Cell-type–based model explaining coexpression patterns of genes in the brain

Pascal Grange[a,1], Jason W. Bohland[b], Benjamin W. Okaty[c], Ken Sugino[d], Hemant Bokil[e], Sacha B. Nelson[f], Lydia Ng[g], Michael Hawrylycz[g], and Partha P. Mitra[h]

[a]Department of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; [b]College of Health and Rehabilitation Sciences, Boston University, Boston, MA 02215; [c]Department of Genetics, Harvard Medical School, Boston, MA 02215; [d]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147; [e]Boston Scientific, Valencia, CA 91355, [f]Department of Biology and Center for Behavioral Genomics, Brandeis University, Waltham, MA; [g]Allen Institute for Brain Science, Seattle, WA 98103; [h]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited by Curtis G. Callan, Jr., Princeton University, Princeton, NJ, and approved February 20, 2014 (received for review June 28, 2013)

Spatial patterns of gene expression in the vertebrate brain are not independent, as pairs of genes can exhibit complex patterns of coexpression. Two genes may be similarly expressed in one region, but differentially expressed in other regions. These correlations have been studied quantitatively, particularly for the Allen Atlas of the adult mouse brain, but their biological meaning remains obscure. We propose a simple model of the coexpression patterns in terms of spatial distributions of underlying cell types and establish its plausibility using independently measured cell-type–specific transcriptomes. The model allows us to predict the spatial distribution of cell types in the mouse brain.

neuroscience | bioinformatics | neuroanatomy

Brain-wide and genome-wide maps of gene expression are now available (1, 2), due to the development of high-throughput neuroanatomical methods (3–6). This has enabled analysis of the spatial correlation structure of gene expression (7–12). In the Allen Brain Atlas (ABA) of the adult mouse, the brain is divided up into cubic voxels of size 200 μm. The expression energies of up to 20,000 genes in the adult C57BL/6J mouse are given by automatic processing of in situ hybridized (ISH) brain sections, coregistered to the Allen Reference Atlas (ARA) (13). Coexpression of two genes in a voxel may arise from two sources: (*i*) both genes are expressed within the same cell type or (*ii*) the two genes are expressed in two different cell types, both present in the voxel. These two possibilities cannot be disentangled solely on the basis of the ABA. Ideally, gene expression profiles should be experimentally obtained for each cell type in the brain, and indeed such transcriptome profiles are now available (14–21). This cell-based approach to the study of gene expression is complementary to the gene-based approach of the ABA. In ref. 22, the data of ref. 19 were used to extract neuron-specific genes, astrocyte-specific genes, and oligodendrocyte-specific genes, which resulted in three combinations of brain-wide maps from the ABA, whose clustering showed anatomical signatures of major brain subdivisions (see also ref. 23 for estimates of neuron-specific and oligodendrocyte-specific expression patterns, both in the mouse and in the human brain). The present paper goes beyond the broad classification of cell types into three classes and attempts to estimate density profiles of every cell type known by its transcriptome profile. To study the genes collectively we use a voxel-by-gene data matrix E, corresponding to $V = 49{,}742$ cubic voxels, and 3,041 genes, as in refs. 24–26. The entry $E(v,g)$ is the expression energy of the gene labeled g in the voxel labeled v [a measure representing the level of mRNA in situ hybridization (1, 10)]. We combine the ABA with cell-type–specific transcriptome profiles, to gain biological understanding of the coexpression patterns of the genes. Our model is based on $G = 2{,}131$ genes found in all these datasets and in the coronal ABA. The model proposed to estimate the brain-wide density profiles of cell types can be compared with the deconvolution techniques (27) in the context of microarray data and cellular types in the blood, but the brain-wide nature of the ABA allows us to interpret the results in terms of the region specificity of cell types.

Results

Brain-Wide Correlation Profiles Between the ABA and Cell-Type–Specific Transcriptomes Reveal Neuroanatomical Patterns. The singular-value decomposition (SVD) of the matrix E shows that 271 principal components (PCs) are required to explain 90% of the variance of the ABA, illustrating the large number of degrees of freedom present in the data (Fig. 1*A*). A projection onto the subspace of gene space spanned by the first three PCs shows that the voxels fall into large clusters corresponding to major brain regions, such as cerebellum and striatum (Fig. 1*B*). The first three spatial PCs also highlight these regions (Fig. 1 *C1–C3*). Brain regions can be approximately recovered by spatial clustering of the ABA (26, 28, 29–32), but estimating the density profile of cell types characterized by their transcription profile can provide a data-driven definition of some brain regions, on which there is no universal agreement (26, 28, 29).

The differences in gene expression underlying the phenotypic diversity of cell types (33, 34) have been investigated, leading to cell-type–specific microarray data. However, the region specificity of the transcriptome of cell types remains an open problem. We considered the transcriptome profiles of $T = 64$ cell types, collectively analyzed in ref. 35 and corresponding to data sets gathered in refs. 14–21. For each of these cell types, we computed the Pearson correlation coefficient of its transcriptome vector C_t with the gene expression data of the ABA at each voxel (*Methods*, Eq. **3**). In many cases, some of which are illustrated in Fig. 2, the correlation pattern is related to the brain region from which the cell type was extracted. Cerebellar granule cells (20) are found to be most correlated to the ABA in the cerebellum; medium spiny neurons (20) in the striatum; cortical pyramidal

Significance

Neuroanatomy is experiencing a renaissance due to the study of gene-expression data covering the entire mouse brain and the entire genome that have been recently released (the Allen Atlas). On the other hand, some cell types extracted from the mouse brain have been characterized by their genetic activity. However, given a cell type, it is not known in which brain regions it can be found. We propose a computational model using the Allen Atlas to solve this problem, thus estimating previously unidentified cell-type–specific maps of the mouse brain. The model can be used to define brain regions through genetic data.

