How low can you go? Calling robust ATAC-seq peaks through read down-sampling

Authors: Jayon Lihm[1], Sandra Ahrens[1], Sara Ballouz[1], Hayan Lee[2,3], Megan Crow[1], Jessica Tollkuhn[1], Shane McCarthy[1], Bo Li[1], W.R. McCombie[1], Jesse Gillis[1]*

Affiliations:
1 The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA
2 Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
3 Department of Genetics, Stanford University School of Medicine, Stanford University, Stanford, CA 94304, USA

* Corresponding author: Dr J Gillis, The Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor 11724, NY, USA

Contact Info:
Correspondence: Jesse Gillis (JGillis@cshl.edu)

Abstract

Chromatin accessibility provides an important window into the regulation of gene expression. Recently, the Assay of Transposase Accessible Chromatin with sequencing (ATAC-seq) was developed to profile genome-wide chromatin accessibility. While pipelines have been developed for the analysis of individual data sets, there has been little evaluation across data sets to determine best practices. In this work we analyzed 193 ATAC-seq data samples from 12 studies to determine the robustness of peaks, as well as their specificity across studies. To determine robustness, we employ a read-downsampling approach and find that even samples which have high average read-depth yield poorly powered peaks using conventional pipelines. To detect specificity, we compare peaks across our diverse corpus and find that peaks are promiscuously identified in most samples, with approximately 31K peaks per sample and 164 genes with peaks at their transcription start site in all samples. We evaluate the properties of these genes in detail, including mean expression across a very large collection of studies, as well as functional characteristics. Finding that many peaks are sensitive to slight perturbations in read-sampling, we develop an approach to improve the robustness of peak signals and apply this novel method to detect differential accessibility between the amygdala and cortex. Finally, using this approach, we identify 274 genes that are robustly differentially accessible between amygdala and cortex.

Keywords

ATAC-seq, meta-analysis, bootstrapping, resampling, aggregation, differential peaks, cortex, amygdala

Author Summary

ATAC-seq is a technology which enables estimation of the accessibility of chromatin more easily than previously possible, thus providing a new window into gene regulation. In this work, we perform a broad-based assessment of 12 previous studies, defining practices to enable productive use of ATAC-seq data, focusing on the specificity of peaks called within ATAC-seq data. We find genes commonly associated with accessible peaks and characterize the expression of these genes in a large corpus of data. We perform a series of downsampling experiments to estimate the robustness of results and, having defined practices yielding robust results, apply our methods to novel data within the mouse cortex and amygdala. Altogether, our results suggest that current peak calling protocols may be sensitive to subtle variation in the data but that this problem can be ameliorated by careful application of conventional statistical techniques.

Introduction

Compacted chromatin in eukaryotic cells constantly changes state between open and closed to harbor transcription factors for the regulation of gene expression [1, 2]. These dynamic changes in chromatin accessibility are a useful source of information to understand biological processes such as cell differentiation and development [3-5]. Advances in sequencing assays into epigenetic states, such as DNase-seq [6] (DNase I hypersensitive sites sequencing) and more recently ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) [7], have made it feasible to profile global patterns of chromatin accessibility at a genome-wide scale. ATAC-seq, in particular, has rapidly gained popularity due to its straightforward protocol and low input requirements [7, 8]. In essence, ATAC-seq exploits the tendency of transposons (through transposase) to incorporate into nucleosome-free regions of the genome, generating DNA fragments from enzyme ligation sites, which are then sequenced. Sequenced

reads show enrichment of accessible sites, piling-up as peaks, as transposase ligates to nucleosome-free regions at higher frequencies.

Due to its novelty, ATAC-seq data analysis has mostly relied on repurposed tools from other domains, such as ChIP-seq, DNase-seq and FAIRE-seq [9-11], with conventional peak-calling software being used to detect the regions of the genome with read “pile-ups”. At its core, peak-calling uses relative read density to define a peak, first estimating appropriate bin widths from the data and then looking for bins some fold-change higher than the background. Like many successful methodologies, there are a number of tailored steps which have emerged over time and which contribute to the utility of the method in ways which may not generalize (for a discussion of comparative performance in ChIP-seq data of common tools, see [12]). How best to determine a bin size, for instance, or how high a fold-change defines a difference to the background, may change dramatically in a new technology. Thus, repurposing methods to ATAC-seq might raise some room for concern, particularly given the potential for noise in a biological approach optimized for low input. However, through a comparative meta-analysis, one can assess the relative impact of method choices on the ability to obtain robust and distinguishable peak-calls.

In order to estimate the robustness of features within ATAC-seq data, we performed a comparative analysis across 193 mouse samples from twelve previously published studies. Originally developed for ChIP-seq data, MACS2 [9] is peak-calling software commonly applied to ATAC-seq, which we use and assess here. We show that peaks vary dramatically in their degree of robustness, with some peaks highly sensitive to slight alterations in read distributions and other peaks robust to dramatic losses of data. We determine 164 constitutively accessible genes across all samples, regardless of study design or biological condition. Based on the results from our meta-analysis, we propose a novel approach to call robust peaks through downsampling of reads. We then

apply this approach to our own data, where we detect epigenetic changes between the mouse amygdala and cortex, focusing on somatostatin positive (SOM+) neurons in fear-conditioned mice.

Results

We downloaded raw sequence files of 193 samples from 12 studies from the sequence read archive (SRA) [12] (**Table 1**). For all our analyses, we uniformly reprocessed the data, using the most commonly applied default parameters across the 12 studies (described in **S1 Table**). In addition, to assess the stability of our findings, we calculated results using the ENCODE pipeline [13, 14], BAMPE mode for paired-end reads [15], and tested the robustness of results to the use of their broad or narrow peak calling in MACS2 (see **S1-3 Fig**). The downloaded FASTQ files were processed through a unified pipeline described in **Fig 1** from mapping to peak calling (described in more detail under **Methods**).

Table 1. Summary of 12 studies.

Authors	Year	GSE ID	Number of biological samples	Total number of samples	ATAC-seq Samples
Atianand et al	2016	GSE78873	12	12	Bone-marrow derived macrophage cells
DellOrso et al	2016	GSE76010	4	4	Skeletal muscles C2C12 cell line - myoblasts and myotubes
Denny et al	2016	GSE81255	59	175	Small cell lung cancer from primary tumors and metastases
Dieuleveult et al	2016	GSE64825	10	20	Embryonic stem cells
George et al	2016	GSE74688	8	8	Acute myeloid leukemia cells
Hay et al	2016	GSE78800	19	19	Erythroid cells
Jiang et al	2016	GSE80272	8	8	B and T lymphocytes
Lara-Astiaso et al	2014	GSE59992	11	25	Hematopoietic differentiation stages
Mostafavi et al	2016	GSE75262	4	4	B-cells

Shih et al	2016	GSE77695	46	46	Innate lymphoid cells
Tong et al	2016	GSE74191	8	8	Bone-marrow-derived macrophages
Wang et al	2016	GSE68288	4	4	Hair follicle stem cells
Total			**193**	**333**	

Listed are the studies used in our analysis, as well as the number and type of sample, as well as the number of replicates.

