Published in final edited form as:

Nat Neurosci. 2014 November ; 17(11): 1476–1490. doi:10.1038/nn.3816.

Nature Neuroscience Review:

Analytical Tools and Current Challenges in the Modern Era of Neuroepigenomics

Ian Maze¹, Li Shen², Bin Zhang³, Benjamin A. Garcia⁴, Ningyi Shao², Amanda Mitchell^{2,5}, HaoSheng Sun², Schahram Akbarian^{2,5}, C. David Allis¹, and Eric J. Nestler²

¹Laboratory of Chromatin Biology and Epigenetics, The Rockefeller University, New York, NY 10065, USA

²Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

³Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

⁴Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

⁵Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

Abstract

Over the past decade, rapid advances in epigenomics research have extensively characterized critical roles for chromatin regulatory events during normal periods of eukaryotic cell development and plasticity, as well as part of aberrant processes implicated in human disease. Application of such approaches to studies of the central nervous system (CNS), however, is more recent. Here, we provide a comprehensive overview of currently available tools to analyze neuroepigenomics data, as well as a discussion of pending challenges specific to the field of neuroscience. Integration of numerous unbiased genome-wide and proteomic approaches will be necessary to fully understand the neuroepigenome and the extraordinarily complex nature of the human brain. This will be critical to the development of future diagnostic and therapeutic strategies aimed at alleviating the vast array of heterogeneous and genetically distinct disorders of the CNS.

Introduction

Our current understanding of how the brain adapts and responds over time to a host of environmental challenges, both under normal conditions and in a range of neurological and psychiatric disease states, is incomplete. Although candidate gene approaches have been useful, too little is still known to select the best candidate genes for future investigations.

*Correspondence: Eric J. Nestler (eric.nestler@mssm.edu).

Competing Financial Interests

The authors declare no competing financial interests

Unbiased approaches are therefore essential to reveal fundamentally new insights into these questions.

Genome-wide studies of expressed RNAs are powerful but not sufficient. This is because many adaptations and maladaptations do not involve alterations in steady state levels of RNAs. Instead, they involve “molecular scars”—chromatin structural alterations at specific genes that alter their inducibility (e.g., priming or desensitization) in response to subsequent challenges^{1,2}. Studies of chromatin are thus required to identify genes affected by this latent form of regulation. Likewise, studies of chromatin endpoints are the primary means of exploring the detailed molecular mechanisms by which the steady state expression or inducibility of genes is affected. Prior to chromatin studies, all efforts to understand mechanisms focused on cell culture, despite the fact that what happens in cultured cells—even cultured neurons—is not always an accurate reflection of what happens within the fully differentiated adult brain. Analogous to the developmental biology and cancer biology fields, where certain epigenomic modifications are seemingly permanent, studies of chromatin in brain have the potential to identify how environmental experiences/challenges lead to life-long changes in neuronal or glial function and in behavior, including disease susceptibility or resilience. Finally, an increasing number of CNS disorders are being shown to be caused by primary abnormalities in chromatin regulatory proteins. Increased knowledge of brain adaptations and disease pathogenesis resulting from explorations of epigenomic mechanisms³⁻¹⁸ has led to the possibility that such information can be mined to generate better diagnostic tests and treatments for a large variety of disabling nervous system disorders (see Table 1 for an update on progress in neuroepigenomics research).

A host of genome-wide methods have become available over the past decade leading to increasingly powerful tools for characterizing changes in RNA expression and chromatin modifications, as well as relating the two phenomena. The reader is referred to a companion review of these experimental approaches¹⁹. However, application of such methods to the brain, given its distinctive cellular heterogeneity and the need to focus on *in vivo* models, involves many unique challenges with regards to data analysis. In this review, we provide an overview of such challenges and highlight ways of overcoming them to derive the extraordinary advances promised by epigenomic studies of the nervous system.

RNA Expression Analysis

Genome-wide epigenomic studies typically begin with measures of RNA expression, since ultimately it is the regulation of such expression that serves as the functional readout of epigenomic modifications. Over the past decade, genome-wide RNA expression analysis in brain has served as a powerful tool for identifying transcriptional ‘signatures’ associated with normal neurodevelopment as well as pathological disease states. Historically, such investigations have relied on microarray technology as the primary means of generating transcriptome data in brain; however, since its development, RNA-seq²⁰⁻²² has proven to be a more powerful tool for assessing transcriptional outputs for a number of reasons. 1) Whereas microarray technology limits researchers to detecting and analyzing transcripts that correspond to existing genomic sequence information, RNA-seq allows for studies of both known and novel transcripts, an approach that is ideal for discovery-based experiments. 2)

Since RNA-seq allows for unambiguous mapping of obtained DNA sequences to unique regions of the genome, as opposed to cross-hybridization procedures inherent to microarray technologies, signal to noise ratios are significantly improved. 3) RNA-seq quantifies absolute rather than relative values, thereby allowing for assessments of a large dynamic range of expression levels²³⁻²⁵. Given these considerations, we focus exclusively here on RNA-seq, which provides the most complete and accurate assessment of all expressed RNAs in a given tissue¹⁹. Despite the potential power of this approach, the analysis of RNA-seq data is still far from routine and involves numerous bioinformatics challenges, which we review here.

RNA-seq: Initial Methods of Data Processing and Annotation

The raw data produced by RNA-seq (see Figure 1 for initial pipelines of RNA-seq data analysis) is—for each biological sample—tens to hundreds of million short sequences (called reads, typically 50-100 bp) that correspond to random fragments of expressed RNAs present in the original tissue. The first step in analyzing such data is to assess the quality of these reads, which greatly influences downstream bioinformatics outputs. For that purpose, fastqc (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) can be used. Fastqc is a lightweight, highly efficient and low profile (i.e., requires reduced memory and yields less excessive outputs) program that simply requires raw sequencing reads in fastq format for initial quality control (QC) assessments. Fastqc, however, does not address RNA-seq specific questions, such as exonic vs. intronic alignments, transcript detection rates, etc., making QC determinations for subsequent downstream analyses difficult. To do so, RNA-SeQC²⁶ is often used after fastqc, allowing investigators to examine numerous additional parameters associated with RNA-seq sample quality including yield, alignment and duplication rates, GC bias, contaminating rRNA content, regions of alignment (exon vs. intron vs. intragenic), continuity of coverage, 3'/5' biases and counts of detectable transcripts, among others. Although RNA-SeQC is a comprehensive program that addresses most QC issues, its usage can be cumbersome requiring a great deal of computational power. In addition, ngs.plot²⁷ can be used in connection with RNA-SeQC or fastqc to generate gene body plots for RNA-seq data, thereby providing an intuitive visualization tool for investigators to examine overall coverage patterns of sequencing reads from transcription start sites (TSSs) to transcription end sites (TESs). This tool allows for the identification of common sequencing abnormalities, such as strong 3' biases. Although underreported, such QC data are extremely important to interpretations of RNA expression in brain, as differences observed in transcript abundance between control and experimental conditions, although important, are often small in magnitude, owing to a variety of unique challenges encountered when working with neural tissues (Box 1). Since low quality sequencing results in decreased signal to noise, as well as potential biases, such small differences may be inadvertently masked or amplified, thereby leading to high false positive and negative rates.

Following initial QC assessments, short sequence reads are mapped to appropriate transcriptomes using splicing-aware alignment tools. Such mapping requires both a reference genome sequence and a description of transcripts (e.g., a format such as GFF) rendering alignment results highly specific to a particular version of the transcriptome, which one can refer to using a database name, as well as a release number (e.g., Ensembl

67). When working with a species in which a comprehensive reference genome is not available, *de novo* transcriptome assembly can be performed using tools such as Cufflinks²⁸ and Trinity²⁹. A popular choice for RNA-seq data alignment is Tophat³⁰, which wraps around Bowtie as the basic alignment tool but provides additional workflows to accomplish tasks specific to RNA-seq data analysis (e.g., splicing detection). Since spliced alignment is critical to RNA-seq analysis, it has attracted much research effort in recent years. For example, a new alignment program, referred to as STAR³¹, has been gaining significant market share and attention. In a recent study comparing 11 programs for RNA-seq alignment³², STAR achieved impressive performance across numerous benchmarks including alignment yield, basewise accuracy, mismatch and gap placement, exon junction discovery and suitability for transcript reconstruction. STAR is an ultra-fast, sensitive and precise tool that reduces alignment time from 1-2 days to a few hours for RNA-seq samples containing ~100 million reads, based on 4-core workstation processing capabilities. STAR is limited, however, in that its memory footprint can easily reach 32 GB or more, which is prohibitive for many common servers.

Analysis of Transcripts, Splice Variants and Non-coding RNAs

The next step in data analysis is to infer expression levels of individual transcripts from aligned reads. Simply put, if tissue samples derived from control conditions generate on average 1,000 reads for a given RNA, and samples from an experimental condition generate 2,000 reads, one can conclude a 2-fold induction of that gene's expression. Such analyses are rarely so simple, because of the expression of multiple transcripts from a single gene and because such transcripts share a majority of their sequences, often differing by only a few or dozens of base pairs. Cufflinks³³ solves this problem through the use of a statistical learning approach that assigns a fraction of the total read count to each transcript based on a maximum likelihood principle. In addition, Cufflinks attempts to quantify splicing events by grouping transcripts into TSS groups. A TSS group is a collection of transcripts that share TSSs. Such grouping is based on the rationale that alternative splicing events, such as exon skipping, will only happen between transcripts sharing the same TSS. Two types of splicing events are thereby characterized: 1) alternative splicing, which is defined between different transcripts that are within the same TSS group; and 2) alternative promoter usage, which is defined between transcripts with different TSSs. It should be noted that Cufflinks only reports these events at the level of TSS groups, as well as entire transcripts, and does not provide detailed information regarding which exons demonstrate alternative splicing. In order to obtain more detailed splicing information, Mixture of Isoforms (MISO) model³⁴ can be used. MISO is based on a different design from that of Cufflinks and works on a predefined set of splicing events, such as exon skipping, intron retention, mutual exclusion and alternative 3' UTR. MISO uses a Bayesian approach, which allows researchers to combine prior knowledge with obtained data for more reliable inferences, to derive relative abundance levels of spliced transcripts. Such detailed information provides an advantage when performing integrative analyses between alternative splicing and other forms of epigenomic regulation, such as the contribution of histone modification states and transcription factor binding events to regulation of alternative splicing. Additional information on alternative transcripts can be obtained from sequencing nuclear RNA as opposed to total cellular RNA. Nuclear RNA contains much larger amounts of sequenced

introns, which can provide invaluable information about splicing mechanisms³⁵. RNA-seq analysis of brain tissue has revealed an order of magnitude more alternative transcript production compared to that inferred from older microarray and related technologies¹⁶.

