

# A tool for identification of genes expressed in patterns of interest using the Allen Brain Atlas

December 23, 2008

Fred P. Davis and Sean R. Eddy  
HHMI Janelia Farm Research Campus  
19700 Helix Dr, Ashburn, VA 20147  
tel: 571.209.4000 x3037  
E-mail: [davisf@janelia.hhmi.org](mailto:davisf@janelia.hhmi.org)  
WWW: <http://research.janelia.org/davis>

Short title: Mining expression patterns in the mouse brain  
Keywords: spatial expression pattern, Allen Brain Atlas, mouse brain, neuronal cell types, regional specificity, expression gradient

Gene expression patterns can be useful in understanding the structural organization of the brain and the regulatory logic that governs its myriad cell types. A particularly rich source of spatial expression data is the Allen Brain Atlas (ABA), a comprehensive genome-wide *in situ* hybridization study of the adult mouse brain. Here we present an open-source program, ALLENMINER, that searches the ABA for genes that are expressed, enriched, patterned, or graded in a user-specified region of interest. Regionally enriched genes identified by ALLENMINER accurately reflect the *in situ* data (95-99% concordance with manual curation) and compare to regional microarray studies as expected from previous comparisons (61-80% concordance). We demonstrate the utility of ALLENMINER by identifying genes that exhibit patterned expression in the caudoputamen and neocortex. We discuss general characteristics of gene expression in the mouse brain and the potential application of ALLENMINER to design strategies for specific genetic access to brain regions and cell types. ALLENMINER is freely available on the Internet at <http://research.janelia.org/davis/allenminer>.

## 1 Introduction

The mouse brain is a complex tissue containing many neuronal and non-neuronal cell types organized in intricate three-dimensional structures. Defining the functional roles of these cell types in the context of their higher-order arrangements is an important challenge in neuroscience. Neuronal cell types have been traditionally defined by cell morphology, electrophysiological properties, and cell surface markers. Recent studies suggest that genomic transcriptome measurement is also a feasible route to defining functional cell types [1]. Besides its utility for classifying cell types, genomics also provides a bridge to genetic strategies to specifically target individual neuronal cell types for optogenetic or pharmacological perturbation [2]. A variety of gene expression technologies including microarrays [1, 3–5], serial analysis of gene expression (SAGE) [6], bacterial artificial chromosome (BAC) transgenics [7], and *in situ* hybridization (ISH) [8] has been used to characterize the expression profiles of brain regions and specific neuronal cell types.

The *in situ* hybridization data collected in the Allen Brain Atlas (ABA) offers the most spatially resolved ( $\sim 300 \mu\text{m}$ ) description of gene expression in the adult mouse brain available to date [8, 9]. The ABA contains sagittal *in situ* images for  $\sim 20$  thousand genes, coronal images for a subset of

~4 thousand genes, and three dimensional registration of these images by projection onto a reference atlas [9]. The registered expression data can be accessed in a number of ways including interactive visualization (BrainExplorer [10]), a summary description of expression levels in 17 anatomical brain regions, a tool (NeuroBLAST) to identify genes with expression patterns that are spatially similar to that of a query gene, a tool (AGEA) that defines a series of spatial regions based only on gene expression similarity and identifies genes that are enriched in these regions, and manually curated lists of genes that are enriched in 75 “fine structures”, such as small brain nuclei and layers of cortex.

The ABA expression data can answer a wide range of neurobiological questions. However, these questions often surpass the currently available query mechanisms. For example, we wished to identify genes that were differentially expressed between the medial and lateral caudoputamen, as behavioral studies have shown that these regions play distinct roles in learning [11, 12]. To improve the utility of the ABA in answering these kinds of queries, we have developed an open-source program, ALLENMINER, that searches for genes that are expressed, enriched, patterned, or graded in regions of interest. The search criteria are flexible so that it is generally applicable to expression analyses of the adult mouse brain.

We first describe the search strategy implemented in ALLENMINER. We next assess its accuracy by comparing query results to (i) manually curated and (ii) microarray identified lists of regionally-enriched genes. We then apply ALLENMINER to identify genes with patterned expression in the caudoputamen and neocortex. Finally, we discuss general characteristics of gene expression in the mouse brain and discuss ALLENMINER’s utility for designing genetic strategies to access restricted brain regions and cell types.

## 2 Results

### 2.1 Identification of regionally-enriched genes

The user specifies a region of interest (ROI) by either (i) describing the edges of a cuboid ( $x,y,z$  minimum and maximum) or (ii) listing the voxels in the integer coordinate system used by the ABA BrainExplorer [10], or by (iii) listing individual or boolean combinations of ABA Reference Atlas brain regions (Fig 1). An example of this last form of ROI definition is “CTX=on,HPF=off,OLF=off”, which selects all

voxels in the cortex (CTX) that are not also part of the hippocampal formation (HPF) or olfactory area (OLF). The examples presented in this paper use ROIs based on atlas regions, either as individual regions, combinations of regions, or sub-regions.

ALLENMINER processes the ABA three-dimensional expression files (.XPR files), which contain the registration results of the *in situ* series (Materials and Methods). These files contain the 3D coordinates of voxels that correspond to *in situ* image pixels with detectable expression. The expression in each voxel is quantified by a series of measures, including the estimated number of expressing cells, cell diameter, grid area spanned by expression, and total expression level.

We developed a score to quantify the enrichment of expression in the ROI. ALLENMINER iterates through each 3D expression file and quantifies the expression level, specificity, and enrichment in the user-specified ROI (*roi\_list* run mode). The program can compute the expression level in the ROI ( $expr(roi)$ ) as a sum or per-voxel average of any of the voxel expression measures, or as the number of component voxels with detectable expression. In the analyses presented here, the ROI expression level is defined as the sum of expression levels called for each component voxel. Specificity is computed as the fraction of a gene's total brain expression that occurs in the ROI (Eqn 1).

$$\text{Specificity}(roi) = \frac{expr(roi)}{expr(total)} \quad (1)$$

Enrichment is computed as the specificity normalized for the size (number of voxels) of the ROI relative to the whole brain (Eqn 2).

$$\text{Enrichment}(roi) = \frac{\text{Specificity}(roi)}{\frac{size(roi)}{size(total)}} \quad (2)$$

In the case of atlas region queries, the total levels ( $expr(total)$  and  $size(total)$ ) refer to the left-hemisphere "Brain" structure defined by the ABA; otherwise, they refer to the entire brain. Atlas region queries are accelerated by using a precomputed file containing expression statistics for all genes in all atlas regions (*fast\_query* run mode; 5 minutes on a single 2.0 GHz Intel Xeon processor). A similar indexing strategy for non-atlas ROI uses a uniform 3D gridding of the atlas to restrict searches to XPR files with expression in the grid sections corresponding to the ROI. For comparison, a non-indexed search completes in 5-10 minutes when run in parallel on fifty 3.0 GHz Intel Xeon processors.

To demonstrate an ALLENMINER query, we searched for genes enriched in the ventromedial hypothalamus (VMH) compared to the rest of the hypothalamus (HY). As expected, the results suggest that most genes are expressed in the VMH at levels similar to the rest of the HY, although a spectrum is observed that ranges from VMH-depletion to VMH-enrichment (Fig 2). Examples of genes that are enriched in the VMH include Fez family zinc finger 1 (Fezf1; Fig 2c) and Patched-2 (Ptchd2; Fig 2d).

