ALLEN INSTITUTE for BRAIN SCIENCE

Allen Cell Types Database

TECHNICAL WHITE PAPER: TRANSCRIPTOMICS

OVERVIEW

This Technical White Paper describes the transcriptomics profiling of the dorsal part of the lateral geniculate complex (LGd). Part of the initial goal of the Allen Institute's cell type project is to characterize, in a systematic and standardized manner, individual neurons in the LGd and primary visual cortex (VISp or V1) of the young adult laboratory mouse. Descriptions of the tissue preparation, RNA amplification and library preparation for RNA-Seq, RNA-Seq data processing, and clustering are provided. Slices were prepared from P53-P59 mice and were sectioned (250 µm) using a vibrating microtome. Each slice was imaged to aid in brain region identification and registration to the Allen Mouse Common Coordinate Framework (CCF). For more information on the CCF please see the whitepaper in the Documentation tab. Regions of interest were microdissected under a fluorescence dissecting microscope. Dissected tissue pieces were treated with protease and subsequently triturated. Single cells were then isolated by fluorescence-activated cell sorting (FACS). cDNA amplification and library construction were performed using SMART-Seq v4 (Clontech) and Nextera XT (Illumina) kits. Single cell libraries were sequenced on HiSeq (Illumina) to generate 50 base-pair paired-end reads. The reads were aligned and after alignment, quality control was performed and two independent methods were utilized to identify sets of clusters.

The core and shell regions of mouse dorsal part of the lateral geniculate complex (LGd) receive different inputs from the retina: on-off direction-selective retinal ganglion cells project to the shell but not core region (Piscopo *et al*. 2013). In addition, approximately 20% of neurons in rodent LGd are GABAergic (Gabbott *et al*. 1986) and 20% are located within the shell region (Piscopo *et al*. 2013). Our sampling strategy for LGd leveraged four mouse Cre lines to capture excitatory (*Snap25*, *Slc17a6*) and inhibitory (*Gad2*, *Slc32a1*) neurons from core and shell regions of LGd based on their relative proportions *in vivo*. Monte Carlo simulations were used to estimate the number of cells needed to capture (with 95% confidence) at least 16 cells of a cell type as rare as 2% of all LGd neurons. With 16 cells, we expected to be able to discriminate between cell types as similar as two Sst interneuron subtypes recently identified in mouse primary visual cortex (Tasic *et al*. 2016).

TISSUE PREPARATION

Male Cre driver mice crossed to the tdTomato (Ai14) reporter line between the ages of P53-P59 were anesthetized with 5% isoflurane and intracardially perfused with 50 ml of ice cold, oxygenated artificial cerebral spinal fluid (ACSF.I) at a flow rate of 9 ml per minute until the liver appeared clear, or the full 50 ml had been flushed through the vasculature. The brain was then rapidly dissected and mounted for coronal slice preparation (rostral end at base) on the chuck of a Compresstome VF-300 vibrating microtome (Precisionary Instruments) (see **Figure 1** for entire workflow). Using a custom designed photodocumentation configuration (Mako G125B PoE camera with custom integrated software), a blockface image was acquired before each section was sliced at 250 µm intervals. The slice was then hemisected along the midline, and the left hemisphere transferred to chilled, oxygenated solution (ACSF.I).

Each slice-hemisphere was transferred into a Sylgard-coated dissection dish containing 3 ml of chilled, oxygenated ACSF.I. Brightfield and fluorescent images between 4X and 20X were obtained of the intact tissue with a Nikon Digital Sight DS-Fi1 mounted to a Nikon SMZ1500 dissecting microscope. To guide anatomical targeting for dissection, boundaries were identified by trained anatomists, comparing the blockface image and the slice image to a matched plane of the Allen Reference Atlas. For LGd, samples for RNA-Seq were targeted for either core or shell enrichment (see **Table 1**). In general, three to four slices were sufficient to capture the targeted region of interest, allowing for expression analysis along the anterior/posterior axis. The region of interest was then dissected and both brightfield and fluorescent images of the dissections were acquired for secondary verification. The dissected regions were transferred in ACSF.I to a microcentrifuge tube, and stored on ice. This process was repeated for all slices containing the target region of interest, with each region of interest deposited into a new microcentrifuge tube.

