



OPEN

MESIA: multi-epigenome sample integration approach for precise peak calling

Seung Gwa Park^{1,2,3}, Woo-Jin Kim^{1,2,3}, Jae-I Moon^{1,2}, Ki-Tae Kim^{1,2,3}✉ & Hyun-Mo Ryou^{1,2,3}✉

The assay for transposase-accessible chromatin with sequencing (ATAC-seq) is the most widely used method for measuring chromatin accessibility. Researchers have included multi-sample replication in ATAC-seq experimental designs. In epigenomic analysis, researchers should measure subtle changes in the peak by considering the read depth of individual samples. It is important to determine whether the peaks of each replication have an integrative meaning for the region of interest observed during multi-sample integration. We developed multi-epigenome sample integration approach for precise peak calling (MESIA), which integrates replication with high representativeness and reproducibility in multi-sample replication and determines the optimal peak. After identifying the reproducibility between all replications, our method integrated multiple samples determined as representative replicates. MESIA detected 6.06 times more peaks, and the value of the peaks was 1.32 times higher than the previously used method. MESIA is a shell-script-based open-source code that provides researchers involved in the epigenome with comprehensive insights.

Next-generation sequencing (NGS) growth has paved the way for epigenetics analysis of post-transcriptional changes and various related analysis methods have been developed. Recently, many studies have shown higher quality results by combining¹ transcriptomic analysis through RNA-seq² and chromatin accessibility analysis^{3–5}.

Among the many methods for measuring chromatin accessibility, the assay for transposase-accessible chromatin with sequencing (ATAC-seq), which requires relatively little time and a smaller amount of samples for analysis, is widely used⁴. Researchers supplement the lack of consistency by including multi-sample replication in experimental designs. In this process, it is important to integrate the information from biological replicates. In epigenome analyses, including ATAC-seq, which require the extraction of peaks from individual samples to measure subtle changes in specific regions, require integrated analysis to determine whether the results of the measured replication are consistent. Furthermore, the measurement of peak epigenome levels via NGS is accompanied by the process of extracting significant signals compared to background signals using complex algorithms, such as hidden Markov models⁶. Therefore, it is important to determine an optimal peak within a region of interest to ensure the reliability of the research results by measuring the similarity of individual samples rather than by simple merging.

Currently, methods used for ATAC-seq replication integration can be divided into two categories. One is the length overlap-based approach and the other is the statistical-based approach⁷. The length overlap-based approach is a method of determining the same peak based on the degree of overlap between the peaks of each replication. On the other hand, a statistical-based approach uses the reliable boundaries of reproducibility through statistical parameters to determine the reproducibility of peaks. In the case of the length-overlap-based approach, integration is possible regardless of the number of replications by focusing on speed and convenience; however, because reproducibility is derived based on the degree of overlap in length, the process of confirming statistical significance is not carried out. In the case of a statistical-based approach, replication integration is performed after determining the presentation and reproducibility based on the statistical background; however, only two replications are possible⁷. That means, no statistical background-based methods that are not significantly constrained by the number of replicates exist.

Here, we developed a novel replication integration method called Multi-Epigenome Sample Integration Approach for Precise Peak Calling (MESIA). MESIA determines replication with high representativeness and

¹Department of Molecular Genetics & Dental Pharmacology, School of Dentistry and Dental Multiomics Center, Dental Research Institute, Seoul National University, Seoul, South Korea. ²Epigenetic Regulation of Aged Skeleto-Muscular System Laboratory, School of Dentistry and Dental Research Institute, Seoul National University, Seoul, South Korea. ³These authors contributed equally: Seung Gwa Park and Woo-Jin Kim. ✉email: kিতae@snu.ac.kr; hmryoo@snu.ac.kr

reproducibility in multi-sample replication and then integrates them to extract more robust and less likely false positive peaks. For researchers whose work may have been difficult to analyze due to a lack of computational background, a shell-script-based open-source code was disclosed to GitHub (<https://github.com/ERASMUSlab/MESIA>).

Materials and methods

Assay for transposase-accessible chromatin using ATAC-seq workflow

The GM12878 human lymphoblastoid cell line⁴ was mainly used to verify the pipeline for ATAC-seq (Fig. 1A), and mouse embryonic fibroblasts⁸ (MEFs) were also used to verify the pipeline on other datasets. Sequence trimming was performed using Cutadapt⁹ and Trimmomatic¹⁰. Subsequently, the filtered fastq was aligned to the hg19 genome assembly for GM12878 and mm10 for MEFs using Bowtie2¹¹. The generated BAM file was sorted, filtered by MAPQ quality, and PCR duplicates and mitochondrial reads were removed using Samtools¹², and multi-mapping control was performed using assign_multimappers.py¹³, provided by Encode. Finally, the filtered BAM file was divided by histone status using Sambamba and bedpeTn5shift¹⁴. MESIA takes BAM files as input and uses MACS2¹⁵(v2.2.7.1) for peak calling. Through a series of processes, MESIA determines and integrates the reproducibility of multi-epigenomic samples. (Fig. 1B).

RNA-sequencing data processing

RNA-seq data obtained from two biological replicates were analyzed on the GM12878 human lymphoblastoid cell line⁴ and three biological replicates were analyzed on the MEFs⁸. The median read length was 101 bp. The reads were aligned hg19 and mm10 using Bowtie2¹¹(v2.2.5), Samtools¹²(v1.13), Bamtools¹⁶(v2.5.1), Bedtools¹⁷(v2.30.0), Biobambam¹⁸(v2.0.183), Cutadapt⁹(v4.0), and Sambamba¹⁹(v0.8.2). Expression calling was performed using Salmon²⁰(v1.7.0) and Kallisto²¹(v0.46.2).

Multi-epigenome sample integration approach for precise peak calling (MESIA) workflow and pseudocode

MESIA is a method of selecting only reproducible replications in consideration of representativeness in multi-replication ATAC-seq and then deriving an optimal peak within replications (Fig. 2).

$$R_{ij} = \begin{cases} 1 & \left(Nt_{ij}/Np_{ij} \leq 2 \right) \\ 0 & \left(Nt_{ij}/Np_{ij} > 2 \right) \end{cases} \quad (1)$$

$$SubS = \begin{cases} 1 & \left(Rep - pseudo\ rep1/rep2 \leq 2 \right) \\ 0 & \left(Rep - pseudo\ rep1/rep2 > 2 \right) \end{cases} \quad (2)$$

$$MS1pass_{ij} = P(R_{ij} = 1 | (SubS_i + SubS_j = 2)) \quad (3)$$

Subsequently, it goes through MESIA stage 1 based on the IDR²². MESIA stage 1 is process of determining whether two replications are reproducible to each other, indicating that Nt called “peaks consistent between true replicates” and Np called “peaks consistent between poured pseudoreps” are less than twice that of each other. Furthermore, an analysis is conducted to examine potential biases within the sample by generating pseudo-replicates for each replicate through random read splitting. These conditions are reflected in constraints (1) and (2), where the variables i and j can range from 1 to the number of replications being considered as inputs. Constraint (3), MS1pass, refers to the replication set that passed MESIA Stage 1, which is the process of deriving a replication set with reproducibility based on the previous analysis.

