

The Poisson Margin Test for Normalization-Free Significance Analysis of NGS Data

ADAM KOWALCZYK, JUSTIN BEDO, THOMAS CONWAY, and BRYAN BERESFORD-SMITH

ABSTRACT

The current methods for the determination of the statistical significance of peaks and regions in next generation sequencing (NGS) data require an explicit normalization step to compensate for (global or local) imbalances in the sizes of sequenced and mapped libraries. There are no canonical methods for performing such compensations; hence, a number of different procedures serving this goal in different ways can be found in the literature. Unfortunately, the normalization has a significant impact on the final results. Different methods yield very different numbers of detected “significant peaks” even in the simplest scenario of ChIP-Seq experiments that compare the enrichment in a single sample relative to a matching control. This becomes an even more acute issue in the more general case of the comparison of multiple samples, where a number of arbitrary design choices will be required in the data analysis stage, each option resulting in possibly (significantly) different outcomes. In this article, we investigate a principled statistical procedure that eliminates the need for a normalization step. We outline its basic properties, in particular the scaling upon depth of sequencing. For the sake of illustration and comparison, we report the results of re-analyzing a ChIP-Seq experiment for transcription factor binding site detection. In order to quantify the differences between outcomes, we use a novel method based on the accuracy of *in silico* prediction by support vector machine (SVM) models trained on part of the genome and tested on the remainder. See Kowalczyk et al. (2009) for supplementary material.

Key words: algorithms, computational molecular biology.

1. INTRODUCTION

CURRENT SHORT READ SEQUENCING TECHNOLOGY, routinely referred to as next generation sequencing (NGS), allows for genome wide scans for various phenomena of interest, such as methylation and transcription factor binding sites. In order to derive meaningful results from the mapping of short reads (or tags) to the reference genome, a number of statistical filters based on the binomial distribution, Poisson distribution and their variants have been proposed in the literature (Rozowsky et al., 2009; Nix et al., 2008; Robertson et al., 2007; Kowalczyk, 2009). These methods are also very similar to SAGE data analysis (Baggerly et al., 2003; Robinson and Smyth, 2007), which also deals with short sequence data.

NICTA, Victoria Research Laboratory, The University of Melbourne, Parkville, Australia.

In order to discern a meaningful signal from tags mapped to a reference genome, a number of biases have to be dealt with and corrected for, as the signal “is actually the convolution of a number of effects: the density of mappable bases in a region, the underlying chromatin structure and the actual signal from transcription factor binding” (Rozowsky et al., 2009). The natural way to mitigate these issues is to introduce control samples, so that the detected signal is in the form of local enrichment of tag counts with respect to the control. The closer the preparation and processing of the control to the target sample, the more reliable the mitigation of biases.

However, experimenters cannot ensure that the target and control samples are prepared and processed in a completely equivalent manner and in practice the number of tags derived from two separate sequencing reaction can differ by a significant factor. This situation becomes endemic if one attempts to develop local models (Rozowsky et al., 2009), compensating for local biases along the reference DNA. In the simplest situation such as ChIP-Seq experimental detection of binding sites for a transcription factor (Rozowsky et al., 2009), where one attempts to detect enrichments in the target sample with respect to a carefully prepared control, there is a plausible argument for scaling the control sample counts to the level of the target, especially when considering pre-filtered narrow regions of significant enrichment.

In practice, there are other scenarios where such scaling is less applicable. An example is where one looks for the differential peaks between two different tissue samples, for example, differences in methylation between two cell lineages (Bloushtain-Qimron et al., 2008; Kowalczyk, 2009). Here, even the direction of scaling (local or global) is not obvious. Moreover, as argued in Kowalczyk (2009), the common level to which the sample counts are adjusted has a profound impact on the statistical significance of peaks when either Poisson or binomial models are used (see Section 3.3); thus, the number of detected peaks depends significantly on the choices of scaling strategies.

The situation becomes even more cumbersome for experiments that involve multiple samples, for example, when quantifying the difference between two cell lineages using pairs of samples collected from multiple subjects in order to account for patient specific heterogeneity. One possibility is to adjust all counts to a common size across the whole collection. As we have noted, this common size impacts on the number of “significant peaks” as in practice scaling could be by a factor of 2 or more with the resulting variation in p -values being by a factor of 4 or more. Moreover, the addition of new samples to the analysis could lead logically to readjustment of the updated “common size” distorting the previous results. The ad hoc nature of some of these adjustments undermines the principled statistical analysis, introducing arbitrary design choices and obscure data driven adjustments.

