Bioinformatics, 32(12), 2016, 1766–1772
doi: 10.1093/bioinformatics/btw058
Advance Access Publication Date: 9 February 2016
Original Paper

OXFORD

Genome analysis

Measuring the spatial correlations of protein binding sites

Yingying Wei[1,*] and Hao Wu[2,*]

[1]Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong and [2]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on 23 August 2015; revised on 4 January 2016; accepted on 25 January 2016

Abstract

Motivation: Understanding the interactions of different DNA binding proteins is a crucial first step toward deciphering gene regulatory mechanism. With advances of high-throughput sequencing technology such as ChIP-seq, the genome-wide binding sites of many proteins have been profiled under different biological contexts. It is of great interest to quantify the spatial correlations of the binding sites, such as their overlaps, to provide information for the interactions of proteins. Analyses of the overlapping patterns of binding sites have been widely performed, mostly based on *ad hoc* methods. Due to the heterogeneity and the tremendous size of the genome, such methods often lead to biased even erroneous results.

Results: In this work, we discover a Simpson's paradox phenomenon in assessing the genome-wide spatial correlation of protein binding sites. Leveraging information from publicly available data, we propose a testing procedure for evaluating the significance of overlapping from a pair of proteins, which accounts for background artifacts and genome heterogeneity. Real data analyses demonstrate that the proposed method provide more biologically meaningful results.

Availability and implementation: An R package is available at http://www.sta.cuhk.edu.hk/YWei/ChIPCor.html.

Contacts: ywei@sta.cuhk.edu.hk or hao.wu@emory.edu.

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

1 Introduction

Understanding the regulatory mechanism of gene expression is an important goal in functional genomics research. It is widely believed that the regulation of transcriptional process involves the combinatory controls of different regulation including protein bindings, histone modification and DNA methylation (Brivanlou and Darnell, 2002; Jaenisch and Bird, 2003; Mitchell and Tjian, 1989). DNA-binding proteins such as transcription factors (TFs) bind to the DNA and thereby either activate or repress the expressions of nearby genes. There are extensive interactions and cooperations among different DNA-binding proteins in order to provide precise gene expression regulation (Cheng *et al.*, 2012). The interactions among proteins can often be inferred from exploring the spatial correlation of their binding sites. For example, if the binding sites of two proteins significantly overlap, it is very likely that they interact in some manner. Thus, exploring the spatial correlation of binding sites is an important first step toward understanding the interaction of proteins.

The genome-wide profiling of protein binding sites has become an easy procedure, thanks to the rapid advances of high-throughput sequencing technologies such as ChIP-seq (Johnson *et al.*, 2007). With the continuous reduction of sequencing cost, it has become a common task for a scientist to map the binding sites of a few proteins under biological conditions of interest, and then compare them with each other or with public data. These results provide information for potential interactions and co-bindings of the proteins. Tasks like this require methods to assess and quantify the spatial correlations of binding sites.

Given the binding sites (in the form of genomic regions) of two proteins, it is often of interest to know whether the binding sites significantly overlap. A straightforward approach is to perform statistical test based on a 2×2 table. To do so, one would segment the genome into small bins (such as a few hundred base paris), and use a binary vector to represent the presence/absence of binding for a protein. Then the co-occurrence pattern of the two binding sites can be represented by a 2×2 table. The statistical significance of the overlaps can be assessed using Pearson's χ^2 test or Fisher's exact test, and the overlaps can be visualized by a Venn Diagrams (Chen *et al.*, 2008; Khushi *et al.*, 2014; Zhu *et al.*, 2010). Unfortunately, such an approach is strongly influenced by the heterogeneity and tremendous size of the genome (details provided in next section). As a result, almost all pair-wise comparisons give significant results. Although constraining analysis to a smaller subset of the genomic regions can reduce the overall genome size and decrease the genome heterogeneity to certain degree implicitly (Garber *et al.*, 2012), simple procedures such as Fisher's exact tests targeting at limited regions still lack systematic treatment of the genome heterogeneity and fails to account for the background artifacts. Approaches based on distance metrics (Chikina and Troyanskaya, 2012; Favorov *et al.*, 2012) have also been proposed to measure the similarity of two lists of genomic regions. These approaches, however, don't consider the genome heterogeneity, thus also suffering from inflated significance.

Considering the genome heterogeneity, Bickel *et al.* (2010) developed a subsampling randomization test procedure and a software package Genome Structure Correction (GSC) to assess the overlaps between two region lists. The method assumes that the genome is block stationary, and the overlapping ratios are different in each block. The genome segmentation is estimated from data using a dyadic segmentation approach. However, the segmentation completely depends on the overlapping of the two given lists of regions and thus could provide unstable results. Moreover, the validity of GSC relies on the assumption that the ratio of the number of segments and the total length of the genome goes to zero. That requires the length of each stationary block to be very large, thus the genome segmentation is rather coarse, and the adjustment of genome heterogeneity is insufficient. Additional works on analyzing relationships of genomic regions are available, for example, using multivariate hidden Markov model (HMM) to characterize the combinatory histone patterns across the genome (Ernst and Kellis, 2012; Ernst *et al.*, 2011), or assessing the reproducibility of two replicates for a single TF using a copula mixture model (Li *et al.*, 2011). These methods, however, are not directly applicable in assessing the strength of correlation of two lists of binding sites.