## 2.2 Comparison to regionally enriched genes manually curated from the ABA

To assess the validity of the ALLENMINER enrichment score, we first compared it to lists of the hundred genes manually curated from the ABA *in situ* data to be the most enriched in 12 brain regions: cerebellum, cortex, hippocampus, hypothalamus, midbrain, medulla, olfactory bulb, pallidum, pons, retrohippocampal region, striatum, and thalamus (Supp Table 3 in [8]). This comparison assessed how the ALLENMINER enrichment score performs relative to manual curation using the same expression data. Receiver-operator characteristic (ROC) analysis was performed to characterize the ability of the score to distinguish between genes curated as enriched (positive set) and the remaining ABA genes (negative set). The optimal score threshold was defined as that which gave rise to the point on the ROC curve closest to the perfect discriminator (true positive rate = 1, false positive rate = 0). In cases where multiple expression files were available for a gene, the highest enrichment score was used.

The areas under the ROC curves (AUC) demonstrate a 95-99% concordance between ALLENMINER results and the manually curated lists (Fig 3a, Table 1). Restricting the ROC analysis to coronal *in situ* data degraded concordance to 80-96% (data not shown). This comparison suggests that ALLENMINER enrichment queries return accurate results expected from the *in situ* data.

## 2.3 Comparison to regionally enriched genes identified by microarray studies

Next, we compared ALLENMINER to two published microarray studies of regional enrichment in the mouse brain [13, 14]. This comparison characterizes not only ALLENMINER accuracy, but also the similarity of regional enrichment detection by the two expression technologies. As the technologies suffer from different kinds of errors, we expected poorer concordance than the previous comparison to manually curated *in situ* data, although we still expected general agreement.

The microarray identified genes were mapped to the ABA through UCSC genome tables (Materials and Methods). The first microarray dataset identified 299 ABA genes that were enriched at least 3.5 fold in the amygdala, cerebellum, hippocampus, olfactory bulb, or periaqueductal gray of 3 week old female mice [13]. The second study identified 37 ABA genes that were enriched in the fetal ventromedial hypothalamus (VMH) compared to the rest of the hypothalamus (HY) [14]. ALLENMINER enrichment queries were performed using the corresponding atlas regions (AMY, CB, HIP, MOB, PAG; VMH vs HY).

ROC analysis demonstrated a 61-80% concordance between ALLENMINER results and the microarray data (Table 1). This agrees well with a published genome-wide comparison of ABA and microarray expression data in the GNF [3] and Terragenomics [15] compendia of mouse tissue expression data that found an absence/presence call agreement of 58-72% [16]. Surprisingly, the adult ALLENMINER results exhibited better concordance with the fetal VMH study than with the adult brain regions analyzed in the Zirlinger study [13].

## 2.4 False negatives

Visual inspection of *in situ* images for the 12 microarray identified genes that exhibited an ALLENMINER VMH-vs-HY enrichment score of less than 1.5 suggested four reasons for these apparent false negatives (Table 5; Fig 4). First, several genes exhibited VMH expression below the threshold required by the ABA pipeline to call many expressing pixels, and therefore had no registered VMH voxels (Kcnj5, Mst1r, Nmbr, Ntn2l, Sema3a; Fig 4a). Second, tissue damage, bubbles, and other *in situ* artifacts produced several false negatives (Mst1r, Nmbr; Fig 4b).

Third, the VMH was often very sparsely sampled by *in situ* sections or even skipped entirely (Fig 4c). In two image series (A2bp1, Satb2), a gene appeared to express strongly in the VMH, but the sagittal slice was 3D registered onto the edge of the atlas VMH definition rather than the VMH itself. However, it is not clear from the images if there was indeed VMH expression, or rather expression in an adjacent structure, such as the tubercular nuclei that lies ventral to the VMH.

Finally, it is possible that several of the apparent false negatives are actually true negatives caused by variation of gene expression across development. For instance, Vgll2, the most enriched gene in the

fetal VMH (8.2 fold), exhibited no detectable adult expression (Fig 4d). This agrees with a published *in situ* study that found a dramatic reduction in expression from the significant enrichment observed at birth (developmental stage P0) to subtle expression at stage P7, and complete abolition at stage P21 [14]. Although it is difficult to clearly distinguish true from false negatives, expression was also not evident on the *in situ* images for B3gnt3, B530002L08, and Card14. Similarly, Titf1 expression appears to be strong throughout the hypothalamus and not particularly enriched in the VMH.

## 2.5 False positives

Visual inspection of *in situ* images for genes identified by ALLENMINER, but not by microarray, to be VMH-enriched (18.5% estimated false positive rate; Table 1) suggested several potential causes of false positives (Fig 5). The most prevalent cause appears to be *in situ* image artifacts that range from small particulate aggregates registered as single voxels (*eg* Fbxl7; Fig 5a) to fluid bubble artifacts registered as a curve of expressing voxels (*eg* Gna15; Fig 5b). Particularly in the absence of expression in the surrounding tissue, even a small number of artefactual voxels registered in the ROI produces a high enrichment score. Imposing a minimum threshold on the number of expressing voxels or the expression level in the ROI may minimize the effects of these artifacts. However, this strategy is problematic for small nuclei since a small number of artefactual voxels is difficult to distinguish from true low level expression, especially given the sparse sampling in small ROI. For example, Cholecystokinin B receptor (Cckbr; 11 VMH voxels, enrichment score = 12.53), Malic enzyme 2 (Me2; 7 voxels, enrichment score = 18.34), and Calcium binding protein 39-like (Cab39l; 4 VMH voxels, enrichment score = 21.46) all express in few voxels yet appear to be truly enriched on the *in situ* images (data not shown). In other instances, a folded tissue edge (*eg* Gnal; enrichment score = 11.88; Fig 5c) or tissue debris lying in the ROI (*eg* Tctex1d1; enrichment score = 16.09; Fig 5d) is registered as expression.

Finally, a portion of the apparent false positives are likely true positives that are either differentially expressed across development, or were missed by the microarray experiment. It is not feasible to quantify the fraction of the “false positives” that are true expressors, but many genes appear to be enriched in the ABA although not identified by the microarray study. For example, Cylindromatosis (Cyld; score = 27.38; Fig 5e) and Patched-2 (Ptchd2, score = 23.65; Fig 2d) are clearly enriched in the ABA, but

neither is identified by the fetal microarray study.

## 2.6 Identification of genes with patterned or graded expression

To identify genes that are expressed non-uniformly or in a graded fashion across an ROI, for example the mediolateral axis of the caudoputamen, (1) partitions are first defined along an axis (rostrocaudal, dorsoventral, or mediolateral; `roi_partition` run mode), (2) expression levels are quantified in each partition (`roi_list` mode), and (3) these results are used to compute cross-axis gradient and regional patterning scores (`calc_entropy` mode). The gradient score describes how much of the total expression change from partition to partition occurs in the same numerical direction (Eqn 3). A gene that expresses in ever-decreasing amounts across an axis receives a gradient score of -1, while ever-increasing amounts receives a gradient score of 1.

$$\text{Gradient score} = \frac{\sum_{i=2}^n \text{Enrichment}(roi_i) - \text{Enrichment}(roi_{i-1})}{\sum_{i=2}^n |\text{Enrichment}(roi_i) - \text{Enrichment}(roi_{i-1})|} \quad (3)$$

The regional patterning score (RPS) is a Shannon entropy-like score that sums the enrichment observed in each ROI partition, weighted by the occupancy of each partition (Eqn 4). The RPS score decreases as a gene expresses more uniformly across the defined ROI partitions, and can become slightly negative if a gene is depleted in an ROI partition.

$$\text{Regional patterning score} = \sum_{i=1}^n \text{Specificity}(roi_i) \log_2(\text{Enrichment}(roi_i)) \quad (4)$$

We performed regional patterning searches in the neocortex and caudoputamen to identify genes that were differentially expressed across these structures. Five rostrocaudal (R-C) bins were defined across the neocortex, as defined by the ABA cortex region after removal of voxels from the hippocampal formation and olfactory area. Similarly, five R-C and five mediolateral (M-L) bins were defined across the caudoputamen. The M-L bins were constructed to adjust to the width of the caudoputamen, which is greatest in the rostral and smallest in the caudal aspects. Gradient (Eqn 3) and regional patterning scores (Eqn 4) were computed for each gene to determine the gradient and non-uniformity of expression across the axis of interest, respectively.