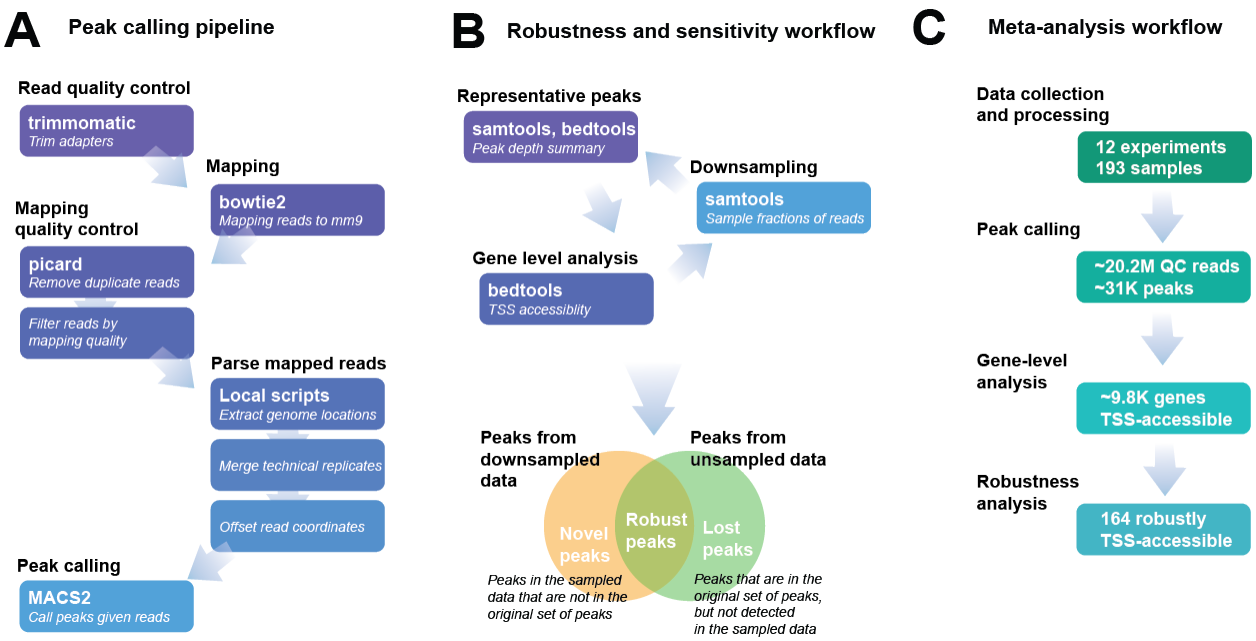


Figure 1 Schematic overview of methods (A) Peak calling pipeline. In brief, the reads are mapped following quality control and then MACS2 is used to call peaks. (B) Robustness and sensitivity workflow to discover stable peaks. We call robust peaks those that are not lost (compared to original data) or gained in the downsampling process. (C) Meta-analysis workflow. We took 12 studies totaling 193 samples and tested each sample for robust peaks and TSS-accessible genes.

Read depth varies dramatically within and across ATAC-seq studies

From a very basic perspective, we see that studies are highly heterogeneous. For example, within the 12 studies we assessed, the number of reads varies dramatically, often by orders of magnitude across samples (**Fig 2A**), averaging 5.7-fold between the highest and lowest depth samples. This is also true across studies, where the number of reads also varies dramatically, ranging from 8.0M - 44.7M with fold-change within study from 1.6 to 14.2-fold (**S2 and S3 Tables**). This variation is mirrored by large variation in

the number of peaks detected per experiment (narrow mean 31,022 +/- SD 18,751; **Fig 2A**). Variation of peak counts was modestly lower within experiments, although still large, with an average fold change of 2.6 between least and most detected sample within an experiment, and ranging from 1.1 to 8.6-fold, except for one outlier case with 39.6-fold differences in Denny et al's studies [16] (**S3 Table**). This large degree of variation was consistently shown within each of the three parameter settings, regardless of peak modes. (**S1 Fig)**

