Genetic influences on hub connectivity of the human connectome

Aurina Arnatkevičiūtė[1✉], Ben D. Fulcher*[1,2], Stuart Oldham*[1], Jeggan Tiego[1], Casey Paquola[4], Zachary Gerring[3], Kevin Aquino[1], Ziarih Hawi[1], Beth Johnson[1], Gareth Ball[5], Marieke Klein[6,7], Gustavo Deco[1,8,9,10,11], Barbara Franke[6,12], Mark Bellgrove[1], and Alex Fornito[1]

[*]Contributed equally to this work
[1]The Turner Institute for Brain and Mental Health, School of Psychological Sciences, and Monash Biomedical Imaging, Monash University, Victoria, Australia.
[2]School of Physics, The University of Sydney, Camperdown, NSW 2006, Australia.
[3]Translational Neurogenomics Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia.
[4]McConnell Brain Imaging Centre, Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada.
[5]Murdoch Children's Research Institute, Melbourne, Victoria, Australia.
[6]Department of Human Genetics, Radboud University Medical Center, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[7]Department of Psychiatry, University of California San Diego, La Jolla, CA, USA
[8]Center for Brain and Cognition, Computational Neuroscience Group, Department of Information and Communication Technologies
[9]Universitat Pompeu Fabra, Roc Boronat 138, Barcelona, 08018, Spain
[10]Institució Catalana de la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, 08010, Spain
[11]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany
[12]Department of Psychiatry, Radboud University Medical Center, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

Brain network hubs are both highly connected and highly inter-connected, forming a critical communication backbone for coherent neural dynamics. The mechanisms driving this organization are poorly understood. Using diffusion-weighted imaging in twins, we identify a major role for genes in shaping hub connectivity of the human connectome, showing that genes preferentially influence connectivity strength between network hubs. In two independent samples, we show that DNA variants preferentially related to hub connectivity are expression quantitative trait loci for genes that overlap with those implicated in intelligence, schizophrenia, and metabolism. Using transcriptomic atlas data, we show that connected hubs demonstrate tight coupling of transcriptional activity related to metabolic and cytoarchitectonic similarity. Finally, comparing thirteen generative models of network growth, we show that stochastic processes cannot explain the spatial distribution, and thus the precise wiring pattern, of hub connectivity. Together, our findings indicate that genetic influences on brain connectivity are not uniformly distributed throughout the brain, but are instead concentrated on the functionally valuable, metabolically costly connections between connectome hubs.

gene expression | connectome | hub | heritability | transcriptome | GWAS | MRI

Correspondence: ***aurina.arnatkeviciute@monash.edu.au***

Introduction

Nervous systems are intricately connected networks with complex wiring patterns that are neither completely random nor completely ordered (1, 2). Numerous studies, conducted in species as diverse as the nematode worm *Caenorhabditis elegans*, mouse, macaque, and human, and at scales ranging from the cellular to the macroscopic, have shown that this complex organization is, in part, attributable to a heterogeneous distribution of connectivity across neural elements, such that a large fraction of network connections are concentrated on a small subset of network nodes called hubs (3–7). These hubs are more strongly interconnected with each other than expected by chance, forming a rich-club (3–5, 7) that is topologically positioned to integrate functionally diverse neural systems and mediate a large proportion of inter-regional communication (5, 8).

In human cortex, hubs are predominantly located in multimodal association areas (6, 9) and are among the most metabolically expensive elements of the connectome (10), with rich-club connections between hubs accounting for a disproportionate fraction of axonal wiring costs (3–5, 7, 11). Association hubs of the human brain also show marked inter-individual variability in connectivity and function that relates to a diverse array of behaviors (6, 12–14). These brain regions are disproportionately expanded in individuals with larger brains (15) and in human compared to non-human primates (16). They also show greater topological centrality and evolutionary divergence in the human connectome when compared to chimpanzee (17). These findings support the view that rapid expansion of multimodal association hubs, and the costly, valuable rich-club connections between them, underlies the enhanced cognitive capacity of humans compared to other species (18).

What influences the way in which hub regions connect to each other? The rapid evolution of network hubs in humans, coupled with evidence supporting the heritability of many different aspects of brain organization (19), suggests an important role for genes. In the developing brain, neurons can innervate precise targets, even over long anatomical distances, by following genetically regulated molecular cues (20, 21). However, it is unknown whether genetic influences are preferentially exerted across specific classes of connections, such as the costly and functionally valuable links between network hubs. Preliminary evidence from human twin re-

search suggests that certain properties of hub functional connectivity are strongly heritable (22), and analyses of *C. elegans*, mouse, and human data suggest that hub connectivity is associated with a distinct transcriptional signature related to metabolic function (7, 11, 23–25). Alternatively, some have suggested that the protracted maturation of hub regions (16, 26, 27) may endow these areas with enhanced plasticity (12), suggesting a prominent role for environmental influences. Moreover, recent computational models of whole-brain connectome wiring suggest that it is possible to grow networks with complex topological properties, including hubs, that mimic actual brains using simple, stochastic wiring rules based on geometric constraints (28–31) or trade-offs between the wiring cost and functional value of a connection (32–34). These findings imply that the emergence of network hubs may not require precise genetic control and may instead result from random processes shaped by generic physical and/or functional properties.

Here we use a multifaceted strategy to test between these competing views and characterize genetic influences on hub connectivity of the human cortical connectome. First, using a connectome-wide heritability analysis (Fig. 1A-B), we show that genetic influences on phenotypic variance in connectivity strength are not distributed homogeneously throughout the brain, but are instead preferentially concentrated on links between network hubs. Second, through analysis of genome-wide single nucleotide polymorphisms (SNPs; Fig. 1C), we demonstrate that variants preferentially related to hub connectivity are expression quantitative trait loci (eQTLs) for genes implicated in risk for schizophrenia, intelligence, and energy metabolism. Third, as previously demonstrated in *C. elegans* (11) and mouse (7), we show that connected pairs of hubs in the human brain exhibit tightly coupled gene expression related to the metabolic demand and cytoarchitectonic similarity of these areas (Fig. 1D). Finally, we use computational modeling to show that stochastic network wiring models can indeed generate networks with brain-like properties, but they fail to capture the spatial distribution of hub regions and, by extension, the precise pattern of wiring between network hubs.

Collectively, these findings demonstrate a direct link between molecular function and the large-scale network organization of the human connectome and highlight a prominent role for genes in shaping the costly and functionally valuable connections between network hubs.

Results

Using diffusion weighted imaging (DWI) data for 972 subjects acquired through the Human Connectome Project (HCP) (36) we generated a representative group-level connectivity matrix [see Online methods] containing 12 924 unique connections between 360 brain regions defined by the HCPMMP1 atlas (37). This network contains a set of highly connected regions, quantified using the measure of node degree (k), which represent network hubs and which span sensorimotor, paracentral, mid-cingulate ($k > 105$), insula, posterior cingulate, lateral parietal, and dorsolateral prefrontal cortices ($k > 145$) (Fig. 2A). As shown previously (5, 6, 8, 10), the network also exhibits rich-club organization, with hubs being more densely and strongly interconnected than expected by chance (Fig. S1). Rich-club connections also have higher average wiring cost and communicability (Fig. S1), indicating that they are amongst the most topologically central and costly elements of the connectome.

Genetic influences on brain connectivity are concentrated in the rich club. To investigate whether genes preferentially influence certain classes of connections in the human brain, we performed a connectome-wide heritability analysis of twin data acquired through the Human Connectome Project. For each of 234 monozygotic (MZ) twins and their 69 non-twin siblings as well as 120 dizygotic (DZ) twins 48 of their non-twin siblings, we reconstructed macroscale cortical connectomes using DWI [see Online methods].

For each connection in the representative group connectome, we used the classic ACTE model to estimate the proportion of variance in connectivity strength that is attributable to additive genetic factors (narrow-sense heritability, annotated as h^2, see Online methods).

Using the average fractional anisotropy (FA) of each connecting fibre bundle to quantify connectivity strength [see Online methods], we observe a wide range of heritability estimates across connections, spanning 0 to 0.99 ($h^2{}_{\text{mean}} = 0.45$, $h^2{}_{\text{SD}} = 0.2$). Non-trivial genetic influences, quantified using the A component of the ACTE model, were observed for the majority of connections, with the AE model showing the best fit for 86.7% of edges, ACTE for 4.3%, and ACE for 1.3%. A total of 7.7% of connections were influenced only by environmental factors (CE model 6.8%, E model 0.9%).

To examine whether genetic influences are preferentially concentrated on specific types of inter-regional connections, we distinguish between hub and nonhub regions, resulting in three possible types of connections: rich (hub-to-hub), feeder (between a hub and a nonhub), and peripheral (nonhub-to-nonhub) links [see schematic in Fig. 1A, (38)].

We find that mean heritability derived from the best-fitting biometric model is highest for rich, intermediate for feeder, and lowest for peripheral connections across nearly all values of k (Fig. 2B–C). The increase in heritability for rich links as a function of the hub-defining threshold, k, indicates that genetic influences are, on average, stronger for connections between the most highly connected brain regions (see also Fig. 1E). The same pattern is replicated when taking genetic parameters from the full ACTE model, confirming that this result is not an artefact of our model-selection procedure (Fig. S2).

Contrasting with rich links between hubs, phenotypic

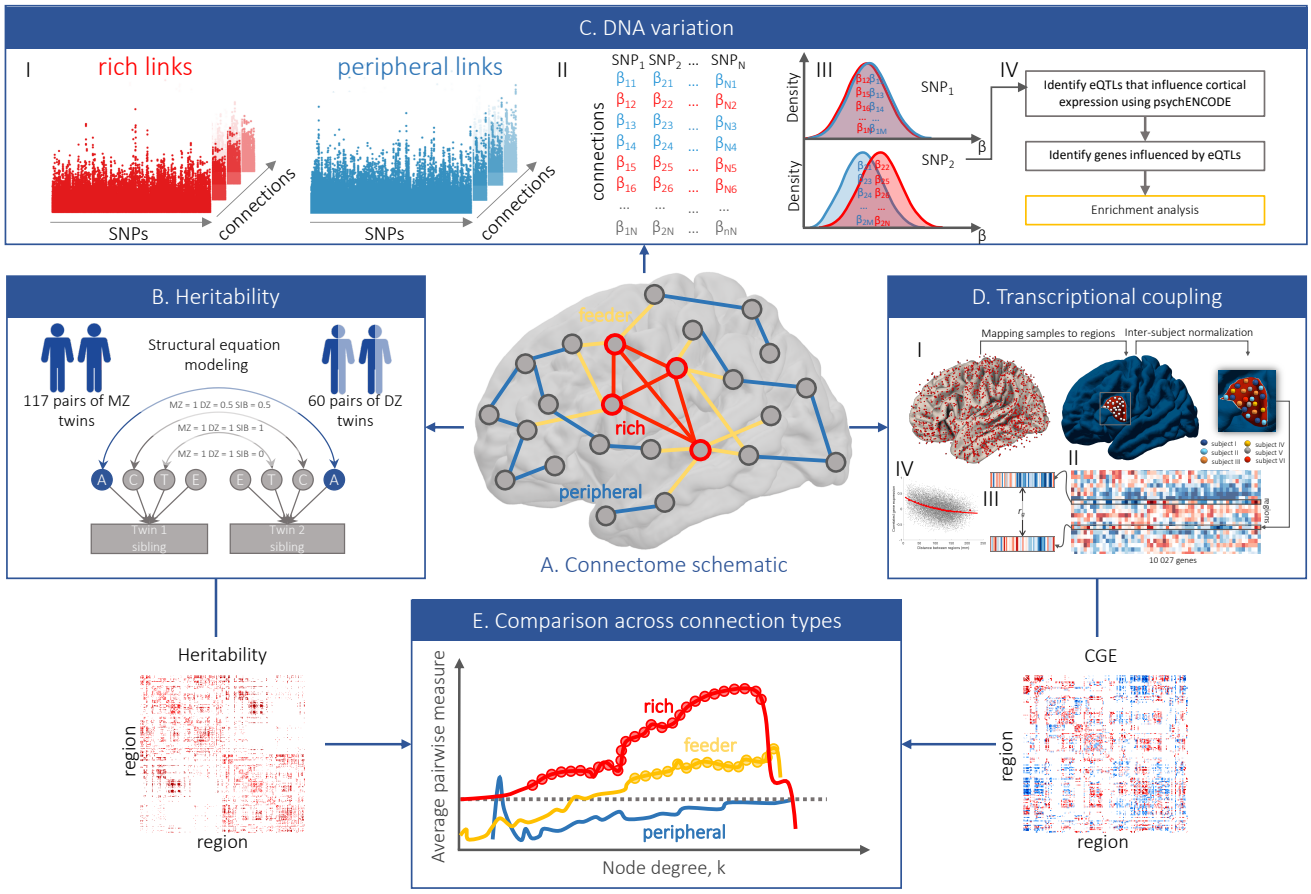

Fig. 1. Workflows used to characterize genetic influences on hub connectivity. (A) A schematic representation of the connectome showing different connection types in the brain. Given a distinction between hub nodes (red outline) and nonhub nodes (grey outline), we can distinguish three classes of connections: rich links, which are connections between two hubs (red); feeder links, which are connections between a hub and a nonhub (yellow); and peripheral links, which are connections between two nonhubs (blue). **(B)** Connectome-wide heritability analysis. We use structural equation modeling to fit a classic ACTE biometric model to every connection within the brain, resulting in estimates of genetic and environmental influences for each link. **(C)** DNA variation related to hub connectivity. (I) We perform a connectome-wide, genome-wide association analysis (GWAS) to quantify the association between polymorphic variation at each of 5 235 153 variants and connectivity strength at each connection. (II) The resulting beta estimates, quantifying the associations between connection weights and SNPs, are sorted by connection type. (III) t-tests are used to identify SNPs more strongly associated with rich than peripheral links. (IV) We identify which of a set of prioritized SNPs are eQTLs and map them to the genes whose expression they influence. We then test for enrichment of these gene lists with respect to genes implicated in intelligence, mental illness, and annotated gene categories. **(D)** Analysis of transcriptional coupling. (I) Each of 3702 tissue samples in the Allen Human Brain Atlas (AHBA) is mapped to a given region in our brain parcellation. (II) Expression values are then subjected to a quality control and processing pipeline (35) to construct a region $\times$ gene matrix of expression values. (III) We estimate correlated gene expression (CGE) between each pair of brain regions as the Pearson correlation between region-specific gene-expression profiles. (IV) Inter-regional CGE is corrected for spatial autocorrelation of the expression data via regression of the exponential distance trend (35). **(E)** Schematic representation of how values assigned to each edge are compared across connection types. We compare the mean of edge-level (pairwise) measures of heritability and CGE across rich, feeder, and peripheral links across all possible hub-defining thresholds (horizontal axis). As k increases, the definition of a hub becomes more stringent and identifies the actual hubs of the network. Thus, if a given effect is stronger for rich links, we expect the pairwise estimates to increase as a function of k, with the increase for rich links being particularly large relative to feeder and peripheral links.

