SISSRs User Manual

Raja Jothi
National Institutes of Health (NIH)
jothi@mail.nih.gov

Last Modified: November 25, 2008

Contents

1 Introduction

This document is intended as a user guide to the SISSRs (Site Identification from Short Sequence Reads) application (versions 1.0 to 1.4). It does not provide details on the underlying SISSRs algorithm or its implementation. For details on the SISSRs algorithm, we refer the reader to Jothi *et al* [1].

1.1 What is SISSRs?

SISSRs is a software application for precise identification of genome-wide transcription factor binding sites from ChIP-Seq data. It is essentially a `perl` implementation of the SISSRs algorithm outlined in Jothi *et al* [1], with several new features that were not fully described in the original paper.

ChIP-Seq, which combines chromatin immunoprecipitation (ChIP) with next generation massively parallel sequencing, is a powerful experimental technique to determine whether proteins including, but not limited to, transcription factors bind to specific regions on chromatin *in vivo*. In ChIP-Seq, the DNA fragments obtained from ChIP are directly sequenced using the next generation genome sequencers such as Illumina Genome Analyzers. Although the lengths of the input DNA could be anywhere between ~200 bp and ~1 kb, typically, only the first ~25–50 nt from the DNA ends are sequenced. The resulting short *reads* are mapped back to a reference genome, and only those reads that map to an unique genomic locus in the reference genome are considered for further analysis. Mapped reads are commonly referred to as *tags* (henceforth, 'reads' and 'tags' are used interchangeably).

A binding site is a region on the DNA to which specific proteins including, but not limited to, transcription factors bind *in vivo*. A typical binding site could be anywhere between ~5-20 nucleotides in length.

1.2 Getting SISSRs

The latest version of the `perl` implementation of the SISSRs algorithm is available for free download at http://sissrs.rajajothi.com and http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/. Since constant efforts are made to improve the SISSRs application, users are recommended to check back for newer versions. After downloading the zipped archive, users may save the extracted `sissrs.pl` program onto their working directory (to run the program from the working directory) or to their `/bin` directory (`/usr/bin/` or `home-directory/bin/`; to run the program from anywhere in the home directory).

2 Preparing the Input

SISSRs takes as input a data file containing genomic coordinates of the mapped reads for your protein of interest and an optional background file containing control data (IgG, input DNA, etc.) in BED file format [2]. In BED file format, each line contains 6 tab- or space-separated terms in the following format:

```
chr1   1234561   1234585   U0   1   +
```

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

3 Running SISSRs

Since SISSRs application was written in `perl`, in order to run SISSRs, you will need `perl` installed on your computer. If you are running a Linux system (or most UNIX systems, including Mac OS X), you probably already have an installation of `perl` that was packaged with it. If you are running Microsoft Windows OS, you may download `perl` from http://ww.perl.org.

Typing just the name of the executable (`sissrs.pl` or `./sissrs.pl`) on the command line displays the help menu as shown on the next page. A detailed description of each option is provided in subsequent pages.

```
================================================================

USAGE:

 sissrs.pl -i <input-file> -o <output-file> -s <genome-size> [OPTIONS]

 [-i <file>]    input file containing tags/reads in BED format
 [-o <file>]    output file into which results will be stored
 [-s <int>]     genome size (number of bases or nucleotides)
 [-a]           only one read is kept if multiple reads align to the
                same genomic coordinate (minimizes amplification bias)
 [-F <int>]     average length of DNA fragments that were sequenced
                (default: estimated from reads)
 [-D <real>]    false discovery rate (default: 0.001) if random
                background model based on Poisson probabilities need
                to be used as control (also check option -b below)
 [-b <real>]    background file containing tags in BED format to be
                used as control; -e and -p can be set to desired
                values to control specificity and specificity, resp.
 [-e <real>]    e-value (>=0); it is the number of binding sites
                one might expect to infer by chance (default: 10);
                this option is irrelevant if -b option is NOT used
 [-p <real>]    p-value threshold for fold enrichment of ChIP tags
                at a binding site location compared to that at the
                same location in the control data (default: 0.001);
                this option is irrelevant if -b option is NOT used
 [-m <int>]     fraction of genome mappable by reads (default: 0.8
                for hg18, assuming ELAND was used to map the reads;
                could be different for different organisms and other
                alignment algorithms)
 [-w <int>]     scanning window size (even number > 1), which
                controls for noise (default: 20)
 [-E <int>]     min number of 'directional' reads required on each
                side of the inferred binding site (>0); (default: 2)
 [-L <int>]     upper-bound on the DNA fragment length (default: 500)
 [-q <file>]    file containing genomic regions to exclude; reads
                mapped to these regions will be ignored; file
                format: 'chr startCoord endCoord'
 [-t]           reports each binding site as a single genomic
                coordinate (transition point t in Fig 1 [1])
 [-r]           reports each binding site as an X-bp binding region
                centered on inferred binding coordinate; X denotes
                the distance from the start of the right-most red
                bar (see Fig 1A in the manuscript [1]; lower-left)
                to the end of the left-most blue bar surrounding the
                actual binding site (transition point t in Fig 1A)
 [-c]           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; this is the default option.
 [-u]           (also) reports binding sites supported only by reads
                mapped to either sense or anti-sense strand; this
                option will recover binding sites whose sense or
                anti-sense reads were not mapped for some reason
                (e.g., falls in unmappable/repetitive regions)
 [-x]           do not print progress report (default: prints report)

================================================================
```

