BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 607–613
doi:10.1093/bioinformatics/bts009

Genome analysis

Advance Access publication January 19, 2012

An effective statistical evaluation of ChIPseq dataset similarity

Maria D. Chikina[1] and Olga G. Troyanskaya[2,*]

[1]Department of Neurology, Mount Sinai School of Medicine, New York, NY 10029 and [2]Lewis-Sigler Institute for Integrative Genomics, Department of Computer Science and Molecular Biology, Princeton University, Princeton, NJ 08540, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: ChIPseq is rapidly becoming a common technique for investigating protein–DNA interactions. However, results from individual experiments provide a limited understanding of chromatin structure, as various chromatin factors cooperate in complex ways to orchestrate transcription. In order to quantify chromtain interactions, it is thus necessary to devise a robust similarity metric applicable to ChIPseq data. Unfortunately, moving past simple overlap calculations to give statistically rigorous comparisons of ChIPseq datasets often involves arbitrary choices of distance metrics, with significance being estimated by computationally intensive permutation tests whose statistical power may be sensitive to non-biological experimental and post-processing variation.

Results: We show that it is in fact possible to compare ChIPseq datasets through the efficient computation of exact *P*-values for proximity. Our method is insensitive to non-biological variation in datasets such as peak width, and can rigorously model peak location biases by evaluating similarity conditioned on a restricted set of genomic regions (such as mappable genome or promoter regions).

Applying our method to the well-studied dataset of Chen *et al.* (2008), we elucidate novel interactions which conform well with our biological understanding. By comparing ChIPseq data in an asymmetric way, we are able to observe clear interaction differences between cofactors such as p300 and factors that bind DNA directly.

Availability: Source code is available for download at http://sonorus.princeton.edu/IntervalStats/IntervalStats.tar.gz

Contact: ogt@cs.princeton.edu

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

Received on August 2, 2011; revised on January 2, 2012; accepted on January 4, 2012

1 INTRODUCTION

With the advent of genome sequencing, a central question of genome biology concerns how genetic instructions in the genome translate into biochemical states of a cell, and how completely different cellular phenotypes can arise from the same genetic background. In particular, how do the various histone modifications, transcription factors and other chromatin proteins orchestrate a particular transcriptional profile?

Sequencing technology now makes it possible to map the location of the various chromatin proteins with great precision. ChIPseq studies can thoroughly characterize the binding sites of transcription factors, giving a global view of the genes they regulate, making the identification of specific binding motifs much easier, and uncovering previously uncharacterized variation (Johnson *et al.*, 2007; Wederell *et al.*, 2008). ChIPseq has also been used to profile histone modifications, increasing our understanding of the role of such modifications in transcription and differentiation (Hoffman *et al.*, 2010; Mikkelsen *et al.*, 2007; Rugg-Gunn *et al.*, 2010) as well as allowing for the discovery of novel genes not detected by other methods (Guttman *et al.*, 2009).

Ultimately, a complete understanding of chromatin organization must involve not only actual characterization of the locations of individual chromatin proteins but also an understanding of how they interact with each other to regulate transcription. Gene regulations in metazoan organisms can be quite complex, with transcription factors interacting combinatorially with each other as well as with histone modifications and the various chromatin modifiers (Cuddapah *et al.*, 2009; Hoffman *et al.*, 2010). Although ChIPseq experiments hold the promise of investigating these effects on a system-wide level, suitable methods of analysis are still being developed [e.g. Pepke *et al.* (2009), Park (2009)]. An initial step in building genome-wide understanding of chromatin organization is developing methods that infer biologically meaningful co-occurrence of chromatin components from a series of ChIPseq experiments. The simplest approach to this problem [used in (Chen *et al.*, 2008; Johnson *et al.*, 2007; Wederell *et al.*, 2008)] is to compute the total overlap of peak-enrichment regions from different experiments. While this method has a strong intuitive basis, it has several limitations. One problem simply concerns the variability in size of enrichment regions: transcription factors appear as distinct punctate peaks, for example, while chromatin modifications produce broad enrichment regions (Park, 2009). Moreover, general variations in peak distribution (such as peak length, peak number, etc.) can arise anywhere in the experimental pipeline, as well as in subsequent data processing, making direct comparisons between datasets, especially those produced by different protocols, difficult (Teytelman *et al.*, 2009).

More recently, statistical methods to assess co-occurrence have been proposed. These methods involve the computation of some metric of similarity between datasets such as correlation (Zhang *et al.*, 2007), number of clustered transcription factors (Chen *et al.*, 2008), distance-based measures (Carstensen *et al.*, 2010; Huen and Russell, 2010), etc., and then the estimation of the probability the observed metric would occur by chance, either by simulation (Chen *et al.*, 2008; Huen and Russell, 2010; Zhang *et al.*, 2007) or analytically using bounds from parametric models (Cuddapah *et al.*, 2009). [See Fu and Adryan (2009) for a

*To whom correspondence should be addressed.

comprehensive review.] With these methods, rankings of similarity of pairs of ChIPseq experiments depend both on the chosen metric and on the effectiveness of the significance estimates for the particular outcomes observed. These dependencies can lead to a lack of robustness with respect to variation in simple parameters such as enrichment-region length, which can easily result from small differences in experimental setup and post-processing steps (fragment size selection, peak calling algorithm parameters, etc.). Additionally, when compared with the simple overlap method, these approaches can be quite technical due to the methods used to obtain significance bounds, or even because of the similarity metric used, making it possible that some experimental biologists might be reluctant to adopt the more rigorous methods.

