METHODS
published: 18 February 2022
doi: 10.3389/fgene.2022.818344

Integration of Count Difference and Curve Similarity in Negative Regulatory Element Detection

Na He[1,2]*†, Wenjing Wang[3]†, Chao Fang[4], Yongjian Tan[2], Li Li[5,6] and Chunhui Hou[2]*

[1]Harbin Institute of Technology, Harbin, China, [2]Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China, [3]School of Life Science and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China, [4]Cancer Centre, Faculty of Health Sciences, University of Macau, Macao, China, [5]Department of Bioinformatics, Huazhong Agricultural University, Wuhan, China, [6]Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China

OPEN ACCESS

Edited by:
Liang Guo,
Chinese Academy of Fishery Sciences (CAFS), China

Reviewed by:
Sebastiaan Meijsing,
Max Planck Society, Germany
Zhihua Zhang,
Beijing Institute of Genomics (CAS), China

***Correspondence:**
Chunhui Hou
houch@sustech.edu.cn
Na He
11849492@mail.sustech.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 19 November 2021
Accepted: 20 January 2022
Published: 18 February 2022

Citation:
He N, Wang W, Fang C, Tan Y, Li L and Hou C (2022) Integration of Count Difference and Curve Similarity in Negative Regulatory Element Detection.
Front. Genet. 13:818344.
doi: 10.3389/fgene.2022.818344

Negative regulatory elements (NREs) down-regulate gene expression by inhibiting the activities of promoters or enhancers. The repressing activity of NREs can be measured globally by massively parallel reporter assays (MPRAs). However, most existing algorithms are designed for the statistical detection of positively enriched signals in MPRA datasets. To identify reduced signals in MPRA experiments, we designed a NRE identification program, fast-NR, by integrating the count and graphic features of sequenced reads to detect NREs using datasets generated by experiments of self-transcribing active regulatory region sequencing (STARR-seq). Fast-NR identified hundreds of silencers in human K562 cells that can be validated by independent methods.

Keywords: count difference, curve similarity, silencer, silencer identification, negative regulatory element

INTRODUCTION

Eukaryotic gene expression is tightly controlled by various types of *cis*-regulatory elements (CREs) that are different in regulatory function, genetic, and epigenetic characteristics (Maston et al., 2006). Promoters and enhancers are positive CREs that initiate and enhance transcription, respectively. Enhancers act locally or over long genomic distances through chromatin looping to regulate their target genes (Shlyueva et al., 2014; Haberle and Stark, 2018; Schoenfelder and Fraser, 2019). In contrast, silencers are negative regulatory elements (NRE) that suppress gene expression through mechanisms that are not completely understood (Ogbourne and Antalis, 1998). Mutations in human CREs associate frequently with tumorigenesis, neurodegeneration, and metabolic diseases (Maston et al., 2006), highlighting the functional importance of transcription control in cells.

In eukaryotic genomes, silencers had not been as vigorously investigated as enhancers (Della Rosa and Spivakov, 2020). Most silencers in the database of silencerDB are predicted (Zeng et al., 2021). Potential silencers were also predicted in cell lines (Doni Jayavelu et al., 2020) by gkmSVM which utilizes sequence features of known silencers (Ghandi et al., 2016). Different from enhancers, silencer prediction is currently infeasible because that whether silencers carry ubiquitous epigenetic signatures is unknown. Genome-wide characterization of functional silencers is thus critical to unveil the genetic and epigenetic features of silencers. Genomic sequences of regulatory activity can be systematically assessed by STARR-seq, a widely used MPRA method initially designed for enhancer identification (Melnikov et al., 2012; Arnold et al., 2013; Crocker and Stern, 2013; Gisselbrecht et al., 2013; Mogno et al., 2013; Vanhille et al., 2015; Wang et al., 2018; Sun et al., 2019). Theoretically, STARR-seq measures silencer activity as well. Actually, Doni Jayavelu et al. had successfully used STARR-seq to measure the transcription-repressing activity of silencers that were predicted by epigenetic features (Doni Jayavelu et al., 2020). Recently, several studies had reported catalogs of silencers that had been predicted or identified by different methods in different model systems at small scales (Petrykowska et al., 2008; Huang et al., 2019; Doni Jayavelu et al., 2020; Pang and Snyder, 2020).

For MPRAs, a statistical method specially designed for silencer identification is needed to facilitate the investigations into silencers' identity and their roles in transcription regulation. To design a silencer identification program, developers need to consider the functional differences between enhancers and silencers (Zhang et al., 2008; Heinz et al., 2010; Lee et al., 2020). Doni Jayavelu et al. measured silencer activities in selected accessible chromatin regions by comparing the sequenced reads of the reporter cDNA to the reads of the input insert DNA using a one-tail *t* test (Doni Jayavelu et al., 2020). While Pang et al. used a model-based method MAGeCK (Li et al., 2014) after counting reads with the method of HTSeq (Anders et al., 2015; Pang and Snyder, 2020). MAGeCK is similar to edgeR (Robinson et al., 2010) and DESeq (Anders and Huber, 2010) in their design strategies, but it is different from the other two methods in its intended usage. MAGeCK is originally used in CRISPR/Cas9 knockout screen assays. Different from these small-scale assays, genome-wide sequenced reads follow a negative binomial distribution. Potentially, methods designed for the detection of differentially methylated regions (DMRs) or differential chromatin modifications (Shen et al., 2013; Lienhard et al., 2014; Zhang et al., 2014; Lun and Smyth, 2016; Gaspar and Hart, 2017) can be used to identify silencers. However, the specificity, robustness, accuracy, and resolution of these programs have not been evaluated for silencer identification. CRADLE, a recently published method, is designed for enhancer identification (Kim et al., 2021). Theoretically, CRADLE can detect silencers as well. Nevertheless, a computational method designed specifically for the identification of silencers has not been reported.

In this research, we provide a program Fast-NR that is designed for the identification of silencers using STARR-seq-generated datasets by integrating the sequenced read count and signal shape features which are considered in the design of many ChIP-seq peak callers including Polyapeak, PICS, and CLC (Thomas et al., 2017; Hower et al., 2011; Cremona et al., 2019; Yan et al., 2020) (Zhang et al., 2011; Wu and Ji, 2014; Strino and Lappe, 2016). Fast-NR is available at https://github.com/Na-He/Fast-NR. We tested this program on simulated and STARR-seq datasets (Johnson et al., 2018; Doni Jayavelu et al., 2020), compared the performance of Fast-NR with several other programs, and show here that Fast-NR can detect NREs under different conditions.

