Computational neuroanatomy and co-expression of genes in the adult mouse brain, analysis tools for the Allen Brain Atlas

Pascal Grange [1], Michael Hawrylycz [2] and Partha P. Mitra [1]

[1] *Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, New York 11724, United States*

[2] *Allen Institute for Brain Science, Seattle, Washington 98103, United States*

Abstract

We review quantitative methods and software developed to analyze genome-scale, brain-wide spatially-mapped gene-expression data. We expose new methods based on the underlying high-dimensional geometry of voxel space and gene space, and on simulations of the distribution of co-expression networks of a given size. We apply them to the Allen Atlas of the adult mouse brain, and to the co-expression network of a set of genes related to nicotine addiction retrieved from the NicSNP database. The computational methods are implemented in `BrainGeneExpressionAnalysis`, a Matlab toolbox available for download.

Contents

1 Introduction and background

The mammalian brain is a structure of daunting complexity, whose study started millenia ago and has been recently renewed by molecular biology and computational imaging [1]. The Allen Brain Atlas, the first Web-based, genome-wide atlas of gene expression in the adult mouse brain, was a large-scale experimental effort [2, 4, 3, 5, 6, 7]. The resulting dataset consists of co-registered *in situ* hybridization (ISH) image series for thousands of genes. It is now available to neuroscientists world-wide, and has given rise to the development of quantitative techniques and software for data analysis. The present paper reviews recent developments that have been applied to co-expression studies in the mouse brain and are publicly available for use on the Web [8] and on the desktop [9].

On the other hand, lists of condition-related genes are now available from databases that pool results of different studies [10, 11]. As these studies employ different methods and result in lists of hundreds of genes, it is important to investigate any possible order (or lack of it) in these lists. The Allen Brain Atlas provides ways to do this, by stuying brain-wide co-expression of genes, and by enabling to compare gene expression to classical neuroanatomy, in a genome-scale dataset based on a unified protocol.

Advanced data exploration tools have already been developed for the Allen Brain Atlas. NeuroBlast allows users to explore the correlation structure between genes in the ABA. was inspired by the Basic Local Alignment Research Tool [12], which derives lists of similar genes to a given gene at the level of sequences, and transposed the technique to the analysis of similarity between patterns of gene expression in the brain [13]. The Anatomic Gene Expression Atlas [14] was launched in 2007. It is based on the spatial correlation of the atlas. The user can explore three-dimensional correlation maps based on correlations between voxels, computed using thousands of genes, and retrieve hierarchical data-driven parcellations of the brain.

The Weighted Gene Co-Expression Network Analysis framework (WGCNA) has been used to isolate clusters of genes from correlations between multiple microarray samples. In this approach the gene networks are typically constructed from the correlation coefficients of microarray data, from which graphs are constructed and thresholded at a value chosen as as to satisfy certain statistical criteria [15, 16]. However, in the case of the Allen Brain Atlas, gene-expression data are scaffolded by classical neuroanatomy, since ISH data are co-registered to the Allen Reference Atlas (ARA) [17]. The whole brain is voxelized, and the voxels are are annotated according to the brain region to which they belong, which allows to compare the expression of sets of genes to brain regions (see Figure 6 and [18]). Hence we developed computational methods to:
1. study the whole range of co-expression values between pairs of genes;
2. use the Allen Atlas as a probabilistic universe to estimate the distribution of co-expression networks;
3. compare the expression patterns of highly co-expressed sets of genes to classical neuroanatomy.

These methods are implemented in `BrainGeneExpressionAnalysis` (BGEA), a Matlab toolbox downloadable from `www.brainarchitecture.org`. They are applied to a set of 288 genes extracted from the NicSNP database, which have been linked to nicotine dependence, based on the statistical significance of allele frequency difference between cases and controls, and for which mouse orthologs are found in the coronal Allen Atlas.

2 Methods

The spatial frequency of tissue-sectioning in the experimental pipeline of the Allen Brain Atlas corresponds to slices with a thickness of 100 micrometers. Each section was registered to a grid with a resolution of 100 microns [20, 21]. The induced three-dimensional grid was sub-sampled to a resolution of 200 microns in order to increase the overlap between different experiments. This procedure results in a partition of the mouse brain into $V = 49,742$ cubic voxels. We focus on the co-registered quantities obtained at a spatial resolution of 200 micrometers, for several thousands of genes, after subsampling.

In particular, the expression energy of each gene labelled g in the Atlas was defined and computed [14] at each voxel labelled v in the mouse brain:

$$E(v,g) = \frac{\sum_{p\in v} M(p)I(p)}{\sum_{p\in v} 1}, \tag{1}$$

where p is a pixel index, and the denominator counts the pixels that are contained in the voxel v for the ISH image series of gene g. The quantity $M(p)$ is a Boolean segmentation mask that takes value 1 at pixels classified as expressing the gene, and 0 at other pixels. The quantity $I(p)$ is the grayscale value of the pixels in ISH images. The present paper uses the voxel-by-gene matrix of expression energies E as the digitized version of the Allen Brain Atlas. The expression energies of the genes in the full coronal and sagittal atlas can be downloaded using the Web service provided by the Allen Institute [22].

2.1 Brain-wide co-expression networks: graph properties

The statistical study of brain-wide co-expression networks using BGEA is summarized in the flowchart of Figure 1. Detailed examples of the use of the software are provided in toolbox manual [9].

The columns of the matrix E of expression energies of Equation 1 are naturally identified to vectors in a V-dimensional space (the voxel space). Given two genes, the two corresponding columns of the matrix E span a two-dimensional vector of voxel space. The simplest geometric quantity to study for this system is the angle between the two vectors. As all the entries of the matrix E are positive by construction, this angle is between 0 and $\pi/2$. The angle between the two vectors is therefore completely characterized by its cosine, which is readily expressed in terms of expression energies. This cosine similarity, defined in Equation 2, for genes labelled g and g', is called the co-expression of genes g and g'.

$$\mathrm{coExpr}(g,g') = \sum_{v=1}^{V} \frac{E(v,g)E(v,g')}{\sqrt{\sum_{u=1}^{V} E(u,g)^2 \sum_{w=1}^{V} E(w,g)^2}}. \tag{2}$$

The more co-expressed g and g' are in the brain, the closer their cosine similarity is to 1.

Once the co-expressions have been computed for all pairs of genes in the Allen Brain Atlas, they are naturally arranged in a matrix, denoted by C^{atlas}, with the genes arranged in the same order as the list of genes in the atlas:

$$C^{\mathrm{atlas}}(g,g') = \mathrm{coExpr}(g,g') \quad 1 \le g,g' \le G_{\mathrm{atlas}}, \tag{3}$$

where G_{atlas} is the total number of genes included in the dataset (see the next section for more details on this choice). The matrix C^{atlas} is symmetric and its diagonal entries are all

