RECAP reveals the true statistical significance of ChIP-seq peak calls

Justin G. Chitpin [1,2], Aseel Awdeh[2,3] and Theodore J. Perkins [2,3,4,*]

February 5, 2018

[1]Translational and Molecular Medicine Program, University of Ottawa, Ottawa, ON, K1H8M5, Canada
[2]Regenerative Medicine Program, Ottawa Hospital Research Institute, Ottawa, ON, K1H8L6, Canada
[3]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, K1N6N5, Canada
[4]Department of Biochemistry, Microbiology and Immunology, University of Ottawa, Ottawa, ON, K1H8M5, Canada
[*] Correspondence to tperkins@ohri.ca

Abstract

Motivation: ChIP-seq is used extensively to identify sites of transcription factor binding or regions of epigenetic modifications to the genome. The fundamental bioinformatics problem is to take ChIP-seq read data and data representing some kind of control, and determine genomic regions that are enriched in the ChIP-seq versus the control, also called "peak calling." While many programs have been designed to solve this task, nearly all fall into the statistical trap of using the data twice—once to determine candidate enriched regions, and a second time to assess enrichment by methods of classical statistical hypothesis testing. This double use of the data has the potential to invalidate the statistical significance assigned to enriched regions, or "peaks", and as a consequence, to invalidate false discovery rate estimates. Thus, the true significance or reliability of peak calls remains unknown.
Results: We show, through extensive simulation studies of null hypothesis data, that three well-known peak callers, MACS, SICER and diffReps, output optimistically biased p-values, and therefore optimistic false discovery rate estimates—in some cases, orders of magnitude optimistic. We also propose a new wrapper algorithm called RECAP, that uses resampling of ChIP-seq and control data to estimate and correct for biases built into peak calling algorithms. RECAP also enables for the first time local false discovery rate analysis, so that the likelihood of individual peaks being true positives or false positives can be estimated based on their re-calibrated p-values. RECAP is a powerful new tool for assessing the true statistical significance of ChIP-seq peak calls.
Availability: The RECAP software is available at www.perkinslab.ca.

1 Introduction

Chromatin Immunopreciptation followed by high-throughput sequencing, or ChIP-seq, has become a central approach to mapping transcription factor-DNA binding sites and studying the epigenome [16, 12, 21]. ChIP-seq is the primary technique employed by a number of highly successful large-scale genomics projects, including ENCODE [6, 3], modENCODE [5, 19], the NIH Roadmap Epigenomics Project [2, 13], and the International Human Epigenome Consortium [23]. Collectively, these projects have generated over 10,000 ChIP-seq data sets at a cost of 10s or 100s of millions of dollars, while other, smaller-scale projects have generated many more. Thus, understanding exactly how much information we can or should extract from such data is a question of paramount importance.

Bioinformatics analysis of ChIP-seq data is a multi-stage process [14], with the end goal of identifying genomic regions enriched for ChIP-seq signal—the regions that represent locations of possible transcription factor (TF)-DNA binding, or histone positions, or chromatin marks, etc. There are numerous algorithms for identifying ChIP-seq enriched regions, or peak-calling (e.g., [8, 25, 30, 22, 24, 29, 11, 18, 27, 1, 20]). Because ChIP-seq data is noisy, virtually all peak calling algorithms output peaks (i.e., enriched regions) with associated p-values. These p-values are useful for ranking peaks in decreasing order of confidence, and estimating false discovery rates at different significance thresholds. But how well can we trust the p-values produced by peak callers?

Although approaches to peak calling differ in a number of ways, many follow a common two-stage pattern: First, candidate peaks are identified by analyzing the ChIP-seq data, and second, those candidate peaks are evaluate for significance by comparing ChIP-seq data with some kind of control data. (Or, in the case of differential enriched region detection, two ChIP-seqs may be compared to each other.) The problem with this design, as already pointed out by Lun and Smyth [15], is that it commits the statistical sin of using the data twice. More specifically, the ChIP-seq data is used to construct hypotheses to test (the candidate peaks), and then the same ChIP-seq data, along with control data, is used to test those hypotheses by means of classical statistical hypothesis testing. In general, when the hypothesis and the test both depend on the same data, classical p-values cannot be trusted.

To be more concrete, let us consider three specific algorithms that we chose to focus on in this paper: MACS [30, 10, 9], SICER [29, 28], and diffReps [20]. We chose to study MACS because it is, at present, the most highly cited peak caller, and it is used by the ENCODE and modENCODE consortia for analysis of their data. SICER is another widely used and high-cited algorithm, but one designed more for the detection of the broad, regional enrichment characteristic of certain chromatin marks. This suits some of our experiments below, although MACS is also able to detect such regions, particularly when used in "broad peak" mode. diffReps is designed to solve the differential enrichment problem—the comparison of two ChIP-seqs instead of a ChIP-seq and a control—which again comes up in certain of our experiments.

Let us consider why MACS [30, 10, 9] may produce biased p-values. After fragment size estimation and read shifting, MACS scans a fixed-width window across the genome, counting ChIP-seq reads. An initial p-value is assigned to each window by comparing its read count to the expectation under a Poisson model with rate parameter that depends on window size, genome size, and total reads in the ChIP-seq data set. If that initial p-value is less than or equal to 10^{-5} (a parameter the user can specify), the window is deemed enriched compared to a flat background model. Overlapping enriched windows are then merged, resulting in candidate peak regions. The candidate peak regions are then tested for enrichment versus the control, and that p-value is attached to the candidate peak. Both the initial selection for enriched windows and the merging of windows tend to result in candidate peaks that have substantial numbers of reads. In particular, these candidates are far more likely than some randomly selected region to be enriched in comparison with the control. In other words, by construction, the candidate peaks are very likely to not conform to some null hypothesis of no enrichment in ChIP-seq versus control. As such, we expect the output p-values to be unduly biased towards apparent statistical significance, even if there is actually no underlying difference between ChIP-seq and control data distributions.

