PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data

Yanxiang Zhang\textsuperscript{1,*}, Yu-Hsuan Lin\textsuperscript{1,*}, Timothy D. Johnson\textsuperscript{2}, Laura S. Rozeck\textsuperscript{3}, Maureen A. Sartor\textsuperscript{1,2,*}

\textsuperscript{1}Department of Computational Medicine and Bioinformatics, \textsuperscript{2}Department of Biostatistics and \textsuperscript{3}Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI, 48109, United States

ABSTRACT

Motivation: ChIP-Seq is the standard method to identify genome-wide DNA-binding sites for transcription factors (TFs) and histone modifications. There is a growing need to analyze experiments with biological replicates, especially for epigenomic experiments where variation among biological samples can be substantial. However, tools that can perform group comparisons are currently lacking.

Results: We present a Peak-calling Prioritization pipeline (PePr) for identifying consistent or differential binding sites in ChIP-Seq experiments with biological replicates. PePr models read counts across the genome among biological samples with a negative binomial distribution and uses a local variance estimation method, ranking consistent or differential binding sites more favorably than sites with greater variability. We compared PePr to commonly-used and recently proposed approaches on eight TF datasets and show that PePr uniquely identifies consistent regions with enriched read counts, high motif occurrence rate and known characteristics of TF binding based on visual inspection. For histone modification data with broadly enriched regions, PePr identified differential regions that are consistent within groups, and outperformed other methods in scaling FDR analysis.

Availability: http://code.google.com/p/pepr-chip-seq/

Contact: sartorma@umich.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) is the standard technique to identify the genome wide occurrences of transcription factor (TF) binding sites and histone modifications \textit{in vivo}. Over the last few years, there has been tremendous development of analysis methods for ChIP-Seq data, with tens of “peak finders” published (Blahnik, et al., 2010; Boyle, et al., 2008; Fejes, et al., 2008; Jothi, et al., 2008; Kornacker, et al., 2012; Qin, et al., 2010; Rashid, et al., 2011; Rozowsky, et al., 2009; Song and Smith, 2011; Valouev, et al., 2008; Wang, et al., 2013; Xu, et al., 2010; Zang, et al., 2009; Zhang, et al., 2008). Over this course, several characteristics of ChIP-Seq datasets, such as enrichment profile features (peak width, signal-to-noise ratio, location relative to genomic features) of different types of TFs and histone modifications, sources of artifacts, and the commonly observed statistical distributions of read counts, have been gradually revealed. (Park, 2009; Pepke, et al., 2009; Rye, et al., 2011).

As sequencing cost decreases, use of biological replicates is emerging and may eventually become standard practice for ChIP-Seq studies. Most of the ENCODE consortium data was performed in duplicate (Landt, et al., 2012). Furthermore, as researchers shift from performing ChIP-Seq experiments that address mechanistic questions to those that hypothesize differential and/or context-specific binding in a disease, treatment, or epidemiologic setting, the use of replicates to account for individual variability becomes crucial. We expect that this will lead to more analyses comparing a group of ChIP samples with a group of controls, or two groups of ChIP samples, with or without controls, run under different experimental conditions.

ChIP-Seq peak finders that perform direct group comparisons within the peak-calling pipeline are currently lacking. When biological replicates are available, researchers may choose to combine the replicates (CR) in each group and run one-ChIP-vs-one-control analysis to identify all possible peaks. Alternatively, they can pair ChIP and control samples, conduct a separate analysis (SA) for each pair and then stipulate rules to combine the peak-finding results, such as requiring the peaks to be found in all pairwise comparisons. The CR approach is often used in TF ChIP-Seq studies to identify all possible binding sites. However, if the goal is to find consistent binding among replicates, many false positives may occur where binding is present in only one or a subset of the samples. The SA approach is more sophisticated in the sense that it does not lose all information regarding sample-to-sample variability and is more applicable to experiments that have a natural pairing of samples. However, because it evaluates peaks for each replicate separately, the effects of false negatives across replicates may become compounded. Thus, the SA approach is more likely to miss moderate, yet consistent differences in binding. When there is no inherent pairing between test and control samples, as often occurs with differential binding analyses, the SA approach may call a peak because in each one-vs-one analysis one sample has greater enrichment than its paired sample. Yet if the pairs were construct-
ed differently, some “peaks” may no longer exist. An alternative approach is the Irreproducible Discovery Rate (IDR) (Landt, et al., 2012) approach recommended by ENCODE. IDR can be considered a sophisticated CR approach, which assesses the consistency of peak rankings in replicates to find an optimum significance cutoff for determining the final peak list.

