DRAFT

Data exploration, quality control and statistical analysis of ChIP-exo experiments

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

ChIP-exo is a modification of the ChIP-Seq protocol for high resolution mapping of transcription factor binding sites. Although many aspect of the ChIP-exo data analysis are similar to those of ChIP-Seq, ChIP-exo presents a number of unique challenges. We present a quality control pipeline that analyzes ChIP-exo's strand imbalance and library complexity. Assessment of these biases and artifacts are facilitated through diagnostic plots and summary statistics calculated over regions of the genome with varying levels of coverage.

We systematically evaluated diverse aspects of ChIP-exo and found the following characteristics: First, ChIP-exo's background is quite different from ChIP-Seq's. Second, although often assumed in ChIP-exo data analysis methods, the "peak pair" assumptions does not hold locally in actual ChIP-exo data. Third, we for the first time compared Paired End (PE) ChIP-Seq with ChIP-exo and found that both protocols are comparable in resolutions and sensitivity for closely located binding events, but as the distance between binding events increases ChIP-exo shows higher sensitivity that PE ChIP-Seq. Finally, at fixed sequencing depths, ChIP-exo provides higher sensitivity, specificity and spatial resolution than PE ChIP-Seq.

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

1 Background

ChIP-exo (Chromatin Immunoprecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh, 2011 [1] 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 introduced to trim the 5' end of each DNA fragment to a fixed distance from the bound protein compared to ChIP-Seq. This step is unique to ChIP-exo and could potentially provide significantly higher spatial resolution compared to ChIP-Seq. Finally, high throughput sequencing of a small region (25 to 100 bp) at the 5' end of each fragment generates millions of reads.

While the number of ChIP-exo data keeps increasing,characteristics of ChIP-exo data are not fully investigated yet. First, DNA libraries generated by the ChIP-exo protocol seem to be less complex than the libraries generated by ChIP-Seq (Mahony et al., 2015, [2]), i.e. the possible number of positions to which the reads can be aligned has been reduced due to the exonuclease digestion. Second, although there are roughly the same amount of read in both strands, locally there may be more reads in one strand than in the other. To address this challenges, we suggest a collection of quality control visualizations to understand which of this biases are present in a ChIP-exo experiment and globally asses the enrichment and library complexity of a ChIP-exo sample. We gathered ChIP-exo data from diverse organisms: CTCF factor in human [1]; ER factor in human and FoxA1 factor in mouse (Serandour et al., 2013 [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 (courtesy of Professor Robert Landick's lab).

Most of current ChIP-Seq quality control (QC) guidelines (Landt et al., [4]) may not be applicable on ChIP-exo, additionally to our knowledge there are not established QC pipelines for ChIP-exo. Previous ChIP-exo analyses used ChIP-Seq samples to compare the resolution between experiments ([1], [5], [3]); Carroll et al., 2014 [6] studied the use of the Strand-Cross Correlation (SCC) (Kharchenko et al., 2008 [7]) and showed that by filtering blacklisted regions the estimation of the SCC is improved. However, using the SCC may not be helpful since the peaks that are attained at the read and fragment lengths are confused in a typical ChIP-exo SCC curve. Additionally this method requires to know blacklisted regions in advance which may not be available. In our pipeline we propose an out-the-shelf analysis that contrasts the enrichment of the experiment against its library complexity.

In order to obtain 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, 2011 [1] discussed that reads in the forward and reverse strand might construct peak pairs around bound proteins, 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 sites.

Mace (Wang et al, 20114 [8]), CexoR (Madrigal, 2015 [9]) and peakzilla (Bardet et al., 2013 [10]), recently developed ChIP-exo data analysis methods are also based on this peak pair assumption. However, appropriateness of such assumption was not fully evaluated in the literature yet. Furthermore, it is still unknown which factors could affect protein binding site identification 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 by Wilbanks and Facciotti, 2012 [11] and Pepke and Wold, 2009 [12]), such as MACS (Zhang et al., 2008 [13]), CisGenome (Ji et al., 2008 [14]) and MOSAiCS (Kuan et al., 2009 [15]), 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 appropriate to evaluate the potential of ChIP-exo data for high resolution identification of protein binding sites. More recently, deconvolution algorithms such as Deconvolution (Lun et al., 2009 [16]), GEM (Guo et al., 2012 [17], an improved version of Guo et al., 2010 [18]) and PICS (Zhang et al., 2010 [19]) 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 PE and SE ChIP-Seq data in a unified framework and as a result, currently available methods are not appropriate for fair comparison between ChIP-exo and ChIP-Seq. To address these limitations, we developed an improved of dPeak (Chung et al., 2013 [20]), a high resolution binding site identification (deconvolution) algorithm that we previously developed for PE and SE 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 work we demonstrate that the "peak pair" assumption of Rhee and Pugh, 2013 [5] 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. We evaluated several methods to identify binding events and dPeak performs competitively respect to GEM and MACE when analyzing ChIP-exo data. More importantly, when comparable number of reads is used for both ChIP-exo and ChIP-Seq , dPeak coupled with ChIP-exo data provides resolution comparable to PE ChIP-Seq and both significantly improve the resolution of protein binding identification compared to SE-based analysis with any of the available methods.

