Marker Genes for Anatomical Regions in the Brain: Insights from the Allen Gene Expression Atlas

Pascal Grange, Partha P. Mitra

Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, New York 11724, USA

∗ **E-mail: pascal.grange@polytechnique.org**

Abstract

Quantitative criteria are proposed to identify genes (and sets of genes) whose expression marks a specific brain region (or a set of brain regions). Gene-expression energies, obtained for thousands of mouse genes by numerization of *in situ* hybridization images in the Allen Gene Expression Atlas, are used to test these methods in the mouse brain. Individual genes are ranked using integrals of their expression energies across brain regions. The ranking is generalized to sets of genes and the problem of optimal markers of a classical region receives a linear-algebraic solution. Moreover, the goodness of the fitting of the expression profile of a gene to the profile of a brain region is closely related to the co-expression of genes. The geometric interpretation of this fact leads to a quantitative criterion to detect markers of pairs of brain regions. Local properties of the gene-expression profiles are also used to detect genes that separate a given grain region from its environment.

Contents

1 Introduction

Neuroanatomy is experiencing a renaissance under the influence of molecular biology and computational methods. Brain regions can be delineated on stained sections of brain tissue. The set of boundaries between brain regions defined on sections can be registered in order to obtain a three-dimensional atlas. Conflicts exist between the various nomenclatures of brain regions. The present paper will consider brain regions defined by classical anatomy as in the Allen Reference Atlas [1]. Gene-expression energies are positive quantities defined at every point in the brain (or rather at every cubic voxel of side equal to the resolution, which is 200 microns in the present paper). With contemporary techniques of *in situ* hybridization, such data were produced by the Allen Institute for thousands of genes in the mouse brain [2, 3]. This makes the ISH data much higher-dimensional than classical neuroanatomy. Given an anatomical atlas, it is therefore natural to ask if the patterns formed by the expression energy of single genes and/or sets of genes can delineate and/or separate brain regions.

The structure of the paper is as follows. We will first formalize the notion of marker genes by defining quantitative criteria that allow to rank individual genes by computing scores. The *localization score* measures how much of the expression energy of a gene is contained in the region of interest. The *fitting score* measures how close the expression-energy profile is to the characteristic function of the region. The associated rankings of genes are computed. The genes ranked as the top few markers make sense optically, but there are conflicts between the two rankings. The two criteria are then used to rank sets of genes as markers of brain regions. The localization score gives rise to a generalized eigenvalue problem, and the solutions can have much higher localization scores than individual genes, but they are difficult to interpret because the sets of genes are very large and weighted by coefficients of alternating signs. The fitting score of sets of genes gives rise to sparse sets of markers. These two scores are easy to compute and to generalize, but they are both global in nature and they penalize genes that fit well the centermost part of a brain region, have low expression around the region, and high expression in remote parts of the brain. But such genes are interesting to detect, as they separate brain regions from their environment. A local fitting criterion is proposed, using the eikonal function in order to formalize the situation described above. Markers of pairs of regions are also investigated. They may be of special interest to evolution, especially for pairs of regions that are not equally well-identified in other species.

2 Methods and models

2.1 Gene expression energies and classical neuroanatomy

The gene expression energies we analyzed were drawn from the Allen Gene Expression Atlas [2]. The steps taken in an automatized pipeline to produce those data for each gene can be summarized as follows:
1. Colorimetric *in situ* hybridization;

2. Automatic processing of the resulting images. Find tissue area eliminating artifacts, look for cell-shaped objects of size $\simeq 10 - 30$ microns to minimize artefacts;
3. Aggregation of the raw pixel data into a grid.

The mouse brain is partitioned into cubic voxels of side 200 microns (the whole brain consists of $\simeq 50,000$ voxels). For every voxel v, the *expression energy* of the gene g is defined as a weighted sum of the greyscale-value intensities I evaluated at the pixels p intersecting the voxel:

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

where $M(p)$ is a Boolean mask worked out at step 2 with value 1 if the pixel is expressing and 0 if it is non-expressing.
Partitions of the brain (or of the left-hemisphere) of various degrees of coraseness in terms of classically-defined neuroanatomical regions were also published in the Allen Reference Atlas ([1], see also the white paper http://mouse.brain-map.org/documentation/index.html for the definition of expression energies).

The present analysis is focused on a subset of the genes for which sagittal and coronal data are available from the Allen data. We computed the correlation coefficients between sagittal and coronal data and selected the genes in the top-three quartiles of correlation (this makes for 3041 genes) for further analysis. Of course the quantitative methods can be tested against larger datasets or different reference atlases, but the genes we selected are already numerous enough to motivate the use of computational methods to detect markers.

2.2 Individual genes: localization scores

Given a brain region of interest, say the cerebral cortex (call it ω), define the localization score of a gene g as the fraction of the L^2 norm of its expression energy contained in the region:

$$\lambda_\omega(g) = \frac{\int_\omega E(v,g)^2 dv}{\int_\Omega E(v,g)^2 dv},$$

where Ω is the whole brain.

We computed the localization score of every gene in every region, for a given annotation of the brain. These numbers induce a ranking of genes as markers of each region of the brain, the better markers having higher localization scores. A perfect marker of ω according to this criterion would have a score of 1. Going from a region to another region, one has to be careful when comparing the values of the localization scores: as the volumes of the brain regions vary across the annotation, the localization score is biased by the sizes of the region. We need a reference in order to estimate how good a localization score is compared to what could be expected for a given brain region. We can use two references:
• **Uniform reference.** Consider an *indifferent* marker that would be expressed uniformly across the brain. Its localization score in region ω is simply the relative volume of the region:

$$\lambda_\omega^{\text{unif}} = \frac{\int_\omega dv}{\int_\Omega dv} = \frac{\text{Vol } \omega}{\text{Vol } \Omega}.$$

A gene is a better marker of ω than expected from a uniform expression if its score $\lambda_\omega(g)$ is larger than $\lambda_\omega^{\text{unif}}$.

• **Average (data-driven) reference.** A more realistic reference is given by the gene-expression profile averaged across all the genes:

$$E^{\text{average}}(v) = \frac{1}{G} \sum_{g=1}^{G} E(v, g).$$

The corresponding localization score in a given region ω is:

$$\lambda_{\omega}^{\text{average}} = \frac{\int_{\omega} E^{\text{average}}(v)^2 dv}{\int_{\Omega} E^{\text{average}}(v)^2 dv}.$$

A gene is a better marker of ω than expected from an average expression if its score $\lambda_{\omega}(g)$ is larger than $\lambda_{\omega}^{\text{average}}$.

The values of these references, and the rankings of genes for ω taken from the list of 12 largest regions in the left hemisphere, are presented in the results section and in appendices.

2.3 Individual genes: fitting scores

The criterion defined above does not take into account the repartition of the signal inside the region of interest: the localization score for a given gene in region ω is invariant under a transformation that moves the whole expression energy into a single voxel within ω, leaving all other voxels in ω with a zero signal. It is therefore desirable to have another ranking of genes as markers, that compares the gene-expression profiles to characteristic functions of brain regions.

This criterion compares the shape of the expression energy profile of a gene and the shape of the region of interest. The fitting score $\phi_{\omega}(g)$ of gene g in region ω is defined as follows:

$$\phi_{\omega}(g) = 1 - \frac{1}{2} \int_{\omega} \left(\tilde{E}_{\text{norm}}(v, g)^2 - \chi_{\omega}(v) \right)^2 dv = \int_{\omega} \tilde{E}_{\text{norm}}(v, g) \chi_{\omega}(v) dv.$$

where χ_{ω} is the characteristic function of ω normalized in the L^2 sense, and $\tilde{E}_{\text{norm}}$ is a normalized version of the expression energy (the columns of the matrix $\tilde{E}_{\text{norm}}$ are the columns of the matrix E, normalized in the L^2 sense):

$$\tilde{E}_{\text{norm}}(v, g) = \frac{E(v, g)}{\sqrt{\int_{\Omega} E(v, g)^2 dv}}, \quad \chi_{\omega} = \frac{\mathbf{1}_{\omega}}{\sqrt{\text{Vol}(\omega)}}.$$

A perfect marker of the region ω would be a gene with fitting score equal to 1. The geometric interpretation of this coefficient is as the cosine of the angle between the unitary vectors $\tilde{E}_g$ and χ_{ω} in voxel space. A perfect marker of region ω is a gene whose expression profile is colinear with the characteristic function of region ω. Again, this error function can be evaluated for all the genes in the dataset, and induces a ranking of genes (see the results section and appendices).

2.4 Sets of genes: optimal localization scores as a generalized eigenvalue problem

Consider the problem of optimizing the localization score of a set of genes, whose collective expression energy is taken to be a linear combination of the expression energies in our dataset:

$$E_{\alpha}(v) := \sum_{g=1}^{G} \alpha_g E(v, g).$$

where $G = 3041$ is the number of genes in our dataset. The previous analysis corresponded to vectors α with only one non-zero coordinate.

