RESOURCE

nature neuroscience

# Canonical genetic signatures of the adult human brain

Michael Hawrylycz[1,14], Jeremy A Miller[1,14], Vilas Menon[1,14], David Feng[1], Tim Dolbeare[1], Angela L Guillozet-Bongaarts[1], Anil G Jegga[2], Bruce J Aronow[2], Chang-Kyu Lee[1], Amy Bernard[1], Matthew F Glasser[3], Donna L Dierker[3], Jörg Menche[4–6], Aaron Szafer[1], Forrest Collman[1], Pascal Grange[7], Kenneth A Berman[8], Stefan Mihalas[1], Zizhen Yao[1], Lance Stewart[9], Albert-László Barabási[4–6,10,11], Jay Schulkin[12], John Phillips[1], Lydia Ng[1], Chinh Dang[1], David R Haynor[13], Allan Jones[1], David C Van Essen[3], Christof Koch[1] & Ed Lein[1]

**The structure and function of the human brain are highly stereotyped, implying a conserved molecular program responsible for its development, cellular structure and function. We applied a correlation-based metric called differential stability to assess reproducibility of gene expression patterning across 132 structures in six individual brains, revealing mesoscale genetic organization. The genes with the highest differential stability are highly biologically relevant, with enrichment for brain-related annotations, disease associations, drug targets and literature citations. Using genes with high differential stability, we identified 32 anatomically diverse and reproducible gene expression signatures, which represent distinct cell types, intracellular components and/or associations with neurodevelopmental and neurodegenerative disorders. Genes in neuron-associated compared to non-neuronal networks showed higher preservation between human and mouse; however, many diversely patterned genes displayed marked shifts in regulation between species. Finally, highly consistent transcriptional architecture in neocortex is correlated with resting state functional connectivity, suggesting a link between conserved gene expression and functionally relevant circuitry.**



The adult human brain is composed of many regions, with distinct distributions of cell types and patterns of functional connectivity. Underlying this complexity is differential transcription, whereby different brain regions and their constituent cell types express unique combinations of genes during their developmental specification and maturation and in their mature functional state. Despite a range of brain sizes across individuals and variation in sulcal patterning in the neocortex, the general anatomical positioning of and connectivity between regions is highly stereotyped between individuals, suggesting that a substantial proportion of the transcriptional coding for this common architecture is conserved across the human population.

In contrast to numerous studies that explore genetic variants associated with disease traits by analyzing enormous sample sizes in population studies[1,2], we aimed to identify the core or 'canonical' transcriptional machinery conserved across individuals. If common expression relationships can be identified with high confidence in modest sample sizes and with good anatomical coverage of various brain regions, the resulting 'default gene network' could provide a base template for understanding the genetic underpinnings of highly conserved features of brain organization and a baseline from which deviations in individuals may be measured and associated with diseases such as autism, schizophrenia, epilepsy and major depression. While previous studies have identified gene networks associated with normal and diseased brain architecture in limited brain regions[3–7], the new availability of a data set with vastly enhanced structural coverage allows an explicit approach aimed at identifying network structure common across individuals that is related to structural and functional organization of the entire brain.

We approached this problem by identifying genes with highly consistent patterning across anatomical structures in six independent human brains of the Allen Human Brain Atlas (http://human.brain-map.org/) using the concept of differential stability (DS), which we define as the tendency for a gene to exhibit reproducible differential expression relationships across brain structures[8]. To understand large-scale transcriptome organization, we apply weighted gene coexpression network analysis (WGCNA)[9,10] to sets of high-DS genes. This and other quantitative network-based approaches have proven to be powerful tools for elucidating cell type, anatomic and species-specific patterning. Studies using these methods suggest that, largely because of their nonparametric, statistically robust nature, conserved differential expression relationships may be more descriptive of transcriptome organization than absolute magnitudes of expression

[1]The Allen Institute for Brain Science, Seattle, Washington, USA. [2]Division of Biomedical Informatics, Cincinnati Children's Hospital and Medical Center, Cincinnati, Ohio, USA. [3]Department of Anatomy and Neurobiology, Washington University, St. Louis, Missouri, USA. [4]Center for Complex Networks Research, Northeastern University, Boston, Massachusetts, USA. [5]Department of Physics, Northeastern University, Boston, Massachusetts, USA. [6]Center for Network Science, Central European University, Budapest, Hungary. [7]Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Jiangsu, China. [8]Department of Electrical Engineering and Computing Systems, University of Cincinnati, Cincinnati, Ohio, USA. [9]Institute for Protein Design, University of Washington, Seattle, Washington, USA. [10]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [11]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [12]Department of Neuroscience, Georgetown University, Washington, DC, USA. [13]Department of Radiology, The University of Washington, Seattle, Washington, USA. [14]These authors contributed equally to this work. Correspondence should be addressed to M.H. (mikeh@alleninstitute.org) or E.L. (edl@alleninstitute.org).

Received 22 August; accepted 16 October; published online 16 November 2015; doi:10.1038/nn.4171

levels[3,5,11–13]. We found that high DS genes and the gene networks involving them showed highly significant enrichment of functional ontology, disease and drug association terms, as well as strong relationships to anatomical structure and functional connectivity, indicating that they may represent essential transcriptional features of the human brain.

## RESULTS

### Conserved transcriptional patterning in adult human brain

To identify genes with highly conserved patterning across brain regions, we analyzed the complete data set from the Allen Human Brain Atlas, consisting of six neurotypical adult whole brains. The brains were from three males of European ancestry, two African-American males and one woman of European ancestry, the first two of which were part of an initial report on the project[3]. For each brain, 345–911 samples spanning one ($n = 4$) or both ($n = 2$) hemispheres were analyzed using whole-genome Agilent microarrays. In total, samples from 232 discrete brain structures were sampled at least once in at least one brain. We first focused on comparing expression patterns for a smaller set of 96 brain regions that were sampled at least twice in at least five brains, pooling across hemispheres (**Supplementary Table 1**). **Figure 1a** shows the number of genes (out of the 81.8% expressed in all six brains that were differentially expressed between pairs of these regions in at least five of six specimens. Large transcriptional differences are seen between major structures. Heterogeneity within subdivisions is also evident, such as the distinctive patterning of primary visual cortex compared to other cortical regions[3,14] and complex differentiation of nuclei in the brainstem. In contrast, cortex, cerebellum and amygdala were notably homogenous across their constituent subdivisions. This representation highlights the magnitude of mesoscale (fine but not cellular resolution) similarities and differences between brain regions. In addition to the open access data resource (http://www.brain-map.org/), this new map, its associated genes, metadata and anatomic guide are available for dynamic online browsing (http://casestudies.brain-map.org/ggb/).

In addition to globally conserved differential relationships, the patterning for any given gene across structures in different individuals was often well conserved. The CNS-specific cell adhesion protocadherin family member *PCDH8*, implicated in the development of schizophrenia[15], had a distinctive brain-wide pattern highly consistent across all six brains (**Fig. 1b**). High conservation of anatomical patterning across individuals has been proposed to explain common core functionality of the human brain[16,17], and we sought to test that hypothesis with this data set. So as to not overly bias the analysis toward the homogeneous cortex and cerebellum, here we treated all samples from the same cortical or cerebellar lobes as coming from a single brain region (**Supplementary Table 1**). Finally, by including only structures available in all brains, we retained a set of 132 of the original 232 structures, which are used for the remainder of the analysis.

### Genes with conserved patterning relate to brain function

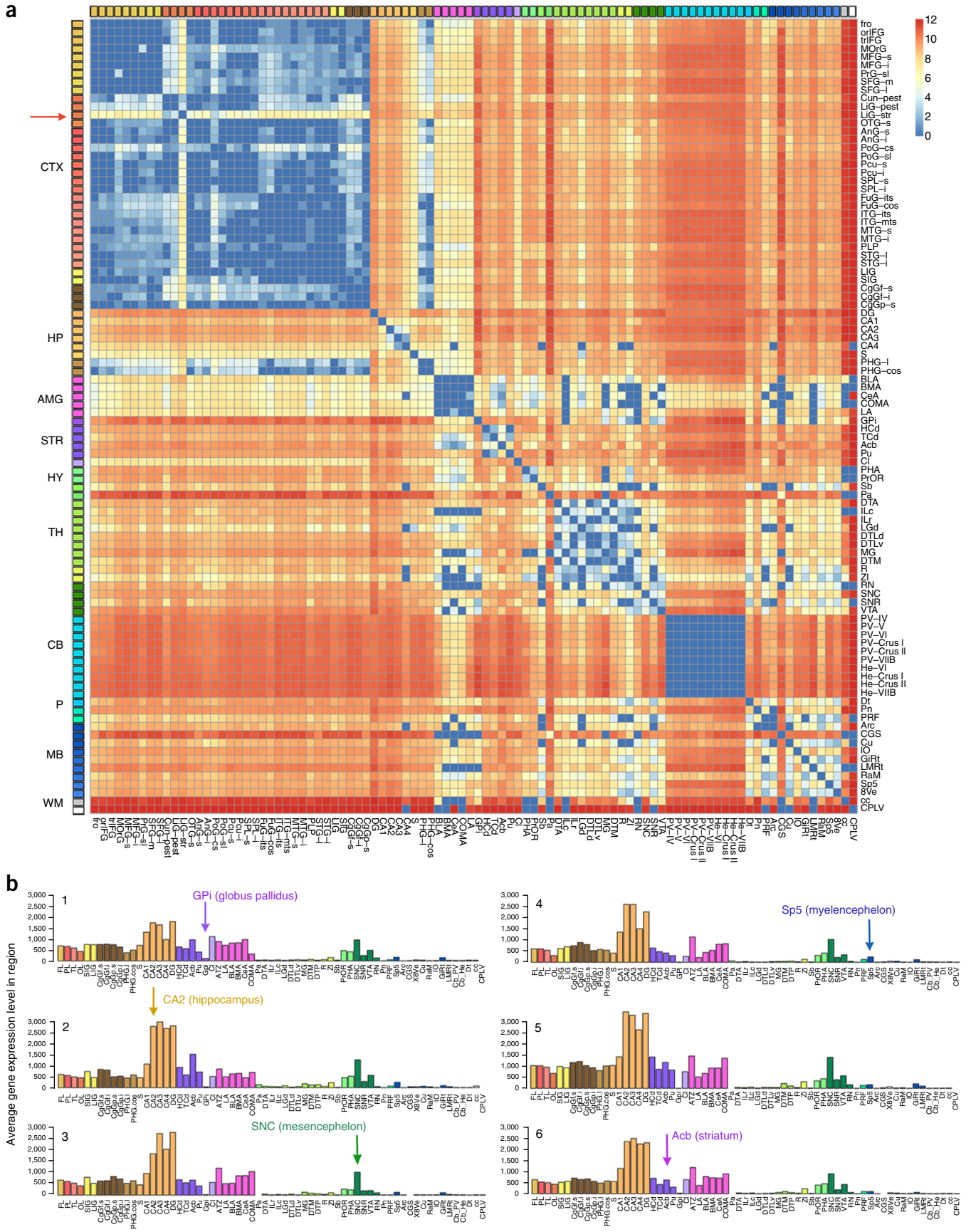
To quantify reproducibility of expression, we measured the tendency of any one gene to have consistently higher expression in one region than another across all brains sampled. We call this concept DS, which we quantify as the average Pearson correlation ρ between pairs of brain specimens over a set of anatomic regions. This concept of reproducibility of structural patterning is well defined and can be captured nearly equivalently by a variety of other metrics (see **Supplementary Analysis**). DS effectively measures the fraction of preserved differential relationships between regions, with a bias for stronger differentials (**Fig. 2a**). We denote the DS of a gene $g$ over a set of anatomic regions $S$ as $\Delta_S(g)$. Brain DS ($\Delta_{BR}$) values ranged between $\Delta_{BR} = 0.970$ and $-0.057$ for the 17,348 uniquely annotated human transcripts represented on the microarrays (**Supplementary Table 2**).

