

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) and two mouse neocortical areas: primary visual cortex (VISp) and anterior lateral motor area (ALM). 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 different cortical areas 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 μ m) 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. For LGd, 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). The 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 *Sst+* interneuron subtypes (*Sst Cbln4* versus a mixture of *Sst Myh8*, *Sst Cdk6*, and *Sst Th*) recently identified in mouse primary visual cortex (Tasic *et al.* 2016).

An initial study of cell diversity in mouse primary visual cortex identified 42 neuronal types and 7 non-neuronal types (Tasic *et al.* 2016). The current cortical survey builds upon this initial study to provide a more comprehensive overview of cell types in VISp and comparison with a prefrontal cortical region (ALM). The sampling strategy for neocortex leveraged four mouse Cre lines to capture glutamatergic (*Snap25*, *Slc17a6*) and GABAergic (*Gad2*, *Slc32a1*) neurons from 4-5 layers of cortex (L1, L2/3, L4 (in V1 only), L5, and L6) based on their relative proportions *in vivo*. Additional cells were isolated from transgenic Cre recombinase lines selected to enrich for broad classes of cell types that were underrepresented in the initial 4 Cre lines. Similar to the preliminary study, the number of cells captured were estimates of what were needed to capture (with 95% confidence) at least 16 cells of a cell type as rare as 0.4% of glutamatergic neurons and 0.1% of GABAergic neurons.

TISSUE PREPARATION

Tissue samples were obtained from adult (postnatal day P53-P59) mice (males for LGd, and both males and females for cortical samples) carrying a Cre transgene and the tdTomato (*Ai14*) Cre-reporter transgene (see **Table 1** and **Table 2**). Mice were anesthetized with 5% isoflurane and intracardially perfused with either 25 or 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 volume of perfusate had been flushed through the vasculature. The brain was then rapidly dissected and mounted for coronal slice preparation (rostral end at base for LGd and VISp collections, caudal end at base for ALM collections) on the chuck of a Compressstome 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 μ m intervals. The slice was then hemisected along the midline, and the left hemisphere for LGd or both hemispheres for cortical samples were then 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 or a Sentech STC-SC500POE camera 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 five 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.

