BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 18 2014, pages 2568–2575
doi:10.1093/bioinformatics/btu372

Genome analysis

Advance Access publication June 3, 2014

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

Yanxiao Zhang[1,†], Yu-Hsuan Lin[1,†], Timothy D. Johnson[2], Laura S. Rozek[3] and Maureen A. Sartor[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, [2]Department of Biostatistics and [3]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

Associate Editor: Michael Brudno

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 with 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 False Discovery Rate (FDR) analysis.

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

Contact: sartorma@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 6, 2014; revised on May 11, 2014; accepted on May 29, 2014

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 *in vivo*. Over the past 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 and 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 the standard practice for ChIP-Seq studies. Most of the Encyclopedia of DNA Elements (ENCODE) consortium data were 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-versus-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-versus-one analysis, one sample has greater enrichment than its paired sample. Yet, if

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

the pairs were constructed 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 with 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 with the other tested approaches, prioritizing peaks that reflect stronger enrichment fold and higher consistency among samples.

2 METHODS

2.1 Datasets

2.1.1 ATF4 data ATF4 data were previously published (Han *et al.*, 2013). Briefly, samples were obtained from mouse embryonic fibroblasts from transgenic mice after 8 h treatment with tunicamycin, 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 Supplementary Table S1.

2.1.3 H3K27me3 data ChIP-Seq using two human papillomavirus (HPV)-positive and two HPV-negative squamous cell carcinoma (SCC) cell lines were performed. Cell lines were cultured as previously described (Sartor *et al.*, 2011), and chromatin immunoprecipitation using HistonePath$^{\mathrm{TM}}$ (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 were 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 Supplementary Table S2. The numbers of peaks called are listed in Supplementary Table S3.

2.2 PePr algorithm

2.2.1 Data preprocessing First, all reads are shifted toward 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 pre-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 (owing to differences in ChIP efficiency) among ChIP samples. With the exception of normalization, at each step the user has the option of whether 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 $\mu_{\mathrm{test}} = \gamma \cdot \mu_{\mathrm{control}}$. To obtain stable variance estimates even with 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: $\gamma \le 1$ versus H_1:$\gamma > 1$ using a Wald's test. Using the log transformation, we can define

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

where $\hat{\gamma}$ is the observed ratio of group means, $\gamma_0 = 1$, and is the variance estimate for $\hat{\sigma}\hat{\gamma}^2$. 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 cutoff for *P*-value. Nearby enriched windows are merged to form continuous

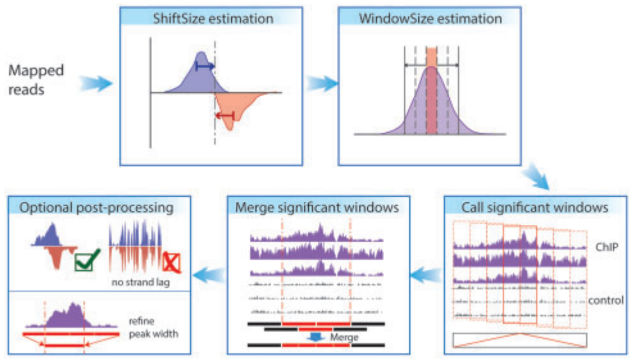

Fig. 1. Workflow of PePr

regions, using slightly different parameters for narrow or broad peaks. Optional post-processing steps are available to (i) 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 (ii) refine peak boundaries for TF binding sites (see Supplementary Methods for more details).

2.3 Motif analysis

Multiple Em for Motif Elicitation (MEME) (Bailey and Elkan, 1994) was used to identify overrepresented motifs in the binding sites. For TF datasets, peaks that were found in all programs were used, and the region within 150 bp of the peak mode was used as input to MEME. The most significant motifs identified by MEME are listed in Supplementary Figure S1, and were consistent with previous reports (Han *et al.*, 2013; Jothi *et al.*, 2008). On obtaining the motif position-specific score matrix for each TF from MEME, Find Individual Motif Occurrences (FIMO) (Grant *et al.*, 2011) was used to find motif matches in the regions within 150 bp 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 cutoffs used for each program are provided in Supplementary Methods (Supplementary 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. Because the number of unique peaks was often highly imbalanced, we examined the same numbers of top unique peaks with a maximum of 500. If too few unique peaks (<150) were identified, then the top 500 peaks identified by each but with the highest difference in rank were used as a surrogate to unique peaks. The heatmaps of unique peaks were generated using Hypergeometric Optimization of Motif EnRichment (HOMER) (Heinz *et al.*, 2010) and visualized with Java TreeView (Saldanha, 2004).

