Methods 50 (2010) 105–112

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy

Jason W. Bohland [a,*,1], Hemant Bokil [a,*,1], Sayan D. Pathak [b], Chang-Kyu Lee [c], Lydia Ng [c], Christopher Lau [c], Chihchau Kuan [c], Michael Hawrylycz [c], Partha P. Mitra [a]

[a] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States
[b] Microsoft Imaging Research and Development, Redmond, WA 98052, United States
[c] Allen Institute for Brain Science, Seattle, WA 98103, United States

ARTICLE INFO

Article history:
Accepted 1 September 2009
Available online 3 September 2009

Keywords:
Gene expression
Mouse
Brain Atlas
Neuroanatomy
Clustering
Exploratory data analysis
Singular value decomposition

ABSTRACT

Spatial gene expression profiles provide a novel means of exploring the structural organization of the brain. Computational analysis of these patterns is made possible by genome-scale mapping of the C57BL/6J mouse brain in the Allen Brain Atlas. Here we describe methodology used to explore the spatial structure of gene expression patterns across a set of 3041 genes chosen on the basis of consistency across experimental observations ($N = 2$). The analysis was performed on smoothed, co-registered 3D expression volumes for each gene obtained by aggregating cellular resolution image data. Following dimensionality and noise reduction, voxels were clustered according to similarity of expression across the gene set. We illustrate the resulting parcellations of the mouse brain for different numbers of clusters (K) and quantitatively compare these parcellations with a classically-defined anatomical reference atlas at different levels of granularity, revealing a high degree of correspondence. These observations suggest that spatial localization of gene expression offers substantial promise in connecting knowledge at the molecular level with higher-level information about brain organization.

1. Introduction

Understanding the brain's basic structural organization is of critical importance across areas of study in neuroscience, but to date there remains significant uncertainty about how the nervous system can be broken down in terms of regions, sub-regions, cell groups, and so forth [1,2]. Over time, many methods have been used to derive atlases that partition the brain into individual regions on the basis of differential spatial patterns in cyto-, myelo-, or chemo-architecture or that are based on other structural or functional characteristics [3,4]. On some level, it is clear that the expression of particular genes drives the organization of the brain into differentiated regions and varied cell types [5–8]. Individual genes with expression patterns restricted to classically-defined brain structures have been identified using a range of different techniques [9–16]. Further, microarray-based analysis of expression profiles in tissues from 24 brain areas has indicated that the adult mouse brain maintains an imprinted developmental genetic program established during embryogenesis [15]. Genes have also been found to be differentially expressed among different classes of neurons, even within the same brain areas [17].

Studies such as these raise the question of whether neuroanatomically distinct entities, as classically defined, can be delineated on the basis of gene expression patterns alone. Examining this question, and more generally, the patterns of spatial correlation among gene expression profiles in the brain, motivated the present study. We examined these patterns using multivariate exploratory data analysis performed on a large set of spatially co-registered gene expression volumes derived from the Allen Brain Atlas (ABA) [10]. The ABA project has made it possible to examine gene expression in the adult mouse brain at an unprecedented scale and level of resolution. By registering data for each gene to a common atlas space [18], it becomes possible to examine the "genetic signature" of different locations in the brain, and to examine how these signatures change across brain areas. In this paper, we present a straightforward approach to determine data-driven partitions of the brain using the singular value decomposition and cluster analysis, and we demonstrate procedures for comparing expression-based clusters with regions delineated from a classically-defined anatomical reference atlas [19]. Our intent is to demonstrate the usefulness of these techniques for the study of high-resolution spatial gene expression at a large scale, and to begin to determine if

* Corresponding authors. Address: Boston University, Sargent College of Health and Rehabilitation Sciences, 635 Commonwealth Ave, Room 403, Boston, MA 02215, United States. Fax: +1 617 353 7567.
E-mail address: jbohland@bu.edu
[1] J.W.B. and H.B. are contributed equally to this work.

1046-2023/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2009.09.001

the brain's molecular organization can be reconciled with classical neuroanatomical parcellations at a particular spatial scale.

2. Description of method

2.1. Allen Brain Atlas

The Allen Brain Atlas (http://mouse.brain-map.org) provides cellular resolution expression profiles for ~20,000 genes in the male, 56-day old C57BL/6J mouse brain [10]. Primary image data were generated by highly automated and methodical application of non-isotopic *in situ* hybridization (ISH) procedures [20] using custom gene-specific probes, followed by automated slide scanning, informatics, and image processing stages [18]. Genome-wide coverage is available in sagitally-oriented sections (25 μm thickness) at uniform 200 μm inter-slice intervals (with 1.07 μm/pixel in-plane resolution), resulting in ~20 sections across a single hemisphere. For ~4000 genes that exhibited restricted expression patterns or that were deemed to be of high neurobiological interest, replicate data, for the full brain, were produced by processing a more numerous set of coronal sections (~56 sections, also with 200 μm spacing). The coronal data set generally excludes ubiquitously expressed genes, or so-called 'housekeeping' genes, and has some selection bias toward genes with cortical and/or hippocampal expression patterns [21].

To enable a meaningful and tractable comparison of spatial expression patterns across genes, the high (in-plane) resolution primary data from each experiment were reconstructed in 3D and registered to a *de novo* age and gender-matched Nissl stain-based reference atlas (Allen Reference Atlas; ARA) [19]. For each gene, the data were then aggregated into isotropic voxels defined by a uniform 200 μm grid in the reference space. The expression of a gene within each 200 μm voxel was summarized by a measure of *smoothed expression energy* [10], defined as the average intensity of pixels in the pre-processed ISH image (see Ref. [18] for details) intersecting that voxel. The resulting data consist of spatially aligned $67 \times 41 \times 58$ (rostral–caudal, dorsal–ventral, left–right) volumes for each gene. The ARA is also provided in the common volumetric space as a series of masks (in a single hemisphere) for each defined anatomical region. The study presented here is based on the brain-wide coronal image volumes in this 200 μm volumetric data set.

