ALLEN INSTITUTE
for BRAIN SCIENCE
Fueling Discovery

TECHNICAL WHITE PAPER:
INFORMATICS DATA PROCESSING FOR THE ALLEN DEVELOPING MOUSE BRAIN ATLAS

The Allen Developing Mouse Brain Atlas provides *in situ* hybridization (ISH) data for ~2000 genes over embryonic and postnatal timepoints, and many of these genes display very restricted spatial expression patterns that change over time. Events that shape the development of the brain from an undifferentiated set of precursors to a mature, functioning organ occur at different times in different regions, and thus the ability to localize gene expression at specific stages of development is highly desirable. The informatics data processing pipeline developed by the Allen Institute enables the navigation and analysis of this large and complex dataset to identify gene expression with precise spatial and temporal regulation.

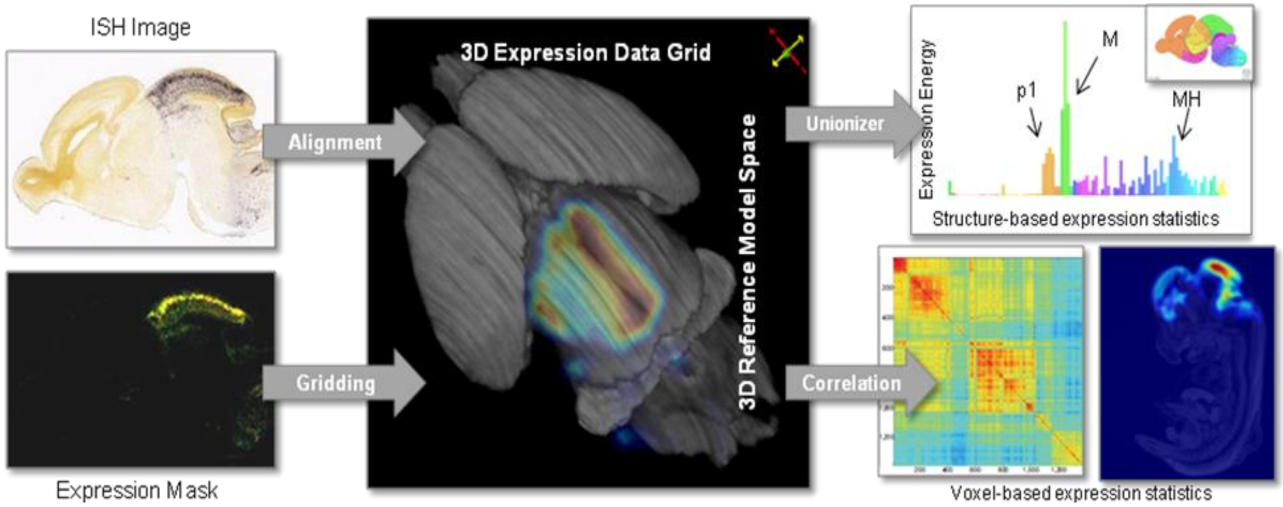


Figure 1. The informatics data processing pipeline. The Alignment module registers each ISH image to the common coordinates of a 3D reference model. The Expression Gridding module produces an expression summary in 3D for downstream analysis. The Structure Unionizer module generates structure-based statistics by combining or "unionizing" grid voxels with the same 3D structural label from the hierarchical reference atlas. Further downstream the grid data is use to compute gene-to-gene correlations and voxel-to-voxel correlations to support NeuroBlast (similarity search) and Developmental AGEA functions. The ISH image shows gene *Tcfap2b* at age E18.5.

March 2010

In particular, the informatics data processing supports the following features in the web application:

1. The **“Expression Summary”** which is a heatmap representation of gene expression for a given gene by age and by atlas structure.

2. A cross-plane and cross-time, point-based “**Synchronize**” feature in the Zoom and Pan (Zap) Image Viewer allows multiple image series to be synced to the same approximate position in the brain based on a linear alignment of the images to a set of 3D reference models. An image series is an indexed set of images spanning a single specimen where sections are treated with the same stain, such as an ISH for a particular gene or a Nissl stain.

3. Visualization of gene expression in a 3D format using “**Brain Explorer 2**”.

4. The “**Anatomic Search**” feature enables users to discover genes that are predominantly enriched within a brain structure at a specific age.

5. The “**Temporal Search”** feature which allows users to search for genes that exhibit higher expression at a particular age for a specific brain region.

6. The “**Developmental AGEA**”, or “Developmental Anatomic Gene Expression Atlas” through which users can explore the spatial and temporal relationships in the developing brain based on gene expression, and search for genes expressed at a given voxel in the brain.

7. The “**Neuroblast**” feature allows the user to search for genes whose expression patterns are highly correlated to the seed gene.