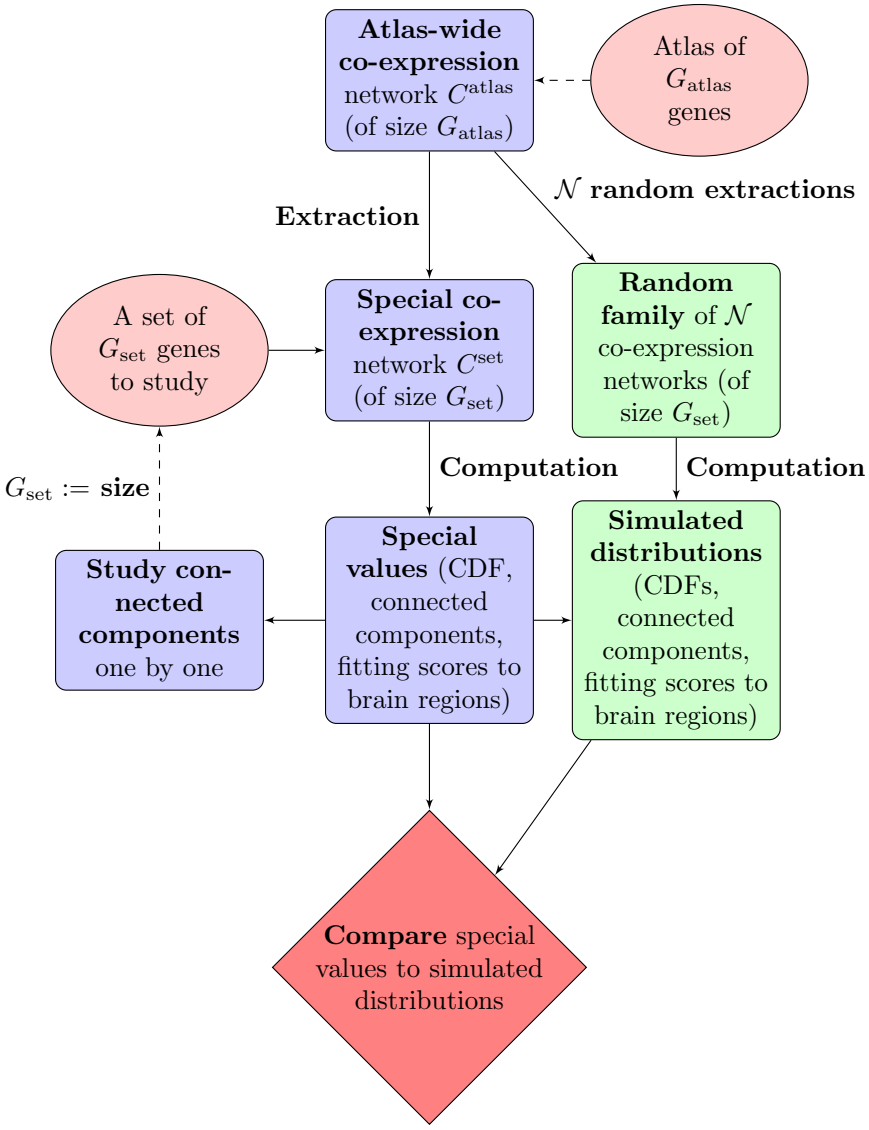


Figure 1: The flowchart of computational analysis of the collective neuroanatomical properties of a set of genes in the `BrainGeneExpressionAnalysis` toolbox. The steps marked as random extractions and computations are described in supplementary materials.

equal to one. This diagonal is trivial in the sense that it expresses the perfect alignment of any vector in voxel space with itself. When we consider the distribution of the entries of the co-expression matrix, we really mean the distribution of the upper-diagonal coefficients.

Given a set of genes (with G_{set} elements) curated from the literature, possibly coming from different studies, one may ask if the brain-wide expression profiles of these genes (or a subset thereof) are closer to each other than expected by chance, using the full atlas as a probabilistic universe. The set of genes for which brain-wide expression data are available from the Allen Atlas of the adult mouse brain consists of 4,104 genes, which is of the same order of magnitude as the total number of genes in the mouse genome. The number of sets of genes of a given size that can be drawn from the atlas therefore grows quickly with the size of the set. To study the co-expression properties of the chosen set of genes, a G_{set}-by-G_{set} matrix C^{set} can be extracted from the whole co-expression matrix C^{atlas}. A set of strongly co-expressed genes corresponds to a matrix C^{set} *with large coefficients*. To formalise this idea, we propose to study the matrix in terms of the underlying graph. There are $G_{\text{atlas}}!$ ways of ordering the genes in the Atlas. They give rise to different co-expression matrices, related by similarity transformation. But the *sets* of highly co-expressed genes are invariant under these transformations. The co-expression matrix can be mapped to a weighted graph in a straightforward way. The vertices of the graph are the genes, and the edges are as follows:
- genes g and g' are linked by an edge if their co-expression $\text{coExpr}(g, g')$ is strictly positive.
- If an edge exists, it has weight $\text{coExpr}(g, g')$.

We have to define large co-expression matrices in relative terms, using thresholds on the value of co-expression that describe the whole set of possible values. The entries of the co-expression matrices are numbers between 0 and 1 by construction. We define the following thresholding procedure on co-expression graphs: given a threshold ρ between 0 and 1, and a co-expression matrix (which can come from any set of genes in the Allen Atlas), put to zero all the coefficients that are lower than the threshold (see Figure (2) for an illustration on a toy-model with 9 genes).

The graph corresponding to a co-expression matrix has connected components, and each connected component has a certain number of genes in it. The graph properties of C^{set} can be studied by computing the average and maximal size of the connected components at every value of the threshold. This induces functions of the threshold that can be compared to those obtained from $\mathcal{N}$ random sets of genes of the same size G_{set} (these computations on random sets of genes correspond to the two green boxes in Figure 1,see Supplementary Materials S2 for mathematical details).

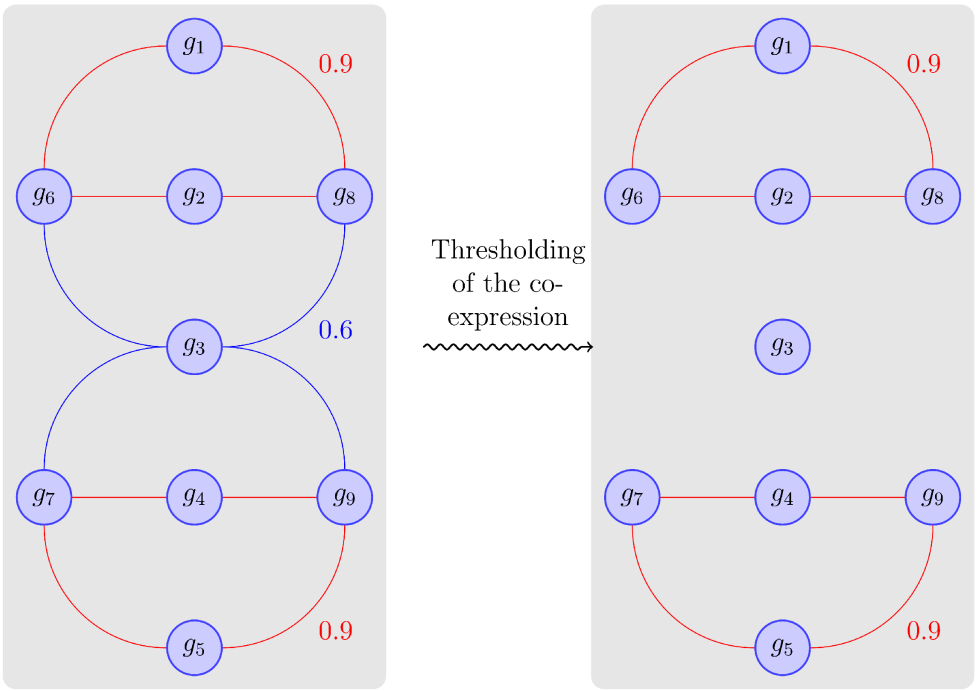


Figure 2: A toy model with 9 genes, and only three distinct values of co-expression, 0, 0.6 and 0.9, for simplicity. Before any thresholding procedure is applied (on the left-hand side of the figure), there is one connected component. The average and maximum size of connected components are both 9. The graph on the right-hand side is obtained by a thresholding procedure at a threshold of 0.6. There are three connected connected components, the maximal size is 4, and the average size if 3.

2.2 Cumulative distribution functions of co-expression

complement the graph-theoretic approach, we can study the cumulative distribution function of the entries of the co-expression matrix of the set of genes to study, and compare it to the one resulting from random sets of genes of the same size (see Supplementary Materials S2 and S3 for mathematical details). For every number between 0 and 1, the empirical cumulative distribution function of C^{set}, denoted by CDF^{set} is defined as the fraction of the entries of the upper-diagonal part of the co-expression matrix that are smaller than this number.

