Chapter 20

ChIP-Seq Data Analysis: Identification of Protein–DNA Binding Sites with SISSRs Peak-Finder

Leelavati Narlikar and Raja Jothi

Abstract

Protein–DNA interactions play key roles in determining gene-expression programs during cellular development and differentiation. Chromatin immunoprecipitation (ChIP) is the most widely used assay for probing such interactions. With recent advances in sequencing technology, ChIP-Seq, an approach that combines ChIP and next-generation parallel sequencing is fast becoming the method of choice for mapping protein–DNA interactions on a genome-wide scale. Here, we briefly review the ChIP-Seq approach for mapping protein–DNA interactions and describe the use of the SISSRs peak-finder, a software tool for precise identification of protein–DNA binding sites from sequencing data generated using ChIP-Seq.

Key words: ChIP-Seq, SISSRs, Protein–DNA interaction, Binding sites, Transcription factor, Next-generation sequencing, Genomics

1. Introduction

DNA-binding proteins are essential for the proper functioning of several cellular processes such as transcriptional regulation, which is primarily mediated by interactions between proteins called transcription factors and specific regions on the DNA. These interactions play key roles in determining gene-expression programs during development, differentiation, proliferation, and lineage-specification (1–5). Besides regulating transcription, DNA-binding proteins are essential for DNA replication (6), DNA repair (7), and chromosomal stability (8). Identification of regions targeted by such proteins is therefore crucial for a better understanding of these cellular processes.

Originally developed to investigate protein–DNA binding at a Drosophila locus (9), chromatin immunoprecipitation (ChIP) has become the most widely used assay for determining DNA regions

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978–1-61779–400–1_20,

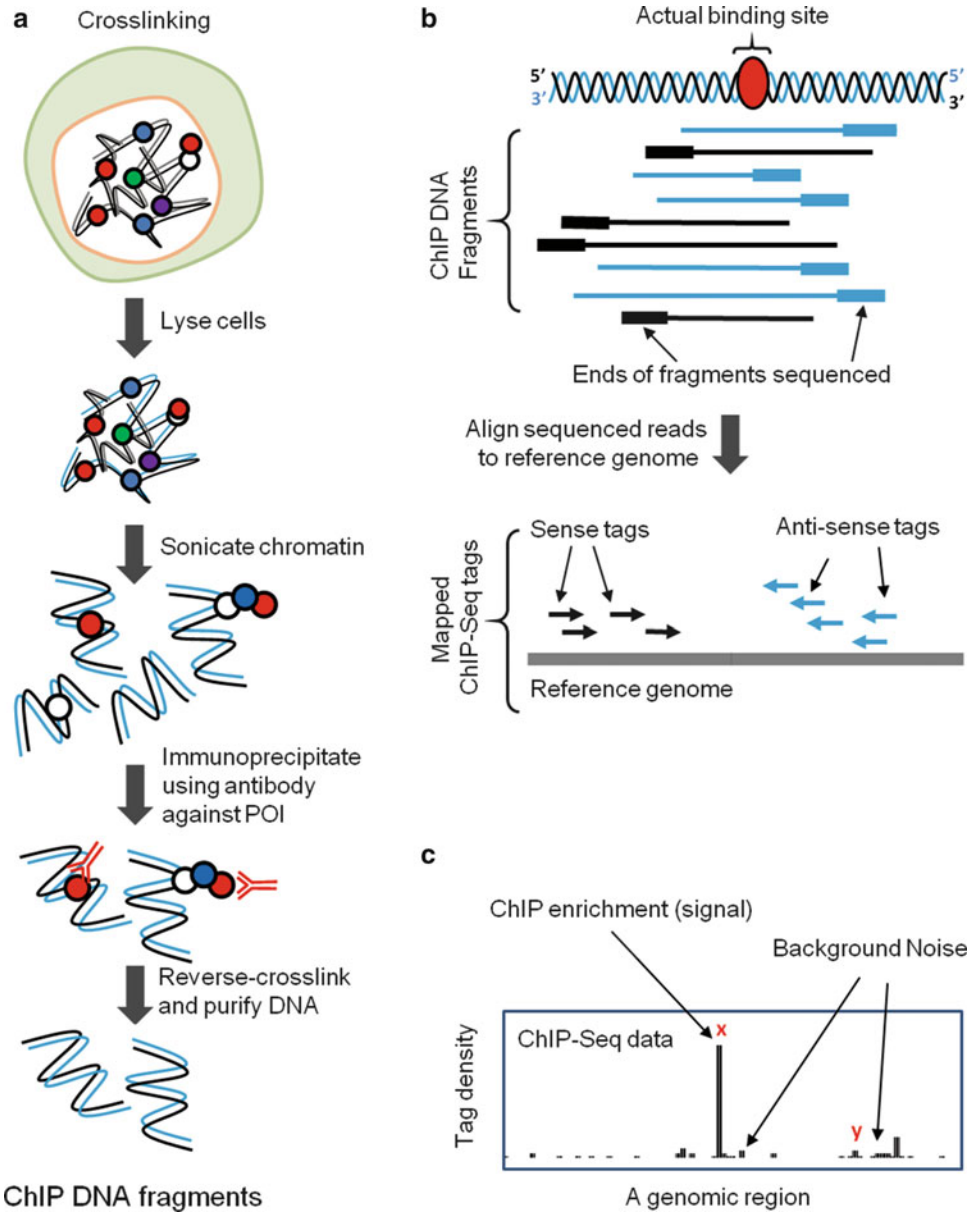

Fig. 1. ChIP-Seq experiment and data. (**a**) Steps involved in chromatin immunoprecipitation (ChIP). Proteins are represented as *circles*. The antibody used in the immunoprecipitation step is represented as a Y-shaped structure. (**b**) Ends of DNA fragments obtained from ChIP are sequenced and aligned back to the reference genome (*arrows* represent the sequenced portion of the ChIP DNA fragment). (**c**) Tags mapped to a genomic region are visualized as a histogram of tag density. Regions with signal and noise are marked with *x* and *y*, respectively.

