BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 753–760
doi:10.1093/bioinformatics/btt200

Genome analysis

Advance Access publication May 10, 2013

A statistical framework for power calculations in ChIP-seq experiments

Chandler Zuo[1] and Sündüz Keleş[1,2,*]

[1]Department of Statistics, and [2]Department of Biostatistics and Medical Informatics, 1300 University Avenue, Madison, WI 53706, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: ChIP-seq technology enables investigators to study genome-wide binding of transcription factors and mapping of epigenomic marks. Although the availability of basic analysis tools for ChIP-seq data is rapidly increasing, there has not been much progress on the related design issues. A challenging question for designing a ChIP-seq experiment is how deeply should the ChIP and the control samples be sequenced? The answer depends on multiple factors some of which can be set by the experimenter based on pilot/preliminary data. The sequencing depth of a ChIP-seq experiment is one of the key factors that determine whether all the underlying targets (e.g. binding locations or epigenomic profiles) can be identified with a targeted power.

Results: We developed a statistical framework named CSSP (ChIP-seq Statistical Power) for power calculations in ChIP-seq experiments by considering a local Poisson model, which is commonly adopted by many peak callers. Evaluations with simulations and data-driven computational experiments demonstrate that this framework can reliably estimate the power of a ChIP-seq experiment at different sequencing depths based on pilot data. Furthermore, it provides an analytical approach for calculating the required depth for a targeted power while controlling the false discovery rate at a user-specified level. Hence, our results enable researchers to use their own or publicly available data for determining required sequencing depths of their ChIP-seq experiments and potentially make better use of the multiplexing functionality of the sequencers. Evaluation of power for multiple public ChIP-seq datasets indicate that, currently, typical ChIP-seq studies are powered well for detecting large fold changes of ChIP enrichment over the control sample, but they have considerably less power for detecting smaller fold changes.

Availability: Available at www.stat.wisc.edu/~zuo/CSSP.

Contact: keles@stat.wisc.edu

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

Received on December 4, 2012; revised on April 19, 2013; accepted on April 22, 2013

1 INTRODUCTION

Next-generation sequencing technologies produce tens of millions of sequence reads during each instrument run and are used to answer questions central to human diseases. Multiple NIH consortia (ENCODE, modENCODE, 1000 Genomes, Roadmap Epigenome) are pursuing mapping of transcription factor (TF) binding and epigenome in multiple tissues and developmental stages with ChIP-seq applications (Gerstein *et al*., 2010; Kasowski *et al*., 2010; McDaniell *et al*., 2010; Myers *et al*., 2011; Roy *et al*., 2010). Analysis of ChIP-seq experiments involves comparing sequence reads from a ChIP sample to an appropriate control sample (e.g. chromatin input) to identify genomic loci/regions that exhibit enrichment in the ChIP sample compared with the input sample. Although there are >30 algorithms for analyzing data from ChIP-seq experiments (reviewed in Wilbanks and Facciotti, 2010), there has been little and mostly empirically driven efforts on the design of these experiments (Ho *et al*., 2011). Identification of biologically interesting genomic regions can be hindered by background noise. Detection of these regions can be improved by sequencing more reads. The total number of reads from a sequencing experiment is referred to as the sequencing depth. Sequencing depths so far have been set empirically because of lack of a formal statistical framework, e.g. the ENCODE Consortium suggested using a minimum sequencing depth of 20 million (M) mapped reads for sequence-specific TFs (Myers *et al*., 2011); however, Chen *et al*. (2012) recently concluded with empirical studies that the regularly adopted sequencing depth of 15–20 M reads in humans may not be high enough.

Kharchenko *et al*. (2008) and Rozowsky *et al*. (2009) explored the impact of sequencing depths of ChIP samples using saturation analysis. This analysis evaluated the effect of sequencing depth on the number of peaks discovered by identifying peaks from reads sub-sampled at varying proportions from the original ChIP sample. The proportion of sub-sample peaks that overlap the peaks from the full set is plotted against the sub-sample depth. When this curve reaches a horizontal asymptote, it indicates that the set of detected enrichment sites has stabilized at the current depth. Although this computational approach is useful for evaluating the available sequencing depth of a ChIP sample, it has three major drawbacks: (i) it is not suited for addressing how many more reads are needed if saturation has not been reached at the available depth [e.g. in a recent ENCODE publication (Myers *et al*., 2011), RNA Pol II, which mainly interacts with DNA across genes, exhibited a nearly linear gain in the number of peaks through 50 M reads with no indication of how many more reads are needed for saturation]; (ii) it only evaluates saturation based on the ChIP sample and discards the control sample; and (iii) it only allows investigating saturation from the point of either a minimum fold-enrichment or false discovery rate (FDR), but not both.

*To whom correspondence should be addressed.

Addressing the question of sequencing depth requires (i) defining a statistical criterion that can quantify the information loss of an experiment because of its apparent sequencing depth and (ii) determining the sequencing depth needed to control the information loss based on a pilot, possibly under-sequenced, dataset. From a statistical point of view, ChIP-seq peak calling procedures can be cast as multiple testing problems because they aim to assess whether data for each candidate locus is supported by the background noise distribution or the ChIP signal. Therefore, the information loss is naturally connected to the concept of the testing power. As a result, both of the aforementioned issues can be considered within a power calculation framework where the sequencing depth plays the role of sample size.

