ALLEN INSTITUTE for BRAIN SCIENCE
Fueling Discovery

ALLEN **Mouse Brain Atlas**

TECHNICAL WHITE PAPER: INFORMATICS DATA PROCESSING

OVERVIEW

The Allen Mouse Brain Atlas provides genome-wide *in situ* hybridization (ISH) data for approximately 20,000 genes in 56 day old male "black" mice. The informatics data processing pipeline developed by the Allen Institute (Ng et al, 2007, Dang et al, 2007) enables the navigation, analysis and visualization of this large and complex dataset to identify gene expression of interest and relationships between genes and between spatial regions.

The informatics data processing pipeline consists of the following components: a Preprocessing module, a 3-D reference model, an Alignment module, an Expression Detection module, an Expression Gridding module and a Structure Unionizer module. The output of the pipeline is quantified expression values at a grid "voxel" level and a structure level according to the Allen Reference Atlas ontology. The grid level data are used downstream to provide an on-the-fly differential and correlative gene search service and to support visualization of spatial relationships.

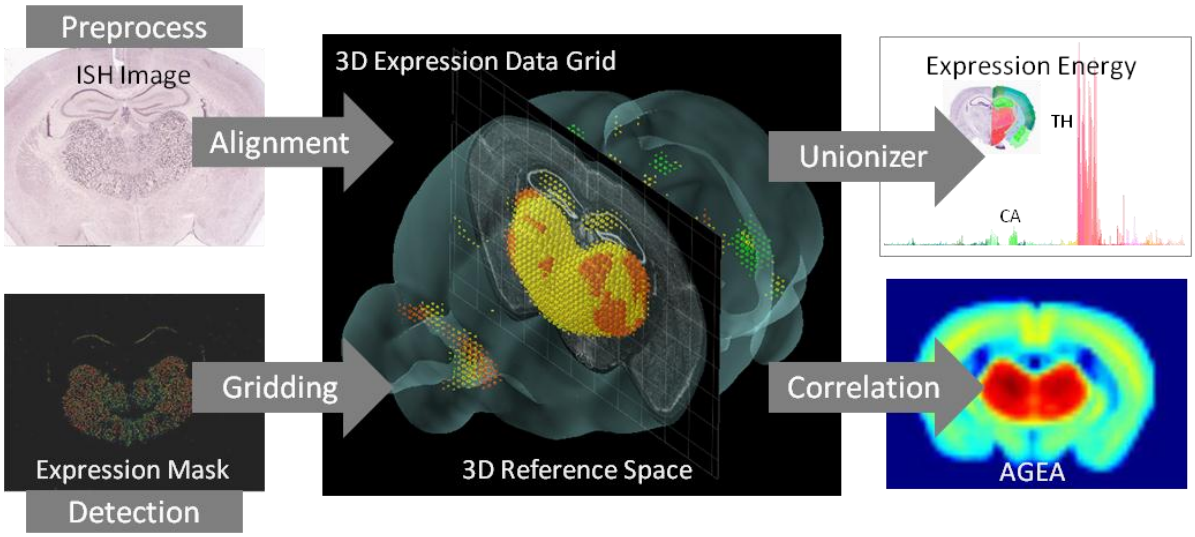

Figure 1. The informatics data processing pipeline.
The Alignment module registers each ISH image to the common coordinates of a 3-D reference model. The Expression Gridding module produces an expression summary in 3-D for downstream analysis. The Structure Unionizer module generates structure-based statistics by combining or "unionizing" grid voxels with the same 3-D structural label from the hierarchical reference atlas. Further downstream, the grid data are used to compute gene-to-gene correlations and voxel-to-voxel correlations to support NeuroBlast (similarity search) and AGEA functions. The ISH image shows the gene *Plekhg1*.

