Overcoming bias in gene-set enrichment analyses of brain-wide transcriptomic data.

Ben D. Fulcher[1,2,✉], **Aurina Arnatkevičiūtė**[2], **and Alex Fornito**[2]

[1]School of Physics, The University of Sydney, Camperdown, NSW 2006, Australia.
[2]The Turner Institute for Brain and Mental Health, School of Psychological Sciences and Monash Biomedical Imaging, Monash University, Victoria, Australia.

The recent availability of whole-brain atlases of gene expression, which quantify the transcriptional activity of thousands of genes across many different brain regions, has opened new opportunities to understand how gene-expression patterns relate to spatially varying properties of brain structure and function. To aid interpretation of a given neural phenotype, gene-set enrichment analysis (GSEA) has become a standard statistical methodology to identify functionally related groups of genes, annotated using systems such as the Gene Ontology (GO), that are associated with a given phenotype. While GSEA has identified groups of genes related to diverse aspects of brain structure and function in mouse and human, here we show that these results are affected by substantial statistical biases. Quantifying the false-positive rates of individual GO categories across an ensemble of random phenotypic maps, we found an average 875-fold inflation of significant findings relative to expectation in mouse, and a 582-fold inflation in human, with some categories being judged as significant for over 20% of random phenotypes. Concerningly, the probability of a GO category being reported as significant in the extant literature increases with its estimated false-positive rate, suggesting that published reports are strongly affected by the reporting of false-positive bias. We show that the bias is primarily driven by within-category gene–gene coexpression and spatial autocorrelation, which are not accounted for in conventional GSEA nulls, and we introduce flexible ensemble-based null models that can account for these effects. Testing a range of structural connectivity and cell density phenotypes in mouse and human, we demonstrate that many GO categories that would conventionally be judged as highly significant are in fact consistent with ensembles of random phenotypes. Our results highlight major pitfalls with applying standard GSEA to brain-wide transcriptomic data and outline solutions to this pervasive problem, which is made available as an open toolbox.

transcriptome | gene ontology | gene enrichment | genome | MRI
Correspondence: ***ben.fulcher@sydney.edu.au***

Introduction

The brain's multi-scale organization spans at least five orders of magnitude in space and time (1, 2). Understanding how these distinct scales relate to each other has proven challenging, largely because typical assays face a trade-off between spatial resolution and anatomical coverage. For example, *in vivo* brain-imaging techniques like magnetic resonance imaging (MRI) afford an unparalleled capacity to measure diverse aspects of structure and function across the entire brain, but have a limited spatial resolution that rarely surpasses $1\,\mathrm{mm}^3$. By contrast, invasive physiological recordings enable the measurement of neural structure and dynamics with cellular resolution, but traditionally only within small subregions of the brain. In recent years, our capacity to understand variations in molecular function across the entire brain has been greatly enhanced by industrialized, high-throughput transcriptome profiling, which has yielded genome-wide quantification of expression levels across the entire brain (3, 4). Two of the most influential genome-wide resources are the Allen Human Brain Atlas (AHBA), which encompasses microarray measurements of expression in $> 20\,000$ genes across 3702 tissue samples from six postmortem brains (4), and the Allen Mouse Brain Atlas (AMBA), which comprises *in situ* hybridization measures of expression for $> 17\,000$ genes at cellular resolution (3).

The availability of these genome-wide expression atlases has generated new opportunities to bridge spatial scales and uncover the microscale molecular correlates of macroscale brain organization. This correspondence has commonly been assessed by comparing regional variations in molecular function (from gene-expression maps) to independently measured macroscale structural and functional properties, e.g., from imaging techniques like magnetic resonance imaging (MRI) (5, 6). Prior work has characterized a relationship between gene expression and structural connectivity in *Caenorhabditis elegans* (7–10), mouse (11–18), and human (19, 20). Human research has further characterized links between gene-expression patterns and correlations of neural dynamics (functional connectivity) estimated from functional MRI (fMRI) (21–23) and electrocorticography (ECoG) (24); brain morphometry and microstructure (25–30); neurotransmitter receptor densities (31); and disease-related changes in brain structure and function (32–35). In these analyses, genes are typically scored according to the correspondence between their spatial expression map

and anatomical variations of the independently measured phenotype, most commonly quantified as a Pearson correlation across a set of spatial locations. Instead of performing inference on thousands of genes directly, statistical tests can be performed at the level of groups of functionally annotated sets of genes, such as those involved in different types of biological processes. This process of gene-set enrichment analysis (GSEA) investigates which gene categories are preferentially related to a spatial brain phenotype of interest, reducing the genome-wide multiple comparison burden and facilitating biological interpretation. GSEA uses a statistical hypothesis-testing framework to assess which categories of genes are most strongly related to a given phenotype, leveraging annotations of genes to categories from open ontologies like the Gene Ontology (GO) (36) and KEGG (37). GSEA has thus been applied extensively across species, scales, and diverse aspects of brain structure and function to gain insight into the biological processes that are mediated by genes with similar spatial expression profiles. Note that here we use the term 'gene-set enrichment analysis' (GSEA) in a very general way to denote the class of algorithms that aim to quantify enrichment in some quantity across labeled sets of genes, which is not to be confused with the specific 'GSEA' algorithm developed by Subramanian et al. (38).

In the brain-imaging literature, a wide variety of publicly available tools have been used to perform GSEA (39–48). While software packages make enrichment analysis easy to run, there are numerous challenges to correctly interpreting the results of GSEA. This requires careful consideration of: (i) the reliability of gene annotations (including the rate of false-positive annotations (49)); (ii) the reproducibility of results (different software packages use different methods to perform enrichment with respect to the GO hierarchy; both annotations and GO terms are updated daily (50)); and (iii) the statistical inference procedure (the design of multiple-hypothesis testing across many non-independent and often hierarchically structured tests (49, 51)). Methodological variability can lead to vast discrepancies in results obtained using different software packages. For example, one test (using identical inputs to different software packages) revealed a variation in p-values spanning several orders of magnitude for some GO categories (49). Methodological developments of GSEA are ongoing (52, 53).

Applications of GSEA, from the interpretation of genome-wide association studies (GWAS) (54), to case–control comparisons of microarray data, are subject to the general statistical challenges outlined above. But the growing applications of GSEA to whole-brain transcriptional atlas data are associated with unique challenges due to the spatial embedding of the atlas data that have yet to be characterized or explored. First, spatial embedding introduces coexpression between genes with similar spatial patterning of expression (4, 55, 56). This results in gene–gene dependencies that are characteristic of the expression dataset. Second, transcriptional maps are strongly spatially autocorrelated, such that nearby anatomical regions have more similar patterns of gene expression than distant regions, as has been observed in head neurons of *C. elegans* (10), the mouse brain (12, 18), and human cortex (5, 6, 55, 57–59). The neural phenotypes that are matched to gene-expression atlas data are also commonly spatially autocorrelated (10, 18, 60–64). Two spatial maps with similar spatial autocorrelation structure have a greater chance of exhibiting a high correlation to each other than two random spatial maps. Issues related to spatial autocorrelation of brain data have been highlighted in other contexts, with researchers developing methods to better estimate null distributions in the presence of spatial autocorrelation, e.g., using spatial permutation methods like spatial-lag models (30, 65) and spin tests (66) to test against an ensemble of surrogate spatial maps, or by removing the effect of physical distance through regression (5, 12, 14, 16–18, 24, 67). However, the precise impact of gene–gene coexpression and spatial autocorrelation of expression profiles on GSEA results has not been systematically characterized.

In this work we characterize the statistical biases involved in applying standard GSEA methodologies to brain-wide transcriptomic data. We demonstrate that the rate at which a GO category is judged as significantly correlated to a random phenotype is far higher than statistical expectation, exceeding 20% for some GO categories. We show that the literature is compatible with the reporting of this false-positive bias, finding a progressive increase in the rate of GO categories being reported as significant with their false-positive rate under random phenotypes. We show that these biases can be overcome using new ensemble-based null models that assess statistical significance relative to ensembles of randomized phenotypes (rather than constructing nulls by randomizing genes). Using case studies applying GSEA to a range of structural connectome and cell density phenotypes in mouse and human brains, we show that highly significant categories under conventional GSEA are often consistent with ensembles of randomized phenotypes. Our ensemble-based approach to GSEA overcomes biases in investigating the transcriptomic correlates of spatially varying neural phenotypes, and is made available as a Matlab toolbox.

Results

We first describe a typical GSEA pipeline applied to a spatial brain phenotype (SBP) of interest, depicted in Fig. 1A. The SBP is a spatial map of some measurement taken across brain areas, such as gray-matter volume, functional connectivity strength, or case–control differences in some property. Each gene in the atlas is scored using a measure of similarity between its expression pattern and the SBP (e.g., a correlation coefficient), yielding high scores for genes with expression maps that

closely resemble the SBP. GSEA then tests whether high-scoring genes are concentrated in particular types of GO categories by comparing the GO-category scores obtained from the data to those obtained from a null model. GSEA algorithms typically use a 'random-gene null' to assess whether a GO category's score is higher than would be expected if genes were annotated to GO categories randomly (49). This is similar to shuffling gene identities, as depicted in Figs 1B and C. This general procedure captures the essence of applying GSEA to expression-atlas datasets. While here we focus on these types of spatial comparisons, we note that some studies instead focus on pairwise inter-regional phenotypes, such as the presence of a connection between brain areas and a measure of correlated gene expression (6, 18, 21). We also note that some pipelines for performing GSEA have been proposed that randomize phenotypes rather than genes (38, 68), but to our knowledge they have not been applied in the setting of spatially embedded transcriptomics.

Throughout this work, we perform GSEA on biological process GO categories (36) using brain-wide gene-expression data from the AMBA (3) for the mouse brain, and the AHBA (4) for the human cortex. As described in Methods, enrichment is performed using gene-score resampling (39) using open Matlab software that we developed. Given the sensitivity of GSEA results to specific methodological pipelines (49), our aim here is to highlight general issues with applying any GSEA pipeline to spatially embedded transcriptional-atlas data. Thus, while the quantitative results presented here vary across different GSEA packages and parameters, the statistical biases we characterize will apply to all current GSEA pipelines that assess significance relative to random-gene null models.

