Benchmarking Paper

strandngs

Streamlining NGS Data Management & Analysis

Calling narrow and broad peaks from ChIP-Seq data in Strand NGS

Rohit Gupta and Anita Sathyanarayanan
Strand Life Sciences

www.strand-ngs.com

Analyze | Visualize | Annotate | Discover

strand
New Generation Healthcare

Benchmarking Paper

Calling narrow and broad peaks from ChIP-Seq data in Strand NGS

Rohit Gupta and Anita Sathyanarayanan*

Strand Life Sciences

March 2015

* *Corresponding Author (rohit@strandls.com)*

ChIP-Seq is a widely used approach for studying transcription factor binding sites or histone modifications and their role in gene regulation in multi-cellular organisms. This benchmarking paper describes some of the algorithms implemented in Strand NGS and illustrates their efficacy in detecting both narrow and broad peaks/regions from the ChIP-Seq data.

1 Introduction

Chromatin immunoprecipitation (ChIP) followed by high throughput sequencing (ChIP-Seq) is one of the widely used approaches for elucidating interactions between DNA and proteins. It provides an essential tool for researchers to understand the role of transcription factors (TFs) or histone modifications (HMs) in gene regulation. Briefly, in the first step of chromatin immunoprecipitation, certain DNA fragments are enriched using antibodies for a specific TF. In the second step, enriched DNA fragments are sequenced using massively parallel DNA sequencing technology. The output of this process is a collection of many short sequencing reads or tags.

ChIP-seq is used to study either the cistrome of TFs (identification of all cis-acting targets or binding sites of a TF) [2] or the epigenome profile, especially the histone modification status [1]. Typically the distribution of sequencing reads for these cases is very different. While for most TFs, enriched regions are generally discrete and form sharp peaks covering short regions of DNA (10s or 100s of bases), the distribution of reads for many types of histone modification events follows a continuous property and span regions of up to several hundred kilobases.

While ChIP-Seq technology provides a useful way to study transcription factor binding sites and histone modification events, it has several challenges such as the presence of noisy background tag counts with local biases, dependence of ChIP-Seq sensitivity and specificity on sequencing depth and number of replicates, difficulty in discovering shorter regions for transcription factor binding sites - larger regions can lead to more false positive motifs in the downstream analysis, etc. Numerous approaches have been proposed in the past to address some of these challenges. While MACS [14], PICS [13], PeakSeq [8] and SISSRS [6] are some of the popular methods for narrow peak detection, Broadpeaks [10], SICER [12], ZINBA [7], CCAT [11], and RSEG [9] are some of the broad peak callers.

In this study, we wanted to assess the applicability of the algorithms implemented in Strand NGS to detect binding sites of a transcription factor (narrow peaks) as well as enriched regions corresponding to histone modification (broad peaks). Several state-of-the-art algorithms such as PICS [13] and MACS [14] have been implemented in Strand NGS for the detection of transcription factor binding sites (TFBS). In addition, we have also implemented a sliding window based algorithm called 'Find Enriched Regions (FER)', which detects enriched regions showing high coverage with respect to a control sample. For broad peak detection, one of the recent studies [3] suggested running MACS with an advanced parameter '-nomodel' (no shift model building). Further, when control sample is not available, one can also use the parameter 'nolambda' (no background estimation). MACS with these advanced parameters turns out to be a good alternative for broad peaks detection and was also found to outperform several other broad peak callers [10]. Further, a new version of MACS (version 2.0) has been released recently and has an additional parameter '-broad' for the detection of broad peaks. In Strand NGS, MACS (version 1.4) is implemented without the advanced parameter '-nomodel'.

We hypothesize that for most practical purposes, the performance of MACS with the advanced parameter '-nomodel' is similar to that of 'FER'. In this study, we used 3 histone modification data sets to test this hypothesis. MACS (both version 1.4 and version 2.0) was downloaded and executed from the command-line option. In addition

a human transcription factor binding data set for FoxA1 is also used to show similarity in results for narrow peaks obtained from Strand NGS implementation of MACS and command-line option of MACS from the original source.

2 Methods

2.1 Algorithms

The algorithms used in this study are described below. MACS1.4 (with only model building option) and FER are implemented in Strand NGS. MACS1.4 (with advanced parameters) and MACS2.0 were downloaded and executed from the command-line.

2.1.1 *Find Enriched Regions (FER)*

Description: The enriched region detection is a simple procedure for quickly estimating regions with high coverage. A window of user-specified size slides through the treatment and control sample and is considered enriched if the following criterion is satisfied:

$$\frac{N_t/S_t}{N_c/S_c} > e \tag{1}$$

where N_t and N_c denote the # of reads in the treatment and control sample window respectively; S_t and S_c denote the total # of reads in the treatment and control sample respectively.

Equation (1) ensures that the treatment sample window has $e - fold$ more normalized read count compared to the respective control sample window. In cases where control sample is not available, a window is considered enriched if the number of reads exceeds a given threshold. Finally to obtain an enriched region, consecutive enriched windows are merged together.