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

1. An "**Expression Mask**" for each ISH image showing areas of detected expression with color-coding of the expression level.
2. An "**Expression Summary**" for each image series as a bar plot representation of gene expression over coarse level spatial regions.
3. A cross-plane, 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 3-D reference model. 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.
4. Visualization of gene expression in a 3-D format as pre-rendered views and in the "**Brain Explorer® 2**" desktop software application.
5. The "**Differential Search**" feature enables users to discover genes that are relatively enriched in one brain structure when compared to another structure.
6. The "**NeuroBlast**" feature allows the user to search for genes whose expression patterns are highly correlated to a seed gene.
7. The "**AGEA**", or "Anatomic Gene Expression Atlas" through which users can explore spatial relationships in the brain based on gene expression, and search for genes locally enriched at a given voxel in the brain.

PREPROCESSING MODULE

Scanned image tiles are first stitched to form a single large high-resolution image. The first step in preprocessing is to white balance and intensity normalize the image for better display. To identify individual tissue sections on an image, a global adaptive thresholding method is applied to obtain a rough separation of the background and foreground. This step is followed by morphological filtering and connected component analysis to remove noise and connect broken segments. The output of this step is a bounding box for each tissue section which is used to track each the individual tissue section in the database.

A rigorous manual QC protocol has also been established which includes decisions on failing an experiment due to production issues, discarding damaged images, verifying and adjusting the tissue bounding boxes and identification of "dark" artifacts such as bubbles and tears.

3-D REFERENCE MODELS

The cornerstone of the automated pipeline is an annotated 3-D reference space. For this purpose, we based our reference space on the same specimen used for the coronal plates of the Allen Reference Atlas. The brain was sectioned to span a nearly complete specimen resulting in 528 sections, each 25 µm thick.

A brain volume was reconstructed from the section images based on a combination of high-frequency section-to-section histology registration with low-frequency histology to (ex-cranio) MRI registration (Yushkevich, 2006). This first-stage reconstructed volume was then aligned with a sagittally sectioned specimen. Once a straight mid-sagittal plane was achieved, a synthetic symmetric space was created by reflecting one hemisphere (the annotated side) to the other side of the volume.

In 2011, the reprocessing of the Allen Reference Atlas enabled the extraction of over 800 structures from the 2-D annotation, which were then interpolated to create 3-D annotations. The end result is an updated symmetric, fully annotated reference space with a more consistent and deeper level of annotation. (See Appendix for further details.)