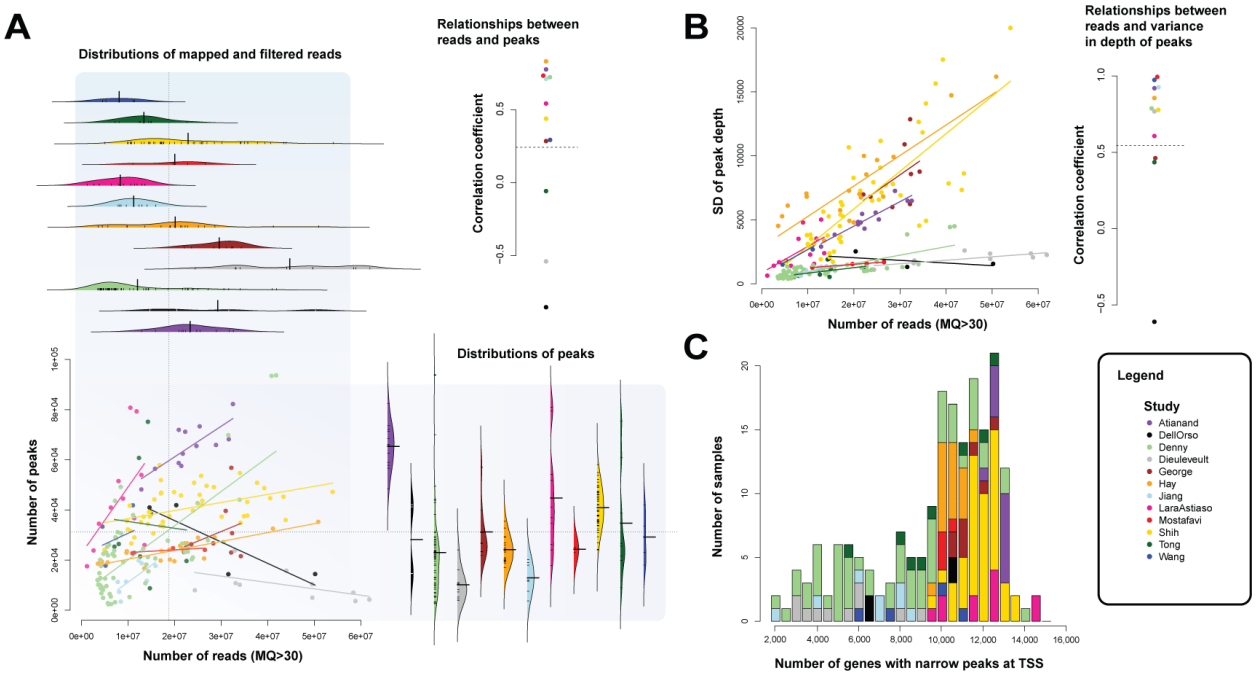


Figure 2 ATAC-seq data is highly variable in read depth and peak counts (A) Read depth and number of peaks called are strongly related. The x-axis is the number of mapped reads to genome with MQ30 or greater. The violin plots at the top show the distribution of reads per study, which vary within and across studies. The number of peaks also varies within and across studies, shown in the violin plots on the right. The correlations are also study dependent. (B) Reads versus variability of reads per peak within each sample. Read-depth per peak region is computed as sum of read depth within peak region and standard deviation is calculated within a sample. (C) Stacked histogram of TSS-accessible genes. Number of TSS-accessible genes varies largely with average at 9.8K.

To determine if some of this variability diminishes by focusing on genomic locations of likely biological importance, we look to the transcription start sites (TSS) as they require open chromatin for transcriptional machinery to bind. In order to measure a TSS localized signal, we first count the number of peaks within +/- 1,000 base pairs around a TSS, and collapse these peaks into a "TSS region". We then call genes "TSS-

accessible" if any peaks are present within a gene's TSS region. The number of TSS-accessible genes also varies substantially across samples, ranging from 1,784 to 14,765 (**Fig 2C and S2 Fig**). Even if much of this variation is biological, the breadth of it will make interpretation challenging. On average, ~9.8 thousand genes were TSS-accessible, constituting 40.37% of RefSeq genes (mm9) in the mouse genome.

Increased read depth yields more peaks but also more variable peaks

Having observed dramatic differences in read depth within and across studies, we wished to quantify whether this variability had an effect upon the observed number of peaks. We assessed the correlation between the input read depth and the number of output peaks (**Fig 2A** and **S2 Table**). Overall, we observe a positive trend between the numbers of reads and peaks (Pearson correlation coefficient r=0.24, p<0.001) though the exact trend varied substantially within each study (**Fig 2A**). Interestingly, two of the three studies with the highest mean reads (of mapping quality (MQ) 30 or greater), Dieuleveult et al (2016) [17] and Dell'Orso et al (2016) [18], showed negative correlations between peaks and read depth (r = -0.54 and -0.86, respectively), suggesting that the removal of false positive peaks, rather than the ascertainment of new peaks, is the main advantage to be obtained from sequencing to higher depths. We note that while we cannot definitively rule out biological variability as a driving factor for the changes in number of peaks obtained within studies, the relationship with read depth makes a technical explanation seem likelier.

To test how increasing coverage contributes to the variability of signals, we measured the standard deviation of peak intensities within a sample (**Fig 2B, Table S2**). The peak intensity is defined as the sum of reads underlying each peak. As the total read depth increases, the distribution of reads associated with peaks becomes steadily broader (i.e., higher standard deviation). The overall trend in each study is linear and

strongly positive (r= 0.54, p< 5E-16), although the slope of the line varies between studies. In essence, samples with more reads have increasingly variable peak intensities; this is likely to drive differences in the robustness of called peaks.

Because the peaks in TSS regions are more likely to be biologically interpretable, we again assessed the relationship between read depth and the counts of TSS regions with peaks (**Fig 2C**). Within almost every study, as the number of reads increased the number of TSS-accessible genes increased, but this time plateauing to approximately 12K genes at 30M reads (**S2 Fig**). This suggests that higher depth (i.e., minimum 30M MQ>30 reads) is sufficient to obtain good gene coverage in peak ascertainment for TSS regions. On average, 19.2% of 193 samples met this criterion. The utility of this criterion within the TSS-regions is also captured by the number of peaks within a sample with fold changes below 2, where fold change is calculated as the average depth of each peak divided by the overall average depth. This fraction follows a roughly exponential decay (**S3 Fig)** with fewer than 10% of peaks failing to meet this criterion at 30 million reads or more.

Peak signals are significantly disrupted by changing coverage

Our previous results suggest an important dependency for peak detection on read depth within and across samples. In order to test the sensitivity of peaks within each sample, we down-sampled the original reads and re-assessed the peaks called (see schematic, **Fig 1B**). For each read-sampling, ranging from 10% to 90% of the original reads, we re-processed and generated peak sets to compare with the original set of peaks. For the purposes of this analysis, we considered peaks that only became detectable after down-sampling to likely reflect false positives. In general, for each study, the fewer reads that are present the fewer peaks that are detected, and so a relatively small number of novel peaks are detected at 10% sampling, with some exceptions (study-average 97 +/- 248

new peaks, **S4 Table**). This rises slightly to an average of 3.8% novel peaks at the 90% sampling (**Fig 3A**).

