Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

APPLICATION NOTE

AIAP: A Quality Control and Integrative Analysis Package to Improve ATAC-seq Data Analysis

Shaopeng Liu[1,#], Daofeng Li[2,#], Cheng Lyu[1,#], Paul M. Gontarz[1], Benpeng Miao[1,2], Pamela A.F. Madden[3], Ting Wang[2,*], Bo Zhang[1,*]

[1]*Department of Developmental Biology, Center of Regenerative Medicine, Washington University School of Medicine, St. Louis, MO 63108, USA*

[2]*Department of Genetics, Center for Genomic Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108, USA*

[3]*Department of Psychiatry, Washington University School of Medicine, St. Louis, MO 63108, USA*

Received 17 February 2020; revised 28 May 2020; accepted 27 October 2020
Available online 15 July 2021

Handled by Zhihua Zhang

Abstract Assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) is a technique widely used to investigate genome-wide chromatin accessibility. The recently published Omni-ATAC-seq protocol substantially improves the signal/noise ratio and reduces the input cell number. High-quality data are critical to ensure accurate analysis. Several tools have been developed for assessing sequencing quality and insertion size distribution for ATAC-seq data; however, key quality control (QC) metrics have not yet been established to accurately determine the quality of ATAC-seq data. Here, we optimized the analysis strategy for ATAC-seq and defined a series of QC metrics for ATAC-seq data, including reads under peak ratio (RUPr), background (BG), promoter enrichment (ProEn), subsampling enrichment (SubEn), and other measurements. We incorporated these QC tests into our recently developed ATAC-seq Integrative Analysis Package (AIAP) to provide a complete ATAC-seq analysis system, including quality assurance, improved peak calling, and downstream differential analysis. We demonstrated a significant improvement of sensitivity (20%–60%) in both peak calling and differential analysis by processing paired-end ATAC-seq datasets using AIAP. AIAP is compiled into Docker/Singularity, and it can be executed by one command line to generate a comprehensive QC report. We used ENCODE ATAC-seq data to benchmark and generate QC recommendations, and developed *qATACViewer* for the user-friendly interaction with the QC report. The software, source code, and documentation of AIAP are freely available at https://github.com/Zhang-lab/ATAC-seq_QC_analysis.

KEYWORDS ATAC-seq; Quality control; Chromatin accessibility; Differential analysis; Data visualization

Introduction

To regulate the transcription of a eukaryotic genome, chromatin must remain in an accessible state to allow binding of transcription factors and initiation of transcription activation [1–4]. Several sequencing-based methods have been developed to assess chromatin accessibility and nucleosome positioning, including DNase I hypersensitive sites sequencing (DNase-seq) [5], formaldehyde-assisted

*Corresponding authors.
E-mail: twang@wustl.edu (Wang T), bzhang29@wustl.edu (Zhang B).
[#]Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.
https://doi.org/10.1016/j.gpb.2020.06.025

isolation of regulatory elements with sequencing (FAIRE-seq) [6], micrococcal nuclease digestion with deep sequencing (MNase-seq) [7], and the recently developed Assay for Transposase-accessible Chromatin with high-throughput sequencing (ATAC-seq) [8]. ATAC-seq can detect the accessible regions of a genome by identifying open chromatin regions (OCRs) using a prokaryotic Tn5 transposase [8,9], and the technology features an easy experimental protocol, a reduced requirement of input material, and a high signal/noise ratio. These unique advantages have propelled ATAC-seq technology to quickly become a widely-used method to define chromatin accessibility, especially in several large consortiums focusing on functional genomics profiling, including ENCODE [10], TaRGET II [11], and IHEC [12].

The ATAC-seq analysis strategy is primarily adopted from ChIP-seq data analysis. After aligning sequencing reads to the genome, peak calling tools, such as MACS2 [13], are commonly used to identify highly enriched ATAC-seq signals across the genome. Unlike ChIP-seq, an ATAC-seq experiment does not normally require input control. Thus, accurately assessing the quality of ATAC-seq data is a critical step influencing downstream analysis. Several software packages were developed for ATAC-seq quality control (QC) and data analysis [14–16]. These tools provide general QC metrics of sequencing data, including read quality score, sequencing depth, duplication rate, and library insert fragment size distribution. Many tools also provide analysis functions, including footprinting analysis, motif analysis, and library complexity analysis.

Here, we present ATAC-seq Integrative Analysis Package (AIAP), a software package containing an optimized ATAC-seq data QC and analysis pipeline. Along with general QC metrics, such as library insert fragment size distribution, we specifically introduced a series of QC metrics for ATAC-seq, including reads under peak ratio (RUPr), background (BG), promoter enrichment (ProEn), subsampling enrichment (SubEn), and other measurements. By applying AIAP, we demonstrated a significant improvement in both peak calling and differential analysis by processing the paired-end sequencing data in single-end mode: more than 20% of ATAC-seq peaks can be identified using AIAP, and over 30% more differentially accessible regions (DARs) can be identified by AIAP in downstream analysis. We applied AIAP to reanalyze 70 mouse ENCODE [17] ATAC-seq datasets and determined the general QC recommendations for ATAC-seq data analysis. We also developed *qATACViewer*, a visualization tool included in AIAP, for user-friendly visualization of QC reports. AIAP is compiled into a Docker/Singularity image to allow maximized compatibility on different operating systems and computing platforms. The software, source code, and documentation are freely available at https://github.com/Zhang-lab/ATAC-seq_QC_analysis.