2.2. Higher-consistency data set

While substantial quality control measures were employed in the creation of the ABA, it remains possible that data for individual genes contain artifacts or other problems that result in poor reproducibility. For example, tissue folding, dust, or air bubbles on slides could potentially give rise to incorrect determination of expressing cells, or particular genes or probes may simply produce highly variable results. For this study, we sought to use a suitably large set of genes while maintaining at least some control over the quality of the individual data volumes (e.g. how representative a particular expression volume is of the population). Because expression volumes for each gene mapped in the coronal series were also available from the sagittal series (in a single hemisphere), we could make a single estimate of reproducibility by comparing coronal and sagittal volumes. For each unique gene, we used the Pearson correlation coefficient to compare the expression patterns across voxels in a single hemisphere for available coronal and sagittal volumes. For some genes, more than one volume was available for an orientation; we summarized the data consistency for that gene as the maximum correlation coefficient between a coronal volume and any available sagittal volume. The distribution of coronal–sagittal correlation values for the 4104 unique genes in our coronal data set is shown in Fig. 1. For further analysis, we discarded the 25% of genes for which these correlations were lowest. This particular threshold was varied somewhat during exploratory analysis, and the precise value did not appear to qualitatively impact the overall results. The full-brain coronal volumes for the remaining 3041 genes were considered a higher consistency, more reliable dataset and examined further. The full list of 4104 coronal data sets, with indication of which were included in the higher consistency group, is included in supplementary materials.

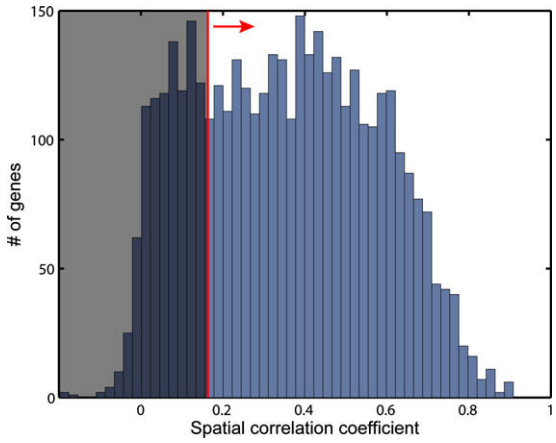

Fig. 1. Histogram of Pearson correlation coefficients obtained from comparing data from coronal and sagittal sections of the same gene. The genes corresponding to the top 75% of correlation values (to the right of the red line) were selected for further analysis.

2.3. Singular value decomposition

The expression energy values at each brain voxel were extracted from volumes for each gene in the higher-consistency dataset. These values were concatenated to form a large ($49{,}742 \times 3041$) *voxel* × *gene* matrix, $E(v, g)$, where v denotes voxels and g denotes genes. Each column of the matrix thus contains the expression energy for one gene across all brain voxels. This matrix representation makes the dataset amenable to classical multivariate data analysis methods, including the singular value decomposition (SVD), a robust technique that is valuable for denoising and reducing the dimensionality of high-dimensional datasets.

The SVD is a decomposition of any $p \times q$ matrix M into a product $M = USV^{\dagger}$ where U and V are unitary matrices ($UU^{\dagger} = VV^{\dagger} = I$), and S is a diagonal matrix with real entries. Here, U is a $p \times q$ matrix, and S and V are $q \times q$ matrices. The columns of U and V are known as the left and right singular vectors, respectively, and entries along the diagonal of S are known as singular values. Note that when M is centered (row and column means are zero), the left singular vectors are eigenvectors of the covariance matrix $M^{\dagger}M$, the right singular vectors are eigenvectors of the covariance matrix $MM^{\dagger}$, and the square of a singular value is the variance of the corresponding eigenvector. Therefore, a projection of the data matrix M to a d-dimensional subspace with the largest variance may be obtained by using $MV = US$, retaining only the d largest singular values and corresponding singular vectors.

The SVD was applied to the voxel × gene matrix $E(v, g)$, transformed so that both the row and column means were equal to zero. Because $E(v, g)$ is quite large ($p = 49742$, $q = 3041$), efficient computation of the SVD necessitated the use of ScaLAPACK, a parallel version of the LAPACK library for linear algebra [22]. The decomposition of $E(v, g)$ can be written as:

$$E(v, g) = \sum_{i=1}^{q} \lambda_i S_i(v) G_i(g) \tag{1}$$

where λ_i are the sorted singular values, $S_i(v)$ are the left singular vectors (referred to here as *spatial modes*) and $G_i(g)$ are the right singular vectors (*gene modes*).

Fig. 2(A) shows the distribution of singular values obtained from the decomposition of $E(v, g)$. The first 67 modes account for more than 80% of the variance, and the subspace spanned by the first 271 modes contains more than 90% of the variance. This indicates that, while there is considerable correlation within spatial expression profiles, the space is still of relatively high rank, with rich structure across genes and voxels. Interestingly, the dimensionality of the signal subspace appears to be comparable to some estimates of the number of anatomically distinct brain regions in the rodent brain [1]. This observation indicates that expression in the mouse brain varies in many orthogonal directions in the gene space, leading to the hypothesis that these different combinations of genes may define distinct anatomical entities; this question is explicitly addressed through cluster analysis, described below.