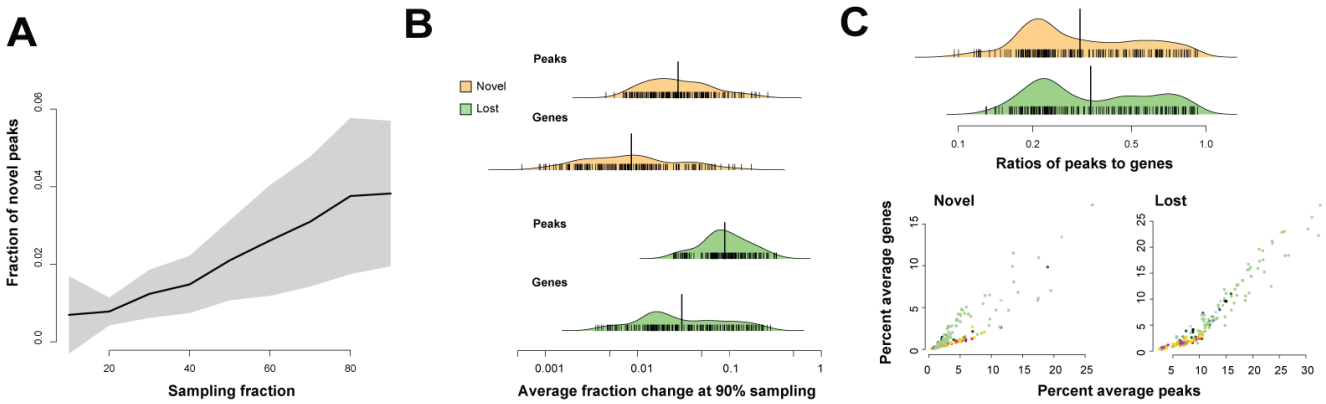


Figure 3 Peaks calls vary strongly in robustness to downsampling (A) Fraction of novel peaks called across sampling fractions. The x-axis is the percentage of sampling ranging from 10% to 90%. The y-axis shows the average fraction of novel peaks out of number of peaks called per sample at a sampling percentage. (B) Histogram of average fraction of novel (orange) and lost (green) peaks and genes at 90% sampling. The y-axis represents the frequency of samples across all studies. With 90% of data, the fraction of lost peaks ranges to 30%. (C) Comparisons of changes in fractions of peaks and TSS-accessible genes lost and gained at 90% sampling. The top distributions show ratios of peaks to genes for novel (top in orange) and lost (bottom in green). The two scatterplots below show the percent of peaks (x-axis) compared to percent of genes (y-axis), colored by study. The peaks to genes relationships are correlated, but study dependent.

While the creation of peaks from removing data is surprising, their disappearance is a likelier outcome. We next assessed the sensitivity of peaks to small changes in read coverage, defined as a 10% loss of reads. Empirically, this is a small perturbation for any given sample within an experiment since it usually leaves a sample's rank by number of reads unchanged within its experiment (a rank shift requires a 12% change in number of reads, on average). Sampling 90% of reads results in the loss of 2.4%-45.9% of peaks per sample, with an average of 11.8% (**Fig 3B**). Within each experiment this ranges between 6.4%-34.0%. In order to investigate whether the lost peaks were poorly powered peaks (i.e. q-values of these peaks were right below q-value cutoff line), we checked q-value distributions of lost peaks aggregated across samples in the original peak sets. Aggregated distribution showed that 33.8% of lost peaks had q-values between 0.001 and 0.0001, suggesting many of lost peaks were relatively well powered. (**S4 Fig**)

Interestingly, while slight read perturbations have an impact on peak detection globally, the impact on TSS regions was lower, with 6.0% of TSS-accessible peaks being lost on average. In addition, the ratio of globally lost peaks to lost TSS-accessible peaks tended to be strongly consistent across samples, and even more so within experiments (**Fig 3C**) suggesting that specific assessment of functional targets is a promising strategy to distinguish noise from signal within the data. This greater stability of peaks within the TSS also suggests that peaks disrupted by read-downsampling are likely false positives, and that bootstrapping or parallel approaches can exploit this to yield more robust results, a possibility considered in more detail below (see section "Downsampling strategy to detect robust peak differences").

Commonly accessible genes shared by all samples

These results suggest that TSS-accessible ATAC-seq signals are comparatively reliable although also common across the genome, with the vast majority of samples exhibiting many thousands of peaks. Our evaluation across experiments allows us to estimate the probability of a gene exhibiting a TSS-accessible peak, shown in **Fig 4A**. Of the 24K genes assessed, ~5,678 genes (23.3%) exhibited no peak in any of samples in the TSS region and ~3,603 genes (14.8%) showed no peak within 1Kbp of a gene body. Approximately 11K genes (43.6%) exhibit a peak within the TSS region in more than half of samples. ~3,127 genes (12.8%) showed peaks in 90% or more samples in the TSS region. Thus, ATAC-seq peaks are relatively common.

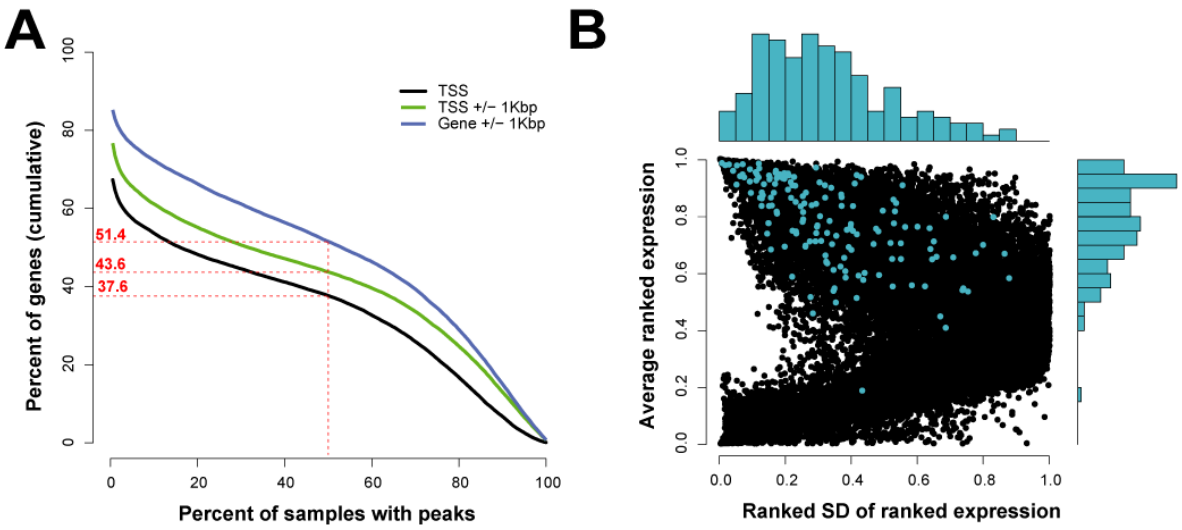


Figure 4 Properties of commonly accessible genes (A) Probability of a gene exhibiting a TSS-accessible peak across 193 samples. Black line is for peaks overlapping with TSS, green line is for peaks overlapping in TSS regions, and blue line is for peaks overlapping with genic regions (+/- 1kb of gene body). 50% or more samples have peaks in roughly 51% of genic regions in RefSeq. (B) Ranked mean expression (y-axis) of genes across compared to their variation in expression (SD, x-axis). TSS-accessible genes are highlighted in blue. The histogram on the top shows the distribution of SDs for the subset of 158 TSS-accessible genes, and the histogram on the right shows the distribution of ranked expression.