METHODS

Algorithm

DNA fragments of NRE activity reduce their own expression levels in the STARR-seq reporter cDNA library. To identify NREs, we first calculate *p* values for each nucleotide covered by the reporter cDNA and the input insert DNA across the genome. If a *p* value is below an arbitrary threshold, the corresponding genomic region is considered as a potential NRE. We then plot the numbers of the reporter cDNA and input DNA reads as curves and measure the distances between them to determine whether they are similar by using several different methods. For NREs, the similarity scores are supposed to be low. By integrating count number difference and curve similarity features, we designed a computational method, Fast-NR (**Figure 1**), and tested its NRE detection power on simulated and real STARR-seq datasets, respectively. Basically, we first screened nucleotides which had the number of reporter cDNA reads smaller than the input insert reads by at least 12, corresponding to the *p* value threshold (10^{-5}) we set. We calculate *p* values using cumulative density function (CDF) of negative binomial distribution (NBD). Next, we use the single nucleotides that pass the initial screen as anchors and extend the genomic window to upstream and downstream to a total of 601 bp. We further examine the *p* values of each nucleotide in each 601bp window and keep only windows in which 3/4 of all nucleotides are with a *p* value below 10^{-5}. If two windows overlap, we keep the one in the shared region with smaller *p* values. Then, we compare the similarity between the curves of reporter cDNA and the input insert DNA reads, and discard any window with a curve similarity score higher than the arbitrary threshold. Finally, we correct *p* values for each window of identified NRE by Benjamini–Hochberg (BH) test and keep only these with a corrected *p* value smaller than 10^{-5}.

p Value Calculation

We calculate *p* values by cumulative density function (CDF) of the negative binomial distribution for the sequenced reporter cDNA reads. The probability mass function of the number of *k* times failure for a negative binomial distribution is

$$\boldsymbol{CDF(m,n,p) = P(x_n \le m) = \sum_{i=0}^{m} \binom{i+n-1}{n-1} p^n (1-p)^i},$$

where $CDF(m,n,p)$ returns the probability that is fewer than m times failure before the n th times success, with a single success probability p. Here, the m is treatCount which comes from a negative binomial distribution, n is treatTotal-treatCount, and p is (controlTotal-controlCount)/controlTotal. treatTotal and controlTotal are the total fragment numbers of the reporter cDNA and the input insert DNA in the sequenced libraries, respectively. treatCount and controlCount are the count numbers of reporter cDNA and input insert DNA covering each nucleotide, respectively.

Curve Similarity

We compare the shape of the curves of the reporter cDNA and the input insert DNA reads. Cosine, Pearson, Euclidean, and an in-house method gradient (linear slope correlation) are used to calculate the curve similarity in this research.

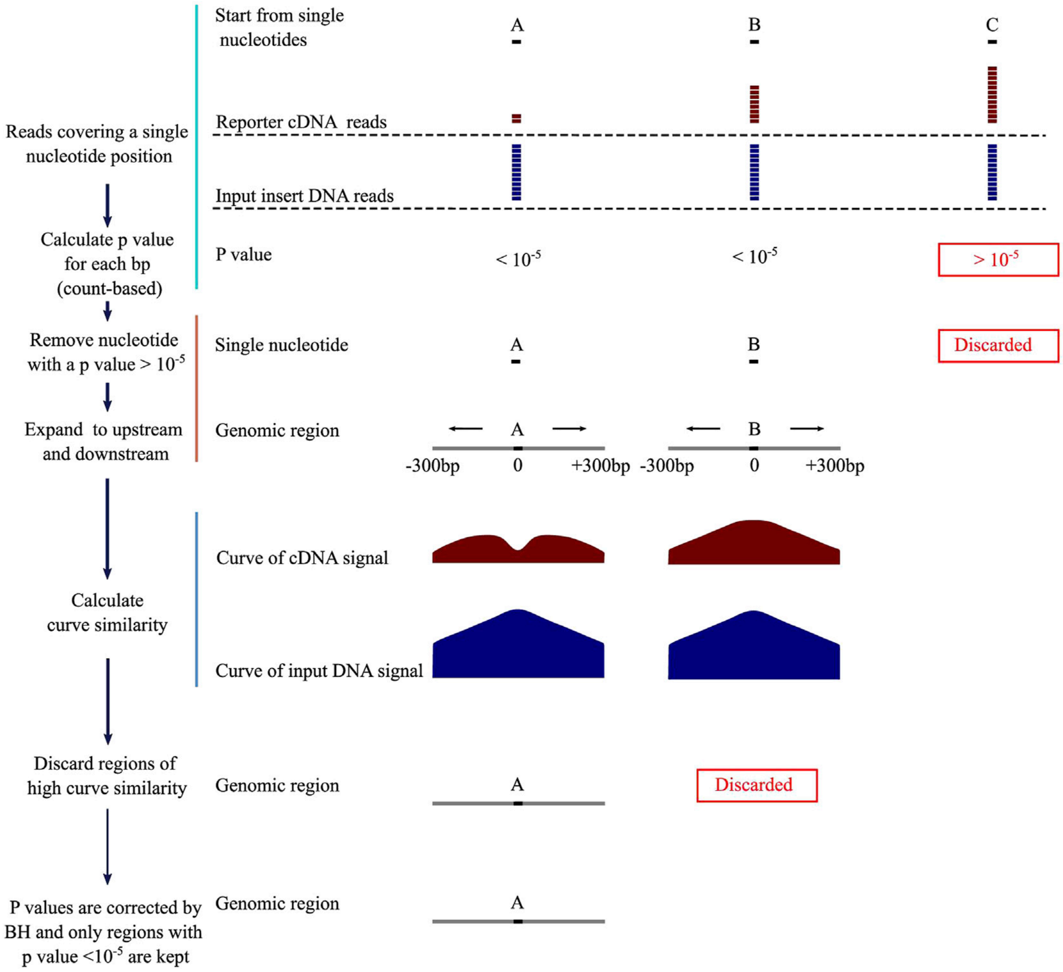
Start from single nucleotides
A
B
C
Reads covering a single nucleotide position
Reporter cDNA reads
Input insert DNA reads
Calculate p value for each bp (count-based)
P value
< 10^-5
< 10^-5
> 10^-5
Remove nucleotide with a p value > 10^-5
Single nucleotide
A
B
Discarded
Expand to upstream and downstream
Genomic region
A
B
-300bp
0
+300bp
-300bp
0
+300bp
Calculate curve similarity
Curve of cDNA signal
Curve of input DNA signal
Discard regions of high curve similarity
Genomic region
A
Discarded
P values are corrected by BH and only regions with p value <10^-5 are kept
Genomic region
A

FIGURE 1 | Negative regulatory element identification pipeline. BH, Benjamini–Hochberg correction.

