 1 + the maximum distance observed between connected genes.

Significance calculation for Φ

Because Φ reflects a partitioning of a subset of genes, a significance value can be calculated by Fisher's Exact Test. As a specific example: the overlap of genes between two studies is 15902 genes. After computing network distances, the top decile contains

1590 genes. Imagine that the core gene set (after excluding non-coding, non-regulatory genes) contains 32 genes, and 12 of these overlap the set of 1590. The contingency table $\begin{pmatrix} 14312 & 1590 \\ 20 & 12 \end{pmatrix}$ reflects this observation, and has a p-value of 0.00003.

[1] Mostafavi, S.; Battle, A.; Zhu, X.; Urban, A. E.; Levinson, D.; Montgomery, S. B.; & Koller, D. Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge. PLoS ONE, 8(7), e68141. **2013**

[2] Cope, L., Naiman, D. Q., & Parmigiani, G.. Integrative correlation: Properties and relation to canonical correlations. Journal of Multivariate Analysis, 123, 270–280. **2014**

[3]Csardi, G. & Nepusz, T. The igraph software package for complex network research *InterJournal,* **2006***, Complex Systems*, 1695

[4] Liu, H.; Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2, Curran Associates Inc.,* **2010**, 1432-1440

[5]Hornik, K. & Grün, B. movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions *Journal of Statistical Software,* **2014**

[6] Zeisel, A.; Hochgerner, H.; Lönnerberg, P.; Johnsson, A.; Memic, F.; van der Zwan, J.; Häring, M.; Braun, E.; Borm, L. E.; Manno, G. L.; Codeluppi, S.; Furlan, A.; Lee, K.; Skene, N.; Harris, K. D.; Hjerling-Leffler, J.; Arenas, E.; Ernfors, P.; Marklund, U. & Linnarsson, S. Molecular Architecture of the Mouse Nervous System *Cell, Elsevier Inc.,* **2018**

[7] Lake, B. B.; Chen, S.; Sos, B. C.; Fan, J.; Kaeser, G. E.; Yung, Y. C.; Duong, T. E.; Gao, D.; Chun, J.; Kharchenko, P. V. & Zhang, K. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain *Nature Biotechnology,* **2018***, 36*

[8] Zhang, Y.; Chen, K.; Sloan, S. A.; Bennett, M. L.; Scholze, A. R.; O'Keeffe, S.; Phatnani, H. P.; Guarnieri, P.; Caneda, C.; Ruderisch, N.; Deng, S.; Liddelow, S. A.; Zhang, C.; Daneman, R.; Maniatis, T.; Barres, B. A. & Wu, J. Q.
An RNA-Sequencing Transcriptome and Splicing Database of Glia, Neurons, and Vascular Cells of the Cerebral Cortex *Journal of Neuroscience, Society for Neuroscience,* **2014***, 34*, 11929-11947

[9] Zhang, Y.; Sloan, S.; Clarke, L.; Caneda, C.; Plaza, C.; Blumenthal, P.; Vogel, H.; Steinberg, G.; Edwards, M.; Li, G.; John A. Duncan, III; Cheshier, S.; Shuer, L.; Chang, E.; Grant, G.; Gephart, M. & Barres, B. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse *Neuron, Elsevier Inc.,* **2016**

[10] Jeremy A. Miller Steve Horvath & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways *PNAS,* **2010**

[11] Mancarci, B. O.; Toker, L.; Tripathy, S. J.; Li, B.; Rocco, B.; Sibille, E. & Pavlidis, P.
Cross-Laboratory Analysis of Brain Cell Type Transcriptomes with Applications to Interpretation of Bulk Tissue Data *eneuro, Society for Neuroscience,* **2017***, 4*, ENEURO.0212-17.2017

[12] Romanov, R. A.; Zeisel, A.; Bakker, J.; Girach, F.; Hellysaz, A.; Tomer, R.; Alpár, A.; Mulder, J.; Clotman, F.; Keimpema, E.; Hsueh, B.; Crow, A. K.; Martens, H.; Schwindling, C.; Calvigioni, D.; Bains, J. S.; Máté, Z.; Szabó, G.; Yanagawa, Y.; Zhang, M.-D.; Rendeiro, A.; Farlik, M.; Uhlén, M.; Wulff, P.; Bock, C.; Broberger, C.; Deisseroth, K.; Hökfelt, T.; Linnarsson, S.; Horvath, T. L. & Harkany, T. Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes *Nature Neuroscience, Springer Nature,* **2016***, 20*, 176-188

[13]Tasic B, Menon V, Nguyen TN, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci*. 2016;19(2):335–346. doi:10.1038/nn.4216

[14] Heintz, N. Gene Expression Nervous System Atlas (GENSAT) *Nature Neuroscience,* **2004**

[15] Kelley, KW; Inoue, H; Molofsky, AV & Oldham, MC. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nature Neuroscience,* **2018**

[16] Oldham, M. C. et al. Functional organization of the transcriptome in human brain. Nature Neuroscience **11**, 1271–1282 (2008).