RNA-seq also enables the detection and quantification of several types of non-coding RNAs, which are proving to play crucial roles in biological regulation. A subset of long non-coding RNAs (lncRNAs)—similar to protein-coding genes—contain polyadenylated (polyA) tails, which allow for their detection by RNA-seq regardless of the RNA purification procedure used (e.g., ribozero or polyA selection). Identification of non-polyA lncRNAs, however, can only be accomplished through purification procedures preserving total RNA (i.e., polyA+ and polyA-). Since lncRNAs exist in high abundance in mammals, the Ensembl database has recently incorporated a large number of lncRNAs into its gene collection. Thus, predefined lncRNAs can now be analyzed alongside of protein-coding genes in the same sample. If investigators wish to predict a large number of novel, and still un-annotated, lncRNAs, however, this can be accomplished by utilizing histone modification state data to define candidate regions. To do so, ChIP (chromatin immunoprecipitation)-seq data for euchromatic H3K4me3 (trimethylated Lys4 of histone H3) and H3K36me3 need to first be obtained, as described in the next section. Using the intersection of H3K4me3 and H3K36me3 (so called ‘K4-K36 domains,’ as defined through “peak calling”—see below), RNA-seq reads at these domains can be extracted using programs such as bedtools³⁶ to estimate lncRNA abundance levels.

MicroRNA (miRNA) sequencing experiments, which must be run separately from standard RNA-seq experiments, investigate mature miRNAs that are typically 22 nucleotides in length. Since most high-throughput sequencing machines produce reads that are significantly longer than mature miRNAs, the short reads obtained from miR sequencing often contain a portion of adapter sequences. The first step of miRNA analysis is thus to remove adapter sequences using tools such as fastx (http://hannonlab.cshl.edu/fastx_toolkit/). As an alternative, the Cutadapt program³⁷ can be used for these purposes. Cutadapt is useful in that it supports a larger range of sequencing platforms, including color-space reads, which allow basepairs to be encoded in color to reduce sequencing error rates (<https://www.biostars.org/p/43855/>), in comparison to fastx. Following adapter removal, short reads can be mapped to the reference genome similar to that of data obtained from long RNA sequencing. miRBase³⁸ provides a comprehensive collection of miRNA sequences. Bedtools can then be used to extract read counts for each of the annotated miRs obtained from alignment. If novel prediction of miRNAs from a given sample is desired, the miRDeep³⁹ program can be used. Many additional families of small RNAs (Table 2) can be identified and analyzed with similar approaches^{40-46,38,47}.

Differential Analysis: Approaches, Advantages and Disadvantages

Differential analysis refers to the process of identifying differences in RNA expression levels of individual genes, or of individual splice variants of a single gene, between control and experimental samples. This is not straightforward, as there are numerous tools available, each of which is associated with high rates of false positive or false negative discovery and generates very different lists of regulated genes when applied to the same sequencing data.

Generation of ideal differential analytical tools is therefore a focus of great interest in the field.

The first step in this process often involves summarizing gene- or gene variant-level read counts using a popular python program called htseq (<http://www-huber.embl.de/users/anders/HTSeq/>). According to individual needs, bedtools can also be used for gene count summarization. All read counts across genes and samples are then imported into a data matrix so that each row represents a gene, and each column represents a sample. This data matrix serves as the input for downstream differential analyses. Numerous differential analysis tools have been developed in recent years. Two popular choices are DESeq⁴⁸ and edgeR⁴⁹, with both methods based on negative binomial testing, which provide an exact test (generalization of the Poisson distribution model) that is ideal for modeling biological variances of read count data. Variance estimates are often problematic for RNA-seq datasets, as sample sizes from animal models are typically small with an N=3 (three biological replicates) used per condition in most experiments. Therefore, numerous statistical methods are used to exploit the relationship between mean- and variance-related information obtained from neighboring genes to stabilize variance estimations. DESeq uses local regression to model mean-variance relationships, while edgeR uses Bayesian methods to “borrow” information from neighboring genes. Both methods generate satisfactory results, and their lists are often consistent. Recently, a new method called voom⁵⁰ has been developed, which does not depend on negative binomial testing and instead models mean-variance relationships on log-transformed read counts, thereby assigning a precise ‘weight’ for each gene. These weights are then entered into the limma⁵¹ empirical Bayes analysis pipeline. This approach provides access to a large collection of analysis tools developed originally for microarrays, which makes “voom” a particularly attractive option. Furthermore, DESeq has been criticized for its extremely conservative outputs, and its authors have acknowledged that their method yields high false negative rates⁵². Recently, the statistical power of the program has been improved with the introduction of DESeq2 (<http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>). Nevertheless, experience with RNA-seq suggests that a larger number of biological replicates are needed to derive the most reliable data^{53,54}.

Cuffdiff, another popular tool for RNA-seq differential analysis, and part of the cufflinks pipeline, is a natural choice for investigators using cufflinks for initial identification of transcripts and splice variants. However, we, along with numerous other groups, have found cuffdiff to generate a high degree of false positives. The authors of cufflinks claim to have improved the reliability of detection in cuffdiff⁵⁵, but a recent comparison pointed out that cuffdiff2 may be too conservative⁵⁶. Another reason for avoiding cuffdiff is its restrictive workflow. Cuffdiff only accepts short read alignments and performs read counts, GC content adjustments, 3’ bias corrections and differential analyses in one integrated workflow; based on a standard four-core workstation, it would take 1-2 days to complete analysis for a 20 sample (10 vs. 10) dataset and may crash for datasets with >80 samples. Using htseq for gene counts in connection with limma or DESeq greatly reduces the analysis time to <1 hour.

Performing all of the steps of RNA-seq analysis, including quality assessment, alignment, gene count summarization and differential analysis, can be tedious if investigators have hundreds/thousands of samples (e.g., from studies of clinical populations), as well as dozens of comparisons, to run and examine. Therefore, a new Python-based computational pipeline has been developed SPEctRA (scalable PipELine for RNA-seq Analysis) (<https://github.com/shenlab-sinai/complete-seq/>) to perform all these tasks with a single command accepting various parameter settings as a configuration file. This pipeline is designed to function within different computational environments. For example, when performing analyses under a portable batch system (PBS) cluster, computer software that allocates computational tasks, SPEctRA will generate batch scripts, automatically submit them and wait for them to finish. A new feature of this pipeline, which will allow it to run in Amazon cloud where both CPU power and memory can be configured on demand for each run, is currently being developed.

Epigenomic Analyses of the Nervous System

ChIP-seq: Initial Methods of Data Processing and Annotation

As with RNA-seq, the raw data obtained from ChIP-seq are tens to hundreds of million short reads per sample that correspond to genomic regions bound to the DNA-binding protein subjected to ChIP (e.g., a modified histone, transcription factor, etc.)¹⁹. Quality control and sequence alignment are the first steps in analyzing such ChIP-seq data. Any of several short read aligners, such as bowtie⁵⁷ or BWA⁵⁸, can be used to align reads to a reference genome and assess the enrichment of the DNA-binding protein of interest. Alignments are exported as BAM⁵⁹ files for further analysis. Sequencing quality can then be assessed using fastqc, as described above. If ChIP-seq samples are generated from sonication-based fragmentation methods, then duplicated alignments need to be removed, a procedure easily accomplished with samtools⁵⁹ or picard (<http://picard.sourceforge.net/>). In doing so, PCR duplicates are identified as reads aligned to the same genomic location or on the same strand of fragmented DNA. During sonication, genomic DNA is theoretically fragmented at random, thereby rendering it unlikely that two unique reads will align to the same location or on the same strand of DNA. Since micrococcal nuclease (MNase) digestion, as used in “native” ChIP, preferentially digests genomic sequences containing specific nucleotide sequences (i.e., the rate of MNase cleavage 5’ of A or T nucleotides occurs at a 30 times greater rate in comparison to cleavage at G or C nucleotides), the probability of obtaining fragments from the same location or strand is increased. Therefore, removal of these reads with native ChIP samples, unless excessively high (thresholds can be determined using DANPOS⁶⁰, as described below), is not typically performed. Removal of repeat reads can be problematic when analyzing repeat regions of the genome, which are now known to be important in biological regulation, thus framing a challenge for current bioinformatics innovations. The number of unique reads per sample is an important criterion for the quality of ChIP-seq data. Since antibodies play a key role in the success of any ChIP-seq experiment¹⁹, it is vitally important to determine the efficiency and specificity of the antibody being used. For this purpose, the phantomPeak⁶¹ tool can be employed, as suggested by the Encyclopedia of DNA Elements (ENCODE) consortium⁶², to examine the distribution of cross-correlations between represented strands of DNA to determine peak enrichments (see next section).

It is also helpful to visualize ChIP-seq enrichment profiles to ensure quality and diagnose any problems that may exist. Two approaches—localized inspection and global visualization—are used. For local inspection, the IGV genome browser⁶³, which runs on all platforms without uploading data, is popular. With the tools provided by the IGV, BAM file are first converted to TDF files and then loaded onto the browser to display coverage between two chromosomal coordinates. For global visualization, ngs.plot (<https://code.google.com/p/ngsplot/>)²⁷, which allows for inspection of both average and ‘laid out’ coverages as curves or heatmaps, respectively, at various functional genomic regions, such as TSS, TES, gene body, CpG islands, enhancers, exons and DNase I hypersensitive sites (DHSs). ngs.plot, which is simple to use and supports numerous genomes, can accommodate large alignment files with relatively small memory footprints. A ChIP-seq quality assessment pipeline has been developed (https://github.com/nyshao/chip-seq_preprocess) to perform all the above steps with a single command. This pipeline uses a similar design to the RNA-seq pipeline described previously. When dealing with large sample numbers and multiple comparisons, this pipeline saves significant amounts of time and effort.

Another important quality control consideration during the initial analysis/alignment of ChIP-seq data requires an in depth understanding of the differences between ‘uniquely mapped’ vs. ‘unique reads/tags,’ two terms that are often used interchangeably in the field but are distinctly defined. ‘Uniquely mapped’ reads are identified by all alignment software programs and exclude sequences that align to multiple genomic locations. These excluded sequences likely represent repetitive regions of the genome, or non-repetitive genomic loci that are extremely similar in sequence (although the latter becomes increasingly unlikely with greater read lengths). ‘Unique reads/tags,’ however, refer to reads remaining after PCR de-duplication (i.e., the non-redundant fraction or NRF) using tools such as samtools. PCR de-duplication essentially excludes reads (all copies but one) that align to the same genomic location. These non-unique reads are often PCR duplicates resulting from over-amplification during library processing, most likely due to low starting material or poor antibody efficiency. These reads are thus often removed to avoid PCR amplification bias. However, arguments now exist^{64,65} to suggest that these duplicative reads may provide increased dynamic range to ChIP signals. For example, if one were to imagine two 100 bp regions of the genome (A and B), where factor X binds strongly to region A and weakly to region B, and if all duplicated reads mapping to these regions were removed (assuming a sequencing read length of 100 bp), the resulting effect would be two reads for each of these regions (one mapping to each strand of DNA), an output that would prevent the investigator from distinguishing the differential binding strengths of factor X to these regions. Therefore, although removing duplicates is a highly conservative measure to prevent PCR amplification bias, it is also likely that this process similarly removes real signals that might be informative for determination of binding interactions. Having said this, without excluding these duplicates, one runs the risk of generating high levels of false positive findings due to increased signal. In our experience, keeping two reads for any identified duplicates (instead of only one) dramatically increases the sensitivity of downstream data analysis and genome browser viewing with IGV, while retaining low false positive discovery rates.

