nature neuroscience

Analytical tools and current challenges in the modern era of neuroepigenomics

Ian Maze[1,2], Li Shen[2], Bin Zhang[3], Benjamin A Garcia[4], Ningyi Shao[2], Amanda Mitchell[2,5], HaoSheng Sun[2], Schahram Akbarian[2,5], C David Allis[6] & Eric J Nestler[2]

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 CNS, however, is more recent. Here we provide a comprehensive overview of available tools for analyzing 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.

Our 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. 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 (for example, 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. Before chromatin studies, all efforts to understand mechanisms focused on cell culture, even though what happens in cultured cells—even cultured neurons—is not always an accurate reflection of what happens in the fully differentiated adult brain. Analogous to studies in 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 or 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 mechanisms[3–19] 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 (**Table 1**).

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 approaches[20]. However, application of such methods to the brain, given its distinctive cellular heterogeneity and the need to focus on *in vivo* models, involves many specialized 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 benefits 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-seq[21–23] has proven to be a more powerful tool for assessing transcriptional outputs for a number of reasons. (i) Whereas microarray technology limits researchers to detecting and analyzing transcripts that correspond to existing genomic sequence information, RNA-seq allows

[1]Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [2]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [4]Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [5]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [6]Laboratory of Chromatin Biology and Epigenetics, The Rockefeller University, New York, New York, USA. Correspondence should be addressed to E.J.N. (eric.nestler@mssm.edu).

Received 28 April; accepted 20 August; published online 28 October 2014; doi:10.1038/nn.3816

Table 1 Progress report of epigenomic data from brain

Year	Species/brain region(s) examined	Modification(s) and DNA binding protein(s) examined	Platform(s)	Variable(s)	Key finding(s)	Ref.
2009	Mouse/embryonic forebrain and midbrain	p300	ChIP-seq	Basal state	Genome-wide map of p300 identifying tissue-specific enhancers	3
2010	Human/prefrontal cortex, neurons versus non-neuronal cells	H3K4me3	ChIP-seq	Age	Age-correlated reorganization of H3K4me3: cell type– and subject-specific regulation	4
2010	Mouse/adult hippocampus	H4K12ac	ChIP-seq, microarray	Fear conditioned learning	Dysregulated H4K12ac and gene expression in aging	5
2011	Mouse/adult hippocampal dentate granule cells	5mC and 5hmC	MethylC-seq, BS-seq/microarray	Electroconvulsive stimulation	Genome-wide, single-base-resolution maps of 5mC and 5hmC	6
2011	Mouse/adult nucleus accumbens	H3K9me3	ChIP-seq	Chronic cocaine	Reduced H3K9me3 at heterochromatic loci and induction of retrotransposable elements after chronic cocaine	7
2011	Mouse/early postnatal and adult hippocampus and cerebellum Human/adult cerebellum	5hmC	Chemical labeling and immunoprecipitation, ChIP-seq	Age, *Mecp2* overexpression and knockout	Genome-wide maps of 5hmC during development and aging, including mouse models of Rett syndrome	8
2011	Human/adult hippocampus	H3K4me3	ChIP-seq, RNA-seq	Cocaine and alcohol addiction	Transcriptional and chromatin changes after cocaine or alcohol exposure	9
2012	Rat/adult hippocampus	H3K9me3	ChIP-seq	Acute restraint stress	Increased H3K9me3 at heterochromatic loci and repression of retrotransposable elements after acute stress	10
2012	Mouse/adult cerebellar Purkinje, granule and Bergmann glial cells	5mC, 5hmC and non-CpG methylation	MeDIP-seq, TRAP-seq	Cell type, *Mecp2* knockout	Cell type–specific relationships between 5hmC, 5mC and gene expression, and evidence that MeCP2 is the main 5hmC-binding protein in brain	11
2012	Human, chimpanzee and macaque/adult prefrontal cortex, neurons versus non-neuronal cells	H3K4me3	ChIP-seq	Species	Insights into human-specific modifications of the neuronal epigenome, with evidence for coordinated regulation across distant sites	12
2012	Mouse/adult nucleus accumbens	H3K9me2	ChIP-seq, RNA-seq	Morphine	Genome-wide map of H3K9me2 and identification of regulated target genes after chronic morphine	13
2013	Human and mouse/frontal cortex, neurons versus non-neuronal cells	5mC and 5hmC	MethylC-seq, RNA-seq	Cell type, age	Genome-wide single-base resolution maps of 5mC throughout the lifespan, showing increased non-CpG methylation during development	14
2013	Mouse/adult hippocampus	H4K5ac	ChIP-seq, microarray	Fear conditioned learning	Insights into mechanisms of gene priming and 'bookmarking' by histone acetylation during memory activation	15
2013	Human/adult	H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K9ac and H3K27ac	ChIP-seq, microarray	Cell and tissue types	Global chromatin state transitions accompanying cell specification during development, as well as age-related changes	16
2014	Mouse/adult nucleus accumbens	H3K4me3, H3K4me1, H3K27me3, H3K9me2, H3K9me3, H3K36me3 and RNA Pol II	ChIP-seq, RNA-seq	Chronic cocaine	Identification of combinations of chromatin changes (signatures) that predict regulation of pre-mRNA splicing by chronic cocaine	17
2014	Mouse/adult hippocampal dentate granule cells Human/brain	5mC (CpG) and non-CpG methylation (CpH)	BS-seq, ChIP-seq, RNA-seq	Age, triple *Dnmt1 Dnmt3a Dnmt3b* knockout	Genome-wide, single-base-resolution maps of the neuronal DNA methylome, identifying high levels of both CpG and CpH methylation	18
2014	Mouse/adult nucleus accumbens	PARP-1 and H3K4me3	ChIP-seq, RNA-seq	Chronic cocaine	Genome-wide map of PARP1 and identification of regulated target genes after chronic cocaine	19

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

studies of both known and new transcripts, an approach that is ideal for discovery-based experiments. (ii) Since RNA-seq allows unambiguous mapping of obtained DNA sequences to unique regions of the genome, as opposed to cross-hybridization procedures inherent in microarray technologies, signal-to-noise ratios are substantially improved. (iii) RNA-seq has a finer granularity of expression measurements, thereby allowing assessments of a large dynamic range of expression levels[24–26]. 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 tissue[20]. 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 (**Fig. 1**) is—for each biological