variance in peripheral connectivity (between nonhubs) is predominantly influenced by unique environment (quantified by the model parameter E, Fig. 2D), while common environmental influences are consistently low across all link types (mean values $\langle C \rangle < 0.08$ and $\langle T \rangle < 0.02$ across all k thresholds). Critically, we obtain similar evidence of preferential genetic influences on hub connectivity when using different methods for parcellating or thresholding our connectomes (Fig. S3) or when evaluating connection strength based on the number of reconstructed streamlines (streamline count, SC) between regions (Fig. S4).

To investigate whether genetic influences are specific to certain functional systems of the brain, we next categorize edges according to the major functional networks that they connect, as defined using a network parcellation (39) of the HCPMMP1 atlas (37) (Fig. 2E). Figure 2F shows the proportion of nodes with degree $> k$ in each functional network. High-degree nodes are present in most networks until $k \approx 120$, beyond which they are predominantly found in multimodal association networks; namely, the fronto-parietal, cinguolo-opercular, and default mode systems.

Across all 12 canonical functional networks, rich links both within and between networks demonstrate significantly higher heritability than other types of connections (Fig. 2G-I, one-sided Welch's t-test, comparing heritability of rich *vs* feeder and rich *vs* peripheral connections,

all $p < 1.9 \times 10^{-12}$), suggesting that the elevated genetic influences observed for rich links cannot be explained by the affiliation of hub nodes to any specific functional network. Moreover, stronger heritability of rich links is evident across different connection distances (Fig. S5), suggesting that preferential genetic influences on hub connectivity cannot be explained simply by the longer average distance of rich links (Fig. S1D).

Together, these findings indicate that genetic influences on phenotypic variance in connectivity strength are not distributed homogeneously throughout the brain, nor are they confined to specific functional networks or long *vs* short-range connections. Instead, they are most strongly concentrated on the connections between network hubs. These hubs are distributed throughout the cortex, with the most highly connected residing in multimodal association networks.

DNA variation related to hub connectivity. We next developed a pipeline for identifying genes whose expression in cortex is influenced by eQTLs preferentially related to hub connectivity [see Online methods]. For each of 12 924 unique connections in the representative group connectome, we run a linear regression-based GWAS in a sub-sample of unrelated Caucasian subjects from the HCP dataset that have both DWI and genotyping data available ($n = 254$, Fig. 1C, I) and use the resulting β weights for each SNP to select the top 10% of variants that show a stronger influence on rich link connectivity strength compared to peripheral connectivity (all $p_{\mathrm{FDR}} < 0.05$).

We then use data from the psychENCODE database (40) for identifying, from our prioritized list of SNPs, eQTLs that influence mRNA transcript levels in prefrontal cortex [see Online methods, Fig. 1C, IV]. Note that our goal in this analysis is to use beta estimates from the GWAS to prioritize SNPs for eQTL mapping and not to conduct inference at the level of individual SNPs with genome-wide significance. We therefore do not have the same multiple comparison burden as traditional GWAS and the statistical power of this analysis is determined by the psychENCODE database. Importantly, to ensure that our results are reliable, we perform the same analysis in an independent sample of 276 unrelated Caucasians (Monash Sample, see Online methods for details).

Using the HCP sample, we find that 883 of our 8207 prioritized SNPs are significant eQTLs, and that these eQTLs influence the expression of 886 annotated genes. For the Monash dataset, we identify 674 significant eQTLs out of 6081 prioritized SNPs, influencing the expression of 644 annotated genes. A total of 96 of these genes are common across the two samples [see supplementary text S3], a statistically significant level of overlap (hypergeometric test, $p = 4 \times 10^{-35}$).

Given the proposed role of rich links in supporting complex, integrated brain function (10, 38, 41) and in mediating diverse behaviors and disorders (6, 12–14), we first test whether the lists of genes we identify as being related to hub connectivity overlap with genes impacted by eQTLs that we identify using prior GWAS of intelligence and mental illness [see Online methods, (42–47)]. In the HCP sample, we find a statistically significant overlap between genes related to hub connectivity and intelligence (hypergeometric test, $p = 1 \times 10^{-27}$) and schizophrenia (hypergeometric test, $p = 3 \times 10^{-18}$, see supplementary text S3), but not major depression, attention-deficit hyperactivity disorder, autism spectrum disorder, or bipolar disorder (all hypergeometric tests $p > 0.008$, Bonferroni correction for $n = 6$ tests). Analysis of the Monash dataset replicated the significant overlap with genes implicated in intelligence (hypergeometric test, $p = 1 \times 10^{-11}$) and schizophrenia ($p = 3 \times 10^{-13}$), while also identifying overlap with major depressive ($p = 1 \times 10^{-7}$) and bipolar disorder ($p = 2 \times 10^{-7}$, see supplementary text S3).

In enrichment analyses using the Kyoto Encyclopedia of Genes and Genomes (KEGG) (48) and Gene Ontology (GO) (49) annotation systems [see Online methods], we find that hub connectivity genes in both the HCP and Monash datasets are over-represented in KEGG metabolic pathways (KEGG:01100) (Table S1, Table S2). We also find qualitatively similar results using the consensus list of 96 genes across both samples [see supplementary text S3 and Table S3]. Together, these findings indicate that DNA variants preferentially related to rich-club connectivity include eQTLs for genes implicated in intelligence, schizophrenia, and energy metabolism.

Transcriptional coupling is elevated between connected hubs. We now investigate the transcriptional correlates of hub connectivity using data from the Allen Human Brain Atlas (AHBA) (50), focusing on expression profiles of 10 027 genes surpassing our quality-control criteria (35) within the 180 cortical regions of the left hemisphere, where spatial coverage in the AHBA is most comprehensive. We quantify transcriptional coupling between different brain regions using spatially-corrected correlated gene expression (CGE) (Fig. 1D, Fig. S6, see also Online methods) and define inter-regional connectivity using a binary group-representative matrix [see Online methods]. The spatial correction is important as prior studies of *C. elegans*, mouse, and human nervous systems have shown that, across the brain, CGE decays exponentially as a function of distance (7, 11, 24, 51). Recent analyses of the mesoscale connectome of the mouse (7) and microscale (cellular) connectome of *C. elegans* (11) indicate that, after correcting for this bulk trend, connected pairs of hubs show the highest CGE, despite being separated by longer anatomical distances, on average, than other neural elements.

Figures 3A–B show that the same effect is observed in humans: CGE is highest for rich, intermediate for feeder, and lowest for peripheral links. We obtain similar results under different connectome processing options

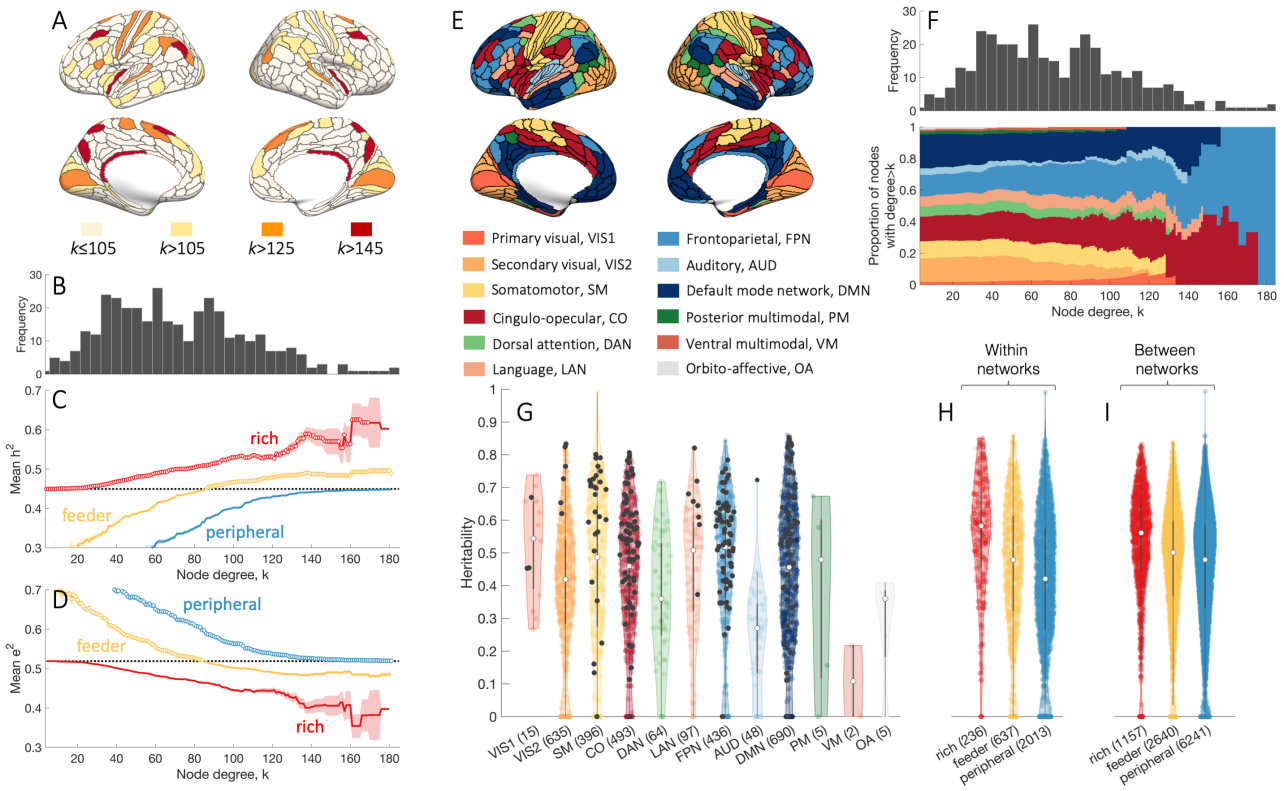

Fig. 2. Genetic influences on connectivity strength are preferentially concentrated on rich-club links. **(A)** Anatomical locations of hubs defined at different levels of k. **(B)** The degree distribution of the representative group-level connectome. Mean genetic **(C)** and unique environmental **(D)** influences for rich (hub-hub), feeder (hub-nonhub), peripheral (nonhub-nonhub) connections as a function of the hub-defining threshold, k. The mean of the corresponding measure across all network links is shown as a dotted black line. Shaded area corresponds to the standard error of the mean, circles indicate a statistically significant increase in mean heritability in a given link type compared to the rest of the network (one-sided Welch's t-test, uncorrected $p < 0.05$). **(E)** Regional assignments to canonical functional network modules (39), represented using color. **(F)** The proportion of nodes with degree > k in each functional network module as a function of k. **(G)** Distributions of heritability estimates across edges within functionally defined networks (39): VIS1 – primary visual; VIS2 – secondary visual; SM – somatomotor; CO – cingulo-opecular; DAN – dorsal attention; LAN – language; FPN – frontoparietal; AUD – auditory; DMN – default mode; PM – posterior multimodal; VM – ventral multimodal; OA – orbito-affective. Rich links within each module are represented as black dots, as defined for $k > 105$. Heritability distributions for edges within **(H)** and between **(I)** functional modules across rich, feeder, and peripheral link types for $k > 105$. Rich links show significantly higher heritability compared to both feeder and peripheral links, within and between functional modules (one-sided Welch's t-test, all $p < 1.9 \times 10^{-12}$).

(Fig. S7), distance ranges (Fig. S8), and when using connectivity data from the Monash Sample (Fig. S9). The consistency of this effect between human, mouse, and *C. elegans* [see Fig. S10 for comparison] is striking given the large physiological differences between species, methods for connectome reconstruction (DWI, viral tract tracing, electron microscopy), analysis resolution [macroscale (mm to cm), mesoscale (microns to mm), microscale (individual cells and synapses)], and gene-expression assays (microarray, *in situ* hybridization, curation of published reports).