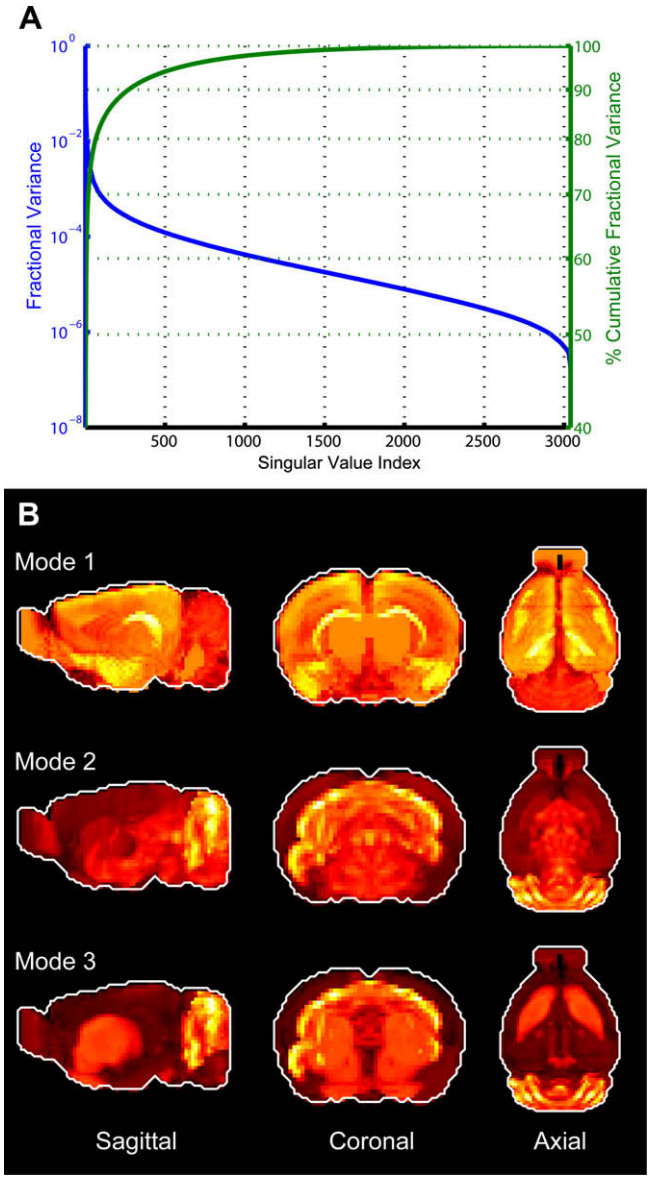

Fig. 2. Results from application of the singular value decomposition. (A) Sorted singular value spectrum from decomposition of voxel × gene matrix $E(v, g)$, plotted as both the fractional variance accounted for by each mode (green) and the cumulative variance (blue). (B) Absolute values of the first three spatial modes plotted as maximum intensity projection images in the three cardinal planes. White outlines indicate approximate boundaries of the mouse brain.

The individual spatial modes can be cast as patterns across brain space. Fig. 2(B) illustrates the absolute values of the first three spatial modes using maximum intensity projection images. The first mode has a very broad distribution across the brain, but shows distinctly high amplitude in the dentate gyrus of the hippocampus. The second mode is also broadly distributed with enhanced specificity for the cerebellum, and the third mode is particularly prominent in the cerebellum and the striatum. Thus, it is clear that the decomposition is able to extract correlated structure in the expression profiles that corresponds generally to broad anatomical subdivisions.

Motivated by these observations, we projected the expression data into the subspace spanned by these first three modes for the purpose of visualization. Fig. 3 shows a scatter plot of the data for 25,155 left hemisphere voxels in the reduced 3-space. Each point (voxel) is color-coded by the gross anatomical area in which it is located, according to the Allen Reference Atlas. In this view, the cerebellum is largely separable from other regions, as is the striatum. To a lesser extent, other regions are clustered in the greatly reduced space, but with considerable overlap. Thus, while distinct anatomical regions do not correspond directly to individual modes from the SVD, it appears that clustering based on similarity of gene expression profiles, even in a very low dimensional subspace, can be expected to group voxels drawn the same areas in the brain. It is noteworthy that the two most separable structures (cerebellum and striatum) each contain large numbers of GABAergic inhibitory neurons relative to other areas. Previous microarray analysis suggested a fundamental separation in the gene expression profiles of GABAergic interneurons and glutamatergic projection neurons in the adult mouse forebrain [17].

2.4. Clustering voxels based on similarity of gene expression

As discussed above, even the first three modes contain spatial structure in rough concordance with classical anatomy. However, it is also apparent that finer structure cannot be resolved with a projection to such a low dimension, which captures less than half of the overall variance in the data. This motivated a more thorough cluster analysis to examine groupings of voxels with similar patterns of expression across the genes in our higher-consistency volumetric data set. For computational efficiency, and as a noise reduction measure, we first projected the original data matrix $E(v, g)$ into the signal subspace which accounts for 90% of the original variance ($n = 271$). We then applied the K-means clustering algorithm to group voxels based on the similarity (defined by Euclidean distance) of the expression profiles projected into this 271-dimensional space. Thus, the procedure sought to obtain a data-driven segmentation of the mouse brain based on patterns of gene expression, with the parameter K (number of clusters) determining the granularity of the segmentation.

The K-means algorithm is a simple and routinely used method for assigning data observations to clusters. The method requires the *a priori* selection of the number of clusters, K, and seeks to assign each data point to one of K clusters such that the sum of squared distances between the observations and the centroid (mean across all data points assigned to a cluster) of the cluster to which they are assigned is minimized. The standard Lloyd's algorithm [23] begins with an arbitrary set of K centroids and assigns each point to the cluster defined by the closest one. Then, new cluster centroids are calculated, and the assignment process is repeated. When iteratively continued, the method is guaranteed to converge to a locally optimal solution. To avoid suboptimal solutions, a common procedure is to apply the algorithm repeatedly, using different initial centroid locations.

