Supplementary Material for

Detecting differential binding of transcription factors with ChIP-seq

Kun Liang and Sündüz Keleş

1 Further Details of DBChIP

1.1 Consensus site clustering

Most ChIP-seq peak-finding programs (SPP by Kharchenko et al. 2008, SISSR by Jothi et al. 2008, QuEST by Valouev et al. 2008, and others) can output a list of predicted binding site locations and their corresponding scores indicating the strength of binding for a given ChIP sample. The preferred score for the clustering purpose is the ChIP sample read count near the binding site or a proxy of it such as the WTD score from SPP.

We group predicted locations from multiple conditions into clusters of close-by locations by using agglomerative (bottom-up) hierarchical clustering with centroid linkage. The centroid of a cluster is the average of predicted locations within the cluster. The clustering can be configured by a parameter `in.distance` (default value 100 bp), which determines the maximum distance between the centroids of any two clusters.

The window size for counting the read at each binding site, denoted as w in the main text, should be set slightly larger than the estimated average fragment length. The average fragment length can be obtained either through experimental protocol or computational estimation (e.g., Zhang et al. 2008, Kharchenko et al. 2008 and Jothi et al. 2008). This parameter is not crucial, and a reasonable estimate of the average fragment length should suffice for the purpose of detecting differential binding. Majority of the ChIP-seq datasets on GEO (Barrett et al., 2011) have an average fragment length of 200-250 bp. We set the default value of w at 250 bp.

1.2 Test for differential binding

Given the sequencing depth (total number of reads) of ChIP sample i, n_i, the read count at site j follows a Binomial distribution, i.e., $x_{ij} \sim \text{Binom}(n_i, \pi_{ij})$ where x_{ij} is the read count for the jth site in the ith condition. When n_i is large and π_{ij} is small, which is the case for ChIP-seq data, x_{ij} approximately follows Poisson distribution with rate $n_i\pi_{ij}$. Then the null hypothesis of non-differential binding is

$$H_0 : \pi_{1j} = \pi_{2j} = \cdots = \pi_{Kj}, \tag{1}$$

which can be tested though a Chi-square test for equal proportions (Agresti, 2002).

However, real data often exhibit dispersion that Poisson distribution can not fully account for. Negative binomial distribution is a commonly used alternative for such cases. More specifically, assume that x_{ij} follows a $NB(\mu_{ij}, \phi)$ distribution with mean

$\mu_{ij} = n_i\pi_{ij}$ and variance $\mu_{ij} + \phi\mu_{ij}^2$, where ϕ is a dispersion parameter. Under this setup, testing the null hypothesis in (1) can be carried out within the well established generalized linear model framework (McCullagh and Nelder, 1989). The main difficulty is the estimation of the dispersion parameter ϕ when there are only a few or even no replicates. In this paper, replicates stand for biological replicates; technical replicates usually can be merged after their consistency is established. In the remainder of this section, we develop proper tests for differential binding depending on the availability of replicates.

1.2.1 Differential binding with replicates

If replicates are available, we use the edgeR package (Robinson et al., 2010) which is popular for RNA-seq data analysis to estimate the dispersion parameter based on the replicates, followed by the test for differential binding under the negative binomial assumption. The default is to use a common dispersion parameter, which is especially encouraged when the number of binding sites is small. When the number of binding sites is large, users can choose to use site-specific dispersion estimates by setting `common.disp=FALSE` in DBChIP. More details can be found in the manual of DBChIP and in Robinson et al. (2010).

1.2.2 Differential binding without replicates

ChIP-seq experiments without replicates are quite common. The dispersion parameters are expected to differ among experiments, and their estimation without any replication is nontrivial. We propose two methods that can effectively account for the potential over-dispersion. When there are more than two conditions, we can combine or "collapse" any two conditions into one condition and treat them as biological replicates. We then obtain an estimate of the common dispersion for each possible collapse through edgeR, and the final dispersion estimate is a certain quantile of all the dispersion estimates. If the binding sites across conditions are believed to be alike, median is the recommended quantile. In contrast, if sites are believed to be very different, we suggest using a smaller quantile, e.g., the lower quartile. We will refer to this strategy as the *collapsed quantile* method. We first evaluated the method on known null cases that are comparisons between biological replicates within the same condition. The test data is from a human ChIP-seq dataset studying NFκB transcription factor binding (Kasowski et al., 2010). We focused on three cell lines (GM10847, GM15510 and GM18505) with the largest number of biological replicates (5 each) in the study. For each cell line, we called peaks using SPP (Kharchenko et al., 2008) and kept only the top 5000 peaks. We performed all possible three-way comparisons within the total 5 replicates of a cell line, i.e., each time we randomly drew a unique combination of 3 replicates from the total 5 and treated them as coming from 3 different conditions. Figure 1 are the histograms for p-values: $(a - c)$ are the results of running the original edgeR, which assumes almost no dispersion when there are no replicates; $(d - f)$ are the results after adopting the collapsed quantile method. The p-values obtained through our collapsed quantile method behave comparably with the p-values obtained through edgeR when replicates are available.