Author contributions: P.G., M.H., and P.P.M. designed research; P.G. performed research; B.W.O., K.S., S.B.N., L.N., and M.H. contributed new reagents/analytic tools; P.G., J.W.B., and H.B. analyzed data; and P.G. and P.P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: pascal.grange@polytechnique.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1312098111/-/DCSupplemental.

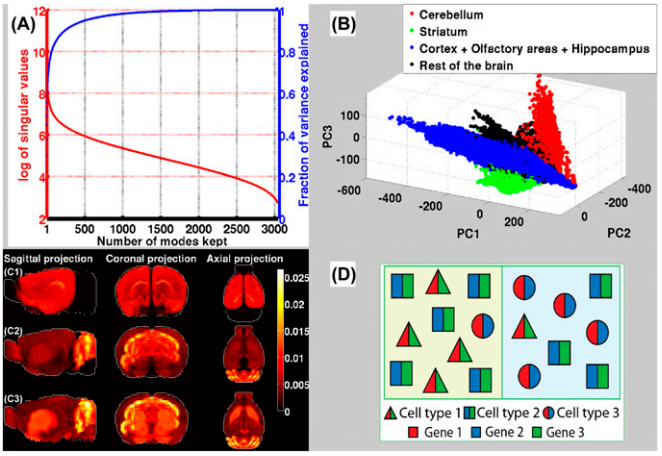

Fig. 1. The ABA in gene space. (*A*) Sorted logarithm of singular values of the ABA data (in red) and fraction of the variance explained (in blue). The 271 largest values explain 90% of the variance. (*B*) Projection of the voxels of the left hemisphere on the space spanned by the first three PCs in gene space. Some brain regions visibly fall into clusters. (*C1–C3*) Maximum-intensity projections of the first three PCs in voxel space. The second vector highlights the cerebellum and the third the striatum and the cerebellum. (*D*) A sketch of the distribution of cell types in two voxels.

neurons (14) in the cerebral cortex; amygdalar pyramidal neurons (14) in the cerebral cortex, with a "hot spot" in the amygdala; and hippocampal pyramidal neurons in the hippocampal region. However, correlation profiles are not identical to the density profiles of cell types, as the vectors C_t are not mutually orthogonal. Hence there is a need for a model to estimate the spatial distributions of cell types. (For presentation of results for all cell types in this study, see ref. 36.)

Region Specificity of Cell Types Can Be Estimated from a Linear Model. We propose a linear model for the voxelized ABA in terms of the transcriptome profiles of cell types and their spatial densities,

$$E(v,g) = \sum_{t=1}^{T} C_t(g)\rho_t(v) + \text{Residual}(v,g), \qquad [1]$$

where the index t labels the tth cell type, with gene expression profile C_t and spatial density $\rho_t(v)$ at the voxel labeled v (see Fig. 1*D* for a diagrammatic illustration). The residual term reflects other cell types (uncharacterized by their transcriptome) as well as measurement noise and nonlinearities. If the same cell type is present in two voxels, then genes highly expressed in this cell type will be highly coexpressed in the voxels. In this model, the row space of the matrix E is spanned by the vectors C_t corresponding to distinct cell types. However, cell types contributing a small amount of variance to the matrix E cannot be detected by SVD analysis. Fitting the model (Eq. **1**) amounts to solving a quadratic programming problem at each voxel with positivity contraints (*Methods*, Eq. **4**). The estimated densities ρ_t are shown (Fig. 3) for the same set of cell types as in Fig. 2 (more examples in Fig. S1). These densities are in agreement with prior biological knowledge. Importantly, the densities in Fig. 3 are significantly sparser in space than the correlation profiles in Fig. 2. This confirms that the fitting of the model (i.e., quadratic optimization with positivity constraint) yields results that cannot be reached by simple projections on the vectors C_t in gene space.

The cell-type–specific data do not sample all of the brain regions in the coarsest annotation of the ARA, which consists of 13 regions (Fig. S2). The cerebral cortex and the cerebellum are represented by more cell types than expected from their volume. Both correlation and density profiles allow us to distinguish layer-specific pyramidal neurons (Figs. 2*C* and 3*C*) from amygdalar pyramidal neurons (Figs. 2*D* and 3*D*). For each cell type, we ranked the brain regions in the ARA according to their contribution to the correlation and density profile (*SI Methods*). For instance, this procedure showed the striatum as the top region for correlation and density of medium spiny neurons; hence the sections through striatum shown in Figs. 2*B* and 3*B*. This ranking is based on the coarsest annotation of the ARA, which covers the left hemisphere only. However, the results for most cell types present a large degree of left–right symmetry. This symmetry was also observed in the expression of neuron-specific and glia-specific genes in ref. 22. We find that the predicted densities of four cerebellar cell types (Purkinje cells, granule cells, stellate basket cells, and cerebellar mature oligodendrocytes) exhibit distinct spatially nonhomogeneous patterns (Fig. 4 *B1–B4*). These patterns are consistent with neuroanatomical expectations: In particular, they define a data-driven granular layer in the cerebellum, even though granule cells are known to be much smaller than the voxel size.

