Briefings in Bioinformatics, 18(3), 2017, 441–450

doi: 10.1093/bib/bbw035
Advance Access Publication Date: 11 May 2016
Paper

Features that define the best ChIP-seq peak calling algorithms

Reuben Thomas, Sean Thomas, Alisha K. Holloway and Katherine S. Pollard

Corresponding author: Katherine S Pollard, Gladstone Institutes, San Francisco, CA 94158, USA. Tel.: 415-734-2711. Fax: 415- 355-0141; E-mail: katherine.pollard@gladstone.ucsf.edu

Abstract

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an important tool for studying gene regulatory proteins, such as transcription factors and histones. Peak calling is one of the first steps in the analysis of these data. Peak calling consists of two sub-problems: identifying candidate peaks and testing candidate peaks for statistical significance. We surveyed 30 methods and identified 12 features of the two sub-problems that distinguish methods from each other. We picked six methods GEM, MACS2, MUSIC, BCP, Threshold-based method (TM) and ZINBA] that span this feature space and used a combination of 300 simulated ChIP-seq data sets, 3 real data sets and mathematical analyses to identify features of methods that allow some to perform better than the others. We prove that methods that explicitly combine the signals from ChIP and input samples are less powerful than methods that do not. Methods that use windows of different sizes are more powerful than the ones that do not. For statistical testing of candidate peaks, methods that use a Poisson test to rank their candidate peaks are more powerful than those that use a Binomial test. BCP and MACS2 have the best operating characteristics on simulated transcription factor binding data. GEM has the highest fraction of the top 500 peaks containing the binding motif of the immunoprecipitated factor, with 50% of its peaks within 10 base pairs of a motif. BCP and MUSIC perform best on histone data. These findings provide guidance and rationale for selecting the best peak caller for a given application.

Key words: ChIP-seq; peak caller; benchmark

Introduction

Regulation of gene expression is one of the fundamental means by which cells adapt to internal and external environments. Many regulatory mechanisms rely on modifying or 'marking' the DNA in particular ways, either through covalent modification or by intermolecular interactions. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) data are generated to provide readouts of these modifications, such as the location and frequency of binding of a transcription factor or the distribution of histone modifications that are used by the cell to establish or maintain specialized chromatin domains.

The data for ChIP-seq peak calling are stacks of aligned reads across a genome. Some of these stacks correspond to the signal of interest (e.g. binding of a transcription factor or modified histone). Many other stacks are regarded as molecular or experimental noise, or as being influenced by a systematically greater accessibility of measurement by the experiment at that particular genomic location. This manuscript deals with the problem of separating signal from noise in the stacks of reads to estimate where the immunoprecipitated protein is bound to the DNA.

Many methods target this problem. The newer methods make claims of superiority over a subset of the existing ones by

Reuben Thomas is a Staff Research Scientist in the Bioinformatics Core at Gladstone Institutes.
Sean Thomas is a Staff Research Scientist in the Bioinformatics Core at Gladstone Institutes.
Alisha K. Holloway is the Director of Bioinformatics at Phylos Biosciences, visiting scientist at Gladstone Institutes and Adjunct Assistant Professor in Biostatistics at the University of California, San Francisco.
Katherine S. Pollard is a Senior Investigator at Gladstone Institutes and Professor of Biostatistics at University of California, San Francisco.
Submitted: 20 January 2016; **Received (in revised form):** 1 March 2016

displaying performance on certain metrics. However, the generalizability of these performance results is unclear given the number of data sets (typically 3–5), methods, parameter settings (often only default settings) and performance metrics used. There have been efforts to benchmark peak calling methods [1–5]. These have the advantage of being independent evaluations, but they have also been limited in scope and sometimes provide conflicting advice owing to differences in numbers and widths of peaks evaluated [6]. For example, Harmanci *et al.* [7], Koohy *et al.* [5] and Micsinai *et al.* [4] disagree about the ranking of Spatial clustering approach for the Identification of ChIP-Enriched Regions (SICER) [8], F-seq [9] and Model-based Analysis for ChIP-Seq (MACS) [10]. Consequently, the field lacks systematic recommendations for calling peaks in different scenarios.

We address these criticisms of benchmarking efforts by first abstracting the peak calling problem into two sub-problems, identifying peaks and testing peaks for significance, respectively. We identify four features of the algorithms that differentiate them in terms of how they address the first sub-problem and eight features that differentiate the algorithms on the second sub-problem. We pick six methods that span most of the possible feature values for the two sub-problems. We then simulate ChIP-seq data for transcription factor binding to generate 100 independent ChIP and input sample pairs at each of three different noise levels to evaluate the operating characteristics of the six methods by varying the respective thresholds for determining peak significance. We also test the methods on data from one transcription factor and two histone mark experiments. Combining results across these data with mathematical analyses, we determine which features of peak calling methods optimize performance.

Methods

Choice of peak calling methods

We surveyed 30 methods [4,7,8, 9–31] in the literature and annotated each of them in terms of how they solve sub-problems 1 and 2 (Table 1 and Supplementary Table S1). We chose Model-based Analysis for ChIP-Seq version 2 (MACS2) [10], MUltiScale enrIchment Calling for ChIP-Seq (MUSIC) [7], Genome wide Event finding and Motif discovery (GEM) [13], Zero-Inflated Negative Binomial Algorithm (ZINBA) [11], Bayesian Change Point (BCP) [14] and TM for further analysis to balance the need to cover most of the feature space of the different methods with that of directly testing a feasible number of the more recent and/or popular methods (Supplementary Text).

Simulations

ChIP-seq data representing transcription factor binding and corresponding input samples are simulated using functions adapted from the ChIPsim [32] Bioconductor [33] package in R [34] that are in turn based on [35]. The data are simulated for pairs of ChIP and corresponding input samples under three noise settings—high noise, medium noise and low noise (details in Supplementary Text).

To ensure that our simulated ChIP and input data resemble real ChIP-seq experiments, we compared them with data from the first 10 million base pairs (bp) of chromosome 1 in a ChIP-seq experiment on the transcription factor Tbx5 in mouse cardiomyocytes [36]. Peaks were identified using MACS2, which is one of the best performing methods on the simulated data. Input was quantified using reads per 1000 bp window. Quantiles of simulated ChIP and input data match the real data well (Figure 1), and browser tracks resemble real data visually (Supplementary Figure S1). The only difference detected is a more extreme number of reads in about 1% of simulated regions, both ChIP and input, compared with the real data.

Evaluation metrics for simulated data

All peak calling methods were run with the lowest significance threshold possible (*P*-value or q-value equal to 1 or for any fold enrichment) to generate the complete list of peaks that could be evaluated for their operating characteristics on each of the simulated data sets. Note, GEM only reports exact genomic locations of binding so a 200 bp window around these identified binding locations was used to define peaks for comparison with other methods. For each method, we varied the *P*-value or q-value threshold to produce a nested set of peaks for evaluating performance across a range of significance levels.