The localization score in the brain region ω of a set of genes is naturally written as

$$\lambda_\omega(\alpha) = \frac{\int_\omega \left(\sum_g \alpha_g E(v,g)\right)^2 dv}{\int_\Omega \left(\sum_g \alpha_g E(v,g)\right)^2 dv} = \frac{\alpha^t J^\omega \alpha}{\alpha^t J^\Omega \alpha},$$

where the quadratic forms J^ω and J^Ω have coefficients given by scalar products of gene expression profiles across ω and the whole brain:

$$J^\omega_{g,h} = \int_\omega E(v,g)E(v,h)dv, \quad J^\Omega_{g,h} = \int_\Omega E(v,g)E(v,h)dv.$$

We can fix an overall dilation invariance by fixing the value of the quadratic form in the denominator, and maximizing the localization factor boils down to a maximization of one quadratic form under a quadratic constraint.

$$\max_{\alpha \in \mathbf{R}^G} \lambda_\omega(\alpha) = \max_{\alpha \in \mathbf{R}^G, \alpha^t J^\Omega \alpha = 1} \alpha^t J^\omega \alpha.$$

Introducing the Lagrange multiplier χ associated to the constraint, we are led to maximizing the following quantity under α:

$$L_{\omega,\sigma}(\alpha) = \alpha^t J^\omega \alpha - \sigma(\alpha^t J^\Omega \alpha - 1).$$

The stationarity condition reads as a generalized eigenvalue problem,

$$J^\omega \alpha = \sigma J^\Omega \alpha,$$

and the Lagrange multiplier is the largest generalized eigenvalue. Maximizing the generalized localization score is therefore equivalent to finding the largest generalized eigenvalue corresponding to the quadratic forms J^ω and J^Ω.

Of course the alternating signs of the coefficients make these sets difficult to interpret. But these algebraic sums provide absolute bests that one could not beat by taking combinations of genes with positive coefficients. The negative coefficients allow to offset the contribution of some genes outside the region of interest.

2.5 Sets of genes: optimized fitting scores for sparse sets of genes

As the optimal set of of markers is very hard to interpret due to alternating signs of components, we can take advantage of the simple quadratic structure of the error function used to compute fitting scores in order to obtain sets of markers with positive coefficients. Optimization of a quadratic form under positivity constraint is all we need to compute the optimal sets of markers. Let us write down the fitting error function for a set of genes and expand it in powers of the coefficients:

$$\text{ErrFit}_\omega(c) = \int_\Omega \left(\sum_g c_g E_g(v) - \chi_\omega(v)\right)^2 dv \tag{1}$$

$$= \sum_{g,h} c_g c_h J_{gh} - 2\sum_g c_g f_g + 1, \tag{2}$$

where ω and Ω respectively denote the brain region of interest and the whole brain. The problem of finding thje best-fitting set of genes therefore boils down to the following quadratic programming problem under positivity constraints:

$$c^{\mathrm{opt}} = \mathrm{argmin}_{c \in \mathbf{R}_+^G} \mathrm{ErrFit}_\omega(c) = \mathrm{argmin}_{c \in \mathbf{R}_+^G} \left(\frac{1}{2} c^t J c - f^t c \right),$$

with the following notations for the quadratic form J and the vector f:

$$J_{gh} = \int_\Omega E(v,g)E(v,h)dv,$$

$$f_g = \int_\Omega E(v,g)\chi_\omega(v)dv.$$

The set of genes with strictly positive coefficients corresponds to the set of inactive constraints. It happens to be much sparser than the vector encoding the generalized eigenvector for the cortex localization problem (see figure (11)).

However, lots of secondary minima are guaranteed to exist when larger and larger sets of genes are taken into account, and coefficients c of very different norms can be hard to use to construct markers out of digitized data, as the absolute intensity of genes is quite heterogeneous, and a gene with low absolute intensity can happen to be weighted by a large coefficient, thus amplifying noise rather than contributing to a realistic marker.
But we can take advantage of the expression of the fitting score in terms of the scalar product between the gene expression profile and the characteristic function of the brain region:

$$\mathrm{ErrFit}_\omega(c) = 2 \left(1 - \int_\Omega \sum_g c_g E_g(v) \chi_\omega(v) dv \right),$$

$$c^{\mathrm{opt}} = \mathrm{argmax}_{c \in \{0,1\}^G} \int_\Omega \sum_g c_g E_g(v) \chi_\omega(v) dv.$$

Another approach to the optimization problem consists in looking for sets of genes such that the co-expression between the sum of the expression energies of those genes and the characteristic function is larger than that of any individual genes. This can happen, for instance if the characteristic function in voxel space equals the sum of two genes, whose expression energies are two independent vectors in voxel space: the cosine of the angle between any of these two vectors with the characteristic function is strictly smaller than one, but the cosine of the angle between the sum and the characteristic function equals one.

This is a finite problem, even though the number of subsets is extremely large. We impose a maximum G_{max} on the number of genes we want to accept, and adopt a bootstrapping approach: we repeatedly draw random subsets of size G_{max} from our set of genes, and keep the subsets that beat the record fitting score (this record is initialized at the highest fitting score for an individual gene).

2.6   Separation properties

The methods described so far are global in nature in the sense that the error functions involve sums of expression energies over the whole brain. This corresponds to evaluating how a brain region is singled out with respect to the rest of the brain. No attention is paid to the position of the voxels that contribute to the error functions: a voxel with high expression in the cerebellum will penalize a gene as a merker of the striatum, no more but no less than if it was in the ventral pallidum. However it may be interesting

to detect genes that separate some brain region from its environment, without necessarily highlighting these brain regions in an exclusive way. The description of such a situation implies a more *local* error function.

However, when looking at the expression profile of a gene in the neighbourhood of a particular brain regions, one can sometimes notice that the region is well-separated from the rest of the brain, because the expression is high in voxels close to the center of the region, and locally declines around the center. At large distances from the center, the details of the gene-expression profile matter much less, as long as this pattern of decreasing expression from center to boundary is detected. Such genes have good *separation* properties.

The separation property we described above corresponds to the situation where the gene-expression pattern looks like a plateau around the center of the region ω, and gradually fades away when the boundary of the region is crossed. Of course the notion of center of a brain region needs to be defined more precisely. So does the notion of distance to a brain region. The eikonal distance to the boundary of the region is a geometric quantity that is well adapted to this problem, as it measures the minimal distance traveled by light emitted from the boundary of the region [5]. In order to control how far from the center of a region a voxel is, one can therefore solve the eikonal equation with boundary conditions on the boundary of the region:

$$|\nabla h_\omega| = 1,$$
$$h_\omega|_{\partial\omega} = 0.$$

The eikonal distance has been used used to place injections in the brain in a way that preserves the boundaries of regions defined by the Allen Atlas [6,7,9]. It is also a useful tool to evaluate the misalignment of skulls and skull variability in stereotactic protocols [8]. The equation is solved using level-set methods [4].

We define a model function ξ_ω that detects the most central part of the region ω, using the eikonal function as a measure of centrality. The function is positive It is a plateau in the central part of ω, and fades away across voxels that are more peripheric to ω. More specifically, let us define the *eikonal radius* ρ_ω of the region ω as the maximum value of the eikonal function inside the region:

$$\rho_\omega := \max_{v \in \omega} h_\omega(v).$$

Let us first apply a mask to the eikonal function, with negative signs outside the region and positive signs inside:

$$h_\omega^{\text{signed}}(v) := h_\omega(v) \times \left(\mathbf{1}(v \in \omega) - \mathbf{1}(v \notin \omega) \right).$$

Our model function ξ_ω equals one around the center of the region, where h_ω^{signed} is positive and larger than a specified fraction of the eikonal radius, given by a certain fraction ι of the eikonal radius. It equals zero where h_ω^{signed} is negative and larger in absolute value than another specified fraction of the eikonal radius. The values are interpolated between these two regions according to the values of the eikonal function.

Having defined this local characteristic function ξ_ω around the brain region of interest, one can treat the support of ξ_ω as we treated the whole brain Ω in the previous computations, and adapt the various quantitative criteria by making the following substitution:

$$\text{Global} \longleftrightarrow \text{Local},$$

$$\Omega = \text{whole brain} \longleftrightarrow \Omega = \text{Supp}\xi_\omega,$$

$$\chi_\omega \longleftrightarrow \xi_\omega,$$

$$E(v,g) \longleftrightarrow E(v,g)\mathbf{1}(v \in \mathrm{Supp}\xi_\omega).$$

This substitution expresses the fact that the expression energy outside the support of the local characteristic function of ω can be very singular of very intense without affecting the separation properties.

2.7 Co-markers for pairs of brain regions

The various quantitative criteria can be repeated for reunions of brain regions. For instance one can look for a marker of the two brain regions ω_A and ω_B. Ideally one would like the expression profile of a marker gene to look like the sum of the two characteristic functions of regions A and B, normalized in the L^2 sense[1]. But it may be interesting to allow the two characteristic functions to be weighted by coefficients, in order to detect genes whose expression looks like two bumps, one centered around A, one centered around B, with possibly different intensities.

Consider the two characteristic functions χ_A and χ_B, normalized in the L^2 sense, and a linear combination thereof with positive coefficients, normalized in the same way. The coefficients of the linear coimbination can be interpreted geometrically in terms of a single parameter, which is an angle between 0 and $\pi/2$. Let us denote it by θ:

$$\int_\Omega \chi_A(v)^2 dv = 1, \ \mathrm{Supp}\chi_A = \omega_A,$$

$$\int_\Omega \chi_B(v)^2 dv = 1, \ \mathrm{Supp}\chi_B = \omega_B,$$

$$\chi := \alpha\chi_A + \beta\chi_B, \ \alpha \geq 0, \ \beta \geq 0, \ \int_\Omega \chi^2 = \alpha^2 + \beta^2 = 1,$$

$$\alpha = \cos\theta, \ \beta = \sin\theta, \ 0 \leq \theta \leq \frac{\pi}{2},$$

where we have used the fact that the functions χ_A and χ_B are orthogonal, because they have disjoint supporty. Geometrically, the function χ we are trying to fit is the sum of two unit orthogonal vectors in voxel space, that sits on the intersection of the unit circle and the first quadrant in the two-plane spanned by these two vectors. We can compute the fitting error for each gene at fixed angle θ, but it can be optimized wrt the angle:

$$\mathrm{ErrFit}_{A,B}(g, \theta) = \int_\Omega \left(E(v,g) - (\cos\theta\chi_A(v) + \sin\theta\chi_B(v)) \right)^2 dv \tag{3}$$

$$= 2 \left(1 - \cos\theta \int_\Omega E(v,g)\chi_A(v)dv - \sin\theta \int_\Omega E(v,g)\chi_B(v)dv \right). \tag{4}$$

