Normalization of ChIP-seq data with control

Kun Liang and Sündüz Keleş
Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison, 53706
email: liang@biostat.wisc.edu

July 19, 2011

Abstract

Motivation: ChIP-seq has become an important tool for identifying genome-wide protein-DNA interactions, including transcription factor binding and histone modifications. In ChIP-seq experiments, ChIP samples are usually coupled with their matching control samples. Proper normalization of the ChIP and control samples is an important aspect of ChIP-seq data analysis.

Results: We have developed a parameter-free method for estimating the normalization factor between the ChIP and the control samples. Our method, named as NCIS (Normalization of ChIP-seq), can accommodate both low and high sequence depth datasets. We compare statistical properties of NCIS against existing methods in a set of diverse settings, where NCIS exhibits the smallest bias and mean squared error. In addition, we investigate the commonly used sample-swapping approach as a method of controlling FDR when the ChIP and control samples are unbalanced. This investigation reveals that the sample-swapping strategy works well when used in conjunction with a score normalization scheme for mildly unbalanced samples.

Availability: R code (www.r-project.org) for NCIS is available from authors' website (http://pages.cs.wisc.edu/∼kliang/NCIS/) or by request.

1. INTRODUCTION

Genome-wide protein-DNA interactions including transcription factor binding and epigenomic modifications play a crucial role in the programming of cell specific gene expression. Therefore, their

genome-wide mapping with the ChIP-seq (Chromatin immunoprecipitation followed by sequencing) technology can significantly advance our ability to understand biology and diagnose human diseases. ChIP-seq is now routinely used in many applications, e.g., Blow et al. 2010, Ramagopalan et al. 2010 and Smagulova et al. 2011.

In a ChIP-seq experiment, the DNA fragments from binding sites of a target protein or from sites of specific histone modifications are enriched through immunoprecipitation. These sites can be sharp point sources in transcription factor binding, or long and diffused regions in some histone modifications, or combination of both in RNA polymerase-DNA interactions (Park 2009). Then reads that are sequenced end(s) of millions DNA fragments are aligned to a reference genome to identify enrichment sites with over-abundance of reads. In ChIP samples, there are considerable number of fragments coming from non-specific "background" regions throughout the genome. Thus, the reads in a ChIP sample can be considered as a mixture of enrichment signal reads and background noise reads (Xu et al. 2010).

Early studies without the use of control samples have assumed uniform background read distribution when assessing the significance of enrichment sites (Robertson et al. 2007; Boyle et al. 2008). However, regions with high read counts do not necessarily contain enrichment sites. Many follow-up studies have shown that the distribution of reads is far from uniform and is affected by many factors, including GC content (Dohm et al. 2008; Kuan et al. 2011), mappability (Rozowsky et al. 2009), chromatin structure, and copy number variation (Vega et al. 2009), among others. The most effective approach to account for these known and other unknown biases is to include a matching control sample that is generated either from input DNA or by using non-specific antibody.

The ChIP and control samples usually are sequenced at different depths (total number of reads). A common strategy for making the samples "comparable" is to linearly scale according to the sequencing depth ratio. Because of the mixture nature of ChIP sample, it is reasonable to align/normalize only the background component of the ChIP sample with respect to the control sample. Thus, an appropriate normalization involves the estimation of the background read proportion (π_0) among ChIP sample reads and the corresponding ChIP/control normalization factor. The proper estimation of the normalization factor is important for finding weak enrichment sites, especially for those sites whose enrichment ratio is between the sequencing depth ratio and

the true normalization factor. Xu et al. (2010) has experimentally validated the existence of weak enrichment sites and showed that they are biologically meaningful.

The normalization factor is a critical parameter of most ChIP-seq data analysis programs that can utilize control samples. For example, CisGenome (Ji et al. 2008) and PeakSeq (Rozowsky et al. 2009) explicitly use the normalization factor to estimate p-values under Binomial distribution. MACS (Zhang et al. 2008), SPP (Kharchenko et al. 2008), and USeq (Nix et al. 2008), among many others, use the normalization factor to scale the control sample for comparison with the ChIP sample. Furthermore, many programs (MACS, SPP, SISSRs by Jothi et al. 2008 and others) estimate false discovery rate (FDR) using a sample-swapping method as follows. After computing an enrichment statistic for each non-overlapping region, the FDR can be estimated as $R_I(s)/R_C(s)$, where $R_I(s)$ and $R_C(s)$ are the numbers of enriched regions called on the control sample and the ChIP sample, respectively, using the same threshold s on the enrichment statistics. To make the statistics in ChIP and control samples comparable, a normalization factor is implicitly used in the FDR estimation. Thus, the normalization factor is a crucial parameter of enrichment site detection and error rate control in ChIP-seq data analysis.

Last but not least, the estimation of background read proportion π_0 is of scientific interest itself. π_0 can be viewed as an overall quality indicator which is related to the specificity of the antibody used in an experiment, experimental design, and other experimental protocols. We have observed that π_0 can vary from 0.3 to close to 1 in many ChIP-seq datasets. Unless the number of truly enriched regions is small, π_0 close to 1 indicates the scarcity of enrichment reads and the need for better antibody or protocol, or both.

Many ChIP-seq data analysis programs (CisGenome, SPP, PeakSeq and CCAT by Xu et al. 2010) have proposed methods for estimating the normalization factor; however, their performances under diverse set of settings have not been studied. Most of the above methods are intuitively appealing; but many rely on ad-hoc tuning parameters.

In this paper, we develop a parameter-free normalization method and compare it with existing methods through data-driven simulations. In addition, it has been shown that FDR estimation with the sample-swapping method is justifiable when the background component of ChIP sample is exchangeable with respect to the control sample, i.e., the samples are balanced (Xu et al. 2010).

However, ChIP-seq data are usually not balanced. Thus, we propose a new method to approximate FDR estimation when the ChIP and control samples are not balanced and show that our method is computationally simple yet more powerful than existing methods.

2. METHODS

2.1 Existing methods for estimating ChIP to control normalization factor

Suppose there are N_1 and N_2 uniquely aligned reads for the ChIP and the control samples, respectively. According to the signal-noise model proposed by Xu et al. (2010), the reads in the ChIP sample can be decomposed into $\pi_0 N_1$ background and $(1-\pi_0)N_1$ enriched signal reads. Then the correct ChIP/control ratio should be $r = \frac{\pi_0 N_1}{N_2}$. We will refer to r as the normalization factor in the rest of the paper.