As with heritability (Fig. 2C), higher CGE occurs for connections between high-degree nodes distributed across the brain; i.e., the effect is not confined to a single functional network (Fig. 3C). Indeed, connected pairs of hubs demonstrate higher CGE both within (Fig. 3D) and between (Fig. 3E) functionally defined networks (one-sided Welch's t-test, comparing CGE of rich *vs* feeder and rich *vs* peripheral connections, all $p \leq 0.02$).

Expression values in the AHBA are extracted from bulk tissue samples, and thus agglomerate transcriptional information from many different cell types. It is therefore possible that inter-regional CGE may be related to similarity in regional cellular composition [see also (52)]. We thus repeated the CGE analysis, this time using only data from genes showing cell-specific expression for seven canonical cell types: excitatory and inhibitory neurons, oligodendrocyte progenitor cell, astroglia, endothelial cells, microglia, and oligodendrocytes [see Online methods, (53–57)]. We find that all classes of cell-specific genes exhibit an increase in CGE for rich links relative to peripheral (Fig. S11), with oligodendrocyte-related genes showing a significantly stronger contribution to elevated CGE between connected hubs (one-sided Welch's t-test, $p = 2 \times 10^{-11}$, Fig. 3F) compared to all other genes [see Online methods].

These findings suggest that connected hubs have higher cytoarchitectonic similarity than other pairs of regions. We confirm this hypothesis using an analysis of the BigBrain atlas (58), which is a high-resolution Merker-stained histological reconstruction of a post-mortem human brain that provides an opportunity to map regional variations in cellular density as a function of cortical depth. Following Paquola et al. (59), we estimate in-

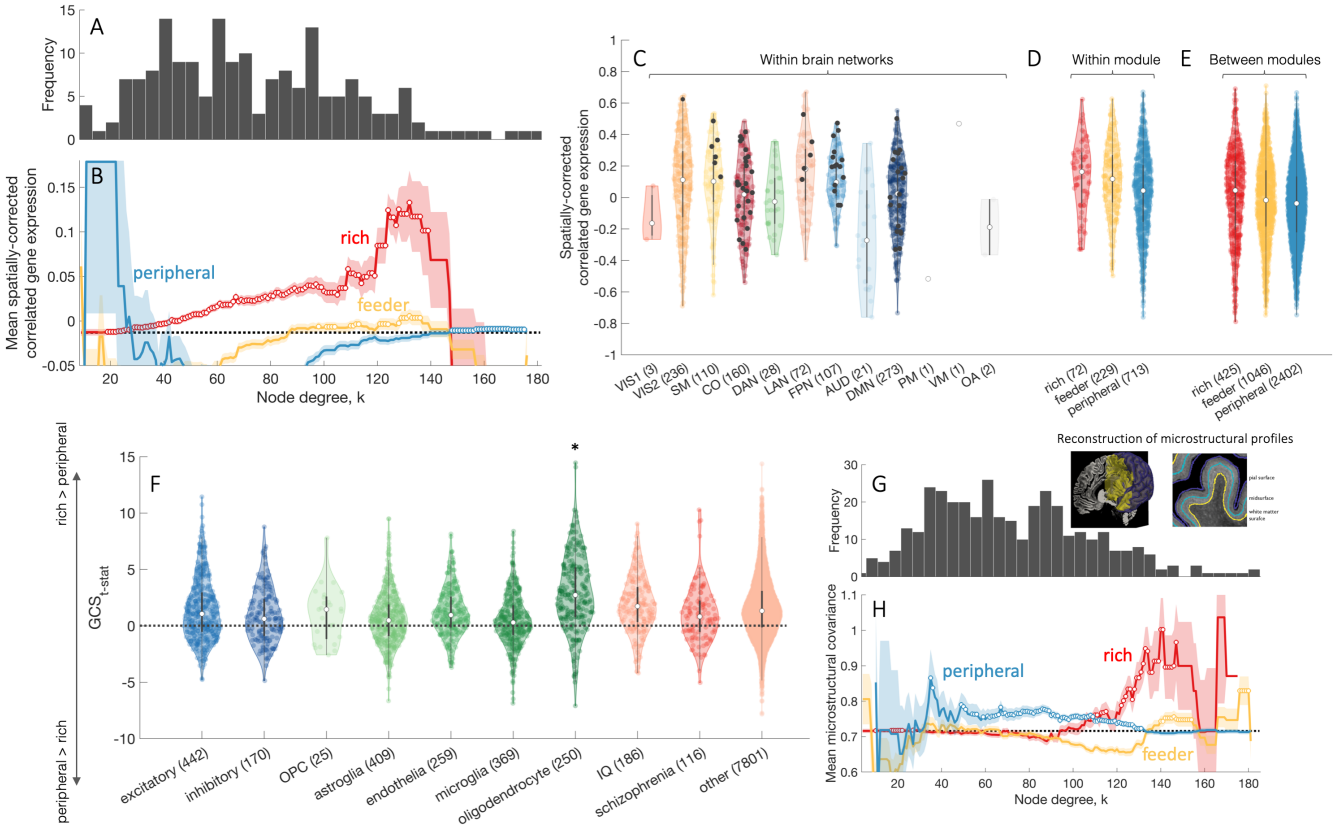

Fig. 3. Transcriptional coupling is elevated for connected brain network hubs. **(A)** The degree distribution of the representative group-level connectome of brain regions in the left cortical hemisphere. Degree is computed from whole-brain connectivity. **(B)** Mean correlated gene expression (CGE) for rich (hub-hub), feeder (hub-nonhub), peripheral (nonhub-nonhub) connections as a function of the degree threshold, k, used to define hubs. The mean CGE across all network links is shown as a dotted black line. The shaded area corresponds to the standard error of the mean, circles indicate a statistically significant increase in CGE in a given link type compared to the rest of the network (one-sided Welch's t-test, uncorrected $p < 0.05$). CGE estimates are corrected for distance effects, as explained in the Online methods. **(C)** CGE within functionally defined networks as in Fig. 2E. Black dots represent CGE values for rich links ($k > 105$). CGE values within **(D)** and between **(E)** functional modules in the left hemisphere across different link types (rich, feeder, and peripheral). Inter-module rich links show significantly higher CGE compared to both feeder (one-sided Welch's t-test, $p = 0.03$) and peripheral links ($p = 1.5 \times 10^{-4}$). Within functional modules, rich links show higher CGE compared to peripheral ($p = 1.2 \times 10^{-4}$) but not to feeder links ($p = 0.5$). **(F)** Gene contribution score t-statistic values ($\mathrm{GCS}_{\mathrm{t-stat}}$) for cell-specific gene groups quantifying the contribution of individual genes towards increased CGE for rich compared to peripheral links. Neuronal gene groups (excitatory – excitatory neurons; inhibitory – inhibitory neurons) are colored blue; glial gene groups (OPC – oligodendrocyte progenitor cells, astroglia, endothelia – endothelial cells, microglia, oligodendrocytes) colored green; IQ and schizophrenia-related gene groups colored pink, $GCS_{\mathrm{t-stat}}$; and values for all other genes presented in light orange. Oligodendrocyte-related genes show a statistically significant increase in GCC compared to all other genes (one-sided Welch's t-test, $p = 2 \times 10^{-11}$). No such effect was observed for schizophrenia or intelligence related genes. **(G)** The degree distribution of the representative group-level cortical connectome. **(H)** Mean microstructural profile covariance (MPC) for rich (hub–hub), feeder (hub–nonhub), peripheral (nonhub–nonhub) connections as a function of degree threshold, k used to define hubs. The MPC across all network links is shown as a dotted black line. Shaded area corresponds to the standard error of the mean, circles indicate a statistically significant increase in MPC in a given link type compared to the rest of the network (one-sided Welch's t-test, uncorrected $p < 0.05$). Inset near the degree distribution shows examples of the intermediate surfaces used to assay microstructure across the cortical depth.

tensity profiles across 16 equivolumetric surfaces placed between the gray/white and pial boundaries of the cortical ribbon and compute the inter-regional microstructural profile covariance (MPC) [see Online methods] as a proxy for cytoarchitectonic similarity. Mirroring the CGE and heritability findings, rich links exhibit elevated MPC compared to feeder and peripheral edges (Fig. 3H). These convergent MPC and cell-specific CGE results indicate that connected hubs have a more similar cytoarchitecture than other pairs of brain regions.

Finally, a gene set enrichment analysis of gene groups related to elevated CGE between hubs (Online methods) identifies significant enrichment of 49 GO categories, notably featuring genes related to oxidative metabolism, ATP synthesis coupled electron transport, and mitochondrial function ($p_{\mathrm{FDR}} < 0.05$, Table S4) that mirror those previously reported in the mouse brain (7). These results align with our eQTL findings to suggest a close genetic link between hub connectivity and metabolic function [for additional considerations see (60) and Online methods].

Stochastic models of brain wiring do not capture the spatial distribution of degree. The results above suggest strong genetic control of hub connectivity, which seems at odds with recent modeling studies suggesting that simple stochastic wiring rules can generate networks with complex, brain-like topologies, including heavy-tailed degree distributions that imply the existence of hubs (29, 30, 32, 33). Critically, our genetic analyses focus on individual variability in the *strength* of connec-

tivity between regions. Investigating variability in the binary topology of connectivity–that is, the specific pattern of wiring between regions–is more challenging, as such variations at the level of macroscale human connectomes are limited. It is thus possible that stochastic processes may give rise to the basic binary topology of hub connectivity, with variations in connectivity strength subsequently being influenced by genetic factors.

To investigate the role of stochastic processes in shaping hub connectivity, we fitted 13 different generative models of network wiring to the HCP connectome data. Under each model, synthetic connectomes are generated using probabilistic wiring rules. The models we consider here have been explored extensively in prior work (33) and have the general form:

$$\theta_{ij} = d_{ij}^{\eta} \times t_{ij}^{\gamma}, \quad \textbf{(1)}$$

where θ_{ij} is a score that weights the probability of connecting nodes i and j, d_{ij} is the Euclidean distance between node pairs, and t_{ij} is a topological property of an edge that may confer functional value to the network. Each of the 13 models substitutes a different topological property for t_{ij} (definitions in Table 2). The exponents η and γ are free parameters fitted to the data to optimally match the topological properties of the actual human connectome, as defined using nodal distributions of degree, clustering, and betweenness, and the edge-level distribution of connection distances [(33), see Online methods].

In line with prior work (33), we find that models in which connections form according to both spatial (saving wiring cost) and topological rules can fit the distributions of empirical network properties better than a model based on wiring cost alone (the 'sptl' model), as shown in Fig. 4A. The best-fitting model, 'deg-avg', modulates a pure wiring cost term by favoring connectivity between pairs of nodes that already have high average degree, and shows a good fit to the data (i.e., all fits, indexed by the Kolmogorov-Smirnov statistic, were KS < 0.18 [see Online methods for an extended discussion]).

Despite this adequate fit to four key network properties of the human connectome (Figs. 4B–E), we find that node degree in the empirical and model networks have very different spatial distributions. As shown in Fig. 4F, hubs in the empirical data are distributed throughout the brain, whereas hubs in the network that demonstrates the best fit to data across 130 000 model runs are predominantly confined to temporal cortex. As a result, the correlation between the degree sequences of the empirical and model networks is very low (e.g., $\rho = -0.001$, Fig. 4G). This low correlation is observed consistently across all models (Fig. 4H), and even when we fit model parameters to explicitly optimize the correlation between empirical and model degree degree sequences [see Fig. S13, Online methods]; across 260 000 model runs, the degree sequence correlation with the empirical data never exceeds $\rho = 0.3$.

Thus, while stochastic models of brain network wiring can capture the statistical properties (node- and edge-level distributions) of connectomes, they cannot reproduce the way in which these properties are spatially embedded and thus do not accurately replicate the precise pattern of wiring between connectome hubs.

Discussion

The complex topology of neural networks is thought to have been sculpted by competitive selection pressures to minimize wiring costs and promote complex, adaptive function (10, 61). Across diverse species, rich-club connections between hubs are among the most costly and topologically central links of the connectome (3–5, 7) and thus play a major role in determining how cost–value trade-offs are negotiated within a given nervous system. Here, we combine a multifaceted genetic analysis with mathematical modeling to examine the mechanisms that shape hub connectivity of the human connectome. We find that: (i) genetic influences on phenotypic variation in connection strength are principally concentrated on the rich links between hubs; (ii) DNA variants that are preferentially related to rich-link microstructure are eQTLs for genes that overlap with those implicated in intelligence, schizophrenia, and metabolism; (iii) connected hubs have highly correlated gene expression patterns that are related to similarity in regional cytoarchitecture and energy metabolism; and (iv) current stochastic models of network growth cannot reproduce the spatial distribution of hubs, consistent with a major role for genes in shaping the rich-club organization of the brain.

Our connectome-wide heritability analysis presents evidence for a non-uniform distribution of genetic influences across the brain, characterized by a gradient in which genetic influences are weak for peripheral connections between nonhubs, intermediate for feeder connections between hubs and nonhubs, and strongest for rich links between hubs. Critically, this effect cannot be attributed to connection distance or network affiliation, suggesting some degree of specificity to hubs located throughout the brain.

