DRAFT

High Resolution Identification of Protein-DNA Binding Events and Quality Control for ChIP-exo data

Dongjun Chung[6], Rene Welch[1], Irene Ong[3], Jeffrey Grass[3,4], Robert Landick[3,4,5] and Sündüz Keleş[1,2*]

*Correspondence:
keles@stat.wisc.edu
[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI
Full list of author information is available at the end of the article

Abstract

Recently, ChIP-exo has been developed to investigate protein-DNA interaction in higher resolution compared to popularly used ChIP-Seq. Although ChIP-exo has drawn much attention and is considered as powerful assay, currently, no systematic studies have yet been conducted to determine optimal strategies for experimental design and analysis of ChIP-exo. In order to address these questions, we evaluated diverse aspects of ChIP-exo and found the following characteristics of ChIP-exo data. First, Background of ChIP-exo data is qute different from that of ChIP-Seq data. However, sequence biases inherently present in ChIP-Seq data still exist in ChIP-exo data. Second, in ChIP-exo data, reads are located around binding sites much more tightly and hence, it has potential for high resolution identification of protein-DNA interaction sites, hence the space to allocate the reads is greatly reduced. Third, although often assumed in the ChIP-exo data analysis methods, the peak pair assumption does not hold well in real ChIP-exo data. Fourth, spatial resolution of ChIP-exo is comparable to that of PET ChIP-Seq and both of them are significantly better than resolution of SET ChIP-Seq. Finally, for given fixed sequencing depth, ChIP-exo provides higher sensitivity, specificity, and spatial resolution than PET ChIP-Seq.

In this article, we provide a quality control pipeline which visually asses ChIP-exo biases and calculates a signal-to-noise measure. Also, we updated dPeak [1], which makes a striking balance in sensitivity, specificity, and spatial resolution for ChIP-exo data analysis.

Keywords: ChIP-exo; QC; TFBS; BS identification on High-Res

* to remove later

Contents

1 Background

ChIP-exo (Chromatin Immunopecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh ([2]) is the state-of-the-art experiment developed to attain single base-pair resolution of protein binding site identification and it is considered as a powerful alternative to popularly used ChIP-Seq (Chromatin Immunoprecipitation coupled with next generation sequencing) assay. ChIP-exo experiments first capture millions of DNA fragments (150 - 250 bp in length) that the protein under study interacts with using random fragmentation of DNA and a protein-specific antibody. Then, exonuclease is introducted to trim 5' end of each DNA fragment to a fixed distance from the bound protein. As a result, boundaries around the protein of interest constructed with 5' ends of fragments are located much closed to bound protein compared to ChIP-Seq. This is the step unique to ChIP-exo that could potentially provide significantly higher spatial resolution compared to ChIP-Seq. Finally, high throughput sequencing of a small region (25 to 100 bp) at 5' end of each fragment generates millions of reads or tags.

While the number of produced ChIP-exo data keeps increasing, characteristics of ChIP-exo data and optimal strategies for experimental design and analysis of ChIP-exo data are not fully investigated yet, including issues of sequence biases inherent to ChIP-exo data, choice of optimal statistical methods, and determination of optimal sequencing deoth. However, currently, the number of available ChIP-exo data is still limited and their sequencing depths are still insufficient for such investigation. To address this limitation we gathered ChIP-exo data from diverse organisms: CTCF factor in human [2]; ER factor in human and FoxA1 factor in mouse from [3]; and generated σ^{70} factor in Escherichia coli (E. Coli) under aerobic ($+O_2$) condition, and treated by rifampicin by 0 and 20 minutes.

DNA libraries generated by the ChIP-exo protocol seem to be less complex than the libraries generated by ChIP-Seq [4]. Hence, most of current guidelines [5] may not be aplicable on ChIP-exo, additionally to our knowledge there are not established quality control pipelines for ChIP-exo. To address this challenge, we suggest a series of quality control visualizations to understand which biases are present in ChIP-exo data. Related to quality control, Previous ChIP-exo analysis used ChIP-Seq samples to compare the resolution between experiments ([2], [6], [3]). In [7], Carroll et. al. studied the use of the Strand-Cross Correlation (SCC) [8]. and showed that by filtering blacklisted regions the estimation of the SCC is improved. However, this method requires to know blacklisted regions in advance which may not be available. In our pipeline we propose two out-the-shelf metrics equivalent to RSC and NSC for the estimation of the signal-to-noise ratio for a ChIP-exo sample.

