Caroline Townsend
December 2012
Biochem 218

A critical review of ChIP-seq enrichment analysis tools

Introduction

Transcriptional regulation, chromatin states, and genome stability pathways are largely governed by interactions between DNA and proteins (Schmidt et al., 2009, Park et al., 2009). To understand these biological processes on a global scale, it is important to map specific protein-DNA interactions throughout the genome. For example, the global mapping of protein-DNA interactions and histone mark modifications has lead to the discovery that monomethylation of specific lysine residues on histone tails are linked to gene activation while certain combinatorial trimethylations of histone tails are linked to gene repression (Barski et al., 2007; Gordon Roberston et al., 2008). This genome-wide protein-DNA interaction data is gathered through the technique of chromatin immunoprecipitation (ChIP) followed by deep sequencing. This method, known as ChIP-seq, relies on immunoprecipitating (IP-ing) a fragment of DNA bound by a protein of interest followed by sequencing of the isolated DNA to determine sites of enrichment throughout the genome (Johnson et al, 2007; Barski et al, 2007; Robertson et al, 2007). Importantly, the data generated by ChIP-seq depends heavily on the computational tools applied to the sequencing analysis. In this review, I will discuss the methodology and data output of ChIP-seq and then focus on describing and critiquing a variety of different programs, known as peak calling algoirhtms, which exist to determine enriched regions of the genome. It is important to appreciate the different programs available for ChIP-seq enrichment analysis, and the strengths and weaknesses of each approach.

Methodology

The ChIP-seq technique relies on being able to detect *in vivo* protein-DNA interactions. To achieve this, researchers rely on covalently cross-linking DNA and protein, commonly with formaldehyde, *in vivo* in order to capture protein dynamics (Solomon et al., 1988). Cells are then lysed and the chromatin is sonicated into small (200-600 base pair) fragments. Researchers must take care to sonicate properly, as too large of fragments will not IP efficiently and create higher background, while fragments

that are too small can disrupt the protein-DNA interaction. At this point, a fraction of the sonciated chromatin is taken as the input sample while the remainder is used for immunoprecipitation (IP). An antibody to a protein of interest is then incubated with the sonicated chromatin. An antibody that is high quality and specific is essential, and it is important to note that cross-linking can mask some target epitopes. Beads are then coupled to the antibody and the antibody is used to pull down both the protein of interest and the corresponding DNA fragment to which the cross-linked protein is bound. Following this pull down, it is important to reverse the cross-links by heating the sample at 65^{o} C overnight and to treat with RNase A and proteinase K such that only the IPed DNA remains. This DNA along with the input DNA is then purified and prepared for sequencing (Figure 1).

The methodology of the library preparation for sequencing varies based on the next generation sequencing (NGS) platform used; described here is the sample preparation for sequencing on an Illuminia Genome Analyzer (Schmidt et al., 2009; Illuminia, 2007). First, the T4 DNA polymerase, Klenow polymerase, and T4 polynucleotide kinase are used for end repair, converting the DNA overhangs into phosphorylated blunt ends. To prepare the samples for adaptor ligation, an A base is added to the newly generated 3' blunt phosphorylated end. This A pairs with the single T overhang on the sequencing adaptor. After the adaptors have been ligated to DNA ends, the library is size selected for the desired template size via gel extraction. These adaptor-containing fragments can then be amplified by PCR. Because this step can create bias as some fragments amplify more efficiently then others, the use of controls is vital. The PCR-amplified library can then be run through a flow cell sequencer.

Proper ChIP controls are important for determining accuracy of the experimental technique, and the computational analysis. For example, it is vital to control for non-uniform sonication that occurs across different regions of the genome due to changes in chromatin state and repetitive sequences (Park, 2009). It is also recommended to control for nonspecific antibody binding, and bias in the amplification of different fragments in library construction (Park, 2009; Bardet et al., 2012). In a simple two-sample experiment, moreover, a negative control is strongly recommended to build a model of background noise (Ji, 2010). Common controls, which are described below and each test for slightly

different artifacts, include an input sample, a mock sample, or a nonspecific antibody sample.

Any enrichment in a ChIP-seq profile should be determined relative to the input sample in the same genomic region. By comparing enrichments of read counts between a ChIP sample and its input sample, it minimizes bias in sonication and in the PCR amplification and consequent sequencing steps (Park, 2009). The steps for normalization to the input are briefly described in this review, in the data output and analysis section. A mock sample, where no antibody is used for the IP, is a control that can minimize background of the IP and sonication efficiency. It is not commonly utilized, however, because the amount of DNA pulled down in the mock is limited and often insufficient for library generation and sequencing (Kharchenko et al., 2008; Bardet et al, 2012). Another control is the non-specific antibody (such as immunoglobulin G or GFP), whose target does not interact with chromatin. The non-specific antibody control is similar to the mock control in artifact detection, but also suffers from the same limitations. Thus, given the range of artifacts input controls for as well as the large amount of DNA the input yields, input is considered to be the most reliable and commonly-used control (Park et al., 2009; Schmidt et al., 2009).

Data Output and Analysis

There are a variety of NGS platforms available that provide high resolution, in-depth coverage for ChIP-seq analysis. Roche454, Life SOLID3, Illumina GAII, Helicos Heliscope, and Pacific Biosystems RS system are common platforms that rely on pyrosequencing, sequencing by ligation, and sequencing by synthesis (Illuminia, Helicos, and Pac Bio) respectively (Zhou et al., 2010). The basic work scheme for data generated from a sequencing platform for ChIP-seq is shown (Figure 2). The first step of analysis is genome alignment, which is another aspect of ChIP-seq processing that requires a great deal of computational work. Because it is not the focus of this review, a brief description is provided. The generated NGS data must first go through platform-dependent image analysis, also known as base calling, where each nucleotide base is identified. For example, the Illumina base-calling program depends on assigning the nucleotide sequences from the fluorescence trace that is created when each fluorescently-labeled nucleotide is incorporated into the complementary strand during sequencing (Ledergerber

and Dessimaz, 2011). These stretches of 30 – 50 base pair sequences assigned through base-calling are known as sequence tags.

Once the sequence tags are generated, they must be aligned back to a reference genome. This alignment is challenging, as many short reads need to be mapped back to a large reference data set while minimizing sequence errors but allowing for genomic variation (Fonseca et al., 2012). Many mapping programs exist, and most rely on the following principle, as described in Fonseca et al., 2012. " Given a set of sequences *Q* (produced by a HTS [high throughput sequencing] technology), a set of reference sequences *R,* a possible set of constraints, and a distance threshold *k,* find all substrings *m* of *R* that respect the constraints and that are within a distance *k* to a sequence *q* in *Q,* i.e., $d(q,m) \leq k$, where *d*() is some distance function. The occurrences *m* in R are called *matches.* The constraints imposed can vary depending upon the HTS application and data type (e.g., whether the data generated are single reads (most common for ChIP-seq, or pair-end reads)." Genome matching software works to find the sequence *q* among all the reference data.

