

Technical White Paper: NIH Blueprint Non-Human Primate (NHP) Atlas

PROJECT OVERVIEW

The NIH Blueprint Non-Human Primate (NHP) Atlas project, funded by the NIH Blueprint for Neuroscience Research and the National Institute of Mental Health (NIMH), presents gene expression data across different neuroanatomical regions and stages of postnatal development in rhesus macaque brain. The atlas is designed to contain an MRI and histological reference series at each developmental timepoint, genome-scale transcriptional profiling using DNA microarrays, and cellular resolution localization of gene expression in specific anatomical regions and cell types using colorimetric *in situ* hybridization (ISH). These data will be publicly accessible via a web-based application that allows viewing of transcriptional data, searchable by gene, brain area, specimen and age.

The NIH Blueprint NHP Atlas project was initiated in October 2008, and is slated for completion October 2011. The initial data release in October 2009 contains ISH data presented as navigable images, and the project as a whole will contain the following components:

- (1) **Cellular resolution in situ hybridization (ISH)**: A focused study of high-interest genes represented in this dataset comprise several categories of broad scientific and clinical interest, including highly restricted marker genes for specific anatomical regions and cell types, gene families important to neural function, disease-related genes and genes important in the comparative genomics field. ISH data will be presented for medial prefrontal cortex, primary visual cortex, hippocampus, amygdala and ventral striatum, with genes for each structure assayed across four timepoints (birth, 3 months, 1 year, 4 years) in three male specimens.
- (2) **Genome-scale transcriptional profiling** (COMING SOON): Genome-wide transcriptional profiles of the same neuroanatomical regions and developmental stages will be generated using DNA microarrays. These data will consist of profiling by
 - a. Grossly dissected neuroanatomical regions of interest
 - b. Laser microdissected cytoarchitectural and functional subdivisions of these regions of interest
- (3) **Stage-specific MRI and histological reference series** (COMING SOON): High resolution MRI and Nissl histology reference series for each developmental timepoint will be presented to provide a developmental neuroanatomical frame of reference for the ISH and microarray data.

The NHP Developing Atlas can be accessed through the Allen Atlas Portal (www.brain-map.org) or directly at www.blueprintnhpatlas.org.

Pipeline Overview:

High-throughput processes for generation of ISH-based gene expression data were developed at the Allen Institute for Brain Science for the production of the Allen Mouse Brain Atlas (<http://www.brain-map.org>), a genome-wide atlas of gene expression in the mouse brain [1]. The process, equipment and workflow for generation of gene expression data in macaque closely follows that described for generation of the Allen Mouse Brain Atlas (see Supplemental Methods 1 in [1]) with some adaptation to manage specific challenges posed by working with developing macaque tissue. For example, modifications were made to ISH methodology (e.g. PK concentration differences at different ages), image acquisition and data processing capacity to accommodate larger tissue sections.

A Laboratory Information Management System (LIMS) was used to manage all information related to experimental design, slide and image tracking as described previously [1]. In addition to modifications to laboratory production processes mentioned above, the LIMS was updated to accommodate information

management needs specific to macaque tissue samples. The LIMS is also used to view raw image data and perform image QC following initial image acquisition.

Briefly, the workflow involved the following: Validated tissue and probes were coordinated in to work packets, where each packet consisted of tissue and probes that were progressed as a discrete unit through the entire process of sectioning, ISH, image acquisition and image processing. QC metrics were established for image and data quality. Images passing QC were then released for public display. This workflow enabled the systematic generation of data for each gene, and where possible data for a particular gene/structure (multiple ages and replicates) were processed simultaneously to allow cross-comparison between samples.

All processes associated with data production, including tissue receipt and storage, solution preparation, probe preparation, colorimetric ISH, and equipment and other laboratory maintenance functions were governed by Allen Institute Standard Operating Procedures (SOPs) with revision control, and complied with appropriate health and safety precautions.