In order to archieve the potential benefits of ChIP-exo on protein binding site identification, it is critical to understand which are the important characteristics of ChIP-exo data and to use algorithms that could fully utilize information available in ChIP-exo data. Rhee and Pugh [2] discussed that reads in the forward and reverse strand might construct peak pairs around bound protein, of which heights were implicitly assumed to be symmetric. Hence, they used the "peak pair method" that predicts the midpoint of two modes of peak pairs as potential binding site. Mace [9], CexoR [10] and peakzilla [11], recently developed ChIP-exo data analysis methods, are also based on this peak pair assumption. However, appropriatness of

such assumption was not fully evaluated in the literature yet. Furthermore, it is still unknown which factors could affect protein binding site ientification using ChIP-exo data. In order to address this problem, we investigated various aspects of ChIP-exo data by contrasting them with their respective ChIP-Seq experiments.

Currently, research on statistical methods for ChIP-exo data is still in its very early stage. Although many methods have been proposed to identify protein binding sites from ChIP-Seq data (reviewed in [12] and [13]), such as MACS [14], CisGenome [15] and MOSAiCS [16], these approaches reveal protein binding sites in lower resolution, i.e., at an interval of hundreds to thousands of base pairs. Furthermore, they report only one "mode" or "predicted binding location" per peak. Hence, these methods are not appropiate to evaluate the potential of ChIP-exo data for high resolution identification of protein binding sites. More recently, deconvolution algorithms such as CSDeconv [17], GEM [18] (an improved version of [19]) and PICS [20] have been proposed to identify binding sites in higher resolution using ChIP-Seq data. However, most of them are still not tailored for ChIP-exo and PET and SET ChIP-Seq data in a unified framework and as a result, currently available methods are not appropiate for fair comparison between ChIP-exo and ChIP-Seq. To address these limitations, we developed an improved of dPeak [1], a high resolution binding site identification (deconvolution) algorithm that we previously developed for PET and SET ChIP-Seq data, so that it can also handle ChIP-exo data. The dPeak algorithm implements a probabilistic model that accurately describes the ChIP-exo and ChIP-Seq data generation process.

In this paper, we demonstrate that the peak pair assumption of Rhee and Pugh [2] does not hold well in real ChIP-exo data. Furthermore, we found that when we analyze ChIP-exo data from eukaryotic genomes, it is important to consider sequence biases inherent to ChIP-exo data, such as mappability and GC content in order to improve sensitivity and specificity of binding site identification. dPeak outperforms or performs competitively with ChIP-exo data analysis such as GEM and MACE. More importantly, when comparable number of reads is used for both ChIP-exo and ChIP-Seq, dPeak couple with ChIP-exo data provides resolution comparable to PET ChIP-Seq and both significantly improve the resolution of protein binding site identification compared to SET-based analysis with any of the available methods.

2 Results and discussion

2.1 Deeply sequenced E.Coli σ^{70} ChIP-exo and ChIP-Seq data

σ^{70} factor is a transcription initiation factor of housekeeping genesin E. coli. In E. coli genomes, many promoters contain multiple transcription start sites (TSS) and these TSS are often closely space (10 150 bp). These closely spaced binding sites are considered to be multiple "switches" that differentially regulate gene expression under diverse growth conditions [21], [22]. Therefore, investigation of ChIP-exo's potential for identification and differentiation of closely spaced binding sites are invaluable for elucidating the transcriptional networks of prokaryotic genomes.

2.2 Current guidelines and quality metrics

To have a first view on the data, we calculated a series of the most commonly used quality indicators for ChIP-Seq data, which are listed in table 1. The PBC, NSC and RSC were calculated as defined in [5] and the SSD was calculated as defined in [23]. Additionally, we calculated the FSR to compare the number of fragments in each strand. A few conclussions that we may get by observing the PBC in table 1 are all the samples have a quite low library complexity, hence it would be suggested to repeat the experiment.

Condition	Rep.	Nr. reads	PBC	SSD	FSR	NSC
Aerobic	1	1,454,566	0.2156	0.0830	0.5012	114.14
Aerobic	2	864,714	0.2471	0.0488	0.4923	160.93
Rif-0min	1	960,256	0.2823	0.0361	0.4998	10.29
Rif-0min	2	2,247,295	0.2656	0.1091	0.5047	25.08
Rif-20min	1	1,940,387	0.2698	0.0820	0.5054	17.69
Rif-20min	2	4,229,574	0.2153	0.1647	0.5071	14.11