The most strongly connected hubs in our connectomes were located in multimodal association networks, which show disproportionate expansion in size and connectivity in human compared to nonhuman primates (16–18). Given the high centrality and cost of these connections [Fig. S1, (3–5, 7)], the preferential genetic influence on rich-club connectivity that we observe supports the hypothesis that natural selection favors wiring patterns that provide high value for low cost and that selection pressures are strongly concentrated on the valuable, costly links between hubs (2, 10). This view is also supported by recent evidence that genes demonstrating accelerated divergence between humans and chimpanzees

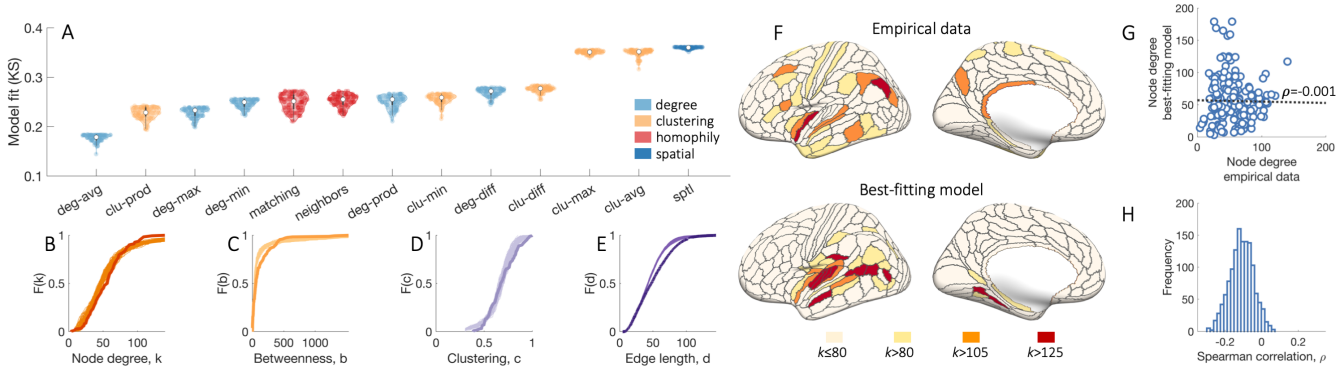

Fig. 4. Generative brain network models do not reproduce the spatial distribution of brain network hubs. (A) Each distribution represents estimates of model fit, as quantified by the maximum KS value of the top 100 networks (out of 10 000) produced by the model optimization procedure. The color of each box indicates conceptually related models, as determined by the specific topology metric used in the model [see Table 2]. Models favoring homophilic connectivity between node pairs are shown in red, those favoring clustering in orange, those based on degree in light blue, and a purely spatial model considering wiring costs alone is in dark blue. The specific wiring-rule names are shown along the horizontal-axis, with formal definitions provided in Table 2. Cumulative distributions of: **(B)** node degree, k; **(C)** betweenness centrality, b; **(D)** clustering coefficient, c; and **(E)** edge length, d, for the empirical connectome (darker line) and 100 runs (lighter lines) of the best-fitting 'deg-avg' model with optimized parameters ($\eta = -2.28, \gamma = 2.89$). **(F)** Anatomical locations of hubs defined for a single hemisphere at selected k thresholds for the empirical data (top) and the single run of the optimized 'deg-avg' generative model demonstrating the best model fit across 10 000 runs (bottom). **(G)** Correlation between the degree sequences of the empirical data and the best-fitting generative model within a single hemisphere (Spearman's $\rho = -0.001$). **(H)** The distribution of correlation values quantifying the relationship between left hemisphere degree sequences of the empirical data and synthetic networks generated using the top 100 best-fitting parameter combinations for each of the 13 considered models, corresponding to the data points shown in **A**.

show elevated expression in multimodal association networks (62).

In contrast to rich-club connectivity, phenotypic variance in peripheral connections between nonhubs is primarily influenced by unique environment. Topologically peripheral connections are more strongly conserved between human and chimpanzee connectomes (17). Moreover, the spatial topography and function of nonhub sensory areas is highly consistent across primates, presumably being specified early in development by evolutionarily conserved transcriptional gradients (18). These conserved gradients may couple with simple physical processes to give rise to predominantly short-range connectivity between topologically peripheral pairs of regions (31, 63). Subsequent modifications to peripheral connectivity may be driven by activity-dependent mechanisms, resulting in a greater environmental influence on phenotypic variance in connection strength.

It has been proposed that evolutionary expansion of multimodal hubs untethers these regions from transcriptional anchors in sensory areas, resulting in distinctive, non-canonical anatomical and functional properties (18). Our findings suggest that, despite this putative untethering, genes still play an important role in shaping phenotypic variance of hub connectivity. This result aligns with evidence that non-conserved network properties reflect evolutionary innovations that are driven by structural variation of DNA, yielding greater phenotypic variation within a species (12) and higher trait heritability when compared to more strongly conserved properties (64).

Our eQTL analysis revealed, in two independent samples, that SNPs preferentially related to variance in rich-link connectivity have a greater influence on the expression of genes that overlap with those implicated in intelligence and schizophrenia. The link to intelligence is consistent with the presumed role of association networks in supporting higher-order cognitive abilities (12, 13, 18). The link to schizophrenia is coherent with evidence that rich-club connectivity is disproportionately affected in this disorder (65, 66). An important future direction involves quantifying the degree to which genetic influences on intelligence and risk for schizophrenia are directly mediated by their effects on hub connectivity.

In addition to being highly heritable, pairs of connected hubs also show the highest levels of transcriptional coupling, as previously observed in the mesoscale mouse connectome (7) and cellular connectome of *C. elegans* (11) [see Fig. S10]. In *C. elegans* this result is not explained by the developmental proximity (i.e., similarity in neuron birth time or cell lineage distance), neurochemical identity, nor anatomical position of neuron pairs, but is instead related to the functional identity of hub neurons, which tend to be command interneurons. Our analysis of cell-specific genes points to a similar result at the regional level in humans, as rich-link CGE was elevated for gene markers of seven different cell types suggesting that network hubs have enhanced similarity in regional cytoarchitecture. This conclusion was supported by our MPC analysis of the BigBrain atlas. Our results align with the structural model of cortical connectivity, in which regions with similar cytoarchitecture are more likely to connect with each other, even over long distances (67). More specifically, when interpreted in light of the structural model, our findings suggest that hub areas are the most similar in their cellular composition, and that this similarity may play a critical role in how genes preferentially sculpt long-range interconnectivity between hubs.

We also show that current stochastic models of network growth, despite capturing key statistical network properties of the connectome, do not reproduce the spa-

tial locations of network hubs. For example, the actual hubs of the human brain have a widespread anatomical distribution, whereas the hubs in the best-fitting model network were concentrated around centrally located regions. In line with this result, recent work has shown that cost-neutral randomizations, in which connections are progressively randomized while preserving total wiring cost and the existence (but not position) of hubs, almost always degrade the functional complexity of the network, disconnect high-cost hubs, and lead to a distinct hub topography in which the most highly connected nodes cluster near the centre of the brain (68). These findings suggest that actual brains are very close to optimally balancing wiring cost with topological complexity, and that hub connectivity plays a critical role in determining how this balance is realized [see also (69)]. Some models suggest that random growth of connections, when coupled with changes in brain geometry and heterochronicity of connection formation across regions, can yield brain-like networks with realistic features (29), including connectivity between regions with similar cytoarchitecture (70, 71). Although genes likely influence heterochronous development, future work extending such models so that they can be directly fitted to empirical data in humans will be important for delineating the precise roles of genetic, environmental, stochastic, and physical mechanisms in shaping connectome architecture.

Online methods

Code for reproducing the results presented here is available at github `https://github.com/BMHLab/GeneticBrainHubs`. Data are available from an associated figshare repository.

Imaging data acquisition. We examined DWI data from two independent cohorts. The first was obtained from the Human Connectome Project [HCP, (36)]. We used the minimally processed DWI and structural data from the HCP for 972 participants (age mean ± standard deviation: 28.7 ± 3.7, 522 females), including a cohort of MZ and DZ twin pairs together with their non-twin siblings (more details presented in Online methods). Data were acquired on a customized Siemens 3T "Connectome Skyra" scanner at Washington University in St Louis, Missouri, USA using a multi-shell protocol for the DWI: $1.25\,\mathrm{mm}^3$ isotropic voxels, repetition time (TR) = 5520 ms, echo time (TE) = 89.5 ms, field-of-view (FOV) of 210×180 mm, 270 directions with b = 1000, 2000, 3000 s/mm^2 (90 per b value), and 18 b = 0 volumes. Structural T1-weighted data were collected using $0.7\,\mathrm{mm}^3$ isotropic voxels, TR = 2400 ms, TE = 2.14 ms, FOV of 224×224 mm. The full details can be found elsewhere (72).

The second DWI dataset came from individuals recruited as part of ongoing research conducted at Monash University. This Monash Sample was primarily used for replication of the connectome-wide GWAS and CGE analyses, and comprised 414 participants with MRI data obtained on a Siemens Skyra 3T scanner at Monash Biomedical Imaging in Clayton, Victoria, Australia using the following parameters: $2.5\,\mathrm{mm}^3$ voxel size, TR = 8800 ms, TE = 110 ms, FOV of 240×240 mm, 60 directions with b = 3000 s/mm^2 and seven b = 0 volumes. In addition, a single b = 0 s/mm^2 was obtained with the reverse-phase encoding so distortion correction could be performed. T1-weighted structural scans were acquired using: $1\,\mathrm{mm}^3$ isotropic voxels, TR = 2300 ms, TE = 2.07 ms, FOV of 256×256 mm. Data for 82 subjects were excluded due to: low connectome density ($n = 11$, connectome density more than 3 standard deviations lower than the mean) or no genotype data ($n = 69$), resulting in a final sample of 334 participants (age mean ± standard deviation: 23.65 ± 5.5, 190 females).

Image pre-processing. HCP DWI data were processed according to the HCP minimal preprocessing pipeline, which included normalization of mean b_0 image across diffusion acquisitions, and correction for EPI susceptibility and signal outliers, eddy-current-induced distortions, slice dropouts, gradient-nonlinearities and subject motion. T1-weighted data were corrected for gradient and readout distortions prior to being processed with Freesurfer [full details can be found in (72)].

Pre-processing for T1-weighted structural images in the Monash Sample consisted of visual screening for gross artefacts followed by the reconstruction of the grey/white matter interface and the pial surface using FreeSurfer v5.3.0 software. Surface reconstructions for each subject were visually inspected, with manual corrections performed as required to generate accurate surface models (72).

Distortions in the Monash DWI data were corrected with TOPUP in FSL, using the forward and reverse phase-encoded $b = 0$ images to estimate the susceptibility-induced off-resonance field (73, 74). We corrected for eddy-current distortions, volume-to-volume head motion, within-volume head motion, and signal outliers using eddy tool in FSL [version 5.0.11; (75–77)]. This implementation of EDDY significantly mitigates motion-related contamination of DWI connectivity estimates (78). DWI data were subsequently corrected for $B1$ field inhomogeneities using FAST in FSL (74, 79).

Connectome reconstruction. For both the HCP and Monash datasets, network nodes for each individual were defined using a recently-developed, data-driven group average HCPMMP1 parcellation of the cortex into 360 regions [180 per hemisphere, (37)]. An advantage of this parcellation is that it uses diverse structural and functional information to derive a consensus partition of the cortex into different areas. Each region has also been assigned to a distinct canonical functional network (39), allowing us to examine results in relation to the organization of these classic systems. However, the resulting areas can vary considerably in size, which can affect regional connectivity estimates since larger regions are able to accommodate more connections. To ensure that our results were not driven by the use of this specific parcellation, we replicated our main findings using a random cortical parcellation consisting of 500 approximately equally sized regions [250 per hemisphere, generated using the approach described in (80); code available at `https://github.com/miykael/parcellation_fragmenter`]. This offers a stringent test of the generalizability of our findings, as the parcellations vary in terms of both method for construction (data-driven *vs* random) and resolution (360 *vs* 500 nodes).

We focus our analysis on cortical connectivity for simplicity, as we know of no unified parcellation of cortical and subcortical areas, positional differences between cortex and subcortex can affect DWI connectivity estimates, and there are major differences between cortical and subcortical patterns of gene expression in the AHBA data that can be difficult to appropriately accommodate (35).

Subsequent processing of the DWI data for both the HCP and Monash data was performed using the MRtrix3 (81) and FMRIB Software Library (82). Tractography was conducted in each participant's T1 space using second order integration over fibre orientation distributions (iFOD2) (83). To further improve the biological accuracy of the structural networks, we also applied Anatomically Constrained Tractography (ACT), which uses a tissue segmentation of the brain into cortical grey matter, subcortical grey matter, white matter, and cerebrospinal fluid to ensure that streamlines are beginning, traversing, and terminating in anatomically plausible locations (84). Tissue types were determined using FSL software (82). A total of 10 million streamlines were generated on a probabilistic basis using a dynamic seeding approach that evaluates the relative difference between the estimated and current reconstruction fibre density and preferentially samples from areas of insufficient density (85). This method helps mitigate biases related to poor reconstruction of tracts from certain parts of the

brain due to insufficient seeding. The resulting tractogram was then combined with the cortical parcellation for each subject to produce a network map of white matter connectivity. Streamline termination points were assigned to the closest region within a 5 mm radius.