This optimization step corresponds to the fact that the angle between a fixed vector in voxel space (corresponding to a gene), can have a lower angle with a two-plane than with any of the vectors of an orthonormal basis of the two-plane. The optimal angle θ^* is given by the equation:

$$\frac{\partial}{\partial\theta}\mathrm{ErrFit}_{A,B}(g, \theta^*) = 0,$$

$$\text{i.e. } -\sin\theta^* \int_\Omega E(v,g)\chi_A(v)dv + \cos\theta^* \int_\Omega E(v,g)\chi_B(v)dv = 0,$$

---

[1]One can as well look for genes that separate regions A and B from their respective environments, by considering the local characteristice functions worked out using the eikonal functions with boundary conditions at the boundaries of A and B, rather than the characteristic functions.

$$\text{i.e. } \theta^* = \arctan\left(\frac{\int_\Omega E(v,g)\chi_B(v)}{\int_\Omega E(v,g)\chi_A(v)}\right).$$

The value of the error function at the optimal angle (meaning the linear combination of the two characteristic function with positive coefficients that is best fit by gene g) is then evaluated in terms of the scalar product between the gene expression and the two characteristic function χ_A and χ_B:

$$\cos\theta^* = \frac{\int_\Omega E(v,g)\chi_A(v)dv}{\sqrt{\left(\int_\Omega E(v,g)\chi_A(v)dv\right)^2 + \left(\int_\Omega E(v,g)\chi_B(v)dv\right)^2}},$$

$$\sin\theta^* = \frac{\int_\Omega E(v,g)\chi_B(v)dv}{\sqrt{\left(\int_\Omega E(v,g)\chi_A(v)dv\right)^2 + \left(\int_\Omega E(v,g)\chi_B(v)dv\right)^2}},$$

so that

$$\phi_{A,B}(g,\theta) = 1 - \frac{1}{2}\text{ErrFit}_{A,B}(g,\theta^*) = \sqrt{\left(\int_\Omega E(v,g)\chi_A(v)dv\right)^2 + \left(\int_\Omega E(v,g)\chi_B(v)dv\right)^2}.$$

This score is comprised between 0 and 1, as the fitting score evaluated for the fitting of a single region by a gene. So, given a non-hierarchical atlas $\mathcal{A}$, one can find better fittings for pairs of regions in the atlas than for single regions.

These scores can be computed for all pairs of regions in a given non-hierarchical atlas. Of course there is no reason why the top co-marker of regions A and B should be especially more impressive than the best marker of A or B. By the look of the expressions of the coefficients $\cos\theta^*$ and $\sin\theta^*$, it is clear that in the case where $\int_\Omega E(v,g)\chi_B(v)dv$ is much smaller than $\int_\Omega E(v,g)\chi_A(v)dv$, its score as a co-marker of regions A and B is slightly larger than its score as a marker of A, but most of the expression will of course be in the A. The value of $\tan\theta^*$ controls the balance between the expression energies in the two regions. The closer it is to 1, the better co-marker we have. Asking for a value of exactly one would amount to trying to fit the sum of the characteristic functions of regions A and B without Once the genes have been ranked as co-markers of A and B, one can filter out the genes for which $\tan\theta^*$ is out of a tolerance zone around 1. This is a balance constraint. The genes at the top of the ranking that do not satisfy it are rather markers of the region (A or B) that has the highest coefficient. The genes that satisfy it are the co-markers we are after, and they are penalized by the localization and fitting criteria, both for region A and for region B:

$$\text{Balance constraint}_\tau \equiv |\tan\theta^* - 1| \leq \tau.$$

3 Results and discussion

3.1 Rankings of genes

A plot of the sorted localization scores of individual genes is shown on figure 1 for the cerebral cortex, as well as a table of the best few marker genes. Tables for all the other brain regions in the coarsest Allen Reference Atlas are included in an appendix. The maximum-intensity projections of the best marker of the cerebral cortex in the left hemisphere, and compared to those of the characteristic function of the cerebral cortex are shown on figure 2. The sorted fitting scores and the list of top genes for the cerebral cortex are shown on figure (3). A coronal section of the ISH data for Satb2 is shown on figure (5). The cerebral cortex indeed appears strikingly on the section. However, Satb2 is not the absolute best gene according to the localization criterion, which is Pak7, but it is still among the best 10 genes by localization scores. A coronal of the ISH data for Pak7 is shown on figure (6). Maximal-intensity projections of the registered 3D data on figures (2) and (4) show indeed that the expression energy of Satb2 is more evenly

	Cerebral cortex
Pak7	0.97
Myl4	0.97
Sytl2	0.97
Tnnc1	0.96
Gtdc1	0.96
2310026E23Rik	0.95
Dact2	0.95
Satb2	0.94
Ier5	0.93
D830014K04Rik*	0.93
Baalc	0.92
1810023C24Rik*	0.92
Ddit4l	0.92
LOC433228	0.91
Rorb	0.91
Gnb4	0.9
A930001M12Rik	0.89
LOC433698	0.89
E430002G05Rik	0.89
TC1460681	0.89
Tox	0.89
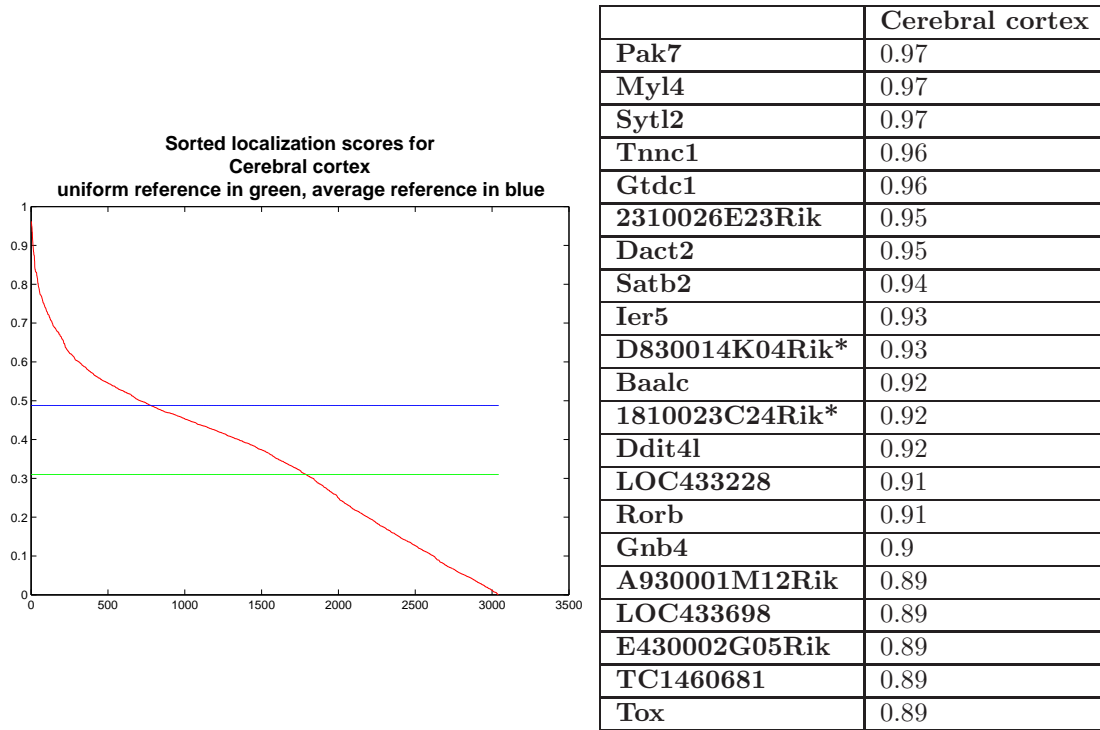
Figure 1. Plot of sorted localization scores in the cerebral cortex (left), with the list of the first few genes with highest localization scores in the cerebral cortex (right).
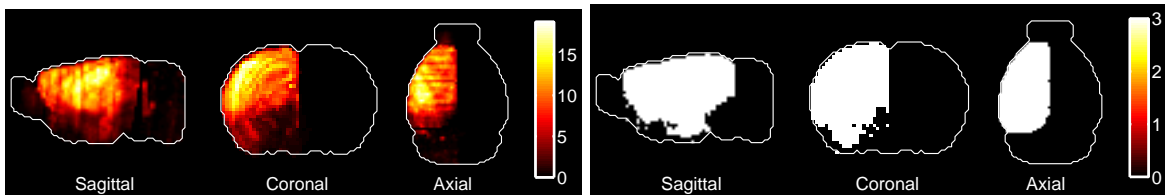
Figure 2. Heat map of the maximum-intensity projection of Pak7 (left), the best marker of the cortex in the sense of localization scores, compared to the characterictic function of the cerebral cortex (right).