We present a method for testing the similarity of peak distributions between two sets of ChIPseq experiments in an asymmetric fashion, comparing each single peak region from a 'query' experiment to the set of peak regions in a 'reference' experiment. Apart from allowing analysis that can capture important asymmetry in the relationships between datasets, this approach eliminates the arbitrary choice of a similarity metric and dependence on the quality of significance estimates, and is the first which evaluates significance with exact *P*-values—i.e. *P*-values which represent the exact probability of a similarity event conditioned on the null hypothesis.

We use the term 'exact' to differentiate the *P*-values we obtain in our method from those obtained by either parametric methods that make assumptions about distributions or methods that employ permutation tests and thus are only able to provide estimates limited by the sample size. The fact that these *P*-values are exact in this sense allows for more robust comparisons among calculated strengths of association as even small differences in *P*-values indeed imply a real difference in the significance of the observed events.

Finally, the principles of our method are intuitive and easy to grasp, as our approach relies on the same biological intuition as the widely used overlap method. We hope this simplicity increases the likelihood of adoption by experimental biologists.

2 METHODS

Our method compares the results of ChIPseq experiments by comparing the sets of genomic interval peaks associated with the enrichment regions. The inputs to our algorithm are: the query set, which consists of enrichment regions from the query experiment (presented in blue in Fig. 1), the reference set, which consists of enrichment regions from the reference experiment (in red), and a third set of intervals, the domain set (in black), representing the line-space of possible interval locations. In the simplest case, the domain set is the entire sequence genome, but by restricting the domain set further, one can make a comparison allowing for known biases in overall interval locations. For example, if we suspect that both the reference and the query are biased towards upstream regions of genes, we can reduce the region of the genome considered by adjusting the domain set. This has the effect of giving a much more conservative estimate on the probability of overlap or proximity between the query and the reference set.

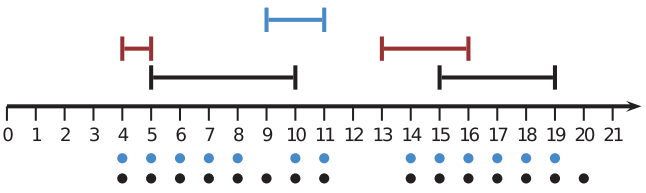

Fig. 1. A hypothetical domain set (black), reference set (red) and query interval (blue). All possible midpoint locations for the query are shown in black dots. Locations where the minimum distance is at most 2 are denoted in blue.

A principal difference between our method and alternative methods for determining statistical significance of peak distribution similarity is that our comparison is essentially asymmetric (the query and reference sets are treated differently). Other methods compare datasets in a symmetric way, assigning a score to a pair of interval sets based on distances between intervals, extent of overlap, etc., and then estimating statistical significance. Consequently, particular comparisons are affected both by the choice of metric and the power of the significance estimates for the metric values observed in the comparison. Our method instead makes an asymmetric one-against-many comparison in which we compare query intervals individually against the whole set of reference intervals. Performing the significance calculation for each interval means that, unlike for many-to-many symmetric comparisons, our algorithm's output would be identical for any distance/overlap-based metric (distance, distance-squared, etc.), eliminating the necessity to make an arbitrary choice of the scoring metric. We thus simply adopt the notion that a query interval of fixed length ℓ which is closer in distance to (or has greater overlap with) the reference set is more similar to it, and then explicitly calculate for each query interval an exact *P*-value representative of the significance of its proximity to the reference set—in other words, we calculate the probability that a randomly located interval of the same length ℓ would have been at least as close to the reference set.

We thus carry out a calculation for each individual query interval separately (we only carry out the calculation for intervals which intersect the domain). Considering Figure 1A, we calculate an exact *P*-value for the proximity of a query interval (the blue interval) to the reference set (the red intervals), conditioned on the query length ℓ (3 in this case, the length is the number of points contained in the interval) as well as intersection with the domain set (the black intervals), if applicable.

The calculation is as follows. After excluding any intervals in the reference set that do not intersect the domain set, we calculate the minimum distance, d, of the query interval to the nearest interval remaining in the reference set. This distance is the minimum distance between endpoints of the intervals (in Fig. 1 it is 2). In the case where intervals overlap, we differentiate different degrees of closeness by setting the distance to a negative number equal to

$$(\text{midpoint distance}) - \frac{\ell + (\text{reference length})}{2}.$$

Note that this is -1 in the case of overlap at a single point, -2 in the case of overlap at 2 points, etc., but can be even less than the negative overlap when one interval completely contains the other.

We then calculate the denominator of the *P*-value. This is simply the number of possible locations on the chromosome where an interval of length ℓ could be located so that it intersects the domain set. If we are not considering the domain set, this would essentially be the size of the sequenced genome (it is not exact because of the edges). In Figure 1, the denominator is 15, because there are 15 different possible locations for query intervals of length 3 which intersect the domain set.