Some mapping algorithms also take into account platform-specific biases. For example, Illuminia sequencing is less accurate as read cycle number increases, making the 3' end of each read less reliable. It has been shown that mismatches toward the 3' end of tags make up 41-75% of total mismatches (Kharchenko et al., 2008). Therefore, algorithms such as Bowtie remove several of the 3' ends from the read (Fonseca et al., 2012). Moreover, many algorithms utilize the base quality score that is produced by the sequencer during base-call analysis to generate a more accurate alignment. Use of the quality score can lower the frequency of alignment errors by assigning bases that have low scores a decreased penalty of mismatch (Li and Homer, 2010; Fonseca et al., 2012). Importantly, only reads that map to single unique location are used (Taslim et al., 2012). A list of the available mapping algorithms, along with read length limits, how the tags are aligned (end-to-end or locally), and whether gaps such as insertions and deletions are allowed in the alignment is shown in Figure 3. The choice of mapping algorithm to use for ChIP-seq alignment thus depends on the sequencing platform used and downstream analysis tools.

One challenge to generating reliable ChIP-seq data is obtaining the appropriate sequencing depth. Because ChIP-seq reads map to only a subset of the genome, many publications require 100x coverage. Moreover, depending on the type of protein being examined in ChIP and the number of corresponding binding sites in the genome, a large number of tags may be necessary to cover each binding site at the same density (Park, 2009). For example, many more reads would be needed for proper ChIP-seq analysis of a histone modification that spans a large portion of the genome than of a transcription factor that binds only several discrete sites along the genome. The best way to determine if sequencing depth is sufficient is to find a 'saturation point;' beyond this point additional reads should not result in any additional enrichment or change in binding sites (Park, 2009).

Genome Density Analysis and Normalization

After tags have been aligned to the reference genome, the data needs to be transformed into count-per-position data, also known as genomic densities. To generate genomic densities, nearby reads are grouped together. A sliding window algorithm, which is commonly applied, calculates the number of tags found in a fixed window across the entire genome (Wilbanks and Facciotti, 2010). Another option is to "[extend] the alignments beyond the tag length and [record] a tag count at each position *N* bases downstream of the alignment start", where *N* is equal to the average insert size (Leleu et al., 2010). This method creates a smoothed density, resulting in largely contiguous regions. The caveat to this approach is that it estimates fragment size and assumes fragments are uniform (Park, 2009). Genomic density data is important for control analysis as contiguous regions generate a more uniform normalization of the ChIP data (Leleu et al., 2010).

Normalization to the ChIP control sample is essential to extract information about binding enrichment of a protein of interest. The data generated from a control and the ChIP data itself is very similar (Auerbach, et al., 2009) and there is often a similar bias in the distribution of sequence reads both in the input and ChIP sample. Therefore, it is important to quantitatively compare the two samples using a linear regression of the genomic densities and "scale them globally by the slope of the regression" (Leleu et al., 2010). Nonlinear normalization is important for comparing several ChIP samples, such as

before and after a specific treatment where only a subset of genes will be affected. In this analysis, the data should also be normalized with respect to the mean and then with respect to the variance (Taslim et al., 2012). Because there are a variety of ways to normalize the data, depending on the type and number of ChIP samples being compared, it is essential to consider the experimental design as well as the desired output prior to normalization.

Peak calling program analysis

The next step after the ChIP reads have been aligned to the genome and normalized with respect to a control sample is to look for enriched regions known as ChIP peaks. These peaks ultimately allow the researcher to localize the binding site of their protein of interest. Peak calling depends on the analysis of tag density profiles. As described by Pepke et al. (2009) and Leleu et al. (2010), these profiles are typically classified into three major categories: punctate/sharp patterns like those of transcription factors, localized but broad profiles like those created by active histone marks, and extended broad regions like those of an inactive histone mark that is found widely throughout the genome. These three major types of peaks present a major challenge for ChIP-seq analysis, as the majority of software is designed to assign only sharp peaks (Pepke et al., 2009).

As of 2010, there were 31 different algorithms available for determining ChIP-seq peaks (Wilbanks and Facciotti, 2010). The basic steps of a peak calling software are outlined here (and shown in Figure 4). The first step of many peak calling software includes determining genomic enrichment (already briefly described in the *Genome Density Analysis and Normalization* section of this review). The program often then provides 1) a signal profile along each chromosome, 2) a background model, 3) peak call criteria, 4) a post-call filtering of artifact peaks, and 5) the associated significance of the called peaks (Pepke et al., 2009). While not all programs take into account the above criteria, the majority of programs rely on similar principles. The peak calling algorithms that this review focuses on are: CisGenome, MACS, PeakSeq, and SISSRs. These programs were selected for analysis because they are commonly used and comparative analysis of these programs has been published (Wilbanks and Facciotti, 2010).

CisGenome Program

The CisGenome algorithm builds a signal profile by moving a sliding window of a fixed width across the genome and summing the tag counts in each window with a summed value in the center of the window (Pepke et al., 2009; Ji et al., 2008). Filtering is done to prevent duplicate tag reads. In a one-sample analysis, CisGenome identifies regions that have tag counts greater than a user-chosen cutoff point. The false discovery rate is then estimated by predicting the read count in "nonbinding windows using a negative binomial distribution." (Ji et al., 2008). The use of the negative binomial distribution is different than the Poisson distribution used by MACS, for example, in that it allows the rate of background reads to vary across the genome (Ji et al., 2008).

The CisGenome algorithm differs slightly for a two-model system (such as a ChIP sample and a control sample). In this case, it uses a conditional binomial model to call regions that are significantly enriched in the ChIP sample relative to the control. Windows of fixed width are again used to identify predicted binding regions via a user-specific false discovery rate (Ji et al., 2008). P-values are provided for enriched peaks. Interestingly, tag shifting (required in MACS and SiSSRs prior to processing) is used only to refine the analysis such that the modes of 5' and 3' tag peaks are then used to define the binding boundaries (Ji et al, 2008). Because tag shifting is thought to make identifying the precise binding site more accurate, some consider this a weakness of the CisGenome program (Wilbanks and Facciotti, 2010). One of the strengths of the CisGenome algorithm is that it is integrated with a user-friendly browser for visualizing mapped data and peaks (Figure 5).