The gene with the highest DS was *FOXG1* ($\Delta_{BR} = 0.970$), a transcription factor regulating the development of the early-born Cajal-Retzius cells in the cerebral cortex[18,19]. *PCDH8* (shown in **Fig. 1b**) also had a high DS, with $\Delta_{BR} = 0.931$. Calcium channels and genes associated with calcium/calmodulin signaling (*NRGN*, *CAMK2A*, *CAMKV*) were overrepresented among high-DS genes ($P < 1.32 \times 10^{-4}$, Benjamini and Hochberg corrected) indicating a strong conservation of these signaling pathways. Potassium channels ($P < 1.70 \times 10^{-12}$) were the most over-represented among high-DS genes, exhibiting diverse and specific expression patterns (**Supplementary Fig. 1**). The DS metric selects for genes with moderate to high expression levels and moderate to high variability across structures; however, there was no significant relationship ($\rho = 0.064$) between mean gene expression level and high DS (**Fig. 2b**).

Previous studies have demonstrated transcriptional similarity based on anatomical proximity[3,14], which may reflect phylogenetic and ontogenetic distance[20]. Consistent with this observation, the relationship between anatomical structures based on these top DS genes generally resembled the physical topography and developmental origins of brain regions, as shown by multidimensional scaling in which distance is determined by correlation between samples (**Fig. 2c**). This ordering and separation between structures was progressively less preserved with decreasing DS (**Supplementary Fig. 2**). In contrast, low-DS genes with ubiquitous expression (bottom quartile in variability across brain regions; **Supplementary Table 3**), such as the ubiquitin isoforms *UBB* and *UBC*, were typically not brain specific. Instead, they appear to provide the basal machinery common to all cells and were enriched for housekeeping functions such as RNA binding ($P < 3.26 \times 10^{-21}$), KEGG spliceosome pathway ($P < 6.4 \times 10^{-13}$) and mitochondrial ribosomal proteins ($P < 1.32 \times 10^{-10}$).

The top decile (highest 10%) of genes ($n = 1{,}735$) ranked by DS was significantly over-represented in many functional categories (**Fig. 2d**), including gene ontology (GO) molecular function, cellular component and biological processes associated with neuronal function and transcription factor binding sites (**Supplementary Table 4**). Furthermore, the top DS genes had a disproportionate number of drug interactions and a monotonically decreasing association with microRNA targets. Specifically, the top DS decile collectively had

**Figure 1** Reproducible differential gene expression between brain regions across six individuals. (**a**) Consensus map of all genes differentially expressed between any pair of 96 regions in at least five of six specimens. Each matrix entry represents the number of genes with at least a fold change >3 in expression level between those two structures (unpaired *t*-test, Benjamini and Hochberg FDR-corrected $q < 0.01$). The scale is $\log_2$, with deep blue indicating no genes that are differentially expressed above threshold. Large differences between main brain structures are apparent, with comparative homogeneity within structures such as cerebellum, amygdala and cerebral cortex. LiG-str (lingual gyrus–striate cortex) contains primary visual cortex (red arrow), which shows differential expression relative to the rest of cerebral cortex. (**b**) Consistent expression pattern of *PCDH8* across individual brains (numbered 1–6), with cortical and cerebellar samples reduced to main lobes (shown for 65 of 132 structures for readability). FL, frontal lobe; OL, occipital lobe; TL, temporal lobe; PL, parietal lobe; CTX, cortex; HP, hippocampus; AMG, amygdala; STR, striatum; HY, hypothalamus; TH, thalamus; CB, cerebellum; P, pons; MB, midbrain; WM, white matter; GPi, globus pallidus internal segment; CA2, hippocampal CA2 subfield; SNC, substantia nigra pars compacta; Sp5, spinal trigeminal nucleus; Acb, nucleus accumbens. See **Supplementary Table 1** for other abbreviations.



4,651 interactions with small molecule drugs and chemicals from principal databases (including DrugBank and the Comparative Toxicogenomic Database; Online Methods), more than 30 times the number in the third highest decile. There was also a strong relationship between DS and the probability a given gene has been studied in the context of the brain (**Fig. 2e**).

Finally, brain-related diseases were associated with high-DS genes. The distribution of genes associated with neurological and psychiatric diseases was shifted toward high DS compared to all genes (**Fig. 2f**), on the basis of 2,289 gene sets from the Autworks database (http://autworks.hms.harvard.edu/)[21]. Genes sets associated with mental disorders were significantly enriched ($P < 8.06 \times 10^{-31}$) in the top 10% of

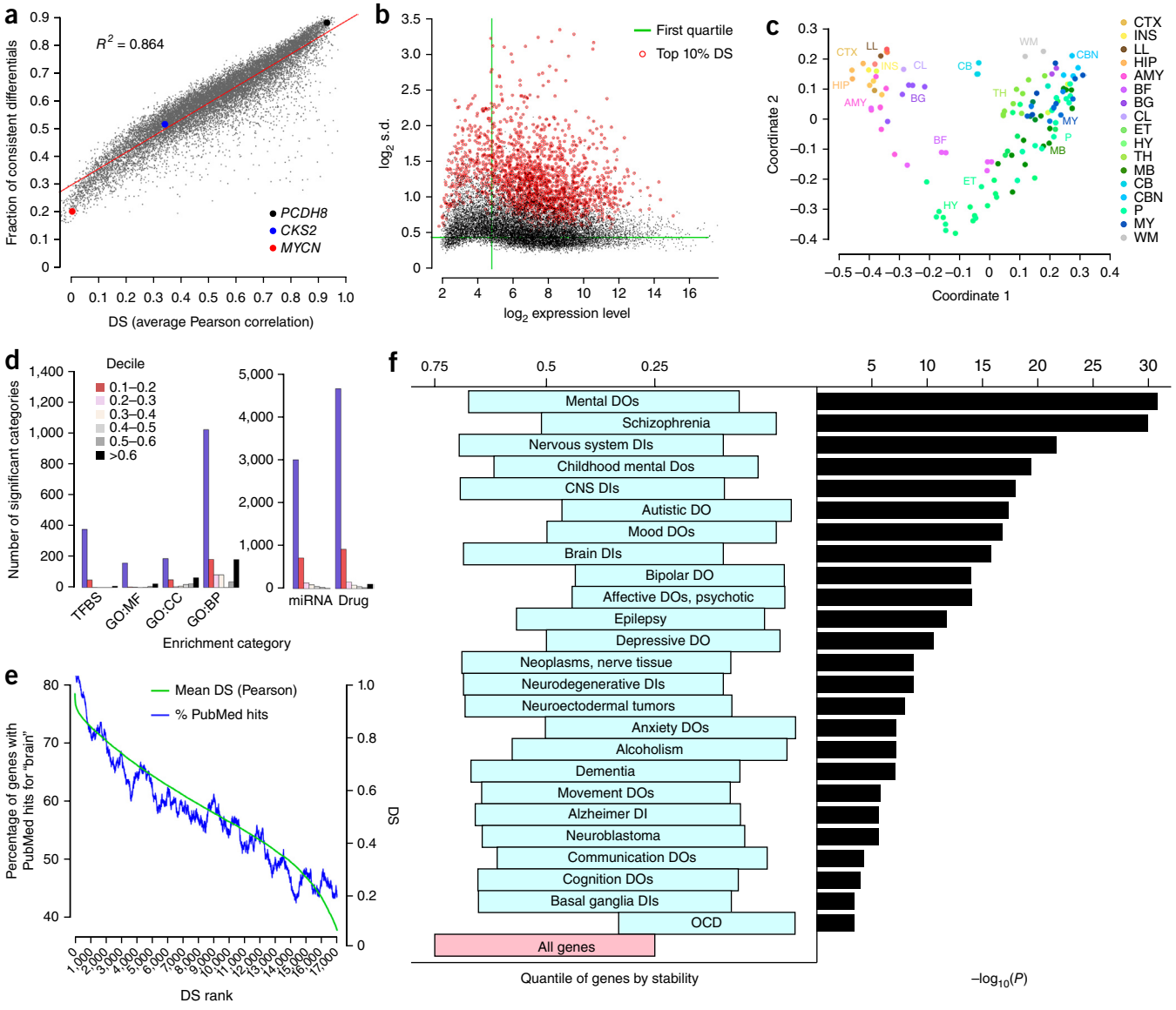



**Figure 2** Relationship of DS to brain structure, function and disease. (**a**) DS is strongly correlated with the fraction of conserved differential gene expression relationships between brain regions (81 regions in at least five of six brains, averaged across each pair of brains). *PCDH8* has high DS, whereas *CKS2* and *MYCN* show greater variability among individuals. (**b**) Variation over 132 structures (s.d.) versus expression level for all 17,348 genes. There is no significant relationship between the highest DS genes (top 10% highlighted in red), and expression level (ρ = 0.064). Green lines indicate first quartile in mean expression level and in s.d. (**c**) Multidimensional scaling analysis to represent transcriptional similarity between anatomical structures using genes in the top decile of DS. Anatomical subdivisions generally cluster according to their membership in major brain structures, common ontogenetic origin and topographic proximity. CTX, cortex; INS, insula; LL, limbic lobe; HIP, hippocampus; AMY, amygdala; BF, basal forebrain; BG, basal ganglia; CL, claustrum; ET, epithalamus; HY, hypothalamus; TH, thalamus; MB, midbrain; CB, cerebellum; CBN, cerebellar nuclei; P, pons; MY, medulla; WM, white matter. (**d**) Highest DS genes (blue bars) are significantly overrepresented (Benjamini and Hochberg corrected, FDR $q < 0.01$) in functional annotations, markedly so for drug targets. TFBS, transcription factor binding site; GO:MF, gene ontology molecular function; CC, cellular component; BP, biological process; miRNA, miRNA binding site. (**e**) Strong relationship between DS rank and frequency of study in the literature. Genes ordered by DS score on *x* axis; the *y* axis shows the fraction of genes in sliding windows of 600 genes (blue) scored for brain citation based on PubMed queries of the form "gene name" + "brain" (scoring 0 for no hit, 1 for positive hits). (**f**) DS distribution of genes associated with brain diseases relative to all genes, showing the interquartile range for each gene set from Autworks (left panel). Diseases are sorted by a hypergeometric test (right panel), assessing significance of overlap between each gene set and genes with a DS in the top 10% ($-\log_{10}$(*P*-value)). All *P*-values shown attain a significance of $10^{-3}$ or better. DO, disorder; DI, disease; OCD, obsessive-compulsive disorder.

DS genes (**Fig. 2f** and **Supplementary Table 5**): in particular, those for schizophrenia ($P < 9.14 \times 10^{-30}$), autism ($P < 3.45 \times 10^{-17}$), epilepsy ($P < 1.00 \times 10^{-11}$), bipolar disorder ($P < 4.94 \times 10^{-14}$) and major depression ($P < 4.60 \times 10^{-10}$). This result was independently reproduced in gene sets from the Simons Foundation Autism Research Initiative Database (https://gene.sfari.org/autdb/Welcome.do; $n = 666$, $P < 1.19 \times 10^{-38}$) and Alzheimer's disease–related genes from the Alzheimer's Forum (http://www.alzforum.org/; $n = 630$; $P < 3.55 \times 10^{-7}$). Taken together, these observations indicate that tightly regulated structural transcription is important for brain function and dysfunction.