Correctly modeling the variation among samples in gene expression studies when testing for differential expression has been shown to be of great importance (Anders and Huber, 2010; Robinson, et al., 2010; Sartor, et al., 2006). For RNA-Seq analysis, several methods (for example, edgeR (Robinson, et al., 2010) and DESeq (Anders and Huber, 2010)) use a negative binomial distribution instead of a Poisson distribution to capture the extra variance among replicates. These approaches can be used with ChIP-Seq data, however they do not perform the first several steps of the ChIP-Seq analysis pipeline nor do they take advantage of local chromosomal information. An exact negative binomial test (diffReps) was recently introduced for ChIP-Seq data and compared to edgeR and DESeq using two histone modification datasets (Shen, et al., 2013). Other approaches to identify differential binding with replicates include the R packages DiffBind (Ross-Innes, et al., 2012) and DBChIP (Liang and Keles, 2012); although these programs take into account sample variation, they rely on other peak callers to generate peak sets for each individual sample first and conduct analysis on the candidate regions that fall within the peak sets.

Here, we introduce a ChIP-Seq Peak-finding and Prioritization (PePr) pipeline that can analyze either a group of ChIP-Seq samples together with controls or compare two groups of ChIP-Seq samples, with or without controls. PePr uses a sliding window approach and models read counts across replicates and between groups with a local negative binomial model. Genomic regions with less variable read counts across replicates are ranked more favorably than regions with greater variability, thus prioritizing consistently enriched regions. We tested PePr on ChIP-Seq data for Activating Transcription factor 4 (ATF4) (Han, et al., 2013), seven ENCODE TF datasets, and one histone modification dataset (H3K27 tri-methylation), and compared the performance of PePr to several ChIP-Seq methods representing different statistical models and using different sources of information: MACS (Zhang, et al., 2008), MACS2 and SPP (Kharchenko, et al., 2008) with IDR (Landt, et al., 2012), ZINBa (Rashid, et al., 2011), SICER (Zang, et al., 2009), diffReps (Shen, et al., 2013), DiffBind (Ross-Innes, et al., 2012) and edgeR (Robinson, et al., 2010). We show that PePr performs favorably compared to the other tested approaches, prioritizing peaks that reflect stronger enrichment fold and higher consistency among samples.

2 METHODS

2.1 Datasets

2.1.1 Activating Transcription factor 4 (ATF4) data: ATF4 data were previously published (Han, et al., 2013). Briefly, samples were obtained from mouse embryonic fibroblasts (MEF) from transgenic mice after 8 hour treatment with tunicamycin (Tm), including three ATF4 wild type ChIP samples and three ATF4 knockout ChIP samples, which served as the controls. Data were obtained from Gene Expression Omnibus (GEO) with the accession number GSE35681.

2.1.2 ENCODE TF data: Neuron-Restrictive Silencer Factor (NRSF), CCCTC-binding factor (CTCF), GA-binding protein (GABP), Nuclear Respiratory factor 1 (NRF1), Structure maintenance of chromosome 3 (SMC3), upstream stimulatory factor 1 (USF1), and USF2 were downloaded from the UCSC collection of ENCODE ChIP-Seq data. Details of the datasets are provided in Table S1.