MACS Program

MACS (model-based analysis of ChIP-seq) is a common peak calling algorithm which analyzes short read sequences and models the shift size of the ChIP-seq tags to improve resolution of predicted binding sites (Figure 6) (Zhang et al., 2008). First, MACS scales the total control tag count to be identical to the tag count from the ChIP experiment and removes all redundancies in sequencing tags so that each location in the genome contains no more than one tag in order to reduce error (Zhang et al., 2008). The algorithm then specifically uses sliding windows across the genome equal to 2 x a given sonication size to search for regions with tags that are considered enriched relative to

random tag distribution patterns. Importantly, the algorithm samples 1,000 of these regions of enrichment and "aligns them by the midpoint between their Watson and Crick tag centers" (Zhang et al., 2008). MACS shifts all the tags by *d*/2, where *d* is the distance between the mode of the Watson and Crick peaks.

After this tag shift has occurred, a sliding window is used to identify peaks with enriched tags based on a Poisson distribution. The MACS model using the Poisson distribution is a better measure of enrichment than simply examining the fold ratio of the ChIP signal relative to the control because it can take into account the statistical significance of the number of samples being described (Park, 2009). As Park (2009) described, determining fold enrichment from a tag ratio does not take into account whether 50 ChIP tags and 10 control tags or 500 tags ChIP tags and 100 control tags were being compared, despite the fact this affects the statistical significance. Moreover, the MACS algorithm can model background tag distribution based on either a random distribution model or through use of a control dataset and can account for regional biases in tag density (Zhang et al., 2008; Wilbanks and Facciotti, 2010). The precise protein binding location, also known as the summit, is then identified as the location that has the most fragment overlap. Candidate peaks with a p-value threshold below a set user defined threshold (often 10^{-5}) are called (Zhang et al., 2008) and fold enrichment is provided.

PeakSeq Program

The PeakSeq algorithm relies on extended tag aggregation to form a fragment density map. Specifically, reads on either strand are extended toward the 3' in order to have the average DNA fragment length (Rozowsky et al., 2008). Peaks are determined through two rounds of analysis (Figure 7). Peaks are initially called by comparing the extended tag aggregation to a simulation for each segment of the genome. A threshold is then determined that meets the user-set false discovery rate, and this threshold is used to find potential target sites (Rozowsky et al., 2008). PeakSeq, unlike MACS, CisGenome, and SiSSRs, applies a level of post-filtering where regions are subdivided into more than one summit call (Pepke et al., 2009). While this is intended to generate more precise genome binding information, no conclusive analysis on this type of filtering has been published. Interestingly, PeakSeq differs from CisGenome, MACS, and SISSRs in that PeakSeq also does not filter out duplicate reads that map to the same region of the

genome (Wilbanks and Facciotti, 2010). While the lack of filtering can create bias in the analysis, other software packages can be used to filter the data.

The peaks are then normalized to the control sample through an algorithm that selects the fraction of peaks to exclude and sums the tag counts in both the ChIP and control sample. A linear regression model is used to determine the scaling factor, which is used for normalizing the number of mapped fragments in the ChIP and control sample (Rozowsky et al., 2008). As a second round of analysis, the fold enrichment of the peaks is then calculated by determining the number of tags relative to the input. Q-values for the called peaks are calculated using a binomial distribution (Pepke et al., 2009).

SISSRs Program

The SISSRs (Site Identification from Short Sequence Reads) package relies on the density and direction of reads along with the DNA fragment length to determine peaks. Peak criteria are determined by a window scanning method (Jothi et al., 2008). The window size is *w* nucleotides, and consecutive windows overlapping by *w/2* are used, similar to the CisGenome algorithm. This model, like the MACS model, mandates a shift size of the ChIP-seq tags to improve resolution of predicted binding sites. The net tag count is for SISSRs is calculated in a given window by subtracting the number of tags mapped to the antisense strand from the number of tags mapped to the sense strand (Figure 8) (Narlikar and Jothi, 2012). When the net-tag count moves from a positive to a negative value, this region is considered a candidate-binding site. The program then determines if the candidate binding site meets a variety of other criteria, which are well-described in Jothi et al. (2008).

SISSRs provides a model to account for background; control data is recommended to be substituted for the default model. The fold enrichment is the calculated by determining the ratio of the ChIP tag number in a given genomic loci to the tag number in the same location of the control data (normalized by total number of tags in each data set) (Jothi et al., 2008). Like many of the algorithms, only peaks with p-values less than the user-set p-value threshold are called as true binding sites. It is interesting to note that because SISSRs relies on strand-specific tag densities where the space between peaks depends on the fragment length (and fixed width peaks are used), SISSR is

considered best for looking at only punctate/sharp binding patterns like transcription factor binding (Wilbanks and Facciotti, 2010).

Comparing Peak Calling Program Performance

While CisGenome, MACS, PeakSeq, and SISSRs each function to find regions of enrichment in ChIP-seq data, the profile, peak criteria, normalization, and filtering options differ. In order to compare the performance of these algorithms to one another, Wilbanks and Facciotti (2010) evaluated three different transcription factor (NRSF, GABP, and FoxA1) ChIP-seq and control datasets using different peak finding algorithms, including CisGenome, MACS, PeakSeq, and SISSRs. Because all three transcription factors used in the study have characterized binding motifs, these factors could be used to monitor peak quality and confidence. The first characteristic, sensitivity, was assayed by determining how many peaks each algorithm identified. MACS, SISSRs, and PeakSeq all identified a relatively large number of peaks for each dataset (Figure 9). For the GABP transcription factor, approximately 15,000 peaks were identified with MACS, SISSRS, and PeakSeq while only approximately 9,000 peaks were found with CisGenome. This likely reflects that CisGenome has more stringent default peak finding settings (Wilbanks and Facciotti, 2010).

In order to determine which peaks were shared between the different algorithms, the authors conducted a pair-wise comparison. The percentage of total peaks that is shared with another algorithm is represented is shown (Figure 10)**.** For the NRSF transcription factor, it is interesting to see that a small peak list from CisGenome is almost completely contained (100%) within a much larger peak algorithm such as MACS (Wilbanks and Facciotti, 2010), while the CisGenome peaks make up only 19% of the total MACs peaks (Figure 10, Wilbanks and Facciotti, 2010). The CisGenome peaks make up 44% of the PeakSeq peaks and 31% of the SISSR peaks. Even SISSRs and PeakSeq selected peaks made up only 45% (46% and 43% respectively) of the total MACS peaks. Similar trends were observed for the other two datasets.

The authors then examined the sensitivity of each peak calling algorithm. Sensitivity is defined by an algorithms ability to call true peaks. By examining quantitative PCR (qPCR) validated true positive binding sites, they found that

CisGenome did miss several true-positive peaks (Figure 11). More surprisingly, however, was that SISSRs (which called nearly as many peaks as MACS) had an even lower sensitivity and missed multiple true-positive peaks. This suggests that the directionality scoring method relied on by SISSRS is less sensitive and perhaps more error prone.