In this article, we propose a different statistical technique that does not require an explicit sample size adjustment and thus functions directly on the original counts. Any adjustments can be used as an additional means for accounting for other biases in the data. An example of this is the known variations in the density of mappable tags (i.e., the effective depth of sequencing) (Rozowsky et al., 2009) along the DNA sequence.

2. BACKGROUND

We now outline a conceptual model that can be used while reading this article. Using a specific protocol (e.g., sonication, enzymatic reactions cutting the DNA at specific locations, protein immunoprecipitation selecting specific fragments of protein bound to them), a library L of DNA fragments is prepared. From this library, a subset R is randomly sampled, and for each of the sampled fragments a part of it, a *short read*, is sequenced providing a *tag*, which is a k -mer of DNA bases. The tag is then mapped to one or many matching locations in an a priori known reference DNA sequence, the human genome in our case, using a specific protocol, for example, only exact matching, or only exact and unique, or only unique with up to one error. We are interested in the reference genome locations where significant over/under-representation of the mapped tags occur, so called *peaks* or *peak ranges*, as these can be interpreted as evidence for some specific property of DNA or its epigenetic modifications. In some experiments such as SAGE-Seq or digital karyotyping (Bloushtain-Qimron et al., 2008; Kowalczyk, 2009), there are natural peak regions, as the tags congregate at specific DNA locations determined by the enzymes used to cut the source DNA. In other cases, such as ChIP-Seq experiments using sonication, the peak regions have to be determined from the data using specific algorithms (Rozowsky et al., 2009; Zang et al., 2009), or perhaps just defined by a simple partitioning of the genome into uniform small blocks, say, on the order of a thousand bases.

In this section, we assume that a set of peak regions of interest is given to us. Let us consider a single peak range r . We denote by $X = X(R)$ a random variable of the count of tags mapped to r , and by x its particular realisation. If we denote by λ , $0 < \lambda < 1$, the *proportion* (fraction) of reads in the library L with tags mappable to r , then X can be modelled as a binomial random variable, $\text{Bin}(\lambda, |R|)$:

$$\mathbb{P}[x = X(R)] = \binom{|R|}{x} \lambda^x (1 - \lambda)^{|R| - x}, \quad (1)$$

for $x = 0, 1, \dots, |R|$. In a typical case of interest $\lambda \ll 1$ and the distribution of X is very well approximated by the Poisson distribution (Keeping, 1995), $\text{Poi}(\mu)$:

$$\mathbb{P}[x = X] = \mu^x \frac{e^{-\mu}}{x!}, \quad (2)$$

where $x = 0, 1, \dots, |R|$ and the *Poisson rate* is defined as $\mu = \lambda|R|$.

3. THE POISSON MARGIN TEST

Now we focus on comparing two libraries, L_A and L_B , from which random samples of mappable tags R_A and R_B were drawn, respectively. Suppose we have observed counts x_A and x_B of tags mapped to the specific peak range r of interest, and λ_A and λ_B are the corresponding proportions of tags in the libraries L_A and L_B mappable to r , respectively. Let $(a, b) \in \{(A, B), (B, A)\}$ and the following relation for empirical proportions holds:

$$\hat{\lambda}_a = \frac{x_a}{|R_a|} < \hat{\lambda}_b = \frac{x_b}{|R_b|}. \quad (3)$$

How strong is this evidence for $\lambda_a < \lambda_b$? We approach this issue in a typical statistical hypothesis testing manner. Namely, we are interested in testing the *alternative hypothesis* ($H1$) that $\lambda_a < \lambda_b$ versus the *null hypothesis* ($H0$) that $\lambda_a \geq \lambda_b$, i.e., that the complementary relation holds. As a natural test statistic we can use the maximal probability of observing at least as extreme counts X_a, X_b under the null hypothesis ($H0$). This probability we shall quantify in two different ways, a test based on the binomial distribution (*Binomial Margin*, $\mathcal{M}_{Bi}$) and its Poisson approximation (*Poisson Margin*, $\mathcal{M}_{Po}$), respectively:

$$\begin{aligned} \mathcal{M}_{Bi}(x_a, x_b) &:= \sup_{\lambda_a \geq \lambda_b > 0} \mathbb{P}[X_a \leq x_a \ \& \ x_b < X_b \mid X_i \sim \text{Bin}(\lambda_i, |R_i|)], \\ \mathcal{M}_{Po}(x_a, x_b) &:= \sup_{\lambda_a \geq \lambda_b > 0} \mathbb{P}[X_a \leq x_a \ \& \ x_b < X_b \mid X_i \sim \text{Poi}(\lambda_i |R_i|)], \end{aligned} \quad (4)$$

assuming the observed proportions relation (3) holds and, otherwise:

$$\mathcal{M}_{Bi}(x_a, x_b) = \mathcal{M}_{Po}(x_a, x_b) := 1. \quad (5)$$

In practice, both tests are numerically equivalent, but $\mathcal{M}_{Po}$ is easier to handle analytically and computationally, and will be the primary focus of the rest of this article.