The query results demonstrate that although most genes are uniformly expressed across the neocortex R-C and caudoputamen R-C and M-L axes, there exists a continuum of patterns, as quantified by the regional patterning and gradient scores (Fig 6, 7). The distributions of gradient scores in all three structures exhibit a shared tri-modal shape that appears to arise for common reasons (Fig 6a, 7a, 7e). The main peak over a gradient score of 0 is due to non-expressors as well as uniform expressors. In addition to genes that are actually expressed in perfect negative and positive gradients, the two smaller peaks at the extrema of -1 and +1 often correspond to expression calls that are from neighboring structures but are registered on the edge of the first and last partitions, respectively. For example, the medial caudoputamen peak corresponds to expression in a neighboring ventricle, and the lateral peak to expression in the cortex, both of which are often registered on the edges of the caudoputamen (Fig 7e).

The patterning scores also suggest that most genes express uniformly across the queried structures. These distributions exhibit a similar shape for all three queries, with a peak at a value of 0, corresponding to uniform expression, and long tails towards higher values, corresponding to patterned expression (Fig 6b, 7b, 7f). The distribution is smoother for the neocortex than for the caudoputamen. There are two likely explanations for this observation. First, the edge effect observed in the gradient distributions of the caudoputamen queries, especially across the M-L axis, are larger in magnitude than that of the neocortex R-C query. These outliers correspond to the small peaks of high patterning scores observed in both the R-C and M-L caudoputamen queries (Fig 7b, 7f). Second, because the caudoputamen is smaller than the neocortex, it is sampled less thoroughly by the *in situ* slices. This reduced sampling leads to a more significant discrete binning effect across the ROI partitions, resulting in more fluctuation in the patterning score distribution.

The gradient distributions may reflect the underlying cellular architecture of the structures. For example, the neocortical R-C gradient distribution is biased towards rostral enrichment (Fig 6a), while the caudoputamen R-C gradient is more symmetric and slightly biased towards caudal enrichment (Fig 7a). This may correspond to the known variation of cell density and morphology along the neocortical R-C axis [17–19]. In contrast, although there are regional neurochemical differences in the caudoputamen, it is often considered a structurally and cytoarchitectonically homogeneous tissue [20, 21], and variability in cellular composition does not exist along its R-C axis.

## 2.7 Genome-wide distribution of expression patterns and the diversity of neuronal cell types

Beyond queries for particular regions and patterns of interest, the comprehensive nature of the ABA also allows unique insight into the diversity of expression patterns and, potentially, of neuronal cell types that exist in the mouse brain. The specificity of gene expression in the brain was quantified for each gene by computing the regional patterning score across all 210 atlas regions (Fig 8a). The initial ABA report used the number of expressing voxels as a measure of specificity, suggesting that most genes expressed in specific patterns [8]. Here, we find that although most genes express relatively sparsely (Fig 8b), they do so in patterns that are not specific with respect to anatomical regions (Fig 8a). Comparing the number of expressing voxels to the RPS demonstrates a range of patterning for genes with comparable voxel occupancy, suggesting that the RPS score is a useful additional dimension to quantify gene expression specificity (Fig 8b). For example, *Rgs5* (Fig 8c,8e) and *Gabra4* (Fig 8d, 8f) have detectable expression in a similar number of voxels, however the former is expressed in a dispersed fashion across several brain regions, while the latter is enriched in the thalamus, caudoputamen, and cortex.

This analysis provides a molecular basis for discussions of neuronal cell type diversity. The spatial expression data in the ABA are most obviously useful for identifying region-specific genes. However, cell types in the brain are not necessarily, or even predominantly, region-specific. If we assume a parsimonious genomic definition of cell types, it is likely that cell types are defined by a single or small number of genes more often than by several genes. Previous microarray analysis of 12 neuronal subpopulations suggest that this is a reasonable assumption [1]. Sugino, *et al* found that although functional cell types often exhibit differential expression of a large battery of genes, as few as 5 genes are necessary to uniquely identify a particular cell type [1]. Following from this assumption and the observed distributions of voxel occupancy and regional patterning (Fig 8a, 8b), it is likely that most cell types are dispersed throughout the brain (*eg* GABA-ergic neurons) rather than restricted to a brain region (*eg* cerebellar Purkinje cells). Although the definition of a cell type is still debated and its catalog in the brain far from complete [22, 23], it is likely that there is a spectrum of regional specificity and a hierarchical resolution to its definition that spans these two extreme scenarios.

## 3 Discussion

We presented a tool for three-dimensional searches of expression data in the Allen Brain Atlas (Fig 1), demonstrated its accuracy and utility in identifying genes with specific (Figs 2, 3, 5, 4; Tables 1, 5) and patterned expression (Fig 6, 7). We will now discuss potential improvements to ALLENMINER performance and the utility of ALLENMINER for designing genetic strategies to access restricted neuronal subtypes.

### 3.1 Future improvements

The comparison of VMH enriched genes identified by ALLENMINER to those identified by microarray analysis suggested properties of the registered ABA data, such as *in situ* artifacts, that affect ALLENMINER queries (Fig 4, 5). Although these artifacts are best handled at the image analysis stage of the ABA processing and registration pipeline, some of them, such as the particulate and bubble artifacts, can likely be identified in the three-dimensional registrations and the corresponding voxels flagged for ALLENMINER analysis.

The distributions of expression gradient in the neocortex and caudoputamen highlighted the edge effect that results from expression in neighboring tissues being registered within the ROI (Fig 6a, 7a, 7e). The precise registration of regional boundaries is a difficult problem that requires single voxel accuracy. However, the impact of these errors can likely be minimized by peeling away the outer layer of voxels from the ROI definition. This would reduce the contribution of voxels that have been mis-registered from neighboring tissues, such as the cortical expression registered as caudoputamen, to the gradient and patterning scores.

The biological and technical non-uniformities in the expression data confound precise quantitative analysis of the spatial expression patterns. For example, the rostral bias in the gradient distribution of the neocortex (Fig 6a) may reflect rostrocaudal variation in cell morphology, size, and density [17–19]. Similarly, the *in situ* sampling of tissue slices across the brain series often varies between genes. ALLENMINER currently does not correct for these non-uniformities. One possible way to correct for cell density is to normalize the observed expression levels or gradient scores using a ubiquitous or pan-neuronal marker, for example that captured by the ABA fractional area metric [8]. However, this

strategy would not correct for the tissue sampling differences across partitions, as the sampling often varies from *in situ* series to series. Similarly, sparse tissue sampling is not currently distinguished from very patterned expression. Non-uniform sampling might be corrected for by computing a normalization factor proportional to the number of tissue slices that correspond to each ROI partition.

An application programming interface (API) to the ABA has recently been released that provides standardized access to the ABA data (<http://mouse.brain-map.org/api/index.html>), facilitating the development of custom software. ALLENMINER development began before the API was released, and so we developed parsers for the ABA data to which access was required, such as the three-dimensional expression profiles. However, the search algorithm we describe here is equally applicable to implementation with the API.

Currently, ALLENMINER only uses the three-dimensional registered expression data in the ABA. However, the original two-dimensional *in situ* images may contain additional information, such as the shape of the hybridization signal in each soma, nucleus, or dendrite. This high resolution information about the shape and distribution of the *in situ* signal may be useful in the classification of neuronal subtypes and can possibly be extracted by applying image analysis techniques to the raw *in situ* images.

### 3.2 Designing neurogenetic strategies

ALLENMINER may also be useful in identifying genes or combinations of genes that express in a specific region or cell type of the mouse brain. Although the resolution of the registered *in situ* data precludes precise definition of cell types, the regionally patterned genes that are identified by ALLENMINER include ion channels, solute carriers, and neurotransmitter machinery that can provide a link between genomic definitions and traditional electrophysiological and neurochemical definitions of cell type. Further integration of ABA expression data with other publicly available databases, such as the list of currently available transgenic mouse lines (MGD; [24]) may highlight brain regions that can be specifically addressed by intersectional strategies [2].