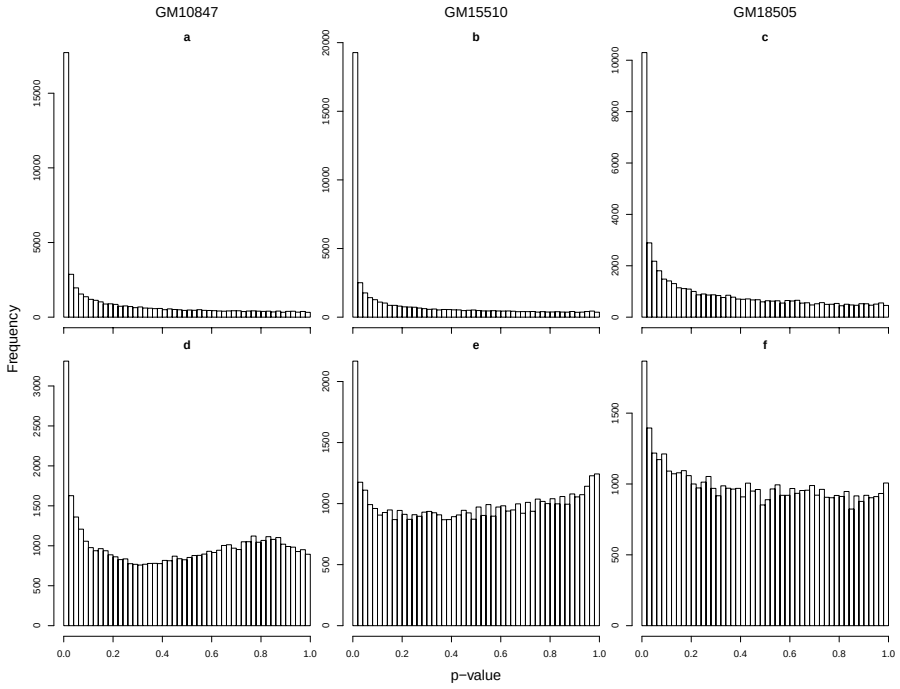

Figure 1: Histograms of true null p-values without replicates: upper row (a-c), ϕ estimated through edgeR; lower row (d-f), ϕ estimated through the collapsed quantile method.

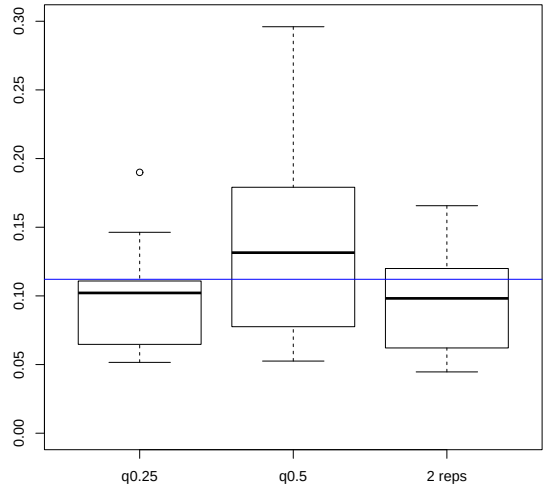

Figure 2: Estimation of the dispersion parameter in the absence of replicates. q0.25, lower quartile; q0.5, median; 2 reps, dispersion parameter estimated when there are 2 replicates in each cell line. All boxplots are based on 10 estimates. Blue horizontal line is the dispersion estimate using all 5 replicates.