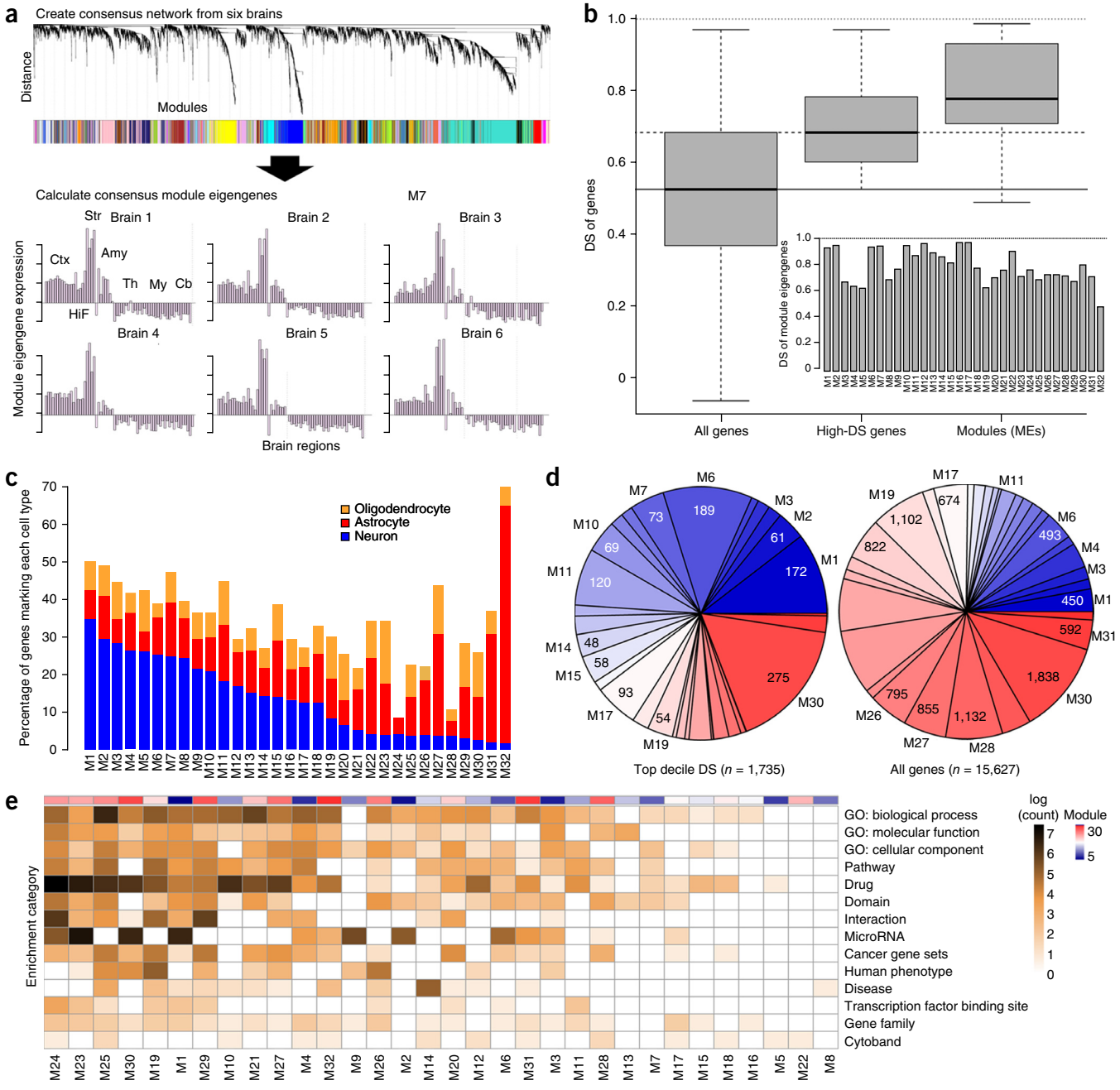



**Figure 3** Consensus coexpression patterns of the adult brain. (**a**) Network construction using consensus WGCNA on high DS genes[22]. Top, assignment of genes to modules. Bottom, module eigengene expression for M7 showing consistent patterning across all six brains, with peak expression in the striatum. Brain regions defined as in **Figure 2**. (**b**) Distribution of DS scores for all genes (left) and genes with $\Delta_{BR} > 0.528$ used for initial network construction (middle). DS metric applied to MEs demonstrates high consistency of module patterns across brains (right). Box plot shows minimum, first quartile, median, third quartile and maximum values. Inset shows DS score for each ME. (**c**) Percentage of known neuron-, astrocyte- and oligodendrocyte-enriched genes in 32 modules, ordered by proportion of neuron-enriched gene membership. Other genes add to 100%. (**d**) Module composition for the top decile of DS genes (left) versus all genes (right), with modules color-coded with uniform scale from high (blue) to low (red) neuron-enriched gene membership. Note the larger proportion of genes assigned to more neuronal modules for high DS genes compared to all genes. (**e**) Most gene annotations are associated with a subset of modules. Log counts of number of significant (Benjamini and Hochberg corrected, FDR $q < 0.05$) enrichments for gene ontology terms, protein interaction, cytoband, gene families, pathways and drug interactions. Modules ordered from most (M24, left) to fewest (M8, right) annotations, with neuron-enriched gene membership indicated in top color bar. Color bar as in **d**.

### A canonical brain transcription coexpression network

Given the relevance of high DS genes to brain function demonstrated above, we sought to understand the gene regulatory landscape by deriving a core set of coexpression modules based on high-DS genes that capture most of the variation between regions. Despite the cellular heterogeneity inherent in different brain regions, network analytic techniques such as WGCNA[10] have revealed biologically meaningful trends in mesoscale-sampled transcriptome data that relate to anatomy and brain function. Indeed, previous analysis using WGCNA on the first human brain from this data set identified coexpression modules that were generally corroborated in the second brain[3].

**Table 1 Ontological and functional associations of the canonical modules**

| Module (hub gene) | Anatomy | Ontology or pathway | *q*-value FDR | Drug or disease | *q*-value FDR |
|---|---|---|---|---|---|
| M01 (*GABRB3*) | Telencephalon | Synaptic transmission<br>Regulation of synaptic plasticity | $q < 1.07 \times 10^{-17}$<br>$q < 1.04 \times 10^{-10}$ (*) | Cocaine-related disorders<br>Nicotine addiction | $q < 4.51 \times 10^{-9}$<br>$q < 5.38 \times 10^{-6}$ (*) |
| M03 (*KCNAB2*) | Hippocampus, thalamus, pons, medulla | Neurotransmitter transport<br>Axon part | $q < 4.28 \times 10^{-6}$ (*)<br>$8.62 \times 10^{-7}$ (*) | Abnormality of pyramidal motor function | $q < 1.99 \times 10^{-3}$ |
| M04 (*GABARAPL1*) | Thalamocortical | Synaptic vesicle cycle<br>Insulin receptor recycling | $q < 4.66 \times 10^{-14}$ (*)<br>$q < 2.72 \times 10^{-5}$ (*) | Huntington's disease | $q < 1.28 \times 10^{-4}$ (*) |
| M06 (*MEF2C*) | Neocortex, claustrum | Postsynaptic membrane<br>Cell signaling | $q < 2.5 \times 10^{-5}$<br>$q < 3.58 \times 10^{-6}$ | Clozapine (schizophrenia, bipolar disease) | $q < 1.87 \times 10^{-3}$ |
| M07 (*NGEF*) | Striatum, neocortex, amygdala | Calcium signaling pathway<br>Dendritic spine membrane | $q < 1.07 \times 10^{-4}$<br>$q < 1.69 \times 10^{-3}$ (*) | Fluoxetine (depression, OCD) | $q < 4.34 \times 10^{-3}$ |
| M09 (*PGAP1*) | Hippocampus, amygdala, hypothalamus | Synaptic membrane<br>Zinc finger, CH-2 | $q < 5.23 \times 10^{-4}$<br>$q < 6.23 \times 10^{-3}$ | Cognitive impairment<br>Amyotrophic lateral sclerosis | $q < 5.99 \times 10^{-4}$<br>$q < 1.74 \times 10^{-3}$ (*) |
| M10 (*ADORA2A*) | Striatum | Monoamine GPCRs<br>Dopamine receptor signaling | $q < 7.27 \times 10^{-5}$ (*)<br>$q < 3.30 \times 10^{-5}$ (*) | Drug-induced dyskinesia<br>Haloperidol (schizophrenia, Tourette's syndrome) | $q < 1.23 \times 10^{-6}$ (*)<br>$q < 9.76 \times 10^{-7}$ |
| M11 (*NTNG1*) | Dorsal thalamus | Cadherin signaling pathway<br>Cholinergic synapse | $q < 2.02 \times 10^{-3}$ (*)<br>$q < 2.45 \times 10^{-4}$ (*) | Alzheimer's disease: presenilin pathway | $q < 2.78 \times 10^{-3}$ (*) |
| M12 (*SLC6A3*) | Substantia nigra, ventral tegmental area | Adrenaline, noradrenaline<br>Dopamine biosynthesis | $q < 5.48 \times 10^{-6}$<br>$q < 8.39 \times 10^{-6}$ (*) | Cocaine addiction<br>Dopamine | $q < 5.64 \times 10^{-5}$<br>$q < 3.68 \times 10^{-6}$ |
| M14 (*TLE6*) | Hypothalamus | Neuropeptide signaling<br>GPCR ligand binding | $q < 9.75 \times 10^{-3}$ (*)<br>$q < 1.76 \times 10^{-4}$ (*) | X-linked mental retardation<br>Prader-Willi syndrome | $q < 2.47 \times 10^{-3}$ (*)<br>$q < 2.47 \times 10^{-3}$ (*) |
| M15 (*NEFH*) | Deep cerebellar nuclei, brainstem | Neuron projection<br>Neurofilament | $q < 3.42 \times 10^{-3}$<br>$q < 3.49 \times 10^{-4}$ (*) | Dexamethasone (cerebral inflammatory) | $q < 4.99 \times 10^{-3}$ |
| M16 (*SLC47A1*) | Dentate gyrus | Protocadherin genes | $q < 5.74 \times 10^{-4}$ (*) | Depressive disorder<br>Parkinson's disease | $8.48 \times 10^{-3}$ (*)<br>$q < 8.48 \times 10^{-3}$ |
| M17 (*CBLN3*) | Cerebellar cortex | Zinc fingers, C2H2 type<br>Spinal cord development | $q < 3.66 \times 10^{-5}$<br>$q < 8.83 \times 10^{-3}$ (*) | | |
| M19 (*VDAC2*) | Thalamus, cerebellar nuclei, brainstem | Mitochondrial<br>Cellular respiration | $q < 5.50 \times 10^{-82}$ (*)<br>$q < 1.07 \times 10^{-35}$ | Ataxia<br>Leigh's syndrome<br>Alzheimer's disease<br>Parkinson's disease | $q < 2.65 \times 10^{-9}$ (*)<br>$q < 2.18 \times 10^{-8}$ (*)<br>$q < 3.98 \times 10^{-22}$ (*)<br>$q < 1.62 \times 10^{-27}$ |
| M20 (*B3GAT1*) | White matter, neocortex, basal ganglia, ventral thalamus | Eukaryotic translation<br>Ribosomal nucleolus | $q < 6.32 \times 10^{-3}$ (*)<br>$q < 4.82 \times 10^{-14}$ | Disease progression<br>Selenium (laboratory value)<br>Abnormal blood glucose | $q < 5.79 \times 10^{-5}$<br>$q < 2.00 \times 10^{-6}$<br>$q < 1.20 \times 10^{-3}$ (*) |
| M21 (*GBP4*) | Sensory-motor nuclei, choroid | Vasculature development | $q < 1.17 \times 10^{-17}$ (*) | Toluene (abuse)<br>Losartan (stroke)<br>Azidothymidine (HIV) | $q < 3.61 \times 10^{-15}$<br>$q < 8.13 \times 10^{-8}$<br>$q < 1.17 \times 10^{-7}$ |
| M24 (*POGZ*) | Cerebellar cortex, dentate gyrus, white matter, basal ganglia | Zinc fingers, C2H2 type<br>Chromatin organization | $q < 7.37 \times 10^{-40}$<br>$q < 2.77 \times 10^{-16}$ (*) | Provocholine (bronchial)<br>Ellipticine (cancer) | $q < 1.53 \times 10^{-16}$<br>$q < 1.32 \times 10^{-9}$ |
| M25 (*RGS10*) | Ependyma, white matter, substantia nigra | Immune system regulation<br>Inflammatory response | $q < 3.71 \times 10^{-35}$ (*)<br>$q < 3.70 \times 10^{-26}$ (*) | Systemic lupus erythematosus<br>Malignant glioblastoma | $q < 8.49 \times 10^{-29}$ (*)<br>$q < 1.01 \times 10^{-9}$ |
| M26 (*MYCBP*) | Ependyma | Cilium organization<br>Axoneme | $q < 1.47 \times 10^{-28}$ (*)<br>$q < 1.79 \times 10^{-33}$ (*) | Breathing dysregulation | $q < 2.25 \times 10^{-5}$ (*) |
| M28 (*SERPINA6*) | Interbrain-hindbrain nuclei | G-protein-coupled receptors<br>Olfactory receptors | $q < 4.06 \times 10^{-7}$ (*)<br>$q < 1.0 \times 10^{-3}$ (*) | | |
| M29 (*GAS5*) | White matter, substantia nigra, globus pallidus | Cytosolic ribosome<br>Translation activity | $q < 2.96 \times 10^{-102}$ (*)<br>$q < 1.81 \times 10^{-85}$ (*) | Influenza life cycle<br>Vigabatrin (seizures, epilepsy) | $q <$ 4.87E-69 (*)<br>$q < 1.70 \times 10^{-18}$ |
| M30 (*VAMP3*) | White matter, ventral thalamus, globus pallidus | Myelination<br>Axon ensheathment | $q < 5.63 \times 10^{-9}$<br>$q < 3.77 \times 10^{-9}$ (*) | Cognitive impairment<br>Hereditary spastic paraplegia | $q < 8.99 \times 10^{-6}$ (*)<br>$q < 7.83 \times 10^{-3}$ (*) |
| M32 (*SLC25A18*) | Striatum, amygdala, substantia nigra | Glial cell differentiation<br>Astrocyte differentiation | $q < 4.04 \times 10^{-5}$<br>$q < 2.42 \times 10^{-4}$ (*) | Dexamethasone (corticosteroid)<br>Deafness | $q < 1.58 \times 10^{-7}$<br>$q < 6.67 \times 10^{-3}$ (*) |

Each module is given with a representative hub gene, anatomic description, ontology and pathway, drug and disease associations. (*) indicates items uniquely associated with the module. Modules not reported are weakly annotated. All *q*-values 0.01 FDR.