Table 1 Usual quality control indicators applied to our σ^{70} samples. PBC stands PCR-bottleneck coefficient, SSD for standarized standard deviation, FSR for forward strand ratio, NSC for normalized strand cross-correlation coefficient. We ommited the RSC (relative strand cross-correlation since ChIP-exo experiments don't usually have an acomnying input file)

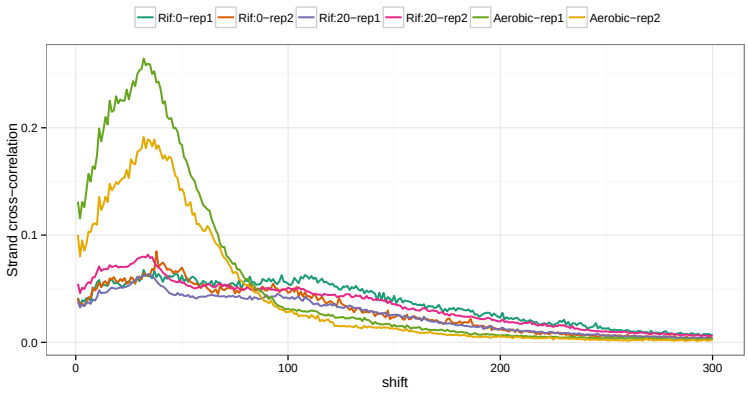

Figure 1 Strand cross-correlation curves for σ^{70} samples. The "phantom peak" and the summit that corresponds to the read and fragment length respectively are confounded, due to the exonuclease enzyme digestion.

In a good ChIP-Seq experiment the strand cross-correlation is maximized at the fragment length and there is also a "phantom peak" at the sample's read length

([5] and [8]). However, in ChIP-exo's case this two peaks are confounded. Hence an enrichment measure based in the strand cross-correlation such as NSC or RSC is harder to interpret. In figure 1 we plotted the SCC curves for σ^{70} ChIP-exo experiments used to calculate the NSC and RSC values in table 1. According to ENCODE's quality metrics, NSC values close to one corresponds to low enrichment samples or bad quality datasets, hence in our case the last four samples would be considered to have higher quality than they actually have. On the other hand, all PBC values are considerably low and correspond to samples with severe bottlenecking issues, in particular we are going to show that for the first two samples we are obtaining a smaller resolution and identifying a comparable number of targets as in ChIP-Seq PET experiments.

2.3 ChIP-exo Quality Control pipeline

Figure 2 shows a flowchart for the ChIPexoQC pipeline. Which basically partitions genome by keeping the non-diggested ChIP-exo regions. Then, for each region it calculates a series of independent statistics. Finally, it creates several visualizations designed to asses the overall quality levels of A ChIP-exo sample.

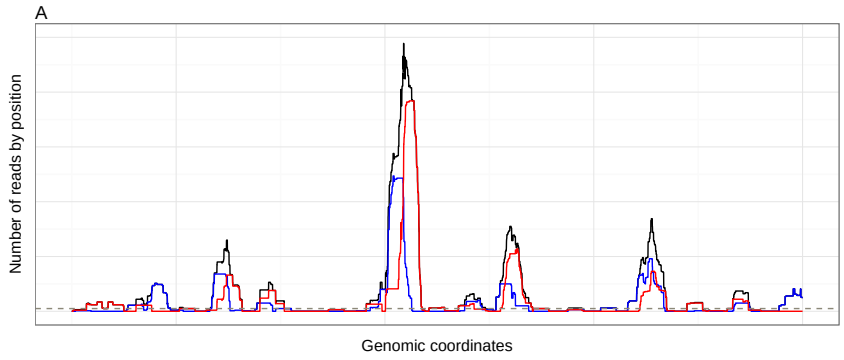

Figure 2 Diagram of the ChIP-exo Quality control pipeline(this is a place-holder). The genome is partitioned into several regions by removing the empty segments and for each regions we calculate a collection of qc indicators.

2.3.1 Enrichment and library complexity in ChIP-exo data

In ChIP-exo experiments, the background is often digested by the exonuclease enzyme, therefore to determine the sample's quality is necesary to address the balance between the enrichment and library complexity. To diagnose this, we considered the Average Read Coverage (ARC) and the Unique Read Coverage Rate (URCR) which are defined as:

$$\text{ARC} = \frac{\text{Nr. of reads in the region}}{\text{Width of the region}}$$

$$\text{URCR} = \frac{\text{Nr. of reads mapped to only one position in the region}}{\text{Nr. of reads in the region}}$$

Figure 3A shows hexbin plots illustrating the interaction between these two quantities. Notice how there are two strong arms in each panel: The first one corresponds to regions with low ARC values and URCR values that can vary accross the whole $[0, 1]$ interval, while the second one shows a decreasing trend in URCR as ARC increases. When an experiment shows a higher degree of enrichment, then the separation of this two arms is more noticeable, since the second arm corresponds to possibly enriched regions (7A and 7B).