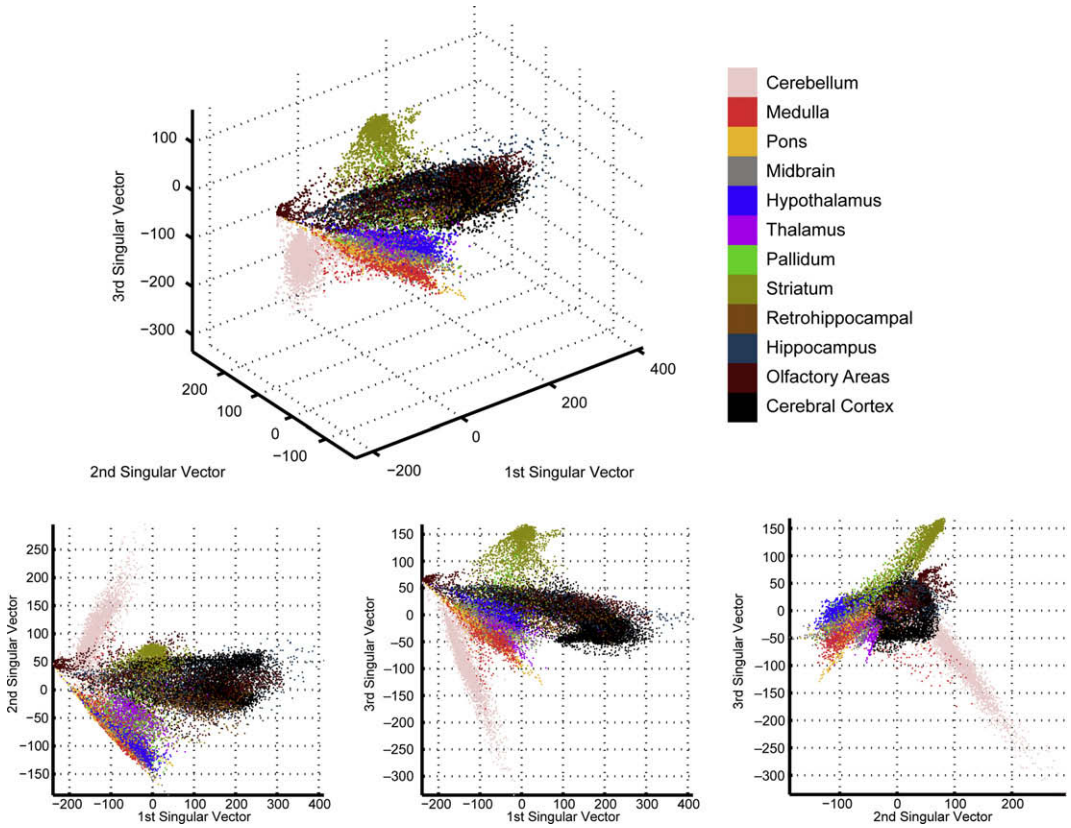

Fig. 3. Projection of expression data for left hemisphere voxels onto the first three right singular vectors. Each dot (voxel) is color-coded to indicate the gross region from which it was drawn according to the Allen Reference Atlas. For clarity, the 2D projections along each pair of modes are shown at bottom.

An implementation of the *K*-means algorithm from the MATLAB® (http://www.mathworks.com) Statistics Toolbox was used to cluster all *left hemisphere* brain voxels into distinct classes. Only left hemisphere voxels were analyzed because we were interested in comparing clustering outcomes to anatomical partitions from the Allen Reference Atlas, which are only explicitly annotated in the left hemisphere; additionally, this reduced computation time considerably. The input to the algorithm was the dimension-

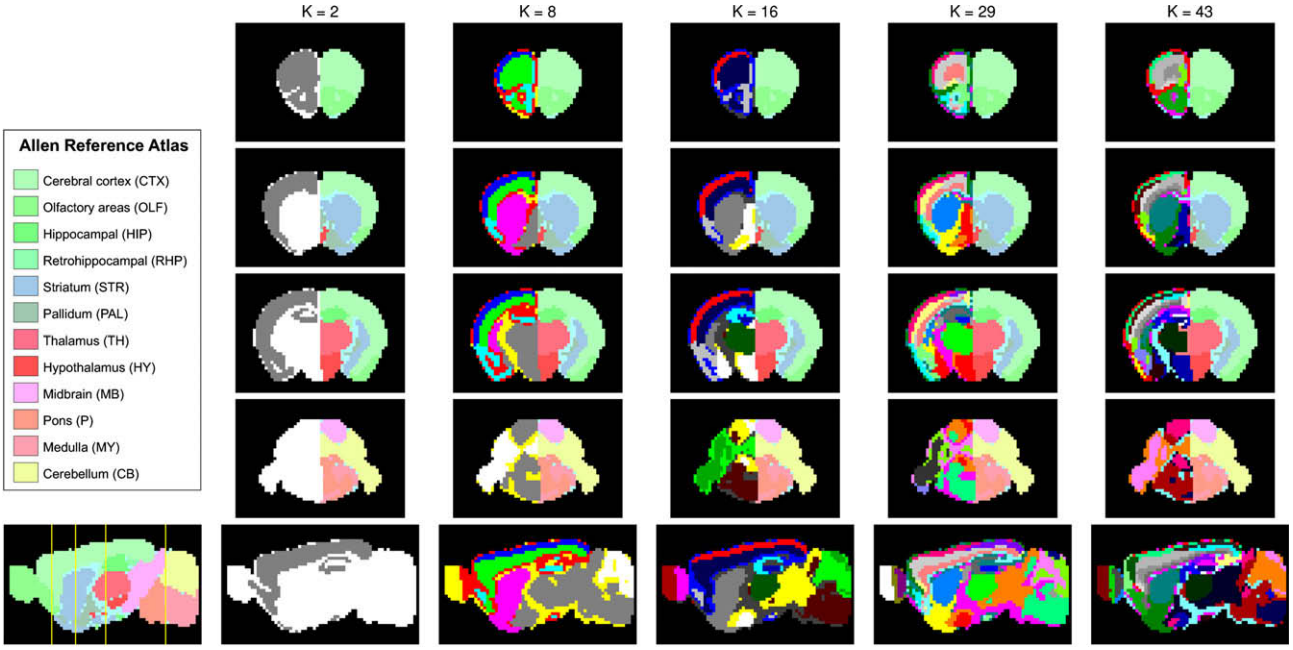

Fig. 4. *K*-means clustering results for various values of *K*. The top four rows in each column show coronal sections from anterior to posterior (at locations specified by yellow lines through the Allen Reference Atlas (ARA) section at lower left). In each coronal view, clustering results are shown in the left hemisphere (color scale is arbitrary), and labeling of the same section from the coarse ARA is shown on the right. A color legend is shown on the left. The bottom row in each column shows a sagittal view of the clustering result.

reduced 25,155 × 271 expression matrix; the algorithm was asked to cluster the rows of the matrix by similarity based on Euclidean distances. Five different initial conditions were used, and for each K the solution with minimum total within-cluster variance was chosen. Robustness of the results was verified by performing additional runs of the algorithm for ~20 different values of K ranging from 2 to 35 using 50 different initial configurations. Results were very similar, with only minor quantitative differences.