$$SigBP_{ij} = \int_{BPpp[(n+1)/2]}^{BPpp[n]} Bpd(BPpp) \cap \int_{BPqv[m]}^{BPqv[(m+1)/2]} Bpd(BPqv) \quad (4)$$

$$SigSP = \int_{SPpp[(n+1)/2]}^{SPpp[n]} SPd(SPpp) \cap \int_{SPqv[m]}^{SPqv[(m+1)/2]} SPd(SPqv) \quad (5)$$

$$MS2pass_{ij} = SigBP_{ij} \cup SigSP_i \cup SigSP_j \quad (6)$$

MESIA stage 2 uses Nt and Np of the replication set with the reproducibility selected in MESIA stage 1. The intersection of Nt and Np is called the base peak of the replication set, and MESIA stage 2, which passes the multi-replication peak, is obtained using the base peak. The base peak is labeled as BP in constraint (4). Similarly, within the MESIA Stage 1 passed replication set, the peak of a single replication is denoted as SP in constraint (5). High-ranking standards for calculating constraint (4, 5, 6) include the peak point, the peak point, distance from the start site to the highest signal site of the peak, and the Q value, which indicates statistical significance. Since base peak have already passed MESIA stage 1, the significant criteria for peak point and Q values were set to the medians of the base peak. This process is described in constraints (4) and (5), where the variable n represents the position of the largest value when sorting the peak points in descending order, and similarly, the variable m represents the position of the smallest value when sorting the q values in ascending order. Using these

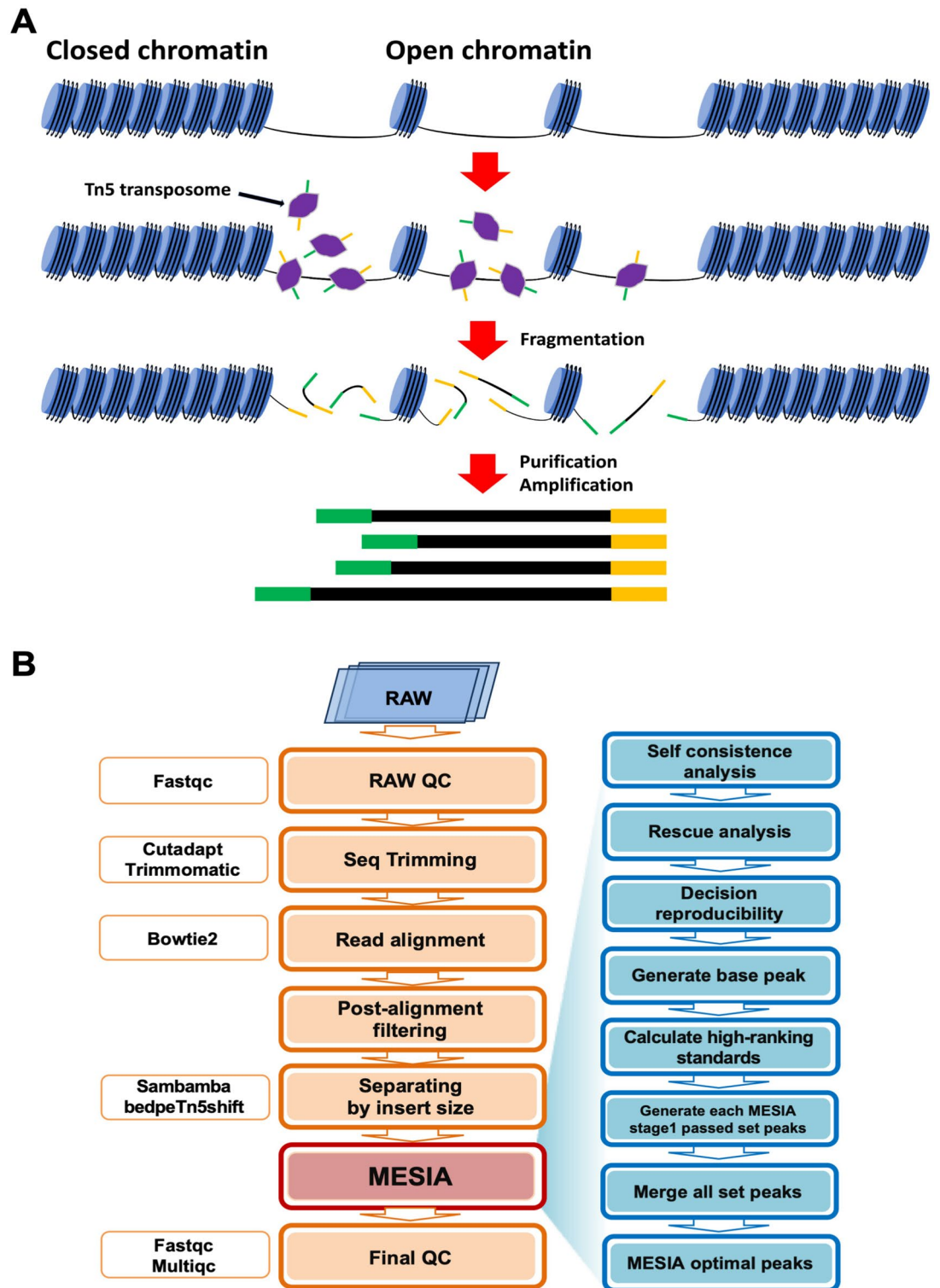


Figure 1. Assay for transposase-accessible chromatin using ATAC-seq workflow. (A) Basic concept and principle of ATAC-seq. (B) Generalized ATAC-seq preprocessing pipeline. MESIA is a tool that takes a BAM file as input and determines and integrates the reproducibility of replication's peak. It first checks the bias within the replication through self-consistency analysis. It then checks the reproducibility between replications through rescue analysis. Finally, it generates optimal peaks based on the Q value and the peak point of the base peak. It repeats this process for each replication and merges the results to produce MESIA optimal peaks.