We evaluated performance using several complementary metrics. First, each set of peaks was compared with the location of the binding features in the ChIP sample using the *findOverlappingPeaks* and *annotatePeakInBatch* functions in the ChIPpeakAnno [37] Bioconductor [33] package in R [34]. The fraction of the true binding features that overlaps with the significant peaks is defined as the 'sensitivity', the fraction of the significant peaks that overlap with the true binding features is defined as the 'precision' and the harmonic mean of the sensitivity and precision is defined as the 'F-score' for each method at the particular significance threshold setting on the given simulated data set. We also computed the distance from the center of each significant peak to the center of closest binding feature and used the median of these distances over all significant peaks as a performance metric (median distance-binding).

Table 1. Features of peak calling methods

	GEM	BCP (TF)	BCP (Histone)	MUSIC	MACS2	ZINBA	TM
Locating the potential peaks							
High resolution	Yes	Yes	No	Yes	Yes	No	Yes
ChIP and input sample signals combined	No	No	No	No	No	Yes	Yes
Multiple alternate window sizes	Yes	Yes	Yes	Yes	No	No	No
Use of variability of local signal	Yes	Yes	Yes	No	Yes	Yes	No
Ranking of peaks							
Binomial test	Yes	No	No	Yes	No	No	No
Poisson test	No	Yes	No	No	Yes	No	No
Normalized difference score	No	No	No	No	No	No	Yes
Use of underlying genome sequence	Yes	No	No	No	No	No	No
Posterior probability of enrichment	No	No	Yes	No	No	Yes	No

Note. See Supplementary Table S1 for a more complete list of features and peak calling methods.

We additionally computed the converse metric, the median distance from each of the binding features to the center of the closest significant peak (median distance-peak).

The different methods have different typical peak widths, and also the estimated *P*-values are not always comparable because they result from tests of different hypotheses. Therefore, we chose two approaches to look at the variation of these five metrics as a function of a common measure of a false-positive rate across the six methods. First, we used the sets of peaks for each method to compare performance at significance levels that produce the same genome coverage ($\log_{10}$ fraction of genome). Second, we limited peak lengths to a 200 bp window around either the peak summit or the center of the peak (when the information on summit position was not available) for all the methods and compared performance at significance levels that produce the same number of 200 bp peaks. Results were qualitatively the same with both approaches, so we focus on performance as a function of $\log_{10}$ number of peaks. Note that all performance evaluations are based on characteristics of the data we used and assumptions about the ground truth, which we discuss further in the Supplementary Text.

Evaluation with Tbx5 ChIP-seq data

We again used the ChIP-seq data sample that measured the binding of the Tbx5 transcription factor to mouse cardiomyocytes [36]. We used the two binding motifs [38] given in Supplementary Figure S2 to represent the *in vitro* sequence binding for Tbx5 in mouse cardiomyocytes. We used the *matchPWM* function in the Bioconductor Biostrings [39] package to identify all the potential binding locations of Tbx5 for each of the two motifs. The threshold likelihood ratio score for defining the potential binding location was fixed at 95% of the maximum possible likelihood ratio value, assuming a zero-order Markov model for a given sequence with prior probabilities for the nucleotides given by their frequencies in the 2 kb regions of all mouse gene promoters. The shortest genomic distance of each significant peak identified by each peak calling method to binding the two motifs was used as a measure of accuracy. We also compared methods based on the fraction of the top *n* peaks (ordered by statistical significance or by fold enrichment for the thresholding method) that are within 100 bp of either motif.

Evaluation with H3K36me3 data

H3K36me3 domains are broad and are known to mark regions of the genome that are being actively transcribed [40]. Encyclopedia of DNA Elements (ENCODE) [41] data for the H3K36me3 histone mark in the GM12878 cell line were used. These files were aligned to the human genome hg19 using Bowtie2 [42], and gene counts were obtained using Htseq [43]. We also obtained ENCODE RNA-sequencing data for GM12878 and used edgeR [42] to convert counts to reads per kilobase exon model per million mapped reads (RPKM) and then estimated gene expression by the average of the RPKM values over the two replicates. We considered a peak as positive if it overlaps an active gene (defined varying RPKM from 0 to 2) and compared methods based on sensitivity, precision and F-score.

Evaluation with H3K4me3 data

H3K4me3 domains mark active and poised promoters [44]. ENCODE [41] data for the H3K4me3 histone mark in the GM12878 cell line were used. We considered a peak as positive if it overlaps the promoter of an expressed gene (RPKM > 0.5). The top 15 000 peak calls from the different methods are ranked by their significance or by their fold enrichment for the thresholding method. We plotted the correct peak fraction (fraction of the top 1000xn peaks that overlap with active promoters) detected as a function of the correct promoter fraction (fraction of the active promoters that overlap with the top 1000xn peaks).

Binomial test versus the Poisson test

One of the problems that most peak callers need to address is to assign significance to a potential peak region. The significance is based on the rejection of the null hypothesis that the proportion of DNA from a given genomic region in the ChIP sample is less than equal to that in the input sample. This is typically tested by either a Poisson or a Binomial test on the number of reads that map to this genomic location in the ChIP and input samples. We compared the operating characteristics of these two tests using a simulation procedure detailed in the Supplementary Text.

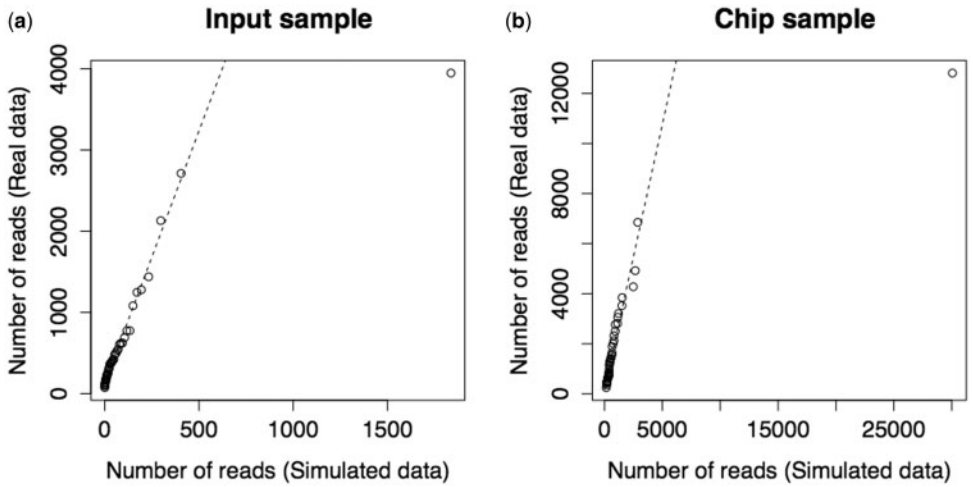

Figure 1. Quantile–quantile plots comparing the distributions of reads in the input and ChIP sample in one of the simulated data sets with those in a real Tbx5 ChIP sample from mouse cardiomyocytes. The dotted lines represent the linear fits to the data excluding the one extreme point in both (**A**) and (**B**). The pearson correlation of the scatter of points modeled by the dotted lines is 0.99 in (**A**) and 0.97 in (**B**).