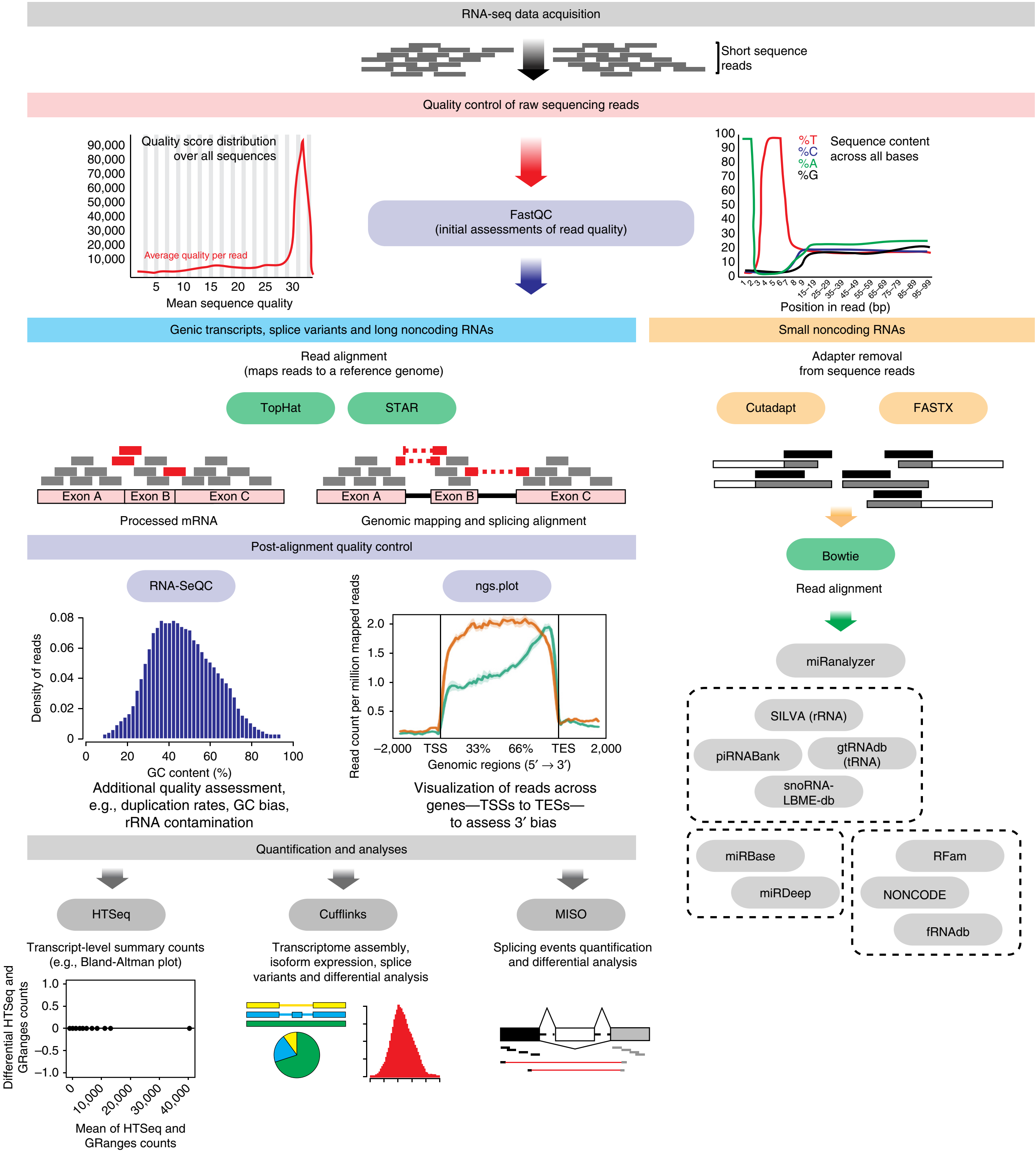
RNA-seq data acquisition
Short sequence reads
Quality control of raw sequencing reads
Quality score distribution over all sequences
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
Average quality per read
5
10
15
20
25
30
Mean sequence quality
FastQC
(initial assessments of read quality)
%T
%C
%A
%G
Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
Position in read (bp)
Genic transcripts, splice variants and long noncoding RNAs
Small noncoding RNAs
Read alignment
(maps reads to a reference genome)
TopHat
STAR
Exon A
Exon B
Exon C
Processed mRNA
Genomic mapping and splicing alignment
Adapter removal
from sequence reads
Cutadapt
FASTX
Bowtie
Read alignment
miRanalyzer
SILVA (rRNA)
piRNABank
gtRNAdb (tRNA)
snoRNA-LBME-db
miRBase
miRDeep
RFam
NONCODE
fRNAdb
Post-alignment quality control
RNA-SeQC
ngs.plot
Density of reads
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
GC content (%)
Additional quality assessment, e.g., duplication rates, GC bias, rRNA contamination
Read count per million mapped reads
2.0
1.5
1.0
0.5
–2,000
TSS
33%
66%
TES
2,000
Genomic regions (5′ → 3′)
Visualization of reads across genes—TSSs to TESs—to assess 3′ bias
Quantification and analyses
HTSeq
Cufflinks
MISO
Transcript-level summary counts (e.g., Bland-Altman plot)
Transcriptome assembly, isoform expression, splice variants and differential analysis
Splicing events quantification and differential analysis
Differential HTSeq and GRanges counts
1.0
0.5
0
–0.5
–1.0
0
10,000
20,000
30,000
40,000
Mean of HTSeq and GRanges counts

Figure 1 Initial pipelines of RNA-seq data analysis. Following data acquisition, RNA-seq analyses typically begin with quality control assessments using analytical tools such as FastQC. Next, for analysis of genic transcripts, splice variants and lncRNAs, short sequencing reads can be aligned to a reference genome using programs such as TopHat or STAR. After 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 from ref. 28. 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, after 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 miRanalyzer[139]. Additional ncRNA-specific analyses can similarly be integrated into miRanalyzer's pipeline, or independent databases (for example, SILVA, piRNABank, etc.) can be used.

sample—tens to hundreds of millions of 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 program (that is, it requires relatively little memory and yields outputs that are not excessive) that simply requires raw sequencing reads in FASTQ format for initial quality control assessments. FastQC, however, does not address RNA-seq-specific questions, such as exonic versus intronic alignments and transcript detection rates, making quality control determinations for subsequent downstream analyses difficult. To address these, RNA-SeQC[27] 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 ribosomal RNA content, regions of alignment (exon versus intron versus intragenic), continuity of coverage, 3′/5′ biases and counts of detectable transcripts, among others. Although RNA-SeQC is a comprehensive program that addresses most quality control issues, its usage can be cumbersome, requiring a great deal of computational power. In addition, ngs.plot[28] 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 the identification of common sequencing abnormalities, such as strong 3′ biases. Although underreported, such quality control 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 challenges specific to working with neural tissues (**Box 1**). Since low-quality sequencing results in decreased signal-to-noise ratio, as well as potential biases, such small differences may be inadvertently masked or amplified, thereby leading to high false positive and negative rates.