Parameters: Table 1 shows the parameters of the FER algorithm along with their default values in Strand NGS.

Parameter	Description	Default value in Strand NGS
Enrichment factor (e)	Ratio of normalized coverage in treatment and control window	5
Window size	Width of the sliding window	100bp
Window slide size	Length by which the window slides in a step	50bp
Min # reads in window	Minimum # of reads that should start in the window	10
Min region size	Lower cut-off on the size of the detected enriched region	150
Min # reads in region	Minimum # of reads in the enriched region	15
Upstream padding distance for genes	Defines the extent of the upstream region for a gene for annotating enriched regions	5000bp

Table 1: *Description of parameters used in FER algorithm.*

Based on different parameter settings, we discuss 3 variations of FER in this paper. FER-PS1 uses all the default parameters; FER-PS2 uses a *window size* of 150bp and default values for the rest of the parameters; FER-PS3 uses a *window size* of 400bp, *window slide size* of 100bp, *Min region size* of 100bp, *Min # reads in region* of 10, and default values for the rest of the parameters.

2.1.2 *MACS (ver1.4 and ver2.0)*

Description: Model-based Analysis of ChIP-Seq (MACS) estimates the region of DNA-protein interaction sites (transcription factor binding sites) or epigenetic modification (histone modification) regions by building a shift model where the tags are shifted to the 3' direction. The density of the ChIP-Seq fragments show a bimodal distribution around the region of interest as they are sequenced from both ends equally. MACS estimates the fragment length as the distance between the two modes and hence the shift distance of the tags to precisely identify the region of interest.

Following are the brief steps in MACS:

1. Select 1000 regions with a $10-$ to $30-fold$ enrichment relative to the random tag genome distribution.
2. Build a model and estimate the DNA fragment size (d). To build a model, separate the Watson and Crick tags for the above selected regions. Next, align these tags by the midpoint between the Watson and Crick tag centers. The distance between the modes of these positive and negative peaks in the alignment is defined as the fragment length, d.
3. Shift the tags towards the 3' end by $d/2$.
4. For experiments with control, linearly scale the control and the treatment libraries.
5. Using a window of size $2d$, slide across the genome to find candidate peaks with significant enrichment relative to genome background.
6. Model the number of reads from a genomic region as a Poisson distribution with dynamic parameter λ_{local}. The λ_{local} for a candidate peak is defined as:

$$\lambda_{local} = max(\lambda_{BG}, [\lambda_{Region}, \lambda_{1k}], \lambda_{5k}, \lambda_{10k}) \tag{2}$$

   Here, λ_{BG} is a constant estimated from genome background. λ_{Region} is from candidate region in the control sample and remaining λ_x values are from x-bp window centred at candidate region in the control sample. When control is not available, λ_{local} is calculated using the treatment sample.
7. The candidate peaks that pass the user defined $p - value$ cutoff based on λ_{local} (default $10e-5$) are called out as the final peaks.

The above steps are followed by MACS to detect narrow peaks typically reflective of transcription factor binding sites. However as mentioned earlier, the reads distribution of histone modification ChIP-seq data is usually continuous. Hence the peaks are much broader than those of the TF ChIP-seq data.

Parameters: Below is the description of the important parameters of the MACS algorithm along with their default values.

- *–g (genome size):* specifies the genome size.
- *–shiftsize:* the shiftsize specified to shift the tags. If the model is built, it is computed as half of the estimated fragment length. If the model is not built, default size of 100bp is used.
- *–mfold:* this parameter is used for narrow peak detection and specifies the lower and upper limits of the enrichment factor (ratio of treatment to control). The regions with in $mfold$ range are selected for model building. The default value is [10,30].

To identify the broad peaks, following two additional parameters are recommended to be used with MACS1.4 ([3]):

- $-nomodel$: Due to the increase in the data for histone modification, it is difficult to build a robust shifting model and hence it is recommended not to build the model. It uses default shiftsize 100bp and fragment size 200bp.
- $-nolambda$: When control is not available, the local background estimation using the treatment sample should be skipped. In this case, MACS will not estimate dynamic lambda but rather use constant background to predict peaks.

In addition to using parameters $-nomodel$ and $-nolambda$, the new version of the MACS algorithm (MACS2.0) uses the parameter $-broad$ for detecting broad peaks. With this additional parameter, MACS2.0 links the nearby highly enriched regions. The maximum linking region is 4 times of the fragment length. This should enable MACS2.0 to find larger regions, which may be more appropriate for capturing the signal from histone modification events.

2.2 Data Sets

Table 2 briefly describe the data sets used in this study. We used FoxA1 ChIP-Seq data in human MCF7 cell line [4] and its corresponding control sample for evaluating the algorithms for transcription factor binding sites. In addition, to assess the performance of algorithms in detecting histone modification sites, we used H3K4me3 (GSM307618), H3K27me3 (GSM307619) and H3K36me3 (GSM307620) data sets obtained from mouse ES cell [5]. These data sets can be downloaded from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12241.