The ABA is a rich collection of spatially resolved three-dimensional gene expression patterns in the mouse brain. ALLENMINER enables an unlimited repertoire of “virtual” experiments using the ABA, and we expect this to facilitate the generation and validation of neurobiological hypotheses.

## 4 Materials and Methods

### 4.1 Obtaining Allen Brain Atlas data

The list of genes available from the ABA was downloaded (<http://mouse.brain-map.org/pdf/allGenes.csv>; Nov 5, 2007). For each gene, an XML file was downloaded that describes the available *in situ* series and the riboprobes used to generate them. The image series identifiers were extracted from the XML files and the corresponding 3D expression (.XPR) files were downloaded (25,636 XPR files corresponding to 21,201 genes). The 3D reference atlas that describes the brain structures to which each voxel belongs, as well as the neuroanatomical hierarchy of these structures, was retrieved from the Annotation100.sva and BrainStructures.csv files, respectively, of the publicly available API package (March 27, 2008).

The XPR files were linked to the knownGenes in the mouse mm9 UCSC genome database through riboprobe identifiers and the knownToAllenBrain table ([25], <http://genome.ucsc.edu>). In addition, knownGene, knownToLocusLink, knownToRefSeq, all\_est, estOrientInfo, gbCdnalInfo, all\_mrna, kgXref, tables were obtained from the UCSC mm9 mouse genome database. The *in situ* images presented here can be accessed using their ABA image identifiers. For example, image 73429295 is available at <http://mouse.brain-map.org/viewImage.do?imageId=73429295>.

### 4.2 Mapping microarray probes to ABA XPR files

The microarray study of amygdala, cerebellum, hippocampus, olfactory bulb, and periaqueductal gray in 3 week old female mice identified 452 probes enriched at least 3.5 fold in one of the 5 regions [13]. Enriched Affymetrix probe set ids (Mu11Ka, Mu11Kb, Mu19Ka, Mu19Kb, Mu19Kc platforms) were mapped to mm9 UCSC knownGenes using the Entrez Gene ID from the Affymetrix annotation file (<http://www.affymetrix.com>) and UCSC genome browser KnownToLocusLink table. In cases where the Entrez GeneID was not available in the annotation file or the knownToLocusLink table, the aligned chromosome location (available for Mu11k, but not Mu19k) was used to search the knownGene table for the gene that contained the probe in the correct orientation. In total, 410 probes were mapped to 377 knownGenes, 299 of which mapped to the ABA (via knownToAllenBrain).

The microarray study of the fetal ventromedial hypothalamus (VMH) identified fifty genes that are

the most enriched compared to the rest of the hypothalamus (HY) [14]. Forty-six of these genes mapped to UCSC knownGenes table and 37 also mapped to the ABA through the UCSC knownToAllenBrain table.

### 4.3 Availability

ALLENMINER is implemented in Perl except for a single routine written in Python. It is freely available under the GPL v3 license at <http://research.janelia.org/davis/allenminer>. The results of enrichment analysis for all ABA reference brain regions, as well as all the results presented in this paper, are also available for download.

## 5 Acknowledgments

We are grateful to the Allen Brain Institute, in particular Susan Sunkin and Mike Hawrylycz, for making the Allen Brain Atlas expression data and API publicly available. We thank Lee Henry, Alla Karpova, and Albert Lee (HHMI) for useful discussion and Goran Ceric for managing Janelia's high performance computing resources.

## References

- [1] Sugino, K., Hempel, C. M., Miller, M. N., Hattox, A. M., Shapiro, P., Wu, C., Huang, Z. J., and Nelson, S. B. (2006) Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nat Neurosci*, **9**(1), 99–107.
- [2] Luo, L., Callaway, E. M., and Svoboda, K. (2008) Genetic dissection of neural circuits. *Neuron*, **57**(5), 634–660.
- [3] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**(16), 6062–6067.
- [4] Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing,

- Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A., et al. (2008) A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci*, **28**(1), 264–278.
- [5] Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., and Geschwind, D. H. (2008) Functional organization of the transcriptome in human brain. *Nat Neurosci*, **11**(11), 1271–1282.
- [6] Khattra, J., Delaney, A. D., Zhao, Y., Siddiqui, A., Asano, J., McDonald, H., Pandoh, P., Dhalla, N., Prabhu, A. L., Ma, K., et al. (2007) Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome Res*, **17**(1), 108–116.
- [7] Gong, S., Zheng, C., Doughty, M. L., Losos, K., Didkovsky, N., Schambra, U. B., Nowak, N. J., Joyner, A., Leblanc, G., Hatten, M. E., and Heintz, N. (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature*, **425**(6961), 917–925.
- [8] Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**(7124), 168–176.
- [9] Ng, L., Pathak, S. D., Kuan, C., Lau, C., Dong, H., Sodt, A., Dang, C., Avants, B., Yushkevich, P., Gee, J. C., et al. (2007) Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Trans Comput Biol Bioinform*, **4**(3), 382–393.
- [10] Lau, C., Ng, L., Thompson, C., Pathak, S., Kuan, L., Jones, A., and Hawrylycz, M. (2008) Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, **9**, 153.
- [11] Corbit, L. H. and Janak, P. H. (2007) Inactivation of the lateral but not medial dorsal striatum eliminates the excitatory impact of pavlovian stimuli on instrumental responding. *J Neurosci*, **27**(51), 13977–13981.
- [12] Grahn, J. A., Parkinson, J. A., and Owen, A. M. (2008) The cognitive functions of the caudate nucleus. *Prog Neurobiol*, **86**(3), 141–155.

- [13] Zirlinger, M., Kreiman, G., and Anderson, D. J. (2001) Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei. *Proc Natl Acad Sci U S A*, **98**(9), 5270–5275.
- [14] Kurrasch, D. M., Cheung, C. C., Lee, F. Y., Tran, P. V., Hata, K., and Ingraham, H. A. (2007) The neonatal ventromedial hypothalamus transcriptome reveals novel markers with spatially distinct patterning. *J Neurosci*, **27**(50), 13624–13634.
- [15] Zapala, M. A., Hovatta, I., Ellison, J. A., Wodicka, L., Rio, J. A. D., Tennant, R., Tynan, W., Broide, R. S., Helton, R., Stoveken, B. S., et al. (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. *Proc Natl Acad Sci U S A*, **102**(29), 10357–10362.
- [16] Lee, C. K., Sunkin, S. M., Kuan, C., Thompson, C. L., Pathak, S., Ng, L., Lau, C., Fischer, S., Mortrud, M., Slaughterbeck, C., et al. (2008) Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome Biol*, **9**(1), R23.
- [17] Elston, G. N. (2003) Cortex, cognition and the cell: new insights into the pyramidal neuron and prefrontal function. *Cereb Cortex*, **13**(11), 1124–1138.
- [18] Benavides-Piccione, R., Hamzei-Sichani, F., Ballesteros-Yanez, I., DeFelipe, J., and Yuste, R. (2006) Dendritic size of pyramidal neurons differs among mouse cortical regions. *Cereb Cortex*, **16**(7), 990–1001.
- [19] Schuz, A. and Palm, G. (1989) Density of neurons and synapses in the cerebral cortex of the mouse. *J Comp Neurol*, **286**(4), 442–455.
- [20] Glynn, G. and Ahmad, S. O. (2002) Three-dimensional electrophysiological topography of the rat corticostriatal system. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, **188**(9), 695–703.
- [21] Nieuwenhuys, R. (1998) *The Central Nervous System of Vertebrates*, Springer, Berlin.
- [22] Masland, R. H. (2004) Neuronal cell types. *Curr Biol*, **14**(13), R497–R500.
- [23] Yuste, R. (2005) Origin and classification of neocortical interneurons. *Neuron*, **48**(4), 524–527.