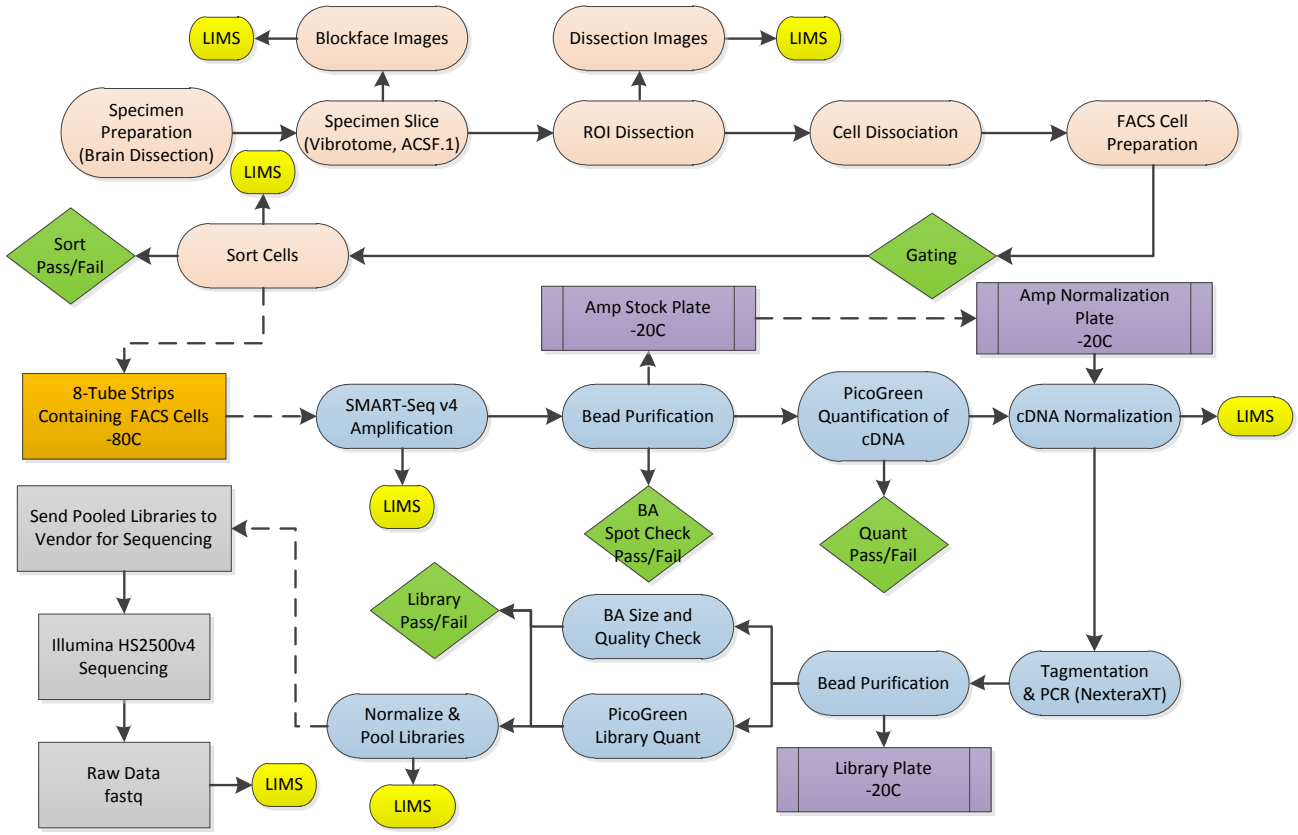

Figure 1. Workflow for tissue preparation and RNA-Seq data generation.
The main steps of the entire workflow include brain dissection, Region of Interest (ROI) dissection, cell sorting, SMART-Seq v4 Ultra amplification, bead purification, PicoGreen quantification, cDNA normalization, tagmentation, PCR, bead purification, and library normalization and pooling. Quality control checkpoints are indicated in green. Interface points with the Laboratory Information Management System (LIMS) are shown in yellow. Abbreviations: BA, Bioanalyzer; Quant, quantification.