Method

Here, we describe AIAP for processing and analyzing ATAC-seq data. The AIAP workflow typically consists of four steps, as shown in **Figure 1**: 1) Data Processing; 2) QC; 3) Integrative Analysis; and 4) Data Visualization. Below,

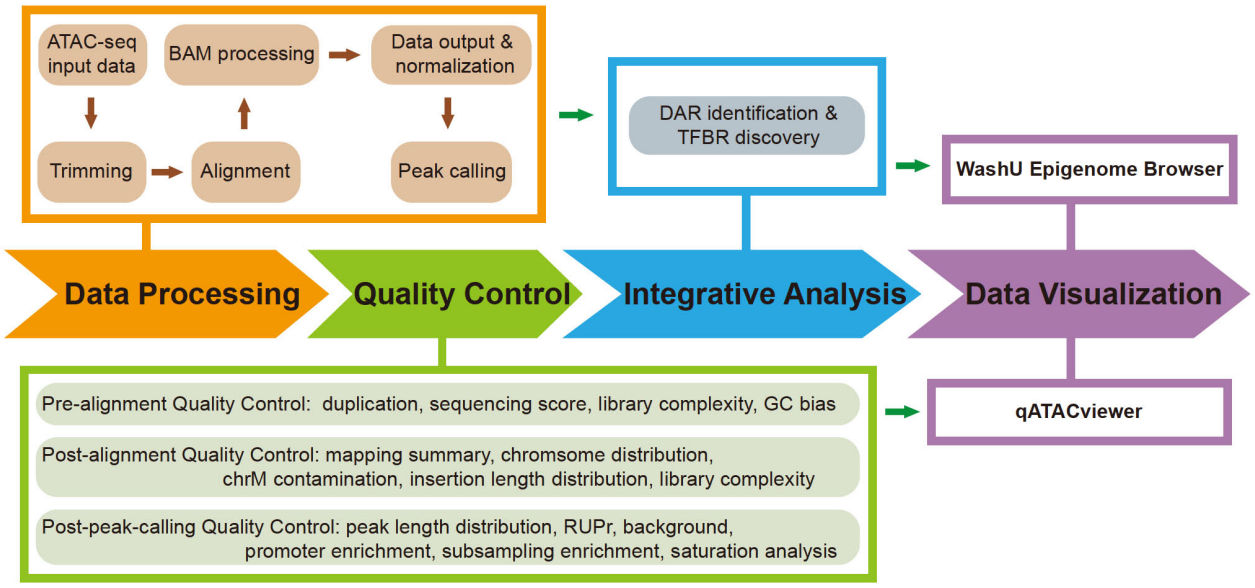

Figure 1 Schematic representation of AIAP
The schema reports the four analytical steps, namely, Data Processing, Quality Control, Integrative Analysis, and Data Visualization. AIAP, ATAC-seq Integrative Analysis Package; RUPr, reads under peak ratio; DAR, differentially accessible region; TFBR, transcription factor binding region.

we introduce the technical details of AIAP. Detailed documentation is available at https://github.com/Zhang-lab/ATAC-seq_QC_analysis/.

ATAC-seq data processing

The data processing step first configures the working path. The ATAC-seq paired-end (PE) raw-read FASTQ files are trimmed by *Cutadapt* and aligned to the reference genome by *BWA* [18]. The BAM file is further processed by *methylQA* [19] in the ATAC mode. The *methylQA* first filters unmapped and low-quality mapped PE reads and then identifies the Tn5 insertion position at each read end by shifting +4 bp/−5 bp on the positive/negative strands. *methylQA* further extends 75 bp in both directions around the Tn5 insertion position to create two pseudo single-end (SE) mapped reads with the length of 150 bp in PE as SE (PE-asSE) mode. Next, AIAP compiles different files for downstream analysis (.bed files) and normalized visualization (.bigWig files). The bed file is used to perform peak calling by MACS2 [13] with a q value cut-off of 0.01 and the following setting: *--keep-dup 1000 --nomodel --shift 0 --extsize 150*.

In PE-noShift mode, AIAP first filters unmapped and low-quality mapped PE reads and then identifies the Tn5 insertion position at each read end by shifting +4 bp/−5 bp on the positive/negative strands, then isolates the whole fragment between two insertion positions to create one PE mapped fragment for downstream analysis (.bed files) and normalized visualization (.bigWig files). The bed file is used to perform peak calling by MACS2 [13] with a q value cut-off of 0.01 and the following setting: *--keep-dup 1000 --nomodel --shift 0 --extsize 0*.

ATAC-seq data QC

AIAP performs a series of quality checking steps before and after alignment. AIAP calls *FastQC* to check the sequencing quality, duplication rate, and GC bias before alignment. After alignment, AIAP generates the mapping statistics summary, chromosome distribution of uniquely mapped reads, mitochondrial genome (chrM) contamination rate, library insert fragment size distribution, and library complexity. AIAP also performs a series of post-peak calling quality checks, including peak width distribution, RUPr, BG, ProEN, and SubEn. AIAP also provides saturation analysis, promoter peak distribution, and signal ranking analysis. AIAP reports the quality metrics in a JSON file, which can be visualized using *qATACViewer*. The default QC standard of AIAP is determined by the QC metrics of mouse ATAC-seq data generated by ENCODE consortium (Good standard: Mean; Acceptable standard: Mean−SD). We defined the key QC metrics below.

Alignment QC

There are 5 QC metrics for alignment. 1) Non-redundant uniquely mapped reads refer to the reads that are uniquely mapped to the reference genome after removing redundancy. 2) The second is chromosome distribution/chrM contamination rate. The former describes the distribution of uniquely mapped reads across all chromosomes, while for the latter, the number of uniquely mapped reads on chrM is used as a QC metric to measure the quality of the ATAC-seq library. 3) Library insert fragment size distribution is measured as the length of DNA fragment defined by non-redundant uniquely mapped read-pairs. 4) Library complexity is estimated in both duplication rate and predicted yield of distinct reads generated by *preseq* (https://github.com/smithlabcode/preseq). 5) To obtain the total number of useful single ends, each end of a non-redundant uniquely mapped read pair will be shifted +4 bp/−5 bp on the positive/negative strands and then further extended 75 bp in both directions around the Tn5 insertion position.