Fig. 4 shows the results of K-means clustering for several different values of K. From these illustrations and from further manual inspection, it is of note that the vast majority of clusters formed are spatially contiguous within the volume, even though nothing in the methods required this to be true. Indeed, this result reflects the spatial smoothness of the dataset—within distinct regions, expression energy values tend to change slowly across space, while at anatomic region boundaries they can change rapidly. Also of note is the tendency for clusters (left hemisphere in coronal sections and in sagittal sections in the bottom row) to reflect the known anatomy (depicted at a coarse level as a mirror image in the right hemisphere of coronal sections). A series of NIfTI-1 formatted image volumes depicting cluster analysis results for various values of K can be downloaded from the supplementary materials, allowing for further interactive examination of results. Because the truncated SVD method effectively smoothes the data, and therefore may account for the spatially contiguous clusters, we also performed more computationally intensive cluster analysis on the full data set for several values of K. Results were qualitatively similar, and examples are included as supplementary materials.

In the first full column of Fig. 4, we see that dividing the brain into 2 maximally distinct clusters separates the cerebral cortex and portions of the hippocampus (gray) from all other areas (white). As K is increased to 8, the cerebellum and striatum are each clearly segmented, and the cortex is subdivided into distinct layers. At $K = 16$, the thalamus is assigned to its own cluster, cortical layers further differentiated, and the midbrain separated from the hindbrain. For larger K more and more anatomical details are observed, for example the separation of the caudoputamen from

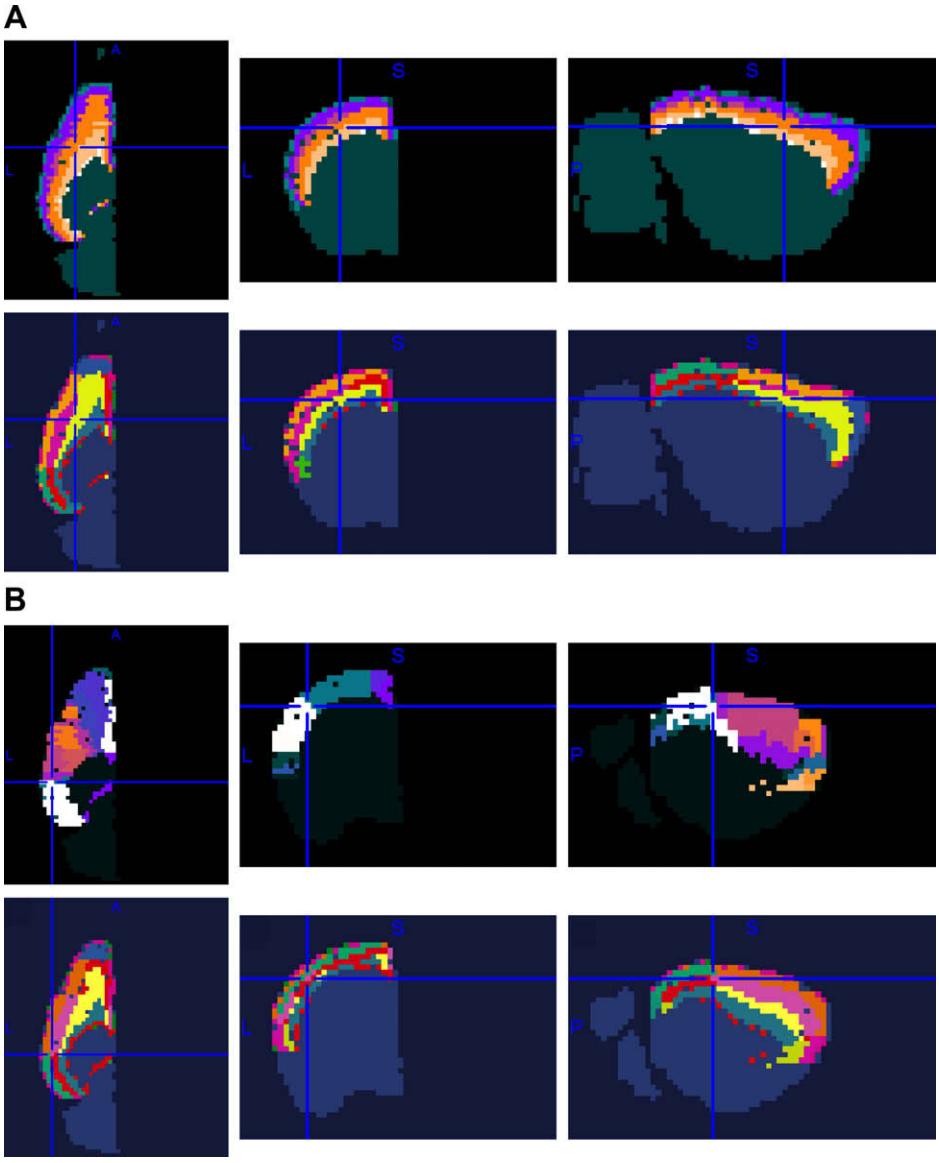

Fig. 5. Laminar and areal clustering in the cerebral cortex. (A) Top: Cortical layer masks from the ARA shown in axial, coronal, and sagittal planes corresponding to a chosen voxel (positioned at the blue crosshairs). Bottom: Clustering result for $K = 40$ at the same location. (B) Top: Cortical area masks from the ARA. Bottom: Clustering result for $K = 40$ at the same location.

the nucleus accumbens and other basal nuclei and divisions within the cortex that not only reflect laminar patterns but also certain areal patterns with borders oriented normal to the cortical surface. For example, in Fig. 5(B), we see laminar clusters broken into distinct groups along the anterior–posterior direction (bottom) at a location that corresponds quite precisely with the border between auditory and somatosensory areas (cortical area masks shown above). In Fig. 5(A) we see, for the same clustering result, that laminar cluster patterns closely follow cortical layer masks delineated by an expert anatomist. These observations, which suggest a dominant clustering of gene expression along cortical layers, with a lesser but still observable clustering within areas, are consistent with the findings of Ng et al. [21]. An interesting consequence of performing cluster analysis for a large range of K values is the ability to determine, for a given structure, at what value of K it emerges as its own cluster. This yields a relative prioritization of anatomical areas based on expression pattern similarity.

We also note the existence of boundary effects, particularly at the edge of the brain and the edges of large anatomical areas. It is unclear at this point if the separation of some voxels at these boundaries into distinct clusters represents truly distinct expression profiles, or if this might instead reflect small registration errors across the gene volumes.