Connection weights were quantified using both streamline count (number of streamlines connecting two regions, SC) and the mean fractional anisotropy (FA) of voxels traversed by streamlines connecting two regions, which is commonly used as a marker of white matter microstructure. We focused on SC and FA as measures of connectivity strength because they are the most widely used in the literature, but note that they can be influenced by numerous factors that are not directly related to physiological measures of communication capacity between two regions (86). Moreover, while diffusion tractography remains the only available tool for *in vivo* connectivity mapping in humans, tractography algorithms can vary in their specificity and sensitivity for tract reconstruction (87, 88). To mitigate these effects, our data processing pipeline has been designed to limit contributions from spurious streamlines (84) and head motion (78). While the accuracy of all tractography methods remains an open challenge for the field (89), we note that any errors in tract reconstruction should reduce our chances of identifying stronger genetic effects for rich links through a heritbility analysis, since noisy connectivity values will inflate estimates of the E parameter (which also accounts for measurement error) in our biometric models, and rich links tend to be long-range connections, which are more prone to tractography errors (90). Our findings may thus provide a conservative estimate of genetic influences on hub connectivity.

Connectome thresholding. As connectomes are estimated with some degree of noise, it is common practice to threshold weak or inconsistent edges to focus on connections that can be more reliably estimated (91). We therefore selected edges that were: i) present in at least 30% of subjects; and ii) were amongst the τ% strongest edges (based on the streamline count) to achieve a desired connectome density. Since the desired connection density is arbitrary, we examined our main results across a range of densities: $\tau = 15\%, 20\%, 25\%$ for 360 region parcellation and $\tau = 5\%, 10\%, 15\%$ for the higher resolution parcellation of 500 regions. We note that the actual connection density of the human connectome remains unknown, and we chose these thresholds to span a range commonly studied in the literature.

The connection matrix resulting from our thresholding procedure was then used as a binary mask for selecting edges for the heritability, connectome-wide GWAS, and gene expression analyses. This masking procedure thus restricted individual variability in the binary topology of connectomes across individuals (indeed, in healthy individuals such topology should be highly conserved). For heritability and connectome-wide GWAS analyses, we used this group-representative connectome as a mask to extract FA-based connection weights. Heritability analyses were also repeated using streamline count as a measure of connection strength.

Rich-club organization. The connectivity of each region (node) in a network can be quantified by counting the number of connections to which it is attached. This measure is known as node degree. At a particular degree threshold, k, nodes can be labelled as hubs (degree $> k$) or nonhubs (degree $\leq k$), subsequently classifying all connections within the network as 'rich' (connection between two hubs), 'feeder' (connection between a hub and a nonhub), and 'peripheral' (connection between two nonhubs) [see, Fig. 1A, (5)]. To quantify the inter-connectivity between hub regions within a binary brain connectivity network, we used the topological rich-club coefficient $\phi(k)$:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k}-1)}, \quad \textbf{(2)}$$

where $N_{>k}$ is the number of nodes with degree $> k$, and $E_{>k}$ is the number of edges between nodes with degree $> k$ (92). Therefore, the rich-club coefficient quantifies the density of the subgraph comprising nodes with degree higher than the hub-defining threshold k.

Since nodes with higher degree make more connections, and can thus be expected to have a higher connection density compared to other nodes, we compared the $\phi(k)$ of the empirical network to the mean value across a 1000 randomized null networks, $\phi_{\text{rand}}(k)$, generated by rewiring the edges of the empirical network while retaining the same degree sequence, using the `randmio_und` function from the Brain Connectivity Toolbox (93), rewiring each edge 50 times per null network.

To assess whether the connections between high-degree nodes were also more likely to have stronger connection weights than expected by chance, we evaluated the weighted rich-club coefficient (94):

$$\phi^w(k) = \frac{W_{>k}}{\sum_{l=1}^{E_{>k}} w_l^{\text{rank}}}, \quad \textbf{(3)}$$

where $W_{>k}$ is the sum of weights in the sub-graph with degree higher than k, and the denominator is the total sum of l strongest weights in the network. As a null model for the weighted rich-club coefficient, we separate the definitions of weighted and topological rich-club coefficients by randomly reassigning weights within the network while preserving the binary topology (95) (instead of rewiring the links).

In both binary and weighted cases, we computed the normalized rich-club coefficient $\phi_{\text{norm}}(k)$ as the ratio between the rich-club coefficient in the empirical network and the mean rich-club coefficient in the set of the corresponding randomized networks:

$$\Phi_{\text{norm}}(k) = \frac{\phi(k)}{\langle \phi_{\text{rand}}(k) \rangle}. \quad \textbf{(4)}$$

Values of $\Phi_{\text{norm}} > 1$ indicate rich-club organization, where high-degree nodes are more densely interconnected (in a case of the topological rich-club) or have higher weights (in a case of the weighted rich-club) than be expected by chance. The statistical significance of the result is assessed by computing a p-value directly from the empirical null distribution of the 1000 randomized networks, $\phi_{\text{rand}}(k)$, as a permutation test (5). We note that in all our analyses, we estimated node degree using SC-weighted connectomes. Where indicated, FA-weighted connectomes were used in analyses of connectivity weights.

Communicability. We investigated the topological centrality of rich links using a measure called communicability (96), estimated across a range of degree thresholds. The communicability, C_{ij}, between a pair of nodes i and j, is calculated by accounting for all possible paths of length l between the nodes, weighted as $1/l!$, so that shorter paths make a stronger contribution to the overall score. The communicability, C_{ij}, for a binary matrix A is formally defined as:

$$C_{ij} = \sum_{l=0}^{\infty} \frac{(A^l)_{ij}}{l!} = (e^A)_{ij}. \quad \textbf{(5)}$$

In a weighted network, communicability ,C_{ij}^w, is defined using a weighted adjacency matrix W:

$$C_{ij}^w = (e^{S^{-\frac{1}{2}} W S^{-\frac{1}{2}}})_{ij}, \quad \textbf{(6)}$$

where $S^{-\frac{1}{2}}$ is the diagonal matrix with elements $\frac{1}{\sqrt{s_i}}$ and s_i is the strength of node i. We estimated the mean binary and weighted communicability for rich links, as a function of the hub-defining threshold k, to evaluate whether rich links are topologically central within the human connectome (Figs. S1E,F). An advantage of communicability is that, unlike classic measures of centrality, it does not assume that information is routed exclusively along shortest paths in the network, which is likely to be an inappropriate assumption for brain networks (2, 97).

Heritability analysis. The HCP diffusion imaging dataset includes 117 pairs of genetically confirmed monozygotic (MZ) twin pairs together with 69 of their non-twin siblings, as well as 60 dizygotic (DZ) same-sex twin pairs and 48 of their non-twin siblings. For each twin pair with more than one non-twin sibling, we selected one sibling at random (demographic details summarized in the Table 1). Only twin pairs where both twins had genetically-verified zygosity were included in the heritability analysis.

Table 1. Demographic data for twin groups and their non-twin siblings. MZ – monozygotic twins, DZ – dizygotic twins. Age is displayed in years: mean ± SD.

Zygosity	Number of subjects	Sex (F/M)	Age
MZ twins	117 pairs	69/48	29.3 ± 3.3
MZ non-twin siblings	69	34/35	29.1 ± 4.2
DZ twins	60 pairs	33/27	28.8 ± 3.5
DZ non-twin siblings	48	24/24	29.1 ± 4.0

Heritability analysis relies on the assumption that both shared genetic factors and common environment contribute to phenotypic similarity between twins within a pair, whereas unique environmental factors and non-shared genetic effects contribute to the differences observed between them. In the classical twin design, MZ twins are assumed to be genetically identical whereas DZ twins on average share half of their DNA, which is similar to non-twin siblings. Structural equation modeling can thus be used to decompose phenotypic variance and covariance in any particular trait into additive genetic (A), common environmental (C), and unique environmental (E) influences. Considering that twins raised together might have experienced a more similar environment compared to their non-twin siblings, including a set of non-twin siblings into the analysis allows us to separate the common environmental contributions into twin-specific (T) and twin non-specific (C) common environmental factors.

We used the binary group-representative cortical connectome mask described above to extract FA-weighted edges and applied standard structural equation modeling (SEM) to every connection in the connectome using OpenMx software (98, 99) in R. The analysis reported in the main text was performed on the 360 region (37) cortical connectome at 20% density (12 924 unique connections) using FA as a connection weight. The analyses were subsequently reproduced using SC (Fig. S4) and at different connectome densities (Fig. S3A-C) and using a higher resolution 500-region random cortical parcellation at 5%, 10% and 15% densities (Fig. S3D–F).

A range of biometric models – ACTE, ACE, AE, CE, E – were fitted to each edge defined by the group connectome mask in order to find connection-specific maximum likelihood estimates of additive genetic (A), twin-specific common environmental (T), twin non-specific common environmental (C) and unique environment (E) factors, using age and sex as covariates. Outlying connection weights for each analysis were removed using the `boxplot` function in R by keeping data points (w) in a range $Q1 - 1.5 \times IQR < w < Q3 + 1.5 \times IQR$ where $Q1$ and $Q3$ are the first and third quartiles respectively and IQR is the interquartile range. The Akaike information criterion (AIC) (100) was used to compare the goodness of fit of all tested models in order to find the most parsimonious model. For each edge, the model with the lowest AIC was selected. Consequently, the narrow-sense heritability (the proportion of variance attributable to additive genetic factors, referred to as heritability) was estimated for each connection using the best-fitting model. We also show heritability results using parameter estimates from the full ACTE model to ensure that our findings cannot be explained by our model selection procedure (Fig. S2).

Connectome-wide GWAS: discovery cohort. DNA samples from non-transformed and non-amplified white blood cells for 1141 individuals from the HCP were genotyped using the Infinium Multi-Ethnic Genotyping Array (MEGA). We restricted our analysis to a randomly selected a subset of unrelated individuals of self-reported Caucasian ancestry resulting in a set of 312 subjects. We performed standard quality control (QC) using PLINK 1.9 software at both the individual subject and SNP level. Initially, we checked if any subjects had a very low-genotyping score (≥ 0.10 of missing data, $n = 0$) and excluded SNPs with genotyping call rate $< 90\%$ and with a minor allele frequency (MAF) < 0.01. Further, several subject-level QC steps were performed by checking for and removing individuals: i) with disparities between the recorded and observed sex status as determined through X-chromosome homozygosity ($n = 1$); ii) with low genotyping score (≥ 0.05 of missing data, $n = 0$); iii) with cryptic relatedness higher than 0.25 ($n = 0$); iv) displaying outlying mean heterozygosity (greater than ± 3 SDs from the sample mean, $n = 6$). In order to identify any potential sources of population stratification in the sample we performed multidimensional scaling (MDS) using HapMap3 dataset (101) and saved first 10 principal components for future use as covariates. The data for a total of 305 subjects were used for imputation. SNPs with low genotyping call rate $< 95\%$, MAF< 0.01 and significantly departing from Hardy–Weinberg (H–W) equilibrium ($p < 10^{-7}$) were excluded leaving 502 027 variants in the final set taken forward to imputation. We used MaCH (102) and Minimac2 (103) for phasing and genotype imputation respectively employing the 1000 Genomes (phase one release three) reference panel. After imputation, SNPs with $0.05 < MAF < 0.98$ were selected. To eliminate poorly imputed variants, SNPs with imputation quality $r^2 < 0.8$ were excluded, resulting in a total of 5 235 153 variants used for the connectome-wide GWAS.

We performed a GWAS for each connection ($n = 12\,924$) in the group-representative 360 region (37) cortical connectome at 20% density. We used a binary mask of a group-representative connectome to extract FA-weighted edges across HCP subjects that had both DWI and genotype data available ($n = 254$). For each edge in the binary mask, subjects with an absent connection were removed and then connection weights were z-score normalized across subjects. We used the fastGWAS tool (104) implemented in the *gcta* software package (version 1.92.3beta3) to perform a linear regression-based association analysis (fastGWA-lr) for every edge controlling for age, sex and the first 10 principal components of the MDS as covariates. For each SNP, we extracted the effect sizes (β) associated with each connection and compared them between rich and peripheral links using a one-sided Welch's t-test. We then selected the top 10% of SNPs showing the largest differences between those two link types (all passing Bonferroni correction, p-threshold $= 0.05/12\,924$), thus representing SNPs that showed a preferential influence on rich compared to peripheral connectivity (Fig. 1C).

To reduce the SNP-set to an array of independent variants we performed a clumping procedure using the summary statistics from a GWAS of the mean FA across all available links in PLINK ($p_1 = 0.01$, $p_2 = 0.05$, $r^2 = 0.5$, $kb = 250$). To accurately estimate LD, we used the 1000 genomes phase 3 dataset (European ancestry sample, `https://sites.google.com/a/broadinstitute.org/ricopili/reference-panels`). The reduced set of SNPs consisted of 8207 variants that were subsequently used for eQTL mapping [see Online methods].

It is important to emphasize that our goal in conducting this connectome-wide GWAS is not to identify individual SNPs influencing hub connectivity at a genome-wide significant level. Instead, we use our GWAS to obtain beta estimates that are used to prioritize SNPs for eQTL mapping. We therefore do not suffer the same multiple comparison burden and sample size constraints as in a conventional GWAS. To ensure that our sample size was sufficient to reliably characterize the effects of interest, we replicated our analysis in an independent cohort of individuals, as described in the following section.