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 (Supplementary Fig. S2). Unlike the RNA-Seq methods, however, we estimate the dispersion parameter φ (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 φ using local genomic information is that estimates of φ from one window are often unstable owing 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 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 (Supplementary Fig. S3). Given the high autocorrelation observed for φ estimates along the genome, including information from nearby windows effectively increases the sample size, improving the robustness of the estimator.

3.2 Comparisons with other methods

We assess the performance of PePr by applying it to eight TF ChIP-Seq datasets: NRSF, ATF4, CTCF, GABP, NRF1, SMC3, USF1 and USF2, and to a histone modification dataset: tri-methylation of Histone 3 lysine 27 (H3K27me3). The performance was compared with MACS, MACS2, SPP, ZINBA, edgeR and diffReps for TF data, and to SICER, ZINBA, edgeR, DiffBind and diffReps for histone data. MACS and SICER are among the favorite choices for sharp peaks and broad peaks, respectively. ZINBA uses a similar sliding window approach, and is geared toward either sharp or broad peaks. MACS and ZINBA were run with both the CR and SA approaches described above. IDR was incorporated with MACS2 and SPP, as is recommended by the ENCODE project. EdgeR was performed in two ways to distinguish the effects of the core statistical model from the effects of the pre- and post-processing steps: adopting all of PePr's pre- and post-processing steps, and following a basic processing procedure (see Supplementary Methods for details). We denote them as edgeR-plus and edgeR-basic, respectively, in the main text.

3.2.1 Comparison of PePr and alternative methods using NRSF ChIP-Seq data The NRSF data consists of two ChIP and two input DNA samples, each sample having 14.3–26.6 million mapped reads. PePr identified a total of 5284 peaks, comparable with diffReps, edgeR, SA and IDR-based approaches (Fig. 2 B–F(i), Supplementary Fig. S4). CR-based approaches identified significantly more peaks, as expected, because they gained coverage by pooling samples in the same group. This trend in number of peaks was also observed for the other TFs (Supplementary Table S3). Comparing the ranks of peaks among the methods, we observed high correlation between PePr and MACS-CR (Pearson's $r = 0.73$), MACS-SA ($r = 0.79$),