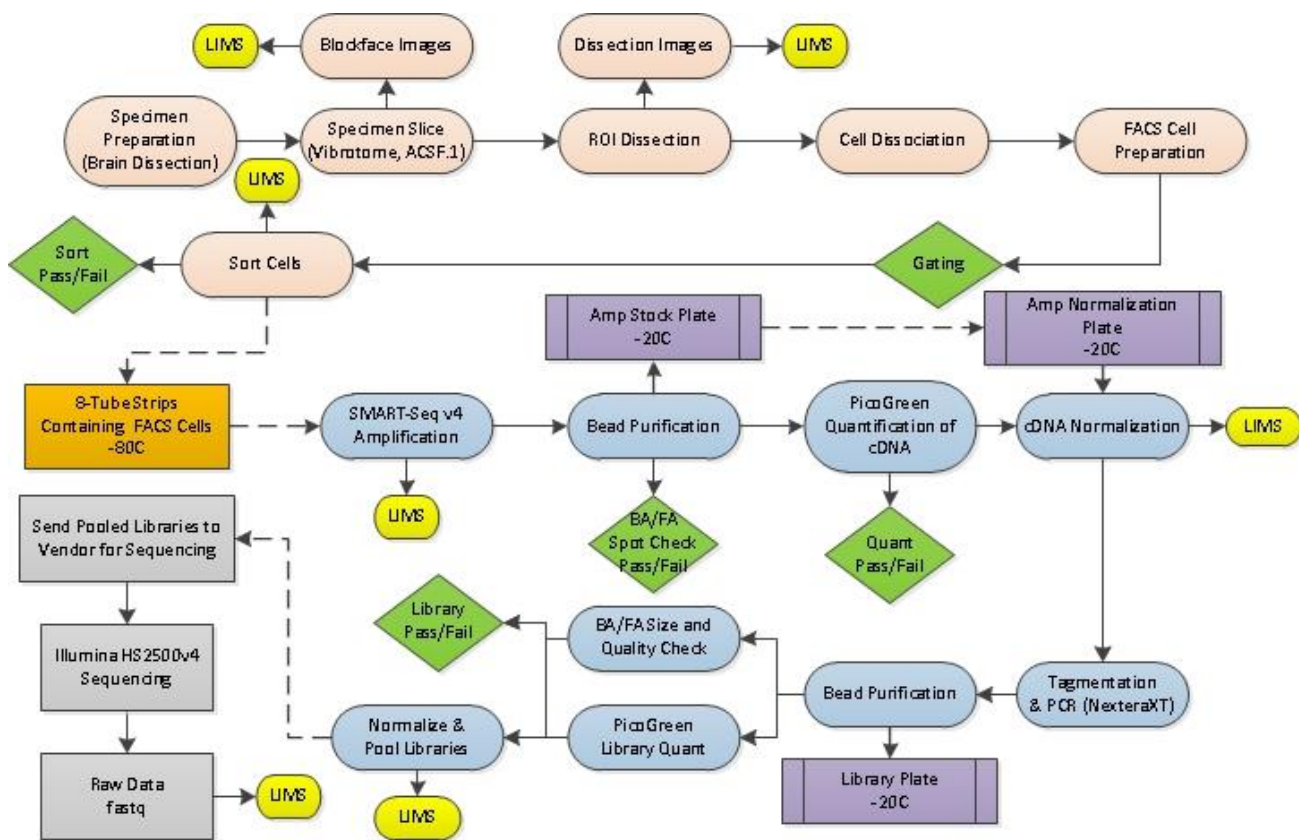


Figure 1. Workflow for tissue preparation and RNA-Seq data generation.

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/FA, Bioanalyzer/Fragment Analyzer; Quant, quantification.

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

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

Table 2. Mouse lines used for cortical samples.

Driver Line	Abbreviation
Chat-IRES-Cre	Chat
Chrna2-Cre_OE25	Chrna2
Chrna2-Cre_OE25_and_Pvalb-2A-Dre	Chrna2-Pvalb
Ctgf-2A-dgCre	Ctgf
Cux2-CreERT2	Cux2
Gad2-IRES-Cre	Gad2
Htr3a-Cre_NO152	Htr3a
Ndnf-IRES2-dgCre	Ndnf
Nos1-CreERT2	Nos1
Nos1-CreERT2_and_Sst-IRES-FlpO	Nos1-Sst
Nr5a1-Cre	Nr5a1
Ntsr1-Cre_GN220	Ntsr1
Pvalb-IRES-Cre	Pvalb
Rbp4-Cre_KL100	Rbp4
Scnn1a-Tg2-Cre	Scnn1a-Tg2
Sim1-Cre_KJ18	Sim1
Slc17a7-IRES2-Cre	Slc17a7
Slc32a1-IRES-Cre	Slc32a1
Snap25-IRES2-Cre	Snap25
Sst-IRES-Cre	Sst
Tlx3-Cre_PL56	Tlx3
Vip-IRES-Cre	Vip

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 with the third wash being 500 μ l for final sample volume. The sample was then triturated using fire-polished glass pipettes of decreasing bore sizes (600, 350 and 150 μ m for LGd, and 300 and 150 μ m for cortical samples). For triturating larger region of interest (ROI's), pipets with three decreasing bore sizes were used (600, 300 and 150 μ m). The cell suspension was incubated on ice in preparation for fluorescence-activated cell sorting (FACS).

FACS preparation involved adding 2 μ l of 4'-6-diamidino-2-phenylindole (DAPI) (2 mg/ml) to the 1.0 ml or 500 μ l ACSF.I-1% FBS cell suspension. The suspension was then filtered through a fine-mesh cell strainer (35 μ m) 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 11.5 μ l of collection buffer (SMART-Seq v4 lysis buffer 0.83x, (Clontech #634894), RNase Inhibitor (0.17U/ μ l) and ERCCs (External RNA Controls Consortium) (Baker *et al.*; Risso *et al.*) (MIX1 at 1×10^{-8})). After sorting, the cells were subjected to centrifugation and then stored at -80°C.