Finally, we calculate the numerator for the *P*-value. This is the number of possible locations on the chromosome where an interval of length ℓ could be located so that it both intersects the domain set, and is at distance at most d from the reference set. In the example shown in Figure 1, where $\ell = 3$ and $d = 2$, the numerator of the *P*-value is 13, because there are 13 possible locations for query intervals of length 3 which *both* intersect the domain set, *and* are within distance ≤ 2 of a reference interval. The *P*-value is the ratio of the calculated numerator and denominator; in the example from Figure 1, it is $\frac{13}{15} \approx 0.87$.

The algorithm scales linearly with the number of query intervals as well as the number of endpoints used for the calculation (approximately the size of the domain set plus the size of the reference set filtered for domain intersection). Although working through the example from Figure 1 might

give the impression that determining the exact numerator for the *P*-value can be tricky, our algorithm implements the calculation with simple and efficient operations on intervals, allowing very fast output. For example, the running time is about 5 s per 1000 query intervals on a personal computer when the total number of reference and domain intervals is 3500 and about 50 s per 1000 query intervals when the total number is 40 000.

Since the algorithm computes the exact *P*-value for each proximity event, as expected, it generates the null distribution if the interval sets are distributed completely randomly, as demonstrated by simulated results in Figure 2A. On the other hand, real data from interacting chromatin factors produce distributions that are skewed towards small *P*-values (Fig. 2B).

3 RESULTS

3.1 Results overview

Our algorithm provides a highly intuitive way of evaluating ChIPseq dataset similarity while maintaining statistical rigor. It is robust to non-biological variation that may arise from the experimental technology, such as total genome coverage and average peak width. Additionally, the method allows us to explicitly account for global chromatin distribution biases.

Applying our method to reanalyse the Chen *et al.* dataset (we use processed data reported in GSE11431) of chromatin factors involved in stem cell maintenance, we are able to capture previously established conclusions, such as that spacial clustering of stem cell factors occurs independently of biases towards promoter regions, within a simple rigorous framework. We also resolve some disagreements in prior literature and identify several novel observations. For example, our similarity metric predicts that Nanog is much more tightly associated with Smad1 than other members of the cluster (Sox2, Oct4, p300) which is consistent with the known direct interaction between these two proteins. Additionally,

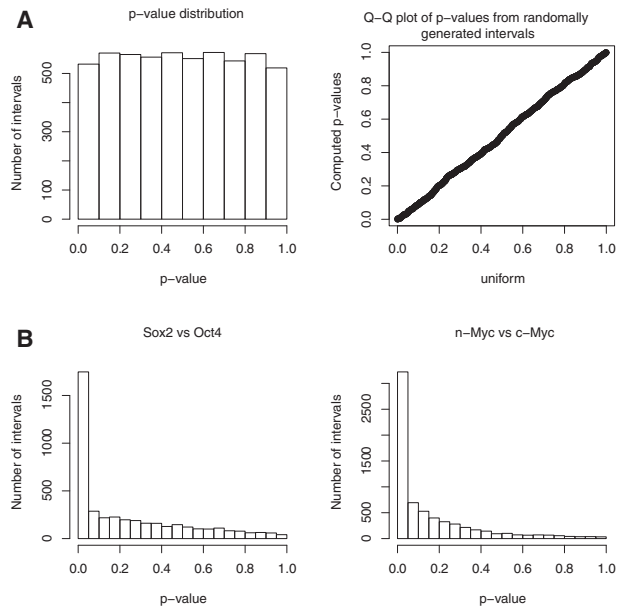

Fig. 2. (**A**) *P*-value histogram and Q–Q plot generated from intervals randomly placed on mouse chromosomes. (**B**) *P*-value histograms generated from real data in Chen *et al.* (2008).

the asymmetric nature of our comparison method (as switching the roles of query and reference produces distinct results) provides important information about global chromatin organization.

3.2 Mapping minimal distance to *P*-value

ChIPseq datasets are often compared by counting overlapping peaks which is a good first pass way of assessing which DNA binding proteins co-occur. However, overlap counts depend on parameters of the peak distribution that may not be biologically meaningful, such as the total coverage and the distribution of peak widths. Intuitively, our method may be viewed as applying the biological intuition of the overlap approach, with two important differences. First, our method uses distance as well as overlap to evaluate similarity, allowing greater sensitivity when total overlap may be small. More importantly, the essential feature of our method is that it allows statistically rigorous evaluation of similarity of interval sets, by mapping minimum distances between intervals, which are not comparable across datasets, to exact *P*-values, which can be easily compared.

The important point is that, depending on the particular characteristics of a reference set, it may be 'easier' or 'harder' to be close to the reference set—that is, statistical significance for a specific proximity event may be weaker, or stronger, respectively. To demonstrate this effect, we generate a reference set-specific function, $P_{\text{ref}}(d)$ that maps minimal distances (including overlap) to *P*-values for some fixed query interval length. By analysing the function $P_{\text{ref}}(d)$ for different sets, we can evaluate the benefits of using our statistical comparison method, rather than trying to compare overlap or distance statistics directly.