Power computations require modeling distribution of both the background reads and ChIP signal in a way that reflects the stochastic nature of read accumulation at each genomic locus as a function of sequencing depth. Although a number of models were proposed for locus-specific read counts, none of them explicitly accounted for read accumulation. Zhang *et al.* (2008) and Ji *et al.* (2008) considered models with locally Poisson distributed background and did not model ChIP signal. Kuan *et al.* (2011) proposed a flexible model taking into account the genome structure and overdispersion. However, this model used the input sample as a covariate and did not explicitly parametrize the model in terms of sequencing depths. Zhang *et al.* (2011) proposed a hierarchical Bayesian t-mixture model to identify local concentration of directional reads, but did not consider the relationship between read accumulation and sequencing depth. Xu *et al.* (2010) adopted a signal-to-noise model, parameters of which followed some arbitrary prior distribution to account for intrinsic read bias. Although such a prior distribution, if estimated, could be utilized to model the background distribution at varying sequencing depths, the work of Xu *et al.* (2010) exclusively focused on the normalization aspect of ChIP-seq analysis.

We developed CSSP (ChIP-seq Statistical Power) framework for statistical power calculation by considering a local Poisson model for the read generation process. We assume that background reads in the ChIP and the input samples are generated by local Poisson processes with shared Gamma prior distributions. The corresponding Gamma parameters are modeled as functions of the local genome structure, including mappability and GC content. The local Poisson parameters for the enrichment signals follow convolution of Gamma distributions. This model preserves the local structure of the Xu *et al.* (2010) model while keeping the negative binomial distribution as the marginal signal distribution as in Kuan *et al.* (2011). Such a local structure is key for capturing dynamics of the counting process for individual genomic locus as a result of increasing sequencing depths. We introduce a conditional power definition that uses the practically used notion of fold change of ChIP signal over the control input sample. We show with data-driven computational experiments that our approach can be used to determine (i) the apparent conditional power for a given sequencing depth; (ii) the required sequencing depth to achieve a target power while controlling the FDR at a specified level. Simulation experiments based on a deeply sequenced *Escherichia coli* dataset indicate that power predictions of our model agree well with the observed empirical power. Using data from pilot studies, we can reliably estimate power for larger sequencing depths; thus, the CSSP framework has significant implications for designing ChIP-seq experiments with the multiplexing functionality. Finally, we study the power of multiple ENCODE datasets with varying sequencing depths. Our results illustrate that, although the power varies considerably with the signal-to-noise ratios of the datasets, the current sequencing depths have high power for protein–DNA interactions with large effect sizes and are generally adequate for smaller effect sizes. Our calculations are further supported by the data quality metrics proposed by the ENCODE project (Harrow *et al.,* 2012).

2 METHOD

2.1 The CSSP framework

Our CSSP framework models read counts from ChIP-seq data as Poisson processes with Gamma prior distributions. We assume that the uniquely mapping reads of both the ChIP and the input samples are pre-processed by the commonly adopted method of extension to the average fragment length provided by the experimental design (Kuan *et al.*, 2011; Rozowsky *et al.*, 2009). For modeling purposes, we divide the reference genome into n non-overlapping intervals, e.g. bins as in Ji *et al.* (2008), with sizes set to the average fragment length. Let X_i and Y_i denote the number of extended input and ChIP sample reads overlapping the ith bin, respectively. Let N_x and N_y denote the sequencing depths for input and ChIP samples. We assume that X_i and Y_i follow Poisson distributions (Zhang *et al.*, 2008; Xu *et al.*, 2010):

$$(X_i \mid \lambda_i^x) \sim Pois(\lambda_i^x N_x), \quad (Y_i \mid \lambda_i^y) \sim Pois(\lambda_i^y N_y), \tag{1}$$

where λ_i^x and λ_i^y are bin-specific rate parameters for the input and ChIP samples, satisfying $E[\sum_i \lambda_i^x] = 1$ and $E[\sum_i \lambda_i^y] = 1$, where the expectations are with respect to the prior distributions that we introduce later in the text. This formulation models bin counts as Poisson processes with fixed intensities. Let Z_i be the vector containing local genomic information, such as mappability and GC-content as in Kuan *et al.* (2011) and Rashid *et al.* (2011) for the ith bin. We consider the following bin-specific prior distributions for local Poisson intensities of the input sample:

$$\begin{gathered}(\lambda_i^x \mid Z_i = z_i) \sim \Gamma\left(a, \frac{a}{\mu(z_i)}\right),\\ \mu(z_i) = \exp\{\gamma_0 + f_\gamma(z_i)\},\end{gathered} \tag{2}$$

where γ_0 is a normalization constant such that $\sum_{i=1}^n \mu(z_i) = 1$ and $f_\gamma(.)$ is a function of local genomic information. We adopt the flexible smoothing spline framework as in Kuan *et al.* (2011) for capturing the effect of mappability and GC by $f_\gamma(.)$ on the input read counts.

For the ChIP sample, we define an unobserved variable B_i to indicate enrichment state of bin i, e.g. $B_i = 0$, for background bins. For enriched bins, we allow J different states to reflect levels of enrichment strengths (e.g. $J = 2$ broadly captures low- and high-affinity binding for TFs), and correspondingly $B_i = j,\ j = 1, \cdots, J$. The prior distributions for each state are

$$\begin{aligned}(\lambda_i^y \mid Z_i = z_i, B_i = 0) &\sim \Gamma\left(b, \frac{b}{e_0\mu_i}\right),\\ (\lambda_i^y \mid Z_i = z_i, B_i = j) &\sim \Gamma\left(b^j, \frac{b^j}{\nu^j}\right), \quad j = 1, \cdots, J,\end{aligned} \tag{3}$$

where $e_0 \in (0, 1)$ is a normalizing factor reflecting the proportion of background reads in the ChIP sample (Liang and Keleş, 2012; Xu *et al.*, 2010). For brevity, we denote $\mu(z_i)$ by μ_i by suppressing its dependence on z_i, which is fixed for a genome at given read and fragment lengths and bin size. Under this model specification, the marginal distributions of X_i and Y_i are negative binomials given by:

$$\begin{aligned} X_i &\sim NB\left(a, \frac{a}{\mu_i N_x}\right), \\ (Y_i \mid B_i = 0) &\sim NB\left(b, \frac{b}{e_0 \mu_i N_y}\right), \\ (Y_i \mid B_i = j) &\sim NB\left(b^j, \frac{b^j}{\nu^j N_y}\right), \quad j = 1, \cdots, J. \end{aligned} \tag{4}$$