The informatics data processing pipeline consists of the following components: a set of 3D reference models, an Alignment module, an Expression Gridding module, a Structure Unionizer module, an Anatomic Search strategy, a Temporal Search strategy, a Gene-to-Gene Correlation module (to support NeuroBlast similarity search) and a Voxel-to-Voxel Correlation module (to support Developmental AGEA) (Figure 1). These are described below.

3D REFERENCE MODELS

The cornerstone of the automated pipeline is a set of 3D reference models. For each timepoint, a specimen is sectioned to span a nearly complete specimen and the slides are either Nissl or Feulgen-HP yellow stained to form one high density image series. The images are reassembled to form a consistent 3D volume. Structural delineation from the 2D reference atlas images are inserted into the 3D model and interpolated to created 3D structural delineations. The 3D reference spaces are then co-registered and scaled into a common space such that brains of different ages can be roughly compared for the purpose of the “Synchronize” feature.

ALIGNMENT MODULE

The Alignment module operates on a per-specimen basis where all image series from a specimen are combined as one super series. Based on maximization of image correlation, the module interleaves the sections from different gene image series, reconstructing the specimen as a consistent 3D volume with co-registration to the 3D reference model. Once registration is achieved, information from the 3D reference model can be transferred to the reconstructed specimen and vice versa. The resulting transform information is saved in the database to support the image synchronization feature in the Zap viewer. Because reference models for each timepoint are also co-registered, synchronization is possible between specimens of different ages (Figure 2).

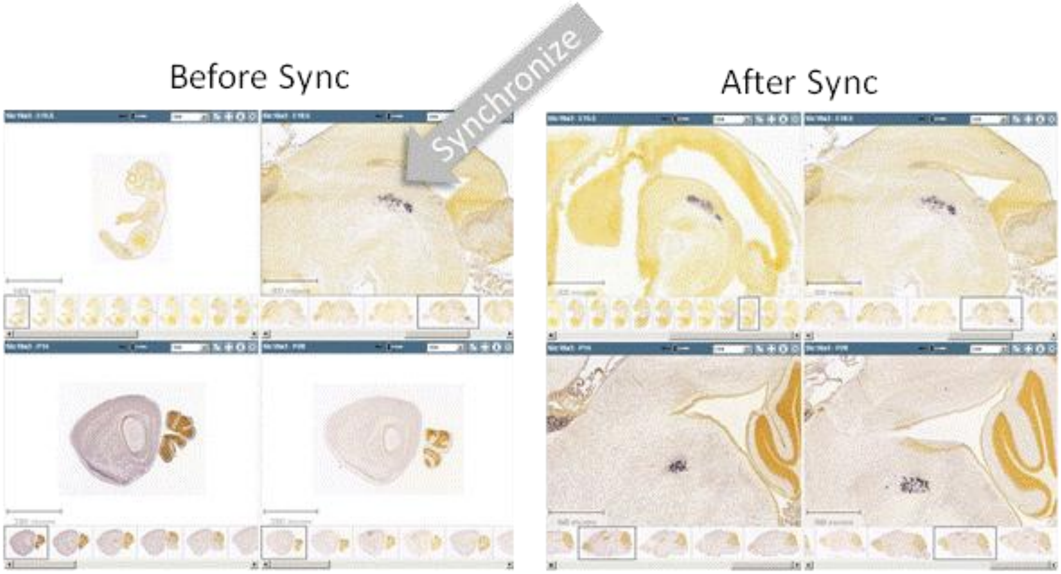
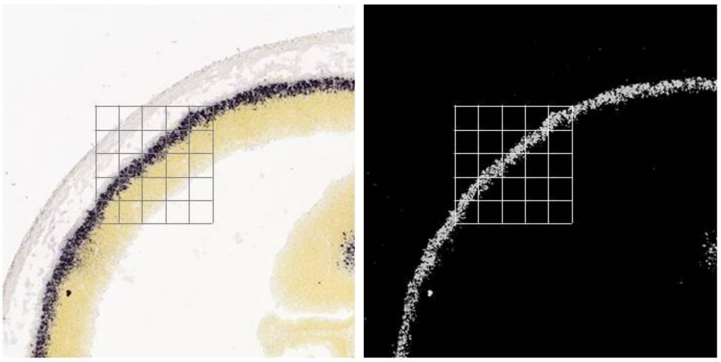


Figure 2. Point-based image synchronization. Multiple image-series in the Zap viewer can be synchronized to same approximate location within and across timepoints. Before and after synchronization screenshots showing gene *Slc18a3* at ages E15.5, E18.5, P14 and P28.

