nature neuroscience

RESOURCE

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

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

Christopher L. Hartl[1,2], Gokul Ramaswami[2], William G. Pembroke[2], Sandrine Muller[3,4], Greta Pintacuda[3,5], Ashis Saha[6], Princy Parsana[6], Alexis Battle[6,7], Kasper Lage[3,4,8] and Daniel H. Geschwind[2,9,10,11] ✉

Gene networks have yielded numerous neurobiological insights, yet an integrated view across brain regions is lacking. We leverage RNA sequencing in 864 samples representing 12 brain regions to robustly identify 12 brain-wide, 50 cross-regional and 114 region-specific coexpression modules. Nearly 40% of genes fall into brain-wide modules, while 25% comprise region-specific modules reflecting regional biology, such as oxytocin signaling in the hypothalamus, or addiction pathways in the nucleus accumbens. Schizophrenia and autism genetic risk are enriched in brain-wide and multiregional modules, indicative of broad impact; these modules implicate neuronal proliferation and activity-dependent processes, including endocytosis and splicing, in disease pathophysiology. We find that cell-type-specific long noncoding RNA and gene isoforms contribute substantially to regional synaptic diversity and that constrained, mutation-intolerant genes are primarily enriched in neurons. We leverage these data using an omnigenic-inspired network framework to characterize how coexpression and gene regulatory networks reflect neuropsychiatric disease risk, supporting polygenic models.

Neuropsychiatric diseases are genetically complex, adhering to a polygenic architecture consisting of thousands of risk-conferring variants and genes[1]. In contrast to Mendelian disorders, where generalizable mechanistic insight can be obtained from the analysis of a single gene, the etiology of complex genetic disorders is organized around functional groups of genes, or pathways[1]. Genes within these groups are expected to be co-regulated and expressed at levels that permit the pathway to function[2,3]. RNA coexpression and protein–protein interaction (PPI) networks provide a powerful conceptual framework for understanding how such groups of genes are organized, with predictive power to prioritize disease-associated variation in polygenic disorders[4–6]. This framework aids in characterizing relevant biological pathways by arranging genes into smaller, tractable and coherent sets of modules for experimental analysis. Additionally, gene coexpression networks can further our understanding of complex, polygenic disorders by linking together genes that covary across prevalent cell types and cell states within the tissue of interest[3,7].

To inform our understanding of molecular mechanisms in the human brain and their disease relevance, we create an atlas of coexpression networks across 12 human brain regions from the Genotype-Tissue Expression (GTEx) project[8]. We compare different network construction methods and demonstrate that the coexpression relationships defined in these networks are robustly identified using multiple methods and orthogonal brain datasets. These networks constitute a new resource for understanding convergent pathways and brain regions affected by disease-associated variation in adult brain. We use this resource to address several biological questions. We show that coexpression is hierarchically organized into signatures ranging from those that are brain wide, to those that pertain to multiple or specific regions. For both autism spectrum disorder (ASD) and schizophrenia (SCZ), three major types of genetic or genomic signals—differential expression, rare high-impact variants and common low-effect variants—converge on cross-regional networks that implicate neuronal and neural progenitor cell (NPC) types. Lastly, we incorporate our networks into a model of genetic architecture, asking whether these networks exhibit a core–periphery structure that follows the recently framed omnigenic hypothesis[9]. We provide a web browser, HUBgene, to facilitate access to these data.

Results

Building robust human coexpression networks. To explore the molecular anatomy of the human brain, we used RNA-sequencing (RNA-seq) data from the GTEx Consortium (v7), focusing on the 12 major brain regions profiled: cerebellum (CBL), cerebellar hemisphere (CBH), dorsolateral prefrontal cortex (PFC), Brodman area 9 (BA9), Brodman area 24 (BA24), hippocampus (HIP), amygdala (AMY), hypothalamus (HYP), substantia nigra (SNA), nucleus accumbens (ACC), caudate nucleus (CDT) and putamen (PUT;

[1]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, CA, USA. [2]Program in Neurogenetics, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Surgery, Massachusetts General Hospital, Boston, MA, USA. [5]Department of Stem Cell and Regenerative Medicine, Harvard University, Cambridge, MA, USA. [6]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [7]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [8]Institute for Biological Psychiatry, Mental Health Center Sct. Hans, University of Copenhagen, Roskilde, Denmark. [9]Department of Psychiatry and Biobehavioral Sciences, Semel Institue, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [10]Center for Autism Research and Treatment, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [11]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. ✉e-mail: dhg@mednet.ucla.edu

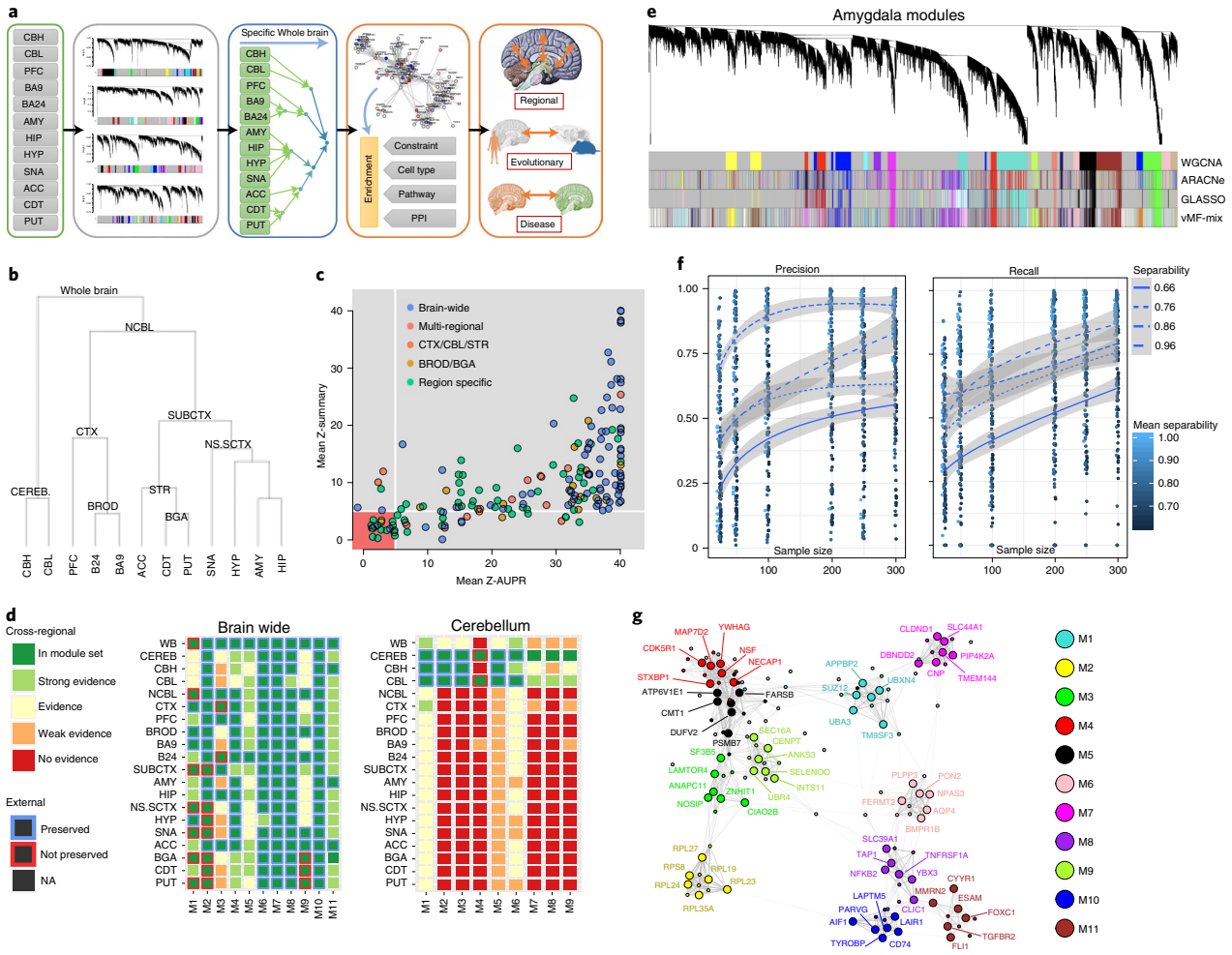


Fig. 1 | Human whole-brain coexpression atlas. a, Overview of module construction, annotation and downstream analysis. **b**, Module-merging hierarchy based on median within-region expression. **c**, Module preservation in external datasets, with the area containing weakly preserved and non-preserved modules (highlighted in red; $z < 5$). **d**, Module evidence across all regions of the human brain, for brain-wide and cerebellar module sets. Strong evidence: $z > 8$; evidence: $z > 5$; weak evidence: $z > 3$; no evidence: $z < 3$. NA, not applicable. **e**, Dendrogram from rWGCNA in AMY, showing a high degree of overlap between four methods of network construction and module identification. Colors under the dendrogram are default WGCNA colors, ordered by module size. **f**, Precision and recall of co-clustering a gene with the hub gene of its true module, as a function of module separability and sample size. Lines represent the mean, and bands denote the 95% confidence interval (CI). **g**, Example hub gene network of whole-brain modules. The top six hub genes by module kME were extracted (large circles, labeled) along with 80 randomly selected genes to inform the embedding (small circles, unlabeled). The edges are the topological overlap, and the network is embedded using the Fruchterman–Reingold algorithm.

Fig. 1a). Using a tissue hierarchy to structure consensus coexpression (Fig. 1b and Methods), we used robust weighted gene coexpression network analysis (rWGCNA) to create 311 coexpression modules for 20 hierarchical expression categories: 12 brain-region-specific categories (corresponding to each sampled region), 7 multiregional categories (corresponding to multiple, structurally linked regions; Fig. 1b) and a brain-wide category, while correcting for known technical factors (Extended Data Fig. 1), sample outliers and brains impacted by inflammation at time of death (Methods). We find that 87% (173/199) of the region-level modules are highly preserved in independent datasets (Fig. 1c and Supplementary Note) and that region-specific networks show low preservation in other brain regions (Fig. 1d). To further demonstrate the robustness of coexpression relationships to methodological factors, we show that modules were robust to multiple alternative network methods (Fig. 1e) and aggregation methods (Supplementary Fig. 1 and Supplementary Note). By downsampling our dataset, we establish that we have power to identify all module hub genes (Extended Data Fig. 1), and that the co-membership of gene pairs is identified with reasonable accuracy (Fig. 1f,g).

We summarized our analyses at the whole-brain, multiregional and region-specific levels, structuring our analysis in terms of 48 module sets, based on merging modules—within the tissue hierarchy—by their similarity (Methods and Supplementary Table 1). As expected, the most physiologically distinct regions, HYP, CBL and SNA, showed the largest number of region-specific modules (Fig. 1d). We were also able to identify modules representing components shared between specific regions, such as ependymal and choroid epithelial cells constituting the choroid plexus, which is juxtaposed in striatum and parts of the telencephalon, providing evidence that our modules reflect the biologically correct placement of this cell-type-specific module (Fig. 2a).

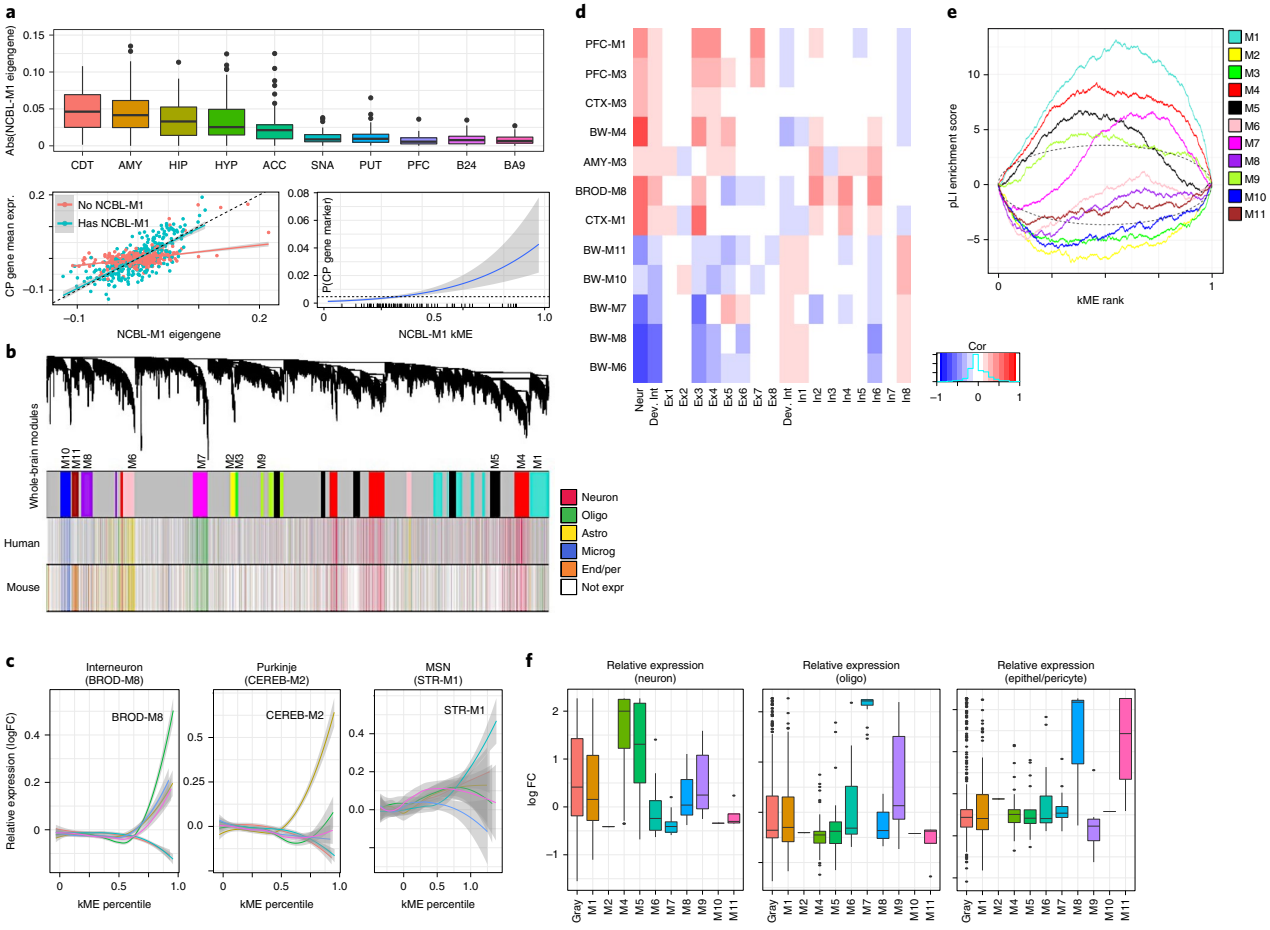


Fig. 2 | Cell-type heterogeneity relates to coexpression modules, mutation intolerance and evolution. a, Top, standard box plots (box displays quartiles, whiskers represent 1.5 times the interquartile range (IQR)) of absolute value of the eigengene of module NCBL.M1 plotted across regions, showing higher variance in regions adjacent to or accessible through ventricles. Region sample sizes are as specified in the main text. Left, relationship between NCBL.M1 eigengene and mean expression of choroid plexus (CP) marker genes in regions with and without an NCBL-M1 module (solid lines represent the least-squares fit, band denotes the 95% CI). Right, marginal probability (via logistic regression) of a gene being a choroid plexus marker, as a function of NCBL-M1 soft membership. **b**, Brain-wide modules largely correspond to cell class. WGCNA dendrogram at the whole-brain level, labeled by module and colored by human cell-type markers, and mouse cell-type markers (key: right). **c**, Relative expression in log fold change (FC) values (*y* axis; Methods) of neuronal marker genes for modules BW-M4, BROD-M8, CEREB-M2 and STR-M1 within interneurons from cortical single-cell sequencing, Purkinje neurons from cerebellar single-cell sequencing and medium spiny neurons from mouse striatal single-cell sequencing, as a function of module kME (*x* axis; Methods). **d**, Factorization-based decomposition of bulk expression from aggregated cortical single-cell sequencing (Methods). Pearson correlations between module eigengenes and cell-type factors for BW, CTX and PFC modules came from decomposition of dorsolateral PFC expression, for AMY from decomposition of AMY bulk expression and for BROD from decomposition of B24 bulk expression (Methods; blue, negatively correlated; red, positively correlated). **e**, Gene-set enrichment analysis plots for LoF-intolerant genes[14] (pLI > 0.9) for all whole-brain modules, colored by module. The extent of deviation from the horizontal reflects the degree of enrichment (+) or depletion (−) at a given centrality (kME). Black dotted line: 95% confidence band based on the Brownian bridge. **f**, Standard box plot (box displays quartiles, whiskers represent 1.5 times the IQR) of lncRNA relative expression in single-cell data, grouped by the imputed module in RNA-seq data from BA9. Overlapping module sizes (number of lncRNA): gray (518), M1 (116), M2 (1), M4 (68), M5 (46), M6 (16), M7 (17), M8 (4), M9 (13), M10 (1) and M11 (4).