After initial quality control 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 (for example, 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 and release number (for example, Ensembl 67). When working with a species for which a comprehensive reference genome is not available, *de novo* transcriptome assembly can be performed using tools such as Cufflinks[29] and Trinity[30]. A popular choice for RNA-seq data alignment is TopHat[31], which wraps around Bowtie[32] as the basic alignment tool but provides additional workflows to accomplish tasks specific to RNA-seq data analysis (for example, 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, STAR[33], has been gaining market share and attention. In a recent study comparing 11 programs for RNA-seq alignment[34], 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 ultrafast, sensitive and precise tool that reduces alignment time from 1–2 d to a few hours for RNA-seq samples containing ~100 million reads, based on four-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.

Box 1 Challenges specific to neuroepigenomics

Epigenomic studies of the brain are more challenging as a result of 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 mm^3 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 many 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 (for example, 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 tags[140]—represents one promising approach for the future. However, it is not yet feasible to obtain from smaller brain regions the numbers of isolated cells or nuclei that are required for most epigenomic analyses. Refinement of techniques that make ChIP-seq and related procedures possible with a far smaller numbers of cells or nuclei will advance the field dramatically[4,141,142].

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 twofold cutoff, whereas studies of brain reveal few changes of this magnitude. This is not solely due to the confounding effect of multiple cell types, because similar findings have been obtained from analysis of single cell types. Rather, it is likely that fully differentiated neurons in 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 (for example, 20%)—which are demonstrably functionally important[1,2,15]—adds another special burden on bioinformatics analysis of neuroepigenomic data.

Analysis of transcripts, splice variants and noncoding 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 twofold induction of that gene's expression. Such analyses are rarely so simple, however, 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. Cufflinks[29] solves this problem through the use of a statistical learning approach that assigns a fraction of the total read count to each transcript on the basis of 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: (i) alternative splicing, which is defined between different transcripts that are in the same TSS group; and (ii) 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. To obtain more detailed splicing information, the Mixture of Isoforms (MISO) model[35] 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 the Bayesian theorem to iteratively infer splicing ratios, both among isoforms of the same gene and between two conditions for the same isoform, thereby allowing researchers to derive relative abundances of spliced transcripts. Such detailed information offers an advantage when performing integrative analyses between alternative splicing and other forms of epigenomic regulation, such as the contributions 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 mechanisms[36]. 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 technologies[17].

RNA-seq also enables the detection and quantification of several types of noncoding RNAs, which are proving crucial in biological regulation. A subset of long noncoding RNAs (lncRNAs)—as in protein-coding genes—contain polyadenylated (poly(A)) tails, which allow their detection by RNA-seq regardless of the RNA purification procedure used (for example, ribozero—an extraction protocol for isolating total RNA with removal of cytoplasmic rRNA—or poly(A) selection). Identification of non-poly(A) lncRNAs, however, can only be accomplished through purification procedures preserving total (that is, poly(A)$^+$ and poly(A)$^-$) RNA. Since lncRNAs exist in high abundance in mammals, the Ensembl database has been incorporating many lncRNAs into its gene collection. Thus, predefined lncRNAs can now be analyzed alongside protein-coding genes in the same sample. If investigators wish to predict a large number of novel, and still unannotated, lncRNAs, however, this can be accomplished by using 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 BEDTools[37] to estimate lncRNA abundance.

MicroRNA sequencing experiments, which must be run separately from standard RNA-seq experiments, investigate mature miRNAs that are around 22 nucleotides in length. Since most high-throughput sequencing machines produce reads that are significantly longer than mature miRNAs, the short reads obtained from miRNA sequencing often contain portions 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 program[38] can be used for these purposes. Cutadapt is useful in that, in comparison to FASTX, it supports a larger range of sequencing platforms, including color-space reads, which allow base pairs to be encoded in color to reduce sequencing error rates (https://www.biostars.org/p/43855/). Following adapter removal, short reads can be mapped to the reference genome similarly to data obtained from long RNA sequencing. miRBase[39] provides a comprehensive collection of miRNA sequences. BEDTools can then be used to extract read counts for each of the annotated miRNAs obtained from alignment. If *de novo* prediction of miRNAs from a given sample is desired, the miRDeep[40] program can be used. Many more families of small RNAs (**Table 2**) can be identified and analyzed with similar approaches[39,41–48].

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. Many differential analysis tools have been developed in recent years. Two popular choices are DESeq[49] and edgeR[50], with both methods based on negative binomial testing, which provides 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 data sets, as sample sizes from animal models are typically small, with an $N = 3$ (three biological replicates) used per condition in most experiments. Therefore, many 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 voom[51] has been developed. It does not depend on negative binomial testing

Table 2 Existing databases for analysis of ncRNAs

ncRNAs	Database	Description	Ref.
General purpose	Rfam (http://rfam.xfam.org/)	Sister database of Pfam[a], collections of ncRNA families	46
	NONCODE (http://www.noncode.org/)	Database of ncRNA, excluding rRNA and tRNA	48
	fRNAdb (http://www.ncrna.org/frnadb/)	Database of functional ncRNAs	44
miRNA	miRBase (http://www.mirbase.org/)	Database for miRNAs	39
piRNA	piRNABank (http://pirnabank.ibab.ac.in/)	Database for piRNAs; clustering information provided	42
snoRNA	snoRNABase (https://www-snorna.biotoul.fr/)	Comprehensive database for human snoRNAs	41
lncRNA	lncRNAdb (http://www.lncrnadb.org/)	Database for lncRNAs	45
rRNA	SILVA (http://www.arb-silva.de/)	Curated database for rRNAs	47
tRNA	GtRNAdb (http://gtrnadb.ucsc.edu/)	Genome-wide tRNA database predicted by the program tRNAscan-SE	43

This list is not comprehensive but aims to provide an overview of available databases for analyzing ncRNAs from biological samples. Recently identified types of ncRNA, such as circular RNAs, which may have critical functions in the CNS, are not included, as analysis tools do not yet exist. piRNA, piwi-interacting RNA.

[a] PFam: a database widely used in the analysis and annotation of protein sequences that uses hidden Markov model–based multiple sequence alignments to provide detailed information about protein families and domains.

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 limma[52] 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 rates[53]. 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 larger numbers of biological replicates are needed to derive the most reliable data[54,55].

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 many 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 Cuffdiff 2 (ref. 56), but a recent comparison pointed out that Cuffdiff 2 may be too conservative[57]. 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; with a standard four-core workstation, it would take 1–2 d to complete analysis for a 20-sample (10 versus 10) data set and may crash for data sets with >80 samples. Using HTSeq for gene counts in connection with limma or DESeq greatly reduces the analysis time, to <1 h.