Some of the predicted density profiles could be tested by further experimentation. For instance, a class of pyramidal neurons studied in ref. 14 was extracted from the prelimbic and infralimbic areas of the cerebral cortex. Our prediction for the spatial density of these pyramidal neurons is highly concentrated in the cerebral cortex, but the infralimbic and prelimbic areas rank sixth and seventh of 40 cortical regions by their contribution to the density (Fig. 4 *A1* and *A2*). The regions contributing higher fractions include the primary and secondary motor areas, which are close to the prelimbic areas, but also the ectorhinal area and temporal association areas, which are much more caudal. As the classification of cell types is a hierarchical problem, refitting the model to a dataset including the transcriptome profile of pyramidal neurons extracted from the ectorhinal area and temporal association areas would lead to a splitting of the density between these regions.

Similarity Between Groups of Cell Types Is Reflected in Additive Properties of the Estimated Densities. The transcriptome profiles C_t are not mutually orthogonal in gene space. If some of them are

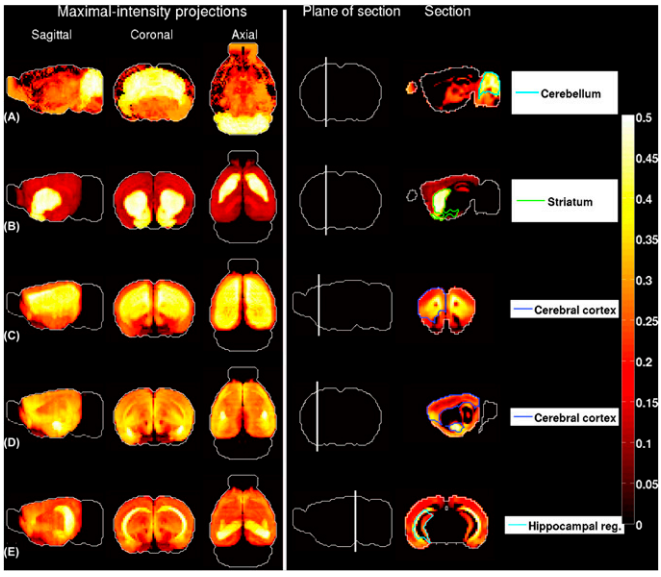

Fig. 2. Brain-wide correlation profiles between the ABA and cell-type transcriptome profiles. The color map represents all negative correlations in black. The first three columns are heat maps of maximal-intensity projections of correlation profiles, and the fifth column consists of sections through the region in the ARA with largest average correlation profile. The plane of section is depicted in the fourth column. (*A*) Granule cells [extracted from the cerebellar cortex (20)]. (*B*) Medium spiny neurons [extracted from the striatum (20)]. (*C*) Pyramidal neurons [extracted from layers 5–6 of cingulate cortex (14)]. (*D*) Pyramidal neurons [extracted from the amygdala (14)]. (*E*) Pyramidal neurons [extracted from the hippocampus (14)].

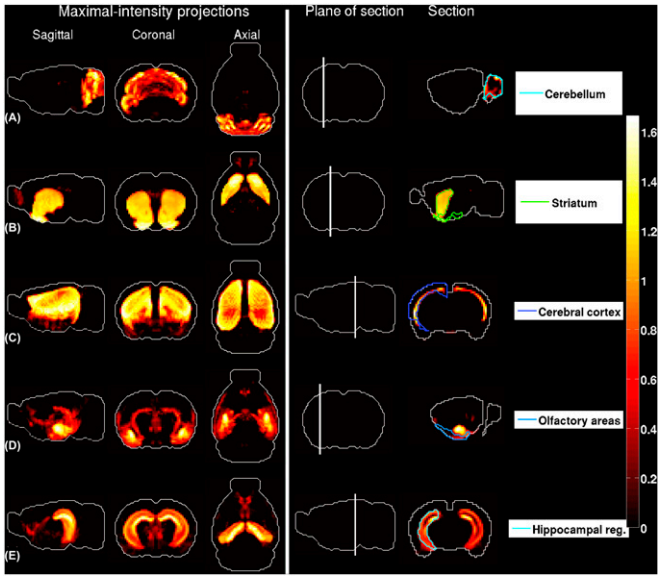

Fig. 3. Brain-wide density profiles of cell types estimated by fitting the model. The first three columns are heat maps of maximal-intensity projections of density profiles, and the fifth column consists of sections through the region in the ARA with largest contribution to the density profile. The plane of section is depicted in the fourth column. (*A*) Granule cells [extracted from the cerebellar cortex (20)]. (*B*) Medium spiny neurons [extracted from the striatum (20)]. (*C*) Pyramidal neurons [extracted from layers 5–6 of cingulate cortex (14)]. (*D*) Pyramidal neurons [extracted from the amygdala (14)]. (*E*) Pyramidal neurons [extracted from the hippocampus (14)].

almost colinear, the densities estimated in our linear model can be expected to be negatively correlated. We tested this idea in the absence of new microarray data by refitting the model to new panels of microarray data consisting of fewer cell types than C. We conducted two analyses corresponding to two choices of panels.