To compare the co-expression network of interest C^{set} to random networks of the same size, the procedure is exactly the same as with the thresholded matrices, except that the quantities computed from the $\mathcal{N}$ random draws are cumulative distribution functions rather than connected components (see Supplementary Materials for mathematical details). For each random set of G_{set} genes drawn from the Allen Atlas, one can compute the empirical distribution function of the corresponding submatrix of C^{atlas}, and average over the draws. The average over the draws converges towards the one of a typical network of G_{set} genes when the number of random draws is sufficiently large.

2.3 Comparison to classical neuroanatomy

Given a brain region ω_r, $1 \leq r \leq R$, where R is the number of brain regions in the Allen Reference Atlas [17] (to which gene expression data are registered), the fitting score of a brain-wide function f in this region, or $\phi_r(f)$ can be defined [18] as the cosine distance between this function and the characteristic function of the region. It is formally the same as the co-expression of a gene whose expression energy would be the brain-wide function, and a another gene that would be uniformly expressed in the region, and nowhere else:

$$\phi_r(f) = \sum_{v \in \Omega} \frac{f(v)\mathbf{1}(v \in \omega_r)}{\sqrt{\sum_{u \in \Omega} f(u)^2}\sqrt{\sum_{w \in \Omega} \mathbf{1}(w \in \omega_r)^2}}. \tag{4}$$

The distribution of fitting scores in all the brain regions for sets of G_{set} genes can be simulated by the Monte Carlo methods described in Supplementary Materials S4.

Even though clustering methods [19] have shown that the correspondence between large sets of genes and brain regions in the Allen Atlas is not perfect, it is possible to detect small subsets of a set of genes curated from the literature to have exceptionally good fitting properties in some brain regions (see Figure 8 for an example of a set of 3 genes detected to fit the striatum significantly better than expected by chance).

3 Applications

3.1 Choice of genes: coronal and sagittal atlases

The notion of an atlas of gene expression in the adult mouse brain rests on the assumption that there is a constant component across all brains at the final stage of development (the developmental atlas adresses the challenge of measurement of this component at earlier stages [23]). For an account of the standardization process that began in 2001 and led to the data generation and release of the Allen Brain Atlas, see [6].

The issue of reproducibility of ISH data can been addressed in several ways during the analysis of data. In NeuroBlast, the user can specify a given image series as input. The `BrainGeneExpressionAnalysis` toolbox (BGEA) is based on the analysis of the matrix of expression energies 1, whose columns consist of brain-wide gene-expression data. This restricts the choice of genes to be analyzed in by BGEA to the 4,104 genes for which a brain-wide, coronal atlas was developed. For these genes, sagittal, registered data are also available in the left hemisphere. We computed correlation coefficients between sagittal and coronal data. The left-right correlation coefficients are not all positive. Sagittal datasets usually come from brain sections taken from the left hemisphere only. Hence the computation of correlation between (co-registered) sagittal and coronal data has to be restricted to the voxels belonging to the left hemisphere. For each gene g in the coronal atlas, we computed the following correlation coefficient between sagittal and coronal data:

$$\rho_{\text{sagittal/coronal}}(g) = \frac{\sum_{v\in\text{left hemisphere}} E_{\text{sagittal}}(v,g)E_{\text{coronal}}(v,g)}{\sqrt{\sum_{v\in\text{left hemisphere}} E_{\text{sagittal}}(v,g)^2 \sum_{v\in\text{left hemisphere}} E_{\text{coronal}}(v,g)^2}}, \qquad (5)$$

where E_{sagittal} and E_{coronal} are the voxel-by-gene matrices of Equation 1 for sagittal and coronal data respectively. The results are shown on Figure 3. Some genes have negative correlation between sagittal and coronal data. The gene with highest value of $\rho_{\text{sagittal/coronal}}$ is *Tcf7l2*. The present study focuses on genes for which the correlation is larger than the 25th percentile of the distribution of $\rho_{\text{sagittal/coronal}}$.

This set of $GAtlas := 3,041$ genes serves as a reference set to which special sets of genes can be compared using the methods described above. In particular, this choice excludes all the genes with negative correlation. Other user-defined choices of genes are possible within the coronal atlas. They can be implemented by modifying the data matrix 1 and the list of genes corresponding to its columns in BGEA.

The sorted entries of the upper-diagonal part of the induced co-expression matrix C^{atlas} are plotted on Figure 4(a). The pair of genes with highest co-expression are *Atp6v0c* and *Atp2a2*, whose expression profiles are plotted on Figures 4(b,c). The profile of co-expressions is fairly linear, except at the end of the spectrum, which motivates a uniform exploration of