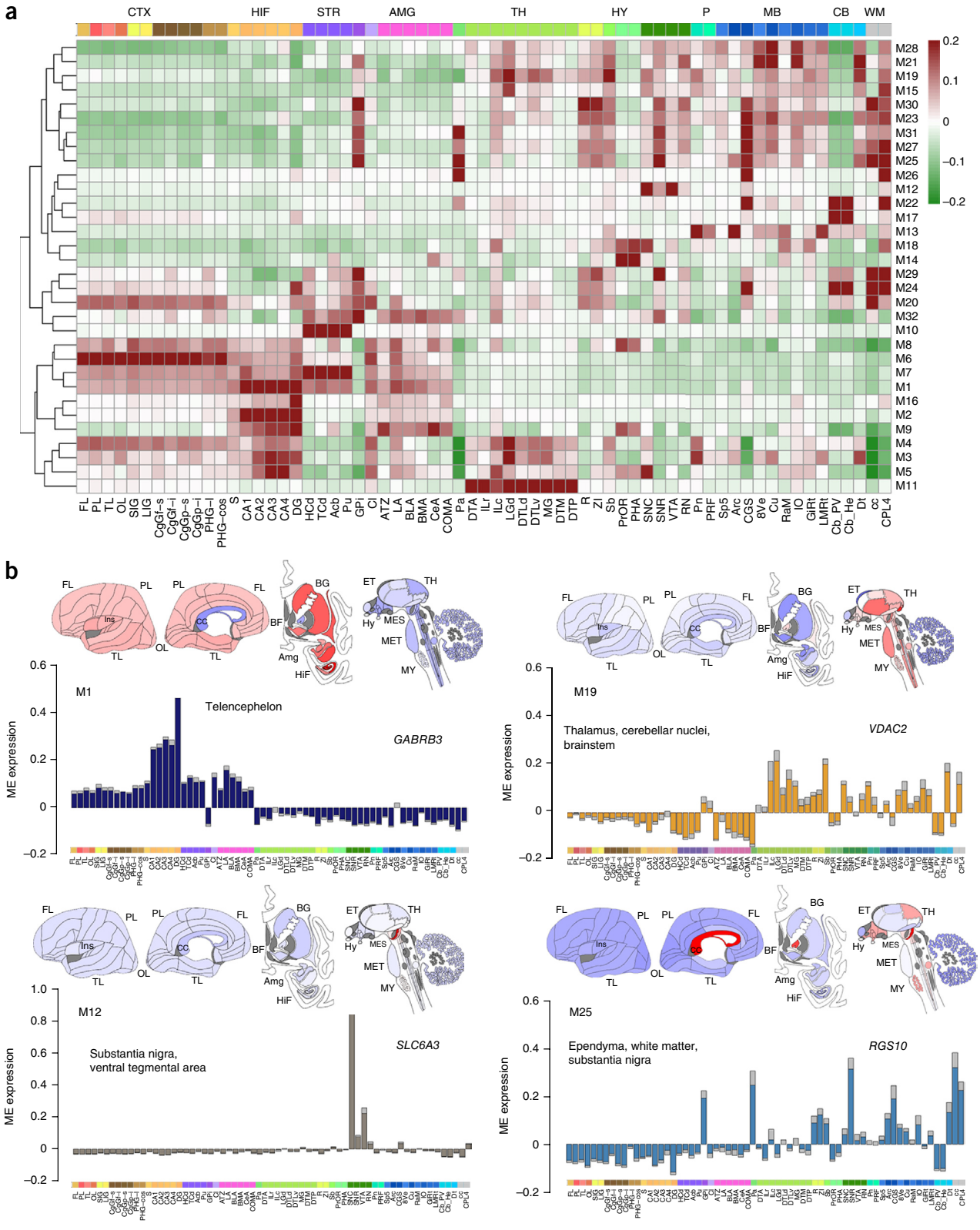



**Figure 4** Anatomical specificity of module expression. (**a**) Hierarchical clustering of the modules M1–M32 with anatomic ordering fixed from cortex to brainstem. Individual modules are enriched in specific structures, with a partition between predominantly telencephalic versus deeper brain region enrichment. See **Figure 1** for abbreviations. (**b**) Anatomical patterning of four representative modules with average ME expression values plotted on brain diagrams and on bar plots (gray regions, s.e.m.), with a representative hub gene. Anatomical structures (color-coded by major structure); bar below ordered from neocortex (left) to brainstem (right). ME expression range scaled from minimum (blue) to maximum (red) in brain structure diagrams above. Plots for all modules provided in **Supplementary Figures 5–8**.

Using the six brains now available, we identified a consensus network[22] common to all brains (**Fig. 3a**), using the top half DS genes ($\Delta_{BR} > 0.5284$, $g = 8{,}674$; Online Methods). This network identified 32 main transcriptional patterns, or modules, each represented by a characteristic expression pattern across brain structures, the module eigengene (ME), in each brain (**Fig. 3a**). We then assigned every gene to one of the 32 modules if its average correlation to the corresponding ME across all six brains was greater than 0.4. Most genes (90.1%; 15,627) correlated with this relatively small set of patterns (**Supplementary Table 6**), and the MEs derived from these genes were highly stable (**Fig. 3b**). Furthermore, these patterns are likely to be representative of a larger population because they were highly consistent across all six individuals, and patterns identified using any five brains could be found reproducibly in the sixth (**Supplementary Analysis** and **Supplementary Fig. 3**).

Previous studies demonstrated that brain gene modules often correspond to expression in main cell classes, including neurons and glia[3,5]. To ask how these 32 modules represent expression in specific cell types, we assessed the distribution of genes within each module that are known to be enriched (at least 1.5-fold) for neurons, oligodendrocytes and astrocytes in cultured preparations from mouse[23] (**Supplementary Table 7**). While not a perfect comparison, given the differences between species and between *in vivo* versus *in vitro* conditions, there was a clear ordering of modules that could be seen on the basis of predominance of neuronal gene expression (**Fig. 3c**). These proportions ranged widely from those with largely neuronal signatures (M1–M15) to essentially non-neuronal (M16–M32). Notably, the proportion of genes assigned to modules with high neuronal expression (M1–M15; **Fig. 3d**) was significantly greater for the highest DS gene set (top decile).

To characterize each module functionally, we used the ToppGene portal[24] (https://toppgene.cchmc.org/; Online Methods) to identify significant enrichments in gene ontology, pathways, cytoband, disease association, transcription factor binding sites, miRNAs, drug targets and protein-protein interactions (**Fig. 3e** and **Supplementary Table 8**). Over 96% of all annotations were associated with just 17 of the 32 modules (**Fig. 3e**), including those restricted to the highly studied nigro-striatal pathway structures associated with Parkinson's disease (M10 and M12) as well as a set of predominantly non-neuronal modules (M24, M23, M25, M30; **Table 1**). Consistent with the overlap with neuronal genes, modules M1–M6 contained the most neuronal function–related annotations, including synapse structure and transmission,

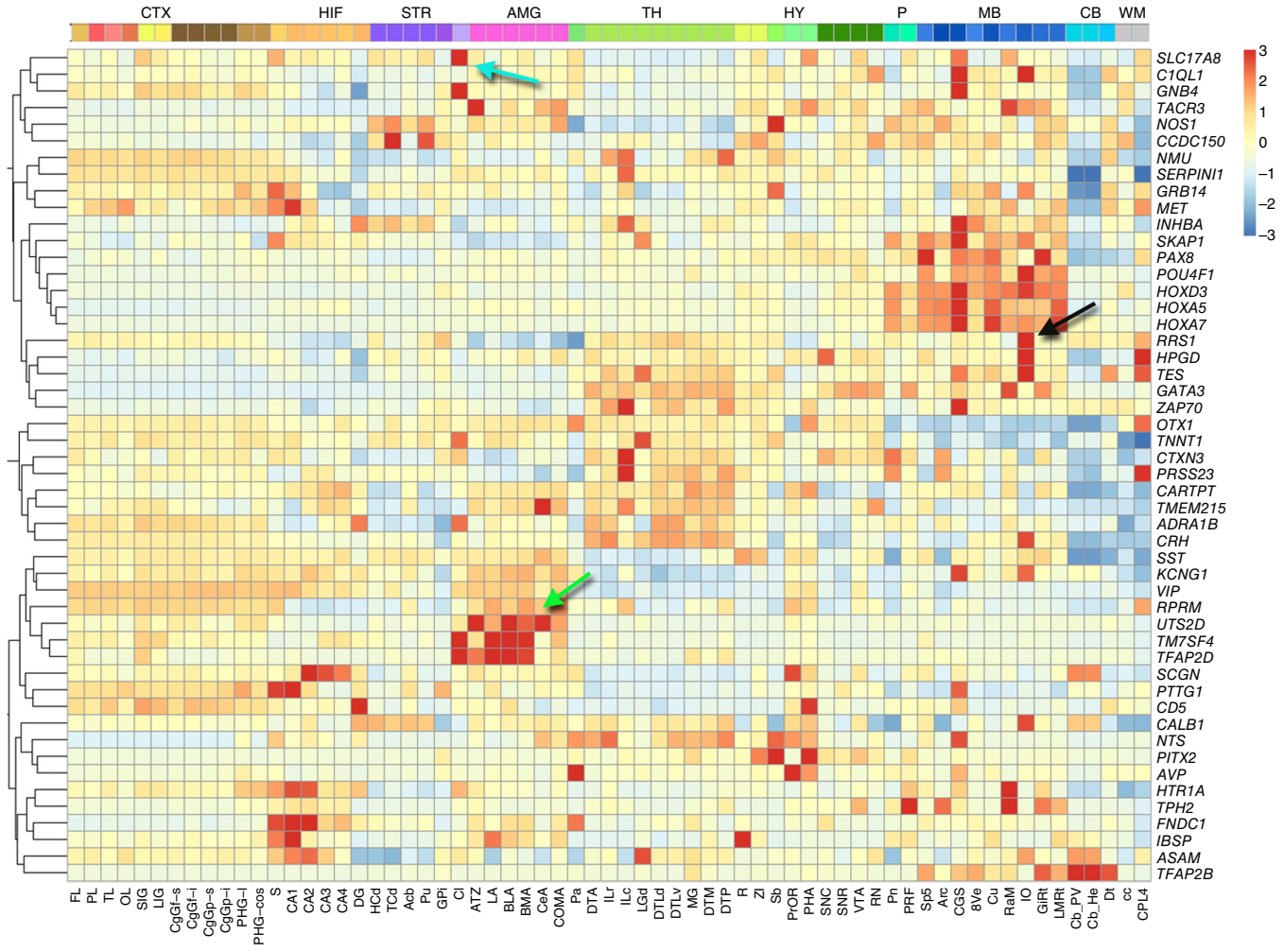



**Figure 5** Unique anatomical patterning of a subset of high-DS genes. Relative (*z*-score-normalized) expression of 50 genes (rows) across anatomical structures (columns), ordered from cortex to brainstem and clustered by expression pattern. The solute carrier family 17, member 8 (*SLC17A8*) (also known as *VGluT3*) gene is primarily expressed in the claustrum (Cl, cyan arrow), ribosome biogenesis regulator homolog (*RRS1*) is enriched in the inferior olivary nucleus (IO, black arrow), and urotensin 2D (*UTS2D*, green arrow) is enriched in the amygdala. These anatomically specific genes have low correlation with any of the 32 main modules.

neurogenesis and neural projection, and channel dynamics. The most non-neuronal modules (M21–M32) showed strong and unique associations with vasculature development (M21, $q < 1.17 \times 10^{-17}$, false discovery rate (FDR) 0.01), immune system function (M25, $q < 3.71 \times 10^{-35}$) and myelination and glial ensheathment (M30, $q < 1.48 \times 10^{-6}$). While by construction there were no common genes between any pairs of modules, there were common functional interactions between modules and enrichments based on ontology, pathway and function. These relationships generally reflect neuronal versus non-neuronal module content and are summarized in **Supplementary Figure 4**.