Sorted fitting scores for Cerebral cortex

	Cerebral cortex
Satb2	0.89
Kcnh7	0.85
Ephb6	0.84
3110035E14Rik	0.83
Homer1	0.83
Fhl2	0.83
Pak7	0.83
Klf10	0.83
Dusp3	0.83
Cckbr	0.83
1110008P14Rik	0.82
Tbr1	0.82
Igsf9b	0.82
Stx1a	0.82
A230097P14Rik*	0.81
D430041D05Rik	0.81
Mal2	0.81
Igfbp6	0.81
Nlk	0.81
Arc	0.81
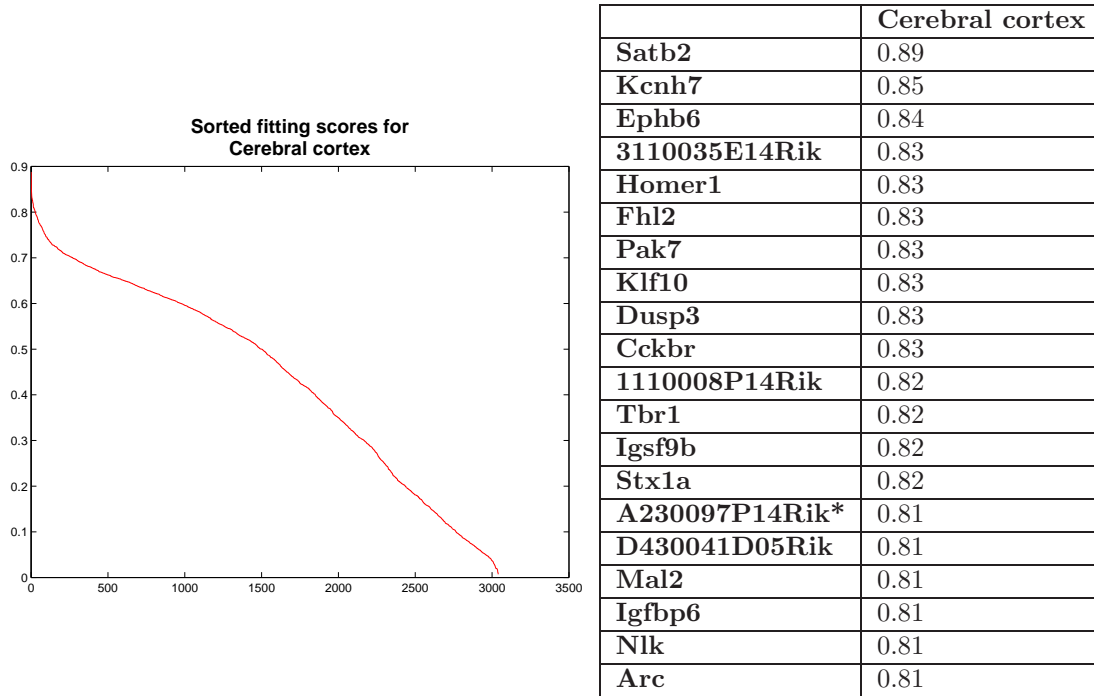
Figure 3. Plot of sorted fitting scores in the cerebral cortex (left), with the list of the first few genes with highest fitting scores in the cerebral cortex (right).
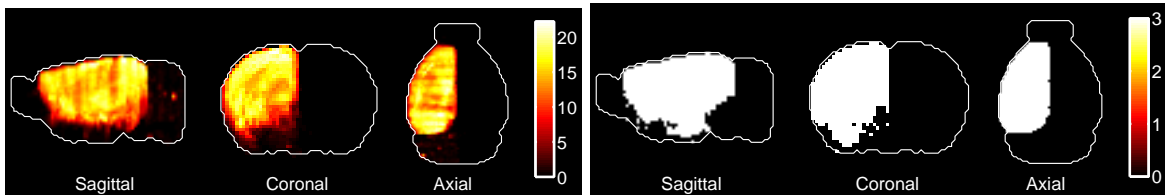
Figure 4. Heat map of the maximum-intensity projection of Satb2 (left), best marker of the cortex in the sense of fitting scores, compared to the characterictic function of the cerebral cortex (right).
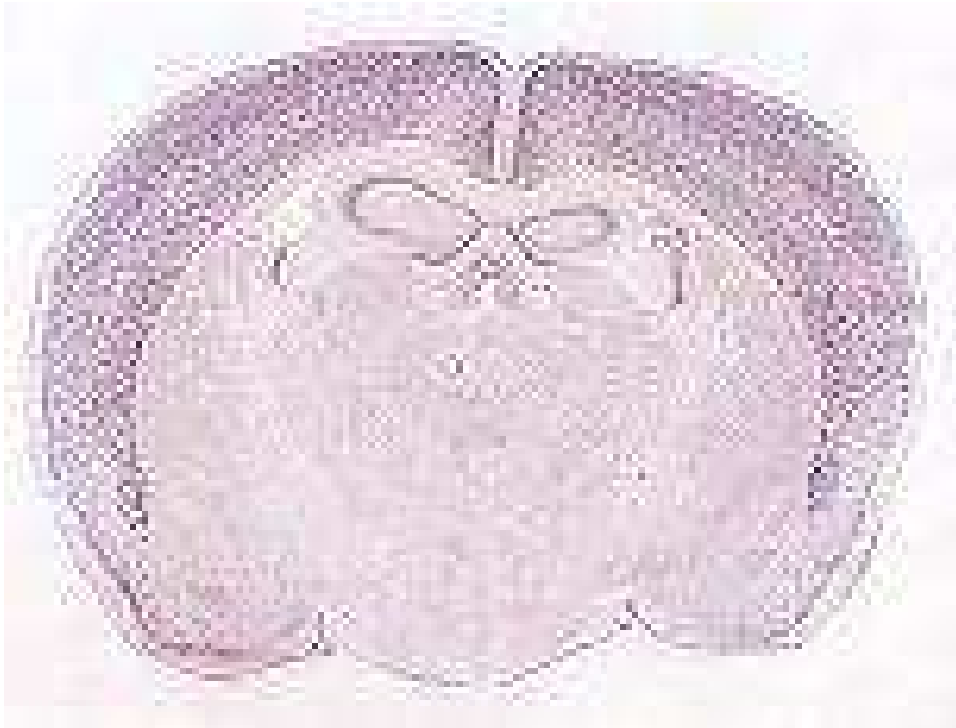
Figure 5. A coronal section of the ISH of Satb2. Satb2 has the highest localization score in the cerebral cortex. The concentration of blue precipitate in the region is manifest.

distributed across cortex than the one on Pak7, which makes Satb2 closer to the characteristic function of the cerebral cortex.

3.2   Rankings of regions

For each region ω, we can count the number of markers as the number of genes whose localization score in the region is larger than the fraction of the brain occupied by the region, as defined by the uniform of average references. The results are illustrated on figure (7).

Moreover, the two reference values defined above using either the volumes of the brain regions or the average of the expression of all genes in the dataset show important distorsions, as can be seen from the table (18). In particular, the gene-expression profiles are biased towards the cerebral cortex and the hippocampal region, as the value of $\lambda_{\mathrm{cerebralcortex}}^{\mathrm{average}}$ is higher than 47 percent, while the value of $\lambda_{\mathrm{cerebralcortex}}^{\mathrm{uniform}}$ is lower than 30 percent. This distorsion is manifest of the maximal-intensity projection of the sum of all expression-energy profiles across the dataset, shown on figure (8).

The highest scores across all genes and brain regions are for the cerebellum, both by localization and by fitting. This comparison across all regions and genes makes more sense for fitting scores than for localization scores, as the fitting scores is not biased by the volume of the region. Both genes make good sense optically as markers of the cerebellum. The best-localized gene is Gabra6, at 98 percent localization score (see figure (9)). It is the 73rd best-fitted gene to the cerebellum. The best-fitted gene is 3110001A13Rik, at 89 percent fitting score (see figure (10)). It is the 2nd best-fitted gene to the
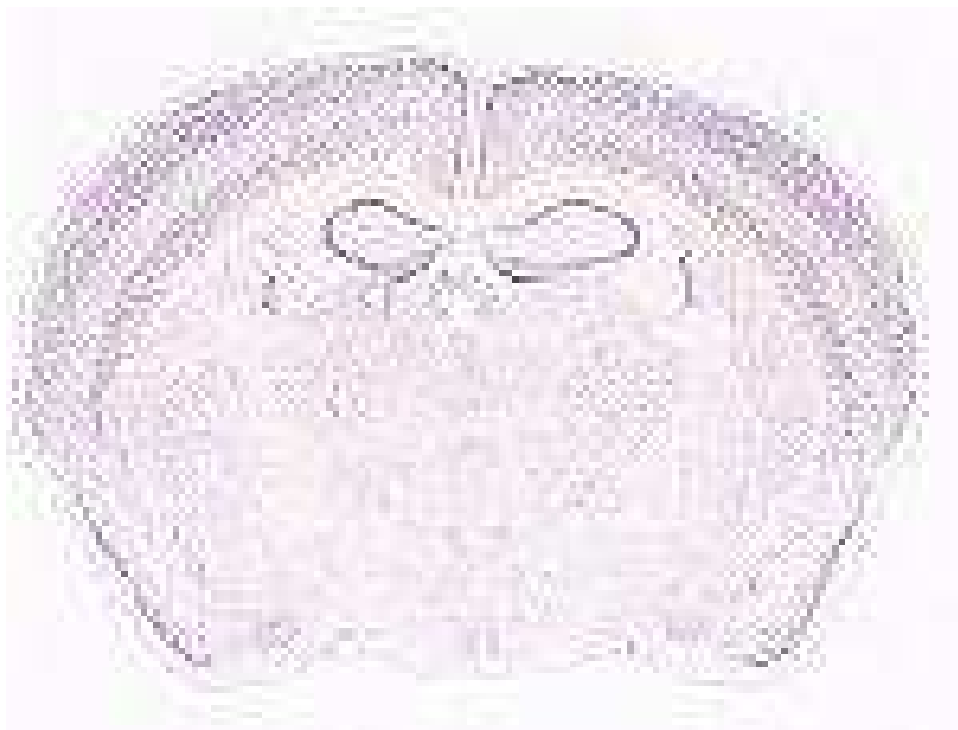
Figure 6. A coronal section of the ISH of Pak7. Pak7 has the highest fitting score in the cerebral cortex. The concentration of blue precipitate in the region is manifest.