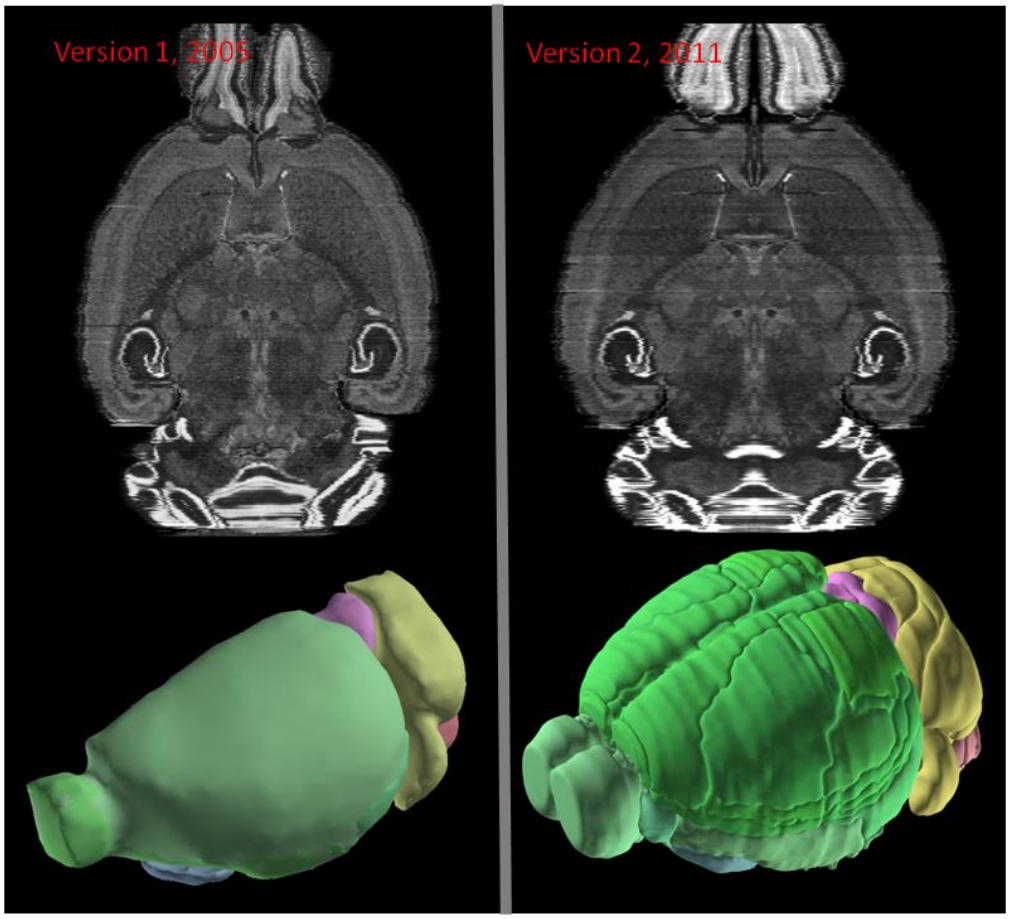

Figure 2. Updated 3-D reference space.
The new reference space is symmetric and annotated on both hemispheres with a more consistent and deeper level of 3-D annotation. The revised 3-D space addresses issues of an error in scaling and a "twist" in the reconstruction.

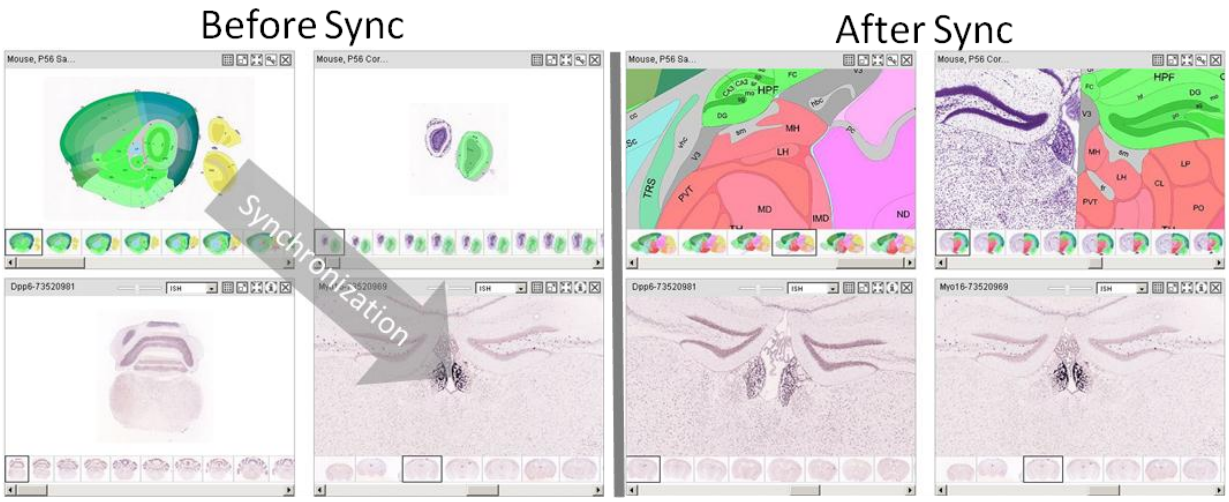

Figure 3. Point-based image synchronization.
Multiple image-series in the Zoom-and-Pan (Zap) viewer can be synchronized to the same approximate location both in sagittal and coronal planes. Before and after synchronization screenshots showing genes *Dpp6* and *Myo16* and the relevant coronal and sagittal plates of the Allen Reference Atlas. Gene *Myo16* shows enriched expression in the medial habenula (MH).