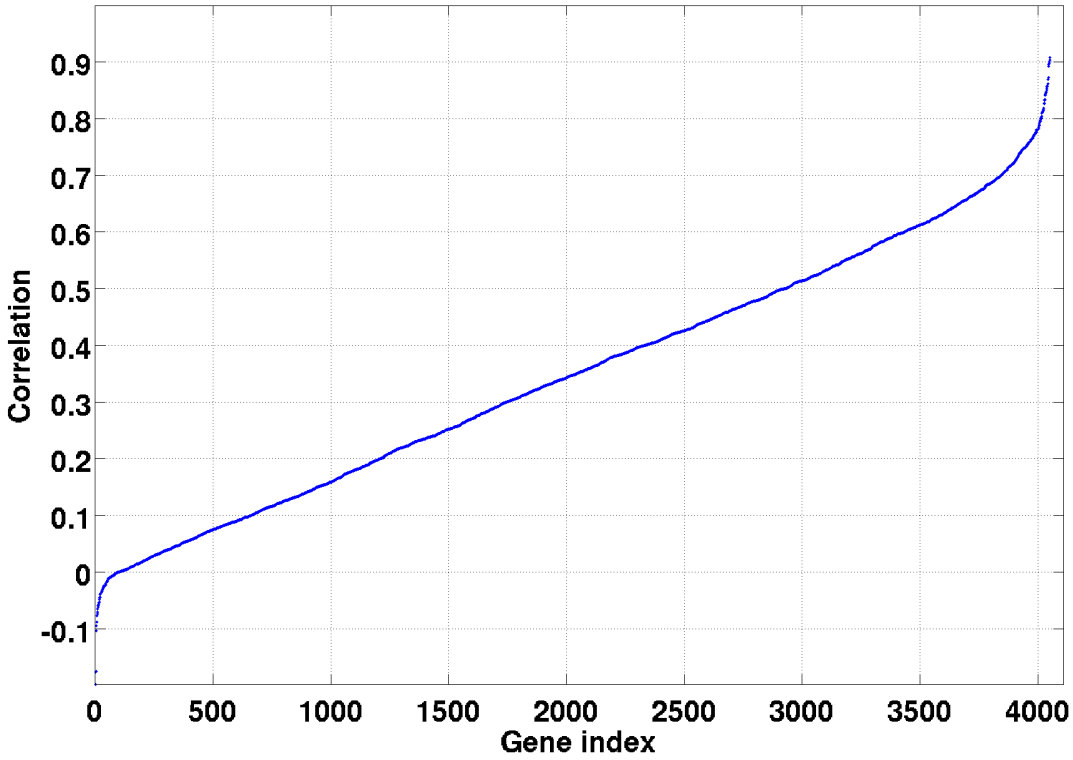


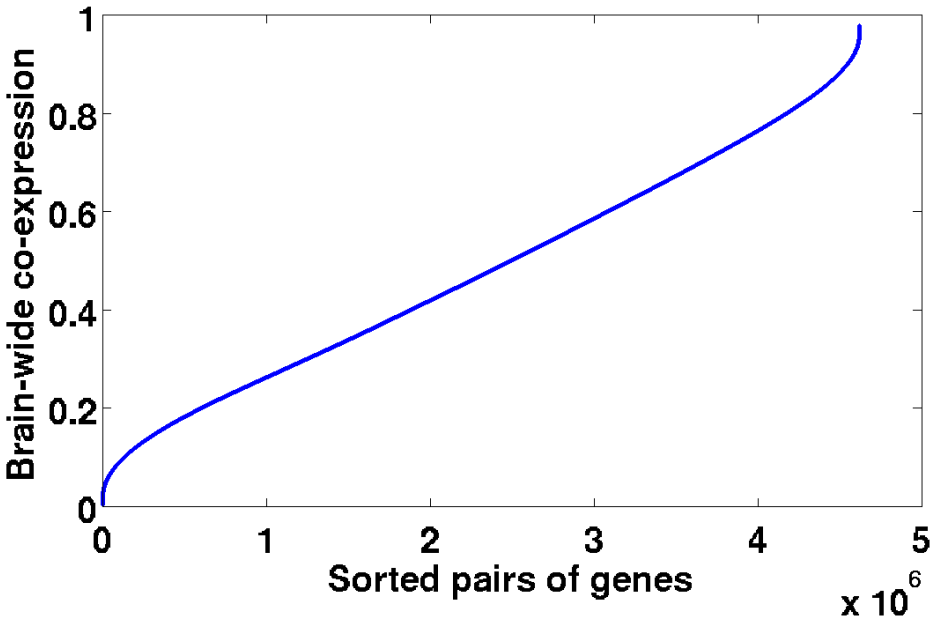
Figure 3: Sorted correlation coefficients between expression energies evaluated from sagittal and coronal sections in the left hemisphere of the mouse brain.

the interval $[0, 1]$ when studying co-expression networks (see the pseudocode in Supplementary Materials S2).

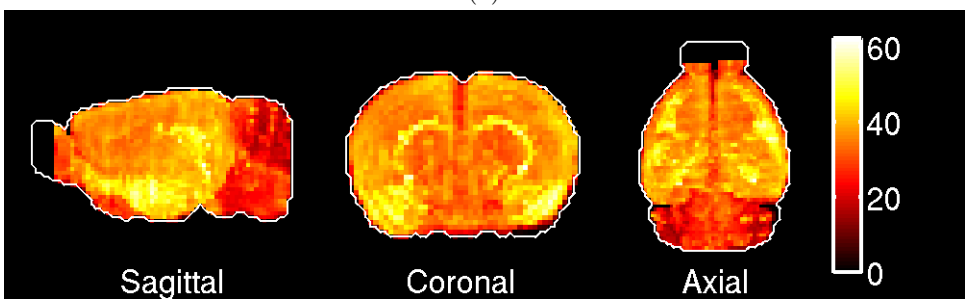
3.2 Application to a set of addiction-related genes

The methods reviewed above were applied to a set of 288 genes related to nicotine addiction [11], retrieved from the NicSNP database[1]. The simulation of the cumulative distribution function of co-expression networks of size 288 can be compared to the one of the special set, and plotted together on 5. Since the CDF of the special sets is larger than average at low values of co-expression, the special set is not more co-expressed as a whole than expected by chance. This is confirmed by the statistics of graph properties of networks of 288 genes (Figures 6 and 7). See [24] for a set of autism-related genes that is more co-expressed in the brain than expected by chance).

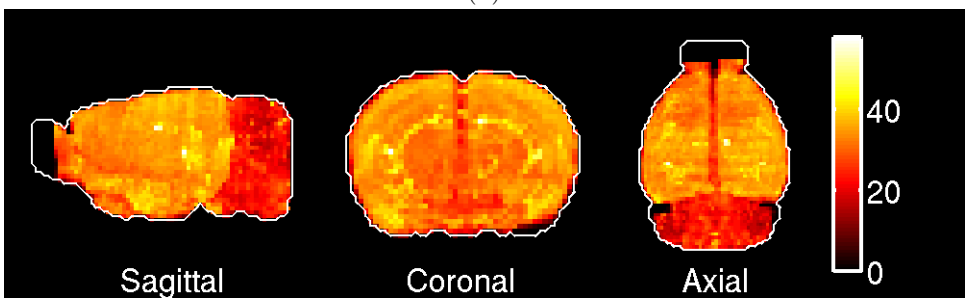
[1] `http://zork.wustl.edu/nida/Results/data1.html`



(a)



(b)



(c)

Figure 4: (a) Sorted elements of the upper-diagonal part of the co-expression matrix of the coronal atlas, C^{atlas}. (b) Maximal-intensity projection of the expression energy of *Atp6v0c*. (c) Maximal-intensity projection of the expression energy of *Atp2a2*. The pair of genes (*Atp6v0c*,*Atp2a2*) has the highest co-expression in the coronal atlas, 0.9781.

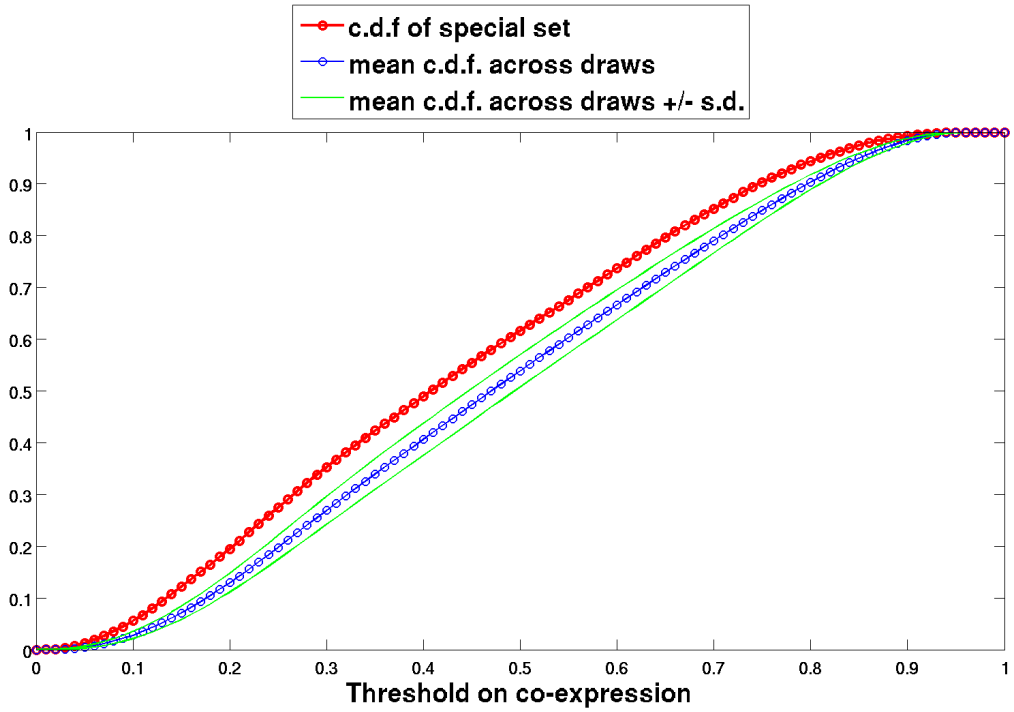


Figure 5: Cumulated distribution function of the upper-diagonal entries of the co-expression matrix of 288 genes (the special co-expression network C^{set} of the flowchart 1) from the NicSNP database, for which mouse orthologs are found in the Allen Atlas of the adult mouse brain. As the red curve (or CDF^{set}, Equation 15) sits above the simulated average of the simulated mean of CDFs (or $\langle \text{CDF} \rangle$, Equation 17) of co-expression networks of 288 genes, at low values of the threshold, the special set as a whole appears to be less co-expressed than expected by chance.

