ALLEN INSTITUTE for BRAIN SCIENCE
Fueling Discovery

ALLEN **Mouse Brain Atlas**

TECHNICAL WHITE PAPER: FINE STRUCTURE ANNOTATION

Identification of genes that are enriched within or predominantly specific to a brain structure or nuclei is potentially of high biological interest. Fine Structure Annotation lists are sets of approximately 50 genes found to exhibit particularly specific expression patterns in the named structure. The most efficient way to find genes that are enhanced in a given region is to use the mapped expression data with a combination of informatics methods and hand curated validation of the resulting genes. As these lists are generated, the Allen Mouse Brain Atlas is making them available for public data viewing and research. Methods

Methods

The Allen Mouse Brain Atlas dataset consists of approximately 20,000 genes. A standard sagittal images series consists of 20 images at 1.07μm^2 resolution spanning one hemisphere and ~210 million "on tissue" image pixels. To make mining of the Allen Mouse Brain Atlas data tractable, expression statistics are aggregated by structure into 200μm^2 grids.

We first start by making a subdivision of the 3D informatics reference atlas volume into isotropic 100μm^3 cubes. Each cube is assigned a unique identifier and serves as the highest resolution for quantifying expression information for the purpose of search and comparison. The reference atlas grid and corresponding structural labels are mapped onto the ISH images using the transform parameters computed in the registration process. The resulting deformed polyhedra define the spatial extent over which local statistics are aggregated for each cube, enabling systematic mapping of gene expression statistics of every ISH image into the common coordinate system. Every cube is also labeled with the anatomic structure they intersect. Thus, statistics at a 100 micron grid level can be further aggregated for each brain structure. The 100 micron grid is smoothed and post-processed to be more consistent with the out of plane resolution of the data (200 microns for sagittal sections, 100 or 200 microns for coronal sections) to produce a 200 micron grid over which correlation searches in gene expression can be performed.

Since every image series is spatially mapped to the same 3D atlas, we can compare expression statistics in approximately the same 200 micron spatial extent for all image series in the Allen Mouse Brain Atlas dataset. We have used the 200 micron grid structure data to mine for genes that are differentially enriched in given regions of interest. While many different statistics and features can be derived from the expression segmentation mask (for example, number of expressing cells, average cell size, average intensity, etc.), in the search for genes with enriched expression patterns in particular structures, we use a measure of "expression energy".

Expression energy E(C) for a cube C is defined as follows:

$$E\ (C) = \frac{\sum_{\forall p \in} M(p) \times\ I(p)}{\#C}$$

where p is an image pixel that intersects cube C, #C is the total number of pixels which intersects C, M(p) is the expression segmentation mask which is either 1 (“expressing” pixel) or 0 (“non-expressing” pixel) and I(p) is the grayscale value of the ISH image intensity.

The advantages of using this measure are that it:

1. Can be robustly computed over all regions of the brain.
2. Is easily normalized to account for different image resolution and/or section sampling frequency.
3. Is amenable to smoothing for generating the down sampled 200 micron grid.
4. Combines the features of expression intensity and expression density as a single measurement.

Structure enriched gene search
One approach to measuring specificity is the energy ratio:

$$S(A,B) = \frac{E(A)}{E(B)}$$

where A is the structure of interest, B is an enclosing structure or arbitrarily defined region such that A⊂ B, and E(A), E(B) are respectively the expression energy for A and B. Typically B will be chosen to be the parent structure of A in the given ontology; for example if A is “lateral septal complex”, B might be chosen to be “striatum.” Values of S range from 0.0 (not expressing in A) to 1.0 (ideal specificity to A). In practice, false positives can result from dark bubble-like artifacts and/or from registration inaccuracies causing leakage of expression from neighboring structures of A.

Curated Annotation Lists
The Fine Structure Annotation lists provided on www.brain-map.org were identified by the process described above and then confirmed by manual expert curation.