3.1 Required Parameters

-i The name of the file containing the input data. Each line in this file should be of the following (BED [2]) format:

chr19 12345 12370 XXX YYY +
chr19 12345 12370 XXX YYY -

The first column contains the name of the chromosome, the second and third columns contains the chromosomal start and end coordinates of the mapped read (tag), respectively, and the sixth column contains either a + or a – denoting the strand (sense or anti-sense, respectively) onto which the read was mapped. The data in the fourth and fifth columns are not used by SISSRs and are thus irrelevant.

More information about the BED format can be found at
http://genome.ucsc.edu/FAQ/FAQformat#format1

-o The name of the file onto which the output needs to be stored.

-s Genome size (or length). Number of bases or nucleotides in the reference genome. For eg., 3080000000 for hg18.

3.2 Optional Parameters

-a If this option is set, only one read is kept if multiple reads align to the same genomic coordinate, thus effectively minimizing 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 picking up these pseudo-enriched regions as binding sites, it is advisable to use this option. This may reduce the total number of reads considered for the binding site analysis. The original and the reduced number of reads will be reported in the output file.

-F Average length of the DNA fragments from ChIP, whose ends (typically ~25-50 nt) were sequenced. Typically, DNA fragments of "certain" length are isolated for sequencing. Enter this length (integer), if it is known. This value is one of the critical parameters used during the identification of binding sites. The person that did the ChIP, who would have performed the size-selection of the DNA fragments before sequencing, should have an approximate idea of average length of sequenced DNA fragment lengths. If this information is not available, then the average length of the DNA fragments could be estimated from the reads (the algorithm that estimates this number is given in Jothi *et al* [1]; also check option -L below)
Default: estimated from reads

-D False discovery rate if random background model based on Poisson probabilities need to be used as control (i.e., no background data is available).
Default: 0.001

-b The name of the file containing the background control data (for example: IgG or input DNA). Each line in this file should be in the same format (BED) as the data file. The reads in this file are used as a negative control to minimize the number of false-positives. Section 5 contains a detailed description of how SISSRs uses the background control data to minimize the number of false positives. Users may use -e and -p options (described below) to set the *e*-value and *p*-value thresholds to control sensitivity and specificity, respectively. If no background control file is specified, SISSRs uses a random background model based on Poisson probabilities (in which case, use option -D above to set FDR).

Based on our experience in analyzing ChIP-Seq datasets for several proteins, we have noticed that using a background control data produces a relatively more reliable set of binding sites than when no background control data is used. We suspect that one of the reasons for this could be that there are several open chromatin regions in the genome that bind many proteins in a non-specific manner. By using a background control dataset, one can eliminate including such non-specific binding sites as true binding sites. We recommend using a background control data when available. For questions about how many reads may be necessary for the background control sample, please refer Section 5.