The function $P_{\text{ref}}(d)$ for all interval sets generated in Chen *et al.* (2008) (a dataset of 15 DNA binding proteins involved in stem cell maintenance) is shown in Figure 3A. Each reference set generates a unique function, supporting a key motivation of our approach that minimal distances (including overlaps corresponding to a negative distance) are not directly comparable. However, such analysis suggests a natural question as to what distribution characteristics of the reference set determine $P_{\text{ref}}(d)$. An obvious candidate is the total coverage of the interval set: the greater area of the genome covered, the easier it is to overlap with it (or be close to it). It is important to note that total genome coverage may not have a direct biological meaning (such as being related to the total number of binding events) since variation may be introduced anywhere in the experimental pipeline as well as in the peak calling algorithm. Nevertheless, since coverage does vary widely between datasets, total coverage will affect subsequent processing steps. Indeed, we find that total coverage is the dominant factor that determines the $P_{\text{ref}}(d)$; however, it does not completely explain the explicitly calculated relationship.

For example, as evident in Figure 3A, the various $P_{\text{ref}}(d)$ distance functions can cross; for example $P_{\text{Suz12}}(-10\,\text{bp})$ is smaller than $P_{\text{Smad1}}(-10\,\text{bp})$, but the relationship is reversed at $d = 10\,\text{bp}$. We further illustrate the discordance between total coverage and the $P_{\text{ref}}(d)$ function in Figure 3B, which shows a bar plot of total coverage for the analysed datasets sorted by their *P*-value at 1000 bp. We do observe the expected global pattern: the dataset with the most coverage (Ctcf) is the least significant at 1000 bp, whereas the dataset with the least coverage (p300) is most significant; yet the relationship is not entirely monotone. The reason for this non-monotonic behaviour is that the distance to *P*-value function depends

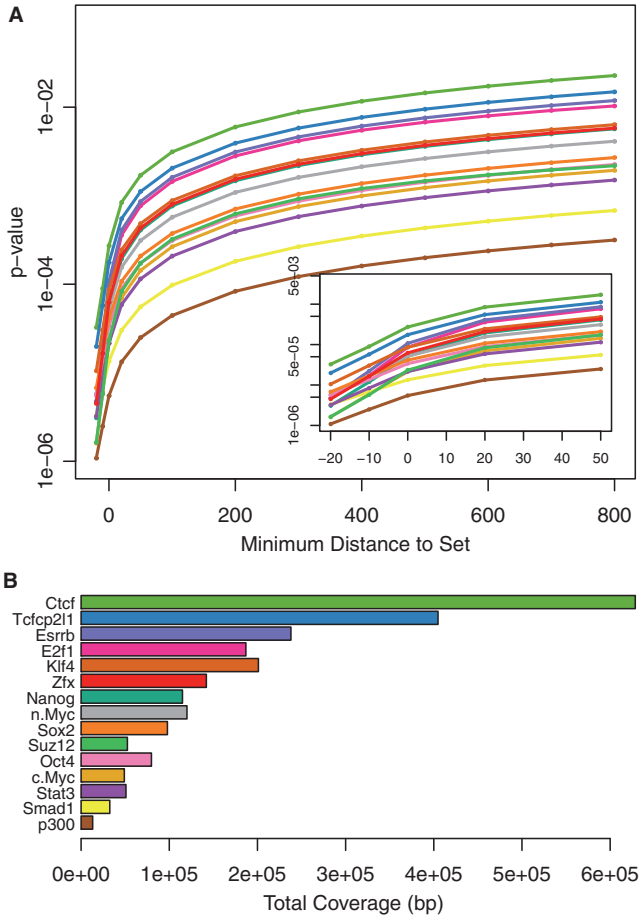

Fig. 3. (**A**) The mapping of proximity to *P*-value for the various ChIPseq experiments from Chen *et al.* (2008). The functions differ significantly demonstrating that proximity statistics are not comparable and the need for a rigorous statistical method. (**B**) Bar graph showing total coverage for the datasets in A sorted by their *P*-value at 1000 bp.

not only on total coverage but also on how that value grows as the peak regions are expanded, which may vary depending on subtle aspects of peak distribution, such as the distribution of lengths and the distribution of intervals relative to one another (e.g. clustering). Thus, since $P_{\text{ref}}(d)$ cannot be accurately parameterized in terms of easily calculated reference distribution characteristics, our empirical approach, which is not computationally intensive, presents a viable alternative.

Since the *P*-values calculated are exact and the query length is explicitly accounted for, our method is robust to variation in interval length that may be introduced (for example, by changing parameters in the peak calling method). In fact, if the intervals are expanded even by a large factor (relative to the size of the interval) equally on both sides, the *P*-values computed are exactly the same (Fig. 4A) when the domain intervals are contracted by the same factor (so that the set of discarded query intervals is identical). This is because expansion affects the distance of each query interval to the reference set equally, leaving the numerator of each *P*-value the same, and the corresponding contraction of the domain set results in

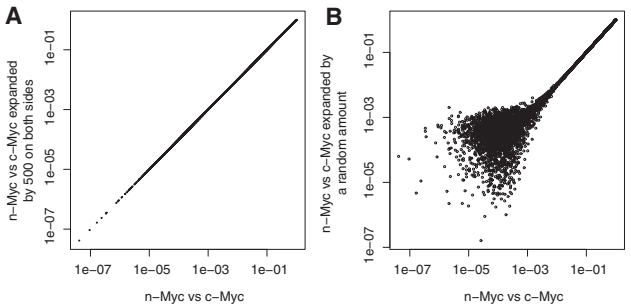

Fig. 4. Robustness of our method to interval expansion. (**A**) Reference (c-Myc) and query (n-Myc) are expanded to 500 bp on both sides, representing a more permissive peak calling parameter. If reference and query are expanded by 500 bp on both sides, the resulting *P*-values are exactly the same. (**B**) When a more realistic perturbation of random expansion (mean 500 bp) is applied, only small *P*-values are affected while the distribution shape remains constant.