EXPRESSION GRIDDING MODULE

A detection algorithm is applied to each ISH image to create a mask identifying pixels in the high resolution image which corresponds with gene expression. The aim of the Gridding module is to create a low resolution 3D summary of the gene expression and project the data to a common coordinate space of the 3D reference model to enable spatial comparison between data from different specimens. The expression data grids are used for downstream search and analysis, and they can also be viewed directly as 3D volumes in **Brain Explorer 2** (similar to Brain Explorer; Lau et al., 2008)**,** alongside the 3D version of the Allen Developing Mouse Brain Reference Atlas. The resolution of the data grids varies with age and corresponds with the sampling density for that time-point: ranging from 80µm for E11.5 to 200µm for P28 (Figure 3).

For the purpose of search and analysis we are collecting pixel-based statistics of sum and average number of expressing pixels and sum and average expression intensity per grid voxel.



Age	E11.5	E13.5	E15.5	E18.5	P4	P14	P28
3D Grid Size (µm/side)	80	100	120	140	160	200	200

Figure 3. Expression grid sizes per age. Images show an example of a 100 µm grid on an E13.5 embryo. Grid sizes are determined by the interval between sections for ISH image series.

STRUCTURE UNIONIZER (EXPRESSION SUMMARY, ANATOMIC SEARCH AND TEMPORAL SEARCH)

Expression statistics can be computed for each structure delineated in the reference atlas by combining or “unionizing” grid voxels with the same 3D structural label. ***Expression energy for brain region R*** is defined as **the sum of expressing pixel intensities in R divided by the total number of pixels that intersects R**. While the reference atlas is annotated at ontological Level 08, statistics at lower levels (Levels 00 to 05) can be obtained by further combining measurements of the hierarchical children to obtain a statistics for the “parent” structure. The computed structure-based expression statistics are then displayed as an expression summary on the imageseries or gene page (Figure 4), and used downstream to enable Anatomic and Temporal Search.

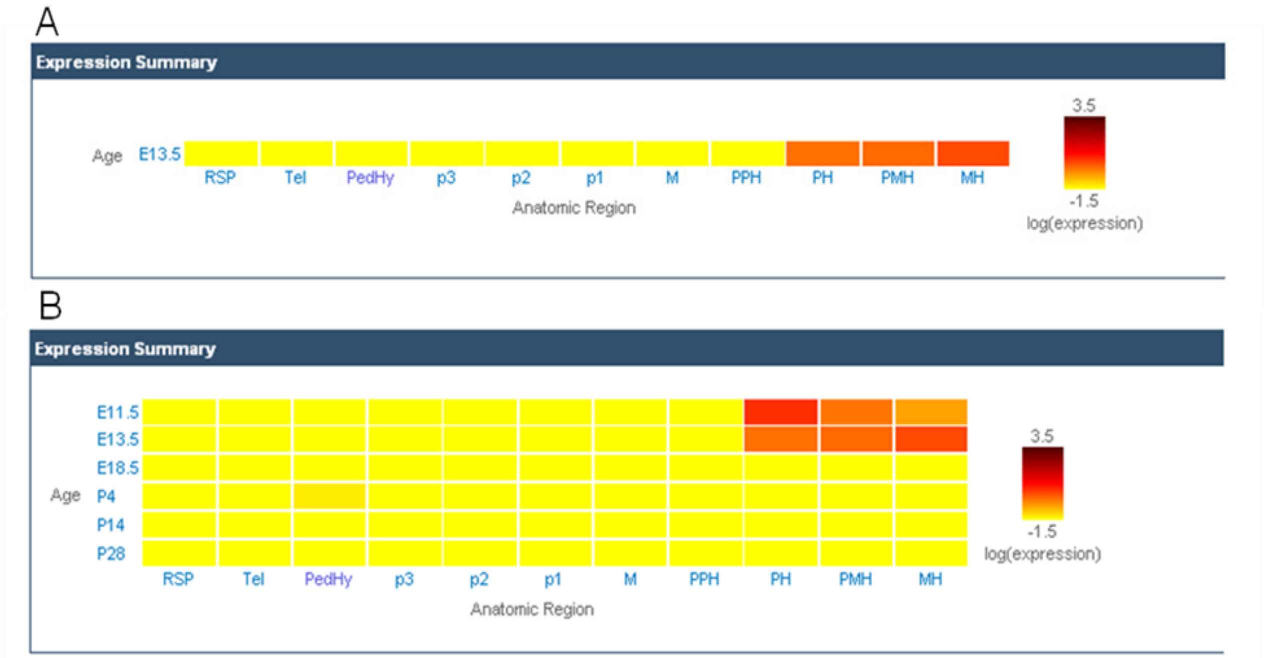


Figure 4. The expression summary provides an overview of gene expression by anatomic structure for an experiment or a gene. The expression summary is accessible either on the imageseries page for an individual experiment (**A**), or on the gene summary page with data across all available ages (**B**). Low expression is shown as yellow with high expression shown as red. The key to expression strength is shown to the right.