Cosine

The method of cosine computes the cosine distance between the 1-D arrays of u and v:

$$\mathbf{Cos}\ (\boldsymbol{u},\ \boldsymbol{v}) = \mathbf{1} - \frac{\sum_{i=0}^{n} \boldsymbol{u}_i \boldsymbol{v}_i}{\sqrt{\sum_{i=0}^{n} \boldsymbol{u}_i^2}\sqrt{\sum_{i=0}^{n} \boldsymbol{v}_i^2}},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Euclidean

Euclidean method computes the Euclidean distance between the 1-D arrays of u and v:

$$\mathbf{Euclidean\,(dis)} = \sqrt{\sum_{i=0}^{n} (\boldsymbol{u}_i - \boldsymbol{v}_i)^2},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Pearson

Pearson computes the Pearson correlation coefficient between the 1-D arrays of u and v:

$$\boldsymbol{cor} = \frac{\sum_{i=0}^{n} (\boldsymbol{u}_i - \bar{\boldsymbol{u}})(\boldsymbol{v}_i - \bar{\boldsymbol{v}})}{\sqrt{\sum_{i=0}^{n} (\boldsymbol{u}_i - \bar{\boldsymbol{u}})^2}\sqrt{\sum_{i=0}^{n} (\boldsymbol{v}_i - \bar{\boldsymbol{v}})^2}},$$

where u_i and v_i are the reporter cDNA and the input insert DNA read count values in the u and v vectors.

Gradient

This algorithm computes the Pearson correlation coefficients between the gradients of the curves of the reporter cDNA and the input insert DNA reads. The coverage and genomic location values of each silencer form a 2-D array, represented by y and x, respectively. We calculate the gradient between two adjacent points in this array as in the following formula:

$$\boldsymbol{G}(\boldsymbol{i}) = \frac{(\boldsymbol{y}_{i+1} - \boldsymbol{y}_i)}{(\boldsymbol{x}_{i+1} - \boldsymbol{x}_i)},$$

where y_i is the coverage value and x_i is the position value of the i point in the 2D array. Gradient curve similarity index is the Pearson correlation coefficient (as mentioned above) between the cDNA and the input insert DNA $G(i)$ arrays.

p Value Correction

Bonferroni and Benjamini–Hochberg (BH) adjustments are applied to correct p values in our program.

Datasets

To test the performance of Fast-NR, we downloaded STARR-seq datasets for silencer identification in human K562 (GSE142207) (Doni Jayavelu et al., 2020) and for enhancer identification in untreated A549 cells (GSE114063) (Johnson et al., 2018), respectively. We downloaded histone modification datasets (H3K4me1 GSE91306; H3K27ac GSE91337; H3K4me3 GSE91218; H3K9me3 GSE91335) from ENCODE (Consortium, 2012). H3K27me3 (GSE75903) was downloaded from NCBI (Sayers et al., 2022). We mapped reads to human reference genome version hg19 (GRCh37) using bowtie2 (Langmead and Salzberg, 2012) with default parameters except "-p 24 -X 2000 --sensitive," then filtered and kept only reads with a value of MAPQ ≥20 by using samtools (Danecek et al., 2021). We kept only unique reads and discarded duplicates by using picardtools (Broad Institute, 2019) except for K562 STARR-seq datasets.

RESULTS

Performance of Fast-NR and Other Potential NRE Identification Methods Tested on Simulated Data

We simulated a pair of STARR-seq datasets to test the NRE identification powers of Fast-NR and other methods including csaw (Lun and Smyth, 2016), MEDIPS (Lienhard et al., 2014), PePr (Zhang et al., 2014), and CRADLE (Kim et al., 2021) (see **Supplementary Table S1**) using default parameters. We used the input insert reads, which are acquired by sequencing plasmids recovered from the transfected cells, and mapped them to chromosome 22 in the enhancer-screening STARR-seq experiment in A549 cells (GSE114063) as the simulation basis. Chromosome 22 is scanned, binned into 400bp windows and only genomic regions covered by at least 100 sequenced reads are kept. We selected 1,000 regions as simulated silencers (true positive silencers) by removing reads from these regions at four different percentage levels (30, 50, 70, and 90%), thus retained fraction of reads at 70, 50, 30, and 10% in the simulated datasets, respectively. These four datasets are used as the reporter cDNA libraries.

Program csaw detects differentially enriched regions in ChIP-seq dataset by using a sliding window strategy. MEDIPS identifies differentially methylated sites in the dataset generated by methylated DNA immunoprecipitation sequencing (MeDIP-seq). MEDIPS fails to detect any silencer at all from libraries in which 30, 50, and 70% reads are retained for the simulated silencers, while csaw identifies less than 100 silencers independent of what percentages of reads are retained (**Figure 2A**). These results suggest that MEDIPS and csaw may lack NRE detection power.

Program PePr is similar to csaw in their differential signal detection power for ChIP-seq datasets. PePr identifies most simulated silencers when reads are retained at three different levels (10, 30, and 50%) (**Figure 2A**), suggesting PePr could potentially be a usable NRE identification method. Program CRADLE calls both positive and negative regulatory elements for STARR-seq datasets. This program identifies approximately 800 silencers (815, 819, 821, and 812, respectively) independent of the percentages of reads retained (**Figure 2A**). Fast-NR detects 897, 871, 712, and 317 simulated silencers at 10, 30, 50, and 70% retained read levels (**Figure 2A**), suggesting its NRE identification power correlates positively with the read removal percentages. These results together suggest that PePr, CRADLE, and Fast-NR may all be usable NRE identification methods. Also, Fast-NR is more sensitive to signal reduction levels than other programs.

