Scaling of gene transcriptional gradients with brain size across mouse development

Lau Hoi Yan Gladys[1,2], **Alex Fornito**[1], **and Ben D. Fulcher**[2,✉]

[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, NSW 2006, Australia.

The structure of the adult brain is the result of complex physical mechanisms acting through development. These physical processes, acting in three-dimensional space, mean that the brain's spatial embedding plays a key role in its organization, including the gradient-like patterning of gene expression that encodes the molecular underpinning of functional specialization. However, we do not yet understand how the dramatic changes in brain shape and size that occur in early development influence the brain's transcriptional architecture. Here we investigate the spatial embedding of transcriptional patterns of over 1800 genes across seven time points through mouse-brain development using data from the Allen Developing Mouse Brain Atlas. We find that transcriptional similarity decreases exponentially with separation distance across all developmental time points, with a correlation length scale that follows a power-law scaling relationship with a linear dimension of brain size. This scaling suggests that the mouse brain achieves a characteristic balance between local molecular similarity (homogeneous gene expression within a specialized brain area) and longer-range diversity (between functionally specialized brain areas) throughout its development. Extrapolating this mouse developmental scaling relationship to the human cortex yields a prediction consistent with the value measured from microarray data. We introduce a simple model of brain growth as spatially autocorrelated gene-expression gradients that expand through development, which captures key features of the mouse developmental data. Complementing the well-known exponential distance rule for structural connectivity, our findings characterize an analogous exponential distance rule for transcriptional gradients that scales across mouse brain development, providing new understanding of spatial constraints on the brain's molecular patterning.

gene expression | transcription | gradients | brain development
Correspondence: ***ben.fulcher @sydney.edu.au***

Introduction

Brain structure is the result of physical mechanisms playing out through development, shaped by genetic and environmental factors. In particular, the embedding of neural tissue in physical space places major constraints on neural organization. For example, the connection probability between pairs of neural elements decays with their separation distance in *C. elegans* (1), mouse (2, 3), rat (4), zebrafish (5), non-human primate (6), and human (7), and is frequently characterized as an exponential distance rule (8–10). Indeed, very simple spatial rules can explain much of the high-order statistics of connectome topology, including its modularity, long-tailed degree distribution, local network motif probabilities, and the existence of a network core (7, 9, 11–15). Spatial proximity also plays a strong role in shaping gradients of gene expression, with nearby areas exhibiting more similar gene-expression profiles than more distant areas in the head neurons of *C. elegans* (1), the mouse brain (2, 16), and the human cortex (17–22).

The spatial embedding of gene-expression patterns provides important clues about how the brain's functional specialization is organized. Macroscopic functional gradients across the human cortex maximize the spatial distance between areas functionally involved in sensory perception and those involved in integrative cognition (23, 24). These hierarchical functional gradients are underpinned by a corresponding variation in structural microarchitecture (25–27). But given the constraint of a fixed brain size, how does the brain strike the right balance between local molecular similarity (similar gene-expression patterns supporting regional specialization of function), with inter-regional molecular diversity, which supports functional heterogeneity. Furthermore, how does the spatial embedding of transcriptional gradients emerge through development? Do different spatial constraints dominate early in development compared to later? Although macroscopic brain organization in adult is often attributed to gradients set up during development (e.g., in transcriptional factors) (28), to our knowledge, no study has analyzed the spatial embedding of gene expression through development.

The Allen Developing Mouse Brain Atlas (ADMBA) (29) is a comprehensive database of gene-expression across mouse-brain development. The atlas contains *in situ* hybridization measurements for approximately 2100 genes at seven developmental time points, available in a three-dimensional reference space registered to an anatomical reference atlas at each developmental

time point. Many analyses of gene transcription through mouse-brain development have shed light on how the spatiotemporal variation of specific genes through development shape different aspects of brain structure (29–32). But these developmental atlas data have not previously been analyzed from the viewpoint of spatially embedded gradients of transcriptional similarity. By characterizing gradients, we can address global questions about transcriptional patterning through development in a parcellation-free way that does not involve the difficult task of tracking anatomically defined areas.

Here we study how correlated gene expression (CGE), defined as the similarity in the gene-expression signatures of two points in the brain, is spatially embedded across mouse-brain development. Matching the well-known exponential distance rule for connectivity, we find an exponential decay in CGE at each developmental time point, with a spatial correlation length that scales with a linear dimension of brain size. Extrapolating the relationship fitted on developing mouse-brain data allows us to obtain an estimate consistent with the measured correlation length of CGE in the human cortex. Using a simple model, we show that the patterns are consistent with brain growth in which spatially autocorrelated gene-expression gradients expand with brain size over time.

Methods

Data. We analyze data from the ADMBA, which contains *in situ* hybridization (ISH) expression estimates for approximately 2100 genes in C57Bl/6J mice across seven developmental stages, spanning both embryonic (E11.5, E13.5, E15.5, and E18.5) and postnatal (P4, P14, and P28) stages (29). These genes were selected to encompass relevant transcription factors, neurotransmitters and their receptors, cell-type or regional marker genes, genes relevant to developmental signaling pathways, and genes corresponding to common drug targets or implicated in neurodevelopmental disorders (29). The raw imaging data, encompassing 434 946 brain sections of high-resolution ISH, are registered to a three-dimensional mouse-brain atlas at each time point. We accessed this registered data at the voxel level via the Allen Software Development Kit (SDK, 2015). Our code for retrieving data (in python) and for data processing and analysis (Matlab) is available at `https://github.com/NeuralSystemsAndSignals/DevelopingMouse`. At each time point, we retrieved expression data for all available genes. We used sagittal section data for its comprehensiveness, which had coverage across the left hemisphere at each time point. Expression data for each gene was obtained as 'expression energy' (ISH signal intensity) in each voxel in the three-dimensional grid. Distances between pairs of voxels were computed as Euclidean distances in three-dimensional space using the spatial resolution of the coordinate grid at each time point: E11.5 (0.08 mm), E13.5 (0.1 mm); E15.5 (0.12 mm), E18.5 (0.14 mm), P4 (0.16 mm), P14 (0.2 mm), and P28 (0.2 mm).

Filtering and normalization. At each developmental time point, we first filtered out voxels that were labeled as spinal cord or were unannotated, and labeled voxels according to four anatomical masks: 'forebrain', 'midbrain', 'hindbrain', 'dorsal pallium'. At each time point, we retained genes with valid expression (i.e., not a missing value) for at least 70% of voxels, resulting in different numbers of good-quality genes across time: E11.5 (2038 genes), E13pt5 (2085), E15pt5 (1861), E18pt5 (2053), P4 (2056), P14 (2015), and P28 (2063). These small differences in the number of genes across time did not affect our results: we obtained similar findings when restricting our analysis to a consistent subset of 1861 genes that had good-quality data across all time points. We then retained only voxels with valid expression data for at least 70% of these genes. The number of voxels varied across time as 5034 (E11.5), 9473 (E13.5), 11 314 (E15.5), 11 313 (E18.5), 19 754 (P4), 21 579 (P14), and 24 822 (P28). We generated a voxel × gene expression matrix containing voxels and genes that survived these filtering steps.

Raw expression values measured using *in situ* hybridization are not comparable between genes (33). We accounted for this at each time point by normalizing expression values of each gene using an outlier-robust sigmoidal transformation, $S(x) = [1 + \exp(-(x - x_{\mathrm{med}})/1.35x_{\mathrm{iqr}})]^{-1}$, where x_{med} is the median of the expression vector, x, for a given gene, and x_{iqr} is its interquartile range (17, 34). For visualization, this normalization was followed by linear rescaling to the unit interval. The resulting normalized voxel × gene matrix, G_t, at a given time point, t, was used for further analysis.

Quantifying the spatial embedding of CGE. To understand the spatial embedding of gene-expression patterns, we analyzed the distance-dependence of correlated gene expression (CGE) at each developmental time point. CGE is a measure of gene-expression similarity between a pair of voxels and was computed as the Pearson correlation coefficient between normalized gene-expression vectors (rows of a given voxel × gene matrix, G_t, defined above).