ALIGNMENT MODULE

The Alignment module operates on a per-specimen basis where all image series from a specimen are combined as one series. Based on maximization of image correlation, the module interleaves reconstructing the specimen as a consistent 3-D volume with co-registration to the 3-D reference model. Once registration is achieved, information from the 3-D reference model can be transferred to the reconstructed specimen and vice versa. The resulting transform information (a 2-D affine transform per image and 3-D affine transform per image-series) is saved in the database to support the image synchronization feature in the Zap viewer and generation of grid-level gene expression summaries.

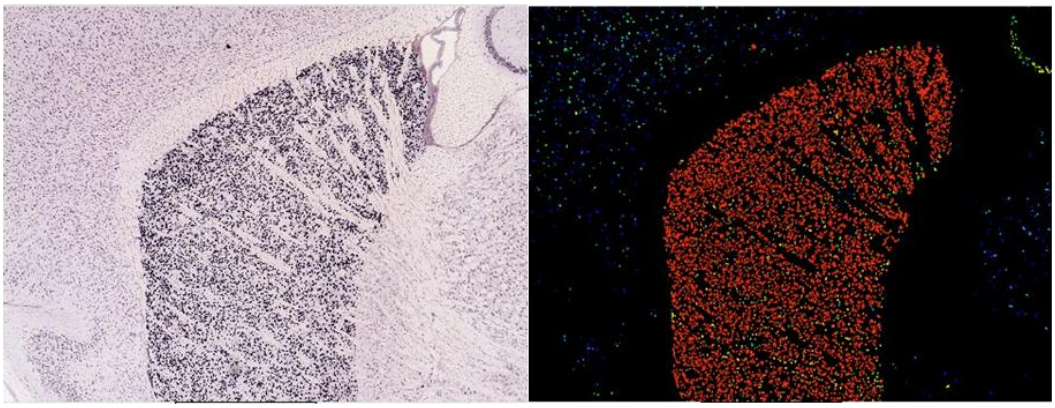

Figure 4. Expression detection for gene Pde10a.
Screenshot of expression detection mask for Pde10a showing dense high expression in the striatum and low expression in the isocortex.

EXPRESSION DETECTION MODULE

A detection algorithm is applied to each ISH image to create a grayscale mask identifying pixels in the high-resolution image that correspond with gene expression. The grayscale intensity represents the average ISH signal within a connected area. For Web presentation, the intensity is color-coded to range from blue (low expression intensity), through green (medium intensity) to red (high intensity).

There are three stages in the expression detection algorithm: 1) tissue region segmentation, 2) small or isolated object detection, and 3) dense cell area object segmentation. The segmentation from each stage is combined into a single output mask (Ng, 2007).

Tissue area segmentation

A tissue area mask is produced using adaptive thresholding and morphology operations in combination with connected component analysis and classification by shape and size of segmented objects. A rule-based system is applied to 1:8 down-sampled images to reject non-tissue objects such as air bubbles and other artifacts in the images. This system includes heuristics such as the expected detection of at most one large object and identification of unlikely smaller tissue fragments in each image. The algorithm is designed to be robust in dealing with the wide range of image variation typically seen in ISH images.

Small and isolated object segmentation

Morphological kernel based spatial filtering is performed on the original resolution images to preferentially segment neural cell-shaped objects of interest (OOI) approximately the diameter of a neuronal cell body (i.e., 10-30 µm diameter) while reducing the non-uniform background. A 31x31 binominal filter was used for signal enhancement over background, followed by an adaptive threshold method that chooses a maximum inter-class variance to separate signal from background. The masked tissue area is then combined with the previous filtered result and an additional edge-enhanced image to create a signal-enhanced image capturing those OOI of lower image contrast but sufficient intensity strength. Statistical characteristics based on object

properties such as size (area), shape (compactness, aspect ratio), intensity (raw and filtered pixel values and integrated contrast value), and spatial information (image coordinate and mean distance) are finally used to reduce artifacts.

Dense and clumped object segmentation

Dense cell regions such as those found in the hippocampus and olfactory bulb contain expressing tissue that is difficult to separate into individual cells. This difficulty can be ameliorated by certain fluorescent methods but remain challenging in colorimetric ISH. In the present algorithm, the detection of these objects is performed in a lower-resolution level of the image pyramid in order to best recognize their essential shape and avoid confusion with artifacts. Local object edge and contrast information can be used to isolate these structures from the non-uniform tissue background.