an unchanged denominator. (Note also that the algorithm correctly handles a transition from proximity to overlap that inevitably results from a large expansion.) If on the other hand the intervals are expanded unequally, representing a more realistic scenario (Fig. 4B), substantial variability arises only in the low *P*-values ($P < 0.01$) and does not affect the overall distribution shape. Repeating the experiment 100 times produced a very tight distribution of summary statistics (fraction of $P < 0.01$) with all but one run being the same to three significant figures. However, it is important to note that while the distribution of *P*-values is nearly unchanged with this perturbation, the rank order of significant proximity events may be quite different.

While it is obvious that simple methods such as overlap or correlation are affected by these perturbations in the data to which our algorithm is robust, sophisticated alternatives can also be vulnerable. For example, 'Cooccur' (Huen and Russell, 2010) reports a *P*-value of 1 for the n-Myc versus c-Myc comparison, while even the exact symmetric expansion by 500 bases on both sides (used to generate Fig. 4A for our method) changes the *P*-value to 0.000999.

3.3 Controlling for global distribution biases

It should be clear that even the most 'unrelated' chromatin proteins will have some clustering of ChIP peaks, for reasons which are both biological and technical. On the technological side, various artifacts may be introduced at any step of the process from library preparation to read alignment, though many of these are markedly reduced with an appropriate control (Park, 2009). On the biological side, ChIPseq regions may be biased towards broad classes of chromosomal locations such as promoters, transcribed regions, GC rich regions, etc. (Teytelman *et al.*, 2009). With our algorithm, such hypotheses can be tested directly by supplying different domain sets to explicitly model these biases and thereby normalize the results against them. To demonstrate the effect of domains with our method, we consider a set of mouse promoters (defined as 5000 upstream to 500 downstream of transcription start site). We simulate two unrelated transcription factors by randomly assigning a binding along the chromosome with sites in promoter regions over-represented over what would be expected by chance. The simulated peak regions are

then used as input to our method. Using whole chromosomes as the domain regions (i.e. not using background correction) produces a *P*-value distribution indicative of a non-random association (Fig. 5, black). However, when promoter regions are used for the domain set, the resulting *P*-value distribution is uniform (Fig. 5, red), as expected, demonstrating that our method correctly computes the conditional probability of proximity events. In practice, knowledge of bias regions is not always perfect, making the normalization problem more difficult. We can simulate incomplete knowledge of bias regions with a set of 'noisy promoters', in which the promoters are perturbed randomly allowing them to move, expand and contract. When we use this noisy promoter set for background correction, the association between the two simulated factors is reduced (fewer small *P*-values are observed at any cutoff). It does not disappear, however, as the distribution is not uniform (Fig. 5, green). A realistic scenario in practice, however, is the ability to make a conservative estimate of the bias region, aiming to describe not the bias region exactly but a subset of it. Our domain-based calculation is ideally suited to this type of correction: considering a set of 'noisy promoters' which are incorrect only in the sense that promoter regions are too small or not represented, i.e. a strict subset of the real promoter regions, our method correctly calculates a uniform *P*-value distribution for the association of the two simulated transcription factors (Fig. 5, blue).

Various biological biases are an important concern when evaluating the similarity among distributions of binding site. For example, Chen *et al.* (2008) took care to show that the clustering of proteins assayed did not simply result from bias towards promoter regions. The authors' approach was to demonstrate that many of the clusters occurred outside of promoter regions and that the number of clustered TFs was not different the between promoter and distal regions. While this analysis is a convincing control for this dataset, it may not generalize well to other cases. Since such analysis is only a statement about the distribution of peaks that lie outside promoter regions, it relies on the fact that the hypothesized bias region (promoter regions) is nearly completely known, and the analysis would be less effective if many real genes were not annotated. Our method on the other hand directly tests if distributions are non-randomly clustered given that we expect that peak regions to occur within promoters (or any other subset). Aside from being statistically motivated, our method has the advantage of testing the distribution of intervals that occur within the domain set and not outside of it, and thus the domain set does not have to be exhaustive.

Using our method, we confirm the Chen *et al.* (2008) conclusion that the assayed TFs cluster independently of their bias towards promoters; however, peak regions generated from the chromatin modifier Suz12, used as a control by the authors, presents an interesting test case for our method. Though the authors concluded that Suz12 does not significantly cluster with any of the other TFs, a later reanalysis of the same dataset in Carstensen *et al.* (2010) reported a significant association between Suz12 and Oct4, which was supported by some experimental evidence of a meaningful association between these factors. Our approach detects the weak association observed by Carstensen *et al.* (2010), but when correcting for promoter regions by restricting the domain set, the association between Suz12 and Oct4 disappears and a uniform probability distribution is observed, suggesting that the association between Suz12 and Oct4 may be explained by the co-occurrence in promoter regions (Fig. 6A and B).