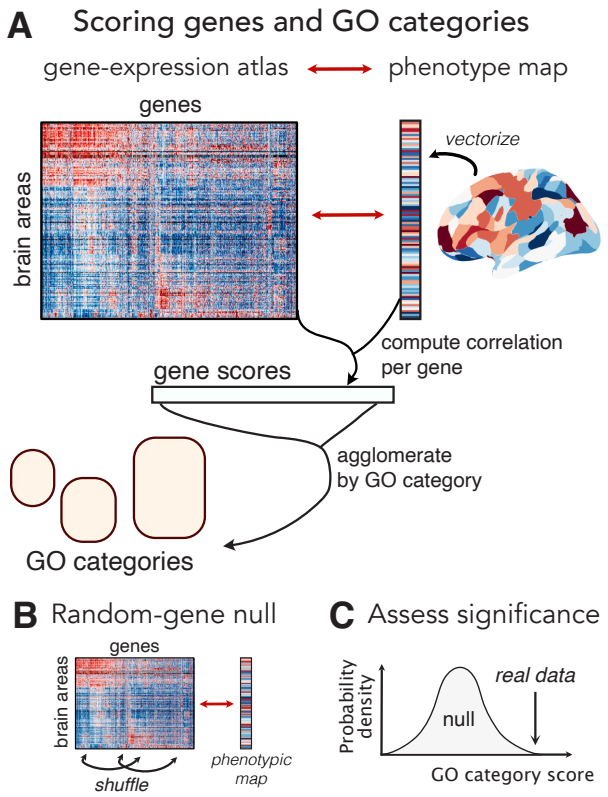

Fig. 1. Pipeline for applying gene set enrichment analysis (GSEA) to brain-wide expression atlas data. **A** Given a phenotype map, we first compute the spatial correlation coefficient between that map and each gene. These gene scores are then agglomerated at the level of categories using an annotation system like the Gene Ontology (GO). For continuous scores, agglomeration is typically performed as the mean score of genes annotated to a category. **B** Statistical significance of a GO category is assessed relative to the random-gene null, which estimates a null distribution for each GO category by annotating genes to GO categories at random. **C** For every GO category, a p-value is estimated using a permutation test, by comparing the GO category score obtained from the real data to the null distribution.

Diverse phenotypes in the literature are enriched for similar GO categories We undertook this study after observing a strong similarity in the results of GSEA analyses of spatially embedded data in the literature, despite these analyses involving very different phenotypes, species, measurement modalities, gene-expression processing pipelines, and software implementations of GSEA. To evaluate this consistency, we surveyed 16 mouse GSEA analyses (taken from eight different studies) and 60 human GSEA analyses (from 23 different studies), all involving spatially embedded atlas data from the AMBA and AHBA. Despite investigating very different phenomena, these studies reported similar sets of significant GO categories, most frequently implicating categories related to metabolic, neuronal, and generic biological and behavioral processes. The most reported GO category in our survey was 'chemical synaptic transmission' (GO:000726), which has been reported in 15 human analyses from 10 different studies (21, 26, 32, 69–75) and four different mouse analyses from three different studies (14, 17, 76). These studies implicate genes involved in chemical synaptic transmission in the organization of human resting-state functional connectivity (21), human adolescent cortical shrinkage and myelination (26), and tract-traced structural connectivity in the mouse brain (17). Some other commonly reported categories include 'potassium ion transmembrane transport' (22, 26, 32, 74, 77, 78), 'learning or memory' (15, 74, 75, 79, 80), and 'electron transport chain' (18, 19, 23, 26, 74) (see Supplementary File `LiteratureAnnotations_p005.csv` for a sorted list of GO categories, annotated by studies). All of these biological processes are sufficiently broad to be plausibly linked to any type of brain-related phenotype, but we aimed to investigate whether the consistency of these findings might instead be driven by common statistical biases in GSEA that favor the selection of some GO categories over others.

The GO enrichment signature of randomized spatial maps We first tested for false-positive bias in the mechanics of applying GSEA to transcriptional atlas data by characterizing the enrichment results for purely random SBPs (generated by assigning a random number to each brain area independently). We performed GSEA sep-

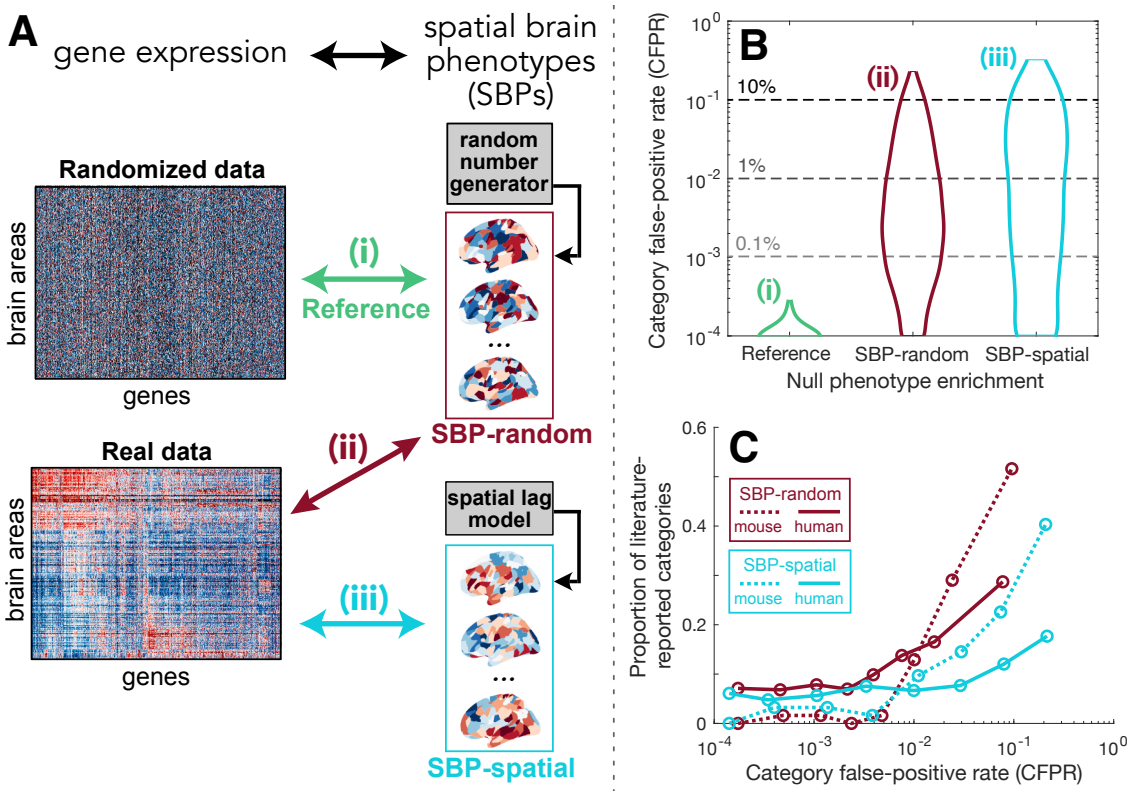

Fig. 2. Some GO categories have far higher false-positive rates under randomized spatial phenotypes than statistical expectation, and these GO categories are more likely to be reported as significant in published studies. A A schematic of three GSEA analyses involving correlations between: (i) 'reference'—an ensemble of random phenotypes and randomized gene-expression data (green), (ii) 'SBP-random'—an ensemble of random phenotypes and real gene-expression data (red), and (iii) 'SBP-spatial'—an ensemble of spatially autocorrelated phenotypes and real gene-expression data (blue). For the human cortex, examples of spatial maps in each ensemble are plotted; relative to the SBP-random maps that have no spatial correlation structure, the SBP-spatial maps are more likely to have similar values in nearby locations. **B** Distributions of the category false-positive rate (CFPR) across all GO categories are shown as violin plots in mouse. Across an ensemble of 10 000 SBP-random or SBP-spatial maps, the CFPR is computed for each GO category as the proportion of phenotypes for which that GO category was found to be significant. Results are shown for the three analyses depicted in A: (i) 'reference' (green); (ii) 'SBP-random' (red); and (iii) 'SBP-spatial' (blue). Note the logarithmic vertical scale (and therefore exclusion of GO categories with CFPR $= 0$). **C** The proportion of literature-reported GO categories increases with the false-positive rate expected under ensembles of random phenotypes. GO categories were labeled from a literature surveys of GSEA analyses using atlas-based transcriptional data in human and mouse (see Supplementary Information for survey details). Across eight equiprobable bins of CFPR (i.e., each bin contains the same number of GO categories), we plot the proportion of all literature-reported GO categories that are contained in that bin. Results are shown for the SBP-random (red) and SBP-spatial (blue) ensembles in mouse (dotted) and human (solid). The position of each bin is shown as the mean of its extremities.

arately for each of 10 000 random SBPs, noting which GO categories were significantly related to each SBP (false discovery rate, FDR < 0.05). We then computed the category false-positive rate (CFPR) for each GO category as the proportion of random SBPs for which a statistically significant spatial correlation was identified. This method of computing CFPRs under random SBPs is depicted schematically in Fig. 2A(ii), labeled 'SBP-random'. As random SBPs are uninformative, an unbiased GSEA procedure should produce CFPRs consistent with the expected statistical false-positive level, and all GO categories should have similar CFPRs. We estimated this expected CFPR numerically using a 'reference' case, shown in Fig. 2A(i), in which each gene's expression data was randomized independently. As this randomization destroys gene–gene coexpression structure in the gene-expression matrix, it allows us to isolate the contribution of non-random gene-expression structure to CFPRs through comparison to the 'SBP-random' results. Motivated by the strong spatial autocorrelation observed in many real SBPs (6), we also tested whether the use of spatially autocorrelated random phenotypes would affect CFPRs. These spatially autocorrelated SBPs, labeled 'SBP-spatial' [Fig. 2A(iii)], were generated using a spatial-lag model (81), with parameters determined from the spatial autocorrelation properties of the gene expression data (following Burt et al. (30), see Methods for details).

Distributions of non-zero CFPR across GO categories are shown on a logarithmic scale for the three experiments in Fig. 2B performed using mouse expression data (similar quantitative results were obtained in human, cf. Fig. S1). In the 'reference' case, consistent with the lack of signal from randomized gene expression and random phenotypes, we observed very low CFPR across our ensemble of 10 000 random SBPs. The vast majority of GO categories (84% in mouse and human) were never judged as statistically significant for any of the 10 000 random maps, and the maximum CFPR was 0.03% in both mouse and human (mean CFPR $= 0.002\%$ in both mouse and human).