To evaluate the effectiveness of the collapsed quantile method in a more realistic data, we randomly drew one replicate from each cell line and estimated the dispersion

parameter with quantile of the lower quartile and the median. Here, we focused on the top 5000 peaks in terms of total read count after merging peaks across three cell lines. As a comparison, we also sampled 2 replicates from each cell line and estimated the dispersion across conditions. Both processes were repeated 10 times to draw the boxplots in Figure 2. The blue horizonal line is the dispersion estimate using all 5 replicates and is treated as the reference. This result suggests that the collapsed quantile method performs reasonably well for estimating dispersion when there are no replicates. There are expected differences between the cell lines, and not surprisingly, the median estimates are slightly conservative; a quantile between the lower quartile and the median might have worked better. Despite the success in this simulation, the collapsed quantile method is designed to provide a rough estimate of the dispersion in the absence of replicates. Its performance will depend on the quantile parameter specified. Reliable estimates of the dispersion parameter will require two or more replicates.

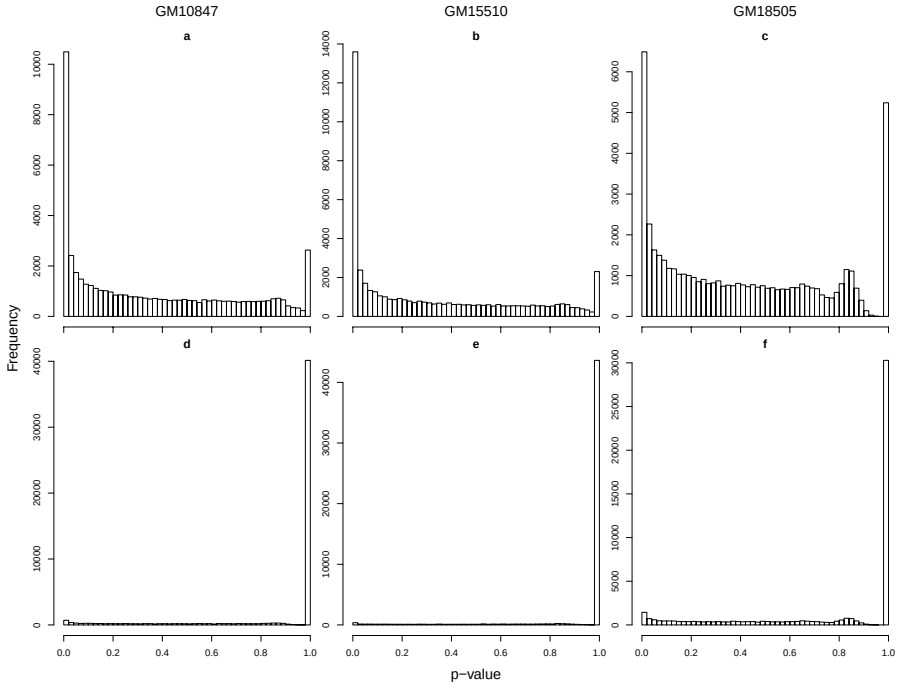

Figure 3: Histograms of p-values from comparisons of two biological replicates that are treated as from two different conditions without replicates: upper row (a-c), assuming Poisson distribution; lower row (d-f), testing composite null with allowable fold change 1.5.

For comparison of two conditions without replicates, we propose another method that tests a composite null instead of a point null under the Poisson distribution assumption. Define $r = n_1/n_2$ to be the expected fold change between the read counts of two conditions under the non-differential binding assumption. Then $x_{1j} \sim \text{Binom}(x_j, \pi_0)$ where $x_j = x_{1j} + x_{2j}$ and $\pi_0 = r/(1+r)$. This binomial model has been used in CisGenome (Ji et al., 2008) for detecting ChIP-seq binding sites. Instead of testing a point null that $r_j = r$ (equivalent to $\pi_{1j} = \pi_{2j}$), we could instead test a relaxed composite null as $H_0 : r/f \leq r_j \leq f \cdot r$ where f is a allowable fold change parameter. Usually researchers are mostly interested in differential binding with a large

fold change that can be validated by the qPCR assays. Thus, setting $f < 2$ is not expected to dampen the power of finding differential binding sites. Users can specify f according to their own preferences. The default value of f is 1.5, which worked well for all possible pairwise comparisons between biological replicates within all three cell lines. Figure 3 displays the histograms of the p-values for testing under the point null (the upper row) and the composite null (the lower row) and supports that testing composite null can effectively reduce the false positives that are due to unaccounted over-dispersion. Note that, unlike when testing a simple null hypothesis, the true null p-values from testing a composite null may not follow Uniform(0, 1) distribution any more.