Figure 3B shows boxplots of the local-NSC stratified by the number of reads mapped exactly to one position for all σ^{70} ChIP-exo experiments; for each stratum, we sampled 100 regions that whenever possible (in the opposite case, all the regions in the category were considered). There is an increasing trend as the number of unique position increases, which means that this is an effective indicator of library complexity since regions where the reads are mapped to more positions are usually more complex.

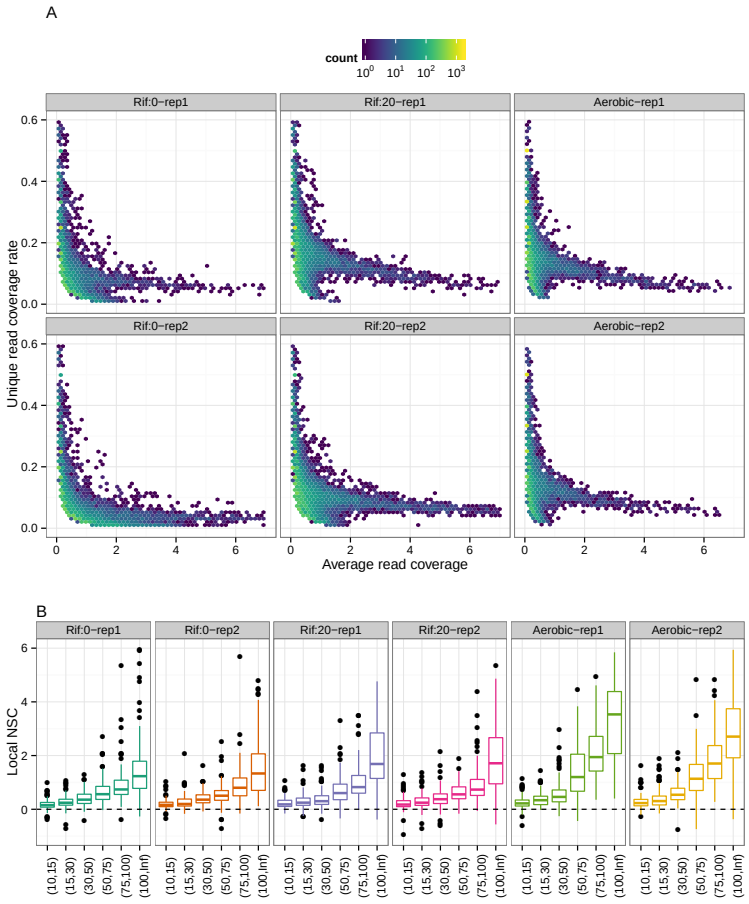

Figure 3 A) Hexbin plots of ARC against URCR and B) boxplots of the local-NSC stratified by nr. of reads mapped to only one position in the regions for the ChIP-exo σ^{70} experiments divided by condition and biological replicate

2.3.2 Strand imbalance in ChIP-exo data

The strand imbalance assesment is based in the observation that the enriched regions usually have a higher concentration of fragments, therefore we examine the FSR (defined as the ratio of the number of forward stranded reads divided by the total number of reads in a given region) as the region with lower depth are being filtered out. This indicator is of particular importance, since several methods rely on the "peak-pair" assumption. In table 1, we calculated the FSR and noticed that for all the samples, it's value is close to 0.5, which means that there are roughly the same amount of reads in both strands. However, figure 4B shows that this assumption don't hold in practice.

2.4 Comparison with ChIP-Seq data

We first compared various factors that could affect binding site identification between ChIP-exo and ChIP-Seq data. in order to compare distribution of signal and background between ChIP-exo and ChIP-Seq data, we calculated ChIP tag counts accross the genome by counting the number of reads mapping to each of 150 non-overlapping window after extending reads by 150 to their 3' end directions. ChIP tag counts in ChIP-exo data were linearly related to ChIP tag counts in ChIP-Seq data for the regions with high ChIP tag counts (Figure 4A). This implies that signals for pontential binding sites are well reproducible between ChIP-exo and ChIP-Seq data. On the other hand, there was clear difference in the background distribution between them. In ChIP-Seq data reads were almost uniformly distributed over background (non-binding) regions and the ChIP tag counts in there regions were significantly larger than zero. In contrast, in ChIP-exo data, there was larger variation in ChIP tag counts among background regions and ChIP tag counts were much lower in these regions compared to ChIP-Seq data. There were also large proportion of regions without any read in ChIP-exo data. These results indicate that background distribution of ChIP-exo data is less homogeneous than that of ChIP-Seq data.