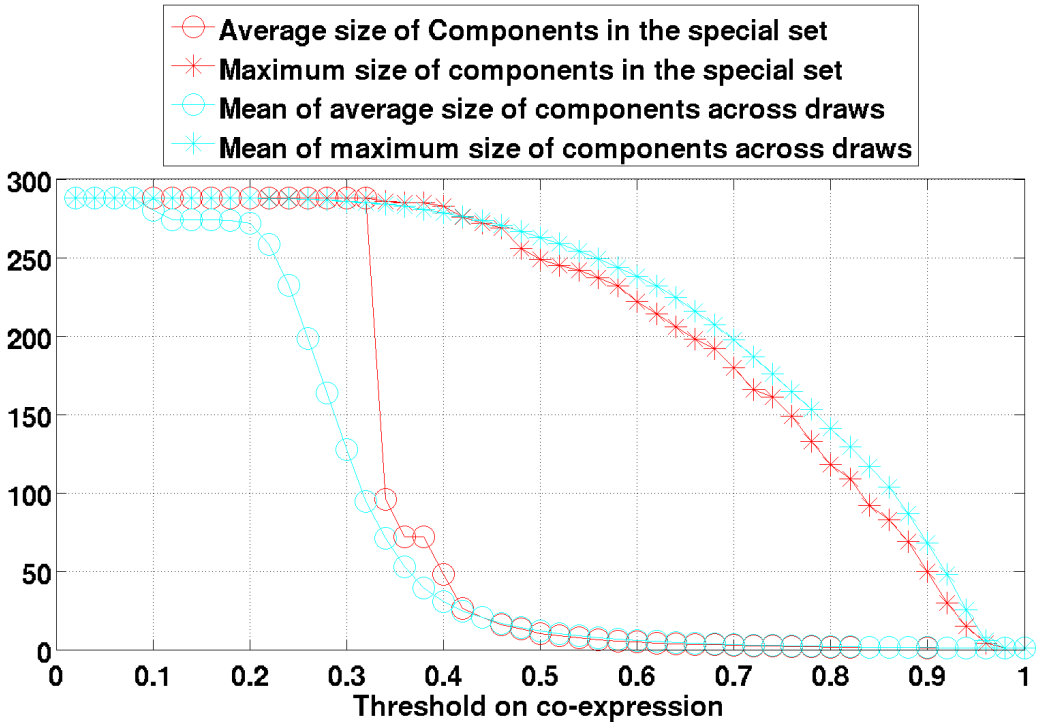


Figure 6: **Monte Carlo analysis of the graph underlying the co-expression matrix of 288 genes from the NicSNP database.** Average and maximum size of connected components as a function of the threshold (the quantities $\mathcal{A}(\rho)$ and $\mathcal{M}(\rho)$ defined in Equations 8 and 9 are plotted in red circles and red stars respectively, the quantities $\langle \mathcal{A}(\rho) \rangle$ and $\langle \mathcal{M}(\rho) \rangle$ defined in Equations 11 and 10 are plotted in cyan circles and cyan stars respectively).

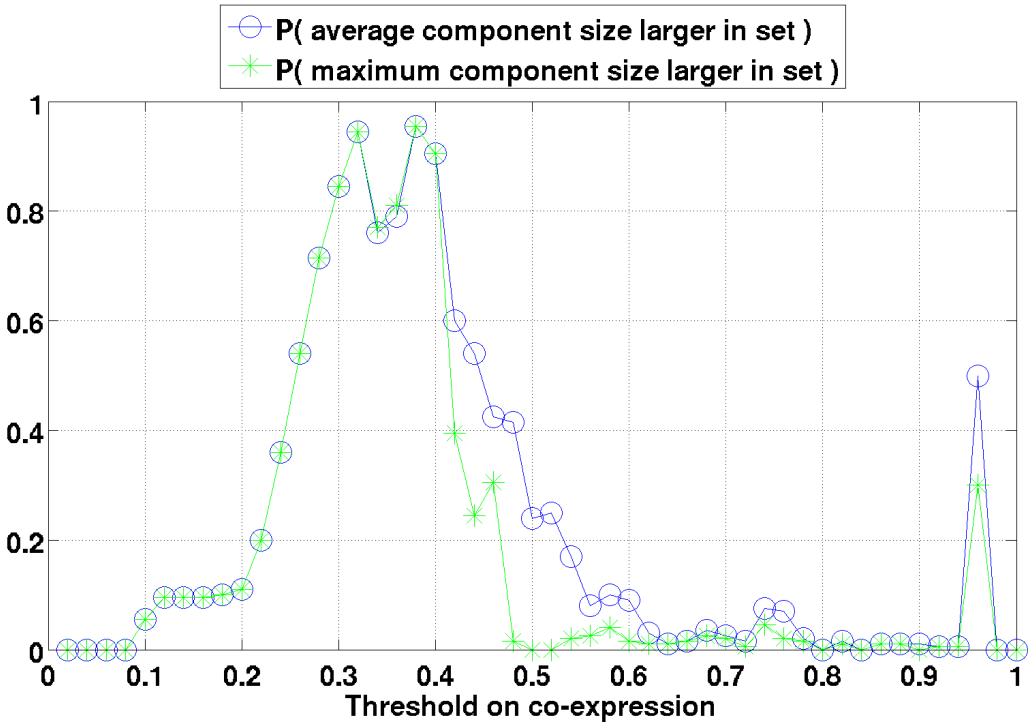


Figure 7: **Monte Carlo analysis of the graph underlying the co-expression matrix of 288 genes from the NicSNP database.** Estimated probabilities for the average and maximum size of connected components to be larger than in random sets of genes of the same size (the quantities defined in Equations 12 and 13 respectively).

However, the graph-based procedure returns special sets when the threshold on co-expression goes from 1 to 0, that may have exceptional neuroanatomical properties compared to sets of the same size, even if this does not affect the distribution of average and maximal size of connected components. For each of the connected components, the sum of expression energies can also be compared to the partition(s) of the brain given by the ARA, inducing fitting scores in each brain regions (see Supplementary Materials S4 for mathematical details). The probability for each connected component of thresholded co-expression networks to have a larger fitting score in a given brain region can be estimated. Imposing a threshold on this probability (99% for instance) returns sets of genes with exceptional anatomical properties. For the coarsest partition of the left hemisphere, a small set of 3 genes (*Rgs9, Drd2, Adora2a*) connected at a co-expression of 0.9, is in the 99th percentile of fitting scores in the striatum (see Figure 8 for a bar diagram of the estimate of P-values of fitting scores, and a maximal-intensity projection of the sum of the expression energies of these genes). Even though this set of genes is not exceptional in terms of its size at this value of the co-expression threshold, it has exceptional anatomical properties.