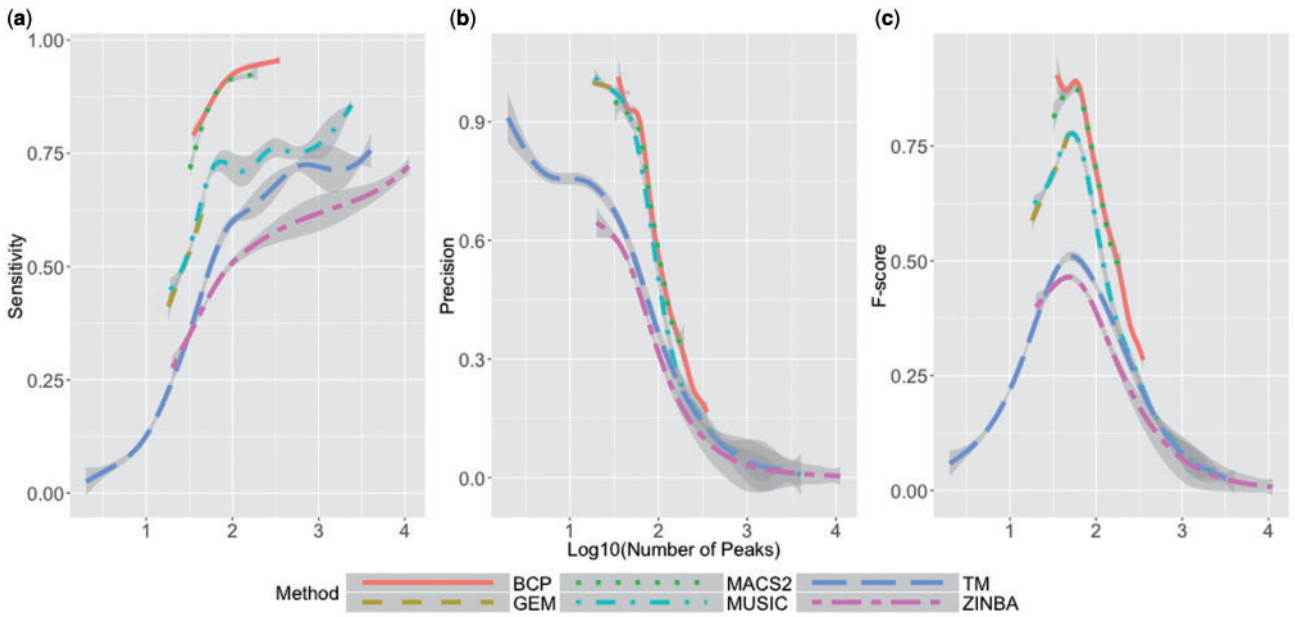

Figure 2. Sensitivity (a), Precision (b) and F-score (c) as a function of the $\log_{10}$ of the number of called peaks for the six peak calling methods on 100 simulated transcription factor ChIP-seq data sets under the medium noise setting. For each method, the means and 95% confidence intervals (dark gray regions around the mean profiles) of the means of a metric are estimated using Generalized Additive Models [51] of variation across the 100 simulated data sets, as a function of a smooth function of $\log_{10}$ of the number of called peaks. Note the overlapping performances of GEM and MUSIC and BCP and MACS2 in (a), (b) and (c) with $\log_{10}$ number of peaks, respectively, between 1.2 and 1.6 and 1.6 and 2.0.

Results

Benchmarking peak calling methods

We benchmarked six peak calling methods representing different features of the approaches to identify candidate peaks and evaluate their statistical significance: GEM, MACS2, MUSIC, BCP, TM and ZINBA. These evaluations used 300 simulated and 3 real ChIP-seq data sets. Performance was compared across a range of significance values representing different number of called peaks.

Simulated transcription factor binding data

We simulated transcription factor ChIP-seq data with three different noise levels in a manner that closely resembles real data (Methods). Simulations have the advantage of allowing us to flexibly explore a range of different scenarios in a situation where the ground truth is known.

BCP and MACS2 perform best by sensitivity, precision and F-score metrics across the low (Supplementary Figure S3), medium (Figure 2) and high (Supplementary Figure S4) noise levels. TM and ZINBA perform worst, and MUSIC is intermediate. Across methods, except for with ZINBA, median distance of the called peaks to the true peaks and of the true peaks to the called peaks is typically within 100 bp regardless of significance threshold across the low (Supplementary Figure S5), medium (Figure 3) and high (Supplementary Figure S6) noise levels. Reduction in simulated noise has the expected effect of improving the sensitivity of BCP, MACS2, GEM and MUSIC, but not of TM and ZINBA (Supplementary Figure S3). The median distance metrics are predictably larger in the high noise settings (Supplementary Figure S6).

Transcription factor Tbx5 binding data

We next evaluated performance of the six methods on data from a Tbx5 ChIP-seq experiment to assess whether trends are similar to those revealed by our simulations. Figure 4 and Supplementary Figure S7 show the fraction of the top n peaks that are within 100 bp of a Tbx5 motif. BCP and GEM do particularly well relative to the other methods for Tbx5 Motif 1 (Figure 4). Figure 5 and Supplementary Figure S8 displays the empirical distribution of the shortest distance of the called peaks to each of the Tbx5 motifs for each method. GEM stands out among all the methods in terms of the fraction of its peaks being closer to a Tbx5 motif than any of the other methods. GEM has the highest fraction of the top 500 peaks with either of two binding motifs of Tbx5, and 50% of its peaks are within 100 bp of a motif (Motif 1; Figure 4). Only 10% of the called peaks of the other methods are within 100 bp of the same Tbx5 motif (Figure 4).

Histone H3K36me3 and H3K4me3 data

Because histones typically have wider peaks than transcription factors and lack DNA motifs, methods perform differently on them. To assess performance on histone ChIP-seq data, we used two sets of experiments from the ENCODE Project [41]. Figure 6 shows the performance of the methods on H3K36me3 data, in terms of how well peaks overlap genes that are actively transcribed (see Methods). MUSIC and BCP perform better than other methods in terms of sensitivity and F-score at a relatively small price in terms of precision. Figure 7 displays the performance of four of the methods on H3K4me3 data, assessed in terms of overlap of peaks with promoters of expressed genes (see Methods). ZINBA failed to run on this data set, giving an error that has failed to be resolved with the authors of the software. The other methods perform comparably on this data set, with MUSIC and BCP again being slightly better than the other methods.

Features of peak calling methods that influence performance

We next investigated which features of peak calling methods drive the differences in their performance. To do so, we identified the features or analysis choices that differentiate the six methods we benchmarked, along with 24 other methods from the literature. Then, we evaluated whether these features drive performance.