In contrast to the Kuan *et al.* (2011) model, marginal distributions in our model arise from two levels of hierarchy: a local Poisson distribution and a prior distribution. The local Poisson structure is critical for modeling counts for each bin on a process level as sequencing depths increase. The prior distribution models the intrinsic read biases, which are supported by arguments in Rozowsky *et al.* (2009) and Xu *et al.* (2010). In the resulting model, although the number of local Poisson parameters λ_i^ys is the same as the number of observations, inference is possible through Bayesian analysis where the posterior distribution of the ChIP counts for each bin given the bin count y_i and enrichment state $B_i = j$ is given by:

$$(\lambda_i^y \mid Z_i = z_i, B_i = 0, Y_i = y_i) \sim \Gamma\left(b + y_i, \frac{b}{e_0 \mu_i} + N_y\right), \tag{5}$$

$$(\lambda_i^y \mid Z_i = z_i, B_i = j, Y_i = y_i) \sim \Gamma\left(b^j + y_i, \frac{b^j}{\nu^j} + N_y\right), \tag{6}$$

$j = 1, 2, \cdots, J$. Determining the number of components J is a model selection problem within the CSSP framework. In practice, we recommend setting J to 2. Kuan *et al.* (2011) observed that when modeling the ChIP signal as a mixture of negative binomial distributions, two components adequately captured both the low- and high-affinity binding. Our R package enables using larger values of J and users can apply model selection criterion, such as Bayesian Information Criterion (Schwarz, 1978) to control model complexity. We used $J = 2$ for the examples presented in this article.

2.2 A multiple testing procedure and power evaluation

There is a plethora of algorithms for assessing whether individual bins are enriched in the ChIP sample compared with input sample (Wilbanks and Facciotti, 2010). Our CSSP framework naturally lands itself into a multiple testing framework. For unenriched bin i, the ChIP count originates from negative binomial distribution, $Y_i \sim NB(b, b/(e_0 \mu_i N_y))$, and ChIP counts for enriched bins have larger values. Therefore, we consider one-sided testing against the null $H_0 : Y_i \sim NB(b, b/(e_0 \mu_i N_y))$. Under the prior distribution for the local Poisson rate, the decision rule based on the marginal distribution achieves optimal Bayes risk. The P-value for $Y_i = y_i$ is thus $Pval(y_i) = P\{Y_i \geq y_i \mid Y_i \sim NB(b, b/(e_0 \mu_i N_y))\}$. Suppose we control the overall Type-I error at q and reject the null hypothesis when $Pval(y_i) < \alpha_q$. The corresponding rejection region for bin R_i is $(Q_i(\alpha_q), \infty)$ where $Q_i(\alpha_q)$ is the $(1 - \alpha_q)$-th percentile of the null distribution. To control the FDR, we set α_q using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995). The mean power of the aforementioned testing procedure across all enriched bins is given by

$$Pow'_q(N_y) = \frac{\sum_{i: B_i \neq 0} P\{Y_i > Q_i(\alpha_q) \mid B_i \neq 0\}}{\#\{i : B_i \neq 0\}}.$$

This definition of power nominally considers all enriched regions regardless of their actual enrichment levels, i.e. effect sizes. However, in practice, investigators expect true protein–DNA interaction regions to not only exhibit statistically higher read counts in the ChIP sample compared with the input sample but also achieve a pre-determined enrichment level. Regions failing to achieve this are usually filtered by peak callers (i.e. SPP and MACS) through post-processing. The statistical implication of such a practice is that in testing the observed ChIP count Y_i against the background distribution, we should impose restriction on the effect size in addition to a P-value threshold. As a result, when evaluating the power across all enriched regions, our attention should be restricted to only the enriched regions with sufficiently large effect sizes.

We introduce quality thresholds, including a fold change threshold r and a minimum intensity threshold τ, to accommodate this practical issue. As a result, the enrichment detection procedure requires that read counts exceed not only the corresponding percentile $Q_i(\alpha_q)$ required for FDR control but are also r fold of the prior mean $e_0 \mu_i N_y$ and exceed minimum intensity of τN_y. Hence, the peak calling threshold for bin i is

$$T_i(\alpha_q, r, \tau) = Q_i(\alpha_q) \vee r e_0 \mu_i N_y \vee \tau N_y,$$

where $x \vee y = \max(x, y)$. As a result, we establish the following conditional power function:

$$Pow_{q,r,\tau}(N_y) = \frac{\sum_{i \in A} P\{Y_i > T_i(\alpha_q, r, \tau) \mid \lambda_i^y > r e_0 \mu_i \vee \tau, B_i \neq 0\}}{\#A}, \tag{7}$$

where $A = \{i : \lambda_i^y > r e_0 \mu_i \vee \tau, B_i \neq 0\}$, and $\#A$ denotes the size of set A. This definition depends on local Poisson parameters, which are usually not estimable. We propose the following conditional posterior power function by plugging in Bayesian estimators of the numerator and denominator of Equation (7), respectively. A numerical algorithm to compute (8) is presented in Supplementary Materials Section A.

$$Pow^B_{q,r,\tau}(N_y) = \frac{\sum_{i=1}^{n} w_i E\left[P\{Y_i > T_i(\alpha_q, r, \tau) \mid \lambda_i^y > r e_0 \mu_i \vee \tau, B_i \neq 0\} \mid Y_i = y_i\right]}{\sum_{i=1}^{n} w_i}, \tag{8}$$

where $w_i = P\{\lambda_i^y > r e_0 \mu_i \vee \tau, B_i \neq 0 \mid Y_i = y_i\}$.