Module sets reflect brain cell types and processes. Previous work has shown that cellular composition is a major driver of gene expression in tissue[10,11]. We therefore expected whole-brain coexpression modules to represent major cell classes, and multiregional or regional modules to represent more specialized cell subtypes. We found that 5 of 11 whole-brain modules (M4, M6, M7, M10 and M11) represented the 5 canonical brain cell classes (Fig. 2b, Supplementary Table 2 and Methods), and two additional modules (M1 and M8) reflected neuronal differentiation and glial activation (both microglia and astrocyte), respectively (Extended Data Fig. 2). The most significantly cell-type-enriched module, BW-M1, enriched for markers of NPCs, neuronal migration and differentiation, and is most preserved in neurogenic regions, suggesting that it corresponds to adult NPCs. BW-M5, another neuronal module, enriched for neurodegenerative disease pathways (Extended Data Fig. 2). Modules in the ACC enriched for morphine addiction and alcoholism terms (Supplementary Table 2 and Supplementary Note), and the region-specific module BRNHYP-M7 enriched for the oxytocin signaling pathway, meeting prior expectations. Thus, we hypothesized that region-specific modules may reflect region-specific

biology, such as unique cellular subtypes. We used single-cell data to confirm (Supplementary Note) that the region-specific modules BROD-M8, CEREB-M1 and STR-M2 correspond to regional cell classes—cortical interneurons, Purkinje cells and medium spiny neurons, respectively (Fig. 2c and Supplementary Table 2). We identify three region-specific excitatory neuron modules (CTX-M3, PFC-M1 and PFC-M3), and an inhibitory neuronal module (Fig. 2e), BROD-M8 (neuropeptide signaling and perception of pain), while PFC-M1 (serine/threonine kinase activity) and PFC-M3 (circadian rhythm) both enriched for a specific excitatory cell type (Supplementary Table 3).

Signatures of mutational intolerance. Brain-expressed genes are under higher levels of purifying selection than average[12] and genes intolerant to loss-of-function (LoF) mutations are expressed disproportionately in the brain[13], but whether mutation intolerance is a general feature of brain-expressed genes is not known, so we explored the relationship of mutation intolerance to specific cell types. Across all modules, the whole-brain module BW-M1 was the most significantly enriched for LoF-intolerant genes (defined as pLI[14] > 0.9), followed by BW-M4 and closely related BW-M5 (Fig. 2d and Extended Data Fig. 2), all of which are neuronal. Several of the regional or cell-type-specific modules, such as BROD-M8, CEREB-M2 and STR-M1, were also enriched in LoF-intolerant genes. Notably, only one glial module, BW-M7 (oligodendrocytes), enriched for LoF-intolerant genes, but its degree of enrichment was lower than that of neurons (Supplementary Table 4). The concentration of LoF intolerance within neuronal modules suggests that disruption of genes in microglia and astrocytes is buffered.

Identifying cell-type-specific long noncoding RNA. Long noncoding RNA (lncRNA) are diverse species that play roles in neurodevelopment and neuropsychiatric disease, yet only 52 known lncRNA species were quantified in the initial GTEx analysis[8]. We used gradient-boosted trees to learn module signatures (Methods), assigning 286 lncRNAs to modules; the majority of lncRNAs associated with neuronal module BW-M4 (66) or the NPC module BW-M1 (109; Supplementary Table 5). Notably, more than 20% (61/286) of the cell-type-specific lncRNAs were previously shown to be dysregulated in neuropsychiatric disease, augmenting previous work on differential expression of lncRNA in ASD (Supplementary Table 5)[15].

Identifying cell-type-specific gene isoforms. We next integrated isoform-level expression with cell-type modules (Fig. 3a and Methods), identifying 1,987 isoforms showing specificity to major cell types, of which 549 were neuronal, 543 were astrocytic and 696 were oligodendroglial (Supplementary Table 6). We validated a subset of these findings in sorted cells, quantified at the isoform-level (Supplementary Note and Fig. 3b) and built cell-specific isoform maps for D1/D2 medium spiny neurons, Purkinje cells, basket cells and inhibitory neurons—cells that had strong enrichments and region-specific modules (Fig. 3c and Supplementary Table 6). All modules enriched for synapse-related functions, suggesting that splicing plays a major role in regional cell-type synapse diversity (Fig. 3d).

A subset of ASD risk genes switches isoforms across cell types. We observed that in 7% of cases, the parent gene of an isoform differed in coexpression relationships from an alternatively spliced derivative (Fig. 3e). We identified 52 genes exhibiting switching between cell-type modules, 11 of which showed neuron/astrocyte switching (BW-M4/BW-M6) and 8 of which showed neuron/oligodendrocyte switching (BW-M4/BW-M7), trends validated in sorted cells (Extended Data Fig. 3). Of the 11 neuron/astrocyte switching genes, *ANK2* and *SCP2* are known autism susceptibility genes (Fig. 3f,g), while two others, *ERGIC3* and *PDE4DIP*, are weaker candidates (AutDB database score of 4; Extended Data Fig. 3; $P < 0.01$, Fisher's exact test). We validated previous observations of isoform switching of ANK2 in ASD and SCZ[16] at the protein level, establishing that the long isoform is primarily neuronal (Fig. 3h and Extended Data Fig. 4). The neuronal *ANK2* transcript includes a giant exon, which is an organizer of initial axon segments and a stabilizer of GABA-A synapses[17], confirming a neuron-specific role for this *ANK2* variant.

Ribosomal genes are downregulated across the cortex. We next sought to understand module-level regulation and its relationship with differential gene expression across brain regions. We developed a regional contrast test (RCT; Fig. 4a) to test a gene's regional enrichment (Methods and Supplementary Table 7), which we examined at varying degrees of granularity (Fig. 4b). Genes upregulated in subcortical regions enriched for nonneuronal cell-type modules ($P < 1 \times 10^{-10}$ for BW-M11, BW-M6, BW-M8, BW-M10 and BW-M7), consistent with a higher glia to neuron ratio (Fig. 4c). Conversely, we found BW-M4 (neuronal) to enrich for the genes upregulated in the cortex compared with subcortical regions. Interestingly, we observed a significant enrichment in BW-M2 ($P = 4.89 \times 10^{-3}$), a module dominated by RNA from small and large ribosomal subunits (Fig. 4d,e), for subcortical upregulated genes. This is consistent with the observation that ribosomal turnover drastically increases in cultures with a higher proportion of glial cells[18].

Regional specificity of neuropsychiatric disorder networks. We next assessed the regional specificity of disease-associated transcriptomic modules by reevaluating changes identified in postmortem tissue from 11 publications representing multiple childhood and adult brain disorders (Supplementary Note). We found that a common set of modules are involved in overlaps across every coexpression study: BW-M1, BW-M3, BW-M4, BW-M6 and BW-M10 (Extended Data Fig. 5). At least one, and in some cases every, disease-significant module overlapped with at least one of our whole-brain or multiregional modules (Supplementary Table 8). Thus, although each study was performed in a specific brain region, the modules that associate with disease largely reflect brain-wide coexpression signatures. While this definitely does not rule out region-specific components for each disease, it does suggest that genetic risk has brain-wide impact in neuropsychiatric disorders.

Brain-wide and regional pathways in neuropsychiatric disease. We next investigated whether genetic risk for neuropsychiatric disease converges onto region-specific or cross-regional modules, identifying two whole-brain modules, BW-M4 (neuron) and BW-M1 (NPC) that enriched for ASD-linked rare variants (Fig. 5a), SCZ GWAS signal (Fig. 5b), and that manifest disrupted expression in ASD postmortem brain relative to controls (Fig. 5c–g). We also identify two regional modules, CTX-M3 (activity-dependent regulation and endocytosis) and CEREB-M1 (mRNA binding), that show ASD rare-variant and SCZ genome-wide association study (GWAS) enrichment. Remarkably, both modules showed significant preservation in control brain, but not in ASD postmortem brain (Fig. 5g), consistent with the disruption of these modules in ASD.

BW-M4 enriched for Gene Ontology (GO) terms related to membrane organization and ion transport, consistent with convergence of risk onto synaptic signaling pathways[19,20] (Extended Data Fig. 6 and Supplementary Note). BW-M1 contained genes and pathways corresponding to neurogenesis, differentiation, migration (Fig. 6a) and RNA splicing (Fig. 6b,c). Genes within BW-M1 are strongly LoF intolerant, and the genes in this module were upregulated in ASD cortex (Fig. 5d), including the transforming growth factor beta signaling pathway (false discovery rate (FDR) = 0.0047; STRING)[21], key *REST* co-repressors *CTDSPL* and *RCOR1*, as well as

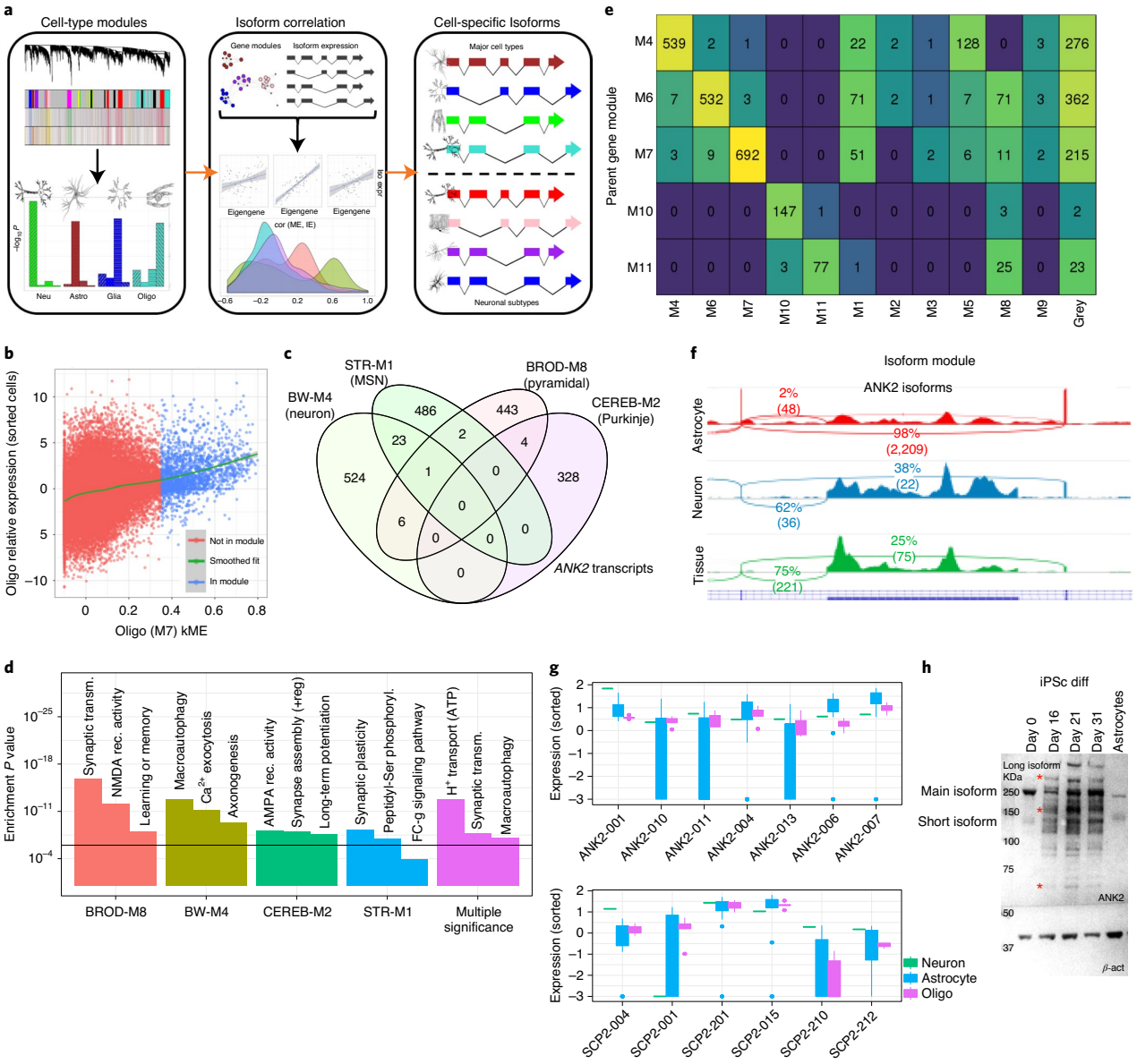


Fig. 3 | Creating a catalog of cell-type-specific isoforms. a, Overview of isoform assignment on the basis of kME to cell-type modules. Isoforms were correlated with module eigengenes to identify cell-type-specific isoforms. **b**, Isoform relative expression (log fold change of transcripts per million (TPM)) in oligodendrocytes plotted against isoform kME to BW-M7 showing a significant positive relationship ($P = 5 \times 10^{-7}$, linear regression likelihood-ratio test, two-sided). **c**, Venn diagram of isoforms assigned to neuronal subtypes showing extremely high specificity. **d**, GO enrichment of parent genes of cell-subtype-specific isoforms identified cell-type-specific pathways. Top module-specific terms are shown, followed by terms that were significant across multiple subtypes (minimum *P* value shown). **e**, Assignment of daughter isoforms of genes with membership to a whole-brain cell-type module, showing that most daughter isoforms are either assigned to the parent gene module, or to the gray (un-clustered) module. **f**, Integrative Genomics Viewer visualization of the event differentiating the astrocyte and neuron isoforms of ANK2, the inclusion of the giant exon, in sorted-cell data. **g**, Standard box plot (box displays quartiles, whiskers represent 1.5 times the IQR) of expression of ANK2 and SCP2 transcripts in sorted-cell data, showing isoform switching between neurons and astrocytes. **h**, Western blot of ANK2 across induced pluripotent stem cell (iPSc) differentiation into neurons (first four columns), and within astrocytes (column 5), demonstrating the presence of two long isoforms specific to neurons (red asterisk).

differentiation repressors *ADH5*, *TLR3*, *SOX5*, *SOX6*, *PROS1* and *SPRED1* (ref. [22]). Module trajectories showed prenatal upregulation, with continuing postnatal expression into early adulthood (Fig. 6d,e), evidence that one component of ASD may be brain-wide changes in neuronal proliferation/differentiation/maturation balance beginning in early development and that persist[23–25].