1.2.3 Library size

When estimating library size for each ChIP sample, it is tempting to use the sequencing depth of each ChIP sample as the library size, n_i. However, sequencing depth can be highly influenced by high affinity differential binding sites. In many ChIP-seq datasets, the number of reads at the strongest binding site can be thousands of times larger than that of a weak binding site or a background region, and any potential differences in these strong binding sites will contribute substantially to the sequencing depth and lead to bias. DBChIP uses the median ratio strategy proposed by Anders and Huber (2010) to normalize the ChIP samples. More specifically, the ratio between any two samples is the median of the ratios between binding site counts and is robust to the fluctuations of the high affinity differential sites.

1.3 Control samples

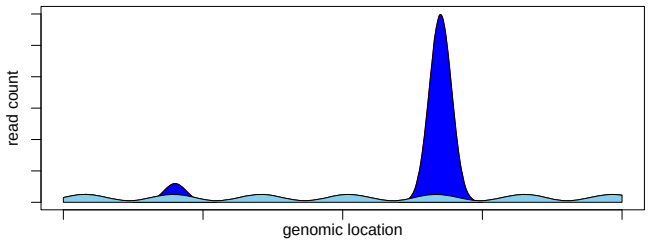

Figure 4: Pictorial depiction of signal-noise model where reads at a binding site are composed of binding signal reads (dark blue) and background noise reads (light blue). One weak binding site is shown on the left, and a strong binding site is on the right.

Figure 4 shows that ChIP sample reads at a binding site are composed of binding signal reads and background noise reads. The binding signal reads are due to the immunoprecipitation of the transcription factor of interest while the background noise reads are present even without the immunoprecipitation process. For more discussion of the signal-noise model, interested reader can refer to Kuan et al. (2011). The matching control samples provide natural estimates of the background read counts, and we can estimate the signal read count at a binding site as the difference between the ChIP sample read count and the corresponding normalized control sample read

count. To compensate for the differences between the matching ChIP and control samples, the control sample read counts need to be scaled by a normalization factor, which is routinely computed during peak-calling (Liang and Keleş, 2011). More specifically, we compute $y_{ij} = x_{ij} - f_i z_{ij}$, where f_i is the normalization factor between the ith ChIP sample and its matching control sample and z_{ij} is the corresponding read count at the location of the jth binding site in the ith control sample. Then the values of y_{ij} are rounded to their closest non-negative integers and treated as the estimate of the signal read count for site j of ChIP sample i. The normalization factors $(f_i, i = 1, \ldots, K)$ are computed by an estimator named NCIS (Normalization of ChIP-seq), which has been shown to have good statistical properties for estimating normalization factor (Liang and Keleş, 2011).

2 Details of the *C.elegans* Application

Zhong et al. (2010) identified binding sites using PeakSeq (Rozowsky et al., 2009) with p-value threshold of $1e$-5. PeakSeq outputs start and end positions of identified binding sites and does not report a most likely position for each binding event. We considered the PeakSeq peaks identified in the original publication and computed the most likely binding location for each peak as follows. We calculated the WTD score proposed in SPP (Kharchenko et al., 2008) for each nucleotide position within the region and picked the position with the highest WTD score as the predicted binding location. The WTD score is essentially a geometric mean of the read count on the positive strand within a upstream window of a putative position and the read count on the negative strand within a downstream window. In the subsequent cluster analysis, the WTD scores were used as the weights for binding sites.

Using the two replicates available per condition, we estimated the dispersion parameter to be 0.16. Our analysis indicated that 139 binding sites, which were identified as peaks under both conditions by PeakSeq, showed strong evidence of differential binding (FDR control at level 0.05). To further investigate this observation, we computed a fold change for each binding sites as the binding signal in embryonic condition divided by the binding signal in L1 condition. The sites showing favorable binding in embryonic condition had a median fold change of 3.7 whereas the median fold change was 0.28 for the sites with favorable binding in L1 starvation condition (hence L1/embryo fold change is 3.5). In summary, the median of fold change in both directions were both larger than 3.5 fold.