We observe that the above definitions do not require that the sample sizes $|R_A|$ and $|R_B|$ be equal. By definition, the Poisson margin $\mathcal{M}_{Po}$ is the tightest universal upper bound on the following probability

$$\mathbb{P}[X_a \leq x_a \ \& \ x_b < X_b \ \& \ \lambda_a \geq \lambda_b \mid X_i \sim \text{Poi}(\lambda_i |R_i|)] \leq \mathcal{M}_{Po}(x_a, x_b).$$

In this sense, it is a very conservative p -value, corresponding to the worst case scenario test.

3.1. Computation of Poisson margin

The following result facilitates the efficient numerical evaluation of $\mathcal{M}_{Po}$; the proof is presented in the supplementary material in Kowalczyk et al. (2009). Let

$$\rho := |R_a|/|R_b|, \text{ and } \chi := x_a/x_b.$$

Theorem 1. *If the empirical relation for proportions (3) holds, then*

$$\mathcal{M}_{Po}(x_a, x_b) = \sup_{0 < \mu} e^{-2\mu} \sum_{i=0}^{x_a} \frac{(2\mu)^i}{i!(1+\rho)^i} \sum_{j>x_b} \frac{(2\mu\rho)^j}{j!(1+\rho)^j}, \quad (6)$$

where the supremum is achieved for $\mu = \mu_*$, the only solution of the reduced critical equation:

$$0 = E(\mu) := \rho + \rho \sum_{i=1}^{x_a} \prod_{j=0}^{i-1} \frac{(1+\rho)(x_a-j)}{2\mu} - \sum_{i=1}^{\infty} \prod_{j=1}^i \frac{2\mu\rho}{(1+\rho)(x_b+j)} \quad (7)$$

with the function E monotonically decreasing for $\mu > 0$, from $+\infty$ to $-\infty$. Moreover, if Eqn. 3 holds, then

$$\frac{\rho - \chi + \sqrt{(\rho - \chi)^2 + 4\chi\rho(1+\rho)}}{4\rho} \leq \frac{\mu_*}{x_b} \leq \frac{(1+\chi)(1+\rho)}{2\rho+1} + O\left(\frac{1}{x_b}\right), \quad (8)$$

where we use the “ O ”-notation for the negligible rounding term such that

$$\lim_{\varepsilon \rightarrow 0} |O(\varepsilon)/\varepsilon| < \chi/4 + 1/2.$$

Equation (7) is easy to solve numerically, using Newton’s method for example, even for large counts where $x_a, x_b \sim 10^5$. The function $E(\mu)$ is monotonically decreasing and the bounds (8) can be used for initialization of the solver iterations. The sums in (7) have quickly decaying terms, so in practice they are reduced to a summation of only a few terms. One of the aims of our derivation was to develop such a simplification and to remove some very small nuisance factors that are below the computer precision, say with $\log_{10}$ below -308 (=the limit of IEEE-754 double precision).

Proof outline. We express (4) explicitly as a two-dimensional optimization task:

$$\mathcal{M}_{Po}(x_a, x_b) = \sup_{0 < \lambda_b \leq \lambda_a} e^{-\mu_a - \mu_b} \sum_{i=0}^{x_a} \frac{\mu_a^i}{i!} \sum_{j>x_b} \frac{\mu_b^j}{j!} \Bigg|_{\substack{\mu_a = \lambda_a |R_a| \\ \mu_b = \lambda_b |R_b|}}, \quad (9)$$

which can be reduced to the one-dimensional optimization

$$\mathcal{M}_{Po}(x_a, x_b) = \sup_{0 < \lambda} e^{-\mu_a - \mu_b} \sum_{i=0}^{x_a} \frac{\mu_a^i}{i!} \sum_{j>x_b} \frac{\mu_b^j}{j!} \Bigg|_{\substack{\mu_a = \lambda |R_a| \\ \mu_b = \lambda |R_b|}}.$$

The latter task can be solved by finding a solution λ_* of the critical equation for the function of λ under “sup” above, which, after some simplifications, introduction of variable $\mu := \lambda \frac{|R_a| + |R_b|}{2}$, and removal of small positive factors, reduces to (7). The details are available in Kowalczyk et al. (2009). ■

Figure 2A shows the numerical evaluation of $\mathcal{M}_{Po}$ across a range of counts that occur in practice. This figure clearly shows where the effects of the precision limits of IEEE-754 become apparent: the significant, dark blue shaded part of the plot corresponds to p -values $< 10^{-400}$. The thousands of peaks in the experimental data discussed in Section 4 fall into this region (Kowalczyk et al., 2009). Note that the truncation of $\log_{10} \mathcal{M}_{Po}$ at -500 is used in Figure 2A purely for the purpose of visualization.