RNA SEQUENCING AND ANALYSIS OF LGd SAMPLES

RNA Amplification and Library Preparation for LGd 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 3** 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 Dynamics 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 3. Library prep conditions applied to the LGd 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

LGd 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 4**). Gene-level quantification files (TPM, FPKM, and number of reads) are available as part of the resource from the RNA-SEQ page.

Table 4. 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 of LGd RNA-Seq Data

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.

RNA SEQUENCING AND ANALYSIS OF CORTICAL SAMPLES

RNA Amplification and Library Preparation for RNA-Seq of Cortical Samples

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 11.5 μl of collection buffer (SMART-Seq v4 lysis buffer at 0.83x, RNase Inhibitor at 0.17 U/ μl , and ERCC MIX1 at final 1×10^{-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 either 4 wells of 10 pg of Mouse Whole Brain Total RNA (Zyagen, MR-201) or 2 wells of 10 pg of Mouse Whole Brain Total RNA (Zyagen, MR-201) and 2 wells of 10 pg Control RNA provided in the Clontech kit. AMPure XP Bead (Agencourt AMPure beads XP PCR, Beckman Coulter A63881) purification was done using the Agilent Bravo NGS Option A instrument. A bead ratio of 1x was used (50 μl of AMPure XP beads to 50 μl cDNA PCR product with 1 μl of 10x lysis buffer added, as per Clontech instructions), and purified cDNA was eluted in 17 μl 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 either run on the Agilent Bioanalyzer 2100 using High Sensitivity DNA chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment Analysis Kit (1bp-6000bp) to qualify cDNA size distribution. An average of 7.3 ng 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 V2 Set A (FC-131-2001). NexteraXT DNA Library prep was done at 0.5x volume (applied to input and all reagents), but otherwise followed manufacturer's instructions. An aliquot of all amplified cDNA samples was first normalized to 30 pg/ μl with Nuclease-Free Water (Ambion), then this normalized sample aliquot was used as input material into the NexteraXT DNA Library Prep (for a total of 75pg input). AMPure XP bead purification was done using the Agilent Bravo NGS Option A instrument. A bead ratio of 0.9x was used (22.5 μl of AMPure XP beads to 25 μl library product, as per Illumina protocol), and all samples were eluted in 22 μl of Resuspension Buffer (Illumina). All samples were run on either the Agilent Bioanalyzer 2100 using High Sensitivity DNA chips or the Advanced Analytics Fragment Analyzer (96) using the High Sensitivity NGS Fragment Analysis Kit (1bp-6000bp) to for sizing. All samples were quantitated using PicoGreen using Molecular Dynamics M2 SpectraMax instrument. Molarity was calculated for each sample using average size as reported by Bioanalyzer or Fragment Analyzer and pg/ μl concentration as determined by PicoGreen. Samples (5 μl aliquot) were normalized to 2-10 nM with Nuclease-free Water (Ambion), then 2 μl from each sample within one 96-index set was pooled to a total of 192 μl at 2-10 nM concentration. A portion of this library pool was sent to an outside vendor for sequencing on an Illumina HS2500. All of the library pools were run using Illumina High Output V4 chemistry. Covance Genomics Laboratory, Seattle subsidiary of LabCorp Group of Holdings performed the RNA-Sequencing services. An average of 229 M reads were obtained per pool, with an average of 2.0-3.1 M reads/cell across the entire data set.

RNA-Seq Data Processing of Cortical Samples

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 Entrez gene entries from the gtf reference file for RSEM processing. 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 and the *E.coli* genome (version ASM584v2). The final results files included quantification of the mapped reads (raw read

counts, FPKM, and TPM 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 *E.coli* (see **Tables 5-9, 11**). Gene-level quantification files (TPM, FPKM, and number of reads) are available as part of the resource from the download page.

Table 5. QC metrics of VISp pan-neuronal lines.