For SICER [29, 28] there are similar sources of bias. SICER counts ChIP-seq reads in predefined, non-overlapping windows, and initially marks each window as eligible (statistically enriched compared to a uniform background model) or ineligible (not enriched). It then constructs islands out of a sequence of eligible windows interrupted by at most a fixed number of ineligible windows. It again tests these islands for enrichment compared to a uniform background model, discarding those that are not significantly enriched. Finally, the remaining islands, which constitute the candidate peaks, are assigned p-values by comparing ChIP-seq reads to control reads. The construction process twice biases attention towards regions enriched for ChIP-seq reads, so that the resulting candidates are highly likely to appear enriched versus control.

In diffReps [20], as in MACS, a fixed-size window is stepped across the genome, with low-count windows being discarded. (The exact definition of low-count is complex; we refer the reader to their paper for details.) For remaining windows, enrichment versus control is assessed by comparing ChIP-seq reads versus control reads by one of several possible statistical tests, and again, windows that are not enriched are discarded. Finally, any remaining windows that overlap are merged to form candidate peaks, and one final enrichment test of ChIP-seq versus control is performed to generate a p-value for that peak. In this case, two initial stages of selection—one versus a uniform model of sorts, and one versus the control itself—highly bias attention towards regions where ChIP-seq appears enriched versus control. Again, even if the underlying ChIP-seq and control distributions are the same, this focusing of attention on differences before the final p-value calculation can make it falsely appear as if there are regions of significant enrichment.

When peaks' p-values are wrong, it creates a host of other problems as a side effect. For one thing, we no longer have a good basis for choosing a p-value cut off for reporting results. Relatedly, we do not know how much we can trust any given peak, or even the set of peaks as a whole. If a peak has a p-value of 10^{-10}, we might feel that is very likely to be indicating true transcription factor binding or epigenetic modification. But if the peak caller is biased, so that the real statistical significance of such a peak is only 10^{-1}, then perhaps we should not put much stock in it after all. False discovery rate estimates, which are also reported by most peak callers, are virtually meaningless when based on p-values that are themselves incorrect. Another problem arises if we try to compare results from different peak callers. To make comparisons "fair", we might restrict both peak callers to the same raw p-value (or false discovery rate) cut-off. But if one algorithm has highly biased p-values and the other does not, then this comparison will hardly be fair.

One approach to unbiased peak-calling would be to develop a new peak calling approach from scratch, in a way that avoids double use of the data. However, given that there are already many programs available that seem largely satisfying in terms of identifying and ranking candidate peaks, and only their significance is in question, we chose a different approach. We asked whether the p-values of peaks generated by these programs could be recalibrated, to correct their bias. Happily, we found this to be largely possible through the new RECAP method that we introduce. RECAP stands both for the goal or our approach, recalibrating p-values, and the method by which it is done, resampling the read data and calling peaks again.

RECAP is a wrapper algorithm that is compatible with most any peak caller, and in particular MACS, SICER and diffReps, for which we provide wrapping scripts. RECAP repeatedly resamples from ChIP-seq and control data according to a null hypothesis mixture. It then applies the peak caller to the resampled data, estimating the distribution of p-values under the null hypothesis of no difference between ChIP-seq and control. It uses the CDF of that estimated distribution to adjust the p-values produced by the peak caller on the original (not resampled) data.

We show that on a variety of different types of null hypothesis ChIP-seq data, where there is no actual enrichment, this produces p-values that are approximately uniformly distributed between zero and one—as should be the case for well-calibrated statistical hypothesis testing. RECAP also allows local false discovery rate analysis. This means that for each peak, we can assess the likelihood that it is a true positive or a false positive, based on its p-value. This gives a more intuitive way of choosing a significance cut-off for peak calling, and allows us to look at whether default cutoffs (such as the 10^{-5} raw p-value cutoff in MACS) are overly conservative or still too loose. In summary, RECAP allows for much more rigorous and rational analysis of enrichment in ChIP-seq data, while allowing researchers to continue to use the peak calling algorithms they already prefer and have come to depend on.

2 Results

2.1 MACS, SICER and diffReps produce biased p-values

To test whether peak-callers produce biased p-values, we generated 10 simulated null hypothesis data sets. In each data set, both ChIP-seq and control data comprise foreground regions and background regions. Foreground regions are 500bp long and spread approximately 20-25kb apart along a hg38-sized genome, and are the same for both ChIP-seq and control. Each ChIP-seq and control data set had 30,882,698 reads—one per 100 basepairs of the genome on average. 10% of the reads were placed uniformly randomly within the foreground regions, while the remainder were placed uniformly randomly within the background regions. Figure 1A shows a zoom-in on part of one of the randomly generated ChIP-seq data sets and its matching control.

We ran MACS, SICER and diffReps on these data sets, using default parameters with one exception. We set p-value or FDR cut-off thresholds at or as close as possible to 1.0, so that all candidate peaks, regardless of significance, would be reported. Figure 1B shows histograms of the p-values of the peaks produced by each program, for one of the 10 simulated ChIP-seq–control data set pairs. These histograms show that the distributions of p-values are complex. They are rough with "spikes" of varying heights at different locations. This is especially visible for diffReps, but it is true of all three peak callers. Spikes in the distributions occur when multiple candidate peaks have the exact same numbers of ChIP-seq and control reads, so that the assessed statistical significance is exactly the same. The rough appearance of the distributions is observable at many scales. For example, if one restricts attention to peaks with significance $p \leq 10^{-5}$ or some other such conservative threshold, one continues to see irregularly spaced spikes of varying heights where multiple peaks have the exact same p-value.

In addition to being complex, the p-value distributions are also far from uniform (which they should be for this null hypothesis data, if p-values were well-calibrated). This is visually clear from Figure 1B, where the horizontal dashed lines indicate the uniform distribution, and from Figure 1C, where we plot the empirical cumulative distribution functions (CDFs) of the p-values of the three programs. Well-calibrated p-values should have empirical CDF close to the thin black diagonal line. Although we will momentarily introduce a different statistic for quantifying deviation from uniformity, a simple KS-test shows that the three p-value distributions of the programs are statistically significantly different from the uniform distribution ($p \approx 0$ incalculable small for all three).