Peak calling QC

RUPr

RUPr is defined as the percentage of all useful ends (E_{total}) that fall into the called peak regions with at least 50% overlap (E_{under_peaks}). RUPr is calculated as follows.

$$RUPr = \frac{E_{under\_peaks}}{E_{total}} \quad (1)$$

Background

In total, 50,000 genomic regions (500 bp each) are randomly selected from the genome outside of ATAC-seq peaks. The ATAC-seq signal in each region is calculated as reads per kilobase per million mapped reads (RPKM). The percentage of all such regions with the ATAC-seq signal over the theoretical threshold (RPKM = 0.377) is considered high-background and used as a QC metric to indicate the background noise.

$$Background = \frac{number\_of\_regions\_with\_high\_background}{50,000} \quad (2)$$

ProEN

The promoters, the regions +/− 1 kb around transcription start sites (TSSs) of active genes, provide a positive control for OCRs. The ATAC-seq useful ends enriched on detected promoters (ATAC-seq peaks) are used as a QC metric to

measure the signal enrichment calculated as follows.

$$ProEn = \frac{E_{under\_promoter\_peaks} / Length_{total\_promoter\_peaks}}{E_{total} / Length_{genome}} \quad (3)$$

SubEn

The ATAC-seq signal (useful ends) enriched on the detected ATAC-seq peaks is used as a QC metric to measure the signal enrichment at the genome-wide level. To avoid sequencing-depth bias, 10 million useful ends are sampled from the complete dataset, and peak calling is performed to identify the OCRs. SubEn is calculated after 10 million pseudo counts are added into the calculation as background, which can avoid calculation failure caused by the low sequencing depth of testing ATAC-seq library.

$$SubEn = \frac{\frac{E_{under\_peaks}}{Length_{total\_peaks}} + \frac{10\,million}{Length_{genome}}}{(E_{under\_peaks} + 10\,million) / (Length_{genome} - Length_{total\_peaks})} \quad (4)$$

Saturation analysis

MACS2 is used to call narrow peaks for a series of subsampling from complete useful ends with a step of 10% of total sequencing depth. The length of identified peaks covering genomic regions at each subsampling are used to calculate the recovery (percentage) of complete peaks covering genomic regions when using complete useful ends.

Signal ranking analysis

The ATAC-seq peak signals are ranked, and the percentage of promoter peaks in each quantile is determined.

ATAC-seq data integrative analysis

AIAP includes two downstream analysis components: 1) analysis of DARs between two groups of samples and 2) discovery of transcription factor binding regions (TFBRs). AIAP calculates the read counts for all peaks identified under all conditions after peak calling, and a pair-wise comparison is performed by querying the *R* package *DESeq2* [20] based on the design table. AIAP will further identify potential TFBRs under ATAC-seq peaks by implementing the Wellington algorithm [21].

ATAC-seq data and QC report visualization

AIAP generates a collection of files for visualizing the ATAC-seq data on a genome browser [22–24], including the normalized signal density file (normalized to 10 million total reads) in bigwig format, the Tn5 insertion position file in bigwig format, the peak file in bed format, and the footprint position file in bed format. AIAP generates a JSON QC report that can be visualized with the embedded *qATACViewer* (Figure S1).

Calculation of DNase I hypersensitive sites and histone modification signal

The raw data FASTQ files for ATAC-seq and histone ChIP-seq were downloaded from ENCODE data portal (https://www.encodeproject.org/), and listed in Table S1. The ATAC-seq FASTQ files were processed by AIAP as described above. The ChIP-seq FASTQ files were aligned to the mouse genome (mm10 assembly) and were further processed by *methylQA*. Methylation calling of whole-genome bisulfite sequencing (WGBS) data was downloaded from the ENCODE data portal. The averaged signals of ATAC-seq, ChIP-seq, and WGBS were calculated at 100-bp windows within 5 kb around the center of the ATAC-seq peaks and were plotted in *R*. The processed DNase I hypersensitive site (DHS) data were downloaded from the ENCODE data portal.

DAR identification

The DARs of each tissue were identified between different mouse developmental stages, embryonic day 11.5 (E11.5) and postnatal day 0 (P0), to evaluate the performance of AIAP. The ATAC-seq peaks generated by AIAP were used as test regions, the read counts were calculated in both PE-asSE and PE-noShift modes, and the DARs were identified as described above with adjusted $P < 0.01$ and absolute $\log_2 FC > 1$. In PE-asSE mode, two Tn5 insertion events of one read-pair were considered independent of each other, and one read-pair was divided into two SE fragments to represent two Tn5 insertion events. In PE-noShift modes, one read-pair was used as one single fragment for downstream analysis.

Results

Defining the QC metrics of ATAC-seq data

Conducting QC checks at different steps of data processing and correctly interpreting QC metrics are crucial to ensure a successful and meaningful analysis. Different QC metrics report important information regarding different aspects of genomic data; thus, it is essential to define the key QC metrics for ATAC-seq data before performing an analysis. In addition to the traditional QC metrics shown in Figure 1, we specifically chose RUPr, BG, and ProEn as key QC metrics to measure the quality of ATAC-seq data. RUPr is an essential QC metric for ChIP-seq experiments [25] and is widely adopted to measure ATAC-seq data. The ENCODE

consortium recommends that at least 20% of non-redundant uniquely mapped reads be located in peak regions. A higher RUPr usually indicates a high signal-to-noise ratio. Similar to RUPr, a higher ProEn also indicates a high signal-to-noise ratio. ProEn is calculated to indicate the enrichment of the ATAC-seq signal over gene promoters, which are usually in open chromatin across different tissue and cell types. We used the ENCODE ATAC-seq data as a benchmark, and we determined that RUPr and ProEn directly reflect the quality of ATAC-seq at comparable sequencing depths (Figure 2A). The ATAC-seq peaks are sharper and stronger when the ATAC-seq data have high RUPr and ProEn values, suggesting better signal enrichment and better quality in the ATAC-seq experiments.

We further defined BG to directly measure the background noise level in the ATAC-seq experiments. We randomly selected 50,000 genomic regions (size: 500 bp) from regions of the genome that do not overlap with ATAC-seq peaks after peak calling. The ATAC-seq signals over each region are calculated as RPKM, and the regions with an ATAC-seq signal over a theoretical threshold (RPKM = 0.377) are considered high-background regions. The percentage of high-background regions within 50,000 randomly selected genomic regions is used as a QC metric to measure the background noise of the ATAC-seq data. We noticed that different background noise levels were directly