Lastly, positional accuracy of peaks was compared. While CisGenome, PeakSeq, and MACS report peaks of variable width, SISSRs reported narrow fixed-width peaks. To analyze the accuracy of binding positions, the authors calculated the distance between each predicted binding coordinate and the center of binding motifs with a certain distance. The results were that MACS and SISSRs provided the best spatial resolution, likely because the peak search strategy the algorithms employ depends on identifying transition points of tag densities between the two strands (Wilbanks and Facciotti, 2010). Moreover, unlike many other peak calling algorithms, MACS has been shown to work well in calling both punctate/sharp peaks and broad signals associated with histone modifications (Feng et al., 2011). A summary of the approaches used by the four peak calling algorithms is shown (Figure 12). The choice of ChIP-seq peak calling algorithm should depend on a balance of sensitivity and precision, considering the experimental target and downstream analysis.

The Future Challenges of ChIP-seq

ChIP-seq offers many advantages over previous technologies aimed at determining sites of genomic enrichment for a protein of interest. ChIP-seq offers better resolution, increased coverage, and fewer artifacts than ChIP-chip (ChIP followed by microarray). Moreover, while the cost of ChIP-seq was once prohibitive (Park, 2009), the price of using NGS technology has continued to decrease (Davey et al., 2011). While there are technical concerns including sonication uniformity, PCR bias, and antibody specificity, one of the main challenges remains properly analyzing and validating results. Bench-top biologists may not be sufficiently trained in computational approaches to interrogate and apply the proper alignment and peak calling algorithms. While some algorithms, such as CisGenome, offer a user-friendly platform for viewing the data, there should be more focus on making programs accessible. Moreover, while a large number of peak calling algorithms exist, relatively few are designed for peaking up broad and mixed

peaks associated with many epigenetic marks (Pepke et al., 2009; Wilbanks and Facciotti, 2010).

Validating ChIP-seq findings and determining the biological significance is likely the most challenging aspect of ChIP-seq. While all of the algorithms critiqued in this review assign peaks some statistical significance, ChIP-seq peaks can be validated with qPCR and, if they exist, true-positive binding sites can be confirmed. It is more difficult to determine if an identified binding site for a transcription factor has any functional relevance, for example. If a transcription factor requires co-factors for its activation, researchers should also ask by ChIP-seq if those co-factors are enriched at the identified binding sites. If a transcription factor is being analyzed, researchers could also confirm that mutating the identified binding site alters expression. These confirmations, however, are laborious and time intensive. It may become standard practice to integrate ChIP-seq results with other high throughput technologies such as RNA-seq to have a more global perspective (Park, 2009). Thus, it is likely that ChIP-seq experiments and analysis will require generating and analyzing extremely large amounts of data. It is imperative that the computational resources and knowledge exist to best utilize this data.

Figures

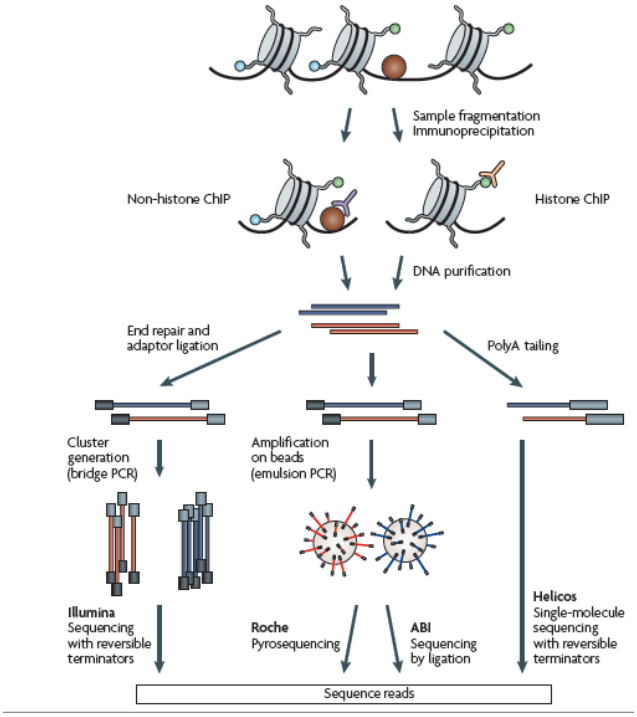

Figure 1. An overview of the ChIP-seq methodology. ChIP is performed on sonicated chromatin with an antibody to a protein of interest and relies on *in vivo* cross-linking of the protein to the DNA prior to IP. After the IP, the cross-links are reversed, the DNA is purified, and prepared for sequencing by generating a library. Many different platforms exist for sequencing, with the most common platforms being represented here. *Figure from Park, 2009.*

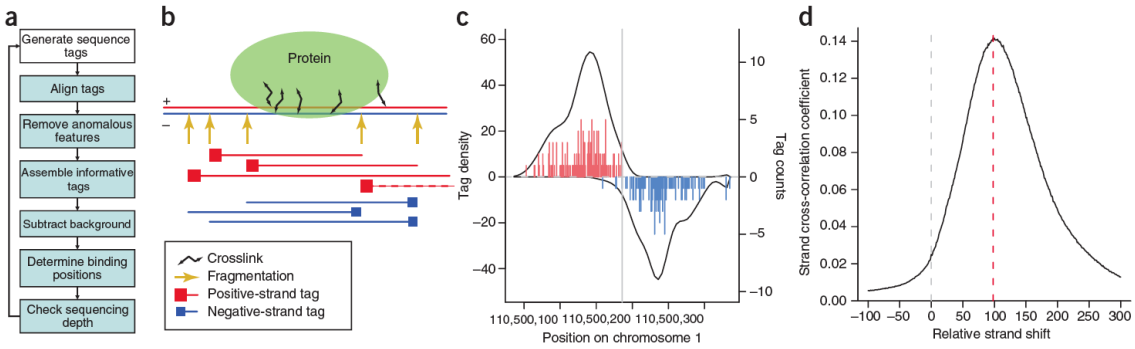

Figure 2. A) An overview of the important computational steps in processing ChIP-seq data. B) Scheme of ChIP-seq measurements. Commonly, the 5' ends (shown in red and blue squares) are sequenced. This creates groups of positive and negative strand tags. C) Illustration of potential tag distribution around a potential binding site. D) Strand cross correlation. The peak represents the distance separating seperating positive and negative strand peaks associated with a binding site. *Figure and adapted legend from Kharchenko et al., 2008.*