The NRE detection power of PePr, CRADLE, and Fast-NR may change when different p value thresholds are applied. Indeed, all these three methods detect fewer silencers as the p value threshold becomes more stringent (**Figure 2B**). Again, CRADLE is insensitive to the read removal percentage. Interestingly, though Fast-NR detects fewer silencers as p value decreases, it identifies more silencers than CRADLE when 10 and 30% of reads are retained. PePr is also sensitive to the change of p value threshold, especially when the fraction of reads retained is 70% (**Figure 2B**). These results show PePr and Fast-NR are more sensitive to the read retained rates than CRADLE. However, these results do not suggest which program outperforms the others.

Performance of Fast-NR and Other Potential NRE Identification Methods Tested on Real STARR-Seq Datasets

Theoretically, a genomic region of repressing activity is supposed to be transcribed less and underrepresented in the reporter cDNA library of STARR-seq. We downloaded STARR-seq datasets for silencer and enhancer identifications in human K562 (Doni Jayavelu et al., 2020) and A549 (Johnson et al., 2018) cells, respectively (**Supplementary Table S2**). STARR-seq in K562 measures the repressing activity of 7,430 sites in the accessible regions that are predicted as potential silencers based on epigenetic states. Differently, STARR-seq in A549 cells measures enhancer activities genome wide. We tested the NRE detection power of the five programs on the datasets generated by these two STARR-seq experiments (**Figure 3A**). Both Fast-NR and CRADLE identified hundreds and thousands NREs. Program csaw identified 2,399 silencers in K562 and only 31 silencers in A549. PePr identified 359 NREs in K562, but unbelievably, 475,797 NREs in A549. MEDIPS nearly failed to identify any NREs in the two STARR-seq experiments. These results confirm that MEDIPS lacks the NRE detection power for either simulated or real experimental data. Programs csaw and PePr perform differently on the two STARR-seq experiments, and their poor

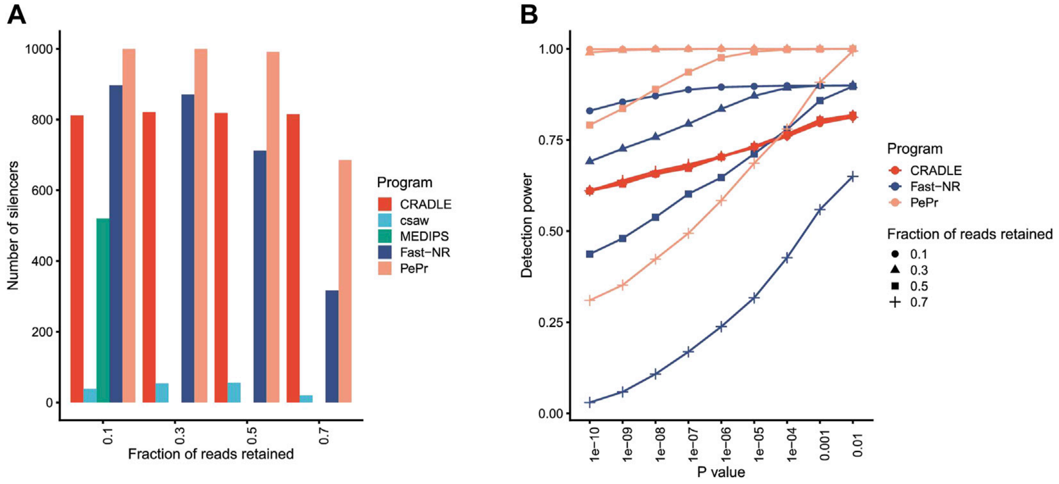

FIGURE 2 | Program performance comparison on simulated datasets. **(A)** The number of silencers identified by different programs. The fraction of reads retained to simulate silencers is shown under *x* axis. *p* value $<10^{-5}$. **(B)** The silencer detection power of different programs at different levels of confidence. Detection power is the ratio between number of identified silencers over total number of simulated silencers.

consistency in performance compromised our confidence to use them for NRE identification. After these comparisons, we kept Fast-NR and CRADLE for more evaluation analyses.

We compared Fast-NR and CRADLE's performance by changing the *p* value threshold for NRE identification. For K562 STARR-seq, Fast-NR identified silencers consistently at high levels and was nearly not affected by the change in the *p* value threshold (**Figure 3B**). In contrast, the number of silencers identified by CRADLE dropped dramatically to only about 10% ($p < 1 \times 10^{-8}$) of these identified at $p < 0.01$. These results provoked us to examine the overlapping rates of silencers identified by Fast-NR and CRADLE in K562. In fact, decent amounts of silencers identified by these two methods overlapped in K562 but not in A549 (**Figure 3C**). Silencers identified were expected to have lower cDNA reads than the input insert DNA reads. We calculated the ratios of (cDNA reads)/(insert reads) for CRADLE-specific and Fast-NR-specific silencers in K562 and A549. The CRADLE-specific silencers showed less reduction in cDNA reads than Fast-NR-specific silencers (**Figures 3D,E**). Over 95% (1,320/1,383) of Fast-NR identified silencers in K562 overlapped with the reported silencers (Doni Jayavelu et al., 2020) (**Figure 3F**). In contrast, only 56% silencers identified by CRADLE overlapped with the reported silencers (**Figure 3F**). These results suggest that many CRADLE-specific silencers were identified because of the heavy correction procedures that are integral to CRADLE (Kim et al., 2021). These CRADLE-specific silencers seemed to be “false-positives” in terms of the reduction rate in the reporter cDNA reads.

To reveal which transcription factors may bind to silencers, we searched through the sequences of silencers and identified a few DNA motifs enriched for transcription factors binding (**Supplementary Table S3**). One of these motifs was the silencing factor REST binding site (Chong et al., 1995), which was particularly enriched in Fast-NR identified silencers. DNA motif for transcription repressor PRDM6 was also enriched (Davis et al., 2006). Histone H4K20 methylation is a mark reported to be associated with silencers (Pang and Snyder, 2020). The binding motif (GC-box sequence) for the transcription factor of Sp1-like factors was also enriched in both Fast-NR and CRADLE silencers. Sp1-like factors activate or repress transcription in response to different physiological and pathological stimuli (Zhao and Meng, 2005). DNA motifs of PAX5 and FOS were enriched at Faste-NR and CATDLE silencers as well. Many transcription factors have dual functional roles in gene regulation, and silencers have been reported to be switchable to enhancers during development and in different cell types (Bessis et al., 1997; Cavalli, 2014; Gisselbrecht et al., 2020). Enrichment of any specific transcription factor's binding motif may not necessarily correlate with the regulatory activity of a CRE in a specific cell type. Nevertheless, silencers are indeed enriched with certain DNA motifs for transcription repressors in our analysis, suggesting that silencers identified by Fast-NR are very likely to be biologically functional.

