BRAINSPAN

ATLAS OF THE DEVELOPING HUMAN BRAIN

TECHNICAL WHITE PAPER: TRANSCRIPTOME PROFILING BY RNA SEQUENCING AND EXON MICROARRAY

BrainSpan, an atlas of the developing human brain, is designed as a foundational resource for studying transcriptional mechanisms involved in human brain development. It is the outcome of three ARRA-funded grants through the National Institutes of Health supporting a consortium consisting of the Allen Institute for Brain Science; Yale University (Nenad Sestan, Mark B. Gerstein); the Zilkha Neurogenetic Institute of the Keck School of Medicine of the University of Southern California (James A. Knowles, Pat Levitt); the Athinoula A. Martinos Center at Massachusetts General Hospital/Harvard Medical School and MIT HST/CSAIL (Bruce Fischl); the University of California, Los Angeles (Daniel H. Geschwind); and the University of Texas Southwestern Medical Center (Hao Huang) with strong collaborative support from the Genes, Cognition and Psychosis Program, which is part of the Intramural Research Program of NIMH, NIH (Thomas M. Hyde, Joel E. Kleinman, Daniel R. Weinberger). All data are publicly accessible via the ALLEN BRAIN ATLAS data portal or directly at www.developinghumanbrain.org.

Two data modalities included in this atlas are RNA sequencing (RNA-Seq) and exon microarray hybridization. These data were generated across 13 developmental stages in 8-16 brain structures. This white paper describes the methods and processes used to generate these gene expression data. Other methods are described in separate technical white papers available via the Documentation tab.

TISSUE ACQUISITION AND QUALIFICATION

This work used postmortem human brain specimens from tissue collections of the Department of Neurobiology at Yale School of Medicine and the National Institute of Mental Health. In addition, specimens were procured from the Human Fetal Tissue Repository at the Albert Einstein College of Medicine, the Brain and Tissue Bank for Developmental Disorders at the University of Maryland, the Birth Defects Research Laboratory at the University of Washington, Advanced Bioscience Resources Inc., and the MRC-Wellcome Trust Human Developmental Biology Resource at the Institute of Human Genetics, University of Newcastle, U.K.

Table 1. Developmental stages profiled.

Stage	Age	Developmental Period
1	4-7 pcw	Embryonic
2A	8-9 pcw	Early prenatal
2B	10-12 pcw	Early prenatal
3A	13-15 pcw	Early mid-prenatal
3B	16-18 pcw	Early mid-prenatal
4	19-24 pcw	Late mid-prenatal
5	25-38 pcw	Late prenatal
6	Birth-5 months	Early infancy
7	6-18 months	Late infancy
8	19 months-5 yrs	Early childhood
9	6-11 yrs	Late childhood
10	12-19 yrs	Adolescence
11	20-60+ yrs	Adulthood

All work was performed according to guidelines for the research use of human brain tissue and with approval by the Human Investigation Committees and Institutional Ethics Committees of each institute from which samples were obtained. Appropriate written informed consent was obtained and all available non-identifying information was recorded for each sample.

Whenever possible, one whole or both hemispheres were collected. Postmortem brains were collected from individuals ranging from 5 to 7 post-conceptional weeks (pcw) to over 40 years of age (see **Table 1**). To ensure that all developmental and adult stages were included, samples were categorized into different stages based upon age and major neurodevelopmental milestones. Prenatal age was extrapolated based upon the date of the mother's last menstruation, characteristics of the fetus noted upon ultrasonographic scanning, and foot length of the fetus.

Selection criteria for tissue qualification

To better ensure consistency between samples and to decrease potential variation due to ante- and postmortem conditions, specific selection criteria were established. The following selection criteria were strictly adhered to:

- Brains with aneuploidy and other large-scale chromosomal abnormalities, detected by karyogram and/or Illumina Human Omni-2.5, were excluded.
- Prenatal and neonatal specimens were excluded if drug or alcohol abuse by the mother during pregnancy was reported or if potassium chloride, salt water, or urea were injected into the amniotic sac during surgical procedure.
- Only brains with no evidence of malformations or lesions were included. Disqualifying characteristics included any abnormality of the neural tube, forebrain, brainstem, cranial nerves, cerebellum, or spinal cord (*i.e*., prominent intraparenchymal hemorrhage, periventricular leukomalacia, abnormal meninges, dysplasia, hypoplasia, alterations in the pial or ventricular surface, white matter lesions).
- Samples were excluded if microscopic analysis revealed clear evidence of neuronal loss, neuronal swelling, glioneuronal heterotopias, or dysmorphic neurons and neurites.
- Samples that tested positive for Hepatitis B, Hepatitis C, or HIV were excluded.
- Postnatal and later stage specimens were excluded if excessive drug or alcohol abuse was reported, if the individual had any known neurological or psychiatric disorders, or if any prolonged agonal conditions (coma, hypoxia, prolonged pyrexia, seizures, prolonged dehydration, hypoglycemia, and multiple organ failure) were reported. Other excluding factors included ingestion of neurotoxic substances at the time of death, suicide, severe head injury, significant hemorrhages, prominent vascular abnormalities, tumors, prominent brain lesions, stroke, congenital neural abnormalities, and signs of neurodegeneration (spongiosis, amyloid plaques, Lewy bodies or amyloid angiopathy).

TISSUE PROCESSING AND DISSECTION

Depending on the condition and stage of the procured specimens, four different macrodissection methods were used. All dissections were video documented using a 10 megapixel digital camera (stage 1-2A) or a Sony HDR-CX150 HD-camera (stages 2B-11).

- Neocortical regions sampled from stage 2B-11 prenatal and postnatal brain specimens included part of the underlying subplate zone and white matter, respectively.
- Neocortical regions collected from all prenatal brain samples included the subplate zone.
- Generative zones were not collected from stage 2B-11 prenatal brains.
- Only neocortical regions from stage 1 and 2A contain the complete neocortical anlage.

Table 2 provides a complete list of all regions collected. Specific dissection protocol depended upon the stage of the sample and the method by which it was preserved. For all brain specimens procured from the Department of Pathology at Yale University School of Medicine and the Human Fetal Tissue Repository at the

AECOM, regions of interest were collected from fresh tissue. For all other cases, regions were collected from frozen tissue slabs or whole specimens stored at -80ºC. To ensure consistency between specimens, all dissections were performed by Nenad Sestan.

Table 2. Brain regions profiled by RNA sequencing and exon microarray technologies.