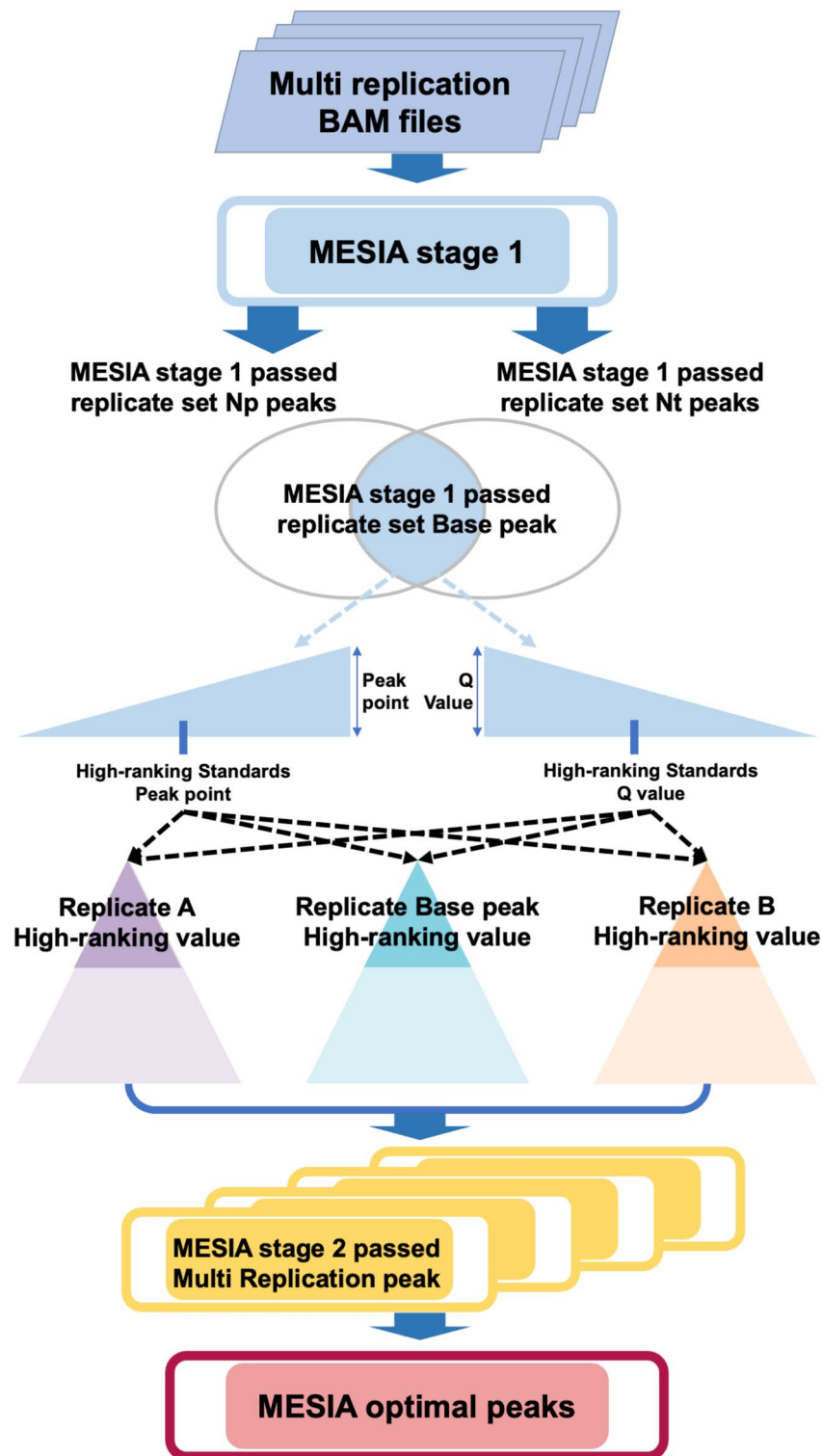


Figure 2. Multi-Epigenome Sample Integration Approach for precise peak calling (MESIA) workflow. MESIA is executed using a two-stage strategy, where stage1 determines the reproducibility of multi-epigenome samples, and stage2 merges reproducible replications. A hypothetical replication A and replication B were grouped together to represent a replication set. If the replication set passes MESIA stage 1, the base peak of the replication set is derived and the replication set base peak, replication A, and replication B are filtered using the median of the peak point and Q value of the base peak. The high-ranking peak sets obtained in this way are merged to produce MESIA stage 2 passed multi replication peak. After repeating this process with several other replication sets and merging them, MESIA optimal peak is produced.

high-ranking standards, we derived a robust peak from the two replicates in the replication set. Constraint (6), MS2pass, refers to the MESIA Stage 2 passed peak.

$$MESIA\ optimal = \sum_{i,j=1}^{i,j=max} MS2pass_{ij} \quad (7)$$

Finally, the MESIA optimal peaks are derived by merging the robust peaks of all replications sets that have passed through MESIA Stage 2. This process is described in constraint (7). Each step mentioned above are summarized into pseudocode (Fig. 3). MESIA stage 1 consists of self-consistency and rescue analysis, which are represented by constraints (2) and (1), respectively. These constraints detect bias within each replication and assess reproducibility between different replications. The base peaks of the replications that pass MESIA stage 1 are found using constraint (3). MESIA stage 2 is the stage that creates optimal peaks using the base peaks from MESIA stage 1. The process of filtering peaks using the peak point and Q value of the base peaks is represented by constraints (4–5). The process of combining these peaks to create optimal peaks is represented by constraints (6–7).

Comparative algorithms related to length-based approaches and statistical-based approaches

Other methods were used for comparison to measure the performance of MESIA. First, the method representing the length overlap-based approach uses NaiveOverlapBroad²³ (NAIVE) (Fig. 4A). NAIVE is a method of pooling all replications to be integrated and then selecting only peaks that overlap by more than half in each replication. This method is not limited to the number of replicates; however, there is a risk of integration between replicates that lack statistical evidence and reproducibility. Statistical-based approaches include IDR²⁴ and ChIP-R²⁵. IDR is a statistical method that is widely used in two-replication integrations and considers the reproducibility between replications; however, it has the disadvantage of not being able to perform multi-sample integration. ChIP-R can perform multi-sample integration, but it does not consider the reproducibility between replications and has the disadvantage of not being robust due to overlapping replication parameters selected by the user. Finally, the non-overlapping maximum signal peak method²⁶ (NOMS) (Fig. 4B) is an integration of the concept of a combination that could maximize the scope of the optimal peak. In terms of stringency, ChIP-R, a statistical-based approach, used several default statistical values as thresholds. The ranking method was *p*-value, and the cut-off value for the set of reproducible peaks was 0.05. Length overlap-based approaches (NAIVE and NOMS, Fig. 4) did not use statistical values. NAIVE selected peaks that overlapped by more than half in each replication, while NOMS integrated the concept of a combination to improve accuracy.

Compiling a basic simulation sample set

We needed a simulation sample with 7 characteristics to perform this simulation. Simulation samples A and B are simulation samples that are not reproducible with each other. Using the GM12878 replication1 ATAC narrowPeaks file, samples A and B with different trends were created. First, adjacent peaks were combined to form small set. Subsequently, the adjacent peaks were randomly sampled and divided into three large set. Select two large set out of three, and then divide the adjacent peak back into a narrowPeaks form. No matter how different the samples tend to be, there are basically shared peaks which were called body peaks. On the other hand, there must also be peaks that distinguish A and B, which we call function peaks. The body peaks and function peaks are combined to form A and B (Fig. 5). The reproducibility of samples A and B produced in this way is shown according to the reading of the body and function peaks. There was a phenomenon in which the number of Nt stabilized when the read of the function peak increased by more than a certain ratio. In consideration of this, sample A and sample B were made based on body peak reads 0.8 M and function peak reads 3.2 M by choosing a case where the result of rescue analysis was greater than 2 (Table 1). After that, the remaining 5 simulation samples were created using simulation sample A, B and the remaining one peak set that was not used to create simulation samples A and B. Simulation samples C and D are samples that do not have reproducibility for A and B. Among them, the simulation sample that is less related to A and B was made as C using the remaining one large set that was not used to create simulation samples A and B. And simulation sample D was randomly sampled at 10% each of sample A and B. Finally, simulation samples E, F, and G are samples that have reproducibility for A and B. The degree of reproducibility for A and B increases in the order of E, F, and G. Simulation samples E, F, and G were randomly selected at 30%, 50%, and 60% of samples A and B, respectively (Supplementary Fig. 9).

Compiling a different state's simulation sample set that highlights the difference in reproducibility

To evaluate the performance differences between the compared algorithms based on the reproducibility of the simulation sets, we defined three different states: standard, hard, and soft. Specifically, in the standard state, we used simulation set F, which has a 50% overlap with either simulation set A or B. However, in the hard state, we used simulation set E, which has only a 30% overlap. Conversely, in the soft state, we used simulation set G, which has a 60% overlap.

Compiling a different state's simulation sample set that highlights the number of similar replications

To evaluate the performance differences of the compared algorithms based on the number of similar replications of the simulation sets, we defined three different states: strong, normal, and weak. In the normal state, we used cases of three and four replication sets that passed through MESIA stage1. In the strong state, we used cases of

```

# MESIA stage 1
for (i in Number of input replication){
  test Self consistence analysis
}

for (i in 1 : Number of input replication C 2){
  test Rescue analysis each set
}

if ( Self consistence analysis result < 2 &&
  Rescue analysis each set result < 2 ){
  the replication set pass MESIA stage 1
}

# MESIA stage 2
Use the replication set that passed MESIA stage 1
for (i in Number of replication set that passed MESIA stage 1){

  # Replication set A
  for (i in length (comparison set) ){
    test Self consistence analysis
    Get Nt peaks_A1 and Nt peaks_A2
  }

  l, length of comparison set
  test Rescue analysis replication set
  Get Np peaks_A
   $Nt\ peaks\_A1 \cap Nt\ peaks\_A2 \cap Np\ peaks$  then that is Base peaks_A
  Get median for Peak point in set A then that is Standard_PP
  Get median for Q value in set A then that is Standard_QV
  Calculate stage 2 Replication set A peak with Standard_PP, Standard_QV

  # Repeat as many times as Number of replication set
  that passed MESIA stage 1
}

for ( i in 1 : Number of replication set that passed MESIA stage 1 ){
   $\sum$  stage of Replication set i peak then that is MESIA Optimal peak
}

```

Figure 3. MESIA Pseudocode. Pseudo-code of the MESIA algorithm.

five and six replication sets that passed through MESIA stage1. Finally, in the weak state, we used the case of one and two replication sets that passed through MESIA stage1.