In this work, taking advantage of the rich collection of public data, we systematically investigate the pairwise overlapping of many protein binding sites. We discover through extensive real data analyses that many seemingly significant overlaps of binding sites are actually results of phenomena in a similar flavor as the Simpson's paradox. We therefore propose a new method to adjust for genome heterogeneity and correct the experimental biases. Real data analyses demonstrate that our proposed method outperforms traditional association tests in determining statistical significance as well as providing more biological relevant ranking.

2 Methods

2.1 Data exploration

We obtained the binding sites (peaks called from ChIP-seq data) of 36 proteins in K562 and GM12878, the two most extensively studied cell lines in the ENCODE project (Consortium *et al.*, 2004). Here we focus on the overlapping patterns of proteins in K562. Results for GM12878 are presented in the supplementary materials, which are shown to have similar patterns.

Table 1. The 2×2 contingency table for a pair of TFs

	TF1-	TF1+	
TF2-	C_{11}	C_{12}	N_{1+}
TF2+	C_{21}	C_{22}	N_{2+}
	N_{+1}	N_{+2}	N

To evaluate the overlapping of a pair of proteins, we first segment the whole genome into bins of 1000 base pairs (bp). For a protein, each bin is annotated as 0 or 1 according to the absence or presence of its binding peaks. Then for a pair of proteins, a 2×2 contingency table in the format of Table 1 is constructed to present the co-occurrence pattern of binding sites. Under this setting, assessing overlapping can be formulated as testing whether the presence of binding sites over the genomic bins for the two TFs are independent, which can be achieved by a χ^2 test. Unfortunately, the χ^2 test claims significance for 627 out of the total 630 pairs of TFs for K562 at significance level of 0.05 (594 out of 595 pairs for GM12878). In the 2×2 table, under independence assumption, the expected value of overlaps C_{22} is $E[C_{22}] = \frac{N_{2+}N_{+2}}{N}$. Figure 1 shows the observed C_{22} versus $E[C_{22}]$ for all pairs of TFs: the observed are systematically greater than the expected overlaps, which leads to the significance results of almost all χ^2 tests.

The patterns showed in Figure 1 clearly demonstrate that a straightforward χ^2 test provides inflated test statistics and exaggerates the significance of correlation, because it is implausible that all pairs of proteins have significant interactions. We carefully examine the data, and discover that the phenomena are mainly caused by three data artifacts. The first artifact is the heterogeneity of the genome. If one examines the correlations at different subsets of the whole genome, two lists of binding sites could be independent at all subsets but correlated marginally. We will provide more in-depth explanation of this point in next section. The second one is the experimental artifact. Some regions tend to have reads clustered and thus easier to be called as peaks. Peaks at these regions artificially increase the overlapping ratio. Thirdly, the sheer size of the genome can turn a tiny effect into statistical significance. Due to the large N, a slight deviation from the balance point of the 2×2 table will result in a tiny P-value, which gives practitioners a wrong impression of TF association strength.

With the understanding of these artifacts, we sought to design a novel method to overcome the biases and provide more accurate results.

2.2 A Simpson's paradox in binding site correlation

Below we provide an in-depth explanation of the effect of genome heterogeneity on the overlapping analysis. Consider the following hypothetical toy example. Suppose the genome consists of two types of segments with 1000 bins and 100 000 bins, respectively. Two lists of binding sites appear independently in each type of bins (the corresponding 2×2 tables are shown in Table 2). Even though the binding sites are completely independent in both types of bins, the marginal table over the whole genome gives significant χ^2 test results with an extremely small P-value ($P < 2.2e-16$).

This phenomenon is similar to the well known **Simpson's paradox** in statistics, and is prevalent in genome-wide association studies (GWAS) and recent epigenome-wide association studies (EWAS).

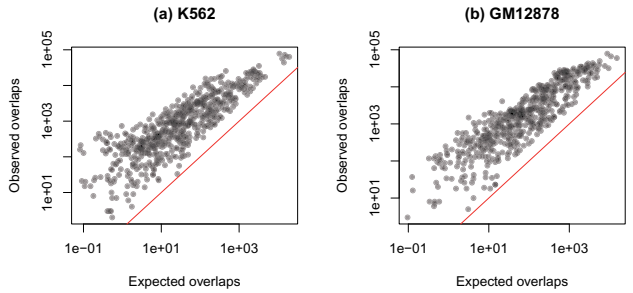

Fig. 1. Observed versus expected number of overlapped binding sites from (**a**) K562 and (**b**) GM12878 cell lines

Table 2. The 2×2 contingency tables for a toy example showing the Simpson's paradox

Type 1 bins	TF1-	TF1+	
TF2-	10	90	100
TF2+	90	810	900
	100	900	1000
Type 2 bins	TF1-	TF1+	
TF2-	81 000	9000	90 000
TF2+	9000	1000	10 000
	90 000	10 000	100 000
Total	TF1-	TF1+	
TF2-	81 010	9090	90 100
TF2+	9090	1810	10 900
	90 100	10 900	110 000

In GWAS, the disease status can be independent of the genetic markers within each subpopulation. However, collectively for a heterogeneous sampling population, spurious associations between disease and makers arise when the population structure is not considered (Marchini *et al.*, 2004; Price *et al.*, 2006). In EWAS, cell-type composition additionally confounds the relationship between epigenetic makers and diseases (Zou *et al.*, 2014). We suspect that similar phenomenon exists in protein binding sites. If one considers the genome heterogeneity and separates the genome into different segments, two lists of proteins binding sites could appear rather independently in each segment, and yet they correlate marginally.