i) Of the 64 cell types in this study, 18 are pyramidal neurons. We combined these 18 transcriptome profiles into their average and refitted the model to a panel consisting of this composite pyramidal neuron and all of the nonpyramidal cell types (*SI Methods*). The density of the composite pyramidal neuron has a 88.75% correlation with the sum of predicted densities of pyramidal neurons in the original model (Fig. S3 shows that the density of the composite pyramidal neuron is highly localized in the cortex and differs from the sum mostly in the hippocampus and in the amygdala, presumably because of the lack of the particular cell types in Fig. 3 *D* and *E*).

ii) We computationally identified pairs of highly similar cell types in the dataset (such as the two medium spiny neurons and a pair of dopaminergic neurons). We refitted the model twice, keeping only one of the two cell types in a panel of $T-1$ cell types. Each refitted density of the remaining medium spiny neuron is highly correlated (more than 99.9%) to the sum of the two original densities (Fig. S4). The other density profiles are quite stable. This confirms that highly similar cell types have negatively correlated densities (the optimum of the model in the case of the two medium spiny neurons appears to be close to degenerate); hence the densities of these cell types should be interpreted together. However, the densities behave as expected when refitting the model.

Purity Issues and Missing Cell Types. The cell-type–specific transcriptomes were collated from different sources, so we examined the robustness of the densities by excluding some of the cell types judged to be possibly contaminated in ref. 28 (restricting the

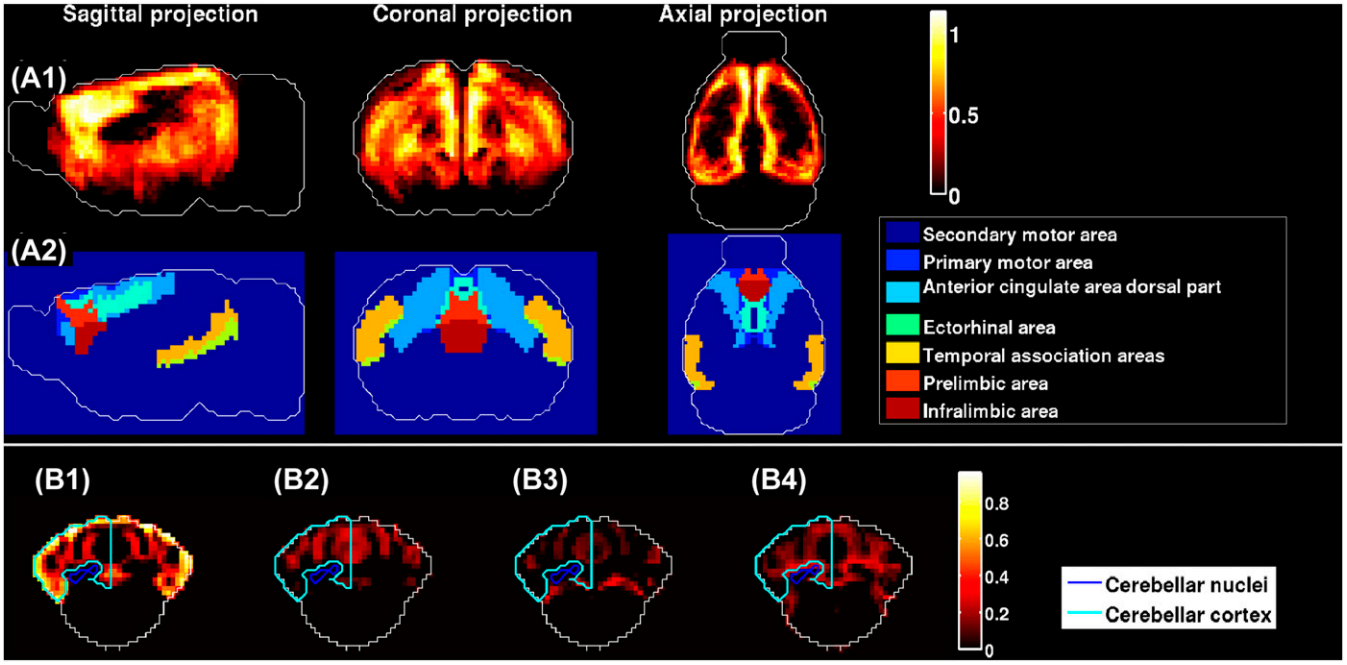

Fig. 4. (*A1*) Maximum-intensity projections of the pyramidal neurons extracted from prelimbic and infralimbic areas (14). These pyramidal neurons are indeed predicted to be found in these regions, but also in five other cortical regions contributing more to the density profile. (*A2*) Maximal-intensity projections of the top seven cortical regions contributing to the density profile. (*B*) Coronal sections through the cerebellum of the predicted density profiles of (*B1*) granule cells, (*B2*) Purkinje cells, (*B3*) stellate basket cells, and (*B4*) mature oligodendrocytes. Even though the cerebellar cortex consists of only one compartment in the digitized version of the ARA, we can define a data-driven granular layer of the cerebellum from *B1*, without having to interpolate between the hardcopy of the ARA and the grid of cubic voxels.

fitting panel to a set of 43 cell types). We found that the predicted densities are largely robust, with the exception of a cerebellar Purkinje cell type (labeled $t=52$). For this case, it was found that using the full panel of 64 cell types produced paradoxical results (the thalamus shows a larger predicted spatial density than the cerebellum, Fig. S1*E*). The predicted spatial density based on the restricted panel yields densities confined to the cerebellum for the cerebellar Purkinje cells. The other density profiles, shown in Fig. 3, did not show such sensitivity to inclusion of the contaminated cell types in the fitting panel.