Gene Selection

For the ISH component of the project, gene selection was based on four major categories constituting thematically interesting datasets for a broad user community. These overlapping classes included the following, with candidate selection biased in favor of genes contained in multiple categories, and with evidence for developmental regulation:

1. ***Gene markers delineating highly regionalized anatomical subdomains, subnuclei and/or cell types.*** Since very little non-human primate transcriptional data was available, these markers were primarily identified through mining of extant data resources in rodents, including the Allen Mouse Brain Atlas (www.brain-map.org), other ISH and microarray-based public data repositories, and literature. These cellular markers generally aim to sample the major cytoarchitectural features of each structure, including, for example, markers for cortical layers, hippocampal subfields or amygdalar subnuclei, as well as markers for specific cell classes with less discrete localization including interneurons, glial and vascular markers.
2. ***Gene families important for neural function.*** These families included ion channels, G-protein-coupled receptors (GPCRs), transporters, synaptic proteins, membrane proteins, and peptide or protein ligands. Particular emphasis was placed on ion channels and genes related to GABAergic neurotransmission.
3. ***Disease-related genes.*** Genes were selected by searching through available literature to identify genes (i) conferring disease susceptibility, (ii) known to be involved in physiological pathways implicated in disease, or (iii) encoding known drug targets. The diseases included autism, schizophrenia, epilepsy, microcephaly, neurodegenerative diseases, depression and anxiety and intellectual and developmental disabilities.
4. ***Comparative genomics.*** This category comprised genes identified in the literature as showing accelerated evolution, being under positive selection, or showing microarray-based gene expression differences either between rodents and primates or humans.

The gene set for which ISH data is available is supplied in Documentation available on the online application.

Tissue Specimens

Frozen postmortem tissue samples from male rhesus macaque (*Macaca Mulatta*) were obtained through the California National Primate Research Center (CNPRC; <http://www.cnprc.ucdavis.edu/>). For the purpose of generating histological and *in situ* hybridization (ISH) data as well as RNA analysis by

microarrays, brain regions were systematically collected from well-characterized rhesus monkeys born and raised in outdoor, ½-acre naturalistic enclosures located at the CNPRC. Animals were born and raised in these enclosures that provide a naturalistic setting and normal social environment. Extensive health, family lineage and dominance information is maintained on all animals in the outdoor enclosures. All procedures were approved by IACUC.

After dissection, brains were sectioned into coronal slabs approximately 1 to 1.5 cm in thickness. These slabs were further dissected to yield blocks containing structures of interest contained within an area that will fit on a standard 1x3 inch microscope slide. These blocks were photodocumented, frozen on dry ice and stored at -80°C. For a subset of animals, regions of interest were selectively dissected from one hemisphere for RNA isolation and microarray analysis and frozen at -80°C until further processing.

Upon receipt at the Allen Institute, tissue was again photodocumented and information entered into LIMS for future tracking throughout the data pipeline. Tissue was stored at -80°C until removed for sectioning. Prior to sectioning for ISH, all tissue samples were tested to confirm that the region of interest was present based on expected cytoarchitecture. Tissue samples that failed these criteria were not used in the study.

Probe Design and Synthesis

For labeling target mRNA in tissue sections using ISH, digoxigenin-labeled riboprobes were designed and synthesized according to specific criteria. In general, the design and synthesis process followed previously described methods used to generate probes for the Allen Mouse Brain Atlas [1] with some modification. Briefly, probes were designed to be between 400-1000 bases in length (optimally >600 bases) and to contain no more than 200 bp with >90% homology to non-target transcripts, using sequence information obtained from NCBI RefSeq (<http://www.ncbi.nlm.nih.gov/RefSeq>), and a semi-automated process using Primer3 software [2]. In addition, to allow comparability of mouse and human gene expression datasets, probes were designed to overlap with the existing Allen Mouse Brain Atlas probe when the mouse and macaque genes were orthologous. Riboprobes were synthesized using standard *in vitro* transcription (IVT) reactions based on PCR templates prepared from cDNA clones (NIH Mammalian Gene Collection, Open Biosystems, Huntsville, AL) or cDNA synthesized from brain total RNA. cDNA was prepared from brain RNA from cerebellum using Superscript III RTS First-Strand cDNA Synthesis Kit (Invitrogen, Carlsbad, CA) to provide templates for PCR.

PCR primers were obtained from Integrated DNA Technologies (Coralville, IA) at a final concentration of 10 µM, and designed with GC content between 42% to 62% and an optimal size of 22nt with lower and upper limits of 18 and 26 nt, respectively. For cDNA clones, the clone sequence was compared with RefSeq sequences, and consensus sequences with >98% homology across 80% of the total length were used to develop probes. When a clone was used as a template, a single PCR was used requiring only a forward and reverse primer with an additional SP6 RNA polymerase binding sequence (GCGATTTAGGTGACTATAG). When using brain cDNA as a template, probes were generated against sequences within a region 3000 bp from the 3' end using 3 primers: forward, reverse, and a nested reverse primer containing the SP6 RNA polymerase binding sequence. cDNA primers underwent a BLAST analysis to verify amplification of only target sequence. All cDNA reactions were run on the Bioanalyzer for quality control.