2 Results and discussion

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

σ^{70} factor is a transcription initiation factor of housekeeping genesin E. Coli. In this organism's genomes, many promoters contain multiple transcription start sites (TSS) and these TSS are often closely spaced ($10 \sim 150$ bp). These closely spaced binding sites are considered to be multiple "switches" that differentially regulate gene expression under diverse growth conditions [21]. 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 ChIP-Seq guidelines and quality metrics on ChIP-exo data

We started our exploration by investigating the suitability of the current state-of-the-art QC pipelines for ChIP-Seq for ChIP-exo. Table 1 contains measures that are commonly calculated for ChIP-Seq samples: PCR Bottleneck Coefficient (PBC) and the Normalized Strand Cross-Correlation (NSC) are calculated as in [4]. Additionally, we calculated the Forward Strand Ratio (FSR) to compare the number of fragments sequenced from each strand.

```
##                files    nreads        pbc      nsc
## 1      edsn931_Sig70  13961493 0.13994203 103.4239
## 2      edsn933_Sig70  14810838 0.16339252 162.8002
## 3      edsn935_Sig70  16108774 0.13535422 153.5088
## 4      edsn937_Sig70  13636541 0.15322537 172.5815
## 5     edsn1311_Sig70    902921 0.26895580 13.76908
## 6     edsn1314_Sig70   1852124 0.25908459 17.91882
## 7     edsn1317_Sig70   2104427 0.25841987  29.6083
## 8     edsn1320_Sig70  11548572 0.15107930 13.08626
## 9               CTCF  48478450 0.45796889 16.02482
## 10         ERR336935  22421729 0.10683407 72.04243
## 11         ERR336942  22210461 0.65622704 21.28197
## 12         ERR336956  23307557 0.79962169 60.42191
## 13 IMR90_GR_chip-exo  47443803 0.29785605 8.860176
## 14  K562_GR_chip-exo 116518000 0.05047634 4.117883
## 15  U2OS_GR_chip-exo   3255111 0.77142141 10.05883
## 16         ERR336933   9289835 0.80825680 19.87523
## 17         ERR336938  12464836 0.82035947 18.72806
## 18         ERR336950  11041833 0.80241200 21.48509
```

Table 1 Usual quality control indicators applied to the gathered ChIP-exo samples. PBC stands for PCR Bottleneck Coefficient (0 - 0.5 is severe bottlenecking, 0.5 - 0.8 is moderate bottlenecking, 0.8 - 0.9 is mild bottlenecking, while 0.9 - 1 is no bottlenecking), NSC for Normalized Strand Cross-Correlation and FSR for Forward Strand Ratio. We omitted the Relative Strand Cross-Correlation (RSC) because a typical ChIP-exo experiment is not accompanied by an input file. σ^{70} samples were given by Dr. Landick's Lab, FoxA1 and ER samples are from Serandour et al., 2013 [3] and CTCF sample is from Rhee and Pugh, 2011 [1]

It is of special importance to notice that the values of PBC are quite low, hence it would be necessary to repeat these experiments by following ChIP-seq guidelines. On the other hand, in a good ChIP-Seq experiment the SCC is maximized at the fragment length, and a "phantom peak" would be also present at the experiment's read length ([4] and [7]). However, in ChIP-exo's case these two peaks are confounded. Hence an enrichment measure based in SCC such as the NSC is harder to interpret. In figure 1 we observe that both local maxima are hard to differentiate in the SCC curves for the σ^{70} ChIP-exo datasets used to calculate the NSC values in table 1.

2.3 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 across 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 2A). This implies that signals for potential 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 for ChIP-exo data a much smaller portion of the genome is expected to be background.

We next evaluated the "peak pair" assumption from Rhee and Pugh, 2011 [1], 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. Wang et al., 2014 [8], Madrigal 2015 [9] and Bardet et al., 2013 [10] proposed method rely in this assumption. In order to evaluate this assumption, we reviewed the proportion of reads in the forward strand 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. 2B) and this indicates that the peak pair assumption might not hold in real ChIP-exo data.

We evaluated ChIP-exo data for CTCF factor from human genome [1] to investigate issues specific to eukaryotic genomes for binding sites identification. Figures 2C and 2D display the bin-level average read counts against mappability and GC content. Each data point is obtained by averaging the read counts across bins with the same mappability of GC content. These results indicate that binding site identification in ChIP-exo sample might also benefit from the use of methods that take into account of apparent sequence biases such as mappability and GC content.

2.4 ChIP-exo Quality Control Pipeline

Figure 3 shows a flowchart for the ChIP-exo QC pipeline. Which basically partitions the genome by keeping the non-digested ChIP-exo regions. Then, for each region calculates a series of summary statistics. Finally it creates several visualizations designed to assess the quality level of ChIP-exo sample.