At each time point, we computed CGE for every pair of voxels. Because the distribution of pairwise distances is concentrated near lower distances, we wanted to prevent the exponential fitting from being biased towards fitting these smaller length scales, but to instead fit well across the full range of distances. To do this, we first performed an equiprobable distance binning of the data using twenty bins, and computed the mean CGE value in each bin (and summarizing each bin as its center). To this binned data, we then fitted a three-parameter exponential function, $\mathrm{CGE}(d) = A\exp(-d/\lambda) + f_0$, using nonlinear least squares and identified the fitted parame-

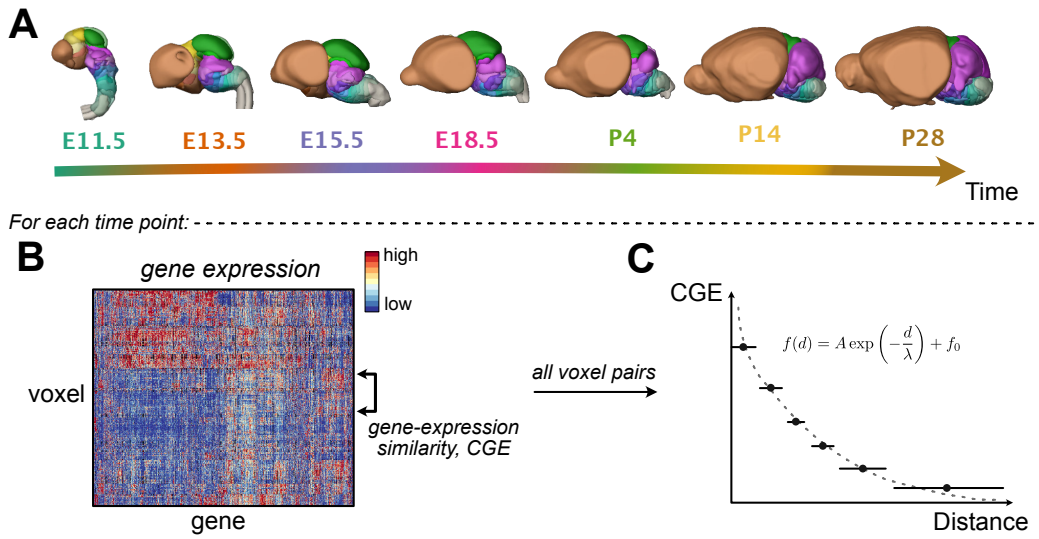

Fig. 1. Schematic analysis pipeline. **A** We analyze gene-expression atlases across seven developmental time points. **B** At each time point, we created a voxel × gene expression matrix (data shown for E11.5). We then compared the correlated gene expression (CGE) between each pair of voxels as a function of their physical separation distance, d. **C** Plotting $\mathrm{CGE}(d)$ across equiprobable distance bins allowed us to analyze the spatial embedding of gene expression at each time point. We fitted an exponential form $\mathrm{CGE}(d) = A\exp(-d/\lambda) + f_0$. Here we focus on the correlation length, λ, and how it scales with brain development.

ters as: spatial correlation length, λ, strength A, and offset, f_0. Adjusted R^2 was used as a goodness-of-fit statistic. Note that for self-correlation, at $d = 0$, $\mathrm{CGE} = 1$, and thus fitting data that includes self-correlations should have a fixed strength parameter $A = 1$. Here we excluded self-correlations and allowed the strength parameter, A, to vary. When characterizing distances relative to brain size, as $d_{\mathrm{rel}} = d/d_{\max}$, we defined $d_{\max}$ as the maximum extent along the anterior–posterior axis of the brain. For computational efficiency, we restricted our analysis to a random subset of 1000 voxels at each developmental time point. Our results do not strongly depend on this spatial downsampling (e.g., λ estimates stabilized for samples of approximately 500 voxels, cf. Fig. S4).

Note that our definition of λ as a length scale (m) is consistent with its conventional use to denote wavelength in physics (units of length), but differs by an inverse relationship from its use as a decay rate with units of inverse length (m^{-1}) used to characterize the exponential distance rule for structural connectivity (4, 8–10).

To estimate and visualize the dominant expression pattern across all genes in matrices, G_t, we used principal components analysis, accounting for missing values using alternating least squares.

Modeling. To better understand patterns found in the empirical data, we simulated simple spatial models of gene expression patterning and brain growth. At each developmental time point, we first simulated gene-expression patterns for each of the 1861 genes. We first defined the geometry by approximating the brain as a rectangular prism with dimensions (anterior–posterior × superior–inferior × left–right) equal to that of the developmental reference atlas at that time point: $3.52 \times 4.72 \times 1.20$ mm (E11.5), $5.70 \times 3.30 \times 1.20$ mm (E13.5), $6.60 \times 3.72 \times 1.68$ mm (E15.5), $7.84 \times 4.06 \times 2.10$ mm (E18.5), $10.56 \times 5.44 \times 3.04$ mm (P4), $12.80 \times 6.60 \times 4.20$ mm (P14), and $13.60 \times 7.40 \times 4.20$ mm (P28). We formed a grid over this idealized geometry by dividing each dimension into 50 equally spaced points. We then generated a random spatially autocorrelated map independently for each of 1861 genes using a spatial-lag model (25, 35, 36). The model defines a dependence between samples using an exponential kernel as $W_{ij} = \exp(-d_{ij}/d_0)$, for pairwise distances, d_{ij}, and a characteristic spatial scale, d_0. From this weighting matrix, W_{ij}, an autocorrelated spatial map can be generated as $x_i = (I + \rho W_{ij})u_j$, where I is the identity and $u_j \sim \mathcal{N}(0,1)$ is i.i.d. Gaussian noise. This defines the second model parameter: the spatial autocorrelation strength, ρ. By setting d_0 as a fixed fraction of brain size, $d_0 = d_{\max}/d_0^{\mathrm{scale}}$, we simulated a brain-size scaling rule consistent with a linear rescaling of space (spatial expansion through development). Values for ρ and d_0^{scale} were set through a gradient descent optimization procedure to best fit the empirical values of λ and A, yielding $\rho = 0.416$ and $d_0^{\mathrm{scale}} = 15.6$. Some illustrative examples of the expression patterns generated are plotted in Fig. 6A.

Note that at small distances, all grid points are included in the CGE calculation. But due to the grid's finite size, beyond a critical distance, only a subset of grid points are included in the CGE calculation (e.g., in the extreme case of the maximum distance, only the corners of the cube are included). This finite-size spatial sampling effect can bias the $\mathrm{CGE}(d)$ curve at large distances. For example, distance bins that are affected by this are labeled pink in Fig. 6B.

For each simulated gene-expression dataset, we performed the same processing and analysis methodology as was applied to the empirical data: (i) a random subsample of 1000 points was taken; (ii) voxel × gene expression data were normalized using a scaled sigmoid; (iii) CGE was computed; (iv) $\mathrm{CGE}(d)$ data were binned using 20 equiprobable bins; (v) exponential decay was fit. We repeated the process 50 times to estimate error bars on each parameter estimate under the sources of stochasticity in the model: the generation of spa-

tially autocorrelated gene-expression maps, and the random spatial subsampling. We were then able to compare the parameter estimates from this model with the empirical data. Code for the modeling is available at https://github.com/NeuralSystemsAndSignals/DevelopmentalExpressionModeling.

Results

Our approach to investigating the spatial embedding of transcriptional gradients across mouse-brain development is shown schematically in Fig. 1. At each of the seven time points in the ADMBA (29), shown in Fig. 1A, we obtained the expression of 1861 genes across voxels of the mouse brain (Fig. 1B). We then computed the correlated gene expression (CGE) for each pair of voxels as the Pearson correlation coefficient between their expression signatures (rows of the matrix in Fig. 1B). We fitted an exponential function to the decay of CGE(d) as a function of separation distance, d, shown in Fig. 1C, to quantify how transcriptional similarity is spatially embedded. Variation in the parameters of this fit, particularly that of the characteristic length scale over which genes are correlated, λ, tell us how the properties of gene-expression gradients vary across mouse-brain development.