Dataset	Organism	Type	Control	Signal
FoxA1	Human	TF	yes	Peak
H3K4me3	Mouse	HM	no	Peak/Region
H3K27me3	Mouse	HM	no	Region
H3K36me3	Mouse	HM	no	Region
TF: Transcription Factor HM: Histone Modification				

Table 2: *Description of data sets.*

2.3 Evaluation Methodology

To detect narrow peaks from FoxA1 ChIP-Seq data, we ran MACS1.4 (from Strand NGS implementation) and MACS1.4 (from command-line using model building option). For broad peak detection, we used FER (from Strand NGS), MACS1.4 (from command-line using advanced parameters, -nomodel and -nolambda), and MACS2.0 (from command-line using additional parameter, -broad). In each case, we evaluated and compared the results in the following ways:

1. *Number and width of peaks*: The peak calling output from different algorithms is first compared based on the number of peaks detected and the distribution of their width.

2. *Overlap (in %bp) between two sets of peaks/regions*: Two sets of peaks and/or regions, which are obtained from either two completely different algorithms or from different parameter settings of the same algorithm, are compared in a comprehensive manner. Let us say, there are two peak sets $P1$ and $P2$, with m and n number of peaks respectively. In order to compare peak sets $P1$ and $P2$, we create a matrix of size $m \times n$, each cell of which will store the overlap between the two peaks. To compute overlap between two individual peaks say $p1$ and $p2$, we used the following two similarity metrics.

   - First similarity metric is given by

   $$Metric1 = \frac{|p1 \cap p2|}{|p1 \cup p2|} \tag{3}$$

   - Second similarity metric is given by

   $$Metric2 = \frac{|p1 \cap p2|}{|p1|} \tag{4}$$

   Here, $|p1 \cap p2|$ represents the number of common bps between peaks $p1$ and $p2$, $|p1 \cup p2|$ represents the union of bps between peaks $p1$ and $p2$, and $|p1|$ represents the width or number of bps in peak $p1$.

   While first similarity metric is more stringent and penalises the similarity score if $p2$ has unique bps which are not in $p1$, second similarity metric does not penalise the score for unique bps in $p2$. For instance, if $|p1|$ is 100bp and $|p2|$ is 200bp with an overlap of 100bp, while score from the first similarity metric will be 0.50, score from the second similarity metric will be 1.0.

3. *Overlap in genes downstream of the peaks/regions*: Two sets of peaks and/or regions are also compared based on the overlap in the genes that are within a distance of 2kb from the center of the peaks/regions. This is done because even if the peaks from the resulting peak sets from different algorithms does not match perfectly, they may still be potentially regulating the same downstream genes.

3 Results and Discussion

Below we present the results obtained on human FoxA1 data set and three mouse histone modification data sets: H3K4me3, H3K27me3 and H3K36me3. As mentioned earlier, while signal from transcription factor binding sites is more discrete and typically depicts distinct positive and negative peaks, signal from histone modification sites is more continuous. Figure 1 and figure 2 shows the behaviour of peaks represented by transcription factor binding sites and histone modification respectively.

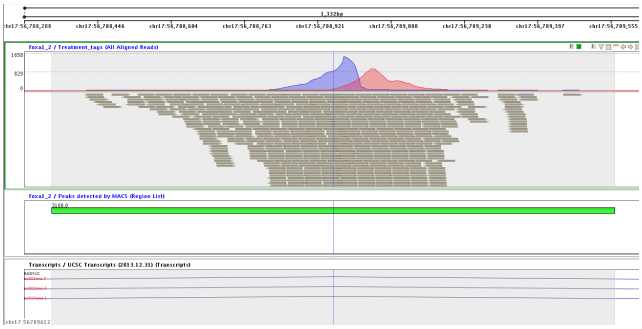

Figure 1: *Characteristics of transcription factor binding sites (narrow peaks).*

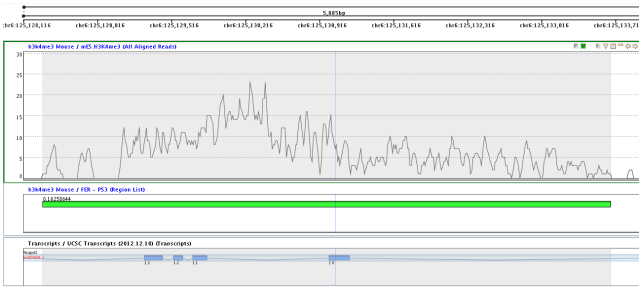

Figure 2: *Characteristics of histone modification sites (broad peaks).*

3.1 Detecting transcription factor binding sites (narrow peaks)

We ran MACS1.4 (Strand NGS implementation) and original MACS1.4 from command-line option on the human FoXA1 data set. The obtained peaks from these two runs were compared in 1) number, 2) peak width distribution and 3) peak-to-peak overlap percentage. In addition, we also compared the genes, which are within a distance of 2kb bp from the peaks obtained from these runs.