Curve Similarity Analyses in Fast-NR

Compared to the other four methods, Fast-NR is the only one that takes into account the similarity between the curves of the reporter cDNA and the input insert DNA signals. We examined to what extent the similarity between the curves of the reporter cDNA and the insert DNA reads could affect the NRE identification. We calculated the similarity index values ($-\log_2$CosineDistance) for the NREs identified in A549 and found that the cosine distances between cDNA and plasmid curves are much higher than the random chosen genomic control regions (**Figure 4A**). Interestingly, the similarity index values correlate negatively with the strengths of silencers (**Figure 4B**), suggesting stronger silencers have low curve similarity. We obtained similar results using other curve similarity calculation methods such as Pearson, Euclidean, and gradient (**Supplementary Figure S1A**).

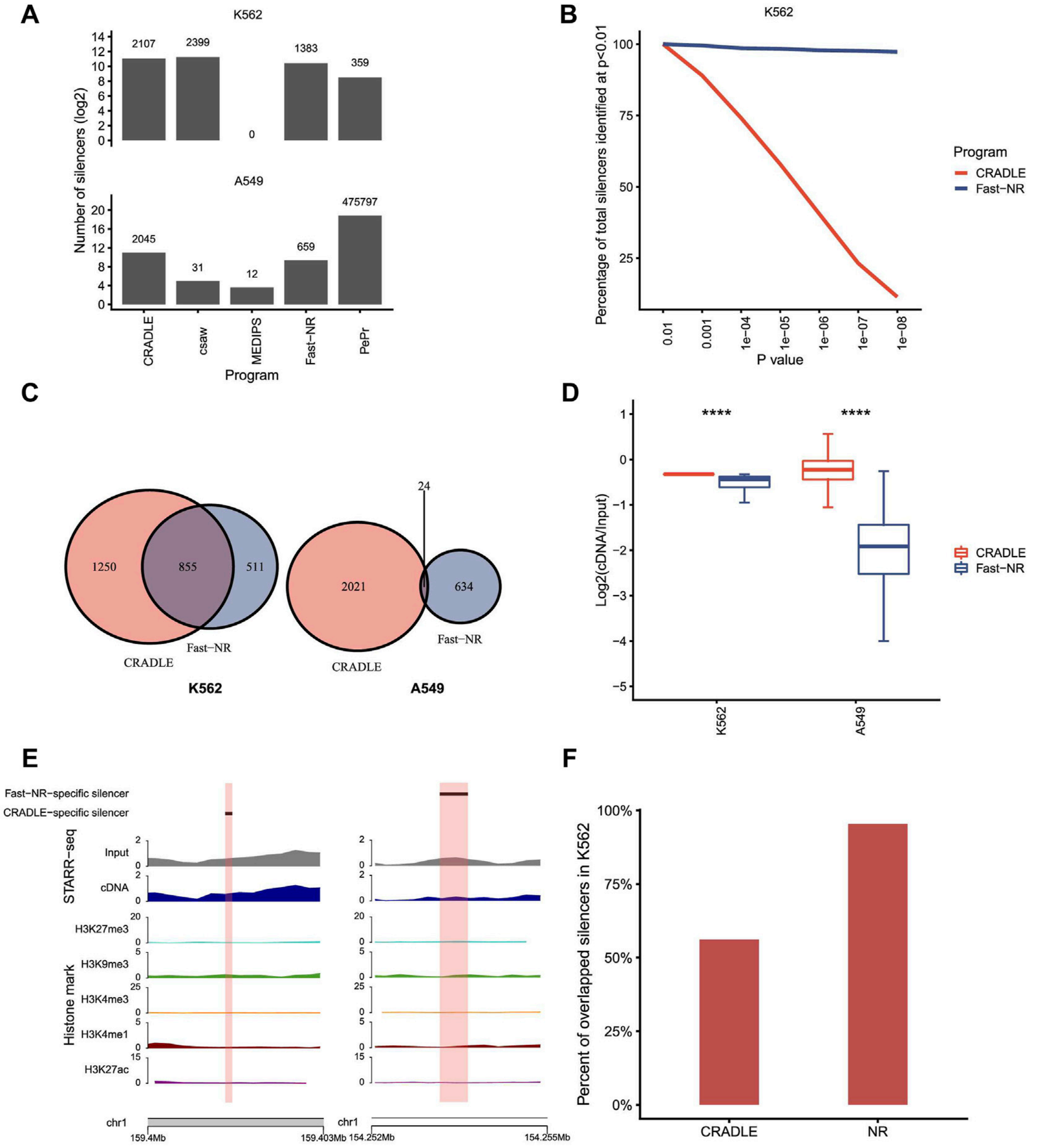

FIGURE 3 | Program performance comparison on STARR-seq datasets. **(A)** The number of silencers identified by different programs. p value $<10^{-5}$. **(B)** The percentage of silencers identified by CRADLE and Fast-NR at different confidence levels. The number of silencers identified at $p < 10^{-2}$ is set at 100%. **(C)** Venn diagrams of silencers identified by Fast-NR and CRADLE in K562 and A549, respectively. p value $<10^{-5}$. **(D)** Reads ratio (reporter cDNA/input inserts) distribution for silencers identified only by CRADLE or Fast-NR in K562 and A549, respectively. **(E)** Exemplary silencers identified only by CRADLE (left) or by Fast-NR (right). **(F)** Percentages of Fast-NR- and CRADLE-identified silencers ($p < 10^{-5}$) reported by Doni Jayavelu et al.

We removed the curve similarity requirement in Fast-NR and identified more silencers (gray dots in **Figure 4B**). These “new” silencers have high curve similarity index values and low silencer strengths compared to the silencers identified with curve similarity considered (blue dots in **Figure 4B**). Interestingly, curve similarity correlated poorly with p values (**Supplementary Figure S1B**), suggesting curve similarity in Fast-NR is a feature independent from the ratio between the reporter cDNA and the input insert DNA reads. The Pearson’s correlation coefficients between the curve similarities and the silencer activities could be positive or negative depending on the method used (**Supplementary Figure S1C**). These results together show that curve similarity comparison is also important for the reliable identification of NREs.