To verify our hypothesis, we segment the whole genome according to the ChromHMM results (Ernst and Kellis, 2012). ChromHMM implements a multivariate hidden Markov model, based on histone modification status, to segment the genome into 15 different states. We obtained the genome segmentations for K562 and GM12878 from ENCODE, and constructed a 2×2 table for each type of genome segments. Then in genome segment type s ($s = 1, \ldots, 15$), C_{22}^s and its expected value $E[C_{22}^s]$ were calculated. Figure 2 plots C_{22}^s versus $E[C_{22}^s]$ for all pairs of proteins in all 15 genome segments. Comparing Figure 2 with Figure 1, it is clear that the discrepancies between observed and expected overlaps are much smaller. The discrepancies are very small in the three types of promoter regions, as the dots are almost on the diagonal lines. The discrepancies still exist in other regions, albeit in a much smaller scale. Thus, we conclude that the genome segmentation does alleviate the discrepancies between observed and expected overlaps compared to on the whole-genome scale. These figures also reveal that there is a heavy left 'tail' in the lower end at some segments, especially the longer ones such as the repressed regions and heterochromatins. These tails are likely to be

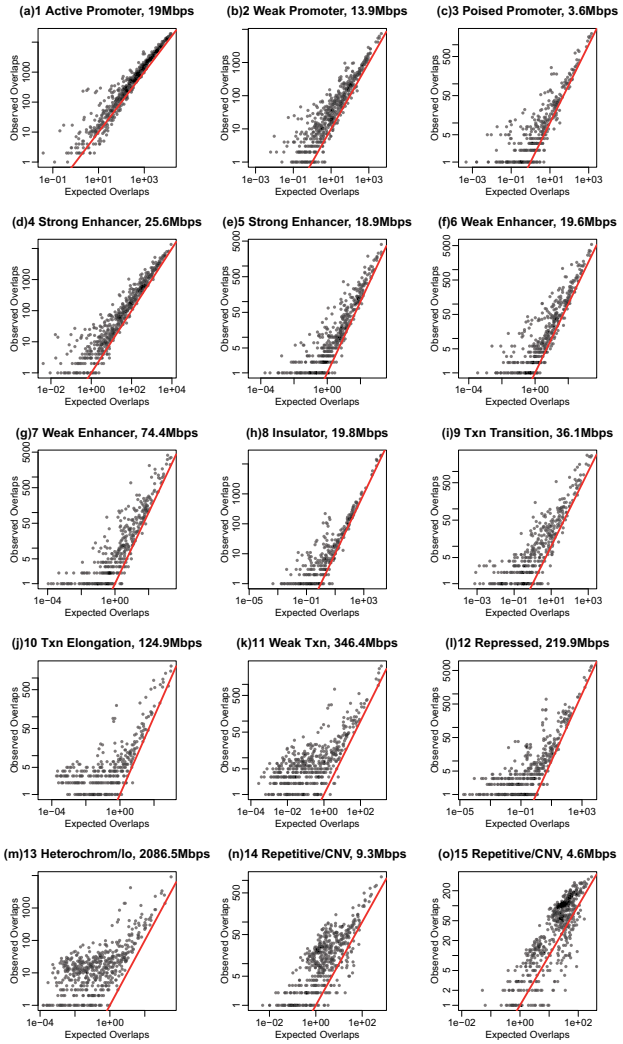

Fig. 2. Observed versus expected number of overlapped binding sites from K562, at different genome segments defined by ChromHMM

caused by the technical artifacts, that is, some regions tend to have reads clustered and be called as peaks. With these understandings, we design a method to correct the background overlaps and propose a statistical testing procedure to more objectively assess the significance of overlaps between the binding sites of a pair of proteins.

2.3 Model the overlaps

We propose the following method to correct the background overlaps of two lists of protein binding sites. The method is based on an assumption that a majority of the protein pairs don't interact. This assumption implies that ideally (after proper correction), the dots in Figure 2 should be around the diagonal lines in all segments. Similar assumption has been the foundation for many different types of differential analysis since early gene expression microarray days, e.g. a majority of the genes are not differentially expressed (Anders and Huber, 2010; Smyth, 2005; Wu *et al.*, 2013). Various methods designed for calling differential protein binding are also based on the assumption that a majority of the binding sites don't change (Chen *et al.*, 2015; Shao *et al.*, 2012). Under this assumption, we design the following method.

To simplify notations, we denote C_{22}^{is} by Y_{si}, representing the observed overlaps for the i^{th} pair of TFs in genome segment type s.