3.1.1 *Number and width of detected peaks*

Table 3 shows the number of peaks obtained from human FoxA1 ChIP-Seq data by MACS1.4(Strand NGS) and MACS1.4 (Command-line option). Default parameters were used in both cases.

The number of peaks obtained from MACS1.4 (Strand NGS) and MACS1.4 (Command-line) are quite comparable, however the small difference in the number of peaks may be attributed to the following reasons. During the first step of MACS where enriched regions are identified to build the model, while MACS1.4 (Command-line) uses a bandwidth (sonication size) of 300bp, MACS1.4 (Strand NGS) uses 400bp. In addition, while MACS1.4 (Command-line) uses a lower and upper cut-off of 10 and 30 respectively for the m-fold cut-off, MACS1.4 (StrandNGS) uses a single parameter 'enrichment factor' with a default value of 32.

Algorithm	Number of peaks
MACS1.4 (Strand NGS)	12,623
MACS1.4 (Command-line)	13,591

Table 3: *Number of Peaks detected by MACS1.4 (Strand NGS and original command-line option) on human FoXA1 data.*

We also looked at the distribution of the width of the peaks detected by MACS1.4 (Strand NGS) and MACS1.4 (Command-line). Figure 3 clearly shows that width of the peaks detected are very similar, indicating that Strand NGS implementation of MACS1.4 resembles very closely with original implementation of MACS1.4.

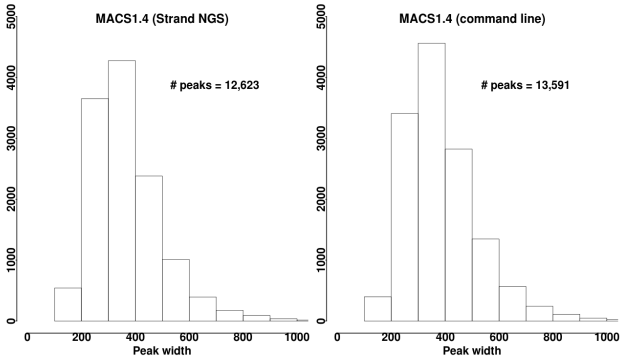

Figure 3: *Distribution of the width of peaks obtained from MACS1.4 (Strand NGS) and MACS1.4 (Command-line).*

3.1.2 *Overlap between the peaks*

We further wanted to compare each peak obtained from MACS1.4 (Strand NGS) with its respective peak detected by MACS1.4 (Command-line) in terms of overlap percentage. As described in section 2.3, two similarity metrics were used to find the answer to the following question: What percentage of peaks detected by MACS1.4 (Strand NGS) have an overlap of x% with at least one of the peaks detected by MACS1.4 (Command-line)?. While the first similarity metric measures $\frac{|A \cap B|}{|A \cup B|}$, the second metric is less stringent and measures $\frac{|A \cap B|}{|A|}$. Figure 4 shows this % of peaks for both the similarity metrics by varying the overlap percent threshold x. It can be clearly seen that there is very high overlap in peaks detected by MACS1.4 (Strand NGS) and MACS1.4 (Command-line). Further the fact that there is a slight drop in % of peaks using the first similarity measure with increase in overlap percentage but practically no drop using the second similarity measure, indicates that peaks by MACS1.4 (Command-line) are slightly wider and

hence MACS1.4 (Strand NGS) peaks are mostly contained within them. Please note that detecting smaller peaks may be more helpful in determining the exact location of the transcription factor binding sites.

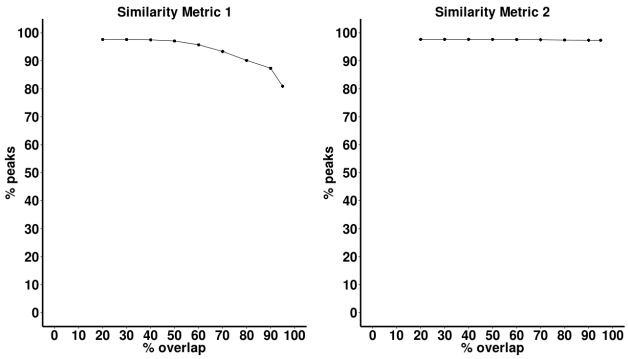

Figure 4: *Percentage of peaks from MACS (Strand NGS) that are overlapping with peaks from MACS (command-line) using two similarity metrics.*

3.1.3 *Genes downstream of transcription factor binding sites*

In addition to comparing the number, width and overlap of the peaks detected by MACS1.4 (Strand NGS) and MACS1.4 (Command-line), we also compared the downstream genes that are within a distance of 2kb bp of the peak centers. We found 4,678 and 4,617 genes corresponding to the peaks detected by MACS1.4 (Command-line) and MACS1.4 (Strand NGS) respectively. The venn diagram in figure 5 shows the comparison of these downstream genes. As expected, there is a high overlap between these two sets with 4,544 common genes.