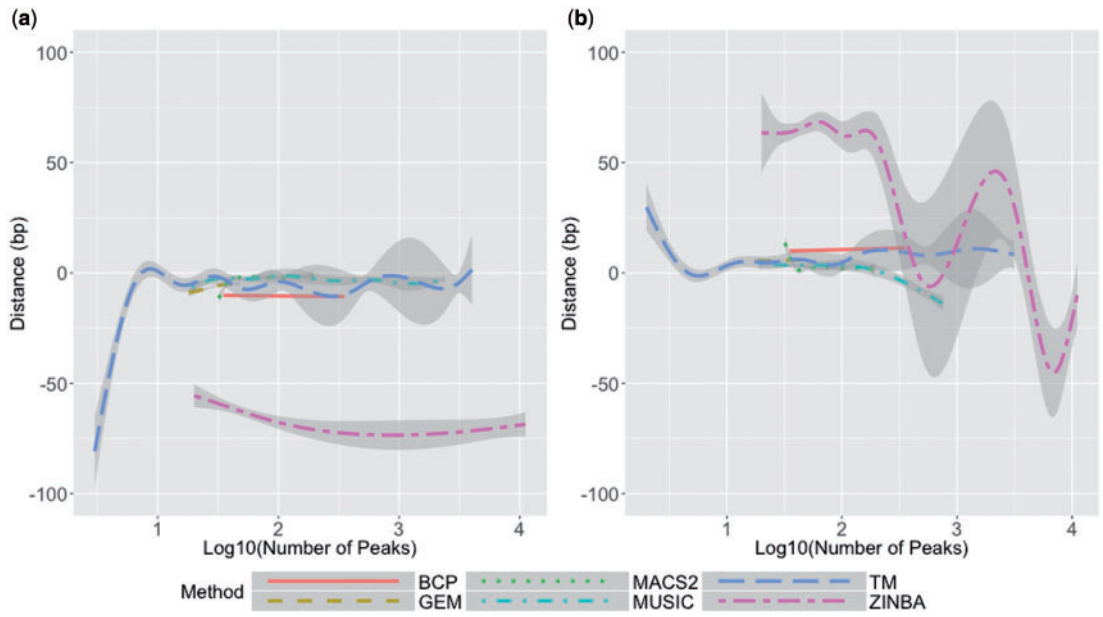

Figure 3. *Median distance-binding* (a) and *Median distance-peak* (b) as a function of the $\log_{10}$ of the number of called peaks for the six peak calling methods on 100 simulated data sets under the medium noise setting. The means and 95% confidence intervals are estimated in a similar manner as is done for Figure 2. Note the overlapping performances of MACS2 and MUSIC in (a), GEM, MUSIC and TM in (b) with $\log_{10}$ number of peaks, respectively, between 1.6 and 2.0 and 1.2 and 2.0.

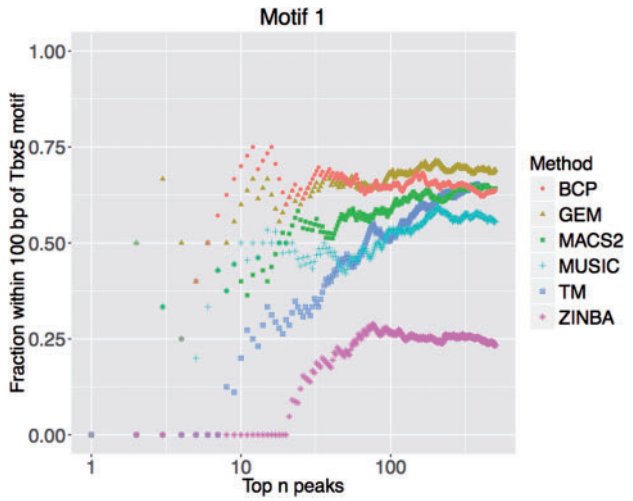

Figure 4. Fraction of top *n* peaks within 100 bp of a Tbx5 motif for the six methods is given. Results are based on Tbx5 Motif 1.

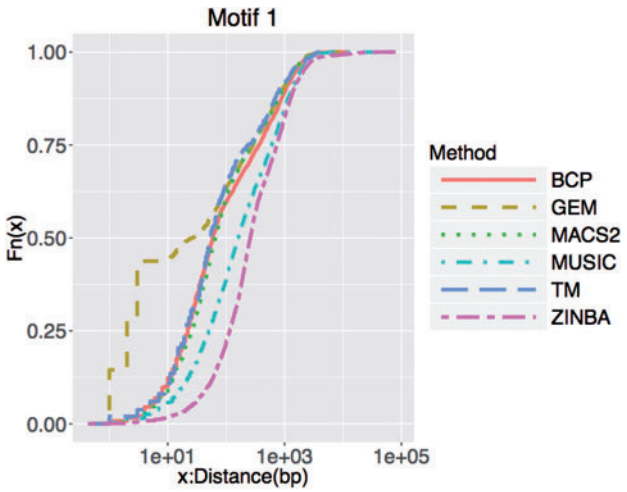

Figure 5. Empirical distribution of the shortest distance to the Tbx5 motif of the significant peaks called by the six methods. Results based on Tbx5 Motif 1. Note the overlapping performances of BCP, MACS2 and TM over the entire range and of GEM, BCP, MACS2 and TM with distance between 100 and 10 000 bp.

Sub-problem 1: detection of candidate peaks

Four features distinguish methods in terms of how they identify candidate peak regions. Most methods (19 of 30, Supplementary Table S1) use the signal in a 'high-resolution' manner, i.e. peaks could be centered on any nucleotide in the genome as opposed to a bin or a region with more than one nucleotide. There are low-resolution methods (11 of 30, Supplementary Table S1) that essentially divide the genome into bins and do not allow signals from one bin to explicitly affect those in the other bins. 'ChIP and input sample signal combined' is the second feature that separates methods. Candidate peak regions are defined using the signal that is an explicit combination of the ChIP and input signals either as a fold-change or as a difference score. Methods are deemed not to have this feature (21 of 30, Supplementary Table S1) either if they do not use the input signal at all or in cases where the input signal is only used to determine the background rate of reads in the region but is not used to modify the ChIP signal in any way. 'Multiple alternate window sizes' (9 of 30, Supplementary Table S1) representing widths of influence of each nucleotide for querying the signal could be explicitly or implicitly used. The 'variability of the local signal' (18 of 30, Supplementary Table S1) is a feature in which the signal is modeled as being generated from a distribution (whose parameters are estimated either using the ChIP or the input signal). Candidate peaks are defined using more than one moment of this distribution and not just the mean.