Interestingly, there were 164 genes which showed peaks in all of our assessed samples, suggesting a broad biological role which may be of use for quality control (as in housekeeping genes). To gain insight into consistent signal within the ATAC-seq data, we looked at the 164 genes whose TSS regions are accessible in all samples (**S5 Table**). The collected twelve studies vary broadly in sample type and experimental conditions, and thus the commonly detected genes are either functionally constitutive or technically prone to recruit reads in their TSS regions; these are both potentially important cases to assess. Because these genes are ubiquitously in an open chromatin state, we can assess their expression level across many conditions. At the least, we might expect genes always exhibiting ATAC-seq peaks to exhibit higher than average expression. We obtained mouse expression data for all genes from a large corpus of studies curated within the Gemma database [19]. Of the 164 genes, 158 were in the database, and almost all of the common TSS-accessible genes rank in the top half of

genes by expression level (**Fig 4B, S5 Fig** with broad peaks). On average, the set of TSS-accessible genes show moderately high expression (mean rank=0.78, i.e., top quartile average rank). Consistent with the view that these may be housekeeping genes, the standard deviation of expression level was also low among the commonly TSS-accessible genes (ranked SD=0.33), exhibiting lower variability even than genes of their expression typically do ($p<0.01$, Mann-Whitney test).

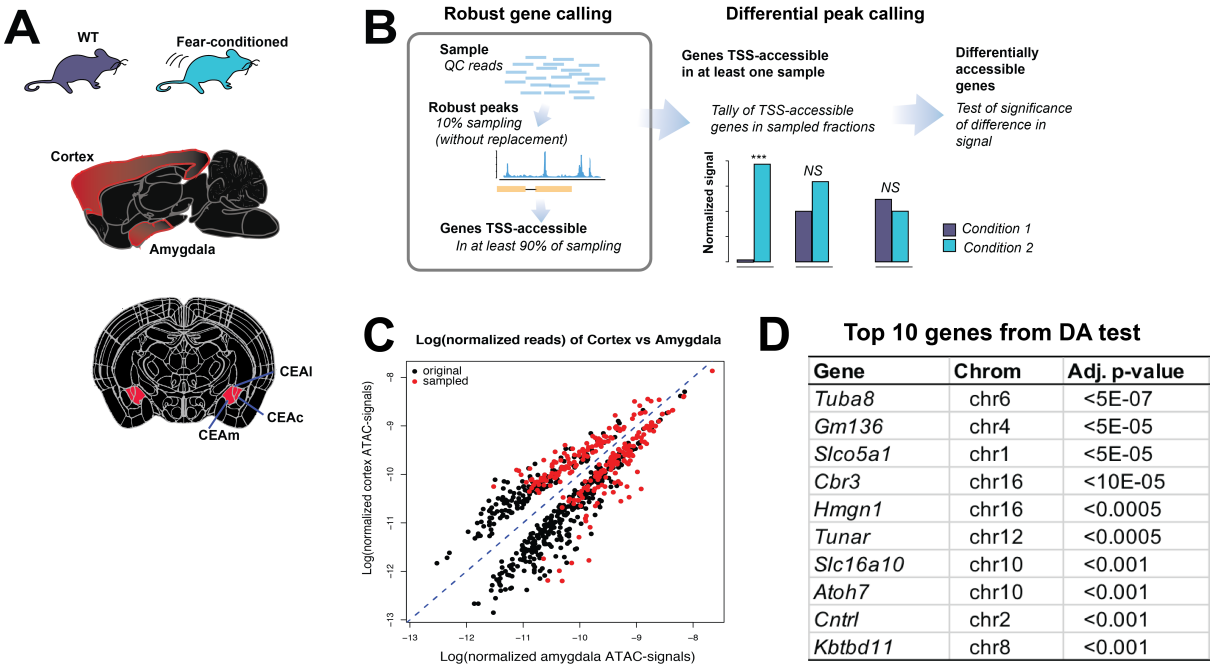
To assess the functional properties of the 164 putative housekeeping genes, we performed functional enrichment using the Gene Ontology (GO). Interestingly, despite the comparative ease with which function enrichment arises in gene sets [20], no GO terms overlapped significantly with the set of TSS-accessible genes, 138 of which did have at least some GO terms associated with them.

Downsampling strategy to detect robust peak differences

ATAC-seq is often done to compare two conditions, in which case peak calling is followed by a differential peak detection analysis. As we have shown, peaks can be noisy, meaning that differences between samples may not be robust. Ultimately, our goal is to apply ATAC-seq to understand condition variability in complex systems. We are particularly interested in profiling the brain, where the promiscuous and challenging character of transcription makes quality control and signal refinement particularly important.

To assess whether robust peak differences exist, we applied ATAC-seq to a specific neuronal type in selected structures in the brain under different behavioral or genetic conditions, focusing on differences between the amygdala and cortex. This partly is a reflection of our experimental interests, but we think is well suited to the question of quality control. Substantial evidence indicates that learning and memory, including the formation of fear memory, involves epigenetic changes in the brain [21, 22].

The amygdala, including its central nucleus (CeA), is known to be involved in processing memory of fear and developing adaptive behaviors [23], and is arguably the most intensively studied brain structure in neuroscience. To explore such global epigenetic changes, we performed ATAC-seq on samples from the CeA and the cortex, across a total of eighteen samples (**Fig 5A**, see **S1 Text** for further experimental details).



Gene	Chrom	Adj. p-value
Tuba8	chr6	<5E-07
Gm136	chr4	<5E-05
Slco5a1	chr1	<5E-05
Cbr3	chr16	<10E-05
Hmgn1	chr16	<0.0005
Tunar	chr12	<0.0005
Slc16a10	chr10	<0.001
Atoh7	chr10	<0.001
Cntrl	chr2	<0.001
Kbtbd11	chr8	<0.001

Figure 5 ATAC-seq on fear-conditioned mice. (A) Schematic of experiment. Conditions were assessed within each brain region. (B) Downsampled gene/peak distribution and differential peak analysis. (C) Normalized ATAC-seq signals of cortex and amygdala in log scale. Black dots are differentially accessible genes called in the original data without sampling and red dots are the ones after sampling. Blue dotted line is $y=x$. (D) The list of top 10 genes from differential accessibility t-test sorted by unadjusted p-values.