Peak calling—identifying a genomic region that displays a significant level of a DNA-binding protein above background—is an important task in analyzing ChIP-seq data. Currently, dozens of peak calling tools exist, and these methods can generally be separated into two categories: sonication- and MNase-digestion based. For sonication-based peaks, Model-based Analysis of ChIP-Seq (MACS)⁶⁶ and Hypergeometric Optimization of Motif EnRichment (HOMER)⁶⁷ are used. In our experience, both methods work very well for punctuated peaks (e.g., H3K4me3), however, if peaks are broad and diffusive (e.g., total histone H3 or H3K9me2), detection can be challenging, and HOMER is generally recommended, as systematic evaluations of broad peak detection are often difficult. For MNase-digested peaks, DANPOS⁶⁰, which detects not only basal enrichment but also various nucleosomal events (e.g., nucleosomal positioning and occupancy and “fuzziness” between the two), can be used. After peaks are detected, it is useful to determine their location within the genome, such as genes, gene deserts or pericentromeres. A regional analysis tool, which is part of the diffReps⁶⁸ package—more recently extended as a standalone program (https://github.com/shenlabsinai/region_analysis)—has been developed to perform this task. This program features single command usage and can assign peaks to one of eight distinct categories including proximal promoter (+/- 250 bp of a TSS), promoter 1k (1 kb upstream of a TSS), promoter 3k, gene body, gene desert, pericentromere, subtelomere and other intergenic loci. Regional annotation is a very useful feature, and other alternatives exist to perform this function including the ChIPpeakAnno⁶⁹ package. After peaks are annotated, enrichment of specific chromatin marks is easily visualized for their genomic distribution using piecharts or plots.

Differential Analysis: Approaches, Advantages and Disadvantages

Differential analysis of ChIP-seq data aims to identify genomic loci or broader regions that display significant changes in enrichment between control and experimental conditions. Although it is natural for investigators to consider differential analyses as an extension of peak calling, in reality the former cannot be measured accurately using standard peak calling methods. Currently, one differential analysis tool⁷⁰ based on peak calling exists, however, the numerous challenges associated with this approach make it disadvantageous. Differences in antibody efficiency, sequencing biases and manual handling of samples (i.e., sequencing library preparation), can produce very different signal-to-noise ratios between samples, which confounds peak calling. Insufficient sequencing depth (i.e., too few reads per sample) can also result in specific genomic regions having different coverage across different samples and complicate differential analyses. For example, consider the following situation: a 3 kb peak in biological replicate #1 under condition A; two 1 kb peaks with 1 kb gap in biological replicate #2 under condition A; a 2.5 kb peak shifted to the left with lower enrichment in biological replicate 1 under condition B; and a 2 kb peak shifted to the right with higher enrichment in biological replicate 2 under condition B. Such heterogeneity in peak calling, which is common in analyses of brain (see Box 1) makes comparisons difficult.

Further exaggerating this problem is the fact that different peak calling methods, or the same peak calling method with different parameter settings, can yield very different peaks using the same ChIP-seq dataset. Biological replicates are thus essential in neuroepigenomics

research to enhance statistical power and increase precision. However, most peak calling methods to date were not designed with biological replicates in mind. To address these challenges, one can use a sliding window-based strategy to ensure that the entire genome is scanned and scored continuously, and that significant regions can be extracted for further analyses. diffReps⁶⁸, which was developed to address this need, is a Perl-based program that features single command usage and has been applied to several projects. diffReps scans genomes with a predefined window size and performs one of four distinct statistical tests, identifies samples passing pre-defined cutoffs, merges replicates, performs multiple testing corrections and reports results. Using benchmark standards, diffReps is highly sensitive and efficiently controls for false positives (see Figure 2 for examples of ChIP-seq analysis outputs). It is also important to assess the reproducibility among biological replicates as an additional quality control measurement during data processing. To do so, programs such as corrgram⁷¹ can be used, which generates Pearson's correlation coefficients between ChIP-seq signals derived from multiple signals. Alternatively, irreproducible discovery rates (IDRs) can be derived following peak calling to determine the number of 'bound' regions observed between replicates, as described⁷².

Important questions for ChIP-seq experiments are the read depth required to appropriately make assumptions based on aforementioned analyses, and to determine if experimental sequencing depths are sufficient for the questions being asked. In general, researchers can follow the guidelines established by the ENCODE consortium⁷², which indicate a minimum of 10 or 20 million uniquely mapped reads for factors displaying punctuated or broad peaks, respectively. One can also perform saturation analyses for each individual factor to assess the sufficiency of sequencing depth. In essence, such analyses repeatedly perform peak calling on a series of samples from the original ChIP-seq library using increasing sampling rates, followed by plotting the number of peaks assigned vs. the sampling rate to identify plateaus indicative of sequencing depth saturation. One caveat for such analyses, however, is that narrow peak binding factors, such as transcription factors, will typically display increased numbers of binding sites/peaks with increasing numbers of reads obtained. This is due to the fact that such factors rarely saturate at the number of reads that are practical for saturation analysis.

Another challenge in analyzing ChIP-seq data involves appropriately annotating chromatin states while examining combinations of chromatin modification patterns in a given sample. To achieve this, an automated computational system, referred to as ChromHMM⁷³, has been developed to integrate ChIP-seq information from multiple histone modifications, chromatin factors, transcription factors, etc. to accurately assess combinatorial and spatial patterns of marks/binding factors in biological samples. Such analyses provide a novel computational tool for 'learning' chromatin states, characterizing biological functions and achieving genome-wide visualizations of such annotations. Although numerous methods now exist for analyzing ChIP-seq data, independent biological validation remains an essential step to confirm the accuracy of any data derived from genome-wide approaches.

Overlaying ChIP-seq and RNA-seq Data

In an effort to understand the transcriptional and epigenomic mechanisms underlying RNA expression, a major goal of current research is to merge ChIP-seq and RNA-seq datasets. In fact, if one analyzes a sufficiently large number of epigenomic endpoints (numerous histone modifications, chromatin remodeling factors, transcription factors and other chromatin regulatory proteins) it should in theory be possible to identify “chromatin signatures” that reflect specific modes of transcriptional regulation, such as gene activation or repression as well as gene priming or desensitization.

Accomplishing this task remains challenging given the vast amounts of data (terabytes) involved. One initial approach is to first reduce observed ChIP-seq events to individual genes. This thereby reduces the problem to essentially comparing two gene lists—one from differential analysis of ChIP-seq data and the other from differential analysis of RNA-seq data. GeneOverlap (<http://www.bioconductor.org/packages/devel/bioc/html/GeneOverlap.html>), a bioconductor package for testing and visualizing gene overlaps, can be used for this purpose. However, such analyses oversimplify the problem. First, a given histone modification can show opposite changes across the span of a given gene. Also, all histone modifications regulate gene expression in a highly cooperative but complex fashion. For example, the histone modification, H3K4me3, is highly enriched at active gene promoters and is often used as an indication of transcriptional initiation. However, H3K4me3 levels at some genes can increase, in concert with other DNA-binding proteins (e.g., RNA polymerase II [Pol 2] serine 5 phosphorylation), in response to a specific stimulus without increasing the expression of its cognate gene, a phenomenon thought to indicate gene “poising.” And in a much smaller number of examples, H3K4me3 has been linked to gene repression when it associates with ING2 (inhibitor of growth family-2), a chromatin regulatory protein⁷⁴. On the other hand, the histone mark, H3K9me3, which typically enriches within the gene bodies of silenced genes along with numerous other corepressor proteins, may exhibit increased enrichment with distinct binding partners to promote alternative splicing⁷⁵.

Previously, numerous groups^{76,77} have attempted to use histone modifications to predict gene expression at baseline states in cultured cells and have achieved good performance. These groups were able to generate predicted expression levels correlating with real expression levels with coefficients as high as 0.8. Their approach worked by separating gene bodies into dozens of bins, where the relative enrichment of histone marks was determined. In this case, enrichment values serve as observed data and corresponding gene expression levels serve as target variables. Such training data were then fed into machine-based learning methods, such as support vector machines or generalized linear models, to learn how the behavior of histone modifications relates to gene expression. This approach, however, has not yet been applied to brain tissue at rest and, in particular, to data examining the relationship between stimulus-induced changes in epigenomic endpoints and gene expression, perhaps the most essential question facing molecular neuroscientists interested in examining the role of epigenomic mechanisms in mediating the impact of environmental manipulations on the brain.

We recently developed a bioinformatics approach that was successfully applied to studying the actions of repeated cocaine exposure in the mouse nucleus accumbens¹⁶, a key brain reward region important in addiction. We first correlated ChIP-seq data for six histone modifications and for total Pol 2 binding with changes in gene expression as determined by RNA-seq. While we were able to demonstrate >50 distinct chromatin signatures that correlate with altered RNA expression, the correlations were relatively weak and not deterministic. This suggests that a far larger number of chromatin endpoints are needed to make such predictions; indeed, current research is revealing large numbers of previously unappreciated histone modifications (see below). Additional considerations unique to the analysis of brain tissue are also likely involved (see Box 1).

We used a similar approach to determine whether repeated cocaine-induced changes in alternative splicing in nucleus accumbens correlate with changes in chromatin modifications⁵⁰. We focused on genomic regions most closely associated with splicing, such as variant exons, alternative donors, alternative acceptors and their neighboring intronic regions, and examined numerous chromatin modifications at these regions. We applied this information extraction procedure to all known transcripts of the mouse genome and calculated the log2 fold changes between repeated cocaine and saline and loaded the values into a data matrix so that each row represents a transcript and each column represents a histone mark-gene region combination (e.g., a H3K4me3 peak at the proximal promoter of that gene). We also used cufflinks to determine the expression changes of these transcripts. K-means clustering was applied on the chromatin modification matrix to identify clusters of transcripts that show similar patterns. These clusters were then correlated with transcripts' expression changes to determine whether there is any significant overlap. We were able to identify 29 such clusters or chromatin modification signatures. Further analysis of these signatures using motif finding revealed several important splicing factors and transcription factors deduced to be important in their regulation. One of the splicing factors, A2BP1 (Rbfox1/Fox-1), was validated by showing that knocking out A2BP1 in adult nucleus accumbens blocks cocaine's regulation of several of the genes identified in this genome-wide analysis. Moreover, A2BP1 knockout was shown to attenuate behavioral responses to cocaine¹⁶. These findings illustrate how the overlay of ChIP-seq and RNA-seq data can reveal fundamentally new insight into the biological basis of a complex brain disorder.

Overlaying DNA Methylation with Gene Expression Analyses

Methylation of cytosine bases in DNA (5-methylcytosine or 5mC) has historically been viewed as an important mode of gene repression. However, many alternative forms of DNA methylation have been demonstrated in recent years, most importantly, 5-hydroxycytosine (5hmC), which is enriched in brain and associated with gene activation⁷⁸. Despite the implication of DNA methylation in numerous neuropsychiatric phenomena, virtually all studies to date have focused on individual candidate genes, with genome-wide explorations of brain sorely lacking.