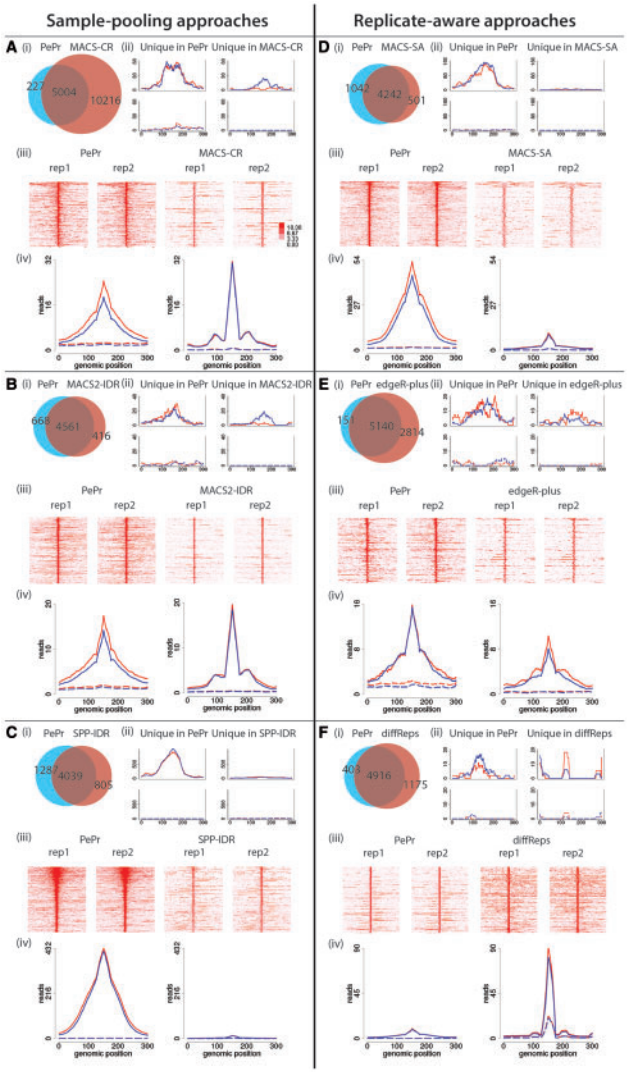

Fig. 2. Comparison of PePr with other approaches on NRSF data. Other approaches are MACS-CR (**A**), MACS2-IDR (**B**), SPP-IDR (**C**), MACS-SA (**D**), edgeR-plus (**E**) and diffReps (**F**). The subplots in each panel are (i) Venn diagram of overlap between peaks found by PePr and the alternative 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, whereas dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Supplementary Figure S4

SPP-IDR ($r = 0.93$), MACS2-IDR ($r = 0.65$), edgeR-basic ($r = 0.78$) and edgeR-plus ($r = 0.84$), but much lower correlation between PePr and ZINBA-CR ($r = 0.14$), ZINBA-SA ($r = 0.16$) and diffReps ($r = –0.25$) (Supplementary Fig. S5). This trend in rank correlations between PePr and the other methods was also observed for the other TFs (see Supplementary Fig. S6 for ATF4).

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 Section 2). In most of the comparisons (except for comparing with diffReps), PePr-unique peaks were more consistent between replicates and showed stronger read intensity (Fig. 2A–E(iii) and Supplementary Fig. S4) than the alternative program. Examining each peak individually [Fig. 2A–F(ii); the top 20 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 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 (Fig. 2A and B(ii) and more examples in Supplementary Fig. 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 [Fig. 2A, B and F(iv)]; however, they were not the expected peak shape, but rather strongly spiked with width close to the read length (Supplementary Fig. 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 [Fig. 2F(ii, iv) and Supplementary Fig. S8]. Although 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 [Fig. 2A and B(iv)].

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 with MACS-CR, ZINBA-CR and MACS2-IDR, and had substantially higher motif occurrence rate than MACS-SA, ZINBA-SA, SPP-IDR, diffReps and edgeR-basic (Table 1). However, the motif rate of PePr's unique peaks is lower than edgeR-plus for NRSF, contrary to their stronger read signals [Fig. 2E(iv)]. The difference between edgeR-plus and edgeR-basic 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 (Supplementary Fig. 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

Table 1. Motif occurrence rate in unique peaks called by PePr or alternative programs for NRSF and ATF4

Program compared	NRSF			ATF4		
	Number of peaks	% motif		Number of peaks	% motif	
		PePr	Alternative		PePr	Alternative
MACS-CR	227	77.1	**83.2**	152	**48.0**	26.3
ZINBA-CR	500[a]	**97.8**	92.6	500[a]	**67.0**	51.4
MACS2-IDR	416	76.2	**78.6**	211	**40.3**	20.0
SPP-IDR	500	**95.2**	66.4	500	**43.0**	39.2
MACS-SA	500	**89.0**	70.4	500	**45.8**	29.8
ZINBA-SA	499	**76.5**	63.9	332	**56.0**	16.9
edger-basic	366	**74.8**	67.5	500	**56.2**	39.4
edgeR-plus	151	67.5	**82.7**	500	**59.0**	45.8
diffReps	403	**79.9**	25.6	500[a]	**75.2**	61.6

Note: Bold values indicate the higher motif occurrence rate in each comparison. [a]Peaks with highest rank difference were used, as explained in Section 2.

occurrence rate in the unique peaks suggests that those peaks with a narrow spiked shape were likely not true NRSF binding sites (Supplementary Fig. 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 (Supplementary 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 owing to PCR duplications) and higher motif occurrence rate than the alternative programs, including edgeR-plus (Fig. 3, Supplementary Fig. S9 and Table 1).

We analyzed six additional ENCODE TF datasets (CTCF, GABP, NRF1, SMC3, USF1 and USF2), which had conserved motifs and both replicated ChIP and control samples. Because 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 with each of the alternative methods that take into account variation/differences among the replicates: diffReps, edgeR-basic, edgeR-plus, MACS-SA and ZINBA-SA, with the same motif analysis (Table 2). In 24 of 30 comparisons, PePr-unique peaks had higher motif occurrence rate than the alternative method.