2.5. Correspondence with classical anatomy

As described above, it was immediately apparent that the results of K-means clustering showed qualitative similarity to the classically-defined anatomical reference atlas. To quantitatively measure concordance with the Allen Reference Atlas (ARA), we used a previously developed set-theoretic framework for comparing different parcellations of the same space [24]. In this framework, a single parcellation R is a set of N clusters (or anatomical regions),

$$R = \{r_1, r_2, \ldots, r_N\}$$

and each region comprises the set of indices of the voxels x that map to that cluster (or anatomical label):

$$r_i = k \in \{1, 2, 3, \ldots M\} : x_k \mapsto r_i$$

We then define a non-symmetric measure of spatial overlap between a region from the ARA and cluster from the K-means result:

$$P_{ij} = \frac{|r_i \cap r_j|}{|r_j|}$$

P_{ij} thus indicates the proportion of voxels that comprise r_j that are contained within r_i, and is bounded on the interval [0, 1]. From the P_{ij} values, computed over all pairs of ARA regions and K-means clusters, we can then derive a global scalar index of similarity between the cluster result and the reference atlas. The index was designed to be relatively insensitive to differences in the granularity of the partitions. That is, it should penalize cluster-to-region relationships that are overlapping but not when one is a pure subset of the other (i.e. hierarchically related). A similar formulation has been applied in the comparison of object segmentations in 2D images [25]. The scalar-valued similarity index S, which takes values between 0 and 1, is computed as follows. First the maximum of the two conditional overlap values relating each ARA region to each cluster is computed:

$$X_{ij} = \max(P_{ij}, P_{ji})$$

along with "weights" for each non-zero X_{ij}, which are dependent on region/cluster volume:

$$U_{ij} = \begin{cases} \min(|r_i|, |r_j|) & \text{if } X_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$W_{ij} = \frac{U_{ij}}{\sum U_{ij}}$$

Finally the weighted maximum conditional overlaps are combined and subtracted from 1 as penalty terms in order to arrive at the final expression for global concordance. The "penalty" is largest when the maximum conditional overlap between a pair of regions is 0.5. The measure thus reflects how precisely clusters can be mapped to regions, or vice versa.

$$S = 1 - 4\sum_{ij} W_{ij} X_{ij} (1 - X_{ij})$$

The S index comparing K-means clustering results with the ARA was computed as a function of K, for both coarse-grained (consisting of 12 regions) and fine-grained (consisting of 94 regions) "flat" (non-hierarchical) versions of the atlas. Fig. 6 shows the results of these comparisons. The overlap between the clusters and the ARA increases until saturating at $K > 30$. Because the measure allows for subset relationships between the two parcellations being compared, this saturation indicates that clusters obtained at larger values of K are generally subdivisions of those obtained at lower values of K.

To compare concordance at the level of individual regions and clusters, it is useful to directly visualize the individual overlap matrices (P_{ij}). For example, Fig. 7 shows the two non-symmetric matrices for $K = 12$, the same number of regions delineated in the coarse version of the ARA. By comparing the two matrices, it is clear, for example, that clusters 1, 2, 3, and 4 together comprise most of the cerebral cortex (they are arranged as layers as illustrated in Figs. 4 and 5). Also, cluster 11 largely corresponds to the thalamus, cluster 9 is wholly contained in the cerebellum, and cluster 10 in the striatum. At this level of granularity, the pallidum (PAL) does not correspond well with any particular cluster, and is divided into several clusters that also contain portions of the midbrain, hypothalamus, and other structures.

These matrices have been subjected to a bandwidth (defined as $\max_{P_{ij} \neq 0}(|i - j|)$) reduction heuristic to re-order the rows and columns in order to bring non-zero entries toward the diagonal [26]. This procedure was implemented as follows: (1) a symmetric matrix was computed by finding the geometric mean of each element in the two conditional overlap matrices; (2) an SVD of the resulting

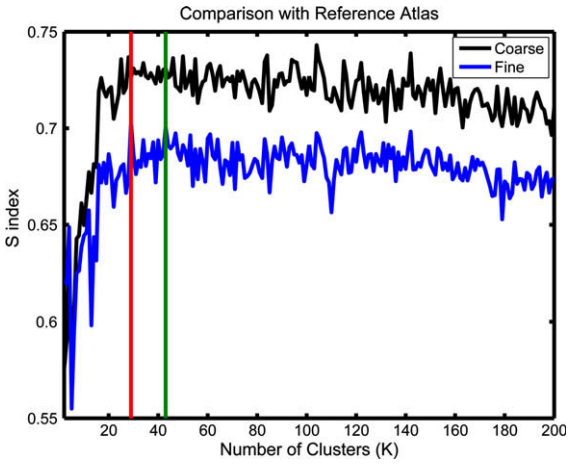

Fig. 6. Similarity index comparing left hemisphere K-means clustering results with coarse (12 regions, black) and fine (94 regions, blue) versions of the Allen Reference Atlas as a function of K. The red and green lines correspond to peak-similarity results for $K = 29$ and $K = 43$, both of which are further depicted in Fig. 4.