	Slc17a7	Snap25	Gad2	Slc32a1
Sample Count	1263	2720	1457	1208
Total Reads	2446891	2553122	2558457	2630308
Reads Aligned	1964791	1992804	2006806	2072548
Reads mRNA	1569210	1545745	1644222	1674312
Reads gDNA	297262	338885	264857	296783
Reads ERCC	4452	4314	5384	5043
Unmapped Reads	482100	560318	551651	557759
Reads <i>E.coli</i>	8802	19556	34182	24862
% Aligned	80.23%	77.96%	78.22%	78.51%
% mRNA	64.00%	60.52%	63.99%	63.24%
% gDNA	12.20%	13.21%	10.39%	11.40%
% ERCC	0.18%	0.17%	0.22%	0.19%
% Unmapped	19.77%	22.04%	21.78%	21.49%
% <i>E.coli</i>	0.37%	0.81%	1.44%	0.99%
FPKM>0	9982	9765	8882	8859
FPKM>1	8352	8118	7296	7231
FPKM>4	6962	6736	6129	6048
FPKM>8	5938	5760	5331	5271
FPKM>16	4518	4442	4227	4228
FPKM>32	2985	2975	2925	2960
FPKM>64	1769	1765	1784	1812
Excluded Cells	9	12	14	2

Excluded cells are samples removed from totals and averages.

Table 6. QC metrics of VISp targeted lines.

	Chat	Chrna	Chrna2-Pvalb	Ctgf	Ndnf	Nos1	Nos1-Sst	Nr5a1	Pvalb	Rbp4	Scnn1a-Tg2	
Sample Count	36	96	4	206	119	35	5	151	255	732	41	
Mean Values	Total Reads	2748104	2762254	2443737	2718094	2874604	2450097	2995074	2483824	3086232	2294452	2429437
	Reads Aligned	2180336	2284101	1741280	2191973	2237526	1905000	2527767	1886383	2488574	1838976	1808235
	Reads mRNA	1726233	1953967	1169959	1871951	1854011	1535222	2261341	1334716	2076584	1474141	1164935
	Reads gDNA	337886	232798	453488	223048	269262	273996	175982	437656	294730	279086	536279
	Reads ERCC	7377	4278	6730	9236	13661	4414	1704	4869	5406	3817	3662
	Unmapped Reads	567768	478154	702457	526121	637078	545097	467307	597441	597658	455476	621202
	Reads <i>E.coli</i>	14351	4418	24064	20903	42273	7652	3120	7319	7600	3816	4600
	% aligned	79.20%	82.72%	70.78%	80.33%	77.65%	77.43%	84.38%	75.86%	81.15%	79.88%	74.01%
	% mRNA	62.70%	70.85%	46.98%	68.60%	64.37%	62.20%	75.44%	53.67%	68.48%	64.06%	47.54%
	% gDNA	12.27%	8.35%	18.98%	8.13%	9.31%	11.30%	5.90%	17.60%	9.00%	12.11%	22.06%
	% ERCC	0.27%	0.16%	0.29%	0.35%	0.49%	0.19%	0.06%	0.20%	0.17%	0.17%	0.15%
	% Unmapped	20.80%	17.28%	29.23%	19.67%	22.35%	22.57%	15.62%	24.14%	18.85%	20.12%	25.99%
	% <i>E.coli</i>	0.53%	0.16%	1.04%	0.82%	1.53%	0.36%	0.11%	0.30%	0.23%	0.16%	0.19%
	FPKM>0	8567	9350	8473	8332	7094	8381	10797	9377	10077	10376	10031
	FPKM>1	6907	7667	7007	6811	5578	6836	8846	7888	8253	8765	8553
	FPKM>4	5849	6453	5600	5656	4770	5677	7070	6561	6844	7194	6967
	FPKM>8	5205	5575	4704	4792	4262	4914	5588	5651	5777	6014	5843
	FPKM>16	4343	4383	3575	3679	3545	3952	3993	4416	4351	4474	4407
	FPKM>32	3118	2978	2412	2618	2628	2836	2641	2956	2900	2928	2865
	FPKM>64	1872	1822	1490	1733	1766	1806	1616	1695	1769	1742	1607
Excluded Cells	0	0	4	1	1	2	0	2	4	21	0	

Excluded cells are samples removed from totals and averages.

Table 7. QC metrics of ALM pan-neuronal lines.