2.4.1 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 necessary to address the balance between the enrichment and library complexity of an experiment. To diagnose this, we considered the Average Read Coverage (ARC) and the Unique Read Coverage Rate (URCR) which are defined as follows:

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

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

Using the FoxA1 in mouse liver cell lines from [3] and these two quantities, the relationship between library complexity and experiment enrichment was explored. In figure 4A we can observe both statistics for the regions that are common for the three replicates. This figure shows typical patterns for ChIP-exo experiment, there are two strong arms: The one on the left with low ARC and varying URCR corresponds to ChIP-exo's background, usually regions composed by scattered reads that were not digested during the exonuclease step; and the one on the right that corresponds to regions that are usually enriched, for these regions the URCR measures how the fragments are allocated into the possible positions in a region. Hence we would expect the first replicate to have have more enriched regions and the third replicate's library complexity to be lower than the other two replicates library complexities. To verify this statements, we extracted the sequences around high confidence binding events and look for the FoxA1 motif using FIMO [22]. Figure 4B shows the number of candidate regions, which shows that the first replicate is being allocated into more enriched regions than the other ones. Figure 4C shows that for the first and third replicates, the FoxA1 motif is being detected in roughly the same proportion of sites, and finally in 4D we observe that the first and third replicate can detect the FoxA1 motif with the same significance, while the second replicate does not.

2.4.2 Strand imbalance in ChIP-exo data

The strand imbalance assessment in based in the observation that the enriched regions usually are composed of a higher quantity of reads, therefore we examined the FSR (defined as the ratio of number of forward stranded reads divided by the total number of reads in a given region) as the regions 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 global FSR and noticed that for all experiments is roughly 0.5, which means there are roughly the same amount of reads in both strands. However figure 5 shows that the global FSR does not represent the experiment's local strand imbalance, hence the "peak pair" assumption may not hold well in real ChIP-exo data.

In order to asses the strand imbalance we created the visualization shown in figure 5: Figure 5A presents the FSR's behavior as the lower depth regions are being filtered out, while B) shows which percentage of the regions are composed by reads in both strand or only one (forward or backward). In a good data set, it would be expected that all quantiles shown to be quickly converging towards the median (in panel A) or the regions composed of reads in one strand being made of few fragments (in panel B). For each replicate, we divided the partitioned regions by asking whether they overlap a set of high quality ChIP-exo peaks, and then we tested if the strand imbalance's distribution were the same for both classes. For regions composed by a higher amount of reads, is is harder to distinguish their peaks by considering only the strand imbalance, and in an experiment with lower quality it is harder to make this distinction.

2.5 Comparison with ChIP-Seq data using dPeak

Figure 6 shows different comparisons among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. A RegulonDB annotation (Salgado et al, 2012 [21]) was considered identified

if the distance between it and a dPeak binding site estimate was at most of 20 bp. That way, the sensitivity is defined as the proportion of RegulonDB annotations identified in a peak and the resolution is defined as the minimum distance between a RegulonDB annotation and the dPeak binding site estimates. The left panel of figure 6 shows that the sensitivity increases as the mean distance between binding events increases. Despite the when the binding events in a peak are closer to each other, both ChIP-exo and PE ChIP-Seq are comparable, as the distance increases ChIP-exo identifies a higher proportion of the RegulonDB annotations; additionally SE ChIP-Seq is significantly less sensitive than both ChIP-exo and PE ChIP-Seq. The right panel shows that ChIP-exo and PE ChIP-Seq are comparable in resolution, while both protocols significantly outperform SE ChIP-Seq.

2.6 Systematic comparison of ChIP-Seq vs ChIP-exo under varying sequencing depth

Previously, ChIP-exo and SE ChIP-Seq have been compared at a fixed depth level in the literature, but this comparisons did not include PE ChIP-Seq. Hence, we sampled a fixed amount of reads for each of the ChIP-exo, PE ChIP-Seq and SE ChIP-Seq datasets of the σ^{70} samples. For each sampled dataset we applied out lower-to-higher resolution pipeline by calling peaks with MOSAiCS [15] and then deconvolving the binding events by using dPeak [20]. For the ChIP-exo datasets we called peaks by using the GC-content and mappability models with MOSAiCS, and for the ChIP-Seq datasets we used their respective Input samples.

Figures 7 shows the behavior of each data type in σ^{70} experiment under aerobic condition when their depth is fixed. It is remarkable that even when the number of candidate peaks or the number of predicted events is lower for ChIP-exo, it outperforms both PE and SE ChIP-Seq in number of identified targets and resolution.

This may suggest that with ChIP-exo less positive peaks are being called and that when the targets are being identified, dPeak estimates binding locations closer to the true location. Additionally, we can see that as the read depth increases, all four indicator seem to stabilize and hit a plateau, which may indicate that with ChIP-exo a smaller amount of reads is necessary to identify a higher number of targets, but it may be also possible that this is an artifact occurring due to ChIP-exo's lower library complexity. Additionally, it is worth noting that for PE ChIP-Seq we sampled both ends of the fragment, hence for each sequencing depth we are sampling the half amount of pairs for PE ChIP-Seq than for ChIP-exo or SE ChIP-Seq.