2.1.3 H3K27me3 data: ChIP-Seq using two Human Papillomavirus (HPV)-positive and two HPV-negative squamous cell carcinoma cell lines were performed. Cell lines were cultured as previously described (Sartor, et al., 2011), and chromatin immunoprecipitation using HistonePath™ (Active Motif) for the commercial-quality antibody pull downs for H3K27 tri-methylation and library preparation were performed by GenPathway, Inc. (part of Active Motif, Carlsbad, CA). DNA was amplified according to the Illumina ChiP-Seq library construction protocol, and a region of 250-350 bp was excised from the preparative Agarose gel. Sequencing of the four immunoprecipitated samples and four input DNA samples was performed at the University of Michigan DNA sequencing core using the Illumina HiSeq with 50 base single-end reads. Data was deposited in GEO with accession number GSE 38629. Raw reads were aligned to hg19 using BWA (Li and Durbin, 2009) with the default parameters.

The total number of aligned reads for each dataset is listed in Table S2. The numbers of peaks called are listed in Table S3.

2.2 PePr algorithm

2.2.1 Data preprocessing: First, all reads are shifted towards their 3’ direction by half of the empirically estimated DNA fragment length, so that forward and reverse strand reads are properly aligned. A recommended window width is estimated based on the average width of the top candidate peaks. PePr then divides the genome into consecutive windows of the chosen width that overlap by 50% and counts the reads in each window for every sample. The read counts for each window are linearly scaled using a normalization constant estimated for each sample; the normalization methods adjust both for total read count among ChIP and control samples, and relative average peak heights (due to differences in ChIP efficiency) among ChIP samples. With the exception of normalization, at each step the user has the option of whether or not to accept the PePr default (see Supplementary Methods for a more detailed description of each of the preprocessing steps).

2.2.2 Negative binomial model and hypothesis testing: For each genomic window, PePr models the read counts with a local negative binomial distribution, assuming that the test and control groups share the same dispersion parameter φ, but possibly different means, such that $\mu_{test} = \gamma \mu_{control}$. To obtain stable variance estimates even with very small sample sizes, φ is estimated from the current window and windows in the local genomic region using a triangular kernel function.

To test if each window is enriched in the test group, we test the hypothesis: $H_0: \gamma \leq 1 \text{ vs } H_1: \gamma > 1$ using a Wald’s test. Using the log transformation, we can define

$$Z = \frac{[\log(\hat{\gamma}) - \log(\gamma_0)]^2}{\hat{\sigma}_\gamma^2}$$

where $\gamma$ is the observed ratio of group means, $\gamma_0=1$, and $\hat{\sigma}_\gamma^2$ is the variance estimate for $\gamma$. Further details of the negative binomial model and the estimation of φ are provided in Supplementary Methods. For differential binding experiments, the normalized input reads are subtracted from the normalized ChIP reads for each window, and both directions are tested with one ChIP group being the test and the other being control. For either type of testing, p-values are then calculated, and significantly enriched windows
are determined based on a user/program specified cut-off for p-value. Nearby enriched windows are merged to form continuous regions, using slightly different parameters for narrow or broad peaks. Optional post-processing steps are available to (1) remove spurious peaks identified by having a high PCR-duplication level and not having a shift between the forward- and reverse-strand peaks (strand lag), and (2) refine peak boundaries for TF binding sites (see Supplementary Methods for more details).

### 2.3 Motif analysis

MEME (Bailey and Elkan, 1994) was used to identify over-represented motifs in the binding sites. For TF datasets, peaks that were found in all programs were used, and the region within 150bp of the peak mode was used as input to MEME. The most significant motifs identified by MEME are listed in Figure S1, and were consistent with previous reports (Han, et al., 2013; Jothi, et al., 2008). Upon obtaining the motif position specific score matrix (PSSM) for each TF from MEME, FIMO (Grant, et al., 2011) was used to find motif matches in the regions within 150bp of the peak mode found by each program to identify the motif occurrences in the peaks.