To estimate the normalization factor, the commonly used set-up is to divide reference genome into non-overlapping bins of width w, numbered from 1 to m. Let n_{1i} and n_{2i} denote the total number of reads in the ith bin in the ChIP sample and the control sample, respectively, and $n_i = n_{1i} + n_{2i}$ denote the total number of ChIP and control reads for bin i. If the knowledge of which bins are within background regions ware given to us by an oracle, then a natural estimator of ChIP/control ratio would be

$$\hat{r} = \frac{\sum_{i \in B} n_{1i}}{\sum_{i \in B} n_{2i}}, \tag{1}$$

where B represents the index set of the background bins provided by the oracle. Each existing normalization factor estimation method employs a different approach for estimating B. Given that enrichment sites tend to have high read counts, bins with small total counts are more likely to belong to background. CisGenome sets bin width $w = 100$ bp and uses the bins with low total counts as background. Specifically, $B_w(t) = \{i : n_i \leq t\}$ and the total threshold t is set to 1. As implied by this definition, $B_w(t)$ depends on the choice of w and t. Another idea, similar in spirit but operating on the opposite direction, is to exclude bins with high read counts. SPP estimates the background regions by excluding highly "enriched" regions with a small p-value either in the ChIP sample or the control sample under uniformity assumption on the reads. Specifically, SPP sets $w = 1$ Kbp and $B = \{i : \min(p_{1i}, p_{2i}) > c\}$, where p_{1i} and p_{2i} are the Poisson p-values for testing whether the ith ChIP and control bin read counts are generated from an uniform background read

distribution, and the threshold c is set to 10^{-5}.

CCAT estimates B and the normalization factor in an iterative fashion where B is estimated based on reads from the positive strand and r is updated using reads from the negative strand through (1). More specifically, in the jth iteration, $B = \{i : n_{1i+} < \hat{r}^{(j)} n_{2i+}\}$, where $\hat{r}^{(j)}$ is the current estimate of the normalization factor which is initialized at the sequencing depth ratio and n_{1i+} and n_{2i+} are ChIP and control positive strand read counts in bin i with bin width $w = 1$ Kbp. The algorithm iterates till convergence.

A related method, PeakSeq, first defines enriched regions by using a certain threshold of FDR on the height of the ChIP sample read profile. Instead of using (1), PeakSeq then excludes a proportion (P_f) of bins that overlap with putative enrichment sites defined in the first step and utilizes the slope of linear regression of ChIP against control bin counts (with $w = 10$ Kbp) as the normalization factor.

All the above methods attempt to approximate the background region in some intuitive way; however they rely on tuning parameters which are set in an ad-hoc fashion. The suggested bin width w ranges from 100 bp to 10 Kbp. Utilized definitions of the background regions B depend on arbitrary thresholds on an array of parameters, e.g., total count, p-value, and FDR. The same procedure with different tuning parameters may lead to drastically different estimates. In an application of PeakSeq (Rozowsky et al. 2009), the estimates of the normalization factor changes from 1.24 to 0.96 when the exclusion proportion P_f changes from 0 to 1. Furthermore, there aren't any established guidelines for optimally setting the tuning parameters.

2.2 Estimating normalization factor: NCIS

We propose a parameter-free method named as NCIS for Normalization of ChIP-seq. Our method extends CisGenome's estimator by choosing the optimal value of bin width w and the threshold of total read counts t in a data-adaptive manner. In general, the smaller the total count threshold t, the more likely that bins with small total counts, i.e., $n_i \leq t$, are from background regions, and thus, the normalization factor estimated from (1) by treating these bins as background tends to have smaller bias. CisGenome sets t to 1, the smallest possible non-zero total count, so that bias can be minimized. On the other hand, using larger t will increase the size of $B_w(t)$ and reduce

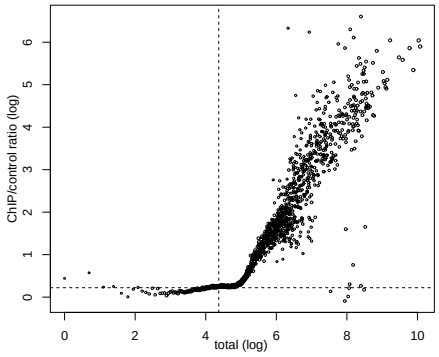

Figure 1: Marginal ChIP/control ratio against total count, both in natural log scale, from a yeast ChIP-seq dataset of transcription factor Ste12 (Zheng et al. 2010). Sizes of the plotting circles are proportional to the base 10 log of numbers of reads. Vertical dash line marks the median of the total counts. Horizontal dash line marks the normalization factor estimate from NCIS.

variance. Statistically, the choice of t represents the trade-off between bias and variance.

We next provide additional motivation for our method using ChIP-seq data from a yeast study (Zheng et al. 2010). Figure 1 shows the marginal ChIP/control ratio ($r_m(t) = \sum_{i:n_i=t} n_{1i} / \sum_{i:n_i=t} n_{2i}$) against the total count (t) with $w = 100$ bp for segregant 1 (SEG1) of a set of yeast ChIP-seq data of transcription factor Ste12 (Zheng et al. 2010). On the left half of the figure where t is small, the ratio estimates almost fall on a horizontal line and the variability increases as t gets small. This observation confirms that the reads from bins with small total counts are mostly from background regions and their marginal ChIP/control ratios are similar. There is an obvious ascent starting from a total count around 150 ($log(t) \approx 5$) which indicates the significant infusion of enrichment signal reads into the ChIP reads. Noticeably, there are some data points at the right bottom corner of the Figure 1 at the level of low total count ratio. These points arise as a result of the bins with both high ChIP and high control counts. The existence of such bins first confirms that the correct normalization factor should be at the level of the low count ratio; secondly, it indicates the importance of using a control sample because otherwise these regions are likely to be classified as enriched based on their high ChIP read counts alone.