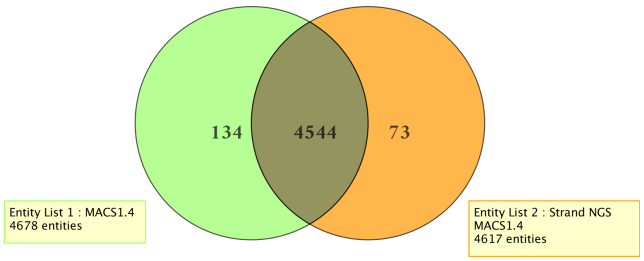

Figure 5: *Comparison of genes annotated by MACS1.4 (Strand NGS) and MACS1.4 (command-line) peaks.*

3.2 Detecting histone modification sites (broad peaks)

We ran FER (implemented in Strand NGS), MACS1.4 (with advanced parameters, -nomodel and -nolambda using the command-line option) and MACS2.0 (with additional parameter, -broad using the command-line option) on three mouse histone modification data sets mentioned in table 2. Please note that for transcription factor binding site detection, MACS1.4 was run **without** advanced parameters, -nomodel and -nolambda. Since the signal from histone modification sites is continuous, we will refer to the detected broad peaks as histone modification enriched regions (HMERs). For evaluation, HMERs obtained from different runs were compared in number, width distribution and overlap percentage. Similar to the previous case, we also compared the downstream genes that are within a distance of 2kb bp from the center of the detected HMERs.

3.2.1 *Number and width of detected HMERs*

Table 4 shows the number of HMERs detected by different algorithms on all the histone modification data sets. Both MACS1.4 and MACS2.0 were run from command-line option with appropriate parameters for broad peak detection.

	Data		
Algorithm	H3K4me3	H3K27me3	H3K36me3
FER-PS1	21,111	2,854	862
FER-PS2	22,495	5,662	5,631
FER-PS3	36,540	19,494	54,467
MACS1.4	36,292	23,903	55,253
MACS2.0	32,807	2,899	248

Table 4: *Number of HMERs detected by different algorithms*

One can make several important observations from table 4. First, the number of HMERs increases in all data sets with parameter setting 1 (PS1) to parameter setting 3 (PS3) of FER. Second, number of HMERs detected by FER-PS3 is comparable to those detected by MACS1.4. Third, MACS2.0 as expected gives smaller number of HMERs compared to MACS1.4 because of merging different regions. However, the difference in the number of HMERs detected by MACS1.4 and MACS2.0 is much more in H3K36me3 or even in H3K27me3 data compared to H3K4me3 data. MACS2.0 is a recent version and may require more in-depth study to understand how each parameter, particularly the newly introduced '-broad' parameter is being used. Therefore, in this study, we will focus more on the comparison of FER-PS3 and MACS1.4. In addition to the number, if the HMERs obtained from FER-PS3 and MACS1.4 also compare well on other evaluation measures, Strand NGS can be confidently recommended for histone modification studies where detection of broad peaks is the goal.

With the above objective in mind, we looked at the distribution of the width of HMERs detected by different algorithms on all three data sets. Figure 6 shows this distribution. Overall the width distribution is very similar, however a slight left shift can be observed in HMERs detected by FER-PS3. This indicates that HMERs detected by FER-PS3 are slightly smaller compared to HMERs detected by MACS1.4. Another interesting point to note is that MACS2.0 not only produces smaller number of HMERs, they are smaller in width as well. This is counter-intuitive to some extent as we were hoping to see smaller number but wider HMERs. We included MACS2.0 to give an initial glimpse into the new parameter and results involved however as mentioned earlier, to understand all the parameters and results of MACS2.0 and compare them appropriately with other algorithms, a separate in-depth

study may be needed. Therefore, detailed comparison of MACS2.0 results is outside the scope of this paper.

3.2.2 *Overlap between HMERs obtained from FER-PS3 and MACS1.4*

Each HMER obtained from FER-PS3 is compared to HMERs obtained from MACS1.4 in terms of the overlap percentage. Question is: What percentage of HMERs detected by FER-PS3 have an overlap of x% with at least one of the HMERs detected by MACS1.4?. Again we use the same two similarity metrics: $\frac{|A \cap B|}{|A \cup B|}$ and the less stringent one $\frac{|A \cap B|}{|A|}$. By varying overlap percentage x from 20 to 100, we show in figure 7 (A1, A2, and A3) the % of HMERs for the first similarity metrics and in figure 7 (B1, B2, and B3) the % of HMERs for the second similarity metric for three data sets.