### 2.4 Unique peak analysis

The versions, parameters and significance cut-offs used for each program are provided in Supplementary Methods (Tables S4 and S5). The unique peaks for each program were defined as the peaks not overlapping any peak from the alternative program being compared. Since the number of unique peaks was often highly imbalanced, we examined the same numbers of top processing steps to generate the final peak calls. For ChIP-Seq data, we conjectured that close genomic regions share a similar microenvironment, and thus their behavior across samples may be correlated. Especially for histone marks like H3K27me3, that result in broad peaks, the estimated dispersion parameters from adjacent regions show strong correlation (Figure S3). Given the high auto-correlation observed for \( \phi \) estimates along the genome, including information from nearby windows effectively increases the sample size, improving the robustness of the estimator.

### 3 RESULTS

#### 3.1 Overview of the PePr method

A schematic overview of the PePr pipeline is shown in Figure 1. After shifting forward and reverse strand reads to achieve proper alignment, PePr estimates a recommended window width based on the median peak width among top pre-candidate peaks to optimize statistical power. This is in contrast to most peak-finders, which use a fixed or user-specified window size. Motivated by the importance of modeling variation in RNA-Seq data, we model read counts with a negative binomial distribution to account for extra-variation beyond that of the Poisson distribution observed in replicated ChIP-Seq data (Figure S2). Unlike the RNA-Seq methods, however, we estimate the dispersion parameter \( \phi \) (which accounts for extra-variation) from the local genomic area. After calculation of p-values, PePr merges adjacent significant windows to form continuous peak regions, which then undergo multiple post-processing steps to generate the final peak calls.

The motivation for estimating \( \phi \) using local genomic information is that estimates of \( \phi \) from one window are often unstable due to the small sample sizes commonly used in ChIP-Seq studies. This problem of unstable variance estimates in experiments with small sample size has been studied extensively in the context of microarray data analysis, where using information from other genes has shown significant improvement (Sartor, et al., 2006; Smyth, 2004). For ChIP-Seq data, we conjectured that close genomic regions share a similar microenvironment, and thus their behavior across
Table 1. Motif occurrence rate in unique peaks called by PePr or alternative programs for NRSF and ATF4. *peaks with highest rank difference were used, as explained in Methods.

<table>
<thead>
<tr>
<th>program compared</th>
<th>NRSF</th>
<th>ATF4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#peaks</td>
<td>% motif</td>
</tr>
<tr>
<td></td>
<td>PePr</td>
<td>alternative</td>
</tr>
<tr>
<td>MACS-CR</td>
<td>227</td>
<td>83.2</td>
</tr>
<tr>
<td>ZINBA-CR</td>
<td>500*</td>
<td>97.8</td>
</tr>
<tr>
<td>MACS2-IDR</td>
<td>416</td>
<td>78.6</td>
</tr>
<tr>
<td>SPP-IDR</td>
<td>500</td>
<td>66.4</td>
</tr>
<tr>
<td>MACS-SA</td>
<td>500</td>
<td>70.4</td>
</tr>
<tr>
<td>ZINBA-SA</td>
<td>499</td>
<td>63.9</td>
</tr>
<tr>
<td>edgeR-basic</td>
<td>366</td>
<td>74.8</td>
</tr>
<tr>
<td>edgeR-plus</td>
<td>151</td>
<td>67.5</td>
</tr>
<tr>
<td>diffReps</td>
<td>403</td>
<td>79.9</td>
</tr>
</tbody>
</table>

The most direct assessment of peak-calling results that has been used is visual inspection of the shape and read coverage of the peak regions (Landt, et al., 2012; Rye, et al., 2011), however because this evaluation process cannot be fully automated, it is often overlooked in the evaluation of ChIP-Seq methods. Instead, much of the literature depends on motif occurrence rate as the main performance measure, which can be inaccurate when the goal of the analysis is to identify differential or consistent binding sites under a specific biological context. Thus, we present visual inspections of the peak profile results, as well as the motif occurrence rates in light of these results.