From a global anatomical perspective, the 32 gene modules were predominantly enriched either in the telencephalon or in deeper, non-telencephalic brain structures (**Fig. 4a**). Individual modules were enriched in distinct combinations of structures (**Fig. 4b** and **Supplementary Figs. 5–8**) and were associated with distinct functional pathways, disease and drug interactions (**Table 1** and **Supplementary Fig. 9**). For example, M1 is highly enriched in cortex, hippocampus, amygdala and striatum, with genes associated with synaptic transmission and genes downregulated in Alzheimer's disease. Some modules were remarkably selective for particular brain regions, such as M6 in cortex, M10 in striatum, M11 in thalamus, M12 in the substantia nigra, M14 in hypothalamus, M16 in dentate gyrus and M17 in the cerebellum. In agreement with known function, M12 was uniquely associated with dopamine biosynthesis ($q < 8.39 \times 10^{-6}$) and cocaine addiction ($q < 5.64 \times 10^{-5}$). Other modules showed more complex anatomical patterning, such as non-telencephalic expression (M19), the inverse of M1. M19 was associated with mitochondria ($q < 5.50 \times 10^{-82}$) and

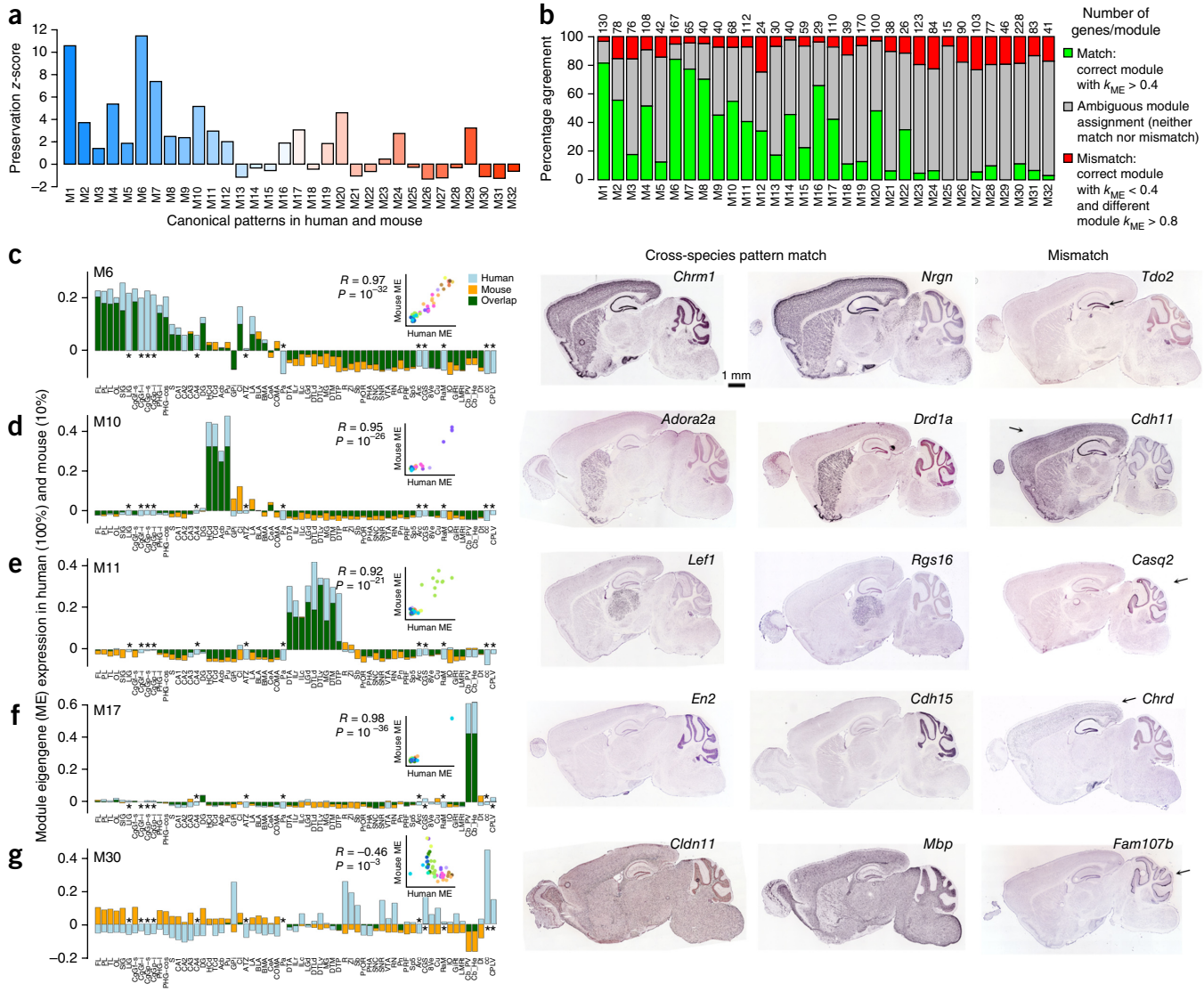



**Figure 6** Module preservation from human to mouse. (**a**) Mouse-human module preservation index, which measures conserved within-module gene coexpression in an anatomy-independent fashion, shows the highest preservation for some of the most neuronal modules (M1, M6, M7). (**b**) Conservation of anatomical patterning, defined as the proportion of mouse genes correlated at ρ > 0.4 to the corresponding human ME (green bars, bottom). A subset of genes in each module are both poorly correlated to the human eigengene (gray bars, middle) and very highly correlated to a different human module eigengene (ρ > 0.8, red bars, top). (**c**–**g**) Correspondence of ME anatomical patterning between human and mouse. Left, histogram representation of ME pattern in human (light blue) and mouse (orange), with overlap in green, demonstrate highly conserved patterns for M6, M10, M11 and M17, while M30 is weakly anticorrelated. Inset panels show correlation between mouse and human. Asterisks indicate samples present in human but not mouse. Right, mouse *in situ* hybridization for genes with matching and mismatched patterns in right panels, representing genes in the green and red categories in **b**. Arrows indicate areas of differential regulation in mouse. All hybridization images from the Allen Mouse Brain Atlas.

demyelination ($q < 2.57 \times 10^{-4}$) and had strong disease associations with Alzheimer's disease ($q < 3.98 \times 10^{-22}$), Parkinson's disease ($q < 1.62 \times 10^{-27}$) and the neurodevelopmental disorder Leigh's syndrome ($q < 2.18 \times 10^{-8}$). Consistent with its non-neuronal enrichment noted above, M25 was enriched in the glia-rich corpus callosum and internal portion of the globus pallidus and was associated with immune response ($q < 3.71 \times 10^{-35}$) and systemic lupus erythematosus ($q < 8.49 \times 10^{-29}$). The complete set of module patterns are provided in **Supplementary Figures 5–8** and the complete functional annotations in **Supplementary Table 8**.

Finally, a small number (302) of high-DS genes appeared to have unique spatial expression patterns and were not strongly correlated to any of the 32 module eigengenes. **Figure 5** shows 50 of these genes, illustrating the diversity of their expression patterns. Surprisingly, among these outlier patterns are several extensively studied genes, including somatostatin (*SST*), CART prepropeptide (*CARTPT*), Met proto-oncogene (*MET*) and calbindin 1 (*CALB1*), as well as regional markers (*UTS2D*, amygdala; *SLC17A8* (*VGluT3*), claustrum; and *RRS1*, inferior olive). These uniquely patterned yet stable genes were enriched in the G-protein-coupled receptor (GPCR) pathway and signaling events ($q < 2.59 \times 10^{-9}$; **Supplementary Table 9**). This set was overrepresented for genes identified through association analysis of regulatory and non-synonymous SNPs involved in neurotransmission and neurodevelopment[25] ($q < 2.11 \times 10^{-15}$) and that confer susceptibility to bipolar and obsessive-compulsive disorders and major depression, as well as anorexia nervosa[26] ($q < 7.23 \times 10^{-12}$). Thus, while most genes are correlated with a small number of global expression patterns across the brain, a few functionally important genes have patterns of spatial regulation that are distinct from these modules.

### Differential gene module conservation from human to mouse

Are these core transcriptional modules unique to the human brain or do they reflect mammalian brain architecture more generally? To address this question, we compared anatomical distributions and within-module coexpression metrics in human to those in mouse, a dominant model organism in biomedical research. We created an expression matrix for a set of 53 mouse structures unequivocally matching those in the human microarray data, using spatially quantified *in situ* hybridization data for the 2,651 genes with data for $n \geq 2$ brains in the Allen Mouse Brain Atlas[27] (Online Methods and **Supplementary Table 1**). We then created a mouse module eigengene for each of the 32 modules defined in human, using the subset of genes present in both species. We used a previously described module preservation index[22] to assess the degree to which the coexpression structure of genes in these mouse modules was conserved. Some modules were well preserved (**Fig. 6a**), including the neuron-enriched modules M1, M6 and M7, whereas many of the most non-neuronal modules (M25–M32) were poorly preserved. Because this metric is an aggregate score of all genes in a module, we then scored the percentage of genes in each module whose mouse