Calculating peak distribution and gene expression

Peaks formed in the promoter region derived from each method were used. Performance was compared based on the difference in the degree of gene expression related to the peak formed in the promoter. Gene expression information was derived from GM12878 polyA- and RNA-seq. it was aligned to the hg19 genome assembly.

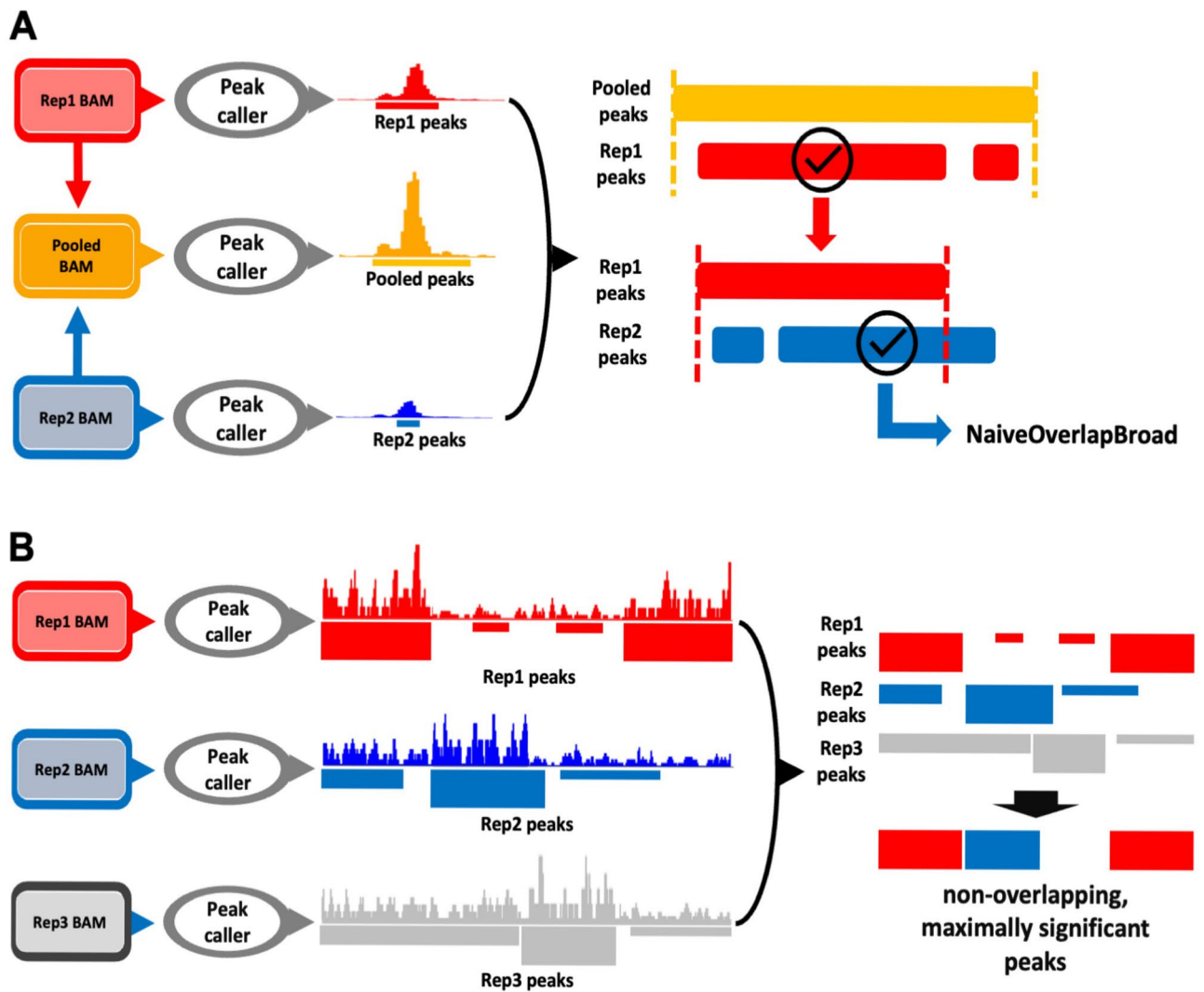


Figure 4. Comparative algorithms related to length-based approaches and statistical based approaches. **(A)** a schematic diagram of NAIVE, the length-based approaches algorithm. **(B)** a schematic diagram of NOMS, union approaches algorithm.

Results

Differences in peak contents between comparative algorithms

When comparing the changes in the number of peaks between the soft and hard states based on the standard state, MESIA exhibited a 10.55% increase in the soft state and a 20.47% decrease in the hard state. This method had the smallest amount of variation among all methods, except for NOMS like union method (Fig. 6A). To investigate the differences in the number of peaks based on the degree of reproducibility, we identified the regions where the peaks formed in each state using org.Hs.eg.db²⁷ (v4.3). In the hard state, MESIA formed the fewest peaks in the distal intergenic region, while forming the most peaks in the promoter region when the region within 3 KB from the transcription start site was defined as the promoter region (Fig. 6B). This trend was consistent in both soft and standard states (Supplementary Fig. 1).

Simulation results

We compared the performance of the methods depending on a degree of reproducibility, based on the three previously defined states. First, we present the distribution of genes in which optimal peaks were found in the promoter region for each method. In the hard state, for the other three algorithms that used the significant integration approach, MESIA identified 4,716 genes, which was 1.78 times more than ChIP-R and 3.37 times more than NAIVE (Fig. 7A). This trend was also observed for the soft and standard states (Supplementary Fig. 2).

We compared the performance of these algorithms in detecting genes based on gene expression levels. The expression of genes detected by MESIA was significantly different from that detected by all the comparative algorithms ($p=0.04402$, $p=0.004593$, $p<2.2e-16$) (Fig. 7B). We present the distribution of genes detected by each significant integration approach algorithm using a Venn diagram. In the hard state, among the three algorithms, MESIA detected the highest number of genes (2425). This was 6.06 times higher than the number detected by