Figure 5. Screenshot of experiment details page for gene Pdyn.

(A) Rotating thumbnail of expression grid rendered by maximum density projection. (B) Expression summary over 12 coarse-level structures. (C) Image preview "syncs" with bar graph by using computed transform to display corresponding position to the structure's 3-D center.

EXPRESSION GRIDDING MODULE

The aim of the Gridding module is to create a low-resolution 3-D summary of the gene expression and project the data to the common coordinate space of the 3-D 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 3-D volumes in the Brain Explorer 2 3-D viewer (similar to the Brain Explorer viewer; Lau et al., 2008)**,** alongside the 3-D version of the Allen Reference Atlas.

The Gridding module operates on a per image-series basis. Each image is divided into a 200 µm x 200 µm grid. For each division, we collect pixel-based statistics of the sum of the number of expressing pixels and sum of expressing pixel intensity. From these statistics we obtain measures for:

- *expression density* = sum of expressing pixels / sum of all pixels in division
- *expression intensity* = sum of expressing pixel intensity / sum of expressing pixels
- *expression energy* = sum of expressing pixel intensity / sum of all pixels in division

In the previous step, the Alignment module computes the transforms that rotates each 2-D image to form a consistent 3-D volume per specimen. Each per-image 2-D expression grid is smoothed and rotated to form a 3-D grid. Finally, z-direction smoothing is applied to the 3-D grid which is then transformed into the standard reference space.

The expression data grid can be viewed in the Brain Explorer 2 desktop program where each 200 µm per side grid voxel is rendered as a sphere where the diameter of the sphere represents expression energy. The color of the sphere in expression level mode represents expression intensity with the same color-coding scheme as used in the expression mask presentation. Additionally, a preview of the expression data grid is shown on the web application as a maximum density projection rotating thumbnail.

STRUCTURE UNIONIZER MODULE

Expression statistics can be computed for each structure delineated in the reference atlas by combining or "unionizing" grid voxels with the same 3-D structural label. While the reference atlas is typically annotated at the lowest level of the ontology tree, statistics at upper level structures can be obtained by combining measurements of the hierarchical children to obtain statistics for the "parent" structure. The end result is per structure expression density, intensity and energy values for each image-series which are stored in the database and used in the web application to display expression summary bar graphs.

EXPRESSION GRID SEARCH SERVICE

A novel on-the-fly expression grid search service has been implemented to allow users to instantly search over the ~25,000 image series to find genes with specific expression patterns:

- The "**Differential Search**" function allows users to find genes which have higher expression in one structure (or set of structures) compared to another structure (or set of structures).
- The "**NeuroBlast**" function enables the user to find genes that have a similar spatial expression profile to a seed gene when compared over a user-specified domain.

In order to perform these computations quickly over the entire data set, a subset of voxels are loaded in memory. The full expression grid is 67x41x58=159,326 voxels spanning both hemispheres and includes background voxels. To load all voxels for all image series into memory would require 14GB of RAM. To reduce memory requirements and increase the efficiency of calculations, voxels spanning over 80% of all experiments were identified. Only these ~26,000 voxels were then used in the search service requiring 4 GB of RAM and partially spanning one hemisphere.

To activate a "Differential Search", a user specifies a set of target structures and a set of contrast structures. In the service, the set of voxels belonging to any of the target structures forms the target voxel set, and voxels belonging to any of the contrast structures form the contrast voxel set. For each image series, a fold change is computed as the ratio of average expression energy in the target voxel set over the average expression energy in the contrast voxel set. The dataset is then sorted in descending order by fold-change and displayed on the Web application.