Low-dimensional spatial gradients. To better understand the dominant patterns of gene expression in the mouse brain, we first visualized voxel $\times$ gene expression matrices for each time point, and projected their leading principal components into brain space. Examples for two selected time points are shown in Fig. 2 (see Fig. S1 for all time points). Gene-expression matrices have been reordered to put genes with similar expression close to each other (columns of the expression matrices in Figs 2A,C) and, within three main divisions—forebrain, midbrain, and hindbrain—voxels with similar expression signatures close to each other. Visually, we find interesting gene-expression patterns that clearly differ between the three major brain divisions. The similarity-based ordering of columns in these expression matrices makes the low-dimensionality of the expression data visually clearer: we see how large numbers of genes display highly correlated expression patterns. The leading principal component of the expression data (computed using alternating least squares to account for a small proportion of missing data) is annotated to the left of Figs 2A,C, and visually follows the types of dominant gene-expression patterns that capture the most variance across all genes.

To determine whether these low-dimensional components are spatially embedded as gradients, we plotted them in Figs 2B,D for E15.5 and P28, respectively. We find clear spatial autocorrelation in these low-dimensional spatial gradients, an effect that is also seen across the other developmental time points (Fig. S1), and for lower-order principal components (not shown). Having built some intuition for the existence of low-dimensional spatial gradients of gene-expression in the developing mouse brain, in the remainder of this study we aim to quantitatively these gradients across development using CGE(d) curves.

Distance dependence of correlated gene expression. Binned CGE(d) data with exponential fits are shown for each of the seven developmental time points in Figs 3A–G. The three-parameter exponential form captured the data well at all time points (all adjusted $R^2 > 0.96$). This exponential distance-dependence of gene-expression similarity provides a quantitative way to characterize the gradient-like expression patterns shown in Figs 2B,D.

We next investigated whether the CGE(d) curves shown in Fig. 3, computed using all voxels, vary as a function of anatomical division, focusing on the forebrain, midbrain, and hindbrain. We recomputed CGE(d) curves for each type of voxel pair: forebrain–forebrain, forebrain–midbrain, midbrain–midbrain, midbrain–hindbrain, or hindbrain–hindbrain, shown in Fig. S2. Most of these anatomically distinguished CGE(d) curves decayed with distance, but with different length scales and strengths. There were some deviations, such as an increase in midbrain–midbrain CGE at $\sim 3\,\mathrm{mm}$ at E18.5, which causes a slight 'hump-like' deviation from an exponential form for the bulk CGE(d) curve (Fig. 3D). Consistent with an increased gene-expression similarity within each anatomical division, within-division CGE was higher than between-division CGE at a given separation distance, d, across all developmental time points. Indeed, within-division CGE was almost always positive, even at the longest separation distances, whereas between-division CGE often became negative (i.e., gene-expression signatures are anti-correlated), even at intermediate separation distances. This between-division anti-correlation contributes to the overall negative baseline, f_0, of the exponential fits to bulk CGE(d) in Fig. 3. The increase in CGE towards ~ 0 at the longest separation distance bin—seen for the postnatal times P4, P14, and P28—can mostly be attributed to a decrease in anti-correlation of expression profiles between forebrain and hindbrain voxels. Postnatal time points were also distinguished by high expression similarity within forebrain voxels, even at the longest separation distances (Figs S2E–G). While the focus of this study is to investigate CGE(d) curves using voxels from across the whole brain, we note that this global gradient-like spatial embedding is not isotropic and non-specific, but indeed varies with anatomy.

We next investigated whether the CGE(d) curves vary across subsets of neuron-enriched, astrocyte-enriched, and oligodendrocyte-enriched genes (defined as being more than 1.5-fold enriched using data from Cahoy et al. (37)). Compared to using all genes, CGE(d) curves and relationships between fitted CGE decay parameters with brain size, remained similar for all gene subsets, indicating that the spatial embedding of gene expression does

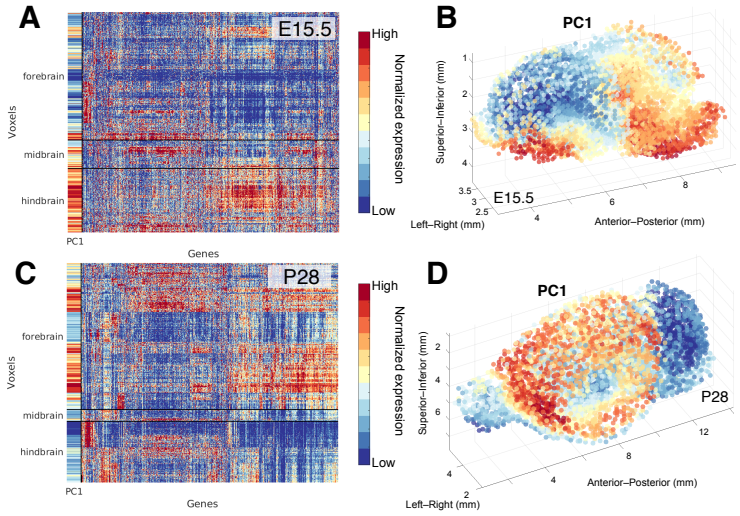

Fig. 2. Developmental mouse gene-expression patterns display low-dimensional spatial gradients. Here we plot results for E15.5 and P28 (see Fig. S1 for all time points). We show voxel × gene expression matrices as heat maps for E15.5 and P28 in **A** and **C**, respectively. All brain voxels and genes passing quality-control criteria are shown ($11\,314 \times 1851$ for E15.5, and $24\,797 \times 1846$ for P28). Each gene has been normalized according to an outlier-robust sigmoidal transform, followed by rescaling to the unit interval, from relative low expression for a given gene (blue) to relative high values (red). To visualize the structure in these data, genes have been reordered by their correlation similarity, and within each labeled subdivision (forebrain, midbrain, hindbrain), voxels have been similarly reordered. Horizontal black lines distinguish voxels from these three major subdivisions. The leading principal component of each matrix is labeled as 'PC1', and visually follows the dominant expression pattern across genes. These leading principal components are plotted in three-dimensional brain space in **B** and **D** for E15.5 and P28, respectively. To aid visualization of the three-dimensional structure, a random subset of 5000 voxels is plotted with transparency. There is clear spatial autocorrelation in these low-dimensional expression gradients in the developing mouse brain, which we quantify in this work.