Region name	Nb of genes above average ref.	Nb of genes above volume ref.
Basic cell groups and regions	3041	3041
Medulla	1418	920
Pons	1315	600
Cerebellum	1230	675
Olfactory areas	1210	1241
Thalamus	1144	614
Midbrain	1126	381
Hippocampal region	1055	1544
Pallidum	1038	276
Hypothalamus	1007	456
Retrohippocampal region	998	1626
Striatum	855	479
Cerebral cortex	782	1791

Figure 7. Ranking of regions by decreasing number of genes with localization score above average reference.
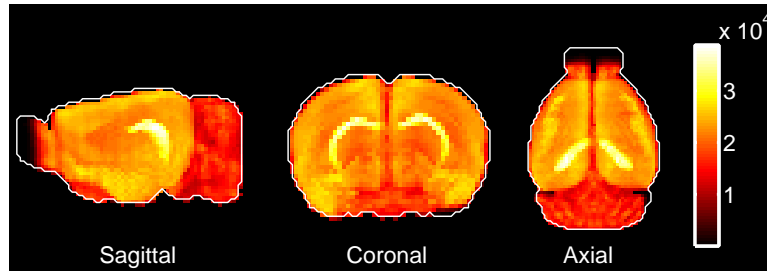
Figure 8. Heat map of the maximal-intensity projections of the sum of the expression energies across all genes in the dataset. The cerebral cortex and the hippocampal region are clearly visible.
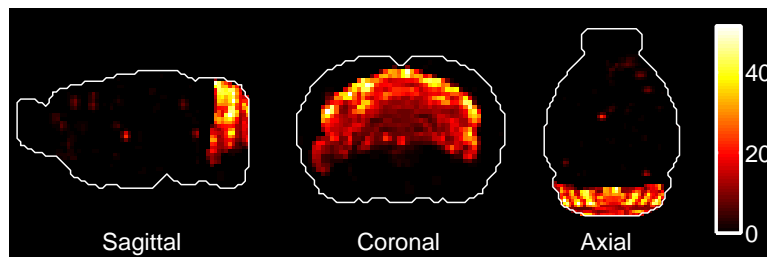
Figure 9. Absolute best localization across all genes and regions. Heat map of the maximal-intensity projections of the expression energy of Gabra6.

cerebellum. The distorsion is much larger for Gabra6 because its expression profile shows inhomogeneities inside the cerebellum. For 3110001A13Rik, the localization is optically extremely good.

3.3   Sets of genes

Sorting the coefficients of the generalized eigenvector associated to the largest generalized eigenvalue for the cerebral cortex yields a profile (illustrated in figure (11)), which has coefficients of both signs (as above ω is chosen to be the cerebral cortex), with a localization score of 0.979 (higher than the optimum for single genes, as it should be).

 It is interesting to note that in some regions the gene-expression profile correponding to the generalized eigenvector looks much more coherent as a marker than the best-localized gene. More examples can be found in figure (22). Taking combinations of genes therefore enables one to get closer to the characteirstic
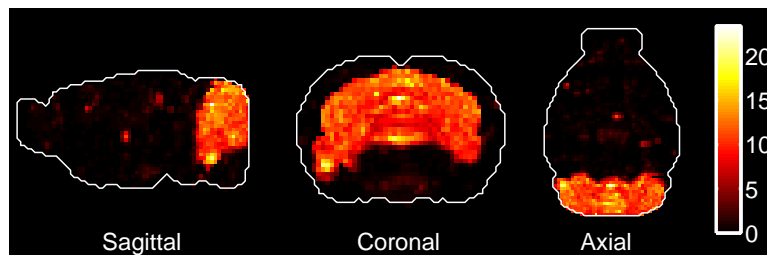
Figure 10. Absolute best fitting across all genes and regions. Heat map of the maximal-intensity projections of the expression energy of 3110001A13Rik.
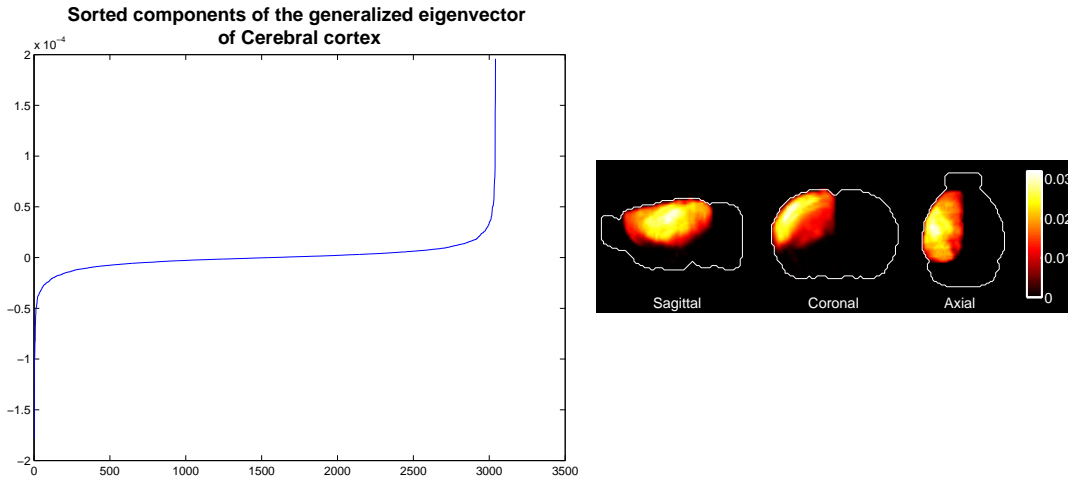
Figure 11. Optimal set of marker genes in the sense of generalized localization scores. Sorted components (left), heat map of the maximal-intensity projections of the expression energies of the genes weighted by these coefficients (right).

function of the regions. It is therefore tempting to go back to the fitting scores and to adapt it to sets of genes, with positivity constraints that would produce sparser sets of genes. Midbrain is a brain region for which the generalized eigenvector gives a much better visual impression of the whole structure than the best-localized gene. The generalized fitting score, in the case of midbrain, gets rid of much of the signal outside the region but does not achieve much homogeneity inside. Generally speaking, and not too surprisingly, the sparse sets of genes are much less impressive markers than the generalized eigenvectors, as they rely on much fewer degrees of freedom for optimization. On the other hand, the best individual marker returned by the global fitting criterion is much more convincing than the one returned by the localization score.

3.4 Good separators

For every region in the coarsest annotation of the left hemisphere, we ran an algorithm with a range of values of the internal an external paramater for the model function and for the local mask. As the boundary of the striatum does not have too much overlap with the boundary of the brain, it is easier to visualize than the cerebral cortex in a maximum-intensity projection and we chose it for illustration (see figure (24) for the first 10 genes returned by the algorithm, none of which has better rank than 218 for localization and 70).

Since the maximum-intensity projection can hide some well-sepearated regions, it is not as reliable as sections to evaluate separation properties, but still it is instructive to see how this local criterion can return genes that score low for localization and/or fitting but still have a distinct pattern around striatum. Slc32a1 has a clear but inhomogeneous pattern in striatum, and a high expression in the main olfactory bulb. Both features penalize the global fitting score, but only the second one penalizes the global localization score, which is consistent with the fitting rank being much lower than the localization rank. Ptpn5 exhibits a less contrasted but more homogeneous pattern in the striatum (it rather follows the caudoputamen than the striatum), but the expression in also quite high in the cerebral cortex. Caudoputamen is still striking and the gene was rescued by the local algorithm, even though the expression in the cerebral cortex severely penalizes Ptpn5 both for glocal localization (according to which it is ranked

	Fitting rank	Fitting scores, score of set 0.89656
Satb2	1	0.89
Ephb6	3	0.84
Igfbp6	18	0.81
Map3k5	38	0.79
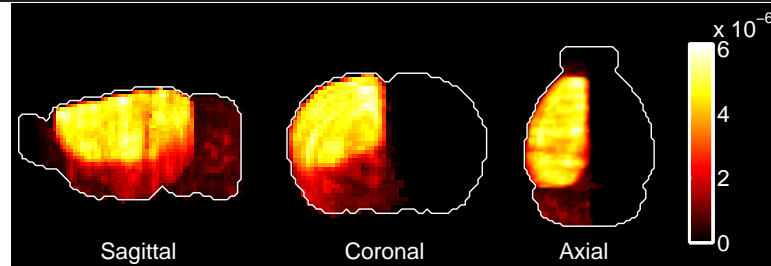
Figure 12. Optimal set of marker genes in the sense of generalized fitting scores. Heat map of the maximal-intensity projections of the (normalized) sum of the expression energies of the genes.
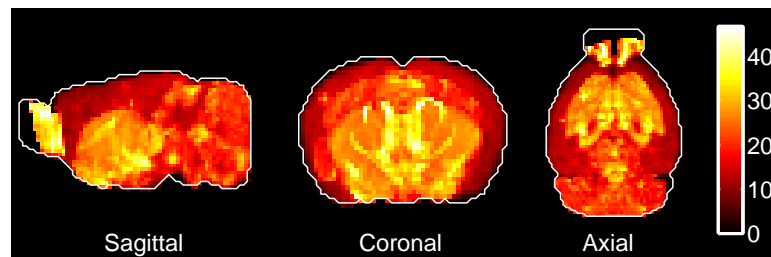
Figure 13. Best separator of striatum. Heat map of the maximal-intensity projection of Slc32a1.

2056) and global fitting (according to which it is ranked 634). It would be interesting, when repeating the experiment for the same gene, either in the same species or in different species, to see if the local separation property is as the global properties measured by the localization and fitting scores. A sagittal section drawn from the ISH data (figure (15)) confirms that Ptpn5 is highly expressed in the striatum, but also in the cortex, albeit to a lesser extent. The separation between cerebral cortex and striatum is clearly visible on the section. This is the separation property that our local criterion is supposed to detect.