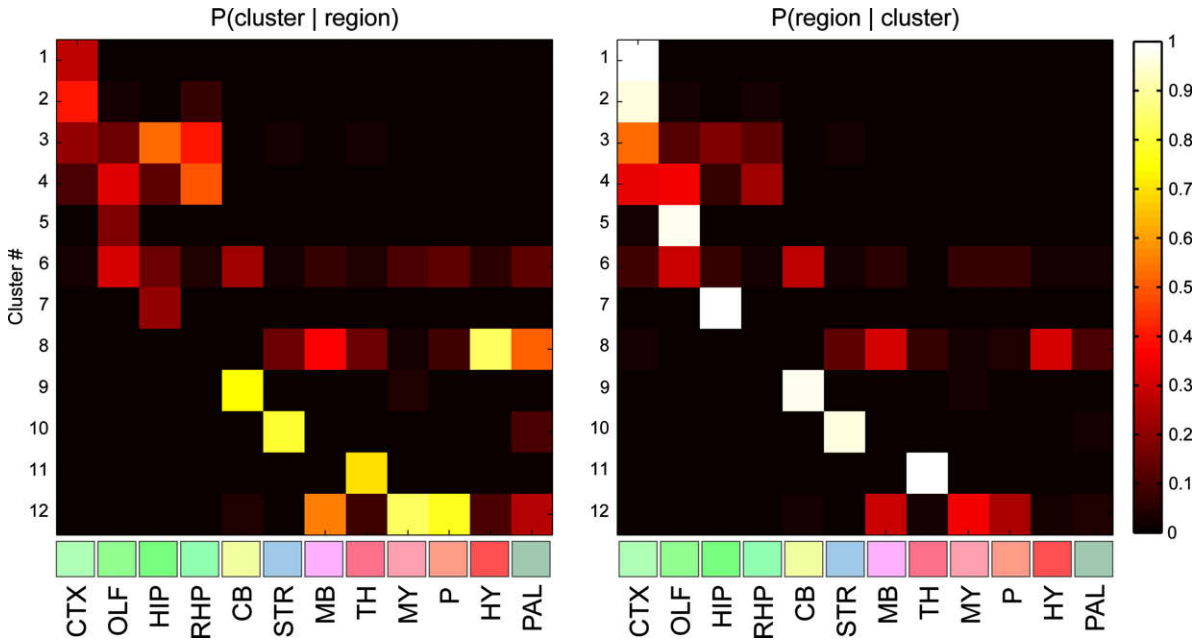

Fig. 7. Overlap matrices relating data-driven clusters for $K = 12$ to the 12 regions in the coarse ARA. Left: Fraction of each region contained within each cluster. Right: Fraction of each cluster contained within each region. Color scale is identical for both matrices. Region labels (bottom) include color squares corresponding to the colors of regions in ARA annotations in Fig. 4.

matrix was computed; (3) the first left and right singular vectors were sorted, and the sorted indices were used as permutations of the rows and columns for both conditional matrices. This procedure results in the non-zero entries in each block being "moved" toward the diagonal, yielding a more suitable visualization than the arbitrarily ordered *P*. In general, lower resulting bandwidth is then indicative of better correspondence between the cluster assignment and the reference atlas.

3. Concluding remarks

The open availability of large-scale spatially referenced gene expression profiles enables novel modes of data-driven analysis that promise to reveal hidden structure in the nervous system and lead to new experimental hypotheses. Exploratory analysis of the correlation structure within a large set of genes, made possible by application of standard multivariate techniques such as the SVD and cluster analysis, provides a data-driven framework for investigating the brain's molecular/genetic architecture. The analyses described here use simple methods with minimal assumptions about the underlying data. A somewhat similar approach was taken by Ng et al., who have developed an online tool (the Anatomic Gene Expression Atlas; AGEA) for the ABA that allows users to investigate expression-based correlations across the brain [27]. The present results differ in several ways. First, we sub-selected a set of gene volumes based on reproducibility across sagittal and coronal datasets, and used the SVD to reduce noise and dimensionality. Additionally, various details of the clustering algorithms including the distance measures were different. Further, we quantitatively compared clustering results to the reference atlas using a similarity measure that is based on spatial overlap, and that allows for region refinement without penalty.

The application of these methods to the higher-consistency coronal data set from the ABA led to several key insights. The SVD showed that the expression signal subspace in this dataset is rich, but with considerable correlation structure, with dimensionality on the order of a few hundred. Cluster analysis used the *K*-means algorithm, a simple and intuitive method chosen to indicate the robustness of data clusters. Here *K* could be used to vary the granularity of clusters, and visually exploring the results for increasing *K* proved to be a useful technique for comparison with classical anatomical subdivisions, which are themselves multiscale. Quantitative comparison with the ARA, using a custom similarity index showed that cluster results matched the reference atlas best with approximately 30–40 clusters, but similarity remained stable for increasing *K*. This suggests that the finer-grained data-driven partitions largely subdivide classically-defined anatomical regions into areas with more similar expression patterns. In future work it will be useful to examine where the *dissimilarities* are between classical region boundaries and cluster boundaries, which may help to inform future experimental work. Further, in addition to the set-theoretic procedures we used here to assess concordance, it will be interesting to consider the similarity of the geometric and/or topological properties of expression clusters in comparison with atlas-based parcellations.

The observation that clusters of voxels with similar patterns of expression corresponded strongly with classically-defined anatomical areas offers a compelling argument that *localization* may help form a bridge between the molecular level and higher levels of functional organization in the normal and diseased brain. The size of the voxels in the present dataset (200 μm on a side) is large relative to individual cell bodies, and in many cases voxels will contain a mixture of several cell types. The smoothed expression energy at that voxel thus reflects that mixture, and provides a unique expression signature for discrete brain locations with different combinations of cell types. These initial results offer reason to believe that spatial co-expression (cf. co-expression within the same cell) may be a powerful indicator of functionally-related or interacting genes and a promising research direction. The field of neurogenomics is undoubtedly ripe with interesting possibilities [28].

Acknowledgments

This work was supported by the W.M. Keck Foundation (PI: P.P. Mitra) and the Crick-Clay professorship. The authors thank the

Allen Institute founders P.G. Allen and J. Patton. We also thank Daniel Herrera and Fernando Osorio-Duque for their assistance in interpreting results.

