Supplementary Information S1 (box)
Methodology and analysis for cross-disorder transcriptomic analysis.

For the analysis in Figure 3, we re-analysed raw data from three published studies including ASD[1], schizophrenia[2], and Alzheimer's disease[3]. Each study was processed and analysed using the general flow described in **Figure 2a**.

Sample description: Raw microarray data from post-mortem gene expression studies and metadata were obtained from the published studies via GEO (GSE25821, GSE21138, GSE48350). Only data from cortical samples was used (ASD – Brodmann area (BA) 9 and BA41/2, schizophrenia – BA46, Alzheimer's disease – BA3 and BA11).

Normalization: Data were log2 transformed and quantile normalized using the affy and lumi packages in R[4,5].

Quality control: We balanced case/control status across all available biological and technical covariates (e.g., age, sex, batch, RIN, post-mortem interval, pH) so that, for each study, the case/control status was not significantly associated with any measured covariate ($p > 0.05$, minimizing correlations between case-control status and biological and technical factors) (**Figure S1**).

Outlier and batch effect removal: Outliers were defined as samples with standardized sample network connectivity Z scores < -2 (see Oldham et al., 2012[6]), and were removed. Batch effects were defined by array ID or hybridization date depending on the study, and were regressed out using the ComBat function of the sva package in R[7].

DGE analysis: Multiple array probes to the same gene were collapsed to unique genes using Ensembl 75 gene IDs by taking the maximum mean signal across all probes available for a given gene via the CollapseRows function[8]. DGE log fold changes were calculated using the limma package, with linear models including all available covariates[9]. Spearman's ρ was used to compare DGE signatures across disease.

Gene Co-expression network analysis: Consensus network analysis[10], a meta-analytic approach, was performed with the WGCNA package using signed network analysis[11]. A soft-threshold power of 16 was used for all studies, and consensus modules were defined based on a consensus quantile threshold of 0.2 using biweight midcorrelation (bicor), with minimum module size of 200, deepSplit parameter of 2, a merge threshold of 0.2, and without re-assigning genes using partitioning around medians[12] (pamStage = FALSE). Modules were summarized by their 1st PC (e.g., eigengenes) and eigengene-disease associations were evaluated with a two-tailed t-test. Functional enrichment of Gene Ontology pathways was assessed with GO-Elite[13].