Several methods, described in the companion review¹⁹, are used to obtain a genome-wide map of 5mc and 5hmc, but doing so at single nucleotide resolution remains a challenge due to the sequencing costs involved. These methods include genome-wide bisulfite sequencing

(BS-seq)—including oxidized BS-seq (oxBS-seq) to distinguish 5mc and 5hmc, RRBS (reduced representation bisulfite sequencing) which focuses on CG rich regions, and meDIP-seq which involves immunoprecipitation of genomic DNA fragments with an antibody directed against 5mC or 5hmC followed by deep sequencing. The human “methylome” can be obtained from a chip-based method, however, available chips do not distinguish between 5mC and 5hmC.

The first step in analyzing BS-seq data is to trim the adaptors and low quality 3' ends. Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which wraps around Cutadapt and fastqc to provide extra functionality for RRBS libraries, can be used for this purpose; it reports the sequencing QC by calling fastqc after trimming. Then the reads are aligned to the reference genome. Generally, the aligners need a preprocessing step to tolerate the conversion of C to T. Two strategies of aligners are available. One is a wild-card aligner, such as BSMAP (<https://code.google.com/p/bsmap/>)⁷⁹, as all Cs in a reference genome are replaced by a wild-card letter (e.g., “Y”), with both C and T possibly aligned to Y. Another approach is a three-letter aligner, such as Bismark (<http://www.bioinformatics.babraham.ac.uk/projects/bismark/>)⁸⁰; here, all Cs in reads and a reference genome are converted to Ts, whereas Gs are converted to As equivalent to C-to-A conversions on the reverse strand, and the aligners process alignments in only three-letter alphabets (A, G and T). Wild-card aligners may reach better genomic coverage while potentially introducing bias to increase DNA methylation levels, compared with three-letter aligners⁸¹. After the alignment, the quantification of absolute DNA methylation can be achieved by counting the alignments, referred to as “methylation calling”.

The next step is to detect sites of differential methylated regions (DMRs) in experimental vs. control conditions. Basic analyses like t-tests or Wilcoxon rank tests can be applied; details about more advanced methods have been reviewed recently⁸¹. oxbs-sequencing-qc (<https://code.google.com/p/oxbs-sequencing-qc>) is a pipeline based on Bismark, including trimming of adaptors and low quality, alignment and methylation calling. Users need to implement the detection of DMRs by other analytical tools. MOABS (<http://code.google.com/p/moabs/>)⁸² is another pipeline, developed by the authors of BSMAP. It calls and accepts the alignment results from BSMAP or Bismark, and implements the detection of DMRs based on a Beta-Binomial hierarchical model.

As the final step of data analysis, DMR events detected by BS-seq or oxBS-seq are annotated by region_analysis or ChIPpeakAnno, and then integrated with differential expressed genes detected by RNA-seq, or with ChIP-seq characterizations of other epigenomic endpoints, as described in the ChIP-seq section above. Approaches similar to ChIP-seq are used to analyze sequencing data obtained from meDIP-seq experiments. Reads are first evaluated for QC and then aligned to a reference genome using the various tools outlined above, with the number of aligned reads used for methylation calling⁸³. In addition to canonical CpG methylation, highly conserved non-CpG methylation (mCH, where H is either A, T or C) may also play an important role in the central nervous system. It was recently reported that non-CpG methylation accumulates in neurons, but not glia, of the cerebral cortex during development¹³.

Reconstruction of Multi-Scale Biological Networks

A high priority for current research is to optimize ways of uniting genome-wide measures of RNA expression and chromatin modifications with human DNA sequence and other clinical data. Several genome-wide methods aimed at assessing individual differences in DNA sequence—including genome-wide association studies (GWAS) or whole exome or genome sequencing—are uncovering large numbers of genetic loci associated with neuropsychiatric disease states such as Alzheimer's disease, Parkinson's Disease, autism and schizophrenia/bipolar disorder among many others. While these studies are identifying common and rare sequence variations, available data only explain a small portion of the genetic contribution of most of these illnesses⁸⁴. Also, gene sequence analysis alone is insufficient to uncover causal mechanisms regarding gene/pathway dysregulation that gives rise to the disease. Increasingly available large scale genomic, epigenomic and clinical data informing neuropsychiatric disorders, derived from humans and animal models, is now making it possible for the first time to more comprehensively uncover key mechanisms and regulators of these diseases. For example, the nearly completed Cancer Genome Atlas project (TCGA) aims to provide a comprehensive genomic, epigenomic and pathophysiological characterization of over 30 different types of cancers including glioblastoma⁸⁵. The ongoing Genotype Tissue Expression project (GTEx) is generating a compendium of genotypic and gene expression data in many human tissues including cerebral cortex and cerebellum⁸⁶. For such large-scale molecular datasets, several systems/network approaches have been developed to identify and dissect the underlying “interactomes” for the discovery of key mechanisms and causal regulators in normal or pathological biological systems.

Different methods of analyzing correlated gene regulation, such as weighted gene co-expression network analysis (WGCNA)⁸⁷, aims to capture total interactions among genes in a more comprehensive manner than traditional un-weighted approaches and has been used to identify coexpressed gene modules and key genes associated with glioblastoma⁸⁸, late-onset Alzheimer's disease (LOAD)⁸⁹⁻⁹¹ and autism^{92,93}. Given the correlative nature of WGCNA, it can neither predict causal relationships nor identify causal regulators. For instance, to identify master regulators of LOAD, Rhinn et al. developed a differential co-expression analysis (DCA) approach to integrate differential expression and differential correlation to identify *APOE* $\epsilon 4$ effectors⁹⁴. While DCA can pinpoint master regulators, DCA by itself does not provide a context of causal networks to further understand the underlying subnetworks and pathways controlled by master regulators.

Recently, a more comprehensive effort was made to build multiscale gene regulatory networks from large scale genetic, genomic and pathophysiological data for identification of key mechanisms and causal regulators in LOAD⁹⁵. A key component of this multiscale network analysis (MNA) approach is to integrate gene coexpression and causal network analyses so as to make full use of the multiscale biological data. Gene modules that are comprised of highly interacting genes through WGCNA are rank ordered based on their correlations with clinical outcomes and enrichment for differential expression in LOAD when compared with normal controls. The causal relationships among the genes in each module are then determined by causal network inference that integrates gene expression data and prior information derived from expression-associated gene sequence variations such as

single nucleotide polymorphisms (SNPs) and quantitative trait loci (QTL). Bayesian networks are then used to identify key causal regulators based on network connectivity.

MNA is in line with the recent finding that no single network inference method is universally best across multiple datasets and the best strategy is to integrate multiple networks constructed from different approaches⁹⁶. MNA is a natural framework to integrate genomic, epigenomic and clinical data. Gene expression, copy number variations (CNV) and CpG sites can be put together to create comprehensive interaction networks using WGCNA or other approaches to fully capture interactions and coordination among gene expression, DNA methylation and DNA sequence variations. Regulatory relationships identified through the analysis of ChIP-seq data, expression-associated CNVs (eCNVs) and expression-associated DNA methylation sites (eMSs) can be taken as priors for causal network inference⁹⁷. In addition, such analyses can reveal how DNA sequence variations across individuals help to determine epigenomic modifications (including DNA methylation, histone modifications and many others), both within those regions (“cis”) or at distant interacting genomic regions (“trans”). Genetic and epigenetic markers that are identified in conjunction with altered gene expression reflect molecular features related to phenotypic changes. Several approaches have been developed to formally disentangle the causality among differences in DNA sequence, epigenetic modifications, and gene expression across numerous genes^{98,99}. With increasingly available large scale molecular profiling and clinical data, and the rapid evolution of network inference approaches (see Figure 3), we expect to derive a far more comprehensive picture of molecular pathways and key regulators underlying neuropsychiatric diseases in the next decade.

MNA can generate a large number of hypotheses (i.e., models) concerning the underlying molecular mechanisms or biomarkers for complex diseases, such as neuropsychiatric disorders. Validation of MNA-derived models and mechanisms can only be achieved through perturbation (overexpression, knockdown, pharmacological manipulation, etc.) of individual key causal regulators in *in vitro* and eventually *in vivo* experiments from which phenotypic changes are examined and compared with predicted outcomes^{98,88,97,94,95}. Full validation of network-based mechanisms for complex diseases, however, is still prohibitive due to high costs, intensive labor efforts and a lack of suitable *in vitro* and *in vivo* models for many common nervous system disorders. Recent progress in stem cell research has made it possible to model aspects of neuropsychiatric diseases using induced pluripotent stem cells (iPSCs) derived from reprogrammed somatic cells. Such advances will likely pave a way to systematically investigate genetic and epigenetic endpoints associated with disease states in such *in vitro* systems, which could then be integrated with data obtained from human postmortem brain tissue and from animal models.

Proteomic Exploration of Epigenomic Mechanisms

Mass Spectrometry in Epigenomics Research

Mass spectrometry (MS) has become a powerful tool for the analyses of DNA methylation and histone modifications¹⁰⁰ with some newer applications to small non-coding RNAs¹⁰¹. However, much of this has been at the global chromatin level and not at specific genes¹⁰². Identification and quantification of histone PTMs from a variety of cells and tissues have

been the focus of several studies, and typically involve examining the protein by Bottom Up, Middle Down or Top Down MS¹⁰².

Bottom Up MS refers to digestion of the protein into small (<3 kDa) size peptides, and analysis of those peptide typically by nanoLC-MS/MS. These experiments are fairly high-throughput and easier to perform than Middle or Top Down MS. The Bottom Up MS approach has been coupled with many diverse platforms and quantification schemes (e.g., stable isotope labeling approaches)¹⁰³ to compare the histone PTM profiles from various cellular sources and states. This Bottom Up approach has been applied to the global analysis of core histone PTMs and variants from whole brain of adult C57BL/6 mice¹⁰⁴. This large-scale proteomics analysis identified more than 10,000 peptides creating a dataset that containing 146 modification sites on 1475 peptides in various combinatorial patterns on short peptides. Among these histone peptides, 58 novel sites of modification were discovered. Bottom Up MS in fact has led the way to identify novel marks on histones, such as crotonylation, succinylation, malonylation and 2-hydroxyisobutyrylation of lysine residues¹⁰⁵⁻¹⁰⁷ and serine/threonine O-acetylation¹⁰⁸.

Middle Down MS involves interrogation of polypeptides over the 3 kDa molecular weight range. These MS experiments allow for long-range spanning PTMs to be identified on the same peptide to determine a long distance combinatorial code, and has mainly been focused on the N-terminal tails of the histone proteins. Currently, it has been determined by this approach that there exists >200 combinatorially modified N-terminal forms of histone H3, with about half as many N-terminal forms existing on histone H4¹⁰⁹. These numbers are far lower than those that are theoretically possible, seemingly indicating that much specificity exists for creating multiply modified histone forms. Kelleher and co-workers used this approach to identify histone PTM patterns and variants from whole brain, cerebral cortex, cerebellum and hypothalamus from 7-8 week old Sprague–Dawley rats¹¹⁰. Interesting patterns of modifications were found, including more silencing modifications such as H3K9me3 in the cerebellum.

Analysis of intact histone proteins (Top Down MS) is a challenging task, but allows for the characterization of combinations of modifications across the entire protein sequence. Top Down MS has been used to identify over 700 unique histone forms across all the core histones¹¹¹ but still remains the most technically challenging MS experiment to perform.