Stage 1	Stage 2A	Stages 2B-11
Frontal (Anterior) cortical wall	Frontal cortical wall	Dorsolateral prefrontal cortex (FC) (BA9, 46)
		Ventrolateral FC (BA44, 45)
		Medial FC (BA32, 33, 34)
		Orbital FC (BA11)
		Primary motor cortex (M1) (BA4)
Parietal (Middle dorsal) cortical wall	Parietal cortical wall	Primary somatosensory cortex (S1) (BA1-3)
		Posterior inferior parietal cortex (BA40)
Temporal (Postero-lateral) cortical wall	Temporal cortical wall	Primary auditory temporal cortex (A1) (BA41)
		Posterior superior temporal cortex (BA22)
		Inferior temporal cortex (BA20)
Occipital (Posterior) cortical wall	Occipital cortical wall	Primary visual cortex (V1) (BA17)
Hippocampal anlage	Hippocampal anlage	Hippocampus
		Amygdala
Ventral forebrain	Medial ganglionic eminence	Striatum
	Lateral ganglionic eminence	
	Caudal ganglionic eminence	
Dorsal diencephalon	Dorsal thalamic anlage	Mediodorsal nucleus of thalamus
Upper rhombic lip	Upper rhombic lip	Cerebellar cortex

Neuropathological evaluation

All brain specimens, tissue slabs, or tissue sections were comprehensively evaluated by Anita Huttner, M.D., Nenad Sestan, M.D., Ph.D., and Alexander Vortmeyer, M.D., Ph.D., to exclude any confounding pathological factors such as hypoxia, cerebrovascular incidents, tumors, microbial infections, neurodegeneration, demyelination, and metabolic disease.

To prepare tissue sections, samples of the fresh or frozen tissue were fixed in 4% paraformaldehyde, processed and paraffin-embedded. Each sample was sectioned at 5 µm using a microtome. Sections were mounted on glass slides and stained with hematoxylin and eosin. Selected slides were immunostained using the following antibodies: glial fibrillary acidic protein (DAKO; polyclonal; 1:3000), Ki67 (MIB-1; DAKO; 1:400), synaptophysin (ICN; clone SY38; 1:100), Neu-N (Chemicon; 1:10000), and CD45 (DAKO; clone 2B11; 1:250). The Envision Plus detection system (DAKO) was used to visualize all antibody staining.

Histological verification of dissected regions

Small frozen or fresh tissue blocks adjacent to the region of interest were occasionally collected and fixed in 4% paraformaldehyde for 48 hours. Tissue blocks were sectioned at 50 mm thickness using a vibratome and Nissl stained to verify the cytoarchitecture of dissected regions.

Regional sampling

Fresh tissue

Brains were chilled on ice for 15-30 minutes prior to sectioning. Brains were placed ventral side up onto a chilled aluminum plate (1 cm thick) on ice. The brainstem and cerebellum were removed from the cerebrum by making a transverse section at the junction between the diencephalon and midbrain. Next, the cerebrum was divided into left and right hemispheres by cutting along the midline using a long scalpel. The cerebellum was separated from the brainstem by making a cut directly posterior to the brainstem, along the cerebellar peduncles. The regions of interest were dissected using a scalpel blade and immediately frozen in liquid nitrogen. Dissected samples were either immediately processed for RNA extraction or stored at -80°C for later RNA extraction. The remaining brain tissue was cut to obtain 1 cm (for late prenatal and postnatal specimens) or 0.5 cm (for prenatal) thick serial, coronal sections. The tissue slabs were snap frozen in isopentane/dry ice at -30 to -40 ºC and stored at -80ºC.

Frozen tissue

All previously frozen stage 2B to 11 specimens and tissue slabs were microscopically inspected and the desired region was demarcated, then dissected using a dental drill and a Lindemann Bone Cutter H162A.11.016 or diamond disk saw (Dental Burs USA; r=11 mm) on a 1 cm thick aluminum plate over dry ice. Dissected tissue samples were stored in a small freezer bags at -80ºC prior to further processing.

Tissue processed in RNAlater

Frozen stage 1 and 2A specimens were sectioned coronally at approximately 500µm, beginning at the frontal pole using a dental diamond disk saw. For gradual thawing, tissue slabs were transferred from -80°C storage to overnight storage at -20°C in RNA*later* ICE (Ambion). Tissue slabs were visually inspected for neuropathological defects. Next, tissue was microdissected at 4°C under a dissection microscope, and stored at 4°C in a lysis buffer for RNA extraction, which occurred within 1 to 2 hours.

Dissection scoring

A scoring system, described in **Table 3**, was developed to ensure consistency between individual samples of each dissected brain region and to maintain a record of whether the whole region or just an area of interest from a region had been collected.

Table 3. Dissection scoring scheme.

Score	Sample Description
1-2	The region/area of interest was largely absent and thus not collected.
3	The region/area of interest was not complete, but was of suitable quality to collect.
4	The region/area of interest was largely intact but was not histologically verified or could not be collected at precisely the same position from which other corresponding contralateral sample was collected. This was primarily an issue with neocortex, hippocampus, and cerebellar samples.
5	The region/area of interest was fully intact, verified by gross inspection or Nissl staining, and was collected at precisely the same position as corresponding samples.

Tissue pulverization

Frozen tissue samples were pulverized in liquid nitrogen using a ceramic mortar and pestle. Pulverized samples were transferred to wide-mouth cryogenic vials (Nalgene, cat# 03-337-7B) and stored at -80ºC until used for RNA extraction.

RNA AND DNA ANALYSIS

Genotyping and genotype data quality control

Genotyping was done using Illumina Omni-2.5 million SNP arrays. For genotyping analysis, 25mg of brain tissue was homogenized by bullet blender (Next Advance) and lysed in the ATL buffer at 56°C for 3 to 4 hrs. Genomic DNA was isolated from human brains using a nonphenolic procedure (DNeasy Blood & Tissue Kit, Qiagen) with proteinase K and RNase A treatment (Qiagen). The purified DNA sample was quantitated by Fluorospectrophotometry (Thermo Scientific, NanoDrop 3300) using PicoGreen dsDNA assay kit (Invitrogen) and the integrity of DNA was confirmed by agarose gel electrophoresis. Processing of DNA samples was performed according to the Infinium HD Assay Super, Automated Protocol for Human Omni 2.5-Quad Bead Chip (Illumina).