Modeling Cross-Hybridization and Estimating the Errors in Densities. Microarray data are subject to cross-hybridization. No gene in this study has zero signal in any of the cell types, but every gene has zero ISH signal in at least 0.38% of the voxels (and in more than 4% of the voxels on average). Hence the data in C have a larger background value than in E. To offset the background intensity of microarray data, we assumed a uniform value across genes and cell types. [See *SI Methods* and Fig. S5 for a motivation of this calculation. Because the offset in the entries of the matrix C is uniform, the results of the correlation analysis (Fig. 2) are unchanged.] We subtracted from C its minimum entry (bringing the minimum value to 0) and refitted the model, to obtain a new estimate ρ_t^{offset} of the density of cell types. The results can be compared visually to those of the original model. For the cell types illustrated in Fig. 3, there is no dramatic difference, and the contrast of the "hot spot" is improved in the case of amygdalar pyramidal neurons (Fig. S6). All profiles have a positive correlation (75.3% on average) between the original and refitted models. The values of the residual term go down in all of the main regions in the ARA (Fig. S7), which indicates a better fit.

This improvement raises the issue of quantitative estimates of errors [upper bounds on differences between such regression results and solutions corresponding to Gaussian residuals are the object of active statistical research (37), but we take a more empirical approach]. To examine errors induced by missing cell types, we simulated an additional thalamic cell type by letting it deviate from the average transcriptome profile for 200 genes (chosen for their large expression at thalamic voxels), proportionally to the values of C_{52} at these genes. (The number was chosen because it represents about 10% of the number of genes in our dataset, which is also the fraction of the genome covered by our dataset with G genes. This number was chosen for subsampling for the same reason.) Increasing the proportionality coefficient allows us to transfer the thalamic signal from $t=52$ to the simulated type (*SI Methods*, section *10a* and Fig. S8). This suggests that some of the errors induced by missing cell types can be compensated by enriching the dataset and that modifying data on 200 genes can be sufficient in some cases to drive the results to a new optimum.

To examine errors induced by missing genes, we used a subsampling method. We repeatedly drew random sets of 200 genes, refitted the model, and computed the fraction of the new density profile $\rho_t^{(s)}$ (in subsample labeled s) supported in the original density profile:

$$\mathcal{I}(s,t) := \frac{1}{\sum_v \rho_t^{(s)}(v)} \sum_v \mathbf{1}\left(\rho_t^{\text{offset}}(v) > 0\right) \rho_t^{(s)}(v). \qquad [2]$$

The more the distribution of overlaps $\mathcal{I}(.,t)$ is concentrated at values close to 1, the more stable the prediction ρ_t^{offset} is. The average value over samples of the overlap induces a ranking of cell types. The cell types illustrated in Fig. 3 *A*–*E* rank 11th, 2nd, 10th, 19th, and 9th, respectively [see Fig. S9 *A* and *B* for a heat map of the cumulative distribution functions (CDFs) of overlaps and *SI Methods*, section *10b* for their interpretation in terms of confidence thresholds for each cell type]. For example, the profiles of medium spiny neurons ($t=16$) have at least 90% of their signal supported by ρ_t^{offset} with probability 84% (this probability falls at 23% for amygdalar pyramidal neurons, $t=48$). The lowest-ranking cell types tend to have sparser and less striking patterns (the last 32 cell types are supported on 0.98% of the voxels on average, compared with 13.98% for the first 32). The top 15 cell types include the cortical pyramidal neurons and cholinergic interneurons from the spinal cord shown in Fig. S1 *C* and *D*. Similar values of overlaps were obtained by fitting the model to a matrix C containing all genes but corrupted by noise (with the columns mixed according to Gaussian matrices in a suitable regime of noise; *SI Methods*, section *10d*).

These simulations, although showing that missing genes still allow some of the most striking predictions to survive subsampling (with a probability depending on the cell type), do not take into account the voxel-wise variation of errors that can come from missing cell types as well as from missing genes, as the measure of Eq. **2** is uniform over the brain and takes only the support of the original profile into account, rather than the full set of values. To estimate voxel-dependent confidence intervals for a fixed cell type, we used a more symmetric subsampling procedure: We repeatedly split the entire set of G genes into two random sets of equal size (up to one gene as G is odd), refitted the model to the two sets, and computed at each voxel the probability of detecting each cell type from one fitting conditional on detecting it from the other (*SI Methods*, section *10c*). These conditional probabilities turn out to be strongly correlated to the intensity of ρ_t^{offset} (on average for the cell types illustrated in Fig. 3 and Fig. S1). Thresholding the profile of amygdalar pyramidal neurons at 99% of conditional probability, for instance, yields a set of voxels consisting of the amygdalar hot spot, with all hippocampal voxels set to zero, which is also clearly discernable in the average subsampled profile, but is penalized in the measure of overlaps by voxels with weaker signal that are less stable under subsampling (Fig. S9 *D1* and *D2*). This random-splitting simulation shows that neuroanatomically striking features of the results can hold with high probability, most likely at voxels with the highest predicted value of the density. (See section 7 of ref. 38 for detailed presentations of the CDFs of the overlaps and section 8 of ref. 38 for a presentation of conditional probabilities of detection for all cell types, with heat maps of profiles thresholded at 99% and 75% thresholds on conditional probability.)