	Slc17a7	Snap25	Gad2	Slc32a1	
Sample Count	1102	1337	753	1148	
Mean Values	Total Reads	3038102	3010972	2851157	2908923
	Reads Aligned	2514685	2450003	2322636	2404344
	Reads mRNA	2124801	2038308	1913339	2030904
	Reads gDNA	269477	293883	289049	265883
	Reads ERCC	4992	5121	11726	7171
	Unmapped Reads	523417	560968	528522	504579
	Reads <i>E.coli</i>	4899	10292	10236	8674
	% Aligned	82.68%	81.11%	81.46%	82.56%
	% mRNA	3.52%	3.40%	3.56%	69.59%
	% gDNA	9.05%	9.65%	10.15%	9.27%
	% ERCC	0.17%	0.18%	0.40%	0.25%
	% Unmapped	17.32%	18.89%	18.54%	17.44%
	% <i>E.coli</i>	0.18%	0.41%	0.36%	0.30%
	FPKM>0	10671	10175	9031	9316
	FPKM>1	8783	8355	7373	7614
	FPKM>4	7143	6869	6183	6414
	FPKM>8	5898	5755	5340	5550
	FPKM>16	4391	4356	4206	4361
	FPKM>32	2929	2930	2916	2998
	FPKM>64	1791	1773	1804	1826
Excluded Cells	14	7	1	13	

Excluded cells are samples removed from totals and averages.

Table 8. QC metrics of ALM targeted lines.

	Cux2	Htr3a	Ntsr1	Pvalb	Rbp4	Sim1	Sst	Tlx3	Vip
Sample Count	63	152	153	266	384	102	198	103	229
Total Reads	2038998	2771546	2384628	2567188	2349716	2171022	2157132	2472373	1979065
Reads Aligned	1527633	2163340	1649583	2084748	1874562	1611815	1556635	1902595	1368020
Reads mRNA	1153870	1710518	1201220	1741077	1509638	1259778	1147409	1529012	974669
Reads gDNA	283344	340268	320423	251742	265968	263276	319980	275320	307479
Reads ERCC	1339	3155	4672	2404	2359	1218	1406	1493	1996
Unmapped Reads	511365	608206	735045	482440	475154	559207	600498	569778	611045
Reads <i>E.coli</i>	33563	14259	114278	8041	12770	40973	66935	13370	100647
% Aligned	74.88%	78.05%	67.79%	81.16%	79.55%	74.29%	71.81%	76.77%	68.05%
% mRNA	56.43%	61.75%	49.33%	67.69%	64.09%	58.08%	52.74%	61.78%	48.01%
% gDNA	14.01%	12.24%	13.17%	9.88%	11.27%	12.13%	14.91%	11.02%	15.69%
% ERCC	0.06%	0.11%	0.22%	0.09%	0.10%	0.06%	0.06%	0.06%	0.10%
% Unmapped	25.12%	21.95%	32.21%	18.84%	20.45%	25.71%	28.19%	23.23%	31.95%
% <i>E.coli</i>	1.63%	0.51%	5.63%	0.34%	0.60%	1.86%	3.22%	0.62%	5.46%
FPKM>0	11273	9133	6385	10073	11015	11643	9704	11936	8225
FPKM>1	9619	7378	5107	8351	9304	9957	8193	10009	6972
FPKM>4	7699	6155	4182	6913	7513	7933	6654	7862	5755
FPKM>8	6281	5343	3629	5831	6147	6392	5622	6238	5015
FPKM>16	4555	4250	2970	4392	4492	4569	4329	4419	4065
FPKM>32	2933	2957	2255	2924	2913	2920	2933	2852	2929
FPKM>64	1720	1801	1580	1766	1737	1715	1752	1712	1829
Excluded Cells	1	0	14	9	0	1	2	1	2

Excluded cells are samples removed from totals and averages.

Table 9. QC metrics of controls data.