RNA extraction

A bead mill homogenizer (Bullet Blender, Next Advance) was used to lyse the pulverized tissue. Each pulverized tissue sample was transferred to a chilled microcentrifuge tube (safe-lock tubes, Eppendorf). A mass of chilled stainless steel beads (cat# SSB14B, Next Advance) equal to the mass of the tissue was added to the tube. Two volumes of Buffer RLT (Qiagen) were added to the tissue and beads. Samples were mixed in the Bullet Blender for 1 minute at a speed of six. Samples were visually inspected to confirm perfect homogenization, then incubated at 37°C for five minutes. Buffer RLT was added up to 0.6 ml, and samples were mixed in the Bullet Blender for an additional minute. Total RNA was extracted using a non-phenolic procedure (RNeasy Plus Mini Kit, Qiagen), followed by DNase treatment (TURBO DNase, Ambion) as per manufacturers' instructions. Optical density values of extracted RNA were measured by NanoDrop (Themo Scientific) to confirm the ratio of 260/280 was greater than 1.9. RNA quality was confirmed using Bioanalyzer RNA 6000 Nano Kit or Bioanalyzer RNA 6000 Pico Kit (Agilent), depending upon the total amount of RNA.

mRNA LIBRARY PREPARATION AND SEQUENCING

To purify polyA RNA from total RNA, total RNA (5 µg) was diluted with nuclease-free water to 50 µl in a 1.5ml RNase-free non-sticky tube. Total RNA was heated at 65°C for 5 minutes and then placed on ice for 1 minute. Next, 15 µl of (dT) beads were aliquoted into a 1.5 ml RNase-free non-sticky tube. On the magnetic stand, beads were washed twice with 100 µl of Binding Buffer. Then, the beads were resuspended with 50 µl of Binding Buffer, and 50 µl of total RNA was subsequently added. The sample was rotated at room temperature for 5 minutes and the supernatant was removed using a magnetic stand. Next, the sample was washed twice with 200 µl of Washing Buffer. Then, 50 µl of 10mM Tris-HCl was added to the sample and heated at 80°C for 2 minutes. Immediately thereafter, the tube was placed on the magnetic stand and the supernatant was transferred to Binding Buffer (50 µl), and then heated at 65°C for 5 minutes. While the samples were incubating at 65°C, the beads were washed twice with 200 µl of Washing Buffer and the supernatant was removed. Following the 65°C incubation, samples were placed on ice for 1 minute. Next, 100 µl of buffer/sample was added to the beads and rotated for 5 minutes. Then, the supernatant was removed and beads were washed twice with 200 µl of Washing Buffer. After removing the Washing Buffer, 15.5 µl of 10 mM Tris-HCl was added to beads and heated at 80°C for 2 minutes. Then, the tube was immediately put on the magnetic stand and the supernatant was transferred to a fresh 200 µl thin-wall PCR tube for quantitation.

Quaint-IT RiboGreen RNA Assay Kit (Invitrogen) was used to quantitate purified mRNA with the NanoDrop 3300. Following mRNA quantitation, 2.5 µl spike-in master mixes were added per 100 ng of mRNA and nuclease-free water was added to bring the total volume to 16 µl. The addition of RNA spike-in sequences will enable normalization of expression levels between samples. The spike-in RNA mixes are unique for each brain region, and hence also provide a molecular barcode of the sample to enable the detection of sample mix-ups in the subsequent stages of the protocol. Although we have added these to all samples, version 1 of the analysis pipeline does not utilize this information; hence RPKM data released in version 1 was not normalized to these spike-ins.

Prior to cDNA synthesis, the RNA was fragmented to provide a more even distribution of reads across the length of the RNA molecules. To 16 µl of mRNA, 4 µl of 5X Fragmentation Buffer was added and incubated at 94°C for 5 minutes. Next, 2 µl of Stop Buffer was added and the tube was placed on ice for 1 minute. Next, the solution was transferred to a 1.5 ml RNase-free non-sticky tube. The following reagents were added to the tube and incubated at -80°C for 30 minutes: 2 µl of 3M NaOAC, pH 5.2, 2 µl of glycogen, and 60 µl of 100% EtOH (-20°C). Next, the sample was centrifuged at 14,000 rpm (20,200 rcf) for 25 minutes at 4°C. After carefully removing the ethanol by pipetting, the pellet was washed with 300 µl of 70% EtOH (-20°C) and centrifuged at 14,000 rpm for 5 minutes at 4°C. The ethanol was removed by pipetting. Next, the pellets were subjected to a Speed Vac for 5 minutes. Then, RNA was resuspended in 11.1 µl of RNase-free water.

For the first strand cDNA synthesis, in a 200 µl thin wall PCR tube, 11.1 µl mRNA and 1 µl of random primers were added. After incubation at 65°C for 5 minutes, the tube was placed on ice for 1 minute. Then, a master mix was made consisting of the following volume per sample: 4 µl of 5X First Strand Buffer (Invitrogen), 2 µl of 100 mM DTT (Invitrogen), 0.4µl of 25 mM dNTP mix, and 0.5µl of RNase Inhibitor for a total master mix volume of 6.9 µl per sample. A 6.9 µl aliquot was added to each sample, mixed well, and heated at 25°C for 2 minutes. Next, 1 µl of SuperScript II (Invitrogen) was added to the sample and incubated in a thermal cycler for 25°C for 10 minutes, 42°C for 50 minutes, and 70°C for 15 minutes, followed by 4°C. The tube was then placed on ice during preparation of the reagents for the second strand cDNA synthesis.

For the second strand cDNA synthesis, 62.8 µl of water was added to the first strand cDNA synthesis reaction. Then, 10 µl of GEX Second Strand Buffer and 1.2 µl of 25 mM dNTP mix were added, mixed well, and incubated on ice for 5 minutes. Next, 1 µl of RNaseH and 5 µl of DNA Pol I were added for a total volume of 100 µl. The reaction was mixed well and incubated at 16°C for 2.5 hours. The purified DNA was eluted in 50 µl using QIAquick Purification Kit.

Next, the ends from the second strand cDNA synthesis reaction were repaired. The following reagents were prepared in a 1.5 ml RNase-free non-sticky tube: 50 µl of eluted DNA, 27.4 µl water, 10 µl of 10X End Repair Buffer, 1.6 µl 25 mM dNTP Mix, 5 µl of T4 DNA Polymerase, 1 µl of Klenow DNA Polymerase, and 5 µl T4 PNK (10U/µl) for a total volume of 100 µl. The sample was incubated in a heat block at 20°C for 30 minutes. The purified DNA was eluted in 32 µl using QIAquick Purification Kit.