[17] Kang, H. J.; Kawasawa, Y. I.; Cheng, F.; Zhu, Y.; Xu, X.; Li, M.; Sousa, A. M. M.; Pletikos, M.; Meyer, K. A.; Sedmak, G.; Guennel, T.; Shin, Y.; Johnson, M. B.; Krsnik, Z.; Mayer, S.; Fertuzinhos, S.; Umlauf, S.; Lisgo, S. N.; Vortmeyer, A.; Weinberger, D. R.; Mane, S.; Hyde, T. M.; Huttner, A.; Reimers, M.; Kleinman, J. E. & Sˇestan, N. Spatio-temporal transcriptome of the human brain. *Nature,* 2011

[18] Habib, N; Li, Y; Heidenreich, M; Sweich, L; Avraham-Davidi, I; Trombetta, JJ; Hession, C; Zhang, F & Regev, A. Div-Seq: Single nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science,* 2016

[19] Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28, 27–30 (2000). Move this to methods

[20] Szklarczyk, D. et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Research 45, D362–D368 (2016).

[21] 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., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., … Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 2016

[22] Li, T; Wernersson, R; Hansen, RB; Horn, H; Mercer, J; Slodkowicz, G; Workman, CT; Rigina, O; Rapacki, K; Staerfeldt, HH; Brunak, S; Jenson, TS & Lage, K
A scored human protein-protein interaction network to catalyze genomic interpretation
Nature Methods, **2017**

[23] Parikshak, N. N.; Swarup, V.; Belgard, T. G.; Irimia, M.; Ramaswami, G.; Gandal, M. J.; Hartl, C.; Leppa, V.; de la Torre Ubieta, L.; Huang, J.; Lowe, J. K.; Blencowe, B. J.; Horvath, S. & Geschwind, D. H. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature.* 2016

[24] Chen, T., & Guestrin, C. (2016). XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. ACM Press. https://doi.org/10.1145/2939672.2939785

[25] Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Chadhoury, S.R., Aguet, F., Gelfand, E., Ardlie, K., Weitz, D., Rozenblatt-Rosen, O., Zhang F., Regev, A. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature methods* 2017

[26] Zhang, Y.; Sloan, S.; Clarke, L.; Caneda, C.; Plaza, C.; Blumenthal, P.; Vogel, H.; Steinberg, G.; Edwards, M.; Li, G.; John A. Duncan, III; Cheshier, S.; Shuer, L.; Chang, E.; Grant, G.; Gephart, M. & Barres, B. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron,* 2016

[27] Basu, S. N., Kollu, R. & Banerjee-Basu, S. AutDB: a gene reference resource for autism research. Nucleic Acids Research **37**, D832–D836 (2008).

[28] Hawrylycz, M.; Miller, J. A.; Menon, V.; Feng, D.; Dolbeare, T.; Guillozet-Bongaarts, A. L.; Jegga, A. G.; Aronow, B. J.; Lee, C.-K.; Bernard, A.; Glasser, M. F.; Dierker, D. L.; Menche, J.; Szafer, A.; Collman, F.; Grange, P.; Berman, K. A.; Mihalas, S.; Yao, Z.; Stewart, L.; Barabasi, A.-L.; Schulkin, J.; Phillips, J.; Ng, L.; Dang, C.; Haynor, D. R.; Jones, A.; Essen, D. C. V. & Lein, C. K. &. E. Canonical genetic signatures of the adult human brain. *Nature Neuroscience,* 2016

[29] Konopka, G.; Friedrich, T.; Davis-Turak, J.; Winden, K.; Oldham, M.; Gao, F.; Chen, L.; Wang, G.-Z.; Luo, R.; Preuss, T.; Geschwind, D.; Friedrich, T.; Davis-Turak, J.; Winden, K.; Oldham, M.; Gao, F.; Chen, L.; Wang, G.-Z.; Luo, R.; Preuss, T. & Geschwind, D. Human-Specific Transcriptional Networks in the Brain. *Neuron,* 2012

[30] Parikshak, N. N.; Swarup, V.; Belgard, T. G.; Irimia, M.; Ramaswami, G.; Gandal, M. J.; Hartl, C.; Leppa, V.; de la Torre Ubieta, L.; Huang, J.; Lowe, J. K.; Blencowe, B. J.; Horvath, S. & Geschwind, D. H.

Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature,* 2016

[31] Fromer, M.; Roussos, P.; Sieberts, S. K.; Johnson, J. S.; Kavanagh, D. H.; Perumal, T. M.; Ruderfer, D. M.; Oh, E. C.; Topol, A.; Shah, H. R.; Klei, L. L.; Kramer, R.; Pinto, D.; Gumus, Z. H.; Cicek, A. E.; Dang, K. K.; Browne, A.; Lu, C.; Xie, L.; Readhead, B.; Stahl, E. A.; Xiao, J.; Parvizi, M.; Hamamsy, T.; Fullard, J. F.; Wang, Y.-C.; Mahajan, M. C.; Derry, J. M. J.; Dudley, J. T.; Hemby, S. E.; Logsdon, B. A.; Talbot, K.; Raj, T.; Bennett, D. A.; Jager, P. L. D.; Zhu, J.; Zhang, B.; Sullivan, P. F.; Chess, A.; Purcell, S. M.; Shinobu, L. A.; Mangravite, L. M.; Toyoshiba, H.; Gur, R. E.; Hahn, C.-G.; Lewis, D. A.; Haroutunian, V.; Peters, M. A.; Lipska, B. K.; Buxbaum, J. D.; Schadt, E. E.; Hirai, K.; Roeder, K.; Brennand, K. J.; Katsanis, N.; Domenici, E.; Devlin, B. & Sklar, &. P. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nature Neuroscience,* 2016

[32] Radulescu, E.; Jaffe, A. E.; Strau, R. E.; Chen, Q.; Shin, J. H.; Hy, T. M.; Kleinman, J. E. & Weinberger, D. R. Identification and prioritization of gene sets associated with schizophrenia risk by co- expression network analysis in human brain. *BioRXiv,* **2018**

[33] Gandal, M. J.; Haney, J. R.; Parikshak, N. N.; Leppa, V.; Ramaswami, G.; Hartl, C.; Schork, A. J.; Appadurai, V.; Buil, A.; Werge, T. M.; Liu, C.; White, K. P.; Horvath, S. & Geschwind, D. H. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science,* 2018

[34] Wang, M.; Roussos, P.; McKenzie, A.; Zhou, X.; Kajiwara, Y.; Brennand, K. J.; Luca, G. C. D.; Crary, J. F.; Casaccia, P.; Buxbaum, J. D.; Ehrlich, M.; Gandy, S.; Goate, A.; Katsel, P.; Schadt, E.; Haroutunian, V. & Zhang, B. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Medicine, 2016*

[35] Johnson, M. R.; Shkura, K.; Langley, S. R.; Delahaye-Duriez, A.; Srivastava, P.; Hill, W. D.; Rackham, O. J. L.; Davies, G.; Harris, S. E.; andMaxime Rotival, A. M.-M.; Speed, D.; Petrovski, S.; Katz, A.; Hayward, C.; Porteous, D. J.; Smith, B. H.; Padmanabhan, S.; Hocking, L. J.; Starr, J. M.; andAlessia Visconti, D. C. L.; Falchi, M.; Bottolo, L.; Rossetti, T.; Danis, B.; Mazzuferi, M.; Foerch, P.; Grote, A.; Helmstaedter, C.; Becker, A. J.; Kaminski, R. M.; Deary, I. J. & Petretto, &. E. Systems genetics identifies a convergent gene network for cognition and neurodevelopmental disease. *Nature Neuroscience,* 2016

[36] Parikshak, N.; Luo, R.; Zhang, A.; Won, H.; Lowe, J.; Chandran, V.; Horvath, S. & Geschwind, D. Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism *Cell,* 2013

[37] Hormozdiari, F.; Osnat, P.; Borenstein, E. & Eichler, E. E. The discovery of integrated gene networks for autism and related disorders. *Genome Research,* 2015

[38] Mahfouz, A.; Ziats, M. N.; Rennert, O. M.; Lelieveldt, B. P. & Reinders, M. J. Shared Pathways Among Autism Candidate Genes Determined by Co-expression Network Analysis of the Developing Human Brain Transcriptome. *J Mol Neurosci,* 2015

[39] Schanzenbacher, CT; Langer, JD & Schuman, EM. Time- and polarity-dependent proteomic changes associated with homeostatic scaling at central synapses. *eLIFE,* 2018

[40] Li, T; Wernersson, R; Hansen, RB; Horn, H; Mercer, J; Slodkowicz, G; Workman, CT; Rigina, O; Rapacki, K; Staerfeldt, HH; Brunak, S; Jenson, TS & Lage, K
A scored human protein-protein interaction network to catalyze genomic interpretation
Nature Methods, 2017