3.5 Good co-markers

A table of the best few co-markers for striatum and cerebellum (see figure(26)) can be found in an appendix. It was obtained at a value $\tau = 0.5$ of the τ-balance constraint defined above. This value is somewhat arbitrary and and by the look of the tangent coefficients and fitting scores for the genes, there is no monitonicity of the value of the tangent wrt the value of the score. The user of the software can input a higher value of τ in order to explore genes with a higher tolerance on the relative fittings. We do not have a natural optimization criterion to propose to choose τ and therefore leave it as a parameter. It can be noted, however, than the τ-balance constraint at $\tau = 0.5$ for pallidum and cerebellum returns an empty set of markers. Thus, fixing the level of the balance constraint and counting the number of marker genes returned by the algorithm suggests an indication on the degree of solidary between pairs of regions.

The best two co-markers of the cerebellum and the striatum are Id4 (see figure (16)) and D330017J20Rik
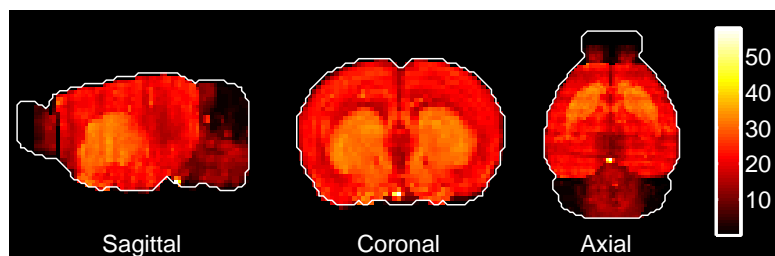
Figure 14. Third best separator of striatum. Heat map of the maximal-intensity projection of Ptpn5.
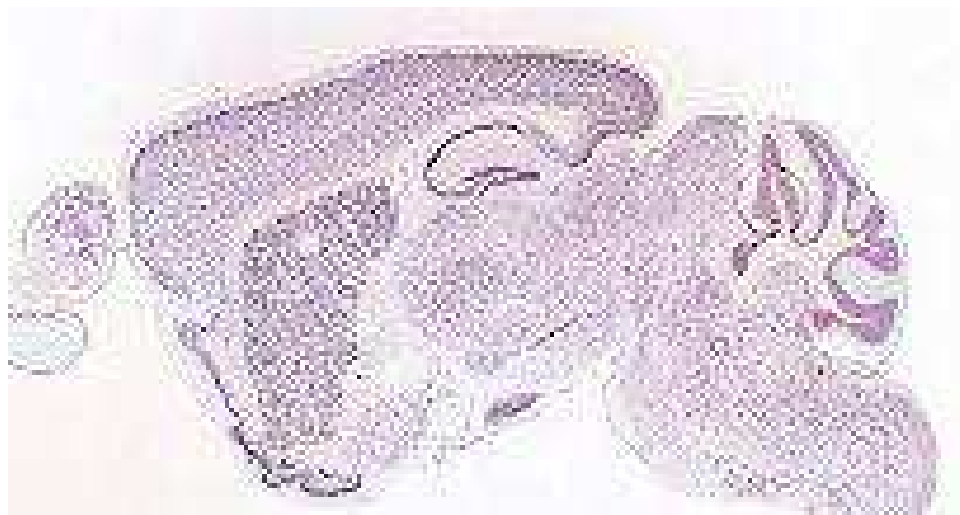
Figure 15. Ptpn5: ISH on a sagittal section for Ptpn5. Ptpn5 is detected as the third best local marker of striatum. Digitized expression energies suggested that this gene is also highly expressed in the cerebral cortex. The two structures are indeed visible and separable.
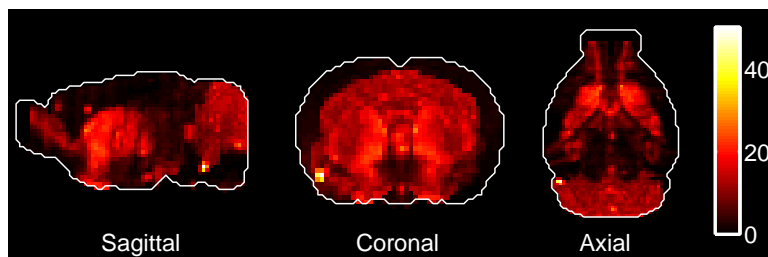
Figure 16. Best co-marker of striatum and cerebellum. Heat map of the maximal-intensity projection of Id4.
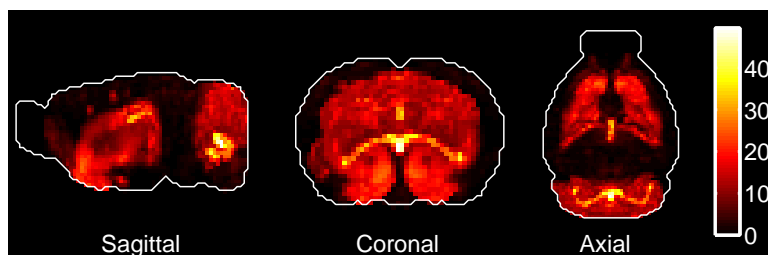
Figure 17. Second Best co-marker of striatum and cerebellum. Heat map of the maximal-intensity projection of D330017J20Rik.

(see figure (17)). The first one has a value of tangent very close to one, the second has a value of tangent close t0 0.86. Id4 has indeed a more homogeneous expression across striatum and cerebellum, and D330017J20Rik has a pattern of higher expression inside cerebellum, hence a tangent more remote from one, but the two genes show a clear pattern for the pair striatum-cerebellum.

Acknowledgments

References

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

2. 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 and M. Hawrylycz, *An anatomic gene expression atlas of the adult mouse brain*, Nature Neuroscience **12**, 356 - 362 (2009).

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

4. J.A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision and Materials Science*, Cambridge University Press, 1999.

5. L.D. Landau and E.M. Lifshitz, *Course of Theoretical Physics: Classical Mechanics*, Vol. 1, Pergamon Press (1976, 3rd ed.).

6. P.Grange and P.P. Mitra, *Algorithmic choice of coordinates for injections into the brain: encoding a neuroanatomical atlas on a grid*, `arXiv:1104.2616`.

7. J.W. Bohland *et al.*, *A proposal for a coordinated effort for the determination of brainwide neuroanatomical connectivity in model organisms at a mesoscopic scale*, PLoS Comput. Biol. **5(3)**, e1000334. PMID: 19325892.

8. P. Grange, T. Pologruto, V. Pinskiy, H. Wang, A. Khabbaz, S, Spring, L. Yu, J. Germann, M. Henkelman and P.P. Mitra, *Computer-guided stereotactic injections for the Mouse Brain Architecture project*, SFN Abstracts (2010), and in preparation.

9. L. French and P. Pavlidis, *Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain*, PLoS Comput. Biol. **7(1)**: e1001049..

4 Appendix: Reference values of the localization scores for the coarsest atlas of the left hemisphere
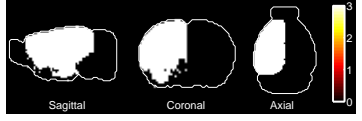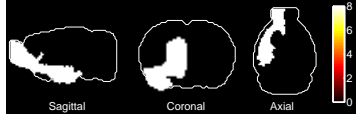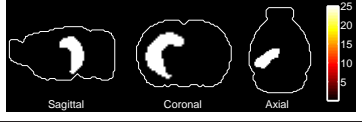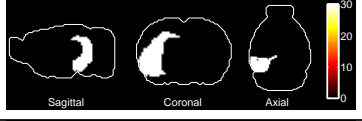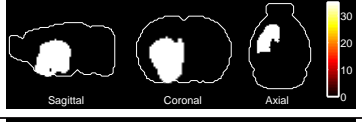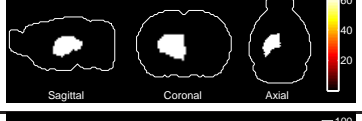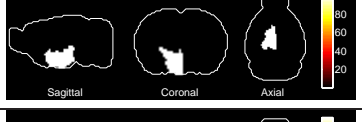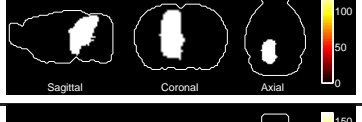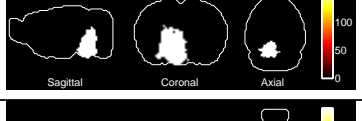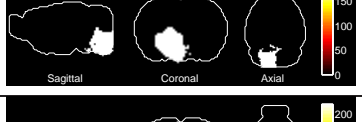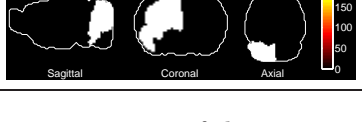
	Uniform reference	Average reference	Region profile
Cerebral cortex	0.295	0.476	
Olfactory areas	0.092	0.096	
Hippocampal region	0.0426	0.0642	
Retrohippocampal region	0.039	0.057	
Striatum	0.086	0.0605	
Pallidum	0.019	0.010	
Thalamus	0.043	0.030	
Hypothalamus	0.034	0.020	
Midbrain	0.078	0.041	
Pons	0.046	0.024	
Medulla	0.061	0.042	
Cerebellum	0.114	0.055	

Figure 18. Values of the uniform and average reference scores for each of the 12 main regions of the left hemisphere.