The next step involved adding 'A' Bases to the 3' end of the DNA fragments. The following reagents were added to a 1.5 ml RNase-free non-sticky tube: 32 µl eluted DNA, 5 µl A-Tailing Buffer, 10 µl 1mM dATP, and 3 µl Klenow Exo- for a total volume of 50 µl. The sample was incubated in a heat block at 37°C for 30 minutes. The purified DNA was eluted in 23 µl using the MinElute Purification Kit. Following DNA purification, adapters were ligated to the DNA fragments. The following reagents were added to a 1.5 ml RNase-free non-sticky tube: 23 µl eluted DNA, 25 µl 2X Rapid T4 DNA Ligase Buffer, 1 µl PE Adapter Oligo Mix, and 1 µl T4 DNA Ligase (600U/µl) for a total volume of 50 µl. The sample was incubated at room temperature for 15 minutes and the DNA was purified in 10 µl using MinElute Purification Kit.

For cDNA template purification, four samples were loaded per one E-Gel or 2% agarose gel. DNA between 200 to 250 bp was collected. Samples were then purified using QIAquick PCR purification kit. Alternatively, the appropriate region of the gel was excised (~225 bp) with a clean razor blade. The purified DNA was eluted to 30 µl using QIAquick Gel Extraction Kit. Next, the gel-purified cDNA templates were enriched by PCR. The following master mix and thermal cycler conditions were used: 10 µl of 5X Phusion Buffer HF, 1 µl of PCR Primer PE 1.0, 1 µl of PCR Primer PE 2.0, 0.5 µl of 25 mM dNTP mix, 0.5 µl of Phusion DNA Polymerase (20U/10 µl), and 7 µl of water for a total volume of 20 µl. Next, 30 µl of cDNA template mix was added to the 200 µl PCR tube and subjected to the following parameters: 30 seconds at 98°C; followed by 15 cycles of 10 seconds at 98°C, 30 seconds at 65°C, and 30 seconds at 72°C; with a final 5 minute extension at 72°C, followed by 4°C.

For quantifying the libraries, DNA dye concentrate and DNA gel matrix (Agilent DNA1000 kit) were equilibrated to room temperature for 30 minutes prior to use. DNA dye concentrate was vortexed and 25 µl of dye was added to a DNA gel matrix vial. The solution was vortexed well, spun down, transferred to a spin filter, and centrifuged at 2240*xg* ± 20% for 15 minutes. The solution was protected from light and stored at 4°C. The gel-dye mix was equilibrated to room temperature for 30 minutes before use. An Agilent DNA chip was loaded with gel-dye, DNA ladder, and samples and analyzed on the Agilent 2100 Bioanalyzer. The library concentration was read from the peak (approximately ~250 bp).

For DNA Sequencing on Illumina Genome Analyzer II's (GAIIx), the library was diluted to 10 nM in 0.1% Tween20. The sequencing library was denatured using the Illumina protocol. The denatured libraries were diluted to ~6-8 pM following the Illumina protocol. Single-end GAIIx flow cells (v4) were clustered using an Illumina cBOT, according to the manufacturer's instructions. An Illumina GAIIx flow cell was run with 76-bp single-end reads (v4 and v5 sequencing kits) according to the manufacturer's instructions.

RNA SEQUENCING ALIGNMENT AND EXPRESSION QUANTIFICATION PIPELINE

Once the sequence of the reads was determined, the reads were mapped to a reference sequence in order to identify their genomic locations. Thus, the alignment tool needed a reference sequence (or sequences) in order to provide the coordinates of the reads. When reads were aligned, it was then possible to measure the expression of any element in the genome (gene, exon, novel transcribed region, etc.) by "counting" the number of reads that map to that element and properly normalizing by sequencing depth and size of the element. The protocols described below are based on the RSEQtools framework[1] (Habegger *et al*. 2011).

When the BrainSpan project was initiated, the consortium used Gencode v3c (Gencode version 3c (July 2009 freeze, GRCh37) - Ensembl 56) for processing the RNA sequencing data. Further information and statistics about the Gencode Reference Gene Set (version 3c) can be found at http://www.gencodegenes.org/releases/3c.html. For the BrainSpan project, Gencode version 3c protein

[1] RSEQtools main website: http://rseqtools.gersteinlab.org/; version 0.5.

coding genes were utilized. Gencode version 3c data has initially displayed in the Developmental Transcriptome Heatmap and is currently only available for download.

Upon completion of the BrainSpan project, the consortium utilized Gencode v10 (Gencode version 10 (July 2011 freeze, GRCh37) - Ensembl 65) for processing the RNA sequencing data. Further information and statistics about the Reference Gene Set (version 10) can be found at http://www.gencodegenes.org/releases/10.html. For the BrainSpan project, all Gencode version 10 genes were utilized. Gencode version 10 data is currently displayed in the Developmental Transcriptome Heatmap and is also available for download.

The next two sections describe the alignment and expression quantification processing that was done on the RNA sequencing data with Gencode v10 and Gencode v3c annotations. For mapping the resulting RPKM gene expression and exon expression data from Gencode v3c to Gencode v10, any orphan genes (genes present in Gencode v3c and absent in Gencode v10) are not available in the resulting Gencode v10 RNA sequencing expression quantification output. Where possible, the Gencode ID (ENSEMBL ID) was utilized to map the genes between Gencode v3c and Gencode v10.

RNA SEQUENCING ALIGNMENT AND EXPRESSION QUANTIFICATION PIPELINE WITH GENCODE V10

Alignment

Alignment of the reads was performed by the free software Tophat (version 1.3.1) (Trapnell *et al*. 2009). In essence, reads were mapped to a reference sequence that needed to be properly indexed. The program "bowtie-build" was used to build the sequence index (Langmead *et al*. 2009). The human genome sequence and spike-in sequence were indexed separately. For the human genome mapping, the gtf format annotation, Gencode (version 10, http://www.gencodegenes.org/releases/10.html), was additionally provided to improve the mapping quality of exon-exon junction reads. As an example, the alignment of sample A was performed as below:

Mapping to human genomic sequence:

```
$ tophat  --solexa1.3-quals -p  8  -G  gencode.v10.annotation.gtf  hg19  A.fq
```

Mapping to spike-in sequence:

```
$ tophat  --solexa1.3-quals -p  8  spike-in  A.fq
```

Here,"hg19" and "spike-in" were the indexed human genome sequences and spike-in sequences. More details about the parameters are available at http://tophat.cbcb.umd.edu/manual.html. The alignment was saved as BAM format and the junction mapping was saved as BED format.