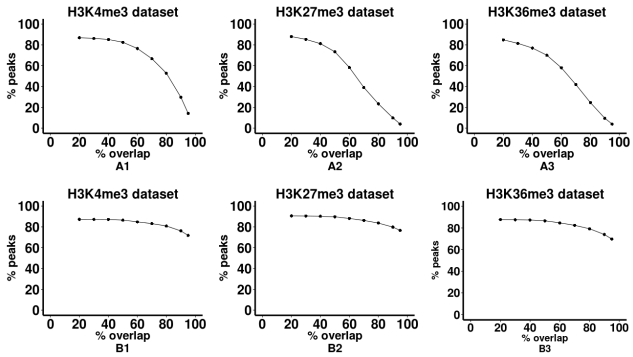

Figure 7: *Comparison of HMERs detected by FER-PS3 with those detected by MACS1.4*

It is clear from figure 7 that with the first similarity metric (equation 3), % of HMERs drops rather rapidly with increase in overlap % while it practically remained constant between 80 - 90 % when second similarity metric (equation 4) is used. This suggests that 80 - 90 % of HMERs detected by FER-PS3 have an overlap of $\geq 80\%$ with at least one of the HMERs detected by MACS1.4 using similarity metric 2. However the fact that % of HMERs drops with increasing overlap % also suggests that HMERs detected by MACS1.4 have several extra bases which are not present in HMERs detected by FER-PS3.

3.2.3 *Overlap between genes downstream of HMERs*

Downstream genes that are within a distance of 2kb from HMERs detected by FER-PS3 and MACS1.4 are also compared. Figure 8, 9 and 10 shows the venn diagram comparing these downstream annotated genes for HMERs obtained by these algorithms from data sets H3K4me3, H3K27me3 and H3K36me3 respectively. For all the 3 data sets, very similar number of downstream genes and high overlap between them again reinforces the hypothesis that FER-PS3 and MACS1.4 identified very similar enriched regions reflective of histone modification sites.

Figure 8: *Comparison of genes annotated using HMERs detected by FER-PS3 and MACS1.4 on H3K4me3 data set.*

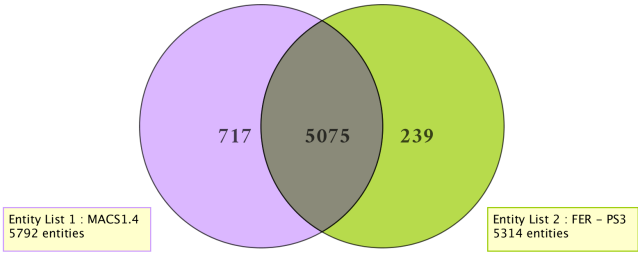

Figure 9: *Comparison of genes annotated using HMERs detected by FER-PS3 and MACS1.4 on H3K27me3 data set.*

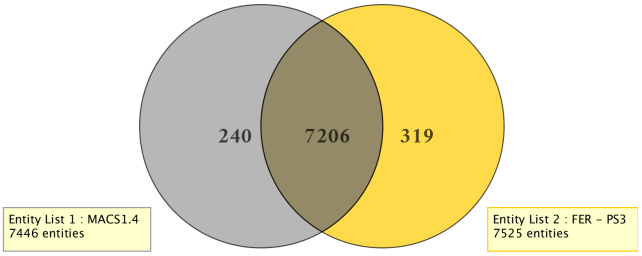

Figure 10: *Comparison of genes annotated using HMERs detected by FER-PS3 and MACS1.4 on H3K36me3 data set.*

3.2.4 *Parameter sensitivity analysis*

Based on the results presented in this study, we've established that for most practical purposes, FER algorithm with appropriate parameters settings is similar to MACS1.4 for detecting broad peaks that are reflective of histone modification sites. Since we recommend using the FER algorithm in Strand NGS for detecting broad peaks, it is also important to understand the sensitivity of the parameters involved on the results. We use H3K27me3 data set and study the sensitivity of the parameters of FER on the number of HMERs detected in the following ways:

- Parameter 'window size (ws)' is varied from 100 bp to 600 bp in increments of 100 bp while keeping other parameters constant.
- Parameter 'slide size (ss)' is varied from 100 bp to 500 bp in increments of 100 bp while keeping other parameters constant.
- Parameter 'minimum reads in a enriched window (mr)' is varied from 10 to 60 in increments of 10 while keeping other parameters constant.