Some of the Anatomical Patterns Emerge from Biclustering of Sets of Localized Genes. We focused on the set of genes with highly localized expression to investigate links between expression patterns and neuroanatomy. Using the Kullback–Leibler divergence, we identified the 153 most localized genes. However, their spatial expressions are not entirely nonoverlapping. We therefore applied a biclustering procedure (see *SI Methods* and ref. 39 for a description of the algorithm and ref. 40 for an application to data mining of adverse drug-related events) to divide the most localized genes (and the voxel supporting their expression) into spatially least-overlapping sets. These biclusters could be associated with nonoverlapping sets of cell types. Indeed the four biclusters illustrated in Fig. 5 *A1*–*A4* resemble the neuroanatomical patterns that were observed in Fig. 2 *A*, *B*, and *E* and Fig. S1*A*. It is natural to ask, for example, whether the bicluster illustrated in Fig. 5*A1* corresponds to medium spiny neurons and whether the bicluster in Fig. 5*A2* corresponds to cell types in the cerebellum. It can indeed be seen that there is a correspondence between the two: The cerebellar and striatal biclusters respectively contain genes that are differentially expressed in cell types detected in the cerebellum and the striatum (Fig. 5*B*). This is consistent with the emergence of striatum and cerebellum in gene space observed in Fig. *1B*, *C2*, and *C3*. However, the restriction to the most localized genes excludes genes whose expression is spread over the cerebral cortex, as the cerebral cortex is a large region, and such expression profiles are therefore too close to a uniform expression to be highly ranked by our localization criterion.

Coexpression Goes Up with Biochemical Interaction. A natural question is whether the coexpression of genes can be associated with other measures of interaction between the genes. According to our model, gene coexpression can occur if genes are expressed in the

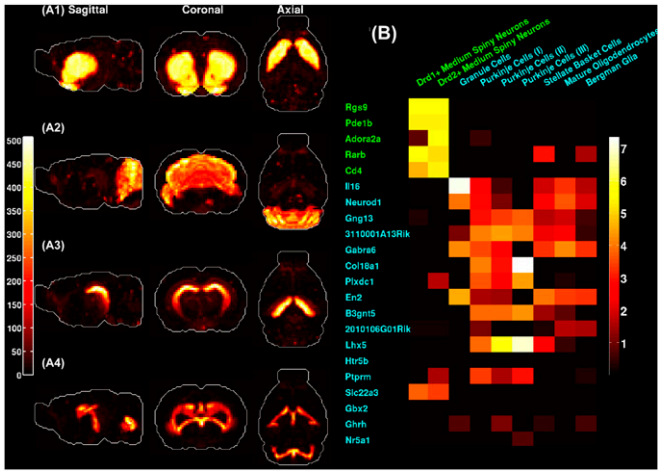

Fig. 5. Biclustering-based discovery of separable brain regions. (*A1–A4*) Maximal-intensity projection of the genes corresponding to four of the biclusters. (*B*) Separation of underlying cell types. A matrix shows the expression of genes in clusters *A1* (green) and *A2* (blue) in cell types that correlate mostly to the striatum (green) and to the cerebellum (blue), respectively.

same cell type. In that case, there is a possibility that they participate in a shared biochemical reaction. We found 921 genes in the intersection of our dataset with the Reactome database (41). We defined the coexpression of two genes as the cosine similarity between their brain-wide expression energies (*SI Methods* and refs. 24 and 25). We calculated the numbers of pairs of genes interacting in at least k reactions, for all values of k (these values are found on the horizontal axis of Fig. 6). We estimated the density of coexpression at constant values of the number of shared reactions, both for the real data encoding the participation of genes in reactions and for 100,000 matrices obtained by randomly permuting the genes and reactions. The density of coexpression is larger than expected by chance when the number of shared reactions is sufficiently large. However, high coexpression is only weakly predictive of involvement in the same biological pathway (Fig. S10).

Discussion

Our analysis shows that the ABA complements cell-based data to make neuroanatomical predictions, even though gene expression varies according to developmental stage and brain state, whereas the expression energies in the ABA represent a single state of one adult brain (42). It would be interesting to compare the developmental ABA (developingmouse.brain-map.org) to transcriptomes of cell types at the same developmental stages. Major neuroanatomical divisions, between cortex, striatum, and cerebellum, can be seen on low-dimensional projections of SVD of the ABA in gene space. Indeed these separations are recovered on correlation and density profiles of cell types extracted from these brain regions. However, some cell types are estimated by our model to have much finer region-specific properties, based on their transcriptomes, which can be far from orthogonal in gene space, unlike the principal components returned by SVD of the ABA.