bound by the protein of interest (POI) in vivo. In this assay, protein–DNA and protein–protein interactions are first crosslinked by treating living cells with formaldehyde (Fig. 1a). This crosslinking step can be omitted in case of proteins such as histones that stably bind DNA. Next, the crosslinked cells are lysed

and then sonicated – a process in which ultrasonic waves are used to shear the chromatin into short fragments of desired length (~0.2–0.5 kb). The sheared chromatin is then immunoprecipitated with a specific antibody against the POI. The antibody may not necessarily target only direct POI–DNA complexes but also those complexes where the POI is indirectly bound to the DNA via its interaction with another protein or protein complex (Fig. 1a). The immunoprecipitated protein–DNA crosslinks are reversed, and the DNA is purified for downstream assays designed to characterize the sequences bound by the POI.

Traditionally, PCR or quantitative/real-time PCR (qPCR) with primers designed to probe regions of interest are used to detect and quantify ChIP-derived DNA in relation to a control input DNA, which is obtained the same way as the ChIP DNA but without the immunoprecipitation step. Although ChIP-qPCR still remains the gold-standard assay for quantifying specific protein–DNA interactions, the necessity to design primers for every region to be probed makes it ill-suited for profiling protein–DNA interactions on a large scale. ChIP-chip (10), an approach that combines ChIP with DNA microarrays, was the most widely used technique for mapping protein–DNA interactions on a global scale until recently (11, 12). Advances in sequencing technology have enabled millions of short DNA fragments to be sequenced within a day or two in a cost-effective manner. These sequences can then be aligned back to the reference genome to determine the source of origin. This is exploited in ChIP-Seq (13–17), where ChIP is combined with next-generation massively parallel sequencing technology to identify DNA regions bound by the POI. Its superior coverage and resolution have resulted in ChIP-Seq replacing ChIP-chip as the method of choice. Readers are referred to ref. 18, 19 for a detailed review on ChIP-Seq.

In ChIP-Seq, ChIP-derived DNA fragments are directly sequenced on a next-generation sequencing platform. Although the length of ChIP DNA fragments can range anywhere between a few hundred and a few thousand nucleotides, sequencing just ~25–75 nucleotides from the ends of the DNA fragments is sufficient to align/map the fragments back to unique locations in the reference genome (Fig. 1b). Bowtie (20), MAQ (21), and ELAND from Illumina are popular tools for aligning short sequence reads back to the reference genome. During the alignment process, reads that map to multiple locations in the reference genome are discarded and only those reads that map to unique genomic locations are retained. Such reads are commonly referred to as tags. Henceforth, "reads" and "tags" are used interchangeably.

The first step in interpreting a ChIP-Seq dataset involves identifying regions bound by (or associated with) the POI using the mapped tags. Hereafter, we will refer to these regions as binding sites/regions. Regions with higher tag densities

compared to the background "noise" are typically good binding site candidates (site *x* compared to site *y* in Fig. 1c). In theory, only the regions bound by the POI are expected to have tags associated with them since these would be the regions immunoprecipitated and sequenced (Fig. 1a, b). In practice, however, sequencing errors can cause some of the incorrectly sequenced reads to get mapped to regions that were not immunoprecipitated, resulting in background noise tags at these regions (Fig. 1c; see Note 1). Noise in the data could also be due to biological reasons, primarily stemming from antibodies that are not specific to the POI. For instance, nonspecific antibodies targeting additional proteins can result in ChIP-derived DNA fragments that bind one of these proteins and not the POI. Since this type of noise is difficult to detect postsequencing, pre-ChIP experiments are typically performed to confirm antibody specificity.

Issues outlined above highlight the need for a systematic approach for the precise identification of binding sites from ChIP-Seq data. Such an approach must not only identify regions bound by the POI but also filter out false-positive regions by evaluating the test dataset (obtained from ChIP DNA) against a control dataset obtained from input DNA or IgG ChIP (see Note 2). In this chapter, we describe a widely used method called SISSRs (22), a peak-finder that leverages the direction of ChIP-Seq tags (mapped to sense/antisense strands) to identify binding sites at a high resolution, typically within few tens of base pairs. We provide a detailed description of the SISSRs software application tool and instructions for using it effectively to identify protein–DNA binding sites from data generated using ChIP-Seq.