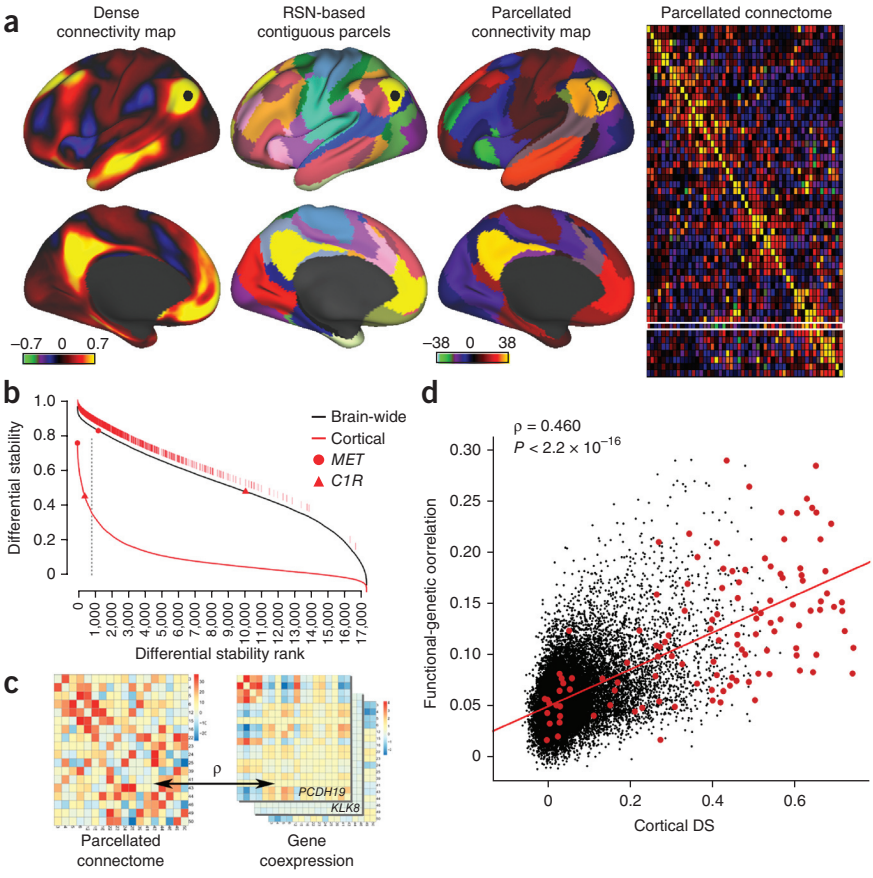



**Figure 7** Cortical DS and functional connectivity. (**a**) Resting-state functional connectivity from Human Connectome Project (HCP) data[29,45] (http://www.humanconnectome.org/). Column 1: functional connectivity from a seed in lateral parietal cortex (black disk), based on a group average of 468 HCP subjects and showing the full correlation results (scale: Fisher *z*-transformed ρ); data set accessible via https://db.humanconnectome.org/data/projects/HCP_500). Column 2: 52 left-hemisphere contiguous parcels (50 $mm^2$ or larger) from the 17 resting-state networks (RSN) identified by Yeo *et al.*[31]. Column 3: parcellated connectivity map for a default-mode network parcel (black outline) containing the selected seed, based on 447 HCP subjects (a subset of the above 468) and using partial correlation (scale: *z*-score). Column 4: group-average ($n = 447$) parcellated connectome showing relative connection strength between regions, also based on partial correlation. White rectangular outline identifies connectivity map shown in column 3. (**b**) Ranked cortical DS genes (red) and brain-wide DS genes (black). Top 5th percentile ($g = 867$) cortex DS genes (vertical line) are shown as hash marks just above the whole-brain DS curve. Red disk and triangle show two genes with differing brain-wide and cortical DS. (**c**) Functional correlation of parcellated RSN connectome (left) is compared with genetic coexpression similarity for each gene in each subject (right), using the same set of cortical parcels and by calculating the Pearson's correlation between the vectorized elements above the diagonal of the matrices. (**d**) DS versus functional-genetic correlation for 17,348 genes. Higher cortical DS genes are more predictive of functional cortical connectivity (ρ = 0.460, $P < 2.2 \times 10^{-16}$), whereas the correlation is substantially weaker for brain-wide DS (ρ = 0.17; data not shown). Red points are 132 genes identified as drivers of functional connectivity using the same post-mortem brain tissue data set as in the present study, whose polymorphisms are significantly associated with resting-state functional connectivity in a large sample of healthy adolescents[28].

patterns were well correlated with the structural expression pattern of the corresponding human module eigengene ($\rho > 0.4$). In general, this percentage varied with neuronal content (**Fig. 6b**) and decreased with decreasing DS (data not shown). In addition, the most neuron-associated modules had the most genes with highly correlated patterns across structures. Notably, each module also contained genes that not only were poorly correlated to the human module, but instead were very highly correlated ($\rho > 0.8$) to a different human module eigengene. **Supplementary Table 10** lists the module assignments for each cross-species gene and whether the gene's patterning was conserved in mouse.

Despite massive differences in size and measurement techniques of these two data modalities, many of the modules had very highly correlated expression across specific anatomical regions (**Fig. 6c–g**). For example, the forebrain-enriched and neuron-associated M6, striatal M10, thalamic M11 and cerebellar M17 had nearly identical patterning. In contrast, the module eigengenes for the non-neuronal module M30 was very poorly correlated. Despite this variation in the overall strength of correlation across species, individual genes with matching anatomical and cellular patterns could be found for each module (**Fig. 6c–g**). Conversely, for each module there were genes whose pattern fundamentally shifted between species—for example, from thalamic to cerebellar in M11 or from glial to neuronal in M30. Thus, it appears that, even in highly preserved modules, some individual genes differ in their patterning across species. The frequency of these changes in gene regulation varied by module, with better general pattern conservation in the more strongly neuronal modules.

### Functional connectivity relates to transcription in cortex

Despite its relative homogeneity compared to other brain regions, human neocortex shows variations in transcriptional patterns that are consistent across individuals and are correlated with topographical position across the cortical sheet[3]. Recent work has revealed that functional brain networks determined using resting-state functional magnetic resonance imaging (fMRI) can be recapitulated using correlated gene expression in post-mortem brain tissue[28]. To explore whether genes with highly consistent cortical patterning across individuals drive this functional organization, we compared transcriptional variation of high-DS genes with recent resting-state functional connectivity data from the Human Connectome Project (http://www.humanconnectome.org/)[29]. We generated a parcellated connectivity matrix **C** (netmat; Online Methods) averaged across 447 subjects (**Fig. 7a**) using the partial correlation[30] among spatially contiguous nodes derived from the resting-state network identified by Yeo *et al.*[31]. We then mapped each cortical region containing transcriptional data onto these corresponding 52 parcels and analyzed parcels containing at least one gene expression sample.

Because the transcriptional differences between cortical areas are generally smaller than between major brain regions, we recalculated DS selectively for cortex using those 52 cortical regions matching the functional parcels above. A smaller percentage of genes have high DS in cortex ($\Delta_{CTX}$; **Fig. 7b**): in contrast to whole brain, the top 5% ($n = 867$) of DS cortex genes had $\Delta_{CTX} > 0.357$, versus $\Delta_{BR} > 0.854$ for whole brain. The rank ordering of genes by whole-brain DS versus cortex-only DS differed markedly, with only 31% of the top 5% of cortical DS genes in the top 5% of brain-wide DS (**Fig. 7b**). For example, the *MET* proto-oncogene, a functional marker of glioblastoma (see http://glioblastoma.alleninstitute.org/)[32] was ranked first in cortex ($\Delta_{CTX} = 0.761$) but 985th in the whole brain. Similarly, the classical complement pathway gene *C1R*, which mediates immune and antibody response and is synthesized by both neurons and glia[33], had cortical rank 569 but a brain-wide rank of 10,094. *C1R* may contribute to inflammatory and degenerative diseases of the CNS, including Alzheimer's disease[33–35]. Thus, the DS metric applied to cortex alone identifies genes having more subtle but consistent variation across functionally defined cortical parcels that is not captured when highly distinct structures across the whole brain are included. Local DS scores across 20 major brain structures are supplied in **Supplementary Table 11**.

To examine which genes have spatial expression patterns that correlate with functional connectivity patterns and whether this is related to cortical DS, we correlated (Pearson $\rho$) the parcellated connectome matrix **C** with gene coexpression matrices **E**($g$) for each of 17,348 genes with data pooled to the same regions (**Fig. 7c** and Online Methods). **Figure 7d** shows that higher cortical DS of a gene predicts a stronger correlation between functional connectivity and cortical gene expression pattern ($\rho = 0.46$, $P < 2.2 \times 10^{-16}$); the correlation with brain-wide DS alone was much weaker ($\rho = 0.17$), indicating that the conserved cortical genes correlate most strongly with cortical functional connectivity. Further, we found that 136 genes whose polymorphisms significantly affect resting-state functional connectivity in a large sample of healthy adolescents[28] had both higher correlation of expression with functional connectivity ($\rho = 0.12$, $P < 2.2 \times 10^{-192}$) and higher DS ($\Delta_{CTX} = 0.40$, $P < 1.2 \times 10^{-325}$) in human cortex (**Fig. 7d**), though they constitute only a minority (15%) of the highest cortical 5th percentile DS genes. Significant ontological categories of this genes in this top 5% included main axon ($P < 7.84 \times 10^{-9}$), neuron projection ($P < 2.2 \times 10^{-4}$), calcium and sodium channel activity ($P < 3.46 \times 10^{-11}$) and axon guidance reactome pathway ($P < 2.22 \times 10^{-4}$). These categories are consistent with a role for these genes in neuronal connectivity and function as suggested in other studies[28]. **Supplementary Table 12** gives the functional genetic correlations and associated $P$-values for all six brains.

## DISCUSSION

While a great deal of human brain genomic research focuses on individual variation to understand the genetic underpinnings of neurological and neuropsychiatric disease, less emphasis has been placed on understanding the core, canonical brain transcriptional regulatory architecture common to all humans. Using the rich microarray profiling data set in the Allen Human Brain Atlas (http://human.brain-map.org/), we found that many genes showed highly consistent patterns of transcriptional regulation across brain regions as quantified using the DS metric. Those genes with the most consistent anatomical patterning across individuals appeared to be particularly significant for brain function and disease. Remarkably, the genes in the top decile of DS brain-wide were substantially overrepresented for known drug targets and highly significantly enriched for functional annotations of many cellular functions and biological pathways. This suggests that these genes represent a functionally critical set whose transcriptional regulation is tightly controlled. While we recognize there may remain ascertainment bias in the selection of regional transcriptional patterns insofar as they may be preferentially annotated in available studies, these findings appear relevant, as we found that genes with nonspecific patterning or ubiquitous expression tended to represent more basal cellular functions. Approximately 9% of this most conserved gene set has no functional annotation to date, thus representing a set of rich targets for future research into human brain function.

Taking this concept of conserved patterning from genes to gene networks, we demonstrate the existence of a relatively small (32) set of consensus coexpression gene networks that explain most (90.1%, $\rho > 0.4$)

transcriptional variation across adult brain regions. Although the anatomical sampling in this study was not cell type specific, some of these networks appear to represent coexpression in broad cell classes such as telencephalic neurons, oligodendrocytes, astrocytes and vasculature, as observed previously[3,5]. A second set of networks shows striking enrichment in specific brain regions such as the striatum, thalamus, cerebellum and substantia nigra, presumably reflecting unique and coherent expression signatures responsible for the functions of specific neuronal types in these regions. The final set of networks were more difficult to interpret, showing consistent but more complex distributed anatomical patterning and enriched in genes associated with organelles or cellular machinery such as mitochondria, the proteasome, ribonucleoprotein complexes and chromatin. These networks may represent core intracellular pathways that tend to be coexpressed and also differentially enriched in particular brain regions. A small set of highly patterned and consistent genes were not correlated to these consensus networks but rather showed unique patterns. Surprisingly, many of those genes are well studied as markers of specific cell types (for example, *VIP* and *SST* for GABAergic interneurons) or developmental zones (homeobox genes) and are not representative of the patterning of most genes.

Previous studies of the adult and developing brain have applied similar network methods to understand coexpression relationships in normal or disease states. Oldham *et al.* identified sets of modules in normal human cortex, caudate nucleus and cerebellum reflecting the composition of these regions (that is, different cell types and cellular processes)[5]. More recent studies examined several brain regions across development. For example, Kang *et al.* created a network using data from six brain structures across pre- and postnatal development and found modules related to different spatiotemporal profiles[4], while Voineagu *et al.* studied changes in expression modules in autism[7]. In our initial assessment of the first two brains in the Allen Human Brain Atlas[3], we identified groups of genes related both to main cell types (for example, astrocytes, choroid plexus, etc.) and to specific brain regions (for example, striatum, cerebellum). This study builds on and complements our initial results in two ways: first, we took steps to equalize sampling depth across brain regions by decreasing representation of cortex and cerebellum in the analysis, and second, we perform a consensus network analysis, which exclusively identifies common patterns across brains. Despite similarities in data type and methodology, these differences in analysis strategies and experimental design make direct comparisons with previous studies challenging. We also note that DS provides complementary information from the network-based gene correlations ($k_{ME}$) presented here and in previous studies, and that both are valuable for understanding brain function (**Supplementary Analysis**).