-e *e*-value threshold. It is the maximum number of "enriched regions" one can expect to see by chance (Poisson probabilities), when analyzing a similar-sized dataset. The value entered for this option is used to estimate the minimum number of reads necessary ('R' in Fig 1 of Jothi *et al* [1]) to identify candidate binding sites. This option controls sensitivity (the -p option explained below controls specificity), and is irrelevant if no background control data is used.
Default: 10

-p *p*-value threshold. For a given *F* value (average DNA fragment length) and a candidate binding site (represented as genomic coordinate *t* in Fig 1 of Jothi *et al* [1]), let *u* be the number of tags mapped to the region [*t* - *F*, *t*] on the sense strand, and let *v* be the number of tags mapped to the region [*t*, *t* + *F*] on the anti-sense strand. The "fold enrichment" for this site is then the ratio of the number of tags supporting the site, which is *u* + *v*, to the number of tags supporting the same site in the control datasets. The fold enrichment is normalized with respect to the number of tags in both the real and the control data. To assess the statistical significance of the observed fold enrichment (the probability that the observed fold enrichment is by chance), a distribution of fold enrichments from at least 1 million random genomic locations, spanning all chromosomes, is used to estimate the *p*-value for each inferred binding site. Only those binding sites with *p*-values at most the *p*-value threshold are reported as true binding sites. This option controls specificity (the -e option explained above controls sensitivity), and is irrelevant if no background control data is used.
Default: 0.001

-m Fraction of genome (0.0 to 1.0) mappable by reads. Typically, sequenced reads could be mapped only to a fraction of the reference genome. That is, there are regions in the genome, containing repetitive elements, which are not mappable. For example, if the ELAND algorithm is used to map the reads to the human genome (hg18), which maps reads to unique positions in the genome allowing for up to two mismatches, reads could be mapped to ~80% of the hg18

assembly. ELAND discards those reads that map to two or more genomic locations and those that map to a unique location but with three or more mismatches. As a result, genomic regions containing repeats are never mapped with reads. The value for this parameter is used to determine the effective (mappable) genome length, which is one of the variables used to estimate Poisson probabilities.
Default: 0.8 for hg18, assuming ELAND was used for mapping. This could be different for different organisms and other mapping algorithms.

-w Size of the overlapping/sliding scanning window (must be an even number >1), which is one of the parameters that attempts to control for noise in the data (see Fig 6 in [1]). The scanning window slides in such a manner that there is a 50% overlap between two consecutive window positions. As a result, the resolution of the indentified binding sites (transition point t in Fig 1 of Jothi *et al* [1]) is w/2. For example, for w = 20, the list of binding sites in your output file (using option -t) will all have genomic coordinates ending with 1 (e.g., 1161, 11101, 764381, etc), and for w = 10, the list of binding sites in your output file (using option -t) will all have genomic coordinates ending with either a 1 or a 6 (e.g., 1161, 11106, 11676, etc). A larger window size reduces the influence of non-specific reads and thus false positives at the cost of lower resolution. A smaller window size provides for increased resolution but may increase the number of false positives if the data contains a high number of non-specific 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 a important factor one needs to consider before setting the size of the window.
Default: 20

-E Number of "directional" reads required within *F* base pairs on either side of the inferred binding site. This is one of the parameters that controls for specificity. The higher the *E*, the more specific (and less sensitive) SISSRs will be, and vice-versa.
Default: 2 (assuming that the data file contains ~2 to ~5 million reads; the value may need to be increased 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 person that did the ChIP, who would have performed the size-selection of the DNA fragments before sequencing, should have an approximate idea of the range of DNA fragment lengths.
Default: 500

-q The name of the file containing genomic regions in simple BED format (chr start end). 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. The center of the binding site is essentially the transition point *t* shown in Fig 1 of Jothi *et al* [1]. If this option is not selected, SISSRs uses the -c option as default (see below).

-r If this option is set, SISSRs, instead of reporting each binding site as a single genomic coordinate (representing the center of the inferred binding site; e.g., chr1 12345), reports each binding site as an *X*-bp region (e.g., chr1 12330 12360), where X denotes distance (12360-12330=30) from the start of the right-most red bar (see Fig 1 in Jothi *et al* [1]; lower-left) to the end of the left-most blue bar surrounding the actual binding site (transition point *t* in Fig 1 of Jothi *et al* [1]). *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. As a result, the number of binding sites reported using the –c option will be slightly less 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 particularly useful if w is set to a small value (*Y* or less), where *Y* depends on the background noise in the data. Typically, it is advisable to use -c option if w is set to 10 or less.
Default: This is the default option, which SISSRs uses to report binding sites.