We next evaluated the "peak pair" assumption of Rhee and Pugh [2], MACE [9], CexoR, [10] and peakzilla [11], i.e. a peak of reads in the forward strand is usually paired with a peak of reads in the reverse strand that is located in the other site of the binding site. In order to evaluate this assumptiom, we reviewed the proportion of reads in the forward streand in candidate regions (i.e. regions with at least one binding site) in σ^{70} ChIP-exo data. We found that strands of reads were much less balanced in ChIP-exo data than in ChIP-Seq data in these regions with potential binding sites (Fig. 4B) and htis indicates that the peak pair assumptuon might not hold in real ChIP-exo data.

2.4.1 Comparison with ChIP-Seq data using dPeak

2.5 Recomendations for the design of ChIP-exo experiments

We sampled a fixed amoung of fragments for each of the ChIP-exo, PET ChIP-Seq and SET ChIP-Seq datasets of the σ^{70} sample in aerobic conditions. For each sampled dataset we applied our lower-to-higher resolution pipeline by calling peaks with MOSAiCS [16] and then deconvolving the binding events by using dPeak [1].

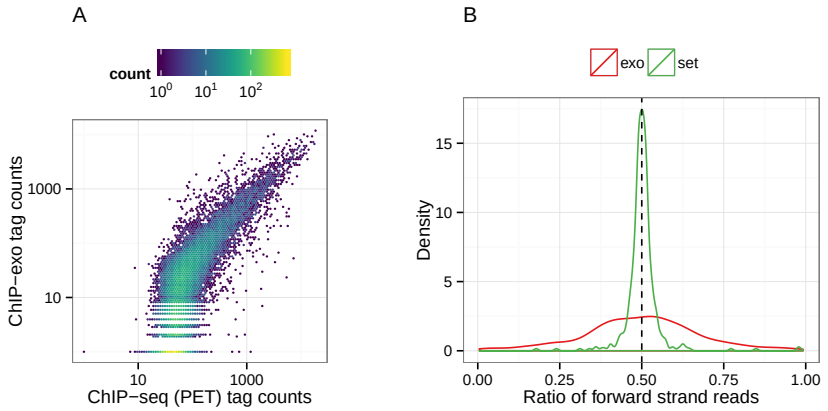

Figure 4 A) ChIP signal in ChIP-exo was linearly related to that in PET ChIP-Seq data in the region with high ChIP tag counts. In contrast, there wer clear differences in their background distribution, where several ChIP-exo background regions were empty. B) Density of the proportion of the ratio of forward strand reads in a ChIP-exo and ChIP-Seq SET data.

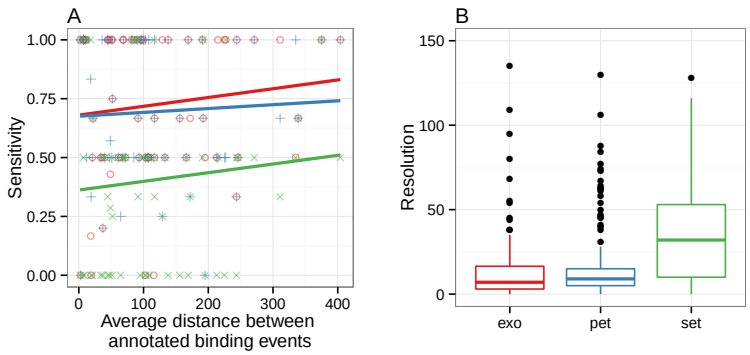

Figure 5 Comparison of (A) sensitivity and (B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the propotion of regulonDB annotations identificed using each data. Resolution is defined as the distance between regulonDB annotation and its closest prediction.

For the ChIP-exo datasets we called peaks by using GC-content and mappability with MOSAiCS, and for the ChIP-Seq datasets we used their respective Input samples.

Figure 6 shows the behaviour of each data type when their depth is fixed. It is remarkable that even when the number of candiate peaks or the number of predicted events in quite lower than for both ChIP-Seq cases, it outperforms them in the number of identified targets and in resolution. This may suggest that with ChIP-exo less false positive peaks are being called and that when the targets are being identified, dpeak is able to estimate binding locations closer to the actual true cases. Additionally, we can see that as the read depth increases all four indicators do so as well, which may indicate that with ChIP-exo a smaller amount of reads is necessary to identify a higher number of targets, but it may also be possible that this is an artifact occurring due to ChIP-exo's lower library complexity.