Many differences in gene regulation across species have been previously described[36–39]. For example, a study of ~700 genes in human versus mouse cortex by *in situ* hybridization found approximately 25% of genes to have at least subtle differences in areal or cellular patterning[39], and the frequency of differences appears to vary as a function of evolutionary distance[37]. Here as well, we found that the different coexpression networks varied in their conservation between species. In general, both the coexpression statistics and anatomical patterning of neuron-associated modules were better preserved than glia-associated modules, a result similar to findings from a meta-analysis of microarray studies that showed better preservation of neuronal gene coexpression compared to that of glial and disease-related genes[36]. Why glial expression would be less well conserved is unclear, although major differences in glial structure and function between human and mice have been described[38]. One possible explanation is that result may reflect the difference between clear anatomical parcellation of neuronal types into specific nuclei in the two species compared to more widely distributed patterning of glia; therefore, differences in glial distribution across species would make glial modules appear less conserved. However, this does not seem a likely explanation for this difference. First, glial modules are among the most stable and robust in the brain despite the lack of strict anatomical segregation. Furthermore, this lack of conservation for glial modules was observed both anatomically and also in the within-module correlation of glial genes, which is independent of anatomical distribution. Notably, many individual genes shifted their global expression from one main pattern to another. This may represent a mechanism for species differentiation whereby mutations lead to marked shifts in gene regulation across brain regions.

Functional connectivity as measured using fMRI is a powerful but indirect measure of network organization that can provide high spatial resolution but should not be considered a surrogate for direct anatomical connectivity[29,30]. The canonical transcriptional architecture of the human brain may in some respects be considered a genetic analog to group average anatomical or functional architecture revealed by MRI[40–42]. As before[28], we report here a significant correlation between the transcriptional patterning in the neocortex and functional connectivity (as manifested by the partial correlation functional connectivity matrix), and we obtained this result with an independent imaging data set (the Human Connectome Project). This correlation was only evident for genes having consistent cortical patterning across individuals (that is, high DS in cortex), in part because the high cortical DS genes showed more variation across cortex and in part because these genes represent the consistent transcriptional features. There might be a causal relationship, as suggested for mice, in which genes associated with neuronal connectivity are particularly likely to also be correlated with long-range connectivity patterns[43,44]. Exploring the nature of such relationships will be a worthwhile avenue of research, facilitated by our identification of many genes involved.

Transcriptional regulation patterns are informative about human brain organization at multiple scales. The DS metric, when applied only to neocortex, where transcriptional variation is much less than between major brain regions, identified genes that varied consistently across cortical regions, and these genes were preferentially correlated with resting-state functional cortical architecture. Since many of the core coexpression networks described here at mesoscale appear to reflect the underlying cellular architecture, it is likely that discrete cell types similarly display highly conserved transcriptional regulation critical to cellular function. The recent explosion of cell-type and single-cell profiling techniques, coupled with the emphasis on cellular characterization through the BRAIN Initiative and other efforts, will provide the means to explore the canonical cell type architecture of normal and diseased human brains in the near future.

## METHODS

Methods and any associated references are available in the online version of the paper.

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

**ACKNOWLEDGMENTS**

The authors thank the Allen Institute for Brain Science founders, Paul G. Allen and Jody Allen, for their vision, encouragement and support. Research was supported by the

Allen Institute for Brain Science. We also gratefully acknowledge support from the US National Institute of Drug Abuse, grant 4R33DA027644; D. Wall of Stanford University School of Medicine; and 1U54MH091657 (NIH Blueprint for Neuroscience Research).

## AUTHOR CONTRIBUTIONS

M.H., J.A.M. and V.M. performed the primary analyses, with supporting analyses by A.L.G.-B., F.C., K.A.B., P.G., Z.Y., L.S., A.-L.B. and J.S. Graphics and networks analysis were done by D.F., T.D., L.N., C.D. and J.A.M. Annotation analysis was done by A.G.J. and B.J.A. M.F.G. and D.L.D. performed resting state analysis with V.M., S.M. and D.C.V.E. Data processing and normalization were done by C.-K.L., L.N., C.D., A.B., J.P., A.S. and M.H. J.A.M., D.R.H., D.C.V.E., A.J., C.K. and E.L. wrote the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.




1. Manolio, T.A. & Collins, F.S. The HapMap and genome-wide association studies in diagnosis and therapy. *Annu. Rev. Med.* **60**, 443–456 (2009).
2. McCarthy, M.I. & Hirschhorn, J.N. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* **17**, R156–R165 (2008).
3. Hawrylycz, M.J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391–399 (2012).
4. Kang, H.J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
5. Oldham, M.C. *et al.* Functional organization of the transcriptome in human brain. *Nat. Neurosci.* **11**, 1271–1282 (2008).
6. Roth, R.B. *et al.* Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics* **7**, 67–80 (2006).
7. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
8. Shaw, G.T., Shih, E.S., Chen, C.H. & Hwang, M.J. Preservation of ranking order in the expression of human Housekeeping genes. *PLoS ONE* **6**, e29314 (2011).
9. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
10. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, 17 (2005).
11. Miller, J.A. *et al.* Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
12. Oldham, M.C., Horvath, S. & Geschwind, D.H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. USA* **103**, 17973–17978 (2006).
13. Ponomarev, I., Wang, S., Zhang, L., Harris, R.A. & Mayfield, R.D. Gene coexpression networks in human brain identify epigenetic modifications in alcohol dependence. *J. Neurosci.* **32**, 1884–1897 (2012).
14. Bernard, A. *et al.* Transcriptional architecture of the primate neocortex. *Neuron* **73**, 1083–1099 (2012).
15. Bray, N.J. *et al.* Screening the human protocadherin 8 (PCDH8) gene in schizophrenia. *Genes Brain Behav.* **1**, 187–191 (2002).
16. Geschwind, D.H. & Konopka, G. Neuroscience in the era of functional genomics and systems biology. *Nature* **461**, 908–915 (2009).
17. Konopka, G. *et al.* Human-specific transcriptional networks in the brain. *Neuron* **75**, 601–617 (2012).
18. Hanashima, C., Fernandes, M., Hebert, J.M. & Fishell, G. The role of Foxg1 and dorsal midline signaling in the generation of Cajal-Retzius subtypes. *J. Neurosci.* **27**, 11103–11111 (2007).
19. Siegenthaler, J.A. & Miller, M.W. Generation of Cajal-Retzius neurons in mouse forebrain is regulated by transforming growth factor beta-Fox signaling pathways. *Dev. Biol.* **313**, 35–46 (2008).
20. Zapala, M.A. *et al.* Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc. Natl. Acad. Sci. USA* **102**, 10357–10362 (2005).
21. Nelson, T.H. *et al.* Autworks: a cross-disease network biology application for autism and related disorders. *BMC Med. Genomics* **5**, 56 (2012).
22. Langfelder, P., Luo, R., Oldham, M. & Horvath, S. Is my network module preserved and reproducible? *PLoS Comput. Biol.* **7**, e1001057 (2011).
23. Cahoy, J. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
24. Chen, J., Bardes, E.E., Aronow, B.J. & Jegga, A.G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37**, W305–W311 (2009).
25. Gratacos, M. *et al.* Identification of new putative susceptibility genes for several psychiatric disorders by association analysis of regulatory and non-synonymous SNPs of 306 genes involved in neurotransmission and neurodevelopment. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **150B**, 808–816 (2009).
26. Pinheiro, A.P. *et al.* Association study of 182 candidate genes in anorexia nervosa. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **153B**, 1070–1080 (2010).
27. Lein, E.S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
28. Richiardi, J. *et al.* Correlated gene expression supports synchronous activity in brain networks. *Science* **348**, 1241–1244 (2015).
29. Van Essen, D.C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
30. Smith, S.M. *et al.* Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17**, 666–682 (2013).
31. Yeo, B.T. *et al.* The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
32. De Bacco, F. *et al.* The MET oncogene is a functional marker of a glioblastoma stem cell subtype. *Cancer Res.* **72**, 4537–4550 (2012).
33. Hosokawa, M., Klegeris, A., Maguire, J. & McGeer, P.L. Expression of complement messenger RNAs and proteins by human oligodendroglial cells. *Glia* **42**, 417–423 (2003).
34. Rosenmann, H. *et al.* A polymorphism in the complement component C1r is not associated with sporadic Alzheimer's disease. *Neurosci. Lett.* **336**, 101–104 (2003).
35. Walker, D.G., Dalsing-Hernandez, J.E. & Lue, L.F. Human postmortem brain-derived cerebrovascular smooth muscle cells express all genes of the classical complement pathway: a potential mechanism for vascular damage in cerebral amyloid angiopathy and Alzheimer's disease. *Microvasc. Res.* **75**, 411–419 (2008).
36. 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).
37. Morris, J.A. *et al.* Divergent and nonuniform gene expression patterns in mouse brain. *Proc. Natl. Acad. Sci. USA* **107**, 19049–19054 (2010).
38. Oberheim, N.A., Wang, X., Goldman, S. & Nedergaard, M. Astrocytic complexity distinguishes the human brain. *Trends Neurosci.* **29**, 547–553 (2006).
39. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483–496 (2012).
40. Damoiseaux, J.S. *et al.* Consistent resting-state networks across healthy subjects. *Proc. Natl. Acad. Sci. USA* **103**, 13848–13853 (2006).
41. Greicius, M.D., Krasnow, B., Reiss, A.L. & Menon, V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* **100**, 253–258 (2003).
42. Richiardi, J. *et al.* Correlated gene expression supports synchronous activity in brain networks. *Science* **348**, 1241–1244 (2015).
43. French, L. & Pavlidis, P. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Comput. Biol.* **7**, e1001049 (2011).
44. French, L., Tan, P.P. & Pavlidis, P. Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data integration. *Front. Neuroinform.* **5**, 12 (2011).
45. Smith, S.M. *et al.* Resting-state fMRI in the Human Connectome Project. *Neuroimage* **80**, 144–168 (2013).

## ONLINE METHODS

**AHBA data and normalization.** The study used the complete data set of six human brains from the Allen Human Brain Atlas (http://human.brain-map.org/). Briefly, approximately 500 anatomically discrete samples were collected from cortex, subcortex, cerebellum and brainstem of each brain and profiled for genome-wide gene expression using a custom Agilent 8 × 60K cDNA array chip. Two methods were used to dissect samples: a scalpel-based manual macrodissection method primarily for cortical and other relatively large, uniform samples and laser microdissection for small or oddly shaped structures such as subcortical or brainstem areas. Over the multiyear course of the project, the six brains were processed serially (that is, expression profiles for the first brain were completed before profiling the second brain), and multiple batches of samples were submitted per brain. Because of the complexity and timing of the entire process, assessing and minimizing nonbiological systematic biases among different batches within a brain and across all brains while keeping biological variations was crucial to the data analysis. The same control samples were included in all batches to serve this purpose. Within each batch, array-specific biases characterized by probe GC content, location in the chip and experiment-wise mean intensity were corrected. The 75th-percentile expression levels of all samples in each batch were then aligned. Within a brain, cross-batch correction was done so that the mean expression levels of control samples in the batches could be aligned. Adjustment was made for the bias introduced in sample quality and expression level by different dissection methods. Across brains, brain-wise mean expressions as well as control samples' mean expression were aligned. More details on the microarray data normalization are available at http://help.brain-map.org/display/humanbrain/documentation/.

**Annotation.** In addition to the Allen Human Brain Atlas and Mouse Atlas databases, the ToppGene (https://toppgene.cchmc.org/) portal[24] was used to access an extensive list of databases and to calculate *P*-values from hypergeometric tests, corrected for multiple comparisons. Those used in this study included the following: gene ontology (GO) annotations (biological process, cellular component and molecular function), pathway annotations (Biosystems, BIOCYC, KEGG and REACTOME), miRNA targets (the full set indicated on https://toppgene.cchmc.org/navigation/database.jsp, including MicroRNA.org, Pic Tar and TargetScan), computational cancer-related gene sets (from Broad MSigDb; http://www.broadinstitute.org/gsea/msigdb/collections.jsp), and drug annotations (from DrugBank, Comparative Toxicogenomics Database, including marker and therapeutic, and Broad Institute CMAP). Significant genes were identified using the ToppFun application. For disease annotations, we used the Autworks database (http://autworks.hms.harvard.edu/), listing genes associated with 2,288 disease states, including autism. Gene sets from the Simons Foundation Autism Research Initiative Database (https://gene.sfari.org/autdb/Welcome.do) were also compared for autism relevance. Alzheimer's disease–related genes were retrieved from the Alzheimer's Forum (http://www.alzforum.org/).