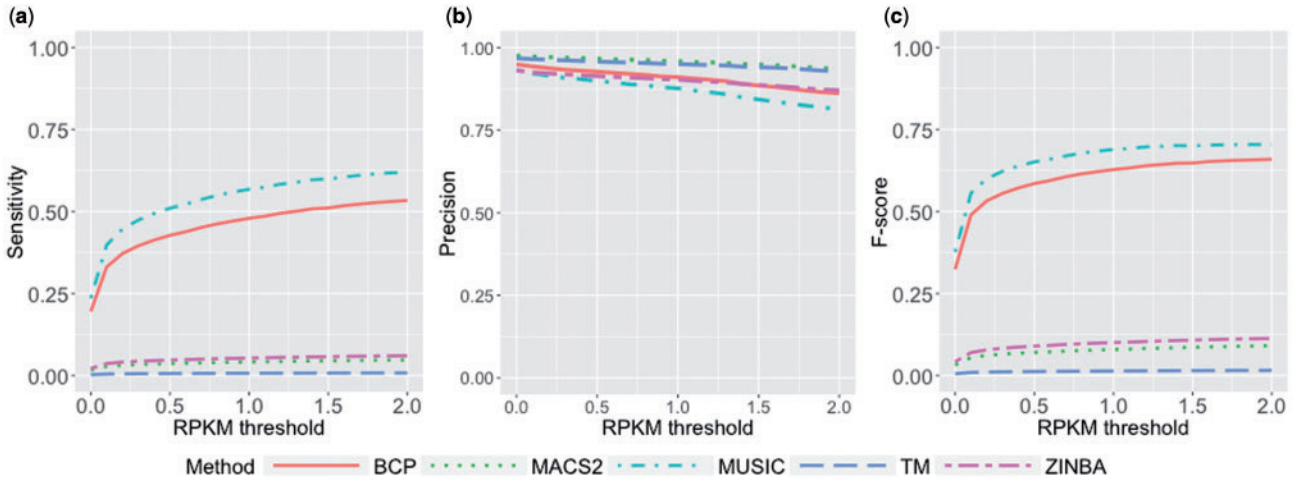

Figure 6. Sensitivity (**A**), precision (**B**) and F-score (**C**) of the overlap of the called significant peak regions with active gene bodies for H3K36me3 data. The threshold for defining active genes was varied from 0 to 2 RPKM. Note the overlapping performances of MACS2 and ZINBA in (**A**), MACS2 and TM in (**B**) and BCP and ZINBA in (**B**).

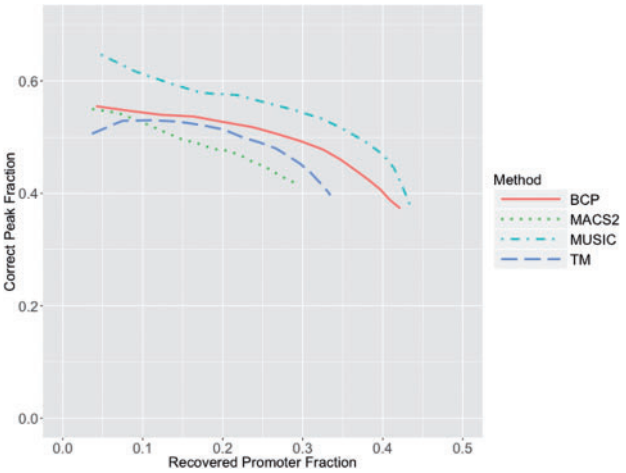

Figure 7. Correct peak fraction (fraction of top called 1000*n* peaks that overlap with the promoters of active genes, genes with expression > 0.5 RPKM) as a function of recovered promoter fraction (fraction of promoters of active genes that overlap with the top called 1000n peaks) for H3K4me3 data is given. The peaks for each method were ranked by their assigned significance.

Peak detection is reduced when ChIP and input signals are explicitly combined

We mathematically analyzed the probability of detecting a peak when combining ChIP and input data versus not. This analysis was motived by the behavior of TM and ZINBA relative to the other four methods, none of which explicitly combine ChIP and input (Table 1) at the stage of identifying candidate peaks.

Consider a region of the genome with a 'true' peak or binding event. Assume that the number of reads from the ChIP sample in this region, denoted by X, comes from a Poisson distribution with parameter λ_1. The number of reads from the input control sample in this region, denoted by Y, comes from a Poisson distribution with parameter λ_0. The reads in the ChIP and input sample are independent of each other and have been normalized for potential differences in sequencing depths. We are interested in identifying this 'true' peak. We will compare a measure of the error probability (of the inability to identify this peak) under two scenarios. In the first scenario, only the ChIP signal is used, whereas in the second one, the ChIP and input signal are explicitly combined as a difference. Let H_0 and H_1 denote the null and alternate hypothesis under the two scenarios. (Note: the difference of two Poisson random variables follows a Skellam distribution. The mean and the variance of a Poisson random variable with parameter λ are both equal to λ. The mean and the variance of a Skellam distribution with parameters λ_1 and λ_0 are $\lambda_1-\lambda_0$ and $\lambda_1+\lambda_0$, respectively.)

1. Compare $H_0 : X \sim Pois(\lambda_0)$ versus $H_1 : X \sim Pois(\lambda_1)$
2. Compare $H_0 : X - Y \sim Skellam(\lambda_0, \lambda_0)$ versus $H_0 : X - Y \sim Skellam(\lambda_0, \lambda_1)$

Let $P_e(i, \pi)$ denote the error probability of the inability to distinguish the distributions under H_0 and H_1, with prior probabilities $\pi = (\pi_0, \pi_1)$under scenario $i = 1, 2$.

If we can show $P_e(1, \pi) < P_e(2, \pi)$, then using Scenario 1 is preferable to using Scenario 2 to identify the peak. We will use the distance between the underlying distributions of H_0 and H_1 as an approximation of this error probability. The intuition is that the farther apart the distributions are the less likely it is to make an error in being unable to distinguish the two distributions. Following Kailath [46], we will use Bhattacharyya distance [47], B, as a measure of distance between two probability distributions. Let $B(i)$ denote the distance between the probability distributions associated with H_0 and H_1 under scenario $i = 1, 2$. Kailath [46] showed that if $B(1) > B(2)$ then there exists prior probabilities πsuch that $P_e(1, \pi) < P_e(2, \pi)$.

The mean number of reads in the binding regions is typically >20 in a typical experiment [35]. Using this observation and the Central Limit Theorem [48], the normal distribution reasonably approximates both Poisson (Supplementary Figure S9) and Skellam (Supplementary Figure S10) distributions. The Bhattacharyya distance between two normal distributions N_1 and N_2 with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) is given by the following equation:

$$B(N_1, N_2) = \frac{1}{4} ln\left(\frac{1}{4}\left(\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right) \quad (1)$$

Theorem 1: If the Poisson and the Skellam probability distributions can be replaced by Normal distributions with the

corresponding means and variances of the respective distributions, then there exists prior probabilities π such that $P_e(1,\pi) < P_e(2,\pi)$.