Reference Sequence

The alignment tool needed a reference sequence (or sequences) in order to provide the coordinates of the reads. The human genome sequence (hg19, GRCh37) was used as the genomic reference. From the full reference, minor haplotypes, random and unknown sequences were excluded. Hence, the 22 chromosomes, and chromosomes X, Y and M were considered for the genomic alignment. The FASTA files were retrieved from the UCSC genome browser and are available at http://archive.gersteinlab.org/proj/brainseq/genome/. In addition, a set of standard spike-in RNA sequences was employed as reference for spike-in alignment. These spike-in RNAs were used both to track the brain regions and sequencing quality control. Each sample was tagged by adding a pair of spike-in RNAs unique to the region from which the sample was taken, and three common spike-ins were added for sequencing quality control. Spike-in sequences are available at: http://archive.gersteinlab.org/proj/brainseq/spike_in/spike_in.fa

Reads processing and quality filtering

The sequenced reads were processed and quality filtered prior to alignment to the references. Initially, the first base was trimmed from each read during the Genome Analyzer analysis pipeline, *i.e*., USE_BASES nY* (see the CASAVA 1.7 User Guide for more details), aiming at removing potential primer contamination because the errors at the start of the read extremely affect the mapping quality. Second, the sequenced reads were filtered by the base-calling quality score. In the Illumina GAIIx platform, the ASCII character "B" encoding a Phred-like quality score indicated a larger probability that a base is called incorrectly (see the CASAVA 1.7 User Guide for more details). Therefore, sequenced reads with quality scores of “B” (or “B”s) were discarded.

Measuring Expression

After the reads were mapped to the reference sequences, the expression level of genes, exons, and spike-in RNAs were measured in the commonly used units of RPKM (reads per kilobase of exon model per million mapped reads) (Mortazavi *et al*. 2008). SAMtools and RSEQtools software packages were used to perform this task (Li, *et al*. 2009; Habegger *et al*. 2011). The manual for SAMtools is available at http://samtools.sourceforge.net/samtools.shtml and the manual for RSEQtools is available at http://info.gersteinlab.org/RSEQtools. First, the BAM format alignment was converted into SAM format alignment by using the "view" function in SAMtools, and then the "sam2mrf" function in RSEQtools was used to convert the SAM format to Mapped Read Format (MRF), a compact format to represent the mapped reads (for details see RSEQtools manual). Notably, the downstream analysis was restricted to using uniquely mapped reads because of the uncertainty of the reads with multiple mappable sites. In addition, mitochondrial reads were excluded due to their deep coverage and large variation across different individuals. After filtering, RPKM values were computed using "mrfQuantifier" function in RSEQtools. This program required an annotation set, which includes elements whose expression level was to be measured. For the spike-in RNAs, it was easy to quantify the expression level. As for the human genes, the presence of alternative transcripts constituted an issue for the quantification of expression levels because the assignment of reads to specific transcripts is not straightforward. Hence, a composite model of a gene was defined that summarizes its exonic regions. The composite model of a gene was the union of all exonic nucleotides across all of its transcripts (**Figure 1**).

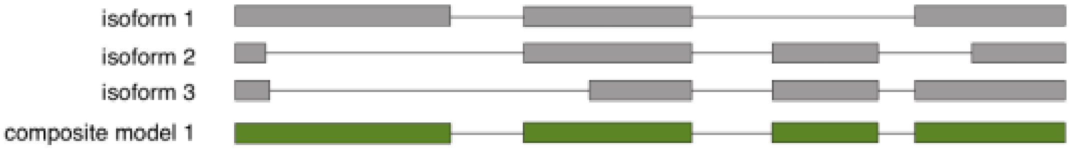

Figure 1. Example of a composite model for a gene and its three isoforms.

In RSEQtools, "mergeTranscript" function was used to generate the composite model from Gencode (version 10, http://www.gencodegenes.org/releases/10.html). From the composite model of genes, the composite model exons were also extracted. Given a set of mapped reads in MRF and an annotation set (representing exons or gene models) in interval format (see details in RSEQtools manual), the "mrfQuantifier" program calculated an expression value for each annotation entry by counting all nucleotides from reads that overlap with a given annotation entry. Subsequently, this value was normalized per million mapped nucleotides and the length of the annotation item per kb. In summary, here were the steps to compute the various expression levels for sample A:

Convert BAM format alignment to SAM format alignment:

```
$ samtools  view  A.bam  >  A.sam
```

Notably, a custom Perl script was additionally performed to remove reads having multiple mappable hits, as to restrict the entire analyses to using uniquely mapped reads.

Convert SAM format to MRF:

```
$sam2mrf  <  A.sam  >  A.mrf
```

Build composite model of gene:

```
$mergeTranscripts  knownIsoforms.txt  transcript.interval  compositeModel  >  geneComposite.interval
```

Calculate expression of any assigned element:

```
$mrfQuantifier  annotation.interval  multipleOverlap  <  A.mrf  >  A.expression
```

where "knownIsoforms.txt" determined which transcript isoforms belong together; "transcript.interval" was the interval format annotation of transcript isoforms; "geneComposite.interval" was the interval format annotation of gene composite model, from which the interval format annotation of exon composite model was extracted; "annotation.interval" was the interval format annotations for gene composite model, exon composite model and spike-in RNAs; "multipleOverlap" indicated reads that overlap with multiple annotated features were assigned to all of them. More details are also available in the RSEQtools manual.

Conversion to Entrez IDs

The Gencode annotation set was mapped to Entrez Gene IDs in order to be compatible with previous data sets and exon arrays. Biomart was employed to extract the EntrezIDs that correspond to Gencode elements (Haider *et al*. 2009). The current mapping was created in August 2010. For those Gencode IDs having more than one Entrez ID, they were combined into a comma separated list. For those Gencode IDs having no Entrez ID, that column was left blank. Finally, the format of the expression files consisted of the following:

Gene expression file was a composite of 3 columns:

```
Column1:Entrez ID
Column2:Gencode ID
Column3:RPKM (see Measuring Expression)

Example:

EntrezID    GencodeID                       RPKM
7105        ENSG00000000003|TSPAN6          2.550667
64102       ENSG00000000005|TNMD            0.061325
8813        ENSG00000000419|DPM1            11.499898
57147       ENSG00000000457|SCYL3           2.385437
55732       ENSG00000000460|C1orf112        0.602042
2268        ENSG00000000938|FGR             1.259009
3075        ENSG00000000971|CFH             3.011607
…           …                               …
```