When repeating the procedure using the same random

SBPs but now with real gene-expression data ('SBP-random', Fig. 2A), the mean CFPR increased 875-fold in mouse (from 0.002% to 1.5%) and 582-fold in human (from 0.002% to 1.0%). The maximal CFPR jumped dramatically from the reference level of 0.03% to 23% (in mouse; 25% in human). When the ensemble of phenotypes was constrained to exhibit generic spatial autocorrelation structure ('SBP-spatial'), most GO categories decreased their CFPR (60% in mouse and 68% in human). However, some categories exhibited large increases in CFPR driving an increase in the average CFPR under 'SBP-spatial' phenotypes, from 1.5% to 2.8% in mouse (maximal CFPR increased from 23% to 37%) and from 1.0% to 2.2% in human (maximal CFPR increased from 25% to 36%). The differences between the SBP-random and SBP-spatial ensembles will be explored in more detail later.

We thus find major differences in significance testing between controlling false-positive rates relative to randomized genes (conventional GSEA) and controlling false-positive rates relative to randomized phenotypes in the setting of transcriptomic atlas data. Conventional GSEA yields significant false-positive relationships to SBPs made up of random numbers, far beyond statistical expectation, indicating a striking methodological bias. This false-positive bias does not affect all categories equally: some GO categories exhibit reference-consistent CFPR ($\leq 0.03\%$) while others are judged to be significantly correlated to random spatial maps more than 20% of the time. Concerningly, the categories with the highest CFPRs in mouse and human are dominated by brain-related function: neurons (e.g., 'regulation of synaptic plasticity': 23% in mouse, 10% in human), metabolism (e.g., 'respiratory electron transport chain': 19% in mouse and 10% in human), and behavior (e.g., 'learning': 20% in mouse and 5% in human), as listed in Table S1 (full list in Supplementary File `CFPRTable.csv`).

To assess whether GO categories commonly reported as significant in the published literature might be affected by this false-positive bias, we investigated whether these published categories were more likely to have high CFPR. We performed an equiprobable binning of GO categories by their CFPR (placing an equal number of categories in each bin) and then analyzed the proportion of all categories in each bin that had been reported in the literature. As shown in Fig. 2C, for both ensembles of null phenotypes and in both mouse and human, we find a progressive increase in the frequency of reporting of a GO category as significant with its CFPR. This increase in literature reporting with CFPR was confirmed when incorporating information about the number of literature studies (Fig. S2). The increase in the reporting of GO categories with their false-positive rate under ensembles of randomized phenotypes suggests that the existing literature is heavily affected by the false-positive biases of conventional GSEA methodology.

Within-category coexpression drives false-positive bias Since randomizing the gene coexpression structure ('reference') dramatically reduces CFPRs, we reasoned that dependencies between genes within a GO category must drive the vast differences in CFPR between categories.

We explored how categories of the same size differ in their gene–gene coexpression using the example of a low-coexpression category ('zymogen activation', CFPR= 0.07%) and a high-coexpression category ('regulation of dendritic spine morphogenesis', CFPR= 13%) (Sec. S1.1). We found that the high-coexpression category is more likely to exhibit high category-wide correlations to a random SBP than random genes, because a chance correlation between the random SBP and any single gene is amplified through a similar correlation with many other genes in the category. This mechanism, through which gene-gene dependencies drive deviations from a random-gene null, results in the dramatic differences between false-positive rates estimated from conventional random-gene GSEA and those estimated relative to ensembles of random phenotypes for high-coexpression GO categories.

To test this explanation across all GO categories, we defined a simple measure of within-category coexpression, $\langle r \rangle$, as the mean coexpression across all pairs of genes in a GO category. As shown in Fig. 3A, $\langle r \rangle$ and CFPR (using the SBP-random ensemble) are positively correlated in both mouse and human. Consistent with their coordinated function in the brain, genes associated with metabolic and neuronal functioning in both mouse and human, display characteristic expression patterns resulting in high within-category coexpression, $\langle r \rangle$. For example, amongst the GO categories with the highest $\langle r \rangle$ are: 'regulation of short-term neuronal synaptic plasticity' ($\langle r \rangle_{\text{mouse}} = 0.38$, $\langle r \rangle_{\text{human}} = 0.29$) and 'ATP synthesis coupled proton transport' ($\langle r \rangle_{\text{mouse}} = 0.34$, $\langle r \rangle_{\text{human}} = 0.29$) (see Supplementary File `WithinCategoryCoexp.csv` for a full list).

Our results demonstrate that within-category coexpression plays a key role in driving false-positive bias when applying GSEA to transcriptomic data. And because brain-related categories exhibit high within-category coexpression, they are most prone to this effect, thus driving deceptively sensible brain-related enrichment results (evidenced in published GSEAs across diverse brain phenotypes, Fig. 2C).

The role of spatial autocorrelation We next turn to the role of spatial autocorrelation in SBPs by investigating the SBP-spatial ensemble of spatially autocorrelated phenotypes. Testing the same GO categories as above (Sec. S1.1), we found that adding the constraint of spatial autocorrelation (i.e., SBP-random $\rightarrow$ SBP-spatial) can be considered a perturbation on a category's CFPR under SBP-random, depending on the spatial autocorrelation structure of genes in that category. For example, if the genes in a category exhibit similar spatial autocorrelation structure to that of the phenotype (here, the

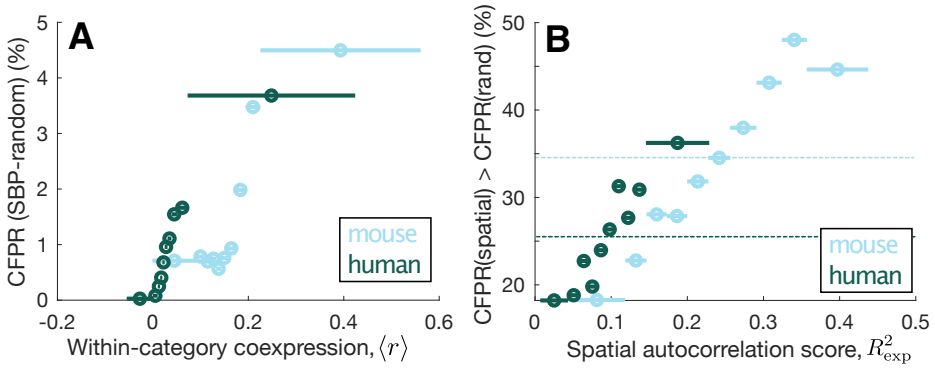

Fig. 3. Category false-positive significance rates (CFPRs) vary with within-category gene–gene coexpression and spatial autocorrelation in mouse and human. **A** CFPR (%) computed from the SBP-rand ensemble of random spatial maps, increases with a measure of mean within-category coexpression, $\langle r \rangle$, across ten equiprobable bins in mouse (blue) and human (green). The extent of each bin is displayed as a horizontal line. **B** The percentage of GO categories that exhibited an increase in CFPR when using the SBP-spatial ensemble relative to the SBP-random ensemble, across ten equiprobable bins of the spatial autocorrelation score, R^2_{exp}. This score captures the goodness of fit of each GO category's correlated gene expression to an exponential function with distance. More spatially autocorrelated GO categories are more likely to exhibit an increase in CFPR for spatially autocorrelated phenotypes (the SBP-spatial ensemble). The average value across all GO categories is shown as a horizontal dotted line.

SBP-spatial ensemble), then that category's CFPR can be substantially inflated.

To demonstrate this, we computed a simple measure of spatial autocorrelation for each GO category, R^2_{exp}, that measures the exponential distance-dependence of each category's correlated gene expression (see Methods). As shown in Fig. 3B, categories containing genes with stronger spatial autocorrelation scores, R^2_{exp}, are more likely to increase their CFPR under spatially autocorrelated phenotypes (SBP-spatial) relative to random phenotypes (SBP-random). This effect is strongest for categories with an autocorrelation length scale close to that of the SBP-spatial ensemble (Fig. S4), confirming the intuition that a high correlation between a GO category and phenotype is more likely for categories of genes with similar spatial autocorrelation properties as the phenotype. As this effect is not taken into account by conventional GSEA, it can drive increased CFPR, in addition to the effects of gene-gene coexpression.

Ensemble-based null models for spatial expression data. The above results reveal clear biases in the application of conventional GO-category enrichment methods to transcriptional-atlas data, in which controlling the false-positive rate under random genes induces a strong false-positive bias under random phenotypes. This bias is strongest in categories of genes that exhibit high within-category correlation and similar spatial autocorrelation properties as the phenotype. Since brain-related GO categories are characterized by high within-category coexpression and strong spatial autocorrelation, this false-positive bias misleadingly favors brain-related GO categories. These methodological issues highlight the need for new statistical methods to enable valid inference and interpretation of GSEA for spatially embedded transcriptional data.

In introducing new null models, we imagine a researcher who wishes to investigate the transcriptional correlates of their spatial brain phenotype, asking: *'Which GO categories contain genes that are significantly correlated to my phenotype?'* The conventional GSEA methodology evaluates the significance of a GO category against a null of random genes, depicted in Fig. 4A (cf. Fig. 1B). This generates null samples by randomizing the annotations of genes to GO categories, and asks the question of each GO category: *'Are genes in this category more strongly correlated to my phenotype than a random set of genes?'* Through the mechanisms described above, this randomization of gene-to-category annotations destroys within-category gene–gene coexpression structure, thereby driving high CFPRs.