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 or thousands of samples (for example, from studies of clinical populations), as well as dozens of comparisons, to run and examine. Therefore, a new Python-based computational pipeline, SPEctRA (Scalable Pipeline for RNA-seq Analysis: https://github.com/shenlab-sinai/SPEctRA/), has been developed to perform all these tasks with a single command, accepting various parameter settings as a configuration file. This pipeline is designed to function in different computational environments. For example, when performing analyses under a portable batch system 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 that will allow it to run in the Amazon cloud, where both CPU power and memory can be configured on demand for each run, is 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 (for example, a modified histone or transcription factor)[20]. Quality control and sequence alignment are the first steps in analyzing such ChIP-seq data. Any of several short read aligners, such as Bowtie[32] or BWA[58], can be used to align reads to a reference genome and assess the enrichment of the DNA-binding protein of interest. Sequencing quality can then be assessed using FastQC, as described above. Alignments are next exported as BAM[59] files for further analysis. If ChIP-seq samples are generated from sonication-based fragmentation methods, then duplicated alignments need to be removed, a procedure easily accomplished with SAMtools[59] 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 so-called native ChIP, preferentially digests genomic sequences containing specific nucleotide sequences (that is, the rate of MNase cleavage 5′ of A or T nucleotides is 30 times greater than the rate of cleavage 5′ of G or C nucleotides), the probability of obtaining fragments from the same location or strand is increased. Therefore, removal of these reads when analyzing native ChIP samples, unless excessively high (thresholds can be determined using DANPOS[60], 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 bioinformatics innovations. The number of unique reads per sample is an important criterion for the quality of ChIP-seq data. Since antibodies are key to success in any ChIP-seq experiment[20], it is vitally important to determine the efficiency and specificity of the antibody being used. For this purpose, the phantomPeak[61] tool can be used, as suggested by the Encyclopedia of DNA Elements (ENCODE) Consortium[62], 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 browser[63], 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/)[28] allows inspection of both average and 'laid out' coverages as curves or heat maps, 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 many genomes, can accommodate large alignment files with relatively small memory footprints. A ChIP-seq quality assessment pipeline has been developed (https://github.com/ny-shao/chip-seq_preprocess/) to perform all the above steps with a single command. This pipeline uses a similar design to that of the RNA-seq pipeline described above. When dealing with large sample numbers and multiple comparisons, this pipeline saves substantial amounts of time and effort.

Another important quality control consideration during the initial analysis and alignment of ChIP-seq data requires an in-depth understanding of the differences between "uniquely mapped" and "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 nonrepetitive 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 (that is, the nonredundant fraction) 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 overamplification during library processing, most likely as a result of 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, let us imagine that we are examining a small genome of 1,000 bp, where the read length obtained is 100 bp. If our goal is to evaluate the overall enrichment of a DNA-binding protein in this genome, then the total number of unique binding locations is (1,000 – 100 + 1) × 2 = 901 × 2 = 1,802, considering both strands. If we were to have a library size of one million reads with no PCR duplicates, then the number of unique locations would be saturated, thereby preventing the investigator from distinguishing differential binding strengths to these regions. If only one alignment at each unique location were to be preserved, then the dynamic range would be 0 to 1,802, with all remaining reads being thrown out. However, if two PCR duplicates were to be kept, then the dynamic range of enrichment would double (that is, it would be 0 to 3,604), and so on. 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. Dozens of peak calling tools exist, and these methods can generally be separated into two categories: sonication- and MNase 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 (for example, H3K4me3). However, if peaks are broad and diffuse (for example, 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, DANPOS[60], which detects not only basal enrichment but also various nucleosomal events (for example, nucleosomal positioning and occupancy and 'fuzziness' between the two), can be used. After peaks are detected, it is useful to determine their type of location within the genome, such as genes, gene deserts or pericentromeres. A regional analysis tool, which is part of the diffReps[68] package—more recently extended as a standalone program (https://github.com/shenlab-sinai/region_analysis/)—has been developed to perform this task. This program features single-command use and can assign genomic regions to one of eight distinct categories: proximal promoter (within 250 bp of a TSS), promoter 250 bp to 1 kb upstream of a TSS, promoter 1 to 3 kb upstream of a TSS, 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 ChIPpeakAnno[69] package. After peaks are annotated, enrichment of specific chromatin marks is easily visualized for their genomic distribution using pie charts 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. One differential analysis tool[70] 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 (sequencing library preparation), can produce very different signal-to-noise ratios between samples, which confounds peak calling. Insufficient sequencing depth (that is, 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 a 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 data set. Biological replicates are thus essential in neuroepigenomics research to enhance statistical power and increase precision. However, most peak calling methods have not been 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. diffReps[68], which was developed to address this need, is a Perl-based program that features single-command use 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 predefined cutoffs, merges replicates, performs multiple-testing corrections and reports results. Using benchmark standards, diffReps is highly sensitive and efficiently controls for false positives (**Fig. 2**). It is also important to assess the reproducibility among biological replicates as an additional quality control measurement during data processing. To do so, one may use programs such as corrgram[71], which generates Pearson's correlation coefficients between ChIP-seq signals derived from multiple signals. Alternatively, irreproducible discovery rates (IDRs) can be derived after peak calling to determine the number of enriched regions observed between replicates, as described[72].

Important questions for ChIP-seq experiments are the read depth required to appropriately make assumptions on the basis of the aforementioned analyses and to determine whether experimental sequencing depths are sufficient for the questions being asked. In general, researchers may follow the guidelines established by the ENCODE Consortium[72], which indicate a minimum of 10 or 20 million uniquely mapped reads for factors displaying punctuated or broad peaks, respectively. One may 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 versus 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 because 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 ChromHMM[73], 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 or binding factors in biological samples. Such analyses provide a new type of computational tool for 'learning' chromatin states, characterizing biological functions and achieving genome-wide visualizations of such annotations. Although many methods now exist for analyzing ChIP-seq data, independent biological validation remains an essential step for confirming the accuracy of any data derived from genome-wide approaches.