Exon expression file was a composite of 6 columns:

```
Column1:Entrez ID
Column2:Gencode ID
Column3:Chromosome Number
Column4:Start Position
Column5:End Position
Column6:RPKM

Example:

EntrezID    GencodeID                   Chromosome    Start       End         RPKM
7105        ENSG00000000003|TSPAN6      chrX          99883666    99884983     3.528534
7105        ENSG00000000003|TSPAN6      chrX          99885755    99885863     5.925681
7105        ENSG00000000003|TSPAN6      chrX          99887481    99887565     2.53031
7105        ENSG00000000003|TSPAN6      chrX          99888401    99888536     2.093812
7105        ENSG00000000003|TSPAN6      chrX          99888927    99889026     2.673379
7105        ENSG00000000003|TSPAN6      chrX          99890174    99890249     5.324895
7105        ENSG00000000003|TSPAN6      chrX          99890554    99890743     4.355129
7105        ENSG00000000003|TSPAN6      chrX          99891187    99892101     0.325124
7105        ENSG00000000003|TSPAN6      chrX          99894941    99894988     0
…           …                           …              …          …           …
```

In data production, the various invocations of Tophat alignment, SAMtools format conversion, measuring expression by RSEQtools, and the mapping to Entrez IDs were driven by in-house Perl scripts and were run on the Yale High Performance Computing clusters. This processing also compared the region spike-in pairing detected in each sample with the brain region nominally associated with the sample and generated a report that identified any inconsistencies.

RNA SEQUENCING ALIGNMENT AND EXPRESSION QUANTIFICATION PIPELINE WITH GENCODE V3C

Alignment

Alignment of raw reads was performed by ELAND2, the alignment tool provided by Illumina within the Genome Analyzer suite. Reads were mapped to a reference sequence that needs to be properly indexed. Three sets of references were used: human genome sequence, splice junction library, and spike-in sequences. The splice junction library was fundamental to capture reads that span two exons. Indeed, alignment tools typically cannot align portion of reads to genomic location that are far apart. Spike-in sequences were used to label samples and provide controls for expression level normalization.

Human Genome Reference Sequence

The human genome sequence (hg19 – GRCh37) was used as the genomic reference. From the full reference, minor haplotypes and random sequences were excluded. Hence, the 22 chromosomes, and chromosomes X, Y and M were considered for the genomic alignment. The FASTA files are located at http://archive.gersteinlab.org/proj/brainseq/genome/.

Splice Junction Library

The splice junction library is fundamental to RNA-Seq data. Typically, alignment tools require reads to be mapped to contiguous regions of the reference sequences. This constraint enables a fast and efficient mapping of millions of reads. However, the main drawback of this approach is that reads generated from the junction of two exons of a transcript cannot be mapped. In order to address this issue, a splice junction library was used, where the sequences of two exons that are part of the same transcript are included in the library (see **Figure 2**). In order to create this library, a gene annotation set must be used. The splice junction library was based on the Gencode annotation set release 3c (http://archive.gersteinlab.org/proj/brainseq/Gencode[2]). To ensure that reads mapped to elements of this library truly span the exon junctions, a minimum overlap of 10 nucleotides was required. Hence, since the reads are 75 nucleotides long, every element in the library includes 65 nucleotides from each exon for a total of 130 nucleotides.

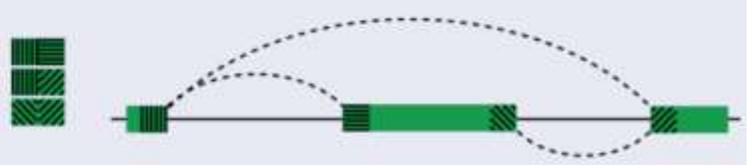

Figure *2*. **Schematic of splice junction library.** An example of splice junctions included from a transcript containing three exons (junctions for exon 1 and 2; 1 and 3; 2 and 3).

The splice junction library was generated with 'createSpliceJunctionLibrary', a utility program from RSEQtools (http://info.gersteinlab.org/RSEQtools#createSpliceJunctionLibrary):

```
$ createSpliceJunctionLibrary hg19.2bit gencode.3c.annotation.interval 65
```

where hg19.2bit was the human genome reference formatted for use with the “faToTwoBit” utility from the BLAT suite[3] (Kent 2002); gencode.3c.annotation.interval was the annotation set in interval format (see http://info.gersteinlab.org/RSEQtools#Interval for format details) and 65 was the number of nucleotides included from each exon. Note that all possible exon combinations were considered, which allows reads from novel junctions of known exons to be mapped. The genome reference hg19.2bit is located at

[2] The original Gencode 3c annotation sets can be found at: ftp://ftp.sanger.ac.uk/pub/gencode/. See the log file for more details about the post-processing of the Gencode 3c annotation set.

[3] http://genome-test.cse.ucsc.edu/~kent/exe/

http://archive.gersteinlab.org/proj/brainseq/splice_junction/hg19.2bit[4]; whereas the gene annotation interval format can be found at http://archive.gersteinlab.org/proj/brainseq/Gencode/[5]. The resulting splice junction library (in FASTA format) is available at: http://archive.gersteinlab.org/proj/brainseq/splice_junction/male/.

Spike-ins

A set of standard spike-in RNA molecules was employed. These were used both to track the brain regions and to normalize expression levels across experiments. Each sample was tagged by adding a pair of spike-ins unique to the region from which the sample was taken. Also, an additional three common spike-ins were added for expression normalization and quality control. Spike-in sequences are available at: http://archive.gersteinlab.org/proj/brainseq/spike_in/spike_in.fa.

Preparing for the alignment

Reference sequences were converted to make the alignment fast and efficient. The genome sequence, the splice junction library and the spike-in sequences were located in one directory such that the reads could be simultaneously mapped to these references. The program squashGenome, provided by Illumina, performed the conversion, *e.g.*:

```
$ squashGenome /path/to/squashed/files/ *.fa
```

See Illumina's CASAVA User Guide for more information. Once completed, this process was only necessary if one of the input sequences changes. All FASTA files that need to be "squashed" are located at http://archive.gersteinlab.org/proj/brainseq/to_be_squashed/

Aligning Reads

A given sample was run through Illumina's standard processing to produce a collection of read sequences. The GERALD stage of Illumina's CASAVA pipeline was used to align these reads by running ELAND and associated processing. To do so, we provided a configuration file to describe the valid bases of each read, the type of ELAND processing to run and the reference to be used (prepared as above):

```
USE_BASES nY*
ANALYSIS eland_extended
ELAND_GENOME /path/to/squashed/files
```

Note that the first base was trimmed from each read. See the CASAVA User Guide for more information.