Table 1. Summary of single cell source for RNA-Seq data generation.

Cre Line	Type	Core	Shell	Total
Slc17a6-IRES-Cre	Excitatory	415	112	527
Snap25-IRES2-Cre	Pan-neuronal	463	320	783
Slc32a1-IRES-Cre	Inhibitory	213	71	284
Gad2-IRES-Cre	Inhibitory	129	93	221
		1227	590	1815

After all regions of interest were dissected, the ACSF.I was removed and 1 ml of a 2 mg/ml pronase in ACSF.I solution was added. Tissue was digested at room temperature (approximately 22°C) for a duration that consisted of adding 15 minutes to the age of the mouse (in days; *i.e.*, P53 specimen had a digestion time of 68 minutes). After digestion, the pronase solution was removed and replaced by 1 ml of ACSF.I supplemented to a concentration of 1% FBS (Fetal Bovine Serum). The tissue was washed two more times with the same solution. The sample was then triturated using fire-polished glass pipettes with three decreasing bore sizes (600, 350 and 150 µm). The cell suspension was incubated on ice in preparation for fluorescence-activated cell sorting (FACS).

FACS preparation involved adding 2 µl of 4'-6-diamidino-2-phenylindole (DAPI) (2 mg/ml) to the 1.0 ml ACSF.I-1% FBS cell suspension. The suspension was then filtered through a fine-mesh cell strainer (35 µm) and sorted by excluding DAPI positive events and debris, and gating to include red fluorescent events (tdTomato-positive cells). Single cells were collected into strip tubes containing 11.5µl of collection buffer (SMART-Seq v4 lysis buffer 0.83x, (Clontech #634894), RNase Inhibitor (0.17U/µl) and ERCCs (External RNA Controls Consortium) (Baker *et al*.; Risso *et al*.) (MIX1 at $1x10^{-8}$)). After sorting, the cells were subjected to centrifugation and then stored at -80°C.

RNA SEQUENCING

RNA Amplification and Library Preparation for RNA-Seq

SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech #634894) was used per manufacturer's instructions for amplification of single cell RNA and subsequent cDNA synthesis. Single cells were stored in 8-strips at -80°C in 11.5µl of collection buffer (SMART-Seq v4 lysis buffer at 0.83x, RNase Inhibitor at 0.17U/µl, and ERCC MIX1 at final $1x10^{-8}$ dilution as described above. Twelve to 24 8-strips were processed at a time (the equivalent of 1-2 96-well plates). At least 1 control strip was used per amplification set, containing 2 wells without cells (termed ERCC), 2 wells without cells or ERCC (termed NTC), and 4 wells of 10pg of Mouse Whole Brain Total RNA (Zyagen, MR-201). AMPure XP Bead (Agencourt AMPure beads XP PCR, Beckman Coulter A63881) purification was done manually for the first amplification set but then was done using the Agilent Bravo NGS Option A instrument. A bead ratio of 1x was used (50µl of AMPure XP beads to 50µl cDNA PCR product with 1µl of 10x lysis buffer added, as per Clontech instructions), and purified cDNA was eluted in 17µl elution buffer provided by Clontech. All samples were quantitated using PicoGreen® on Molecular Dynamics M2 SpectraMax instrument. A portion of the samples, and all controls, were run on the Agilent Bioanalyzer 2100 using High Sensitivity DNA chips to qualify cDNA size distribution. An average of 8.9ng cDNA was synthesized across all non-control samples. Purified cDNA was stored in 96-well plates at -20°C until library preparation.