For each comparison between PePr and an alternative approach, we examined the peaks uniquely found by each (see Methods). In most of the comparisons (except for comparing to diffReps), PePr-unique peaks were more consistent between replicates and showed stronger read intensity (Figure 2A-E (iii) and Figure S4) than the alternative program. Examining each peak individually (Figure 2A-F (ii); the top twenty unique peaks for each method are shown in Supplementary Peak Profiles), PePr-unique peaks exhibited a smooth peak shape and a strand lag, whereas unique peaks found by other approaches either had low read count which formed ambiguous shapes (MACS-SA, ZINBA-SA and edgeR-plus), peak profile shapes suggesting inconsistent binding (MACS-CR, ZINBA-CR, SPP-IDR and MACS2-IDR) or severe PCR-duplications (most notably diffReps and edgeR-basic). As expected by the limitation of the CR approach (including IDR), many of their unique peaks were only observed in one replicate (Figure 2A,B(ii) and more examples in Figure S7).

In the average signal intensity plots, the mode height of MACS-CR, MACS2-IDR and diffReps unique peaks were higher than PePr-unique peaks (Figure 2A,B,F(iv)); however, they were not the expected peak shape, but rather strongly spiked with width close to the read length (Figure S8). This suggests that the reads forming these peaks were mostly PCR duplicates from a limited number of sequences. Some diffReps-unique peaks even had the same peak shape in the control samples, but with fewer reads (Figure 2F(ii,iv) and Figure S8). Whereas the narrow spiked modes are likely false positives with no shift size between strands, the signal levels of the shoulders of these plots likely represent real binding sites, with the expected shift size between strands. The signal in these shoulder regions are higher in PePr than the alternatives (Figure 2A,B(iv)).

Fig. 2. Comparison of PePr to other approaches on NRSF data. Other approaches are: MACS-CR (A); MACS2-IDR (B); SPP-IDR (C); MACS-SA (D); edgeR-plus (E); diffReps (F). The subplots in each panel are: (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Figure S4.

the other methods was also observed for the other TFs (see Figure S6 for ATF4).
We compared the motif rates for peaks uniquely identified by PePr or an alternative approach (Table 1). The peaks uniquely found by PePr had comparable motif occurrence rate to MACSvCR, ZINBAvCR and MACS2vIDR, and had substantially higher motif occurrence rate than MACSvSA, ZINBAvSA, SPPvIDR, diffReps and edgeRvbasic (Table 1). However, the motif rate of PePr’s unique peaks is lower than edgeRvplus for NRSF, contrary to their stronger read signals (Figure 2E (iv)). The difference between edgeRvplus and edgeRvbasic suggests that adopting PePr’s processing steps results in a marked improvement. The CR approaches gained coverage by pooling the samples, resulting in a motif occurrence rate similar to that of PePr. Although the motif is often present, many of their unique peaks only showed enrichment in one replicate (Figure S7), and thus would likely be false positives for identification of consistent binding sites in a specific biological context under study. In the case of diffReps, the low motif occurrence rate in the unique peaks suggests that those peaks with a narrow spiked shape were likely not true NRSF binding sites (Figure S8).

3.2.2 Comparison of PePr and alternative methods using ATF4 and additional ENCODE ChIP-Seq data

We repeated the comparison among methods on ATF4 ChIP-Seq data, which had three samples each of ChIP and control (each having 26.8-30.2 million mapped reads), and the control samples were from chromatin immunoprecipitated ATF4 knockout mice. All methods identified nearly twice as many or more peaks for ATF4 as for NRSF, except diffReps, which identified substantially fewer peaks than all other methods (Table S3). Again, we examined the unique peaks found by PePr versus the other programs. In all comparisons, we observed PePr unique (or ranked higher) peaks had higher read intensities (if we remove the high middle spike which is likely due to PCR duplications) and higher motif occurrence rate than the alternative programs, including edgeRvplus (Figure 3, Figure S9, and Table 1).

We analyzed six additional ENCODE TF datasets (CTCF, GABP, NRF1, SMC3, USF1 and USF2), which had conserved motifs and both duplicated ChIP and control samples. Since we showed in the NRSF comparison that the CR-based methods identify many sites that are inconsistent among samples, in these additional datasets we compared PePr to each of the alternative methods that take into account variation/differences among the replicates.