2. Methods

2.1. SISSRs Algorithm

SISSRs, short for *S*ite *I*dentification from *S*hort *S*equence *R*ead*s*, is a peak-finder algorithm that uses the direction and density of mapped ChIP-Seq tags along with the average length (F) of sequenced DNA fragments to identify protein–DNA binding sites (see Note 3; Fig. 2a). If the user does not know the average fragment length of the ChIP DNA, SISSRs can estimate F from the tags within the dataset (see ref. 22 for details). SISSRs begins by scanning regions mapped with sequence tags in the test data using a sliding window of size w nucleotides with consecutive windows overlapping by $w/2$. For a region i spanned by the sliding window, a measure called "net-tag count" (c_i) is computed by subtracting the number of tags mapped to the antisense strand of i (antisense tags) from the number of tags mapped to the sense strand of i (sense tags). As the window slides along, whenever the

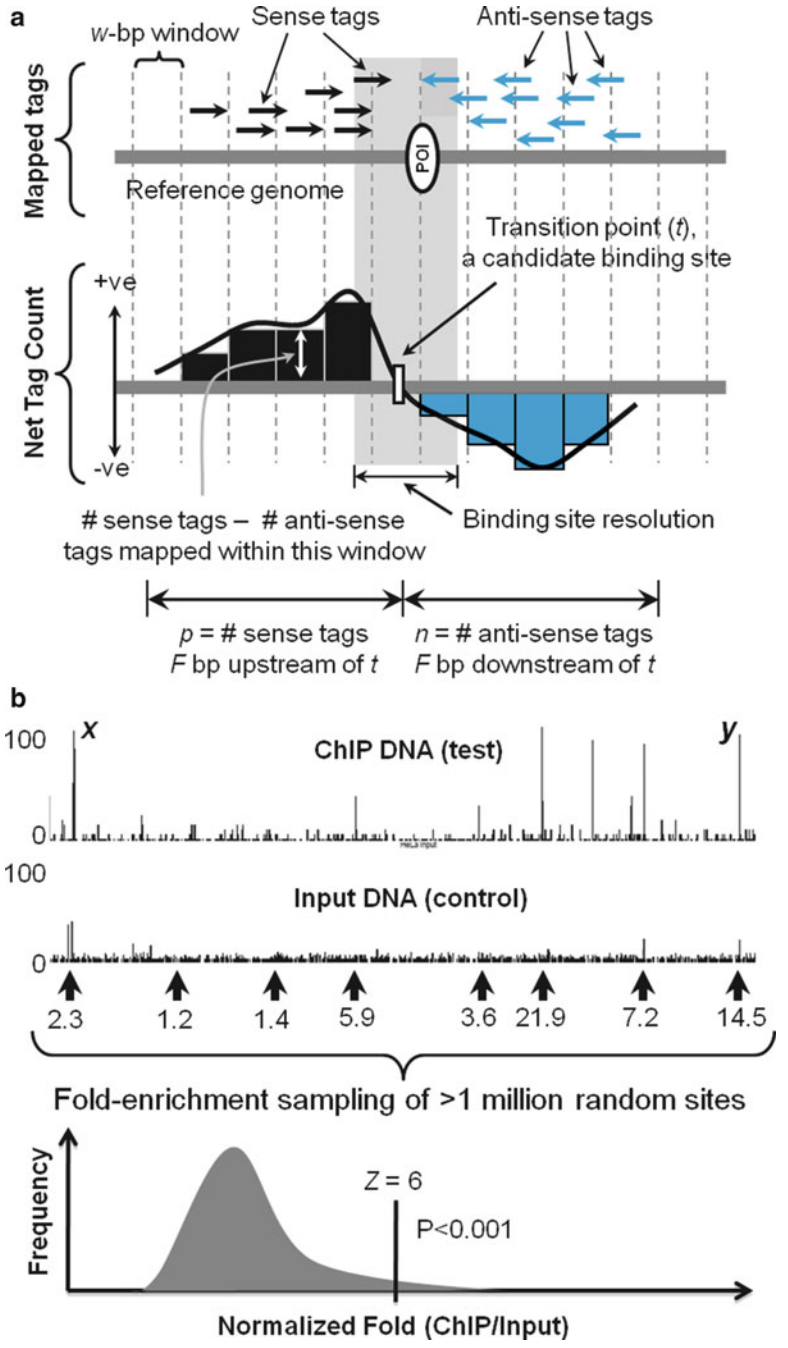

Fig. 2. SISSRs algorithm. (**a**) Typical distribution of tags mapped to sense and antisense strands of a region ChIP-seq using an antibody against the protein of interest (POI), and a schematic showing candidate binding site identification using the direction and density of tags mapped to sense and antisense strands. (**b**) Illustration of how candidate binding sites identified from a test dataset are evaluated against the control dataset to determine the true binding sites. Distribution of fold-enrichment, defined as ratio of the number of tags within a 2*F* bp long region in the test dataset to that within the same region in the control dataset, computed for over one million random sites is used to determine the empirical *p*-values for candidate binding sites. Only those candidate sites with fold-enrichment value greater than or equal to the smallest fold threshold *Z* (with *p*-value not greater than the user-set threshold) are reported as true binding sites. For $Z = 6$, candidate site *y* with 14.5-fold enrichment will be reported as a true binding site, whereas site *x* with a similar ChIP signal but with a smaller fold enrichment over the control (2.3-fold) will not be reported as a true site.

net-tag count transitions from a positive to a negative value, the corresponding transition point marked by genomic coordinate t is recorded as a candidate binding site. Only those candidate binding sites satisfying the following set of conditions are retained and designated as true binding sites.