5 Appendix: Best-localized genes and characteristic functions for the 12 main regions of the left hemisphere
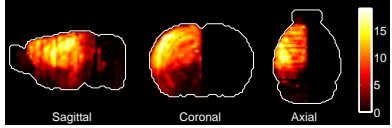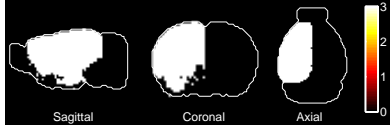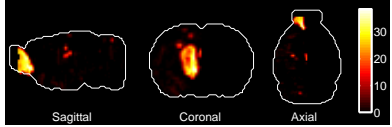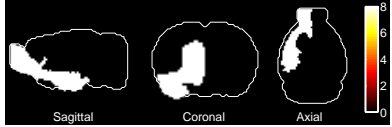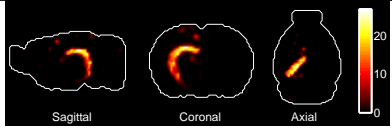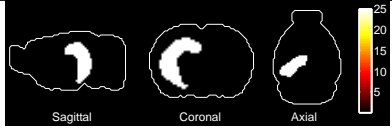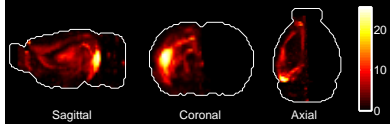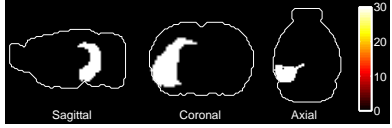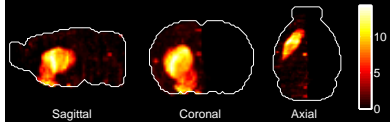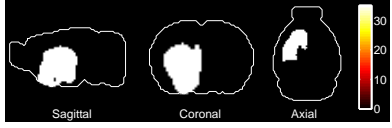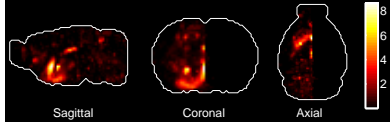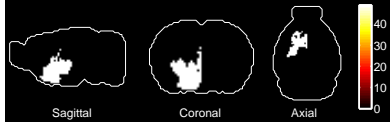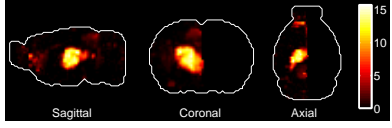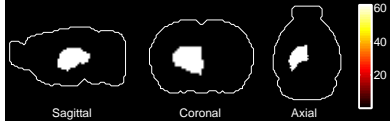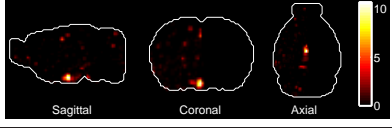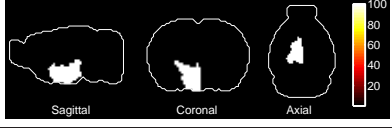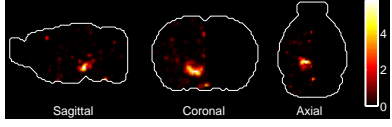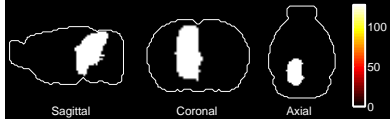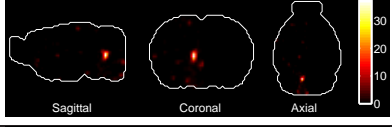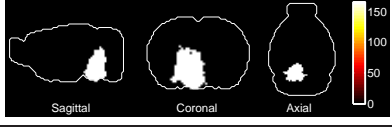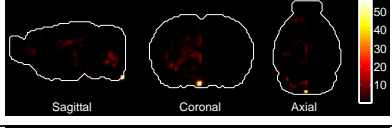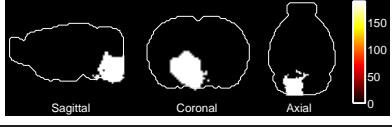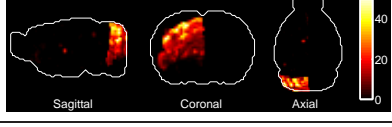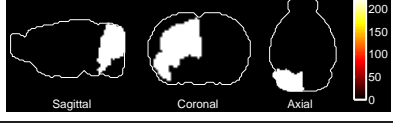
Brain Region	Best marker	Best-localized gene	Region profile
Cerebral cortex	Pak7		
Olfactory areas	C230040D10Rik		
Hippocampal region	Prox1		
Retrohippocampal region	Rxfp1		
Striatum	Ric8b		
Pallidum	Ebf4		
Thalamus	Lef1		
Hypothalamus	Nr5a1		
Midbrain	Ntsr1		
Pons	Dbh		
Medulla	Vil2		
Cerebellum	Gabra6		

Figure 10. The best few genes by localization for each of the 12 main regions of the left hemisphere

6 Appendix: Best-fitted genes and characteristic functions for the 12 main regions of the left hemisphere
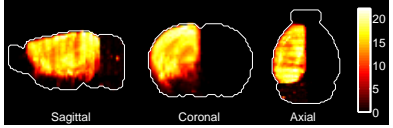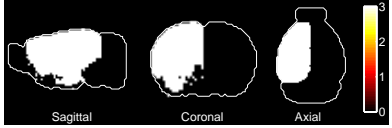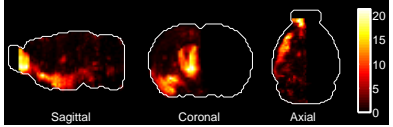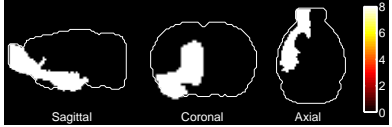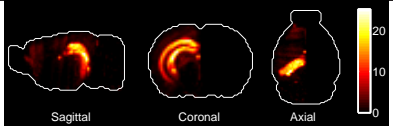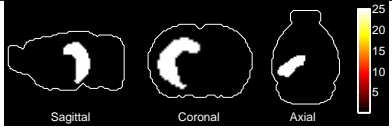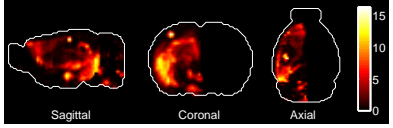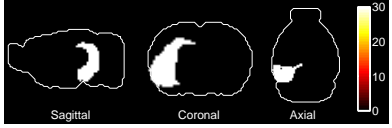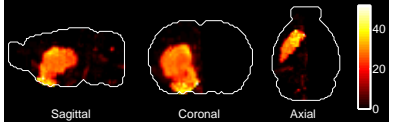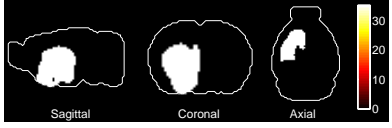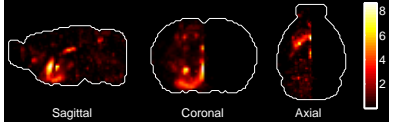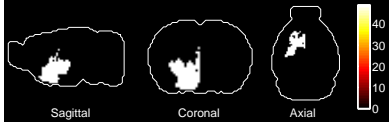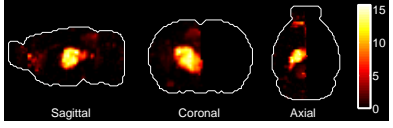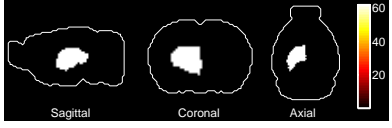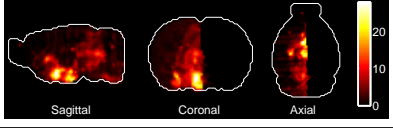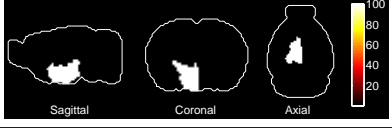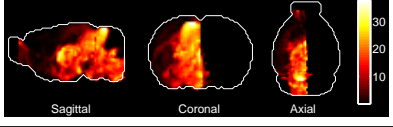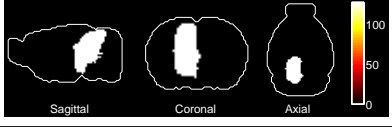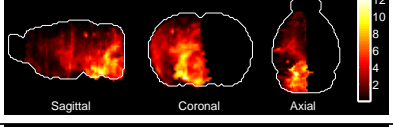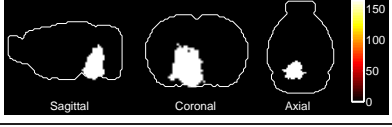
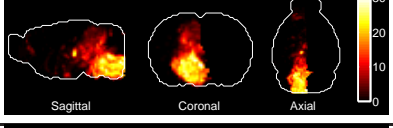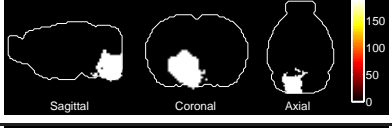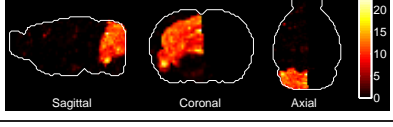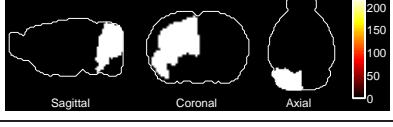| Brain region	Best marker	Heat map of best marker	Region profile
Cerebral cortex	Satb2		
Olfactory areas	Ppfibp1		
Hippocampal region	TC1412430		
Retrohippocampal region	Rspo2		
Striatum	Rgs9		
Pallidum	Ebf4		
Thalamus	Lef1		
Hypothalamus	Gpr165		
Midbrain	Slc17a6		
Pons	Klk6		
Medulla	Glra1		
Cerebellum	3110001A13Rik		

Figure 20. The best few genes by fitting score for each of the 12 main regions of the left hemisphere.