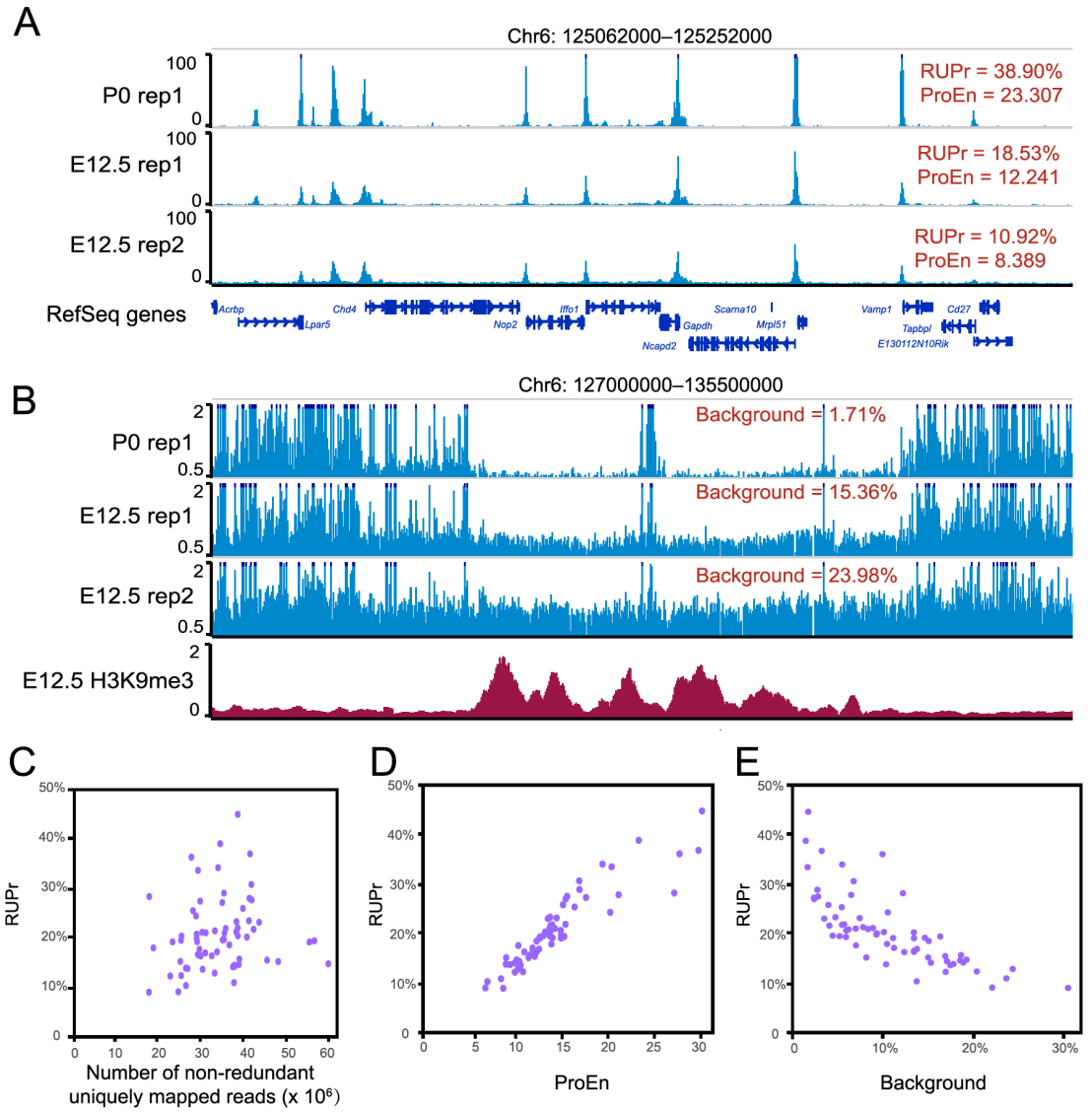

Figure 2 Key QC metrics of ATAC-seq data

A. Forebrain ATAC-seq datasets (sequencing depth normalized to 10 million reads) with distinct RUPr and ProEn (top to bottom: high to low) were visualized on the WashU Epigenome Browser. **B.** Forebrain ATAC-seq datasets (same as A) with distinct background values (top to bottom: low to high) in the heterochromatin regions highly enriched with repressive H3K9me3 histone modification in the mouse forebrain. **C.** Relationship between the number of non-redundant uniquely mapped reads and RUPr in ENOCDE mouse ATAC-seq datasets. **D.** Relationship between ProEn and RUPr in ENOCDE mouse ATAC-seq datasets. **E.** Relationship between background and RUPr in ENOCDE mouse ATAC-seq datasets. QC, quality control; RUPr, reads under peak ratio; ProEn, promoter enrichment; E12.5, embryonic day 12.5; P0, postnatal day 0.

reflected by the QC metric of background, especially in the heterochromatin regions, which are enriched with H3K9me3 signals (Figure 2B).

To further explore the QC metrics, we used AIAP to process 70 ATAC-seq datasets generated by the ENCODE consortium (Table S2). We specifically checked the key metrics, including RUPr, ProEn, SubEn, and BG. We noticed that these metrics were not dependent on sequencing depth (Figure 2C, Figure S2). RUPr was positively correlated with ProEn, proving an accurate measurement of signal enrichment in the ATAC-seq data (Figure 2D). BG was negatively correlated with RUPr, providing a measurement of the noise level in the ATAC-seq data (Figure 2E).

AIAP improves the sensitivity of discovering ATAC-seq peaks

To define ATAC-seq peaks, peak calling strategies adopted from ChIP-seq analysis are widely used to analyze ATAC-seq data. However, unlike ChIP-seq data, an ATAC-seq experiment does not have input control and is usually sequenced with the PE sequencing method to profile the size of DNA fragments. The uniquely aligned PE reads have been used to call open chromatin peaks after alignment in many studies [26–32]. After detecting the distribution of reads with different lengths under the peak regions, we noticed that the medium fragments and long fragments have similar distributions across the genome in ATAC-seq experiments (Table S3). Such evidence indicates that most of the captured fragment represents the open chromatin signal that can be derived from the Tn5 insertions in ATAC-seq experiments. To better represent the Tn5 insertion event, we shifted each end of the non-redundant uniquely mapped read pair +4 bp/−5 bp on the positive/negative strands to define the Tn5 insertion position and then further extended 75 bp in both directions around the Tn5 insertion position. By applying this strategy, one non-redundant uniquely mapped read pair is divided into two SE fragments [PE-asSE], and the sequencing depth doubles compared to that of traditional analysis methods that manage PE fragments without shifting (PE-noShift) (**Figure 3**A).