Mapper	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align. Reported	Alignment	Parallel	QA	PE	Splicing	Data
BFAST		*	Y	Y	Y	B,R,U	G	SM	N	Y	N	DNA
Bismark	16	10K	Score	Score	N	U	-	SM	Y	Y	N	Bisulfite
Blat	11	5000K	Score	Score	Y	B	L	N	N	N	De novo	DNA
Bowtie	4	1K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N	DNA
Bowtie2	4	5000K	Score	Score	Y	A,B,R,S	G L	SM	Y	Y	N	DNA
BS Seeker	-	-	3	0	N	U	-	SM	Y	N	N	Bisulfite
BSMAP	8	144	15	0	N	B,S,U		SM	N	Y	N	Bisulfite
BWA	4	200	Y	8	Y	R,S	G	SM	Y	Y	N	DNA
BWA-SW	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N	DNA
BWT-SW		1K	Score	Score	Y	A		N	N	N	N	DNA
CloudBurst		1K	Y	Y	Y	A,B	G	Cloud	N	N	N	DNA
DynMap	18	8K	5	0	N	B	L	N	N	N	N	DNA
ELAND		32	2	0	N	B		N	N	N	N	DNA
Exonerate	20	*	Score	Score	Y	B,S	G L	N	N	N	De novo	DNA
GEM	0	4294M	1.0	1.0	Y	A, S	G	SM	Y	Y	Lib and de novo	DNA
GenomeMapper	12	2K	10	10	Y	A,B,R	G	SM	N	N	N	DNA
GMAP	8	*	Y	Y	Y	B	G L	SM	N	N	De novo	DNA
GNUMAP	16	1K	Score	Score	Y	B	G	SM/DM	Y	N	N	DNA
GSNAP	8	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and de novo	DNA
MapReads	10	120	Score	0	N	S		N	Y	N	N	DNA
MapSplice	-	-	3		Y	B	-	SM	N	Y	De novo	RNA
MAQ	8	63	Y	Y	N			N	Y	Y	N	DNA
MicroRazerS	10	*	Score	0	N	S	G	N	N	N	N	miRNA
MOM			Y	0	N	A	L	SM	N	Y	N	DNA
MOSAIK	15	1000	Y	Y	Y	A,B	G	SM	Y	Y	N	DNA
mrFAST	25	300	Score	6	N	A,B	G	N	N	Y	N	miRNA
mrsFAST	25	200	Y	0	N	A	G	N	N	Y	N	miRNA
Mummer 3	10	*	Y	Y	Y	A,B	G	N	N	N	N	DNA
Novoalign	30	300	8	2	N	A, B, R, U, S	G	SM/DM/Cloud	Y	Y	Lib	DNA
PASS	23	1K	Y	Y	Y	A,B	G	SM	Y	Y	De novo	DNA
Passion	-	-	Y	Y	Y	U	-	SM	Y	Y	De novo	RNA
PatMaN	1	*	Y	Y	N	A	G	N	N	N	N	miRNA
PerM	20	128	9	0	Y	A,U	G	DM	Y	Y	N	DNA
ProbeMatch	36	50	3	Y	N	A,B		N	N	N	N	DNA
QPALMA	-	-	Y	Y	Y	B	L	N	Y	N	Lib and de novo	RNA
RazerS	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N	DNA
REAL	4	*	Score	N	N	B, U	G	SM	Y	N	N	DNA
RMAP	11	10K	Y	0	N	B,S		N	Y	Y	N	DNA
RNA-Mate	-	-	Y	0	N	S	-	DM	Y	N	Lib	RNA
RUM	-	-	Y	Y	Y	B	-	SM	N	Y	De novo	RNA
SeqMap	15	500	5	3	N	A		SM	N	N	N	DNA
SHRiMP	14	1K	Score	Score	Y	B,S	G	SM	N	Y	N	DNA
SHRiMP 2	30	1K	Y	Score	N	B,U,S	G	SM	Y	Y	N	DNA
Slider		62	3	0	N	B,S		N	Y	Y	N	DNA
Slider II		93	Y		N	B,S		N	N	Y	N	DNA
Smalt	4	2048M	Score	Score	N	A,B,R,U,S	L	SM	Y	Y	N	DNA
SOAP	7	60	5	3	N	B,R S		SM	N	Y	N	DNA
SOAP2	27	1K	2	0	Y	A,B,R	L	SM	N	Y	N	DNA
SOAPSplice	13	3K	5	2	Y	U	-	SM	Y	Y	De novo	RNA
SOCS		64	Y	0	N	A,B		SM	Y	N	N	DNA
SpliceMap	-	-	0.1		Y	A	-	SM	N	Y	Lib and/or de novo	RNA
SSAHA	15	*	Y	Y	Y	B,S	G L	N	N	N	N	DNA
SSAHA2	15	48K	Score	Score	N	B,S	L	N	N	Y	N	DNA
Stampy	4	4K	0.15	30	N	B,R,S	G	N	Y	Y	N	DNA
Supersplat			0	0	Y	A,U	G	N	N	N	De novo	RNA
TopHat	-	-	2	0	N	B,S	-	SM	Y	Y	De novo	RNA
VMATCH			Score	Score	Y	A,B,S	G L	N	N	N	N	DNA
WHAM	5	128	5	3	N	A,B,R,U,S	G	N	Y	Y	De novo	DNA
X-Mate	-	-	Y	0	N	S	-	DM	Y	N	Lib	DNA
ZOOM	12	240	Y	Y	N	B,S,U	G	SM/DM	Y	Y	N	DNA

Figure 3. A table of available mapper programs for mapping reads to the genome following sequencing. *Figure from Fonseca et al., 2012.*

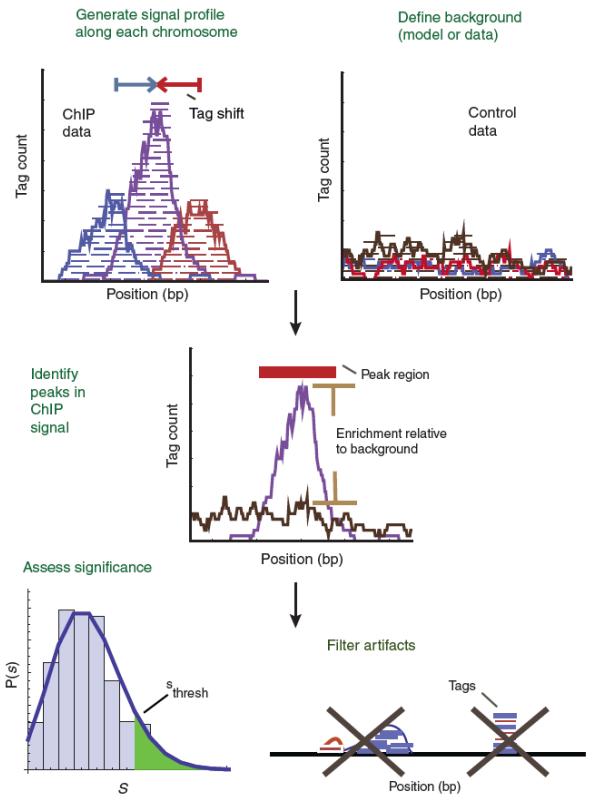

Figure 4. The important steps in peak calling of ChIP-seq analysis. A profile of aligned reads is formed. This could occur by summing the number of reads that overlap each base pair in the genome. The same analysis would then be applied to the control ChIP-seq data, if available. If no control data is available, a random genomic background is often used. Peaks are then filtered and scored with statistical significance. *Figure from Pepke et al., 2009.*