All samples proceeded through NexteraXT DNA Library Preparation (Illumina FC-131-1096) using NexteraXT Index Kit V1 or V2 Set A (FC-131-1002 or FC-131-2001). NexteraXT DNA Library prep was done at either 1x, 0.5x, or 0.25x volume (applied to input and all reagents), but otherwise followed manufacturer's instructions. An aliquot of all samples was first normalized to 30pg/µl with Nuclease-Free Water (Ambion), then this normalized sample aliquot was used as input material into the NexteraXT DNA Library Prep. See **Table 2** for a summary of library prep conditions applied to the samples. AMPure XP bead purification was done using 0.9x bead ratio to sample volume, and all samples were eluted in 22µl of Resuspension Buffer (Illumina). As with the Amplification sets, manual bead purification was done for the first Library set, but thereafter bead purification was performed by Agilent Bravo NGS Option A instrument. All samples were run on Agilent Bioanalyzer 2100 using High Sensitivity DNA chips (for sizing), and all samples were quantitated using PicoGreen using Molecular Bynamics M2 SpectraMax instrument. Molarity was calculated for each sample using average size as reported by Bioanalyzer and pg/µl concentration as determined by PicoGreen. Samples (5µl aliquot) were normalized to 2-5nM with Nuclease-free Water (Ambion), then 2µl from each sample within one 96-index set was pooled to a total of 192µl at 2-5nM concentration. A portion of this library pool was sent to an outside vendor for sequencing on an Illumina HS2500. Most of the library pools were run using Illumina High Output V4 chemistry, although a few sets were sequenced using the Rapid Run V1 chemistry. Covance Genomics Laboratory, Seattle subsidiary of LabCorp Group of Holdings performed the majority of RNA-Sequencing services, with some also provided by EA Genomic Services. An average of 225M reads were obtained per pool, with an average of 2.4M reads/cell across the entire data set.

Table 2. Library prep conditions applied to the samples.

NexteraXT	Number of Samples	cDNA Input (pg)	Average Size (bp)	Average Yield (ng)	Average Yield (fmol)
1x	256	150	545	65	179
0.5x	520	75	523	46	135
0.25x	1056	37.5	443	36	125

RNA-Seq Data Processing

Raw read (fastq) files were aligned to the mm10 mouse genome sequence (Genome Reference Consortium, 2011) with the RefSeq transcriptome version GRCm38.p3 (current as of 01/15/2016) and updated by removing duplicate gene entries from the gtf reference file for consistency with LIMS). For alignment, Illumina sequencing adapters were clipped from the reads using the fastqMCF program (Aronesty et al., 2011). After clipping, the paired-end reads were mapped using RNA-Seq by Expectation-Maximization (RSEM) (Li et al., 2010) using default settings except for two mismatch parameters: bowtie-e (set to 500) and bowtie-m (set to 100). RSEM aligns reads to known isoforms and then calculates gene expression as the sum of isoform expression for a given gene, assigning ambiguous reads to multiple isoforms using a maximum likelihood statistical model. Reads that did not map to the transcriptome were then aligned to the mm10 genome sequence using Bowtie with default settings (Langmead et al., 2009). Reads that mapped to neither the transcriptome with RSEM or to the genome with Bowtie were mapped against the ERCC sequences. The final results files included quantification of the mapped reads (raw read counts, FPKM (Fragments Per Kilobase of transcript per Million mapped reads), and TPM (Transcripts Per Million) values for the transcriptome-mapped reads, and the number of genes detected (FPKM>1). Also, part of the final results files are the percentages of reads mapped to the Refseq transcriptome, to genomic regions not included in the Refseq transcriptome, to ERCC spike-in controls, and to ribosomal and mitochondrial RNA (see **Table 3**). Gene-level quantification files (TPM, FPKM, and number of reads) are available as part of the resource from the download page.

Table 3. QC metrics of LGd RNA-Seq data.