	Control Total RNA	Mouse Whole RNA	ERCC	NTC
Sample Count	113	535	332	332
Total Reads	2692357	2354761	893862	731014
Reads Aligned	1902720	1381075	117187	3752
Reads mRNA	1525491	848088	8257	2550
Reads gDNA	240574	452539	2303	990
Reads ERCC	22247	10758	105867	12
Unmapped Reads	789638	973686	776675	727262
Reads <i>E.coli</i>	14139	59178	211929	232413
% Aligned	70.37%	57.06%	15.46%	0.78%
% mRNA	56.31%	36.72%	1.11%	0.55%
% gDNA	9.00%	16.91%	0.29%	0.18%
% ERCC	0.82%	0.45%	13.96%	0.00%
% Unmapped	29.63%	42.94%	84.54%	99.22%
% <i>E.coli</i>	0.53%	3.32%	23.22%	31.09%
FPKM>0	8179	6920	437	407
FPKM>1	6566	5714	426	404
FPKM>4	5474	4559	407	396
FPKM>8	4811	3933	392	389
FPKM>16	3982	3221	373	375
FPKM>32	2970	2415	345	351
FPKM>64	1833	1614	306	314
Excluded Cells	1	12	332	332

Excluded cells are samples removed from totals and averages.
Abbreviation: NTC, no template control.

Classification of Cortical Cells into Broad Types

Cells were classified as either excitatory (*exc*) or inhibitory (*inh*) neurons or non-neuronal (*non*) based on canonical marker gene expression (see **Table 10**). Cells were classified as excitatory if the maximum expression of excitatory genes (*Slc17a6*, *Slc17a7*) was greater than the maximum expression of inhibitory (*Gad1*, *Gad2*, *Slc32a1*) or non-neuronal (*Olig1*, *Gja1*, *Xdh*, *Ctss*, *Myl9*) genes. Cells were classified as inhibitory if the maximum expression of inhibitory (*Gad1*, *Gad2*, *Slc32a1*) genes was greater than the maximum expression of excitatory (*Slc17a6*, *Slc17a7*) or non-neuronal (*Olig1*, *Gja1*, *Xdh*, *Ctss*, *Myl9*) genes. All remaining cells were classified as non-neuronal. The QC metrics based on classification by broad type are shown in **Table 11**.

Table 10. Gene expression criteria for cell type classification.

Broad Types	Gene Expression Criteria
Glutamatergic	max(<i>Slc17a6</i> , <i>Slc17a7</i>) > max(<i>Gad1</i> , <i>Gad2</i> , <i>Olig1</i> , <i>Gja1</i> , <i>Xdh</i> , <i>Ctss</i> , <i>Myl9</i> , <i>Slc32a1</i>)
GABAergic	max(<i>Gad1</i> , <i>Gad2</i> , <i>Slc32a1</i>) > max(<i>Slc17a7</i> , <i>Slc17a6</i> , <i>Olig1</i> , <i>Gja1</i> , <i>Xdh</i> , <i>Ctss</i> , <i>Myl9</i>)
Non-neuronal	All remaining cells

Table 11. QC metrics by classification.

	Class	Glutamatergic	GABAergic	Non-neuronal
	Cell Wells	7107	6704	508
Mean Values	Total Reads	2620454	2695690	2521028
	Reads Aligned	2094133	2157913	1796635
	Reads mRNA	1678994	1775786	1351499
	Reads gDNA	308941	280089	295819
	Reads ERCC	4106	5349	17635
	Unmapped Reads	526322	537778	724393
	Reads <i>E.coli</i>	11672	21392	82688
	% aligned	79.66%	79.68%	70.18%
	% mRNA	63.68%	65.34%	52.77%
	% gDNA	11.93%	10.54%	11.55%
	% ERCC	0.16%	0.20%	0.70%
	% Unmapped	20.34%	20.32%	29.82%
	% <i>E.coli</i>	0.49%	0.91%	3.76%
	FPKM>0	10343	9212	5436
	FPKM>1	8627	7563	4184
	FPKM>4	7086	6332	3481
	FPKM>8	5944	5476	3107
	FPKM>16	4468	4316	2674
	FPKM>32	2951	2975	2178
	FPKM>64	1760	1807	1607
Excluded Cells	26	46	65	

Excluded cells are samples removed from totals and averages.
There are 14,319 total cells.

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.