2.3 Estimating the model parameters

We present the details of the overall model fitting procedure in Supplementary Materials Section B and highlight a few key points here. The parameters in CSSP are estimated based on the marginal distributions of the ChIP and input read counts in three consecutive steps. First, we estimate hyper parameters in $f_Y(.)$ by fitting a negative binomial regression model to the input sample. This is easily carried out with the glm.nb() function in R. Second, we estimate the normalization factor e_0, the proportion of background bins π_0 and the dispersion parameter b from the ChIP sample. These parameters normalize the ChIP sample against the input sample and are critical for the downstream power evaluation. We observed that the conventional estimating methods, e.g. maximum likelihood and method of moments, lead to poor estimators of π_0 and b (Supplementary Materials Section D). Therefore, we propose the following minimum distance estimator, which is motivated by the minimum distance and robust estimation framework in Parr and Schucany (1980) as an alternative. Let $F_n(.)$ denote the empirical distribution of $Pval(y_i)$ derived from initial estimators of e_0 and b. If we plug in the true values and apply continuity adjustment (Lemma 1, Supplementary Materials Section B), F_n should be a mixture of uniform distribution between 0 and 1 arising from background bins and a point mass concentrated near 0, representing P-values from enriched bins. In other words, when $x > c$ for some tuning parameter c, $F_n(x) \approx \pi_0 x + (1 - \pi_0)$, where π_0 denotes the proportion of unenriched bins. Then minimum distance estimators of e_0, b and π_0 are given by

$$(\hat{e}_0, \hat{b}, \hat{\pi}_0) = \arg \min_{e_0, b, \pi_0} \sup_{x > c} | F_n(x) - x\pi_0 + 1 - \pi_0 |. \tag{9}$$

Finally, we apply a generalized expectation-maximization algorithm as in Kuan *et al.* (2011) to estimate the signal parameters ν^j and b^j, $j = 1, \cdots, J$.

3 RESULTS

We first evaluated our CSSP framework in a simulation study to assess the consistency of our parameter estimates, power and FDR control (Supplementary Materials Section C). Then we performed sub-sampling experiments based on two deeply sequenced datasets [*E.coli* FNR ChIP-seq dataset of Myers *et al.* (2013) and mouse GATA1 ChIP-seq dataset of Wu *et al.* (2011)] to demonstrate the consistency and power of our CSSP framework. We used multiple human CTCF ChIP-seq datasets from the ENCODE consortium to evaluate the impact of laboratory and laboratory-specific batch effects on power estimation. Finally, we investigated eight ENCODE datasets to assess the power of currently available typical ChIP-seq studies.

3.1 Model fit for deeply sequenced data

The *E.coli* FNR dataset consists of 9.07 M 32mer single-end ChIP and 6.45 M input reads. These sequencing depths approximately correspond to 4.9 and 3.5 billion reads for the mappable human genome (Rozowsky *et al.*, 2009). The GATA1 dataset from G1E-ER4 + E2 cell line is also deeply sequenced compared with many available mammalian ChIP-seq datasets and has 106.4 M 55mer and 15.6 M 36mer reads for the ChIP and input samples, respectively. For both of the datasets, we created bin level data by extending aligned reads to the average fragment length provided by their experimental protocols (150 bp for FNR and 250 bp for GATA1) and counting the number of reads overlapping every bin. Fitted probabilities of bin-level counts are compared with their empirical frequencies in Figure 1a and b, for FNR and GATA1 ChIP samples, respectively. Both figures show good agreement between the fitted and empirical frequencies. In addition, we plot the empirical cumulative distribution of the *P*-values obtained under the fitted background distributions in Figure 1c and d. Both of the empirical distributions exhibit an expected mixture pattern where the majority of the *P*-values follow a uniform distribution between 0 and 1. We note here that, for the GATA1 dataset, the ChIP and input samples have different read lengths. This indicates that our assumption of the same background prior distribution for the ChIP and input samples may not hold. However, the resulting model fits, as well as the computational experiments of the latter sections, suggest that the power estimation is robust against this type of mis-specified background estimation. This is partly because GC score is not affected by the read length, and the mappability remains the same for majority of the bins ($\approx 95\%$) between read lengths 36 and 55 bp. We discuss mis-specifications in background estimation in more details in Supplementary Materials Section E.

