

## TECHNICAL WHITE PAPER: SPATIOTEMPORAL SEARCH

A major component of the NIH Blueprint Non-Human Primate (NHP) Atlas is a detailed characterization of cellular gene transcript distribution patterns in five major brain regions across postnatal development. These data are generated using *in situ* (ISH) hybridization to label cells using a particular gene transcript on thin tissue sections spanning the rostral to caudal extent of each brain region and all subnuclei and cell types that make up that structure. Many of the genes analyzed in the NHP atlas were identified on the basis of highly selective expression in rodents, allowing a direct comparison between rodent and primate gene usage through comparison to the Allen Mouse Brain Atlas (available via the Allen Brain Atlas portal at [www.brain-map.org](http://www.brain-map.org)). The NHP atlas will ultimately include ISH data in the medial prefrontal cortex, primary visual cortex, hippocampus, amygdala and ventral striatum, with overlapping but independent gene sets for each structure. For each gene transcript, three replicate data sets are generated at each of four postnatal developmental stages including 0, 3, 12 and 48 months, allowing identification of transcriptional programs differentially active at different stages of brain maturation in neonates, infants, juveniles and adults.

### ABOUT THE SEARCH

A wide variety of complex, highly regionalized and/or developmentally regulated expression patterns are observed in each brain region, and a major goal of the NHP atlas is to allow users to identify genes with specific patterns within the data set as a whole. In order to provide detailed spatiotemporal searching and data mining capabilities to these data, a qualitative scoring strategy was used to discretize specific characteristics of the expression data across a series of well defined subdivisions of each major brain region at each developmental stage. This scoring was designed to provide a single set of expression scores for each individual structure/age for the purpose of allowing searches for differential gene expression between neuroanatomical structures and across ages, and therefore represents the consensus pattern across the three replicate data series for each structure and age rather than an independent set of scores for each replicate. For each major brain region the structure set used for scoring represents major definable subnuclei or laminar compartments, capturing both the functional architecture of the structure as well as the specific cellular components that make up that subdivision (e.g. GABAergic interneurons in non-pyramidal sublayers of CA subfields of the hippocampus).

The scoring paradigm was designed to capture two fundamental aspects of cellular labeling observed in ISH data. The first metric is the cellular expression level, related directly to the staining intensity observed in labeled cells (on a 0-5 scale). As described below, this involved expert scoring of the representative cellular labeling based on algorithmically quantified signal "heat maps," which are also available for visualization for each ISH image in the atlas. The second metric is the density of labeled cells in a particular structure (on a 0-5 scale), designed to capture the proportion of labeled neurons observed within particular brain nuclei. This cellular distribution is highly meaningful, in that sparse and scattered cellular distributions (e.g. scored as density of 1 or 2) typically represent selective expression in glial cells or GABAergic interneurons, whereas higher densities typically represent primary excitatory neurons (or at least include these neurons). The combination of fine neuroanatomical distribution and relative density provide a great deal of predictive power for the cell types using a particular gene transcript, and therefore some indication of that gene's function in the brain region as a whole. Differential expression across the development timepoints assayed in this data set can provide some insight into functional roles for genes in developmental processes associated with maturation of functional circuitry from birth through adulthood.

The search engine is designed to provide two major types of searches to mine these data. The first type allows a search for specific expression patterns in user-selected brain regions, as well as differential expression between two structures. The second search type allows a search for differential expression across developmental stages within a specific neuroanatomical structure. The results of these searches are

a list of ISH image series for genes that meet the criteria defined in the search, allowing rapid examination of the primary data sets.

## DATA SCORING METHODS

To characterize detailed cellular gene expression patterns in the developing macaque brain, qualitative expert annotation of non-isotopic colorimetric *in situ* hybridization (ISH) was performed for each structure across the four developmental timepoints (0, 3, 12 and 48 months). The experimental paradigm for the ISH data generation consists of uniformly spaced sampling for each gene across the structure, with interleaved Nissl-stained sections as anatomical reference series to provide cytoarchitectural delineation of fine nuclear/cellular structure. Each gene is assayed every 1mm, with Nissl staining on interleaved sections at twice this density (500µm spacing). For each major brain region a set of substructures (referred to below as regions of interest or ROIs) clearly identifiable in the Nissl-stained sections were established as the structural basis for the scoring paradigm.

The goal of the scoring was to provide a quantized data set for each structure/age, with a single set of scores representing the expression characteristics of a representative specimen from three independent biological replicate series available for each age. For each specimen at a particular age, ROIs were initially identified based on the Nissl images and available reference literature and atlases. The corresponding ISH gene expression data were then assessed for tissue and stain quality as well as consistency between the replicates at that age, and one of the replicates was selected for scoring as an image series that provides a representative, high quality labeling pattern in a data set that includes all ROIs for that structure. For this image series, the density and intensity of expression within each ROI was assessed using a standardized semi-quantitative scoring system, allowing for direct comparisons across gene, region and/or age. Data for a particular gene was scored by a single annotator, with all timepoints scored together where possible to provide scoring consistency across ages. Clear ISH artifacts and individual data sets displaying different patterns from the other replicates were excluded.

Spatiotemporal search is currently available for the hippocampus and striatum.

## ROI Delineation

Hierarchical structural ontologies for scoring were established for each structure, based on published atlases of the rhesus macaque brain (Paxinos 2008). These structures constitute the set of ROIs for scoring.