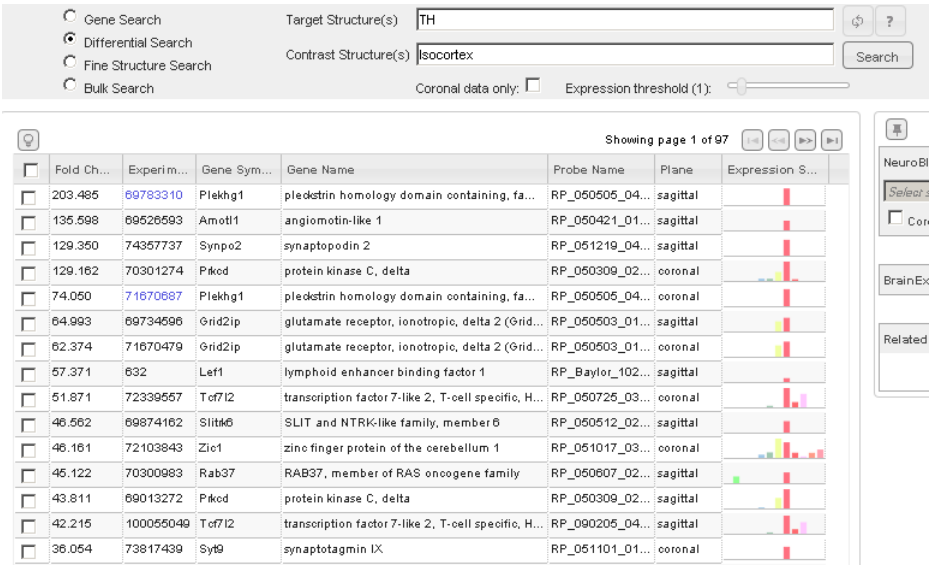

Figure 6. Differential search for genes with higher expression in the thalamus (TH) than the isocortex.
Screenshot of the top 15 returns. Mini-expression summary graphs show enrichment in the thalamus (red) compared to other brain regions.

To start a “NeuroBlast search”, a user selects a seed image-series by selecting a row in any search result and a domain over which the similarity comparison is to be made. All voxels belonging to any of the domain structures forms the domain voxel set. Pearson’s correlation coefficient is computed between the domain voxel set from the seed image series and every other image series in the dataset. The dataset is then sorted by descending correlation coefficient and displayed on the Web application.

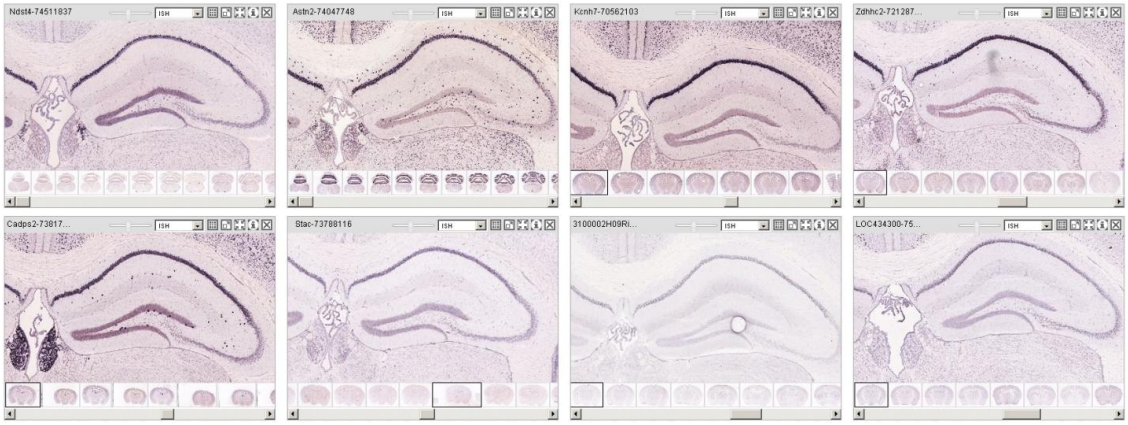

Figure 7. NeuroBlast search for gene with similar expression to Ndst4 in the hippocampus.
Screenshot of the top seven results with the seed gene in the top-left corner. Gene Ndst4 shows relative enrichment of CA1 compared to the rest of the hippocampus.