Figure 1D shows the same empirical CDFs, but plotted on log-log axes. This plot is informative because most p-values are close to zero, and it is difficult to see their distribution on linear axes. Again, this plot shows that all three algorithms produce p-values that are optimistically biased compared to the expectation under a uniform distribution of p-values. But it is now much more clear that diffReps's p-values are the closest to being uniformly distributed, whereas MACS's and SICER's p-value distributions are farther afield. The curve for SICER, in fact, grows worse as p-value get smaller; SICER seems particularly prone to outputting highly significant p-values. Motivated by this log-log plot of empirical CDFs, we propose a measure of deviation from uniformity. For a given set of N p-values, we let N_1/N be the fraction of those p-values in the range $[0.1, 1]$, N_2/N be the fraction in the range $[0.01, 0.1)$, and more generally N_i/N be the fraction in the range $[10^{-i}, 10^{-i+1})$. Then we quantify deviation from uniformity by the statistic

$$D = \text{mean}_{i:N_i>0} |\log_{10}(N_i/N) - \log_{10}(9 \times 10^{-i})|$$

In words, this is the absolute difference between the logarithm of the number of peaks that should be in a p-value bin and the logarithm of the number of peaks that actually are in the bin, averaged over the non-empty bins. If a set of p-values is uniformly distributed on $[0, 1]$, so that 90% of them fall in $[0.1, 1]$, 9% fall in $[0.01, 0.1)$, etc., then D evaluates to zero. Non-uniform distributions produce higher values of D. An advantage of this measure compared, for example, to the statistic used by the KS-test is that it pays equal attention to p-values at many different significance levels. In contrast, the KS-test looks at the maximum difference between the empirical CDF and the theoretical uniform CDF. For the SICER data, for example, this maximum occurs at $p = 1$, where approximately 40% of the peaks are. But the peaks with such high p-values are not of any biological interest, so it is undesirable for a performance metric to emphasize them to the exclusion of all else. If, for instance, the SICER p-values below $p = 0.1$ were well-calibrated, then we would be quite happy to ignore any non-uniformity that occurs above $p = 0.1$. For the present data, the deviations of the three algorithms' p-value distributions evaluate to $D \approx 2.9$ for MACS, $B \approx 4.1$ for SICER, and $D \approx 0.8$ for diffReps.

Although we will quantify bias and its removal more thoroughly in the next section, several important points remain regarding biases in the p-values produced by these programs. First, our results are not an artifact of the precise way the simulated null hypothesis ChIP-seq and control data sets were generated. For example, we also generated data with similar foreground regions but with 20% of reads in the