- [24] Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., and Blake, J. A. (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res*, **36**(Database issue), D724–D728.
- [25] Karolchik, D., Kuhn, R. M., Baertsch, R., Barber, G. P., Clawson, H., Diekhans, M., Giardine, B., Harte, R. A., Hinrichs, A. S., Hsu, F., et al. (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, **36**(Database issue), D773–D779.

## List of Figures

1	ALLENMINER logic. . . . .	18
2	ALLENMINER identified genes enriched in the ventromedial hypothalamus. . . . .	19
3	Receiver-operator characteristic of ALLENMINER enrichment score. . . . .	20
4	Examples of ALLENMINER false negative errors. . . . .	21
5	Examples of ALLENMINER false positive errors. . . . .	22
6	Patterned expression in the neocortex . . . . .	23
7	Patterned expression in the caudoputamen. . . . .	24
8	Distribution of gene expression specificity in the mouse brain. . . . .	25

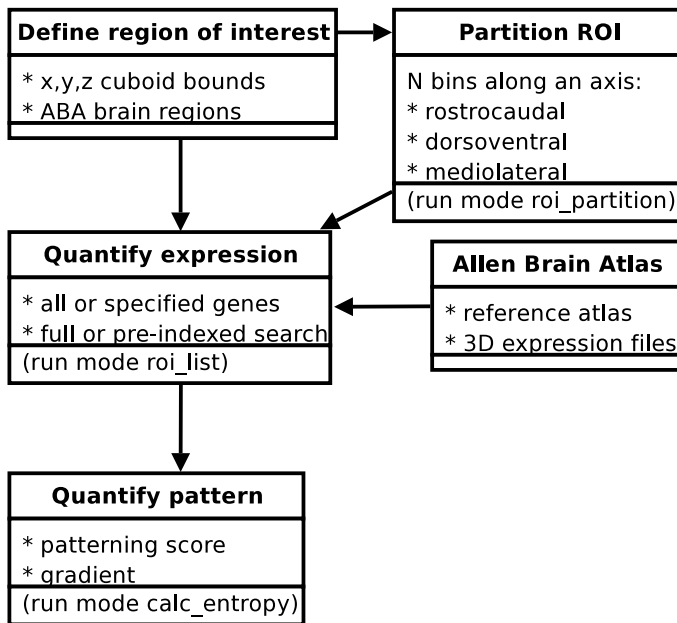
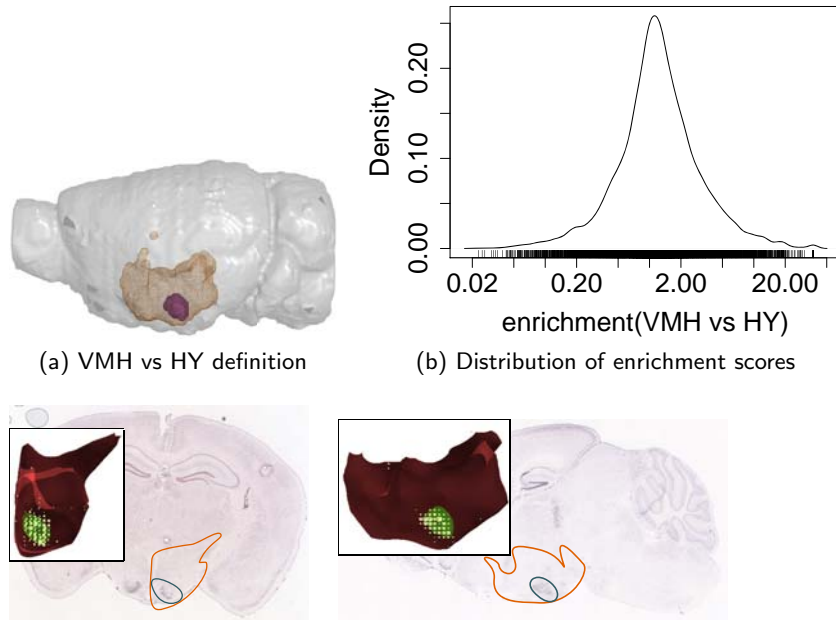


Figure 1: ALLENMINER logic. The program begins with a user-specified region of interest (ROI), optionally partitions it along an axis, and then quantifies expression across the XPR files in the Allen Brain Atlas. If multiple ROI have been defined, for example by partitioning, patterning and gradient scores may also be computed.



(c) Fez family zinc finger 1 (Fezf1; enrichment = 19.16; ABA image 71774899)

Figure 2: ALLENMINER identified genes enriched in the ventromedial hypothalamus. (a) The ventromedial hypothalamus (VMH; purple) is shown in the context of the hypothalamus (HY; orange mesh) in the left hemisphere (grey) (figure produced by PyMol, <http://pymol.org>). (b) Distribution of VMH vs HY enrichment score. *In situ* images, the 3D registered expression in the HY (inset), and ALLENMINER enrichment scores are shown for two genes that are VMH enriched: (c) Fezf1 and (d) Ptchd2. The positions of the HY (orange outline) and VMH (blue outline) are depicted on the *in situ* images. The 3D registered expression (inset, yellow circles) in the HY (red) is shown along with the highlighted VMH nuclei (green) (figure produced by BrainExplorer [10]).

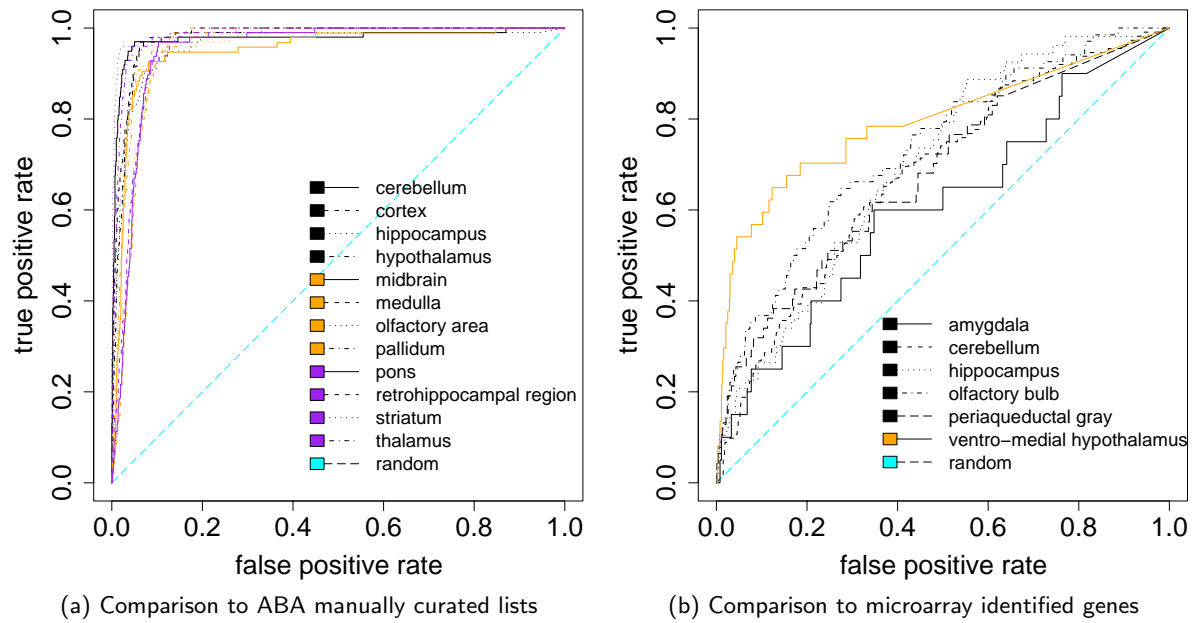
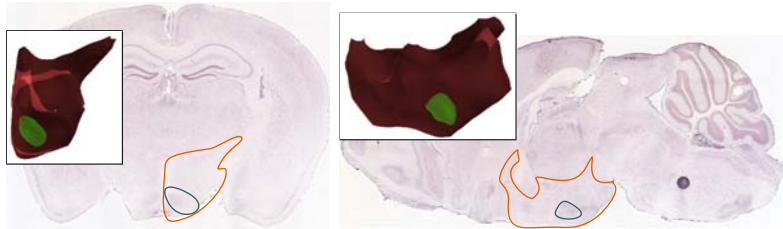
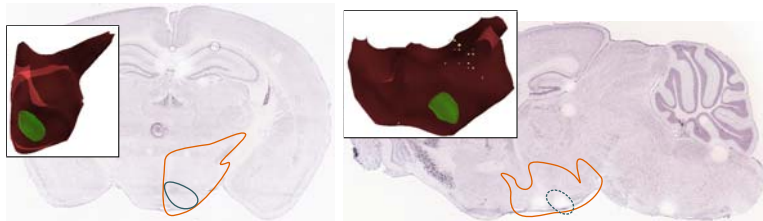