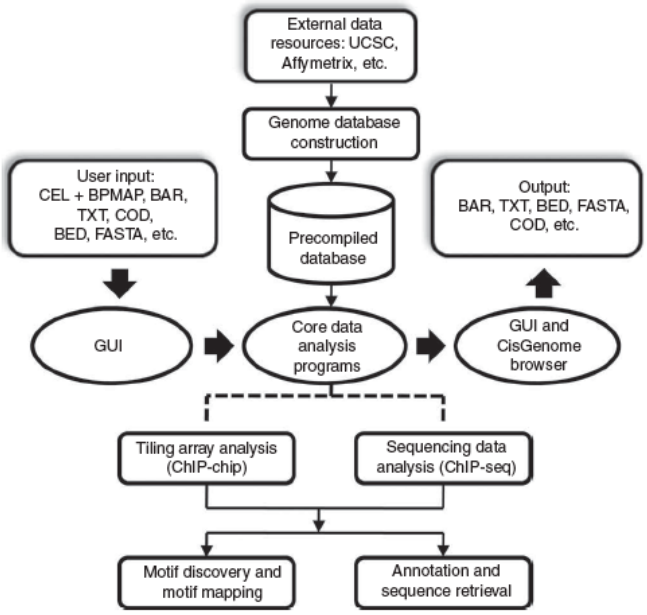

Figure 5. The overview of the workflow for the CisGenome algorithm. "CisGenome contains three core components: a GUI, the built-in CisGenome browser, and of underlying data analysis algorithms. The GUI allows users to load raw data and choose specific analysis functions. Core programs carry out the analysis and results displayed in the CisGenome browser can be exported in various formats." *Legend and Figure from Ji et al., 2008.*

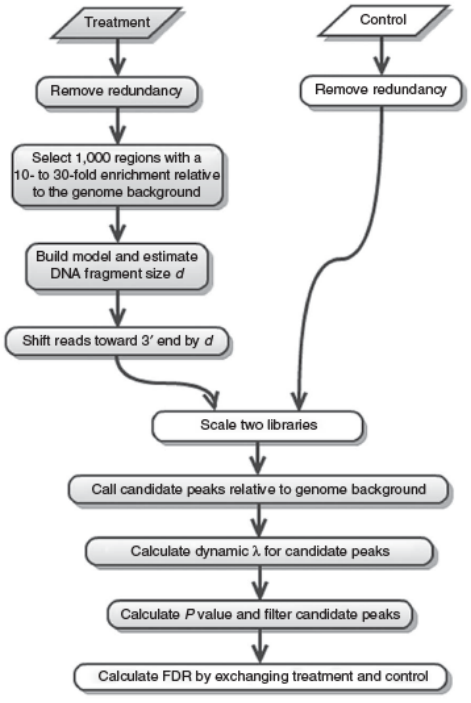

Figure 6. MACS Algorithm workflow overview. *Figure from Feng et al., 2012.*

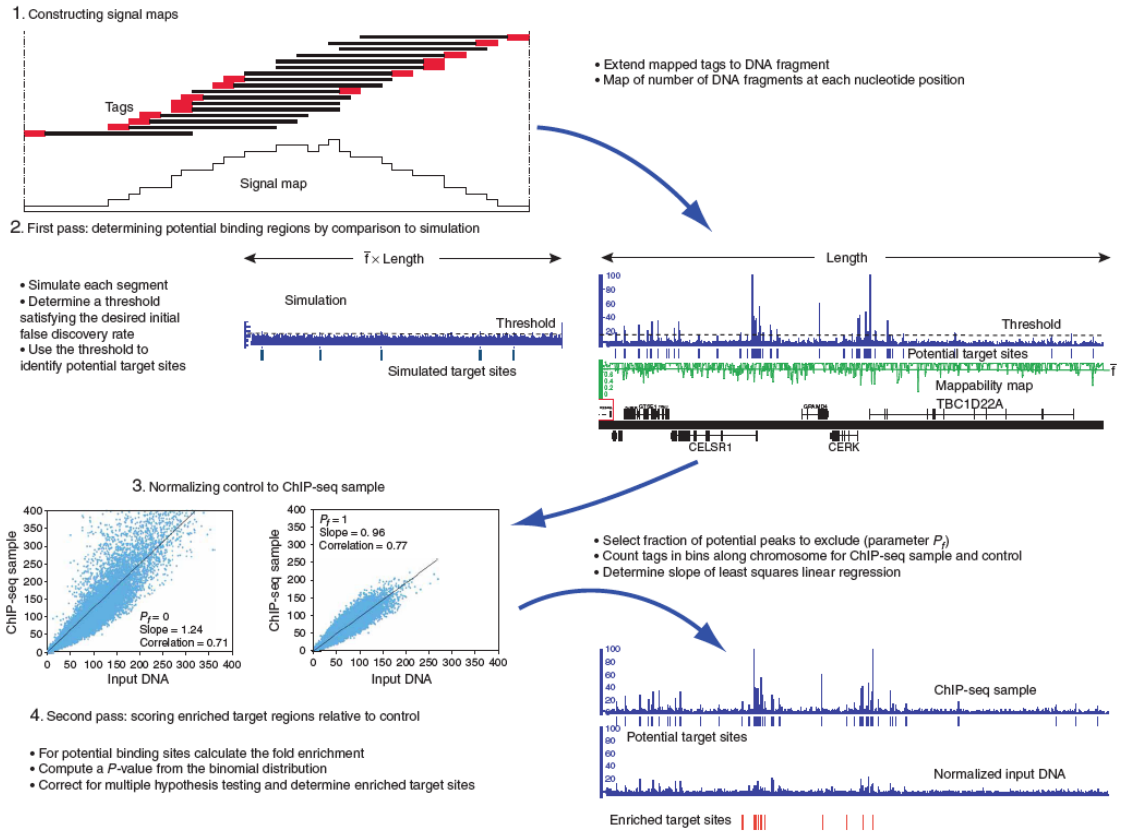

Figure 7. Peak-seq scoring procedure. 1) "Mapped reads are extended to have the average DNA fragment length and then accumulated to form a fragment density signal map." 2) "Potential binding sites are determined in the first pass of the PeakSeq scoring procedure." 3) "After selecting the fraction of potential target sites that should be excluded from the normalization, the scaling factor Pf is determined by linear regression of the ChIP-seq sample against the input-DNA control in 10-KB bins." 4) "Enrichment and significance are computed for putative binding regions." *Abbreviated legend and figure from Rozowsky et al., 2008.*