Our method takes into account the above observations and operates as follows. First, all reads will be shifted towards their 3' end by $l/2$, where l is the average DNA fragment length

available either through experimental protocol or computational estimation (e.g., Zhang et al. 2008, Kharchenko et al. 2008 and Jothi et al. 2008). It is sufficient to use only nonempty bins, that is, we filter out bins with zero total count ($n_i = 0$). For any fixed bin width w, define $\hat{r}_w(t) = \sum_{i \in B_w(t)} n_{1i} / \sum_{i \in B_w(t)} n_{2i}$ as in (1) and $B_w(t) = \{i : n_i \leq t\}$ as in the previous subsection. We search for a total threshold t instead fixing it at a pre-specified constant. In most ChIP experiments, it is reasonable to believe that more than half of the genomic landscape are background regions. Thus, to avoid large variation in estimating $\hat{r}_w(t)$ when t and size of $B_w(t)$ are small, we start searching for t at the median of the non-zero total counts. Specifically, our estimate of the normalization factor with a fixed bin width w is $\hat{r}_w = \hat{r}_w(t_w^*)$ where $t_w^* = \min\{t : \hat{r}_w(t) \geq \hat{r}_w(t-1), |B_w(t)| \geq m_w/2\}$ and m_w is the total number of bins. That is, $\hat{r}_w$ is the first $\hat{r}_w(t)$ estimate that is larger than or equal to its previous one and is based on more than half of the bins.

It is reasonable to set the other tuning parameter, bin width w, close to the width of the enrichment site so that there are clear contrasts between the read counts of the ChIP and the control samples when the bins coincide with enrichment sites. However, without knowledge of the exact locations of enrichment sites and also in the settings where the lengths of enrichment sites vary, it is not possible to put the bin boundaries tightly around enrichment sites. As a result, enriched sites are likely to be split into two or more bins, and it is advantageous to use small bin width to gain resolution if the sequencing depth is high enough. Thus, we search over a grid of bin width $\{w_1, w_2, \ldots, w_n\}$ such that $w_1 < w_2 < \ldots < w_n$ and stop at the first bin width that satisfies $\hat{r}_{w_{i+1}} \geq \hat{r}_{w_i}$ and use $\hat{r}_{w_i}$ as our final estimate $\hat{r}$. That is, $\hat{r} = \hat{r}_{w_{i^*}}$ where $i^* = \min\{i : \hat{r}_{w_{i+1}} \geq \hat{r}_{w_i}\}$. Note that the values of $\hat{r}_{w_i}$ are bound to increase because the normalization factor equals to sequencing depth ratio (the upper limit of normalization factor) when w equals to the total genome size.

2.3 FDR control for unbalanced data

The purpose of normalization in ChIP-seq analysis is to make the ChIP and the control samples comparable. We define the ChIP and the control samples to be balanced when $\pi_0 N_1 = N_2$, or equivalently, $r = 1$. The balance can be judged in practice by checking whether $\hat{\pi}_0 N_1 = N_2$ after

obtaining $\hat{\pi}_0$, an estimate of π_0. The theory of the sample-swapping method for estimating FDR in ChIP-seq data analysis was studied by Xu et al. (2010) under a balanced setting. However, almost all of the ChIP-seq data exhibit imbalance. One strategy to deal with unbalanced data is to subsample the larger of the ChIP and control samples to achieve balance. This strategy has been advocated and practiced in Xu et al. (2010) and Smagulova et al. (2011), among others. The obvious drawback of the subsampling strategy is that part of the samples will not be utilized. To address this issue, Xu et al. (2010) further proposed to resample multiple copies of balanced data and merge the results so that all reads of the samples can have a chance to contribute to the final result. We hypothesize that by incorporating the normalization factor into significance score, the loss of data in the the subsampling strategy and the added computational complexity of resampling strategy can be avoided. Let $g(a_i, b_i, r)$ be a normalized significance score function based on ChIP count a_i, corresponding control count b_i of region i and normalization factor r when comparing the ChIP sample to the control sample. Define E as the collection of nucleotides at which ChIP signal is enriched. We now focus on the positive FDR (pFDR) which is proposed in Storey (2003) and can be defined in ChIP-seq context as

$$\text{pFDR}(s) = \Pr(i \subset \overline{E} | g(a_i, b_i, r) \geq s),$$

where s is a significance threshold.

***Theorem* 1.** Under the conditions:

(a) $\Pr(g(a_i, b_i, r) \geq s | i \subset \overline{E}) \approx \Pr(g(b_i, a_i, 1/r) \geq s | i \subset \overline{E})$ for large s and

(b) $\Pr(g(a_i, b_i, r) \geq s | i \not\subset \overline{E}) \gg \Pr(g(b_i, a_i, 1/r) \geq s | i \not\subset \overline{E})$,

the estimated pFDR at a threshold s can be approximated as

$$\text{pFDR}(s) = \frac{\#\{g(b_i, a_i, 1/r) \geq s\}}{\#\{g(a_i, b_i, r) \geq s\}}$$

for large s.

The first condition requires the normalized significance scores to have similar tail distributions in background regions of the ChIP and control samples. The second condition assumes good separation of the significance scores of $g(a_i, b_i, r)$ and $g(b_i, a_i, 1/r)$ when the region i is not entirely within background regions. When $r = 1$, the approximation in condition (a) is exact. We prove the

above theorem in the Supplementary Material and propose a normalized significance score based on Binomial p-values that works well in simulations when r is not far away from 1 (Section 3.2).

3. SIMULATION STUDIES

3.1 A comparison of statistical properties of normalization factor estimators

We used the yeast ChIP-seq data from the study of Zheng et al. (2010) as a base to simulate data. The yeast genome is about 250 times smaller than that of the human, and thus, high coverage can be easily achieved on yeast with relatively fewer reads. The control sample of segregant 1 (SEG1) has more than 4.2M uniquely aligned reads and is one of the deepest sequenced control samples in the study. With an average fragment length of 200 bp, 4.2M fragments amount to about 70X coverage on the yeast genome. As a comparison, 8M reads (average number of uniquely aligned reads from one lane on an Illumina GA sequencer) for a human sample corresponds to about 0.5X coverage. We randomly split the SEG1 control sample into two halves and subsampled $1/f$ of each. One of the subsamples was treated as control, and the other half was mixed in with simulated reads from p enriched sites and treated as a ChIP sample. Using a high coverage yeast dataset and a subsampling strategy, we can investigate the performance of methods under a spectrum of coverage. For example, the coverage achieved by 8M reads on a human sample corresponds to the coverage on a simulated yeast control sample with $f = 70$. As the cost of sequencing decreases rapidly, we can look into the "future" of ChIP-seq on large mammalian genomes by studying the performances of methods at small values of f.