References

[1] M. Bota, H.W. Dong, L.W. Swanson, Nat. Neurosci. 6 (2003) 795–799.
[2] L.W. Swanson, Trends Neurosci. 23 (2000) 519–527.
[3] A. MacKenzie-Graham, E.F. Lee, I.D. Dinov, M. Bota, D.W. Shattuck, S. Ruffins, H. Yuan, F. Konstantinidis, A. Pitiot, Y. Ding, G. Hu, R.E. Jacobs, A.W. Toga, J. Anat. 204 (2004) 93–102.
[4] A.W. Toga, P.M. Thompson, S. Mori, K. Amunts, K. Zilles, Nat. Rev. Neurosci. 7 (2006) 952–966.
[5] C. Kiecker, A. Lumsden, Nat. Rev. Neurosci. 6 (2005) 553–564.
[6] Y. Nakagawa, D.D. O'Leary, J. Neurosci. 21 (2001) 2711–2725.
[7] D.D. O'Leary, Y. Nakagawa, Curr. Opin. Neurobiol. 12 (2002) 14–25.
[8] J.L. Rubenstein, S. Anderson, L. Shi, E. Miyashita-Lin, A. Bulfone, R. Hevner, Cereb. Cortex 9 (1999) 524–532.
[9] L. Luo, R.C. Salunga, H. Guo, A. Bittner, K.C. Joy, J.E. Galindo, H. Xiao, K.E. Rogers, J.S. Wan, M.R. Jackson, M.G. Erlander, Nat. Med. 5 (1999) 117–122.
[10] E.S. Lein, M.J. Hawrylycz, N. Ao, M. Ayres, A. Bensinger, A. Bernard, A.F. Boe, M.S. Boguski, K.S. Brockway, E.J. Byrnes, L. Chen, L. Chen, T.M. Chen, M.C. Chin, J. Chong, B.E. Crook, A. Czaplinska, C.N. Dang, S. Datta, N.R. Dee, A.L. Desaki, T. Desta, E. Diep, T.A. Dolbeare, M.J. Donelan, H.W. Dong, J.G. Dougherty, B.J. Duncan, A.J. Ebbert, G. Eichele, L.K. Estin, C. Faber, B.A. Facer, R. Fields, S.R. Fischer, T.P. Fliss, C. Frensley, S.N. Gates, K.J. Glattfelder, K.R. Halverson, M.R. Hart, J.G. Hohmann, M.P. Howell, D.P. Jeung, R.A. Johnson, P.T. Karr, R. Kawal, J.M. Kidney, R.H. Knapik, C.L. Kuan, J.H. Lake, A.R. Laramee, K.D. Larsen, C. Lau, T.A. Lemon, A.J. Liang, Y. Liu, L.T. Luong, J. Michaels, J.J. Morgan, R.J. Morgan, M.T. Mortrud, N.F. Mosqueda, L.L. Ng, R. Ng, G.J. Orta, C.C. Overly, T.H. Pak, S.E. Parry, S.D. Pathak, O.C. Pearson, R.B. Puchalski, Z.L. Riley, H.R. Rockett, S.A. Rowland, J.J. Royall, M.J. Ruiz, N.R. Sarno, K. Schaffnit, N.V. Shapovalova, T. Sivisay, C.R. Slaughterbeck, S.C. Smith, K.A. Smith, B.I. Smith, A.J. Sodt, N.N. Stewart, K.R. Stumpf, S.M. Sunkin, M. Sutram, A. Tam, C.D. Teemer, C. Thaller, C.L. Thompson, L.R. Varnam, A. Visel, R.M. Whitlock, P.E. Wohnoutka, C.K. Wolkey, V.Y. Wong, M. Wood, M.B. Yaylaoglu, R.C. Young, B.L. Youngstrom, X.F. Yuan, B. Zhang, T.A. Zwingman, A.R. Jones, Nature 445 (2007) 168–176.
[11] M. Zirlinger, G. Kreiman, D.J. Anderson, Proc. Natl. Acad. Sci. USA 98 (2001) 5270–5275.
[12] X. Zhao, E.S. Lein, A. He, S.C. Smith, C. Aston, F.H. Gage, J. Comp. Neurol. 441 (2001) 187–196.
[13] R. Sandberg, R. Yasuda, D.G. Pankratz, T.A. Carter, J.A. Del Rio, L. Wodicka, M. Mayford, D.J. Lockhart, C. Barlow, Proc. Natl. Acad. Sci. USA 97 (2000) 11038–11043.
[14] M.H. Chin, A.B. Geng, A.H. Khan, W.J. Qian, V.A. Petyuk, J. Boline, S. Levy, A.W. Toga, R.D. Smith, R.M. Leahy, D.J. Smith, Physiol. Genomics 30 (2007) 313–321.
[15] M.A. Zapala, I. Hovatta, J.A. Ellison, L. Wodicka, J.A. Del Rio, R. Tennant, W. Tynan, R.S. Broide, R. Helton, B.S. Stoveken, C. Winrow, D.J. Lockhart, J.F. Reilly, W.G. Young, F.E. Bloom, C. Barlow, Proc. Natl. Acad. Sci. USA 102 (2005) 10357–10362.
[16] S.M. Sunkin, J.G. Hohmann, Hum. Mol. Genet. 16 (Spec. No. 2) (2007) R209–R219.
[17] K. Sugino, C.M. Hempel, M.N. Miller, A.M. Hattox, P. Shapiro, C. Wu, Z.J. Huang, S.B. Nelson, Nat. Neurosci. 9 (2006) 99–107.
[18] L. Ng, S.D. Pathak, C. Kuan, C. Lau, H. Dong, A. Sodt, C. Dang, B. Avants, P. Yushkevich, J.C. Gee, D. Haynor, E. Lein, A. Jones, M. Hawrylycz, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (2007) 382–393.
[19] H.W. Dong, The Allen Atlas: A Digital Brain Atlas of C57BL/6J Male Mouse, Wiley, John & Sons, Incorporated, 2008.
[20] A. Visel, C. Thaller, G. Eichele, Nucleic Acids Res. 32 (2004) D552–D556.
[21] L. Ng, A. Bernard, C. Lau, C.C. Overly, H.W. Dong, C. Kuan, S. Pathak, S.M. Sunkin, C. Dang, J.W. Bohland, H. Bokil, P.P. Mitra, L. Puelles, J. Hohmann, D.J. Anderson, E.S. Lein, A.R. Jones, M. Hawrylycz, Nat. Neurosci. 12 (2009) 356–362.
[22] L.S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, R.C. Whaley, ScaLAPACK Users' Guide, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997.
[23] S.P. Lloyd, IEEE Trans. Inf. Theory 28 (1982) 129–137.
[24] J.W. Bohland, H. Bokil, C.B. Allen, P.P. Mitra, PLoS ONE, in press.
[25] D. Martin, C. Fowlkes, D. Tal, J. Malik, Proc. Int. Conf. Comput. Vis. 2 (2001) 416–425.
[26] C.H. Papadimitriou, Computing 16 (1976) 263–270.
[27] L. Ng, A. Bernard, C. Lau, C.C. Overly, H.W. Dong, L. Kuan, S. Pathak, S.M. Sunkin, C. Dang, J.W. Bohland, H. Bokil, P.P. Mitra, L. Puelles, J. Hohmann, D.J. Anderson, E.S. Lein, A.R. Jones, M. Hawrylycz, Nat. Neurosci. 12 (2009) 356–362.
[28] M.S. Boguski, A.R. Jones, Nat. Neurosci. 7 (2004) 429–433.