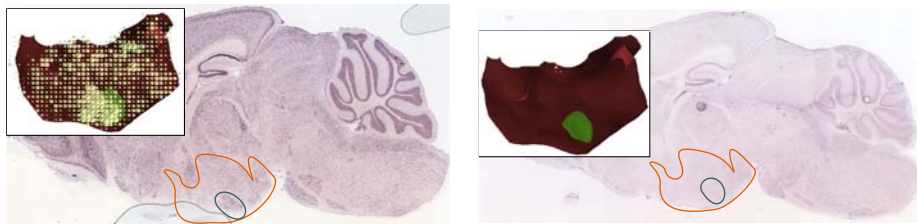
Figure 3: Receiver-operator characteristic of ALLENMINER enrichment score. ALLENMINER results were compared to (a) manually curated lists of genes enriched in 12 ABA brain regions [8] and (b) microarray-identified genes enriched in the cerebellum, amygdala, hippocampus, olfactory bulb, periaqueductal gray [13] and fetal ventromedial hypothalamus [14]. Details of the ROC analysis are presented in Table 1.



(a) G protein-activated inward rectifier potassium channel 4 (Kcnj5; K: 2.22, A:0; ABA images 72040834, 69890146)

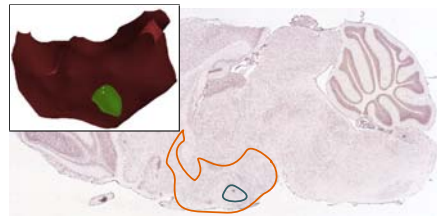


(b) Neuromedin-B receptor (Nmb; K: 3.30, A: 0; ABA images 77328246, 69148481)

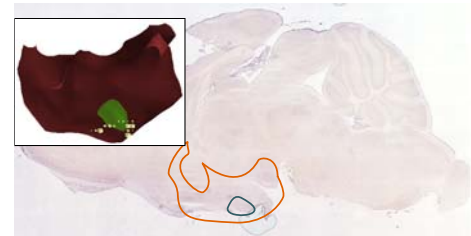


(c) Ataxin-2-binding protein 1 (A2bp1; K:2.4, A:0.95; ABA image 71138095) (d) Transcription cofactor vestigial-like protein 2 (Vgll2; K: 8.42, A: 0; ABA image 71279859)

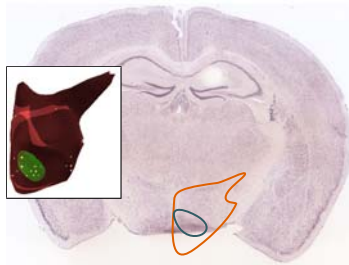
Figure 4: Examples of ALLENMINER false negative errors. Comparison of VMH-enriched genes identified by ALLENMINER to those identified by microarray [14] suggests that ALLENMINER false negatives are caused by: (a) sub-threshold expression, (b) tissue damage, and (c) slice/registration artifacts. Some of the apparent negatives are likely to be true negatives, for example Vgll2 (d) which has been independently shown not to express in the adult VMH [14]. The positions of the hypothalamus (orange outline) and VMH (blue outline) are depicted on the *in situ* images. The 3D registered expression (inset, yellow circles) in the hypothalamus (red) is shown along with the highlighted VMH nuclei (green). The fold enrichment observed by the microarray study (K:, [14]) and the ALLENMINER enrichment score (A:) are listed.



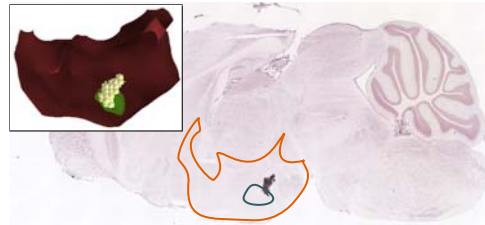
(a) F-box and leucine-rich repeat protein 7 (Fbxl7; A: 37.04; ABA image 73429295)



(b) Guanine nucleotide-binding protein subunit alpha 15 (Gna15; A: 17.31; ABA image 67846082)



(c) Guanine nucleotide-binding protein G(olf) subunit alpha (Gnal; A: 16.09; ABA image 71774899)  
11.88; ABA image 71514991)



(d) Tctex1 domain containing 1 (Tctex1d1; A: 16.09; ABA image 71774899)



(e) Cylindromatosis (Cyld; A: 27.38; ABA image 73968005)

Figure 5: Examples of ALLENMINER false positive errors. Comparison of genes enriched in the ventromedial hypothalamus identified by ALLENMINER to those identified by microarray [14] suggests that ALLENMINER false positives are caused by (a) particulate aggregates, (b) bubble artifacts, (c) tissue damage such as folded tissue, and (d) debris on the slide. In some cases, enrichment is clearly evident on the *in situ* image and may represent a true positive that was missed by the microarray study or that is differentially expressed in the fetal VMH (e). The positions of the hypothalamus (orange outline) and VMH (blue outline) are depicted on the *in situ* images. The 3D registered expression (inset, yellow circles) in the hypothalamus (red) is shown, along with the highlighted VMH nuclei (green). The ALLENMINER enrichment score is listed for each gene (A:).