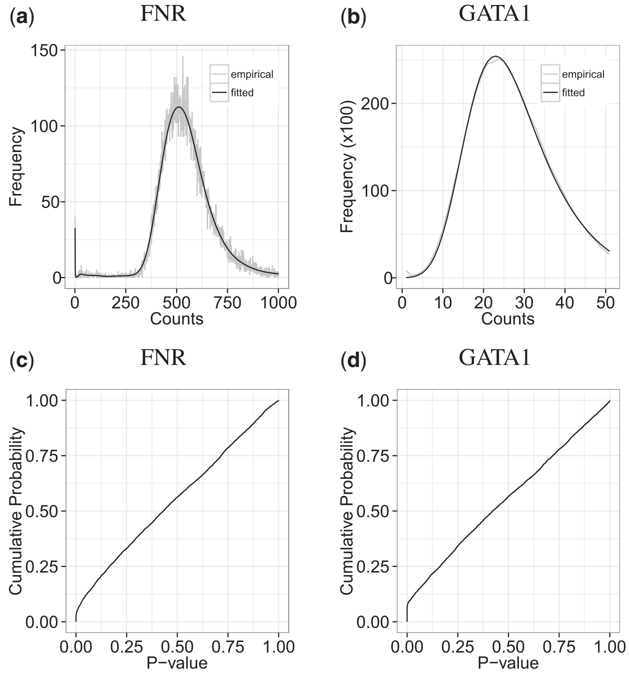

Fig. 1. Evaluating CSSP model fits. (**a**, **b**) Goodness of fit plots for the FNR (a) and GATA1 (b) datasets. (**c**, **d**) Cumulative probability plots of the *P*-values for the FNR (c) and GATA1 (d) datasets

3.2 Evaluating the accuracy of the power curve estimated based on pilot data

Next, we evaluated the consistency of CSSP power estimation when only low-sequenced pilot data are available. For both the FNR and GATA1 datasets, we generated a power curve at various sequencing depths using parameters estimated from the full data. We set the quality thresholds as $r = 2$ and $\tau = 0$. Because both datasets are deeply sequenced, their corresponding power curves can be viewed as oracle or gold-standard curves. To simulate low-sequenced pilot datasets, we sampled 0.5, 2 and 6% (FNR) and 20, 40 and 60% (GATA1) of the available ChIP and input data and refitted the models. The lowest sampling percentages of 0.5 and 20% were chosen because sub-samples with depths lower than these resulted in non-convergent parameter estimators in the CSSP model. Both of these low-depth sub-samples had an average bin-level ChIP count of 2.5. We generated 20 independent sub-samples at each sampling percentage for both datasets. The power estimates from sub-sampled data are compared with the oracle power estimates in Figure 2a and b. We observed that our power estimates based on under-sequenced GATA1 data agreed well with the oracle power curve. For the FNR dataset, when we sub-sampled <6% of the full dataset, the predicted power was biased at low-sequencing depths and agreed well with the oracle curve as the sequencing depth got larger. The mean biases of the power estimates were 0.009 and 0.015 for the 6% FNR and 20% GATA1 sub-sample datasets, respectively. The overall implications of these experiments are significant, as they indicate that our power framework is capable of reliably estimating the required depth for a target power when only undersequenced data are available.

3.3 Predicted power versus empirical power

The aforementioned comparisons thus far relied on theoretical calculations of the power based on our model fit. We next compared our theoretical power predictions with the empirical power

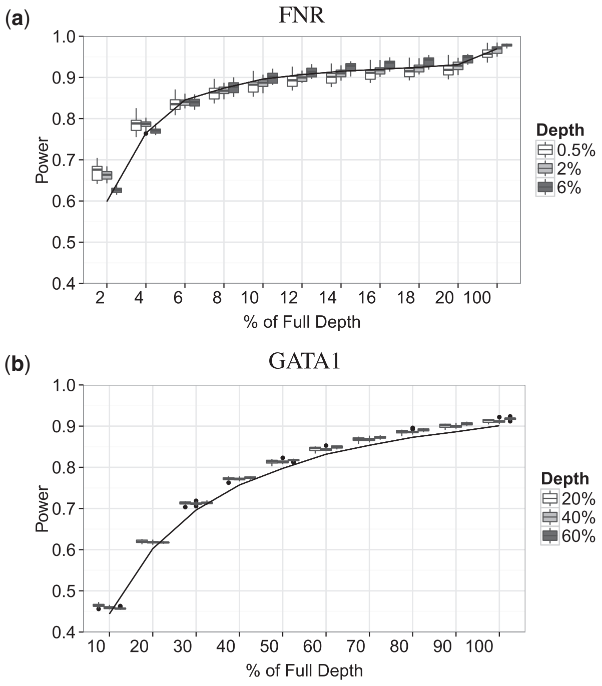

Fig. 2. Accuracy of power estimation based on pilot data. *x*-axis represents the percentage of full sequencing depth. (**a**) Boxplots represent power estimates based on sub-sampled (a) FNR (0.5, 2 and 6% of the full dataset) and (**b**) GATA1 (20, 40 and 60% of the full dataset) datasets over 20 sub-samples. The solid lines indicate the oracle power curve based on the parameters estimated from the full dataset

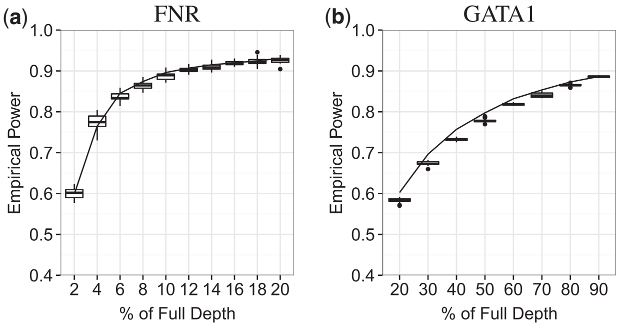

Fig. 3. Predicted power versus empirical power. Solid lines represent the model-based prediction of the power curve. Boxplots represent empirical power observed at sub-sampled (**a**) FNR and (**b**) GATA1 datasets. A total of 20 datasets are sub-sampled at every sequencing depth

observed in undersequenced datasets. This corresponds to assessing whether our Bayesian estimator of power in Equation (8) is a consistent estimator for power defined in Equation (7). Although the true set of enriched regions required by Equation (7) is unknown, we obtained gold-standard peak sets for both datasets using the full dataset with FDR of 0.05, $r = 2$, $\tau = 0$ and considered these gold-standard peak sets as proxies for the true set of enriched regions. Then, we generated 20 independent sub-samples at varying proportions of the full sequencing depth (2–20% for FNR and 20–90% for GATA1). For each sub-sample, we generated the list of enriched regions that were significant at FDR-adjusted significance levels and had at least 2-fold enrichment against the estimated background mean ($r = 2$). We then calculated the proportions of the gold-standard peak sets that overlapped with the sub-sample-based peak sets. We report the empirical power by multiplying these proportions with the power at the full sequencing depths, as the full depth power was used as a proxy for Equation (7).