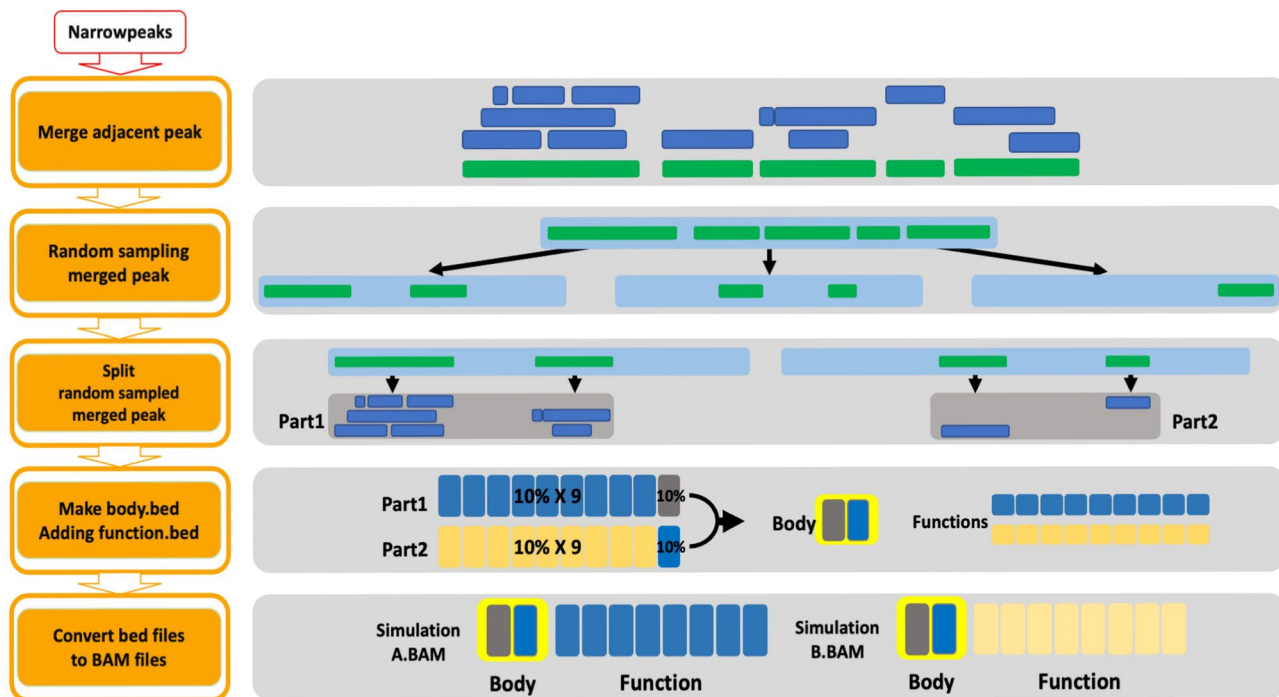


Figure 5. Compiling simulation sample set. Process of compiling a simulation sample set to determine the performance difference between MESIA and the compared algorithms. To ensure nonreproducibility between the simulation sets, samples A and B were created by dividing the peaks obtained from a single ATAC sample.

	Body reads (M)	Function reads (M)	Np peaks	Nt peaks	Rescue analysis result	Decision
0.8	0.4	0.4	9445	47,313	5.009	Unstable
	0.8	0.8	12,238	47,313	3.866	Unstable
	1.2	1.2	14,904	47,314	3.174	Unstable
	1.6	1.6	17,751	47,313	2.665	Unstable
	2	2	20,247	47,313	2.336	Unstable
	2.4	2.4	23,036	47,314	2.053	Unstable
	2.8	2.8	25,882	47,314	1.828	Similar replication set
	3.2	3.2	28,841	10,953	2.633	Different replication set
1.6	0.4	0.4	16,175	94,552	5.845	Unstable
	0.8	0.8	18,517	94,552	5.106	Unstable
	1.2	1.2	21,318	94,554	4.435	Unstable
	1.6	1.6	23,956	94,554	3.946	Unstable
	2	2	26,794	94,554	3.5289	Unstable
	2.4	2.4	29,734	21,901	1.357	Similar replication set
	2.8	2.8	32,495	21,901	1.487	Similar replication set
2.4	0.4	0.4	22,105	141,862	6.417	Unstable
	0.8	0.8	25,049	141,864	5.663	Unstable
	1.2	1.2	24,856	141,862	5.092	Unstable
	1.6	1.6	30,551	32,672	1.069	Similar replication set
	2	2	33,616	32,672	1.028	Similar replication set
	2.4	2.4	36,190	32,672	1.107	Similar replication set
3.2	0.4	0.4	29,193	188,891	6.470	Unstable
	0.8	0.8	31,690	43,977	1.387	Similar replication set
	1.2	1.2	34,510	43,977	1.274	Similar replication set
	1.6	1.6	37,067	43,977	1.186	Similar replication set
	2	2	39,859	43,977	1.103	Similar replication set

Table 1. Summary of intermediate results during compilation of simulation sample set. M, million units; Np, peaks consistent between poured pseudoreps; Nt, peaks consistent between true replicates.

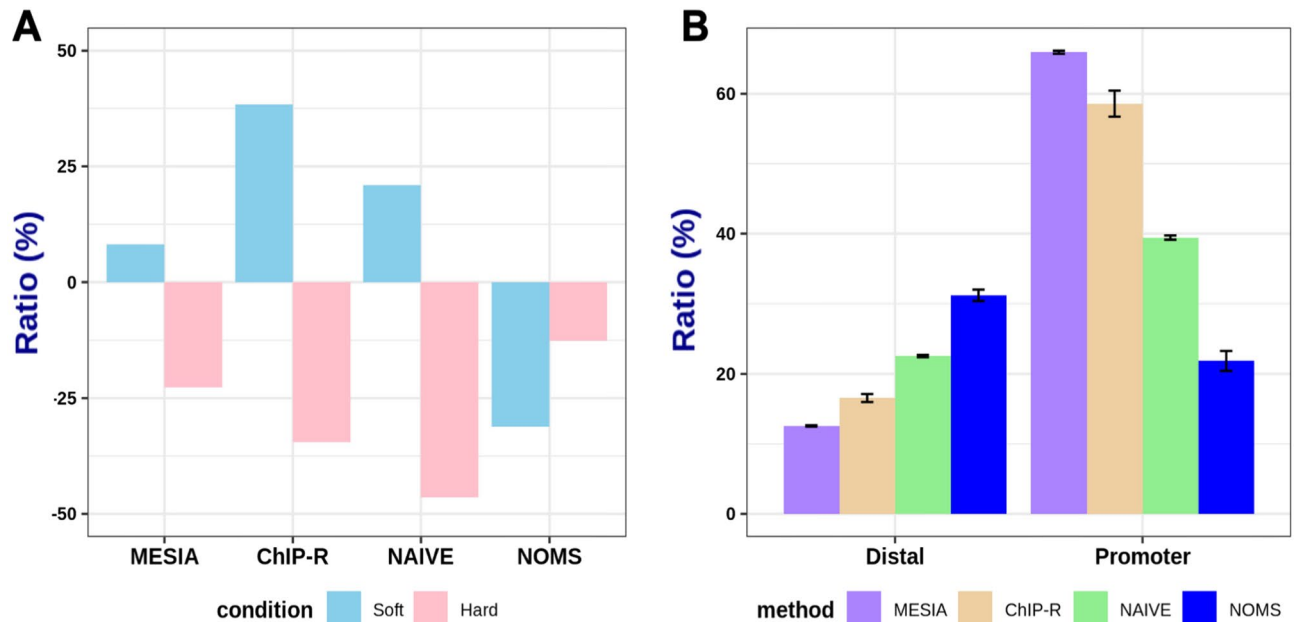


Figure 6. Differences in peak contents between comparative algorithms, depending on a degree of reproducibility. **(A)** Bar plot of difference according to condition across methodologies. In a standard state, we use simulation set 50% overlap with either simulation set A or B. In a hard state, use simulation set 30% overlap. Last, in a soft state, we use simulation set 60% overlap. Compared to the number of peaks in the standard state, an increase and decrease of peak counts in the soft state are represented by pink bars, while the hard state is represented by sky blue bars in the same way. **(B)** Comparative bar plot between distal and promoter regions. Under the hard state conditions, the ratio of genes that form peaks in both the distal intergenic and promoter regions were determined using the optimal peaks of each comparative algorithm. region was set to be 3 KB away from the transcription start site (TSS).