3.2. Related statistical tests

The Poisson Margin test is closely related to some other statistical tests. This relationship has been explored elsewhere (Kowalczyk et al., 2009; Kowalczyk, 2009), and here we mention only the Counts test and the Coin Toss test.

For $x_a < x_b$, the Counts test is defined as

$$T_*(x_a, x_b) := \sup_{\mu_a \geq \mu_b > 0} \mathbb{P}[X_a \leq x_a \ \& \ x_b < X_b \mid X_i \sim \text{Poi}(\mu_i)] \quad (10)$$

and for $x_a < x_b$ and the Coin Toss test as

$$\mathcal{T}_{ct}(x_a, x_b) := \mathbb{P}\left[X \leq x_a \mid X \sim \text{Bin}\left(\frac{1}{2}, x_a + x_b\right)\right]. \quad (11)$$

The Coin Toss test was used in Nix et al. (2008) and Rozowsky et al. (2009). The data from the latter article will be used for our experimental validation; hence, indirectly, we will be comparing against this statistic, and a clarification of its relationship to the Poisson Margin is appropriate here. Both articles have used Coin Toss for computing the statistical significance of the difference between two counts x_a and x_b in the same fashion as in the case of $\mathcal{M}_{Po}$, but for the special case when libraries have equal, or equalized, sizes, namely $|R_a| \approx |R_b|$. In this special case, they are “equivalent” to the $\mathcal{M}_{Po}$ test in the following sense.

Theorem 2. *If $|R_a| = |R_b|$ and the empirical proportion relation (3) holds, then*

$$\mathcal{M}_{Po}(x_a, x_b) = T_*(x_a, x_b) \approx \mathcal{T}_{ct}(x_a, x_b).$$

Proof outline. If $|R_a| = |R_b|$, then $\lambda_a < \lambda_b$ is equivalent to $\mu_a := \lambda_a |R_a| < \mu_b := \lambda_b |R_b|$; hence, the equivalence $\mathcal{M}_{Po}(x_a, x_b) = T_*(x_a, x_b)$ follows from the definitions (4) and (10). The approximation by $\mathcal{T}_{ct}$ has been argued in Kowalczyk (2009); here we demonstrate it by numerical evaluation presented in Figure 2. ■

Figure 2B shows a numerical evaluation of the differences between $\log_{10} \mathcal{M}_{Po}|_{|R_a|=|R_b|}$ and $\log_{10} \mathcal{T}_{ct}$ over a grid of values of x_a and x_b that occur in practice. Note the differences in the shading scales used in the two sub-figures. The difference $|\log_{10} \mathcal{M}_{Po} - \log_{10} \mathcal{T}_{ct}|$ is < 2 ; hence, in the areas of significant values of $\log_{10} \mathcal{M}_{Po}$, say < -100 , it composes practically negligible correction of $\leq 1\%$. This also allows the extension of the scaling properties $\mathcal{T}_{ct}$ discussed below to the case of $\mathcal{M}_{Po}|_{|R_a|=|R_b|}$.

3.3. Scaling properties

The following result can be shown formally (Kowalczyk, 2009).

Theorem 3 (Scaling Power Law). *Let $0 < 2x_a < x_b$ be two integers and $\kappa > 1$. Then*

$$\log \mathcal{T}_{ct}(\kappa x_a, \kappa x_b) = \kappa \log \mathcal{T}_{ct}(x_a, x_b) + \frac{o(x_b)}{x_b}, \quad (12)$$

where $\frac{o(x_b)}{x_b} \rightarrow 0$ for $x_b \rightarrow \infty$ denotes a “negligible” correction.