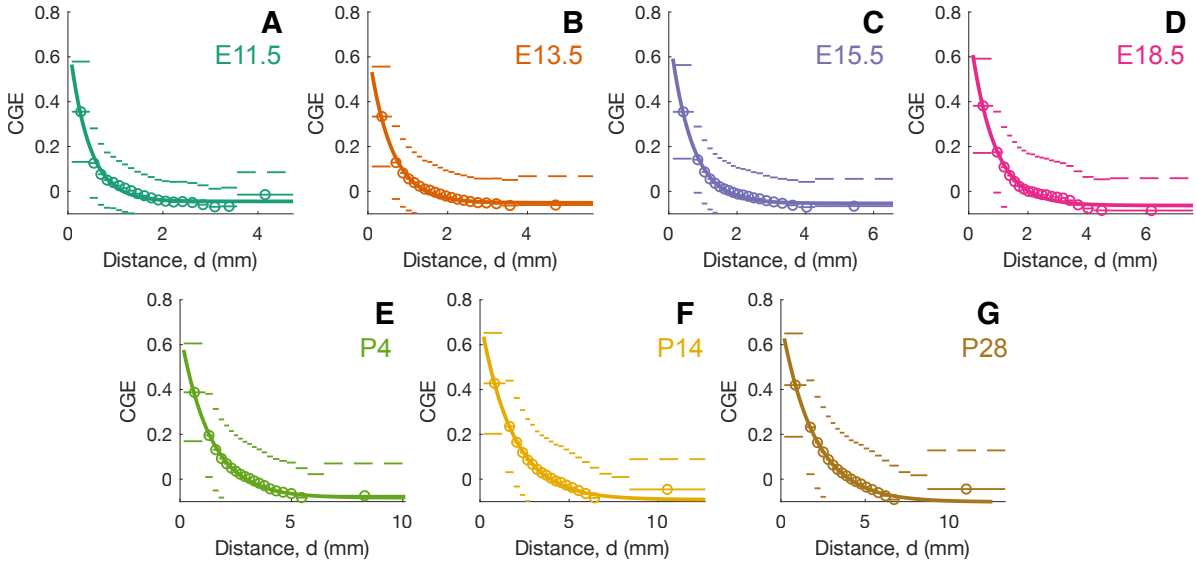

Fig. 3. Correlated gene expression (CGE) exhibits an approximately exponential decay with distance at all developmental time points. A–G: Across twenty equiprobable distance bins, mean CGE of all pairs of voxels in each bin is plotted as circles, with horizontal lines showing the extent of each bin. Dashed lines indicate one standard deviation either side of the mean of each bin. Fitted exponential curves, $\mathrm{CGE}(d) = A\exp(-d/\lambda) + f_0$, are shown solid.

not vary strongly with genes enriched for each of these cell types.

Scaling of transcriptional correlation length with brain size. When plotting the fitted $\mathrm{CGE}(d)$ exponential curves together (Fig. S3A) and rescaling distances by brain size, $d_{\max}$ (maximal anterior–posterior distance), as $\hat{d} = d/d_{\max}$ (Fig. S3B), we found that $\mathrm{CGE}(d)$ approximately collapsed onto a common relative correlation length. To quantify this relationship, we plotted the fitted transcriptional correlation length, λ, as a function of brain size, $d_{\max}$, in Fig. 4A. We find that λ increases with $d_{\max}$ through development (Pearson's $r = 0.98$, $p = 6 \times 10^{-5}$). Having only seven points over less than a decade of brain sizes, $d_{\max}$ prevents rigorous inference of power-law scaling (38), but we have physical reasons to expect a scaling relationship across brain expansion (10). To explore this possibility, we fitted a power-law, $\lambda = 0.073 \times d_{\max}^{1.22}$, shown in Fig. 4B.

The spatial embedding strength, A, controls the strength of the exponential relationship in CGE relative to other effects such as noise. As shown in Fig. 4C, this strength is similar across development, $A \approx 0.75$ (Pearson's $r = 0.65$, $p = 0.1$). The offset, f_0, corresponds to the CGE in the limit of large separation distances, d. As plotted in Fig. 4D, f_0 decreases linearly, becoming more negative across development (Pearson's $r = -0.99$, $p = 3 \times 10^{-5}$). These increasingly anti-correlated gene-expression patterns at the longest separation distances are consistent with an overall increase in transcriptional diversity within the brain across development.

Predicting the transcriptional correlation length in human cortex. We next aimed to test the power-law form for the scaling of transcriptional correlation length, λ, which suggests that the mouse brain maintains a characteristic spatial correlation length of transcriptional simi-

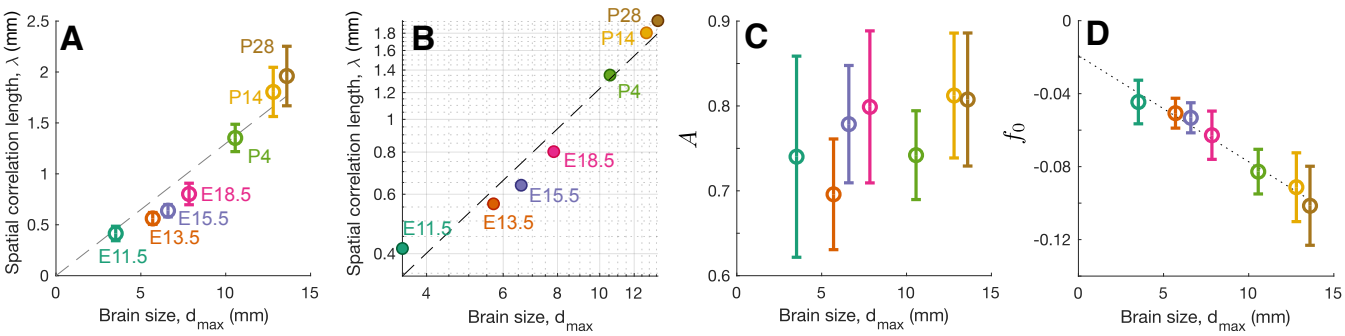

Fig. 4. The correlation length of transcriptional coupling, λ, increases with brain size. The three fitted parameters of the exponential form $\mathrm{CGE}(d) = A\exp(-d/\lambda) + f_0$, vary through brain development, plotted here as a function of brain size, $d_{\max}$ (computed along the anterior–posterior dimension): **A** transcriptional correlation length, λ (mm) (with proportional linear fit dashed); **B** $d_{\max}$–λ as in A but plotted on logarithmic axes with the best power-law fit, $\lambda = 0.0735 d^{1.22}$, shown dashed; **C** spatial embedding strength, A; and **D** offset, f_0 (with linear fit shown dotted). Each data point represents the fitted parameter value from a given time point (labeled in **A** and **B**). Error bars as symmetric 95% confidence intervals are plotted on the three linear-axis plots, A, C, and D.

larity relative to its size. We also wanted to investigate whether the relationship found across mouse-brain development could be extrapolated to make a prediction of the transcriptional correlation length, λ, in human cortex. We used existing data for human cortex, $d_{\max} \approx 148\,\mathrm{mm}$ (maximal anterior–posterior distance in the Glasser et al. (39) parcellation) and $\lambda_{\mathrm{human}}^{\mathrm{meas}} = 61.4\,\mathrm{mm}$ (95% confidence interval: 54.1–71.0 mm (17)). The extrapolation of the power-law relationship fitted to the mouse development data, $0.073 d_{\max}^{1.22}$, and setting $d_{\max} = 148\,\mathrm{mm}$ yields a predicted $\lambda_{\mathrm{human}}^{\mathrm{pred}} = 33.4\,\mathrm{mm}$ (95% CI: 12.6 mm–88.7 mm). This extrapolation is shown in Fig. 5A. The prediction is consistent with the measured value, as shown in Fig. 5B (for all voxels, labeled 'brain'). Thus, despite differences in data type (ISH for < 2000 genes in mouse, microarray for $> 20\,000$ genes in human), spatial discretization (at the level of voxels in mouse and a multimodal parcellation in human), we were able to use mouse developmental data to predict a key spatial property of human gene-expression gradients.

We next investigated how this prediction varied when we restricted the fitting to a subset of voxels (dorsal pallium, forebrain, midbrain, and hindbrain), and when using a proportional linear scaling instead of the power-law form (Fig. 5B). We expected the closest estimate when using dorsal pallium voxels, as this is the structure that best corresponds to the cortex. Indeed, dorsal pallium voxels yielded the best human prediction $\lambda_{\mathrm{human}}^{\mathrm{pred}} = 47.3\,\mathrm{mm}$ (despite the dorsal pallium not being annotated to the reference atlas at E11.5). The worst human prediction was obtained using midbrain voxels, $\lambda_{\mathrm{human}}^{\mathrm{pred}} = 6.96\,\mathrm{mm}$. Proportional linear scaling of λ with brain size, $\lambda \propto d_{\max}$ across mouse development consistently yielded a poorer prediction, supporting a power-law form, $\lambda \propto d_{\max}^{\nu}$, with exponent $\nu > 1$. With just seven points across mouse-brain development, our predictions have wide error bars; it will be important for future work to verify the results of this scaling law, and produce more constrained predictions using more comprehensive transcriptomic data, including across species.