We further denote the expected overlaps computed from the 2×2 table, under independence assumption that two proteins don't interact, by X_{si}. As mentioned, $E[Y_{si}] \neq X_{si}$, so one observes discrepancies shown in Figure 2. Our goal is to estimate the 'true' expected overlaps under independence, denoted as μ_{si}, and then construct a statistical test for independence of two lists of binding sites. We use the following model for μ_{si}, which is assumed to be a function of X_{si}:

$$\mu_{si} = c_s + exp\{f_s(\log X_{si})\}. \quad (1)$$

The model is motivated by the real data observation in Figure 2. Here c_s is a small constant representing the 'baseline' overlap for genome segment type s. It means that when X_{si} is very small (literally no expected overlaps between two lists of binding sites), one still observes some overlaps due to technical artifacts. This constant is used to account for the heavy left tail observed in some regions as shown in Figure 2. When X_{si} is large, c_s becomes negligible. $f_s(\cdot)$ characterizes the relationship between μ_{si} and X_{si} on the logarithm scale. The model parameters are c_s and f_s for each s. When the estimates for these parameters are available, μ_{si} can be estimated given $\log X_{si}$, and then a test statistics can be computed.

We adopt the following procedure to estimate the parameters, based on the assumption that most pairs have no interaction. c_s is estimated from pairs of proteins with small X_{si}. Specifically, let t be the cutoff for the set of small X_{si}, and $A_s = \{i : X_{si} < t\}$ be the resulting set. Then, c_s is estimated as the average observed overlaps in set A_s: $\hat{c}_s = \sum_{i \in A_s} Y_{si}/|A_s|$. By default, we use $t = 1$ in real data analysis. In practice, we find the choice of t does not affect the final result much. Meanwhile, $f_s(\cdot)$ is estimated by locally weighted scatter plot smoothing (lowess) (Cleveland *et al.*, 1992) for pairs with reasonably large X_{si}: $\hat{f}_s = lowess(\log Y_{si} \sim \log X_{si})$ for $i \notin A_s$. With the estimates, we plug in to obtain the estimates for μ_{si}: $\hat{\mu}_{si} = \hat{c}_s + exp\{\hat{f}_s(\log X_{si})\}$.

We applied the procedure to the K562 and GM12878 data. Figure 3 shows the observed (Y_{si}) versus corrected expected overlaps ($\hat{\mu}_{si}$) from K562 data. Similar figure for GM12878 is shown as Supplemental Figure S1. As expected, the dots are *now* scattered around the diagonal lines, indicating that most pairs of proteins don't interact, so that there is no discrepancies between observed and expected overlaps.

2.4 Testing procedure

With $\hat{\mu}_{si}$ available, we use the following testing procedure to assess the overlaps. Denote the total number of bins for segment type s as N_s, we assume that the observed overlap Y_{si} for the i^{th} pair of proteins in segment s follows a Binomial distribution of $Bin(N_s, p_{si})$. p_{si} is the probability of observing overlaps in segment s, and is estimated as $\hat{p}_{si} = \hat{\mu}_{si}/N_s$. Then based on the normal approximation to the Binomial distribution, the test statistic in segment s can be calculated as:

$$t_{si} = \frac{Y_{si} - \hat{\mu}_{si}}{\sqrt{N_s \hat{p}_{si}(1 - \hat{p}_{si})}}. \quad (2)$$

This procedure provides one test statistic for each genome segment. We examine the distribution of these test statistics, and provide a Q–Q plot of t_{si} against standard normal distribution in Supplementary Figures S2 and S3. The figures show that t_{si} follow normal distribution reasonably well in each segment.

To assess the overall association between a pair of proteins, one wants to aggregate the statistics from all segments and come up with a single score. We define the aggregated score, for the i^{th} pair of protein, as $T_i = \sum_s t_{si}^2$. Technically, if t_{si} follow independent normal distributions, T_i should follow a χ^2 distribution under null. However, results show that T_i does not follow a χ^2, partly because of the correlations among t_{si} for different s. We use the following procedure to empirically generate the null distribution of T_i. We assume that no (or very few) pair of proteins 'repel' each other. In other word, the binding sites of proteins do not exclude, or the observed overlaps usually are not significantly lower than expected. Under this assumption, the left parts of t_{si} represent the null distributions for each segment type. This can be empirically seen from Supplementary Figures S2 and S3, where the left tails of the distributions in all segments are much lighter than the right tail. Then for each segment type s, we randomly sampled N_0 items from $\{t_{si} : t_{si} < 0\}$ with replacement. Denote the sample by $\tilde{t}_s^{(j)}, j = 1, \cdots, N_0$, we compute $\tilde{T}^{(j)} = \sum_s (\tilde{t}_s^{(j)})^2$, and use its empirical distribution as the null distribution for T_i. For the i^{th} protein pair, the P-value for the overall association can be calculated as $P(\tilde{T}^{(j)} > T_i)$.