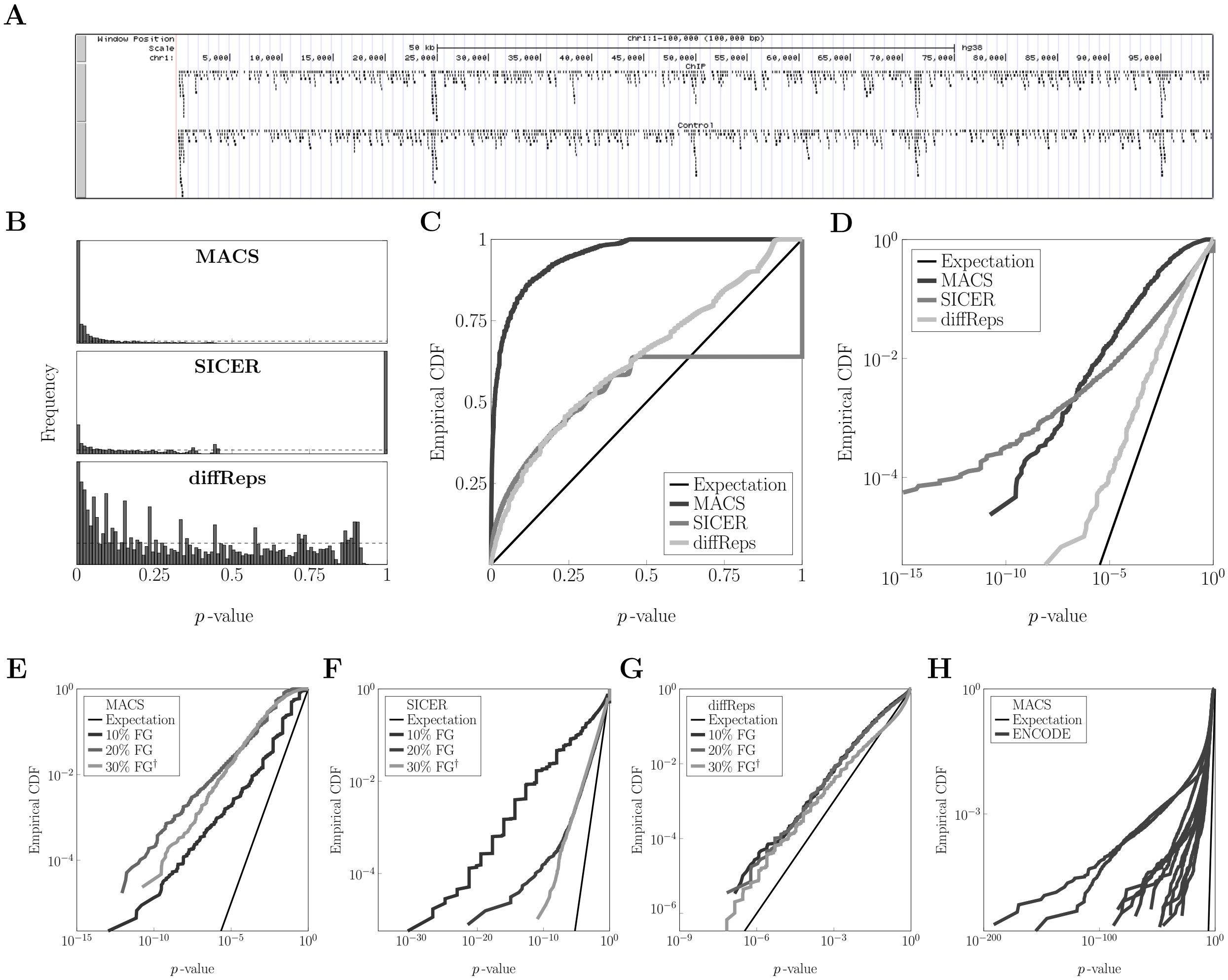

Figure 1: MACS, SICER and diffReps peak callers produce biased p-values. (A) Visualization of part of a simulated ChIP-seq read data set, with 500bp foreground regions every 20-25kbp, where read density is greater. Control data was generated similarly, with matching foreground regions, so a null hypothesis of no enrichment in ChIP-seq versus control is true for every possible genomic region. (B) Peaks called by the three algorithms have p-values that are not uniformly disributed between zero and one, as should be the case for this null hypothesis data if p-values were wellcalibrated. (C,D) Empirical cumulative distribution functions on linear (C) and log (D) axes also show the discrepancy from the uniform distribution. (E,F,G) Empirical CDFs when we vary the percentage of reads in the foreground and background (BG), continue to show bias, although the amount changes. (H) When using 24 real ChIP-seq data sets from the ENCODE project, matched as 12 pairs of replicate ChIP-seqs, we see even greater deviation from the uniform distribution (although this is only an approximation of null hypothesis data).

foreground and 80% in the background. We also generated data with broad foreground regions of 4kbp containing 30% of the reads, leaving 70% for the background. For these data sets, we run MACS in broad peak mode. In all cases, we continue to see deviation from uniformity in the p-value distributions (Figure 1E-G). Second, it is important to notice that the degree of bias in these p-value distributions (again, more careful quantification is coming in the next section) differs for the different types of data and for the different algorithms. This means that there is no universal correction that can be applied to the p-values, to bring them into line. That is, whatever way we can find to remove bias much operate in a way specific to the data being analyzed, and the program being used to call peaks.

Finally, it is important to note that evidence of bias can be seen in real data, not just simulated data. To show this, we turned to ChIP-seq data from the ENCODE consortium [6, 3]. Somewhat arbitrarily, we chose to analyze data sets from the K562 cell line, as this is the cell line for which the most data sets are available. We identified all experiment that included two replicate ChIP-seq experiments and two matching controls (there were 88 such) and arbitrarily chose the first dozen of these for analysis. In an attempt to approximate null hypothesis-like conditions, but using real data, we called peaks on each ChIP-seq data set using its ChIP-seq replicate as control. The resulting p-value CDFs for MACS specifically (the peak caller used by the ENCODE consortium) are shown in log-log format in Figure 1H. As with our simulated data, we see all the CDFs are optimistically biased, in some cases returning dramatic p-values reaching nearly 10^{-200}. Thus, we believe the p-value bias is not just some artificial theoretical concern, but a genuine concern that is observable and should be expected in the analysis of real data.

2.2 RECAP: A wrapper algorithm that removes bias from peak-caller p-values

Our approach to recalibrating p-values is based on empirically estimating an expected CDF for those p-values under a suitable null hypothesis. As shown above, that null hypothesis must be specific to the ChIP-seq and control data sets, as different data sets produce different distributions of p-values. And of course, the recalibration must by different for different peak-calling algorithms, as different algorithms produce different distributions of p-values for the same data. We put forth the null hypothesis that the ChIP-seq and control read data sets are drawn from the same distribution across the genome. That is, if we were to view each read as an i.i.d. sample where different positions on the genome would have different probabilities of being sampled, then we assume the sampling distribution of ChIP-seq and control are identical. Some work [4, 17] has explicitly attempted to estimate such distributions, but we will use a simpler mechanism for our p-value recalibration.

The RECAP algorithm is summarized below. From this point onward, we begin referring to the ChIP-seq data set as the “treatment” data. The reason for this is that the algorithm remixes ChIP-seq and control data into new data sets, and it would be confusing to call such remixed datasets by the name ”ChIP-seq” when really they contain control data. (That said, we continue to call the control data by that name, as we know of no commonly-used term that could take its place.)

The RECAP algorithm

- **Input:** Two read data sets T (treatment) and C (control), peak-calling algorithm A, and repeats number R
- **Call peaks:** Use algorithm A on data sets T and C, to generate peaks P with p-values $p = (p_1, p_2, \ldots, p_n)$
- **Model CDF of p-values under null hypothesis:**
  - Compute the union of all reads $U = T \cup C$
  - For $i = 1$ to R do:
    - ∗ Randomly divide U into mock treatment T_i and control C_i, with the same numbers of reads as T and C respectively
    - ∗ Call peaks using A on data sets T_i and C_i generating peaks with p-values $p^i = (p^i_1, p^i_2, \ldots, p^i_{n_i})$