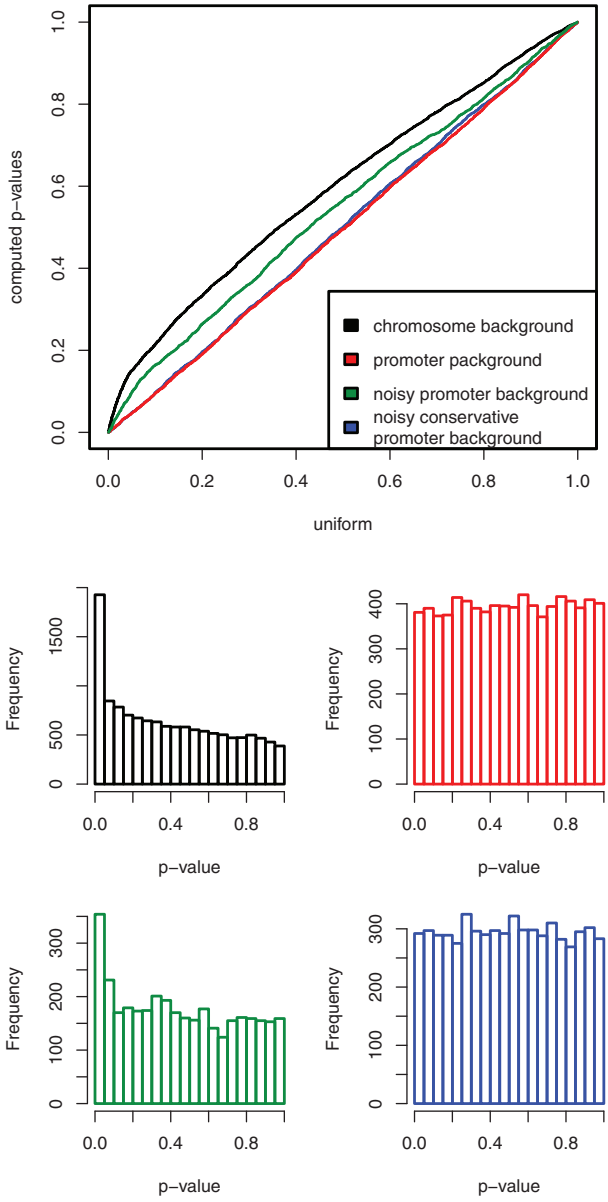

Fig. 5. Effects of applying background correction to simulated data. Two non-interacting transcription factors were simulated by choosing random binding sites along the chromosome with sites in promoters over-represented by a factor of 2.5. The two datasets were tested for association using different backgrounds: chromosome background (black), correct promoter background (red), noisy promoter background, where promoter regions are allowed to shift expand and contract (green), and a conservative noisy promoter set which is a strict subset of the correct set (blue).

Thus, despite the fact that experimental results have shown some cooperation between Suz12 and Oct4, the Chen *et al.* (2008) dataset may not provide independent evidence of an association. Of course, since the normalization in our analysis works by restricting the domain to promoter regions, the possibility remains that Suz12 and Oct4 do not interact inside promoters but do interact elsewhere [though we have also carried out the same analysis with DHS (Wu *et al.*, 2011) regions in place of promoter regions; see Supplementary

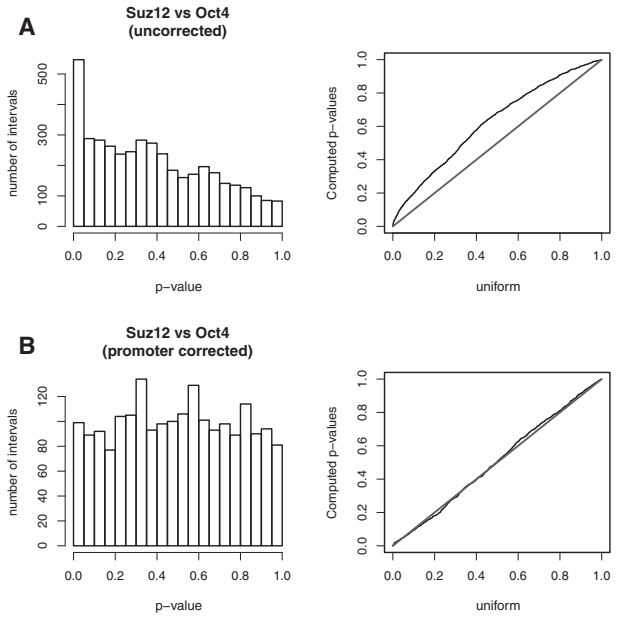

Fig. 6. Effects of applying background correction to real and simulated data. Association between Suz12 and Oct4 seen using the chromosome background (**A**) disappears when the promoter correction is applied (**B**). Corrected *P*-values are near uniform ($P = 0.104$, KS-test).