CTX-M3 and CEREB-M1 also show an enrichment for de novo LoF variants linked to ASD, enrichment for SCZ GWAS risk variants, and are disrupted in postmortem brain from individuals with ASD (Fig. 5). Both modules show region-specific coexpression (Extended Data Fig. 7), enrich for PPI (CEREB-M1 $P < 7 \times 10^{-15}$; CTX-M3 $P < 0.0023$), as well as LoF-intolerant genes, indicating

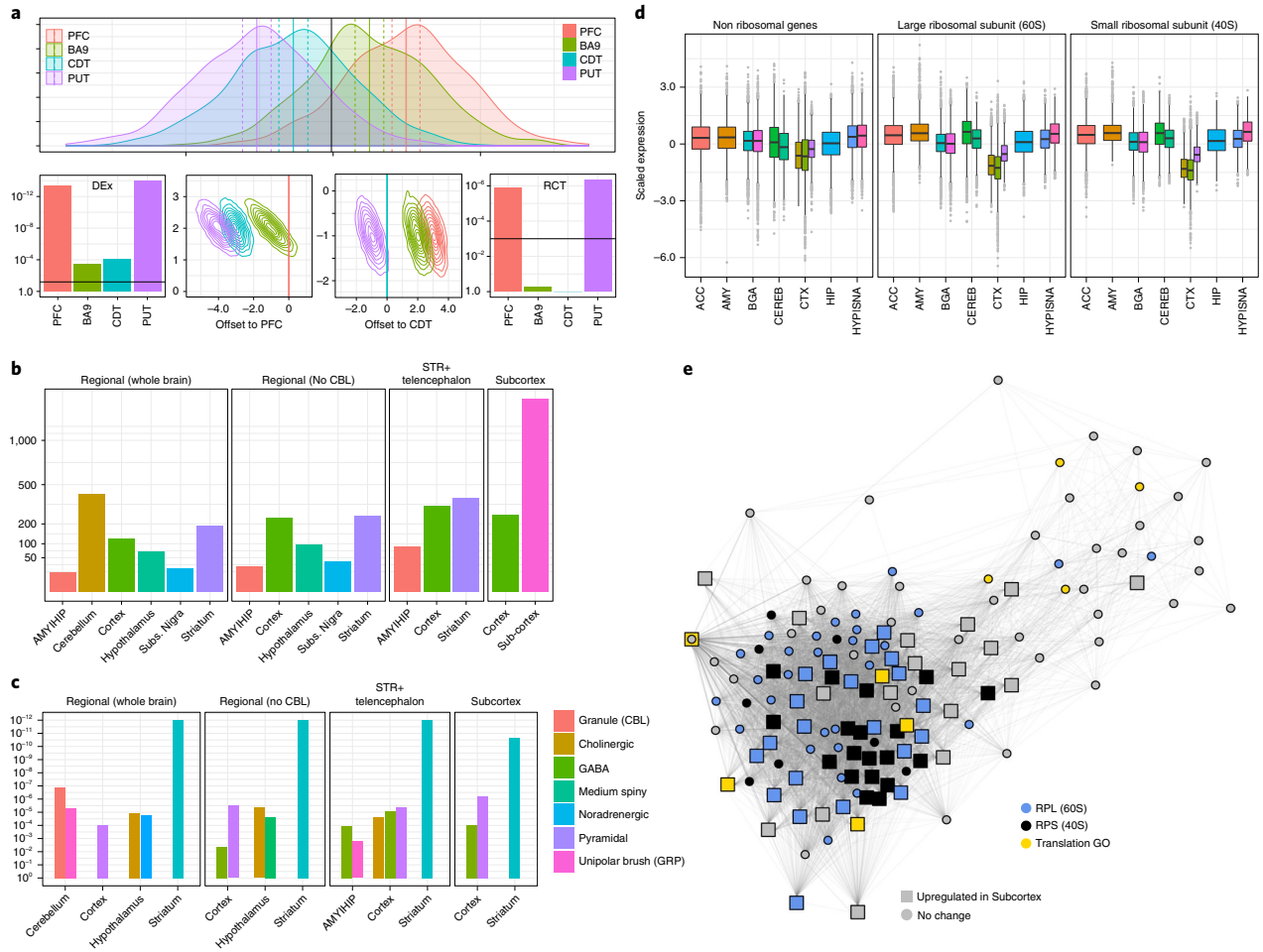


Fig. 4 | Region-specific gene upregulation reflects region-specific cell types and ribosomal turnover. a, Overview of the RCT (Methods); example of heterogeneous data where mean expression within each region differs from the global mean (top; *x* axis, gene expression; *y* axis, density; vertical lines correspond to regional means (solid) and ± two standard errors (dashed)). With only 50 samples, all regions were significantly differentially expressed in a global manner (bottom left, line at $P = 0.01$). Visualization of the RCT statistic for PFC and CDT (bottom middle). The PFC mean (set to 0; red vertical bar) overlaps only a small amount of the confidence region for one other region, while confidence regions straddle the CDT mean, demonstrating that PFC showed extremal expression while CDT did not. The RCT statistic identified the two most distinct tissues (PFC and PUT) with upregulated and downregulated differential gene expression compared to all other regions (right). **b**, Count of genes (*y* axis) significantly upregulated within brain regions, across four contrast backgrounds (labeled on top; $q < 0.1$; FDR-corrected signed RCT). **c**, Cell-type enrichments for the upregulated genes from the corresponding comparisons in **b**. GRP, gastrin releasing peptide–positive. **d**, Standard box plot (box displays quartiles, whiskers represent 1.5 times the IQR) of scaled expression (per gene across tissues) for all genes in BW-M2, showing cortex-specific downregulation of ribosomal subunits. **e**, PPI coexpression network (edge: gene-expression correlation only for known InWeb interactors; Supplementary Note) for genes in BW-M2, embedded by Fruchterman–Reingold. Large squares denote genes upregulated in subcortical regions compared to cortical regions; colors represent ribosomal subunits (blue, black) and GO regulation of translation (yellow), showing a sizeable fraction of the module core, a substantial fraction of large ribosomal subunit (RPL) and nearly all small ribosomal protein subunit (RPS) mRNA are upregulated in subcortical regions.

that they contain essential biological pathways. CTX-M3 enriches for RNA processing and mitochondrial complexes[26] (Fig. 6f). Despite these broad terms, we confirmed the cortical specificity of the coexpression of CTX-M3 hub genes in the Allen Human Brain Atlas (Fig. 6g). The presence of *FMR1*, *ATRX* and others involved in activity-dependent gene regulation in CTX-M3 highlights this process in disease pathophysiology[27]. Indeed, 10% of activity-dependent genes from a published study[28] fell into CTX-M3 ($P = 0.0472$; Fig. 6h and Supplementary Note), as did the mitochondrial ribosome (21 genes, $P < 1.7 \times 10^{-10}$). Other components of this module include alternative polyadenylation and alternative splicing, endocytosis regulation and sorting nexins, consistent with their likely role in supporting neuronal activity-dependent processes that are disrupted in ASD.

Networks and omnigenics in neuropsychiatric disorders. Complex disorders are influenced by large numbers of genetic variants and genes. Gene networks from disease-relevant tissues can capture interactions between these genes and have been hypothesized to inform disease heritability. We sought to incorporate network distance, as defined by brain-wide and regional coexpression networks, into a model of genetic architecture, and examine the role that coexpression networks play in the genetic architecture of neuropsychiatric disorders.

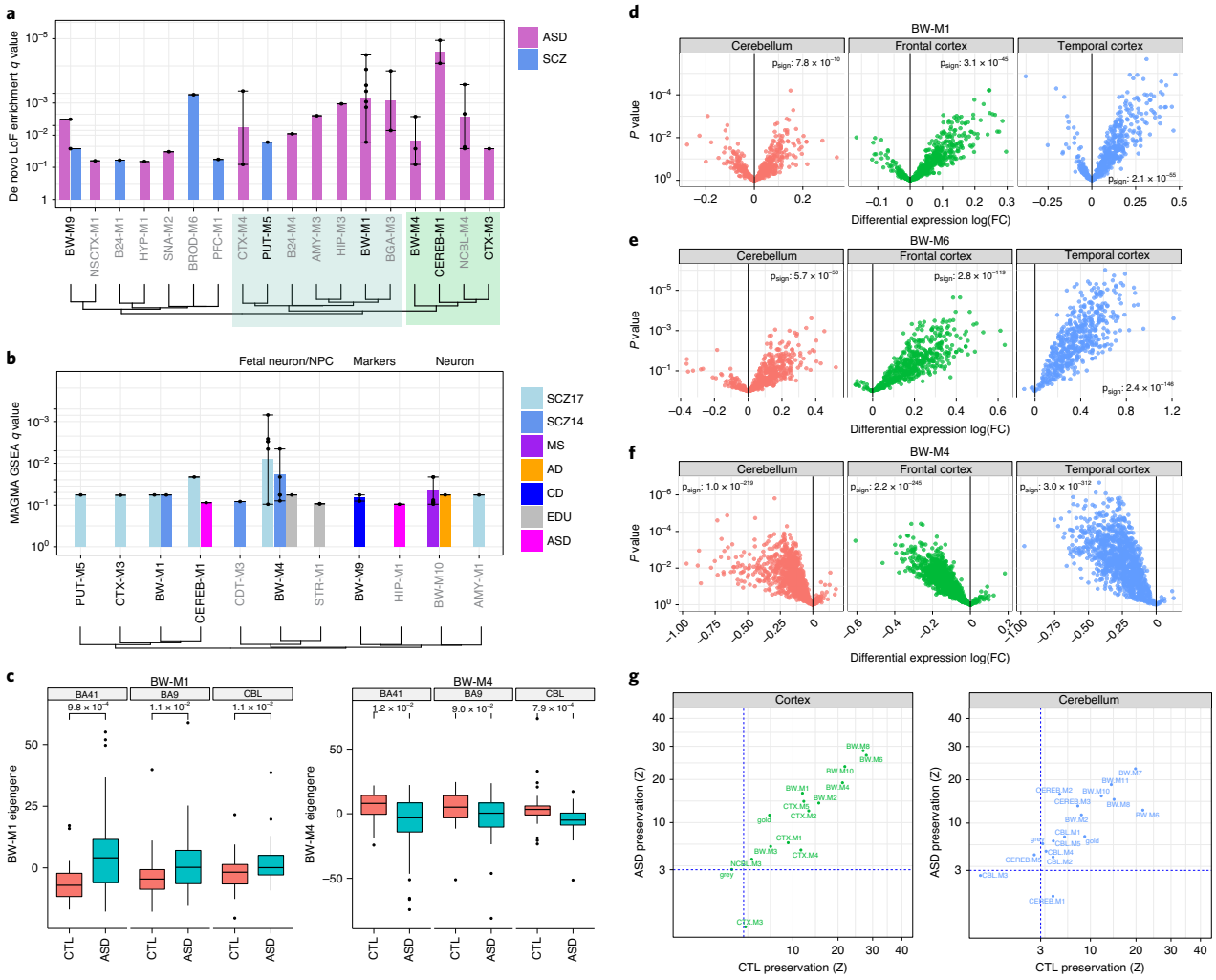


Fig. 5 | Gene-level module enrichment for de novo protein-truncating variants, GWAS summary statistics and differential expression. a, FDR values (Fisher's exact test) for enrichment of de novo LoF variants from ASD and SCZ within modules, summarized to module sets. Bar height denotes the geometric mean of FDR, and whiskers represent the range of significant FDR values for modules within the module set. Modules with bold labels on the *x* axis show enrichment from GWAS summary statistics (**b**; Methods). Module sets were ordered by Jaccard similarity between their index modules. Green region shows modules enriched for neuronal markers. Blue region shows modules enriched for fetal neuron, mitotic progenitor or outer radial glia markers. **b**, Gene-set enrichment analysis (GSEA) FDR values (MAGMA; Methods) for GWAS summary statistics within modules. Method of ordering is identical to **a**. **c**, Standard box plot (box displays quartiles, whiskers represent 1.5 times the IQR) of module eigengene expression for BW-M1 and BW-M4 in ASD and control brains across three regions and associated *P* values from a prior ASD sequencing study. *P* values were derived from a *t*-test (two-tailed). **d**–**f**, Volcano plots for individual genes in modules BW-M1 (NPC), BW-M6 (astrocyte) and BW-M4 (neuron) in a prior ASD sequencing study; the *x* axis shows log fold change values, and the *y* axis shows *P* values (linear mixed model); sign-test *P* values are inset demonstrating an overabundance of upregulated genes in BW-M1, BW-M6 and downregulated genes in BW-M4 (Methods). **g**, Module preservation statistics for ASD and controls calculated separately for cortical and cerebellar modules show highly consistent patterns, except for CTX-M3 and CEREB-M1, which are differentially preserved.

Motivated by the recently proposed omnigenic model of disease[9], wherein disease risk is conferred by the (potentially indirect) disruption of a small number of core genes, we constructed a model whereby allelic effect size is a function of network distance to simulated core genes (network genetic architecture; Methods). We simulated variants to generate a frequency–effect–distance distribution (Methods), observing that the resulting effect size and heritability distributions resembled those derived from the omnigenic hypothesis (Fig. 7a,b)[9], such that high-effect variants fell very near to core genes. We next asked how central genes capture a core–periphery structure for two common neuropsychiatric disorders: ASD and SCZ. We evaluated whether either (1) network-central genes or (2) genes implicated by rare variants behave as 'core' genes under this model (Methods).

We evaluated network-central genes across multiple networks, using whole-blood coexpression as a comparison (Methods). We found that even the largest observed value in blood indicated that only 52% of the likely high-impact genes fall near network core genes—below the simulated baseline (Fig. 7). The largest observed value across cortical networks was even lower at 0.44 (Fig. 7). We observed a significant enrichment for brain (FDR < 0.05, Fisher's exact test) and blood network distances (Supplementary Table 9),