Table 2. Motif results for ENCODE TF data. *peaks with highest rank difference were used, as explained in Methods.

<table>
<thead>
<tr>
<th>Program</th>
<th>MACS-SA</th>
<th>ZINBA-SA</th>
<th>edgeR-plus</th>
<th>edgeR-basic</th>
<th>diffReps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#peaks</td>
<td>% motifs</td>
<td>#peaks</td>
<td>% motifs</td>
<td>#peaks</td>
</tr>
</tbody>
</table>
| CTCF    | 500     | 76.4     | 81.2       | 500*        | 96.2      | 87.6     | 500     | 78.2     | 78.4    | 278      
| GABP    | 500     | 85.8     | 70.2       | 312         | 55.4      | 22.7     | 172     | 54.6     | 8.1     | 370      
| NRF1    | 500     | 82.4     | 74.4       | 500*        | 99.0      | 95.2     | 500*    | 89.0     | 96.0    | 500*     
| SMC3    | 500     | 96.0     | 62.6       | 500         | 90.6      | 24.2     | 500     | 96.8     | 50.4    | 500      
| USF1    | 500     | 68.2     | 78.6       | 500         | 69.2      | 56.8     | 500     | 71.8     | 44.4    | 500      
| USF2    | 500     | 64.6     | 49.6       | 500*        | 79.2      | 73.8     | 173     | 52.0     | 43.3    | 500*     

Fig. 3. Comparison of PePr to other approaches on ATF4 data. Other approaches are: MACSvCR (A); MACS2vIDR (B); SPP-IPDR (C); MACS-SA (D); edgeR-plus (E); diffReps (F). The subplots in each panel are: (i) Venn diagram of overlap between peaks found by PePr and the alternative approach. (ii) Representative genomic view of the unique peaks. Each line represents one of the replicates in the group, with the top window being the test group and the bottom window being the control group. (iii) Heatmaps showing the signal intensity of the test group across the unique peaks. The x-axis denotes the relative chromosomal locations centered at the peak mode; each row denotes one peak. (iv) Average signal intensity of the unique peaks. Solid lines represent the test group, while dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Figure S9.

both, illustrating the substantial level of variation among the samples in each group. PePr, diffReps, and edgeR-plus identified 1015, 17924 and 181 peaks, respectively. EdgeR-plus identified very few significant peaks, possibly due to the high global dispersion parameter estimated from the data, whereas PePr estimated it locally. DiffReps was much more sensitive than edgeR as previously shown for broad peaks (Shen, et al., 2013). For DiffBind, SICER peaksets were generated for each cell line and used as input to the program. The peaksets were merged and a total of 29510 regions were tested. DiffBind has two built-in analysis methods: edgeR and DESeq. DiffBind with edgeR reported no significant peaks (possibly due to the same reason explained above). With DESeq it identified 918 HPV(-) specific sites, which we use for DiffBind below.

Since the number of peaks varied substantially among the programs tested, we evaluated how each program prioritized the peak findings. The top 900 peaks from each program were chosen based on their significance and compared (edgeR-plus was excluded due to finding so few peaks; 900 was chosen because all other methods identified ≥900 sites). Figure 4A shows the overlap between PePr and each of the other five programs. We examined the top ranking peaks that were uniquely identified by PePr, SICER, ZINBA, DiffBind or diffReps. The peaks uniquely identified by PePr were consistent between replicates, while the peaks uniquely found by SICER, ZINBA, or DiffBind often showed large differences in coverage between samples in the same group (Figure 4B, C, D). DiffReps unique peaks seemed consistent in coverage between samples in the same group, however, the ratio in coverage between the test and control groups was smaller than that of PePr (Figure 4E). In addition, when looking at each peak individually, the top diffReps unique peaks had average peak width less than 2 kb, which is much narrower than that expected for H3K27me3.