We simulated reads for enriched regions in two different scenarios. Setting 1 mimics ChIP-seq data of transcription factors where enrichment reads concentrate in sharp peaks. Our set-up is similar to the simulation setting of Ji et al. (2008). More specifically, reads for enrichment sites were simulated from $N(\mu_i, \sigma^2)$ with μ_i randomly assigned along the genome and $\sigma^2 = 900$. The number of reads of each site followed an exponential distribution with mean $c \cdot N_2/p$ so that, on average, we spiked-in c times the total number of control sample reads (N_2). Parameter c, which represents the proportion of signal reads relative to the background was set to $0.2, 0.5$ and 1 to represent weak to strong overall binding signal strength. The value of p was set to 1000, based on the results of Zheng et al. (2010) which identified about 1000 binding sites for the transcription factor Ste12 in

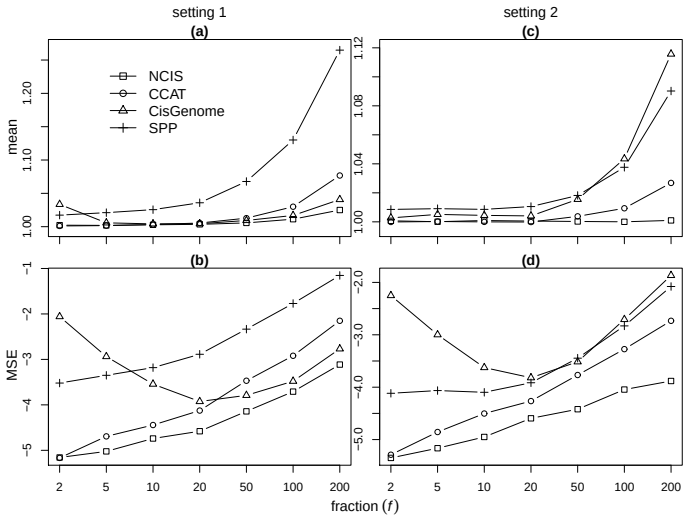

Figure 2: Mean and MSE (log10) for estimating the normalization factor in simulation setting 1 (a and b) and setting 2 (c and d) with c=1. The true value of the normalization factor is 1.

various strains of yeast. The subsampling fraction, f, took values in $\{2, 5, 10, 20, 50, 100, 200\}$ to represent different depths of coverage. The simulation was repeated 100 times for each combination of c and f.

In this simulation study, we compare our estimator (NCIS) with estimators proposed in CisGenome, SPP and CCAT. All the estimators compared were implemented as described in their respective papers. We did not include PeakSeq in our comparisons because it is computationally much intensive compared to other methods and it relies on species and read length specific mappability files. Figure 2a-b display the means and log10 of mean square error (MSE) for setting 1 (transcription factor binding) with c=1. The rest of the results (c = 0.2 and 0.5) for setting 1 are similar and are provided in the Supplementary Figure 1. The true normalization factor is always 1 in these simulations. Overall, our estimate has the smallest bias and MSE over most of the parameter settings. SPP estimator is generally the most conservative. CisGenome estimator has second best MSE when sequencing depth is low; however its performance deteriorates when sequencing depth is high. The performances of all the estimators except CisGenome improve with the increase of sequencing depth.

In setting 2, we simulated the enrichment reads to resemble histone modifications and polymerase binding where enrichment reads are spread out on large regions. We allocated enriched

reads uniformly on $p = 50$ regions. The length of each region was generated uniformly between 5–15 Kbp. The distribution of the number of reads of each region, signal/background proportion c and subsampling fraction f were the same as in above transcription factor set-up. Figure 2c-d display the results for $c = 1$ while the results for $c = 0.2$ and 0.5 are provided in the Supplementary Figure 2. Our method remains the best in terms of bias and MSE. CisGenome's estimator seems to have the worst MSE among all methods.

3.2 FDR control for balanced and unbalanced data

We first evaluate the impact of using different normalization factor estimators on the performance of sample-swapping method when ChIP-seq data is balanced. We focus on finding transcription factor binding sites as in setting 1. We employed a simple two-stage search strategy to find the binding sites in this simulation. In the first stage, we partitioned the yeast genome into 100 bp non-overlapping bins and retained only bins with Binomial p-value smaller than or equal to some loose threshold, like, 0.05. Then in the second stage, we merged nearby retained bins into putative regions and searched each region to locate the 20 bp bin with highest ChIP bin count and used the center of this bin as our prediction of the putative binding site. For each binding site, we extended 110 bp from the site location to both directions and formed a binding region. Then we used the ChIP and control read counts in the binding region to compute a Binomial p-value as the statistic for the binding site. This procedure was first performed with the ChIP sample versus the control sample to obtain a list of putative binding sites and their statistics, and repeated one more time with the control sample versus the ChIP sample to obtain a list of control binding sites and their statistics. Then a number of putative binding sites were declared as true binding sites such that the empirical FDR (#control sites/#ChIP sites) did not exceed certain nominal level (0.05) using the same p-value threshold. A site was classified as false positive if the predicted location was 100 bp away from its closest true binding site. Note that there were few declared binding sites located between 50 bp to 100 bp away from true binding sites, thus any choice between 50 bp and 100 bp would yield similar results.