Measuring Expression

After the reads were mapped to the reference sequences, the expression level of genes and exons was measured in the commonly used units of RPKM (reads per kilobase of exon model per million mapped reads) (Mortazavi *et al*. 2008). RSEQtools was used to perform this task. First, the output of the alignment was converted into Mapped Read Format (MRF – http://info.gersteinlab.org/RSEQtools#Mapped_Read_Format), a compact format to represent the mapped reads. Then, the RPKM values were computed using mrfQuantifier (http://info.gersteinlab.org/RSEQtools#mrfQuantifier). This program required an annotation set, which includes the elements whose expression level is to be measured. The presence of alternative transcripts constituted an issue for the quantitation of expression levels because the assignment of reads to specific transcripts is not straightforward. Hence, a composite model of a gene was defined that summarizes its exonic regions. Indeed, the composite model of a gene was the union of all exonic nucleotides across all of its transcripts (**Figure 1**).
'mergeTranscript' generates the composite model from the Genecode3c annotation set (http://info.gersteinlab.org/RSEQtools#mergeTranscripts). From the composite model of genes, composite model exons were also extracted. These models (in interval format) are located at (http://archive.gersteinlab.org/proj/brainseq/interval_files/)

Given a set of mapped reads in MRF and an annotation set (representing exons or gene models) mrfQuantifier calculated an expression value for each annotation entry by counting all nucleotides from reads

[4] The original hg19.2bit file is available at: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.2bit

[5] See the README and log files therein for more information.

that overlap with a given annotation entry. Subsequently, this value was normalized per million mapped nucleotides and the length of the annotation item per kb.

```
$ mrfQuantifier <file.annotation> <singleOverlap|multipleOverlap>
```

where file.annotation was the composite model (gene or exon); singleOverlap:reads that overlap with multiple annotated features were ignored; multipleOverlap:reads that overlap with multiple annotated features were assigned to all of them. To be comprehensive, mrfQuantifier was run twice using each mode, *i.e.*, singleOverlap and multipleOverlap. Similarly, the expression of the spike-ins and the splice junctions were also measured.

In summary, here were the steps to compute the various expression levels (for lane A):

```
$ singleExport2mrf < s_A_export.txt > s_A.mrf 2> s_A.mrf.log

$ mrfQuantifier gencode.v3c.annotation.GRCh37.exon.gtpc.ttpc.composite.interval singleOverlap < s_A.mrf
>  s_A_export_hg19_gene_so_expression.txt 2> s_A_export_hg19_gene_so_expression.log

$ mrfQuantifier gencode.v3c.annotation.GRCh37.exon.gtpc.ttpc.composite.interval multipleOverlap <
s_A.mrf >  s_A_export_hg19_gene_mo_expression.txt 2> s_A_export_hg19_gene_mo_expression.log

$ mrfQuantifier gencode.v3c.annotation.GRCh37.exon.gtpc.ttpc.composite.interval.singleExon
singleOverlap < s_A.mrf >  s_A_export_hg19_exon_so_expression.txt 2>
s_A_export_hg19_exon_so_expression.log

$ mrfQuantifier gencode.v3c.annotation.GRCh37.exon.gtpc.ttpc.composite.interval.singleExon
multipleOverlap < s_A.mrf >  s_A_export_hg19_exon_mo_expression.txt 2>
s_A_export_hg19_exon_mo_expression.log

$ mrfQuantifier spike-in.interval singleOverlap < s_A.mrf >  s_A_export_si_expression.txt 2>
s_A_export_si_expression.log

$ mrfQuantifier gencode.v3c.annotation.GRCh37.exon.gtpc.ttpc_2x65_spliceJunctions.interval
multipleOverlap < s_A.mrf > s_A_export_hg19_exsp_mo_expression.txt 2>
s_A_export_hg19_exsp_mo_expression.log
```

Conversion to Entrez IDs

The Gencode annotation set was mapped to Entrez Gene IDs in order to be compatible with previous data sets and exon arrays. Biomart was employed to extract the EntrezIDs that correspond to Gencode elements (Haider *et al*. 2009). The current mapping was created in August 2010. For those Gencode IDs having more than one Entrez ID, they were combined into a comma separated list. For those Gencode IDs having no Entrez ID, that column was left blank. Finally, the format of the expression files consisted of the same format as the gene expression file (composite of 3 columns) and exon expression file (composite of 6 columns) as described above in the Gencode v10 section.

In production, the various invocations of mrfQuantifier, the mapping to Entrez IDs, and the conversion of expression data from an internal format to the format presented above was driven by a “make” script (“make” is a Unix utility that automates runs of complex chains of processing steps where there are dependencies from one step to the next). This processing also compared the region spike-in pairing detected in each sample with the brain region nominally associated with the sample and generated a report that identified any inconsistencies. If there were none, the final step assigns semantically rich names (containing species, brain id, and region) to the outputs whose original names are based on Illumina's generic naming scheme (“s_1_...”, “s_2_...”, etc.).

The make script was invoked when the GERALD processing (see above) completes.

EXON MICROARRAY HYBRIDIZATION

Exon microarray hybridizations were performed at the Yale Center for Genome Analysis and Gene Logic, Inc. (Gaithersburg, MD). The Ambion WT Expression kit in combination with the Affymetrix WT Terminal Labeling and control kit was used for target preparation as recommended by Affymetrix. PolyA controls were added to the input RNA to measure efficiency of target amplification. Fragmented, labeled second cycle cDNA (5.5 µg) was added to a hybridization cocktail prior to loading of 200 µl onto individual Affymetrix Human Exon 1.0 ST arrays. Microarrays were hybridized at 45°C for 16- 24 hours, washed, and stained on an Affymetrix FS450 fluidics station according to manufacturer recommendations. Microarrays were scanned on a GeneChip® Scanner 3000 and visually inspected for hybridization artifacts. Exon chip analysis was performed using Affymetrix Power Tools 1.12.0. Probe level data was summarized into probeset level data using the RMA algorithm in combination with a R-script. The raw image files (.DAT files) were analyzed using Affymetrix GeneChip Operating Software (GCOS) to generate .CEL files.

SMALL RNA SEQUENCING AND ANALYSIS

The TruSeq Small RNA Sample Kit (Illumina) was used to prepare cDNA libraries per manufacturer instructions. Briefly, 1 µg of total RNA was ligated with 3'- and then 5'- adapters, followed by reverse transcription and PCR amplification. The PCR utilizes 48 different types of primers that will add 48 different index sequences to the adapters. Samples with distinct indexes were pooled, which allowed subsequent retrieval of each sample from multiplexed sequencing runs. Each pooled library was size selected by gel excision or the LabChip XT DNA chip (Caliper Life Sciences) for cDNA fragments between 145-160 bp, including the ligated 5' and 3' adapters. The final product was assessed for its size distribution and concentration using Bioanalyzer DNA 1000 Kit.