ChIP-R and 4.88 times higher than that detected by NAIVE (Fig. 7C). The expression of genes detected only at MESIA was significantly different from that at ChIP-R and NAIVE ($p = 4.142e-05$ and $p = 7.382e-09$, respectively). (Fig. 7D). Also, we separately compared NAIVE and ChIP-R with MESIA. The trend of MESIA maintaining better performance has continued. (Supplementary Figs. 3 and 4). Similarly, to this approach, we compared the performance of the algorithms according to number of similar replications. In the weak state, for the three algorithms that used the significant integration approach, MESIA identified 1218 genes. This value is in the middle of those of the three algorithms (Supplementary Fig. 5A). The genes detected by MESIA were significantly different from those detected by all the other comparative algorithms. ($p = 0.01707$, $p = 5.369e-05$, $p < 2.2e-16$ in order). (Supplementary Fig. 5B). This trend was also observed for the strong and normal states (Supplementary Fig. 6). In the weak state, among the three algorithms, MESIA ranked in the middle for the number of genes detected exclusively by this method and the expression of genes detected only in MESIA was significantly different from that in ChIP-R and NAIVE ($p = 0.04749$ and $p = 0.00295$, respectively). (Supplementary Fig. 5C and D). MESIA's consistent improvement in performance has remained a prevailing trend. (Supplementary Figs. 7 and 8).

Validation results

The performance was compared using real data. The distribution of the number of peaks in the four GM12878 replications is shown as a Venn diagram (Fig. 8A). For each method, the number of replicates related to each peak was classified and expressed as a circular bar plot. As expected, most peaks overlapped in all four replicates. Interestingly, MESIA showed the smallest difference in the peak according to the number of overlapping replications among the three compared methods. (Fig. 8B). MESIA found that replications 2, 3, and 4 were reproducible among the 4 replications. Therefore, it primarily identified regions where peaks were formed simultaneously in the 3 replications as optimal peaks. (Supplementary Fig. 10). The Venn diagram displays the gene distribution where each method formed optimal peaks in the promoter region. (Fig. 8C). We performed a comparison of the methods used to detect genes that showed significant differences in expression levels. The genes detected by MESIA were significantly different from those detected by NAIVE. (Fig. 8F) ($p = 1.117e-13$). To further analyze this difference, we conducted a comparison between only compared MESIA and NAIVE. We observed that 98.84% of the genes detected by MESIA overlapped with those detected by NAIVE. (Fig. 8D). The expression levels of genes detected by MESIA were higher than the Q3 threshold based on the expression levels of genes in the hg19 reference genome. Genes that were detected only by NAIVE and not by MESIA had expression levels lower than Q3, and this difference was statistically significant compared to the expression levels of genes detected only by MESIA. (Fig. 8G) ($p = 0.0001482$) We also compared MESIA with ChIP-R using a similar approach. The expression levels of genes detected using only MESIA were significantly higher than those detected using only ChIP-R. ($p = 0.0352$), respectively (Fig. 8E and H). These analysis results were also similar in mouse embryonic

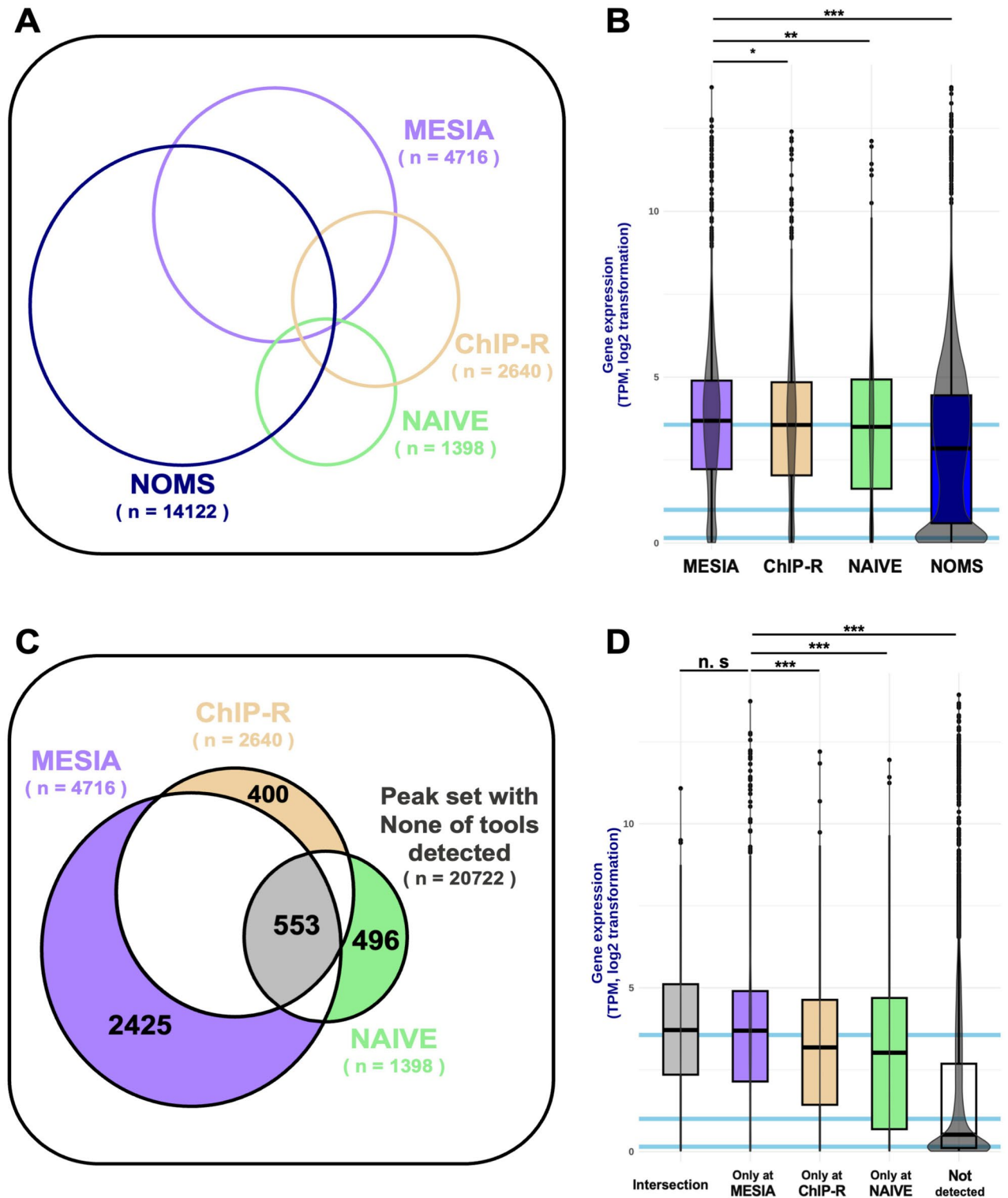


Figure 7. Simulation results. **(A)** Venn diagram of genes with optimal peaks formed in a promoter region across MESIA and compared algorithms. **(B)** Gene expression level of the genes identified by MESIA and compared algorithms. PolyA plus RNA-seq was utilized and it is expressed as a violin plot using log₂ normalization gene expression. The sky-blue lines represent Q3, median, and Q1 in order, based on the expression levels of genes in the hg19 reference genome. **(C)** Venn diagram of genes which detected exclusively by each algorithm in MESIA, CHIP-R and NAIVE. **(D)** Gene expression level of the genes detected exclusively by each algorithm. (n.s., $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $p < 0.0001$).