Standard conditions for PCR and IVT reactions were as described. IVT reactions were diluted to working stocks of 30ng/µl with THE (0.1mM Sodium Citrate pH 6.4, Ambion). Aliquots were stored in single-use volumes to minimize freeze/thaw cycles. IVT dilutions were stored at -80°C. For hybridization, the probe was diluted 1:100 (to 300ng/ml) into *in situ* hybridization buffer (Ambion) in 96-well ISH Probe Plates. Each well provides probe for one ISH slide. Probe plates were stored at -20°C until used in an ISH run.

All PCR and IVT products were run on the Bioanalyzer for size and peak characteristic quality control. Specifically, PCR products that were not of the correct size (+/- 100bp) or that showed multiple products

were not used to generate riboprobes. IVT products that were shorter than their predicted size were not used. It is common to see IVT products that run slightly larger than their predicted molecular weight, or as multiple peaks, due to secondary structure of the RNA.

Tissue Sectioning

Frozen tissue samples were cryosectioned using Leica CM3050 S cryostats (object temperature, -10°C; chamber temperature, -15°C) at 18 µm (amygdala and ventral striatum) or 20 µm (hippocampus, cortical areas) thickness in the coronal plane from anterior to posterior. One or two sections were placed on a positively charged Superfrost Plus™ 1" x 3" microscope slide (Erie Scientific Co, Portsmouth, NH), pre-printed with a unique identifying barcode for tracking. Specimen numbers were also printed on to the slides for tracking purposes. For samples in which two sections/slide were used, sections were serially sectioned onto sets of 50 slides, such that sections 1 and 51 go on slide 1, sections 2 and 52 go on slide 2 ... sections 50 and 100 go on slide 50, sections 101 and 151 go on slide 51 and so on. Each gene was processed on a slide series, representing the same position on sequential sets of 50 slides (e.g. slide 1, 51, 101), thereby giving uniform sampling at ~1 mm spacing across the sample block (e.g. 50 serial sections x 20 µm = 1 mm). In practice, each series consisted of 1 to 5 slides, determined by the tissue sample thickness. For anatomical and cytoarchitectural reference, two of the 50 sectioning series (series 2 and series 26) were designated for Nissl staining so that a Nissl reference was available every ~500 µm throughout the tissue block. Two additional series were used for positive control genes and two series were reserved as backups for repeating data that might fail QC. Each of the remaining 44 series was hybridized with a single gene probe.

Following sectioning, slides designated for ISH were allowed to air dry and tissue was fixed, acetylated and dehydrated according to standard protocols as described [1]. Briefly, tissue was fixed for 20 minutes in 4% neutral buffered paraformaldehyde (PFA) and rinsed in 1x PBS, acetylated for 10 minutes in 0.1M triethanolamine with 0.25% acetic anhydride, and subsequently dehydrated using a graded series of 50%, 70%, 95% and 100% ethanol. Slides that passed section quality QC checks were stored at room temperature in Parafilm™-sealed slides boxes until use.

Standardized high throughput colorimetric ISH

An automated technology platform for standardized high throughput histological slide processing was used to generate cellular resolution ISH data on rhesus macaque tissue sections. Digoxigenin-based riboprobe labeling, coupled with TSA amplification and alkaline phosphatase-based colorimetric detection was used to label target mRNAs in expressing cells. Detailed descriptions of the high-throughput platform, protocols, and reagent preparation are available elsewhere ([1], Supplemental Methods 1).

Briefly, slides containing tissue sections were placed in flow-through chambers on temperature-controlled racks on computer-controlled Tecan Genesis liquid handling platforms for sequential application of solutions. Initial steps in the protocol block endogenous peroxidase activity and permeabilize the tissue for probe penetration. Digoxigenin-labeled probes are subsequently hybridized to target mRNA, and after a series of washes to eliminate excess probe the remaining bound probe is subjected to a series of enzymatic reaction steps to detect and amplify the digoxigenin signal. First, a horseradish peroxidase (HRP)-conjugated anti-digoxigenin antibody is added, followed by biotin-coupled tyramide that is converted by HRP to an intermediate that binds to cell-associated proteins at or near the HRP-linked probe. Neutravidin conjugated with alkaline phosphatase (AP) is then bound to biotin and BCIP/NBT is added for colorimetric detection. A blue/purple particulate precipitate forms as a result of the enzymatic cleavage of BCIP by AP and subsequent indole reaction with NBT. Finally, the colorimetric reaction is terminated by washing with EDTA and fixation with 4% PFA. This entire process occurs over the course of approximately 23.5 hours on the Tecan automated platform.