Xu et al. (2010) performed simulations for FDR estimation of the sample-swapping method. Our simulation differs from theirs in two major ways. First, simulations in Xu et al. (2010) only

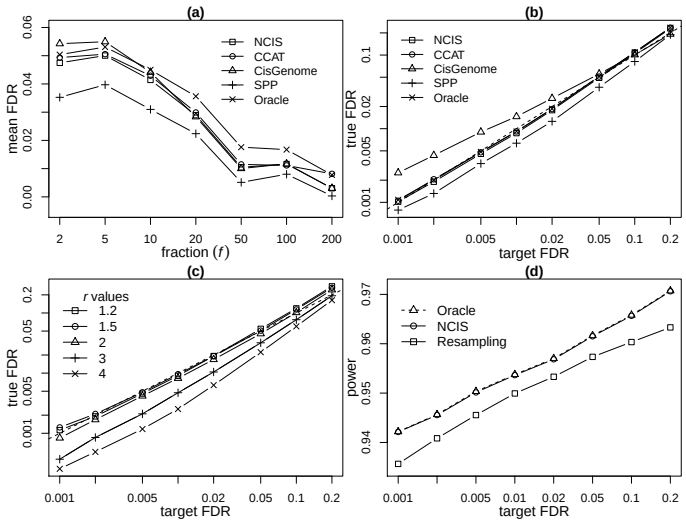

Figure 3: FDR control with the sample-swapping method under setting 1 and c=1. (a) and (b) compare different normalization factor estimators: (a) True FDR after FDR control at 0.05 level, (b) True FDR over a range of target FDR levels with $f = 2$. (c) Unbalanced ChIP-seq data with NCIS: dash line represents the 45 degree line, and different colored solid lines represent different true normalization factors. (d) Power comparison between normalized significance score and the resampling strategy for unbalanced data with a true normalization factor of 3.

evaluated the FDR estimation over a range of nominal FDR levels with a fixed sequencing depth, while our simulations evaluate FDR control over varying sequencing depths; Second, Xu et al. (2010) computed FDR on the basis of 1 Kbp non-overlapping regions, while we classified binding sites by their distance to their closest true binding sites. Our false positive criteria is more accurate and relevant because a single 1 Kbp region can hold more than one binding site and one binding site can be at the boundary of two 1 Kbp regions such that both regions would be regarded as true binding regions.

For each normalization factor estimator, the Binomial probability p when comparing the ChIP sample to the control sample is computed as $\hat{r}/(1+\hat{r})$, where $\hat{r}$ is their respective estimate of normalization factor. As a comparison, we also performed peak calling when the normalization factor is set to its true value of 1 and referred to this method as the Oracle. The mean and MSE of FDR for various methods are shown in Figure 3a. As we can see from the plot, all methods largely control FDR below the nominal level of 0.05 and show a trend of conservativeness when sequencing

depth is low. There are some evidence suggests FDR control with the CisGenome estimator can be liberal when sequencing depth is high. If we look closely at FDR control over a range of FDR thresholds in Figure 3b when $f = 2$ and $c = 1$, the true FDR using the NCIS, CCAT, and Oracle estimators are close to nominal levels while it becomes liberal with the CisGenome estimator and conservative with the SPP estimator. In general, conservative estimation of the normalization factor leads to conservative FDR control, which is the case for the SPP estimator. The CisGenome estimator is conservative when sequencing depth is high, but it also has high variance which leads to its anti-conservativeness of FDR control (Benjamini et al. 2006).

We next study FDR control using sample-swapping method for unbalanced ChIP-seq data. To generate unbalanced data, we first simulated data according to setting 1 and chose $f = 2$ and $c = 1$. We further subsampled $1/r$ of control sample such that the true normalization factor is r. We subsampled the control sample instead of the ChIP sample because the latter is commonly sequenced deeper in practice. We varied the values of r in $\{1.2, 1.5, 2, 3 \text{ and } 4\}$. We used Binomial p-value as the binding region significance score then utilized the sample-swapping method to control FDR. Figure 3c displays the true FDR versus target FDR for various d using the NCIS estimator. When the data is close to be balanced ($r \leq 2$), the empirical FDR stays close to the nominal levels. When the unbalance is large ($r = 3$ and 4), the sample-swapping method tends to be conservative. Using the Oracle yielded almost identical FDR estimates as using NCIS and is not shown here. Note that the performance of the sample-swapping method may depend on the normalized significance score used, and Binomial p-value seems tolerant to mild imbalance between the ChIP sample and the control sample.

To compare the performance of the sample-swapping method with the resampling strategy (Xu et al. 2010), we randomly split the ChIP sample into three equal size subsamples when $r = 3$ such that each subsample is balanced with regard to the control sample. Note that we implicitly assumed that the Oracle estimate of normalization factor ($r = 3$) was available to resampling strategy in this setup. In practice, many more resamples are needed because the estimated $\hat{r}$ is unlikely to be exactly an integer. Then we performed peak calling on each subsample and summarized their results by the median of their Binomial p-values as suggested in Xu et al. (2010). We show in Figure 3d that using normalized significance score on the original data results in better

power for detecting binding sites than using the resampling strategy on balanced subsamples. There is no distinguishable power difference between FDR control using the NCIS and Oracle estimators, and thus, the power loss in resampling strategy is not attributable to potential differences of the normalization factors used. Thus, we recommend using the simpler and more powerful normalized significance score in conjunction with the sample-swapping FDR control method over the more computationally intensive resampling strategy when ChIP-seq data is unbalanced.

4. APPLICATION

4.1 Yeast Ste12 data of Zheng et al. (2010)

We applied our method on the ChIP-seq data of yeast strain S96 in Zheng et al. (2010) and estimated π_0, the background proportion in ChIP sample, to be 0.558. The original analysis was preformed by MACS, which assumes the sequencing depth ratio as the normalization factor. To make results comparable with the original analysis, we modified the latest stable version of MACS (1.3.7.1) such that it can utilize user specified normalization factors through an additional input parameter. The estimated normalization factor (r), background proportion (π_0) and number of detected binding sites under two different criterion are listed in Table 1. The first criteria "#peaks by p-value" refers to the number of peaks detected by MACS with its default p-value threshold of $1e$-5, and the second criteria "#peaks by FDR" refers to the number of peaks detected by MACS with FDR controlled at 0.005 level. The other parameters of MACS were set to be the same as in the original analysis. Using different normalization factors has a dramatic impact on the power to detect binding sites and the estimation of FDR. This is because it is difficult to call peaks in ChIP sample but relatively easy to do so in control sample with a conservative normalization constant such as the sequencing depth ratio. For example, MACS detected 1489 peaks with an estimated FDR of 0.028, while the FDR for the most significant 1489 peaks was estimated as 0.0047 using the NCIS estimate. The 6 fold difference in the estimated FDR was caused by less then 2 fold difference in the normalization factor estimates. A recent paper, Smagulova et al. (2011), also pointed out that MACS overestimated FDR 7.5 fold in their study, and that it is highly likely that the incorrect normalization factor is the major contributing factor. CisGenome's estimate of normalization factor is the most conservative besides MACS. This is because the S96 strain was

Table 1: Comparison of methods on yeast strain S96 through the MACS algorithm.