2.5 Software

The proposed method has been implemented as freely R package. We are in the process of submitting the package to Bioconductor

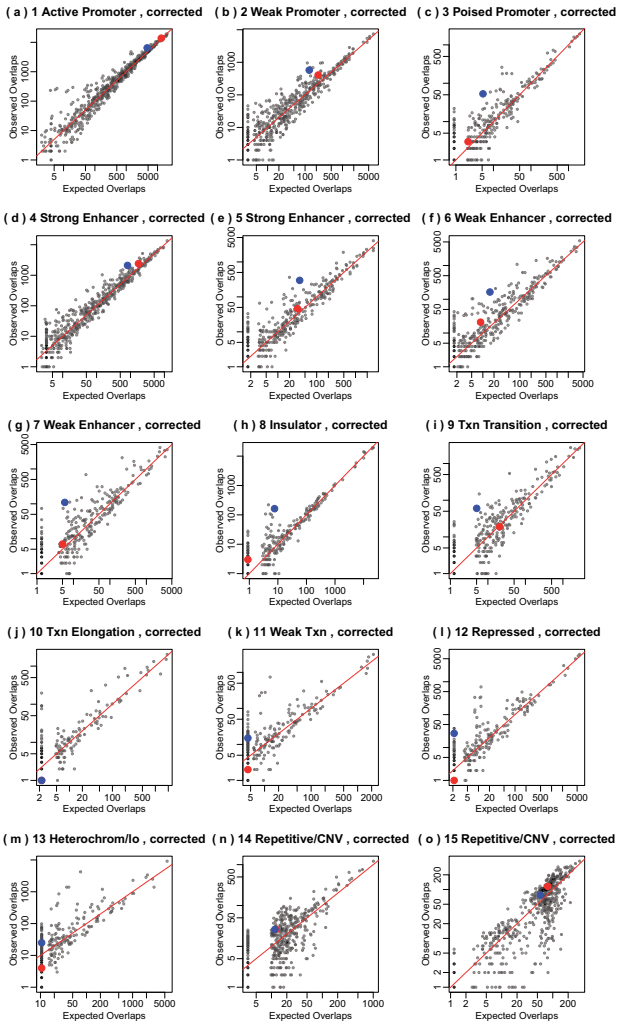

Fig. 3. Observed versus corrected expected number of overlapped binding sites from K562, at different genome segments defined by ChromHMM. The red dots (shallow color) correspond to the pair of Hey1 and c-Myc, and the blue dots (dark color) represent Max and c-Myc (Color version of this figure is available at *Bioinformatics* online.)

(Gentleman *et al.*, 2004). The software package is temporarily available at http://www.sta.cuhk.edu.hk/YWei/ChIPCor.html.

3 Results

We first look at the distribution of the test statistics and p-values from K562 data. Figure 4(a) shows the *Q*–*Q* plot of T_i versus $\tilde{T}^{(j)}$. The lower part of the T_i distribution follows the null very well. The heavy tail in the upper part represents the protein pairs that exhibit significant overlaps. Overall, the background correlation with the statistical test procedures provide reasonable results. Figure 4(b) shows the histogram of the resulting *P*-values $P(\tilde{T}^{(j)} > T_i)$, which behave reasonably well with a spike close to 0 and the rest approximately following a uniform distribution. Finally, we turn *P*-values into *q*-values, the minimum false discovery rate to call a test significant (Storey, 2003), to control for multiple comparison.

We further assess the binding site correlations for proteins from K562 and GM12878 cell lines. In all analyses, we set the cutoff t for choosing small X_{si} as 1, and the number of random samples to construct the empirical null distribution N_0 as 10 000. We apply the above procedure to all pairs of proteins to assess their spatial correlations. Here, we discuss the results for several important proteins.

We first look at MYC (c-Myc), which is an important TF in cell cycle progression, apoptosis and cellular transformation. We calculate the correlations of MYC with all the other proteins with data available from ENCODE using different methods, including our proposed method and the traditional Pearson's χ^2 test. We also intended to compare our results with GSC. Unfortunately, GSC is unavailable from ENCODE anymore, thus the comparison cannot be performed. Table 3 provides results for all proteins, where the proteins are ranked based on their association strengths with c-Myc. The rankings are based on χ^2 test statistics and the proposed T_i statistics, respectively.

First, results from χ^2 test indicate that all proteins are significantly overlapped with c-Myc (χ^2 *P*-values are not provided since they are all very small). From the proposed method, however, only 7 proteins show significant overlap with c-Myc. We believe that the proposed method provides more meaningful results since it is more plausible that only a few instead of all proteins interact with c-Myc. We further look at the rankings of the proteins, which provide important information for the degree of interactions. From χ^2 test, Hey1 is the top rank TF. However from our proposed method, it only ranks at the 11th with *q*-value 0.14, becoming insignificant. On the other hand, the top rank TF by our proposed method is Max (Myc Associated factor X), which is well known to interact with c-Myc (Blackwood and Eisenman, 1991). Figure 3 highlights the Hey1–c-Myc and Max–c-Myc pairs (in red and blue dots respectively) in the scatterplots. It can be seen clearly that Max has substantial deviations from the adjusted expected overlaps in most segment types. In contrast, Hey1 is much closer to the adjusted expected overlaps under the independence null assumption. These demonstrate that it is more reasonable that Max has stronger correlation with c-Myc. The figures also show that the most significant correlations between Max and c-Myc happen at poised promoters and various enhancer regions.