DISCUSSION

In this study, we presented a program of Fast-NR, in which both read counts and shape similarity are considered, for the detection

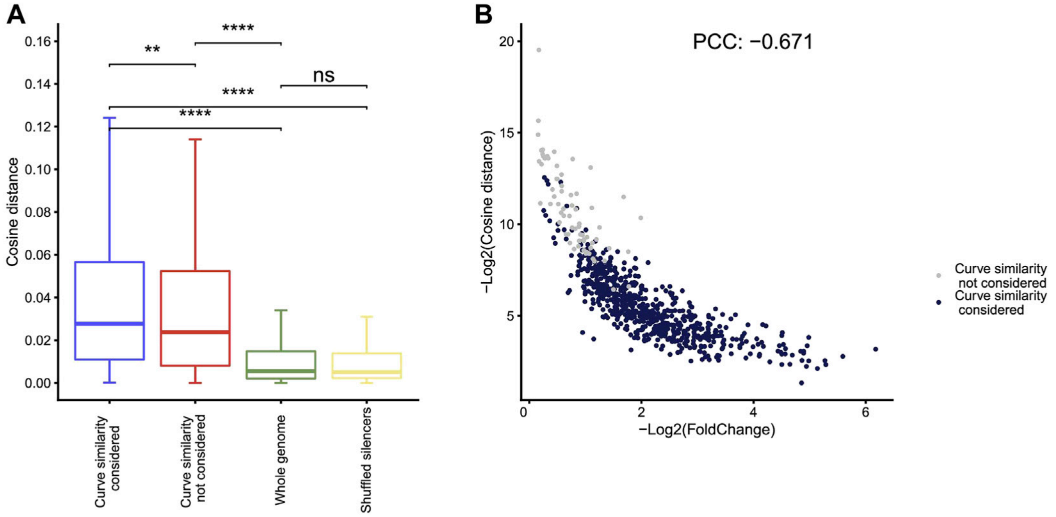

FIGURE 4 | Curve similarity effect on silencer identification. **(A)** The cosine distance distribution for silencers identified by Fast-NR with similarity considered, similarity not considered, and controls of whole genome regions with 400 bp size and shuffled silencer regions. Distance negatively correlates with similarity. $**p < 10^{-3}$, $***p < 10^{-4}$, Wilcoxon rank sum test. **(B)** The correlation between silencer strength and curve similarity. The *X* axis shows the value of $-\log_2$ (cDNA reads/insert DNA reads). The *Y* axis shows the curve similarity index, $-\log_2$ (Cosine distance) of silencers calculated by the method of cosine. Blue dots are silencers that pass curve similarity threshold of 0.9, and gray dots are silencers that do not pass curve similarity threshold.

of NREs using STARR-seq datasets and compare its performance with other four programs of csaw (Lun and Smyth, 2016), MEDIPS (Lienhard et al., 2014), PePr (Zhang et al., 2014), and CRADLE (Kim et al., 2021). Among them, MEDIPS, designed for DNA methylation analysis, shows worst compatibility with silencer identification on either simulated or experimentally generated datasets. Programs csaw and PePr detect significantly differential regions in ChIP-seq data. Neither of them performs consistently when being applied to different types of datasets. Besides methods tested in this research, other methods designed for the identification of differentially enriched signals are not suitable for silencer identification either. For example, DMRfinder (Gaspar and Hart, 2017), DSS (Park and Wu, 2016), and HMST-Seq-Analyzer (Farooq et al., 2020) require specific input data format that are not compatible with NRE analysis.

Fast-NR and CRADLE seem to be good choices for both simulated and experimentally generated datasets. However, many silencers identified by CRADLE showed only small reduction in the reporter cDNA signal than the input insert DNA, and curves of these signals were highly similar. CRADLE uses the GLM approach to correct four types of bias, the DNA structure affecting shear force, Gibbs free-energy affecting PCR efficiency, read sequences mappability, and G-quadruplex affecting DNA polymerase processivity (Kim et al., 2021). CRADLE treats the corrected signals as normal distribution and uses Welch's *t*-test to search for differences. As shown in our analysis, these corrections lead to the detection of "silencers" that cannot be identified based on the differences in the read counts of the reporter cDNA and the input insert DNA.

Being different from methods using sliding window strategy, Fast-NR detects the difference in the number of reporter cDNA and input insert DNA reads at single base-resolution. It is potentially possible to use Fast-NR to reveal the precise locations of regulatory elements and the binding sites of transcription factors.

STARR-seq tests silencers activity in episomal reporter plasmids independent from the endogenous chromatin environment. Ideally, regulatory activities of potential CREs can be tested in their proper chromatin context. Methods for endogenous CREs analysis, such as multiplexed editing regulatory assay (MERA) (Rajagopal et al., 2016) and thousands of reporters integrated in parallel (TRIP) (Akhtar et al., 2013), can be used to measure the regulatory activities of genomic regions in the native cellular context. However, these methods are generally not applicable for unbiased genome-wide analysis of CREs. Nevertheless, the combination of these methods and STARR-seq will help to achieve a global identification, and at the same time, a large scale endogenous validation of CREs.

Another issue we would like to point out is the promoter used in the reporter plasmids. In STARR-seq and related methods, promoter choice could affect the outcome because the promoter used may be irresponsive to some CREs. We speculate that using promoters of house-keeping genes and cell type-specific genes may allow the identification of more CREs that may prefer to regulate different types of promoters. To save the computation time, we filtered potential NREs by applying thresholds on both read counts and *p* values sequentially, which may also, theoretically, reduce false positive rate. However, the thresholds applied could be too strict and exclude some true silencers. We recommend the users to test the threshold effects and choose appropriate thresholds for their own analysis.

Though STARR-seq measures regulatory activity of tested DNA fragment in episomal environment, it provides a catalogue of CREs that can be further tested at their endogenous loci by alternative methods. We did not take sequencing bias into consideration because experimental data that can be used to determine to what extent biases may affect NRE identification were not available. In summary, by combining read count-based negative binomial test and shape similarity comparison, we have shown that Fast-NR is potentially usable for silencer identification, thus providing a powerful and robust computational method for NRE identification.