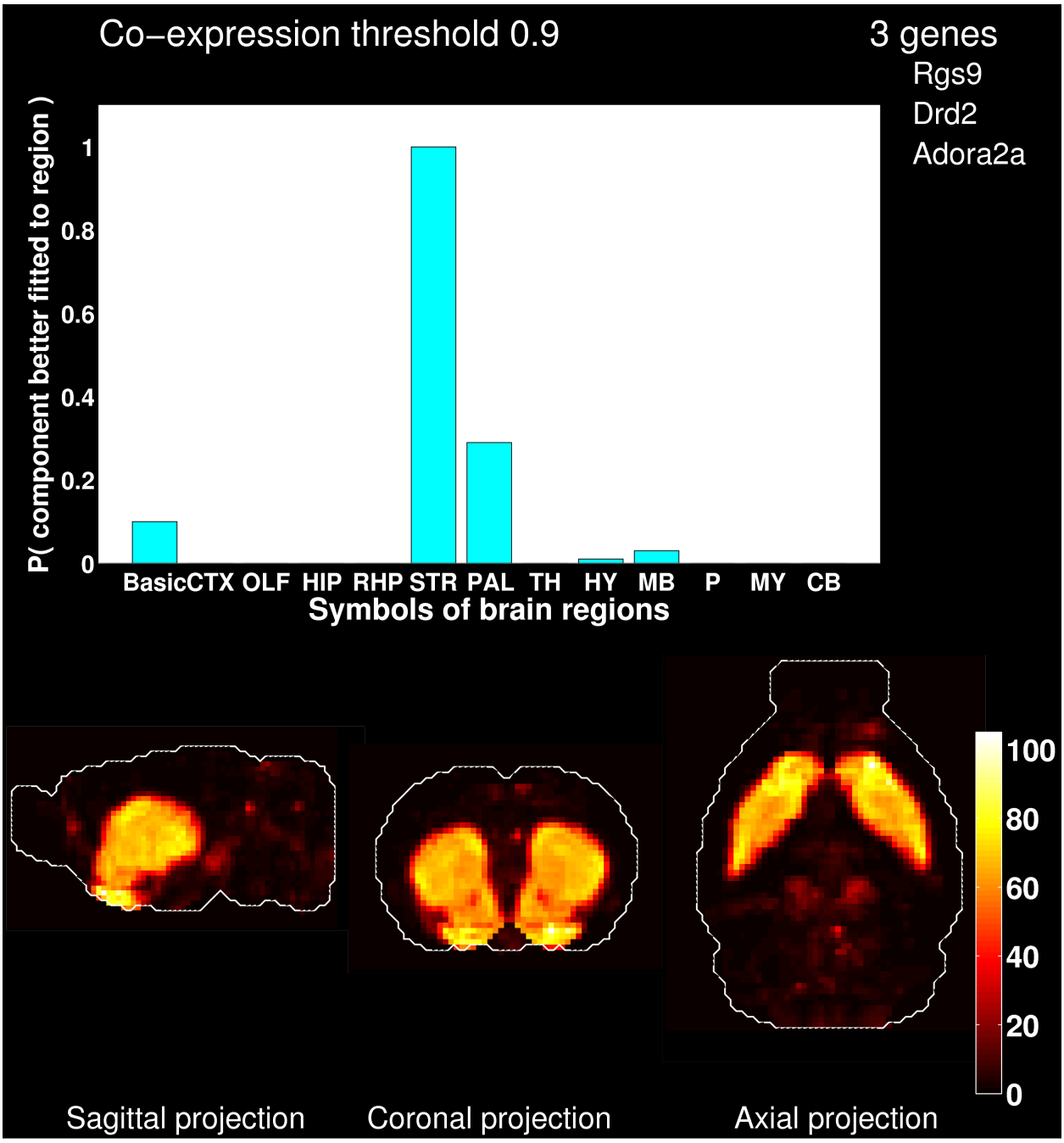


Figure 8: One of the connected components of the co-expression network, at co-expression threshold 0.9. It is better-fitted to the striatum (STR) than more than 99% of the set of three genes drawn from the coronal Allen Atlas of the adult mouse brain. The symbols for other brain regions read as follows: Basic = 'basic cell groups', CTX = cortex, OLF = olfactory areas, HIP = hippocampal region, RHP = retrohippocampal region, PAL = pallidum, TH = thalamus, HY = hypothalamus, MB = midbrain, P = pons, MY = medulla, CB = cerebellum.

4 Conclusion and outlook

The restriction of the first release `BrainGeneExpressionAnalysis` toolbox to the coronal atlas of the adult mouse brain corresponds to a restriction to genes for which brain-wide data are available. However, the sagittal atlas of the adult mouse brain contains more than $20,000$ genes, which are included in the Neuroblast and AGEA tools. The second release of BGEA will include these genes and restrict the Allen Reference Atlas to voxels where all the genes have ISH data (these voxels correspond to the left hemisphere of the brain). It would also be interesting to estimate the variability of the results under changes of probabilistic universe C^{atlas} (by substuting the sagittal atlas to the coronal atlas, and by choosing different image series to construct the data matrix).

Furthermore, the development of large-scale neuroscience is making comparable atlases available to the research community for other species (see [25] for the Allen Atlas of the human brain, and [26, 27] for ZEBrA, the Zebra Finch Expression Brain Atlas), and the development of computational resources for the analysis of large datasets can be adapted from the Allen Atlas of the adult mouse brain to other atlases, allowing insights into evolution and into the validity of animal models ([28]).

Moreover, the size of voxels in the Allen Brain Atlas is large in scale of brain cells, and each voxel may contain cells of different types. Several studies [33, 34, 35, 30, 31, 32, 29] have obtained cell-type-specific transcriptional profiles using microarray experiments. Comparison between ISH and microarray data is an ongoing challenge [36], and steps were taken in [37] to estimate the brain-wide density profiles of cell types by combining the Allen Atlas to the transcriptional profiles of cell types. This sheds light on the cellular origin of co-expression brain-wide co-expression patterns of genes. The corresponding Matlab code will be included in the second release of BGEA.

5 Acknowledgments

We thank Sharmila Banerjee-Basu, Idan Menashe, Eric C. Larsen, Hemant Bokil and Jason W. Bohland for discussions and collaboration. This research is supported by the NIH-NIDA Grant 1R21DA027644-01, *Computational analysis of co-expression networks in the mouse and human brain*.

References

[1] M. Bota, H.-W. Dong, L.W. Swanson, *From gene networks to brain networks*, Nature neuroscience (2003) **6** (8), 795–9.

[2] E.S. Lein, M. 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, *et al.*, *Genome-wide atlas of gene expression in the adult mouse brain.* Nature **445**, 168 176 (2007).

[3] S.M. Sunkin and J.G. Hohmann, *Insights from spatially mapped gene expression in the mouse brain*, Human Molecular Genetics, 2007, Vol. 16, Review Issue 2.

[4] L. Ng, M. Hawrylycz, D. Haynor, *Automated high-throughput registration for localizing 3D mouse brain gene expression using ITK*, Insight-Journal (2005).

[5] 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 and M. Hawrylycz, *Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain*, IEEE/ACM Trans. Comput. Biol. Bioinform. (2007), Jul-Sep **4(3)** 382–93.

[6] A.R. Jones, C.C. Overly and S.M. Sunkin, *The Allen Brain Atlas: 5 years and beyond*, Nature Reviews (Neuroscience), Volume **10** (November 2009), **1**.

[7] M. Hawrylycz, L. Ng, D. Page, J. Morris, C. Lau, S. Faber, V. Faber, S. Sunkin, V. Menon, E.S. Lein, A. Jones, *Multi-scale correlation structure of gene expression in the brain*, Neural Networks **24** (2011) 933–942.

[8] Computational analysis of user-defined sets of genes from the Allen Atlas of mouse and human brain can be conducted online at `addiction.brainarchitecture.org`

[9] P. Grange, J.W. Bohland, M. Hawrylycz and P.P. Mitra, *Brain Gene Expression Analysis: a MATLAB toolbox for the analysis of brain-wide gene-expression data*, downloadable at `www.brainarchitecture.org`.

[10] C.Y. Li, X. Mao, L. Wei (2008) Genes and (common) pathways underlying drug addiction. PLoS Comput Biol 4(1):e2. doi:10.1371/journal.pcbi.0040002. Data can be retrieved from `http://karg.cbi.pku.edu.cn/`

[11] S.F. Saccone, N.L. Saccone, G.E. Swan, P.A.F. Madden, A.M. Goate, J.P. Rice and L.J. Bierut, *Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence*, Bioinformatics (2008), **24**, 1805–1811.

[12] S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, *Basic local alignment search tool*, J. Mol. Biol. **215**, 403–410 (1990).

[13] L. Ng *et al.*, *NeuroBlast: a 3D spatial homology search tool for gene expression*, BMC Neuroscience 2007, **8**(Suppl 2):P11.

[14] 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, *An anatomic gene expression atlas of the adult mouse brain*, Nature Neuroscience **12**, 356 - 362 (2009).

[15] B. Zhang and S. Horvath, *A general framework for weighted gene co-expression network analysis*.