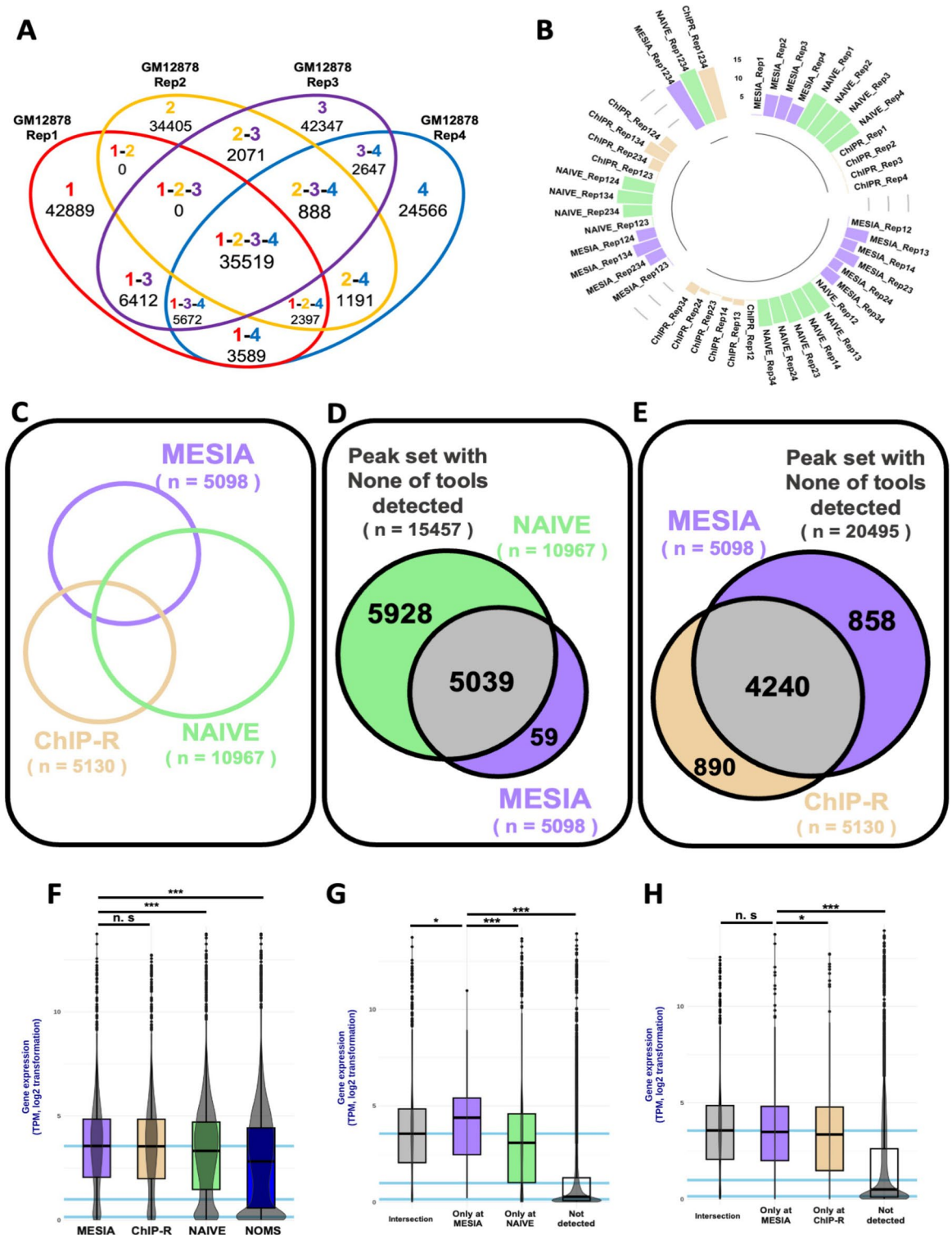


Figure 8. Validation Results. (A) Venn diagram of genes with optimal peaks formed in a promoter region across MESIA and compared algorithms. (B) Circular bar plot visualized number of replications related to peak. (C) Venn diagram of genes with optimal peaks formed in a promoter region across MESIA and compared algorithms. (D) Venn diagram of genes with optimal peaks formed in a promoter region which detected exclusively by each algorithm in MESIA and NAIVE. (E) Venn diagram of genes with optimal peaks formed in a promoter region which detected exclusively by each algorithm in MESIA and ChIP-R. (F) Gene expression level of the genes identified by MESIA and compared algorithms. PolyA plus RNA-seq was utilized and it is expressed as a violin plot using log2 normalization gene expression. The sky-blue lines represent Q3, median, and Q1 in order, based on the expression levels of genes in the hg19 reference genome. (G) Gene expression level of the genes detected exclusively by each MESIA and NAIVE. (H) Gene expression level of the genes detected exclusively by each MESIA and ChIP-R. (n.s., $p > 0.05$; *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$; ***, $p < 0.0001$).

fibroblasts datasets⁸, which are not human datasets. Therefore, it was revealed that MESIA maintains good performance even in datasets that are not human lymphoblastoid datasets. (Supplementary Fig. 11). We evaluated the similarity between the analysis using the latest reference genome, hg38, and the analysis using hg19, employing the Jaccard similarity. The resulting value of 0.9743 was exceptionally high, indicating that the analyses using the two reference genomes produced virtually identical results.

Discussion

The challenges of the diversity and integration of peaks discovered in multi-sample epigenomic analyses still exist. In this study, we proposed a new integrative method for multi-sample epigenomic analysis. This method combines epigenetic data from multiple samples to identify consistent peak patterns and will be valuable for genomic analysis of high-resolution genome and epigenome data.

Previous studies have shown that shared peaks across different cell types and tissues are associated with regulatory elements, such as enhancers and promoters, and play important roles in the regulation of gene expression^{28, 29}. Our MESIA analysis identified conserved peak patterns associated with known regulatory elements, suggesting that our method can effectively identify biologically relevant peaks shared across multiple samples. Several studies integrated epigenomic signals from various disease models and computational approaches.

Epigenome peak analysis has emerged as a powerful tool for investigating the molecular mechanisms underlying common diseases, and numerous studies have used this approach to identify disease-associated epigenetic changes. Recent studies have used epigenomic peak analysis to identify novel therapeutic targets for Alzheimer's disease, autism, and type2 diabetes. Researchers have identified unique peak patterns that are associated with the disease and used this information to develop a new diagnostic tool for disease^{30–32}. Moreover, in cancer research, many researchers have used multi-sample epigenome peak analysis to identify epigenetic changes associated with cancer development and progression. Researchers have found that certain epigenetic modifications are more prevalent in cancer patients and have developed new therapeutic strategies targeting these modifications^{33–35}.

There are several methods for integrating epigenomic peaks. For example, ChromVAR uses a latent variable model to identify shared and sample-specific regulatory elements across multiple epigenomic datasets. It can also perform differential analyses and prioritize disease-associated variants³⁶. The scABC is designed for single-cell epigenomic data and uses a Bayesian clustering approach to identify clusters of cells with similar epigenomic profiles across multiple samples or conditions. It can be used to identify cell type-specific and shared regulatory elements, and to study cellular heterogeneity³⁷. Moreover, Episom uses a multi-sample peak-calling approach to identify shared and sample-specific peaks across multiple epigenomic datasets. It can also be integrated with other genomic data types, such as RNA-seq and ChIP-seq, to identify candidate regulatory regions and their target genes³⁸. EpiTensor uses a tensor decomposition approach to identify shared and sample-specific epigenomic patterns across multiple samples and conditions. They can be used to identify regulatory elements and gene regulatory networks that are conserved or variable across different contexts³⁹.

Although some methods use statistical approaches, most use a length overlap-based approach. However, to the best of our knowledge, there are no methods that can integrate multiple samples while considering reproducibility between replicates. Specifically, most methods do not consider the reproducibility between replicates while integrating the peaks. Consequently, a statistical approach is vulnerable to noise caused by significant differences between samples. Our research aims to extract more accurate and reliable peaks by utilizing a statistical methodology that is more active than traditional multi-sample integration methods. To achieve this, we statistically extracted the significance of the individual peaks and integrated the peak information from multiple samples to determine whether each peak was significant. As a result, compared to other existing algorithms in situations where the degree of reproducibility is not sufficiently ensured, MESIA was able to detect peaks in up to 3.37 times more genes. Furthermore, the genes detected only by MESIA showed significantly higher expression levels than those detected by the other algorithms. In situations where the number of similar replications was insufficient, MESIA also identified genes that were significantly more highly expressed. These results show that MESIA, in addition to detecting genes with significantly higher expression by finding strong signals, also filters out false positive peaks that other algorithms miss. It does this by selecting peaks from reproducible replications. This methodology can be applied regardless of the sample size, allowing us to obtain valid results even with small sample sizes. To do this, appropriate statistical adjustments and simulations are necessary. For example, statistical adjustments were performed for all possible sample combinations and accurate results were obtained through simulations of individual combinations. As a result, we were able to obtain robust peaks that were not strongly affected by the degree of reproducibility and the number of similar replicates. Our study has some limitations. MESIA considers the reproducibility between replications before proceeding with Multi-Sample Integration, which may take a considerable amount of time and generate intermediate products proportional to the number of replications. As with all other methods, further validation using additional genomic data or experiments is necessary to increase the reliability of the results.