We further explore the results for c-Fos. Results are provided in Supplementary Table S1. Again, χ^2 test indicates that all proteins are significantly overlapping c-Fos, whereas the number of significant proteins is 11 from the proposed method. Looking at the rank of the proteins, Pol2, c-Myc and Hey1 are ranked among the top 10 from χ^2 test. From the proposed method, they rank 14, 21 and 22 respectively, and none of them is significant. All four proteins (c-Fos, Pol2, c-Myc and Hey1) heavily bind to the promoter regions, thus their overlaps are significant on the genome-wide scale. However, our results indicate that there's no statistical evidence that these proteins have interaction or co-binding. We conduct literature reviews on these proteins. Even though there are reports on the co-appearance on these proteins, no conclusive evidence has been shown for their interactions. Furthermore, we look at the results for CTCF, an insulator-binding protein (Supplementary Table S2). The proposed method indicate that only Rad21 and Znf263 show significant

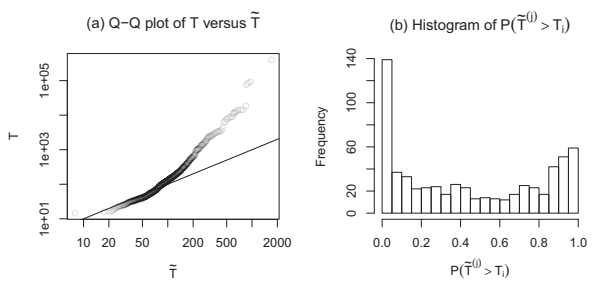

Fig. 4. (**a**) Q-Q plot of T_i versus $\tilde{T}^{(j)}$ (**b**) Histogram of $P(\tilde{T}^{(j)} > T_i)$

Table 3. The ranking of TFs in terms of association with C-myc in K562 by Pearson's chi-squared test and our proposed method

Rank	χ^2	χ^2 *P*-values	Proposed	Proposed *P*-values	Proposed *q*-values
1	Hey1	0.00E+00	Max	0.00E+00	0.00E+00
2	Taf1	0.00E+00	Usf1	0.00E+00	0.00E+00
3	Pol2(Yale)	0.00E+00	Cjun	0.00E+00	0.00E+00
4	Max	0.00E+00	DNase	1.00E−04	1.00E−03
5	Pol2(HA)	0.00E+00	Pol2(Yale)	1.00E−04	1.00E−03
6	Cjun	0.00E+00	Taf1	1.00E−04	1.00E−03
7	Gtf2b	0.00E+00	Sirt6	2.00E−04	1.90E−03
8	DNase	0.00E+00	FAIRE	8.90E−03	7.00E−02
9	FAIRE	0.00E+00	Cfos	1.10E−02	8.20E−02
10	Cfos	0.00E+00	Rad21	1.10E−02	8.20E−02
11	Gabp	0.00E+00	Hey1	2.20E−02	1.40E−01
12	Six5	0.00E+00	Pol2(HA)	2.50E−02	1.50E−01
13	Usf1	0.00E+00	Gtf2b	3.70E−02	1.90E−01
14	Pu1	0.00E+00	Six5	2.00E−01	5.40E−01
15	Tfiiic	0.00E+00	Atf3	2.00E−01	5.40E−01
16	Nfya	0.00E+00	Znf263	2.50E−01	6.20E−01
17	Nfyb	0.00E+00	Nfe2	2.70E−01	6.40E−01
18	Znf263	0.00E+00	Pu1	2.80E−01	6.50E−01
19	Egr1	0.00E+00	Jund	3.10E−01	6.90E−01
20	Sirt6	0.00E+00	Nelfe	3.60E−01	7.50E−01
21	Nelfe	0.00E+00	Ctcf	3.70E−01	7.50E−01
22	Ctcf	0.00E+00	Gabp	3.80E−01	7.60E−01
23	Sin3ak20	0.00E+00	Tfiiic	4.40E−01	8.20E−01
24	Atf3	0.00E+00	Nfya	4.40E−01	8.20E−01
25	Nfe2	0.00E+00	Bdp1	4.40E−01	8.20E−01
26	Jund	0.00E+00	Rpc155	4.60E−01	8.30E−01
27	Rad21	0.00E+00	Pol3	5.00E−01	8.80E−01
28	Rpc155	0.00E+00	Egr1	5.70E−01	9.60E−01
29	Srf	0.00E+00	Sin3ak20	6.80E−01	1.00E+00
30	Brf1	0.00E+00	Xrcc4	7.10E−01	1.00E+00
31	Bdp1	0.00E+00	Brf2	7.20E−01	1.00E+00
32	Pol3	0.00E+00	Nfyb	7.80E−01	1.00E+00
33	Nrsf	2.80E-216	Srf	8.80E−01	1.00E+00
34	Xrcc4	9.30E-187	Brf1	9.20E−01	1.00E+00
35	Brf2	2.80E-96	Nrsf	9.70E−01	1.00E+00

'HA' refers to data generated from Hudson Alpha Institute.

overlaps with CTCF. Rad21 is well known to co-bind with CTCF. From χ^2 test, proteins like c-Myc, Hey1 and Pol2 are shown very significant overlaps with CTCF. They are, however, ranked rather low from the proposed method.