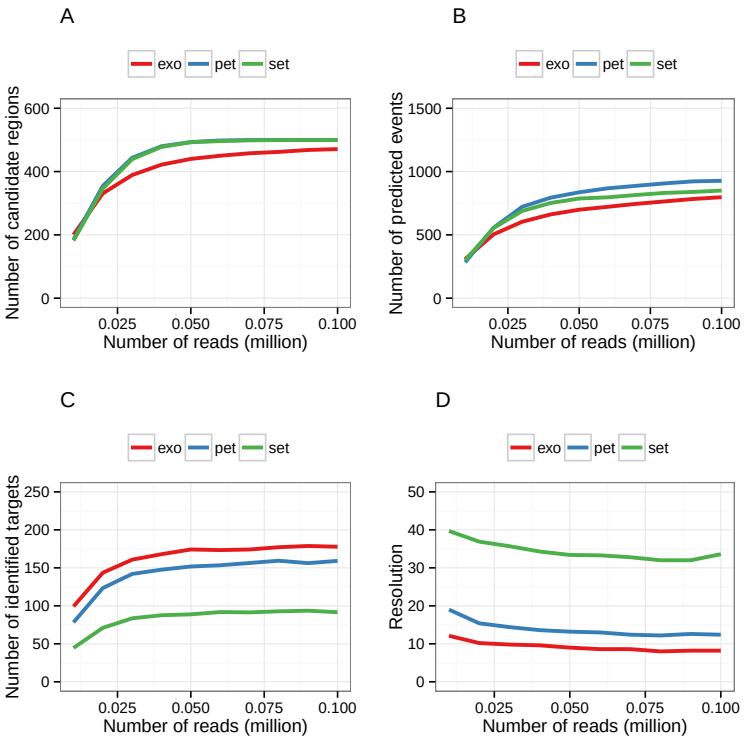

Figure 6 Comparison of the number of candidate regions (A), predicted events (B), identified targets (C) and resolution (D) among ChIP-exo, PET ChIP-Seq and SET ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard was considered identified if a BS was estimated inside a 15 bp window around it

3 Methods

3.1 Growth conditions

3.2 ChIP-exo experiments

3.3 Library preparation, sequencing and mapping of sequencing reads

3.4 Method comparison with ChIP-exo

We considered dPeak Chung et al. [1], GEM Guo et al. [18] and MACE Wang et al. [9] for ChIP-exo data analysis. For the dPeak algorithm, we used the R package *dPeak* version 1.0.0, which is available from http://dongjunchung.github.io/dpeak/. For the GEM algorithm we used its Java implementation version 0.9, which is available from http://cgs.csail.mit.edu/gem/. For the MACE algorithm, we used ity Python implementation version 1.0, which is available from http://dldcc-web.brc.bcm.edu/lilab/MACE/docs/html/. Candidate regions for dPeak were identified for each replicate of ChIP-exo data using the MOSAiCS algorithm Kuan et al. [16] (one sample analysis using false discovery rate of 0.0001) implemented as an R package *mosaics* 2.4 (available from *Bioconductor* https://www.bioconductor.org/packages/release/bioc/html/mosaics.html). We further filtered out candidate region with average ChIP tag count less than 3,000 to avoid potential false positives based in exploratory analysis. These regions were also explicitly provided to the GEM algorithm as candidate regions. Default tuning parameters were used duging model fitting for all methods.

3.5 Local-NSC

The local strand cross-correlation is defined as:

$$\text{local-NSC} = \frac{max_{x_\delta} f(x_\delta)}{\sigma}$$
$$y_\delta = f(x_\delta) + \sigma\epsilon_\delta$$

where f is calculated by fitting a local polynomial regression model over the strand cross-correlation calculated using exclusively the fragments that overlap a pre-defined genomic region, y_δ is the correlation of the forward and backward coverages when one is shifted x_δ bp towards their 3' end.

The local polynomial model was fitted using the *loess* function from the R software version 3.2.1 using the default tunning parameters.

Author details
[1]Department of Statistics, University of Wisconsin Madison, 1300 University Avenue, Madison, WI. [2]Department of Biostatistics and Medical Informatics, University of Wisconsin Madison, 600 Highland Avenue, Madison, WI. [3] Great Lakes Bioenergy Research Center, University of Wisconsin Madison, 1552 University Avenue, Madison, WI. [4] Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI. [5] Department of Bacteriology, University of Wisconsin Madison, 1550 Linden Drive, Madison, WI. [6] Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Charleston, SC.