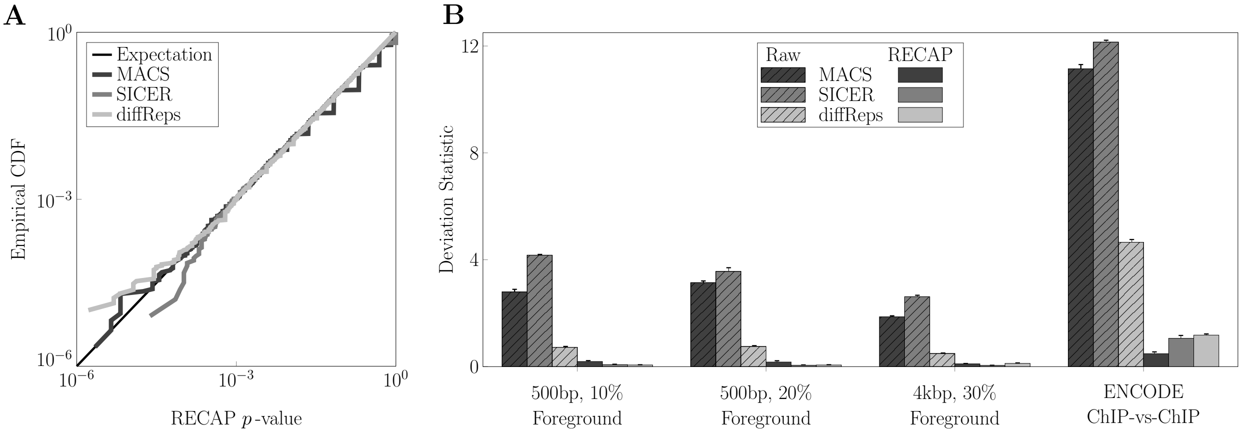

Figure 2: RECAP recalibrates peak callers' p-values to a near-uniform distribution. (A) Log-log plot of the empirical CDF of recalibrated p-values for MACS, SICER and diffReps, on the simulated, 10% foreground, null-hypothesis data. (B) Reductions in deviation statistic, which measured difference from uniform distribution, for the RECAP recalibrated p-values for several types of simulated data (10 data sets each) and 12 matched pairs of ENCODE replicate ChIP-seq data.

  - Combine all re-sampled p-values to estimate null CDF

$$F(x) = \frac{|(i,j) : p^i_j \leq x|}{\sum_i n_i}$$

- **Output:** Original peak set P with recalibrated p-values $p' = (F(p_1), F(p_2), \ldots, F(p_n))$

The intuition behind the algorithm is that if the null hypothesis holds, we can simulate new-but-similar treatment and control data sets by resampling from the combined reads of the original treatment and control. If we do that repeatedly, and call peaks each time, we can estimate an average-case distribution of p-values for similarly-distributed data. Implicitly, this approach makes several assumptions. One assumption is that there even exists some notion of p-value distribution, given by F, that can be estimated. In principle, it is possible that every resampling of the data would generate peaks with radically different p-values or produce no peaks at all. If this were true, the "average" p-value distribution would not exist or would not be meaningful as a point of comparison for the original p-values p. In preliminary testing of all three algorithms, we found that while the *numbers* of peaks called could vary considerably between different resamples (particularly for MACS), the distributions of p-values were largely the same. Furthermore, a peak caller that did generate wildly different p-values for similar data sets would probably not be considered a good algorithm, due to lack of robustness. Second, our method assumes that every peak's p-value in each of the R resamples can be viewed as i.i.d. samples from that distribution—justifying the standard empirical CDF estimate we use for F. In principle, because peak-calling relies in part on local read densities, it is possible for nearby peaks to have non-statistically independent p-values. However, because these dependencies typically do not span a large portion of the genome, we expect the independence assumption is reasonable.

We tested RECAP's ability to correct bias in peak p-values on a variety of simulated and real null hypothesis data sets. Figure 2A shows the results for the same 10%-reads, 500bp foreground region data set used for Figure 1B-D. Comparing particularly Figure 2A with Figure 1D, we see that RECAP has very substantially removed the bias. A quantitative assessment of bias before and after recalibration by RECAP for peaks with greater read density, broader peaks, and replicate ENCODE data sets is in Figure 2B. In all cases, we see that p-value distribution bias, as quantified by our deviation statistic D, is very substantially reduced. It is especially succesfully for the simulated datasets, which we know definitively do obey the null hypothesis of no enrichment between treatment and control. For the replicate ENCODE data sets, the imperfection correction may indicate some small degree of genuine differences between replicates. This would not be surprising, as the whole reason biological replicate ChIP-seq experiments are performed is because we know enriched regions can appear somewhat different in each experiment. Nevertheless, we claim that the results in Figure 2B show that RECAP successfully removes (or greatly reduced) p-value bias for a variety of types of data for MACS, SICER and diffReps.

2.3 RECAP enables local false discovery rate analysis

When peaks are called for non-null hypothesis treatment and control data—that is, for data where the treatment contains some genuine regions of enrichment compared to control—it is expected that some of those peaks will correctly reflect regions of enrichment (true positives), but some may not (false positives). Indeed, our results above with simulated null-hypothesis data show that the mere existence of called peaks does not imply any genuine regions of enrichment, regardless of the p-value. Nevertheless, we might expect that called peaks with smaller (i.e., more significant) p-values are more likely to be true postives than false positives. The methods of local false discovery rate analysis provide one way that this intuition can be formalized, so that we can estimate the probability of any called peak being a true positive versus a false positive.

To demonstrate this idea, we first generated non-null hypothesis simulated data for analysis. The treatment data had 10% of its reads randomly placed in 500bp foreground regions, which were randomly spaced throughout a simulated human genome 20-25kbp apart. The remaining 90% of reads were spread uniformly through the remainder of the genome. The control data, however, had no such foreground regions, and rather had 100% of its reads spread uniformly over the human genome. We applied MACS, SICER and diffReps to this data. Figure 3A shows the distribution of raw p-values for each of the algorithms. As we have seen before, many peaks are called with highly significant p-values, reaching approximately 10^{-15}, and the fractions of peaks found with such small p-values far exceeds what one would expect from null hypothesis data. Indeed, for SICER in particular, so many of its peaks are called with p-values in the range 10^{-15} to 10^{-5} that there are hardly any peaks with less significant p-values such as 10^{-2} or 10^{-1}.

We then appiled RECAP to remove bias from those p-values, with the results shown in Figure 3B. Unlike what we saw above, where RECAP brought the p-value distribution into line with the null hypothesis expectation (see Fig. 2), on this data even the recalibrated p-values are substantially enriched with values near zero (especially 10^{-6} to 10^{-4}). This, of course, is because of the genuinely-enriched regions in our simulated treatment data set. In Figure 3C, in the leftmost set of bars, we can confirm that although re-calibrating p-values substantially reduces their deviation from the null hypothesis expectation, there remains a significant deviation, due to the truly enriched regions.