The computational validation of this result and implied practical extension to the whole range of values $x_b > 0$ is presented in Figure 1B, which is sufficient for our discussion below. Plots there show clearly the

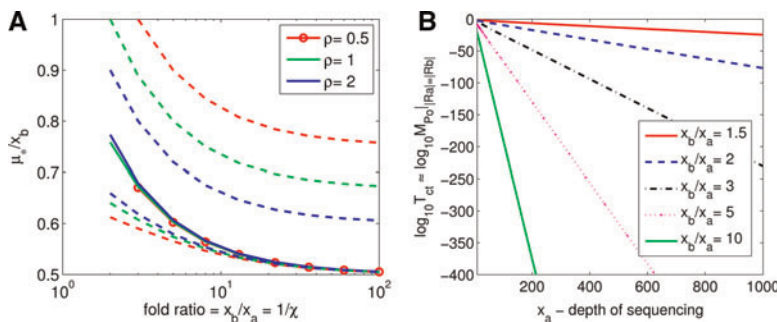
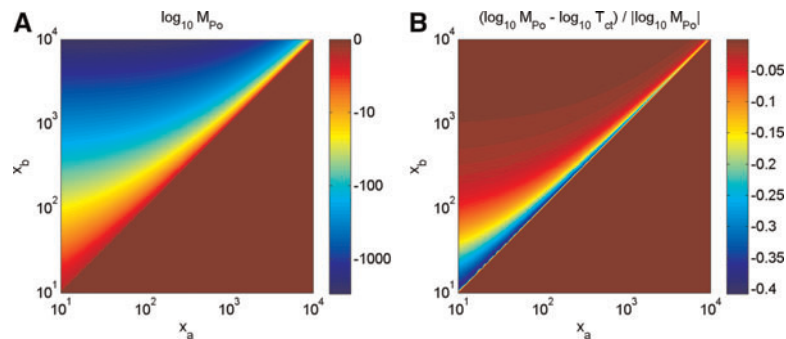


FIG. 1. (A) Lower and upper bounds on μ_* given by Eqn. (8) of Theorem 1 (broken lines) and compared to the exact values μ_* given by solution of Eqn. 7; here we show averages for μ_* as solid lines, the values for evaluation over x_a -grid of 100 values logarithmically spaced between 1 and 1000 and corresponding $x_b := x_a/\chi < 10,000$. (B) Computational validation of the Scaling Power Law given by Theorem 3.

The plots show clearly the asymptotical power scaling law (12): $\mathcal{T}_{ct}(\kappa x_a, \kappa x_b) \approx \mathcal{T}_{ct}(x_a, x_b)^\kappa$ translating to the linear dependence in the plots of the form $x_a \mapsto \log_{10} \mathcal{T}_{ct}(x_a, x_b/\chi) \approx x_a \times A^{-1} \log_{10} \mathcal{T}_{ct}(A, A/\chi)$, where $A \gg 1$ is a constant.

FIG. 2. (A) Numerical evaluation of $\log_{10} \mathcal{M}_{Po}(x_a, x_b)$ for the case of $|R_a| = |R_b|$. (B) Numerical evaluation of the relative difference $(\log_{10} \mathcal{M}_{Po} - \log_{10} \mathcal{T}_{ct}) / \log_{10} \mathcal{M}_{Po}$. The evaluation has been done for the regular logarithmic grid of count values $1 \leq x_a, x_b \leq 10,000$.



asymptotical power scaling law (12): $\mathcal{T}_{ct}(\kappa x_a, \kappa x_b) \approx \mathcal{T}_{ct}(x_a, x_b)^\kappa$ translating to the linear dependence in the plots of the form $\log_{10} \mathcal{T}_{ct}(x_a, x_b/\chi) \approx x_a \times A^{-1} \log_{10} \mathcal{T}_{ct}(A, A/\chi)$, where $A \gg 1$ is a constant.

The above Theorem and Figure 1B facilitate a discussion of two important issues in the analysis of the NGS data, namely, (i) the impact of the number of lanes use by the sequencing machine to map the library and (ii) scaling of the counts, in the case of typically unequal sizes of the sequenced and then mapped tag sets. First, they tell us that having κ times more reads sequenced, for example, using κ lanes in the sequencing machine rather than one, will provide exponentially stronger (i.e., smaller, exponentiated by κ) p -values, asymptotically for large x_b . This can be used, for example, as guidance for selection of more or fewer lanes in an NGS experiment.

However, those results also point to fragility of “count scaling.” More precisely, if the numbers of reads actually sequenced and mapped are significantly different, $|R_a| \neq |R_b|$, then either we need statistical tests that are intrinsically immune to such differences, or we need to normalize the counts in a candidate peak region to make them comparable. The $\mathcal{M}_{Po}$ test (4) falls in the first category, whereas the $\mathcal{T}_{ct}$ test represents the latter.