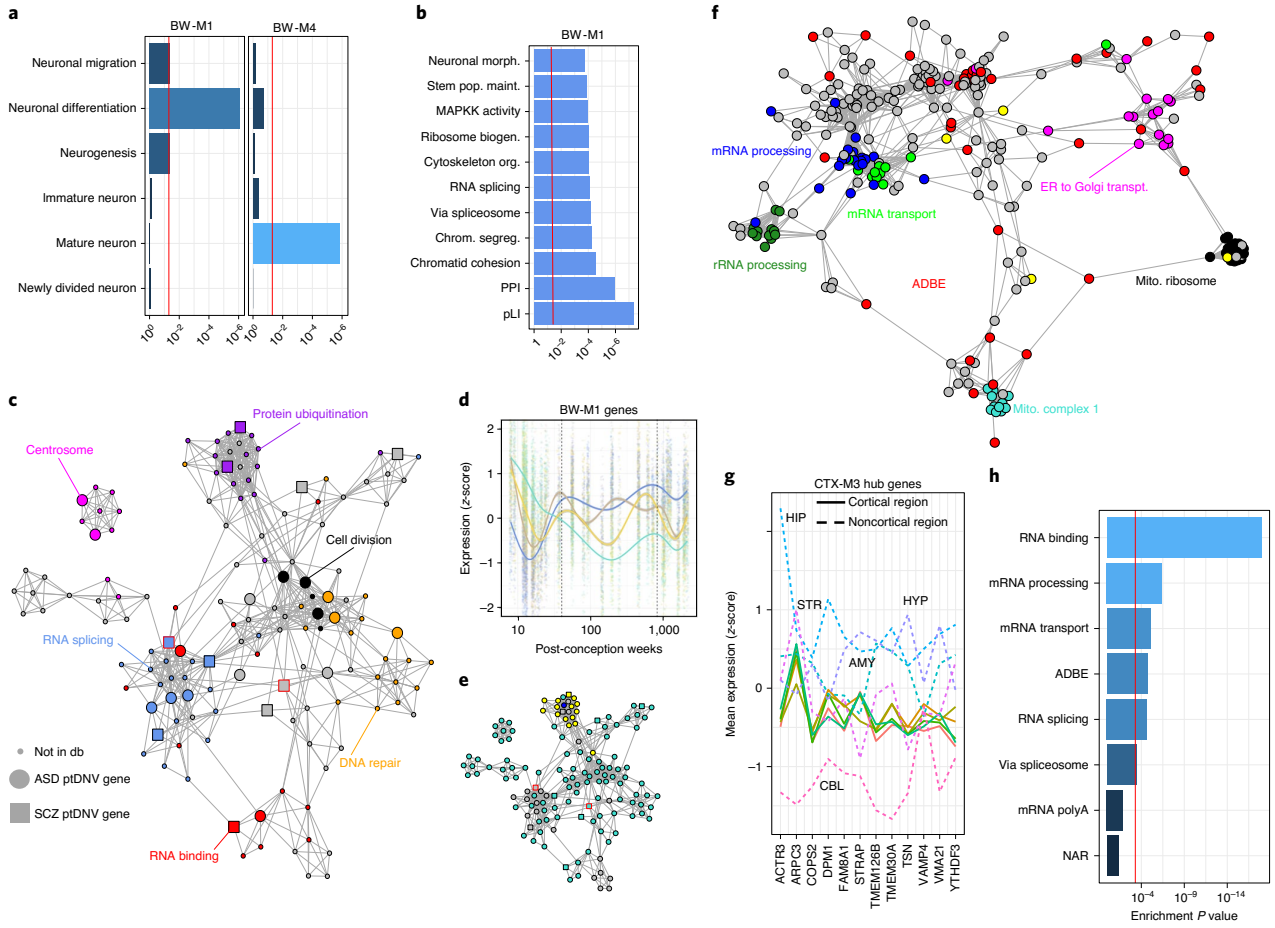


Fig. 6 | Ontologies, PPI networks and expression profiles of ASD-associated modules. a, Enrichment *P* values (Fisher's exact test) for neuron-related ontologies in two ASD-associated whole-brain modules. **b**, Combined (geometric mean) enrichment *P* values of ontologies for all modules in module set BW-M1 that showed enrichment for ASD-implicated de novo LoF mutations. **c**, Coexpression and PPI network of BW-M1 (edge: gene-expression correlation only for known InWeb interactors, embedded with Fruchterman–Reingold layout) highlighting de novo LoF mutations (large nodes) and ontologies (colors). ptDNV, protein-truncating de novo variant. **d**, Expression of BW-M1 across developmental time points, split into subclusters of four component modules using WGCNA (Supplementary Note) showing a cluster of genes downregulated after conception (teal). The scattered gray module is not shown. **e**, Assignment of network nodes in **c** to the subclusters in **d** via label propagation, demonstrating that the ubiquitination-related component of BW-M1 is maintained into adulthood (yellow), but that the bulk of the module is downregulated (teal). **f**, Coexpression and PPI network for CTX-M3, colored by enriched gene ontology sets. ER, endoplasmic reticulum. **g**, Expression profile of CTX-M3 hub genes across brain regions, demonstrating tight co-regulation in cortical regions (solid lines) by virtue of small variance, and highly variable coexpression across noncortical regions (dashed lines). **h**, Enrichment *P* values (Fisher's exact test) of the CTX-M3 module for gene ontologies, including bulk endocytosis genes.

demonstrating that the genetic architecture of these diseases reflects network distances, but still does not clearly separate core and peripheral genes, as defined by coexpression.

It may be that coexpression networks capture the correct notion of gene–gene distance, but network-central genes are not the correct core genes. We therefore used genes implicated by major effect size rare variants both to define core gene sets and to compute the test statistic (Methods). As in the network-central gene analysis, we found that the core genes defined in this manner were not clearly separated from periphery (Fig. 7d). It also is possible that bulk coexpression data failed to capture the appropriate core–periphery relationships. Therefore, we used the InWeb PPI network from brain, empirical gene regulatory (Tf-driven) networks from brain tissue and cell types (Methods)[29], and coexpression methods based on partial correlation (Supplementary Note), repeating the analyses above using high-connectivity genes as central genes. We found that the core–periphery structures in these other networks also did not mirror the expectations of our omnigenic-like model (Extended Data Fig. 8).

The inability of the coexpression networks to separate a core and periphery according to our test may reflect any of several explanations: (1) that the disorders assessed do not have a core–periphery gene structure, indicating that the omnigenic model does not explain their architecture[30], (2) or that peripheral master regulators are somewhat common among the candidate core genes (for example, de novo LoF genes) tested above, but are not an appropriate core set. However, in our analysis, we excluded known transcription factors, DNA-binding proteins, and RNA-binding proteins and noncoding genes, so master regulators present among the candidate core genes would need to regulate expression without directly binding DNA or RNA (Methods). Finally, it could be that our chosen metrics of centrality are not the correct property for assessing

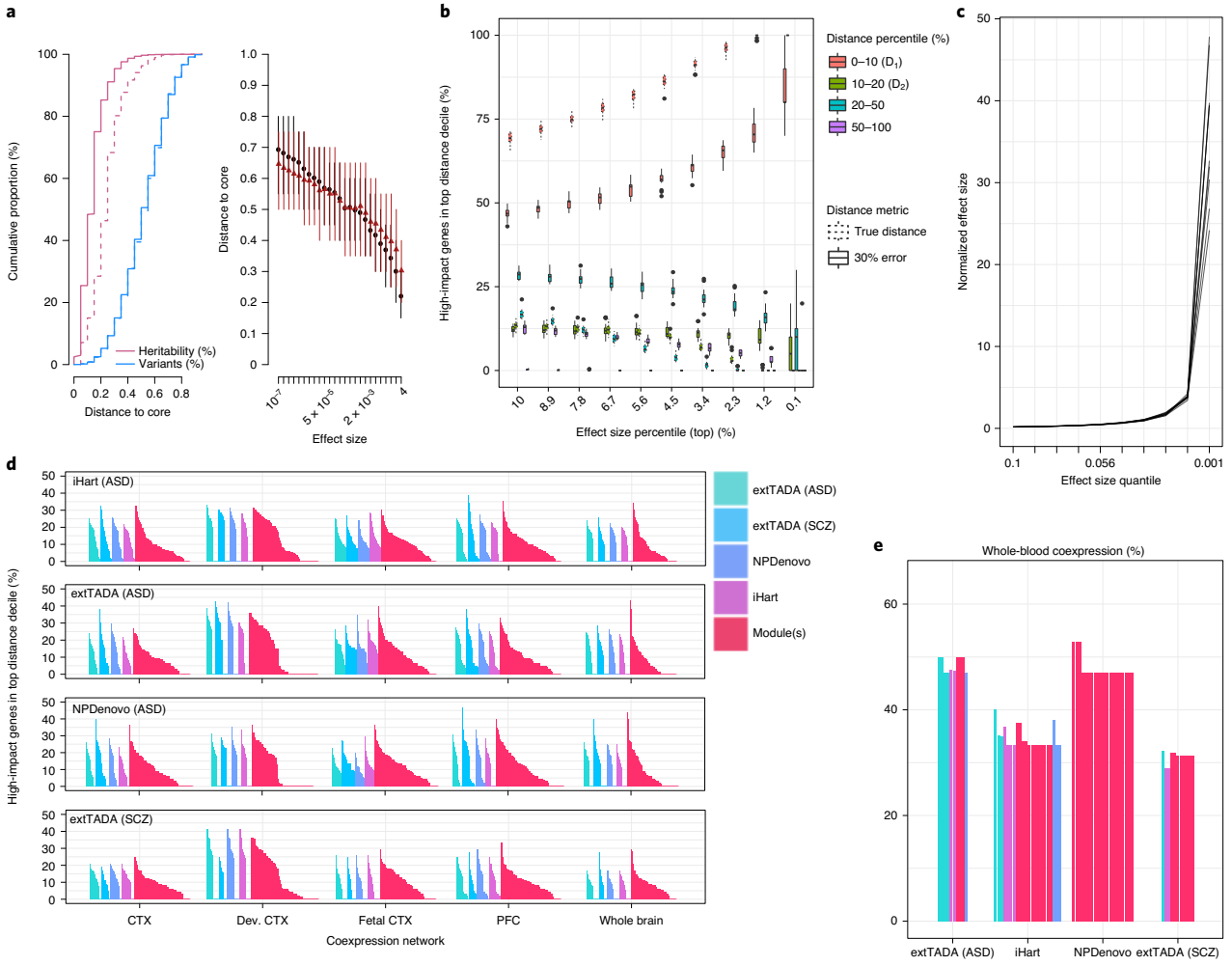


Fig. 7 | Characterizing core–periphery structure of high-impact neuropsychiatric disease genes across multiple networks. a, Example simulation of network genetic architecture, where the variant effect size decayed rapidly with distance to the core. Left, cumulative proportion of genes (blue) and heritability (pink) along the distance distribution. The dashed line shows the cumulative heritability when the true distance was replaced by a corrupted (30% error) distance. Right, the relationship between core distance and effect size resulted in high-effect variants only appearing very close to core genes as predicted by the omnigenic model. Points indicate the mean, and lines extend to the minimum and maximum values. Distance was computed from 50 simulations. **b**, High-impact genes are defined by the effect-size percentile on the *x* axis, and the percentage of genes falling into the core-distance decile (phi) is plotted on the *y* axis. This plot encompasses 20 simulations. Dashed boxes represent the expected values for Φ when the distance is error free, while solid boxes represent the case where distance is 30% corrupted by error. Boxes are standard (box displays quartiles, whiskers represent 1.5 times the IQR). **c**, Validation of the effect-size distribution. The effect size of each quantile was normalized to the effect size for which a balanced GWAS of 10,000 samples had 80% power; the highest-impact variants were only 20–50 times stronger than empowered variants. **d**, All values of Φ across distance metrics, core set size, module definitions and brain coexpression networks, demonstrating that no value of Φ exceeded 50%. **e**, Top ten Φ values (per core set) for the GTEx whole-blood coexpression network, demonstrating that coexpression networks from brain have similar Φ values to non-brain networks, reinforcing the notion that brain coexpression networks fail to reflect an omnigenic-like structure.

omnigenic architecture. While we cannot definitively assert which of these influenced our results, it is clear that genetic effects appear to be more continuously spread across coexpression, PPI or transcriptional regulatory networks, consistent with polygenic models, rather than separating 'core' genes from a periphery.

Discussion

Gene coexpression networks provide a powerful organizing framework for studying the nervous system[31–34]. That gene expression markers for major cell classes can be identified from bulk tissue coexpression is now well established[7,10,35]. However, most studies have not assessed whether such networks were specific to the brain regions studied, or more generalizable. Here, we construct a robust resource aimed at establishing common and region-specific aspects of gene coexpression within the brain. We identified 11 whole-brain coexpression modules, corresponding to common cellular components such as major neuron and glial types, and regional modules capturing signatures of cell subtypes. We demonstrate that: (1) the convergence of genetic risk in ASD and SCZ is primarily reflected in pathways common across brain regions, rather than specific to a single region; (2) disease risk in ASD and SCZ is enriched in downregulated neuronal and neurogenesis modules (several of these modules implicate downregulation of activity-dependent transcriptional programs in the cerebral cortex, a broad regional effect);

(3) cell-type-specific lncRNA and isoform co-regulation are included in networks, and isoform-level analysis is likely essential to interpret disease associations; and (4) brain RNA coexpression, PPI and co-regulatory networks do not cleanly capture the dichotomous core–periphery structure proposed by the omnigenic model, but rather support a continuous model. We provide a browser, HUBgene, to facilitate access to these networks and permit their broader exploration (http://geschwindlab.org/gclabapps/hubgene/home/).

We developed two methods for imputing coexpression networks in new data, and applied these to lncRNA and isoform quantification to identify cell-type-specific expression from bulk tissue measurements. Here, we provide a first generation set of 1,987 cell-type-specific isoforms for major cell classes in the brain, of which 549 are neuronal, 543 are astrocytic and 696 are oligodendrocytic. Remarkably, several of these isoforms, including four ASD risk genes, manifest isoform switching between neurons and glia. We showed that synaptic isoforms represent a major source of regional transcriptomic diversity among neuronal subtypes, and that neuropsychiatric risk genes are expressed at the synapse of multiple neuronal subtypes[36].

Our findings that ASD-linked de novo LoF mutations and SCZ GWAS signals enrich in brain-wide neuronal and neurogenesis modules underscore previous findings linking both common and de novo variation to synaptic genes[37,38], neuronal genes[39,40], developmentally expressed genes[41,42] and neurogenesis pathways[43,44]. We show that the pattern of enrichment in most cases is not region specific, implying likely widespread effects of these genetic risk variants on brain function.

The only region-specific modules with convergent evidence across disease and modality were CTX-M3 and CEREB-M1, which appear to reflect activity-dependent transcriptional profiles identified in previous studies. *VAMP4*, present in CTX-M3, encodes an essential molecule for activity-dependent bulk endocytosis (ADBE)[45], and several module proteins overlap with the ADBE proteome[46]. This suggests a parsimonious explanation that this module reflects the maintenance of organelles and proteins required for long-term neuronal activity, (that is, mitostasis and ADBE proteostasis), through activity-dependent mRNA transcription and neuropil targeting[47].

Incorporating gene networks into models of genetic architecture remains a major challenge. The omnigenic hypothesis does not specify a concrete network model, but to practically use the model to understand the etiology of disease, it would seem useful to connect it to quantifiable relationships between genes. Our approach comes from a unifying hypothesis: that there is a relationship between mutational effect size and network distance, with omnigenic and polygenic architectures representing the strong and weak extremes of that relationship. For the three distinct network types tested, we find that the network structures do not strongly distinguish peripheral genes from core genes as predicted by an omnigenic model. However, there are many other natural network topologies to test, and it will be important to further explore cellular-level or other types of gene networks. The model underlying our analysis is broadly applicable as it provides a means to relate total effect—direct and indirect—to network structure. Future work extending this model provides a means of assessing the proportion of heritability explained by network interactions to characterize the network architecture of disease.

Finally, we acknowledge limitations, including the need for extending these studies to include single-cell, multiregional and multi-omics data. Based on our observation that methods for removing unwanted variance also remove coexpression signal, development of new methods for hidden artifact correction would be useful. Another limitation of network algorithms is that they can produce qualitatively different modules. Although we compared four distinct approaches to show that our findings are stable across methods, we strongly support additional work to benchmark network methods and to develop new methods. Capturing the broadest scope of functionally relevant coexpression relationships will definitely require application of a diverse set of network approaches.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00887-5.

Received: 30 April 2020; Accepted: 7 June 2021;
Published online: 22 July 2021

References

1. Smoller, J. W. et al. Psychiatric genetics and the structure of psychopathology. *Mol. Psychiatry* **24**, 409–420 (2018).
2. Félix, M.-A. & Barkoulas, M. Pervasive robustness in biological systems. *Nat. Rev. Genet.* **16**, 483–496 (2015).
3. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. *Science* **349**, 1489–1494 (2015).
4. Gandal, M. J. et al. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693–697 (2018).
5. Horn, H. et al. NetSig: network-based discovery from cancer genomes. *Nat. Methods* **15**, 61–66 (2017).
6. Mostafavi, S. et al. Parsing the interferon transcriptional network and its disease associations. *Cell* **164**, 564–578 (2016).
7. Oldham, M. C. et al. Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).
8. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
9. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
10. Kelley, K. W., Nakao-Inoue, H., Molofsky, A. V. & Oldham, M. C. Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**, 1171–1184 (2018).
11. McKenzie, A. T. et al. Brain cell-type-specific gene expression and coexpression network architectures. *Sci. Rep.* **8**, 8868 (2018).
12. Wang, H.-Y. et al. Rate of evolution in brain-expressed genes in humans and other primates. *PLoS Biol.* **5**, e13 (2006).
13. Shohat, S., Ben-David, E. & Shifman, S. Varying intolerance of gene pathways to mutational classes explain genetic convergence across neuropsychiatric disorders. *Cell Rep.* **18**, 2217–2227 (2017).
14. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
15. Parikshak, N. N. et al. Genome-wide changes in lncRNA, splicing and regional gene expression patterns in autism. *Nature* **540**, 423–427 (2016).
16. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia and bipolar disorder. *Science* **362**, eaat8127 (2018).
17. Bennett, V. & Lorenzo, D. N. An adaptable spectrin/ankyrin-based mechanism for long-range organization of plasma membranes in vertebrate tissues. *Curr. Top. Membr.* **77**, 143–184 (2016).
18. Dörrbaum, A. R., Kochen, L., Langer, J. D. & Schuman, E. M. Local and global influences on protein turnover in neurons and glia. *eLife* **7**, e34202 (2018).
19. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
20. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
21. Battista, D., Ferrari, C. C., Gage, F. H. & Pitossi, F. J. Neurogenic niche modulation by activated microglia: transforming growth factor-β increases neurogenesis in the adult dentate gyrus. *Eur. J. Neurosci.* **23**, 83–93 (2006).
22. Phoenix, T. N. & Temple, S. Spred1, a negative regulator of Ras-MAPK-ERK, is enriched in CNS germinal zones, dampens NSC proliferation, and maintains ventricular zone structure. *Genes Dev.* **24**, 45–56 (2010).
23. Parikshak, N. et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
24. Selimbeyoglu, A. et al. Modulation of prefrontal cortex excitation/inhibition balance rescues social behavior inCNTNAP2-deficient mice. *Sci. Transl. Med.* **9**, eaah6733 (2017).
25. Nelson, S. B. & Valakh, V. Excitatory/inhibitory balance and circuit homeostasis in autism spectrum disorders. *Neuron* **87**, 684–698 (2015).

26. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
27. Boulting, G. L. et al. Activity-dependent regulome of human GABAergic neurons reveals new patterns of gene regulation and neurological disease heritability. *Nat. Neurosci.* **24**, 437–448 (2021).
28. Schanzenbächer, C. T., Langer, J. D. & Schuman, E. M. Time- and polarity-dependent proteomic changes associated with homeostatic scaling at central synapses. *eLife* **7**, e33322 (2018).
29. Marbach, D. et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* **13**, 366–370 (2016).
30. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
31. Miller, J. A., Oldham, M. C. & Geschwind, D. H. A systems-level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J. Neurosci.* **28**, 1410–1420 (2008).
32. Voineagu, I., et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
33. Fromer, M., et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
34. Wang, Q. et al. The landscape of multiscale transcriptomic networks and key regulators in Parkinson's disease. *Nat Commun.* **10**, 5234 (2019).
35. Miller, J. A., Horvath, S. & Geschwind, D. H. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl Acad. Sci. USA* **107**, 12698–12703 (2010).
36. Howrigan, D. P. et al. Exome sequencing in schizophrenia-affected parent–offspring trios reveals risk conferred by protein-coding de novo mutations. *Nat. Neurosci.* **23**, 185–193 (2020).
37. Pers, T. H. et al. Comprehensive analysis of schizophrenia-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Hum. Mol. Genet.* **25**, 1247–1254 (2016).
38. Sanders, S. J. et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
39. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833 (2018).
40. Ruzzo, E. K. et al. Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850–866 (2019).
41. Wang, Q. et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* **22**, 691–699 (2019).
42. Walker, R. L. et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* **179**, 750–771 (2019).
43. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
44. Polioudakis, D. et al. A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801 (2019).
45. Nicholson-Fish, J. C., Kokotos, A. C., Gillingwater, T. H., Smillie, K. J. & Cousin, M. A. VAMP4 is an essential cargo molecule for activity-dependent bulk endocytosis. *Neuron* **88**, 973–984 (2015).
46. Kokotos, A. C., Peltier, J., Davenport, E. C., Trost, M. & Cousin, M. A. Activity-dependent bulk endocytosis proteome reveals a key presynaptic role for the monomeric GTPase Rab11. *Proc. Natl Acad. Sci. USA* **115**, E10177–E10186 (2018).
47. Doll, C. A. & Broadie, K. Impaired activity-dependent neural circuit assembly and refinement in autism spectrum disorder genetic models. *Front. Cell. Neurosci.* **8**, 30 (2014).

Methods

Ethical statement. As a retroactive reanalysis of data obtained from dbGAP, this study did not require direct formal consent.

Statistics and reproducibility. This study was designed as a retrospective study of bulk RNA expression from human brain. As such, no power analysis was performed to determine sample size, nor was any blinding or randomization applied. All brain regions were profiled with a sufficient number of samples to perform standard WGCNA analysis (>30), and some individuals were excluded on the basis of potentially confounding biological or technical factors (see below).

Expression quantification, quality control and covariate correction. Reads were aligned using STAR[48] in standard two-pass fashion. GENCODE v25 transcripts (hg19/b37) were used as the reference transcriptome and genome for alignment. Transcripts were quantified using the RSEM command line tool to produce gene-level and isoform-level TPM values. The analyzed TPM values were log transformed $\log(0.005+x)$, resulting in approximate normality. A gene was included in the network analysis if it met the following criteria across brain regions:

- The gene must be non-missing in all regions
- The median read count must be >12 in at least one region
- In all regions, 80% of samples must have at least one count
- In all regions, the variance of gene expression must be >0

These thresholds resulted in 15,895 genes, of which 929 had a mean TPM < 0.5 and thus an additional TPM threshold was not applied.

We examined known coexpression artifacts for potential contamination with non-brain tissues[49] and found that pancreas-specific genes *PRSS1* (ENSG00000204983), *PNLIP* (ENSG00000175535), *CLPS* (ENSG00000137392) and/or *CELA3A* (ENSG00000142789) did not pass our coverage thresholds. Further, the genes *KRT4* (ENSG00000170477) and *GP2* (ENSG00000169347), listed as 'inappropriate' for brain, also were not sufficiently covered. We therefore regarded these specific instances of cross-tissue contamination as nonexistent in our analyses. While there may be a more general and subtle effect of contamination from non-pancreatic tissues, it does not appear strong enough to generate sample outliers or modules related to non-brain cell types.

Sample-specific and individual-specific covariates were downloaded from the GTEx website, and supplemented with technical alignment information from the STAR alignment and PicardTools quality-control metrics of the resulting BAM files.

Individuals were excluded if they were positive for any of the following phenotypes: 'MHALS', 'MHALZDMT', 'MHDMNTIA', 'MHENCEPHA', 'MHFLU', 'MHJAKOB', 'MHMS', 'MHPRKNSN', 'MHREYES', 'MHSCHZ', 'MHSEPSIS', 'MHDPRSSN', 'MHLUPUS', 'MHCVD', 'MHHIVCT', 'MHCANCERC', 'MHPNMIAB', 'MHPNMNIA', 'MHABNWBC', 'MHFVRU', 'MHPSBLDCLT' or 'MHOPPINF'. The individual-specific covariates 'GENDER', 'AGE', 'RACE', 'ETHNICITY', 'TRISCH', 'TRISCHD', 'DTHCODD', 'SMRIN', 'SMNABTCH', 'SMGEBTCH', 'SMTSISCH' and 'SMTSPAX' were extracted. The 'DTHCODD' variable was binned into the following categories: 'UNKNOWN', '0to2h', '2hto10h', '10hto3d', '3dto3w' and '3wplus'. Individuals were also excluded if they appeared as outliers for expression principal components (that is, if the sum of the square of the scaled principal components were less than 6). The final sample counts after this process were: NAC (85), AMY (52), CBH (78), cerebellar cortex (91), CDT (85), PFC (85), cortex BA24 (60), cortex BA9 (76), HIP (71), HYP (67), PUT (74) and SNA (43). While small, all counts are above the recommended size (30) for WGCNA, and in many regions of the brain, the GTEx data reflect the largest available sample sizes for sequenced data.

STAR alignment metrics and PicardTools quality-control metrics were subset to non-excluded samples, and outliers were flagged and removed via a chi-squared test ($P < 10^{-5}$). The PicardTools metrics were log scaled, and the top five principal components extracted using the principal-component analysis class from scikit-learn[50] ('seq-PC'). The STAR alignment covariates were subset to those with 'splice' in the feature name, and the top three principal components were similarly extracted ('STAR-PC').

Given the gene expression and covariate matrices, features that explained a significant proportion of expression variance in a nontrivial subset of genes were extracted using a forward–backward regression approach (see below). This approach identified 'seq_pc1', 'seq_pc2', 'seq_pc3', 'SMRIN', 'SMEXNCRT', 'number_of_splices_GT/AG', 'TRISCHD' and 'DTHCODD' (categorical encoding) as significant features, with no significant interactions between these features or between any of these covariates and tissue type.

While there were no significant cross-terms between tissue and covariate, in our comparisons to latent-factor-based approaches (see below), we found that hidden factors were not (nearly) orthogonal rotations of one another, leading us to run all correction methods within each tissue. Following this approach, we used a linear model (expression ~ covariates − 1) to remove covariate effects from within each region. Because we corrected for covariates within each brain region, we did not model individual-specific (cross-regional) effects.

Tissue hierarchy. The median expression of all genes across a given tissue was taken as the exemplar of said tissue. These exemplars (12 in all) were hierarchically clustered into the tissue hierarchy observed in Fig. 1 using Euclidean distance and single-linkage hierarchical clustering.

Module construction. *Robust WGCNA.* rWGCNA[51] was applied to each brain tissue independently. Briefly, the power parameter is selected as the smallest power (between 6 and 20) that achieves a truncated r^2 of >0.8 and a negative slope. Then, 50 signed coexpression networks are generated on 50 independent bootstraps of the samples; each coexpression network uses the same estimated power parameter. These 50 topological overlap matrices are then combined edge-wise by taking the median of each edge across all bootstraps.

The topological overlap matrices are then clustered hierarchically using average linkage hierarchical clustering (using '1 − TOM' (topological overlap measure) as a dis-similarity measure). The bootstraps are used to determine cut height as follows: multiple cut heights are considered (0.9 to 0.999, by 0.005), and for each cut, the within-module correlation of TOMs is considered. For the top eight modules by size (fewer if fewer modules are produced), the consensus and each bootstrap TOM is subset to the genes within each module, and the correlation between bootstrap and consensus is computed. The median (within module, across bootstraps) of these consensuses is computed, and the mean of these summaries is taken to be a measure of 'goodness' for the cut. The cut height that maximizes this metric is taken to define the initial modules.

These initial modules are then merged via 'mergeCloseModules' in WGCNA, which hierarchically reclusters modules based on the module eigengenes, using the correlation-based adjacency as a dis-similarity matrix. Modules with a distance of <0.35 are merged together into a combined module.

Aggregating coexpression. At each merge of the hierarchy, a single round of consensus topological overlap is performed. Each pair of genes has two descendent edges, and the parent edge is estimated as the 80th percentile between the two (that is, for $x < y$; $P = 0.2x + 0.8y$)). This process proceeds up the tissue hierarchy until a single-network TOM remains.

Consensus labeling. After construction of coexpression networks from all tissues and splits, modules were defined for a total of 21 groups (including BRNACC-BRNSNA, BROD, CTX, CBL, BGA, STR, NS-SCTX, SCTX, NCBL and WHOLE-BRAIN), yielding over 300 overlapping modules. The overlapping nature of these modules motivated labeling each module in terms of a hierarchy group, allowing one to identify, for example, BRNHYP-M2 and BRNCTX-M7 with the module group WHOLE-BRAIN-M3.

To perform this labeling, similarity matrices are computed. First, the module eigengenes for all modules (regardless of origin) are computed within every tissue, and the correlation matrix (using 'bicor') is computed for each module for each tissue. This produces an (all modules) × (all modules) matrix for each tissue. The consensus eigengene similarity ('E') between two modules is chosen as the component-wise maximum of all of these matrices. The second similarity matrix is the standard Jaccard similarity ('J') between module gene lists. These similarities are combined into a dis-similarity matrix $D = 1 - (E + 3 \times J)/4$, which is used to hierarchically cluster (average linkage) these modules.

Module groups are defined by cutting the dendrogram at a height of 0.35. This process results in a set of module clusters, each of which has a 'level' in the brain tissue hierarchy (for instance, a cluster of BRNCTXBA9-M4, BRNCTXB24-M2 and CTX-M7 would have the level 'CTX' as the top level of the tree represented is CTX). The 'representative' of the module group is taken to be the module at the highest (most rootward) level of the tree, and if there are two, the larger of the two is chosen. A second round of clustering is performed by removing all modules in the group (except for its representative) from the dis-similarity matrix, and reclustering only the group representatives. This process repeats until there are no additional merges. Finally, each module is labeled with its group representative; for instance, 'BRNCTXBA9-M4' would receive the label 'CTX-M7', because it shares its highest similarity with the consensus cortex module M7.

In addition, we renamed and abbreviated the modules: 'BW' for brain-wide, 'NCBL' for non-cerebellar, 'NS.SCTX' for non-striatal subcortex and 'CEREB' for cerebellum. The GTEx tissue names were abbreviated to clear region codes: ACC, AMY, B24, BA9, CBH, CBL, CDT, HIP, HYP, PFC, PUT and SNA.

Preservation. We considered two module preservation statistics: the classical Z-summary[52] and a leave-one-gene-out neighbor statistic. For the classical Z-summary, module statistics such as the mean gene–gene correlation in the module, the correlation-of-correlations across datasets, the variance explained by the first module principal component, and other metrics were computed for each module (in both the original and comparison dataset) and compared to 100 random (via permutation) modules of identical size. Each observed statistic was converted to a z-score, and these were averaged to generate a final summary, for which large z-scores are indicative of replication of the underlying biological signal.

The neighbor statistic ('Z-AUPR', z-score of the area under the precision-recall curve) is strongly influenced by the single-cell statistic MetaNeighbor[53]. Briefly, a k-nearest-neighbor network is built in the comparison dataset (we used $k = 15$),

and we imposed the module labels from the reference dataset. For each gene, we computed the proportion of its neighbors (again, in the comparison dataset) whose labels matched its own. Note that if this proportion is >0.5, then this gene would be assigned the same label in the comparison dataset as the reference dataset under a neighbor-voting scheme. Using these scores, we computed an AUPR for each module. We repeated this approach for 100 permuted modules (and, unlike the WGCNA permutation, we split genes into connectivity deciles, and permuted only within each decile), and used this baseline to convert observed AUPRs to z-scores. As with the classical Z-summary, high Z-AUPR values are indicative of replication of underlying biological signal.

Learning curves. To examine how module identification and specificity changes as a function of the number of samples, we combined samples from similar tissues to increase the maximum N value as follows: we combined the cerebellar samples into one larger group ($N=122$), and we also grouped the cortical samples (PFC, B24 and BA9) together with hippocampal samples into a second group ($N=304$).

'Reference' modules for these groups were determined by applying rWGCNA to the full dataset. We downsampled the group to a smaller set of samples of size $n=25, 50, \ldots, N$, and performed rWGCNA on the smaller set. We repeated this process ten times, generating ten networks and module assignments for each subsampling of the full dataset.

Because two clusterings should be considered identical up to renaming the labels in one or the other dataset, we use module co-clustering as a measure for accuracy, precision and recall. Within the reference (whole group) dataset, we extracted the top 'hub' gene from each of the modules, and the list of genes co-clustered with that hub gene (that is, the other members of its module). For a given reference module, within a subsampled dataset, we have:

Recall = (no. of ref. hub co-clustered genes also co-clustered in subsample)/(no. of ref. hub co-clustered genes)

Precision = (no. of ref. hub co-clustered genes also co-clustered in subsample)/(no. of subsample co-clustered genes)

In effect, these are precision/recall statistics for the hub gene co-clustering indicators. If two reference modules fail to separate in a subsample (a typical failure mode), the result is slightly higher recall, but far worse precision.

Regional contrast test. The RCT is a multivariate test of significance for:

$$H_0 : \beta_i \leq \max\left(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n\right)$$

$$H_a : \beta_i > \max\left(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n\right)$$

This statistic corresponds to a multidimensional integral, with infinite limits on all coefficients other than β_i, and taking the maximum of $(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_n) < \beta_i < \infty$. Because of the large numbers of degrees of freedom in this regression, we treated the variance–covariance matrix ($\Sigma_\beta^{(ML)}$) of the $\boldsymbol{\beta}$ vector as giving the true sampling covariance of these parameters, and performed Monte-Carlo integration by drawing 50,000,000 samples from the multivariate normal distribution $N(\boldsymbol{\beta}, \Sigma_\beta^{(ML)})$ using the R package 'fastmvn'.