Because we generated the treatment and control data ourselves, we know for each peak whether it overlaps a treatment foreground region (true positive) or does not (false positive). We divided the recalibrated p-values into half decades: 10^{0} to $10^{-0.5}$, $10^{-0.5}$ to 10^{-1}, 10^{-1} to $10^{-1.5}$, etc. In each of these bins we calculated the number of true positive and false positive peaks, and we calculated from those the empirical *local* false discovery rate—i.e. the number of false positives divided by the total number of peaks within that p-value bin. The results are shown Figure 3D. For all three algorithms, we see that peaks with the smallest p-values are almost entirely true positives, or equivalently, the local false discovery rate is nearly zero. However, as the p-value increases, the local false discovery rate increases. So for example, for MACS at a recalibrated p-value of approximately $10^{-2.5}$, about half of the peaks are true positives, but the other half are false positives. If we were analyzing real data, local false discovery rate information would be useful in telling us how much we should believe in any individual peak, and perhaps also for choosing an appropriate re-calibrated p-value cutoff for reporting peaks. However, for real data we cannot perform this same local false discovery rate computation, because we do not know what the truly enriched regions of the treatment data are. The question then becomes, how can we compute or estimate local false discovery rates when the ground truth is not known?

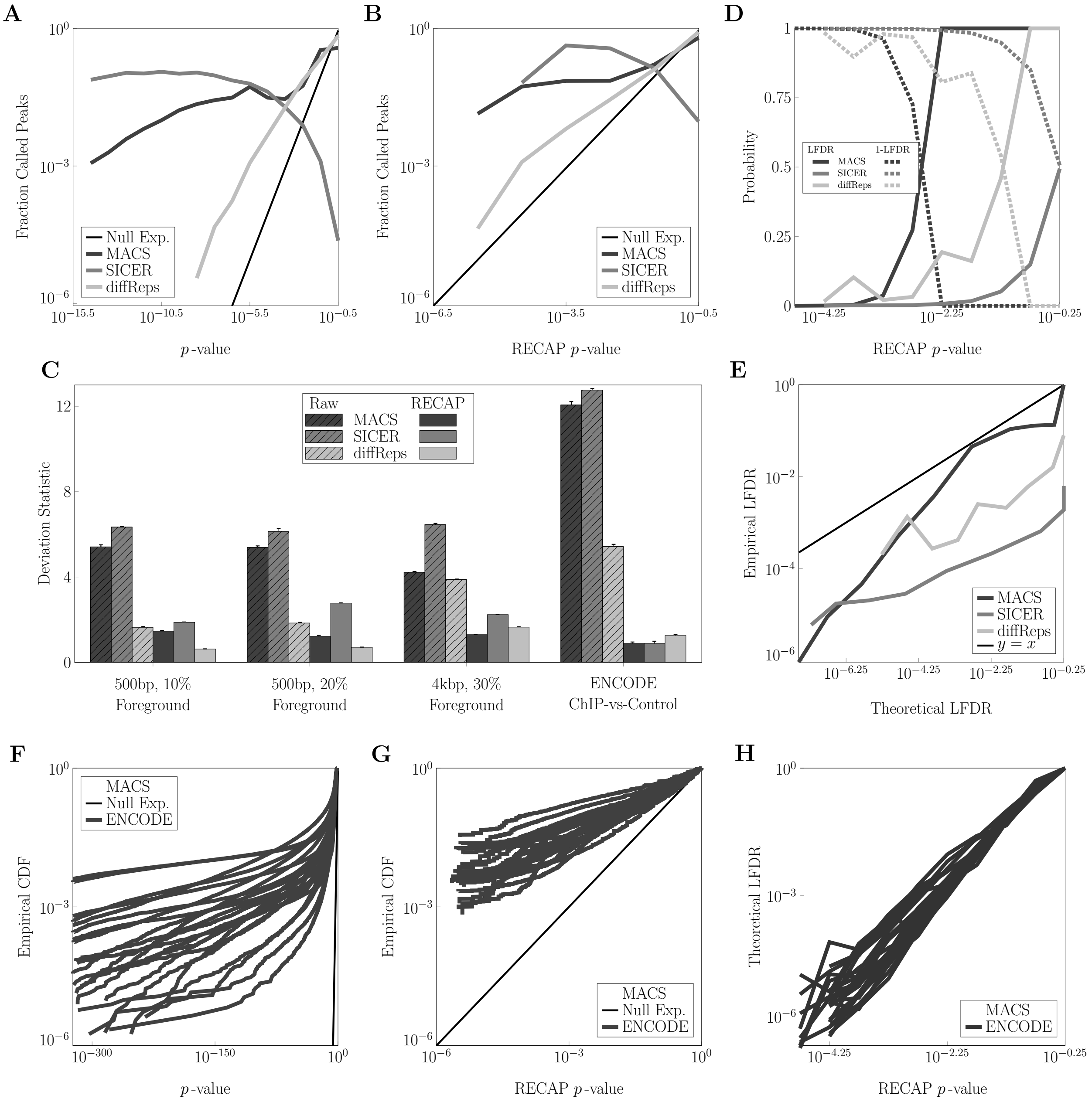

Figure 3: RECAP enables local false discovery rate analysis, which provides probabilities of peaks being true positives or false positives, based on their recalibrated p-values. (A) Fractions of peaks called at different p-value levels (binned into powers of 10), for MACS, SICER and diffReps, on 10% foreground data that includes genuine enrichment of treatment versus control. (B) Fractions of peaks found at different recalibrated p-values. (C) Deviation statistic measuring non-uniformity in raw and recalibrated p-values, for different types of data. (D) Empirical local false discovery rates at different p-value levels (binned into half powers of 10), for the three algorithms on 10% foreground data. (E) Comparison of empirical local false discovery rates with theoretically estimated rates, based on a Bayesian two-class analysis. (F,G) Empirical CDFs for raw (F) and recalibrated (G) p-values for 24 ENCODE Chip-seq dataset called against their matched controls. (H) Theoretical local false discovery rate estimates, based on RECAP recalibrated p-values.