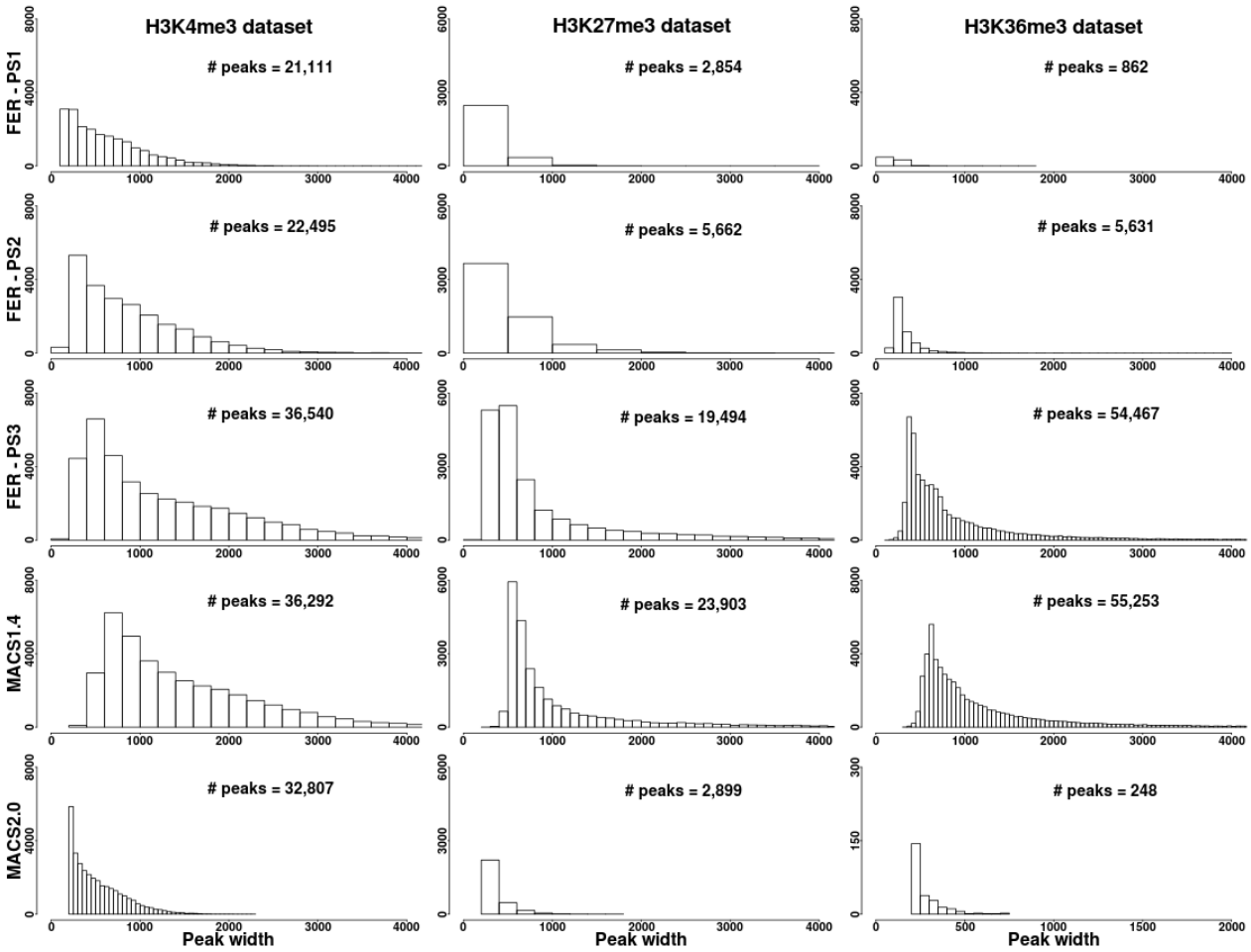

Figure 6: *Distribution of the width of the detected HMERs*

- Parameters, 'ws' and 'mr' are varied together. They are varied from {100 bp, 10} to {600 bp, 60} with 'ws' and 'mr' incrementing by 100 bp and 10 respectively in each step. This is more realistic as the requirement for minimum reads in an enriched window should go up as window size increases.

Intuitively, when we increase the 'ws' keeping 'ss' and 'mr' constant, more windows will be called out as enriched and hence overall we expect to detect more enriched regions. Similarly, with increase in 'ss' or 'mr', we expect to detect less enriched regions because with increase in 'ss', less windows will be considered and with increase in 'mr', less windows will satisfy the threshold. It is interesting to see how the number of enriched regions changes when we vary 'ws' and 'mr' together as they have opposite effect when varied individually. Figure 11 shows the sensitivity of different parameters of FER on the number of enriched regions detected. As expected number of enriched regions increased with increase in 'ws' and decreased with increase in 'ss' and 'mr', however the rate of change appears to be different. While the change in the number of enriched regions is more linear with the change in 'ws' and 'ss', it seems more exponential when 'mr' is varied. Finally, it is interesting to see that when we vary 'ws' and 'mr' together, number of enriched regions decreased, however the change is not exponential as seen with variation in 'mr' alone, but is more linear.

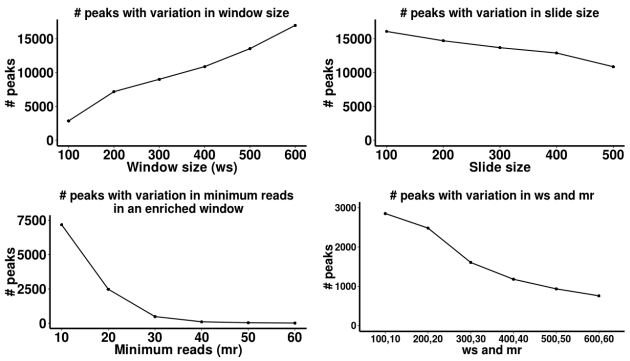

Figure 11: *Sensitivity Analysis*

4 Conclusions

In this study, we discussed the algorithms implemented in Strand NGS for analysing ChIP-Seq data. Specifically we assessed the applicability of the peak calling algorithms implemented in Strand NGS to detect both binding sites of

a transcription factor (narrow peaks) and enriched regions of a histone modification (broad peaks).