1. Number of sense tags (p) within the F bp region upstream of t is at least E.
2. Number of antisense tags (n) within the F bp downstream of t is at least E.
3. The sum of p and n is at least R, which is estimated based on a user-defined false discovery rate (FDR) D (when no control dataset is available) or e-value threshold (when a control dataset is provided).
4. The fold-enrichment, defined as the ratio of the number of tags supporting the candidate site in the test data ($p + n$) to the number of tags supporting the exact same site in the control data, is at least Z, which is determined based on an empirical distribution of fold-enrichment values of at least a million randomly selected sites and a chosen p-value threshold (Fig. 2b; see Note 4).

Condition 4 applies only when a control dataset is available to evaluate the enrichment of tags supporting the binding site in the test versus the control. When no control dataset is available, the background tag distribution is modeled using a Poisson distribution.

E is set to 2 by default and can be changed by the user. The value of R is estimated as follows. The FDR is defined as the ratio of the number of $2F$-bp long regions with V or more tags that the background model indicates should occur by chance (e^V) to the number observed in the real data. If no control dataset is available, R is equal to the smallest V corresponding to FDR $< D$, otherwise R is equal to the largest V such that $e^V < e$. The expected number of tags (λ) within a window of length $2F$ bp is given by $2F$ times the number of tags in the dataset divided by the mappable genome length M (which is roughly 0.8 times the actual genome length for the human and mouse genomes). The probability of observing a binding site supported by at least R tags by chance is given by a sum of Poisson probabilities as $1 - \sum_{n=0}^{R-1} (e^{-\lambda}\lambda^n)/n!$ SISSRs allows users to set their own values for all of the parameters discussed above. This provides the users the leverage to control sensitivity, specificity, resolution, and noise subtraction.

Identified binding sites are reported by their chromosomal coordinates (e.g., chr1:123450–123490). The resolution of each reported binding site is essentially the distance between the sense tag immediately upstream of the identified site and the antisense tag immediately downstream of this site (Fig. 2a; see Note 5). For additional details on the SISSRs algorithm, the reader may refer to ref. 22.

2.2. Identification of Protein–DNA Binding Sites Using SISSRs

This section gives detailed instructions for installing and using SISSRs on a ChIP-Seq dataset.

2.2.1. Getting and Installing SISSRs

A perl implementation of the SISSRs peak-finding algorithm is freely available at refs. 23, 24. Users with Linux operating system (or most UNIX systems, including Mac OS X) typically have an installation of perl. Users with other operating systems can download the latest version of perl for free using ref. 25. After downloading the SISSRs zipped archive, users should save the extracted sissrs.pl executable either onto their working directory (to run it from the working directory) or to a directory containing executables (to enable execution of sissrs.pl from anywhere within the home directory).

2.2.2. Preparing the Input Data Files

SISSRs takes as input data file(s) containing genomic coordinates of the mapped reads or tags in BED file format (26). In BED file format, each line contains six tab-separated terms as follows:

```
chr1    1234501    1234536    U0    1    +
chr2    9876540    9876585    U0    1    -
chr7    4567825    4567860    U0    1    -
chr3    1472585    1472620    U0    1    +
...
...
```

The first term denotes the chromosome, and the second and third terms denote the chromosomal start and end coordinates of the mapped read, respectively. The sixth term denotes the DNA strand onto which the read was mapped (+ and – for sense and antisense strand, respectively). The fourth and the fifth terms are not used by SISSRs.

2.2.3. Running SISSRs

Typing the name of the executable (sissrs.pl or ./sissrs.pl or perl sissrsl.pl) on the command line displays the help menu. A simple execution of the SISSRs application on a ChIP-Seq dataset (without a control dataset) requires three parameters outlined below with optional parameters discussed next.

-i The name of the input file containing the mapped tags in BED file format.

-o The name of the file onto which the output from SISSRs will be stored.

-s Size or length of the reference genome (number of bases/nucleotides) onto which the sequenced reads were mapped. For example, 3080436051 for the human genome (hg18 assembly). If analyzing data for a specific chromosome (or a set of chromosomes), then this would be the length of that chromosome (or sum of the lengths of those chromosomes).

If a control dataset is available, option -b, described below, should be used (see Note 2). Various other options available on SISSRs application are listed below. Some of these parameters are preset to default values, which the users can reset to their desired values. Users are recommended to set the -a option, which controls false positives due to amplification or sequencing biases.

-a Setting this option allows only one read per genomic coordinate to be retained even if multiple reads align to the same coordinate, thus effectively minimizing the effects of sequencing and/or PCR amplification bias. During PCR amplification, certain DNA fragments may be amplified into several orders of magnitude in a biased fashion, which after sequencing and mapping will show up as regions enriched with inordinate number of tags. To avoid calling these pseudo-enriched regions as binding sites, we strongly recommend using this option when running SISSRs.

-F Average length of the DNA fragments from ChIP. Typically, DNA fragments of certain length are size-selected for sequencing. Set *F* to this length (integer), if it is known. The individual performing the ChIP experiment and size-selection usually has a good estimate of the average length of sequenced DNA fragments. If this information is not available, this parameter can be left unset in which case SISSRs estimates this measure from the tags in the dataset (also check option -L below; see ref. 22 for details on length estimation).
Default: estimated from tags.

-D FDR if random background model based on Poisson probabilities needs to be used as control. This parameter is relevant only when a control data (e.g., input DNA or nonspecific IgG control) is not provided using the -b option.
Default: 0.001.