Figure 12 shows an analogous analysis but using the σ^{70} replicates with and without rif treatment. The left, middle and right columns shows the fixed depth against the number of predicted events, identified targets and resolution being compared at a fixed depth level. The behavior of this quantities seems to be opposite to the one as in figure 7, hence we used the ChIP-exo QC pipeline in the fixed depth ChIP-exo experiments. Figure 8 shows hexbin of ARC vs URCR for several fixed sample sizes, as the fixed depth increases the two arm pattern becomes more distinctive while for lower depth, it seems that the majority of the sampled reads were aligned to enriched regions. On the other hand in figures 13 to 16, we used the ChIP-exo QC pipeline on the samples that are outperformed by PE and SE ChIP-Seq. For a fixed low depth, we can see that the majority of the reads are being aligned to non-enriched regions since the vertical arm seems to be stronger for all 2 conditions

and replicates; for higher depth we can see low URCR values being predominant, which indicates that the majority of the regions being formed by few positions with a higher read concentration. In low complexity regions, the reads are being aligned to fewer positions but there is no control over the amount of reads mapped. Hence, those regions are more likely to being strand-imbalance which in turn may bias the binding site estimate and therefore decrease the number of identified targets or increase an experiments resolution.

2.7 dPeak outperforms competing methods in discovering closely spaced binding events from ChIP-exo and ChIP-Seq data

Figure 9 compares the resolution defined as the minimum distance between a RegulonDB annotation and a binding site predicted by either Peakzilla [10], MACE [8], GEM [17] or dPeak [20]. In a good dataset such as both of the ChIP-exo experiments under aerobic (panel 933) or anaerobic (panel 931) conditions, all the methods are comparable in resolution, and dPeak slightly outperforms the rest. On the other hand, for experiments with low library complexity the resolution calculated with dPeak's predictions is smaller than both Mace and Peakzilla, but it is greater than Gem. This may be due that the fact that Gem uses sequence information in addition to the aligned 5' end counts that dPeak uses.[1]

3 Conclusions

We made a systematic exploration of several ChIP-exo experiments and provide a list of factors that reflect the quality of a ChIP-exo experiment and we provide a QC pipeline which is capable of assessing the balance between the enrichment and the library complexity of a ChIP-exo experiment. Additionally, a set of diagnostics was established to assess the an experiment's quality. The QC pipeline only requires a set of aligned reads to give a global overview of a ChIP-exo experiment, this overview coincides with more elaborate analysis, such as motif detection in a set of high quality regions or resolution analysis given a set of annotations as gold-standard.

To the extent of our knowledge, we made for the fist time a comparison between ChIP-exo and PE ChIP-Seq. Using a set of annotations as gold-standard, we showed that both protocols are comparable in resolution and that for regions with more than one binding site, ChIP-exo is more sensitive than both SE and PE ChIP-Seq. We made a systematic comparison between fixed depth ChIP-exo, PE ChIP-Seq and SE ChIP-Seq, and we probed that for sufficiently complex libraries, ChIP-exo experiments can outperform PE and SE ChIP-Seq in number of identified targets and resolution. Using the ChIP-exo QC pipeline, we show how to diagnose if the library complexity of ChIP-exo experiment is adequate.

4 Methods

Growth conditions.

ChIP-exo experiments.

Construction of a SE ChIP-Seq experiment.

For the rif-treatment ChIP-Seq experiments, we sampled SE ChIP-Seq experiment from the PE ones by taking one of both ends randomly following a Ber(0.5) model.

[1]For here we may probably use only 933 as part of the main article and keep the rest for the supplement

Saturation analysis of ChIP-exo, PE ChIP-Seq and SE ChIP-Seq.

ChIP-Seq quality control metrics and Strand Cross-Correlation.

The statistics in table 1 and the SCC curves from figure 1 where calculated with the **ChIPUtils** package version 0.99.0, available in https://github.com/welch16/ChIPUtils.

ChIP-exo quality control pipeline.

We used the R package **ChIPexoQual** to assess the quality of the ChIP-exo datasets by following the steps described in figure 3. We used version 1.0, and it is available in https://github.com/welch16/ChIPexoqual.

dPeak analysis of σ^{70} ChIP-exo and ChIP-Seq data.

Method comparison for ChIP-exo.