Proof: Using Equation (1) and the mean and variances for the Poisson and Skellam distributions,

$$B(1) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\lambda_1}{\lambda_0}+\frac{\lambda_0}{\lambda_1}+2\right)\right)+\frac{1}{4}\left(\frac{(\lambda_1-\lambda_0)^2}{\lambda_1+\lambda_0}\right)$$

$$B(2) = \frac{1}{4}\ln\left(\frac{1}{4}\left(\frac{\lambda_1+\lambda_0}{2\lambda_0}+\frac{2\lambda_0}{\lambda_1+\lambda_0}+2\right)\right)+\frac{1}{4}\left(\frac{(\lambda_1-\lambda_0)^2}{\lambda_1+3\lambda_0}\right)$$

Proving $B(1) > B(2)$

$$\iff \left(\frac{\lambda_1}{\lambda_0}+\frac{\lambda_0}{\lambda_1}+2\right) > \left(\frac{\lambda_1+\lambda_0}{2\lambda_0}+\frac{2\lambda_0}{\lambda_1+\lambda_0}+2\right)$$

$$\iff \left(\frac{\lambda_1-\lambda_0}{2\lambda_0}+\frac{\lambda_0(\lambda_1-\lambda_0)}{\lambda_1(\lambda_1+\lambda_0)}\right) > 0$$

$$\iff \left(\frac{(\lambda_1-\lambda_0)^2(\lambda_1+2\lambda_0)}{2\lambda_0\lambda_1(\lambda_1+\lambda_0)}\right) > 0$$

which is true because by definition the Poisson parameters λ_1 and λ_0 are positive. Using the result in Kailath [46], there exists prior probabilities πsuch that $P_e(1,\pi) < P_e(2,\pi)$.

Therefore, procedures like TM and ZINBA that explicitly combine the ChIP and input signals are expected to be less powerful at identifying true binding than ones that do not.

Using input signal to filter candidate peaks

Input is also used to filter regions. GEM implements a filter that removes candidate peak regions with 3-fold or fewer reads in ChIP relative to input, whereas BCP and MACS2 do not implement a similar filter. We observed that the sensitivity of GEM does not improve beyond a certain level irrespective of how much the significance threshold increases (Figure 2, Supplementary Figure S3 and S4). We hypothesize that filtering may be responsible for this leveling off of performance.

Window size

Methods use different window sizes to scan the genome for candidate peaks. In our benchmark, TM used a 75 bp sliding window, whereas MACS2 and MUSIC used 150 bp windows. To check whether this difference drives the relative performance of these methods, we implemented MACS2 using window sizes of 75 bp and 200 bp. The performance of MACS2 with 75 bp windows is worse than with 150 bp and 200 bp windows (Supplementary Figure S11), suggesting that longer windows are preferable at least for the scenarios we simulated. This raises the question of whether the optimal window size is different for narrow versus broad peaks. To explore this, we mathematically analyzed variation in the likelihood of detecting a peak of length l using a window size w to scan the genome.

Let γ_1 be the Poisson rate parameter for the number of reads per base in the peak corresponding to a binding event. So the number of reads in a w-bp wide interval in the peak is distributed as a Poisson random variable with parameter $\gamma_1 w$. Also, let γ_0 be the Poisson rate parameter for the number of reads per base in the input sample. Let X represent the number of reads arising from a window of width w bp.

Consider two situations, $w \le l$ and $w > l$.
For the situation $w \le l$,

$$\text{Compare } H_0 : X \sim Pois(\gamma_0 w) \text{ versus } H_1 : X \sim Pois(\gamma_1 w) \quad (2)$$

For the situation $w > l$,

$$\text{Compare } H_0 : X \sim Pois(\gamma_0 w) \text{versus } H_1 : X \sim Pois(\gamma_1 l + (w-l)\gamma_0) \quad (3)$$

We will again use the Bhattacharyya distance as a measure of the ability to distinguish between the two distributions in Equations (2) and (3). The Bhattacharyya distance between two Poisson distributions of rate parameters, (λ_1, λ_2), is given by the following equation:

$$B(\lambda_1,\lambda_2) = \frac{1}{2}\left(\sqrt{\lambda_1}-\sqrt{\lambda_2}\right)^2 \quad (4)$$

Let $B(w)$denote this distance measure as a function of the size w of the window used to scan the genome. Then using the Poisson rate parameters in Equations (2) and (3) and the formula for distance in Equation (4),

$$\begin{aligned} B(w) &= \frac{w}{2}\left(\sqrt{\gamma_1}-\sqrt{\gamma_0}\right)^2 && w \le l \\ &= \frac{w}{2}\left(\sqrt{\gamma_1\left(\frac{l}{w}\right)+\gamma_0\left(1-\frac{l}{w}\right)}-\sqrt{\gamma_0}\right)^2 && w > l \end{aligned} \quad (5)$$

Therefore, the distance between the probability distributions associated with the null and alternate hypothesis increases linearly with w until it reaches the actual peak length and thereafter decreases asymptotically to zero. Denote the error probability of the inability to distinguish the two signals in the peak and background as $P_e(w,\pi)$ that is a function of w and prior probabilities π for the two hypotheses. Using the result from [46], there exists prior probabilities πsuch that the value of $P_e(w,\pi)$ decreases to a minimum when $w = l$ after starting from a window of size one base and then again increases asymptotically to 1 with increasing window width (Supplementary Figure S12).

This result that the optimal window size for scanning the genome is the true peak width suggests an explanation for the performance results we observed with the histone data (Figures 6 and 7). Harmanci *et al.* [7] present an estimate of the spectrum of peak lengths associated with a transcription factor and different histone marks including H3K36me3 and H3K4me3. A characteristic of most of these spectra is the presence of peaks that vary in lengths across orders of magnitudes. Therefore, methods that work only with one window size are biased to pick peaks of length only of comparable magnitude.

Incorporating variability of the local signal

The difference in the operating characteristics of MUSIC as compared with MACS2 and BCP can potentially be explained by the fact that in identifying their candidate peak regions MACS2 and BCP uses the variability of the local signal whereas MUSIC does not. The peak regions are identified by local minima of the smoothed signals by MUSIC, whereas MACS2 checks whether the number of reads in a candidate region is different from the expected assuming a given background rate using a Poisson test, and BCP implicitly takes into account the variability of the local signal by its estimation of the parameters associated with its hidden Markov model.

Sub-problem 2: Statistical significance of candidate peaks

Once the candidate peaks have been identified, the different methods typically rank candidates by their significance of a hypothesis test that compares the counts in the corresponding genomic regions of the ChIP and input samples. This test has mostly been implemented as a 'Poisson' (9 of 30, Supplementary Table S1) test, a 'Binomial' (7 of 30, Supplementary Table S1) test or by 'fold-change or Normalized Difference' (5 of 30, Supplementary Table S1). There are methods that rank the peaks by some 'posterior' measure (6 of 30, Supplementary Table S1) that could be the posterior probability of binding at a given genomic region or the posterior rate of counts in a given genomic region. Additionally, there are methods that explicitly use the underlying genome sequence (1 of 30, Supplementary Table S1) or the shape of the candidate peaks (2 of 30, Supplementary Table S1) in assigning significance values.