To validate the sensitivity of our analysis strategy, we downloaded published ATAC-seq data of GM12878 cells generated by the Greenleaf laboratory with the Omni-ATAC-seq protocol [31]. We first processed the data by following the classical method based on non-redundant uniquely mapped PE reads (PE-noShift) and performed peak calling. In PE-noShift mode, we identified 92,058 narrow peaks. In parallel, we applied AIAP to process the same data in PE-asSE mode and performed peak calling with identical parameters (see Method), and 112,848 peaks were identified. Compared to PE-noShift mode, PE-asSE mode reported ~ 99.9% of PE-noShift peaks and identified ~ 23% additional peaks (20,918) (Figure 3B). By visually inspecting the signal density on the genome browser, we noticed that most of the PE-asSE-specific peaks overlapped with known ENCODE DHSs (Figure 3C). We examined the ATAC-seq peaks and known GM12878 DHSs obtained from the ENCODE data portal. We noticed that PE-asSE mode identified more ATAC-seq peaks at different sequencing depths (Figure 3D). When analyzing the full GM12878 Omni-ATAC-seq dataset, PE-noShift mode identified ~ 80% of DHSs, and PE-asSE mode identified ~ 85% of DHSs (Figure 3E). We further used merged DHSs of 95 cell lines to measure the specificity of identified ATAC-seq peaks. We found that nearly 98% of common peaks identified by both PE-asSE and PE-noShift modes overlapped with known DHSs. A total of 19,040 out of 20,918 peaks identified only by PE-asSE mode overlapped with known DHSs, and 112 out of 128 peaks identified only by PE-noShift mode overlapped with known DHSs (Figure 3F). We also used DHS data to estimate the Type-I and Type-II errors of AIAP. The ATAC-seq peaks that cannot be validated by known DHSs are considered as potential false positive. In total 3575 out of 112,848 ATAC-seq peaks in AIAP PE-asSE mode and 1713 out of 92,058 ATAC-seq peaks in AIAP PE-noShift mode are considered as potential Type-I errors (false discovery rate: 3.17% and 1.86%, respectively). However, considering the relatively lower sensitivity of DNase-seq when comparing to ATAC-seq assay, we believe the Type-I error calculated here was overestimated. We further defined the 39,205 GM12878 DHSs that were commonly identified in two independent replicates as true positive DHSs. AIAP PE-asSE can identify 38,078 of them, and 1127 DHSs were considered as negative (Type-II errors) with a false negative rate of 2.87%. Meanwhile, AIAP PE-noShift mode predicted 1827 DHSs as negative (Type-II errors) with a false negative rate of 4.66%. These results suggest that PE-asSE mode can greatly improve the sensitivity of the OCR discovery.

ATAC-seq peaks are generally considered regulatory elements that are enriched for specific histone modifications. We downloaded ChIP-seq data of GM12878 histone modifications (H3K4me3, H3K4me1, and H3K27ac) that are associated with promoter and enhancer activities to validate the functionality of PE-asSE-specific ATAC-seq peaks. Compared to randomly selected genomic regions, the PE-asSE-specific ATAC-seq peaks were highly enriched in all active histone modifications (Figure 3G–I). These results suggest that the PE-asSE-specific ATAC-seq peaks are functional regulatory elements rather than false positives. We further utilized AIAP to analyze the ATAC-seq data of multiple tissues. Compared to the classic PE-noShift mode, the PE-asSE mode resulted in a 28%–55% increase in sensitivity when processing the ATAC-seq data (Table S4). These results indicate that AIAP can dramatically enhance