We first examined the number of peaks from the full data to determine if this new set of data aligns with the general results obtained from our meta-analysis of other mouse studies. On average, 13K genes were called TSS-accessible in the 18 samples, ranging from 11K to 14K genes (**S6 Table**). To increase the power of differential peak detection, we repurposed our approach to assess peak robustness as a method to ascertain robust peaks, in close parallel to bootstrapping, and focusing on the most

reliable peaks that are consistently called after downsampling to 10% (**Fig 5B**). Using our downsampling approach, the number of genes with peaks was reduced to an average of 4,982 genes (**S6 Table**). While many of the omitted positives from this read sampling approach might still be true positives, almost all of the commonly accessible genes identified earlier still exhibit peaks within this data (i.e., averaging 157 out of 164), which suggests that the most salient peaks have been maintained. We note that two samples fail these QC checks, showing high variability in peaks with downsampling and MQ30 reads < 30M (**S6 Fig**); we excluded these two samples from subsequent analysis.

Identifying genes with differential TSS-accessibility between amygdala and the cortex

Next we tested differential chromatin accessibility for each gene across brain regions, experimental conditions, and different mouse models. The identification of robust peaks allows for a comparatively simple test for differential peak intensities. We simply performed Student's t-test on normalized ATAC-seq read signals in TSS regions of genes called to determine whether chromatin accessibility signals are enriched in samples from one condition over the other. We first made consensus gene sets across 16 samples from the full data and from our downsampling approach, respectively. Here consensus gene lists are TSS-accessible genes that are called in at least one sample. Compared to the full data, the number of regions tested is substantially dropped from 15.7K to 10.4K. Differential t-test on these gene sets detected 274 genes with the sampling approach at FDR<0.05, reduced from 723 genes with the full data. The 274 differential genes are all subset of the 723 genes but with higher coverage (**Fig 5C** and **S7 Fig**). We checked functional properties of the top ranked genes among 274 genes. The top 10 genes included potentially interesting genes in brain and optic nerve development as well as a gene associated with chromatin structure, although their

specific roles in fear controlling need to be further studied. (**Fig 5D** and **S7 Table**) Tubulin Alpha 8 (*Tuba8*) gene (adjusted p-value <10E-06) is associated with the development of brain structure, specifically forming many small gyri before birth (called polymicrogyria) and the underdevelopment of the optic nerves (called optic nerve hypoplasia) [24]. Atonal BHLH Transcription Factor 7 (*Atoh7*, adjusted p-value<10E-06) also plays a role in controlling retinal ganglion cell and optic nerve formation [25]. *Tunar* (TCL1 Upstream Neural Differentiation-Associated RNA, adjusted p-value<10E-06) produces a long non-coding RNA that is involved in neural differentiation of embryonic stem cells. This gene is also known to be associated with Huntington Disease and Glioma, a type of brain tumor [26]. Interestingly, *Hmgn1* (High Mobility Group Nucleosome Binding Domain 1) is a protein coding gene that is associated with transcriptionally active chromatin [27]. The protein from this gene might be associated with maintaining open chromatin structure around transcribable genes. Other genes were associated with more general properties of cell functions such as the maturation of centrosome (*Cntrl*) [28], transporter activities (*Slco5a1, Slc16a10*) [29, 30], or metabolism pathway (*Cbr3*) [31]. Out of the 10 genes, 5 genes showed relatively high mean expression across tissues (>50th percentile) in Gemma database, 3 genes were in medium level (25-50th percentile), and the other genes, *Gm136* and *Atoh7* showed relatively low expression level (5th, 7th percentile, respectively).

No significant differences were observed between the behavioral conditions we assessed (fear-conditioning and control), our genetic target (*SOM*$^{Erbb4\text{-}KO}$ and wild-type mice) after the correction of p-values.

Discussion

In this work, we performed a meta-analysis across ATAC-seq studies, identifying features associated with robust peaks, including peak distributions in TSS regions, replicability of peak calling in down-sampled reads, and potential housekeeping peaks. We used this information to perform a novel differential peak analysis within our own experimental data, comparing amygdala to cortex, and identified differentially accessible genes between the brain regions. Our work provides general guidance on how to obtain robust results within ATAC-seq data, particularly by employing sensible baseline approaches, such as read downsampling. Estimating the robustness of results is particularly essential for any differential peak analysis. In essence, each peak is an attempt to observe a significant difference from baseline, but comparisons between samples as to whether those significant observations are present or not are, themselves, not necessarily significant [32]. However, selecting those peaks which are robust still yields a clear differential signal when assessed for significance directly.

As with any sequencing based methods, it is not a surprise to find that the determination of appropriate read coverage is essential in ATAC-seq analysis. Peak-calling depends on determining a significant pile-up of read density relative to the background, and ascertainment of peaks is therefore expected to be dependent on read depth. This is appreciated to some degree within existing analyses, and normalization after peak detection is often employed to compare amplitudes across samples. However, what our works highlights is that the performance of peak detection itself is largely uncontrolled for read depth, both within and across samples. This has a critical impact on which regions will be assessed for differences in amplitude. Our analysis suggests that simple renormalization is unlikely to prove effective, since the distribution

of peak regions, used to select regions for the downstream amplitude assessment, will remain very biased by read coverage.

As in expression analysis, many potential biases are best addressed through cross-laboratory comparison similar to the MAQC evaluations [33]. Our own meta-analysis is a step in this direction, although our recommendations are likely conservative since we are sampling across real biological variability in addition to technical variability. Given the likely utility and popularity of single cell ATAC-seq, where technical challenges are likely to be more severe, we suspect some degree of methodological conservatism is likely helpful. More importantly, our results on brain region comparisons suggest the biological signals in ATAC-seq data survive stringent filtering. The results from differential peak test yields much smaller candidate lists, useful for targeted experimental evaluation. We look forward to the accumulation of precise gene-specific results from careful ATAC-seq analysis in the ongoing effort to understand functional relationships across the genome.

Materials and Methods

Data and data processing

Original FASTQ files for the 12 studies were downloaded from SRA [34] (https://www.ncbi.nlm.nih.gov/sra). We trimmed adapters that were in Buenrostro et al (2013) [7] in both forward and reverse complement sequences. The trimmed reads were mapped to mm9 using bowtie2 [35, 36] (v2.2.3). After the mapping, duplicates were removed with Picard (v1.88). Reads whose mapping quality was below 30 were excluded and those mapped to mouse genome chromosomes (chr1-19,X and Y) were taken into the analysis. We then merged the resulting BAM files from the same biological

sample. Lastly, the coordinates of reads were adjusted as in Qu et al (2015)[37]. Peaks are called by MACS2 [9, 38] (v2.1.0.20140616). We used the sum of read depth per peak for the calculation of standard deviation with both top and bottom 5% trimmed.