For narrow peak detection, both PICS and MACS are implemented in Strand NGS. We demonstrated using a human FoXA1 ChIP-Seq data that results obtained from the Strand NGS implementation of MACS are very similar to those obtained from the original MACS implementation (downloaded and executed via command line). For broad peak detection, we demonstrated using three mouse histone modification data sets that the broad regions detected by Find Enriched Regions (FER) algorithm are very similar to the histone modification events detected by MACS (with advanced parameters). This is encouraging because MACS has already been shown to either compare well or even outperform many existing broad peak callers.

Overall, using the benchmarking results presented in this paper, we concluded that Strand NGS offers a comprehensive and solid approach to analyse ChIP-Seq data. The algorithms implemented in Strand NGS can be used to detect both narrow peaks and broad regions, there by providing a way to study both, cistrome of transcription factors as well as histone modification events and their role in gene regulation.

References

[1] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[2] Jason S Carroll, Clifford A Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Jérôme Eeckhoute, Alexander S Brodsky, Erika Krasnickas Keeton, Kirsten C Fertuck, Giles F Hall, et al. Genome-wide analysis of estrogen receptor binding sites. *Nature genetics*, 38(11):1289–1297, 2006.

[3] Jianxing Feng, Tao Liu, and Yong Zhang. Using macs to identify peaks from chip-seq data. *Current Protocols in Bioinformatics*, pages 2–14, 2011.

[4] Mathieu Lupien, Jérôme Eeckhoute, Clifford A Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S Carroll, X Shirley Liu, and Myles Brown. Foxa1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 132(6):958–970, 2008.

[5] Tarjei S Mikkelsen, Manching Ku, David B Jaffe, Biju Issac, Erez Lieberman, Georgia Giannoukos, Pablo Alvarez, William Brockman, Tae-Kyung Kim, Richard P Koche, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

[6] Leelavati Narlikar and Raja Jothi. Chip-seq data analysis: identification of protein–dna binding sites with sissrs peak-finder. In *Next Generation Microarray Bioinformatics*, pages 305–322. Springer, 2012.

[7] Naim U Rashid, Paul G Giresi, Joseph G Ibrahim, Wei Sun, and Jason D Lieb. Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol*, 12(7):R67, 2011.

[8] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, 27(1):66–75, 2009.

[9] Qiang Song and Andrew D Smith. Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6):870–871, 2011.

[10] Jianrong Wang, Victoria V Lunyak, and I King Jordan. Broadpeak: a novel algorithm for identifying broad peaks in diffuse chip-seq datasets. *Bioinformatics*, page bts722, 2013.

[11] Han Xu, Lusy Handoko, Xueliang Wei, Chaopeng Ye, Jianpeng Sheng, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. A signal–noise model for significance analysis of chip-seq with negative control. *Bioinformatics*, 26(9):1199–1204, 2010.

[12] Chongzhi Zang, Dustin E Schones, Chen Zeng, Kairong Cui, Keji Zhao, and Weiqun Peng. A clustering approach for identification of enriched domains from histone modification chip-seq data. *Bioinformatics*, 25(15):1952–1958, 2009.

[13] Xuekui Zhang, Gordon Robertson, Martin Krzywinski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo. Pics: Probabilistic inference for chip-seq. *Biometrics*, 67(1):151–163, 2011.

[14] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

strand

New Generation Healthcare

Strand was founded in 2000 by computer science and mathematics professors from India's prestigious Indian Institute of Science who recognized the need to automate and integrate life science data analysis through an algorithmic and computational approach. Strand's segue into the life sciences was through informatics products and services for research biologists, chemists, and toxicologists that combine advanced visualization, predictive systems modeling, data integration and scientific content management - over 2000 research laboratories worldwide (about 30% of global market share) are licensees of Strand's technology products, including leading pharmaceutical and biotechnology companies, research hospitals and academic institutions. With a recent investment by Biomark Capital, Strand has grown its established team to over 200 employees, many with multidisciplinary backgrounds that transcend computation and biology.

Since 2012, Strand has been expanding its focus to include clinical genomics, spanning sequencing, data interpretation, reporting and counseling. Strand operates a 10,000 square foot laboratory space with state-of-the-art clinical genomics capabilities and is also establishing Strand Centers for Genomics and Personalized Medicine in several hospitals around the world to serve as outreach points for genomic counseling. Based on the experience gained from sequencing, analyzing, interpreting and reporting on clinical samples over a wide variety of clinical indications, Strand has developed an end-to-end solution for clinical labs that handles all stagesfrom analysis to reporting. The interpretation and reporting software platform has been designed and developed specifically for the medical professional, ranging from the molecular pathologist to the physician. By enhancing sequence-based diagnostics and clinical genomic data interpretation using a strong foundation of computational, scientific, and medical expertise, Strand is bringing individualized medicine to the world.

For more information about Strand, please visit **www.strandls.com**,
or follow us on twitter @StrandLife.

INDIA
5th Floor, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560024

USA
548 Market Street, Suite 82804, San Francisco, CA 94104