-b The name of the file containing the control data (e.g., input DNA or nonspecific IgG control; see Note 2). This file should be in the BED format. The tags in this file are used as a negative control. Subheading 2.2 contains a detailed description of how SISSRs uses the control data to minimize the number of false positives. Users may use -e and -p options (see below) to set the *e*-value and *p*-value thresholds to control sensitivity and specificity, respectively. If no control data is available, SISSRs

uses a random background model based on Poisson probabilities (in which case, use option -D to set the FDR).

-e *e*-Value threshold. It is the expected number of enriched regions (based on Poisson probabilities) in a similar-sized dataset. The value entered for this parameter is used to estimate the minimum number of reads (R) necessary to identify candidate binding sites. This option controls sensitivity (the -p option explained below controls specificity), and is ignored if -b option is not used (no control data).
Default: 10.

-p *p*-Value threshold. For a given F value (average DNA fragment length), the fold/ChIP enrichment for a candidate binding site is the ratio of the number of tags supporting the site, which is $p + n$ (Fig. 2a), to the number of tags supporting the same site in the control dataset. This fold enrichment is normalized with respect to the number of tags in both the test and the control datasets. To assess the statistical significance of the observed fold enrichment (the probability that the observed fold enrichment is by chance), an empirical distribution of fold enrichments from at least one million random sites, spanning the set of all chromosomes in the test dataset, is used to estimate the p-value for each candidate binding site. Only those sites with p-values not over the p-value threshold are reported as true binding sites. This option controls specificity (the -e option explained above controls sensitivity), and is ignored if -b option is not used (no control data).
Default: 0.001.

-m Fraction of genome (0.0–1.0) mappable by reads. Typically, not all sequenced reads map to unique genomic locations. Portions of the genome containing repetitive elements, which account for roughly 20% of the genome, are not mappable. The value entered for this parameter is used to estimate Poisson probabilities.
Default: 0.8.

-w Size of the scanning window (must be an even number >1), which is one of the parameters that attempts to control for noise in the data. The scanning window slides so that there is a 50% overlap between two consecutive window positions. As a result, the resolution of the identified binding sites (t in Fig. 2a) is $w/2$. For example, for $w = 20$, each binding site in the output file (with default -c option) will have a starting and ending coordinate with 1 and 0 in the Units position, respectively (e.g., 1234561–1234620). A larger window size reduces the influence of nonspecific reads and thus false positives at the cost of resolution. A smaller window size provides for increased resolution but may increase the number of false

positives if the data is noisy (contains a high number of nonspecific reads). In other words, smaller window size makes for higher sensitivity possibly at the cost of lower specificity, and larger window size makes for higher specificity possibly at the cost of lower sensitivity. The amount of background noise in the data is an important factor one needs to consider before setting a value for -w.
Default: 20.

-E Threshold for the number of tags mapped within *F* bp upstream or downstream of the center of the inferred binding site (*t* in Fig. 2a). This is one of the parameters that controls for specificity to a small degree. The higher the *E*, the more specific (and slightly less sensitive) SISSRs will be, and vice versa.
Default: 2 (assuming that the data file contains ~5–10 million reads; the user may consider increasing this value if the total number of reads is much larger).

-L Upper-bound on the DNA fragment length. It is the approximate length/size of the longest DNA fragment that was sequenced. This value is one of the critical parameters used during the estimation of average DNA fragment length. The individual who performed the ChIP and size-selection of the DNA fragments before sequencing should have a good estimate on of the upper-bound for the DNA fragment length.
Default: 500 (assuming that DNA fragments of length <500 bp were size-selected).

-q The name of the file containing genomic regions in simple three-column tab-separated format (chr start-coordinate end-coordinate). Reads falling within these regions will not be considered for the analysis.

-t If this option is set, each binding site is reported as a single genomic coordinate representing the center of the inferred binding site (*t* in Fig. 2a). If this option is not selected, SISSRs uses the -c option (see below).

-r If this option is set, SISSRs, instead of reporting each binding site as a single genomic coordinate (representing the center *t* of the inferred binding site; e.g., chr1 12345), each binding site is reported as an *X*-bp binding region, where *X* represents the resolution of the identified site (Fig. 2a). *X* varies for each binding site depending upon the availability of tags supporting the site. If this option is not selected, SISSRs uses the -c option as default (see below).

-c This option is same as the -r option, except that it reports binding sites that are clustered within *F*-bp of each other as a single binding region by merging those sites. As a result, the number of binding sites reported using this option could be

typically fewer than that reported using the -r option. For each binding region reported in the output file, the entry in the "NumTags" column indicates the number of tags supporting the strongest binding site in the reported binding region. The -c option is the recommended option especially if *w* is set to smaller values (ten or less).

Default: This is the default option, which SISSRs is used to report binding sites.

-u If this option is set, SISSRs also reports binding sites supported only by reads mapped to either sense or antisense strand. This option will recover binding sites whose sense or antisense reads were not mapped for some reason, e.g., the actual binding site lies right next to a repetitive region in which case reads aligning to the repetitive side were not mapped because they also align to other region(s) in the genome (see ref. 22 for details).

-x If this option is set, the summary and the progress report are not displayed on the terminal during the execution of the application.