Anatomic Search

The goal of Anatomic Search is to enable users to discover genes that are predominantly enriched within a particular brain region, with results provided for a specific developmental age. Our approach is to define an enrichment measure that will permit the ranking of different genes for their specificity in the brain structure of interest as compared to a “contrast” brain region (Figure 5).

Specifically, we define a set of non-overlapping brain structures as the “numerator” set. Typically, the “numerator” set will simply be the brain structure of interest. The flexibility to incorporate other areas to the numerator is useful for example for the E11.5 embryos where in many areas the brain is just a thin wall surrounding large ventricles. Slight misalignment may cause the expression to be excluded from a structure; the inclusion of adjacent ventricle areas (indicated here with a “v_” prefix) in the numerator may help to mitigate alignment errors.

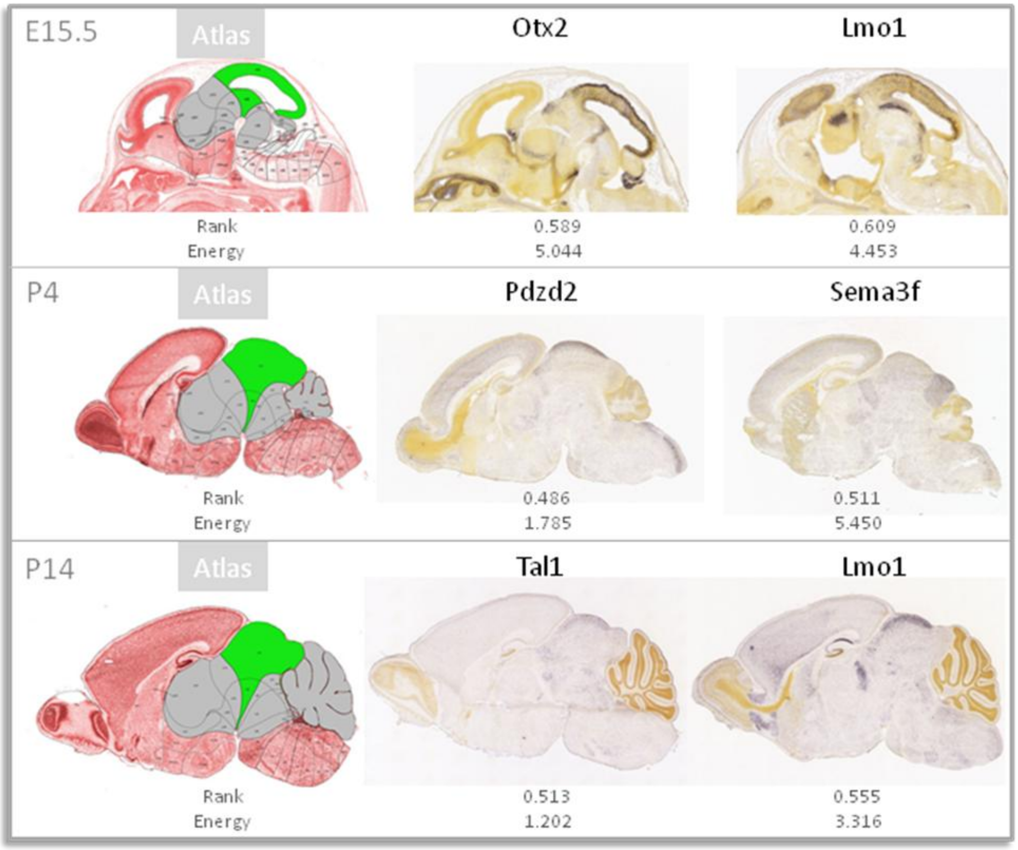


Figure 5. Example of Anatomic Search. Genes are ranked by specificity to midbrain by ratio of expression in the midbrain (green in atlas image) over expression in denominator set of diencephalon, midbrain and prepontine hindbrain (green plus gray in atlas image). Atlas schematics (left column) shown are false-colored images. Gene names are shown at the top of each ISH image (two right columns).

For each search, we also define a set of non-overlapping brain structures as the “denominator” set. Many genes exhibit region-specific enrichment, albeit in multiple areas; thus using the whole brain (or neural plate) as a “contrast” brain region does not necessarily provide a full list of genes, due to local specificity of the gene in multiple regions of the brain. In order to identify genes with local specificity to an anatomic region, a “contrast” region is used as the denominator to determine local specificity. The spatial span of the “numerator” set must be within the spatial span of the “denominator” set, or more simply, the denominator region is inclusive of the numerator. For each gene, we compute the specificity ***rank*** defined as **the ratio of sum of expressing pixel intensities in the numerator set over the sum of expressing pixel intensities in the denominator set**. Theoretically, rank can range from 0 (no expression in the numerator) to 1 (ideal specificity to the numerator). Genes are sorted in descending rank order to generate the Anatomic Search return lists. The maximum observed rank varies per structure and age.