[16] P.K. Olszewski, J. Cederna F. Olsson, A.S. Levine and H.B. Schioth, *Analysis of the network of feeding neuroregulators using the Allen Brain Atlas*, Neurosci. Biobehav. Rev. **32**, 945–956 (2008).

[17] H.-W. Dong, *The Allen reference atlas: a digital brain atlas of the C57BL/6J male mouse*, Wiley, 2007.

[18] P. Grange and P.P. Mitra, *Computational neuroanatomy and gene expression: Optimal sets of marker genes for brain regions*, IEEE, in CISS 2012, 46th annual conference on Information Science and Systems (Princeton), `arXiv:1205.2721 [q-bio.QM]`.

[19] J.W. Bohland, H. Bokil, C.-K. Lee, L. Ng, C. Lau, C. Kuan, M. Hawrylycz, P.P. Mitra, *Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy*, Methods, Volume **50**, Issue 2, February 2010, Pages 105-112.

[20] C. Lau, L. Ng, C. Thompson, S. Pathak, L. Kuan, A. Jones and M. Hawrylycz, *Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain*, BMC Bioinformatics 2008, **8**:153.

[21] M. Hawrylycz, R.A. Baldock, A. Burger, T. Hashikawa, G.A. Johnson, M. Martone, L. Ng, C. Lau, S.D. Larsen, J. Nissanov, L. Puelles, S. Ruffins, F. Verbeek, I. Zaslavsky1, J. Boline, *Digital Atlasing and Standardization in the Mouse Brain*, PLoS Computational Biology **7** (2) (2011).

[22] The Allen Brain Atlas can be used online at `www.brain-map.org/`.

[23] The developmental atlas of the mouse brain is available from `http://developingmouse.brain-map.org/`

[24] I. Menashe, P. Grange, E.C. Larsen, S. Banerjee-Basu and P.P. Mitra, *Co-expression profiling of autism genes in the mouse brain*, SFN Abstracts 2012, and in preparation.

[25] M. Hawrylycz *et al.*, *An anatomically comprehensive atlas of the adult human brain transcriptome*, Nature **489**, 391399.

[26] W.C. Warren, D.F. Clayton, H. Ellegren, A.P. Arnold, L.W. Hillier, A. Kunstner, S. Searle, S. White, A.J. Vilella, S. Fairley *et al.* (2010), *The genome of a songbird*. Nature, 464, 757762.

[27] Data can be retrieved from the ZEBrA database. (Oregon Health and Science University, Portland, OR 97239; `http://www.zebrafinchatlas.org`).

[28] J.A. Miller, S. Horvath and D.H. Geschwind, *Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways*, Proc. Natl. Acad. Sci. U.S.A. (2010) bf107(28):12698-703.

[29] K. Sugino, C.M. Hempel, M.N. Miller, A.M. Hattox, P. Shapiro, C. Wu, Z.J. Huang, S.B. Nelson, *Molecular taxonomy of major neuronal classes in the adult mouse forebrain*, Nature Neuroscience **9**, 99-107 (2005).

[30] C.Y. Chung, H. Seo, K.C. Sonntag, A. Brooks, L. Lin, O. Isacson *Cell-type-specific gene expression of midbrain dopaminergic neurons reveals molecules involved in their vulnerability and protection*. Hum. Mol. Genet. (2005) **14**: 1709–1725.

[31] P. Arlotta, B.J. Molyneaux, J. Chen, J. Inoue, R. Kominami *et al.* (2005) *Neuronal subtype-specific genes that control corticospinal motor neuron development in vivo*, Neuron **45**: 207–221.

[32] M. Heiman, A. Schaefer, S. Gong, J.D. Peterson, M. Day, K.E. Ramsey, M. Surez-Farias, C. Schwarz, D.A. Stephan, D.J. Surmeier, P. Greengard, N. Heintz, (2008) *A translational profiling approach for the molecular characterization of CNS cell types*, Cell **135**: 738–748.

[33] M.J. Rossner, J. Hirrlinger, S.P. Wichert, C. Boehm, D. Newrzella, H. Hiemisch, G. Eisenhardt, C. Stuenkel, O. von Ahsen, K.A. Nave, *Global transcriptome analysis of genetically identified neurons in the adult cortex*, J. Neurosci. 2006 **26(39)** 9956-66.

[34] J.D. Cahoy, B. Emery, A. Kaushal, L.C. Foo, J.L. Zamanian, K.S. Christopherson, Y. Xing, J.L. Lubischer, P.A. Krieg, S.A. Krupenko, W.J. Thompson WJ, B.A. Barres, *A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function*, J. Neurosci. 2008 **28(1)** 264-78.

[35] J.P. Doyle, J.D. Dougherty, M. Heiman, E.F. Schmidt, T.R. Stevens, G. Ma, S. Bupp, P. Shrestha, R.D. Shah, M.L. Doughty, S. Gong, P. Greengard, N. Heintz, *Application of a translational profiling approach for the comparative analysis of CNS cell types*, Cell (2008) **135(4)** 749-62.

[36] C.K. Lee *et al.*, *Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data*, Genome Biol. **9**, R23 (2008).

[37] P. Grange, J. Bohland, H. Bokil, S. Nelson, B. Okaty, K. Sugino, L. Ng, M. Hawrylycz and P.P. Mitra, *A cell-type based model explaining co-expression patterns of genes in the brain*, `arXiv:1111.6217 [q-bio.QM]`.

[38] R. E. Tarjan, *Depth first search and linear graph algorithms*, SIAM Journal on Computing, **1(2)**:146-160, 1972.

6 Supplementary Materials

6.1 S1: Co-expression networks, graph properties

Consider a set of genes of size G_{set}, as in the ellipse on the left-hand-side of the flowchart 1. They correspond to indices $(g_1, \ldots, g_{G_{\text{set}}})$ in the columns of the voxel-by-gene matrix E of expression energies. We can construct the co-expression matrix by extracting the coefficients of the co-expression matrix of the atlas corresponding to these genes. Let us denote this matrix by C^{set}:

$$C^{\text{set}}(i,j) = C^{\text{set}}(g_i, g_j). \tag{6}$$

After applying the thresholding procedure, the co-expression matrix is mapped to a matrix C^{set}_ρ:

$$C^{\text{set}}_\rho(i,j) = C^{\text{set}}(i,j) \times \mathbf{1}(C^{\text{set}}(i,j) \geq \rho). \tag{7}$$

Then for every integer k between 1 and G_{set} we can count the number $N_\rho(k)$ of connected components of C_ρ that have exactly k genes in them (using Tarjan's algorithm [38], implemented as the function `graphconncomp.m` in Matlab). We can study the average size of connected components of thresholded co-expression networks and the size of the largest connected component:

$$\mathcal{A}(\rho) = \frac{\sum_{k=1}^{G_{\text{set}}} k N_\rho(k)}{\sum_{k=1}^{G} N_\rho(k)}, \tag{8}$$

$$\mathcal{M}(\rho) = \max\left\{k \in [1..G_{\text{set}}], N_\rho(k) > 0\right\}, \tag{9}$$

as a function of the threshold ρ. $\mathcal{A}(0)$ and $\mathcal{M}(0)$ both equal the size of the set of genes, as the whole set is connected before any thresholding procedure is applied. At large thresholds every single singe is disconnected from the other genes, as having co-expression equal to one is equivalent to having exactly the same expression across the whole brain. So at threshold 1 all the connected components have size one, and $\mathcal{A}(1) = \mathcal{M}(1) = 1$.

6.2 S2: Monte Carlo study of gene networks

We would like to compare the properties of the matrix C^{set} to the ones of C^{atlas}. In order to eliminate the sample-size bias, we are going to study some properties of the graph underlying C^{set}, and to compare them to the properties of the graphs underlying submatrices of C^{atlas} of the same size, G_{set}.