Gene level analysis

For our gene annotations, we used the mm9 RefSeq gene table from the UCSC genome browser. Multiple promoter regions within a single gene are merged if they overlap. If a single promoter region has multiple gene annotations, all the genes mapped to that region are considered TSS-accessible.

RNA expression data of commonly accessible genes

We collected 199 mouse RNA-seq experiments (with a total of 4,059 samples) from the Gemma database [19]. Each experiment was mapped using bowtie2 [35] and quantified using RSEM (version 1.2.5) [36] with annotations from “mm10_ensembl_72”. A total of 38,223 mouse genes and transcripts were detected, with 22,393 genes with RefSeq IDs. For each sample, we ranked the TPM expression levels, and then averaged across the samples, giving an averaged ranked expression value for each gene. Genes not detected within a given sample were treated as not expressed. To test for the level of expression variability of our commonly accessible gene list, we compared the distribution of ranked standard deviations of these genes to a matched control set. The difference between our set and control gene set is tested by Wilcoxon test.

Enrichment analysis of commonly accessible genes

Using Bioconductor package EGAD [20], we downloaded associated GO data. GO terms with IEA evidence were filtered and genes overlapping with RefSeq database were kept in the further analysis. We selected GO terms with 10 to 300 genes, resulting in 5,579 GO terms. After hypergeometric test, resulting p-values were corrected for

multiple hypothesis testing by the Benjamini-Hochberg method. We also performed an extended functional enrichment test by the Mann-Whitney U test to detect differences in ranks between genes that are listed in a GO term and genes that are not related to the GO term. The p-values were adjusted for multiple tests by the Benjamini-Hochberg method.

Mouse sample collection

The SOM-IRES-Cre mice [39], the *Rosa26-loxP-STOP-loxP-H2B-GFP* reporter line [40] and the *ErbB4*$^{lox/lox}$ mice [41] were generated as described in Taniguchi et al. (2011) [39], He et al. (2012) [40], and Golub et al. (2004) [41]. All procedures involving animals were approved by the Institute Animal Care and Use Committees of Cold Spring Harbor Laboratory. For detailed wet-lab protocols, see **Supplementary Text.**

Differential accessibility test between two conditions

With the full data before read-sampling, we called a gene is TSS-accessible if any peak overlaps with the gene's TSS region. After read-down sampling, a gene was called TSS-accessible if there were overlapping peaks in at least 9 replicates. Then we generated the consensus gene sets for the full data and sampled data separately by collecting genes that are called TSS-accessible in any of 16 samples. Differential accessibility was tested for those consensus gene's TSS regions based on the normalized ATAC-seq signals. We computed the sum of depth per position for each TSS region (+/- 1Kb of TSS). The summed signal is normalized by the total sum of depth of all TSS regions per sample. Differential t-test was performed on the normalized signals between two conditions we tested.

Availability of data and material

The datasets generated and/or analysed during the current study are available in the GITHUB repository, https://github.com/jlihm-seq/atac-seq. Raw data and summary files of 18 mouse brain samples have been uploaded to GEO and SRA (accession GSE111021, secure token: axmpigwkbvinrol).

Acknowledgements

The authors would like to thank Sanja Rogic and Paul Pavlidis for their assistance with the Gemma RNA-seq data. This research used resources of the NERSC (National Energy Research Scientific Computing Center), a DOE Office of Science User Facility supported by the Office of Science of the U.S. The 18 mouse brain samples were sequenced at the CSHL Cancer Center Next Generation Sequencing Shared Resource, supported by Cancer Center Grant #5P30CA045508. Research reported in this publication was supported by the National Institutes of Health R01LM012736 and R01MH113005 to JG as well as a gift from T. and V. Stanley. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Cuvier O, Fierz B. Dynamic chromatin technologies: from individual molecules to epigenomic regulation in cells. Nat Rev Genet. 2017;18(8):457-72.
2. Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. Science. 2017;357(6348).
3. Zhou VW, Goren A, Bernstein BE. Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet. 2011;12(1):7-18.
4. Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. Epigenetics & chromatin. 2014;7(1):33.

5. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100.
6. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor protocols. 2010;2010(2):pdb prot5384.
7. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods. 2013;10(12):1213-8.
8. Madrigal P. On Accounting for Sequence-Specific Bias in Genome-Wide Chromatin Accessibility Experiments: Recent Advances and Contradictions. Frontiers in bioengineering and biotechnology. 2015;3:144.
9. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008;9(9):R137.
10. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome biology. 2011;12(7):R67.
11. Kumar V, Muratani M, Rayan NA, Kraus P, Lufkin T, Ng HH, et al. Uniform, optimal signal processing of mapped deep-sequencing data. Nature biotechnology. 2013;31(7):615-22.
12. Thomas R, Thomas S, Holloway AK, Pollard KS. Features that define the best ChIP-seq peak calling algorithms. Briefings in Bioinformatics. 2017;18(3):441-50.
13. Jin Wook Lee AK. ATACSeq Pipeline 2017 [Available from: https://docs.google.com/document/d/1f0Cm4vRyDQDu0bMehHD7P7KOMxTOP-HiNoIvL1VcBt8/edit#heading=h.sk5jgu4nmrox].
14. Jin Wook Lee CSF, Daniel Kim, Nathan Boley, Anshul Kundaje. ATAC-Seq/DNase-Seq Pipeline [Available from: https://github.com/kundajelab/atac_dnase_pipelines].
15. Evan D Tarbell TL. HMMRATAC, the Hidden Markov ModeleR for ATAC-seq. bioRxiv. 2018.
16. Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Gruner BM, et al. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell. 2016;166(2):328-42.
17. de Dieuleveult M, Yen K, Hmitou I, Depaux A, Boussouar F, Dargham DB, et al. Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature. 2016;530(7588):113-6.
18. Dell'Orso S, Wang AH, Shih H-Y, Saso K, Berghella L, Gutierrez-Cruz G, et al. The Histone Variant MacroH2A1.2 is Necessary for the Activation of Muscle Enhancers and Recruitment of the Transcription Factor Pbx1. Cell reports. 2016;14(5):1156-68.
19. Zoubarev A, Hamer KM, Keshav KD, McCarthy EL, Santos JR, Van Rossum T, et al. Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. Bioinformatics. 2012;28(17):2272-3.