To minimize false positive due to artifacts, genes with expression in the numerator below a specified expression energy threshold are excluded from the return list. Each of the search returns is then verified in a quality control step, and search returns resulting from artifacts are removed from the search. Table 1 lists the numerator, denominator and expression energy threshold for the anatomic searches available via the web application.

Table 1. Calculations used to identify genes expressed in 12 brain regions.

Age	Structure	Numerator	Denominator	Energy Threshold
E13.5, E15.5, E18.5, P4, P14, P28	Tel	Tel	F	1
E13.5, E15.5, E18.5, P4, P14, P28	PedHy	PedHy	p3, PedHy, RSP	1
E13.5, E15.5, E18.5, P4, P14, P28	CSPall	CSPall	Tel	1
E13.5, E15.5, E18.5, P4, P14, P28	DPall	DPall	Tel	1
E13.5, E15.5, E18.5, P4, P14, P28	MPall	MPall	Tel	1
E13.5, E15.5, E18.5, P4, P14, P28	p3	p3	p3, p2	1
E13.5, E15.5, E18.5, P4, P14, P28	p2	p2	F	1
E13.5, E15.5, E18.5, P4, P14, P28	p1	p1	p2, p1, M	1
E13.5, E15.5, E18.5, P4, P14, P28	M	M	D, M, PPH	1
E13.5, E15.5, E18.5, P4, P14, P28	PPH	PPH	M, PPH, PH	1
E13.5, E15.5, E18.5, P4, P14, P28	PH	PH	PPH, PH, PMH	1
E13.5, E15.5, E18.5, P4, P14, P28	PMH	PMH	PH, PMH, MH	1
E13.5, E15.5, E18.5, P4, P14, P28	MH	MH	PMH, MH, SpC	1
E11.5	Tel	Tel, v_Tel	RSP, Tel, p3, v_RSP, v_Tel, v_p3	1
E11.5	PedHy	PedHy, v_PedHy	p3, PedHy, RSP, v_p3, v_PedHy, v_RSP	1
E11.5	D	D, v_D	SP, v_SP, D, v_D, M, v_M	1
E11.5	M	M, v_M	p1, M, PPH, v_p1, v_M, v_PPH	1
E11.5	H	H, v_H	NP, ventricles	1

Abbreviations: CSPall, central subpallium (striatum/pallidum); D, diencephalon; DPall, dorsal pallium (isocortex and entorhinal cortex); F, forebrain; H, hindbrain; M, midbrain; MH, medullary hindbrain (medulla); MPall, medial pallium (hippocampus, taenia tecta, subiculum); NP, neural plate; p1, prosomere 1 (pretectum); p2, prosomere 2 (thalamus); p3, prosomere 3 (prethalamus); PedHy, peduncular hypothalamus; PH, pontine hindbrain (pons proper); PMH, pontomedullary hindbrain; PPH, prepontine hindbrain; RSP, rostral secondary prosencephalon; SP, secondary prosencephalon; SpC, spinal cord; Tel, telencephalic vesicle.

Temporal Search

The goal of Temporal Search is to allow users to search for genes that exhibit higher expression at a particular age, with results returned for a specific brain region. Note that while the temporal search provides results for a particular anatomic region, the results are provided regardless of the anatomic specificity of the gene expression. For each gene and brain region of interest R, a simple ranking metric is computed at age A where rank is defined as **the ratio of the expression energy of R at age A over the sum of expression energy of R over all ages** (the seven standard timepoints). Theoretically rank can range from 0 (no expression at age A) to 1 (ideal specificity to the age A). For each age, genes are sorted in descending rank order to generate the Temporal Search return lists. Temporal search lists are provided for Tel, D, M and H for all seven standard ages.

To minimize false positive due to artifacts, genes with expression energy below the specified threshold or genes with “widespread” expression are excluded from the return list. In order to identify and remove “widespread” genes, a metric based on the coefficient of variation (standard deviation/mean) of the expression energy of voxels spanning the whole brain is used. The threshold for removing “widespread” genes was determined for each individual timepoint by manual assessment of search returns.

The Temporal Search results provided on the website are manually verified in a quality control process by a team of data analysts.

NEUROBLAST

NeuroBlast is a search tool to help identify genes with similar 3D spatial gene expression profiles. While searching for genes using the conventional “anatomic search” strategy is a natural approach to identify genes of interest expressed in a particular region, greater search power may sometimes be obtained by starting with a particular expression pattern and inquiring whether there exist other genes with a similar pattern of expression. For example, in order to identify genes expressed in a particular cell type which is distributed throughout the brain (e.g., astrocytes, oligodendrocytes), a region-based approach may not be useful. Instead, one might use a gene which is a canonical cell type marker to initiate a correlational search to identify genes with a similar expression pattern; the results may be enriched in genes also expressed in the desired cell type.