CONCLUSION

Silencers are negative regulatory elements that control the precise gene expression during cell proliferation and differentiation. The increasing needs for global silencer characterization require a reliable and user-friendly computational method. Our method Fast-NR integrates single nucleotide read count information and graphic information to detect silencers genome widely. Fast-NR identifies NREs at single base resolution. The wide application of Fast-NR will accelerate the genetic and epigenetic studies of the intriguing functional mechanisms of silencers.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

NH conceived the study. NH and WW designed the program. NH analyzed the data. CF, YT, and LL participated in the program development. CH supervised the project. NH and CH wrote the article.

FUNDING

We acknowledge financial supports from the National Key R&D Program of China (no. 2018YFC1004500), Shenzhen Science and Technology Innovation Commission (no. 20200925153547003), and support from the Center for Computational Science and Engineering of Southern University of Science and Technology.

ACKNOWLEDGMENTS

We thank Dr. Longjian Niu for technical assistance and Dr. Kunfeng Bai for comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.818344/full#supplementary-material

REFERENCES

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., et al. (2013). Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell* 154, 914–927. doi:10.1016/j.cell.2013.07.018

Anders, S., and Huber, W. (2010). Differential Expression Analysis for Sequence Count Data. *Genome Biol.* 11, R106. doi:10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python Framework to Work with High-Throughput Sequencing Data. *Bioinformatics* 31, 166–169. doi:10.1093/bioinformatics/btu638

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide Quantitative Enhancer Activity Maps Identified by STARR-Seq. *Science* 339, 1074–1077. doi:10.1126/science.1232542

Bessis, A., Champtiaux, N., Chatelin, L., and Changeux, J.-P. (1997). The Neuron-Restrictive Silencer Element: a Dual Enhancer/silencer Crucial for Patterned Expression of a Nicotinic Receptor Gene in the Brain. *Proc. Natl. Acad. Sci.* 94, 5906–5911. doi:10.1073/pnas.94.11.5906

Broad Institute (2019). Available at: http://broadinstitute.github.io/picard/ (Accessed June 24, 2020).

Cavalli, G. (2014). A RING to Rule Them All: RING1 as Silencer and Activator. *Dev. Cel* 28, 1–2. doi:10.1016/j.devcel.2013.12.015

Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., et al. (1995). REST: A Mammalian Silencer Protein that Restricts Sodium Channel Gene Expression to Neurons. *Cell* 80, 949–957. doi:10.1016/0092-8674(95)90298-8

Consortium, E. P. (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* 489, 57–74. doi:10.1038/nature11247

Cremona, M. A., Xu, H., Makova, K. D., Reimherr, M., Chiaromonte, F., and Madrigal, P. (2019). Functional Data Analysis for Computational Biology. *Bioinformatics* 35, 3211–3213. doi:10.1093/bioinformatics/btz045

Crocker, J., and Stern, D. L. (2013). TALE-mediated Modulation of Transcriptional Enhancers *In Vivo*. *Nat. Methods* 10, 762–767. doi:10.1038/nmeth.2543

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve Years of SAMtools and BCFtools. *Gigascience* 10, giab008. doi:10.1093/gigascience/giab008

Davis, C. A., Haberland, M., Arnold, M. A., Sutherland, L. B., McDonald, O. G., Richardson, J. A., et al. (2006). PRISM/PRDM6, a Transcriptional Repressor that Promotes the Proliferative Gene Program in Smooth Muscle Cells. *Mol. Cel Biol* 26, 2626–2636. doi:10.1128/mcb.26.7.2626-2636.2006

Della Rosa, M., and Spivakov, M. (2020). Silencers in the Spotlight. *Nat. Genet.* 52, 244–245. doi:10.1038/s41588-020-0583-8

Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R. D. (2020). Candidate Silencer Elements for the Human and Mouse Genomes. *Nat. Commun.* 11, 1061. doi:10.1038/s41467-020-14853-5

Farooq, A., Grønmyr, S., Ali, O., Rognes, T., Scheffler, K., Bjørås, M., et al. (2020). HMST-Seq-Analyzer: A New python Tool for Differential Methylation and Hydroxymethylation Analysis in Various DNA Methylation Sequencing Data. *Comput. Struct. Biotechnol. J.* 18, 2877–2889. doi:10.1016/j.csbj.2020.09.038

Gaspar, J. M., and Hart, R. P. (2017). DMRfinder: Efficiently Identifying Differentially Methylated Regions from MethylC-Seq Data. *BMC Bioinformatics* 18, 528. doi:10.1186/s12859-017-1909-0

Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmSVM: an R Package for Gapped-Kmer SVM. *Bioinformatics* 32, 2205–2207. doi:10.1093/bioinformatics/btw203

Gisselbrecht, S. S., Barrera, L. A., Porsch, M., Aboukhalil, A., Estep, P. W., 3rd, Vedenko, A., et al. (2013). Highly Parallel Assays of Tissue-specific Enhancers in Whole Drosophila Embryos. *Nat. Methods* 10, 774–780. doi:10.1038/nmeth.2558

Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., et al. (2020). Transcriptional Silencers in Drosophila Serve a Dual Role as

Transcriptional Enhancers in Alternate Cellular Contexts. *Mol. Cel* 77, 324–337. doi:10.1016/j.molcel.2019.10.004

Haberle, V., and Stark, A. (2018). Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation. *Nat. Rev. Mol. Cel Biol* 19, 621–637. doi:10.1038/s41580-018-0028-8

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cel* 38, 576–589. doi:10.1016/j.molcel.2010.05.004

Hower, V., Evans, S. N., and Pachter, L. (2011). Shape-based Peak Identification for ChIP-Seq. *BMC Bioinformatics* 12, 15. doi:10.1186/1471-2105-12-15

Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L., and Ovcharenko, I. (2019). Identification of Human Silencers by Correlating Cross-Tissue Epigenetic Profiles and Gene Expression. *Genome Res.* 29, 657–667. doi:10.1101/gr.247007.118

Johnson, G. D., Barrera, A., McDowell, I. C., D'Ippolito, A. M., Majoros, W. H., Vockley, C. M., et al. (2018). Human Genome-wide Measurement of Drug-Responsive Regulatory Activity. *Nat. Commun.* 9, 5317. doi:10.1038/s41467-018-07607-x