We considered dPeak Chung et al., 2013 [20], GEM Guo et al., 2012 [17], MACE Wang et al., 2014 [8] and Peazilla Bardet et al., 2013 [10] for the ChIP-exo data analysis. for the dPeak algorithm we used the R package **dPeak** version 2.0.1 which is available from https://github.com/dongjunchung/dpeak. For the GEM algorithm, we used it's Java implementation version 2.6 which is available from http://groups.csail.mit.edu/cgs/gem/. For the Mace algorithm, we used it Python implementation version 1.2, which is available from http://dldcc-web.brc.bcm.edu/lilab/MACE/docs/html/. For the Peakzilla algorithm, we used the version available in https://github.com/steinmann/peakzilla. Candidate regions for **dPeak** were identified for each replicate of ChIP-exo data using the **MOSAiCS** algorithm Kuan et al., 2011 [15] (one sample analysis using false discovery rate of 0.01%) implemented as an R package **mosaics** version 2.9.7 (available from *bioconductor*). We further filtered out candidate regions by using the 300 peaks with higher average ChIP tag count to avoid potential false positive based on the exploratory analysis. These regions were also explicitly provided to the GEM algorithm as candidate regions. Default tuning parameters were used during model fitting for all methods.

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. Rhee, H.S., Pugh, F.: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell (2011)
2. Mahony, S., Franklin, P.B.: Protein-DNA binding in high-resolution. Critical Reviews in Biochemistry and Molecular Biology (2015)
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. 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.: ChIP-Seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research (2012)
5. 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)
6. 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)
7. Kharchenko, P., Tolstorukov, M., Park, P.: Design and Analysis of ChIP-Seq Experiments for DNA-binding Proteins
8. 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)
9. Madrigal, P.: CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. EMBnet.journal (2015)
10. 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)
11. Wilbanks, E., Facciotti, M.: Evaluation of algorithm performance in ChIP-Seq peak detection. PLOS One (2012)
12. Pepke, S., Wold, B., Ali, M.: Computation for ChIP-seq and RNA-seq studies. Nature (2009)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. Zhang, X., Robertson, G., Krzwinski, M., Ning, K., Droit, A., Jones, S., Gottardo, R.: PICS: Probabilistic inference for ChIP-Seq. Biometrics (2010)
20. 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. PIOS, Computational Biology (2013)
21. 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. Nucleic Acids Research (2012)
22. Grant, C., Bailey, T., Noble, W.S.: FIMO: Scanning for occurrences of a given motif. Bioinformatics (2011)

5 Figures

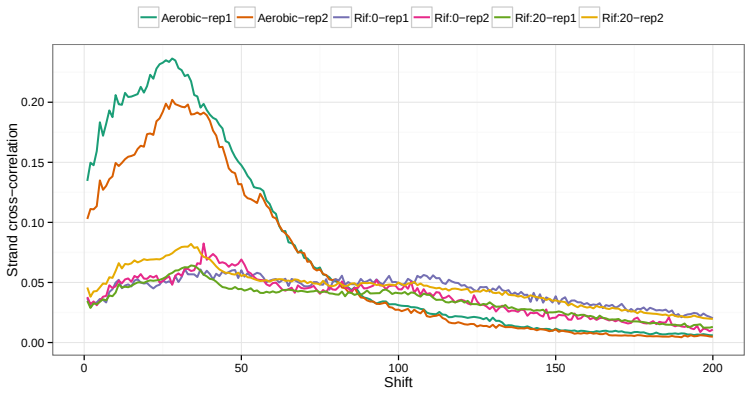


Figure 1 SCC curves for σ^{70} samples. The "phantom peak" and the summit that corresponds to the read and fragment length respectively are confounded due to the exonuclease digestion.