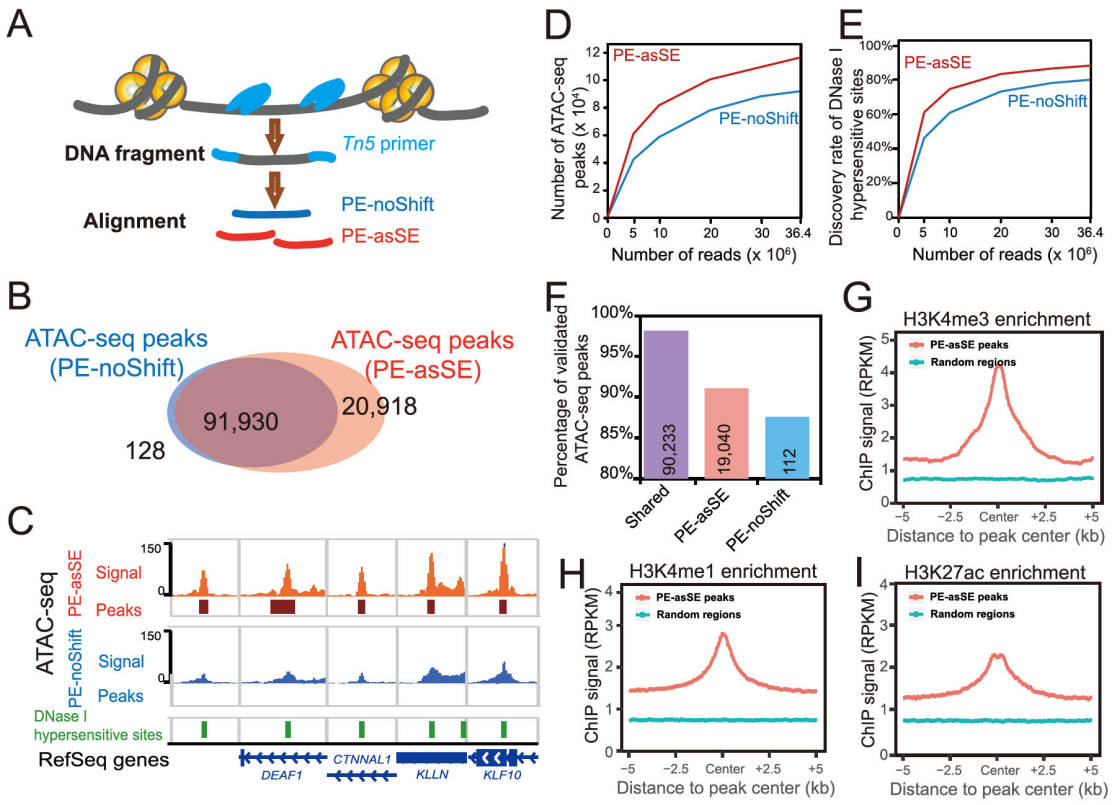

Figure 3 AIAP data processing strategy (PE-asSE) performs better than classic method (PE-noShift) in OCR identification
A. Schematic representation of PE-noShift and PE-asSE data processing strategy. PE-noShift representing the one complete DNA fragment insertion was generated as in previous studies [26–32]. PE-asSE representing two DNA fragment insertions was generated by AIAP. **B.** Venn diagram showing ATAC-seq peaks identified with PE-noShift and PE-asSE in the GM12878 cell line. **C.** ATAC-seq data of GM12878 cell line were independently processed as PE-asSE (red) and PE-noShift (blue), and visualized together with DHSs of GM12878 cell line (green) on the WashU Epigenome Browser. Five regions were randomly selected that contain the peaks identified only in PE-asSE mode but not in PE-noShift mode. **D.** Number of ATAC-seq peaks identified with PE-noShift mode and PE-asSE mode at different sequencing depths by randomly sampling the Omni-ATAC-seq data of GM12878 cell line [31]. **E.** Discovery rate of DHSs of GM12878 cell line by PE-asSE and PE-noShift at different sequencing depths (randomly sampling same as in D). **F.** Percentage of shared, PE-asSE-specific, and PE-noShift-specific ATAC-seq peaks validated by DHSs of GM12878 cell line identified by ENCODE. PE-asSE-specific ATAC-seq peaks were highly enriched with active histone modifications H3K4me3 (**G**), H3K4me1 (**H**), and H3K27ac (**I**), when compared to 20,000 regions randomly selected from the genome (green). DHS, DNase I hypersensitive site.

the sensitivity of OCR discovery with high specificity.

AIAP improves the sensitivity of DAR identification

Chromatin accessibility is dynamically associated with cellular responses to developmental cues, disease progression, and environmental stimuli. The identification of DARs has become an important approach to monitor the activity changes of regulatory elements [33]. Since AIAP dramatically increased the sensitivity of OCR discovery, we further tested the sensitivity of AIAP in identifying DARs. We downloaded ATAC-seq dataset of mouse liver E11.5 and P0 stages from the ENCODE data portal and processed these data in both PE-asSE and PE-noShift modes. As expected, PE-asSE mode identified 30% more ATAC-seq peaks than PE-noShift mode (Table S4). To test the sensitivity of DAR identification, we used the complete set of ATAC-seq peaks identified in PE-asSE mode and calculated the read counts based on both PE-asSE and PE-noShift modes (see Method). A total of 11,040 E11.5-specific and 9584 P0-specific DARs were identified by both modes (shared DARs). We also identified 4213 E11.5-specific and 2819 P0-specific DARs by using only PE-asSE mode (**Figure 4**A). Correspondingly, only 107 E11.5-specific and 72 P0-specific DARs were found by using only PE-noShift mode. Compared to PE-noShift mode, PE-asSE mode resulted in an ~ 35% increase in the number of DARs identified. We further tested AIAP on other tissues at two developmental stages and found that AIAP identified 32%–168% more DARs in different tissues (Table S5).

We examined the genomic distribution of DARs and noticed that the distribution of PE-asSE-specific DARs had a similar distribution as shared DARs, that is, most DARs were located in intergenic and intronic regions, which is consistent with their potential enhancer functionality (Figure 4B). Because dynamic changes in chromatin accessibility accompany the alteration in epigenetic modification synchronously [1,8,28,34], we further used epigenetic data