2.3. Examples

Example 1: A simple example with no control dataset:
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs
SISSRs identify binding sites based on the reads in the test data file ctcf.bed. Since no control data file was provided (−b option), the default background model based on Poisson probabilities and the default FDR (0.001) will be used to determine statistically significant number of tags (R in Fig. 2) necessary to identify binding sites. SISSRs automatically use the default values for other parameters.

Example 2: Using the -a option, which considers only one read per genomic position:
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs -a
This is same as Example 1, except that only one read per genomic position is kept even if multiple reads get mapped to the same gnomic position.

Example 3: Using a control dataset:
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs -b control.bed -a
This is same as Example 2, except that a background control file is used as negative control (replacing the default random model based on Poisson probabilities). Default values are used for other parameters including the -e and -p parameters, which assume the default values 10 and 0.001, respectively.

Example 4: Ignoring reads that fall within certain genomic regions:
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs -b control.bed -a -q repeatsFile.txt
This is same as Example 3, except that the input reads that fall within the genome regions listed in the repeatsFile.txt will be ignored during the analysis. Effectively, this may reduce the number of binding sites reported compared to that reported in the case of Example 3.

Example 5: General run with no control data (relevant options listed using separate square brackets [ ]):
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs [−a] [−F 200] [−D 0.001] [−m 0.8] [−w 20] [−E 2] [−L 500] [−q repeatsFile.txt] [−t]/[−r]/[−c] [−u] [−x]

Example 6: General run with a control dataset (relevant options listed using separate square brackets [ ]):
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.sissrs [−a] [−F 200] [−b bg.bed] [−e 10] [−p 0.001] [−m 0.8] [−w 20] [−E 2] [−L 500] [−q repeatsFile.txt] [−t]/[−r]/[−c] [−u] [−x]

2.4. SISSRs Output, Interpretation, and Downstream Analyses

The results from a SISSRs run are stored under the file name that was provided by the user with the -o parameter. This output file contains the summary of the test and control datasets, the list of command line and estimated parameters which SISSRs used to process the data, and the list of binding sites identified using the statistical thresholds chosen by the user. A typical SISSRs output is shown in Fig. 3. Each identified binding site is listed as a genomic region along with the number of tags supporting that site. If a background control data was used, fold enrichment over the control data along with a *p*-value accompanies each reported site.

The first term denotes the chromosome on which the binding site resides. The second and the third terms denote the chromosomal start and end coordinates of the binding site, respectively. The fourth term "NumTags" denotes the number of tags supporting the identified binding site, which is equal to $p + n$ in Fig. 2a. The fifth and the sixth terms "Fold" and "*p*-value," respectively, are reported only if a background control data was used. Fold denotes fold-enrichment, which is the ratio of NumTags to the number of tags supporting the exact same site in the background control data (see Note 6). While computing the fold enrichment, the number of tags supporting the binding site in the test and control data is normalized by the total number of tags in the test and control data. The *p*-value denotes the probability that one would expect to see this fold-enrichment between the test and the control data just by chance, which is computed based on the empirical distribution of fold-enrichment values for one million or more random sites (Fig. 2b). Only those binding sites with fold-enrichment *p*-value less than or equal to the *p*-value threshold (set by the user using the -p option) are reported in the results file.

Typical downstream analyses of SISSRs-reported binding sites include de novo motif analysis to identify the consensus sequence within the identified binding sites/regions. De novo motif analysis is an unbiased search for a consensus sequence motif present within the identified binding sites (Fig. 4; see Note 7). Software tools such as PRIORITY (27), MEME (28), and GADEM (29)

```
======================================================================
SISSRs: A tool to identify binding sites from ChIP-Seq data
======================================================================
SISSRs version 1.4 (Release date: Mon, 24 November 2008)

...
...

======================================================================
COMMAND LINE SUMMARY & ESTIMATED PARAMETERS
======================================================================
Data file (i)                          : hg18-test.bed
Number of tags in the data file        : 17977403
Number of tags selected for analysis   : 17381505 (96.69%)
Tags mapped to sense strand            : 8690727 (50.00%)
Tags mapped to anti-sense strand       : 8690778 (50.00%)
Background model (Negative Control)    : hh18-control-input.bed
Number of tags in the control file     : 18655942
Genome length (s)                      : 3080436051
Fraction of genome mappable by reads (m): 0.80
Effective Genome length  (s*m)         : 2464348840
E-value (e)                            : 10
P-value (p)                            : 0.001
Scanning window size (w)               : 20
Average DNA fragment length (f)        : 202
Minimum number of 'directional' tags
  required on each side of the inferred
  binding site (E)                     : 2
Keep one tag per genomic position (a)  : YES
Also reports binding sites supported
  only by reads mapped to either sense
  or anti-sense strand (u)             : NO
======================================================================


======================================================================
BINDING SITES
======================================================================
Tags necessary to identify binding sites: 15 (Fold >= 6.00)
(at least E=2 tags on each side)

Chr     cStart     cEnd      NumTags  Fold    p-value
---     ------     ----      -------  ----    -------
chr1    6397351    6397391   91       19.53   7.0e-06
chr1    9164791    9165031   109      29.25   2.0e-06
chr1    11890651   11890951  98       17.53   1.0e-05
chr1    11892391   11892491  40       10.73   5.0e-05
...
```

Fig. 3. A typical SISSRs output file.

can be used to identify the consensus sequence, if any, present within identified sites (see Note 8). If the DNA binding preference for the POI is known, then the identified consensus sequence is expected to match the known binding sequence. Otherwise, the user needs to investigate at least two possible scenarios with regard