Data Processing for Histone PTM Analyses

As can be inferred from the above overview, the data analysis of histones is very complicated due to the many different types of modifications that can be detected, the distinct combinations of these PTMs that can be found and the low level stoichiometry of many histone PTMs, which can often give rise to false positive assignments. Currently, there are numerous computational approaches for both the qualitative and quantitative analysis of histone proteomics datasets. The general workflow for proteomics data is that peptides are resolved by liquid chromatography, typically electrosprayed ionized into the mass spectrometer and full MS spectra of the precursor ions collected. The precursor peptide ions are then selected and fragmented to obtain tandem mass spectra (MS/MS spectra) by various approaches including high energy C-trap dissociation (HCD), collision induced dissociation

(CID) or electron transfer dissociation (ETD). Peptides and associated modifications are identified by searching the MS/MS spectra against an organism specific database.

As histone PTM patterns are challenging, several in-house developed algorithms to deal with these datasets have been developed. (e.g., MILP or PTMap)¹¹²⁻¹¹⁴. However, several other database searching engines are commercially available for analyzing histone proteomics data with the most common being Mascot¹¹⁵ or SEQUEST¹¹⁶ but with many others emerging with different specificities or advantages¹¹⁷⁻¹¹⁹. Searching for a large number of PTMs in one search drastically increases the number of candidate peptides, slowing computational searching speed, increasing false positive IDs and even identifying less spectra. Therefore, it has been shown with histone data to best search for classes of modifications one at a time to reduce false positive misassignments. Additionally and equally important to these analyses for histone PTM proteomics datasets is the use of a site localization scoring algorithm¹²⁰. As histone PTMs occur on peptides with multiple acceptor amino acids (e.g., several lysine residues within a single peptide), it is necessary to either manually confirm the fragment ions or use a localization scoring algorithm to confidently determine on which residue the modification site occurs.

Combining Proteomic with Genomic and Epigenomic Analyses

Most histone proteomics data have been generated from global chromatin extractions, and hence the genomic loci involved are usually lost. The strength in being able to combine genomics-epigenomics level information with proteomics data is that an unbiased assessment of histone PTMs could be performed and mapped back to genomic loci. Several approaches based on tagging of proteins or using antibodies to isolate nucleosomes have been performed¹²¹⁻¹²³ and have resulted in a proteomic portrait of the genomic landscape. For example, histone PTMs and other proteins associated with MSL (male-specific lethal) associated proteins in *Drosophila* S2 cells were recently mapped by quantitative proteomics¹²³. A ChIP-seq and proteomics study also found that Brd (bromodomain) proteins are found on active *Hox* genes in cultured HEK cells, with distinct patterns of histone H4 acetylation¹²⁴. However, both of these studies map histone PTMs, not to a specific locus, but rather over a distribution of enriched genes. Experiments that isolate DNA specific sequences have still greater power to establish protein profiles and histone PTM patterns at a distinct gene. One of the first attempts at this was by Kingston and co-workers using PiCh (*Proteomics of isolated Chromatin segments*) to isolate telomeric repeats¹²⁵. Other approaches such as ChAP-MS and iChIP have also all been shown to be very useful for isolation of genomic material for mass spectrometry interrogation^{126,127}, however, it remains to be seen if they will be highly adopted.

Human Brain Evolution and the Challenge of Functional Genomics

Cognitive abilities and neurological and psychiatric diseases apparently unique to modern humans may result from genomic features distinguishing our brains from those of other primates. Because protein coding sequences for synaptic and other neuron-specific genes are extremely highly conserved within the Order Primates¹²⁸, a significant portion of hominid evolution could be due to DNA sequence changes involving regulatory and non-coding DNA¹²⁹. However, exploring human-specific regulatory DNA is a daunting task considering

that the chimpanzee-human genome comparison alone reveals close to 35×10^6 single bp and 5×10^6 multi-bp substitutions and insertion/deletion events¹³⁰, with the majority of DNA in the human genome encoding functionally active transcripts¹³¹.

Thus, given that transcription is the first step connecting genetic information to phenotype, the initial studies on deep sequencing of human and non-human primate (mostly chimpanzee and macaque) brains are providing a treasure trove of human-specific gene expression signatures, with hundreds or thousands of human-specific transcripts, particularly in the genome's non-repetitive intergenic regions¹³², with 16% of the estimated 8,000 adenosine-to-inosine exonic RNA editing sites in the cerebral and cerebellar cortex defined by human-specific substitution patterns¹³³. Deep sequencing of human and non-human primate transcriptomes becomes particularly powerful in conjunction with comparative genome analyses. For example, a recent study demonstrated that gene expression divergences in lipid catabolism pathways in the brain of modern humans are driven by shared ancestry with Neanderthals¹³⁴. In contrast to the rapidly increasing number of RNA-seq studies in primate brains, the deep sequencing of chromatin, including DNA methylation and histone modification profiles, has barely begun. For example, the recent deep sequencing of H3K4me3 revealed hundreds of regulatory sequences with human-specific epigenomic regulation in prefrontal cortical neurons¹².

Future Challenges

Characterizing the epigenome of the brain is a daunting task. The very large number of distinct types of epigenomic regulation—including several different types of DNA methylation and the great diversity and number of histone modifications and chromatin regulatory proteins—highlights the vast amount of ChIP-seq and related data that must be generated to capture the complete epigenome. Moreover, there is not one “brain epigenome” but a distinct epigenome for each neuronal and glia cell type in the brain, which likely is in the thousands. This is based on increasing evidence that patterns of DNA methylation and histone modifications at promoters and enhancers are highly specific for cell type and tissue^{135,4}. A key challenge for the field therefore is to develop a consensus for how many such epigenomes must be generated to drive an improved understanding of the brain and its diseases.

Ultimately, this expanding knowledge of the brain's epigenomes will shed unique light onto mechanisms underlying transcriptional regulation. Such analyses should incorporate richer RNA-seq datasets that provide not simply a single cross-section of RNA expression levels, but provide detailed time courses of how patterns of RNA expression shift dynamically over time in disease models. As well, work in animal models should be integrated increasingly with studies of postmortem human brain tissue, and even with human gene sequencing data and phenotypic characterizations, to drive translational discoveries. The major bioinformatics challenge is thus to optimize the tools to overlay these different and each vast datasets to derive maximal insight into transcriptional control in health and disease. A crucial step is to ensure that all genome-wide data are placed within the public domain in ways that allow raw datasets to be reanalyzed as advances proceed (Box 2).

The particular complexity of the genetic risk for most common brain disorders adds yet another challenge. For example, a recent GWAS of the Psychiatric Genomics Consortium, involving approximately 40,000 subjects, estimated that 8,300 independent SNPs, positioned mostly in intergenic and (protein)-non-coding sequences, contributed to the genetic risk for schizophrenia¹³⁶. Strikingly, however, there was no significant enrichment for DNase I hypersensitive sites (which broadly define open chromatin primarily at sites of active promoters, enhancers and expressed genes) identified by the ENCODE project¹³⁷. This contrasts with the robust enrichment for DNase I hypersensitive sites of the general GWAS catalog of the National Genome Research Institute¹³⁸. Because the ENCODE database is exclusively built on epigenomic mappings from peripheral cell lines and tissues, it is clear that similar efforts, focused on brain and some of its surrogates, such as pluripotent stem cell-derived neural cultures¹³⁹ or cerebral organoids¹⁴⁰, are now needed to obtain a deeper understanding of the genetic risk architectures of these complex disorders. Furthermore, extending these analyses to disease-linked alterations in chromatin structure over large genomic regions in brain (Box 3) promises to provide great insights into the molecular causes of these heterogeneous syndromes.

To this end, the National Institutes of Health-based initiatives that include epigenomic mappings of human brain, including the Epigenomics Roadmap or, more recently, PsychENCODE, will provide critical starting points to what should be the ultimate goal of modern neuroepigenomics: whole genome, high resolution (single base pair) maps for a large number of epigenomic marks for key anatomically defined brain regions and their important subpopulations of neurons and glia. This type of resource is likely needed to expedite the dramatic, game-changing discoveries from integrated studies of animal models and human patients. Such transformational advances are sorely needed to finally overcome the limited success in diagnostics and drug discovery efforts over the past several decades and to develop truly improved diagnostic tests and treatments for a host of severe neurological and psychiatric disorders.

Text Box 1

Unique Challenges in Neuroepigenomics

Epigenomic studies of the brain are more challenging due to several considerations. In contrast to studies of cultured cells or most peripheral tissues, a discrete brain region of interest must be isolated by microdissection, which adds considerable variability to any study of the brain. Moreover, because many brain regions of interest are very small (~1 mm³ in a mouse), tissue is highly limiting. This necessitates multiple analyses to be performed on different collections of tissue from different cohorts of animals, which further increases the imprecision inherent in brain investigations.

Brain tissue is also highly heterogeneous, with all brain regions containing numerous types of neuronal, glial and vascular cells. The fact that each cell type has a distinct epigenome further increases the “noise” of brain epigenomic data compared to that from simpler systems. In addition, certain diseases might involve shifts in cell type (e.g., loss of neurons in neurodegenerative disorders or invasion of immune cells in neuroinflammatory disorders), which complicates the interpretation of epigenomic data.

Epigenomic analysis of individual cell types—the isolation of which or their nuclei is becoming increasingly possible by use of genetic or viral tags¹⁴¹—represents one promising approach for the future. However, it is not yet feasible to obtain the number of isolated cells or nuclei from smaller brain regions that are required for most epigenomic analyses. Refinement of techniques that make ChIP-seq and related procedures possible with a far smaller number of cells or nuclei will advance the field dramatically^{4,142,143}.

It has long been known that regulation of neural phenomena often involves changes in protein levels or activities that are far smaller than those seen in studies of other systems. As just one example, earlier microarray studies of cultured cells, peripheral tissues or tumors might have a 2-fold cutoff, while studies of brain reveal few changes of this magnitude. This is not solely due to the confound of multiple cell types, because similar findings have been obtained from analysis of single cell types. Rather, it is likely that fully differentiated neurons within the context of an intricate circuit do not show the same degree of adaptation displayed by other cell types. This requirement to reliably detect changes of relatively small magnitude (e.g., 20%)—which are demonstrably functionally important^{e.g., 1,2,15}—adds another unique burden on bioinformatics analysis of neuroepigenomic data.

Text Box 2

Data Deposition

The sequence read archive (SRA)¹⁴⁴ was established as a joint effort between multiple nations and has become the central depository for next-generation sequencing data. It can be accessed at several different URLs depending on the user's physical location and preference:

Provider	URL
NCBI	http://www.ncbi.nlm.nih.gov/Traces/sra
EBI	http://www.ebi.ac.uk/ena
DDBJ	http://www.ddbj.nig.ac.jp/sub/trace_sra-e.html

For functional genomics studies involving short sequence data, such as transcriptomics, small RNA profiling, ChIP-seq and BS-seq, users can upload their data through Gene Expression Omnibus (GEO) at: <http://www.ncbi.nlm.nih.gov/geo/info/seq.html>.