Connectome-wide GWAS: replication cohort. DNA samples for 715 individuals of European descent from the Monash Sample were genotyped using the Illumina Infinium PsychArray-24v1.2 BeadChip at Path West's Diagnostic Genomics Laboratory in Western Australia. The Illumina Psych-Chip comprising of 510 000 markers: 265 000 tagging SNPs from the Infinium Core-24 BeadChip and 245 000 markers from the Infinium Exome-24 BeadChip. Illumina Psych-Chip was developed in collaboration with the Psychiatric Genomics Consortium (PGC) and supplemented with an additional 50 000 SNPs implicated in psychiatric and neurodevelopmental disorders.

SNP and subject-level quality control procedures were identical to the ones applied in the Discovery cohort. Subjects were removed due to: disparities between the recorded and observed sex status ($n = 10$); low genotyping score (≥ 0.05 of missing data, $n = 2$); cryptic relatedness higher than 0.25 ($n = 13$); outlying mean heterozygosity (greater than ± 3 SDs from the sample mean, $n = 10$). In order to identify any potential sources of popula-

tion stratification in the sample we performed multidimensional scaling (MDS) using HapMap3 dataset (101) and excluded subjects exceeding ±2 SD on the 1st or 2nd principal components ($n = 15$) and saved the first 4 PCs for future use as covariates, leaving a total of 665 subjects for the imputation. SNPs with low genotyping call rate < 95%, MAF< 0.01 and significantly departing from Hardy–Weinberg (H–W) equilibrium ($p < 10^{-7}$) were excluded leaving 291 107 variants that were used for the imputation utilising 1000 Genomes (phase one release three) reference panel. After applying post-imputation quality control using the same criteria as in the Discovery cohort, a total of 5 723 894 variants used for the replication of the connectome-wide GWAS [see Connectome-wide GWAS: discovery cohort].

As in the HCP cohort, a group-representative connectome was generated by keeping the connections present in at least 30% subjects and retaining the strongest links based on the streamline count to achieve 20% connectome density, as done for the HCP sample. The resulting connectome was then used as a binary mask to select FA-weighted edges for the genome wide association analysis. Individuals with absent connections for a given edge were removed and then connection weights were normalized across subjects using z-score transformation. The final replication sample consisted of 276 subjects that had available both connectivity and genotyping data surpassing our quality control criteria. Edge-wise GWAS were performed following identical procedures outlined in the section Connectome-wide GWAS: discovery cohort, resulting in a set of 6081 prioritized SNPs used for eQTL mapping.

eQTL mapping. We sought to characterize the functional effects of our prioritized SNP list for hub connectivity by identifying which individual SNPs impact cortical gene expression (i.e., act as eQTLs) and mapping the specific genes that they influence. To maximize the power of eQTL mapping for prioritized SNPs identified separately in the HCP and Monash samples, we downloaded significant ($FDR < 0.05$) cis-eQTLs for gene expression in dorsolateral prefrontal cortex from psychENCODE (`http://resource.psychencode.org/`), a large publicly available genomic compendium with eQTL information measured from the brains of 1866 individuals (40). The psychENCODE eQTL dataset was filtered to genes with expression > 0.1 Fragments Per Kilobase of transcript (FPKM) in at least 10 samples, and contained 2 542 908 SNP-gene pairs with false discovery rate, $FDR < 0.05$ (105). For the prioritized SNPs from the connectome-wide GWAS, we extracted the paired eQTL target gene in psychENCODE from the total of 25 699 genes available in the database. This procedure thus resulted in a list of genes whose expression is affected by SNPs that are preferentially related to hub connectivity microstructure. We next sought to determine whether this gene list overlapped with genes related to intelligence and mental illness. We thus obtained summary results from prior GWAS of IQ (42), schizophrenia (43), bipolar disorder (44), major depression (45), autism spectrum disorder (46), and attention-deficit hyperactivity disorder (47) and used eMAGMA (106) to identify genes associated with each phenotype. eMAGMA is a validated pipeline that uses tissue-specific SNP-gene associations from the psychENCODE database (40) to assign SNPs to genes based on their association with gene expression. The SNP-gene associations are then aggregated in a gene-based test, while adjusting for linkage disequilibrium and correlated gene expression (107). We again used eQTL data from the prefrontal cortex in psychENCODE in our gene-based analyses. To identify genes for each disorder/trait, we controlled the family-wise error at 0.05 using Bonferroni correction (i.e., setting a p-value threshold equal to 0.05 divided by the number of identified genes). We used a hypergeometric test to test for overlap between gene lists, considering a total of $n = 25\,699$ possible genes (based on the number of genes in the eQTL mapping) by quantifying the probability to find a higher number of matches.

Over-representation enrichment analysis. We complemented our hypothesis-driven enrichment analysis with a more global test for enrichment of our hub gene list for certain functional classes of genes. Specifically, we performed over-representation enrichment analysis for GO categories and KEGG pathways on `g:Profiler` (108). We used the default g:SCS algorithm for multiple comparison correction while accounting for the hierarchical relationships between categories to provide a more reliable threshold for significance compared to standard hierarchy-naive correction methods (109).

Gene expression data. We used brain-wide gene expression data from the Allen Human Brain Atlas (AHBA), which consists of microarray expression measures in 3702 spatially distinct tissue samples taken from six neurotypical postmortem adult brains (50). Different brain regions were sampled across each of the six AHBA donors to maximise spatial coverage, resulting in approximately 400–500 tissue samples in each brain. The samples were distributed across cortical, subcortical, brainstem and cerebellar regions, measuring the expression levels of 58 692 probes quantifying the transcriptional activity of 20 737 genes. Considering that only two out of six brains were sampled from both left and right hemispheres whereas the other four brains had samples collected only from the left hemisphere, we focused our analyses on the left cortex only.

The pre-processing procedures applied to the data are outlined below and the choices detailed in (35). Briefly, probe-to-gene annotations were first updated using the Re-Annotator toolbox (110) resulting in the selection of 45 821 probes corresponding to the total of 20 232 genes. Second, tissue samples annotated to the brainstem and cerebellum were removed. Then, intensity-based filtering (35) was applied in order to exclude probes that do not exceed background noise in more than 50% of samples, excluding 13 844 probes corresponding to 4486 unique genes. Afterwards, a representative probe for each gene was selected based on the highest correlation to RNA sequencing data in two of the six brains (111). Gene expression samples were assigned to regions-of-interest by generating donor-specific grey matter parcellations and assigning samples located within 2 mm of the parcellation voxels. To increase the accuracy of assigning samples to regions, the samples were first divided into four separate groups based on their location: hemisphere (left/right) and structure assignment (cortex/subcortex), so samples listed as coming from the left cortical hemisphere in the AHBA ontology are only mapped to left cortical voxels of the parcellation (applying a 2 mm distance threshold, almost 90% of all cortical and subcortical samples were assigned to a non-zero voxel of the parcellation). Then, samples assigned to the subcortical regions as well as the right hemisphere were removed. Finally, gene-expression measures within a given brain were normalized first by applying a scaled robust sigmoid normalization for every sample across genes and then for every gene across samples in order to evaluate the relative expression of each gene across regions, while controlling for donor-specific differences in gene expression [see (35) for a validation]. Normalized expression measures in samples assigned to the same region were averaged within each donor brain and aggregated into a region by gene × matrix consisting of expression measures for 10 027 genes over 180 (left hemisphere, HCP parcellation) and 250 regions (left hemisphere of the random parcellation), respectively.

Distances between region pairs that were subsequently used to account for the spatial effects on transcriptional coupling were estimated on the cortical surface (pial) using the annotation files for each parcellation mapped onto the spherical representation of the `fsaverage` cortical surface. First, all the vertices that correspond to a particular region of interest in the spherical representation were identified and their centroid coordinates were calculated. Then the centroid coordinates were mapped to the `fsaverage` cortical surface and the distances between each pair of regions were calculated using the toolbox `fast_marching_toolbox` in MATLAB.

Transcriptional coupling. The result of the above mapping of AHBA data was an expression profile for each brain region, quantifying transcriptional activity across 10 027 genes. We used these profiles to quantify transcriptional coupling, or correlated gene expression (CGE), between every pair of regions. We defined CGE as the Pearson correlation between the normalized expression measures of the genes available after pre-processing ($n = 10\,027$). As

shown in Fig. S6A and described in (35), CGE exhibits a strong spatial autocorrelation that can be approximated as an exponential relationship with separation distance. To investigate whether CGE differs between different topological classes of connections beyond any low-order spatial effect, we fitted an exponential function with form $r(d) = Ae^{-d/n} + B$. The parameters $A = 0.64$, $B = -0.19$ and $n = 90.4$ capture the trend well, allowing us to retain the residuals for further analysis (Fig. S6B), defined as $\widehat{CGE}_{ij} = CGE_{ij} - r(d_{ij})$. These distance-corrected residual CGE values were used in all CGE analyses.

To evaluate transcriptional coupling for different connection types, for every edge within the connectivity matrix, we assigned a distance-corrected CGE measure. At each degree threshold, k, for defining hubs (nodes with degree$> k$), we then computed the average CGE of rich, feeder, and peripheral links. Significant increases in the CGE for a given link type compared to the rest of the network were evaluated using a one-sided Welch's t-test ($p < 0.05$).

Gene contribution score. To determine which functional gene groups contribute the most to any observed differences in CGE across different link types in the brain, we quantified the degree to which each gene contributes to the overall CGE between a pair of regions, following prior work (7):

$$\widehat{CGE}_{ij} = CGE_{ij} - r(d_{ij}) = \frac{1}{N}\sum_{a=1}^{N}[\widetilde{g}_i^a \widetilde{g}_j^a - r(d_{ij})] = \frac{1}{N}\sum_{a=1}^{N} GCS_{ij}^a, \quad (7)$$

where N is the number of genes ($N = 10\,027$), $\widetilde{g}_i^a \widetilde{g}_j^a$ the product of the z-score normalized expression values for gene a in regions i and j, and $r(d_{ij})$ is the previously defined spatial autocorrelation effect approximated as an exponential line (Fig. S6). Therefore, the gene contribution score between a pair of regions i and j for gene a was defined as $GCS_{ij}^a = \widetilde{g}_i^a \widetilde{g}_j^a - r(d_{ij})$.

We then assigned each gene a t-statistic quantifying the increase in GCS for rich compared to peripheral links ($\mathrm{GCS}_{\mathrm{t-stat}}$), as these two groups constitute the most distinct link types. A high value indicates increased CGE in rich compared to peripheral links. These t-statistic measures were used in the enrichment analyses as gene scores for determining whether any functional gene groups made a stronger contribution to CGE than others.

Cell-specific genes. Given that the AHBA assays gene expression using bulk tissue samples, it is possible that regional variations in cellular architecture drive differences in CGE between different link types. To test this hypothesis, we conducted a second CGE analysis focused on subsets of genes that have previously been identified as cell-specific markers. The set of cell-specific genes was compiled based on data from five different single-cell studies that used postmortem cortical samples of human postnatal subjects. Genes identified in each study as a cell-specific marker or as specifically enriched within a cell type were aggregated into study-specific lists (54–57). In the case of (53) where the normalized gene expression values were available for each cell type, we identified enriched genes as those with an average Fragments per kilobase million, FPKM> 5 and at least a four-fold enrichment over other cell types, as per authors recommendations. We then assigned genes within each of the resulting study-specific gene lists to one of seven canonical cell classes: astroglia, endothelial cells, excitatory neurons, inhibitory neurons, oligodendrocytes, and oligodendrocyte progenitor cell restricting each cell-class list to only contain genes unique to that class.

Gene-set enrichment analysis using gene score resampling. Gene-set enrichment analyses assess whether any functionally related groups of genes, annotated using Gene Ontology (GO), are associated with a selected phenotype. Every gene in our sample ($n = 10\,027$ genes) was assigned a t-statistic score quantifying its contribution towards the increase in GCS for rich links relative to peripheral ($\mathrm{GCS}_{\mathrm{t-stat}}$). Using these scores, we determined which specific functional groups of genes contribute to the observed increase in correlated gene expression. Functional gene group analysis was performed using version 3.1.2 of ErmineJ software (112). Gene ontology (113) annotations were obtained from GEMMA (114) as `Generic_human_ensemblIds_noParents.an.txt` on December 9, 2019. Gene Ontology terms and definitions were acquired from the `archive.geneontology.org/latest-termdb/go_daily-termdb.rdf-xml.gz` on January 13 2020. We performed gene score resampling (GSR) analysis on the $GCS_{\mathrm{t-stat}}$ scores testing the biological process GO categories with 5 to 100 genes available using the mean t-statistic score across genes to summarize the GO category and applying full resampling with 10^6 iterations. The resulting p-values were corrected across 6201 GO categories, controlling the false discovery rate (FDR) at 0.05 using the method of Benjamini and Hochberg (115). Recent work indicates that the default null models used in such analyses are insufficiently constrained for spatially embedded transcriptomic atlas data (60). This problem can lead to inflated significance for some GO categories when testing for spatial correlations between regional variations in gene expression patterns and measures of brain structure or function. The extent to which this problem generalizes to phenotypes defined for pairs of regions, such as the connectivity metrics considered here, is unclear. We nonetheless suggest caution in interpreting the findings of this analysis, as appropriate null models for the analysis of pairwise phenotypes have not yet been developed. We report the enrichment findings to test for consistency with prior findings in the mouse (7), and note that the enrichment of metabolic categories that we observe aligns with the findings of our eQTL analysis, which does not suffer from the same biases.