7 Appendix: Best sets of genes for localization in the twelve main regions of the left hemisphere

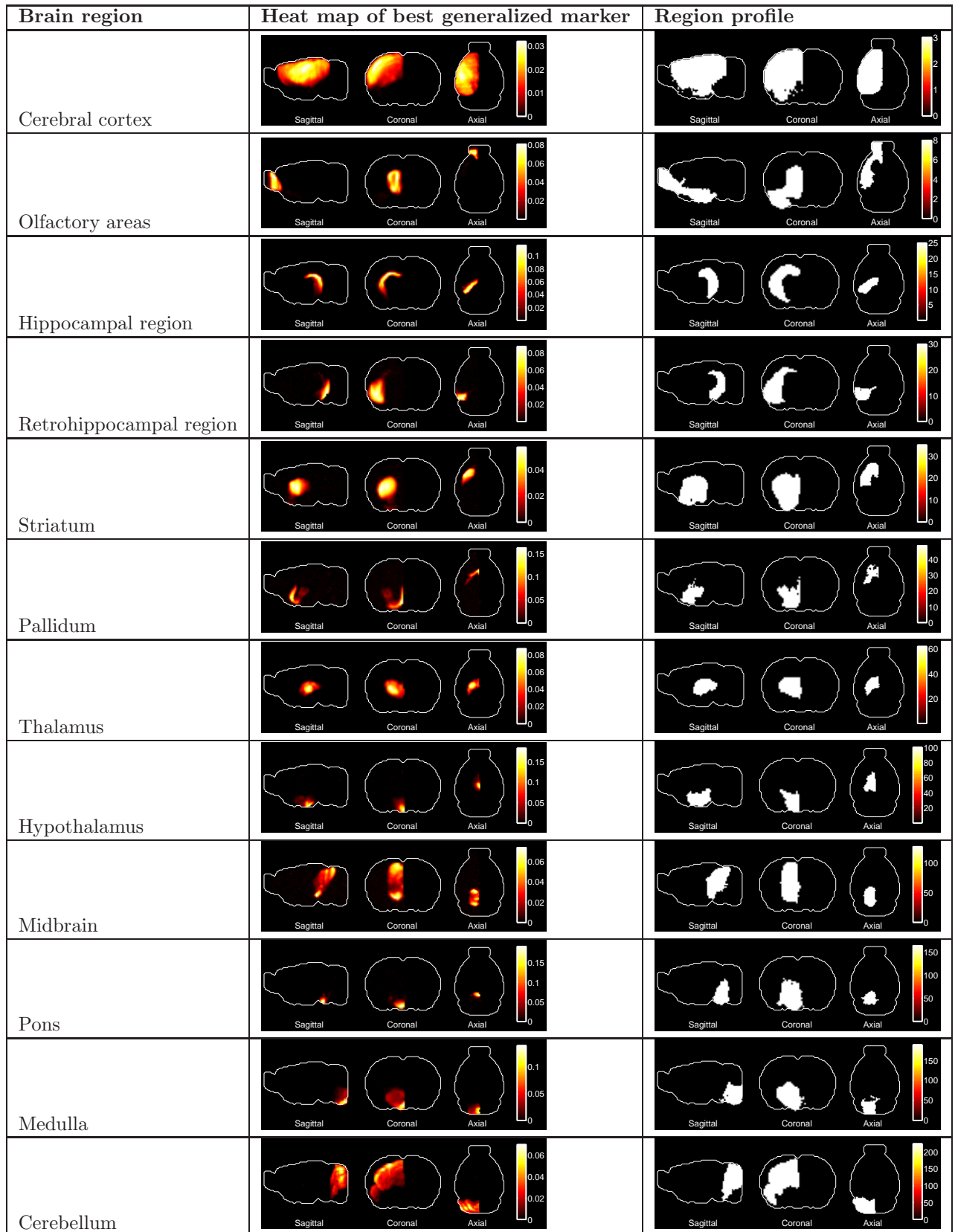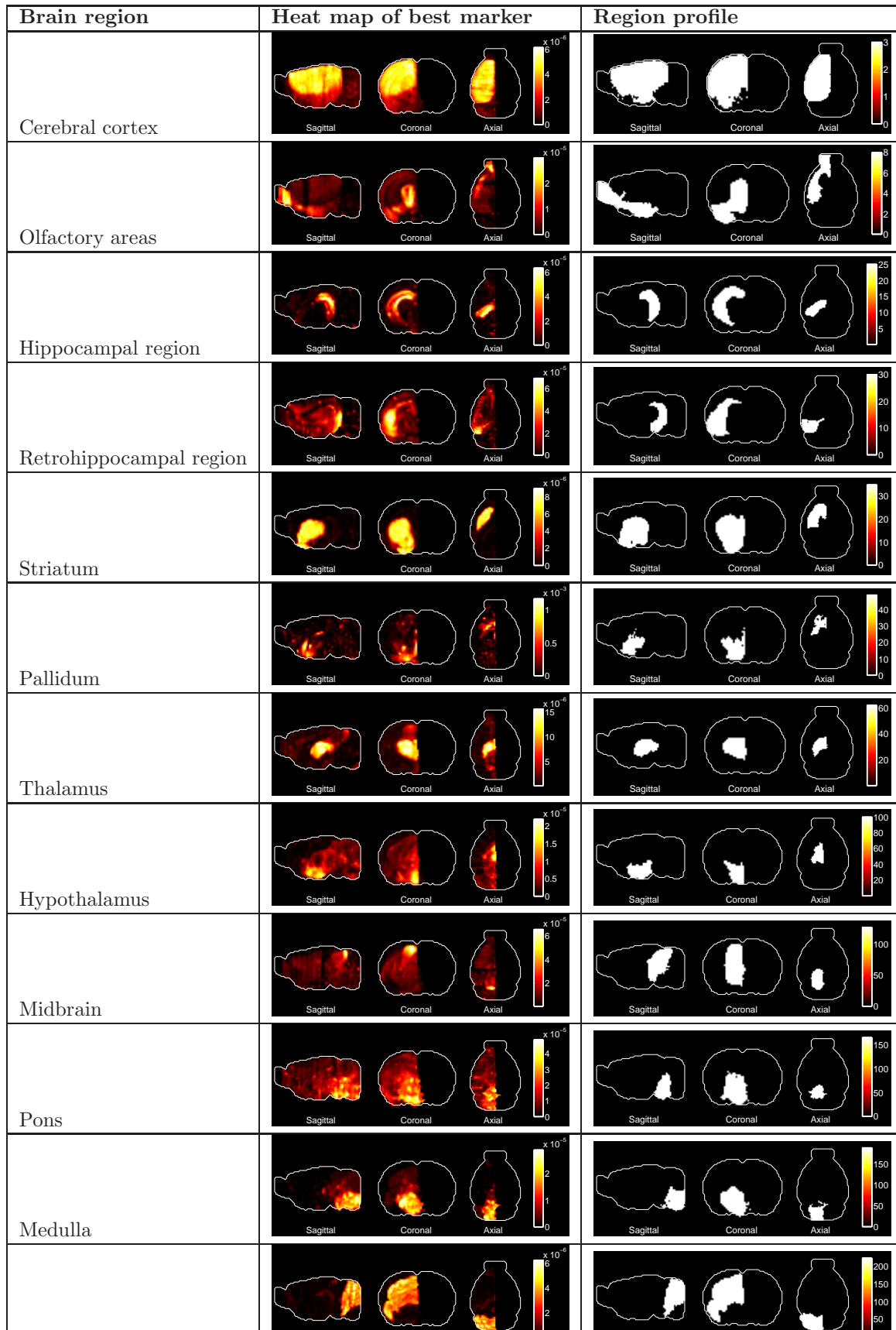| Brain region	Heat map of best generalized marker	Region profile
Cerebral cortex		
Olfactory areas		
Hippocampal region		
Retrohippocampal region		
Striatum		
Pallidum		
Thalamus		
Hypothalamus		
Midbrain		
Pons		
Medulla		
Cerebellum		

Figure 21. Expression profiles of the sets of genes that maximize the localization score for each of the

8 Appendix: Best sets of genes for fitting in the twelve main regions of the left hemisphere

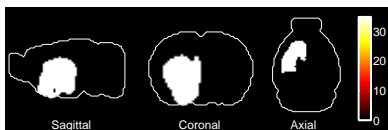| Brain region	Heat map of best marker	Region profile
Cerebral cortex		
Olfactory areas		
Hippocampal region		
Retrohippocampal region		
Striatum		
Pallidum		
Thalamus		
Hypothalamus		
Midbrain		
Pons		
Medulla		

9 Appendix: Best separators

Figure 23. Characteristic function of the reunion of the striatum and the cerebellum.

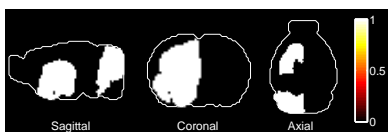|	Fitting rank	Localization rank	Heat map
Slc32a1	2327	218	
Abat	979	2997	
Ptpn5	634	2056	
Serpine2	107	807	
Gad2	898	2267	
Alcam	651	424	
Pcp4l1	2839	968	
Grik2	70	629	
Ckb	1733	2794	
Gad1	861	722	

10   Appendix: Best co-markers

Figure 25. Characteristic function of the reunion of the striatum and the cerebellum.

	Score for pair	Tangent coefficient	Heat map
Id4	0.54781	1.0085	
D330017J20Rik	0.52928	0.85764	
Grik2	0.45338	0.81055	
Vldlr	0.42963	0.80894	
8430415E04Rik	0.42156	0.80471	
Gad1	0.41918	1.0063	
Dpy19l3	0.39955	1.1353	
D13Bwg1146e	0.39028	1.1011	
Ppp2r5d	0.37326	1.0788	
Itpr1	0.36654	0.96994	

Plot of coefficients of the genes in the best−fitted set of genes in Cerebral cortex