The above statistic works for testing each tissue against all others. A grouped version of the test is a simple extension, which considers several β values in tandem. For simplicity, we assumed the indexes for the group were the first k coefficients, then the comparison becomes:

$$H_0 : \min(\beta_{1,\ldots,}\beta_k) \leq \max\left(\beta_{k+1}, \ldots, \beta_n\right)$$

$$H_a : \min(\beta_{1,\ldots,}\beta_k) > \max\left(\beta_{k+1}, \ldots, \beta_n\right)$$

This only changes the integration limits to (for $j \leq k$) to the maximum of $(\beta_{k+1}, \ldots, \beta_n) < \beta_j < \infty$, and we used the same Monte-Carlo approach as before.

Western blot isoform analysis. Human induced pluripotent stem cells were differentiated into cortical glutamatergic-pattern neurons, according to work by Nehme et al.[54], and samples were extracted at days 0, 16, 21 and 31. Human astrocytes were used as an outgroup. Immunoprecipitation was performed using ANK2-specific monoclonal antibody S105-17.

De novo variant enrichment. Denovo-DB[55] was used to extract lists of genes harboring de novo variation linked to ASD and SCZ. Version 1.5 of the database was obtained on 17 February 2018, and we filtered for 'PrimaryPhenotype = autism' (or, separately, 'PrimaryPhenotype = schizophrenia') and 'FunctionClass' as one of 'frameshift', 'frameshift-near-splice', 'splice-acceptor', 'splice-donor', 'start-lost', 'stop-gained', 'stop-gained-near-splice' or 'stop-lost.'

Module enrichments were calculated with Fisher's exact test, using the contingency table formed by cross-tabulating module presence/absence with presence/absence on the denovo-db gene list.

As the denovo-db is a broad collection of de novo mutations in affected individuals and does not curate these variant lists on the basis of total evidence, we considered two additional data sources for alternative enrichment scores. First, we considered the curated list of SFARI genes of rank S, 1, 2 or 3, and performed enrichment analysis on the resulting list. Second, recent work from our laboratory[56] computed transmission and de novo association Bayes factors for 18,472 genes. We regressed the log Bayes factor against module presence/absence and looked for a significant, positive coefficient.

GWAS variant enrichment. Enrichment for GWAS signal was performed through the use of MAGMA[57] gene-set analysis. Briefly, variants were mapped to genes on the basis of genomic distance while taking chromatin contact maps from adult brain Hi-C analyses[58] into account. MAGMA was used to generate gene scores and linkage disequilibrium-based covariances. Subsequently, MAGMA's gene-set analysis was used to compare the distribution of gene scores between modules and the background set of 'gray' genes.

Eight GWAS studies were considered in this analysis: the iPsych and Psychiatric Genomics Consortium cross-disorder GWAS studies (accounting for ASD, SCZ and cross-disorders), Alzheimer's disease, multiple sclerosis and educational attainment[59,60–63].

Defining genes likely to harbor high-impact rare variants. We identified sets of genes likely to harbor high-impact rare variation for both ASD and SCZ by using the top implicated genes from each of three previous rare and de novo studies of neuropsychiatric disease: extTADA[64], iHART and NPDenovo[65]. These studies produced Bayes factors for confidence of association for a particular gene. We used the Bayes Factors as a ranking ('Hub genes and empirical core genes').

Core–periphery enrichment within networks. *Simulation of network genetic architecture.* In total, 10,000 causal variants were simulated with frequency parameters estimated from human populations[66], and distances were drawn from a binned beta distribution:

$$p_i \sim \text{beta}\,(0.14, 0.7)$$

$$d_i \sim \frac{\lfloor k_d \text{beta}(a_d, b_d) \rfloor}{k_d}$$

$$\beta_i | d_i, p_i \sim N\left(0, \sigma_g^2 \left(2p_i\left(1-p_i\right)\right)^{\gamma_1}\left(1+\delta d_i\right)^{\gamma_2}\right)$$

σ_g^2 is arbitrary and set to 1; k_d is arbitrary so long as it is greater than about 5, and is set to $k_d = 12$; a_d, b_d, γ_1, γ_2 and δ are model parameters. Recent results from the UK Biobank suggest that a value of $\gamma_1 = -0.4$ is reasonable for a polygenic trait (height = −0.45; education = −0.32; blood pressure = −0.39) and is fixed to this value. Architectures were simulated on a grid of a_d, $b_d = 1, 1.5, \ldots 6$; $\delta = 1, 1.2, \ldots, 2.6$; $\gamma_2 = -15, -10, -7, -5$ and -2. Notably, for any values of a_d, b_d, δ a γ_2 can be found such that the network term explains >40% of the heritability. For errors-in-distance, here the above simulation of distance is replaced by a normal copula (where 20% error corresponds to $r = 0.8$—this is a purposeful underestimate, as $r^2 = 0.64$, so the latent error is ~36%):

$$Z \sim N\left(0, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right)$$

$$d_{true} = \frac{k_d \Phi^{-1}_{\text{beta}(a_d, b_d)}\left(\Phi_{N(0,1)}\left(Z_1\right)\right)}{k_d}$$

$$d_{meas} = \frac{k_d \Phi^{-1}_{\text{beta}(a_d, b_d)}\left(\Phi_{N(0,1)}\left(Z_2\right)\right)}{k_d}$$

When simulated from a network, first a set of $K = 1, \ldots, 10$ hub genes were simulated with the constraint that no pair can be directly connected by an edge. These formed initial communities of size 1. For the remaining 40 core genes, a community was selected at random, a community member was selected at random and a neighbor was selected at random and added to the community and to the set of core genes. These formed the basis of d_{true}, the minimal path distance to any core gene. For d_{meas}, the communities were distorted by removing $M = 1, \ldots, 10$ core genes at random, or by adding $K = 5, 10, \ldots, 25$ non-core genes at random.

For normalized effect sizes, the identification of effect size of an empowered 5% frequency GWAS variant happened through three steps: (i) estimating the liability distribution, (ii) mapping case–control frequency differences to effect sizes and (iii) estimating power.

i. Liability distribution: a 5,000 × 10,000 genotype matrix X was sampled independently, with frequencies given by the previously simulated vector f, and 5,000 genetic liabilities were generated by $l_g = X\boldsymbol{\beta}$. These liabilities were used to estimate parameters for a T-distribution using 'fitdist' from the R package 'MASS'; the degrees of freedom were reduced by 25% to account partially for

rare variants not sampled in this population of 5,000, and these parameters were used to generate 400,000 genetic liability scores. These were converted to total liability scores by adding noise $l = l_g + N(0, \sigma_e)$, with σ_e chosen so that the heritability is 0.85.

ii. Frequency-ratio-to-effect: the goal is to estimate the ratio p_{aff}/p_{unaff} for a variant with a frequency p_i and effect β_i. The genetic liabilities $l_{new} = l + x\beta_i$ with $x \sim$ binomial $(2, p_i)$ were computed for 400,000 simulated individuals. As 10,000 variants contributed to l, the addition of $x\beta_i$ was assumed to have a minimal effect on heritability. Case–control labels were defined by $l_{new} \geq$ quantile(l_{new}, 0.95) so that the disease prevalence was 5%, and the empirical frequency mean(x_{aff})/mean(x_{unaff}) was taken as an estimate of the ratio p_{aff}/p_{unaff}. Fixing $p_i = 0.05$ and varying β_i produced an empirical and invertible map from variant effect-to-frequency ratio.

iii. Estimating power: given an effect size β_i, the case and control frequencies for a $p = 0.05$ variant were obtained from step (ii). A total of 5,000 case and 5,000 control genotypes were sampled according to the corresponding frequencies, and a two-sided t-test was performed by 't.test' in R. Then, 1,000 simulations were performed, and the number of times the t-test P value achieved a Bonferroni-corrected P value of 0.1/10,000 (the number of causal variants) was tabulated.

Network construction and computation of d(G). *Coexpression networks.* Within coexpression networks, the raw coexpression (cosine) distance is used to define gene–gene distances. In addition, a sparse $\varepsilon = 2.5\% + 1 - NN$ graph was calculated as follows: the cosine distance graph was subset to only the 2.5% smallest edges, and any singleton genes were connected to their closest neighbor. This graph was treated as unweighted, and not necessarily connected. Cross-component distances were treated as 1 + the maximum observed within-component distance. This is referred to as 'sparse distance.'

Module hub genes were defined as the 2.5% of module genes with largest k-within values (minimum of 5). Distances between a gene and a module were computed as (1) 1 − kME, (2) mean cosine distance to a module hub, (3) minimum cosine distance to a module hub, (4) mean sparse distance to a module hub and (5) minimum sparse distance to a module hub. When using arbitrary gene sets as core genes, steps 2–4 were computed with respect to the gene set in place of module hubs. For other network types, see the Supplementary Note.

Hub genes and empirical core genes. Core gene sets that defined distance ('proposal set') were taken to be either collections of network hub genes or the top 10 or 20 genes (by Bayes factor) from each of the three studies (separately). The core gene sets that defined the statistic Φ ('evaluation set') were taken to be the top 25, 35, 50, 75 or 100 genes from each of the three studies. To restrict attention to directly causal (for example, nonregulatory) genes, as the omnigenic model suggests, the core genes were also filtered to remove known transcription factors[67], DNA-binding proteins, RNA-binding proteins and noncoding RNA. Without this filtering, values of Φ still fell below 50% for brain coexpression networks, but achieved 70% for blood coexpression. Φ statistics were calculated only for evaluation sets where, after excluding those genes also in the proposal set, noncoding genes, DNA-binding proteins, known transcription factors and RNA-binding proteins, at least 15 genes remained.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Processed data are available at http://geschwindlab.org/gclabapps/hubgene/home/.

Code availability

Supporting code for network construction and network genetic analysis is available at https://github.com/dhglab/multiregional-networks/.

References

48. Dobin, A. et al. STAR: ultrafast and universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
49. Nieuwenhuis, T. O. et al. Consistent RNA-sequencing contamination in GTEx and other datasets. *Nat. Comm.* **11**, 1933 (2020).
50. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *JMLR* **12**, 2825–2830 (2011).
51. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
52. Langfelder, P., Luo, R., Oldham, M. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7**, e1001057 (2011).
53. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single-cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
54. Nehme, R. et al. Combining NGN2 programming with developmental patterning generates human excitatory neurons with NMDAR-mediated synaptic transmission. *Cell Rep.* **23**, 2509–2523 (2018).
55. Tychele N. T. et al. denovo-db: a compendium of human de novo variants *Nucleic Acids Res.* **45**, D804–D811 (2016).
56. Ruzzo, E. K. et al. Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850–866 (2019).
57. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLOS Comp. Biol.* **11**, e1004219 (2015).
58. Won, H., Huang, J., Opland, C. K., Hartl, C. L. & Geschwind, D. H. Human evolved regulatory elements modulate genes involved in cortical expansion and neurodevelopmental disease susceptibility. *Nat. Commun.* **10**, 2396 (2019).
59. Schork, A. J. et al. A genome-wide association study of shared risk across psychiatric disorders implicates gene regulation during fetal neurodevelopment. *Nat. Neurosci.* **22**, 353–361 (2019).
60. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* https://doi.org/10.1016/S0140-6736(12)62129-1 (2013).
61. Lambert, J. C. et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
62. Andlauer, T. F. et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Sci. Adv.* **2**, e1501678 (2016).
63. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
64. Nguyen, H. T. et al. Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med.* https://doi.org/10.1186/s13073-017-0497-y (2017).
65. Du, Y. et al. Nonrandom occurrence of multiple de novo coding variants in a proband indicates the existence of an oligogenic model in autism. *Genet. Med.* **22**, 170–180 (2019).
66. Ionita-Laza, I., Lange, C. & Laird, M. N. Estimating the number of unseen variants in the human genome. *Proc. Natl Acad. Sci. USA* **106**, 5008–5013 (2009).
67. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).

Acknowledgements

D.H.G., C.L.H. and G.R. were supported by the National Institute of Mental Health (NIMH; R01MH110927 and U01MH115746) and the Simons Foundation Autism Research Initiative. G.R. was supported by NIMH award 1F32MH114620. K.L. and G.P. are supported by institutional funding from the Stanley Center for Psychiatric Research at the Broad Institute, the NIMH (R01 MH109903 and U01 MH121499), the Simons Foundation Autism Research Initiative (awards 515064 and 735604), the Lundbeck Foundation (R223-2016-721 and R350-2020-963). A.B., A.S. and P.P. were supported by NIMH award R01MH110927. We thank members of the laboratories of D.H.G., K.L. and A.B. for stimulating discussions. We are also grateful for data made publicly available by the GTEx consortium and S. Deverasetty for creating the HUBgene web browser.

Author contributions

D.H.G., C.L.H., A.B. and K.L. conceived the study design. A.B., A.S. and P.P. reprocessed RNA-seq data. C.L.H., S.M., P.P., W.G.P., G.R. and A.S. built networks and performed analyses. G.P. differentiated cell lines and performed the ANK2 western blot. C.L.H. and D.H.G. prepared the manuscript and figures. A.B. and K.L. contributed to editing.

Competing interests

The authors declare no competing interests.

Additional information

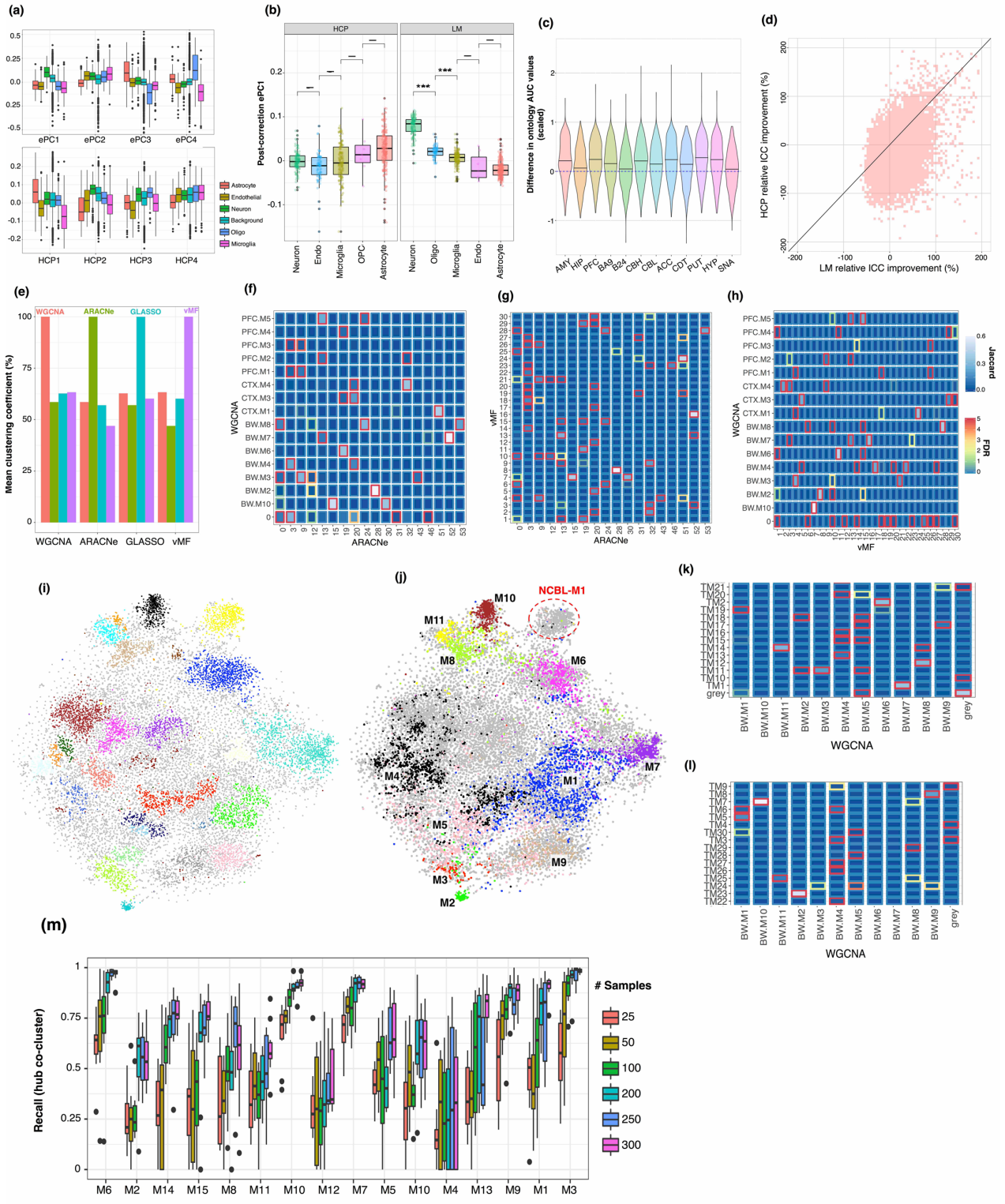
Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00887-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00887-5.