Modeling brain growth. To understand potential mechanisms underlying the scaling of transcriptional gradients through brain growth, we developed a simple spatial model of transcriptional patterning and tested its predictions against the empirical data characterized above. As described in Methods, we generated a random spatially autocorrelated map independently for each gene's expression pattern, evaluated in an idealized prism geometry that matches the brain's three-dimensional spatial extent at each developmental time point. To test whether the observed patterns are consistent with a simple rescaling of space in these expression maps, we rescaled the spatial autocorrelation length, d_0, as a function of brain size, $d_{\max}$. The model is therefore similar to uniform, isotropic stretching of three-dimensional gene-expression patterns that exist at the earliest developmental time point (it reduces to this case when all brains have the same relative dimensions). The types of spatially autocorrelated patterns that were generated are illustrated in Fig. 6A.

We first verified that our model reproduces an approximately exponential dependence of CGE with distance at a given time point, shown in Fig. 6B for E11.5. Apart from a slight decrease of $\mathrm{CGE} < 0$ at moderate distances, and an increase back to $\mathrm{CGE} \approx 0$ at the highest distances, we find a good fit to an exponential form. Note that the distance bins labeled pink in Fig. 6B are sufficiently large relative to brain size that they involve restricted sampling of brain space—only voxels towards the edge of the brain are able to contribute to the CGE calculation at these large distances (whereas all voxels contribute at shorter distances). The qualitative similarity between the simulated $\mathrm{CGE}(d)$ curve and the measured $\mathrm{CGE}(d)$ curves from the ADMBA data, suggest that an ensemble of genes with independent random, spatially autocorrelated transcriptional patterns can provide a good approximation for understanding how these patterns adjust through development.

After fitting the two parameters of our model, ρ and d_0^{scale}, to the empirical data (see Methods), the model approximately reproduces the scaling of the correlation

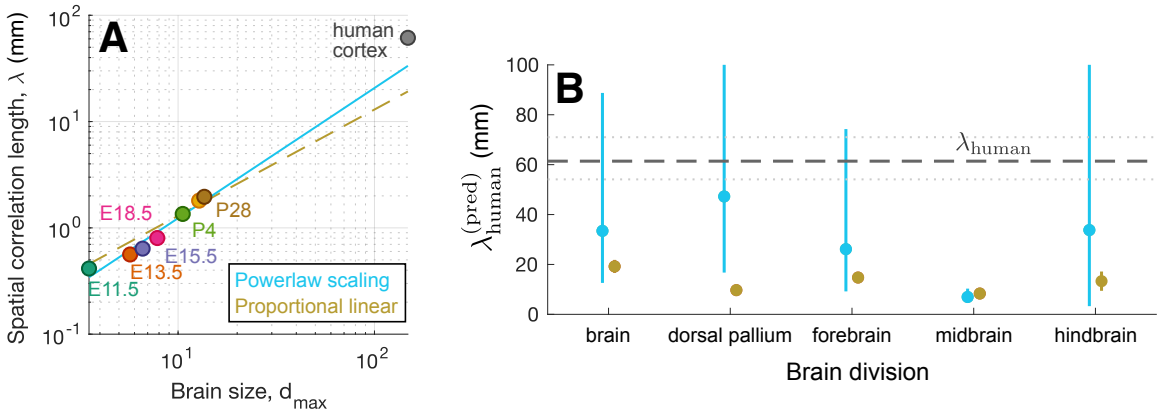

Fig. 5. The transcriptional correlation length in human cortex, λ_{human}, can be predicted by extrapolating the scaling relationship between λ and brain size through mouse-brain development. **A** Estimates of λ plotted as a function of brain size, $d_{\max}$, using all voxels at each of the mouse developmental stages. Extrapolation to the human cortex ($d_{\max} = 148$ mm) is shown using a fitted power-law scaling relationship, $\lambda = 0.073 d_{\max}^{1.22}$ (blue), and a proportional linear relationship, $\lambda = 0.13 d_{\max}$ (brown dashed). **B** We plot predictions trained from fitting to: all brain voxels, dorsal pallium (isocortex) voxels, forebrain voxels, midbrain voxels, and hindbrain voxels. We also compare the predictions of power-law scaling (blue) to those from a proportional linear scaling (brown), with 95% confidence intervals annotated.

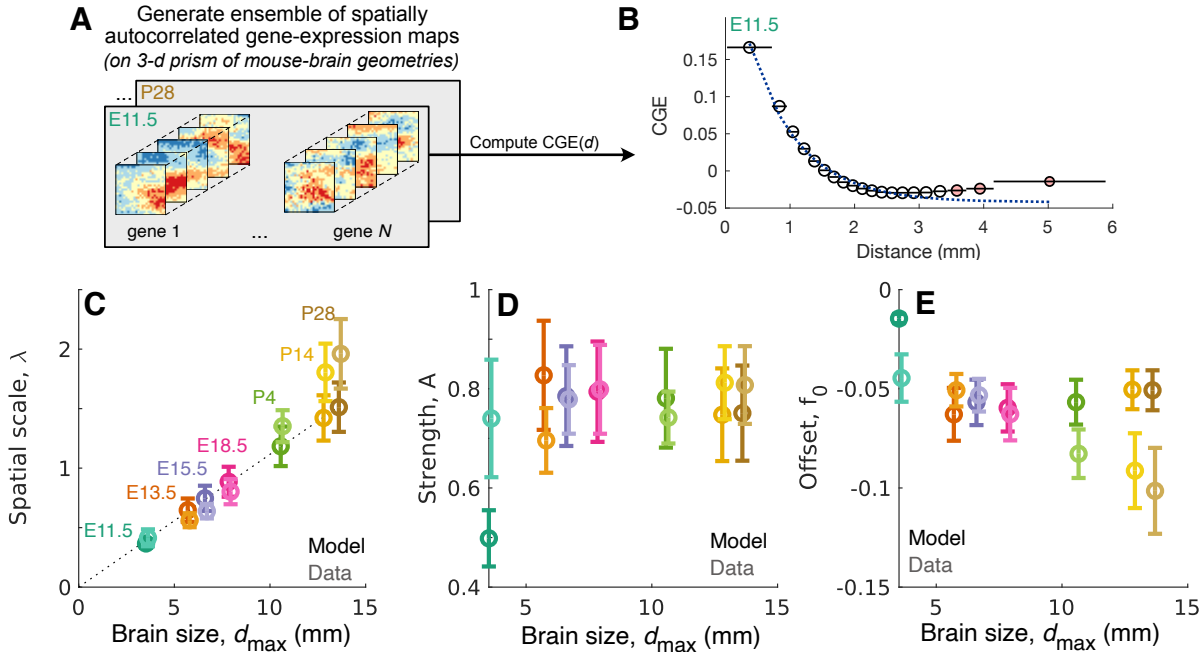

Fig. 6. A simple spatial model for gene expression through development reproduces key features of the empirical data. **A** Illustrative ensemble of independent, spatially autocorrelated gene-expression maps evaluated on three-dimensional prisms matching the dimensions of the brain at each point in development. Examples are shown for two genes in a three-dimensional volume (expression levels shown as color). **B** Each ensemble exhibits an approximately exponential decay of correlated gene expression (CGE) with distance, d, as $\text{CGE}(d) = A\exp(-d/\lambda) + f_0$. Binned, model-simulated data (black) and fit (dotted blue) are plotted here for E11.5. Distance bins for which not all grid points were sampled (due to finite-size effects) are filled pink. Exponential fit parameters of the model-simulated data (shown with dark colors) and empirical data (shown with light colors) are plotted across development for: **C** spatial correlation length, λ; **D** strength, A; and **E** offset, f_0. To aid visibility, empirical values are slightly offset horizontally.

length, λ, with brain size, $d_{\max}$, seen in empirical data (Fig. 6C), and the strength, A (Fig. 6D).