As discussed in the main text, we also identified many sites that were declared differentially bound based on PeakSeq output but did not pass our formal differential binding test. Figure 5 displays one such example where a site is declared bound only in the embryonic condition but not in L1 condition in the original paper. However, there is little evidence to reject the null hypothesis of non-differential binding at this site (p-value 0.98). In fact, this binding site is in the list of putative binding regions under L1 condition in the original PeakSeq output with a p-value of $5.9e$-5, missing the p-value threshold of $1e$-5. Thus, this site is likely a binding site in L1 condition

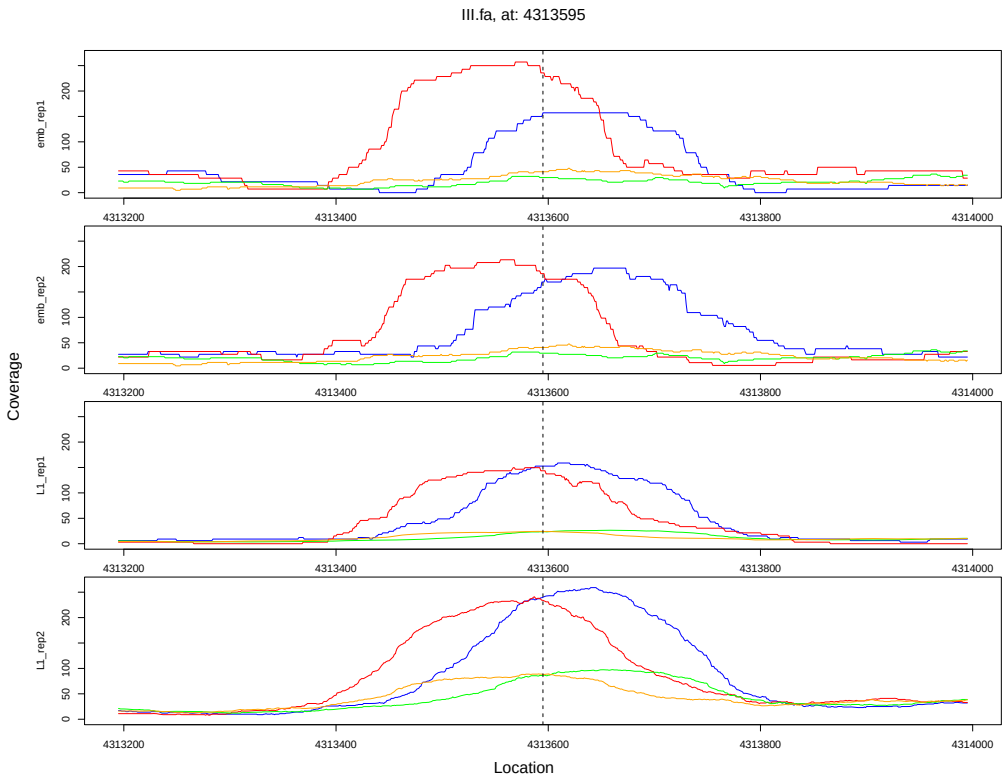

Figure 5: Comparison of coverage plots for a binding site at location 4313595 of chromosome III. This site was found to be differentially bound by Zhong et al. (2010) but shows no evidence of differentially binding by DBChIP. Each read is extended by 200 bp (default average fragment length) from its $5'$ end towards its $3'$ end. The coverage at each nucleotide is defined as the number of extended reads covering the position and is computed separately for ChIP sample forward strand (blue), ChIP sample reverse strand (red), control sample forward strand (green), control sample reverse strand (orange). The coverages in the plots are properly normalized with respect to their library sizes.

and does not exhibit differential binding between the conditions.

In contrast, the binding strengths of the sites that are common to both conditions can differ drastically. Figure 6 displays such an example where the p-value for testing non-differential binding is $8e$-5 and the fold change is estimated as 5.6 (L1 vs embryonic). The biological implications of such instances awaits further investigation.

3 Human NFκB Application

We next illustrate an application of DBChIP for studying differential binding in more than two conditions using a human NFκB ChIP-seq dataset (Kasowski et al., 2010). The original study has ten cell lines, and we focus on three of the cell lines (GM10847, GM15510, and GM18505) for illustration purposes. These three cell lines have the largest number of biological replicates in the study (5 each). We applied the SPP program on each cell line with an FDR level of 0.002 to obtain a list of binding sites.