To estimate local false discovery rates for an arbitrary pair of treatment and control data set, after peak calling and recalibrating p-values, we propose a two-class Bayesian approach that is well established in the statistical literature [7]. We present this approach in a formulation assuming our re-calibrated p-values have been grouped into N bins, with bin boundaries $p_0 = 0 < p_1 < p_2 < \ldots < p_{N-1} < p_N = 1$. We use $P_i = (p_{i-1}, p_i]$ to denote the i^{th} p-value bin. We view the total set of peaks as a mixture of true positives and false positives, with the a priori probability that a peak is a false positive being π_0, and the a priori proabibility of a peak being a true positive being $\pi_1 = 1 - \pi_0$. Further, we imagine that false positive peaks have p-values distributed according to some density f_0 on $[0, 1]$, while true positive peaks have p-values with a presumably-different distribution f_1 on $[0, 1]$. These densities imply the probability of a false positive peak having a p-value p in bin P_i, which we write as

$$P(p \in P_i | FP) = \int_{p'=p_{i-1}}^{p_i} f_0(p')dp' \ .$$

And similarly for the p-value of a true positive peak we can write

$$P(p \in P_i | TP) = \int_{p'=p_{i-1}}^{p_i} f_1(p')dp' \ .$$

If all these quantities, π_0, π_1, f_0, and f_1 where known, then for any given p-value bin P_i we could compute a local false discovery rate—that is, the chance that a peak with p-value p in that bin is a false positive, as follows:

$$\begin{aligned} LFDR(p) &= Pr(FP | p \in P_i) \\ &= \frac{Pr(p \in P_i | FP) Pr(FP)}{Pr(p \in P_i)} \\ &= \frac{Pr(p \in P_i | FP)\pi_0}{Pr(p \in P_i | TP)\pi_1 + Pr(p \in P_i | FP)\pi_0} \ . \end{aligned}$$

Lacking knowledge of the probabilities in this model, we can instead resort to a combination of theoretical and empirical approximations. First, if one believes, as we have tried to establish above, that in the case of null hypothesis data our false positive peaks' re-calibrated p-values are approximately uniformly distributed on $[0, 1]$, then we can estimate:

$$Pr(p \in P_i | FP) \approx p_i - p_{i-1} \ .$$

This helps with the numerator of our equation, and depends crucially on the recalibration of p-values by RECAP. The same approximation can not be made for raw, biased p-values.

For the denominator, we can estimate simply by the empirical fraction of peaks with p-values falling in a certain bin.

$$\frac{Pr(p \in P_i | FP)\pi_0}{Pr(p \in P_i | TP)\pi_1 + Pr(p \in P_i | FP)\pi_0} = Pr(p \in Pi) \approx \frac{n_i}{n} \ ,$$

where we have a set of n peaks called for our data, and n_i is the number of those peaks with p-values in bin P_i. With these approximations, our local false discovery rate estimate becomes

$$LFDR(p) = \frac{(p_i - p_{i-1})\pi_0 n}{n_i} \ .$$

This still leaves us without an estimate for the multiplicative factor π_0. Although more sophisticated approaches may be possible, here we recommend the simple expedient of ignoring the term (or equivalently setting $\pi_0 = 1$, which can be viewed therefore as an upper bound on the true local false discovery rate, or a good approximation in the case that many false positive peaks are called over all).

We applied this approach to estimate theoretical local false discovery rates, and compared them with the empirical local false discovery rates in our simulated data (Figure 3E). The results show that for this data set at least, the theoetical and empirical numbers are well-correlated across a wide range of p-values, with no more than an order of magnitude or two error in any p-value bin for any algorithm. Although this degree of error is not trivial, it is far better than raw p-values might lead one to believe, and shows for the first time that we can establish reasonable bounds on the probabilities of individual peaks being true positive or false positives.

We then sought to extend our results to the real ENCODE data, to see the effect of p-value recalibration and local false discovery rate estimation there. We used all 24 ChIP-seq data sets describe above (which happen to be in 12 pairs of two replicates) and 24 matched controls as specified by the ENCODE data portal. We focused exclusively on MACS for this analysis, as otherwise the combinations of so many datasets and peak callers becomes overwhelming. Figure 3F shows the empirical CDFs of raw p-values, which contain peaks of extraordinarily significant p-values. After recalibration of p-values by RECAP, the empirical CDFs come closer to a uniform distribution (Figure 3G), but are still not close to uniform—a result of genuine regions of ChIP-seq enrichment versus control. Finally, we computed local false discovery rates as a function of half-decade-binned recalibrated p-values, as described above (Figure 3H). These local false discovery rates range from approximately 10^{-6} for peaks with the smallest recalibrated p-values (near 10^{-5}) up to nearly one (i.e., almost all false positives) for peaks with clearly non-significant recalibrated p-values. Because these are real data sets, we lack ground truth, and so cannot say for certain how accurate the local false discovery rate estimates are. However, one interesting note we can make is that the default MACS raw p-value cut-off of 10^{-5} is recalibrated by RECAP to an average value of about $p = 0.0203$. Combined with our observation above that MACS on simulated data was generating empirical false discovery rates above 50% at a recalibrated p-value of $p = 10^{-2.5} \approx 0.0031$ (Figure 3D), this should at the very least give us pause. It is possible that MACS's default and seemingly stringent raw p-value cut off of 10^{-5} is not stringent enough to avoid significant numbers of false positive peaks.

3 Discussion

In this paper we have looked at the question of how statistically significant are peak calls in ChIP-seq data. We argued that, for various reasons, a range of peak-callers likely have optimistic biases built into them, such that the actual significance of called peaks is not clear. Using simulated null hypothesis data with different amounts of background noise and with either narrow or broad foreground regions—regions where read densities are higher than in the rest of the genome, but equivalently high in treament and control, so that there is no differential enrichment—we documented this optimistic bias in three widely-used peak-callers, MACS, SICER and diffReps. Also importantly, we showed that the amount of bias differs between algorithms and between data sets, so that there is no simple, universal correction that can be applied to correct the problem. With such miscalibration of p-values, we have no real, accurate knowledge of the statistical significance of any given peak, and, although this was not a focus of our paper, no way of comparing the significance of results from different approaches.