Although SRA also provides an online submission portal, it is often more convenient to submit data (especially for large studies) using GEO. Creating a GEO submission typically requires compiling all raw data, processed data and a metadata spreadsheet into a folder. These data are then transferred to GEO using FTP. For the metadata spreadsheet, submitters are required to briefly describe their study, list all samples including their properties and associated raw data and processed data files, describe experimental protocols, and elaborate on data processing steps and programs used to analyze their data, as well as parameter settings. After data have been successfully transferred, submitters can send an email to GEO to initiate a review of the submission.

A GEO curator will examine uploaded files and provide feedback. A GEO accession number will then be created and provided to the submitter. The submitter can then log into their account and review submissions. At this point, submissions remain private. However, a reviewer link can be created to share submissions with journal reviewers. Once manuscripts associated with submissions have been accepted for publication, submitters can return to their login page and make the data public.

The deposition of all genome-wide data into a single resource worldwide is essential for financial reasons: each study is so expensive and so many studies are needed that it is essential to avoid redundant efforts. Furthermore, such datasets will increasingly be of great use to investigators worldwide as raw data can be reanalyzed with newer, more powerful tools and as interpretation of a new epigenomic mark can be vastly improved by overlays of maps of many available marks.

Text Box 3

HiC: A Sequencing Approach to Characterize Interactions Over Large Genomic Regions

HiC is an unbiased high throughput method to detect chromatin-looping interactions between all loci in the genome¹⁴⁵. HiC experiments have shown that genome structure and function are linked. In yeast, looping interaction groups form between highly transcribed genes, co-regulated genes (associated with different motifs) and gene ontology groups^{146,147}. Human HiC studies cluster looping interactions into three groups: self-interacting active gene rich clusters, non-active centromere proximal clusters and non-active centromere distal clusters¹⁴⁸. In bacteria, yeast, and humans, chromatin loops place enhancers and promoters in close spatial proximity, which are thought to exist in topological domains bound by insulator proteins.

HiC libraries are created using proximity ligation. Tissue or cells are homogenized and crosslinked using 1% formaldehyde and restricted using an enzyme of choice. The 5' overhangs are filled in with a biotinylated CTP residue and blunt-end ligation is performed under dilute conditions. The library is sheared for 300-500 bp size selection and biotinylated fragments are immunoprecipitated using streptavidin beads. The library is pair-end sequenced and aligned to the corresponding genome. Reads that map to two locations are scored as interactions and an interaction matrix from all interactions in the experiment is created. An expected interaction matrix is used to calculate the statistical significance of all scored experimental interactions.

Current HiC algorithms model looping interactions probabilistically by taking into account mappability, fragment length and GC percentage (which is correlated with gene density, banding patterns, repetitive content and chromatin marks)¹⁴⁸. The best algorithm to date can resolve a human HiC map from cell lines to 5-10 kb using 63% mappable reads (300 million reads mapped from 500 million total reads) using four lanes of Illumina sequencing¹⁴⁹. Problems from these algorithms arise from the probabilistic nature of proximity ligation. HiC detects the average chromatin structure within a population of cells. Failure to detect looping interactions does not mean they do not exist,

but rather that the current method does not detect them¹⁴⁸. In time, the efficiency of algorithms will further increase, and the cost of sequencing will decrease, making HiC a more amenable approach for cell-type specific studies in brain.

Acknowledgments

We would like to thank Dr. Alexey Soshnev for help with illustrations.. This work was supported by grants from the National Institute of Mental Health (NIMH): 5R01 MH094698 and P50 MH096890 and the National Institute on Drug Abuse: P01 DA008227.

References

1. Renthal W, Carle TL, et al. Delta FosB mediates epigenetic desensitization of the c-fos gene after chronic amphetamine exposure. *J Neurosci*. 2008; 28:7344–9. [PubMed: 18632938]
2. Maze I, Covington HE 3rd, et al. Essential role of the histone methyltransferase G9a in cocaine-induced plasticity. *Science*. 2010; 327:213–6. [PubMed: 20056891]
3. Visel A, Blow MJ, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009; 457:854–8. [PubMed: 19212405]
4. Cheung I, Shulha HP, et al. Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107:8824–9. [PubMed: 20421462]
5. Peleg S, Sananbenesi F, et al. Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science*. 2010; 328:753–6. [PubMed: 20448184]
6. Guo JU, Ma DK, et al. Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nature Neuroscience*. 2011; 14:1345–51.
7. Maze I, Feng J, et al. Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:3035–40. [PubMed: 21300862]
8. Szulwach KE, Li X, et al. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature Neuroscience*. 2011; 14:1607–16.
9. Zhou Z, Yuan Q, et al. Substance-specific and shared transcription and epigenetic changes in the human hippocampus chronically exposed to cocaine and alcohol. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:6626–31. [PubMed: 21464311]
10. Hunter RG, Murakami G, et al. Acute stress and hippocampal histone H3 lysine 9 trimethylation, a retrotransposon silencing response. *Proceedings of the National Academy of Sciences of the United States of America*. 2012; 109:17657–62. [PubMed: 23043114]
11. Mellen M, Ayata P, et al. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*. 2012; 151:1417–30. [PubMed: 23260135]
12. Shulha HP, Crisci JL, et al. Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. *PLoS Biology*. 2012; 10:e1001427. [PubMed: 23185133]
13. Lister R, Mukamel EA, et al. Global epigenomic reconfiguration during mammalian brain development. *Science*. 2013; 341:1237905. [PubMed: 23828890]
14. Park CS, Rehrauer H, et al. Genome-wide analysis of H4K5 acetylation associated with fear memory in mice. *BMC Genomics*. 2013; 14:539. [PubMed: 23927422]
15. Zhu J, Adli M, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*. 2013; 152:642–54. [PubMed: 23333102]
16. Feng J, Wilkinson M, et al. Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biology*. 2014; 15:R65. [PubMed: 24758366]
17. Guo JU, Su Y, et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nature Neuroscience*. 2014; 17:215–22.

18. Scobie KN, Damez-Werno D, et al. Essential role of poly(ADP-ribosyl)ation in cocaine action. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111:2005–10. [PubMed: 24449909]
19. TBD. Other review on epigenetic methods in this special issue. *Nature Neuroscience*. 2014
20. Mortazavi A, Williams BA, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*. 2008; 5:621–8. [PubMed: 18516045]
21. Nagalakshmi U, Wang Z, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008; 320:1344–9. [PubMed: 18451266]
22. Wilhelm BT, Marguerat S, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008; 453:1239–43. [PubMed: 18488015]
23. Marioni JC, Mason CE, et al. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*. 2008; 18:1509–17. [PubMed: 18550803]
24. Nookaew I, Papini M, et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Research*. 2012; 40:10084–97. [PubMed: 22965124]
25. Zhao S, Fung-Leung WP, et al. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*. 2014; 9:e78644. [PubMed: 24454679]
26. DeLuca DS, Levin JZ, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012; 28:1530–1532. [PubMed: 22539670]
27. Shen L, Shao N, et al. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC genomics*. 2014; 15:284. [PubMed: 24735413]
28. Trapnell C, Williams BA, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. 2010; 28:511–5.
29. Grabherr MG, Haas BJ, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. 2011; 29:644–52.
30. Trapnell C, Pachter L, et al. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009; 25:1105–1111. [PubMed: 19289445]
31. Dobin A, Davis CA, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013; 29:15–21. [PubMed: 23104886]
32. Engstrom PG, Steijger T, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*. 2013; 10:1185–91. [PubMed: 24185836]
33. Trapnell C, Williams BA, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*. 2010; 28:511–515.
34. Katz Y, Wang ET, et al. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. 2010; 7:1009–15. [PubMed: 21057496]
35. Deal RB, Henikoff S. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Developmental Cell*. 2010; 18:1030–40. [PubMed: 20627084]
36. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010; 26:841–2. [PubMed: 20110278]
37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. 2011; 17:10–12.
38. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*. 2014; 42:D68–73. [PubMed: 24275495]
39. Friedlander MR, Chen W, et al. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotech*. 2008; 26:407–415.
40. Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research*. 2006; 34:D158–62. [PubMed: 16381836]
41. Sai Lakshmi S, Agrawal S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Research*. 2008; 36:D173–7. [PubMed: 17881367]

42. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*. 2009; 37:D93–7. [PubMed: 18984615]
43. Mituyama T, Yamada K, et al. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Research*. 2009; 37:D89–92. [PubMed: 18948287]
44. Amaral PP, Clark MB, et al. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Research*. 2011; 39:D146–51. [PubMed: 21112873]
45. Burge SW, Daub J, et al. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*. 2013; 41:D226–32. [PubMed: 23125362]
46. Quast C, Pruesse E, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*. 2013; 41:D590–6. [PubMed: 23193283]
47. Xie C, Yuan J, et al. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Research*. 2014; 42:D98–103. [PubMed: 24285305]
48. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010; 11:R106. [PubMed: 20979621]
49. Robinson MD, McCarthy DJ, et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010; 26:139–40. [PubMed: 19910308]
50. Law C, Chen Y, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. 2014; 15:R29. [PubMed: 24485249]
51. Smyth, GK. limma: Linear Models for Microarray Data. In: Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R.; Dudoit, S., editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer; New York: 2005. p. 397–420.
52. Love MI, Huber W, et al. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv beta*. 2014
53. Rapaport F, Khanin R, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*. 2013; 14:R95. [PubMed: 24020486]
54. Liu Y, Zhou J, et al. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*. 2014; 30:301–4. [PubMed: 24319002]
55. Trapnell C, Hendrickson DG, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotech*. 2013; 31:46–53.
56. Seyednasrollah F, Laiho A, et al. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*. 2013
57. Langmead B, Trapnell C, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 2009; 10:R25. [PubMed: 19261174]
58. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–60. [PubMed: 19451168]
59. Li H, Handsaker B, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25:2078–9. [PubMed: 19505943]
60. Chen K, Xi Y, et al. DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research*. 2013; 23:341–51. [PubMed: 23193179]
61. Kundaje, A. Computes quick but highly informative enrichment and quality measures for ChIP-seq/DNase-seq/FAIRE-seq/MNase-seq data. Computer Science Dept., Stanford University, ENCODE Consortium; 2010.
62. Landt SG, Marinov GK, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*. 2012; 22:1813–1831. [PubMed: 22955991]
63. Robinson JT, Thorvaldsdóttir H, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011; 29:24–26.
64. Chen Y, Negre N, et al. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature methods*. 2012; 9:609–14. [PubMed: 22522655]
65. Carroll TS, Liang Z, et al. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Frontiers in genetics*. 2014; 5:75. [PubMed: 24782889]
66. Zhang Y, Liu T, et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*. 2008; 9