To support NeuroBlast, Pearson's correlation coefficient was computed for each pair of image series (at the same age) using the 3D expression data grids. In particular, correlation was computed using expression energy of each voxel and over five region of interest: NP, Tel, D, M and H. For database efficiency, the top 250 most similar image series are archived and presented on the web application. NeuroBlast is accessible on the gene search results page to the right of each imageseries.

DEVELOPMENTAL AGEA

The Developmental Anatomic Gene Expression Atlas (Developmental AGEA) is a new relational atlas that allows users to explore the spatiotemporal relationships in the developing mouse brain based on the expression patterns of ~2000 genes. Similar to the AGEA for the adult mouse brain (Ng et al., 2009), Developmental AGEA is based on interactive visualization of 3D correlation maps rendered as false color images. The value at a spatial location (voxel) of a map represents the Pearson's correlation coefficient (cc) of the voxel with respect to a "seed" voxel. Correlation is computed over a "gene vector" whose elements represent the expression energy for a gene at the voxel of interest. 3D correlation maps are generated for each possible seed voxel (265,621 in total over 7 ages).

Figure 6 illustrates the construction of an intra-age (within the same timepoint) and a inter-age (across two timepoints) correlation map. In the figure, the red cross within the isocortex in the P4 Nissl image represents the "seed" voxel. The "gene vector" at this location is correlated with the corresponding values at two other intra-age locations (light blue and dark blue crosses). Scatter plots of the expression energy over ~2000 genes between the "seed" and "target" voxels shows that the cortical target (dark blue) is more correlated (cc=0.94) to the seed location than the striatal target (light blue) (cc= 0.74). For the chosen seed, correlation is computed at every other voxel in the P4 brain and visualized simultaneously as a false color map that can be thresholded for significance by the user. Cool colors represent lower correlation values while warmer colors represent higher correlation values. The P4 correlation map show strong correlation between the seed and cortical areas and lower correlation with subcortical areas.

A similar correlation map can be generated between the "seed" voxel (at P4) and the voxels in the P14 mouse brain. Scatter plots of the expression energy between the "seed" and the inter-age targets also shows that the cortical target (orange cross) is more correlated (cc=0.69) to the seed location that the striatal target (yellow cross) (cc=0.54). As before, a complete 3D map can be generated by computing the correlation at every P14 voxel with respect to the "seed" voxel. The P14 correlation map shows a similar stronger correlation between the seed and cortical areas and lower correlation with subcortical areas.

Correlation values in intra-age maps are typically of higher value than those in inter-age maps. In an intra-age correlation computation, corresponding elements in the gene vector are derived from the same ISH experiment (image-series) while in an inter-age computation, corresponding elements are necessarily derived from different experiments from specimens of different ages. The semi-quantitative nature of ISH, inter-experiment variability and natural developmental differences results in lower correlation values in inter-age maps. Typically inter-age maps should be interpreted with respect to relative correlation within the brain of the "map" age and not absolute comparisons between ages.