To take advantage the data on both hemispheres in coronal data, a second “coronal only” search service is also available as an option. The coronal service spans both hemispheres covering 58,387 voxels and searches over the ~4,000 coronal image series.

It should be noted that this on-the-fly search service is derived from a fully automated processing pipeline. False positive and false negative results can occur due to artifacts on the tissue section or slide and/or algorithmic inaccuracies. Users should confirm results with visual inspection of the ISH images.

AGEA: ANATOMIC GENE EXPRESSION ATLAS

The Anatomic Gene Expression Atlas (AGEA) is a relational atlas that allows users to explore spatial relationships in the adult mouse brain based on the expression patterns of ~4,000 genes, which comprise the set of coronal image-series in the Allen Mouse Brain Atlas. In "**Correlation**" mode, the AGEA is an interactive visualization of 3-D correlation maps rendered as false color images. The value at a spatial location (voxel) on a map represents the Pearson's correlation coefficient (cc). Correlation is computed over a "gene vector" whose elements represent the expression energy for a gene at the voxel of interest. 3-D correlation maps are generated for each possible seed voxel (~51000).

In "**Clusters**" mode, AGEA displays a data-driven hierarchical spatial organization of the brain computed from the AGEA correlation maps. The spectrum of gene expression patterns in the brain is complex, displaying both intra-structure widespread expression and various regional specificity. A simple binary-tree clustering approach was used to capture the various scales of spatial co-expression. To initialize, all voxels were assigned to the root node of the tree. As we descend the tree, a node is bifurcated into two nodes to achieve the maximal dissimilarity between the two groups of voxels based on correlation values. In "**Gene Finder**" mode, users can search for genes with local regionality as defined by AGEA correlation maps. See (Ng et al, 2010) for further computation and analysis details.

Note that in the November, 2011 release, AGEA is computed and displayed with respect to version 1 of the 3-D atlas Allen Reference Atlas space. AGEA will be converted to version 2 of the reference space in a future release.