References

1. Chung, D., Park, D., Myers, K., Grass, J., Kiley, P., Landick, R., Keleş, S.: dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. PlOS, Computational Biology (2013)
2. Rhee, H.S., Pugh, F.: Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. Cell (2011)
3. Serandour, A., Gordon, B., Cohen, J., Carroll, J.: Development of and illumina-based chip-exonuclease method provides insight into foxa1-dna binding properties. Genome Biology (2013)
4. Mahony, S., Franklin, P.B.: Protein-dna binding in high-resolution. Critical Reviews in Biochemistry and Molecular Biology (2015)
5. Landt, S., Marinov, G., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B., Bickel, P., Brown, J., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, C., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A., Hoffman, M., Iyer, V., Jung, Y., Karmakar, S., Kellis, M., Kharchenko, P., Li, Q., Liu, T., Liu, S., Ma, L., Milosavljevic, A., Myers, R., Park, P., Pazin, M., Perry, M., Raha, D., Reddy, T., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J., Tolstorukov, M., White, K., Xi, S., Farnham, P., Lieb, J., Wold, B., Snyder, M. Genome Research (2012)
6. Rhee, H.S., Pugh, F.: Chip-exo a method to identify genomic location of dna-binding proteins at near single nucleotide accuracy. Current Protocols in Molecular Biology (2012)
7. Carroll, T., Liang, Z., Salama, R., Stark, R., de Santiago, I.: Impact of artifact removal on chip quality metrics in chip-seq and chip-exo data. Frontiers in Genetics, Bioinformatics and Computational Biology (2014)
8. Kharchenko, P., Tolstorukov, M., Park, P.: Design and Analysis of ChIP-seq Experiments for DNA-binding Proteins
9. Wang, L., Chen, J., Wang, C., Uusküla-Reimand, L., Chen, K., Medina-Rivera, A., Young, E.J., Zimmermann, M.T., Yan, H., Sun, Z., Zhang, Y., Wu, S.T., Huang, H., Wilson, M.D., Kocher, J.-P.A., Li, W.: Mace: model based analysis of chip-exo. Nucleic Acids Research (2014)
10. Madrigal, P. EMBnet.journal (2015)
11. Bardet, A.F., Steinmann, J., Bafna, S., Knoblich, J.A., Zeitlinger, J., Stark, A.: Identification of transcription factor binding sites from chip-seq data at high resolution. Bioinformatics (2013)
12. Wilbanks, E., Facciotti, M.: Evaluation of algorithm performance in chip-seq peak detection. PlOS One (2012)
13. Pepke, S., Wold, B., Ali, M.: Computation for chip-seq nad rna-seq studies. Nature (2009)
14. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D., Bernstein, B., Nausbam, C., Myers, R.M., Brown, M., Li, W., Liu, X.S.: Model-based analysis of chip-seq (macs). Genome Biology (2008)
15. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M., Wong, W.H.: An integrated software system for analyzing chip-chip and chip-seq data. Nature biotechnology (2008)
16. Kuan, P.F., Chung, D., Pan, G., Thomson, J.A., Stewart, R., Keleş, S.: A statistical framework for the analysis of chip-seq data. Journal of the American Statistical Association (2009)
17. Lun, D.S., Sherrid, A., Weined, B., Sherman, D.R., Galagan, J.E.: A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from chip-seq data. Genome Biology (2009)
18. Guo, Y., Mahony, S., Gifford, D.K.: High resolution genome wide binding event finding and motif discovery reals transcription factor spatial bindings constraints. PlOS, Computational Biology (2012)
19. Guo, Y., Papachristoudis, G., Altshuler, R.C., Gerber, G.K., Jaakkola, T.S., Gifford, D.K., Mahony, S.: Discovering homotypic binding events at high spatial resolution. Bioinformatics (2010)
20. Zhang, X., Robertson, G., Krzwinski, M., Ning, K., Droit, A., Jones, S., Gottardo, R.: Pics: Probabilistic inference for chip-seq. Biometrics (2010)
21. Mendoza-Vargas, A., Olvera, L., Olvera, M., Grande, R., Vega-Alvarado, L., Taboada, B., Jimenez-Jacinto, V., Salgado, H., Juárez, K., Contreras-Moreira, B., Huerta, A.M., Collado-Vides, J., Morett, E.: Genome-wide identification of transcription start sites, promotes and transcription factor binding sites in e. coli. PLOS one (2009)
22. Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muñiz-Rascado, L., García-Sotelo, J.s., Weiss, V., Solano-Lira, H., Martínez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernández, S., Alquicira-Hernández, K., López-Fuentes, P.-S.L. Alejandra, Huerta, A.M., Bonavides-Martínez, C., Balderas-Martínez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jiménez-Jacinto, V., Vega-Alvarado, L., del Moral-Chávez, V., Hernández-Alvarez, A., Morett, E., Collado-Vides, J.: Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more
23. Planet, E., Attolini, C.S.-O., Reina, O., Rosell, D.: htseqtools: high-throughput sequencing quality control, processing and visualization in r. Bioinformatics (2011)
24. Mendenhall, E.M., Bernstein, B.E.: Dna-protein interactions in high definition. Genome Biology (2012)
25. Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., Speed, T.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (2003)
26. Bolstad, B., Irizarry, R., Âstrand, M., Speed, T.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics (2003)