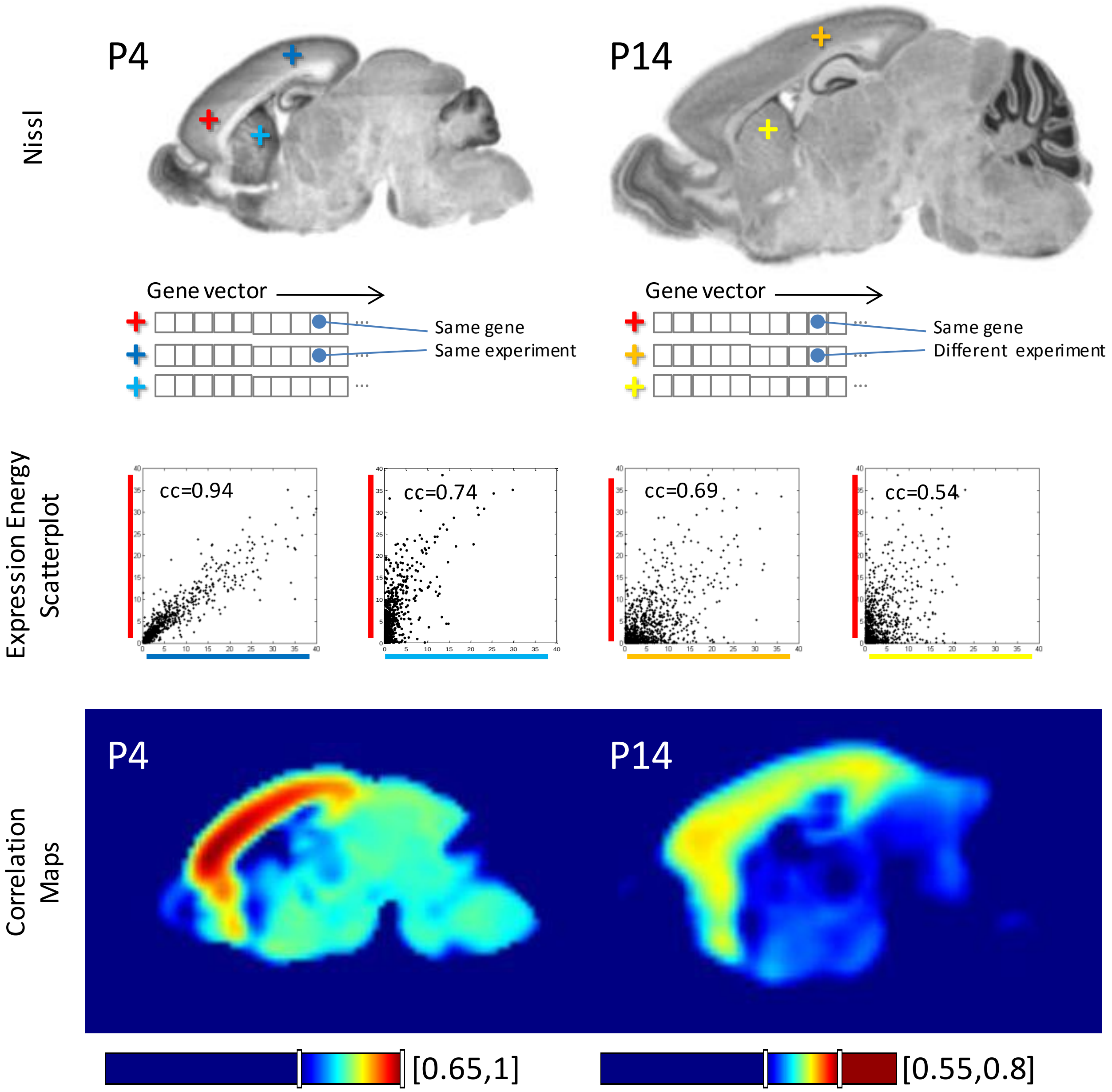


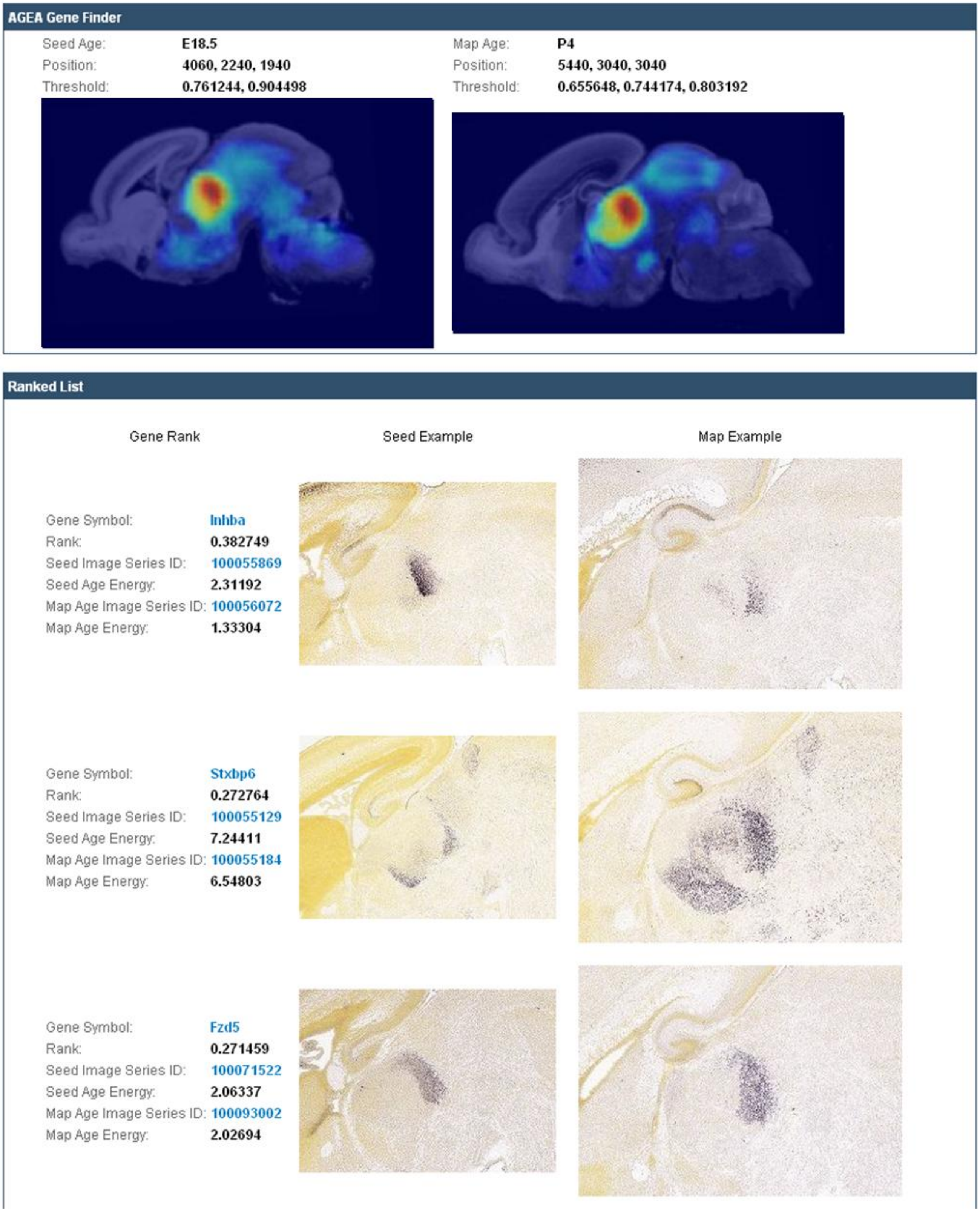
Figure 6. Construction of a intra-age(P4) and inter-age (P4-P14) AGEA correlation map for a seed voxel in the isocortex.