To overcome this bias, we introduce a new way of performing GSEA that, instead of estimating statistical significance relative to randomized gene-to-category assignments, estimates the significance of a given SBP relative to ensembles of randomized SBPs. Two such ensembles of randomized phenotypes are the SBP-random and SBP-spatial ensembles defined above (Fig. 2A). By estimating each GO category's null distribution relative to an ensemble of randomized phenotypes, ensemble-based nulls test whether a given SBP is significantly correlated to genes in a GO category beyond what would be expected from the null phenotype ensemble. For example, in testing a measured SBP relative to the SBP-random ensemble, the researcher can ask the question of each GO category: *'Are genes in this GO category more correlated to my phenotype than they would be to a random phenotype?'* (Fig. 4B). Ensemble-based nulls allow for additional constraints (e.g., spatial autocorrelation) to be incorporated straightforwardly. For example, testing against the SBP-spatial ensemble allows the researcher to ask: *'Are genes in this GO category more correlated to my phenotype than they would be to a random spatially autocorrelated phenotype?'*, shown in Fig. 4C. While here we use the generic SBP-spatial ensemble, the spatial autocorrelation properties of the ensemble could be fitted to match the phenotype of interest, to ensure that the spatially autocorrelated ensem-

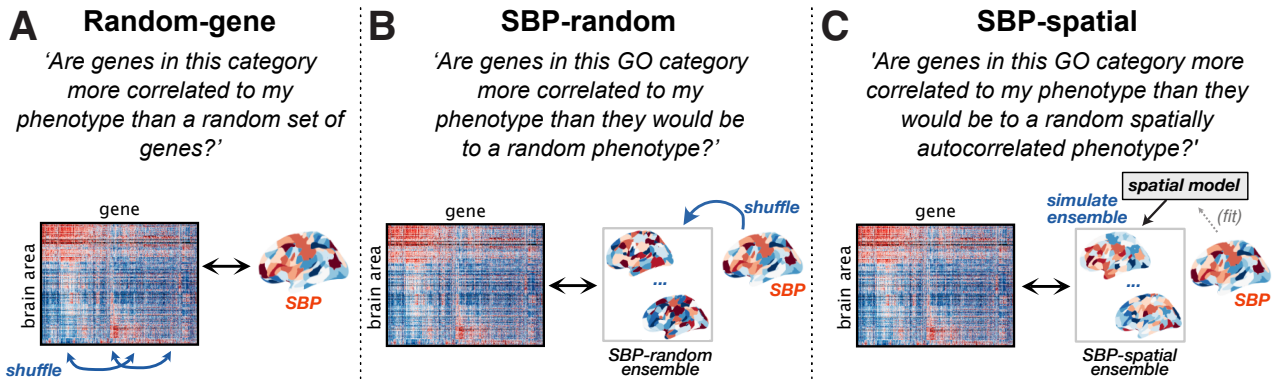

Fig. 4. The statistical significance of a GO category can be quantified relative to conventional random-gene nulls, or ensemble-based null models introduced here. For a given spatial brain phenotype (SBP) of interest, we depict the process through which null samples are generated for estimating statistical significance for GSEA across three different null models. **A** The conventional random-gene null tests whether the observed result is more extreme than if genes were assigned to GO categories at random (similar to the illustrated shuffling of gene identities). As this destroys within-category gene–gene correlation structure, it leads to high category false-positive rates for random phenotypes. An alternative is to compute null distributions for each category based on an ensemble of null phenotypes: **B** The SBP-random null tests whether the observed result is more extreme than if the phenotype of interest was a random spatial map. **C** The SBP-spatial null tests whether the observed result is more extreme than if the phenotype of interest was a random spatially autocorrelated map.

ble exhibits a similar spatial autocorrelation structure as the phenotype of interest, for which a range of existing methods have been developed (82). Importantly, ensemble-based nulls preserve the properties of the transcriptomic data (e.g., within-category coexpression and spatial autocorrelation properties of GO categories) that can inflate CFPRs when performing GSEA using conventional gene randomization.

Case studies: Spatial brain phenotype enrichment. In this section, we present some case studies to demonstrate how GSEA results can change when assessing significance relative to different null models. We tested two types of spatial brain phenotypes for their GO-category enrichment: (i) regional connectivity metrics, degree, k (defined as the number of connections attached to each region), and betweenness, B (defined as the number of shortest paths on the network passing through each region), derived from binary structural connectomes in mouse and human; and (ii) cell-density maps of PV+, SST+, and VIP+ interneurons from mouse cell-mapping experiments (83), and estimated density maps of oligodendrocytes, astrocytes, glia, microglia, neurons, and excitatory and inhibitory neurons (84). Our central aim was to understand how the enrichment signatures of these phenotypes differ across the null models introduced above. Full results tables for all GSEA analyses are provided as supplementary files (see Supplementary Information).

Across the fourteen GSEA analyses performed on mouse (12) and human (2) cortical maps, the conventional (random-gene) null yielded significantly enriched categories for nine of them, as shown in Fig. 5A. However, when assessing significance relative to random phenotypes (SBP-rand), only inhibitory cell density (mouse cortex) exhibited significant GO enrichment. When assessing significance relative to random spatially autocorrelated maps (SBP-spatial), not a single phenotype was significantly enriched for one or more GO categories.

To illustrate how estimated p-values could differ substantially across the three null-testing approaches, we plotted results for oligodendrocyte cell densities across the mouse cortex (84) in Fig. 5B. For example, 'aerobic respiration' (GO:0009060) has uncorrected p-value estimates of 5×10^{-7} (random-gene null), 0.01 (SBP-random), and 0.06 (SBP-spatial). We therefore find that GSEA significance can change markedly across null models and, in most cases investigated here, GO categories considered significantly correlated to a phenotype (relative to random-gene nulls) are in fact statistically consistent with expectation from random phenotypes. Many of the GO categories that are assessed as being significant under the conventional random-gene null are non-neuronal and hard to feasibly interpret, e.g., 'keratinization' correlates with degree, k, in human cortex ($q_{\mathrm{FDR}} = 1 \times 10^{-7}$), 'regulation of kidney development' correlates with neuron density in mouse cortex ($q_{\mathrm{FDR}} = 0.02$), and 'response to pheremone' correlates with VIP+ neuron density in mouse cortex ($q_{\mathrm{FDR}} = 0.002$). Thus, the randomized phenotype nulls can eliminate the statistical significance of physiologically inappropriate and spurious GO categories observed under the conventional random-gene null model.

While p-values are mostly higher under the phenotype-based nulls, this is not true for inhibitory cell density across mouse cortex (84). For this phenotype, the random-phenotype null (SBP-rand) flagged more significant categories (149) than the conventional random-gene null (just eight). That is, of the GO categories that were significantly more correlated to inhibitory cell density than to a random cortical phenotype, most were not more significantly correlated than would be expected from a random set of genes. This discrepancy occurs when an SBP, like inhibitory cell density, aligns strongly with a dominant gene-expression gradient, such that a random gene is far more likely to be correlated to that SBP than to a random SBP. This drives an offset in the random-gene null and increases estimated p-

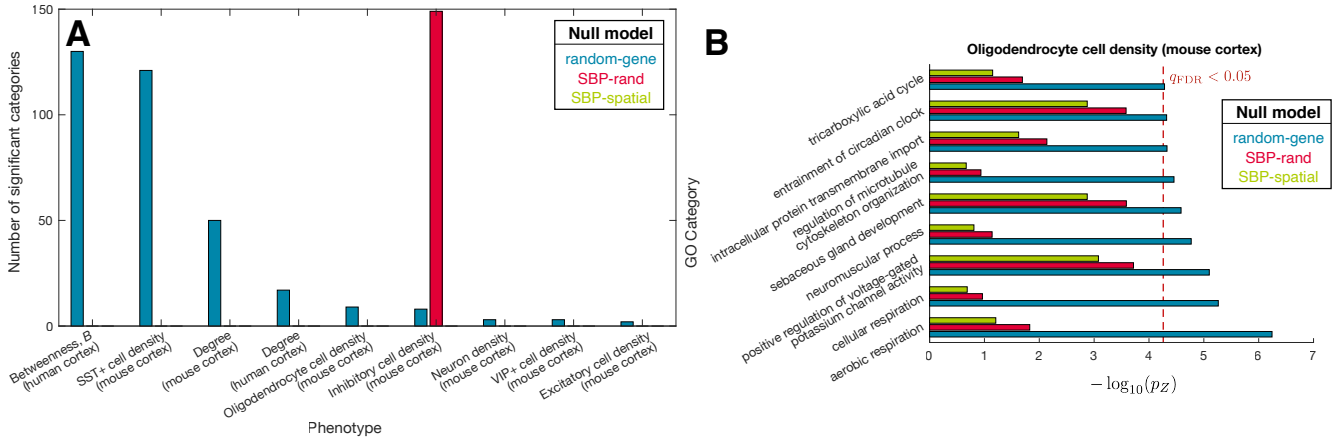

Fig. 5. GO category enrichment results depend strongly on the null ensemble. A Across the range of structural connectome nodal metrics (mouse and human) and cell density phenotypes (mouse), we show those nine phenotypes that individually exhibited categories with significant enrichment according to at least one of the null models. In all but one case, enrichment under the random-gene null is not significant under either of the random-phenotype nulls. **B** Picking an example enrichment analysis—oligodendrocyte cell density, which has 9 significant categories under the random-gene null—we plot the variation in estimated p-values (uncorrected) across the three null models (estimated from a Gaussian fit to the null distribution as p_Z). The corrected significance threshold, $q_{\mathrm{FDR}} = 0.05$, for the random-gene null is shown as a dashed red line; bars to the right of this line are considered significant at a false-discovery rate of 0.05.

values. This effect is important given the dominance of low-dimensional gradients of gene expression in the brain (30). Indeed, when expanding our mouse cortical analyses (38 regions) to the whole mouse brain (213 regions), we found a similar mechanism at play, in which the strong cortex–non-cortex expression difference (exhibited by the majority of genes: 73%, 14 137/19 417, $q_{\mathrm{FDR}} < 0.05$, Wilcoxon rank-sum test), dominates the enrichment signature of any phenotype that is differentially distributed between cortical and non-cortical areas. As the GO categories with the strongest cortical expression specificity are related to neuronal, metabolic, and behavioral function, GSEA results are correspondingly related to these plausible brain-related functions. This represents another potential source of misinterpretation of applying GSEA to transcriptional atlas data—it is easy to mistakenly assign specificity of enriched GO categories to a given phenotype (e.g., connectivity degree, which is higher in the isocortex), when this enrichment signature is actually driven by a non-specific expression difference between cortical and non-cortical areas.

Our results reveal that the significance of many phenotypes in GSEA strongly depends on the null model, with most significant findings (relative to random genes) being consistent with statistical expectation (relative to random phenotypes). In the presence of dominant transcriptional gradients, the reverse can also occur: GO categories that are not surprising relative to random-gene annotations (e.g., when the phenotype follows the gradient) can be more surprising relative to random phenotypes. In these cases, a null model matching spatial autocorrelation may offer more interpretable inference. Overall, these results highlight the need for proper comparison and careful consideration of the brain's spatial embedding when interpreting the results of GSEA.