Microstructural profiles. Our CGE analysis of cell-specific genes indicated that connected hubs have more similar cellular composition than other region pairs. To independently verify this result, we estimated the microstructural profile covariance (MPC) between each pair of regions using the BigBrain atlas, which is a Merker-stained 3D volumetric histological reconstruction of a human brain (58, 59). MPC was estimated using methods described in [(59), see `https://github.com/MICA-MNI/micaopen/tree/master/MPC`]. In brief, the MPC procedure involved constructing 16 equivolumetric surfaces between the pial and white matter boundaries, followed by systematic sampling of the intensity values along these surfaces at 163 842 matched vertices per hemisphere. The intensity profiles, reflecting depth-wise changes in cellular density and soma size, were corrected for the midsurface y-coordinate to account for an anterior–posterior increase in intensity values across the BigBrain related to coronal slicing and reconstruction. Standardized residual intensity profiles were averaged within areas of the HCPMMP1 ($n = 360$) (37) and random ($n = 500$) parcellations. We quantified cytoarchitectural similarity between cortical areas by correlating areal intensity profiles (covarying for cortex-wide mean intensity profile), thresholding to retain only positive values ($r > 0$) and applying a log transformation, resulting in the measure of microstructural profile covariance (MPC) (59). We repeated the same analysis using the 500-region random parcellation. Notably, this analysis did not replicate the elevated MPC for rich links seen with the HCPMMP1 atlas (compare Fig. 3H with Fig. S12). This discrepancy likely reflects the fact that the HCPMMP1 parcellation more closely approximates boundaries between functional zones of the cortex, as it is based on a fusion of multimodal imaging data (37). The random parcellation makes no attempt to capture such boundaries and may blur different cytoarchitectonic regions within the same network node, thus resulting in noisier MPC estimates. In this way, the MPC results appear to depend on accurate approximation of cytoarchitectonic boundaries in cortex.

Models of brain network wiring. To evaluate the role of stochastic processes in shaping connectome architecture, we evaluated a series of generative models of network wiring that have the general form:

$$\theta_{ij} = d_{ij}^{\eta} \times t_{ij}^{\gamma}, \quad (8)$$

where θ_{ij} is a score that weights the probability of connecting nodes i and j, d_{ij} is the Euclidean distance between nodes i and j, and t_{ij} is some topological property of nodes i and j or an edge between them. This topological term modulates the probabilities

of forming an edge along with wiring cost (operationalized as d_{ij}). Numerous topological properties have been evaluated for t_{ij} in past work (32, 33), and we consider these same models here. A summary is provided in Table 2.

Table 2. A list of topological terms, t_{ij}, used in the generative models [see Eq. 8].

Name	t_{ij}
clu-avg	$\frac{c_i}{2}+\frac{c_j}{2}$
clu-diff	$\lvert c_i - c_j \rvert$
clu-max	$\max[c_i, c_j]$
clu-min	$\min[c_i, c_j]$
clu-prod	$c_i c_j$
deg-avg	$\frac{k_i}{2}+\frac{k_j}{2}$
deg-diff	$\lvert k_i - k_j \rvert$
deg-max	$\max[k_i, k_j]$
deg-min	$\min[k_i, k_j]$
deg-prod	$k_i k_j$
matching	$\frac{\lvert N_{i\setminus j} \cap N_{j\setminus i} \rvert}{\lvert N_{i\setminus j} \cup N_{j\setminus i} \rvert}$
neighbors	$\sum_l A_{il} A_{jl}$
sptl	1

where c is the local clustering coefficient, k represents node degree, A - adjacency matrix and $N_{i\setminus j}$ - neighbors of the node i excluding node j. The exponents, η and γ, act as weights on the distance and topological terms, respectively (32, 33). At each iteration, the computed connection score, θ_{ij}, is used to calculate the probability of a given edge, (i,j), being formed in that iteration, as $P_{ij} = \theta_{ij}/\Theta$, where Θ is the sum of θ_{ij} over all edges that have not yet been formed. Thus, at a given iteration, the model calculates the probability of each edge forming based on its distance and the current value of its topological term, t_{ij}. This topological value is recalculated at each iteration. Edges are added iteratively until the total number of edges is equal to the number of edges in the empirical connectome. Due to computational burden, and in line with prior analyses (33), we fitted models to a single (left) hemisphere connectome defined using the HCPMMP1 (37) parcellation, containing 5025 unique edges (20% whole cortex connectome density).

As per prior work (33), we quantified model performance using the Kolmogorov-Smirnov (KS) statistic. The KS statistic quantifies the distances between distributions of key topological statistics of the network; as such, lower values indicate better model fit. We focused on four key metrics: node-level distributions of degree, clustering, and betweenness, and the edge-level distribution of connection distance (33). The quality of model fit was defined as the maximum KS value across all four distributions; that is, model performance is defined by the property that is fitted most poorly. In principle, any number of other topological parameters could be used in this objective function, but these are some of the most widely used to characterize brain networks and were employed in prior work evaluating the same models (32, 33)

The free parameters η and γ were optimized as done previously (33). Specifically, we randomly sample the parameter space and evaluate the fits of the resulting networks. After sampling 2500 points in this space, Voronoi tessellation is then used to identify areas – or cells – of this space where the parameters produce networks with the best fits, as defined by the KS statistic. We then preferentially sample from cells with better fits. This procedure is repeated four times so that the algorithm gradually converges on an optimum. We ran each generative model on the group connectome 10 000 times and then evaluated each different model by comparing the 100 lowest energy values obtained from the optimization procedure.

For our analysis, we draw a critical distinction between the distribution and sequence of a topological property. The distribution refers to how a property is statistically distributed across the nodes of the network. The sequence refers to the exact assignment of a particular value to individual nodes or edges; in other words, how the property is spatially embedded in the brain. It is possible that two networks may have similar distributions for a given property, but with very different underlying sequences.

The models we consider here are optimized to match distributions, not sequences. As we are specifically interested in understanding the mechanisms driving the precise way in which hubs are connected, and given evidence that the specific anatomical location of network hubs has important implications for network dynamics (68), we seek to evaluate whether the models can not only generate networks with hubs, which would be shown by accurate fitting of the degree distribution, but also whether they also yield hubs the same anatomical regions as the empirical data, which would be shown by accurate fitting of the degree sequence. To this end, we also evaluate the correlation between the degree sequences of the empirical and synthetic networks using the Spearman correlation coefficient, ρ. A high correlation between the model and data implies that hubs are located in the same anatomical regions across the two networks. Conversely, a low correlation indicates that the model does not accurately capture the spatial embedding of connectivity in the connectome. Put simply, a low correlation implies that the hubs in the model network reside in anatomical locations that differ from the actual connectome.

An important consideration is that the correlation between model and empirical degree sequences was not a part of the objective function used in model fitting. We fitted the models using topological distributions and then evaluated their performance in capturing the empirical degree sequence. This procedure allows us to examine how well these models, as traditionally implemented, capture spatial properties of hub connectivity. However, this procedure also raises the question of whether it is possible to obtain a higher degree sequence correlation if model parameters are chosen to optimize this specific quantity. We therefore repeated the analysis after replacing the objective function with one that maximized the similarity between model and empirical degree sequences. Specifically, we optimized the Spearman correlation between model and empirical degree sequences with no other constraints to give the models the best possible chance of reproducing the empirically observed spatial topography of hub regions. The results of this analysis are shown in Fig. S13. Qualitatively similar results were obtained when optimizing the Pearson correlation between model and empirical degree sequences.

We note that across the 13 models evaluated in our analysis, we find that the best-fitting model is the 'deg-avg' model, which involves a trade-off between forming connections between highly connected nodes (i.e., node pairs with high average degree) and penalizing long-range connections (i.e., minimizing wiring cost). This result differs from past work, in which a homophilic attachment trade-off model that balances wiring cost with a preference for forming connections between nodes with similar neighbors offered the best fit to empirical connectome data (32, 33). This discrepancy may be related to our use of a higher resolution network parcellation, a connectome mapped at a different connection density, a different tractography algorithm, and/or a different diffusion MRI processing pipeline. Investigating the effect of these factors on modeling results is an important direction of future work, but these potential effects do not change the substantive point of our results that current stochastic models do a poor job of reproducing the spatial embedding of hub connectivity, as across 260 000 runs of the 13 different models considered, no degree sequence correlation exceeded $\rho = 0.3$. We also note that our model fits (Figs 4B–E) are in the same range as those reported by Betzel et al. (33), indicating that our discrepant results are not due to differences in model accuracy.

Finally, while our analysis mimics prior comprehensive evaluations of generative network models based on cost-value trade-offs (32, 33), other formulations and approaches are also possible (116–118) and we cannot completely rule out a role for stochastic processes in shaping hub connectivity. In particular, more abstract models suggest that stochastic wiring, acting in concert with developmental changes in brain geometry and heterogeneous timing of

connection formation across regions, can indeed generate networks with brain-like properties (34, 71, 119). However, a framework for directly fitting such models to human DWI data has not yet been developed.

ACKNOWLEDGEMENTS
We would like to thank Richard Betzel for sharing the optimization code for the generative modeling. This work was supported by the MASSIVE HPC facility (www.massive.org.au).