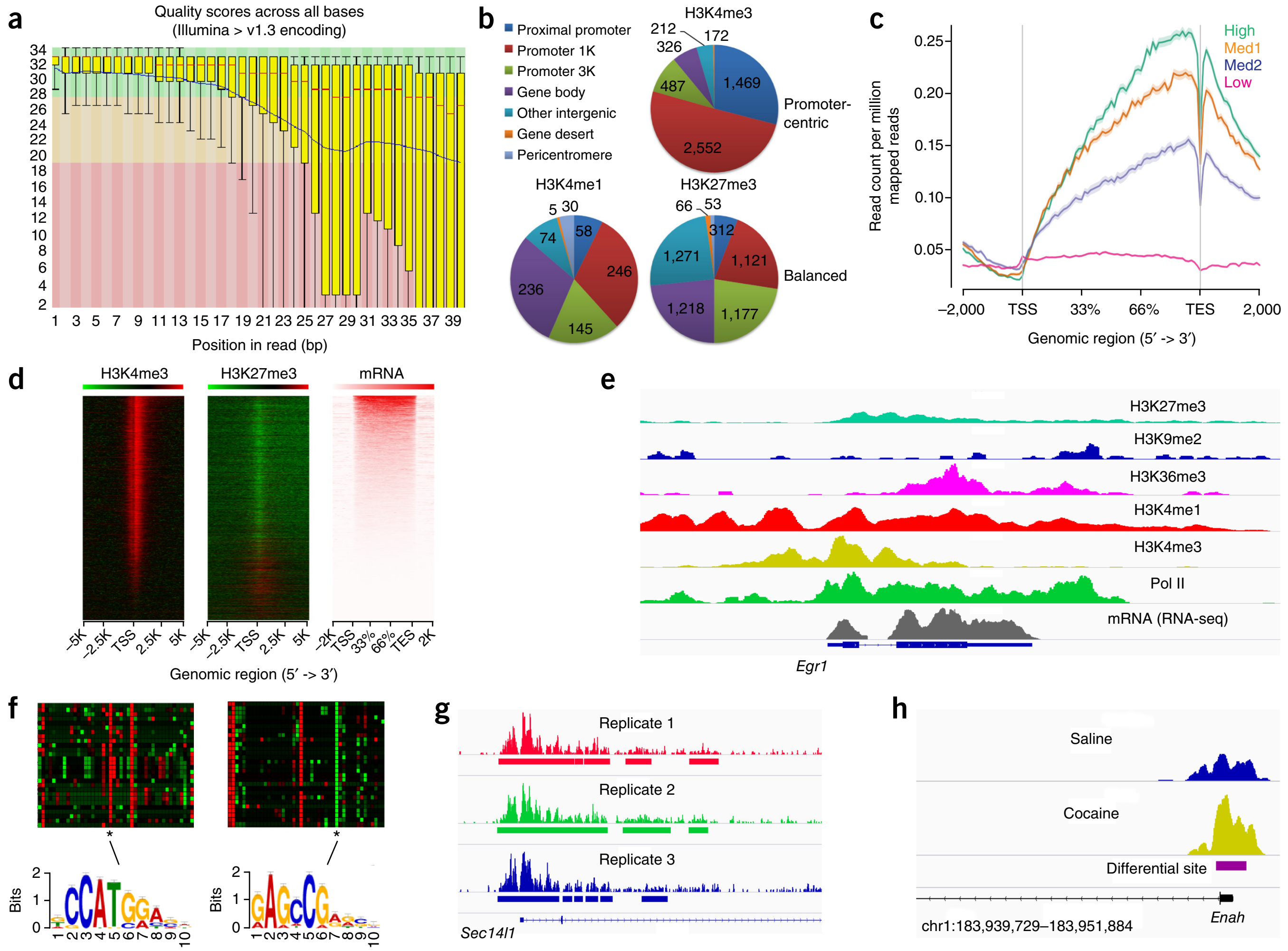

Figure 2 ChIP-seq analysis of brain. (**a**) Assessment of ChIP-seq sample quality, represented here as the sequence quality score versus base pair position using FastQC, is a critical first step in ChIP-seq data analysis and is required for appropriate interpretations of subsequent downstream analyses. 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- versus chronic cocaine–treated mice[16]. 1K, 0.25–1 kb upstream of a TSS; 3K, 1–3 kb upstream of a TSS. Reprinted with permission from ref. 17. (**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 $\log_2$ enrichment heat maps generated in ngs.plot for several histone marks versus DNA input in mouse NAc at TSS ± 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, as well as RNA-seq data from poly(A)-selected RNA. The genomic region displayed represents the TSS and ~20 kb downstream of the *Egr1* gene. (**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 (for example, 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. Although the three biological replicates appear to be generally similar in size and distribution, MACS identified discordant peaks among the samples owing to intrinsic variations in peak location. Unlike MACS, diffReps uses a sliding window approach that allows one to focus on a region of a fixed size across all samples (for example, 1 kb), thus allowing 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- versus cocaine-treated mice at the TSS of the *Enah* gene.

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 data sets. In fact, if one analyzes enough epigenomic endpoints (the many 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 (for example, RNA polymerase II (Pol 2) Ser5 phosphorylation), in response to a specific stimulus without increasing the expression of its cognate gene, a phenomenon thought to indicate gene 'poising'. And in much fewer examples, H3K4me3 has been linked to gene repression when it associates with ING2 (inhibitor of growth family-2), a chromatin regulatory protein[74]. Conversely, the histone mark H3K9me3, which typically enriches within the gene bodies of silenced genes along with corepressor proteins, may exhibit increased enrichment with distinct binding partners to promote alternative splicing[75].

Many 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 in which the relative enrichment of histone marks was determined. In this case, enrichment values served as observed data and corresponding gene expression levels served 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 accumbens[17], 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, new research is revealing many 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 modifications[50]. 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 $\log_2$ fold changes between repeated cocaine and saline, loading the values into a data matrix in which each row represents a transcript and each column represents a histone mark–gene region combination (for example, a H3K4me3 peak at the proximal promoter of a given 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 or 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 cocaine[17]. These findings illustrate how the overlay of ChIP-seq and RNA-seq data can yield 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 activation[78]. Despite the implication of DNA methylation in numerous neuropsychiatric phenomena, virtually all studies so far have focused on individual candidate genes, with genome-wide explorations of brain sorely lacking.

Several methods, described in the companion review[20], are used to obtain a genome-wide map of 5mC and 5hmC, but doing so at single nucleotide resolution remains a challenge owing 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 off sequences representing the adapters 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 quality control 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 aligner strategies are available. One is a wild-card aligner, such as BSMAP (https://code.google.com/p/bsmap/)[79], in which all instances of C in a reference genome are replaced by a wild-card letter (for example, 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 instances of C in reads and a reference genome are converted to T, whereas instances of G are converted to A, 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 as compared with three-letter aligners[81]. 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 versus control conditions. Basic analyses such as *t*-tests or Wilcoxon rank tests can be applied; details about more advanced methods have been reviewed recently[81]. oxbs-sequencing-qc (https://code.google.com/p/oxbs-sequencing-qc/) is a pipeline based on Bismark. It includes trimming of adapters and low-quality portions, 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 differentially 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 quality control and then aligned to a reference genome using the various tools outlined above, with the number of aligned reads used for methylation calling[83]. In addition to canonical CpG methylation, highly conserved non-CpG methylation (mCH, where H is either A, T or C) may also be important in the CNS. It was recently reported that non-CpG methylation accumulates in neurons, but not glia, of the cerebral cortex during development[14].

Reconstruction of multiscale biological networks

A high priority 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 and 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 to most of these illnesses[84]. Also, gene sequence analysis alone is insufficient to uncover causal mechanisms regarding the gene or 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, are 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 glioblastoma[85]. 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 cerebellum[86]. For such large-scale molecular data sets, 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 coexpression network analysis (WGCNA)[87], aim to capture total interactions among genes in a more comprehensive manner than traditional unweighted approaches and have been used to identify coexpressed gene modules and key genes associated with glioblastoma[88], 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 coexpression analysis (DCA) approach to integrate differential expression and differential correlation to identify *APOE* ε4 effectors[94]. 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 LOAD[95]. 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 comprised highly interacting genes through WGCNA are rank ordered on the basis of 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 and quantitative trait loci. Bayesian networks are then used to identify key causal regulators on the basis of network connectivity.