Discussion

The ability to perform GSEA to identify groups of genes putatively related to spatially varying neural phenotypes has been facilitated by open neuroinformatics tools for accessing genome- and brain-wide transcriptional atlases in mouse and human (3–5, 85), and software tools for performing GSEA (39–43, 47, 48). While previous work has emphasized the need for care when applying GSEA for general applications—including substantial inter-software methodological variance, the reliability of gene-to-category annotations, and ongoing statistical improvements (49–51)—here we present a detailed analysis of methodological issues specifically associated with applications involving transcriptional atlas data, which are a new application for GSEA. We show that, when assessing the significance of a GO category with respect to spatial correlations of its member genes to an independent phenotype, it makes a substantial difference whether the null model is defined relative to randomized genes (as is conventional) or randomized phenotypes. While both gene- and phenotype-randomization has been proposed previously in the literature (38, 68), to our knowledge, only random-gene nulls have been used in the context of transcriptional atlas data, and no previous study has investigated the impact of this choice. We show that the conventional random-gene null judges some GO categories as significantly correlated with a surprisingly large fraction of randomized spatial phenotypes. GO categories with the highest CFPRs have high gene–gene coexpression, and CFPR is further increased when GO categories have a similar spatial autocorrelation structure to the phenotype of interest. Concerningly, high-CFPR categories are more likely to be reported as significant in the literature, suggesting that many published reports may be consistent with the expectations of randomized phenotypes. The flex-

ible new ensemble-based framework that we introduce for generating GO-category null distributions will enable researchers to more accurately interpret the results of applying GSEA to spatially embedded transcriptional data and brain phenotypes.

Applying conventional GSEA to an ensemble of randomized phenotypes led to high rates of estimated significance: relative to the reference level, false-positive rates increased by an average of 875-fold in mouse and 630-fold in human. GO categories varied widely in their CFPRs, with some brain-related categories exhibiting CFPRs over 20% (relative to a maximum CFPR of 0.03% in the reference case). We show that this false-positive bias is primarily driven by within-category gene–gene coexpression: a generic property of the expression atlas, not of the specific spatial brain phenotype being analyzed. Categories of genes involved in brain function tend to exhibit spatially coordinated expression patterns (compared to the less correlated expression patterns of non-brain genes), yielding high gene–gene coexpression and thus high CFPR. This leads to the dangerous consequence that applying conventional GSEA to transcriptional atlas data can misleadingly yield significant brain-related GO categories that are feasibly connected to the neural phenotype of interest. For example, GO categories related to oxidative metabolism have been linked to hub connectivity (18, 23), a result that can be plausibly interpreted in light of the known metabolic expense of brain-network hubs inferred from other modalities (86–88). While results of conventional GSEA reported the literature are not necessarily spurious, we must take care to validate these results against alternative nulls, like the ensemble-based approach introduced here.

Ensemble-based hypothesis testing for GSEA is a flexible framework with which to perform valid inference of the enrichment of spatial phenotypes. In the SBP-random null model, random numbers are assigned to all brain regions, and in the SBP-spatial null model we instead generate an ensemble of phenotypes with a given spatial autocorrelation structure. The framework is flexible to other types of generative null models to test more specific questions. For example, the SBP-spatial ensemble used here takes parameters from the average spatial correlation properties of transcriptional atlas data, but they could instead be fitted to match the spatial autocorrelation properties of the phenotype of interest, e.g., using the *brainSMASH* toolbox (65). For data that are sparsely sampled spatially, such that the parameters of these models are difficult to estimate, either the SBP-random null model can be used (which accounts for the major bias of gene–gene coexpression), or the null ensemble could be further constrained by, e.g., additional anatomical factors to test more specific hypotheses. For example, a null ensemble of phenotypes could be generated by randomizing the target phenotype separately within specific anatomical divisions—e.g., the ensemble could preserve differences in expression between cortical and non-cortical areas by randomizing cortical areas separately from subcortical areas. This flexibility to test against different null ensembles allows inference, and subsequent interpretation, to be tailored to the specific scientific question at hand. Furthermore, once a null ensemble has been generated, it can be applied not just to the univariate setting explored in this work, but similarly to multivariate analyses such as those that score genes according to a partial least squares (PLS) or canonical correlation analysis (CCA).

Ensemble-based nulls are expensive to compute, as GSEA often requires estimating very small p-values and new SBPs must be generated for each null sample. To ease the computational burden, here we used a Gaussian approximation for the 40 000 sample null distribution computed for each GO category, allowing us to estimate small p-values from the fitted Gaussian to a greater precision than a direct permutation test. This approximation was appropriate here for characterizing the general mechanisms and trends, but for a specific analysis, a full permutation-based procedure (that does not make strong parametric assumptions on the null distribution) would be necessary. We note that while the initial computation of null distributions for each GO category is expensive, it is a one-off computation for ensembles that are independent of the phenotype being tested (like the SBP-random and SBP-spatial ensembles studied here). That is, once the initial generic null computation has been computed, ensemble-based GSEA can then be applied efficiently for any new phenotype.

Our case studies of a range of spatial phenotypes across mouse and human cortex demonstrate the need for caution in interpreting the results of GSEA applied to transcriptional brain atlas data. We found that estimated p-values varied substantially between the different null models, most typically resulting in GO categories that are surprising relative to random gene annotations, but not surprising relative to randomized phenotypes. This difference highlights the important confounding role of variations in within-category gene–gene coexpression. Statistical sensitivity could be improved by reducing the number of comparisons across GO categories by considering a narrower range of brain-related GO categories of a given size, or by filtering on GO categories containing genes that are not spatially coherent. To this latter point, we found that many GO categories are made up of genes that do not show coherent, characteristic spatial patterning across the brain, and are therefore problematic to interpret as a functionally homogeneous 'set'. We also found that the effect of non-specific spatial effects, like cortical versus subcortical expression, can dominate GSEA findings for phenotypes that, e.g., differ between cortical and non-cortical areas, giving a false illusion of specificity and a plausible set of brain-related functional gene categories (that overlap strongly with cortical genes). For phenotypes that follow these types of spatial gradients, random genes are more likely

to correlate with the phenotype, resulting in p-values that can be higher than those obtained relative to random phenotypes. These discrepancies highlight the need for careful, nuanced interpretation of GSEA results with respect to appropriate null models.

Much of the brain's structural organization can be accounted for by its physical embedding, which imposes a spatial autocorrelation structure on many brain phenotypes. Spatial autocorrelation means that samples are not independent, and requires corrections to any resulting statistical inference, as has been noted in neuroimaging applications (30, 65, 66, 89), in other fields like geography (90), and time-series analysis (temporal autocorrelation) (91–93). The effect of spatial autocorrelation on GSEA is not straightforward: relative to independent random numbers (SBP-random), applying an autocorrelation constraint (SBP-spatial) yielded a large increase for a minority of GO categories, driving an overall increase in average CFPR. We showed that GO categories with a more similar spatial autocorrelation strength and length scale as the SBP-spatial ensemble have the greatest increase in CFPR relative to the SBP-random ensemble (Fig. 3B, Fig. S4). Differences in CFPRs determined by different null ensembles can be understood by viewing each generative null model as defining a probability distribution in the space of all possible SBPs. For categories that have similar spatial autocorrelation properties as the SBP-spatial ensemble, the SBP-spatial ensemble will define a higher probability density around SBPs that are correlated to its genes, yielding CFPR(SBP-spatial) > CFPR(SBP-random). Because here we use an SBP-spatial ensemble with a fixed strength and length scale, only a minority of categories match this quite restricted set of phenotypes better than SBP-random (40% in mouse and 32% in human), but for categories that match closely, CFPR can increase substantially. For example, in mouse, the CFPR of 'ionotropic glutamate receptor signaling pathway' increases from 8.8% (SBP-random) to 29.2% (SBP-spatial) and 'dopamine receptor signaling pathway' increases from 6.5% to 26.2%. Interestingly, maps with strong spatial autocorrelation, but at a different length scales, can be less correlated to each other than to random maps; this effect could explain the large drops in CFPR under SBP-spatial of some highly spatially autocorrelated GO categories such as 'ATP metabolic process' (CFPR in mouse decreased from 20.1% to 10.7% under the SBP-spatial ensemble, despite a high spatial autocorrelation score, $R^2 = 0.37$). Thus, whether the SBP-spatial ensemble is more or less conservative than the SBP-random ensemble depends on the specific spatial autocorrelation properties of each given GO category and how they relate to the properties of the SBP-spatial ensemble.

Our main aim here was to characterize the mechanisms of false-positive bias in the application of GSEA to transcriptional data, rather than to verify specific quantitative results of GSEA. Future work will be important to demonstrate how robust and accurate the results of GSEA are to methodological parameters of the GSEA analysis, as well as the choices made in processing gene-expression data (5). Refinements to the GSEA pipeline will also be important, and may include: developing inferential methods that account for the hierarchical organization of GO categories (49, 53); pruning the set of categories analyzed to those that are brain-relevant; and extending the current work to consider analyses of pairwise inter-regional (e.g., connectivity-based) phenotypes (6, 10, 18). We also note that while GO annotations were used here, the same arguments apply to other annotation systems, like the annotation of gene markers to cell types (94–96), where similar cautions should be noted.

Finally, we make some recommendations to improve the transparency and reproducibility of applying GSEA to transcriptomic atlas data. In reporting on GSEA analyses, we note that GO annotations are updated daily, such that GSEA tools can produce different results to identical inputs when run at different times. We thus recommend that researchers share the gene-score or gene-list inputs they used as input to the GSEA pipeline with a detailed methodological description including all settings used in performing enrichment. This would allow future studies to test for reproducibility through time, as gene annotations, categories, and category structures are updated (50), and would enable testing for robustness to other methodological implementations of GSEA (e.g., from different software packages). In reporting the results of GSEA, we discourage the practice of only listing manually selected GO categories in manuscript text but, rather, full outputs of the enrichment procedure should be provided as supplementary files that include gene scores and raw (and corrected) p-values.