Discussion

Here we used data from the ADMBA to characterize how gene transcriptional gradients are spatially embedded across mouse brain development. Complementing the well-known exponential distance rules for structural connectivity in mammals (3, 4, 8, 9) that are connected by a scaling rule (10), we show that correlated gene expression is also well characterized by an exponential distance rule at each point in mouse brain development. Furthermore, the characteristic correlation length of this embedding, λ, was found to scale with a linear dimension of brain size, $d_{\max}$, as $\lambda \propto d_{\max}^{1.22}$ (Fig. 4). More functionally specialized brains may be expected to encode more molecularly distinctive brain areas within the constraints of its brain size, corresponding to a shorter relative λ. This scaling relationship suggests that the brain maintains a spatial trade-off between local molecular coordination (uniform gene expression within a specialized brain area) and longer-range molecular differentiation (between molecularly distinct, functionally spe-

cialized areas) across development. We were able to use this scaling law to predict the transcriptional correlation length in the adult human cortex (Fig. 5). Given that the brain-related genes exhibit strong consistency in their hierarchical patterning between mouse and human (27), it will be interesting for future work to further investigate these types of cross-species correspondences, including characterizing whether specific classes of genes drive similarities or differences in the spatial organization of gene expression across species.

A benefit of our method of characterizing CGE(d) curves from voxel-wise expression maps is that it does not rely on any assumptions about pre-defined anatomical parcellations or areal boundaries. Our visualizations of the voxelwise expression as clustered voxel $\times$ gene expression matrices, accompanied by their leading principal component (Figs 2A,C), show that there are low-order expression patterns that explain much of the variance across all available genes. When plotting this leading principal component in space (Figs 2B,D), we see strong gradient-like spatial autocorrelation that is quantified here (across all genes) using CGE(d) curves. While these analyses provide clear evidence for spatial expression gradients in the brain, they are not simple, non-specific spatial patterns, but vary markedly across anatomical divisions. This is seen most clearly when distinguishing CGE(d) by pairs of voxels that lie within or between each of the three major anatomical subdivisions: forebrain, midbrain, and hindbrain (Fig. S2). Consistent with characteristic expression patterns for different divisions, seen visually in Figs 2A,C, we found that within-division CGE is higher than between-division CGE at a given separation distance. Future work could take these types of analyses further to combine the benefits of a parcellation-free characterization that facilitates comparisons across development, while also investigating how properties of these spatial gradients vary as a function of anatomy. This includes the ability to score and interpret specific types of genes that contribute most to interesting patterns, such as characterizing differences between individual genes in the strength and spatial scale of their gradient-like patterning, and how these properties vary across development. The type of quantitative spatial characterization used here could also be used to construct new, parcellation-free definitions for 'differentially expressed genes' to test whether the hourglass pattern of transcriptional variation holds in mouse (40).

Our analysis was performed at the voxel level using *in situ* hybridization data from the ADMBA. This allowed us to explore CGE(d) at a much finer spatial resolution than previous analyses of correlated gene expression, performed at the level of parcellations in adult mouse brain (2) and human cortex (17). Nevertheless, each spatial element in our analysis still contains transcriptional contributions from a large number of molecularly distinct brain cells. The advent of three-dimensional *in situ* single-cell transcriptome profiling (41–44), which preserves spatial and transcriptomic information at the cellular level, will allow future analyses of correlated gene expression to be performed at even finer spatial scales. For example, this could allow testing of whether the bulk scaling rule for CGE characterized here applies at the finer scale of neuronal circuits. As transcriptomic atlas data become available for other species, such as *Drosophila melanogaster* (45, 46) and zebrafish (47), we will also be able to more comprehensively compare the spatial organization of the brain's molecular diversity, and assess the cross-species validity of the scaling rule proposed here.

A simple model of independent, spatially autocorrelated transcriptional maps that scale with brain size (similar to the isotropic expansion of transcriptional patterns set up early in development) reproduced key features of CGE(d) scaling observed empirically. We kept the spatial autocorrelation scale for our simulated gene-expression maps at a fixed proportion of brain size, consistent with linear, isotropic brain growth (but allowing us to use realistic brain dimensions from empirical measurements). Our results are thus consistent with molecular specialization occurring very early in development, with brain expansion stretching the correlation length through development. While our simple isotropic model reproduces much of the CGE(d) relationship, our model is only evaluated against these coarse-grained properties, having blurred the contribution of the many individual genes that display precise developmental spatiotemporal programs, and the variation in these CGE(d) curves as a function of anatomy (Fig. S2). It will thus be important for future work to precisely investigate and model these important properties of brain development not considered here. By characterizing a generic variation in CGE, the current work may have a role to play in this research effort. For example, genes that most strongly counter the generic trends in CGE(d) could be considered as candidates for playing a more specific, targeted role in shaping brain development, facilitating data-driven identification of genes that play important developmental roles.

ACKNOWLEDGEMENTS
The authors would like to thank Aurina Arnatkevičiūtė for assisting with human expression data.

References

1. A. Arnatkeviciūtė, 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).
2. B. D. Fulcher and A. Fornito. A transcriptional signature of hub connectivity in the mouse connectome. *Proc. Natl. Acad. Sci. USA* **113**, 1435 (2016).
3. R. Gămănuț, H. Kennedy, Z. Toroczkai, et al. The Mouse Cortical Connectome, Characterized by an Ultra-Dense Cortical Graph, Maintains Specificity by Distinct Connectivity Profiles. *Neuron* **97**, 698 (2018).
4. H. R. Noori, J. Schöttler, M. Ercsey-Ravasz, et al. A multiscale cerebral neurochemical connectome of the rat brain. *PLoS Biol.* **15**, e2002612 (2017).
5. M. Kunst, E. Laurell, N. Mokayes, et al. A Cellular-Resolution Atlas of the Larval Zebrafish Brain. *Neuron* **103** (2019).
6. N. T. Markov, M. Ercsey-Ravasz, D. C. Van Essen, et al. Cortical High-Density Counterstream Architectures. *Science* **342**, 1238406 (2013).
7. J. A. Roberts, A. Perry, A. R. Lord, et al. The contribution of geometry to the human connectome. *NeuroImage* **124**, 379 (2016).