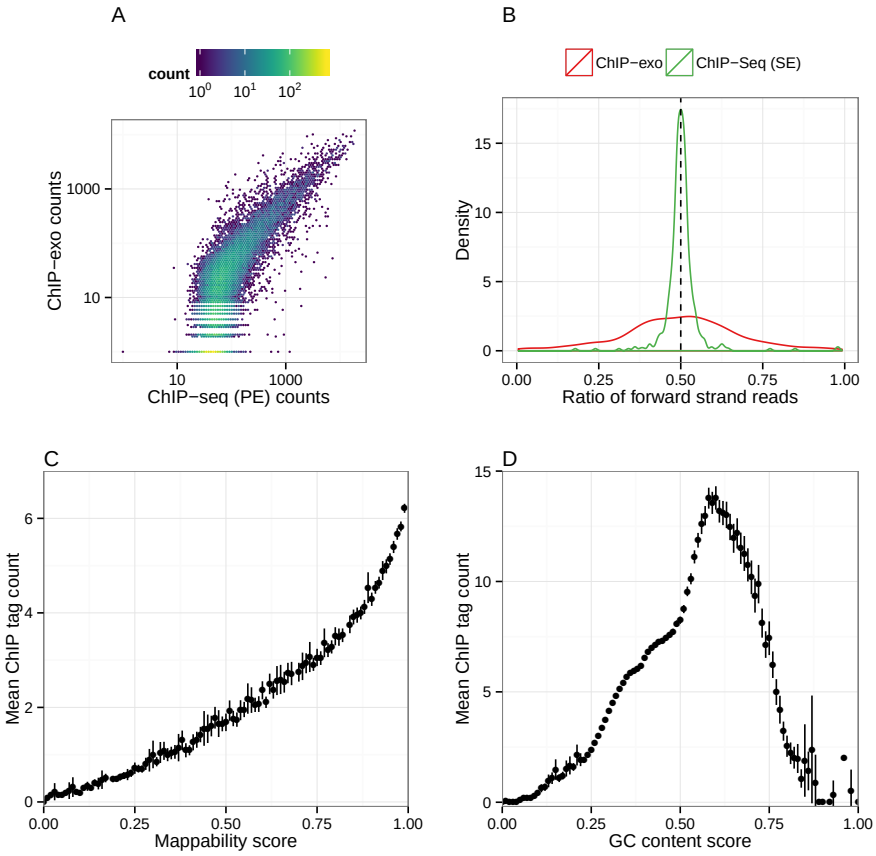


Figure 2 A) Hexbin plot of PE ChIP-Seq bin counts VS ChIP-exo bin counts. B) Forward Strand Ratio densities for SE ChIP-Seq and ChIP-exo peaks. C) ChIP tag counts increase linearly as mappability scores increases. D) ChIP tag counts increase linearly as GC content score increases when GC content is less than 0.6 and then ChIP tag counts decrease as GC content increases. D) Percentage of sites with at least p.value.

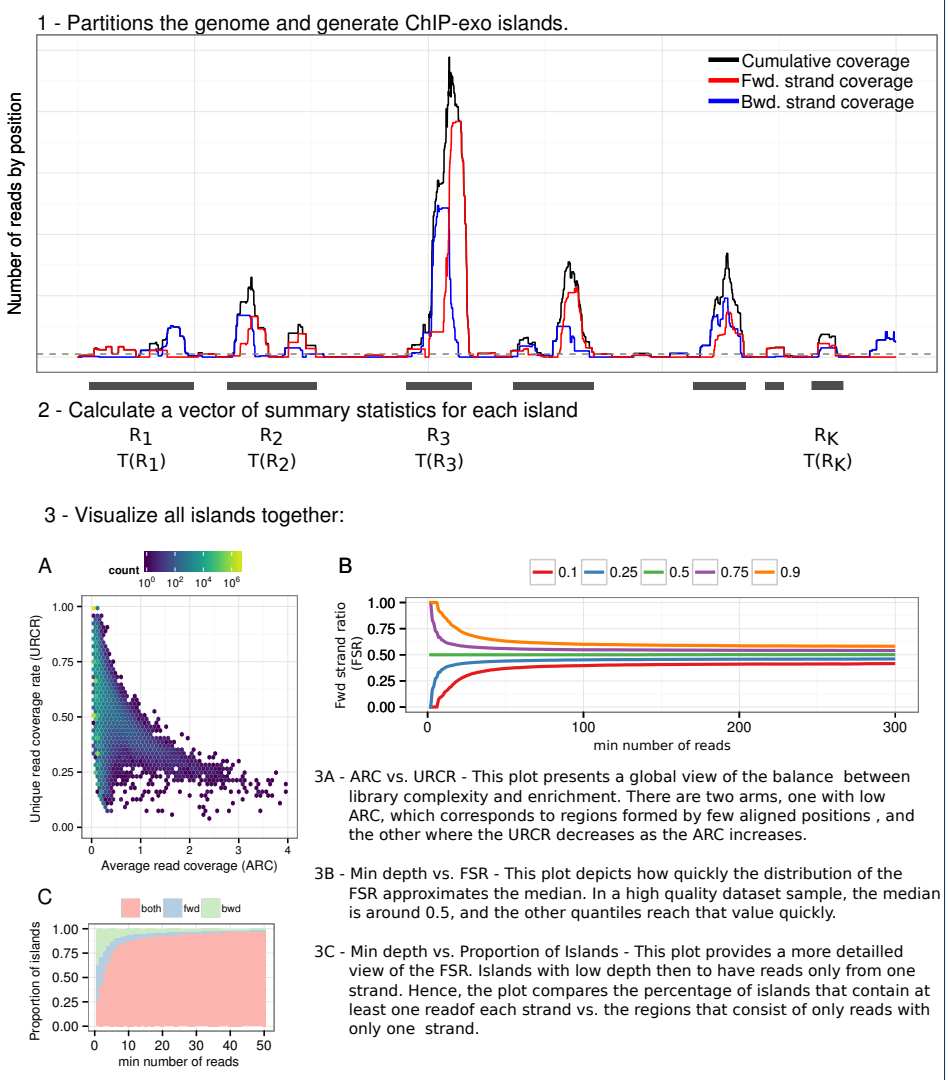


Figure 3 Flowchart of the ChIP-exo quality control pipeline.