Figure 3a and b displays the boxplots of empirical power as a function of sequencing depth and compares the empirical power with the oracle power curve. In both cases, the empirical power follows the CSSP power estimates closely.

3.4 Impact of the control sample on power calculations

The sequencing depth of the input library is an important factor that influences the power of ChIP-seq experiments. Our computational experiments thus far varied ChIP and the input samples simultaneously. Ho *et al.* (2011) observed identification of more peaks when a ChIP-seq dataset was normalized against a more deeply sequenced input dataset. Furthermore, this study also observed that deeply sequenced input datasets correlated well with the GC content. Chen *et al.* (2012) concluded that deeper sequencing of the input sample led to better detection specificity. These studies also established that the dependence on the sequencing depth of the input sample varied substantially between different algorithms. For example, MACS (Zhang *et al.*, 2008) achieved best performance when the ChIP and the input samples were balanced in terms of sequencing depths, whereas USeq (Nix *et al.*, 2008) performed better when more input was sequenced compared with ChIP sample.

To assess how the CSSP power estimates are influenced by the sequencing depth of the input sample, we evaluated the power from FNR sub-samples by varying the input sequencing depth as a percentage of the full depth at three levels of 0.4, 2 and 20% and fixing the ChIP sample depth at 6%. In Figure 4, we compare the resulting estimated power curves to the oracle power curve. We performed a similar experiment with the GATA1 dataset where we used 1, 2 and 4% of the input sample and 20% of the ChIP sample. We observed that varying depths of the input sample had little effect on our power calculations as long as the model parameters are reliably estimated. This is due to the fact that the input sample is only used to estimate prior mean for the background intensity, whereas estimation of the parameters regarding the ChIP signal intensity (ν^j) and normalizing effects (e_0, b) mostly rely on the ChIP sample. We observed that the estimation algorithm encountered convergence problems at extremely small depths, which would be considered as low-quality data by the currently used ChIP-seq data quality standards. Our analysis suggest that if the normalization is done in a similarly efficient fashion for other peak callers, the effect of input on their performances might also be minimized, as the background distribution alone can be captured using lower depth input samples. To illustrate this point, we compared the set of enriched bins identified at sub-sampled ChIP and input data at different combinations of depths at an FDR of 0.05 with the quality thresholds set as $r = 2$ and $\tau = 0$. For fixed ChIP samples, the set of identified enriched regions remained consistent at varying input depths (Supplementary Fig. S6). Overall, this

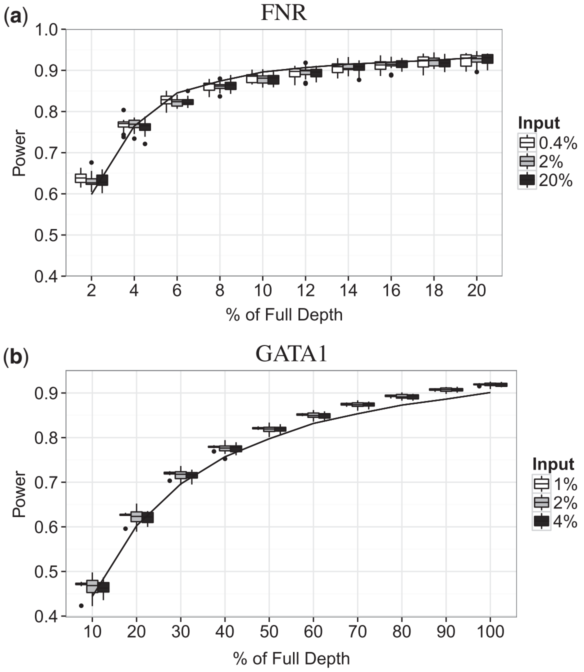

Fig. 4. Comparison of power estimation at varying input sequencing depths. ChIP sample depths are fixed at 6 and 20% of the full sample depth for the (**a**) FNR and (**b**) GATA1 samples, respectively. The oracle power based on parameters estimated from full ChIP and input data are displayed by the solid curves

supports that increasing the sequencing depth of the input beyond what is required for estimating the background parameters does not lead to power gain.

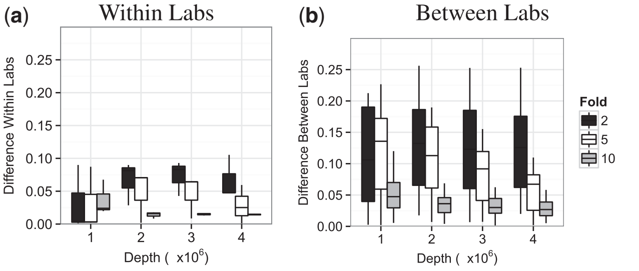

Fig. 5. Comparison of power prediction within and across laboratories. Absolute differences in predicted power are computed at various sequencing depths for CTCF ChIP-seq samples from (**a**) same laboratory and (**b**) different laboratories