Kim, Y.-S., Johnson, G. D., Seo, J., Barrera, A., Cowart, T. N., Majoros, W. H., et al. (2021). Correcting Signal Biases and Detecting Regulatory Elements in STARR-Seq Data. *Genome Res.* 31, 877–889. doi:10.1101/gr.269209.120

Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., et al. (2020). STARRPeaker: Uniform Processing and Accurate Identification of STARR-Seq Active Regions. *Genome Biol.* 21, 298. doi:10.1186/s13059-020-02194-x

Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., et al. (2014). MAGeCK Enables Robust Identification of Essential Genes from Genome-Scale CRISPR/Cas9 Knockout Screens. *Genome Biol.* 15, 554. doi:10.1186/s13059-014-0554-4

Lienhard, M., Grimm, C., Morkel, M., Herwig, R., and Chavez, L. (2014). MEDIPS: Genome-wide Differential Coverage Analysis of Sequencing Data Derived from DNA Enrichment Experiments. *Bioinformatics* 30, 284–286. doi:10.1093/bioinformatics/btt650

Lun, A. T. L., and Smyth, G. K. (2016). Csaw: a Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows. *Nucleic Acids Res.* 44, e45. doi:10.1093/nar/gkv1191

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genom. Hum. Genet.* 7, 29–59. doi:10.1146/annurev.genom.7.080505.115623

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., et al. (2012). Systematic Dissection and Optimization of Inducible Enhancers in Human Cells Using a Massively Parallel Reporter Assay. *Nat. Biotechnol.* 30, 271–277. doi:10.1038/nbt.2137

Mogno, I., Kwasnieski, J. C., and Cohen, B. A. (2013). Massively Parallel Synthetic Promoter Assays Reveal the *In Vivo* Effects of Binding Site Variants. *Genome Res.* 23, 1908–1915. doi:10.1101/gr.157891.113

Ogbourne, S., and Antalis, T. M. (1998). Transcriptional Control and the Role of Silencers in Transcriptional Regulation in Eukaryotes. *Biochem. J.* 331, 1–14.

Pang, B., and Snyder, M. P. (2020). Systematic Identification of Silencers in Human Cells. *Nat. Genet.* 52, 254–263. doi:10.1038/s41588-020-0578-5

Park, Y., and Wu, H. (2016). Differential Methylation Analysis for BS-Seq Data under General Experimental Design. *Bioinformatics* 32, 1446–1453. doi:10.1093/bioinformatics/btw026

Petrykowska, H. M., Vockley, C. M., and Elnitski, L. (2008). Detection and Characterization of Silencers and Enhancer-Blockers in the Greater CFTR Locus. *Genome Res.* 18, 1238–1246. doi:10.1101/gr.073817.107

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., et al. (2016). High-throughput Mapping of Regulatory DNA. *Nat. Biotechnol.* 34, 167–174. doi:10.1038/nbt.3468

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., and Comeau, D. C. (2022). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 50, D20–D26. doi:10.1093/nar/gkab1112

Schoenfelder, S., and Fraser, P. (2019). Long-range Enhancer-Promoter Contacts in Gene Expression Control. *Nat. Rev. Genet.* 20, 437–455. doi:10.1038/s41576-019-0128-0

Shen, L., Shao, N.-Y., Liu, X., Maze, I., Feng, J., and Nestler, E. J. (2013). diffReps: Detecting Differential Chromatin Modification Sites from ChIP-Seq Data with Biological Replicates. *PLoS One* 8, e65598. doi:10.1371/journal.pone.0065598

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional Enhancers: from Properties to Genome-wide Predictions. *Nat. Rev. Genet.* 15, 272–286. doi:10.1038/nrg3682

Strino, F., and Lappe, M. (2016). Identifying Peaks in *-seq Data Using Shape Information. *BMC Bioinformatics* 17 (Suppl. 5), 206. doi:10.1186/s12859-016-1042-5

Sun, J., He, N., Niu, L., Huang, Y., Shen, W., Zhang, Y., et al. (2019). Global Quantitative Mapping of Enhancers in Rice by STARR-Seq. *Genomics, Proteomics & Bioinformatics* 17, 140–153. doi:10.1016/j.gpb.2018.11.003

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that Define the Best ChIP-Seq Peak Calling Algorithms. *Brief Bioinform* 18, 441–450. doi:10.1093/bib/bbw035

Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., et al. (2015). High-throughput and Quantitative Assessment of Enhancer Activity in Mammals by CapStarr-Seq. *Nat. Commun.* 6, 6905. doi:10.1038/ncomms7905

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., et al. (2018). High-resolution Genome-wide Functional Dissection of Transcriptional Regulatory Regions and Nucleotides in Human. *Nat. Commun.* 9, 5380. doi:10.1038/s41467-018-07746-1

Wu, H., and Ji, H. (2014). PolyaPeak: Detecting Transcription Factor Binding Sites from ChIP-Seq Using Peak Shape Information. *PLoS One* 9, e89694. doi:10.1371/journal.pone.0089694

Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From Reads to Insight: a Hitchhiker's Guide to ATAC-Seq Data Analysis. *Genome Biol.* 21, 22. doi:10.1186/s13059-020-1929-3

Zeng, W., Chen, S., Cui, X., Chen, X., Gao, Z., and Jiang, R. (2021). SilencerDB: a Comprehensive Database of Silencers. *Nucleic Acids Res.* 49, D221–D228. doi:10.1093/nar/gkaa839

Zhang, X., Robertson, G., Krzywinski, M., Ning, K., Droit, A., Jones, S., et al. (2011). PICS: Probabilistic Inference for ChIP-Seq. *Biometrics* 67, 151–163. doi:10.1111/j.1541-0420.2010.01441.x

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S., and Sartor, M. A. (2014). PePr: a Peak-Calling Prioritization Pipeline to Identify Consistent or Differential Peaks from Replicated ChIP-Seq Data. *Bioinformatics* 30, 2568–2575. doi:10.1093/bioinformatics/btu372

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi:10.1186/gb-2008-9-9-r137

Zhao, C., and Meng, A. (2005). Sp1-like Transcription Factors Are Regulators of Embryonic Development in Vertebrates. *Dev. Growth Differ.* 47, 201–211. doi:10.1111/j.1440-169x.2005.00797.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.