-u If this option is set, SISSRs also reports binding sites supported only by reads mapped to either sense or anti-sense strand. This option will recover binding sites whose sense or anti-sense reads were not mapped for some reason (e.g., falls in unmappable/repetitive regions; see Fig 6B in [1]).

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

3.3 Examples

1) A simple example

```
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.bsites
```

SISSRs identifies binding sites based on the reads in the file `ctcf.bed`, which is in BED format. The default false discovery rate (0.001) and the default background model based on Poisson probabilities will be used to determine statistically significant number of tags necessary to identify binding sites. For all other parameters, SISSRs automatically uses the default values.

2) Using the -a option, which considers only one read per genomic position

```
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.bsites -a
```

This is same as the previous example, except that only one read per genomic position is kept even if multiple reads get mapped to the same gnomic position.

3) Using a background control file

```
./sissrs.pl –i ctcf.bed -s 3080436051 -o ctcf.bsites -b bg.bed
```

This is same as the first example, except that a background file is used as negative control (instead of the default random model based on Poisson probabilities). The default e- and p-value thresholds (10 and 0.001, respectively) will be used.

4) Ignoring the reads that fall within certain genomic regions

```
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.bsites -q repeatsFile
```

This is same as the first example, except that the input reads that fall within the genome regions listed in the `repeatsFile` will be ignored

5) General run with **no** background control data (relevant options listed using separate square brackets [])

```
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.bsites [-a] [-F 200]
[-D 0.001] [-m 0.8] [-w 20] [-E 2] [-L 500] [-q repeatsFile]
[-t]/[-r]/[-c] [-u] [-x]
```

6) General run with background control data (relevant options listed using separate square brackets [])

```
./sissrs.pl -i ctcf.bed -s 3080436051 -o ctcf.bsites [-a] [-F 200]
[-b bg.bed] [-e 10] [-p 0.001] [-m 0.8] [-w 20] [-E 2] [-L 500] [-q
repeatsFile] [-t]/[-r]/[-c] [-u] [-x]
```

4 Interpreting the Output

The results/output of the run are stored onto the file that was stated (using the -o parameter). The output file contains the following information:

- the version of SISSRs application used, and the date it was released
- the reference to cite if you used SISSRs in your research
- data summary, and the list of command-line and estimated parameters, and
- the list of identified binding sites.

Depending on whether -t, -r, or -c (default) options was chosen, the list of binding sites in the output file will be formatted differently.

If -r or -c option was chosen, each binding site will be listed in the following format:

chr binding-site-start-position binding-site-end-position NumTags [Fold] [p-value]

The first term denotes the chromosome on which the binding site was identified. The second and the third terms denote the chromosomal start and end coordinates of the binding site (or region), respectively. The fourth term “NumTags” denotes the number of tags/reads supporting the identified binding site, which is

equal to $p + n$ in Fig 1B of Jothi *et al* [1]. Here, p is the number of reads mapped to the region $[t - F, t]$ on the sense strand, and n is the number of reads mapped to the region $[t, t + F]$ on the anti-sense strand. In other words, p is the number of reads mapped to the sense strand of the F-bp region upstream of transition point t, and n is the number of reads mapped to the anti-sense strand of the F-bp region downstream of transition point t. The fifth and the sixth terms "Fold" and "p-value" will be reported only if a background control data file was used. Fold denotes "fold enrichment" for the binding site, which is equal to the ratio of NumTags to the number of tags supporting the exact same site in the background control data. While computing the fold enrichment, the number of tags supporting the binding site in the real and control data is normalized by the total number of tags in the real and control data. The p-value denotes the probability that one would expect to see such fold enrichment between the real and the control data just by chance. Only those binding sites with fold enrichment p-values greater than or equal to the p-value threshold (set using the -p option) are reported in the results file.