Bibliography

1. E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186 (2009).
2. A. Fornito, A. Zalesky, and E. Bullmore. *Fundamentals of brain network analysis*. Academic Press (2016).
3. L. Harriger, M. P. van den Heuvel, and O. Sporns. Rich club organization of macaque cerebral cortex and its role in network communication. *PLoS One* **7**, e46497 (2012).
4. E. K. Towlson, P. E. Vértes, S. E. Ahnert, W. R. Schafer, and E. T. Bullmore. The rich club of the C. elegans neuronal connectome. *J. Neurosci.* **33**, 6380 (2013).
5. M. P. van den Heuvel and O. Sporns. Rich-club organization of the human connectome. *J. Neurosci.* **31**, 15775 (2011).
6. M. P. van den Heuvel and O. Sporns. Network hubs in the human brain. *Trends Cogn. Sci.* **17**, 683 (2013).
7. B. D. Fulcher and A. Fornito. A transcriptional signature of hub connectivity in the mouse connectome. *Proc. Natl. Acad. Sci.* **113**, 1513302113 (2016).
8. B. Mišić, R. F. Betzel, M. A. de Reus, et al. Network-level structure-function relationships in human neocortex. *Cereb. Cortex* **26**, 3285 (2016).
9. R. L. Buckner, J. Sepulcre, T. Talukdar, et al. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability, and relation to Alzheimer's disease. *J. Neurosci.* **29**, 1860 (2009).
10. E. Bullmore and O. Sporns. The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336 (2012).
11. A. Arnatkevičiūtė, B. D. Fulcher, R. Pocock, and A. Fornito. Hub connectivity, neuronal diversity, and gene expression in the Caenorhabditis elegans connectome. *PLoS Comput. Biol.* **14**, e1005989 (2018).
12. S. Mueller, D. Wang, M. Fox D., et al. Individual Variability in Functional Connectivity Architecture of the Human Brain. *Neuron* **77**, 586 (2013).
13. R. E. Jung and R. J. Haier. The Parieto-Frontal Integration Theory (P-FIT) of Intelligence: Converging Neuroimaging Evidence. *Behavioral and Brain Sciences* **30**, 135 (2007).
14. N. A. Crossley, A. Mechelli, J. Scott, et al. The hubs of the human connectome are generally implicated in the anatomy of brain disorders. *Brain* **137**, 2382 (2014).
15. P. K. Reardon, J. Seidlitz, S. Vandekar, et al. Normative brain size variation and brain shape diversity in humans. *Science* **360**, 1222 (2018).
16. J. Hill, T. Inder, J. Neil, et al. Similar patterns of cortical expansion during human development and evolution. *Proc. Natl. Acad. Sci.* **107**, 13135 (2010).
17. D. J. Ardesch, L. H. Scholtens, L. Li, et al. Evolutionary expansion of connectivity between multimodal association areas in the human brain compared with chimpanzees. *Proc. Natl. Acad. Sci.* **116**, 7101 (2019).
18. R. L. Buckner and F. M. Krienen. The evolution of distributed association networks in the human brain. *Trends Cogn. Sci.* **17**, 648 (2013). Special Issue: The Connectome.
19. P. M. Thompson. Genetic influences on brain structure. *Nat. Neurosci.* **4**, 1253 (2001).
20. N. Yamamoto, A. Tamada, and F. Murakami. Wiring of the brain by a range of guidance cues. *Progress in Neurobiology* **68**, 393 (2003).
21. K. A. L. and T.-L. Marc. Mechanisms and Molecules of Neuronal Wiring: A Primer. *Cold Spring Harb Perspect Biol* **3**, a001727 (2011).
22. A. Fornito, A. Zalesky, D. S. Bassett, et al. Genetic influences on cost-efficient organization of human cortical functional networks. *J. Neurosci.* **31**, 3261 (2011).
23. P. E. Vértes, T. Rittman, K. J. Whitaker, et al. Gene transcription profiles associated with inter-modular hubs and connection distance in human functional magnetic resonance imaging networks. *Philos. Trans. R. Soc. B* **371**, 735 (2016).
24. A. Fornito, A. Arnatkevičiūtė, and B. D. Fulcher. Bridging the gap between connectome and transcriptome. *Trends Cogn. Sci.* **23**, 34 (2019).
25. A. Arnatkevičiūtė, B. D. Fulcher, and A. Fornito. Uncovering the transcriptional correlates of hub connectivity in neural networks. *Frontiers in Neural Circuits* **13**, 47 (2019).
26. S. T. E. Baker, D. I. Lubman, M. Yücel, et al. Developmental changes in brain network hub connectivity in late adolescence. *J. Neurosci.* **35**, 9078 (2015).
27. S. Oldham and A. Fornito. The development of brain network hubs. *Dev. Cogn. Neurosci.* (2018).
28. M. Ercsey-Ravasz, N. Markov, C. Lamy, et al. A predictive network model of cerebral cortical connectivity based on a distance rule. *Neuron* **80**, 184 (2013).
29. H. F. Song, H. Kennedy, and X.-J. Wang. Spatial embedding of structural similarity in the cerebral cortex. *Proc. Natl. Acad. Sci.* **111**, 16580 (2014).
30. J. a. Henderson and P. a. Robinson. Relations between the geometry of cortical gyrification and white-matter network architecture. *Brain Connect.* **4**, 112 (2014).
31. S. Horvát, R. Gămănuț, M. Ercsey-Ravasz, et al. Spatial embedding and wiring cost constrain the functional layout of the cortical network of rodents and primates. *PLoS Biol.* **14**, e1002512 (2016).
32. P. E. Vértes, A. F. Alexander-Bloch, N. Gogtay, et al. Simple models of human brain functional networks. *Proc. Natl. Acad. Sci.* **109**, 5868 (2012).
33. R. F. Betzel, A. Avena-Koenigsberger, J. Goñi, et al. Generative models of the human connectome. *Neuroimage* **124**, 1054 (2016).
34. A. Goulas, R. F. Betzel, and C. C. Hilgetag. Spatiotemporal ontogeny of brain wiring. *Science Advances* **5**, eaav9694 (2019).
35. A. Arnatkevičiūtė, B. D. Fulcher, and A. Fornito. A practical guide to linking brain-wide gene expression and neuroimaging data. *Neuroimage* (2019).
36. D. C. Van Essen, S. M. Smith, D. M. Barch, et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62 (2013).
37. M. F. Glasser, T. S. Coalson, E. C. Robinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171 (2016).
38. M. P. van den Heuvel, R. S. Kahn, J. Goni, and O. Sporns. High-cost, high-capacity backbone for global brain communication. *Proc. Natl. Acad. Sci.* **109**, 11372 (2012).
39. J. L. Ji, M. Spronk, K. Kulkarni, et al. Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage* **185**, 35 (2019).
40. D. Wang, S. Liu, J. Warrell, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362** (2018).
41. R. F. Betzel and D. S. Bassett. Specificity and robustness of long-distance connections in weighted, interareal connectomes. *Proc. Natl. Acad. Sci.* **115**, E4880 (2018).
42. J. E. Savage, P. R. Jansen, S. Stringer, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912 (2018).
43. S. Ripke, B. M. Neale, A. Corvin, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421 (2014).
44. D. M. Ruderfer, S. Ripke, A. McQuillin, et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705 (2018).
45. N. R. Wray, S. Ripke, M. Mattheisen, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668 (2018).
46. The Autism Spectrum Disorders Working Group of The Psychiatric Genomics Consortium. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* **8**, 21 (2017).
47. D. Demontis, R. K. Walters, J. Martin, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nat. Genet.* **51**, 63 (2018).
48. M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acid. Res.* **28**, 27 (2000).
49. M. Ashburner, C. A. Ball, J. A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
50. M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391 (2012).
51. H. Y. G. Lau, A. Fornito, and B. D. Fulcher. Scaling of gene transcriptional gradients with brain size across mouse development. *bioRxiv* p. 2020.06.04.135525 (2020).
52. J. B. Burt, M. Demirtaş, W. J. Eckner, et al. Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat. Neurosci.* **21**, 1251 (2018).
53. Y. Zhang, S. Sloan A., L. Clarke E., et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37 (2016).
54. M. Li, G. Santpere, Y. Imamura Kawasawa, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362** (2018).
55. B. B. Lake, S. Chen, B. C. Sos, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotech.* **36**, 70 (2018).
56. S. Darmanis, S. A. Sloan, Y. Zhang, et al. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci.* **112**, 7285 (2015).
57. N. Habib, I. Avraham-Davidi, A. Basu, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955 (2017).
58. K. Amunts, C. Lepage, L. Borgeat, et al. Bigbrain: An ultrahigh-resolution 3d human brain model. *Science* **340**, 1472 (2013).
59. C. Paquola, R. V. De Wael, K. Wagstyl, et al. Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLoS Biol.* **17**, e3000284 (2019).
60. B. D. Fulcher, A. Arnatkevičiūtėe, and A. Fornito. Overcoming bias in gene-set enrichment analyses of brain-wide transcriptomic data. *bioRxiv* (2020).
61. S. Ramón y Cajal. *Histology of the nervous system of man and vertebrates*. Oxford University Press (1995).
62. Y. Wei, S. C. de Lange, L. H. Scholtens, et al. Genetic mapping and evolutionary analysis of human-expanded cognitive networks. *Nat. Commun.* **10**, 4839 (2019).
63. J. A. Henderson and P. A. Robinson. Geometric effects on complex network structure in the cortex. *Phys. Rev. Lett.* **107**, 018102 (2011).
64. M. A. Bertolero, A. S. Blevins, G. L. Baum, et al. The human brain's network architecture is genetically encoded by modular pleiotropy (2019).
65. M. P. van den Heuvel, O. Sporns, G. Collin, et al. Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry* **70**, 783 (2013).
66. P. Klauser, S. T. Baker, V. L. Cropley, et al. White matter disruptions in schizophrenia are spatially widespread and topologically converge on brain network hubs. *Schizophr. Bull.* **43**, 425–435 (2016).
67. H. Barbas. General cortical and special prefrontal connections: principles from structure to function. *Annu. Rev. Neurosci.* **38**, 269 (2015).
68. L. L. Gollo, J. A. Roberts, V. L. Cropley, et al. Fragility and volatility of structural hubs in the human connectome. *Nat. Neurosci.* **21**, 1107 (2018).
69. D. S. Bassett, D. L. Greenfield, A. Meyer-Lindenberg, et al. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* **6**, e1000748 (2010).
70. S. F. Beul, H. Barbas, and C. C. Hilgetag. A predictive structural model of the primate connectome. *Sci. Rep.* **7**, 43176 (2017).
71. A. Goulas, P. Majka, M. G. P. Rosa, and C. C. Hilgetag. A blueprint of mammalian cortical connectomes. *PLoS Biol.* **17**, e2005346 (2019).
72. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105 (2013).
73. J. Anderson, S. Skare, and J. Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* **20**, 870

(2003).
74. S. Smith, M. Jenkinson, M. Woolrich, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208 (2004).
75. J. Andersson and S. Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *Neuroimage* **125**, 1063 (2016).
76. J. Andersson, M. Graham, E. Zsoldos, and S. Sotiropoulos. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. *Neuroimage* **141**, 556 (2016).
77. J. Andersson, M. Graham, I. Drobnjak, et al. Towards a comprehensive framework for movement and distortion correction of diffusion MR images: Within volume movement. *Neuroimage* **152**, 450 (2017).
78. S. Oldham, A. Arnatkevičiūtė, E. R. Smith, et al. The efficacy of different preprocessing steps in reducing motion-related confounds in diffusion MRI connectomics. *bioRxiv* (2020).
79. Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* **20**, 45 (2017).
80. M. Notter, K. Eschenburg, T. Serey, Claudio, et al. Parcellation fragmenter (2018).
81. J.-D. Tournier, F. Calamante, and A. Connelly. MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22**, 53 (2012).
82. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. FSL. *Neuroimage* **62**, 782 (2012).
83. J.-D. Tournier, F. Calamante, and A. Connelly. Improved probabilistic streamlines tractography by 2 nd order integration over fibre orientation distributions. *Ismrm* **88**, 2010 (2010).
84. R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly. Anatomically-constrained tractography: improved diffusion MRI streamlines tractography through effective use of anatomical information. *Neuroimage* **62**, 1924 (2012).
85. R. E. Smith, J.-D. Tournier, F. Calamante, and A. Connelly. SIFT2: Enabling dense quantitative assessment of brain white matter connectivity using streamlines tractography. *Neuroimage* **119**, 338 (2015).
86. D. K. Jones. Challenges and limitations of quantifying brain connectivity in vivo with diffusion MRI. *Imaging Med.* **2**, 341 (2010).
87. A. Zalesky, A. Fornito, L. Cocchi, et al. Connectome sensitivity or specificity: which is more important? *Neuroimage* **142**, 407 (2016).
88. C. Thomas, F. Q. Ye, M. O. Irfanoglu, et al. Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. *Proc. Natl. Acad. Sci.* **111**, 16574 (2014).
89. K. H. Maier-Hein, P. F. Neher, J.-C. Houde, M.-A. Cote, and M. Descoteaux. The challenge of mapping the human connectome based on diffusion tractography. *Nat. Commun.* **8**, 1349 (2017).
90. A. Zalesky and A. Fornito. A dti-derived measure of cortico-cortical connectivity. *IEEE Transactions on Medical Imaging* **28**, 1023 (2009).
91. M. A. de Reus and M. P. van den Heuvel. Estimating false positives and negatives in brain networks. *Neuroimage* **70**, 402 (2013).
92. V. Colizza, A. Flammini, M. A. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110 (2006).
93. M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059 (2010).
94. T. Opsahl, V. Colizza, P. Panzarasa, and J. J. Ramasco. Prominence and control: the weighted rich-club effect. *Phys. Rev. Lett.* **101**, 168702 (2008).
95. J. Alstott, P. Panzarasa, M. Rubinov, E. T. Bullmore, and P. E. Vértes. A unifying framework for measuring weighted rich clubs. *Sci. Rep.* **4**, 1 (2014).
96. E. Estrada and N. Hatano. Communicability in complex networks. *Phys. Rev. E* **77**, 036111 (2008).
97. J. Goñi, M. P. van den Heuvel, A. Avena-Koenigsberger, et al. Resting-brain functional connectivity predicted by analytic measures of network communication. *Proc. Natl. Acad. Sci.* **111**, 833 (2014).
98. S. Boker, M. Neale, H. Maes, et al. OpenMx: an open source extended structural equation modeling framework. *Psychometrika* **76**, 306 (2011).
99. M. C. Neale, M. D. Hunter, J. N. Pritikin, et al. OpenMx 2.0: extended structural equation and statistical modeling. *Psychometrika* **81**, 535 (2016).
100. H. Akaike. *Information theory and an extension of the maximum likelihood principle*. Springer, New York, NY (1998).
101. T. I. H. . Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
102. Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816 (2010).
103. C. Fuchsberger, G. R. Abecasis, and D. A. Hinds. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782 (2014).
104. L. Jiang, Z. Zheng, T. Qi, et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749 (2019).
105. Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Roy. Stat. Soc. B* **57**, 289 (1995).
106. Z. F. Gerring, A. Mina-Vargas, and E. M. Derks. emagma: An eqtl-informed method to identify risk genes using genome-wide association study summary statistics. *bioRxiv* (2019).
107. Z. F. Gerring, E. R. Gamazon, E. M. Derks, and for the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. A gene co-expression network-based analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. *PLoS Genetics* **15**, 1 (2019).
108. U. Raudvere, L. Kolberg, I. Kuzmin, et al. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191 (2019).
109. J. Reimand, M. Kull, H. Peterson, J. Hansen, and J. Vilo. g:profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193 (2007).
110. J. Arloth, D. M. Bader, S. Röh, and A. Altmann. Re-Annotator: annotation pipeline for microarray probe sequences. *PLoS One* **10**, e0139516 (2015).
111. J. A. Miller, V. Menon, J. Goldy, et al. Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq. *BMC Genomics* **15**, 154 (2014).
112. J. Gillis, M. Mistry, and P. Pavlidis. Gene function analysis in complex data sets using ErmineJ. *Nat. Protoc.* **5**, 1148 (2010).
113. M. Ashburner, C. A. Ball, J. A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
114. A. Zoubarev, K. M. Hamer, K. D. Keshav, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics* **28**, 2272 (2012).
115. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289 (1995).
116. M. Rubinov. Constraints and spandrels of interareal connectomes. *Nat. Commun.* **7**, 13812 (2016).
117. Y. Chen, S. Wang, C. C. Hilgetag, and C. Zhou. Features of spatial and functional segregation and integration of the primate connectome revealed by trade-off between wiring cost and efficiency. *PLoS Comput. Biol.* **13**, 1 (2017).
118. Y. Chen, S. Wang, C. C. Hilgetag, and C. Zhou. Trade-off between multiple constraints enables simultaneous formation of modules and hubs in neural systems. *PLoS Comput. Biol.* **9**, 1 (2013).
119. S. F. Beul, A. Goulas, and C. C. Hilgetag. Comprehensive computational modelling of the development of mammalian cortical connectivity underlying an architectonic type principle. *PLoS Comput. Biol.* **14**, e1006550 (2018).