8. 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).
9. M. Ercsey-Ravasz, N. T. Markov, C. Lamy, et al. A Predictive Network Model of Cerebral Cortical Connectivity Based on a Distance Rule. *Neuron* **80**, 184 (2013).
10. P. Theodoni, P. Majka, D. H. Reser, et al. Structural attributes and principles of the neocortical connectome in the marmoset monkey. *bioRxiv* p. 2020.02.28.969824 (2020).
11. J. A. Henderson and P. A. Robinson. Geometric effects on complex network structure in the cortex. *Phys. Rev. Lett.* **107**, 018102 (2011).
12. J. A. Henderson and P. A. Robinson. Relations Between the Geometry of Cortical Gyrification and White-Matter Network Architecture. *Brain Conn.* **4**, 112 (2014).
13. H. F. Song, H. Kennedy, and X.-J. Wang. Spatial embedding of structural similarity in the cerebral cortex. *Proc. Natl. Acad. Sci. USA* **111**, 16580 (2014).
14. J. Stiso and D. S. Bassett. Spatial Embedding Imposes Constraints on Neuronal Network Architectures. *Trends Cog. Sci.* (2018).
15. P. A. Robinson. Physical brain connectomics. *Phys. Rev. E* **99**, 012421 (2019).
16. L. French and P. Pavlidis. Relationships between gene expression and brain wiring in the adult rodent brain. *PLoS Comp. Biol.* **7**, e1001049 (2011).
17. A. Arnatkeviciūtė, B. D. Fulcher, and A. Fornito. A practical guide to linking brain-wide gene expression and neuroimaging data. *NeuroImage* **189**, 353 (2019).
18. 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).
19. 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).
20. 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).
21. 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).
22. A. Fornito, A. Arnatkeviciūtė, and B. D. Fulcher. Bridging the Gap between Connectome and Transcriptome. *Trends Cog. Sci.* **23**, 34 (2019).
23. D. S. Margulies, S. S. Ghosh, A. Goulas, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl. Acad. Sci. USA* **113**, 12574 (2016).
24. J. M. Huntenburg, P.-L. Bazin, P. L. Bazin, and D. S. Margulies. Large-Scale Gradients in Human Cortical Organization. *Trends Cog. Sci.* **22**, 21 (2017).
25. 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).
26. 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).
27. B. D. Fulcher, J. D. Murray, V. Zerbi, and X.-J. Wang. Multimodal gradients across mouse cortex. *Proc. Natl. Acad. Sci. USA* **116**, 4689 (2019).
28. L. C. Greig, M. B. Woodworth, M. J. Galazo, H. Padmanabhan, and J. D. Macklis. Molecular logic of neocortical projection neuron specification, development and diversity. *Nature Reviews Neuroscience* **14**, 755 (2013).
29. C. L. Thompson, L. Ng, V. Menon, et al. A High-Resolution Spatiotemporal Atlas of Gene Expression of the Developing Mouse Brain. *Neuron* **83**, 309 (2014).
30. A. A. Dillman, D. N. Hauser, J. R. Gibbs, et al. mrna expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nat. Neurosci.* **16**, 499 (2013).
31. S. Fertuzinhos, M. Li, Y. I. Kawasawa, et al. Laminar and temporal expression dynamics of coding and noncoding rnas in the mouse neocortex. *Cell Rep.* **6**, 938 (2014).
32. M. Mody, Y. Cao, Z. Cui, et al. Genome-wide gene expression profiles of the developing mouse hippocampus. *Proc. Natl. Acad. Sci. USA* **98**, 8862 (2001).
33. C.-K. Lee, S. M. Sunkin, C. Kuan, et al. Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome Biol.* **9**, R23 (2008).
34. 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).
35. L. Anselin. *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media (2013).
36. J. B. Burt, M. Helmer, M. Shinn, A. Anticevic, and J. D. Murray. Generative modeling of brain maps with spatial autocorrelation. *bioRxiv* (2020).
37. J. D. Cahoy, B. Emery, A. Kaushal, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264 (2008).
38. A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review* **51**, 661 (2009).
39. M. F. Glasser, T. S. Coalson, E. C. Robinson, et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171 (2016).
40. M. Pletikos, A. M. M. Sousa, G. Sedmak, et al. Temporal Specification and Bilaterality of Human Neocortical Topographic Gene Expression. *Neuron* **81**, 321 (2014).
41. S. Codeluppi, L. E. Borm, A. Zeisel, et al. Spatial organization of the somatosensory cortex revealed by osmfish. *Nat. Methods* **15**, 932 (2018).
42. J. R. Moffitt, D. Bambah-Mukku, S. W. Eichhorn, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, eaau5324 (2018).
43. X. Wang, W. E. Allen, M. A. Wright, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, eaat5691 (2018).
44. C. Xia, J. Fan, G. Emanuel, J. Hao, and X. Zhuang. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. USA* **116**, 19490 (2019).
45. C. A. Brunet, R. Bruggmann, and S. Sprecher. Single cell transcriptome atlas of the drosophila larval brain. *eLife* **8** (2019).
46. K. Davie, J. Janssens, D. Koldere, et al. A single-cell transcriptome atlas of the aging drosophila brain. *Cell* **174**, 982 (2018).
47. B. Raj, D. E. Wagner, A. McKenna, et al. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotech.* **36**, 442 (2018).

Supplementary Information

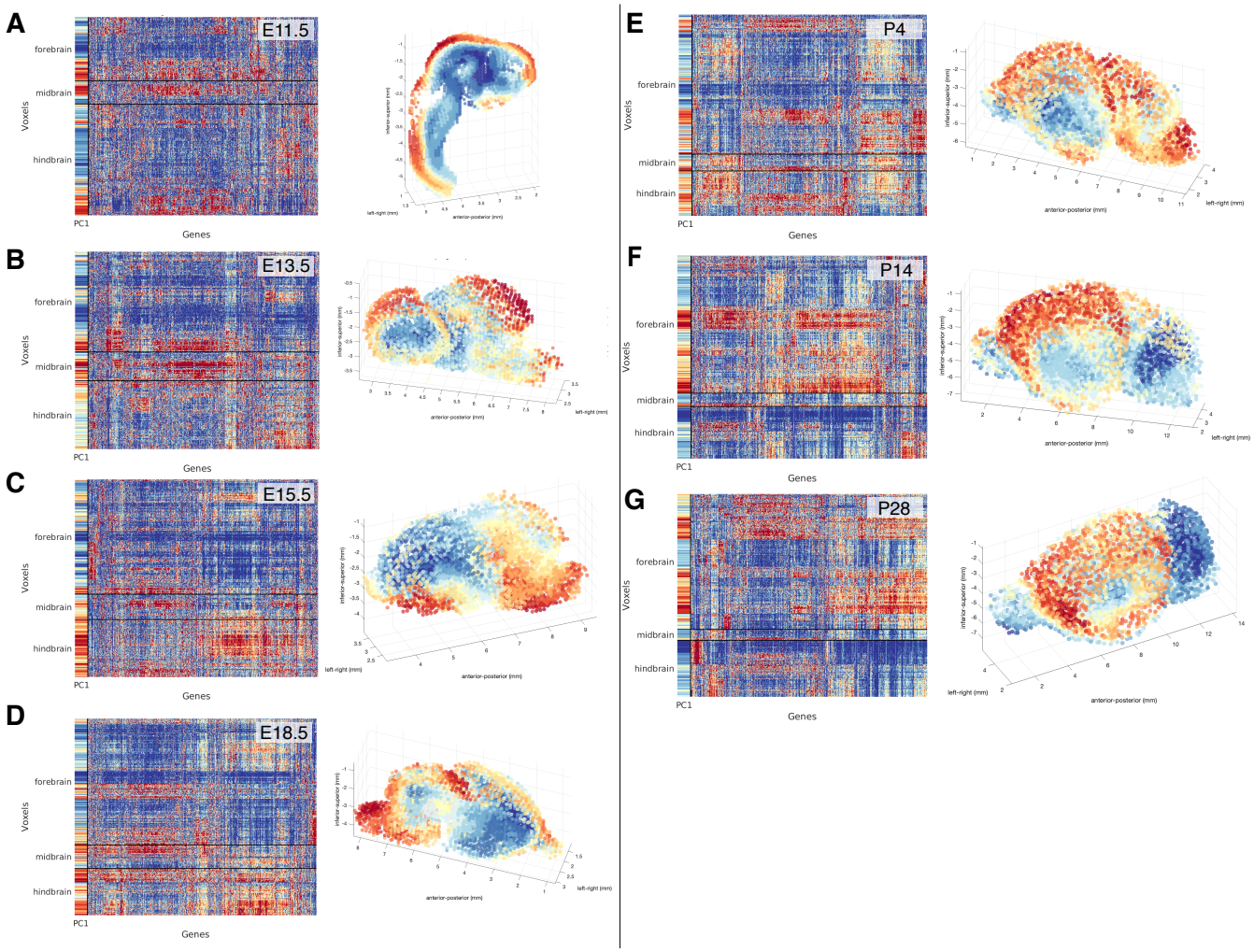

Fig. S1. Voxel × gene expression matrices and spatial projections of their leading principal component at each developmental time point. Plots are as described in Fig. 2.