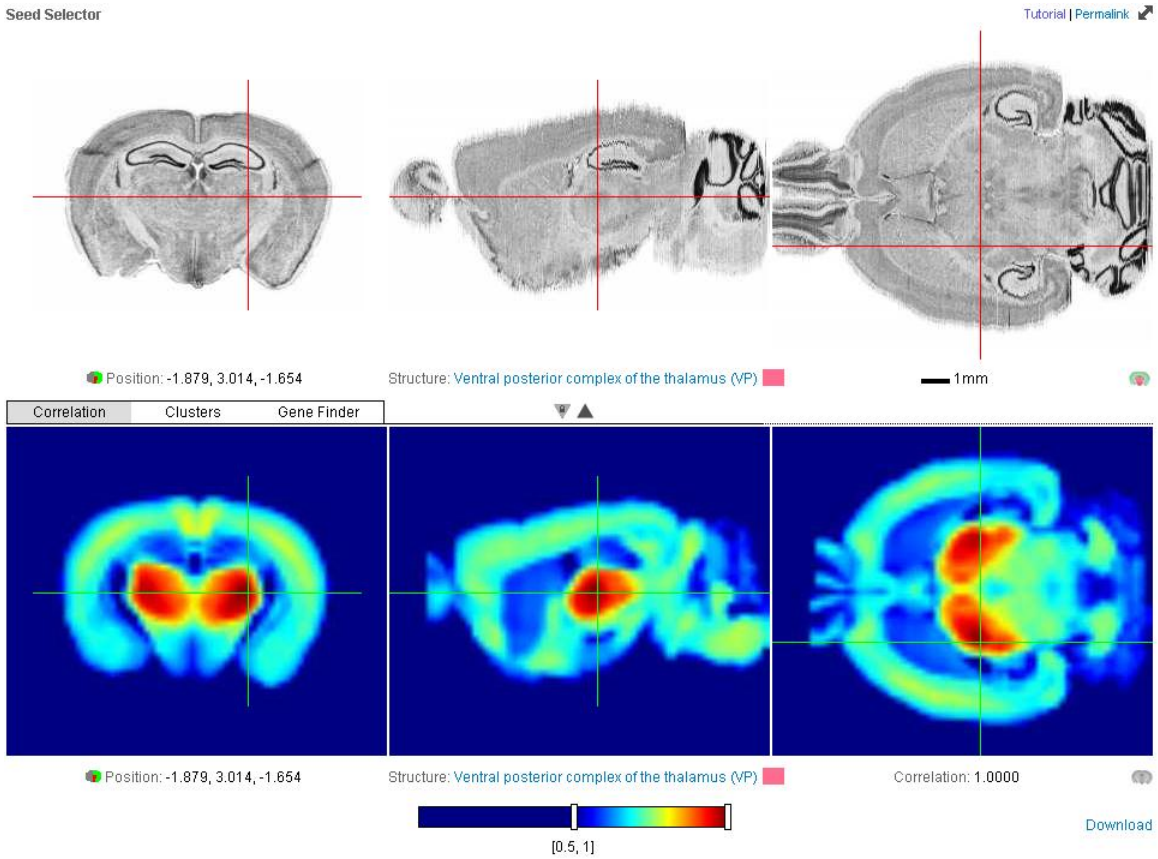

Figure 8. Anatomic Gene Expression Atlas (AGEA) interactive user interface.
Screenshot of AGEA user interface showing spatial correlation map for a seed voxel in the ventral posterior complex of the thalamus (VP). Correlation map shows that the seed location is highly correlated with other regions in the thalamus. Additionally, the seed voxel is more correlation with layer 4 and the retrosplenial area than other areas in the isocortex.

REFERENCES

Yushkevich P, Avants B, Ng L, Hawrylycz M, Burstein P, Zhang H, Gee J (2006) 3-D mouse brain reconstruction from histology using a coarse-to-fine approach. *Third International Workshop on Biomedical Image Registration (WBIR)* 230-237.

Ng L, Pathak SD, Kuan C, Lau C, Dong H, Sodt A, Dang C, Avants B, Yushkevich P, Gee JC, Haynor D, Lein E, Jones A, Hawrylycz M (2007) Neuroinformatics for genome-wide 3-D gene expression mapping in the mouse brain. *IEEE/ACM Trans Comput Biol Bioinform* Jul-Sep;4(3):382-93.

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.

Dang C, Sodt A, Lau C, Youngstrom B, Ng L, Kuan L, Pathak S, Jones A, Hawrylycz M (2007) The Allen Brain Atlas: Delivering neuroscience to the Web on a genome-wide scale. *Data Integration in the Life Sciences* in *Lecture Notes in Computer Science* 4544:17-26.

Ng L et al. (2010) An anatomic gene expression atlas of the adult mouse brain. *Nature Neuroscience* 12(3): 356-362.

APPENDIX

Updated 3-D Atlas Reconstruction

In 2005, the first version of the Allen Reference Atlas 3-D reference space was created using methods described in (Yushkevich, 2006). The reconstruction was based on a combination of high-frequency section-to-section histology registration with low-frequency histology to (ex-cranio) MRI registration. Outlines for approximately 200 structures were extracted from the 2-D annotations of the coronal plates of the Allen Reference Atlas. These were inserted into the 3-D model and interpolated to created 3-D annotations spanning a full hemisphere.

In 2011, a revised 3-D reference space was created to address several major issues with version 1, such as the lack of annotation of the other hemisphere, a twist in the mid-sagittal plane, and errors in scaling. After correcting for scale, the twist in reconstruction was rectified using a reconstructed (fresh frozen) sagittal specimen for guidance. Once a straight mid-sagittal plane was achieved, a synthetic symmetric space was created by reflecting one hemisphere (the annotated side) to the other side of the volume. At the same time, reprocessing of the Allen Reference Atlas enabled the extraction of over 800 structures from the 2-D annotation, which were then interpolated to create 3-D annotations. The end result is a symmetric, fully annotated reference space with a more consistent and deeper level of annotation then the previous version.