Supplement

Additional Enrichment plots for σ^{70}

In figure 7 it is shown the same relationshtip as in figure 3, but considering only regions formed where the reads are being allocated in more than 10 (A) and 30 (B) unique positions respectively. Both plots show that the vertical arms formed by regions with low ARC is formed by low complexity regions, that way suggesting that this segment correspond to the background of a ChIP-exo experiment.

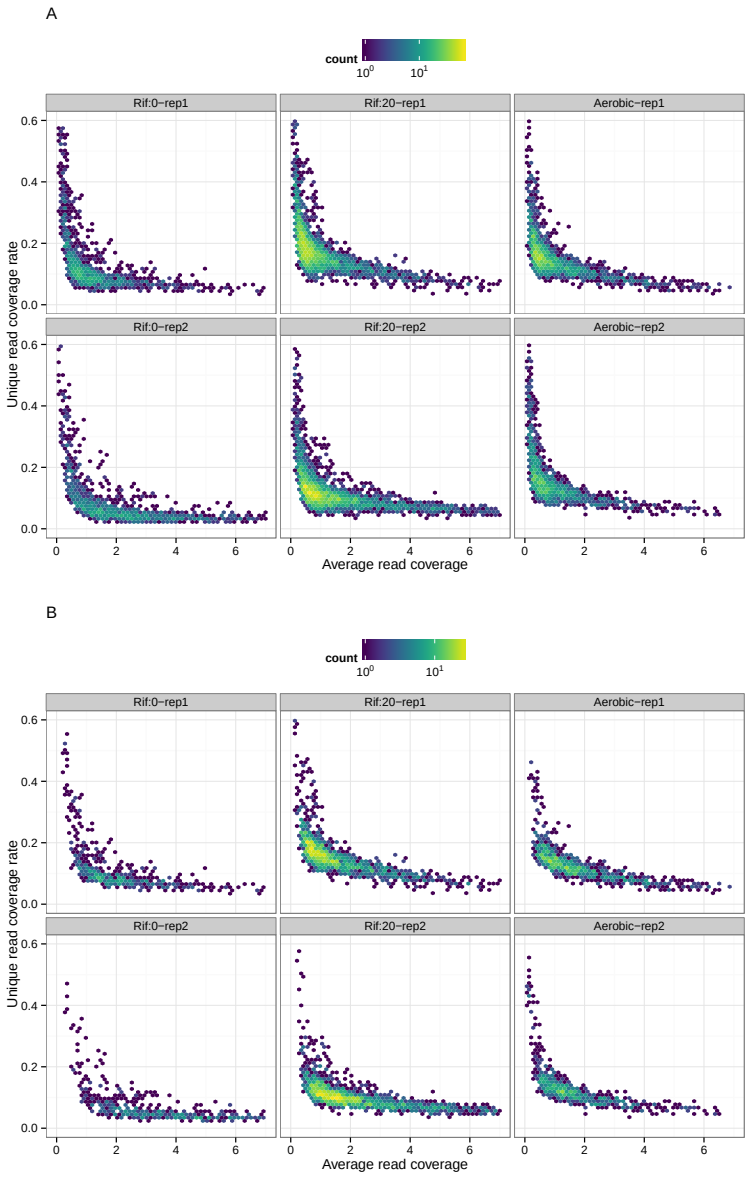

Figure 7 Hexbin plots of ARC vs URCR for each region after partitioning the genome. In A and B the regions with reads mapped to at most 10 and 30 positions respectively where not considered. ARC is defined as the ratio of the nr. of reads and the width of a region and URCR is the ratio of the number of unique position where the reads are being allocated and the number of reads in a region.