3.5 Impact of laboratory and batch effects

Our computational experiments thus far focused on the effect of increasing the sequencing depth while keeping other experimental factors, i.e. laboratory and batch effects, fixed. Such effects almost always exist when pilot data are used for designing future experiments. To investigate their effects on power prediction, we analyzed seven CTCF ChIP-seq datasets from the GM12878 cell line. These datasets were generated by four different laboratories within the ENCODE consortium (The ENCODE Project Consortium, 2012). Each laboratory sequenced two to three individual cultures of GM12878. Samples within a laboratory differed in one or more of the following aspects: person and/or date for preparation of the cell cultures, cross-linked DNA and Illumina sequencing libraries; sequencing machine; and date of sequencing. The total number of reads, as well as the data quality metrics for each dataset, is summarized in Supplementary Table S3. To shorten computation time, we focused on chromosome 1. We sampled ChIP reads from each original dataset so that each pilot dataset had the same number of reads as the lowest sequenced dataset. Then, we extended the reads to average fragment lengths estimated by the R package SPP (Kharchenko *et al*., 2008) and mapped them to bins of size 100 bp. Finally, we paired each ChIP sample to its matching input and fitted the CSSP model. We generated power curves with quality thresholds of $r = 2, 5, 10$ and $\tau = 0$ at sequencing depths ranging from 10^6 to 4×10^6 (Supplementary Fig. S7). We computed the differences in predicted power between samples within the same and across different laboratories at each sequencing depth. Figure 5 illustrates that laboratory effects are more prominent than batch effects within the same laboratory, and both effects decrease when the fold change threshold increases. Supplementary Figure S7 highlights that the minimum fold change at which the batch effects are small is laboratory dependent. However, even at the small fold change of two, the maximum absolute and relative differences in power between different batches from the same laboratory are <10 and 30%, respectively. Although our study of the laboratory and batch effects is hampered by the lack of designed experiments for specifically investigating these effects, this limited study on CTCF confirms that although batch effects are difficult to control, controlling the laboratory effects by using pilot data from the same laboratory should result in better power prediction.

3.6 Power estimation for a set of recent ChIP-seq datasets

We applied the CSSP framework to evaluate the power of eight ChIP-seq experiments generated by the ENCODE project (The ENCODE Project Consortium, 2012). Datasets included ChIP-seq profiling of GATA2, cFos, cJun and Pol II in Huvec and K562 cell lines. We merged replicates within experiments and removed abnormal reads using the R package SPP (Kharchenko *et al.,* 2008). The remaining uniquely mapping reads were extended to 200 bp and mapped to bins of size 200 bp. We note that these datasets have much lower sequencing depths than the FNR and GATA1 experiments, with an average bin count of 5. For computational reasons, we fitted the CSSP model for each chromosome separately. To restrict our attention on high-quality peaks, we set the quality threshold τ to be equivalent to the 99th percentile of the ChIP bin count intensity λ_i^y estimated using the posterior distributions in Equations (5) and (6), i.e. determined by

$$\frac{1}{n}\sum_i Pr\{\lambda_i^y \le \tau \mid Y_i = y_i, Z_i = z_i\} = 0.99.$$

The estimated τN_y ranged between 10 and 20 counts across all experiments. We set the fold change threshold r at different levels of 2, 5 and 10. The numbers of final set of used reads, as well as the genome-wide power are shown in Table 1. Further results on these analyses are available in Supplementary Table S4.

Table 1. Estimated power of selected ENCODE datasets

Cell line	Factor	No. of usable reads	Power		
			2-fold	5-fold	10-fold
Huvec	GATA2	30.4 M	0.894	0.922	0.987
	cFos	36.7 M	0.925	0.932	0.987
	cJun	25.1 M	0.849	0.959	0.992
	Pol II	22.2 M	0.848	0.926	0.987
K562	GATA2	15.9 M	0.785	0.808	0.869
	cFos	9.3 M	0.759	0.794	0.844
	cJun	10.7 M	0.847	0.846	0.869
	Pol II	12.2 M	0.887	0.886	0.911

The estimated powers were generally >80% indicating that >80% of true enriched regions that meet our fold change and minimum intensity thresholds were identified. We evaluated robustness of these power calculations with simulations in Supplementary Materials Section H (Supplementary Table S5). The estimated power, in general, should increase as sequencing depth increases. However, when comparing power estimates across different experiments, quality of the individual datasets should be considered. We investigated the sequencing quality metrics of these datasets provided by the ENCODE consortium (Supplementary Table S6) and corroborated implications of these metrics with our power results. For cFos and cJun, SPOT and PBC values are comparable for both cell lines, reflecting comparable data quality. As a result, power estimates for the deeper sequenced Huvec samples are higher. For Pol II, SPOT values of Huvec, which are reflective of the signal-noise ratio, are a lot smaller than those of K562. Hence, for Pol II, although the sequencing depth of the Huvec dataset is almost twice of the K562, K562 sample has higher or comparable power.

3.7 Power implications for other peak callers

Although the CSSP framework uses a specific peak calling procedure based on testing against background read distribution and quality thresholding, power estimated by CSSP has implications for other peak callers. To illustrate this, we considered one of the commonly used peak callers SPP (Kharchenko *et al.*, 2008), which is also adopted by the ENCODE project. The key to adapting CSSP power estimation to SPP is to identify quality thresholds r and τ that would correspond to the analysis generated by SPP at the same FDR level. To enable this comparison on the ENCODE datasets, we used the enriched regions identified by SPP and set the r and τ parameters based on data from these regions. Specifically, to set τ, we mapped the filtered reads to 200 bp regions surrounding the binding sites identified by SPP, and then set τN_y as the minimum ChIP count across these bins. Similarly, we set rN_x/N_y to the minimum ChIP to input count ratio of these bins to estimate r.