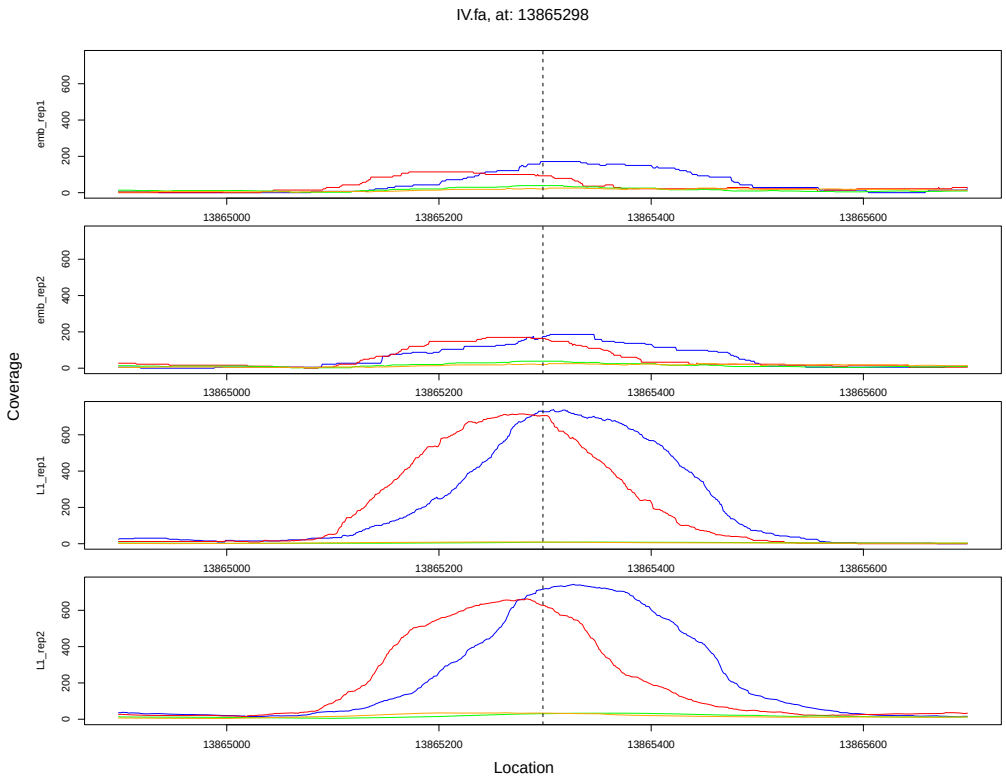

Figure 6: Comparison of coverage plots for a binding site at location 13865298 of chromosome IV. This site was found to be bound in both conditions by Zhong et al. (2010) but shows strong evidence of differentially binding by DBChIP.

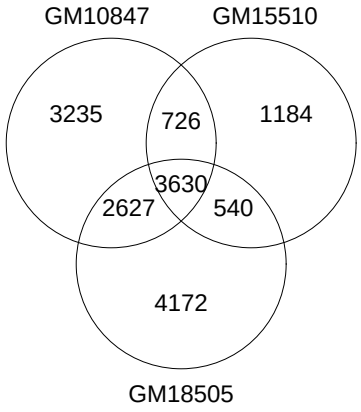

Figure 7: Shared binding sites among three human cell lines.

This resulted in 10532, 6295, and 11270 binding sites for the cell lines GM10847, GM15510, and GM18505, respectively. After merging the sites as described in Section 2 (Methods) of the main text, we compared the overlap of the binding sites by the venn diagram in Figure 7. Based on this simple overlap comparison, the binding patterns of these three cell lines appear to be very different. Among 16114 consensus sites, only 3630 (22.5%) sites are shared with all three cell lines, and the majority of

sites (8591) are declared bound only in one cell line. However, after formal testing of all consensus binding sites, only 1357 sites can be declared differentially bound after controlling FDR at level 0.01. Figure 8 displays the coverage profile of one of the most differentially bound site, in which binding signals are strong in two cell lines but clearly absent in the other. These discoveries can then be followed up for correlation with other genomic elements (e.g., single nucleotide polymorphism) and gene expression to elucidate the regulatory role of the transcription factor. To reveal the relationship between the original overlaps of the binding sites depicted in Figure 7 and the differential binding through DBChIP, we further break down the 1357 sites that are declared to be differentially bound by DBChIP: 1106 of them are originally declared bound in only one cell line, 241 sites in two cell lines, and only 10 sites in all three cell lines. Not surprisingly, the fewer conditions a site is found to be bound, the more likely it is a differential binding site. However, even for those sites found to be bound in only one cell line, the majority of them are not likely to be differentially bound (4854 out of 8591 sites have p-values larger than 0.05).