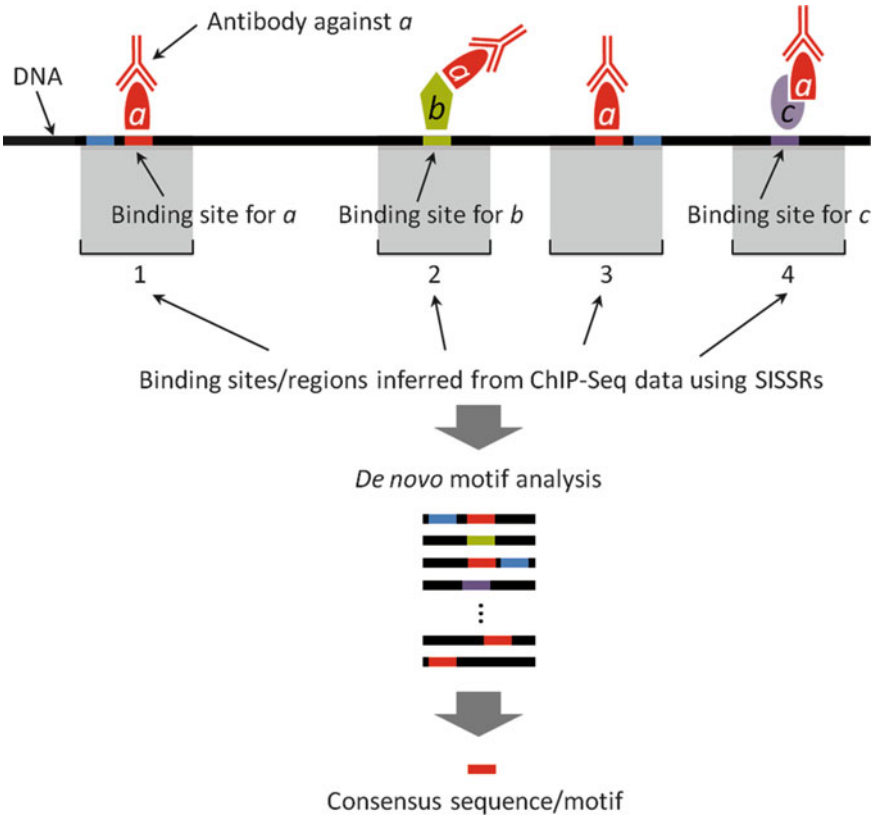

Fig. 4. De novo motif analysis for discovering consensus sequence motif within the identified binding sites.

to the novel consensus sequence: (a) the consensus sequence could characterize an undiscovered novel binding preference of the POI or (b) the POI binds DNA indirectly via another protein, in which case the identified consensus sequence would correspond to the binding preference of that protein.

Other analyses include determining the genomic distribution of identified binding sites in relation to genomic landmarks, and defining a list of genes targeted by the protein being profiled. For a given reference genome and a set of gene annotations, custom software can be written to determine the fraction of identified binding sites that fall within intronic/exonic regions, promoter regions (defined as a few kilo-bases upstream and/or downstream of transcription start sites of known genes), and other genomic landmarks of interest. Given that a binding site may or may not be functional, defining target genes based on the set of identified binding sites alone is not straightforward. But, in practice, genes that contain one of more identified binding sites within a few kilo-bases upstream or downstream of their transcription start sites are defined as targets of protein being profiled.

2.5. SISSRs Running Time

SISSRs running time primarily depends on whether or not a background control data is being used. When no background control data is used, the running time is typically few minutes. Most of this time is spent reading the data files. In general, it takes ~5 min for SISSRs to analyze a test dataset containing approximately ten million reads with default settings and no background control data. If a background control data is used, then SISSRs could take anywhere between ~10 and 30 min for a *p*-value threshold of 0.001, with the additional time spent sampling one million random sites to determine the empirical *p*-value distribution. Setting the *p*-value to smaller values will further increase the running time. Thus, it is recommended that the *p*-value is not set to extremely small values if running time is of primary concern (see Note 9).

3. Notes

1. A high noise-to-signal ratio raises a red flag on the sequencing quality, and it is a good practice to avoid datasets where signal and noise cannot be easily distinguished.
2. Many nucleosome-free (open chromatin) regions in the genome can bind proteins in a nonspecific manner and certain genomic regions are prone to biased amplification/sequencing. These biases in the test dataset can be neutralized to some extent by using a control dataset, which will help reduce the number of nonspecific binding sites inferred as true binding sites. Input DNA and IgG ChIP-derived DNA are the two commonly used controls. Input DNA is prepared the same way as the ChIP DNA without the immunoprecipitation step. IgG ChIP is performed with an antibody against IgG, which binds DNA in a nonspecific manner. If antibody specificity against the POI is not a concern, input DNA serves as a better control for amplification and sequencing bias compared to IgG ChIP DNA. Although not necessary, we strongly recommend using a control data when using SISSRs.
3. SISSRs was designed to identify protein–DNA interaction sites from ChIP-Seq datasets and is not suitable for analyzing histone modification data to identify regions enriched with a specific histone modification. ChIP-Seq data characterizing histone modifications in general have much broader footprints of signal of varying lengths (anywhere from few hundred to several thousand bases) compared to that for protein–DNA interaction sites, which is typically ~200 nucleotides (13). Distinguishing broader footprints of signal from the background noise requires accurate characterization

of boundaries demarcating signal and noise, a task that requires sequencing of the ChIP sample to near saturation. Since samples are rarely sequenced to near saturation, identification of regions with broad footprints of signal (e.g., histone modifications H3K4me1, H3K9me3, H3K27me3, and H3K26me3 (13)) is a relatively difficult task compared to protein–DNA binding sites. We do not recommend SISSRs for analyzing histone modification data in general, but it may be used to analyze histone modification data such as H3K4me3 or H3K9ac (that have ~200–500 bp footprints) with caution.