Correspondence and requests for materials should be addressed to D.H.G.

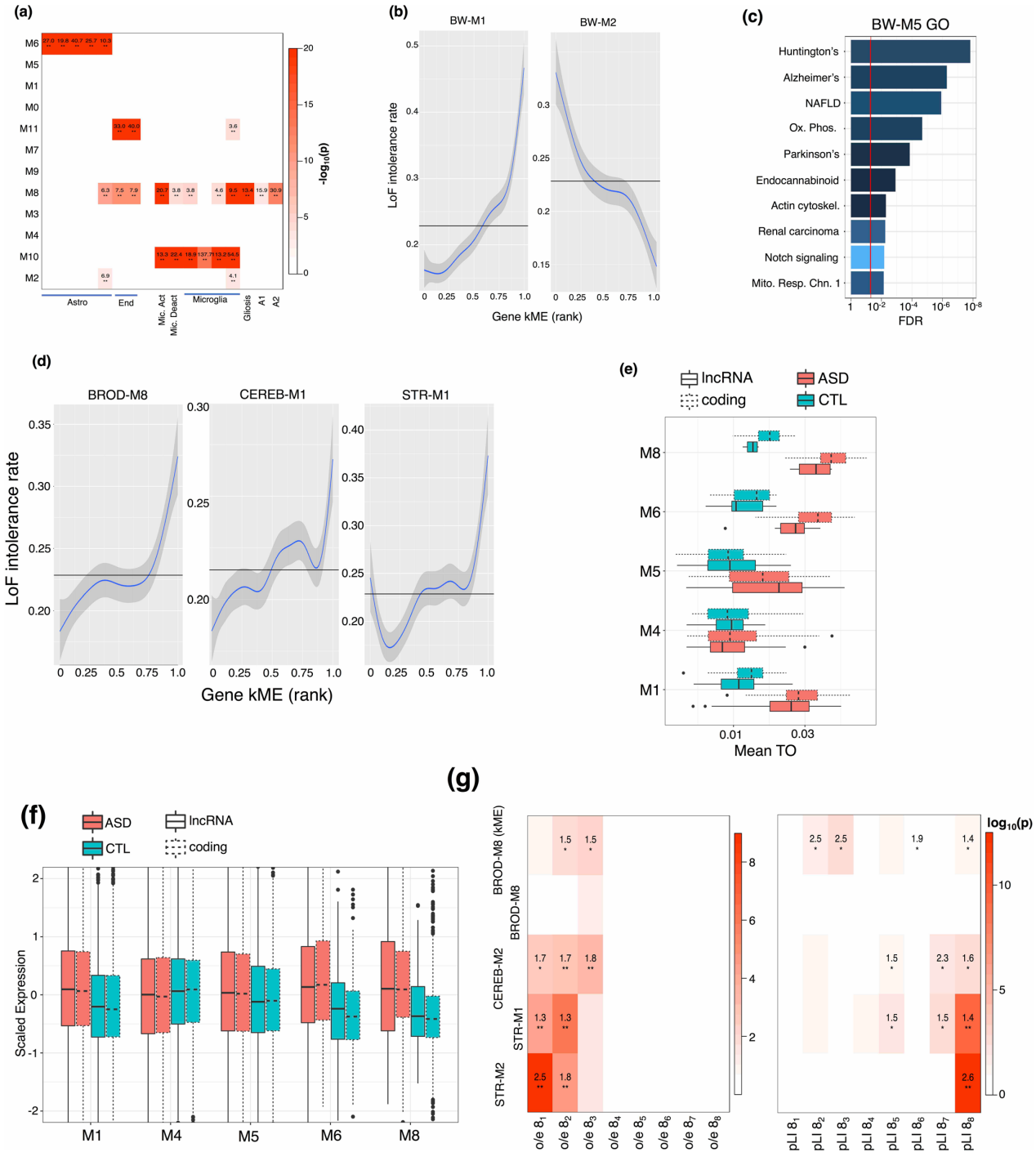
Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

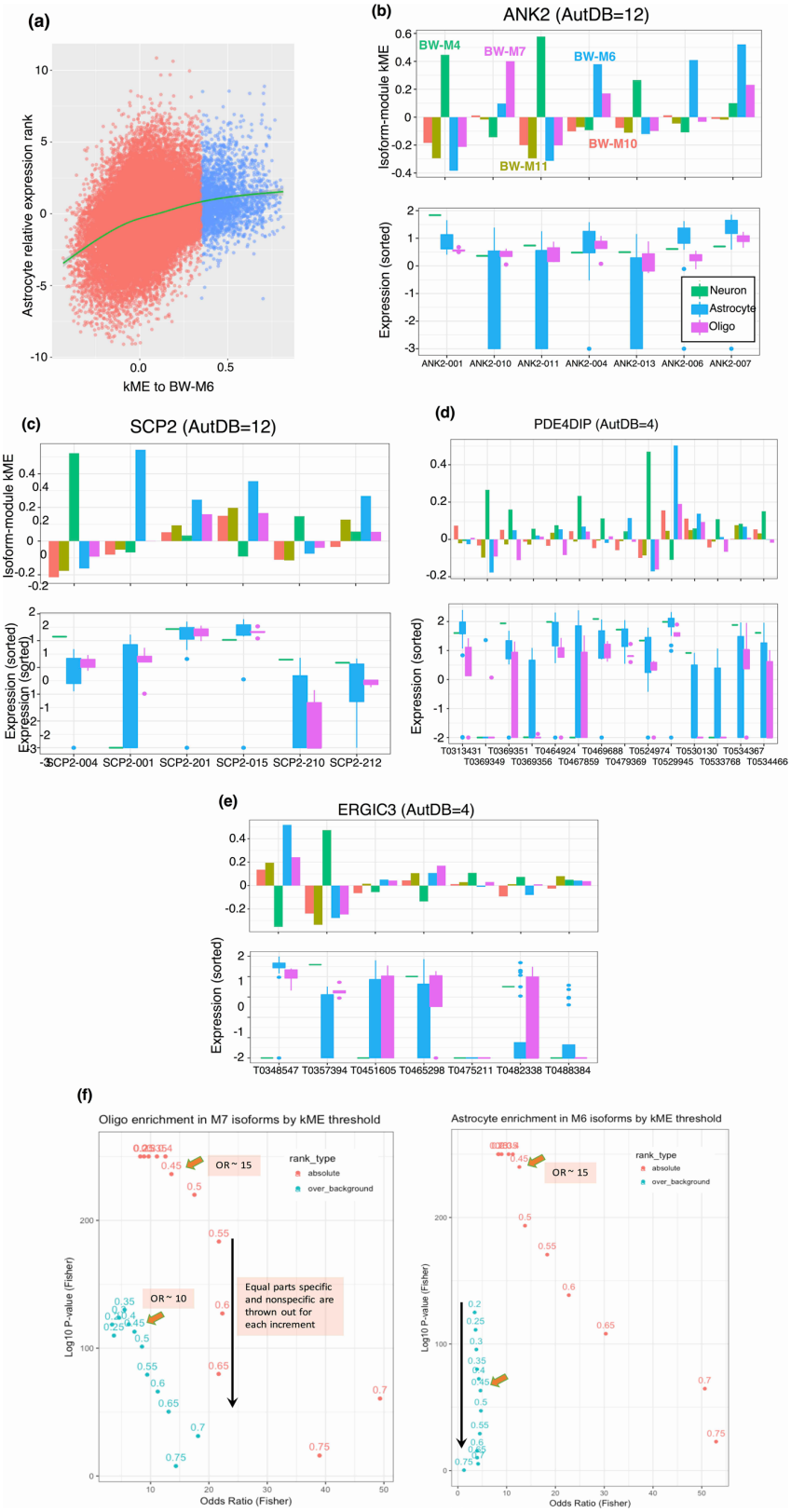


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Brain network QC and algorithm comparisons (related to Fig. 1). a, Standard boxplot (box: quartiles, whiskers: 1.5xIQR) of ePC and HCP loadings onto canonical cell type genes, showing significant heterogeneity of loadings across cell types, N = 114 (Neuron), 79 (Astrocyte), 242 (Microglia), 103 (Oligodendrocyte), 176 (Endothelial). **b**, Standard boxplot (box: quartiles, whiskers: 1.5xIQR) of ePC loadings after covariate correction using HCP and LM base correction, showing that cell type heterogeneity of the 1st component of expression is lost after HCP correction. Gene set sizes as in a; significance (two-sided T-test) ***: < 0.001. **c**, Network-based GO prediction accuracy for each brain region. The same gene holdouts are used in 10-fold cross validation, generating 10 values for the AUC difference of each GO category, which are used to generate a Z-score for the expected AUC difference. **d**, Relative improvement to the integrated correlation coefficient for BRNHYP genes, for linear model and HCP based corrections. **e**, Pairwise co-clustering statistics for the 4 algorithms compared in Fig. 1. X-axis denotes which modules are taken as the reference set. **f-h**, Pairwise module overlaps between 3 of the 4 algorithms compared in Fig. 1 (GLASSO yielded too many modules to visualize here). **i**, t-SNE embedding of gene features from whole-brain tensor decomposition, colored by DBSCAN clusters. **j**, As (**i**), but colored and annotated with whole-brain modules. **k**,**l**, Overlap between whole-brain consensus and tensor-decomposition+DBSCAN modules. Color scheme as in (**f-h**). **m**, Standard boxplot (box: quartiles, whiskers: 1.5xIQR, N = 10 bootstrap re-samplings) of within-module recall values for hub-gene co-clustering, demonstrating that at 100 samples, the recall is above 50% for most modules.

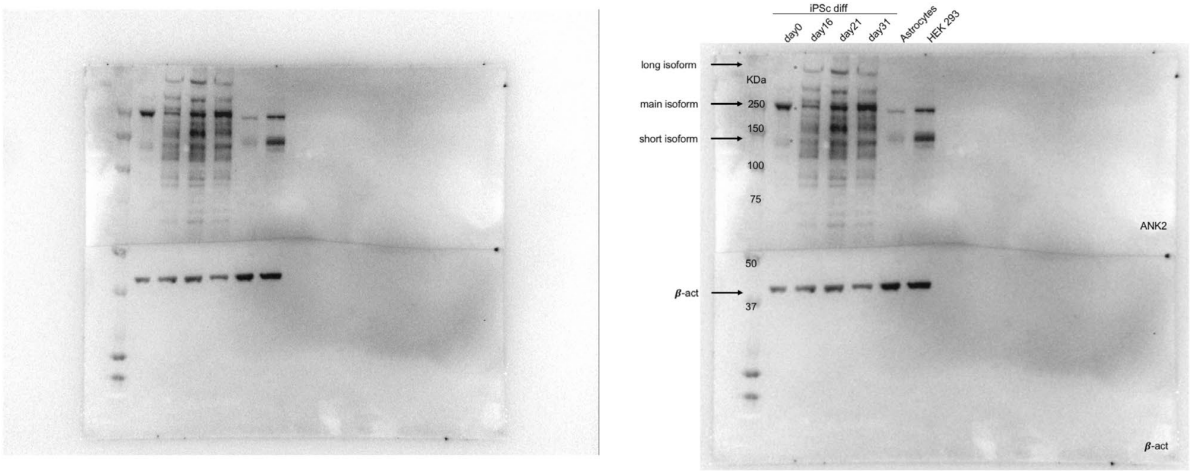


Extended Data Fig. 2 | Exploration of cell-type modules (related to Fig. 2). **a**, Cell-type marker enrichment for brain-wide modules, extended with markers of microglial activation and deactivation, and markers of reactive gliosis and A1/A2 reactive astrocytes. **b**, Plots of the marginal rate (solid: mean, shade: 95% CI of GAM) of LoF-intolerant (pLI>0.9) genes, as a function of BW-M1 (most enriched) and BW-M2 (most depleted) kME. **c**, Gene ontology enrichment for BW-M5. **d**, Marginal LoF-intolerance rates (solid: mean, shade: 95% CI of GAM), by gene kME, for neuronal subtype modules. **e,f**, Standard boxplots (box: quartiles, whiskers: 1.5xIQR) of module mean topological overlap, and gene expression, for 5 whole-brain modules in ASD cases and matched controls (Parikshak 2016). The case/control difference in lncRNA is closely matched by the same difference in randomly-selected, matched coding genes. **g**, LoF-intolerance enrichment for neuronal subtype modules, using pLI and o/e bins as response variables, and a linear model correcting for gene GC and length (logit link, p-values: coefficient T-test). All modules except BROD-M8 show strong enrichment, and BROD-M8 shows enrichment when using soft-membership instead of hard membership.

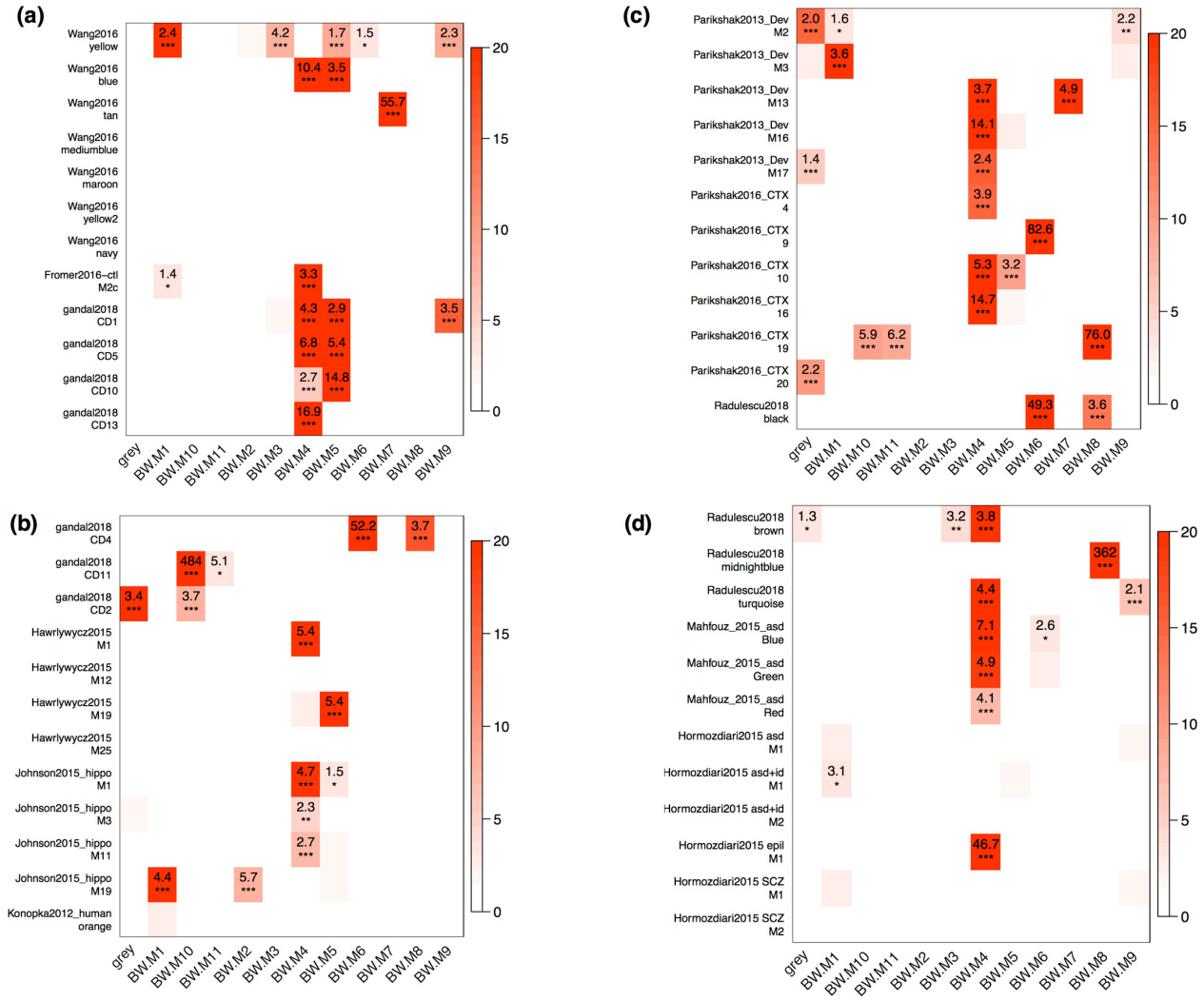


Extended Data Fig. 3 | See next page for caption.