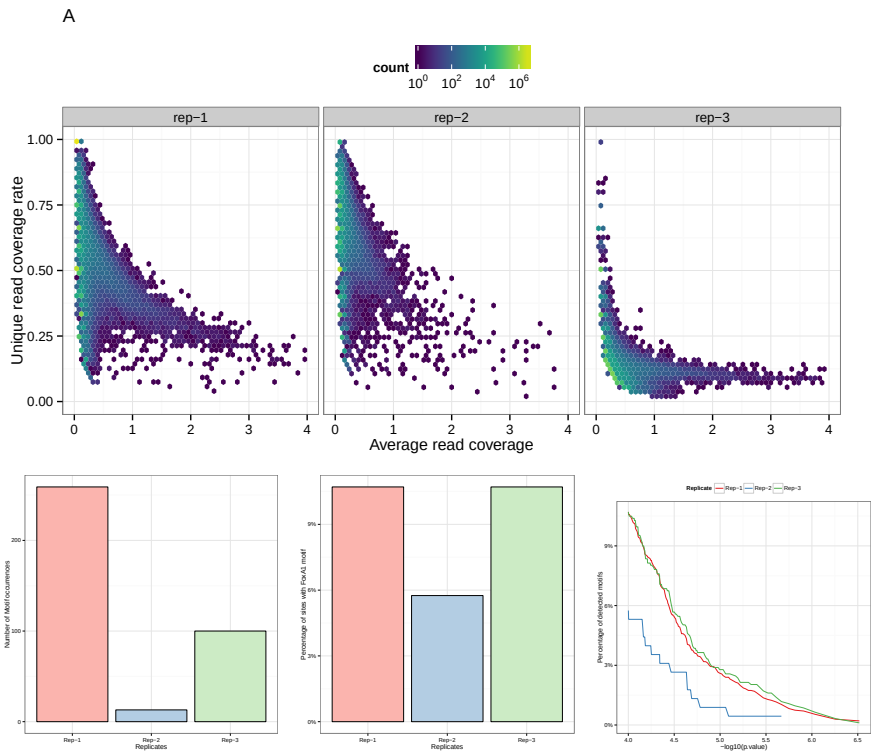


Figure 4 Using the mouse-FoxA1 experiment from [3]: A) Hexbin plots of ARC against URCR, in general we can see a slight separation into two strong arms, one corresponds to low ARC and varying URCR, and for the other URCR decreases as ARC increases. B) Number of candidate sites for each replicate. C) Percentage of candidate sites where the FoxA1 motif was detected. D)