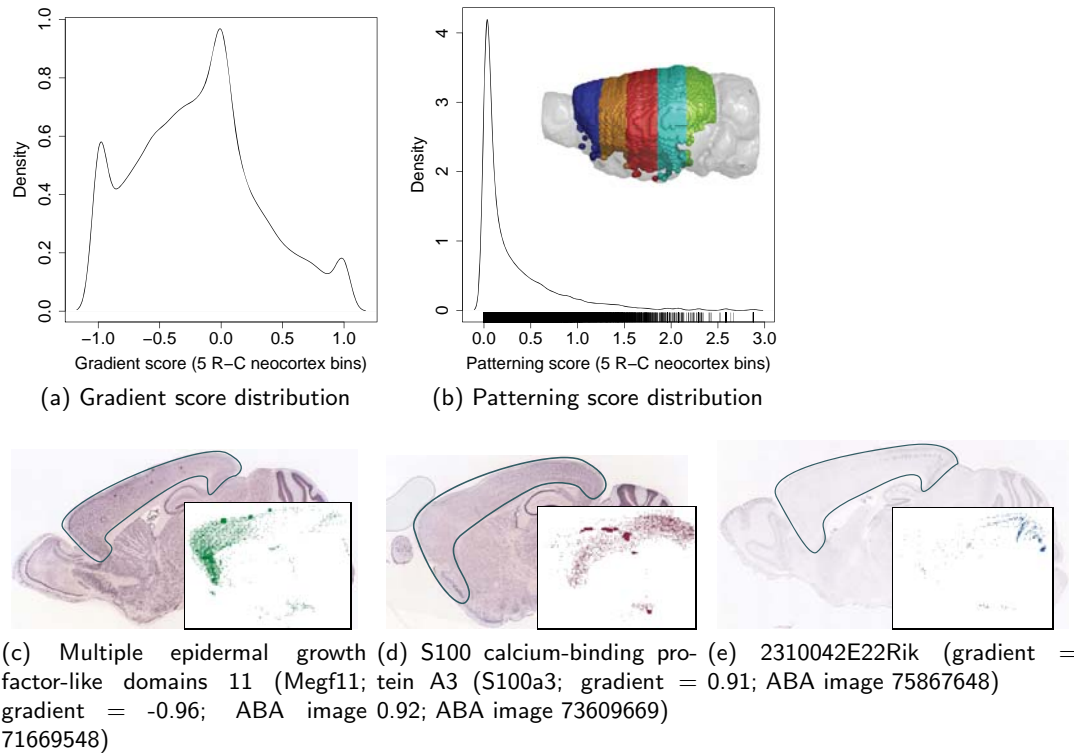


Figure 6: Patterned expression in the neocortex. Gradient (a) and patterning (b) scores were computed across five rostrocaudal neocortex bins (rainbow) for all ABA 3D expression profiles. *In situ* images and 3D neocortex-registered expression (inset) are shown for three genes with graded expression: Megf11 (c) is enriched in the rostral neocortex; S100a3 (d) and 2310042E22Rik (e) are enriched in the caudal neocortex.

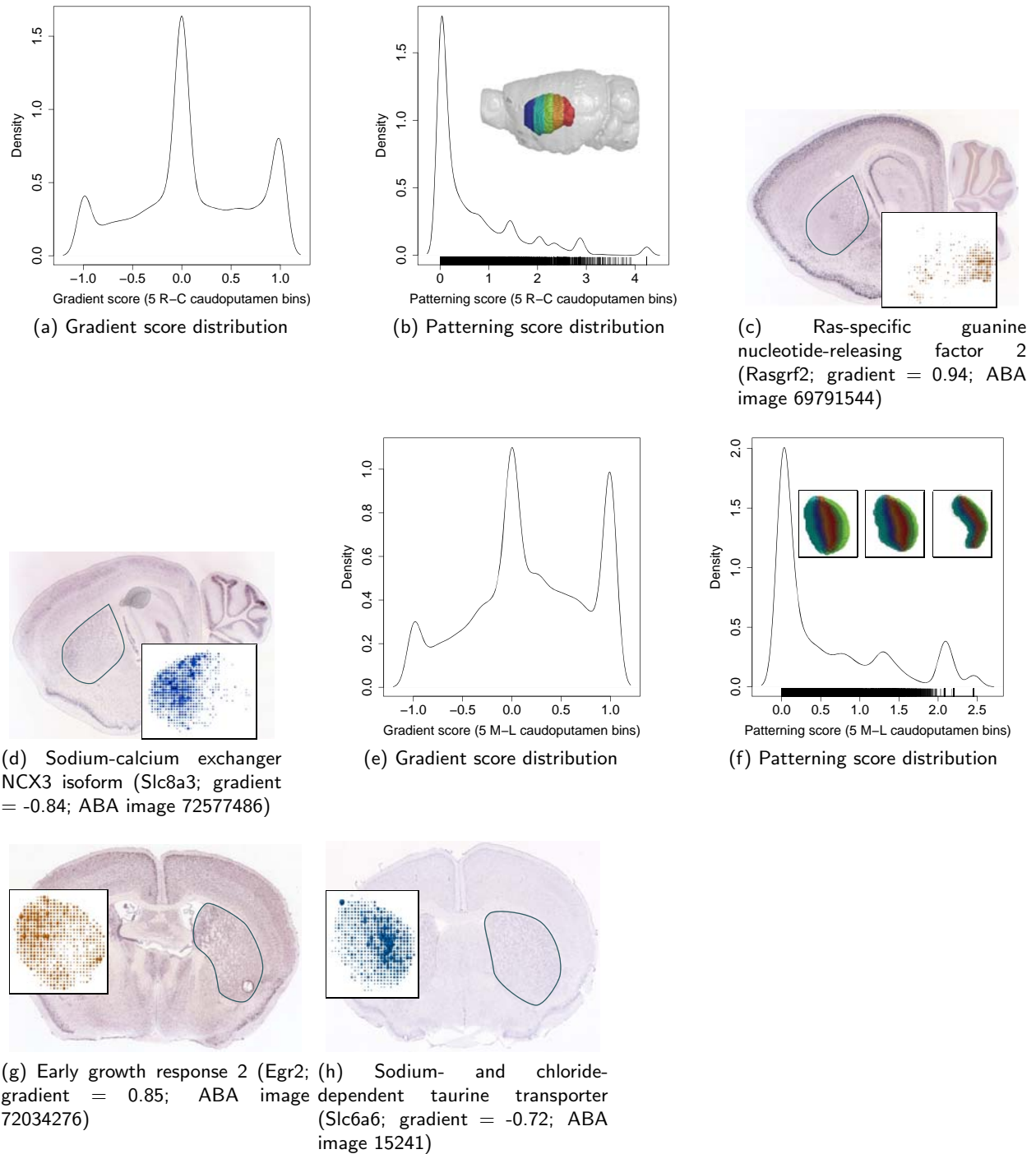


Figure 7: Patterned expression in the caudoputamen. Gradient and patterning scores were computed across five rostrocaudal (a,b) and five mediolateral (e,f) caudoputamen bins (rainbow) for all ABA 3D expression profiles. (f) The mediolateral bins adjust to the varying width of the caudoputamen, as shown by three sections from rostral (left) to caudal (right). *In situ* images and caudoputamen voxel calls (inset) are shown for four genes with graded expression across the rostrocaudal ((c) Rasgrf2, (d) Slc8a3) and mediolateral ((g) Egr2, (h) Slc6a6) axes.