4. EXPERIMENTAL VALIDATION

It is far from clear that the postulated statistical test will provide useful results in practice. Simply put, the worst case scenario embraced in definition (4) may be too conservative and the generated p -values too close to 1 to be informative. In order to address this concern, we have decided to focus on the well-studied NGS application of ChIP-Seq peak calling, which involves comparison of only two libraries (the target and the dedicated control). The more complex problem of differential analysis of multiple libraries will be addressed in future work.

In order to validate our method, we have used public domain ChIP-Seq data. In particular, Rozowsky et al. (2009) provides 36,998 putative locations/regions for binding STAT1 and 24,739 locations/regions for Pol II. Note that both databases in Rozowsky et al. (2009) have been used as the basis of the most recent ENCODE data in their corresponding domains; hence, we have no alternative “gold standards” to evaluate the results of our analysis. In this article, we deal with this obstacle by using a procedure evaluating the results of analysis by checking their internal consistency. This procedure is outlined below in Section 4.3.

Since Rozowsky et al. (2009) also provides the Eland mappings of the tags for the controls, and STAT1 and Pol II data sets, we have performed an additional independent analysis of this data using Poisson Margin outlined in the following two sections.

4.1. Re-ordering

We have used the list of peak ranges, peak locations, and range boundaries exactly as in Rozowsky et al. (2009). For each range, we have extracted (raw) counts, following the protocol described in the original article, and then we used the $\mathcal{M}_{Po}$ statistic to allocate the p -value (see Table 1 for STAT1 and Table 2 for Pol II in Supplementary material in Kowalczyk et al. [2009]). Although the order according to the $\mathcal{M}_{Po}$ method is only slightly different from the original, the overlap is $>90\%$ (Table 1); the differences in performance benchmarks especially for STAT1 are significant.

4.2. *De novo significant range location*

In our experiments described below, we have implemented and run a fixed size sliding window method across the genome comparing the number of tags in each window for the control and target samples, for each sample for both DNA strands pooled together. Regions of significance are then defined by thresholding the p -values obtained from the Poisson Margin test. This resembles the approach in Zhang et al. (2008). This effectively separates the tasks of finding regions of the genome where we believe a peak lies and determining the location of the antibody binding site itself—the former task being done efficiently on the whole genome scale and the latter being done intensively on just those regions identified by the genome wide scan as containing a peak. The genome was scanned sequentially with a window of 200 bp, shifted every four bases, which took about 17 minutes on a workstation with 2GHz Opteron CPUs and 32Gb of main memory.

We have identified 35,229 peak ranges for STAT1 using un-adjusted p -values $< 1E-4$ and 28,890 using un-adjusted p -values $< 1E-6$ for Pol II, with thresholds chosen to match the numbers of peaks in the original publication. The range was defined as a contiguous region of 200BP blocks that passed the threshold.

Such a procedure can be followed by refinements of boundaries, more precise location of the range boundaries, and more precise peak location. Example of such secondary adjustments can be found elsewhere (Rozowsky et al., 2009; Zhang et al., 2008; Ji et al., 2008), but we used none here.

4.3. *Genome annotation test*

We wish to quantify consistency of the putative peak ranges by quantifying ability to predict those locations on the whole genome by a predictor trained on a part of the genome. In our experiments, peak ranges from chromosome 22 were used for training, and those from the remaining chromosomes were used only for testing. For quantifying prediction accuracy, we have adapted protocol 1B in the literature (Sonnenburg et al., 2006; Abeel et al., 2009; Bedo et al., 2009). In brief, the genome is divided into 500bp non-overlapping segments (total of 5,362,342 segments not containing ‘N’s), with each segment labeled as positive if it overlaps a peak range and negative otherwise (the positive segments comprise $< 1\%$ of the total number of segments). These labels were used to build a predictor and verify its performance measured by precision and recall. We recall that “Precision” means ratio of true positive retrieved (for a given decision threshold) to number of retrieved cases, and “Recall” is the number of true positive retrieved divided by the total number of true positive cases.

A linear support vector machine (SVM) was developed to label each 500bp segment independently (Bedo et al., 2009). Each 500bp segment was represented as a feature vector containing frequency counts of 4-mers contained within. Recursive feature elimination was used to reduce the model’s number of features, and other meta-parameters were set to maximize the area under precision-recall curve (PRC) in an internal cross-validation on the training data.

This method works very well for some tasks such as the prediction of transcription start sites and the binding of some transcription factors (e.g., c-Myc) (Bedo et al., 2009) and seems to significantly outperform other methods such as standard position weight matrices (PWM). From our experience, STAT1 is one of the harder transcription factors to predict; however, we still observe much higher performance using the SVM predictor than with PWMs (Bedo et al., 2009).