Our model relies on optimization over the full gene space at each voxel, without choice of a stopping criterion. This is in contrast to the approach of ref. 22, where three sets of genes were selected on the basis of a 10-fold enrichment in each cell type in ref. 19 and declared to be respectively neuron specific (170 genes), astrocyte specific (50 genes), and oligodendrocyte specific (44 genes). The corresponding three sets of genes from the ABA were then clustered using the k-means algorithm, which yielded clusters corresponding to neuroanatomical regions, especially in the case of the neuron-specific genes (the analysis in ref. 22 was based on ISH image series rather than gene-expression energies). Our approach is less dependent on the solidarity between voxels than the clustering approach, because the optimization problem is solved one voxel at a time, and we integrated more cell-type–specific samples into the analysis. Our results associate several major brain regions to specific cell types rather than classes of cell types, thus sharpening the complementary relationship between the gene-based and the cell-based approaches to gene expression in the brain. Pairs of genes belonging to several shared cellular biochemical pathways tend to be more strongly coexpressed (other attempts to leverage the ABA data and functional data include the Gene Ontology-based analysis of ref. 22). The emergence of brain regions from density profiles, despite the heterogeneity of the datasets and the lack of dynamics, is an encouraging confirmation of the validity of the genomic classification of cell types. Moreover, the estimated densities are independent from the annotation of voxels by classical neuroanatomy. This could be of special interest in the case of songbirds (2), as the very existence of some brain regions (such as the cerebral cortex) in the zebra finch is an open problem, whereas the existence of localized cell types could be established in a genetic way (even though the nomenclature of the brain regions from which the cells are isolated is controversial). The availability of the ABA of the human brain (43) gives rise to more technical challenges due the paucity of specimens and to the partial brain coverage. Importantly, our model enables single-cell transcriptome analysis to be extended to the spatial distribution of cell types. The errors in the estimates vary across cell types and voxels because of missing genes and cell types, but confidence intervals are accessible to numerical simulations, and some striking neuroanatomical patterns hold with high probability. The model will remain applicable as the taxonomy of cell types matures.

Methods

Correlations Between Cell-Type–Specific Transcriptomes and ABA. The value at voxel v for cell type t reads as follows:

$$\mathrm{Corr}(v,t)=\frac{\sum_{g=1}^{G}\left(C_t(g)-\sum_{t=1}^{T}(C_t(g)/T)\right)\left(E(v,g)-\sum_{v=1}^{V}(E(v,g)/V)\right)}{\sqrt{\sum_{k=1}^{G}\left(C_t(k)-\sum_{t=1}^{T}(C_t(k)/T)\right)^2\sum_{h=1}^{G}\left(E(v,h)-\sum_{v=1}^{V}(E(v,h)/V)\right)^2}}. \quad [3]$$

Estimating Brain-Wide Density of Cell Types. For a given voxel v, the numbers $\rho_t(v)_{1\le t\le T}$ are the entries of the vector in $\mathbf{R}_+^T$ that is mapped by the matrix C to the best possible approximation of the vector $E(v,g)_{1\le g\le G}$. They minimize a quadratic form under positivity constraint:

$$(\rho_t(v))_{1\le t\le T}=\mathrm{argmin}_{\nu\in\mathbf{R}_+^T}\sum_{g=1}^{G}\left(E(v,g)-\sum_{t=1}^{T}C_t(g)\nu(t)\right)^2. \quad [4]$$

We solved these quadratic optimization problems, one per voxel, using the CVX toolbox (44, 45).

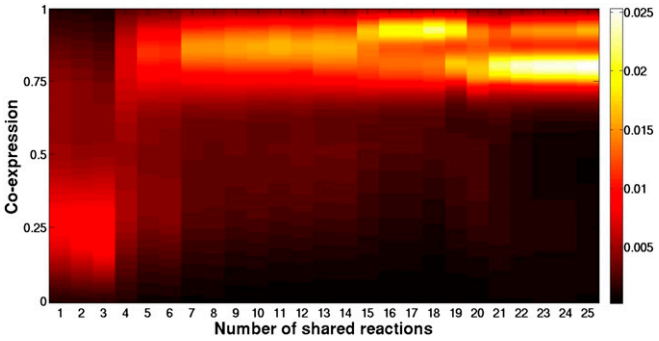

Fig. 6. Coexpression of genes in the ABA and the Reactome database. Shown is a heat map of the density function of coexpression of pairs of genes, as a function of the minimal number of reactions shared by the genes.

ACKNOWLEDGMENTS. We thank Nicolai Meinshausen for correspondence. This research was supported by National Institutes of Health–National Institute on Drug Abuse Grant 1R21DA027644-01, "Computational analysis of coexpression networks in the mouse and human brain."