Raw sequence reads of length 51 nt were obtained from an Illumina HiSeq2000 using the Illumina TruSeq Small RNA-seq protocol. All samples passed quality control using FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), considering a wide range of quality metrics including per-sequence and per-base quality, per-sequence and per-base GC content, per base N content, sequence duplication levels, and overrepresented sequences. Quality metrics were similar for all remaining samples. Sequence reads were clipped to remove the Illumina small RNA 3' adapter (TGGAATTCTCGGGTGCCAAGG). A minimum of 5 adapter bases were required for clipping, resulting in approximately 90% of the reads being shortened. We experimented with using more stringent minimum barcode-nucleotides required for clipping; however values between 4-10 bases yielded very similar numbers and length distributions of clipped reads (data not shown). Clipped reads were collapsed into the multi-FASTA format that contains only unique sequences and associated counts for each read for input into the pipeline constructed around the miRanalyzer miRNA analysis software described in (Hackenberg et al., 2009, Hackenberg et al., 2011).

For the analysis, trimmed reads between 16-36 nt were selected and searched against mature and precursor/hairpin databases in mirBase (release 18) (Kozomara and Griffiths-Jones, 2011). Reads not mapped to known targets in the mirBase reference were searched, in the following order, against piwi-interacting RNA (piRNA) sequences in RNAdb (version 2) (Pang et al., 2005), enhancer RNA (eRNA) sequences from genomic coordinates defined in (Prabhakar et al., 2006) (converted from hg17 to hg19 using UCSC's *liftover* tool), transfer RNA sequences in GtRNAdb (Chan and Lowe, 2009), mRNAs defined in GENCODE (version 10) annotation (Harrow et al., 2006), and entries in the RFam database (version 10.1) (Gardner et al., 2009). Reads mapping to more than 5 distinct entities in any of these data sources were discarded due to poor specificity. All remaining reads were mapped to the human genome (hg19) to identify read clusters corresponding to potential novel miRNAs. All read mapping to miRBase, RNAdb, GtRNAdb, GENCODE, RFam, and the genomic reference was performed using Bowtie (version 0.12.7) (Langmead et al., 2009) using the '-best' and '-strata' arguments, in addition to a maximum 2-base mismatch in sequence alignment. As miRNAs are not a fixed size, a 17 nt seed region is used (corresponding to the minimum length of a known miRNA) and the mapping information was post-processed to identify the best alignment to longer miRNA sequences.

Known miRNA data for all samples were normalized by the total number of mapped reads over all RNA species to reduce any potential bias from very highly [differentially] expressed RNAs; this method of normalization should result in less inflation of expression estimates for lower-expressing RNAs in structures with fewer detected RNAs.

METHYLATION

DNA Purification

Frozen tissue (30 mg) was homogenized for 1 minute using a Bullet blender (Next Advance model BBX24B) on speed 6 and 1.4 mm stainless steel beads. DNA was purified using DNeasy Blood and Tissue Kit (QIAGEN) according to manufacturer's recommendation. Alternatively, DNA was purified from the aqueous phase after QIAzol treatment during RNA purification, using DNeasy Blood and Tissue Kit (QIAGEN). DNA quality and quantity were determined using a NanoDrop 2000.

Infinium HumanMethylation450 Arrays

Genomic DNA samples (1 µg each) were bisulfite converted using Zymo EZ96 DNA methylation kits (Zymo Research, Orange, CA; cat # D5004) according to the manufacturer's instructions. Bisulfite-converted DNA was eluted in a volume of 18 µl and then 3 µl was used for post-bisulfite quality control tests as previously described (Campan et al., 2009). For the Infinium protocol, bisulfite-converted DNA was whole genome amplified (WGA) and then enzymatically fragmented. The bisulfite-converted, WGA-DNA samples were purified and hybridized to the Illumina Infinium HumanMethylation450 arrays (described in Bibikova et al., 2011), in which bisulfite-converted DNA molecules anneal to locus-specific DNA oligomers that are bound to individual bead types. Each CpG locus can hybridize to methylated (CpG) or unmethylated (TpG) oligo bead types. Allele-specific primer annealing was followed by single-base extension using labeled nucleotides. Both unmethylated and methylated bead types for a specific CpG locus incorporate the same labeled nucleotide, as determined by the base immediately preceding the cytosine being interrogated by the assay, and subsequently were detected in a single channel. Each beadchip, containing 12 subarrays, was fluorescently stained after extension, scanned, and the intensities of the methylated (M) and unmethylated (U) bead types for each CpG locus across all samples were measured. Mean non-background corrected M and U signal intensities for each locus were extracted using Illumina BeadStudio (or GenomeStudio) software. The beta value DNA methylation scores for each sample and locus were calculated as (M/(M+U)).

REFERENCES

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, *et al* (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288-95.

Campan M, Weisenberger DJ, Trinh B, Laird PW (2009) MethyLight. *Methods Molecular Biology* 507:325-337.

Chan PP, Lowe TM (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* 37:D93-97.

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, *et al* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Research* 37:D136-140.

Habegger L, Sboner A, Gianoulis TA, Rozowsky JS, Agarwal A, Snyder M, and Gerstein MB (2011) RSEQtools: A modular framework to analyze RNA-Seq data using compact anonymized data summaries. *Bioinformatics* 27:281-283. doi:10.1093/bioinformatics/btq643.

Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research* 37:W68-76.

Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research* 39:W132-138.

Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk (2009) BioMart Central Portal—unified access to biological data. *Nucleic Acids Research* 37:W23-W27. doi:10.1093/nar/gkp265.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, *et al* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biology* 7 Suppl 1:S4 1-9.

Kent JW (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Research* 12:656–664. doi:10.1101/gr.229202.

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research* 39:D152-157.

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 JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, and Church GM (2009) Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* 324(5931):1210-1213.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5:621-628. doi:10.1038/nmeth.1226.

Pang KC, Stephen S, Engstrom PG, Tajul-Arifin K, Chen W, *et al* (2005) RNAdb--a comprehensive mammalian noncoding RNA database. *Nucleic Acids Research* 33:D125-130.

Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. *Science* 314:786.

The External RNA Controls Consortium (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, Ikonom P, Irizarry RA, Kawasaki ES, Kaysser-Kranich T, Kerr K, Gretchen 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, Williams M, Wilmer F, Wilson M, Wolber PK, Wu X, and Zadro R) (2005) The External RNA Controls Consortium: a progress report. *Nature Methods* 2:731-4. doi:10.1038/nmeth1005-731.

Trapnell C, Pachter L, and Salzberg S (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.