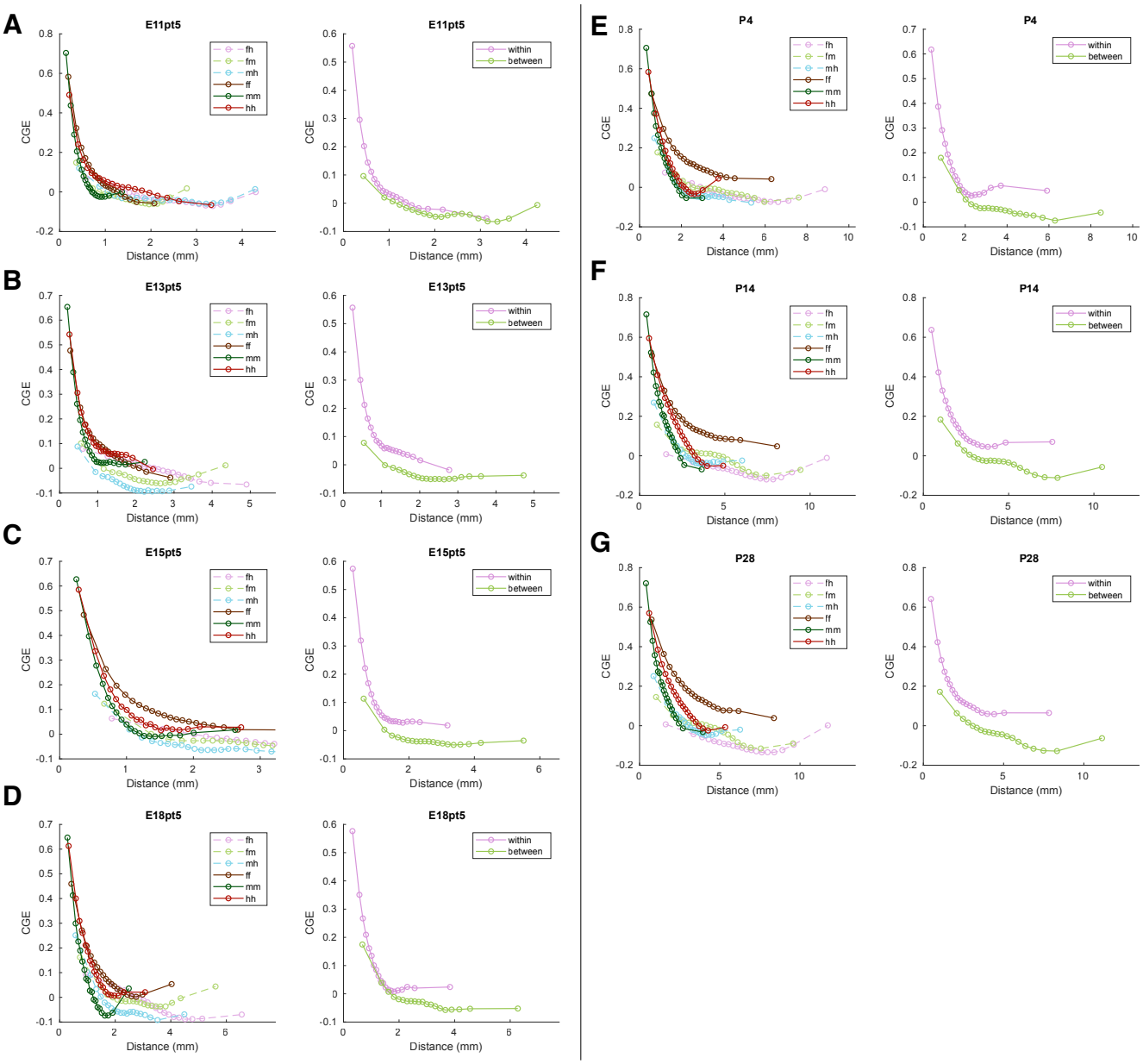

Fig. S2. The distance-dependence of correlated gene expression (CGE) varies with broad anatomical divisions. We show the CGE(d) curves for each time point, as labeled: **A** E11.5, **B** E13.5, **C** E15.5, **D** E18.5, **E** P4, **F** P14, and **G** P28. In each panel, the left plot shows the CGE(d) across 20 equiprobable distance bins for a subset of at most 1000 voxels in each subdivision: forebrain ('f'), midbrain ('m'), and hindbrain ('h'). CGE was computed for each type of voxel pair, the *within-division*: forebrain–forebrain ('ff'), midbrain–midbrain ('mm'), hindbrain–hindbrain ('hh'); and *between-division*: forebrain–hindbrain ('fh'), forebrain–midbrain ('fm'), and midbrain–hindbrain ('mh'). In the right panel of each subplot, within versus between division voxel pairs are agglomerated.

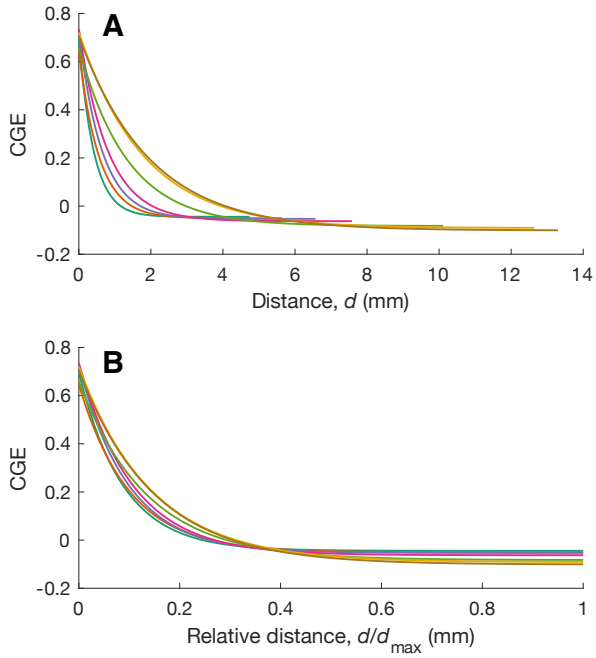

Fig. S3. Rescaling distances by a linear measure of brain size, $d_{\max}$, reveals a similarity in the spatial embedding of correlated gene expression (CGE). Fitted exponential curves are plotted for all seven time points (colored as in Fig. 3) for: **A** Distance, d (mm), and **B** Relative distance, $d/d_{\max}$. The measure of brain size, $d_{\max}$, is taken as the maximum extent along the anterior–posterior axis.

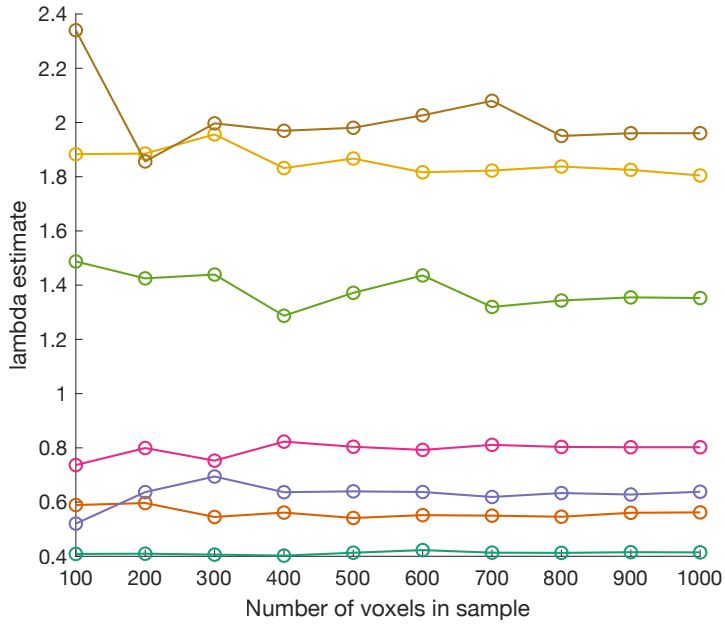

Fig. S4. λ estimates are stable at random samples of 1000 voxels. We sampled different numbers of voxels (100 to 1000 in increments of 100), to estimate the spatial correlation length, λ, from an exponential fit.. Curves of the estimated λ values are plotted for each developmental time point using colors from the main text (see Fig. 3). The λ estimates across random voxel samples plateau around 500 voxels, and are stable at the 1000 voxel samples used here, indicating that 1000 is a sufficiently large sample size to obtain reliable λ estimates.