Binomial versus Poisson Test

BCP and MACS2 have the best operating characteristics on simulated data (Figures 2, Supplementary Figures S3 and S4). Could this be because of how these methods test candidate peaks for statistical significance? BCP and MACS2 use the Poisson test, whereas MUSIC and GEM use the Binomial test. To explore this question, we first used simulations (Methods) to directly compare Poisson and Binomial tests on a predefined set of candidate peaks. These empirical results show that the Poisson test is more powerful at detecting enriched regions, while maintaining a reasonable Type I error rate (Figure 8). Second, we implemented modified versions of BCP and MUSIC that expand the methods to allow either a Binomial or Poisson test. Our Poisson test version of BCP and Binomial test version of MUSIC are essentially the same as in the original methods. In this approach, we are using the read counts in the regions called peaks by each method and performing the statistical test in two ways. The operating characteristics of both versions of each method are hardly distinguishable (Supplementary Figures S13 and S14). However, in real situations, one works with a set of peaks identified by a chosen threshold. Analyzed this way, rather than over all possible thresholds, the Poisson version of each method is clearly better (Supplementary Figures S15 and S16).

Discussion

We performed a benchmarking study and systematic evaluation of the features of ChIP-seq peak calling methods that drive their relative performance.

Our benchmarking analysis included six methods that are representative of the different features of ChIP-seq software tools. Overall, BCP and MACS2 have the best operating characteristics on simulated transcription factor binding data. On real Tbx5 ChIP-seq data, GEM stood out in terms of how close its

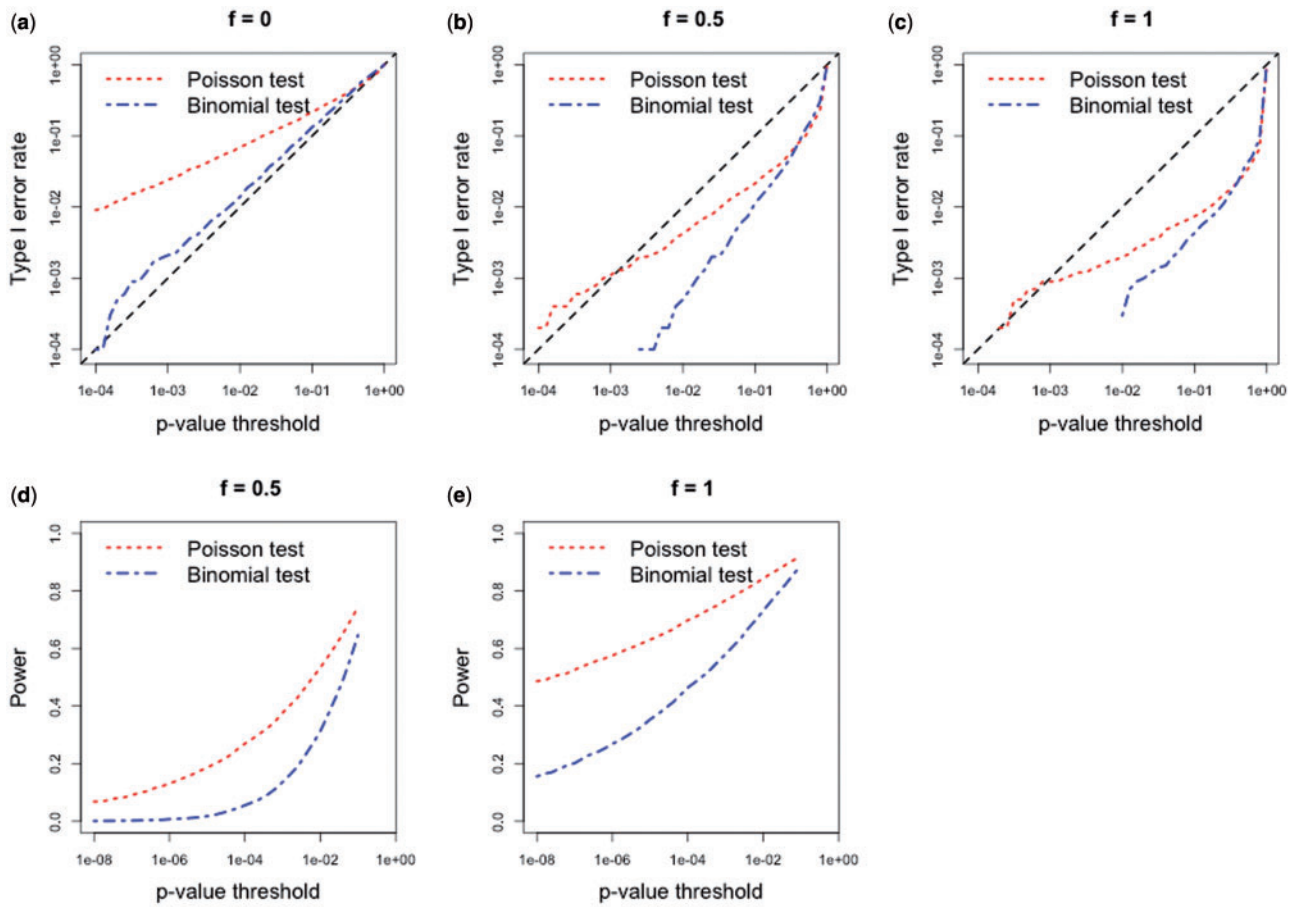

Figure 8. Type I error rate and statistical power comparison between Poisson and Binomial tests is given. *f* is a parameter that controls the increase in the proportion of DNA from a given region in the input relative to the ChIP sample for the Type I error evaluations, (**A**), (**B**) and (**C**), and increase in this proportion for the ChIP relative to the input sample for the power evaluations ((**D**) and (**E**)). The dashed lines in (**A**), (**B**) and (**C**) are the $y = x$ lines.

peaks are to the primary binding motif of Tbx5. BCP and MUSIC perform best on histone ChIP-seq data.

We identified three features that are the most important drivers of performance in peak calling. First, methods that explicitly combine ChIP and input signal in trying to identify candidate peak regions are less powerful than those that only use the signal from the ChIP sample. This does not imply the input sample need not be used in the second sub-problem to rank candidate peak regions. The use of input in any ChIP-seq experiment is essential in filtering out false-positive signals [49] and should be used (when available) at the stage of ranking candidate peaks. Second, methods that use windows of multiple widths to scan the genome for candidate peaks involving histone marks perform better than the others. The use of multiple window sizes can be implemented either explicitly (e.g. MUSIC) or implicitly (e.g. BCP via a hidden Markov model). Finally, the Poisson test is more powerful than the Binomial test for statistically scoring candidate peaks.