In addition, we explore the IRF4 protein from GM12878 cell line. IRF4 belongs to the IRF family and is well known for combinatorially collaborations with other TFs (Garber *et al.*, 2012; Li *et al.*, 2012). From Supplementary Table S3, we can see that 9 TFs overlaps with IRF4 at false discovery rate of 5%. In particular, the proposed method ranks JUND rather high (at the third place) compared to the χ^2 test (at the 10th place), and JUND is recognized as a crucial factor for IRF4-mediated transcription (Li *et al.*, 2012). Once again, the proposed method offers more biologically plausible results.

Moreover, we investigate the robustness of the proposed method to the number of TFs used to train the model for artifacts and genome heterogeneity correction. We take the K562 data, randomly sample a subset of the 36 TFs, and use these TFs to fit the model for correcting artifacts and genome heterogeneity. Based on the estimated model, we then evaluate correlation strengths for the subsampled TFs as well as for all 36 TFs. We repeat such procedure for 10, 20 and 30 TFs. We find that even with only 20 TFs, the results are very similar to the one from using all 36 TFs (comparing Supplementary Figures S4–S9 to Figure 3, the observed versus corrected expected number of overlapped binding sites), and the test statistics t_{si}'s in each segment behave well according to Q–Q plot (Supplementary Figs S10–S12). The ranking of TFs in terms of association with c-Myc by our proposed method based on 20 and 30 TFs are provided in Supplementary Tables S4 and S5, which are very similar to the ones from using all 36 TFs. Nevertheless, when the number of training TFs drops to 10, the estimation of P-values becomes unstable, even though the relative ranking of TFs remains similar to the ones from using more TFs (Supplementary Tables S6).These results demonstrate that our proposed method will work properly even with a reasonable number of TFs. On the other hand, we acknowledge that more training TFs is always preferred for leveraging additional information to better model the background artifacts. For real application, we recommend including at least 20 TFs into the database so that in total around 200 TF pairs will be used to train the model. The rapid accumulation of ChIP-seq experiments in the public repositories will help enhance the power of the proposed method.

Among the ChromHMM segmentation, some regions such as heterochromatin are known to be less informative for assessing TF binding because the peaks in such regions are more likely to be artifacts. To investigate this, we reanalyze the data by excluding the heterochromatin. However, results are almost identical and the ranks of TFs are exactly the same. This indicates that including the less informative regions would not have much impact on the final result, thus we recommend keeping all ChromHMM segment types in analysis to avoid any subjective selection of genomic regions.

Overall, by controlling for genome heterogeneity and correcting for background artifacts, the new testing procedure allows more accurate assessment of the real interaction and collaborative binding patterns of proteins. Compared with traditional method such as χ^2 test, our results are improved in two aspects. First, it provides a better ranking for the association strength with a certain protein. We acknowledge, however, that without gold standard (the true ranking of association strengths), it is very difficult to comprehensively evaluate and compare the performance of ranking. Nevertheless, the ranks from our proposed method agree with the prior biological knowledge better overall. Secondly, the proposed method provides statistical significance is more biologically plausible. This will be particularly important in experiments where only a few proteins are profiled and their relationships are evaluated, and the scientists really care if the proteins have 'significant' interaction. Traditional methods provide statistically significant results (tiny P-values) for almost all pairs of proteins, which is not plausible. In this case the ability to draw better statistical inference from our proposed method will be especially helpful.

4 Discussion

In this work, we focus on assessing the significance of binding sites overlaps. Existing methods mostly fail to consider the genome heterogeneity and experimental artifacts, thus over-estimating the correlations and exaggerate the statistical significance. By carefully exploring the real data, we discover a Simpson's paradox phenomenon in the genome-wide overlapping patterns of protein binding sites, and we show that the genome heterogeneity plays an important role in the seemingly significant results from the traditional methods. We propose a new method to correct for genome heterogeneity and background overlaps, and then develop a statistical test procedure by leveraging information from historical data in large public data repository. Real data analysis shows that the proposed method provides more statistically meaningful and biologically interpretable results.

We want to emphasize that the idea of considering genome heterogeneity in assessing TFBS correlation is not completely absent in literature. The essence of the proposed method is similar to the one proposed in (Bickel *et al.*, 2010) and implemented in GSC, that is, the genome is assumed to be block stationary and the correlations need to be evaluated at different segments of the genome. The difference is that GSC relies on the binding sites of a pair of proteins to perform genome segmentation, whereas our proposed method utilize the genome segmentation provided by ChromHMM. The ChromHMM segmentation leverages information from other data (histone modification ChIP-seq) and provide finer, more biological meaningful segments. These results subsequently improve the inference in assessing binding sites overlaps.

Note that our method depends on the availability of ChromHMM segmentation, which is not always available. A very recent paper ChromImpute (Ernst and Kellis, 2015) supplements ChromHMM by offering an approach to impute epigenomic data and providing genome segmentation based on the imputed data. We believe this method will provide a meaningful segmentation as long as there are some epigenomic data available for a given sample. For biological samples without any available epigenomic data, we recommend segmenting the genome based on other existing data such as genome annotations. The genome can at least be separated into promoter/exonic/intronic/intergenic regions. Other annotations like CpG island and repetitive regions can also be used for better segmentation. Interactions among different regions can be introduced as well. For example, promoter regions can be separated into GC-rich and GC-poor promoters. We believe that finer segmentation and more genomically homogeneous segments will provide better results in assessing protein binding interaction. Another important point is that the data used for genome segmentation must be independent of the protein binding data, so that the confounding can be eliminated. Our software package provides an option for users to provide their own genome segments.