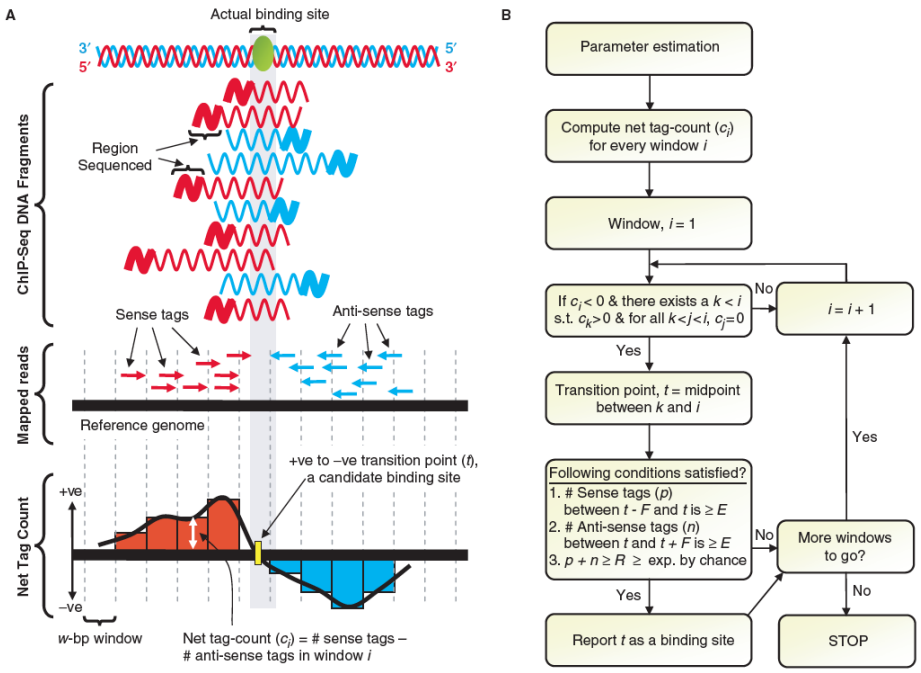

Figure 8. SISSRs algorithm overview. "A) Sequenced short reads (typically 25-50 bp) from ChIP-seq experiments are first mapped onto the reference genome. The mapped reads are then used to estimate statistical parameters, which include the estimation of the average length F of sequenced DNA fragments. B) The entire reference genome along with mapped reads is scanned using overlapping windows of size w base pairs…and the net tag count (ci) for every window i is calculated. Every transition point (t) is a candidate binding site." *Abbreviated legend and figure from Jothi et al., 2008.*

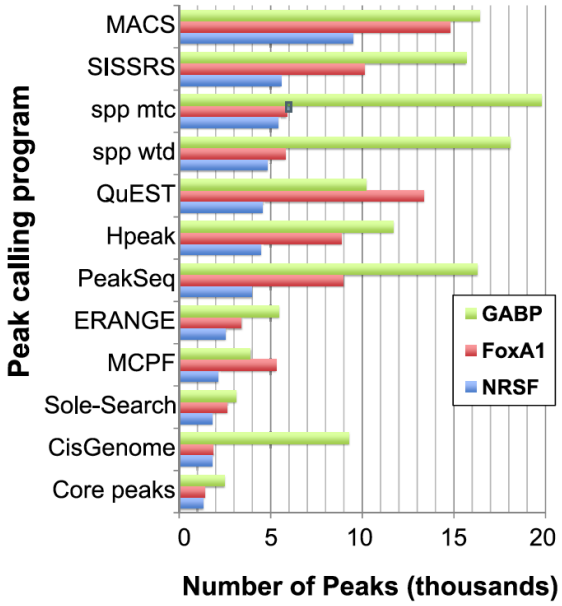

Figure 9. Number of peaks reported from each of eleven different peak calling algorithms run on their default settings. Number of peaks for three different ChIP seq datasets from transcription factors GABP, FoxA1, and NRSF are shown. *Figure from Wilbanks and Facciotti, 2010.*

A

NRSF	CisGenome	Sole-Search	WOLD	ERANGE	PeakSeq	Hpeak	QuEST	wtd	mtc	SISSRS	MACS
CisGenome	X	80	76	64	44	40	36	37	33	31	19
Sole-Search	82	X	81	68	45	40	36	38	34	37	19
MCPF	91	95	X	81	53	48	42	47	41	48	22
ERANGE	91	93	94	X	61	54	47	52	46	49	26
PeakSeq	98	99	100	100	X	85	66	78	69	78	43
Hpeak	98	99	100	100	91	X	69	83	74	80	43
QuEST	91	92	91	89	76	74	X	74	68	76	44
spp wtd	98	99	99	97	87	85	72	X	84	76	45
spp mtc	98	98	99	96	87	86	75	94	X	77	47
SISSRS	97	98	100	99	89	86	75	88	79	X	46
MACS	100	99	100	100	97	94	87	93	88	93	X

Figure 10. Pair-wise comparison of the number of peaks found in each peak calling algorithm (from Figure 9). The pair-wise comparison shown here is for the NRSF transcription factor. "Each panel shows the percentage of total peaks from one method (column) that is sharted with another method (row)." *Portion of legend and figure from Wilbanks and Facciotti, 2010.*