To explore the graph property of the gene network, we have to choose a discrete set of thresholds regularly spaced between 0 and 1, and to apply the procedure of Equation 7 usuing each of these thresholds. Call $\mathcal{N}$ the number of random sets of genes to be drawn. The computations can be described as follows in pseudocode:

1. `Choose a number of thresholds` T `to study.`
2. `Choose a number of draws` $\mathcal{N}$ `to be performed for each value`

of the threshold;
3. For each integer t between 1 and T:
3.a. consider the threshold $\rho_t = \frac{t}{T}$;
3.b. compute the connected components of the thresholded matrix $C^{\text{set}}_{\rho_t}$, as defined in Equation 7; call $\mathcal{M}^{\text{set}}(\rho_t)$ the size of the largest connected component, and $\mathcal{A}^{\text{r}}(\rho_t)$ the average size of connected components;
4. for each integer n between 1 and $\mathcal{N}$:
draw a random set of distinct indices of size G_{set} from [1..G],
extract the corresponding submatrix of C^{atlas};
call it C^n, and repeat step 3 after substituting C^n to C^{set};
call $\mathcal{M}^{\text{n}}(\rho_t)$ the size of the largest connected component of $C^{\text{n}}_{\rho_t}$, and $\mathcal{A}^{\text{n}}(\rho_t)$ the average size of connected components.

At each value ρ of the threshold, we therefore have:
- the size of the maximal connected component $\mathcal{M}^{\text{set}}(\rho)$,
- a distribution of $\mathcal{N}$ numbers, each of wchich is the size of the largest connected components of a random submatrix of the same size as the set of genes to study, thresholded at ρ.

When the number of random draws is sufficiently large, we can estimate the means of the average and maximum sizes of connected components:

$$\langle \mathcal{M}(\rho) \rangle = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \mathcal{M}^{\text{n}}(\rho), \tag{10}$$

$$\langle \mathcal{A}(\rho) \rangle = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \mathcal{A}^{\text{n}}(\rho), \tag{11}$$

We can study where in $\mathcal{M}^{\text{set}}(\rho)$ (resp. $\mathcal{A}^{\text{set}}(\rho)$) sits in the distribution and estimate the probabilities of $\mathcal{M}^{\text{set}}(\rho)$ and $\mathcal{A}^{\text{set}}(\rho)$ being larger than expected by chance:

$$P_{\text{larger}}(\mathcal{A}^{\text{set}}(\rho)) = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \mathbf{1}\left(\mathcal{A}^{\text{set}}(\rho) \geq \mathcal{A}^{\text{n}}(\rho)\right). \tag{12}$$

$$P_{\text{larger}}(\mathcal{M}^{\text{set}}(\rho)) = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \mathbf{1}\left(\mathcal{M}^{\text{set}}(\rho) \geq \mathcal{M}^{\text{n}}(\rho)\right), \tag{13}$$

6.3 S3: Cumulative distribution functions (CDFs)

Given the G_{set}-by-G_{set} co-expression matrix C^{set}, consider the coefficients above the diagonal (which are the meaningful quantities by construction) and arrange them into a vector C_{vec} with $N = G_{\text{set}}(G_{\text{set}} - 1)/2$ components: $C_{\text{vec}} = \{C_{gh}\}_{1 \leq g \leq G_{\text{set}}, h > g}$. The components of this vector are numbers between 0 and 1. For every number between 0 and 1, the cumulative

distribution function of C, denoted by $\mathrm{CDF}^{\mathrm{set}}$ is defined as the fraction of the components of C_{vec} that are smaller than this number:

$$\mathrm{CDF}^{\mathrm{set}} : [0,1] \to [0,1] \tag{14}$$

$$x \mapsto \frac{1}{N} \sum_{k=1}^{N} \delta_{C_{\mathrm{vec}}(k) \leq x}. \tag{15}$$

For any set of genes, $\mathrm{CDF}^{\mathrm{set}}$ is a growing function $\mathrm{CDF}^{\mathrm{special}}(0) = 0$ and $\mathrm{CDF}^{\mathrm{special}}(1) = 1$. For highly co-expressed genes, the growth of $\mathrm{CDF}^{\mathrm{special}}$ is concentrated at high values of the argument (in a situation where all the genes in the special set have the same brain-wide expression vector, all the entries of the co-expression matrix equal 1). To compare the function $\mathrm{CDF}^{\mathrm{special}}$ to what could be expected by chance, let us draw $\mathcal{N}$ random sets of G_{special} genes from the Atlas, compute their co-expression network by extracting the corresponding entries from the full co-expression matrix of the atlas (C^{atlas}). This induces a family of $\mathcal{N}$ growing functions $\mathrm{CDF}_i, 1 \leq i \leq R$ on the interval $[0,1]$:

$$\forall 1 \leq i \leq \mathcal{N}, \qquad \mathrm{CDF}_i : [0,1] \to [0,1]. \tag{16}$$

From this family of functions, we can estimate a mean cumulative distribution function $\langle \mathrm{CDF} \rangle$ of the co-expression of sets of G_{special} genes drawn from the Allen Atlas, by taking the mean of the values of CDF_i across the random draws:

$$\forall x \in [0,1], \quad \langle \mathrm{CDF} \rangle(x) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \mathrm{CDF}_i(x). \tag{17}$$

Standard deviations $\mathrm{CDF}^{\mathrm{dev}}$ of the distribution of CDFs are estimated as follows on the interval $[0,1]$:

$$\forall x \in [0,1], \quad \mathrm{CDF}^{\mathrm{dev}}(x) = \sqrt{\frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \left(\mathrm{CDF}_i(x) - \langle \mathrm{CDF} \rangle(x)\right)^2}. \tag{18}$$

6.4 S4: Comparison to classical neuroanatomy

Consider a system of annotation in the voxelized version of the ARA. Let Ω be the set of voxels in the annotation, let R be the total number of regions in the annotation, and let $\{\omega_1, \ldots, \omega_R\}$ be the regions in the annotation. For the sake of simplicity, the present paper focusses on the coarsest annotation, for which Ω is the left hemisphere, and $R = 13$.

For a set of G_{set} genes, labelled $\{g_1, \ldots, g_{G_{\text{set}}}\}$ in the atlas, the total brain-wide expression energy is

$$E^{\text{set}}(v, \{g_1, \ldots, g_{G_{\text{set}}}\}) = \sum_{k=1}^{G_{\text{set}}} E(v, g_k). \tag{19}$$

Using the same Monte Carlo procedure as in supplementary S2, we draw $\mathcal{N}$ sets of genes from the atlas, and compute the total gene-expression energy defined by Equation 20 for each of these sets. Hence for each random draw (labelled by an integer n in $[1..\mathcal{N}]$) one has a set of genes labelled $\{g_1^n, \ldots, g_{G_{\text{set}}}^n\}$, and the corresponding brain-wide sums of expression energies

$$E^{\text{rand,n}}(v, \{g_1^n, \ldots, g_{G_{\text{set}}}^n\}) = \sum_{k=1}^{G_{\text{set}}} E(v, g_k^n). \tag{20}$$

The fitting scores to each region in the ARA can be computed both for E^{set} and for the each of the random sets of G_{set} genes:

$$\phi_r^{\text{set}} := \phi_r(E^{\text{set}}), \tag{21}$$

$$\phi_r^{\text{rand,n}} := \phi_r(E^{\text{rand,n}}). \tag{22}$$

Hence, one can estimate the position of the fitting score ϕ_r^{set} in the distribution of fitting scores by evaluating the folllowing fraction:

$$P_r^{\text{set}} = \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} \mathbf{1}\left(\phi_r^{\text{set}} \geq \phi_r^{\text{rand,n}}\right) \tag{23}$$

which goes to the probability for ϕ_r^{set} being larger than expected by chance for a sum of G_{set} expression energies. A histogram of 23 is plotted on Figure 8, for the regions of the coarsest annotations of the left hemisphere, and for a set consisting of *Rgs2*, *Drd2* and *Adora2a*.