	$\hat{r}$	$\hat{\pi}_0$	#peaks by p-value	#peaks by FDR
NCIS	2.466	0.558	1980 (0.007)	1697
CCAT	2.495	0.565	1968 (0.007)	1671
SPP	2.901	0.656	1771 (0.011)	993
CisGenome	3.367	0.762	1516 (0.024)	195
MACS	4.419	1	1489 (0.028)	180

The fourth column contains the numbers of detected peaks with p-value $\leq 1e$-5; corresponding estimated values of FDR are in parentheses. The fifth column contains the numbers of detected peaks at FDR level 0.005.

deeply sequenced and as a result, there are only 131 bins whose ChIP and control read count total equal to 1. Thus, CisGenome's estimate is expected to be highly variable due to the small sample size when sequencing depth is high. This is consistent with our observation in Figure 1 and the simulation results.

To further illustrate the differences between different normalization factors, we plotted the ChIP versus the control bin counts with bin width $w = 200$ bp in Figure 4. In this plot, different grey scales indicate different densities of bins which are annotated at the right-hand side. There are many bins with relatively high ChIP counts due to the enrichment signal. The slope of the upper line is the sequencing depth ratio, and majority of bins (80%) appear below this line. We should expect less than 50% of bins to appear below the normalization factor line because binding regions have smaller than 0.5 probability to exhibit a ChIP count/control count ratio below the normalization factor. The NCIS normalization factor is represented by the lower line, which passes right through the densest area of bins and has 33% of bins below the line.

The NCIS and CCAT estimates are similar for this dataset and are supported by Figure 4. This is consistent with our simulation results which suggest that the estimates from the two methods are close to each other when sequencing depth is high, but their performances in lower sequencing depths can be significantly different. To investigate this, we randomly subsampled 1/100 of the S96 ChIP sample reads and 1/50 of the S96 control sample reads 100 times to compare the two

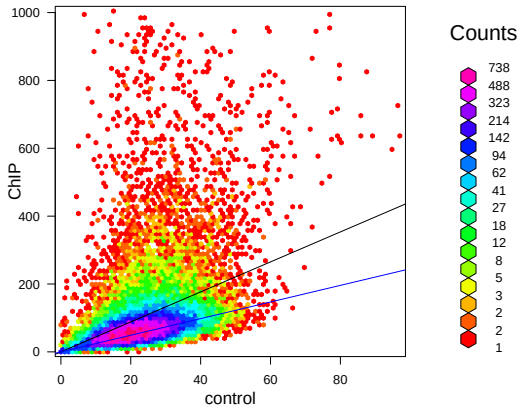

Figure 4: ChIP versus control bin counts for yeast strain S96 with bin width of 200 bp. The upper line represents the sequencing depth ratio, and the lower line the NCIS normalization factor estimate.

estimators in term of precision and the power to detect binding sites. The subsampling fraction is chosen such that the coverage in subsamples are roughly equal to the coverage of 15.4M of ChIP and 7.0M of control reads on the human genome. If we take the average of NCIS and CCAT estimates of r as the "truth", the normalization factor for subsampled data should be $(2.466 + 2.495)/4 = 1.24$. The means of NCIS and CCAT estimates are 1.42 and 1.55, respectively, across 100 replicates of this computational experiment. In addition, the sample variance of CCAT estimates is four times larger than that of the NCIS estimates. The imprecision of CCAT estimator under low sequencing depth leads to 5–14% loss in detection power compared to adopting NCIS estimates when running MACS with various FDR thresholds (see Supplementary Material for details).

4.2 Human NFκB data of Kasowski et al. (2010)

We next compare different normalization estimators on a human NFκB ChIP-seq dataset in Kasowski et al. (2010), where the genome-wide binding of transcription factor NFκB was extensively studied on multiple cell lines. As one of the deepest sequenced cell lines among the data collected, cell line GM12878 has 48.5 and 24.8 million uniquely mapped reads in the ChIP and the control samples, respectively. Table 2 shows estimates of normalization factor from different methods.

Figure 5 displays the marginal ChIP/control ratio against total read counts. We observe that the NFκB data is somewhat noisy compared to the yeast data in Section 4.1 and exhibits violations

Table 2: Comparison of normalization factor estimators on NFκB ChIP-seq data of cell line GM12878

	NCIS	CCAT	SPP	CisGenome
$\hat{r}$	1.743	1.657	1.752	2.123
$\hat{\pi}_0$	0.888	0.844	0.893	1.082

The sequencing depth ratio is 1.963.

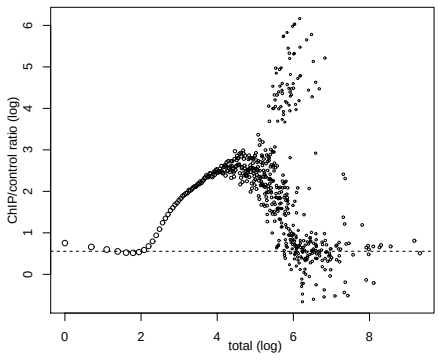

Figure 5: NFκB marginal ChIP/control ratio against total, both in natural log scale. Sizes of the plotting symbols are proportional to the base 10 log of number of reads. Horizontal dash line is NCIS estimate of the normalization factor.

of the signal-noise model (Xu et al. 2010) assumption. That is, some bins have larger control reads than expected as illustrated on the right bottom corner of the plot. This phenomenon can arise due to various artifacts in the ChIP-seq experiments, for example, PCR over-amplification in control sample. The CCAT estimator is susceptible to such artifacts and can have downward bias in estimating the normalization factor. On the other hand, NCIS and CisGenome only utilize bins with low total counts and are robust to such artifacts. SPP is also robust to these artifacts to some degree due to its filtering of bins with large ChIP and control read counts. In this dataset, CisGenome's estimate of normalization factor is larger than the sequencing depth which is an unreasonable outcome for the normalization factor.