A few other features of peak calling methods merit consideration. Our results suggest that methods using variability of the local signal in identifying candidate peak regions are likely to have better operating characteristics than ones that do not. Second, the ability of GEM to call peaks close to motifs of the immunoprecipitated transcription factor points to the benefit of incorporating the underlying genome sequence and knowledge of binding sites at the stage of ranking candidate peaks. There are other methods of ranking candidate peak regions based on their shape characteristics [21, 23] that have not been evaluated in this manuscript, which may also provide performance benefits.

This manuscript focuses on how peak calling methods differ in terms of how they identify candidate peaks and compute their statistical significance. All peak calling methods sequentially implement solutions to these two problems, and then most use a significance threshold to determine a set of peak calls [12]. We note that selecting a threshold is reasonably straight forward; most methods use a false discovery rate-based multiple testing correction and a user-defined proportion of false discoveries that will be tolerated in a given application. We therefore did not compare methods based on a single choice of significance threshold, but instead examined performance across a range of thresholds.

Key Points

- Peak calling using Chip-seq data consists of two sub-problems: identifying candidate peaks and testing candidate peaks for statistical significance.
- Twelve features of the two sub-problems of peak calling methods are identified.
- Methods that explicitly combine the signals from ChIP and input samples to define candidate peaks are less powerful than methods that do not.
- Methods that use windows of different sizes to scan the genome for potential peaks are more powerful than ones that do not.
- Methods that use a Poisson test to rank their candidate peaks are more powerful than those that use a Binomial test.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by the NHLBI Bench-to-Bassinet program (grant #HL098179), NHLBI grant #HL089707 and BioFulcrum: A Gladstone Institutes Enterprise.

References

1. Laajala TD, Raghav S, Tuomela S, *et al*. A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* 2009;**10**:618.
2. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* 2010; **5**: e11471.
3. Rye MB, Saetrom P, Drablos F. A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res* 2011;**39**:e25.
4. Micsinai M, Parisi F, Strino F, *et al*. Picking ChIP-seq peak detectors for analyzing chromatin modification experiments. *Nucleic Acids Res* 2012;**40**:e70.
5. Koohy H, Down TA, Spivakov M, *et al*. A comparison of peak callers used for dnase-seq data. *PLoS ONE* 2014;**9**:e96303–11.
6. Szalkowski AM, Schmid CD. Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts. *BriefBioinform* 2011;**12**:626–33.
7. Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol* 2014;**15**:474
8. Zang C, Schones DE, Zeng C, *et al*. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 2009;**25**:1952–8.
9. Boyle AP, Guinney J, Crawford GE, *et al*. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* 2008;**24**:2537–8.
10. Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137r137.
11. Rashid NU, Giresi PG, Ibrahim JG, *et al*. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol* 2011;**12**:R67.
12. Xu H, Handoko L, Wei X, *et al*. A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* 2010;**26**:1199–204.
13. Guo Y, Mahony S, Gifford DK. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* 2012;**8**:e1002638.
14. Xing H, Mo Y, Liao W, *et al*. Genome-wide localization of protein-dna binding and histone modification by a Bayesian change-point method with Chip-Seq data. *PLoS Comput Biol* 2012;**8**:e1002613.
15. Xu H, Wei CL, Lin F, *et al*. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 2008;**24**:2344–9.
16. Rozowsky J, Euskirchen G, Auerbach RK, *et al*. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* 2009;**27**:66–75.
17. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
18. Jothi R, Cuddapah S, Barski A, *et al*. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 2008;**36**:5221–31.

19. Valouev A, Johnson DS, Sundquist A, *et al*. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* 2008;**5**:829–34.
20. Song Q, Smith AD. Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 2011;**27**:870–1.
21. Hower V, Evans SN, Pachter L. Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics* 2011;**12**:15.
22. Lan X, Bonneville R, Apostolos J, *et al*. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics* 2011;**27**:428–30.
23. Wu H, Ji H. PolyaPeak: detecting transcription factor binding sites from chip-seq using peak shape information. *PLoS ONE* 2014;**9**:e89694.
24. John S, Sabo PJ, Thurman RE, *et al*. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet* 2011;**43**:264–8.
25. Qin ZS, Yu J, Shen J, *et al*. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics* 2010;**11**:369–13.
26. Spyrou C, Stark R, Lynch AG, *et al*. BayesPeak: Bayesian analysis of ChIP-seq data. *BMC Bioinformatics* 2009;**10**:299–17.
27. Kharchenko PV, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 2008;**26**:1351–9.
28. Ji H, Jiang H, Ma W, *et al*. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* 2008;**26**:1293–300.
29. Albert I, Wachi S, Jiang C, *et al*. GeneTrack–a genomic data processing and visualization framework. *Bioinformatics* 2008;**24**:1305–6.
30. Blahnik KR, Dou L, O'Geen H, *et al*. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res* 2010;**38**:e13.
31. Fejes AP, Robertson G, Bilenky M, *et al*. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 2008;**24**:1729–30.
32. Humburg P. ChIPsim: Simulation of ChIP-seq experiments. 2011; R package version 1.22.0.
33. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.
34. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, 2015. http://www.R-project.org/.
35. Zhang ZD, Rozowsky J, Snyder M, *et al*. Modeling ChIP sequencing in silico with applications. *PLoS Comput Biol* 2008;**4**:e1000158.
36. Luna-Zurita L, Stirnimann CU, Glatt S, *et al*. Complex interdependence regulates heterotypic transcription factor distribution and coordinates cardiogenesis. *Cell* 2016;**164**:999–1014.
37. Zhu LJ, Gazin C, Lawson ND, *et al*. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 2010;**11**:237.
38. He A, Kong SW, Ma Q, *et al*. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *PNAS* 2011;**108**:5632–7.
39. Pagès H, Aboyoun P, Gentleman R, *et al*. Biostrings: String objects representing biological sequences, and matching algorithms. 2008; R package version 2.36.4.
40. Kolasinska-Zwierz P, Down T, Latorre I, *et al*. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* 2009;**41**:376–81.
41. Consortium TEP. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 2004;**306**:636–40.
42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.
43. Anders S, Pyl PT, Huber W. HTSeq – A Python framework to work with high-throughput sequencing data. *Bioinformatics* 2014;**31**:btu638–169.
45. Hastie T, Tibshirani R. Generalized additive models. *Statistical Sci* 1986;**1**:297–318.
44. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;**129**:823–37.
46. Kailath T. The divergence and bhattacharyya distance measures in signal selection. *IEEE Trans Commun Technol* 1967;**15**:52–60.
47. Bhattachayya A. On a measure of divergence between two statistical population defined by their population distributions. *Bull Cal Math Soc* 1943;**35**:99–109.
48. Durrett R. Probability Theories and Examples. Duxbury Press, 1996.
49. Landt SG, Marinov GK, Kundaje A, *et al*. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 2012;**22**:1813–31.