Initial QC Metrics	Slc17a6	Snap25	Gad2	Slc32a1	ERCC only	Control RNA*	No Template Control
Sample Count	527	783	221	284	36	88	46
Average Read Count	2.53M	2.33M	2.34M	2.37M	1.1M	2.2M	1.2M
Average % Mapped Reads	72.7%	74.6%	68.9%	71.9%	4.6%	58.8%	0.6%
Average % mRNA	50.4%	54.2%	53.9%	54.8%	0.2%	47.6%	0.4%
Average % gDNA	18.3%	16.4%	11.2%	13.2%	0.2%	5.3%	0.2%
Average % rRNA	0.3%	0.3%	0.5%	0.4%	0.1%	0.8%	0.0%
Average % ERCC	0.2%	0.3%	0.4%	0.5%	4.1%	0.7%	0.0%
Average ERCC Linearity	0.69				0.61	0.67	NA
ERCC Slope	0.90				0.92	0.91	NA
ERCC Limit of Detection	28.7				34.2	26.8	NA
Genes Detected (Raw FPKM>1)	9055	8738	7249	7562	361	5686	391
Exclusion Count	17	7	9	10	36	0	46

*Control RNA indicates 10 pg of Mouse Whole Brain Total RNA (Zyagen, MR-201).
Exclusion Count is all samples with <100,000 transcriptome mapped reads or <1000 genes detected (FPKM>0).

CLUSTERING

To visualize the single-cell data in a meaningful way, the single-cell data was clustered into groups using a consensus approach based on two iterative clustering techniques - iterative weighted gene co-expression network analysis (WGCNA) (as described in Tasic *et al.*, 2016) and an iterative version of Seurat (as described in Macosko *et al*., 2015). A two-layer classification was assigned for each cell: the first classification indicates the broad class (Three classes of GABAergic neurons, one glutamatergic class, one non-neuronal class, and one distinct class), whereas the second classification identifies putative subclasses within each of these broader divisions. The groups in the first and second classification were given names based on a set of marker genes that distinguish them. This two-layer classification data is represented as cell-level metadata in the resource. The Gad2_Syt4 class putatively comprises types outside of LGd, based on expression of *Syt4*, *Fxyd6*, and other genes not detected by *in situ* hybridization within the LGd. In addition, amplification quantification measures, mapped read counts, gene detection counts, and differential gene expression suggest that certain subclasses (in particular the Gad2_Sepp1, Lars2_Kcnmb1, Slc17a6_Pcdhgb4, and Slc17a6_Tcrb subclasses) may not represent biologically meaningful clusters of cells. However, these cells and groupings have been included for the sake of completeness.

REFERENCES

Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing data. Expression Analysis http://code.google.com/p/ea-utils.

Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, Foy C, Fuscoe J, Gao X, Gerhold DL, Gilles P, Goodsaid F, Guo X, Hackett J, Hockett RD, Ikonomi P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Kiser G, Koch WH, Lee KY, Liu C, Liu ZL, Lucas A, Manohar CF, Miyada G, Modrusan Z, Parkes H, Puri RK, Reid L, Ryder TB, Salit M, Samaha RR, Scherf U, Sendera TJ, Setterquist RA, Shi L, Shippy R, Soriano JV, Wagar EA, Warrington JA, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, Zadro R (2005) The External RNA Controls Consortium: a progress report. *Nature Methods* 2**:**731-734.

Gabbott PLA, Somogyi J, Stewart MG, Hámori J (1986) A quantitative investigation of the neuronal composition of the rat dorsal lateral geniculate nucleus using GABA-immunocytochemistry. *Neuroscience* 19:101–111.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 10:R25.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 26:493-500.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202-1214.

Piscopo DM, El-Danaf RN, Huberman AD, Niell CM (2013) Diverse visual features encoded in mouse lateral geniculate nucleus. *Journal of Neuroscience* 33:4642–4656.

Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* 32**:**896-902.

Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, Levi B, Gray LT, Sorensen SA, Dolbeare T, Bertagnolli D, Goldy J, Shapovalova N, Parry S, Lee C, Smith K, Bernard A, Madisen L, Sunkin SM, Hawrylycz M, Koch

C, Zeng H (2016) Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature Neuroscience* 19:335-346.