We then described RECAP, a wrapper algorithm that re-samples from the combined treatment and control data to estimate p-value distributions when a null hypothesis of no differential enrichment is true. RECAP uses that information to compute a data set-specific correction to peak p-values when peaks are called on the treatment versus the control. We showed that RECAP can virtually eliminate bias in p-values generated from null hypothesis data. In turn, this allowed us to contruct, for the first time for ChIP-seq peak-calling, an estimator of local false discovery rate—a Bayesian posterior probability (or bound) that any given peak represents a region of true enrichment or not, based on its re-calibrated p-value. We believe this approach will have profound implications for the assessment of the true statistical signficance of candidate enriched regions, for setting p-value thresholds for reporting enriched regions, and for comparing outputs of alternative programs. And, although we did not provide any details here, we should point out that local false discovery rates can be trivially converted into the (global) false discovery rates with which bioinformaticians are more familiar [7]. So, RECAP allows rigorous false discovery rate analysis for ChIP-seq peak calling as well. Software implementing our approach, and in particular RECAP wrapper scripts that work specifically with the inputs and outputs of MACS, SICER and diffReps, can be found on our lab website at www.perkinslab.ca.

While RECAP is a complete system as it stands, there are a number of possible directions for improvement. For one, we have made a very coarse approximation in the local false discovery rate calculation that the a priori probability of a false positive is close to or bounded by one. If for some algorithms or data sets, large numbers of false positive peaks are not output, despite calling peaks at a loose p-value threshold, then our approximation will tend to overestimate the local false discovery rate; in other words, we will be pessimistic about peaks being false positive as opposed to true positive. Another simplification we have made in our estimates is binning p-values into decades or half-decades for our frequency analyses. We deemed this straightforward and adequate for demonstrating the possibility of recalibrating p-values and local false discovery rate analysis. However, more sophisticated methods for density estimation, such as kernel-based or smoothing methods [26], might yield improvements in the approach. Finally, we have focused here on re-calibration of p-values where one treatment is compared against one control. It is becoming more of a standard practice, including in the ENCODE project in particular, to employ at least two biological-replicate ChIP-seqs and matching controls. Thus, expanding our framework to accommodate multiple treatment and control inputs is another important avenue for improvement.

Finally, although we have focused here on ChIP-seq peak-calling, it is entirely reasonable to think that similar problems with p-value calibration may occur in other areas of high-throughput data analysis. For example, this may occur in DNA variant-calling, where complex conditions of uni- or bi-directional read coverage or other types of pre-filtering are sometimes applied before candidate variants are tested statistically. This double-usage of the data, to both select hypotheses for testing and to compute significance for those hypotheses, is a recipe for biased p-values. Perhaps in such cases, a similar read-resampling scheme could be used to calibrate p-values output by different variant callers.

4 Acknowledgements

This work was supported in part by Discovery Grant number 328154-2014 from the Natural Sciences and Engineering Research Council of Canada (NSERC) to TJP.

References

[1] Anaïs F Bardet, Jonas Steinmann, Sangeeta Bafna, Juergen A Knoblich, Julia Zeitlinger, and Alexander Stark. Identification of transcription factor binding sites from chip-seq data at high resolution. *Bioinformatics*, 29(21):2705–2713, 2013.

[2] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[3] Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146):799–816, 2007.

[4] Alan P Boyle, Justin Guinney, Gregory E Crawford, and Terrence S Furey. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21):2537–2538, 2008.

[5] Susan E Celniker, Laura AL Dillon, Mark B Gerstein, Kristin C Gunsalus, Steven Henikoff, Gary H Karpen, Manolis Kellis, Eric C Lai, Jason D Lieb, David M MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

[6] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[7] Bradley Efron. Size, power and false discovery rates. *Annals of Statistics*, 35(4):1351–1377, 2007.

[8] Anthony P Fejes, Gordon Robertson, Mikhail Bilenky, Richard Varhol, Matthew Bainbridge, and Steven JM Jones. Findpeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008.

[9] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.

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

[11] Xin Feng, Robert Grossman, and Lincoln Stein. Peakranger: a cloud-enabled peak caller for chip-seq data. *BMC bioinformatics*, 12(1):139, 2011.

[12] Terrence S Furey. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

[13] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[14] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

[15] Aaron TL Lun and Gordon K Smyth. De novo detection of differentially bound regions for chip-seq data using peaks and windows: controlling error rates correctly. *Nucleic acids research*, 42(11):e95–e95, 2014.

[16] Peter J Park. Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.

[17] Parameswaran Ramachandran and Theodore J Perkins. Adaptive bandwidth kernel density estimation for next-generation sequencing data. In *BMC proceedings*, volume 7, page S7. BioMed Central, 2013.

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

[19] Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

[20] Li Shen, Ning-Yi Shao, Xiaochuan Liu, Ian Maze, Jian Feng, and Eric J Nestler. diffreps: detecting differential chromatin modification sites from chip-seq data with biological replicates. *PloS one*, 8(6):e65598, 2013.

[21] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. High-throughput sequencing for biology and medicine. *Molecular systems biology*, 9(1):640, 2013.

[22] Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavaré. Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1):299, 2009.

[23] Hendrik G Stunnenberg, Martin Hirst, International Human Epigenome Consortium, et al. The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell*, 167(5):1145–1149, 2016.

[24] Geetu Tuteja, Peter White, Jonathan Schug, and Klaus H Kaestner. Extracting transcription factor targets from chip-seq data. *Nucleic acids research*, 37(17):e113–e113, 2009.

[25] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nature methods*, 5(9):829–834, 2008.

[26] Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

[27] Haipeng Xing, Yifan Mo, Will Liao, and Michael Q Zhang. Genome-wide localization of protein-dna binding and histone modification by a bayesian change-point method with chip-seq data. *PLoS computational biology*, 8(7):e1002613, 2012.

[28] S. Xu, S. Grullon, K. Ge, and W. Peng. Spatial clustering for identification of chip-enriched regions (sicer) to map regions of histone methylation patterns in embryonic stem cells. *Methods in Molecular Biology*, 1150:97–111, 2014.

[29] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng. A clustering approach for identification of enriched domains from histone modification chip-seq data. *Bioinformatics*, 25(15), 2009.

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