We then applied the CSSP model with these r and τ estimates driven by the SPP analysis. We evaluated how well the set of enriched regions from CSSP and SPP agreed with the idea that, for datasets with good agreement, the CSSP power would yield an upper bound for the SPP power (Table 2). In addition to imposing the fold change and minimum count thresholds, SPP further filters the set of enriched regions based on the symmetry of the read distributions around each enrichment site; therefore, it is more conservative than CSSP. For the four datasets where the overlap percentages between CSSP and SPP exceeded 90%, CSSP estimated power presents upper bounds for the SPP power. For the other four experiments, >15% of SPP peaks are not captured by the CSSP model. We noticed that these four experiments either have lower data quality or sequencing depths, and the discrepancy between the two peaks caller might due to low signal-to-noise ratios or different FDR control procedures and requires further investigation.

Table 2. Implications of the CSSP estimated power for SPP

Factor	Huvec power	Overlap	K562 power	Overlap
GATA2	0.410	0.920	0.585	0.759
cFos	0.737	0.964	0.677	0.931
cJun	0.728	0.951	0.645	0.855
PolII	0.656	0.862	0.574	0.566

Overlap proportion was calculated as the proportion of SPP peaks that are among the CSSP peaks. The SPP peaks were constructed by extending each peak of SPP by the estimated 'half window size' (Kharchenko *et al.*, 2008) in both of the 5′ and 3′ directions.

4 DISCUSSION

The sequencing depths of most, if not all, initial published experiments have been limited by practical considerations, such as cost or instrument availability. With decreasing sequencing costs, considerations are shifting from how many sequences should be obtained for a single experiment, to how many experiments one can perform in a single lane. Therefore, power calculations are extremely important for ChIP-seq experiments. We have developed the CSSP framework to enable such power calculations. This framework can be applied to compute power at a wide range of sequencing depths with varying fold change and minimum intensity thresholds. Our extensive computational experiments demonstrated the consistency in predicting power from pilot data and its practical implications. To the best of our knowledge, this is the first model that enables power analysis for ChIP-seq data through an analytical approach.

It is worth noting that although our calculations mostly emphasize the sequencing depth N_y, other parameters including e_0 and ν^j, which indicate the signal-to-noise ratio of the data, as well as the data quality are also important factors of the power analysis. These parameters are fixed when comparing datasets obtained under the same experimental conditions. However, for comparing datasets with different experimental conditions, such as TF and cell line, effects of data quality and strengths of enrichment signals should bear equal emphasis. Our limited investigation of the laboratory and batch effects indicated that laboratory effects are larger than batch effects within a laboratory, and that pilot data from the same laboratory would yield more unbiased power prediction than pilot data from another laboratory.

Although the analytical calculations in the CSSP framework depend on the peak calling procedure implied by our model, the power estimation has broad applications for other peak callers. In general, if the peaks identified by the peak caller can also be identified by CSSP at the same FDR level and at certain fold change and minimum enrichment thresholds, then the power evaluated at these thresholds can serve as the upper bound for that peak caller. When peaks identified by a peak caller are vastly different than that of the CSSP, our power results can not directly be related to that peak caller. Analyzing the power for an arbitrary peak caller requires specialization of our algorithm. Overall, as the CSSP peak calling procedure is simpler than most existing peak callers, our power estimation can serve as a benchmark for other peak callers.

The CSSP framework also enables the investigation of the impact of input sequencing depth on power. Our computational experiments indicate that increasing the input depth does not increase the peak calling power or the accuracy of power prediction. The impact of input depth is, to a large extent, determined by how the ChIP read counts are normalized against input read counts, a procedure that highly varies among peak callers. Overall, our results suggest that if the ChIP sample is normalized efficiently against the input data, the dependence of power on input depths may be reduced.

ACKNOWLEDGEMENT

The authors thank Professor Patricia Kiley and Kevin Myers (UW Madison) for an early access to the FNR ChIP-seq data (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41187).

Funding: This work was supported by National Institutes of Health Grant (HG006716 to S.K.).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.*, **57**, 289–300.

Chen,Y. *et al.* (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. *Nat. Methods*, **9**, 609–614.

Gerstein,M.B. *et al.* (2010) Integrative analysis of the caenorhabditis elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.

Harrow,J. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.

Ho,J. *et al.* (2011) ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*, **12**, 134.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biot.*, **26**, 1293–1300.

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Kharchenko,P.V. *et al.* (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **6**, 1351–1359.

Kuan,P.F. *et al.* (2011) A statistical framework for the analysis of ChIP-Seq data. *J. Am. Stat. Assoc.*, **106**, 891–903.

Liang,K. and Keleş,S. (2012) Normalization of ChIP-seq data with control. *BMC Bioinformatics*, **13**, 199.

McDaniell,R. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

Myers,K.S. *et al.* (2013) Genome-scale analysis of *Escherichia coli* FNR reveals complex features of transcription factor binding. *PLoS Genet.*, **9**, e1003565.

Myers,R.M. *et al.* (2011) A Users Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.*, **9**, 21.

Nix,D.A. *et al.* (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics*, **9**, 523.

Parr,W. and Schucany,W. (1980) Minimum distance and robust estimation. *J. Am. Stat. Assoc.*, **75**, 616–624.

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

Roy,S. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**, 1787–1797.

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

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

The ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.

Wu,W. *et al.* (2011) Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. *Genome Res.*, **21**, 1659–1671.

Xu,H. *et al.* (2010) A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics*, **26**, 1199–204.

Zhang,X. *et al.* (2011) Probabilistic inference for ChIP-seq. *Biometrics*, **67**, 151163.

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