5. DISCUSSION

As the sequencing technology improves rapidly over time, deeply sequenced data sets will become more common. We demonstrated in our simulation and application studies that CisGenome estimator's performance deteriorates when sequencing depth increases. In one unpublished and deeply sequenced *E.coli* dataset (Courtesy of Professor Tricia Kiley, UW Madison), we observed that CisGenome estimator was not applicable because every mappable bin had more than one read. We showed in the human NFκB application that CCAT estimator can exhibit downward bias due to artifacts in the experiment and violation of its signal-noise model assumptions. The implementation by CCAT itself preprocesses the data such that at most one read is kept on each nucleotide per strand, and thus, alleviates the problem of such artifacts on the single nucleotide level. However, CCAT estimator will still be susceptible to other sources of artifacts when control sample can have larger read counts than ChIP sample in some regions. In addition and more importantly, this preprocessing step makes the original CCAT estimator implementation unsuitable for experiments that are deeply sequenced or with strong transcription factor binding sites, where multiple reads on the same nucleotide are expected. All the existing normalization factor estimators rely on some ad-hoc tuning parameters, which may be crucial to the final estimate. Our NCIS method is parameter-free and has fewer bias and smaller MSE than existing methods over wide range of sequencing depths, and for both sharp or diffused ChIP signals.

ACKNOWLEDGEMENT

The authors are grateful to Professor Kiley (Department of Biomolecular Chemistry, UW Madison) for sharing her data to test and evaluate our method on a deeply sequenced ChIP-seq dataset.

Funding: This work was supported by the National Institute of Health [HG03747 to S.K. in parts].

REFERENCES

Benjamini, Y., Krieger, A., and Yekutieli, D. (2006), "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, 93, 491–507.

Blow, M., McCulley, D., Li, Z., Zhang, T., Akiyama, J., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010), "ChIP-Seq identification of weakly conserved heart enhancers," *Nature genetics*, 42, 806–810.

Boyle, A., Guinney, J., Crawford, G., and Furey, T. (2008), "F-Seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics*, 24, 2537.

Dohm, J., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008), "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic acids research*, 36, e105.

Ji, H., Jiang, H., Ma, W., Johnson, D., Myers, R., and Wong, W. (2008), "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, 26, 1293–1300.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008), "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data," *Nucleic acids research*, 36, 5221.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S., Habegger, L., Rozowsky, J., Shi, M., Urban, A., et al. (2010), "Variation in transcription factor binding among humans," *Science*, 328, 232.

Kharchenko, P., Tolstorukov, M., and Park, P. (2008), "Design and analysis of ChIP-seq experiments for DNA-binding proteins," *Nature biotechnology*, 26, 1351–1359.

Kuan, P., Chung, D., Pan, G., Thomson, J., Stewart, R., and Keleş, S. (2011), "A statistical framework for the analysis of ChIP-Seq data," Tech. Rep. 1151, University of Wisconsin, Madison, Department of Statistics, University of Wisconsin, Madison, `http://works.bepress.com/sunduz_keles/19`, To appear in the *Journal of the American Statistical Association*. Accepted May 2011.

Nix, D., Courdy, S., and Boucher, K. (2008), "Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks," *BMC bioinformatics*, 9, 523.

Park, P. (2009), "ChIP–seq: advantages and challenges of a maturing technology," *Nature Reviews Genetics*, 10, 669–680.

Ramagopalan, S., Heger, A., Berlanga, A., Maugeri, N., Lincoln, M., Burrell, A., Handunnetthi, L., Handel, A., Disanto, G., Orton, S., et al. (2010), "A ChIP-seq defined genome-wide map of vitamin D receptor binding: Associations with disease and evolution," *Genome research*, 20, 1352.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007), "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature methods*, 4, 651–657.

Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. (2009), "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nature biotechnology*, 27, 66–75.

Smagulova, F., Gregoretti, I., Brick, K., Khil, P., Camerini-Otero, R., and Petukhova, G. (2011), "Genome-wide analysis reveals novel molecular features of mouse recombination hotspots," *Nature*, 472, 375–378.

Storey, J. (2003), "The positive false discovery rate: A Bayesian interpretation and the q-value," *Annals of Statistics*, 31, 2013–2035.

Vega, V., Cheung, E., Palanisamy, N., and Sung, W. (2009), "Inherent signals in sequencing-based chromatin-immunoprecipitation control libraries," *PLoS One*, 4, e5241.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C., Lin, F., and Sung, W. (2010), "A signal–noise model for significance analysis of ChIP-seq with negative control," *Bioinformatics*, 26, 1199.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., et al. (2008), "Model-based analysis of ChIP-Seq (MACS)," *Genome biology*, 9, R137.

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L., and Snyder, M. (2010), "Genetic analysis of variation in transcription factor binding in yeast," *Nature*, 464, 1187–1191.

Supplement Materials for

Normalization of ChIP-seq data with control

Kun Liang and Sündüz Keleş

1 Proof of Theorem 1

$$\begin{aligned}\text{pFDR}(s) &= \text{P}(i \subset \overline{E} | g(a_i, b_i, r) \geq s) \\ &= \frac{\text{P}(g(a_i, b_i, r) \geq s | i \subset \overline{E}) \text{P}(i \subset \overline{E})}{\text{P}(g(a_i, b_i, r) \geq s)} \\ &\approx \frac{\text{P}(g(b_i, a_i, 1/r) \geq s | i \subset \overline{E}) \text{P}(i \subset \overline{E})}{\text{P}(g(a_i, b_i, r) \geq s)},\end{aligned}$$

where the last step is by condition (a). From condition (b), we also have

$$\begin{aligned}&\frac{\text{P}_{(g(b_i,a_i,1/r)\geq s | i \not\subset \overline{E})} \text{P}_{(i \not\subset \overline{E})}}{\text{P}_{(g(a_i,b_i,r)\geq s)}} \\ \ll\ &\frac{\text{P}_{(g(a_i,b_i,r)\geq s | i \not\subset \overline{E})} \text{P}_{(i \not\subset \overline{E})}}{\text{P}_{(g(a_i,b_i,r)\geq s)}} \leq 1\end{aligned}$$

Thus,

$$\begin{aligned}\text{pFDR}(s) &\approx \frac{\text{P}(g(b_i, a_i, 1/r) \geq s | i \subset \overline{E}) \text{P}(i \subset \overline{E})}{\text{P}(g(a_i, b_i, r) \geq s)} \\ &\quad + \frac{\text{P}(g(b_i, a_i, 1/r) \geq s | i \not\subset \overline{E}) \text{P}(i \not\subset \overline{E})}{\text{P}(g(a_i, b_i, r) \geq s)} \\ &= \frac{\text{P}(g(b_i, a_i, 1/r) \geq s)}{\text{P}(g(a_i, b_i, r) \geq s)},\end{aligned}$$

which can be approximated as

$$\frac{\#\{g(b_i, a_i, 1/r) \geq s\}}{\#\{g(a_i, b_i, r) \geq s\}}$$

2 Proposed Normalized Significance Score

Binomial p-value

$$g(a, b, r) = -\log(\text{P}(x \geq a))$$

assuming $x \sim \text{Binomial}\left(a + b, \frac{r}{1+r}\right)$. Binomial p-value has been used in Ji et al. (2008) and Rozowsky et al. (2009).