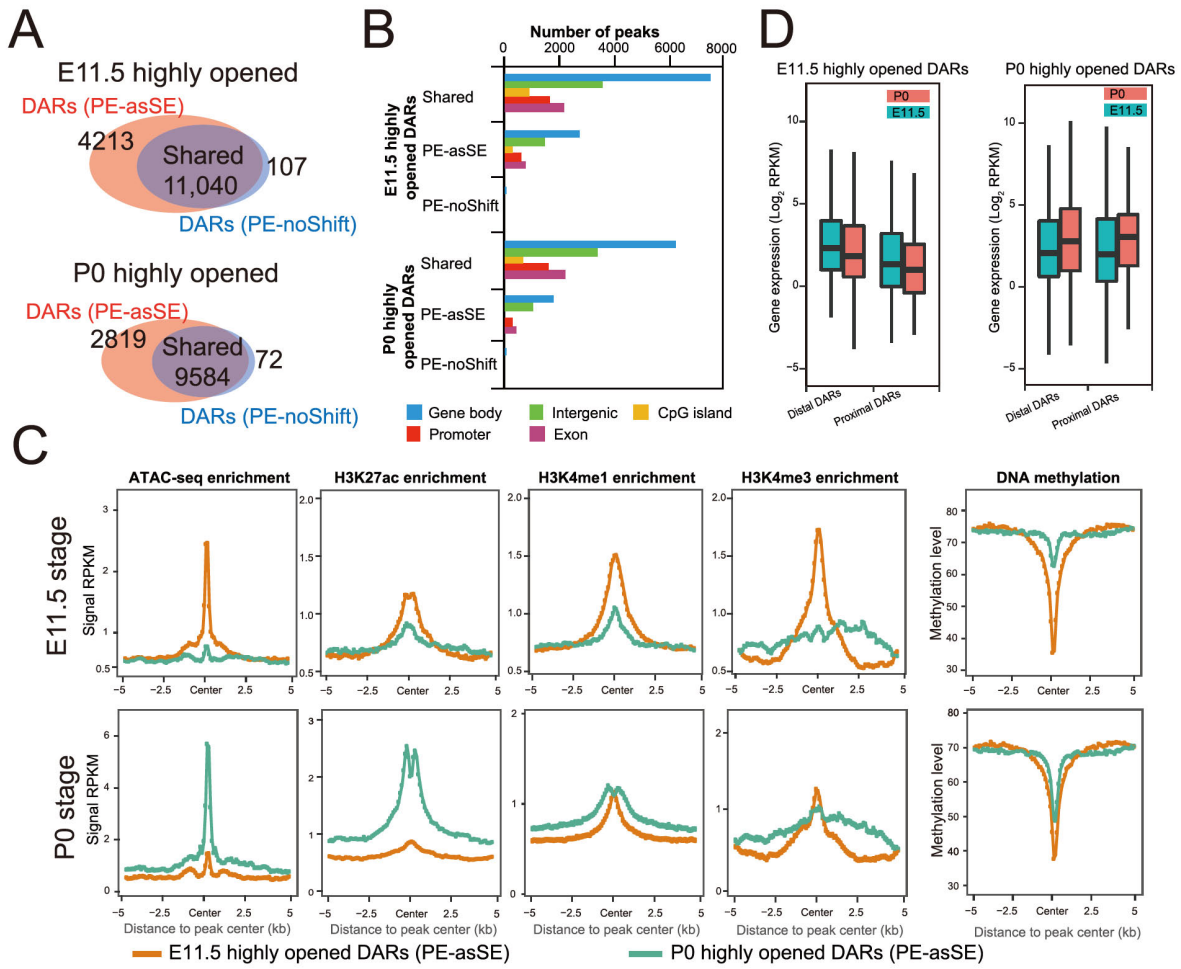

Figure 4 AIAP data processing strategy (PE-asSE) performs better than the classic method (PE-noShift) in DAR identification
A. Venn diagram showing DARs identified with PE-noShift mode and PE-asSE mode when comparing ATAC-seq data collected from mouse liver at E11.5 and those from P0. **B.** Genomic distribution of 20,624 shared, 7032 PE-asSE-specific, and 179 PE-noShift-specific DARs. **C.** Enriched epigenetic modifications (left to right: ATAC-seq, H3K27ac ChIP-seq, H3K4me1 ChIP-seq, H3K4me3 ChIP-seq, and DNA methylation) on the PE-asSE-specific DARs from mouse liver at E11.5 (top) and P0 (bottom). **D.** The expression of genes associated with PE-asSE-specific E11.5 DARs (left) and P0 DARs (right) in proximal (2 kb around TSSs) and distal (2–20 kb around TSSs) at the two developmental stages. TSS, transcription start site.

from the same samples to validate the accuracy of the DARs. We first checked the epigenetic modifications around PE-asSE-specific DARs. Compared to P0-specific PE-asSE-specific DARs, the E11.5-specific DARs recruited highly active histone modifications associated with regulatory elements specifically at the E11.5 stage but not the P0 stage, including H3K27ac, H3K4me1, and H3K4me3 (Figure 4C, top). In contrast, the P0-specific PE-asSE-specific DARs recruited highly active histone modification H3K27ac specifically at the P0 stage but not H3K4me1 or H3K4me3 modification (Figure 4C, bottom). We also noticed that the E11.5-specific DARs were much less methylated at the E11.5 stage, but P0-specific DARs remained methylated at the E11.5 stage. However, E11.5-specific DARs were still un-methylated at the later postnatal stage. The observation of the loss of DNA methylation on regulatory elements during embryo development is consistent with that of a previous study [35]. The epigenetic modifications of shared DARs showed very similar patterns to the PE-asSE-specific DARs (Figure S3). We further examined the expression of genes around the identified DARs. The DARs were assigned to the nearest gene based on distance and were classified into a proximal group (2 kb around the TSS) and a distal group (2–20 kb around the TSS). We noticed that the expression of genes around PE-asSE embryonic DARs was downregulated during liver development. In contrast, the expression of genes around postnatal DARs was upregulated at the same time (Figure 4D).

Discussion

AIAP is a new tool to perform quality assurance and downstream analysis of ATAC-seq data. Comparing with other tools (Table S6), AIAP provides a rapid and reliable data processing solution for ATAC-seq data. Using public

datasets, we systematically tested the QC metrics of ATAC-seq data and established key QC metrics of ATAC-seq data. We determined that RUPr, ProEn, and BG were important measurements to estimate the quality of ATAC-seq data. All three QC metrics directly reflect the quality of library preparation, and the failure of these QC metrics reflect the low-quality of the ATAC-seq data, which cannot be corrected by merely increasing sequencing depth. We found that the RUPr and ProEn reflect the ATAC-seq signal enrichment, and BG indicates the overall background noise of the data. By combining these QC metrics, we obtained an accurate estimation of the quality of ATAC-seq data. We used AIAP to process 54 mouse ATAC-seq datasets to test and evaluate the QC metrics and generate the range of the QC metrics (Table S2). These ranges of QC metrics can be used as a reference to evaluate the success of ATAC-seq experiments.

We optimized the widely used classic analysis methodology and specifically used PE-asSE mode to process the PE sequenced ATAC-seq data. AIAP aligns the PE ATAC-seq data in PE mode to increase the alignment accuracy, and the BAM file is further processed in SE mode for downstream analysis. In PE-asSE mode, AIAP doubles the sequencing depth and dramatically increases the sensitivity of OCR identification. In our test, AIAP identified 20%–40% more ATAC-seq peaks than the widely used classic analysis methods. We further used corresponding DHS data and histone modification data to validate the specificity of newly identified ATAC-seq peaks by AIAP; most of the novel ATAC-seq peaks identified by AIAP were independently validated with DHSs and enriched for active histone modifications. Such a result indicates the high true positive rate resulting from the AIAP analysis strategy.

We also suggest that the PE-asSE strategy can improve the sensitivity of discovering DARs, which are wildly used to measure chromatin dynamics [36]. By using ENCODE ATAC-seq data of liver embryo development, we found that AIAP can identify 32%–168% more chromatin DARs than the classic PE-noShift mode. We further indicated that the novel DARs identified by AIAP were enriched in active histone modifications at different developmental stages. The E14.5-specific DARs were lowly methylated, and the H3K27ac signals were significantly enriched only in the E14.5 stage but not in the P0 stage. In contrast, the P0-specific DARs were highly methylated in the E14.5 stage without active histone modifications and became minimally methylated and recruited strong active H3K27ac signals in the P0 stage. We also noticed that the expression of genes around the developmental stage-specific DARs was associated with the openness of DARs, as other studies reported [37,38]. These results suggest that AIAP can greatly improve the sensitivity to identify DARs with high specificity.