The public availability of gene transcriptional atlases provides an incredible opportunity to bridge scales of analysis in the brain to understand how spatial variations in molecular function relate to large-scale properties of brain structure and function (6, 97). GSEA plays a crucial methodological role in leveraging gene ontologies to interpret these multi-scale relationships. We describe major statistical issues in applying conventional implementations of GSEA to assess the correlation between spatial brain phenotypes and the expression patterns of functionally annotated categories of genes, highlighting striking differences between nulls obtained from randomizing gene-to-category annotations (as is conventional) versus randomizing phenotypes. The new ensemble-based null models that we introduce here will allow researchers to better interpret the results of applying GSEA in using spatially resolved molecular maps to address diverse scientific questions. Extensive data tables for all enrichment analyses, code for reproducing all presented results, and a toolbox for performing ensemble-based enrichment accompany this paper.

Data and Methods

Documented code for reproducing all analyses presented here is available at https://github.com/benfulcher/GSEA_FalsePositives. All raw and processed data are available from an associated figshare repository. A toolbox for performing conventional (random-gene null) and ensemble-based GSEA in Matlab is available at https://github.com/benfulcher/GeneCategoryEnrichmentAnalysis.

Data.

Mouse We used gene-expression data from the Allen Mouse Brain Atlas (AMBA) (3), obtained using the Allen Software Development Kit (SDK)[1] using the 213-region parcellation of Oh et al. (98). Gene transcription in a given brain area was summarized as the 'expression energy' (the mean ISH intensity across voxels of that brain area) (3, 99). Where multiple section datasets were available for the same gene, they were combined by first z-scoring expression data taken across areas, and then computing the expression value of a cortical area as the mean across these z-scored section datasets. For each of 19 419 genes, we applied a 50% quality threshold, first on genes (keeping genes with expression data for at least 50% of brain areas), and then on brain areas (keeping brain areas with expression data for at least 50% of genes). This resulted in a region $\times$ gene expression matrix of size $213 \times 19\,417$ (whole brain), and $38 \times 19\,417$ (isocortical areas).

For the case study involving structural connectivity degree, we used axonal-connectivity data based on 469 anterograde viral microinjection experiments in C57BL/6J male mice at age P56, obtained from the Allen Mouse Brain Connectivity Atlas (AMBCA) (98). We computed degree, k, from binarized ipsilateral axonal connectivity in the right-hemisphere, including edges with $p < 0.05$, as inferred from the whole-brain linear-regression model presented in Oh et al. (98) (see Fulcher and Fornito (18)).

Human We use gene-expression data from the Allen Human Brain Atlas (AHBA) that consists of microarray expression measures across 3702 spatially distinct tissue samples collected from six neurotypical postmortem adult brains (4). Considering that two of the six brains were sampled from both left and right hemispheres while the other four brains had samples collected only from the left hemisphere, we focus our analyses on the left hemisphere only. The data was processed as outlined in Arnatkevičiūtė et al. (5). Specifically: i) probe-to-gene annotations were updated using the Re-Annotator toolbox (100); ii) intensity-based filtering was applied in order to exclude probes that do not exceed background noise in more than 50% of samples; iii) if more than one probe was available for a gene, a probe with the highest differential stability score (56) was selected; iv) gene expression samples from cortical regions were assigned to the regions-of-interest by generating donor-specific gray-matter parcellations (180 regions per hemisphere) and assigning samples located within 2 mm of the parcellation voxels; v) gene expression measures within a given brain were normalized first by applying a scaled robust sigmoid normalization (101) for every sample across genes and then for every gene across samples. This allowed us to evaluate the relative expression of each gene across regions, while controlling for donor-specific differences in gene expression. Normalized expression measures in samples assigned to the same region were averaged within each donor brain and then aggregated into a $180 \times 15\,744$ region $\times$ gene matrix consisting of expression measures for 15 744 genes across 180 cortical regions in the left cortex. Three of the 180 regions had no gene expression samples assigned; all analyses shown here are of the remaining 177 regions.

Structural connectivity was estimated based on the minimally processed diffusion weighted imaging and structural data from the Human Connectome Project (102) for 972 participants ($\text{age}_{\text{mean}} = 28.7 \pm 3.7$, 522 females) (103, 104). Data were acquired on a customized Siemens 3 T 'Connectome Skyra' scanner (Washington University in St Louis, Missouri, USA) using a multi-shell protocol for the DWI: 1.25 mm^3 isotropic voxels, repetition time (TR) = 5.52 s, echo time (TE) = 89.5 ms, field-of-view (FOV) of 210 $\times$ 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 mm^3 isotropic voxels, TR = 2.4 s, TE = 2.14 ms, FOV of 224 $\times$ 224 mm. The full details regarding data acquisition can be found elsewhere (103). For each individual network, nodes were defined using a recently-developed, data-driven group average parcellation of the cortex into 360 regions (180 per hemisphere) (105) using Freesurfer-extracted surfaces and subsequently projected to volumetric space.

Processing of the DWI data was performed using the MRtrix3 (106) and FMRIB Software Library (107). Tractography was performed in each participant's T1 space using second order integration over fibre orientation distributions (iFOD2) – a probabilistic algorithm that improves the quality of tract reconstruction in the presence of crossing fibers and high degree of fiber curvature (108). To further improve the biological accuracy of the structural networks we also applied Anatomically Constrained Tractography (ACT), which delineates the brain into different tissue types (cortical grey matter, subcortical grey matter, white matter, cerebrospinal fluid). This information is then used to ensure that streamlines are beginning, traversing, and terminating in anatomically plausible locations (109). Tissue types were determined using FSL software (107). A total of 10^7 streamlines were generated using a dynamic seeding

[1] ©2010 Allen Institute for Brain Science. Allen SDK. Available from: http://alleninstitute.github.io/AllenSDK/

approach. By evaluating the relative difference between the estimated and current reconstruction fibre density, it preferentially initiates seeding from areas of insufficient density (110).

The resulting tractogram was then combined with the cortical parcellation for each subject assigning streamline termination points to the closest region within 5 mm radius. Connection weights were quantified using streamline count (number of streamlines connecting two regions). Connectomes derived from probabilistic tractography algorithms are often thresholded due to the high probability of false-positive connections (111, 112), therefore a single group-average connectome was aggregated by selecting connections that are present in at least 30% of subjects and retaining the strongest edges (based on the median streamline count across subjects) to achieve a connectome density of 20%.

Spatial phenotypes We computed degree, k, and betweenness, B, of each node from the binary connectomes described above. Node betweenness was computed using the Brain Connectivity Toolbox (113). When estimating these quantities in the mouse cortex, only cortico-cortical connectivity was used. Interneuron subtype densities were measured by *qBrain*, quantitative whole-brain mapping of distributions of fluorescently labeled neural cell types (83). For each brain region, we took the mean cell density (across 10 repeat experiments) for each of parvalbumin-containing (PV), somatostatin-containing (SST), and vasoactive intestinal peptide-containing (VIP) cells, using data provided from the authors. Cell densities were taken from the atlas described in Erö et al. (84), which inferred cell densities computationally from Nissl stains, and used genetic cell-type markers from AMBA data to estimate cell densities of astrocytes, oligodendrocytes, glia, microglia, neurons, and excitatory and inhibitory neurons.

Gene Set Enrichment Analysis (GSEA). We performed GSEA using gene-score resampling, where continuous scores are assigned to each gene, aggregated as the mean across individual GO categories, and then compared to a null distribution of category-level scores to perform statistical inference on GO categories (36).

We first describe our implementation of conventional GSEA using gene-score resampling against random-gene nulls (ensemble-based nulls are described later). We performed GSEA using Matlab-based software that we developed, available at `https://github.com/benfulcher/GeneCategoryEnrichmentAnalysis`. We used GO Term hierarchy and annotation files were downloaded from GO on 17th April 2019. We first matched annotations to GO Terms (excluding NOT and ND annotations (49)). To match genes listed in the GO annotation file (MGI identifiers) to expression data (NCBI identifiers), we used MouseMine (114). In human, identification was made directly from the gene symbol. Annotations to genes that could not be mapped in this way were ignored. Direct gene-to-category annotations were propagated up to parents within the GO-Term hierarchy by taking the `is_a` keyword to indicate parent-child relationships. This process yielded a fully propagated set of GO categories and their gene annotations. We restricted our analyses to GO categories related to biological processes with between 10 and 200 annotations. We performed enrichment across all GO categories meeting these criteria (we did not perform a hierarchical tree-like statistical search that terminates the search at significant leaf nodes (49)).

Null distributions for a GO category of a given size, s, were generated through 40 000 random samples of s gene scores. Achieving stable p-value estimates from the empirical null distribution generated through permutation testing can require millions of null samples due to the small p-values involved in GSEA. As null distributions of category scores tend to be approximately Gaussian, we chose to estimate p-values from a Gaussian distribution fitted to a given set of null samples. That is, p-values were estimated numerically after fitting a Gaussian distribution to the estimated null distribution. We note that null distributions are not exactly Gaussian, and that this approximation leads to a systematic underestimate of p-values relative to a full, permutation-based approach. As our main focus was to elucidate the mechanisms of bias of GSEA (rather than reporting precise p-values), we used this approximation throughout for computational efficiency. Correction for multiple-hypothesis testing was achieved as a false-discovery rate (115), denoted as q_{FDR} and computed using the `mafdr` function in Matlab.

Analysis and Testing.

Generating spatially autocorrelated maps We generated SBP-spatial ensembles of spatially autocorrelated SBPs using the spatial-lag model (30, 81). Code was written in Matlab, adapted from the Murray Lab's `surrogates` package by Joshua Burt, available at `https://bitbucket.org/murraylab/surrogates/src/master/` (30). Note that a more recent *BrainSMASH* package has since been developed: `https://github.com/murraylab/brainsmash` (65). In the spatial-lag model, for a Euclidean distance matrix constructed for all areas, two parameters determine the generated maps: a parameter controlling the strength of the spatial autocorrelation (relative to noise), ρ, and the characteristic spatial scale, d_0. We set $\rho = 0.8$ and estimated d_0 from an exponential fit to (Pearson) correlated gene expression (across all genes) as a function of distance (18), yielding $d_0 = 1.46$ mm in mouse brain, $d_0 = 1.84$ mm in mouse cortex, and $d_0 = 102.2$ mm in human cortex.

The within-category coexpression score, $\langle r \rangle$ To investigate the role of within-category coexpression in driving false-positive significance, we computed a simple measure of within-category coexpression, $\langle r \rangle$. For each GO

category, we computed the gene × gene coexpression matrix, C, for all genes in that category as Spearman correlation coefficients between the spatial expression pattern of each pair of genes. The within-class coexpression metric, $\langle r \rangle$, was then computed as the mean of the unique (upper diagonal) entries of C.