The proposed method operates on the peaks detected from ChIP-seq data, and the raw read counts data are not considered. We believe the correlations in the presence/absence of peaks will be more

robust than the correlations in read count level, which is prone to technical artifacts. One concern is that different ChIP-seq peak caller could potentially affect the correlation results. However, we use a rather large window size (1000 bp) to define overlaps and expect the results to be robust against the choice of ChIP-seq peak callers. Using a relatively larger window size will also include the 'closeness' relationship, because peaks spatially close to each other but not exactly overlapping could fall into the same bin and be counted as overlap.

The method proposed here is specifically designed for measuring the overlaps of short peaks. For longer peaks, or 'blocks' such as the peaks detected from H3K9me3 histone modification data, the method might not be directly applicable. The method can potentially be applied to assess the spatial correlation of other short genomic features, such as euchromatin island (Wen *et al.*, 2012) or DNase I hypersensitive site (Crawford *et al.*, 2006). To comprehensively re-assess the spatially correlation of different genomic features is our research plan in the near future.

The assumption of pairwise independence of most pairs of proteins is the fundamental of the proposed method. However 'most' is a rather loose term, and we believe as long as more than half of the pairs are independent, the method will work well. Even if the assumption is violated, the final result will be on the conservative side (under-estimate the significance), which is more desirable than overly optimistic (over-estimate the significance). Moreover, with the exponential growth of next generation sequencing datasets and their publicly availability, the null distribution can be updated and improved with inclusion of new data.

Acknowledgements

We want to thank the editor and the three reviewers for their thoughtful comments and constructive suggestions.

Funding

Y.W. received financial support CUHK4053139 from the Direct Grant of the Chinese University of Hong Kong.

Conflict of Interest: none declared.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bickel,P.J. *et al.* (2010) Subsampling methods for genomic inference. *The Ann. of Appl. Statist*, **4**, 1660–1697.

Blackwood,E.M. and Eisenman,R.N. (1991) Max: a helix-loop-helix zipper protein that forms a sequence-specific dna-binding complex with myc. *Science*, **251**, 1211–1217.

Brivanlou,A.H. and Darnell,J.E. (2002) Signal transduction and the control of gene expression. *Science*, **295**, 813–818.

Chen,L. *et al.* (2015) A novel statistical method for quantitative comparison of multiple chip-seq datasets. *Bioinformatics*, btv094.

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

Cheng,C. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.

Chikina,M.D. and Troyanskaya,O.G. (2012) An effective statistical evaluation of chipseq dataset similarity. *Bioinformatics*, **28**, 607–613.

Cleveland,W.S. *et al.* (1992) Local regression models. *Stat. Models S*, 309–376.

Consortium,E.P. *et al.* (2004) The encode (encyclopedia of dna elements) project. *Science*, **306**, 636–640.

Crawford,G.E. *et al.* (2006) Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res.*, **16**, 123–131.

Ernst,J. and Kellis,M. (2012) Chromhmm: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Favorov,A. *et al.* (2012) Exploring massive, genome scale datasets with the genometricorr package. *PLoS Comput. Biol.*, **8**, e1002529.

Garber,M. *et al.* (2012) A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell*, **47**, 810–822.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33**, 245–254.

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

Khushi,M. *et al.* (2014) Binding sites analyser (bisa): software for genomic binding sites archiving and overlap analysis. *PloS One*, **9**, e87301.

Li,P. *et al.* (2012) Batf-jun is critical for irf4-mediated transcription in t cells. *Nature*, **490**, 543–546.

Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Statist.*, **5**, 1752–1779.

Marchini,J. *et al.* (2004) The effects of human population structure on large genetic association studies. *Nat. Genet.*, **36**, 512–517.

Mitchell,P.J. and Tjian,R. (1989) Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. *Science*, **245**, 371–378.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Shao,Z. *et al.* (2012) Manorm: a robust model for quantitative comparison of chip-seq data sets. *Genome Biol.*, **13**, R16.

Smyth,G.K. (2005). Limma: linear models for microarray data. In: Robert,G. *et al.* (eds) *Bioinformatics and Computational Biology Solutions Using R And Bioconductor*. New York, Springer, pp. 397–420.

Storey,J. (2003) The positive false discovery rate: a bayesian interpretation and the q-value. *Ann. Stat.*, **31**, 2013C–22035.

Wen,B. *et al.* (2012) Euchromatin islands in large heterochromatin domains are enriched for ctcf binding and differentially dna-methylated regions. *BMC Genomics*, **13**, 566.

Wu,H. *et al.* (2013) A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, **14**, 232–243.

Zhu,L.J. *et al.* (2010) Chippeakanno: a bioconductor package to annotate chip-seq and chip–chip data. *BMC Bioinf.*, **11**, 237.

Zou,J. *et al.* (2014) Epigenome-wide association studies without the need for cell-type composition. *Nat. Methods*, **11**, 309–311.