Finally, we compiled AIAP into a Docker/Singularity image to facilitate the easy operation of AIAP on high-performance computing clusters. AIAP can complete QC checking and file processing in ~ 2 h for one typical ATAC-seq dataset, of 37 million PE reads (Table S7). Besides the high performance, AIAP also has a much better sensitivity and higher specificity (Figure 3F) when comparing to other tools, including ENCODE pipeline (Table S8). AIAP supports multiple genome assemblies, including human (hg19 and hg38), mouse (mm9 and mm10), and rat (rn6). Additionally, each step for QC and data processing is componentized and can be called by advanced users to build pipelines for specialized applications, and different genome assemblies can be easily and directly compiled for ATAC-seq data processing in other species.

Code availability

The software, source code, and documentation of AIAP are freely available at https://github.com/Zhang-lab/ATAC-seq_QC_analysis.

CRediT author statement

Shaopeng Liu: Methodology, Software, Formal analysis, Resources, Writing - original draft. **Daofeng Li:** Methodology, Software, Visualization. **Cheng Lyu:** Methodology, Software, Formal analysis. **Paul M. Gontarz:** Formal analysis, Resources. **Benpeng Miao:** Investigation, Software, Formal analysis. **Pamela A.F. Madden:** Supervision, Project administration. **Ting Wang:** Supervision, Project administration, Writing - original draft, Writing - review & editing. **Bo Zhang:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Supervision, Project administration, Writing - original draft, Writing - review & editing, Visualization. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Institutes of Health (Grant Nos. U24ES026699, U01HG009391, and R25DA027995), the Goldman Sachs Philanthropy Fund (Emerson Collective), and Chan Zuckerberg Initiative, United States.

Supplementary material

Supplementary data to this article can be found online at

https://doi.org/10.1016/j.gpb.2020.06.025.

ORCID

0000-0003-3112-4068 (Shaopeng Liu)
0000-0001-7492-3703 (Daofeng Li)
0000-0002-0345-8143 (Cheng Lyu)
0000-0001-5115-7479 (Paul M. Gontarz)
0000-0002-2070-2339 (Benpeng Miao)
0000-0001-8987-7439 (Pamela A.F. Madden)
0000-0002-6800-242X (Ting Wang)
0000-0003-2962-5314 (Bo Zhang)

References

[1] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.
[2] Lee HJ, Lowdon RF, Maricque B, Zhang B, Stevens M, Li D, et al. Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. Nat Commun 2015;6:6315.
[3] Lowdon RF, Zhang B, Bilenky M, Mauro T, Li D, Gascard P, et al. Regulatory network decoded from epigenomes of surface ectoderm-derived cell types. Nat Commun 2014;5:5442.
[4] Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res 2013;23:1522–40.
[5] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010;2010:pdb.prot5384.
[6] Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc 2012;7:256–67.
[7] Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, et al. Dynamic regulation of nucleosome positioning in the human genome. Cell 2008;132:887–98.
[8] 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. Nat Methods 2013;10:1213–8.
[9] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 2015;523:486–90.
[10] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[11] Wang T, Pehrsson EC, Purushotham D, Li D, Zhuo X, Zhang B, et al. The NIEHS TaRGET II Consortium and environmental epigenomics. Nat Biotechnol 2018;36:225–7.
[12] Stunnenberg HG, International Human Epigenome Consortium, Hirst M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. Cell 2016;167:1897.
[13] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9:R137.
[14] Pranzatelli TJF, Michael DG, Chiorini JA. ATAC2GRN: optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference. BMC Genomics 2018;19:563.
[15] Wei Z, Zhang W, Fang H, Li Y, Wang X. esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis. Bioinformatics 2018;34:2664–5.
[16] Ou J, Liu H, Yu J, Kelliher MA, Castilla LH, Lawson ND, et al. ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. BMC Genomics 2018;19:169.
[17] Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 2014;515:355–64.
[18] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–60.
[19] Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods 2015;72:29–40.
[20] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.
[21] Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res 2013;41:e201.
[22] Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, et al. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. Nat Biotechnol 2015;33:345–6.
[23] Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The human epigenome browser at Washington University. Nat Methods 2011;8:989–90.
[24] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006.
[25] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22:1813–31.
[26] Ackermann AM, Wang Z, Schug J, Naji A, Kaestner KH. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol Metab 2016;5:233–44.
[27] Hay D, Hughes JR, Babbs C, Davies JOJ, Graham BJ, Hanssen LLP, et al. Genetic dissection of the α-globin super-enhancer *in vivo*. Nat Genet 2016;48:895–903.
[28] Mazumdar C, Shen Y, Xavy S, Zhao F, Reinisch A, Li R, et al. Leukemia-associated cohesin mutants dominantly enforce stem cell programs and impair human hematopoietic progenitor differentiation. Cell Stem Cell 2015;17:675–88.
[29] Atianand MK, Hu W, Satpathy AT, Shen Y, Ricci EP, Alvarez-Dominguez JR, et al. A long noncoding RNA lincRNA-EPS acts as a transcriptional brake to restrain inflammation. Cell 2016;165:1672–85.
[30] Shih HY, Sciumè G, Mikami Y, Guo L, Sun HW, Brooks SR, et al. Developmental acquisition of regulomes underlies innate lymphoid cell functionality. Cell 2016;165:1120–33.
[31] Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 2017;14:959–62.
[32] de Dieuleveult M, Yen K, Hmitou I, Depaux A, Boussouar F, Bou Dargham D, et al. Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. Nature 2016;530:113–6.
[33] Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol 2020;21:255.
[34] Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, et al. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res 2011;21:1659–71.
[35] Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, et al. Transcription factor binding dynamics during human ES cell differentiation. Nature 2015;518:344–9.
[36] Toraih EA, El-Wazir A, Ageeli EA, Hussein MH, Eltoukhy MM, Killackey MT, et al. Unleash multifunctional role of long non-

coding RNAs biomarker panel in breast cancer: a predictor classification model. Epigenomics 2020;12:1215–37.

[37] Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci U S A 2017;114:E4914–23.

[38] Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type–specific gene expression from regions of open chromatin. Genome Res 2012;22:1711–22.