MNA is in line with the recent finding that no single network inference method is universally best across data sets and the best strategy is to integrate networks constructed from different approaches[96]. MNA is a natural framework to integrate genomic, epigenomic and clinical data. Gene expression, copy number variations 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 copy number variants and expression-associated DNA methylation sites can be taken as priors for causal network inference[97]. 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 (*in cis*) or at distant interacting genomic regions (*in 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

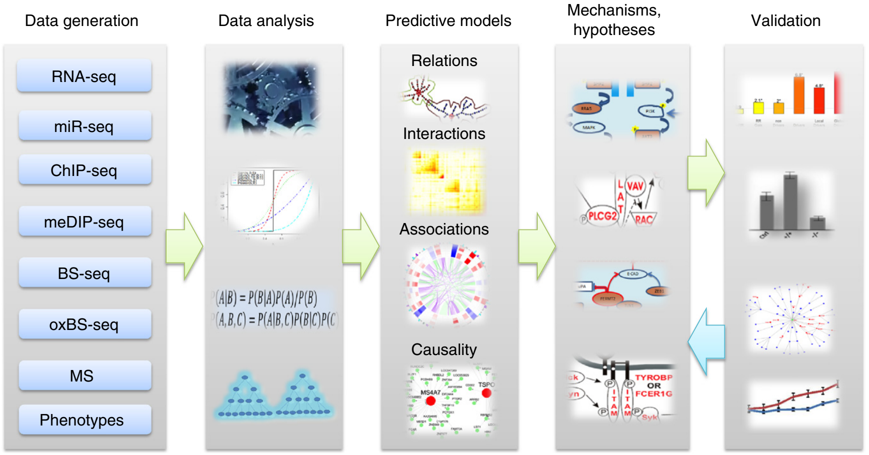

Figure 3 Network inference approaches in neuroscience. Systems/network approaches to integrating large-scale genomic, epigenomic and clinical data to inform investigations of neurological and psychiatric disorders using data derived from humans and animal models. Correlational, interaction and causal analyses can be unified under network-level frameworks to generate data-driven hypotheses (that is, models) concerning the underlying molecular mechanisms or biomarkers resulting in neuronal dysfunction. These models are then further validated experimentally, generating data that can be fed back into modeling processes to refine biological predictions.

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 (**Fig. 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 (that is, 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 in which phenotypic changes are examined and compared with predicted outcomes[88,94,95,97,98]. Full validation of network-based mechanisms for complex diseases, however, is still prohibitive owing 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 derived from reprogrammed somatic cells. Such advances will likely pave the 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 modifications[100], with some newer applications to small ncRNAs[101]. However, much of this has been at the global chromatin level and not at specific genes[102]. Identification and quantification of histone post-translational modifications (PTMs) from a variety of cells and tissues have been the focus of several studies. They typically involve examining the protein by bottom-up, middle-down or top-down MS[102].

Bottom-up MS refers to digestion of the protein into small (<3 kDa) peptides and analysis of those peptide typically by nano-scale liquid chromatography followed by tandem MS (MS/MS). These experiments are fairly high throughput and easier to perform than middle-down or top-down MS. The bottom-up MS approach has been coupled with diverse platforms and quantification schemes (for example, 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 mice[104]. This large-scale proteomics analysis identified more than 10,000 peptides, creating a data set containing 146 modification sites on 1,475 peptides in various combinatorial patterns on short peptides. Among these histone peptides, 58 new sites of modification were discovered. Bottom-up MS in fact has led the way in identifying new marks on histones, such as crotonylation, succinylation, malonylation and 2-hydroxyisobutyrylation of lysine residues[105–107] and serine/threonine O-acetylation[108].

Middle-down MS involves interrogation of polypeptides over the 3 kDa molecular weight range. These MS experiments allow long-range spanning PTMs to be identified on the same peptide to determine a long-distance combinatorial code. They have mainly been focused on the N-terminal tails of the histone proteins. This approach has been used to determine that there are >200 combinatorially modified N-terminal forms of histone H3, with about half as many N-terminal forms of histone H4 (ref. 109). 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 Sprague–Dawley rats 7–8 weeks old[110]. Noteworthy patterns of modifications were found, including a greater prevalence of silencing modifications such as H3K9me3 in the cerebellum.

Analysis of intact histone proteins (top-down MS) is a challenging task but allows 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 histones[111] 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 owing to the many different types of modifications that can be detected, the distinct combinations of these PTMs that can be found and the low stoichiometry levels of many histone PTMs, which can often give rise to false positive assignments. There are many computational approaches for both the qualitative and quantitative analysis of histone proteomics data sets. 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 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 algorithms to deal with these data sets have been developed in an academic setting (for example, MILP or PTMap)[112,113]. However, several other database searching engines are commercially available for analyzing histone proteomics data, with the most common being Mascot[114] and SEQUEST[115] but with many others emerging with different specificities or advantages[116–118]. 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 fewer spectra. Therefore, it has been shown with histone data that it is best to search for classes of modifications one at a time to reduce false positive misassignments. Equally important to these analyses for histone PTM proteomics data sets is the use of a site localization scoring algorithm[119]. As histone PTMs occur on peptides with multiple acceptor amino acids (for example, 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 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- and epigenomics-level information with proteomics data is that an unbiased assessment of histone PTMs can be performed and mapped back to genomic loci. Several approaches based on tagging of proteins or using antibodies to isolate nucleosomes have been performed[120–122] 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 proteomics[122]. A ChIP-seq and proteomics study also found that Brd (bromodomain) proteins are found on active *HOX* genes in cultured human embryonic kidney cells, with distinct patterns of histone H4 acetylation[123]. However, both of these studies mapped histone PTMs, not to a specific locus, but rather over a distribution of enriched genes. Experiments that isolate specific DNA 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 repeats[124]. Other approaches such as chromatin affinity purification (ChAP)-MS and insertional chromatin immunoprecipitation (iChIP) have also been shown to be very useful for isolation of genomic material for MS interrogation[125,126]; 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 primates[127], a significant portion of hominid evolution could be due to DNA sequence changes involving regulatory and noncoding DNA[128]. 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 base pair and 5 × 10^6 multi-base-pair substitutions and insertion/deletion events[129], with the majority of DNA in the human genome encoding functionally active transcripts[130].

Thus, given that transcription is the first step connecting genetic information to phenotype, the initial studies on deep sequencing of human and nonhuman 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 nonrepetitive intergenic regions[131], 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 patterns[132]. Deep sequencing of human and nonhuman 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 Neanderthals[133]. 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 neurons[12].

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.

Box 2 Data deposition

The sequence read archive (SRA)[143] was established as an international effort 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: NCBI, http://www.ncbi.nlm.nih.gov/Traces/sra/; EBI, http://www.ebi.ac.uk/ena/; DDBJ, http://trace.ddbj.nig.ac.jp/dra/index_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, after which the submitter can log in 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 data sets 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.