Each ISH run contained several positive controls. Macaque-specific probes against GAD1 and CALB1 were included as robust and moderately expressed control genes, respectively. A *Drd1a* positive control on mouse brain tissue (the same control used throughout the Allen Mouse Brain Atlas project) was used

to provide verification of a successful ISH run. A negative control (no probe) slide was also included as an indication of background for each ISH run.

To reduce background signal, an acid alcohol wash step was performed after completion of the hybridization process. Slides were rinsed 4 times (1 minute each) in acid alcohol (70%, adjusted to pH = 2.1 with 12N HCl) and rinsed 4 times in milliQ water (1 minute each). Acid alcohol and water solutions were refreshed every fourth rack to ensure that all slides were rinsed in clean solution.

Image Acquisition

Image acquisition was performed using ScanScope® scanners (Aperio Technologies, Inc; Vista, CA). The line scan camera continually adjusts for focus based on a variable number of focus points and provided advantages over tile-based image acquisition platforms for large tissue sections that tended to have more variation in height. The ScanScope scanner uses a 20x objective that is downsampled in software to minimize data volume acquired for this project. The downsampling provides comparable image resolution (approximately 1.00 $\mu\text{m}/\text{pixel}$) to the ICS scanning systems with 10x objectives used for other Allen Brain Atlas projects (3).

Data Processing

Once images were acquired, the Informatics Data Pipeline (IDP) managed image preprocessing, image QC, ISH expression detection and measurement, Nissl processing, annotation QC and public display of information via the Web application. The IDP has been described in detail previously (4), and was modified slightly for processing rhesus macaque histology images.

IDP Cluster Computation Requirements

The large image size resulting from rhesus macaque tissue sections presented a major challenge for the image processing pipeline. To support the processing of macaque tissue images generated by the ScanScope, the informatics processing platform was migrated to a 64-bit Linux platform, including cluster hardware, system software and IDP applications. Cluster blades were configured to operate in 64-bit mode with at least 8 GB of main memory each to provide dedicated blades with 8 – 14 GB of working memory to execute the processing modules.

IDP Processing Modules

Three modules constituted the processing pipeline for macaque images: (1) image preprocessing, (2) ISH expression detection and (3) Nissl processing.

In image preprocessing, scanned ISH and Nissl images were converted and background corrected to provide more consistent white background intensities and orientation across samples. Aperio ScanScope images were first converted to JPEG 2000 format, then orientation-adjusted and white balanced. The final products are images in a JPEG 2000-compressed format for further pipeline processing and analysis.

A major goal of ISH data presentation was to provide users with a quantified representation of the data, similar to the color-coded “heat mask” representation used for the Allen Mouse Brain Atlas (www.brain-map.org). To detect and quantify expressing cells on ISH images of tissue sections, adaptive image processing techniques were applied to 10x full-resolution ISH images. As for imaging, the image sizes posed technological challenges, as the number of detected expressors (cells) on a typical 3 GB macaque neocortex image can reach nearly 2 million. An algorithm based on techniques used for the Allen Mouse Brain Atlas ([1]; Supplemental Methods 2) was significantly redesigned to accommodate the full image resolution needs that resulted from these large image sizes. The resulting module produced a mask of detected expressor objects and a set of numerical values describing the statistical attributes of gene expression. The mask image with measured intensity of expression was then pseudo-color coded and converted to AFF file format for Web display. The raw ISH data, quantified heat mask representation, and closest Nissl-stained section are presented for each tissue section.

QC for web-based data presentation

Additional QC steps during the course of data processing described above ensure the release of accurate, high quality data to the publicly accessible web portal. Once image preprocessing was complete, an image quality control step ensured the images were in good focus, and provided an initial indication of the presence of ISH signal. If focus criteria were not met, the images were failed and the appropriate slides were rescanned. If focus criteria were met, the images were passed and processed through the IDP for ISH expression detection. From time to time, slides may be rescanned to improve image quality. In these cases images in the database will be replaced with the most recent scan of the original slide.

References

1. Lein, E.S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168-76 (2007).
2. Rozen, S. & Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* 365-386 (Humana Press, 2000).
3. Slaughterbeck, C.R. et al. High-throughput automated microscopy platform for the Allen Brain Atlas. *J Assoc Lab Automation* 12, 377-383 (2007).
4. Dang, C.N. et al. The Allen Brain Atlas: Delivering Neuroscience to the Web on a Genome Wide Scale. In *Data Integration in the Life Sciences* 17-26 (Springer Berlin, 2007).