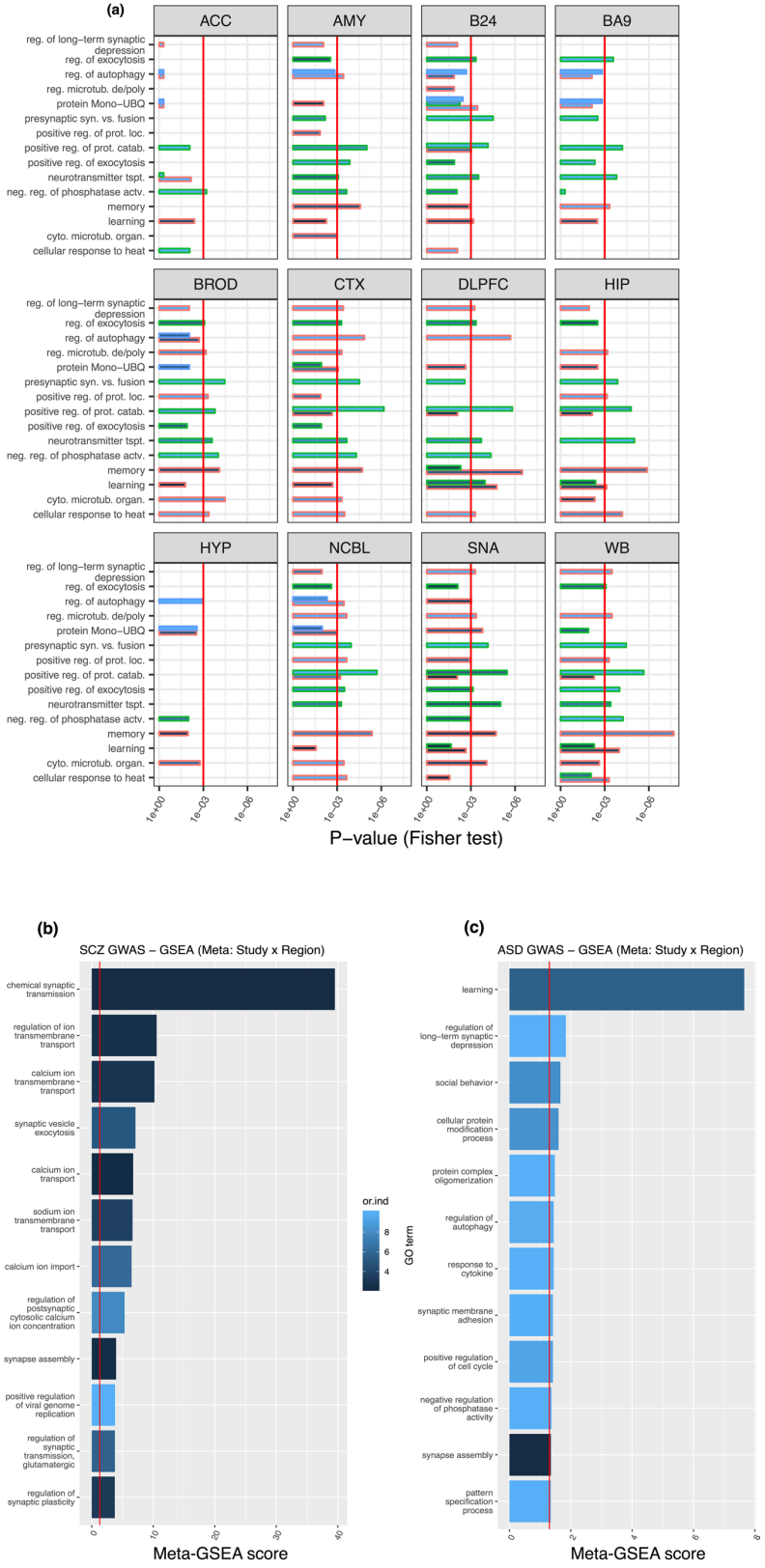
Extended Data Fig. 3 | Glial cell-type isoforms (related to Fig. 3). a, Replicate of main Fig. 3(**b**) in astrocytes, showing a strong positive relationship between astrocyte module membership, and relative expression in astrocyte cells. **b-e**, Relationship between module kME and cell type relative expression for transcripts across 4 neuron/astrocyte isoform switch genes, demonstrating concordance between high kME, and high relative expression. **f**, Unsigned Fisher's exact test of the contingency of 'assigned to module' and 'top-ranked cell type marker' for varying kME thresholds for (left) oligodendrocytes and (right) astrocytes; for marker rankings based on both absolute and relative expression within the cell-sorted data. Thresholds in the range 0.45-0.55 appear to balance significance and odds ratio across absolute and relative rankings.



Extended Data Fig. 4 | Validation of neuron-astrocyte isoform switching (related to Fig. 3). (left) Unmodified Western Blot corresponding to Fig. 3 (right) Same blot, annotated with source of input material and band identities.

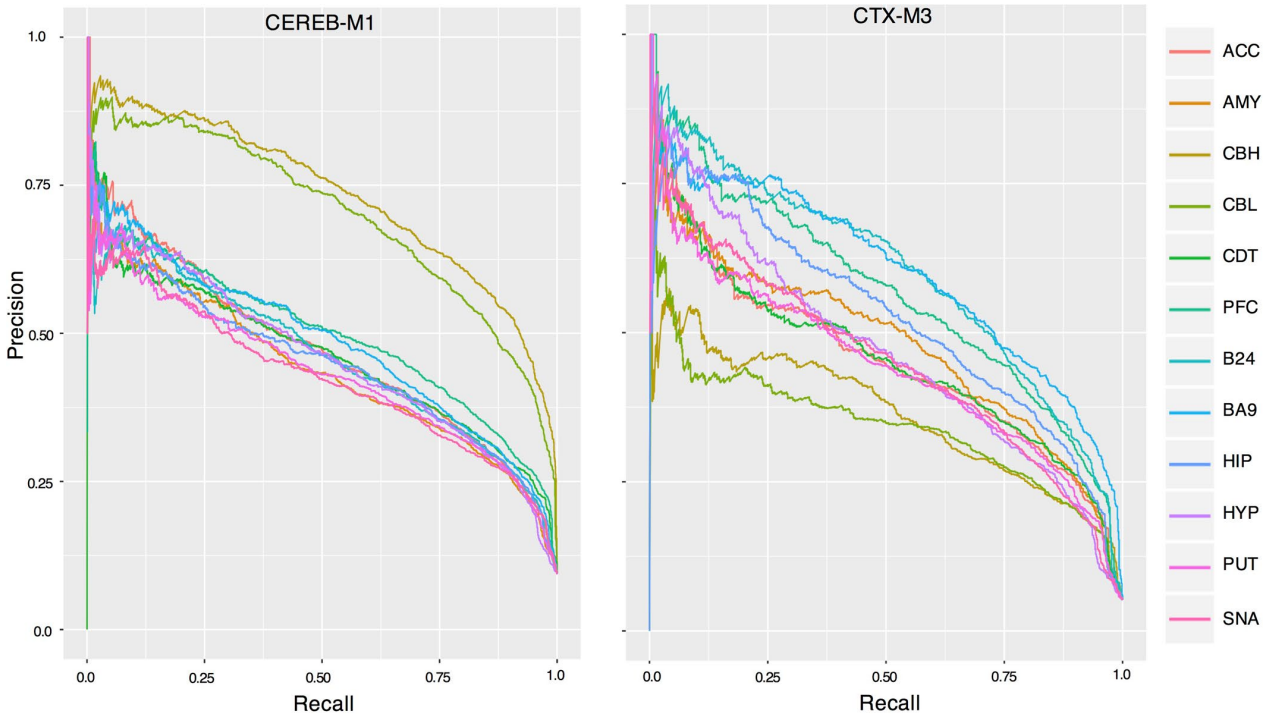


Extended Data Fig. 5 | Overlap with published modules containing disease genes. a–d, Overlaps between published modules and the consensus whole-brain co-expression modules identified in this paper, demonstrating that the majority of modules show a high overlap, particularly to the neuronal module BW-M4. P-values: signed Fisher's exact test. These modules were been selected because of published enrichment for neuropsychiatric disease risk genes. (see Methods).

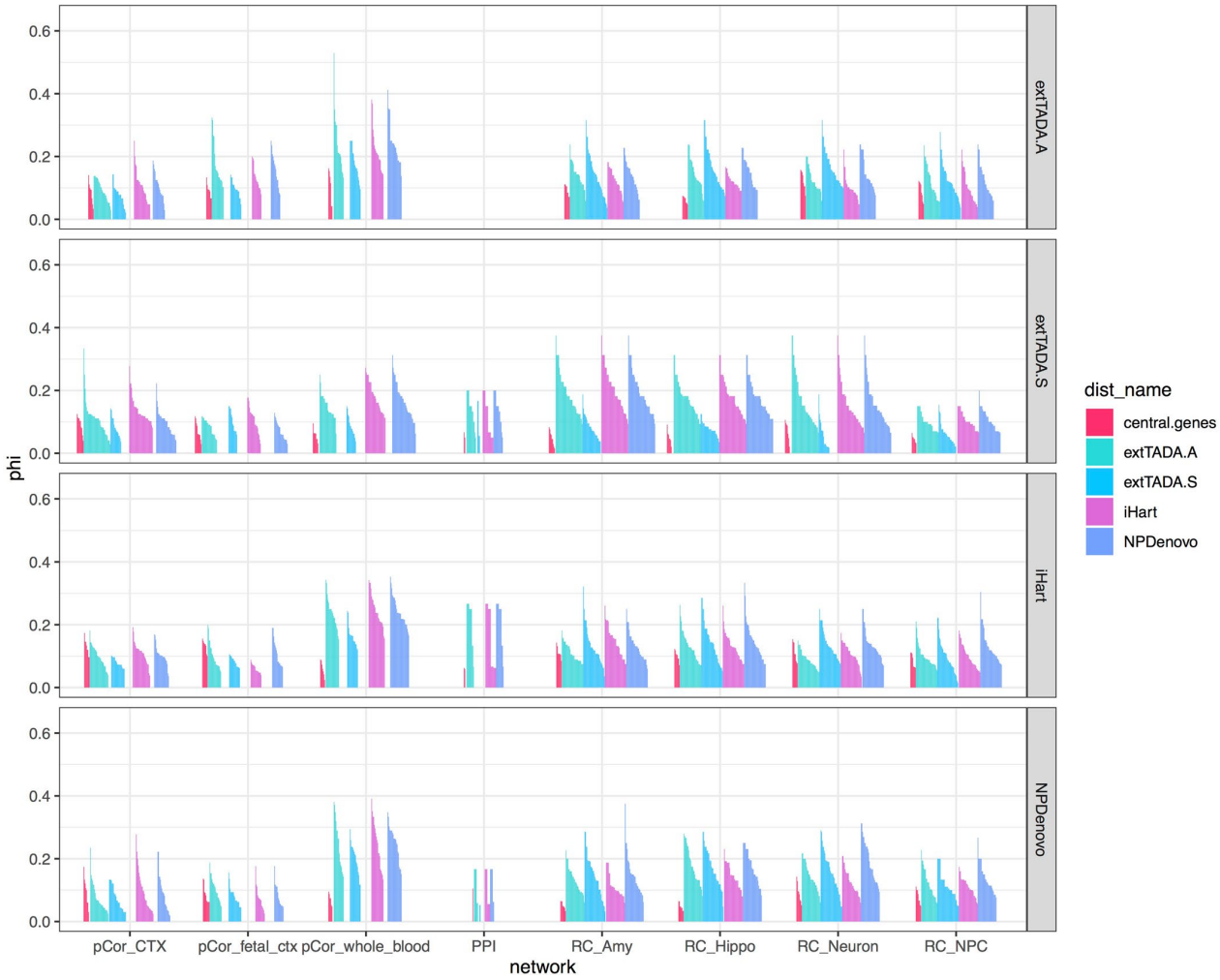


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Module BW-M4 functional annotation (related to Fig. 5). a, Signed gene ontology enrichments (logistic regression controlling for gene length and GC, p-value from coefficient T-test) for MAGMA-significant ASD genes in module set BW-M4 across all regions in which a BW-M4 module is present. Bar color reflects the enrichment odds ratio (black: 5, light blue: 25); outline color reflects the GWAS source (pink: Grove et al., 2017; green: iPsych (Robinson et al., *Nat. Genet*. 2016), blue: PGC-2017). **b**, Meta-GSEA scores for significant MAGMA genes in BW-M4 across all tissues, implicating synaptic transmission and calcium transport as sources of neuronal dysfunction in SCZ. Red line denotes FDR significance level.



Extended Data Fig. 7 | Regional AUPR curves for CEREB-M1 and CTX-M1. Nearest-neighbor precision-recall curves for CEREB-M1 labels across all region-level co-expression networks; showing significantly higher AUPR for cerebellar regions, but substantial AUPR for all remaining regions. Right. Nearest-neighbor precision-recall curves for CTX-M3.



Extended Data Fig. 8 | Exploration of an omnigenic-like model across different networks (related to Fig. 7). Plot of Phi statistics for InWeb brain PPI network ('PPI') and four regulatorycircuits.org ('RC') networks: Hippocampus ('Hippo'), amygdala ('Amy'), NEU+ neurons, astrocytes, and neuroprogenitor cells ('NPC'). Vertical breaks represent the study used to calculate phi, while the colors represent those studies used to define proposed core genes, or network central genes.

nature research

Corresponding author(s): Christopher L Hartl, Daniel H Geschwind

Last updated by author(s): Mar 28, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No custom algorithms were used in collection or quantification of GTEx data. STAR (https://github.com/alexdobin/STAR) was utilized for alignment, and RSEM (https://github.com/deweylab/RSEM) for quantification.
Data analysis	Custom algorithms for network construction, data processing, and omnigenic analysis are provided at https://github.com/dhglab/multiregional-networks . Non-customized basic analyses (such as enrichment analysis using `glm` in R) are described in the methods.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used is documented in prior publications and either publicly available (all but "Parikshak2016") or available upon request ("Parikshak2016"); including: GTEx data (https://gtexportal.org/home/), NABEC (GSE15745, GSE36192, GSE37205), UKBEC/Braineac (http://www.braineac.org), Oldham2008 meta analysis (GSE1572, GSE3790, GSE5392, GSE7540, GSE12649, GSE12654); Kang2011 (www.humanbraintranscriptome.org)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N=30-80 depending on brain region (explicit numbers in text)
Data exclusions	Individual samples excluded due to outlier metric
Replication	Comparisons to the same or closely-matched brain regions, to verify the extent of module replicability across datasets
Randomization	This study does not contain experimental groups
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	ANK2 S105-17 (Abnova)
Validation	ANK2-expressing astrocytes, western blot, used as positive control