To further assess the robustness of the peak-calling methods (as opposed to differential binding methods) in identifying broad peaks, we conducted a scaling FDR analysis as described in (Zang, et al., 2009) for all four ChIP-Seq versus four matching input controls. Briefly, for each replicate, we randomly sample half of the reads to produce several pseudo half-size libraries. The proportion of peaks called only in the half-size library but not in the full-size library is defined as the scaling FDR. Performing this for ten simulations, we observed that PePr had a smaller scaling FDR (mean = 1.66%) than SICER (mean = 4.56%), ZINBA (mean = 11.38%), diffReps (mean=12.54%) and edgeR-plus (mean=5.83%), and thus PePr’s peak prediction was most robust to differences in coverage levels (Figure S11).

### 4 DISCUSSION

Currently, there is a lack of ChIP-Seq analysis programs that account for biological variability within the peak-finding process. We have developed a method and tool, PePr, which uses a local negative binomial model to identify consistent or differential binding sites among ChIP-Seq data, and that additionally estimates the optimal moving window size and offers post-processing steps to reduce false positives and refine peak resolution.

Variation among samples in ChIP-Seq data can sometimes be quite large, such that some binding sites, even for transcription factors, are not reproducible (Landt, et al., 2012). Inconsistent TF peaks among biological samples can exist for many reasons, including differences in accessibility of chromatin regions (e.g., due to
to histone tail modifications or DNA methylation), common sequence variants, competitive or cooperative binding differences with another TF (Whitfield, et al., 2012), or technical artifacts that only occurred in one of the replicates. However, all but the last of these reasons are not significant concerns for most peak-finder programs, the goal of which is to identify all potential binding sites rather than consistent or differential binding sites. In addition, as public datasets from large consortiums such as ENCODE (Consortium, et al., 2012) more comprehensively cover known TF binding in commonly used cell and tissue types, there will be less incentive for individual laboratories to identify all of the potential binding sites for a protein, as many will be available. A more refined hypothesis may be “where does this TF (or histone modification) bind consistently in this specific context (a specific disease, developmental stage, exposure, or treatment)?” Accurately modeling the variation is highly important in population epigenomics studies where substantial variation exists among samples, not only among individuals but also between tissue types (Cui, et al., 2009), developmental time points (Rugg-Gunn, et al., 2010; Sarmento, et al., 2004), and during disease progression (Conte and Altucci, 2014; Jakopovic, et al., 2013).

We compared PePr to five commonly used single-sample peakfinders that use different underlying statistical models (MACS, MACS2, SPP, ZINBA, and SICER), as well as three programs that were designed for replicates (diffReps, DiffBind and edgeR), and found that PePr performed favorably in terms of consistently enriched read counts, motif occurrence rate and known characteristics of TF binding based on visual inspection. For comparison with MACS, ZINBA or SICER, we either performed separate paired analyses and then called peaks in the overlapping regions (SA) or combined the reads for the replicates and called peaks from the concatenated lists (CR). IDR was incorporated with MACS2 and SPP to determine the peak list cut-off, as recommended by the ENCODE consortium. Visual inspection of the peak shape and summarizing the read counts in peaks were extremely valuable in characterizing the unique tendencies of each approach. In particular, MACS was sensitive to detecting regions that had very low background and tended to miss peaks that had a relatively high background (Figure 2A(iii)); visual inspection of the ZINBA and diffReps unique peaks revealed that many had similar peak shape in both ChIP and control samples; SPP had severe false negatives for the NRSF data, which is possibly due to the removal of true binding sites that SPP mistakenly assumed to be artifacts due to having unexpectedly small shift size (i.e., in the “phantom peak” as defined in (Landt, et al., 2012)). When we compared PePr to SPP-IDR and MACS2-IDR, we observed PePr-unique peaks (that are missed by the other two) had high read counts and motif rate.