4.4. *Results*

Table 1 and Figure 3 summarize results for a few variations of genome annotation experiments described above, for STAT1 and Pol II data, respectively. As we recall, the data from chromosome 22 were used for training exclusively, and test results reported are for data from the whole genome. The following four basic variations of the experiment consisted in usage of different sets of peaks:

- **I:** The top 50% of the original list in Rozowsky et al. (2009)
- **II:** The top 50% of the list in Rozowsky et al. (2009) after sorting by $\mathcal{M}_{P_o}$ method
- **III:** The whole 100% list in Rozowsky et al. (2009)
- **IV:** A *de novo* list of peaks derived as outlined in Section 4.2

Additionally, for STAT1 we have also tested a typical PWM from TRANSFAC[®] 7.0. This is reported as row “V” in Table 1. In this case, the “score” per range in Rozowsky et al. (2009) is defined as the max of

TABLE 1. SUMMARY OF ORDERED PEAKS LISTS OVERLAP WITH THE LIST IN ROZOWSKY ET AL. (2009) AND THE ACCURACY PREDICTION OF BINDING SITE ON THE WHOLE GENOME

Experiment	List overlap (%) with Rozowsky et al. (2009)				Prediction for whole genome		
	Number of top peaks				Area under Prec.-Rec. curve	Prec. at recall	
	10%	20%	50%	100%		10%	20%
STAT1							
I: Roz.50%	100	100	100	—	2.6	7.6	4.2
II: Roz. + $\mathcal{M}_{Po}$ 50	96	94	91	—	4.0	12.3	7.5
III: Roz.100%	100	100	100	100	5.5	17.1	8.0
IV: deNovo- $\mathcal{M}_{Po}$	68	69	70	83	5.7	18.2	8.2
V: Roz.100% PWM	3	5	9	12	4.1	8.4	4.4
Pol II							
I: Roz.50%	100	100	100	—	24.4	56.0	50.1
II: Roz. + $\mathcal{M}_{Po}$ 50	53	85	95	—	25.6	57.4	51.7
III: Roz.100%	100	100	100	100	23.2	58.8	51.2
IV: deNovo- $\mathcal{M}_{Po}$	66	71	74	77	26.9	62.4	56.1

In experiments I–IV, we use data on chromosome 22 for training SVM exclusively. All values listed are in %. Boldface marks the most significant results within each group of experiments, I–II and III–IV, respectively.

PWM scores for all positions within a 500BP tile. The overlap in Table 1 was calculated using the top 37k tiles as scored by the PWM.

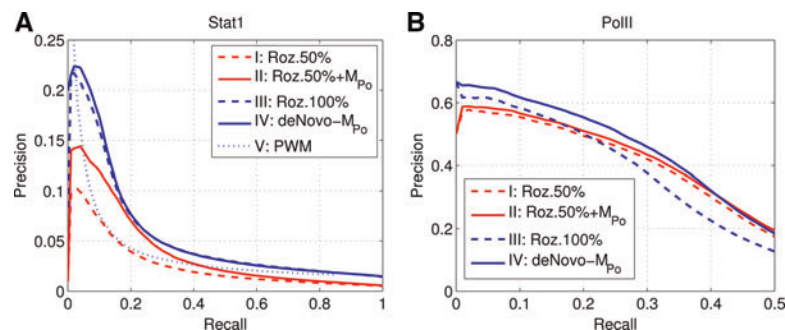
We observe that, for variants I and II for STAT1, the area under the PRC curve (4.0%) is approximately 1.5 times that for the original Rozowsky et al. (2009) ordering (2.6%). For Pol II, the corresponding difference is smaller, due to larger similarity of the ordered lists and much higher accuracy of predictions, but differences between de-novo (blue solid) and 100% list (broken blue) in Rozowsky et al. (2009) in Figure 3B is well pronounced.

5. DISCUSSION

The basic protocol described in this article can be extended in a number of directions. In the experiments, we have used only uniquely mapped reads. The density of such reads varies along the genome, which can affect the relative p -values for peaks at different locations since both $\mathcal{M}_{Po}$ and the Coin Tossing statistic ($\mathcal{T}_{ct}$) (Rozowsky et al., 2009) are sensitive to the sequencing depth (Fig. 1B). A simple way around this obstacle is to scale up the observed counts (per range) inversely to the fraction of mappable tags for the region. Such information for uniquely mapped tags of length 30 is provided in Rozowsky et al. (2009), but information of uniquely mapped up to two mismatches (which we prefer) is not currently available (Rozowsky et al., 2009). We have not used this correction here.