If -t option was chosen, each binding site will be listed in the following format:

chr binding-site-midpoint NumTags [Fold] [p-value]

Instead of reporting the chromosomal start and end coordinates of the binding site, the -t option reports only the center (transition point t in Fig 1 of Jothi *et al* [1]) of the identified binding site. If you had chosen the default scanning window size (w = 20), then you may notice the list of binding sites in your output file will all have genomic coordinates ending with 1 (e.g., 1161, 11101, 764381, etc). The reason for this is that the scanning window slides in a manner that two consecutive windows overlap each other by 50%. So, for w = 20, the window will start scanning the nucleotides in the order 1-20, 11-30, 21-40, and so on. Since the transition point t is identified by computing the midpoint of a scanning window, which in this case will be a number with 1 as the last digit, the coordinates of binding sites end with 1 as the last digit. If you were to set w = 10, then the list of binding sites in your output file will have genomic coordinates ending with a 1 or a 6 as the last digit.

Note: Since the -c option clusters binding sites whose midpoints (transition point t in Fig 1 of Jothi *et al* [1]) are within F-bp of each other, the number of binding sites (or regions) reported using -c option will be less than or equal to that reported using the -r or the -t option.

5 Strategy employed when using background data

This section describes the strategy that SISSRs employs when a background control data is provided as a part of the input. The original paper [1] outlining the SISSRs algorithm did not contain the details of this strategy.

- First, SISSRs uses the values set using -e option to make an initial estimate for R (see Fig 1B in [1]). The estimated value for R is then used to identify binding sites in the same way as that when no background control data was supplied by the user (see Fig 1 in Jothi *et al* [1]).
- Second, for each identified binding site, recorded as a singular genomic coordinate t (denoted as transition point t in Fig 1 of Jothi *et al* [1]), the number of tags $n + p$ supporting the binding site is computed. Here, p is the number of reads mapped to the region $[t - F, t]$ on the sense strand, and n is the number of reads mapped to the region $[t, t + F]$ on the anti-sense strand. In other words, p is

the number of reads mapped to the sense strand of the *F*-bp region upstream of transition point *t*, and *n* is the number of reads mapped to the anti-sense strand of the *F*-bp region downstream of transition point *t*. Similarly, the number of tags *n* + *p* supporting the binding site (exact same genomic coordinate *t*) in control data is also computed.

- Third, for each identified binding site, the fold-enrichment ratio, which is the number of tags supporting the identified site in the real data to the number of tags supporting the exact same site in the control data, is computed. While computing the fold-enrichment ratio, if the denominator is zero, then the ratio is set to the numerator value (instead of infinity).
- Fourth, a distribution of fold-enrichment ratios for at least 1,000,000 random genomic sites is generated, and for the p-value threshold set using the -p option is used to identify the smallest fold-enrichment ratio *Z* such that the probability of obtaining a fold-enrichment ratio of *Z* or higher by chance is at most the p-value threshold.
- Fifth, only those binding sites with fold-enrichment ratios greater than or equal to *Z* are included in the final output as the true binding sites.

Note: The statistics used to determine *Z* is highly dependent on how well saturated the background control data is. If the background control data contains few reads (less than what may be necessary), then the fold-enrichment ratios for random sites (as computed in step three above) will end up being higher that what it is supposed to be. As a result the distribution of fold-enrichment ratios for random sites will be shifted to the right (increased values), thus forcing the approach to pick a higher *Z* (stringent) for given p-value threshold. Consequently, since only those sites with fold-enrichment scores greater than or equal to *Z* is included in the final output, far fewer sites be reported in the output file. Thus, it is important to make sure that the background control data file contains sufficient number of reads (typically, at least ~10 million reads for human or mouse genomes, and preferably ~20 million reads).

7 Running Time

The running time mainly depends upon the size of the dataset and whether or not a background control data is used. In general, it takes ~3 minutes for SISSRs to analyze a dataset containing ~5 million reads with default settings and no background control data. If a background control data were used, then SISSRs takes ~8-9 minutes for the default p-value threshold (0.001). If the p-value threshold is set is to *q* = 0.0001, then the running time will increase dramatically to ~30 minutes (because q/0.001 times 1 Million random sites will be sampled instead of just 1 Million random sites). Thus, it is recommended that the p-value is not set to extremely small values if running time is of primary concern.

6 References

1. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. Nucleic Acids Research (2008) 36(16):5221-31.
2. http://genome.ucsc.edu/FAQ/FAQformat#format1