Gene Finder

The Gene Finder function of the Developmental AGEA uses the spatial correlation maps to generate a search "space" to find genes enriched in the correlation region surrounding a seed voxel. First, for each correlation map, voxels are assigned to numerator and denominator spaces, similar to those defined for Anatomic Search. For each correlation map, let t_m be the maximum correlation value. A "denominator" threshold (t_d) is determined such that the number of voxels greater than t_d spans approximately 1/3 of the brain. A "numerator" threshold (t_n) is defined as ($0.6 * t_m + 0.4 * t_d$). All voxels greater than t_d forms the "denominator" space and all voxels greater than t_n forms the "numerator" space. Note that by definition, the "denominator" space is inclusive of the "numerator" space. For each gene, a specificity rank is defined as the sum of expression energy in the numerator space over the sum of expression energy in the denominator space.

Genes are sorted in descending rank order and the top 100 for each map is archived and presented on the web application. For an inter-age search, the rank between the “seed” age map and “map” age map is averaged to produce a combined rank.

Figure 7 shows an example of a Gene Finder return page for an inter-age query. In this example, the seed voxel is in the diencephalon in the E18.5 brain. The intra-age (E18.5) and inter-age (E18.5-P4) correlation maps has been thresholded with the lower bound set to “denominator” threshold of each map and the upper bound set to the maximum correlation of the map. In the maps the warm-colored areas roughly corresponds with the “numerator” mask with the cool areas being the contrast region for search. Each row in the table corresponds to a gene in rank order. The region of interest in the zoomed in thumbnail was computed using the transform obtained in the Alignment module centered at the corresponding position to the maximum correlation location.



AGEA Gene Finder

Seed Age:	E18.5	Map Age:	P4
Position:	4060, 2240, 1940	Position:	5440, 3040, 3040
Threshold:	0.761244, 0.904498	Threshold:	0.655648, 0.744174, 0.803192

Ranked List

Gene Symbol:	Inhba
Rank:	0.382749
Seed Image Series ID:	100055869
Seed Age Energy:	2.31192
Map Age Image Series ID:	100056072
Map Age Energy:	1.33304

Gene Symbol:	Stxbp6
Rank:	0.272764
Seed Image Series ID:	100055129
Seed Age Energy:	7.24411
Map Age Image Series ID:	100055184
Map Age Energy:	6.54803

Gene Symbol:	Fzd5
Rank:	0.271459
Seed Image Series ID:	100071522
Seed Age Energy:	2.06337
Map Age Image Series ID:	100093002
Map Age Energy:	2.02694

Figure 7. An example of an inter-age Gene Finder return list for a seed voxel in the diencephalon in the E18.5 brain.

REFERENCES

Ng L, Bernard A, Lau C, Overly CC, Dong HW, Kuan C, Pathak S, Sunkin SM, Dang C, Bohland JW, Bokil H, Mitra PP, Puelles L, Hohmann J, Anderson DJ, Lein ES, Jones AR, Hawrylycz M (2009) An anatomic gene expression atlas of the adult mouse brain. *Nat Neurosci* 12: 356-62.

Lau C, Ng L, Thompson C, Pathak S, Kuan L, Jones A, Hawrylycz M (2008) Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics* 9: 153.