We present an alternative to the methods introduced previously (Rozowsky et al., 2009; Robertson et al., 2007; Zhang et al., 2008) (and see Pepke et al. [2009] for a recent review of these and other ChIP-seq computational methods). We have focused on the first of those references, since it is one of the most recent

FIG. 3. Precision–recall curves for STAT1 (A) and Pol II (B) corresponding to Table 1. We report test results for the whole genome for the variants I–V of the genome annotation test experiment described in Section 4.4.



and allows access to good quality of experimental data, included in ENCODE. We have shown that a principled analysis of such data using our method is feasible with minimal need for (arbitrary) design choices and with a minimal number of data-adjustable parameters. (An illustrating example here is the introduction in Rozowsky et al. [2009] of an ad hoc parameter $0 \leq P_f \leq 1$ for the fraction of putative highest peaks to be excluded from regression for local normalization of counts.) Our approach is robust, and can be applied to a wide variety of experimental designs involving different numbers of samples, possibly from different cell lineages. It is applicable precisely because it does not require scaling of the individual libraries of reads. The results in Kowalczyk (2009) and Section 3.3 show that such a scaling, if applied, should be done with extreme caution, if the analysis is to be meaningful.

6. CONCLUSION

We have developed a principled statistical test for the detection of significant read concentrations that is directly applicable to libraries of different (unmatched) sizes without any scaling of read counts and have demonstrated that such a scaling could introduce significant bias in the computed p -values. Although our statistical test targets differential analysis for multiple NGS libraries, the initial validation in this article is restricted to the simplest case of comparison of a target library to a matching reference. Using the recent Encode Chip-Seq data, we have shown that our test delivers non-vacuous results, with peak calling accuracies comparable or even improved with respect to the original dedicated algorithm. The absence of adequate gold standards for benchmarking was circumvented by application of a novel internal consistency check based on the accuracy of generalization of a supervised learning predictor. Demonstration of the utility of that protocol is the second major contribution of this article.

ACKNOWLEDGMENTS

NICTA is funded by the Australian Government's Department of Communications, Information Technology and the Arts, the Australian Research Council through Backing Australia's Ability, and the ICT Centre of Excellence programs.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Abeel, T., Van de Peer, Y., and Saeys, Y. 2009. Toward a gold standard for promoter prediction evaluation. *Bioinformatics* 25, i313–i320.
- Baggerly, K.A., Deng, L., Morris, J.S., et al. 2003. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* 19, 1477–1483.
- Bedo, J., MacIntyre, G., Haviv, I., et al. 2009. Simple SVM based whole-genome segmentation. Available at: <http://dx.doi.org/10.1038/npre.2009.3811.1>. Accessed December 1, 2010.
- Bloushtain-Qimron, N., Yao, J., and Snyder, E. 2008. Cell type-specific DNA methylation patterns in the human breast. *Proc. Natl. Acad. Sci. USA* 105, 14076–14081.
- Ji, H., Jiang, H., Ma, W., et al. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* 26, 1293–1300.
- Keeping, E. 1995. *Introduction to Statistical Inference*. [Reprint of 1962 edition by D. Van Nostrand Co.]. Dover, New York.
- Kowalczyk, A. 2009. Some formal results for significance of short read concentrations. Available at: <http://www.genomics.csse.unimelb.edu.au/shortreadtheory>. Accessed December 1, 2010.
- Kowalczyk, A., Bedo, J., Conway, T., et al. 2009. Poisson margin test for normalisation free significance analysis of NGS data [Supplementary Materials]. Available at: <http://www.genomics.csse.unimelb.edu.au/peakfiltsup>. Accessed December 1, 2010.

- 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.
- Pepke, S., Wold, B., and Mortazavi, A. 2009. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* 6, S22–S32.
- Robertson, G., Hirst, M., Bainbridge, M., et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657.
- Robinson, M., and Smyth, G. 2007. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* 23, 2881–2887.
- Rozowsky, J., Euskirchen, G., Auerbach, R., et al. 2009. Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nat. Biotechnol.* 27, 66–75.
- Sonnenburg, S., Zien, A., and Ratsch, G. 2006. Arts: accurate recognition of transcription starts in human. *Bioinformatics* 22, e423–e480.
- Zang, C., Schones, D.E., Zeng, C., et al. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* 25, 1952–1958.
- Zhang, Y., Liu, T., Meyer, C., et al. 2008. Model-based analysis of ChIP-seq (macs). *Genome Biol.* 9, R137+.

Address correspondence to:

Dr. Adam Kowalczyk

NICTA

Victoria Research Laboratory

The University of Melbourne

Parkville, VIC 3010, Australia

E-mail: Adam.Kowalczyk@nicta.com.au