4. The statistics used to determine Z is highly dependent on how well saturated the control data is. If the control data does not contain sufficient reads (much less than what may be necessary), then using such a dataset as a control is as good as using no control. Thus, it is important to make sure that the control data contains sufficient number of reads. As a rule of thumb, for a genome of length L nucleotides and the average fragment length of F nucleotides, it is desirable that the control dataset contains at least about L/F tags to make reliable inferences.
5. The resolution of the reported binding site is dependent on the number of tags in the dataset. The larger the dataset (more tags), the higher the likelihood of identifying sites with better resolution. Typically, the average resolution of the reported sites is somewhere between 40 and 80 bp, but it could be as much as the length of the average ChIP fragment.
6. The value for ChIP fold-enrichment (when a control is used) or number of tags (when a control is not used) is a good indicator of protein–DNA binding affinity/stability (22). When comparing two or more binding sites, higher (lower) values for these measures can be interpreted as stronger (weaker) binding.
7. If one wishes to perform motif analysis on the DNA sequences corresponding to the reported binding sites, we recommend using the 200 nucleotide sequence centered on the reported binding site. Although the ~5–20 bp DNA sequence bound by a protein is highly likely to be present within the region reported as the binding site, it is quite possible that all or part of this binding sequence is just outside of the reported binding site. And, since the resolution of the reported sites are dependent on the tags that map near these sites, some of which could be noise, there is always a chance that a reported coordinate defining a binding site could be off by a few base

pairs. It is therefore good practice to consider using a 200 nucleotide sequence centered on the reported binding site.

8. Since ChIP using an antibody against POI captures genomic regions bound directly as well as indirectly by POI (Fig. 4), one cannot expect all of the reported binding sites for POI to contain the consensus binding sequence/motif. Thus, a lack of consensus sequence at a site cannot be interpreted as that site being a false-positive.
9. If running time is of concern, do not set the p-value ($-p$) to a number less than 0.0001 (0.001 is the default).

Acknowledgments

This work was supported by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (Project number ES102625–02 to R.J.).

References

1. Boyer LA, Lee TI, Cole MF et al (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947–956.
2. Chen X, Xu H, Yuan P et al (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133:1106–1117.
3. Ho L, Jothi R, Ronan JL et al (2009) An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. Proceedings of the National Academy of Sciences of the United States of America 106:5187–5191.
4. Molkentin JD (2000) The zinc finger-containing transcription factors GATA-4, -5, and −6. Ubiquitously expressed regulators of tissue-specific gene expression. J Biol Chem 275:38949–38952.
5. Hou C, Dale R, Dean A (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proceedings of the National Academy of Sciences of the United States of America 107:3651–3656.
6. Rampakakis E, Gkogkas C, Di Paola D et al (2010) Replication initiation and DNA topology: The twisted life of the origin. J Cell Biochem 110:35–43.
7. Cohn MA, D'Andrea AD (2008) Chromatin recruitment of DNA repair proteins: lessons from the fanconi anemia and double-strand break repair pathways. Mol Cell 32:306–312.
8. Shivji MK, Venkitaraman AR (2004) DNA recombination, chromosomal stability and carcinogenesis: insights into the role of BRCA2. DNA Repair (Amst) 3:835–843.
9. Solomon MJ, Larsen PL, Varshavsky A (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell 53:937–947.
10. Ren B, Robert F, Wyrick JJ et al (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306–2309.
11. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4:613–614.
12. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680.
13. Barski A, Cuddapah S, Cui K et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.
14. Johnson DS, Mortazavi A, Myers RM et al (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502.
15. Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin

immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657.
16. Barski A, Jothi R, Cuddapah S et al (2009) Chromatin poises miRNA- and protein-coding genes for expression. Genome Research 19:1742–1751.
17. Cuddapah S, Jothi R, Schones DE et al (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research 19:24–32.
18. Barski A, Zhao K (2009) Genomic location analysis by ChIP-Seq. J Cell Biochem 107:11–18.
19. Cuddapah S, Barski A, Cui K et al (2009) Native chromatin preparation and Illumina/Solexa library construction. Cold Spring Harb Protoc 2009:pdb prot5237.
20. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
21. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research 18:1851–1858.
22. Jothi R, Cuddapah S, Barski A et al (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Research 36:5221–5231.
23. http://www.rajajothi.com.
24. http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/.
25. http://www.perl.org.
26. http://genome.ucsc.edu/FAQ/FAQformat#format1.
27. Narlikar L, Gordan R, Hartemink AJ (2007) A nucleosome-guided map of transcription factor binding sites in yeast. PLoS Comput Biol 3:e215.
28. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36.
29. Li L (2009) GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. J Comput Biol 16:317–329.