3 Simulation Results for the Normalization Factor

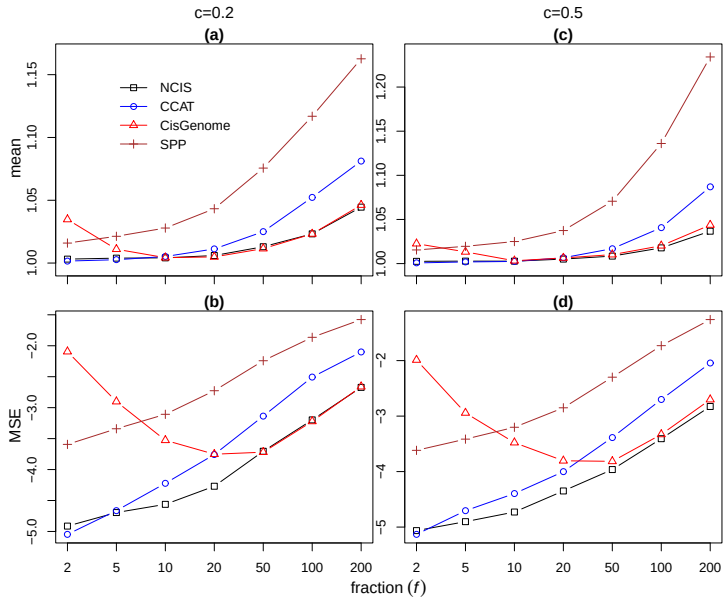

Figure 1: Mean and MSE (log10) for simulation setting 1 with c=0.2 (a and b) and c=0.5 (c and d). The true value of the normalization factor is 1.

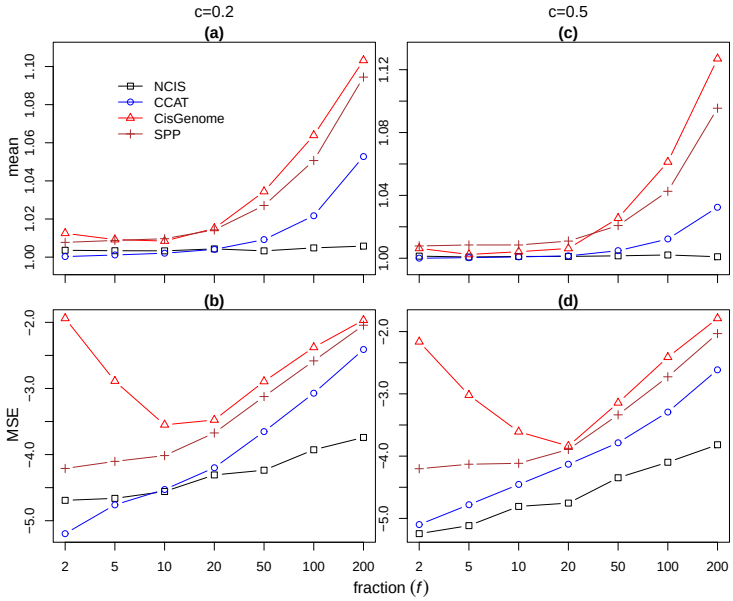

Figure 2: Mean and MSE (log10) for simulation setting 2 with c=0.2 (a and b) and c=0.5 (c and d). The true value of the normalization factor is 1.

4 Power Comparison between NCIS and CCAT when Sequencing Depth is Low

We subsampled 100 times the ChIP-seq data of yeast strain S96 in Zheng et al. (2010) and ran MACS (Zhang et al., 2008) with NCIS and CCAT estimates as described in Section 4.1. The conservative set of 1489 binding regions identified by MACS on the original data were treated as "gold standard". Figure 3 displays the average numbers of detected peaks that overlap with the gold standard by running MACS with the NCIS estimator versus with the CCAT estimator across various FDR thresholds.

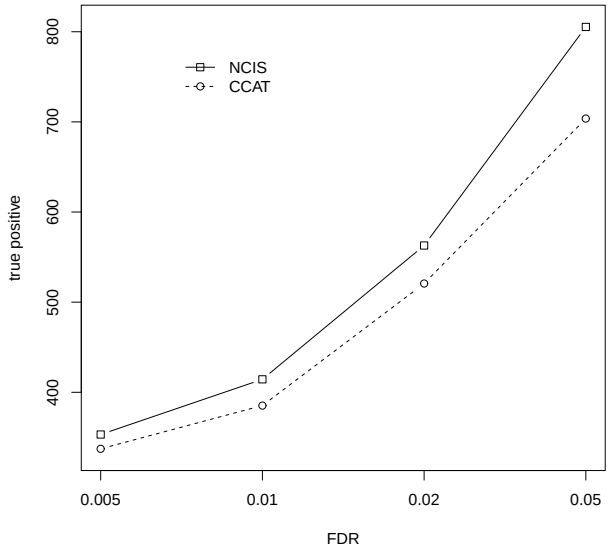

Figure 3: Average true positives resulted from running MACS with the NCIS and CCAT estimators across various FDR thresholds.

References

Ji, H., Jiang, H., Ma, W., Johnson, D., Myers, R., and Wong, W. (2008), "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, 26, 1293–1300.

Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. (2009), "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls," *Nature biotechnology*, 27, 66–75.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W., et al. (2008), "Model-based analysis of ChIP-Seq (MACS)," *Genome biology*, 9, R137.

Zheng, W., Zhao, H., Mancera, E., Steinmetz, L., and Snyder, M. (2010), "Genetic analysis of variation in transcription factor binding in yeast," *Nature*.