3.2.3 Comparison of PePr with alternative methods using a histone modification dataset Oncogenic HPV infection and tobacco use are associated with the etiology of two subtypes of oropharyngeal SCCs (Chung and Gillison, 2009). We generated H3K27me3 ChIP-Seq data from two HPV(+) and two age- and gender-matched HPV(–) SCC cell lines. The aim of the study was to identify candidate differential H3K27me3 sites by HPV status. The H3K27me3 mark exhibits broadly enriched regions in ChIP-Seq data, which we observed to be often highly variable between samples (Supplementary Fig. S10). Owing to the high variation among samples and the goal of identifying consistent differences between HPV(+) and HPV(–) tumors, the CR approach would not be suitable; thus, we compared PePr with the SA approach using two peak callers developed for broad peaks: ZINBA and SICER, as well as diffReps, DiffBind and edgeR-plus.

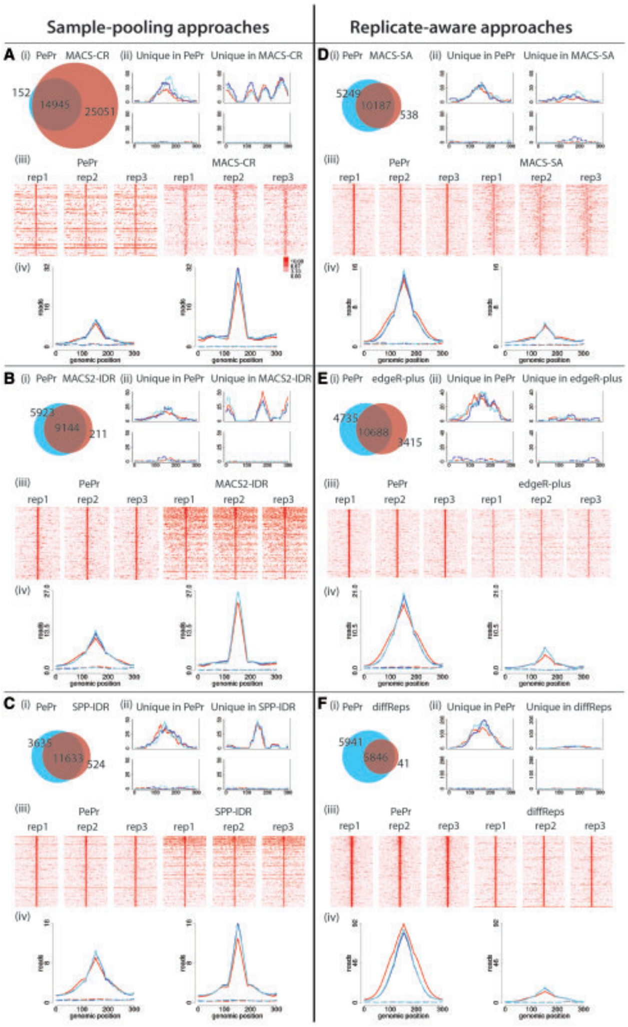

Fig. 3. Comparison of PePr with other approaches on ATF4 data. Other approaches are MACS-CR (**A**), MACS2-IDR (**B**), SPP-IDR (**C**), MACS-SA (**D**), edgeR-plus (**E**) and diffReps (**F**). The subplots in each panel are (i) Venn diagram of overlap between peaks found by PePr and the alternative 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, whereas dashed lines represent the control group. ZINBA and edgeR-basic results are presented in Supplementary Figure S9

To find HPV(–)-specific peaks, we used HPV(–) cell lines as the test samples and compared them with the HPV(+) cell lines.