67. Heinz S, Benner C, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*. 2010; 38:576–589. [PubMed: 20513432]
68. Shen L, Shao N-Y, et al. diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. *PLoS ONE*. 2013; 8:e65598. [PubMed: 23762400]
69. Zhu LJ, Gazin C, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*. 2010; 11:237. [PubMed: 20459804]
70. Liang K, Kele S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*. 2011
71. Wright, K. corrgram: Plot a correlogram. R Package version 1.5. 2013. <http://CRAN.R-project.org/package=corrgram>
72. Landt SG, Marinov GK, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*. 2012; 22:1813–31. [PubMed: 22955991]
73. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*. 2012; 9:215–6. [PubMed: 22373907]
74. Shi X, Hong T, et al. ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature*. 2006; 442:96–9. [PubMed: 16728974]
75. Saint-André V, Batsché E, et al. Histone H3 lysine 9 trimethylation and HP1 γ favor inclusion of alternative exons. *Nat Struct Mol Biol*. 2011; 18:337–344. [PubMed: 21358630]
76. Cheng C, Yan KK, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biology*. 2011; 12
77. Dong X, Greven M, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*. 2012; 13:R53. [PubMed: 22950368]
78. Kriaucionis S, Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*. 2009; 324:929–30. [PubMed: 19372393]
79. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC bioinformatics*. 2009; 10:232. [PubMed: 19635165]
80. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. 2011; 27:1571–2. [PubMed: 21493656]
81. Bock C. Analysing and interpreting DNA methylation data. *Nature reviews. Genetics*. 2012; 13:705–19.
82. Sun D, Xi Y, et al. MOABS: model based analysis of bisulfite sequencing data. *Genome biology*. 2014; 15:R38. [PubMed: 24565500]
83. Down TA, Rakyen VK, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*. 2008; 26:779–85.
84. Goldstein DB. Common genetic variation and human traits. *N Engl J Med*. 2009; 360:1696–8. [PubMed: 19369660]
85. Kandoth C, McLellan MD, et al. Mutational landscape and significance across 12 major cancer types. *Nature*. 2013; 502:333–9. [PubMed: 24132290]
86. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013; 45:580–5. [PubMed: 23715323]
87. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005; 4 Article17.
88. Horvath S, Zhang B, et al. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc Natl Acad Sci U S A*. 2006; 103:17402–7. [PubMed: 17090670]
89. Miller JA, Oldham MC, et al. A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J Neurosci*. 2008; 28:1410–20. [PubMed: 18256261]
90. Miller JA, Horvath S, et al. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc Natl Acad Sci U S A*. 2010; 107:12698–703. [PubMed: 20616000]
91. Miller JA, Woltjer RL, et al. Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med*. 2013; 5:48. [PubMed: 23705665]

92. Luo R, Sanders SJ, et al. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. *Am J Hum Genet.* 2012; 91:38–55. [PubMed: 22726847]
93. Parikshak NN, Luo R, et al. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell.* 2013; 155:1008–21. [PubMed: 24267887]
94. Rhinn H, Fujita R, et al. Integrative genomics identifies APOE epsilon4 effectors in Alzheimer's disease. *Nature.* 2013; 500:45–50. [PubMed: 23883936]
95. Zhang B, Gaiteri C, et al. Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell.* 2013; 153:14. [PubMed: 23540687]
96. Marbach D, Costello JC, et al. Wisdom of crowds for robust gene network inference. *Nat Methods.* 2012; 9:796–804. [PubMed: 22796662]
97. Zhu J, Zhang B, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet.* 2008; 40:854–61. [PubMed: 18552845]
98. Schadt EE, Lamb J, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics.* 2005; 37:710–7. [PubMed: 15965475]
99. Swarup V, Geschwind DH. Alzheimer's disease: From big data to mechanism. *Nature.* 2013; 500:34–5. [PubMed: 23883924]
100. Everitts AG, Zee BM, et al. Modern approaches for investigating epigenetic signaling pathways. *J Appl Physiol (1985).* 2010; 109:927–33. [PubMed: 20110548]
101. Kullolli M, Knouf E, et al. Intact microRNA analysis using high resolution mass spectrometry. *J Am Soc Mass Spectrom.* 2014; 25:80–7. [PubMed: 24174127]
102. Karch KR, Denizio JE, et al. Identification and interrogation of combinatorial histone modifications. *Front Genet.* 2013; 4:264. [PubMed: 24391660]
103. Ong SE, Blagoev B, et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics.* 2002; 1:376–86. [PubMed: 12118079]
104. Tweedie-Cullen RY, Brunner AM, et al. Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain. *PLoS One.* 2012; 7:e36980. [PubMed: 22693562]
105. Tan M, Luo H, et al. Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell.* 2011; 146:1016–28. [PubMed: 21925322]
106. Xie Z, Dai J, et al. Lysine succinylation and lysine malonylation in histones. *Mol Cell Proteomics.* 2012; 11:100–7. [PubMed: 22389435]
107. Dai L, Peng C, et al. Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nature chemical biology.* 2014
108. Britton LM, Newhart A, et al. Initial characterization of histone H3 serine 10 O-acetylation. *Epigenetics.* 2013; 8:1101–13. [PubMed: 23949383]
109. Young NL, DiMaggio PA, et al. High throughput characterization of combinatorial histone codes. *Mol Cell Proteomics.* 2009; 8:2266–84. [PubMed: 19654425]
110. Garcia BA, Siuti N, et al. Characterization of neurohistone variants and post-translational modifications by electron capture dissociation mass spectrometry. *International Journal of Mass Spectrometry.* 2007; 259:184–196.
111. Tian Z, Tolic N, et al. Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.* 2012; 13:R86. [PubMed: 23034525]
112. Frank AM, Pesavento JJ, et al. Interpreting top-down mass spectra using spectral alignment. *Anal Chem.* 2008; 80:2499–505. [PubMed: 18302345]
113. Chen Y, Chen W, et al. PTMap--a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites. *Proc Natl Acad Sci U S A.* 2009; 106:761–6. [PubMed: 19136633]
114. DiMaggio PA Jr. Young NL, et al. A mixed integer linear optimization framework for the identification and quantification of targeted post-translational modifications of highly modified proteins using multiplexed electron transfer dissociation tandem mass spectrometry. *Mol Cell Proteomics.* 2009; 8:2527–43. [PubMed: 19666874]

115. Perkins DN, Pappin DJ, et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999; 20:3551–67. [PubMed: 10612281]
116. Eng JK, McCormack AL, et al. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*. 1994; 5:976–89. [PubMed: 24226387]
117. Geer LY, Markey SP, et al. Open mass spectrometry search algorithm. *J Proteome Res*. 2004; 3:958–64. [PubMed: 15473683]
118. Wang LH, Li DQ, et al. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun Mass Spectrom*. 2007; 21:2985–91. [PubMed: 17702057]
119. Zhang J, Xin L, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics*. 2012; 11 M111 010587.
120. Beausoleil SA, Villen J, et al. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat Biotechnol*. 2006; 24:1285–92. [PubMed: 16964243]
121. Tackett AJ, Dilworth DJ, et al. Proteomic and genomic characterization of chromatin complexes at a boundary. *J Cell Biol*. 2005; 169:35–47. [PubMed: 15824130]
122. Voigt P, LeRoy G, et al. Asymmetrically modified nucleosomes. *Cell*. 2012; 151:181–93. [PubMed: 23021224]
123. Wang CI, Alekseyenko AA, et al. Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in *Drosophila*. *Nat Struct Mol Biol*. 2013; 20:202–9. [PubMed: 23295261]
124. Leroy G, Chepelev I, et al. Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins. *Genome Biol*. 2012; 13:R68. [PubMed: 22897906]
125. Dejardin J, Kingston RE. Purification of proteins associated with specific genomic Loci. *Cell*. 2009; 136:175–86. [PubMed: 19135898]
126. Hoshino A, Fujii H. Insertional chromatin immunoprecipitation: a method for isolating specific genomic regions. *J Biosci Bioeng*. 2009; 108:446–9. [PubMed: 19804873]
127. Byrum SD, Taverna SD, et al. Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Res*. 2013; 41:e195. [PubMed: 24030711]
128. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science*. 1975; 188:107–16. [PubMed: 1090005]
129. McLean CY, Reno PL, et al. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature*. 2011; 471:216–9. [PubMed: 21390129]
130. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 2005; 437:69–87. [PubMed: 16136131]
131. Dunham I, Kundaje A, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
132. Xu AG, He L, et al. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput Biol*. 2010; 6:e1000843. [PubMed: 20617162]
133. Li Z, Bammann H, et al. Evolutionary and ontogenetic changes in RNA editing in human, chimpanzee, and macaque brains. *RNA*. 2013; 19:1693–702. [PubMed: 24152549]
134. Khrameeva EE, Bozek K, et al. Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat Commun*. 2014; 5:3584. [PubMed: 24690587]
135. Heintzman ND, Hon GC, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009; 459:108–12. [PubMed: 19295514]
136. Ripke S, O'Dushlaine C, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*. 2013; 45:1150–9. [PubMed: 23974872]
137. Bernstein BE, Birney E, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
138. Maurano MT, Humbert R, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–5. [PubMed: 22955828]

139. Brennand KJ, Simone A, et al. Modeling psychiatric disorders at the cellular and network levels. *Mol Psychiatry*. 2012; 17:1239–53. [PubMed: 22472874]
140. Lancaster MA, Renner M, et al. Cerebral organoids model human brain development and microcephaly. *Nature*. 2013; 501:373–9. [PubMed: 23995685]
141. Jiang Y, Matevosian A, et al. Isolation of neuronal chromatin from brain tissue. *BMC neuroscience*. 2008; 9:42. [PubMed: 18442397]
142. Evrony GD, Cai X, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*. 2012; 151:483–96. [PubMed: 23101622]
143. Grindberg RV, Yee-Greenbaum JL, et al. RNA-sequencing from single nuclei. *Proceedings of the National Academy of Sciences of the United States of America*. 2013; 110:19802–7. [PubMed: 24248345]
144. Shumway M, Cochrane G, et al. Archiving next generation sequencing data. *Nucleic Acids Research*. 2010; 38:D870–D871. [PubMed: 19965774]
145. Lieberman-Aiden E, van Berkum NL, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009; 326:289–93. [PubMed: 19815776]
146. Tanizawa H, Iwasaki O, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic acids research*. 2010; 38:8164–77. [PubMed: 21030438]
147. Gehlen LR, Gruenert G, et al. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus*. 2012; 3:370–83. [PubMed: 22688649]
148. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*. 2011; 43:1059–65. [PubMed: 22001755]
149. Jin F, Li Y, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. 2013; 503:290–4. [PubMed: 24141950]
150. Hackenberg M, Rodriguez-Ezpeleta N, et al. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research*. 2011; 39:W132–8. [PubMed: 21515631]

Figure 1. Initial Pipelines of RNA-seq Data Analysis

Following data acquisition, RNA-seq analyses typically begin with numerous quality control assessments using analytical tools such as fastqc. Next, for analysis of genic transcripts, splice variants and long non-coding RNAs, short sequencing reads can be aligned to a reference genome using programs such as Tophat or STAR. Following alignment, additional quality control assessments can be made with RNA-SeQC and ngs.plot [orange line: RNA-seq plot from a human postmortem brain sample with a high RNA integrity number (RIN = 7.8); green line: RNA-seq plot from a human postmortem brain sample with a low RIN value (RIN = 3) displaying a clear 3' bias]. ngs.plot image used with permission from²⁷. Finally, to quantify and analyze RNA-seq data, programs such as HTSeq, Cufflinks or MISO are typically used. Depending on experimental purification schemes, researchers may also wish to analyze small ncRNAs from their samples. To do so, following initial quality control analysis, adapters must first be removed from sequence reads using Cutadapt or Fastx, followed by Bowtie alignment to a reference genome and quantitation using a program such as miRanalyzer¹⁵⁰. Additional ncRNA specific analyses can similarly be integrated into miRanalyzer's pipeline, or independent databases (e.g., SILVA, piRNABank, etc.) can be used.