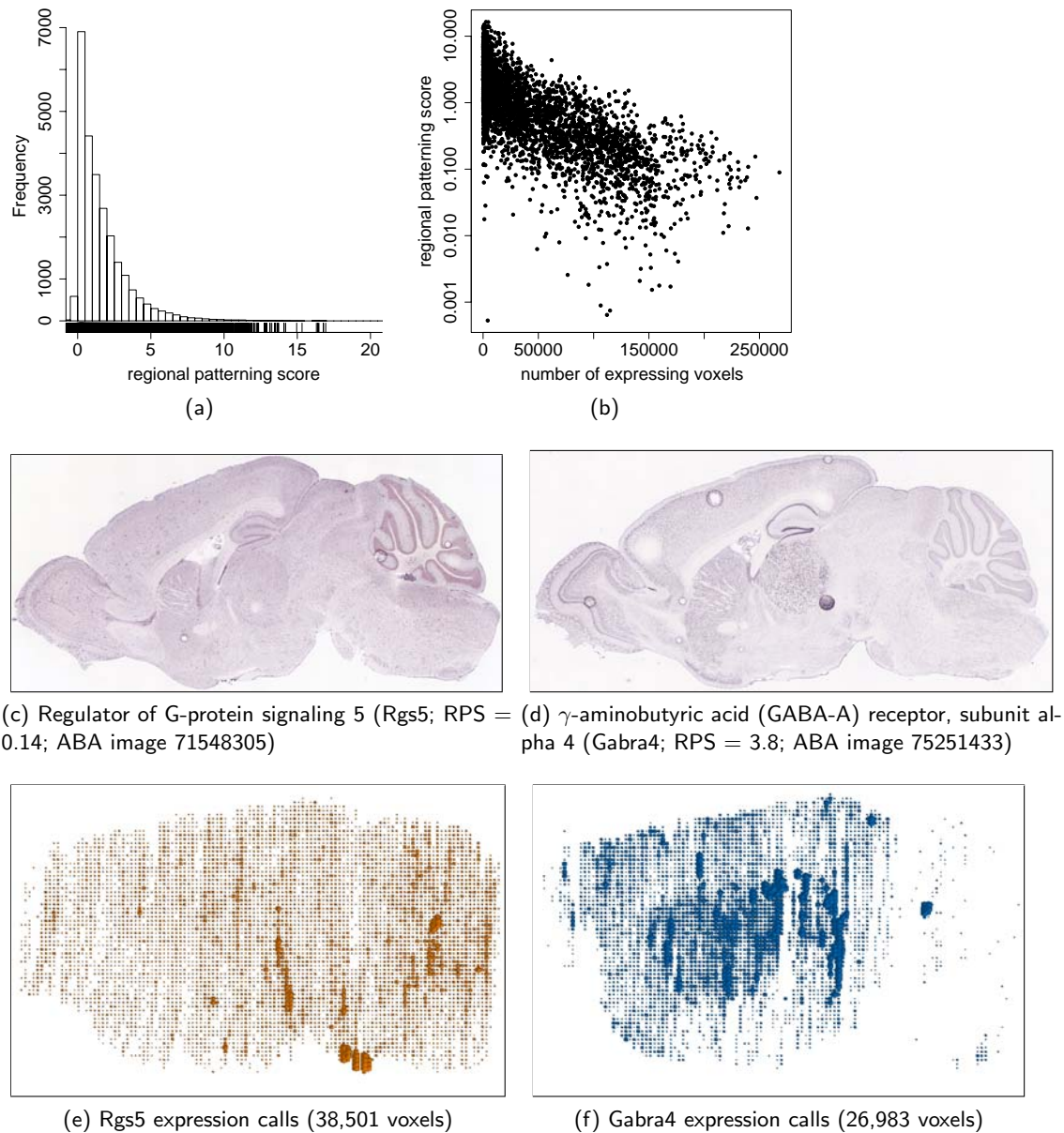


Figure 8: Distribution of gene expression specificity in the mouse brain. (a) Distribution of regional patterning score of all ABA 3D expression profiles (sagittal and coronal). The maximum observed regional patterning score is 39 (not shown). (b) Comparison of the regional patterning score to the number of voxels with detectable expression in the coronal subset of ABA 3D expression profiles. Although they both express in a similar number of voxels, Rgs5 (c,e) is distributed across the brain while Gabra4 (d,f) is more specifically expressed, with enrichment in the thalamus, caudoputamen, and cortex.

**List of Tables**

1 Comparison of ALLENMINER results to published sets of regionally-enriched genes. . . . 27

2 Differences between ALLENMINER and microarray identified genes enriched in the ventromedial hypothalamus . . . . . 28

Benchmark set	num genes	ROC AUC	Score	optimal TPR	FPR
<i>Manually curated lists from the Allen Brain Atlas [8]</i>					
Cerebellum	99	0.977	3.553	0.970	0.050
Cortex	98	0.979	1.715	0.969	0.063
Hippocampus	98	0.961	3.515	0.908	0.072
Hypothalamus	95	0.977	1.530	0.968	0.070
Midbrain	95	0.951	1.352	0.926	0.081
Medulla	95	0.953	2.460	0.958	0.112
Olfactory area	99	0.953	2.489	0.949	0.111
Pallidum	98	0.965	1.299	0.949	0.088
Pons	97	0.951	1.580	0.969	0.103
Retrohippocampal region	99	0.958	1.819	0.929	0.089
Striatum	97	0.991	2.514	0.959	0.019
Thalamus	99	0.982	2.553	0.960	0.045
<i>Published microarray results</i>					
Amygdala [13]	20	0.606	1.478	0.600	0.348
Cerebellum	112	0.679	1.037	0.652	0.369
Hippocampus	53	0.699	1.441	0.660	0.368
Olfactory bulb	68	0.735	1.371	0.647	0.275
Periaqueductal gray	47	0.679	0.609	0.617	0.344
Ventromedial hypothalamus [14]	37	0.792	1.412	0.703	0.185

Table 1: Comparison of ALLENMINER results to published sets of regionally-enriched genes. Receiver-operator characteristic (ROC) analysis quantified the accuracy of the ALLENMINER enrichment score in discriminating regionally enriched genes, as defined by manual curation [8] or microarray analysis [13, 14], from the remaining ABA genes. The “optimal” true positive (TPR) and false positive rates (FPR) for the score threshold closest to the perfect classifier (TPR=1,FPR=0), along with the area under the ROC curve (AUC), are described for each benchmark set.

Gene	Kurrasch	ALLENMINER	
	enrichment	enrichment	rank
1 Gpr149	4.43	20.382	43
2 BC034076	2.27	20.223	45
3 Fezf1	4.38	19.158	54
4 Cckbr	2.60	12.532	145
5 Ddn	4.29	12.174	162
6 Gda	2.81	9.854	227
7 Nkx2-2	2.94	9.744	235
8 Gabra5	2.89	9.388	250
9 Fbxw7	2.62	8.728	280
10 Amigo2	3.13	8.636	287
11 Nr5a1	7.75	7.523	338
12 Pdyn	3.16	6.594	428
13 Col6a3	2.53	6.432	438
14 Bdnf	5.37	5.789	522
15 Cdh4	2.58	5.274	595
16 Ldb2	2.65	5.148	626
17 Nptx2	6.85	5.121	629
18 Grhl1	2.24	4.632	735
19 Calb1	2.21	4.311	819
20 Acvr1c	2.44	3.967	936
21 Adcyap1	3.28	2.742	1,578
22 C030019I05Rik	2.20	2.235	2,082
23 Tmem35	2.20	2.024	2,361
24 Tac1	2.36	1.938	2,499
25 Plp1	3.05	1.624	3,156

(a) ALLENMINER enrichment score greater than 1.5

Table 2: Differences between ALLENMINER and microarray identified [14] genes enriched in the ventromedial hypothalamus. (a) Twenty five of the 37 genes identified by microarray received an ALLENMINER enrichment score of at least 1.5. (b) Visual inspection of *in situ* images for the remaining 12 genes suggested several potential reasons to explain their enrichment scores of less than 1.5.

Gene	Kurrasch enrichment	ALLENMINER enrichment	rank	comments
26 Sema3a	2.52	1.412	3,764	sub-threshold: subtle enrichment on one of three coronal data-sets; few registered voxels
27 Satb2	4.68	0.948	5,786	slice/registration effect: subtle VMH enrichment evident on coronal and sagittal images; no VMH voxels called
28 A2bp1	2.40	0.946	5,796	slice/registration effect: enrichment evident on the <i>in situ</i> image; sagittal slices only catch edge of VMH
29 Titf1	2.27	0.716	6,723	strong expression throughout hypothalamus
30 B3gnt3	2.25	0.000	11,118	no detectable expression
31 B530002L08	6.13	0.000	11,137	no detectable expression
32 Card14	2.32	0.000	11,689	no detectable expression, elongated brain
33 Kcnj5	2.22	0.000	14,603	sub-threshold: slight signal on coronal and sagittal <i>in situ</i> , no voxels registered
34 Mst1r	3.83	0.000	16,453	sub-threshold, <i>in situ</i> artifact: subtle sagittal signal, but close to artifact; no expression called
35 Nmbr	3.30	0.000	16,690	sagittal <i>in situ</i> artifact: missing VMH tissue on sagittal section; coronal sub-threshold: very subtle signal in 1-2 slices
36 Ntn2l	2.21	0.000	16,783	sub-threshold: some low level expression, few called pixels, no registered voxels
37 VglI2	8.42	0.000	19,791	no detectable expression. very low expression on image, few pixels called, no registered voxels

(b) ALLENMINER enrichment score less than 1.5