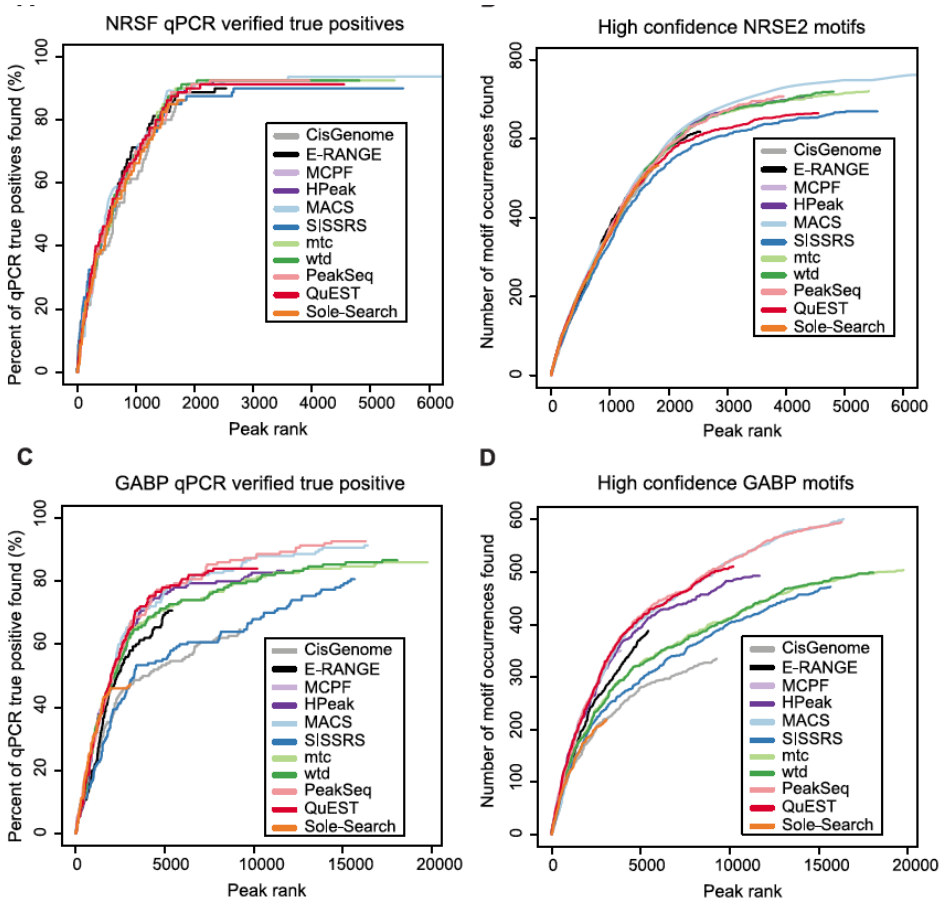

Figure 11. Sensitivities of different peak calling algorithms. "The percentage of qPCR verified positives that were detected by different programs is shown as a function of the increasing number of ranked peaks" examined for the different transcription factor datasets. *Legend and figure from Wilbanks and Facciotti, 2010.*

	Profile	Peak criteria[a]	Tag shift	Control data[b]	Rank by	FDR[c]	User input parameters[d]	Artifact filtering: strand-based/ duplicate[e]	Refs.
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window 2: Number of ChIP reads minus control reads in window	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	Number of reads under peak	1: Negative binomial 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
MACS v1.3.5	Tags shifted then window scan	Local region Poisson *P* value	Estimate from high quality peak pairs	Used for Poisson fit when available	*P* value	1: None 2: $\frac{\#\ \text{control}}{\#\ \text{ChIP}}$	*P*-value threshold, tag length, mfold for shift estimate	No / Yes	13
SiSSRs v1.4	Window scan	$N_+ - N_-$ sign change, $N_+ + N_-$ threshold in region[f]	Average nearest paired tag distance	Used to compute fold-enrichment distribution	*P* value	1: Poisson 2: control distribution	1: FDR 1,2: $N_+ + N_-$ threshold	Yes / Yes	11
PeakSeq	Extended tag aggregation	Local region binomial *P* value	Input tag extension length	Used for significance of sample enrichment with binomial distribution	*q* value	1: Poisson background assumption 2: From binomial for sample plus control	Target FDR	No / No	5

Figure 12. Comparison of CisGenome, MACS, SISSRs, and PeakSeq peak calling algorithms, based on their profile, peak criteria, use of tag shift, control data and significance. *Figure modified from Pepke et al., 2009.*

Literature Cited

Auerbach R.K., *et al*. 2009. Mapping accessible chromatin regions using Sono-Seq. *PNAS.* 106, 14926-31.

Barski, A., *et al.* 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837.

Davey, J., et al., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics.* 12, 499-510.

Feng, J., *et al*., 2011. Using MACS to Identify Peaks from ChIP-Seq Data. *Current Protocols in Bioinformatics*. 34:2.14.1–2.14.14.

Feng, J. *et al.* 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols.* 7, 1728-1740.

Fonseca, N., *et al*. 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics.* bts605.

Illuminia. 2007. Preparing samples for ChIP Sequencing of DNA. 2007. The Braod Insistue and Illumina. http://www.broadinstitute.org/annotation/tbsysbio/Protocols /ChIPSeq_Protocol_2.pdf

Ji, J., *et al.* 2008. An integrated software system for analyzing ChIp-chip and Chip-seq data. *Nature Biotechnology. 26,* 1293-1300.

Ji, J. 2010. Computatiaonl analysis of ChIP-seq data. *Computational biology of transcription factor binding. Methods in Molecular Biology.* 674, 143-159.

Johnson, D. S., *et al.* 2007. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502.

Jothi, R. *et al.* 2008. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research.* 36, 5221-5231.

Kharchenko, P. V., *et al*. 2008. Design and analysis of ChIP–seq experiments for DNA-binding proteins. *Nature Biotech*. **26**,1351–1359.

Ledegerber, C. and Dessimaz, C., 2011. Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics.* 12, 489-497.

Leleu, M. et al., 2010. Processing and analyzing ChIP-seq data: from short reads to regulatory interactions. *Briefings in Functional Genomics.* 9, 466-476.

Li, H. and Homer, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics.* 11, 473-438.

Narlikar, L. and Jothi, R. 2012 .ChIP-seq data analysis: Identification of protein-DNA binding sites with SISSRs peak finder. *Next Generation Micorarray Bioinformatics: Methods and Protocols. Methods in Molecular Biology.* 802, 305-322.

Pepke, S. *et al.,* 2009. Computation for ChIP-seq and RNA-seq studies. *Nature Methods Supplement.* 6, S22-S32.

Robertson, G., *et al.* 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657.

Robertson, G., *et al.* 2008. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res.* 18, 1906-1917.

Rozowsky, J. *et al*. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology.* 27, 66-75.

Schmidt, D., et al. 2009. ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions. *Methods.* 48, 240-248.

Solomon, M. J., *et al.* 1988. Mapping protein–DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell.* **53**,

937–947

Taslim, C., et al., 2012. Analyzing ChIP-seq data: preprocessing, normalization, differential identification, and binding pattern characterization. *Next Generation Micorarray Bioinformatics: Methods and Protocols. Methods in Molecular Biology.* 802: 275-291.

Wilbanks, E. and Facciotti, M. 2010. Evaluation of algorithm performance in ChIP-seq peak detection. *Plos ONE*. 5, e11471.

Zhang, Y. *et al*. 2008. Method Model-based analysis of ChIP-seq (MACS). *Genome Biology.* 9:R1 37.

Zhou, X., *et al*. 2010. The next-generation sequencing technology and application. *Protein Cell.* 6, 520-536.