1. Lein ES, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124):168–176.
2. Warren WC, et al. (2010) The genome of a songbird. *Nature* 464(7289):757–762.
3. Bota M, Dong HW, Swanson LW (2003) From gene networks to brain networks. *Nat Neurosci* 6(8):795–799.
4. Ng L, Hawrylycz M, Haynor D (2005) Automated high-throughput registration for localizing 3D mouse brain gene expression using ITK. 2005 MICCAI Open-Source Workshop. *Insight J*, Available at http://hdl.handle.net/1926/15.
5. Sunkin SM, Hohmann JG (2007) Insights from spatially mapped gene expression in the mouse brain. *Hum Mol Genet* 16 (Spec No. 2):R209–R219.
6. Hawrylycz M, et al. (2011) Digital atlasing and standardization in the mouse brain. *PLoS Comput Biol* 7(2):e1001065.
7. Ng L, et al. (2007) NeuroBlast: A 3D spatial homology search tool for gene expression. *BMC Neurosci* 8(Suppl 2):11.
8. Ng L, et al. (2007) Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Trans Comput Biol Bioinformatics* 4(3):382–393.
9. Lee CK, et al. (2008) Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome Biol* 9(1):R23.
10. Ng L, et al. (2009) An anatomic gene expression atlas of the adult mouse brain. *Nat Neurosci* 12(3):356–362.
11. Jones AR, Overly CC, Sunkin SM (2009) The Allen Brain Atlas: 5 years and beyond. *Nat Rev Neurosci* 10(11):821–828.
12. Hawrylycz M, et al. (2011) Multi-scale correlation structure of gene expression in the brain. *Neural Netw* 24(9):933–942.
13. Dong HW (2007) *The Allen Reference Atlas: A Digital Brain Atlas of the C57BL/6J Male Mouse* (Wiley, New York).
14. Sugino K, et al. (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nat Neurosci* 9(1):99–107.
15. Chung CY, et al. (2005) Cell type-specific gene expression of midbrain dopaminergic neurons reveals molecules involved in their vulnerability and protection. *Hum Mol Genet* 14(13):1709–1725.
16. Arlotta P, et al. (2005) Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo. *Neuron* 45(2):207–221.
17. Rossner MJ, et al. (2006) Global transcriptome analysis of genetically identified neurons in the adult cortex. *J Neurosci* 26(39):9956–9966.
18. Heiman M, et al. (2008) A translational profiling approach for the molecular characterization of CNS cell types. *Cell* 135(4):738–748.
19. Cahoy JD, et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: A new resource for understanding brain development and function. *J Neurosci* 28(1):264–278.
20. Doyle JP, et al. (2008) Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell* 135(4):749–762.
21. Okaty BW, Miller MN, Sugino K, Hempel CM, Nelson SB (2009) Transcriptional and electrophysiological maturation of neocortical fast-spiking GABAergic interneurons. *J Neurosci* 29(21):7040–7052.
22. Ko Y, et al. (2013) Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain. *Proc Natl Acad Sci USA* 110(8):3095–3100.
23. Tan PPC, French L, Pavlidis P (2013) Neuron-enriched gene expression patterns are regionally anti-correlated with oligodendrocyte-enriched patterns in the adult mouse and human brain. *Front Neurosci* 7:5.
24. Grange P, Hawrylycz M, Mitra PP (2013) Computational neuroanatomy and co-expression of genes in the adult mouse brain, analysis tools for the Allen Brain Atlas. *Quantitative Biology* 1(1):91–100.
25. Menashe I, Grange P, Larsen EC, Banerjee-Basu S, Mitra PP (2013) Co-expression profiling of autism genes in the mouse brain. *PLoS Comput Biol* 9(7):e1003128.
26. Bohland JW, et al. (2010) Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy. *Methods* 50(2):105–112.
27. Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* 4(7):e6098.
28. Grange P, Mitra PP (2012) Computational neuroanatomy and gene expression: Optimal sets of marker genes for brain regions. *IEEE 2012 46th Annual Conference on Information Science and Systems* (IEEE, Princeton, NJ). pp 1–6.
29. Bohland JW, Bokil H, Allen CB, Mitra PP (2009) The brain atlas concordance problem: Quantitative comparison of anatomical parcellations. *PLoS ONE* 4(9):e7200.
30. Swanson LW (2004) *Brain Maps* (Academic, New York).
31. Paxinos G, Franklin KBJ (2008) *The Mouse Brain in Stereotaxic Coordinates* (Elsevier Academic, Amsterdam).
32. Guillery RW (2002) On counting and counting errors. *J Comp Neurol* 447(1):1–7.
33. Mott DD, Dingledine R (2003) Interneuron Diversity series: Interneuron research—challenges and strategies. *Trends Neurosci* 26(9):484–488.
34. Markram H, et al. (2004) Interneurons of the neocortical inhibitory system. *Nat Rev Neurosci* 5(10):793–807.
35. Okaty BW, Sugino K, Nelson SB (2011) A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. *PLoS ONE* 6(1):e16493.
36. Grange P, Hawrylycz M, Mitra PP (2013) Cell-type-specific microarray data and the Allen Atlas: Quantitative analysis of brain-wide patterns of correlation and density. Available at http://xxx.lanl.gov/pdf/1303.0013.pdf.
37. Meinshausen N (2013) Sign-constrained least squares estimation for high-dimensional regression. *Electron J Stat* 7:1607–1631.
38. Grange P, et al. (2014) Cell-type-specific microarray data and the Allen Atlas (II): Discussion of the linear model of brain-wide densities of cell types. Available at arXiv.org/abs/1402.2820.
39. Grady L, Schwartz E (2006) Isoperimetric partitioning: A new algorithm for graph partitioning. *SIAM J Sci Comput* 27(6):1844–1866.
40. Harpaz R, et al. (2011) Biclustering of adverse drug events in FDAs spontaneous reporting system. *Clin Pharmacol Ther* 89(2):243–250.
41. Vastrik I, et al. (2007) Reactome: A knowledge base of biologic pathways and processes. *Genome Biol* 8(3):R39.
42. Mitra PP, Pesaran B (1999) Analysis of dynamic brain imaging data. *Biophys J* 76(2):691–708.
43. Hawrylycz MJ, et al. (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489(7416):391–399.
44. Research, Inc. (2012) *CVX: Matlab Software for Disciplined Convex Programming, Version 2.0 beta*. Available at http://cvxr.com/cvx. Accessed September 2012.
45. Grant M, Boyd S (2008) Graph implementations for nonsmooth convex programs. *Recent Advances in Learning and Control (a tribute to M. Vidyasagar)*, Lecture Notes in Control and Information Sciences, eds Blondel V, Boyd S, Kimura H (Springer, New York), pp 95–110. Available at http://stanford.edu/~boyd/graph_dcp.html.