Table 2. Motif results for ENCODE TF data

Program	MACS-SA			ZINBA-SA			edgeR-plus			edgeR-basic			diffReps		
	Number of peaks	% motifs		Number of peaks	%motifs		Number of peaks	%motifs		Number of peaks	%motifs		Number of peaks	%motifs	
		PePr	Alter.		PePr	Alter.		PePr	Alter.		PePr	Alter.		PePr	Alter.
CTCF	500	76.4	**81.2**	500[a]	**96.2**	87.6	500	78.2	**78.4**	278	73.0	**80.2**	500[a]	**95.8**	87.2
GABP	500[a]	**85.8**	70.2	312	**55.4**	22.7	172	**54.6**	8.1	370	**55.6**	24.6	500[a]	**89.2**	74.8
NRF1	500	**82.4**	74.4	500[a]	**99.0**	95.2	500[a]	89.0	**96.0**	500[a]	90.4	**96.0**	500[a]	**99.6**	95.6
SMC3	500	**96.0**	62.6	500	**90.6**	24.2	500	**96.8**	50.4	500	**97.8**	58.8	165	**97.6**	7.9
USF1	500	68.2	**78.6**	500	**69.2**	56.8	500	**71.8**	44.4	500	**74.0**	46.0	202	**73.8**	29.7
USF2	500	**64.6**	49.6	500[a]	**79.2**	73.8	173	**52.0**	43.3	500 [a]	**67.4**	60.8	216	**69.9**	19.0

Note: Bold values indicate the higher motif occurrence rate in each comparison. [a]Peaks with highest rank difference were used, as explained in Section 2.

For ZINBA and SICER, age- and gender-matched samples were used in each pair. SICER identified 35 403 HPV(–)-specific peaks for the first pair, 20 207 peaks for the second pair and 8823 regions (19%) were found in both. ZINBA identified 13 814 peaks and 22 701 peaks, respectively, and only 1878 regions (5%) were found in both, illustrating the substantial level of variation among the samples in each group. PePr, diffReps and edgeR-plus identified 1015, 17 924 and 181 peaks, respectively. EdgeR-plus identified few significant peaks, possibly because of 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 29 510 regions were tested. DiffBind has two built-in analysis methods: edgeR and DESeq. DiffBind with edgeR reported no significant peaks (possibly because of the same reason explained above). With DESeq it identified 918 HPV(–)-specific sites, which we use for DiffBind below.

Because 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 because of finding 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, whereas the peaks uniquely found by SICER, ZINBA or DiffBind often showed large differences in coverage between samples in the same group (Fig. 4B–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 (Fig. 4E). In addition, when looking at each peak individually, the top diffReps unique peaks had average peak width <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 False Discovery Rate (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 10 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

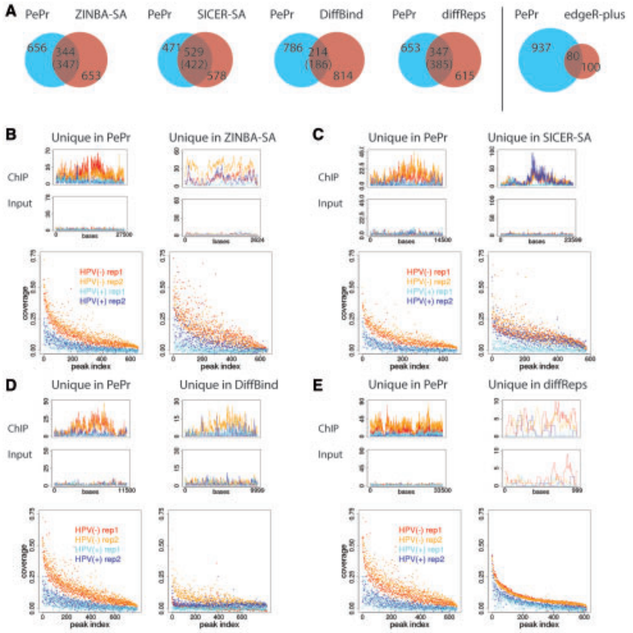

Fig. 4. Comparison of PePr with other approaches for H3K27me3 data. (**A**) Venn diagrams showing the overlap between the top 900 peaks from PePr and alternative approaches (the number in parenthesis shows the number of peaks in the alternative program that overlap with PePr peaks). (**B–E**) Each plot on top shows the genomic view of top-ranking peaks uniquely found by PePr or the alternative approach. The bottom plots show the normalized coverage of reads in unique peaks, sorted by the average coverage of both HPV(–) samples

PePr's peak prediction was most robust to differences in coverage levels (Supplementary Fig. 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 large, such that some binding sites, even for TFs, 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. because of the 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 with five commonly used single-sample peak finders 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 cutoff, 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 low background and tended to miss peaks that had a relatively high background [Fig. 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 because of the removal of true binding sites that SPP mistakenly assumed to be artifacts because of having unexpectedly small shift size [i.e. in the 'phantom peak' as defined in (Landt *et al.*, 2012)]. When we compared PePr with 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 with peaks identified by the 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 with PePr, but many of these were only bound in one replicate on 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 five of the eight 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 owing 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 low detection power, whereas 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 (Supplementary Table S3) but many more for H3K27me3.