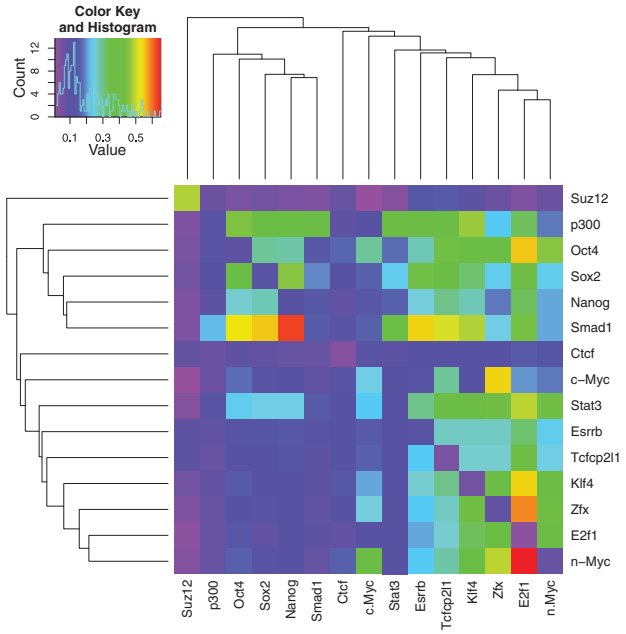

Fig. 7. Heatmap for promoter-corrected similarity values for all factors profiled in Chen *et al.* (2008). Labels on the *y*-axis represent queries, whereas labels on the *x*-axis represent references.

Fig. S1]. It seems likely, however, that the positive result of Carstensen *et al.* (2010) is due to promoter bias, which was not corrected for in their study. Most interesting to us is that these results demonstrate that it is in fact possible to observe uniform *P*-value distributions for associations in ChIPseq experiments by including a fairly crude promoter correction, suggesting that a careful application of our approach can confidently distinguish true biological interaction from coordinated bias.

3.4 A simple summary statistic

As the number of chromatin proteins profiled with ChIPseq methods is rapidly increasing, it will be necessary to move beyond analyses of single experiments to understanding global chromatin organization in terms of interactions. In order to achieve this goal, it is necessary to derive a similarity measure that can be treated much like correlation in gene expression in order to quantify interaction, compare them and ultimately analyse chromatin organization at a global network level. So far our proposed method generates a distribution of *P*-values as opposed to a single number. The fraction of *P*-values below a significance threshold provide a summary statistic that is straightforward to interpret and robust to changes in the threshold parameter. Setting the parameter to 0.05 or 0.01 has little effect on rank order of all pairwise similarity measures. For the Chen *et al.* (2008) data, the resulting ranks have Spearman's correlation of 0.97. In fact, this threshold approach is a natural analogue of counting overlap events: instead of doing it in distance space, we do it in a more principled corrected *P*-value space.

4 DISCUSSION

By using the threshold-based summary statistic, we can visualize an all-against-all comparison of the Chen *et al.* (2008) data (Fig. 7). While our method is similar in principle to overlap-based statistics, the results differ significantly from overlap (Supplementary Fig. S2).

The general similarity structure produced by our analysis agrees well with previous studies (Carstensen *et al.*, 2010; Chen *et al.*, 2008; Ouyang *et al.*, 2009). Two main clusters emerge, one composed of c-Myc, n-Myc, E2f1, Zfx and Klf4, and the second composed of Nanog, Sox2, p300, Oct4 and Smad1. However, there are some notable differences from the previous studies. For example, since our method performs an asymmetric comparison, we observe some one-sided interactions. In particular, all interactions involving p300 are highly asymmetric, implying that while p300 co-occurs with many other factors, the reverse associations are much weaker. The natural interpretation of such asymmetric interactions is that one factor (p300) binds to a subset of genomic locations of another (Oct4, Sox2, etc.). In the case of p300, the pattern of asymmetric interactions with many other factors is indicative of its co-factor function as p300 does not bind DNA on its own but is recruited by other DNA binding proteins (Janknecht and Hunter, 1996).

Another result particular to our analysis is that while Nanog is tightly associated with the Sox2, Oct4, p300 cluster, by far its strongest association is with Smad1. While Nanog has been shown to co-occur in proteins complexes with Oct4 (Wang *et al.*, 2006), Smad1 is the only factor in this dataset that has been shown experimentally to interact directly with Nanog (Suzuki *et al.*, 2006). The interaction serves to block bone morphogenetic protein (BMP) induced differentiation that is meditated by Smad1. The Smad1–Nanog interaction is also highly asymmetric, so that Smad1 co-occurs with Nanog more often than Nanog co-occurs with Smad1. This pattern indicates that the Nanog–Smad1 interaction occurs at a subset of Nanog's genomic locations and is consistent with Nanog having diverse roles in stem cell maintenance that

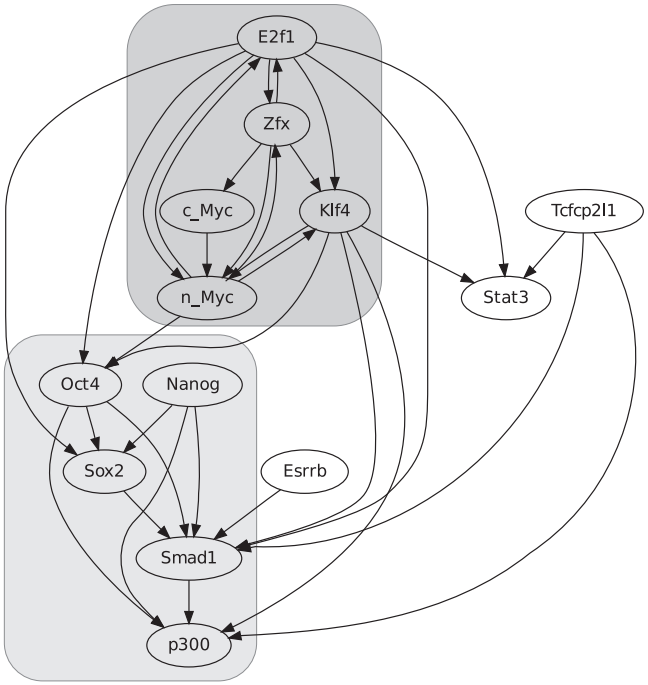

Fig. 8. Graph representation of interactions in Figure 7. Top 35 interactions are included. Two main clusters highlighted in red and blue have an overall hierarchical relationship.