Figure 2. ChIP-seq Analysis of Brain

A) Assessment of ChIP-seq sample quality, represented here as the sequence quality score vs. bp position using fastqc, is a critical first step in ChIP-seq data analysis and is required for appropriate interpretations of subsequent downstream analyses. The example shows relatively poor quality data, given the high error rate and variability toward the 3' end of the reads. B) Example pie charts describing the genome-wide distribution patterns of differential histone modification sites for H3K4me3, H3K4me1 and H3K27me3 (as determined using diffReps) in nucleus accumbens (NAc) of saline vs. chronic cocaine treated mice¹⁶. C) Example gene body plots for H3K36me3 in NAc of control mice derived by ngs.plot. Lines represent average profiles for four gene groups ordered by mRNA expression levels by RNA-seq, defined as "high," "med2," "med1" and "low," and illustrates increasing levels of this histone mark with increasing gene expression. ngs.plot easily generates these kinds of figures for accessible representations of protein enrichment throughout different functional genomic regions. D) Representative log2 enrichment heat maps generated in ngs.plot for several histone marks vs. DNA input in mouse NAc at TSS $\pm$ 5 kb genomic regions. Gene expression levels analyzed by RNA-seq are illustrated by enrichment in the same gene order as ChIP-seq enrichment patterns. Red is high, green is low. E) Integrated genomics viewer (IGV) screenshots for ChIP-seq data of various histone marks in NAc of cocaine treated mice. The genomic region displayed represents the TSS and ~20 kb downstream of the *Rbfox1* (*A2bp1*) gene, a splicing factor which is highly expressed in neurons. F) Chromatin signatures, such as those shown here, can be defined to characterize groups of transcripts displaying similar patterns of chromatin modifications at specific genomic regions following environmental stimuli (e.g., at splicing related regions following cocaine treatment)¹⁶. Motif finding can then be used to identify potential transcriptional and splicing associated factors deduced to regulate these signatures. G) IGV genome browser screenshot for ChIP-seq

coverage of H3K4me3 from NAc of control mice. Three biological replicates are shown as separate tracks with significant peaks identified by MACS depicted as solid bars beneath the tracks. The genomic region depicted is ~ chr11:116,974,000-116,990,000. These data highlight the difficulty of using peak calling-based approaches to identify differential enrichment patterns across samples. Here, although the three biological replicates appear to be generally similar in size and distribution, MACS identified discordant peaks between the samples due to intrinsic variations in peak location. Unlike MACS, diffReps uses a sliding window approach that allows one to focus on a fixed sized region across all samples (e.g., 1 kb), thus allowing for more unified comparisons across samples. H) ChIP-seq differential analysis using diffReps to compare two groups of samples representing two distinct biological conditions and test for significance. Here, diffReps was used to compare differential H3K4me3 enrichment in NAc between saline *vs.* cocaine treated mice at the TSS of the *Enah* gene.

Figure 3. Network Inference Approaches in Neuroscience

Diagrammatic representation of systems/network approaches to integrate large scale genomic, epigenomic and clinical data to inform investigations of neurological and psychiatric disorders using data derived from humans and animal models. Correlational, interaction and causal analyses can be unified under network level frameworks to generate data-driven hypotheses (i.e., models) concerning the underlying molecular mechanisms or biomarkers resulting in neuronal dysfunction. These models are then further validated experimentally, data that can be fed back into modeling processes to refine biological predictions.

Table 1

Progress Report of Epigenomic Data from Brain

Year	Species/Brain region(s) examined	Modification(s)/DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Reference
2009	Mouse/embryonic forebrain and midbrain	p300	ChIP-seq	Basal state	First evidence that in vivo mapping of p300 can be used to accurately identify tissue-specific enhancers and their associated activities	3
2010	Human/prefrontal cortex, neurons vs. non-neuronal cells	H3K4me3	ChIP-seq	Age	This study revealed age-correlated neuronal specific epigenomic patterns of reorganization demonstrating that H3K4me3 in human prefrontal cortex is highly regulated in a cell type- and subject-specific manner	4
2010	Mouse/adult hippocampus	H4K12ac	ChIP-seq/microarray	Fear conditioned learning	During periods of learning, aged mice display deregulated H4K12ac and fail to initiate hippocampal gene expression programs associated with memory consolidation	5
2011	Mouse/adult hippocampal dentate granule cells	5mC and 5hmC	MethylC-seq/BS-seq/microarray	Electroconvulsive stimulation	First <i>in vivo</i> genome-wide single base resolution maps of 5mC and 5hmC in adult hippocampal dentate granule cells. These data indicated a functionally important role for stimulus-dependent alterations in the CNS's DNA methylome in the regulation of activity-dependent gene expression	6
2011	Mouse/adult nucleus accumbens	H3K9me3	ChIP-seq	Chronic cocaine	These data demonstrated that chronic cocaine exposure promotes reduced	7

Year	Species/Brain region(s) examined	Modification(s)/DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Reference
					expression of H3K9me3 within specific heterochromatic loci of the nucleus accumbens epigenome leading to intergenic instability	
2011	Mouse/early postnatal and adult hippocampus and cerebellum Human/adult cerebellum	5hmC	Chemical labeling and immunoprecipitation/ChIP-seq	Age/ <i>Mecp2</i> overexpression and knockout	Provided the first genome-wide maps of 5hmC in mouse hippocampus and cerebellum during periods of development and aging. This work further demonstrated alterations in 5hmC genomic content in mouse models of Rett syndrome indicating a critical role for 5hmC in neurodevelopment and human disease	8
2011	Human/adult hippocampus	H3K4me3	ChIP-seq/RNA-seq	Cocaine and alcohol addiction	This study identified substance-specific and shared transcriptional and chromatin related changes in human hippocampus of individuals chronically exposed to cocaine and alcohol	9
2012	Rat/adult hippocampus	H3K9me3	ChIP-seq	Acute restraint stress	Acute stress was found to result in increased enrichment of heterochromatic H3K9me3 in hippocampus, a finding hypothesized to be important for the repressive maintenance of endogenous retrotransposable elements in brain	10
2012	Mouse/adult cerebellar Purkinje, granule and Bergmann glial cells	5mC, 5hmC and non-CpG methylation	MeDIP-seq/TRAP-seq	Cell type, <i>Mecp2</i> knockout	These data demonstrated that the relationship between 5hmC, genomic	11

Year	Species/Brain region(s) examined	Modification(s)/DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Reference
					distribution of 5mC and gene expression is cell type specific, and that MeCP2 is the major 5hmC-binding protein in brain	
2012	Human, chimpanzee and macaque/ adult prefrontal cortex, neurons vs. non-neuronal cells	H3K4me3	ChIP-seq	Species	The findings presented here provided the first insights into human-specific modifications of the neuronal epigenome, with evidence for coordinated epigenetic regulation of sites separated by megabases of interspersed sequence	12
2013	Human and mouse/frontal cortex, neurons vs. non-neuronal cells	5mC and 5hmC	MethylC-seq/RNA-seq	Cell type, age	This study reported the genome-wide composition, patterning, cell specificity and dynamics of DNA methylation at single-base resolution in human and mouse cortex throughout lifespan, and identified a novel role for accumulation of non-CpG methylation during brain development	13
2013	Mouse/adult hippocampus	H4K5ac	ChIP-seq/microarray	Fear conditioned learning	These data provided insights into the mechanisms of gene priming and 'bookmarking' by histone acetylation following hippocampal memory activation, and proposed that increased H4Kac5 may prime genes for rapid expression following activity	14
2013	Human/adult	H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9ac and H3K27ac	ChIP-seq/microarray	Cell/tissue types	This study described how global chromatin state transitions relate to	15

Year	Species/Brain region(s) examined	Modification(s)/DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Reference
					chromosomal and nuclear architectures, and hypothesized implications for lineage fidelity, cellular senescence and reprogramming.	
2014	Mouse/adult nucleus accumbens	H3K4me3, H3K4me1, H3K27me3, H3K9me2, H3K9me3, H3K36me3 and RNA Pol II	ChIP-seq/RNA-seq	Chronic cocaine	This study demonstrated that repeated cocaine treatment results in an unexpectedly large number of pre-mRNA splicing alterations in nucleus accumbens, and that specific combinations of chromatin changes (i.e., signatures) predict these events	16
2014	Mouse/adult hippocampal dentate granule cells Human brain	5mC (CpG) and non-CpG methylation (CpH)	BS-seq/ChIP-seq/RNA-seq	Age/triple <i>Dnmt1</i> , <i>Dnmt3a</i> and <i>Dnmt3b</i> knockout	This study, using single-base resolution maps of the neuronal DNA methylome, identified high levels of both CpG and CpH methylation, patterns which were shown to be conserved in human brain. Functionally, both mCpGs and mCpHs were shown to repress transcription <i>in vitro</i> and are recognized by MeCP2. Finally, mCpH was demonstrated to be established <i>de novo</i> during neurodevelopment requiring Dnmt3a for active maintenance in neurons	17
2014	Mouse/adult nucleus accumbens	PARP-1 and H3K4me3	ChIP-seq/RNA-seq	Chronic cocaine	This work demonstrated that chronic cocaine exposure results in increases in PARP-1 expression and activity in nucleus accumbens, which mediates the	18

Year	Species/Brain region(s) examined	Modification(s)/DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Reference
					expression of numerous target genes (e.g., SDK1) to promote cocaine related behaviors	

Selected list of genome-wide neuroepigenomic analyses carried out in brains of human or other mammalian species since 2009.

Table 2

Existing Databases for Analysis of ncRNAs

ncRNAs [*]	Database	Description	Reference
General purpose	Rfam (http://rfam.xfam.org/)	Sister database of Pfam ^{**} , collections of ncRNA families	45
	NONCODE (http://www.noncode.org/)	Database of ncRNA, excluding rRNA and tRNA	47
	fRNAdb (http://www.ncrna.org/frnadb/)	Database of functional ncRNAs	43
miRNA	miRBase (http://www.mirbase.org/)	Database for miRNAs	38
piRNA	piRNABank (http://pimabank.ibab.ac.in/)	Database for piRNAs, clustering information provided	41
snoRNA	snoRNA-LBME-db (https://www-snorna.biotoul.fr/)	Comprehensive database for human snoRNAs	40
lncRNA	lncRNAdb (http://www.lncrnadb.org/)	Database for lncRNAs	44
rRNA	SILVA (http://www.arb-silva.de/)	A curated database for rRNAs	46
tRNA	GtRNAdb (http://gtrnadb.ucsc.edu/)	A genome-wide tRNA database predicted by the program tRNAscan-SE	42

^{*} This list is not comprehensive but aims to provide an overview of currently available databases for analyzing ncRNAs from biological samples. The recent identification of novel ncRNAs, such as circRNAs, which may play critical roles in the CNS, are not included, as analysis tools do not currently exist.

^{**} Pfam: a database widely used in the analysis/annotation of protein sequences that uses hidden Markov model based multiple sequence alignments to provide detailed information about protein families and domains.