### Hippocampus:

#### *Ammon's Horn*

- CA1
  - Stratum Oriens
  - Stratum Pyramidale
  - Stratum Radiatum
  - Stratum Moleculare
- CA2
  - Stratum Oriens
  - Stratum Pyramidale
  - Stratum Radiatum
  - Stratum Moleculare
- CA3
  - Stratum Oriens
  - Stratum Pyramidale
  - Stratum Lucidum
  - Stratum Radiatum
  - Stratum Moleculare
- CA4

#### *Dentate gyrus*

- Molecular layer
- Granule cell layer
- Polymorph layer

### Striatum

#### *Dorsal striatum*

- Caudate nucleus (Cd)
- Putamen (Pu)
- Internal capsule (ic)

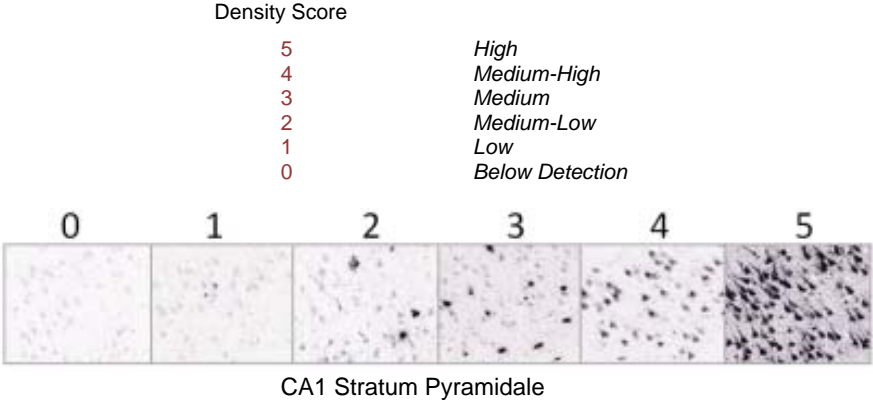
#### *Ventral striatum*

- Nucleus accumbens, Core (AcbC)
- Nucleus accumbens, Shell (AcbS)
- Islands of Calleja (Isl)
- Olfactory tubercle (Tu)

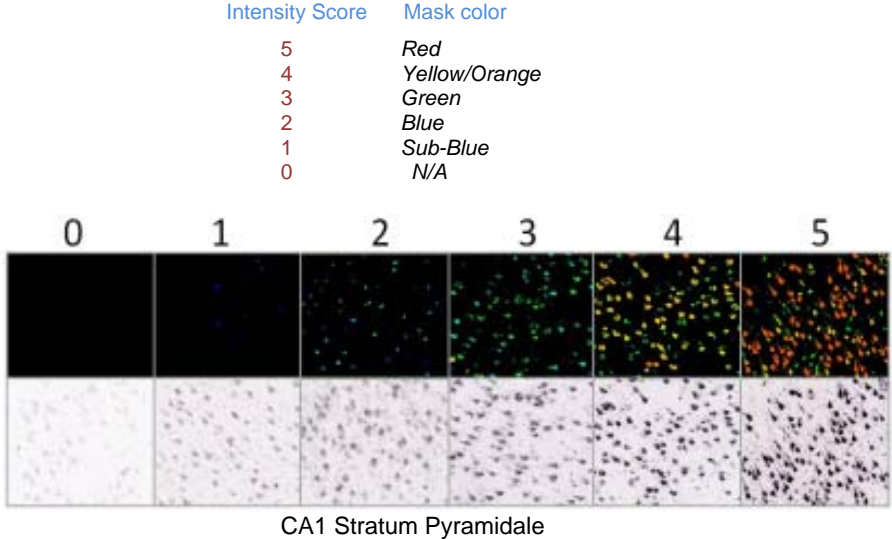
For each specimen, the interleaved Nissl-stained sections were used to delineate the ROIs across the image series constituting the entire structure. These delineations were then used to aid identification of these ROIs for scoring of ISH patterns on nearly adjacent tissue sections, where the cellular labeling patterns are frequently insufficient for structural delineation by themselves. In order to provide a representative score for each structure, all slides containing that structure for a given gene were analyzed to provide a single representative score for that ROI.

**Gene expression scoring**

The density and intensity of gene expression (*D, I*), was assessed using a semi-quantitative scoring method. For expression **density**, scores ranged from 0 (below detection) to 5 (high density), and were assessed manually.



For expression **intensity**, scoring was based on a color-coded heat map image of the ISH data based on an automated quantification algorithm that segments the image to identify clusters of pixels (cells) with signal over background and color-codes these cells by labeling intensity. These color coded “heat masks” are also available for viewing alongside the primary ISH data. Scores ranged from 0 to 5 based on visual integration of representative cellular labeling in the ROI (see key below). N/A indicated either the absence of the ROI in the data set scored, or ambiguity for presence of artifacts.



For each structure, a scoring matrix was constructed based on the structural ontology and density and intensity (D, I) scores as below, and the user interface for structure-based searches rely on selection of specific structures from this structure matrix representation:

Hippocampal Formation Structure Matrix

	Stratum Oriens	Stratum Pyramidale	Stratum Lucidum	Stratum Radiatum	Stratum Moleculare
CA1	(D, I)	(D, I)		(D, I)	(D, I)
CA2	(D, I)	(D, I)		(D, I)	(D, I)
CA3	(D, I)	(D, I)	(D, I)	(D, I)	(D, I)
CA4	Molecular	Granule	Polymorph		
DG	(D, I)	(D, I)	(D, I)		

Striatum Structure Matrix

	Dorsal Striatum
Cd	(D, I)
Pu	(D, I)
ic	(D, I)
	Ventral Striatum
AcbC	(D, I)
AcbS	(D, I)
Isl	(D, I)
Tu	(D, I)