The integration of multi-sample epigenome peak analysis has various applications. One is in the development of new therapeutic strategies. Epigenetic modifications have been shown to play a critical role in the regulation of gene expression, and aberrant epigenetic changes have been implicated in the development and progression of various diseases. By identifying disease-specific peak patterns, researchers can develop new therapeutic strategies that target these epigenetic modifications. For example, drugs that target specific epigenetic enzymes or regulatory proteins can be developed to correct aberrant epigenetic changes and restore normal gene expression⁴⁰.

MESIA is a user-friendly, open-source tool that enables researchers to integrate various epigenetic changes easily and accurately. By shedding light on the sophisticated regulation of gene expression and disease development, this approach can provide valuable and novel biological insights.

Data availability

ATAC-seq and RNA-seq datasets are publicly available on the NCBI Gene Expression Omnibus website (accession numbers GSE47753, GSE78554 and GSE145705).

Code availability

Source codes for MESIA are freely available on GitHub (<https://github.com/ERASMUSlab/MESIA>).

Received: 18 September 2023; Accepted: 20 November 2023

Published online: 27 November 2023

References

1. Itokawa, N. *et al.* Epigenetic traits inscribed in chromatin accessibility in aged hematopoietic stem cells. *Nat. Commun.* **13**, 2691. <https://doi.org/10.1038/s41467-022-30440-2> (2022).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63. <https://doi.org/10.1038/nrg2484> (2009).
3. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1497–1502. <https://doi.org/10.1126/science.1141319> (2007).
4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218. <https://doi.org/10.1038/nmeth.2688> (2013).
5. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322. <https://doi.org/10.1016/j.cell.2007.12.014> (2008).
6. Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinf.* **18**, 441–450 (2017).
7. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
8. Murphy, K. E., Meng, F. W., Makowski, C. E. & Murphy, P. J. Genome-wide chromatin accessibility is restricted by ANP32E. *Nat. Commun.* **11**, 5063. <https://doi.org/10.1038/s41467-020-18821-x> (2020).
9. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
10. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
11. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359. <https://doi.org/10.1038/nmeth.1923> (2012).
12. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
13. Bhattacharyya. `assign_multimappers.py`.
14. Reske, J. `bedpeTn5shift.sh`. (2019).
15. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137. <https://doi.org/10.1186/gb-2008-9-9-r137> (2008).
16. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692. <https://doi.org/10.1093/bioinformatics/btr174> (2011).
17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. <https://doi.org/10.1093/bioinformatics/btq033> (2010).
18. Tischer, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13. <https://doi.org/10.1186/1751-0473-9-13> (2014).
19. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034. <https://doi.org/10.1093/bioinformatics/btv098> (2015).
20. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419. <https://doi.org/10.1038/nmeth.4197> (2017).
21. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527. <https://doi.org/10.1038/nbt.3519> (2016).
22. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. <https://doi.org/10.1101/2023.04.04.535623> (2023).
23. Reske, J. J., Wilson, M. R. & Chandler, R. L. ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigen. Chromatin* **13**, 22. <https://doi.org/10.1186/s13072-020-00342-y> (2020).
24. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831. <https://doi.org/10.1101/gr.136184.111> (2012).
25. Newell, R. *et al.* ChIP-R: Assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates. *Genomics* **113**, 1855–1866. <https://doi.org/10.1016/j.ygeno.2021.04.026> (2021).
26. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203. <https://doi.org/10.1038/ng.3646> (2016).
27. Carlson, M. [org.Hs.eg.db](https://doi.org/10.18129/B9.bioc.org.Hs.eg.db). <https://doi.org/10.18129/B9.bioc.org.Hs.eg.db>
28. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82. <https://doi.org/10.1038/nature11232> (2012).
29. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220. <https://doi.org/10.1038/s41576-018-0089-8> (2019).
30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. <https://doi.org/10.1186/s13059-014-0550-8> (2014).
31. Keller, A. S., Leikauf, J. E., Holt-Gosselin, B., Staveland, B. R. & Williams, L. M. Correction: Paying attention to attention in depression. *Trans. Psychiatry* **10**, 64. <https://doi.org/10.1038/s41398-020-0748-3> (2020).
32. Binder, S. *et al.* Master and servant: LINC00152 – a STAT3-induced long noncoding RNA regulates STAT3 in a positive feedback in human multiple myeloma. *BMC Med. Genom.* **13**, 22. <https://doi.org/10.1186/s12920-020-0692-3> (2020).
33. Mannheimer, J. D., Duval, D. L., Prasad, A. & Gustafson, D. L. A systematic analysis of genomics-based modeling approaches for prediction of drug response to cytotoxic chemotherapies. *BMC Med. Genom.* **12**, 87. <https://doi.org/10.1186/s12920-019-0519-2> (2019).
34. Shirmohammadi, E., Ebrahimi, S.-E.S., Farshchi, A. & Salimi, M. Correction to: The efficacy of etanercept as anti-breast cancer treatment is attenuated by residing macrophages. *BMC Cancer* **20**, 1126. <https://doi.org/10.1186/s12885-020-07641-3> (2020).
35. Zhai, Y. *et al.* Construction of the optimization prognostic model based on differentially expressed immune genes of lung adenocarcinoma. *BMC Cancer* **21**, 213. <https://doi.org/10.1186/s12885-021-07911-8> (2021).

36. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978. <https://doi.org/10.1038/nmeth.4401> (2017).
37. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94. <https://doi.org/10.1038/nbt.4042> (2018).
38. Xie, X. Y. *et al.* Mental health status among children in home confinement during the coronavirus disease 2019 outbreak in Hubei Province. *China. Jama Pediatrics* **174**, 898–900. <https://doi.org/10.1001/jamapediatrics.2020.1619> (2020).
39. Chen, R. E. *et al.* Resistance of SARS-CoV-2 variants to neutralization by monoclonal and serum-derived polyclonal antibodies. *Nat. Med.* **27**, 717–726. <https://doi.org/10.1038/s41591-021-01294-w> (2021).
40. Sulheim, S., Fossheim, F. A., Wentzel, A. & Almaas, E. Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters. *BMC Bioinf.* **22**, 81. <https://doi.org/10.1186/s12859-021-03985-0> (2021).

Acknowledgements

We are grateful to the anonymous reviewers for helpful suggestions. This work was supported by National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) [2020R1A4A1019423, 2021R1C1C2095130, 2022R1I1A1A01062894 and RS-2023-00207971].

Author contributions

S.G.P.: Data curation, Formal analysis, Investigation, Methodology, Software handling, Validation and Drafting the manuscript. W.J.K.: Data curation, Formal analysis, Investigation, Methodology, Software handling, Validation and Drafting the manuscript. J.I.M.: Data curation, Validation and Drafting the manuscript. K.T.K.: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Review and Editing. H.M.R.: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Review and Editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-47948-2>.

Correspondence and requests for materials should be addressed to K.-T.K. or H.-M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023