The spatial autocorrelation score, R^2_{exp} To investigate the role of spatially autocorrelated SBPs in driving differential significance between GO categories, we computed a spatial autocorrelation score, R^2_{exp}, for each GO category. For each category, we computed a correlated gene expression (CGE) value for each pair of brain regions as a Pearson correlation between the expression values of genes in that category. We then fitted a three-parameter exponential function to the variation of CGE with distance, d, as $\text{CGE}(d) = A\exp(-d/\lambda) + B$. The goodness of this fit, measured as R^2_{exp}, provides a measure of how well a given category of genes exhibits a decaying distance-dependent expression similarity. R^2_{exp} will be high for strongly spatially autocorrelated categories of genes. For spatially autocorrelated GO categories, the length scale, λ, provides a measure of the autocorrelation length scale (used in Fig. S4). Note that we obtained similar results when computing a Spearman correlation between d and CGE, which does not assume the exponential functional form. Values of these fitted parameters for each GO category are in Supplementary Files `CategorySpatialScoring_human-cortex.csv` and `CategorySpatialScoring_mouse-all.csv`.

ACKNOWLEDGEMENTS
The authors would like to thank Stuart Oldham for assistance in processing human structural connectivity data and Oliver Cliff for thoughtful discussion and feedback on the manuscript. We thank Colorgorical (116) for assisting with the generation of color palettes. AF was supported by the Sylvia and Charles Viertel Charitable Foundation.

Bibliography

1. G. Buzsáki and A. Draguhn. Neuronal Oscillations in Cortical Networks. *Science* **304**, 1926 (2004).
2. J. W. Lichtman and W. Denk. The Big and the Small: Challenges of Imaging the Brain's Circuits. *Science* **334**, 618 (2011).
3. E. Lein, M. J. Hawrylycz, N. Ao, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168 (2007).
4. M. J. Hawrylycz, E. Lein, A. L. Guillozet-Bongaarts, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489**, 391 (2012).
5. A. Arnatkevičiūtė, B. D. Fulcher, and A. Fornito. A practical guide to linking brain-wide gene expression and neuroimaging data. *NeuroImage* **189**, 353 (2019).
6. A. Fornito, A. Arnatkevičiūtė, and B. D. Fulcher. Bridging the Gap between Connectome and Transcriptome. *Trends in Cognitive Sciences* **23**, 34 (2019).
7. A. Kaufman, G. Dror, I. Meilijson, and E. Ruppin. Gene Expression of Caenorhabditis elegans Neurons Carries Information on Their Synaptic Connectivity. *PLoS Comp. Biol.* **2**, e167 (2006).
8. V. Varadan, D. M. Miller III, and D. Anastassiou. Computational inference of the molecular logic for synaptic connectivity in C. elegans. *Bioinformatics* **22**, e497 (2006).
9. L. Baruch, S. Itzkovitz, M. Golan-Mashiach, E. Shapiro, and E. Segal. Using Expression Profiles of Caenorhabditis elegans Neurons To Identify Genes That Mediate Synaptic Connectivity. *PLoS Comp. Biol.* **4**, e1000120 (2008).
10. A. Arnatkevičiūtė, B. D. Fulcher, R. Pocock, and A. Fornito. Hub connectivity, neuronal diversity, and gene expression in the Caenorhabditis elegans connectome. *PLoS Comp. Biol.* **14**, e1005989 (2018).
11. L. Wolf, C. Goldberg, N. Manor, R. Sharan, and E. Ruppin. Gene Expression in the Rodent Brain is Associated with Its Regional Connectivity. *PLoS Comp. Biol.* **7**, e1002040 (2011).
12. L. French and P. Pavlidis. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Comp. Biol.* **7**, e1001049 (2011).
13. L. French, P. P. C. Tan, and P. Pavlidis. Large-scale analysis of gene expression and connectivity in the rodent brain: insights through data integration. *Front. Neuroinf.* **5**, 12 (2011).
14. S. Ji, A. Fakhry, and H. Deng. Integrative analysis of the connectivity and gene expression atlases in the mouse brain. *NeuroImage* **84**, 245 (2014).
15. M. Rubinov, R. J. F. Ypma, C. Watson, and E. T. Bullmore. Wiring cost and topological participation of the mouse brain connectome. *Proc. Natl. Acad. Sci. USA* **112**, 10032 (2015).
16. A. Fakhry and S. Ji. High-resolution prediction of mouse brain connectivity using gene expression patterns. *Methods* **73**, 71 (2015).
17. A. Fakhry, T. Zeng, H. Peng, and S. Ji. Global analysis of gene expression and projection target correlations in the mouse brain. *Brain Inform.* **2**, 107 (2015).
18. B. D. Fulcher and A. Fornito. A transcriptional signature of hub connectivity in the mouse connectome. *Proc. Natl. Acad. Sci. USA* **113**, 1435 (2016).
19. L. Parkes, B. D. Fulcher, M. Yücel, and A. Fornito. Transcriptional signatures of connectomic subregions of the human striatum. *Genes, Brain and Behavior* **25**, 1176 (2017).
20. P. Goel, A. Kuceyeski, E. LoCastro, and A. Raj. Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain. *Hum. Brain Map.* **35**, 4204 (2014).
21. J. Richiardi, A. Altmann, A.-C. Milazzo, et al. Correlated gene expression supports synchronous activity in brain networks. *Science* **348**, 1241 (2015).
22. I. Diez and J. Sepulcre. Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. *Nat. Comm.* **9**, 1 (2018).
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. *Phil. Trans. Roy. Soc. B* **371**, 20150362 (2016).
24. R. F. Betzel, J. D. Medaglia, A. E. Kahn, et al. Structural, geometric and genetic factors predict interregional brain connectivity patterns probed by electrocorticography. *Nat. Biomed. Eng.* **106**, 1 (2019).
25. J. Ritchie, S. P. Pantazatos, and L. French. Transcriptomic characterization of MRI contrast, focused on the T1-w/T2-w ratio in the cerebral cortex. *NeuroImage* **174**, 504 (2018).
26. K. J. Whitaker, R. Romero-Garcia, F. Váša, et al. Adolescence is associated with genomically patterned consolidation of the hubs of the human brain connectome. *Proc. Natl. Acad. Sci. USA* **113**, 201601745 (2016).
27. J. Seidlitz, F. Váša, M. Shinn, et al. Morphometric Similarity Networks Detect Microscale Cortical Organization and Predict Inter- Individual Cognitive Variation. *Neuron* **97**, 231 (2018).
28. R. Romero-Garcia, K. J. Whitaker, F. Váša, et al. Structural covariance networks are coupled to expression of genes enriched in supragranular layers of the human cortex. *NeuroImage* **171**, 256 (2018).
29. P. K. Reardon, J. Seidlitz, S. Vandekar, et al. Normative brain size variation and brain shape diversity in humans. *Science* **360**, eaar2578 (2018).
30. J. B. Burt, M. Demirtas, W. J. Eckner, et al. Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat. Neurosci.* **27**, 889 (2018).
31. K. H. Preller, J. B. Burt, J. L. Ji, et al. Changes in global and thalamic brain connectivity in lsd-induced altered states of consciousness are attributable to the 5-ht2a receptor. *eLife* **7**, e35082 (2018).
32. I. A. C. Romme, M. A. de Reus, R. A. Ophoff, R. S. Kahn, and M. P. van den Heuvel. Connectome Disconnectivity and Cortical Gene Expression in Patients With Schizophrenia. *Biol. Psychiatry* **81**, 495 (2016).
33. T. Rittman, M. Rubinov, P. E. Vértes, et al. Regional expression of the MAPT gene is associated with loss of hubs in brain networks and cognitive impairment in parkinson disease and progressive supranuclear palsy. *Neurobiology of Aging* **48**, 153 (2016).
34. R. Romero-Garcia, V. Warrier, E. T. Bullmore, S. Baron-Cohen, and R. A. I. Bethlehem. Synaptic and transcriptionally downregulated genes are associated with cortical thickness differences in autism. *Molecular Psychiatry* **24**, 1053 (2019).
35. S. E. Morgan, J. Seidlitz, K. J. Whitaker, et al. Cortical patterning of abnormal morphometric similarity in psychosis is associated with brain expression of schizophrenia-related genes. *Proc. Natl. Acad. Sci. USA* **97**, 201820754 (2019).
36. M. Ashburner, C. A. Ball, J. A. Blake, et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25 (2000).
37. M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acid. Res.* **28**, 27 (2000).
38. A. Subramanian, P. Tamayo, V. K. Mootha, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545 (2005).
39. J. Gillis, M. Mistry, and P. Pavlidis. Gene function analysis in complex data sets using ErmineJ. *Nat. Protoc.* **5**, 1148 (2010).
40. E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
41. G. Dennis Jr, B. T. Sherman, D. A. Hosack, and J. Yang. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* (2003).
42. P. Shannon, A. Markiel, O. Ozier, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498 (2003).
43. H. Mi, A. Muruganujan, and P. D. Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucl. Acid. Res.* **41**, D377 (2013).
44. M. V. Kuleshov, M. R. Jones, A. D. Rouillard, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* **44**, W90 (2016).
45. D. Klopfenstein, L. Zhang, B. S. Pedersen, et al. GOATOOLS: A python library for Gene Ontology analyses. *Scientific Reports* **8**, 10872 (2018).
46. B. Zhang, S. Kirov, and J. Snoddy. Webgestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic acids research* **33**, W741 (2005).
47. S. Carbon, A. Ireland, C. J. Mungall, et al. Amigo: online access to ontology and annotation data. *Bioinformatics* **25**, 288 (2008).
48. J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research* **37**, W305 (2009).
49. S. Y. Rhee, V. Wood, and K. Dolinski. Use and misuse of the gene ontology annotations.