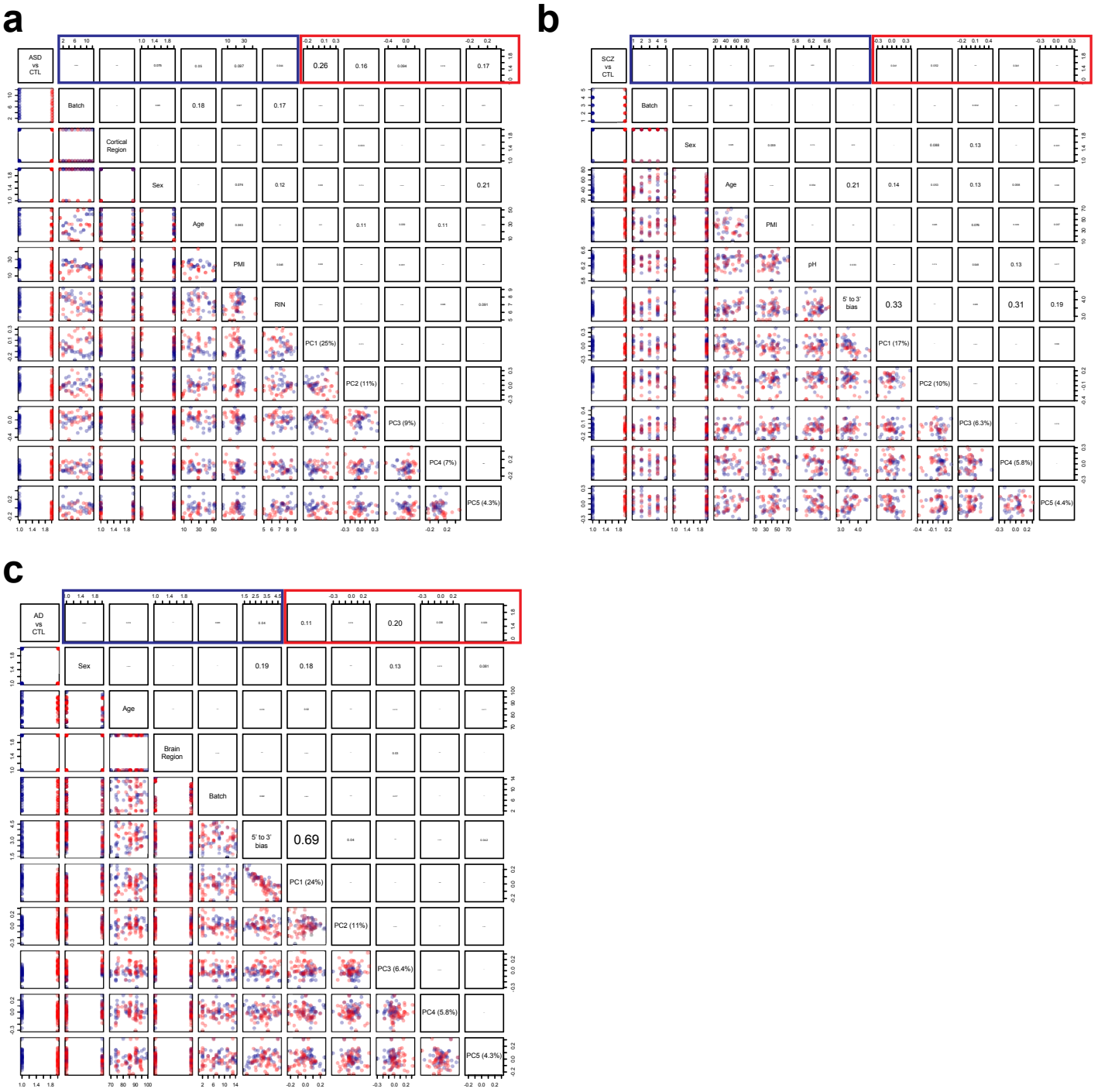

Figure S1.
Correlation panel plots of biological factors, technical factors, and principal components for each of the three studies utilized in Figure 3. The correlation between every biological and technical factor that was measured for **(a)** the ASD dataset, **(b)** the schizophrenia dataset, and **(c)** the Alzheimer's disease dataset. Abbreviations: SCZ, schizophrenia; AD, Alzheimer's disease; CTL, controls. In each panel (**a-c**), the sub-panels on the diagonal contain the names of biological and technical covariates that have been measured, as well as the first five principal components (PCs) of the gene expression data along with the percentage of the variance they explain in the data. The sub-panels in the top-right half of each grid contain correlation values (Spearman's ρ^2 for quantitative or binary variables, the adjusted R^2 for multifactor categorical variables) for each pairwise association, e.g. the pairwise association between ASD vs CTL status and PC1 in the panel **(a)** is 0.25. The font for the correlation values is scaled by the magnitude for display purposes. The sub-panels on the bottom-left are the pairwise scatterplots between variables, with cases in red, and controls in blue. For each panel, the blue rectangle highlights the correlations between case/control status and covariates. In each main panel (**a-c**) the red rectangle highlights the correlation between case/control status and each PC. In

general, good experimental design (randomization) minimizes correlations in the blue rectangle while high correlations in these sub-panels suggests case/control status is confounded by a batch or an unwanted factor. For this analysis, we explicitly balanced case/control status from the studies used here by removing samples that were outliers in age or RIN. Correlations in the red rectangle identify whether case/control status is contributing to a top PC in the gene expression data. Comparing (**a-c**) illustrates that datasets can dramatically differ in how they are affected by biological and technical factors, for example the schizophrenia data's 1st PC is moderately affected by 5' to 3' degradation (0.33), while the Alzheimer's disease dataset's 1st PC is strongly affected (0.69). Note that **(b-c)** were performed using Affymetrix arrays, which enable calculation of the 5' to 3' bias in transcripts, which we prefer use instead of RIN to directly measure RNA degradation. In **(a)** the data were collected on an Illumina array, where this measurement is not available, so we use the RNA integrity number, RIN.

References:

1. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).
2. Narayan, S. *et al.* Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Research* **1239,** 235–248 (2008).
3. Berchtold, N. C. *et al.* Synaptic genes are extensively downregulated across multiple brain regions in normal human aging and Alzheimer's disease. *Neurobiology of Aging* **34,** 1653–1661 (2013).
4. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (2008).
5. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20,** 307–315 (2004).
6. Oldham, M. C., Langfelder, P. & Horvath, S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Syst Biol* **6,** 63 (2012).
7. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127 (2006).
8. Miller, J. A. *et al.* Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics* **12,** 322 (2011).
9. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3,** Article3–25 (2004).
10. Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* **1,** 54 (2007).
11. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
12. Horvath, S. *Weighted Network Analysis: Applications in Genomics and Systems Biology*. (Springer, 2011).
13. Zambon, A. C. *et al.* GO-Elite: a flexible solution for pathway and ontology over-representation. *Bioinformatics* **28,** 2209–2210 (2012).