Although motif occurrence rate is a useful marker for DNA binding, its value as a marker for consistent or differential DNA binding is not as clear. For identification of all DNA binding sites, motif analysis is expected to be specific (if the motif is found within a peak, it is assumed that binding occurs) but not necessarily highly sensitive (indirect binding cooperatively with other protein(s) may not result in a motif occurrence). Because the percent of binding sites without a motif is only expected to vary by DNA binding protein, and not by peak-caller, this is often ignored when comparing peak-callers. However, for consistent or differential binding experiments we can no longer assume specificity; a peak-finder that identifies fewer overall peaks with a motif than an alternative may be correct in not calling the additional peaks as consistently bound or differentially bound. Given these caveats, we nonetheless found motif occurrence rate informative for interpreting our results when used in conjunction with visual inspection of peaks. The large improvement in motif occurrence rates for PePr-unique peaks compared to peaks identified by the Separate Analyses (SA) approaches and edgeR-basic suggests that peaks with higher read intensities and the expected smooth peak shape are more likely to contain a motif (Table 1). The CR approaches, on the other hand, were comparable in motif occurrence rate to PePr, but many of these were only bound in one replicate upon visual inspection, and thus are likely false positives for identification of consistent binding sites in the biological system under study.

Although PePr and edgeR use a similar underlying negative binomial model, edgeR lacks initial steps required for ChIP-Seq peak finding (shifting opposite strand reads, defining and summarizing reads per window, etc), does not incorporate information from neighboring windows which especially benefits histone modification analyses, and does not offer post-processing steps to improve peak resolution or reduce false positives. In 5 of the 8 TF datasets, PePr performed better in motif rate than edgeR if the same PePr-processing steps are performed for edgeR; with the histone data, PePr was more sensitive than edgeR due to estimation of the dispersion parameters locally. PePr’s post-processing steps improved edgeR’s performance when there is a high proportion of PCR-duplicate peaks (6% in NRSF, 1.6% in ATF4, and <0.5% in other datasets). DiffBind was previously shown to work well with differential binding in TF data (Ross-Innes, et al., 2012), however, in H3K27me3 data with broad and highly variable peaks, DiffBind’s edgeR module had very low detection power, while its DESeq module identified 918 peaks, many of which were inconsistent among samples. DiffReps resulted in unpredictable numbers of peaks across the datasets we tested, for example, it identified substantially fewer peaks than all other programs for ATF4 and SMC3 (Table S3), but many more for H3K27me3.

Due to the lack of benchmarks in histone modification datasets, in this manuscript we mainly relied on transcription factor datasets to compare methods. However, PePr is adaptable to datasets with either sharp or broad peaks due to its empirical estimation of the optimal sliding window size, and thus is equally relevant for analysis of histone modification ChIP-Seq datasets as illustrated with our H3K27me3 data. H3K27me3 tends to occur in broad regions several kilobases in length, making consistent peak calling more difficult. Based on our visual inspection of peaks and scaling FDR analysis for the approaches compared, we showed that PePr identified binding regions consistent between groups without being sensitive to changes in read coverage.

One limitation of PePr is that it currently does not perform paired analysis, similar to the limitation of multiple RNA-Seq differential analysis programs (Anders and Huber, 2010; Trapnell, et al., 2013); thus, for example, the paired nature of tumor and patient-matched normal samples, could not be taken into account. For experiments requiring covariates, we currently recommend edgeR. PePr also makes the assumption that the quality of data for each ChIP-Seq experiment is approximately equal, similar to most methods for other types of high-throughput analysis. When this assumption is violated, the result may be a high false negative rate due to missing peak regions in the lower quality sample(s); this may especially be true for experiments with very small sample
size. In this case, users may obtain better performance using a different peak finder on individual samples, and a secondary method to explore options to combine results. Future versions of peak finders for replicated ChIP-Seq data could take into account quality, for example by assigning a weight to each sample.

ACKNOWLEDGEMENTS

Funding; Funding for this work was provided by a Career Development Award from NIH grant SPORE P50 CA097248/CA/NCI, NCI/NIDCR grant ROI CA158286-01, and NIEHS grant P30 ES017885-01.

REFERENCES

Rashid, N.U., et al. (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions, Genome Biol, 12, R67.