Nat. Rev. Genet. (2008).
50. A. Tomczak, J. M. Mortensen, R. Winnenburg, et al. Interpretation of biological experiments changes with evolution of the gene ontology and its annotations. *Scientific Reports* **8**, 1 (2018).
51. S. Ballouz, P. Pavlidis, and J. Gillis. Using predictive specificity to determine when gene set analysis is biologically meaningful. *Nucl. Acid. Res.* **45**, e20 (2017).
52. X. Dong, Y. Hao, X. Wang, and W. Tian. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Sci. Rep.* **6**, 18871 (2016).
53. A. Hukku, C. Quick, F. Luca, R. Pique-Regi, and X. Wen. Bagse: a bayesian hierarchical model approach for gene set enrichment analysis. *Bioinformatics* **36**, 1689 (2020).
54. L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics* **98**, 1 8 (2011).
55. M. J. Hawrylycz, A. Bernard, C. Lau, et al. Areal and laminar differentiation in the mouse neocortex using large scale gene expression data. *Methods* **50**, 113 (2010).
56. M. J. Hawrylycz, J. A. Miller, V. Menon, et al. Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18**, 1832 (2015).
57. F. M. Krienen, B. T. T. Yeo, T. Ge, R. L. Buckner, and C. C. Sherwood. Transcriptional profiles of supragranular-enriched genes associate with corticocortical network architecture in the human brain. *Proc. Natl. Acad. Sci. USA* **113**, E469 (2016).
58. S. P. Pantazatos and X. Li. Commentary: BRAIN NETWORKS. Correlated gene expression supports synchronous activity in brain networks. Science 348, 1241-4. *Front. Neurosci.* **11**, 412 (2017).
59. J. Richiardi, A. Altmann, and M. Greicius. Distance Is Not Everything In Imaging Genomics Of Functional Networks: Reply To A Commentary On Correlated Gene Expression Supports Synchronous Activity In Brain Networks. *bioRxiv* p. 132746 (2017).
60. J. A. Henderson and P. A. Robinson. Geometric effects on complex network structure in the cortex. *Phys. Rev. Lett.* **107**, 018102 (2011).
61. J. A. Henderson and P. A. Robinson. Relations Between the Geometry of Cortical Gyrification and White-Matter Network Architecture. *Brain Conn.* **4**, 112 (2014).
62. J. A. Roberts, A. Perry, A. R. Lord, et al. The contribution of geometry to the human connectome. *NeuroImage* **124**, 379 (2016).
63. J. Stiso and D. S. Bassett. Spatial Embedding Imposes Constraints on Neuronal Network Architectures. *Trends in Cognitive Sciences* (2018).
64. P. A. Robinson. Physical brain connectomics. *Phys. Rev. E* **99**, 012421 (2019).
65. J. B. Burt, M. Helmer, M. Shinn, A. Anticevic, and J. D. Murray. Generative modeling of brain maps with spatial autocorrelation. *NeuroImage* **220**, 117038 (2020).
66. A. F. Alexander-Bloch, H. Shou, S. Liu, et al. On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage* **178**, 540 (2018).
67. P. G. Reddy, R. F. Betzel, A. N. Khambhati, et al. Genetic and Neuroanatomical Support for Functional Brain Network Dynamics in Epilepsy. *arXiv* p. 1809.03934 (2018).
68. W. T. Barry, A. B. Nobel, and F. A. Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**, 1943 (2005).
69. K. M. Anderson, F. M. Krienen, E. Y. Choi, et al. Gene expression links functional networks across cortex and striatum. *Nat. Comm.* **9**, 1428 (2018).
70. Z. Kuncheva, M. L. Krishnan, and G. Montana. Exploring brain transcriptomic patterns: a topological analysis using spatial expression networks. In *Pac. Soc. Biocomput.* (2017).
71. W. Liu, N. Peeters, G. Fernández, and N. Kohn. Common neural and transcriptional correlates of inhibitory control underlie emotion regulation and memory control. *Social Cognitive and Affective Neuroscience* **15**, 523 (2020).
72. D. Vidal-Pineiro, N. Parker, J. Shin, et al. Cellular correlates of cortical thinning throughout the lifespan. *bioRxiv* p. 585786 (2019).
73. P. P. C. Tan, L. French, and P. Pavlidis. Neuron-Enriched Gene Expression Patterns are Regionally Anti-Correlated with Oligodendrocyte-Enriched Patterns in the Adult Mouse and Human Brain. *Front. Psychiat.* **7** (2013).
74. L. French and T. Paus. A FreeSurfer view of the cortical transcriptome generated from the Allen Human Brain Atlas. *Front. Psychiat.* **9**, 25 (2015).
75. M. Meijer, A. Keo, J. M. C. van Leeuwen, et al. Molecular characterization of the stress network in the human brain. *bioRxiv* **526**, 661587 (2019).
76. B. D. Mills, D. S. Grayson, A. Shunmugavel, et al. Correlated gene expression and anatomical communication support synchronized brain activity in the mouse functional connectome. *J. Neurosci.* **38**, 2910 (2018).
77. K. M. Anderson, M. A. Collins, R. Chin, et al. The transcriptional landscape of cortical interneurons underlies in-vivo brain function and schizophrenia risk. *bioRxiv* p. 481036 (2018).
78. J. Liu, M. Xia, X. Wang, X. Liao, and Y. He. Transcriptional Signatures of the Hierarchical Chronnectome Organization in the Human Brain. *bioRxiv* **39**, 637942 (2019).
79. Z. Liu, E. T. Rolls, J. Zhang, et al. The functional and genetic associations of neuroimaging data: a toolbox. *bioRxiv* p. 178640 (2017).
80. X. Yao, J. Yan, S. Kim, et al. Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *Brain Informatics* **4**, 27 (2017).
81. L. Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media (2013).
82. R. Markello and B. Misic. Comparing spatially-constrained null models for parcellated brain maps. *BioRxiv* (2020).
83. Y. Kim, G. R. Yang, K. Pradhan, et al. Brain-wide Maps Reveal Stereotyped Cell-Type-Based Cortical Architecture and Subcortical Sexual Dimorphism. *Cell* **171**, 456 (2017).
84. C. Erö, M.-O. Gewaltig, D. Keller, and H. Markram. A Cell Atlas for the Mouse Brain. *Front. Neuroinf.* **12**, e17727 (2018).
85. L. Ng, S. D. Pathak, C. Kuan, et al. Neuroinformatics for Genome-Wide 3-D Gene Expression Mapping in the Mouse Brain. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **4**, 382 (2007).
86. E. T. Bullmore and O. Sporns. The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 337 (2012).
87. G. Collin, O. Sporns, R. C. W. Mandl, and M. P. van den Heuvel. Structural and Functional Aspects Relating to Cost and Benefit of Rich Club Organization in the Human Cerebral Cortex. *Cereb. Cortex* **24**, 2258 (2014).
88. D. Tomasi, G. J. Wang, and N. D. Volkow. Energetic cost of brain functional connectivity. *Proc. Natl. Acad. Sci. USA* **110**, 13642 (2013).
89. M. Breakspear, M. J. Brammer, E. T. Bullmore, P. Das, and L. M. Williams. Spatiotemporal wavelet resampling for functional neuroimaging data. *Human Brain Mapping* **23**, 1 (2004).
90. M. Cardillo, L. Bromham, and S. J. Greenhill. Links between language diversity and species richness can be confounded by spatial autocorrelation. *Proceedings of the Royal Society B: Biological Sciences* **282**, 20142986 (2015).
91. S. Afyouni, S. M. Smith, and T. E. Nichols. Effective degrees of freedom of the pearson's correlation coefficient under autocorrelation. *NeuroImage* (2019).
92. O. James, H. Park, and S.-G. Kim. Impact of sampling rate on statistical significance for single subject fMRI connectivity analysis. *Human Brain Mapping* (2019).
93. O. M. Cliff, L. Novelli, B. D. Fulcher, J. M. Shine, and J. T. Lizier. Exact inference of linear dependence between multiple autocorrelated time series. *arXiv preprint arXiv:2003.03887* (2020).
94. B. O. Mancarci, L. Toker, S. J. Tripathy, et al. Cross-laboratory analysis of brain cell type transcriptomes with applications to interpretation of bulk tissue data. *eNeuro* **4** (2017).
95. F. Avila Cobos, J. Alquicira-Hernandez, J. E. Powell, P. Mestdagh, and K. De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications* **11**, 5650 (2020). Number: 1 Publisher: Nature Publishing Group.
96. J. Seidlitz, A. Nadig, S. Liu, et al. Transcriptomic and cellular decoding of regional brain vulnerability to neurogenetic disorders. *Nature Communications* **11**, 3358 (2020). Number: 1 Publisher: Nature Publishing Group.
97. B. D. Fulcher. Discovering Conserved Properties of Brain Organization Through Multimodal Integration and Interspecies Comparison:. *J. Exp. Neurosci.* **13**, 117906951986204 (2019).
98. S. W. Oh, J. A. Harris, L. Ng, et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207 (2014).
99. L. Ng, A. Bernard, C. Lau, et al. An anatomic gene expression atlas of the adult mouse brain. *Nat. Neurosci.* **12**, 356 (2009).
100. J. Arloth, D. M. Bader, S. Röh, and A. Altmann. Re-Annotator: annotation pipeline for microarray probe sequences. *PLOS One* **10**, e0139516 (2015).
101. B. D. Fulcher, M. A. Little, and N. S. Jones. Highly comparative time-series analysis: the empirical structure of time series and their methods. *J. Roy. Soc. Interface* **10**, 20130048 (2013).
102. D. C. Van Essen, S. M. Smith, D. M. Barch, et al. The WU-Minn Human Connectome Project: An overview. *NeuroImage* **80**, 62 (2013).
103. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105 (2013).
104. S. N. Sotiropoulos, S. Jbabdi, J. Xu, et al. Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage* **80**, 125 (2013).
105. M. F. Glasser, T. S. Coalson, E. C. Robinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171 (2016).
106. J.-D. Tournier, F. Calamante, and A. Connelly. MRtrix: diffusion tractography in crossing fiber regions. *Int. J. Imaging Syst. Technol.* **22**, 53 (2012).
107. M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. FSL. *NeuroImage* **62**, 782 (2012).
108. 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).
109. 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).
110. 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).
111. T. Sarwar, K. Ramamohanarao, and A. Zalesky. Mapping connectomes with diffusion MRI: deterministic or probabilistic tractography? *Magn. Reson. Med.* **81**, 1368 (2019).
112. S. N. Sotiropoulos and A. Zalesky. Building connectomes using diffusion MRI: why, how and but. *NMR Biomed.* p. e3752 (2017).
113. M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52**, 1059 (2010).
114. H. Motenko, S. B. Neuhauser, M. O'keefe, and J. E. Richardson. Mousemine: a new data warehouse for mgi. *Mammalian Genome* **26**, 325 (2015).
115. 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).
116. C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics* (2017).