Owing to the lack of benchmarks in histone modification datasets, in this manuscript we mainly relied on TF datasets to compare methods. However, PePr is adaptable to datasets with either sharp or broad peaks because of 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 owing to missing peak regions in the lower quality sample(s); this may especially be true for experiments with 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.

Funding: Funding for this work was provided by a Career Development Award from National Institutes of Health (NIH) (SPORE P50 CA097248/CA/NCI), National Cancer Institute (NCI)/National Institute of Dental and Craniofacial Research (NIDCR) (R01CA158286-01) and National Institute of Environmental Health Sciences (NIEHS) (P30 ES017885-01).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol*., **11**, R106.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc*. *Int*. *Conf*. *Intell*. *Syst*. *Mol*. *Biol*., **2**, 28–36.

Blahnik,K.R. *et al*. (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res*., **38**, e13.

Boyle,A.P. *et al*. (2008) F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**, 2537–2538.

Chung,C.H. and Gillison,M.L. (2009) Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. *Clin*. *Cancer Res*., **15**, 6758–6762.

Consortium,E.P. *et al*. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Conte,M. and Altucci,L. (2014) Functions, aberrations, and advances for chromatin modulation in cancer. *Cancer Treat*. *Res*., **159**, 227–239.

Cui,K. *et al*. (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, **4**, 80–93.

Fejes,A.P. *et al*. (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 1729–1730.

Grant,C.E. *et al*. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Han,J. *et al*. (2013) ER-stress-induced transcriptional regulation increases protein synthesis leading to cell death. *Nat*. *Cell Biol*., **15**, 481–490.

Heinz,S. *et al*. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol*. *Cell*, **38**, 576–589.

Jakopovic,M. *et al*. (2013) Targeting the epigenome in lung cancer: expanding approaches to epigenetic therapy. *Front*. *Oncol*., **3**, 261.

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

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

Kornacker,K. *et al*. (2012) The Triform algorithm: improved sensitivity and specificity in ChIP-Seq peak finding. *BMC Bioinformatics*, **13**, 176.

Landt,S.G. *et al*. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*., **22**, 1813–1831.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Liang,K. and Keles,S. (2012) Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics*, **28**, 121–122.

Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat*. *Rev*. *Genet*., **10**, 669–680.

Pepke,S. *et al*. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat*. *Methods*, **6**, S22–S32.

Qin,Z.S. *et al*. (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

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.

Robinson,M.D. *et al*. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Ross-Innes,C.S. *et al*. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Rozowsky,J. *et al*. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat*. *Biotechnol*., **27**, 66–75.

Rugg-Gunn,P.J. *et al*. (2010) Distinct histone modifications in stem cell lines and tissue lineages from the early mouse embryo. *Proc*. *Natl Acad*. *Sci*. *USA*, **107**, 10783–10790.

Rye,M.B. *et al*. (2011) A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res*., **39**, e25.

Saldanha,A.J. (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.

Sarmento,O.F. *et al*. (2004) Dynamic alterations of specific histone modifications during early murine development. *J*. *Cell Sci*., **17**, 4449–4459.

Sartor,M.A. *et al*. (2011) Genome-wide methylation and expression differences in HPV(+) and HPV(-) squamous cell carcinoma cell lines are consistent with divergent mechanisms of carcinogenesis. *Epigenetics*, **6**, 777–787.

Sartor,M.A. *et al*. (2006) Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, **7**, 538.

Shen,L. *et al*. (2013) diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One*, **8**, e65598.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat*. *Appl*. *Genet*. *Mol*. *Biol*., **3**, Article3.

Song,Q. and Smith,A.D. (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics*, **27**, 870–871.

Trapnell,C. *et al*. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat*. *Biotechnol*., **31**, 46–53.

Valouev,A. *et al*. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat*. *Methods*, **5**, 829–834.

Wang,J. *et al*. (2013) BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. *Bioinformatics*, **29**, 492–493.

Whitfield,T.W. *et al*. (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol*., **13**, R50.

Xu,H. *et al*. (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, **26**, 1199–1204.

Zang,C. *et al*. (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, **25**, 1952–1958.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.