20. Ballouz S, Weber M, Pavlidis P, Gillis J. EGAD: ultra-fast functional analysis of gene networks. Bioinformatics. 2017;33(4):612-4.
21. Zovkic IB, Guzman-Karlsson MC, Sweatt JD. Epigenetic regulation of memory formation and maintenance. Learn Mem. 2013;20(2):61-74.
22. Zovkic IB, Sweatt JD. Epigenetic mechanisms in learned fear: implications for PTSD. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology. 2013;38(1):77-93.
23. Penzo MA, Robert V, Li B. Fear conditioning potentiates synaptic transmission onto long-range projection neurons in the lateral subdivision of central amygdala. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2014;34(7):2432-7.
24. Abdollahi MR, Morrison E, Sirey T, Molnar Z, Hayward BE, Carr IM, et al. Mutation of the variant alpha-tubulin TUBA8 results in polymicrogyria with optic nerve hypoplasia. American journal of human genetics. 2009;85(5):737-44.
25. Brown NL, Dagenais SL, Chen CM, Glaser T. Molecular characterization and mapping of ATOH7, a human atonal homolog with a predicted role in retinal ganglion cell development. Mamm Genome. 2002;13(2):95-101.
26. Lin N, Chang KY, Li Z, Gates K, Rana ZA, Dang J, et al. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol Cell. 2014;53(6):1005-19.
27. Rattner BP, Yusufzai T, Kadonaga JT. HMGN proteins act in opposition to ATP-dependent chromatin remodeling factors to restrict nucleosome mobility. Mol Cell. 2009;34(5):620-6.
28. Gromley A, Jurczyk A, Sillibourne J, Halilovic E, Mogensen M, Groisman I, et al. A novel human protein of the maternal centriole is required for the final stages of cytokinesis and entry into S phase. J Cell Biol. 2003;161(3):535-45.
29. Kim DK, Kanai Y, Matsuo H, Kim JY, Chairoungdua A, Kobayashi Y, et al. The human T-type amino acid transporter-1: characterization, gene organization, and chromosomal location. Genomics. 2002;79(1):95-103.
30. Sebastian K, Detro-Dassen S, Rinis N, Fahrenkamp D, Muller-Newen G, Merk HF, et al. Characterization of SLCO5A1/OATP5A1, a solute carrier transport protein with non-classical function. PloS one. 2013;8(12):e83257.
31. Bains OS, Karkling MJ, Lubieniecka JM, Grigliatti TA, Reid RE, Riggs KW. Naturally occurring variants of human CBR3 alter anthracycline in vitro metabolism. J Pharmacol Exp Ther. 2010;332(3):755-63.
32. Gelman A, Stern H. The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. The American Statistician. 2006;60(4):328-31.
33. Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotech. 2014;32(9):903-14.
34. Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C. The Sequence Read Archive. Nucleic Acids Research. 2011;39(Database issue):D19-D21.
35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012;9(4):357-9.

36. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics. 2011;12(1):323.
37. Qu K, Zaba Lisa C, Giresi Paul G, Li R, Longmire M, Kim Youn H, et al. Individuality and Variation of Personal Regulomes in Primary Human T Cells. Cell Systems. 2015;1(1):51-61.
38. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome biology. 2008;9(9):R137.
39. Taniguchi H, He M, Wu P, Kim S, Paik R, Sugino K, et al. A resource of Cre driver lines for genetic targeting of GABAergic neurons in cerebral cortex. Neuron. 2011;71(6):995-1013.
40. He M, Liu Y, Wang X, Zhang MQ, Hannon GJ, Huang ZJ. Cell-type-based analysis of microRNA profiles in the mouse brain. Neuron. 2012;73(1):35-48.
41. Golub MS, Germann SL, Lloyd KC. Behavioral characteristics of a nervous system-specific erbB4 knock-out mouse. Behav Brain Res. 2004;153(1):159-70.

Supporting Information

S1 Text: Detailed methods and experimental designs

S1 Figure: Number of peaks vs. number of filtered reads. MACS2 peaks are generated under **(A,B)** default parameter setting with broad and narrow mode, **(C,D)** ENCODE parameters with p-value cutoff only, **(E,F)** ENCODE parameters with the default q-value restrictions (0.1 for broad, 0.05 for narrow), and **(H,I)** BAMPE parameters.

S2 Figure: Number of TSS-accessible genes vs. number of filtered reads. Each color represents each study. Left panel is generated from *broad* peaks and right panel is from *narrow* peaks.

S3 Figure: Fraction of peaks with fold-change of read signals below 2. Left panel is from *narrow* peaks and right panel is from *broad* peaks.

S4 Figure: Distribution of q-values of lost peaks at 90% sampling. Q-values are transformed to negative log scale. Y-axis is the relative frequency.

S5 Figure: Related to Figure 4, with broad peaks: (A) Probability of a gene exhibiting a TSS-accessible peak across 193 samples. Black line is for peaks overlapping with TSS, green line is for peaks overlapping in TSS regions, and blue line is for peaks overlapping with genic regions (+/- 1kb of gene body). 50% or more samples have peaks in roughly 55% of genic regions in RefSeq. (B) Ranked mean expression (y-axis) of genes across compared to their variation in expression (SD, x-axis). TSS-accessible genes are highlighted in blue. The histogram on the top shows the distribution of SDs for the subset of 438 TSS-accessible genes, and the histogram on the right shows the distribution of ranked expression.

S6 Figure: Number of TSS-accessible genes called after applying 10% downsampling strategy. A gene is called TSS-accessible if called in more than 9 replicates.

S7 Figure: Normalized signals vs. p-value of amygdala and cortex

S1 Table: MACS2 parameter settings in 12 studies

S2 Table: Summary of sequencing reads including number of genomic reads and mitochondrial reads, followed by summary of peak calling and TSS-accessible genes of 193 samples

S3 Table: Summary of sequencing reads and peak calling results per study. Aggregated from S2 Table per study.

S4 Table: Number of novel and lost peaks with 10% and 90% sampling for 193 samples

S5 Table: Commonly accessible genes called by narrow and broad peaks

S6 Table: Description of 18 mouse brain samples, followed by number of reads and TSS-accessible genes

S7 Table: Top 10 differentially accessible genes between amygdala and cortex

List of abbreviations

ATAC-seq: Assay of Transposase Accessible Chromatin with sequencing

TSS: Transcription start site

GO: Gene ontology

MQ: Mapping Quality

FDR: False Discovery Rate

MAQC: Micro Array Quality Control

Authors' contributions

JL performed the computational experiments. SA performed the wet-lab experiments. JG designed the study. JL and JG analyzed the data. JG and JL wrote the manuscript. MC and HL assisted in computational experimental design. JT, SM, BL, and WRM assisted in wet-lab experimental design. All authors read and approved the final manuscript.

Declaration of Interests

The authors declare that they have no competing interests.