involves both activation and repression of target genes (Pan and Thomson, 2007).

Overall, the complete interaction pattern presents a hierarchical structure depicted as a directional interaction graph in Figure 8. In particular, the second cluster (Nanog, Sox2, p300, Oct) interacts more strongly with the first (c-Myc, n-Myc, E2f1, Zfx, Klf4) than vice versa, indicating a subset relationship. This global hierarchical pattern is consistent with the first cluster being more broadly distributed and having a more general role, as several of the proteins in this cluster are involved in proliferation and have important roles outside embryonic stem cells (Leung *et al.*, 2008; Singh and Dalton, 2009; Zajac-Kaye, 2001). On the other hand, the factors in cluster 2 (Nanog, Sox2, Oct4) have a more restricted distribution and are considered to be specific markers of pluripotency (Kunisato *et al.*, 2010; Wernig *et al.*, 2007).

We have developed a method for ChIPseq comparisons that is statistically rigorous, yet fully transparent and thus intuitive to biology researchers. In particular, our method makes simple assumptions and does not rely on any numerical optimizations or permutation tests; our ability to observe uniform *P*-value distributions in simulated and non-interacting real datasets lends confidence to our predictions for real data. Our method naturally extends to the concept of conditional similarity making it easy to investigate how general biases in genomic locations of chromatin factors affect the perceived clustering among them. The transparency of our method should aid in its adoption in the general biology community as the output is very easy to interpret. For every query interval, our method produces the closest reference interval, the distance between them and the corresponding numerator, denominator and *P*-value, making it easy to trace the results back to the proximity events in the original data.

Funding: This research was supported by NSF CAREER award DBI-0546275, NIH grant R01 GM071966, NIH grant R01 HG005998, and partially supported by NIGMS Center of Excellence grant P50 GM071508 and NIH grant T32 HG003284.

Conflict of Interest: none declared.

REFERENCES

Carstensen,L. *et al.* (2010) Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC Bioinformatics*, **11**, 456.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Cuddapah,S. *et al.* (2009) Global analysis of the insulator binding protein ctcf in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.

Fu,A.Q. and Adryan,B. (2009) Scoring overlapping and adjacent signals from genome-wide chip and damid assays. *Mol. Biosyst.*, **5**, 1429–1438.

Guttman,M. *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Hoffman,B.G. *et al.* (2010) Locus co-occupancy, nucleosome positioning, and h3k4me1 regulate the functionality of foxa2-, hnf4a-, and pdx1-bound loci in islets and liver. *Genome Res.*, **20**, 1037–1051.

Huen,D.S. and Russell,S. (2010) On the use of resampling tests for evaluating statistical significance of binding-site co-occurrence. *BMC Bioinformatics*, **11**, 359.

Janknecht,R. and Hunter,T. (1996) Versatile molecular glue. transcriptional control. *Curr. Biol.*, **6**, 951–954.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-dna interactions. *Science*, **316**, 1497–1502.

Kunisato,A. *et al.* (2010) Generation of induced pluripotent stem cells by efficient reprogramming of adult bone marrow cells. *Stem Cells Dev.*, **19**, 229–238.

Leung,J.Y. *et al.* (2008) A role for Myc in facilitating transcription activation by e2f1. *Oncogene*, **27**, 4172–4179.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Ouyang,Z. *et al.* (2009) Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **106**, 21521–21526.

Pan,G. and Thomson,J.A. (2007) Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Res.*, **17**, 42–49.

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

Pepke,S. *et al.* (2009) Computation for chip-seq and rna-seq studies. *Nat. Methods*, **6** (11 Suppl.), S22–S32.

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

Singh,A.M. and Dalton,S. (2009) The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. *Cell Stem Cell*, **5**, 141–149.

Suzuki,A. *et al.* (2006) Nanog binds to smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **103**, 10294–10299.

Teytelman,L. *et al.* (2009) Impact of chromatin structures on dna processing for genomic analyses. *PLoS One*, **4**, e6700.

Wang,J. *et al.* (2006) A protein interaction network for pluripotency of embryonic stem cells. *Nature*, **444**, 364–368.

Wederell,E.D. *et al.* (2008) Global analysis of in vivo foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.

Wernig,M. *et al.* (2007) In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *Nature*, **448**, 318–324.

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

Zajac-Kaye,M. (2001) Myc oncogene: a key component in cell cycle regulation and its implication for lung cancer. *Lung Cancer*, **34** (Suppl. 2), S43–S46.

Zhang,Z.D. *et al.* (2007) Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions. *Genome Res.*, **17**, 787–797.