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 genome[144]. 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[145,146]. Human HiC studies cluster looping interactions into three groups: self-interacting active gene–rich clusters, nonactive centromere-proximal clusters and nonactive centromere-distal clusters[147]. 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 cleaved using a restriction 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 paired-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)[147]. 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 sequencing[148]. 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 them[147]. In time, the efficiency of algorithms will further increase and the cost of sequencing will decrease, making HiC a more attractive approach for cell type–specific studies in brain.

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[4,134]. 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 data sets that provide not simply a single cross-section of RNA expression levels, but 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 data sets to derive maximal insight into transcriptional control in health and disease. A crucial step is to ensure that all genome-wide data are placed in the public domain in ways that allow raw data sets 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 single nucleotide polymorphisms, positioned mostly in intergenic and (protein)-noncoding sequences, contribute to the genetic risk of schizophrenia[135]. Strikingly, however, there is 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 project[130]. This contrasts with the robust enrichment for DNase I hypersensitive sites of the general GWAS catalog of the National Genome Research Institute[136]. 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 cultures[137] or cerebral organoids[138], 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 US 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 for 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 to be 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.

ACKNOWLEDGMENTS
We thank A. Soshnev for help with illustrations. This work was supported by grants from the US National Institute of Mental Health (5R01 MH094698 and P50 MH096890), US National Institute on Drug Abuse (P01 DA008227) and the Hope for Depression Research Foundation (HDRF).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Renthal, W. *et al.* Delta FosB mediates epigenetic desensitization of the c-fos gene after chronic amphetamine exposure. *J. Neurosci.* **28**, 7344–7349 (2008).
2. Maze, I. *et al.* Essential role of the histone methyltransferase G9a in cocaine-induced plasticity. *Science* **327**, 213–216 (2010).
3. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
4. Cheung, I. *et al.* Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. *Proc. Natl. Acad. Sci. USA* **107**, 8824–8829 (2010).
5. Peleg, S. *et al.* Altered histone acetylation is associated with age-dependent memory impairment in mice. *Science* **328**, 753–756 (2010).
6. Guo, J.U. *et al.* Neuronal activity modifies the DNA methylation landscape in the adult brain. *Nat. Neurosci.* **14**, 1345–1351 (2011).
7. Maze, I. *et al.* Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens. *Proc. Natl. Acad. Sci. USA* **108**, 3035–3040 (2011).
8. Szulwach, K.E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.* **14**, 1607–1616 (2011).
9. Zhou, Z. *et al.* Substance-specific and shared transcription and epigenetic changes in the human hippocampus chronically exposed to cocaine and alcohol. *Proc. Natl. Acad. Sci. USA* **108**, 6626–6631 (2011).
10. Hunter, R.G. *et al.* Acute stress and hippocampal histone H3 lysine 9 trimethylation, a retrotransposon silencing response. *Proc. Natl. Acad. Sci. USA* **109**, 17657–17662 (2012).
11. Mellén, M. *et al.* MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417–1430 (2012).

12. Shulha, H.P. *et al.* Human-specific histone methylation signatures at transcription start sites in prefrontal neurons. *PLoS Biol.* **10**, e1001427 (2012).
13. Sun, H. & Maze, I. *et al.* Morphine epigenomically regulates behavior through alterations in histone H3 lysine 9 dimethylation in the nucleus accumbens. *J. Neurosci.* **32**, 17454–17464 (2012).
14. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
15. Park, C.S. *et al.* Genome-wide analysis of H4K5 acetylation associated with fear memory in mice. *BMC Genomics* **14**, 539 (2013).
16. Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642–654 (2013).
17. Feng, J. *et al.* Chronic cocaine-regulated epigenomic changes in mouse nucleus accumbens. *Genome Biol.* **15**, R65 (2014).
18. Guo, J.U. *et al.* Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17**, 215–222 (2014).
19. Scobie, K.N. *et al.* Essential role of poly(ADP-ribosyl)ation in cocaine action. *Proc. Natl. Acad. Sci. USA* **111**, 2005–2010 (2014).
20. Shin, J., Ming, G. & Song, H. Decoding neuronal transcriptomes and epigenomes: high-throughput sequencing for neuroscience. *Nat. Neurosci.* **17**, xxx–yyy (2014).
21. Mortazavi, A. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
22. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
23. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
24. Marioni, J.C. *et al.* RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
25. Nookaew, I. *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 Res.* **40**, 10084–10097 (2012).
26. Zhao, S. *et al.* Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE* **9**, e78644 (2014).
27. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
28. Shen, L. *et al.* ngs.plot: quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 284 (2014).
29. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
30. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
31. Trapnell, C. *et al.* TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
32. Langmead, B. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
34. Engström, P.G. *et al.* Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **10**, 1185–1191 (2013).
35. Katz, Y. *et al.* Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
36. Deal, R.B. & Henikoff, S. A simple method for gene expression and chromatin profiling of individual cell types within a tissue. *Dev. Cell* **18**, 1030–1040 (2010).
37. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
38. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
39. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68–D73 (2014).
40. Friedländer, M.R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407–415 (2008).
41. Lestrade, L. & Weber, M.J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.* **34**, D158–D162 (2006).
42. Sai Lakshmi, S. & Agrawal, S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.* **36**, D173–D177 (2008).
43. Chan, P.P. & Lowe, T.M. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.* **37**, D93–D97 (2009).
44. Mituyama, T. *et al.* The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.* **37**, D89–D92 (2009).
45. Amaral, P.P. *et al.* lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* **39**, D146–D151 (2011).
46. Burge, S.W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* **41**, D226–D232 (2013).
47. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
48. Xie, C. *et al.* NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.* **42**, D98–D103 (2014).
49. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
50. Robinson, M.D. *et al.* edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
51. Law, C.W. *et al.* voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
52. Smyth, G.K. limma: linear models for microarray data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S.) 397–420 (Springer, New York, 2005).
53. Love, M.I. *et al.* Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv beta* doi:10.1101/002832 (2014).
54. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
55. Liu, Y. *et al.* RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30**, 301–304 (2014).
56. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
57. Seyednasrollah, F. *et al.* Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* doi:10.1093/bib/bbt086 (2013).
58. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
59. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **23**, 341–351 (2013).
61. Kundaje, A. Phantompeakqualtools. https://code.google.com/p/phantompeakqualtools/ (2013).
62. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
63. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
64. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods* **9**, 609–614 (2012).
65. Carroll, T.S. *et al.* Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* **5**, 75 (2014).
66. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
67. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
68. Shen, L. *et al.* diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. *PLoS ONE* **8**, e65598 (2013).
69. Zhu, L.J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**, 237 (2010).
70. Liang, K. & Keleş, S. Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* **28**, 121–122 (2012).
71. Wright, K. corrgram: plot a correlogram. *R Package version 1.5* http://CRAN.R-project.org/package=corrgram (2013).
72. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
73. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
74. Shi, X. *et al.* ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature* **442**, 96–99 (2006).
75. Saint-André, V. *et al.* Histone H3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. *Nat. Struct. Mol. Biol.* **18**, 337–344 (2011).
76. Cheng, C. *et al.* A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol.* **12**, R15 (2011).
77. Dong, X. *et al.* Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.* **13**, R53 (2012).
78. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929–930 (2009).
79. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
80. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
81. Bock, C. Analysing and interpreting DNA methylation data. *Nat. Rev. Genet.* **13**, 705–719 (2012).
82. Sun, D. *et al.* MOABS: model based analysis of bisulfite sequencing data. *Genome Biol.* **15**, R38 (2014).
83. Down, T.A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* **26**, 779–785 (2008).
84. Goldstein, D.B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
85. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
86. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
87. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl. Genet. Mol. Biol.* **4**, 17 (2005).

88. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. USA* **103**, 17402–17407 (2006).
89. Miller, J.A. *et al.* A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J. Neurosci.* **28**, 1410–1420 (2008).
90. Miller, J.A. *et al.* Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proc. Natl. Acad. Sci. USA* **107**, 12698–12703 (2010).
91. Miller, J.A. *et al.* Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Med* **5**, 48 (2013).
92. Luo, R. *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.* **91**, 38–55 (2012).
93. Parikshak, N.N. *et al.* Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell* **155**, 1008–1021 (2013).
94. Rhinn, H. *et al.* Integrative genomics identifies APOE epsilon4 effectors in Alzheimer's disease. *Nature* **500**, 45–50 (2013).
95. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013).
96. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804 (2012).
97. Zhu, J. *et al.* Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40**, 854–861 (2008).
98. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
99. Swarup, V. & Geschwind, D.H. Alzheimer's disease: from big data to mechanism. *Nature* **500**, 34–35 (2013).
100. Evertts, A.G. *et al.* Modern approaches for investigating epigenetic signaling pathways. *J. Appl. Physiol. (1985)* **109**, 927–933 (2010).
101. Kullolli, M. *et al.* Intact microRNA analysis using high resolution mass spectrometry. *J. Am. Soc. Mass Spectrom.* **25**, 80–87 (2014).
102. Karch, K.R. *et al.* Identification and interrogation of combinatorial histone modifications. *Front. Genet.* **4**, 264 (2013).
103. Ong, S.E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
104. Tweedie-Cullen, R.Y. *et al.* Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain. *PLoS ONE* **7**, e36980 (2012).
105. Tan, M. *et al.* Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell* **146**, 1016–1028 (2011).
106. Xie, Z. *et al.* Lysine succinylation and lysine malonylation in histones. *Mol. Cell. Proteomics* **11**, 100–107 (2012).
107. Dai, L. *et al.* Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. *Nat. Chem. Biol.* **10**, 365–370 (2014).
108. Britton, L.M. *et al.* Initial characterization of histone H3 serine 10 O-acetylation. *Epigenetics* **8**, 1101–1113 (2013).
109. Young, N.L. *et al.* High throughput characterization of combinatorial histone codes. *Mol. Cell. Proteomics* **8**, 2266–2284 (2009).
110. Garcia, B.A. *et al.* Characterization of neurohistone variants and post-translational modifications by electron capture dissociation mass spectrometry. *Int. J. Mass Spectrom.* **259**, 184–196 (2007).
111. Tian, Z. *et al.* Enhanced top-down characterization of histone post-translational modifications. *Genome Biol.* **13**, R86 (2012).
112. Frank, A.M. *et al.* Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **80**, 2499–2505 (2008).
113. DiMaggio, P.A. Jr. *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* **8**, 2527–2543 (2009).
114. Perkins, D.N. *et al.* Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
115. Eng, J.K. *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.* **5**, 976–989 (1994).
116. Geer, L.Y. *et al.* Open mass spectrometry search algorithm. *J. Proteome Res.* **3**, 958–964 (2004).
117. Wang, L.H. *et al.* pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **21**, 2985–2991 (2007).
118. Zhang, J. & Xin, L. *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol. Cell Proteomics* **11**, M111 010587 (2012).
119. Beausoleil, S.A. *et al.* A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285–1292 (2006).
120. Tackett, A.J. *et al.* Proteomic and genomic characterization of chromatin complexes at a boundary. *J. Cell Biol.* **169**, 35–47 (2005).
121. Voigt, P. *et al.* Asymmetrically modified nucleosomes. *Cell* **151**, 181–193 (2012).
122. Wang, C.I. *et al.* Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in *Drosophila. Nat. Struct. Mol. Biol.* **20**, 202–209 (2013).
123. LeRoy, G. *et al.* Proteogenomic characterization and mapping of nucleosomes decoded by Brd and HP1 proteins. *Genome Biol.* **13**, R68 (2012).
124. Déjardin, J. & Kingston, R.E. Purification of proteins associated with specific genomic Loci. *Cell* **136**, 175–186 (2009).
125. Hoshino, A. & Fujii, H. Insertional chromatin immunoprecipitation: a method for isolating specific genomic regions. *J. Biosci. Bioeng.* **108**, 446–449 (2009).
126. Byrum, S.D. *et al.* Purification of a specific native genomic locus for proteomic analysis. *Nucleic Acids Res.* **41**, e195 (2013).
127. King, M.C. & Wilson, A.C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
128. McLean, C.Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011).
129. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
130. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
131. Xu, A.G. *et al.* Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput. Biol.* **6**, e1000843 (2010).
132. Li, Z. *et al.* Evolutionary and ontogenetic changes in RNA editing in human, chimpanzee, and macaque brains. *RNA* **19**, 1693–1702 (2013).
133. Khrameeva, E.E. *et al.* Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans. *Nat. Commun.* **5**, 3584 (2014).
134. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
135. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
136. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
137. Brennand, K.J. *et al.* Modeling psychiatric disorders at the cellular and network levels. *Mol. Psychiatry* **17**, 1239–1253 (2012).
138. Lancaster, M.A. *et al.* Cerebral organoids model human brain development and microcephaly. *Nature* **501**, 373–379 (2013).
139. Hackenberg, M. *et al.* miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.* **39**, W132–W138 (2011).
140. Jiang, Y. *et al.* Isolation of neuronal chromatin from brain tissue. *BMC Neurosci.* **9**, 42 (2008).
141. Evrony, G.D. *et al.* Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* **151**, 483–496 (2012).
142. Grindberg, R.V. *et al.* RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. USA* **110**, 19802–19807 (2013).
143. Shumway, M. *et al.* Archiving next generation sequencing data. *Nucleic Acids Res.* **38**, D870–D871 (2010).
144. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
145. Tanizawa, H. *et al.* Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* **38**, 8164–8177 (2010).
146. Gehlen, L.R. *et al.* Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. *Nucleus* **3**, 370–383 (2012).
147. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
148. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).