**Correlation metrics and differential stability.** Average Pearson correlation over 15 pairs of 6 brains was used for measuring differential stability between brains and networks. For each gene $g$ expressing in a pair of brains $B_i(g)$, $B_j(g)$ with common anatomic structures $S$, we computed the Pearson correlation $\rho_s[B_i(g), B_j(g)]$. The differential stability of $\Delta_S(g)$ was defined as the average $\rho_S$ over ($n = 15$) pairs of brains. Pearson correlation was used for comparing genes to modules[5]. The R programming language (http://www.r-project.org/) was used for statistical analysis throughout. See **Supplementary Analysis** for comparable metrics.

**WGCNA consensus network construction.** The general framework of WGCNA has been described in detail previously[10], and it has been packaged into a user-friendly R library[9]. In short, Pearson's correlations were calculated for all pairs of genes, and then a signed similarity ($S_{ij}$) parameter was derived: $S_{ij} = [1 + \mathrm{cor}(x_i,x_j)]/2$, where gene expression profiles $x_i$ and $x_j$ consist of the expression of genes $i$ and $j$ across multiple microarray samples. In the signed network, the similarity between genes reflects the sign of the correlation of their expression profiles. The signed similarity ($S_{ij}$) was then raised to power β to represent the connection strength ($a_{ij}$): $a_{ij} = S_{ij}^{\beta}$ (here $\beta = 14$). This step aims to emphasize strong correlations and reduce the emphasis of weak correlations on an exponential scale. The resulting adjacency matrix is then transformed into a topological overlap matrix, defined as $\mathrm{TOM}_{ij} = (\Sigma_u a_{iu}a_{uj} + a_{ij})/(\min(k_i,k_j) + 1 - a_{ij})$, that essentially assesses to what extent each pair of genes has common expression patterns with all other genes in the network. Topological overlap has been shown to be a more biologically relevant metric than adjacency in many cases[10].

Here we constructed a signed network for each brain, summarizing gene expression across the subset of 132 broad brain structures sampled in each of six adult human brains (as described below). Since we are primarily interested in finding expression patterns conserved between brains, these networks included only the 50% of genes with highest differential stability score ($N = 8{,}674$, DS > 0.5284). From these six individual brain networks, we then created a consensus network to identify common expression patterns across individuals, following the published method[22] with a few modifications. Specifically, the consensus network was created by calculating the component-wise minimum values for topologic overlap (TO) across all six brains, after appropriately scaling each individual network. Genes were hierarchically clustered using 1 − TO as the distance measure, and initial module assignments were determined by using a dynamic tree-cutting algorithm (cutreeHybrid, using default parameters except deepSplit = 4, cutHeight = 0.999, minClusterSize = 30 and pamStage = FALSE)[46].

Modules were then iteratively merged until all pairs of module eigengenes (ME; that is, the first principal component of the module) were correlated with $R < 0.8$, using the following strategy: (1) calculate the MEs; (2) identify the most highly correlated pair of MEs; (3) if $R > 0.8$ merge these two modules and repeat steps 1–3. Next, modules were assessed for coherency and incoherent modules were removed from consideration using the following strategy: (1) calculate the MEs; (2) calculate the Pearson correlation between each gene and each ME—referred to as a gene's module membership ($k_{\mathrm{ME}}$)[5]—and reassign each gene to the module to which it is most highly correlated; (3) find the module for which the largest number of genes were reassigned; (4) if <70% of the genes in this module remain in the module after reassignment, remove the module and repeat steps 1–4. Finally, after generation, merging and deletion of modules, all 17,348 expressed genes were assigned to modules based on $k_{\mathrm{ME}}$ (as above), and genes with $k_{\mathrm{ME}} < 0.4$ against all modules were considered unassigned. Module eigengenes were calculated using all available data, but for brevity visualizations include only the 62 structures present in all six brains (**Supplementary Table 1**) in addition to CGS, CPLV and Pa. This method is entirely automated, resulting in an unbiased collection of distinct yet well-defined modules. This same strategy was used to construct networks and define modules for the 50% of genes with the lowest differential stability and for the leave-one-out analysis. Alternative network strategies are discussed in detail in the **Supplementary Analysis**.

**Module enrichment analysis.** Module enrichment analysis was performed using the userListEnrichment R function in the WGCNA library[47]. These tests included enrichment for gene ontology categories, brain cell types, brain regions, disease classes and results from previous network analyses, with primary references cited in the main text as appropriate. All *P*-values in this analysis were Bonferroni corrected for multiple hypergeometric test comparisons. The ToppGene facility (https://toppgene.cchmc.org/)[24] was used for access to annotation databases as discussed above.

**Gene expression quantification in mouse.** Expression levels for mouse *in situ* hybridization data from the Allen Mouse Brain Atlas were quantified using a metric called expression energy (fraction of stained volume × average intensity of stain), as described previously[18,27]. Coronal data for 2,651 genes with expression in human and reproducible coronal and sagittal series in mouse ($\rho > 0.7$) were included for comparison between species. For each of 53 human structures, average expression energy was calculated across all voxels in matched mouse structures. These values were treated the same way as microarray intensities for between-species correlations and network statistics. Genes in mouse were assigned to the same modules as their human orthologs.

**Comparison with mouse *in situ* hybridization data.** The Allen Mouse Brain Atlas has a comprehensive informatics pipeline (see Informatics Data Processing, http://help.brain-map.org/display/mousebrain/Documentation/). While this data set contains replicates for only about 25% of genes, there are 2,651 genes which both had highly correlated expression patterns ($\rho > 0.7$) in the voxel-mapped coronal and sagittal series in mouse ISH data and could also be mapped using the orthologous gene symbols in mouse and human, as mentioned above. In cases where gene symbol matching was ambiguous, the gene was omitted from




doi:10.1038/nn.4171

the analysis. We mapped regions of the human ontology to analogous regions of the mouse (**Supplementary Table 1**). In total 62 structures were represented in all six human brains, and of those 53 had regions that could be defined in the voxel-based mouse ISH data. In some cases this was due to species differences, where a human structure was not present in mouse, and in other cases it was due to technical details as a result of the level of annotation in mouse or human. In human, the networks were created by using all available regions in each brain (with contraction of cortex and cerebellum as noted above; ~100 regions per brain) but then visualized using only 65 structures (as discussed earlier). In mouse, the data were processed by finding the average expression energy (see web documentation above) of the 250 most highly expressed voxels in each of the 53 regions (or all voxels if fewer than 250 in a region) and then treating the resulting 2,651 × 53 matrix as a microarray matrix. Modules eigengenes were calculated by first assigning to each mouse gene the corresponding module label from the human high-DS network and then calculating the MEs by applying the moduleEigengenes function in the WGCNA library to the mouse expression energy matrix. The resulting MEs in mouse are compared with the corresponding values in human as discussed in the main text.

We note that, although we used all available mouse data, these 2,651 genes were initially chosen to be run in coronal section because they were potentially biologically interesting in some way, and they therefore represent a biased set of genes. This said, these genes were distributed across all of the human modules and therefore represent a fairly complete set of expression patterns at the whole-brain level. While the specific proportions of genes that agree between species may not be representative of the transcriptome as a whole, the assignment of genes as either conserved or not conserved in the text should accurately reflect their actual conservation on the basis of these two data sets.

**Functional connectivity.** Resting-state functional connectivity was analyzed using data from the Human Connectome Project (HCP; http://www.humanconnectome.org/documentation/S500/). A group-average functional connectivity matrix ('dense connectome') was generated from 468 subjects with complete 1-h resting-state fMRI data sets (see below). A cortical left-hemisphere parcellation was generated using the 17-network resting-state-network parcellation of ref. 31, subdivided into 52 spatially contiguous left-hemisphere parcels, each at least 50 $mm^2$ in cortical surface area. Parcellated connectomes were based on resting-state data from 447 HCP subjects using the 52 contiguous left hemisphere parcels and using FSLNets (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets/) to estimate partial correlation parcellated connectomes (see below).

**Differential stability in cortex and resting state network analysis.** The resting-state functional connectivity analysis was based on data from the Human Connectome Project (http://www.humanconnectome.org/documentation/S500/). To generate the functional connectivity map shown in **Figure 7a**, a group-average functional connectivity matrix (dense functional connectome) was generated from 468 subjects having complete resting-state fMRI data sets. Each of four 15-min resting state runs for each subject in the group was cleaned (denoised) using 24-parameter motion regression followed by ICA+FIX denoising[30,48]. The resulting cleaned runs were combined across the 468 related subjects, respectively, applying variance normalization of the time series using the same approach as MELODIC[49], followed by a group-PCA approach (MIGP: MELODIC's incremental group PCA) that approximates full temporal concatenation of all subjects' data. PCA was then used to output the strongest 4,500 spatial eigenvectors (PCA components, weighted by the eigenvalues). The spatial eigenvectors of each grayordinate (surface vertex or subcortical voxel) were correlated with those of every other grayordinate to produce a dense connectome.

To generate a parcellation covering the entire left hemisphere with spatially contiguous parcels closely related to functional organization, we started with the 17-network resting-state-network parcellation of Yeo *et al.*[31]. This was subdivided into 52 spatially contiguous left-hemisphere parcels, each at least 50 $mm^2$ in cortical surface area (based on the average surface area of each tile in standard-mesh cortical midthickness surfaces in a group of 196 HCP subjects). To generate the parcellated connectomes that are illustrated in **Figure 7a** (right panel) and were used for the analyses in **Figure 7c**,**d**, we used resting-state data from 447 HCP subjects having complete resting-state (and also task fMRI) scans and used the aforementioned 52 contiguous parcels of the left hemisphere. Cleaned (as above) resting state time series were averaged across the grayordinates within each left hemisphere parcel. FSLNets (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets/) was used to estimate partial correlation parcellated connectomes (52 × 52) in each resting state run of each subject, and these matrices were converted to *z*-scores. The *z*-scores were averaged across runs, and a one-sample *t*-test was performed across subjects (also converted to *z*-scores) to produce a group partial correlation parcellated connectome.

The next step was to map the Allen Human Brain Atlas (AHBA) tissue samples to the HCP 52 region parcellation so that comparison could be made. Using the MNI centroid coordinate of the AHBA samples, and by manually examining each of the AHBA tissue samples using the online tools, one can assign a set of HCP space voxels to each AHBA tissue sample. As each of the 52 parcels is composed of a set of voxels, we now have potentially one-to-many map from AHBA tissue to HCP parcels. If all ABHA tissue samples belong to a common HCP parcel, we average the gene expression of that tissue in the corresponding parcel. However, some of the 52 parcels represent smaller regions of the brain and therefore there is no unique assignment of AHBA gene expression tissue samples to that region. Therefore, if a collection of AHBA tissue samples intersects more than one region, we average the gene expression values as before but fractionally weight the expression contribution to each of the interesting HCP parcels. This has the effect of allowing some assignment of expression without overweighting non-unique samples. **Supplementary Table 12** gives the sample distribution by parcels as well as the uniquely assigned samples.

To obtain the expression correlation matrix for a given gene (**Fig. 7c**, right panel), we transformed the expression values of that gene into *z*-scores over all the sampled brain regions (averaging sample data for those samples contained in the same parcel) and calculated the coexpression as the outer product of this *z*-score vector. Thus, if two regions both show high expression or low expression of the gene of interest, they will have a high positive coexpression value for that gene, whereas if they show opposite expression patterns, they will have a large negative value for that gene. After generating these matrices, we compared each of the 17,348 gene coexpression matrices to the parcellated connectome matrix by calculating the Pearson's correlation between the vectorized elements above the diagonal of the matrices (**Fig. 7d**). We also obtained a significance value for each gene-connectivity comparison using the randomized gene coexpression matrices. **Supplementary Table 12** gives the complete distribution of tissue samples by HCP parcel for the 52 regions and the functional genetic correlations and *P*-values.

**Statistics of the methods.** Standard methods are presented without detailed description. Less standard methods are described in the main text, in the Online Methods or through citations to references where they are described in detail. To the best of our knowledge, statistical tests were chosen such that the data meet the assumptions of the specific statistical test.

A **Supplementary Methods Checklist** is available. Data are available in **Supplementary Data Sets 1** and **2**.


46. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
47. Miller, J.A. *et al.* Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics* **12**, 322 (2011).
48. Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449–468 (2014).
49. Beckmann, C.F. & Smith, S.M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* **23**, 137–152 (2004).