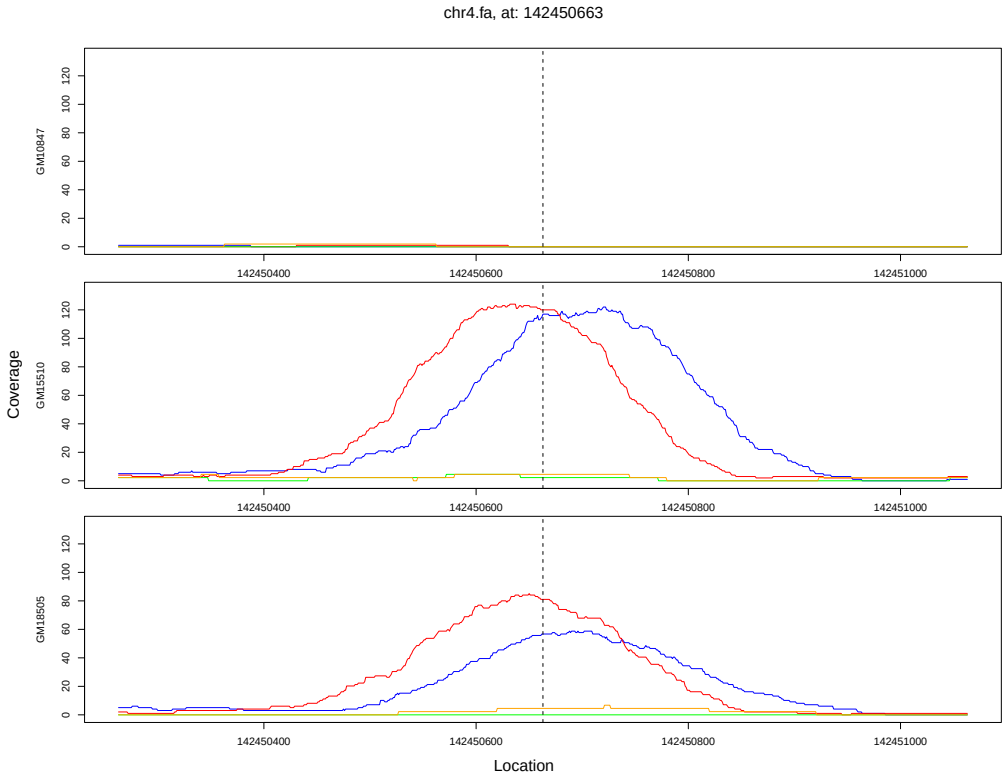

Figure 8: An example differential binding site among the three human cell lines. The five biological replicates within each cell line are combined for this plot.

References

Agresti, A. (2002), *Categorical Data Analysis*, Wiley.

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

Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., Marshall, K., Phillippy, K., Sherman, P., et al. (2011), "NCBI GEO: archive for functional genomics data sets–10 years on," *Nucleic Acids Research*, 39, D1005.

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

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

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S., Habegger, L., Rozowsky, J., Shi, M., Urban, A., et al. (2010), "Variation in transcription factor binding among humans," *Science*, 328, 232–235.

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

Kuan, P., Chung, D., Pan, G., Thomson, J., Stewart, R., and Keleş, S. (2011), "A statistical framework for the analysis of ChIP-Seq data," `http://pubs.amstat.org/doi/abs/10.1198/jasa.2011.ap09706`, to appear in *Journal of the American Statistical Association*.

Liang, K. and Keleş, S. (2011), "Normalization of ChIP-seq data with control," Tech. Rep. 224, University of Wisconsin, Madison, Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, `http://www.biostat.wisc.edu/Tech_Reports/pdf/tr_224.pdf`.

McCullagh, P. and Nelder, J. (1989), *Generalized linear models*, Chapman & Hall/CRC.

Robinson, M., McCarthy, D., and Smyth, G. (2010), "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, 26, 139.

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

Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R., and Sidow, A. (2008), "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data," *Nature methods*, 5, 829–834.

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

Zhong, M., Niu, W., Lu, Z., Sarov, M., Murray, J., Janette, J., Raha, D., Sheaffer, K., Lam, H., Preston, E., et al. (2010), "Genome-wide identification of binding sites defines distinct functions for Caenorhabditis elegans PHA-4/FOXA in development and environmental response," *PLoS Genet*, 6, e1000848.