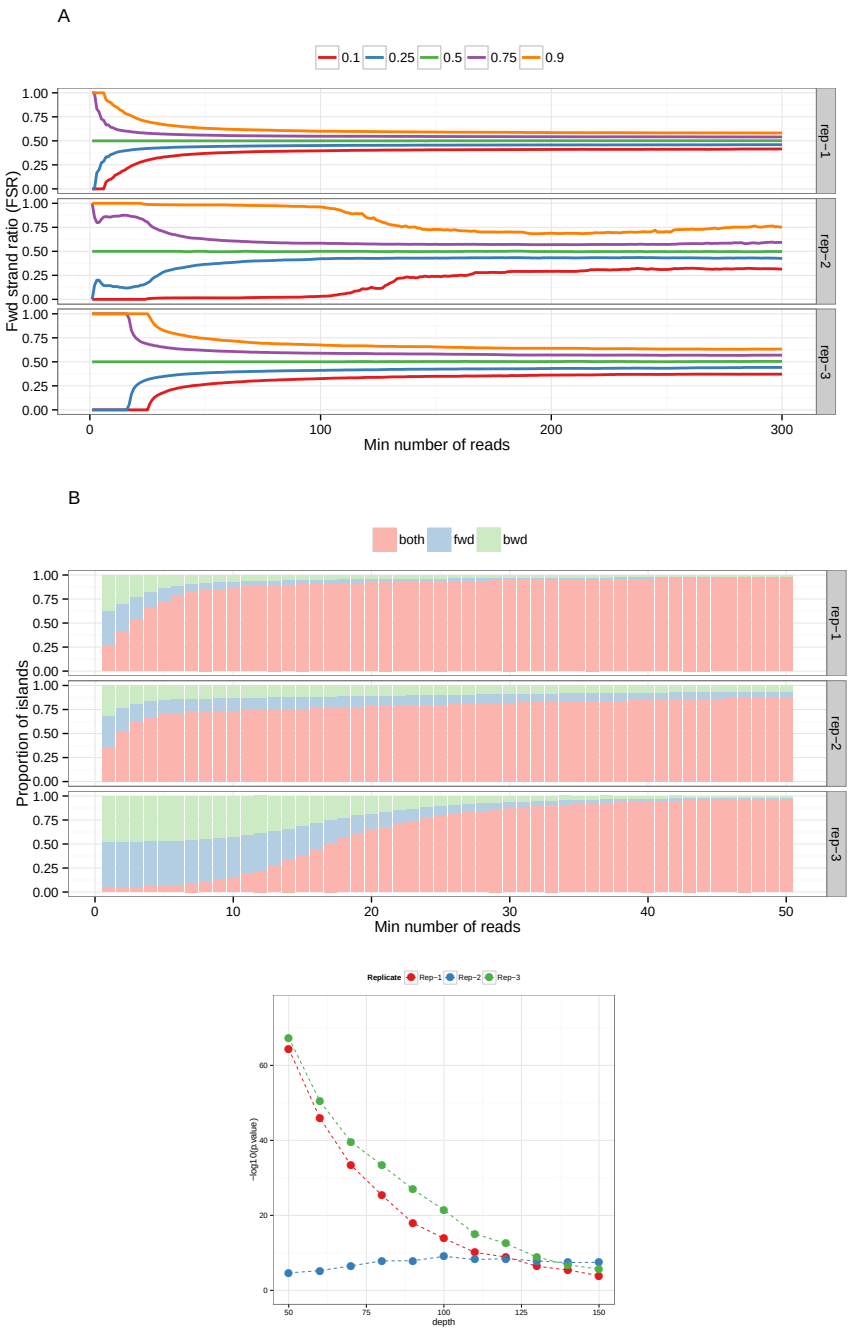


Figure 5 Strand imbalance QC plots for the same data as in Figure 4A. A) FSR distribution quantiles as the lower depth regions are being filtered out, all quantiles approach to the median as the lower bound increases. B) Stacked histogram with the proportion of regions that are formed by two strands or only one, in a good sample the single-stranded regions are going to be filtered out quickly as in the middle row. C) $-log_{10}$(p.value) of testing if the imbalance distributions differs when ChIP-exo regions overlap their peaks.

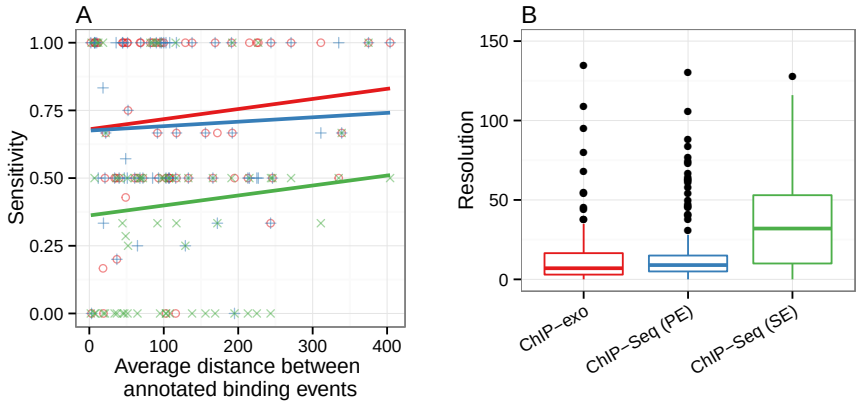


Figure 6 Comparison of (A) sensitivity and (B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction.

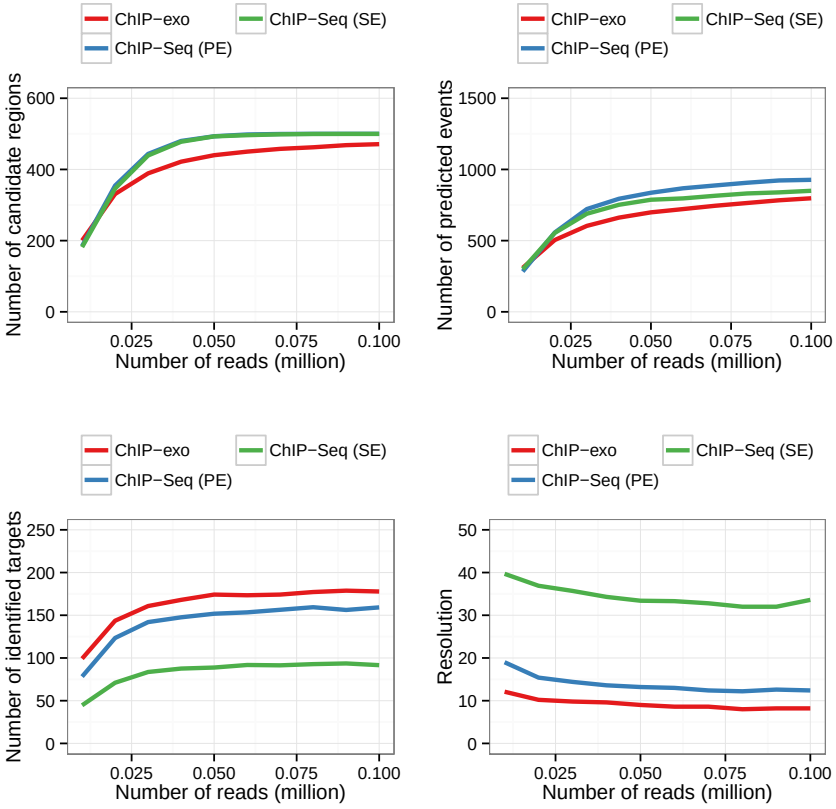


Figure 7 ChIP-exo, PE ChIP-Seq and SE ChIP-Seq comparison at varying sequencing depths. Comparison of the number of candidate regions (A), predicted events (B), identified targets (C) and resolution (D) among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

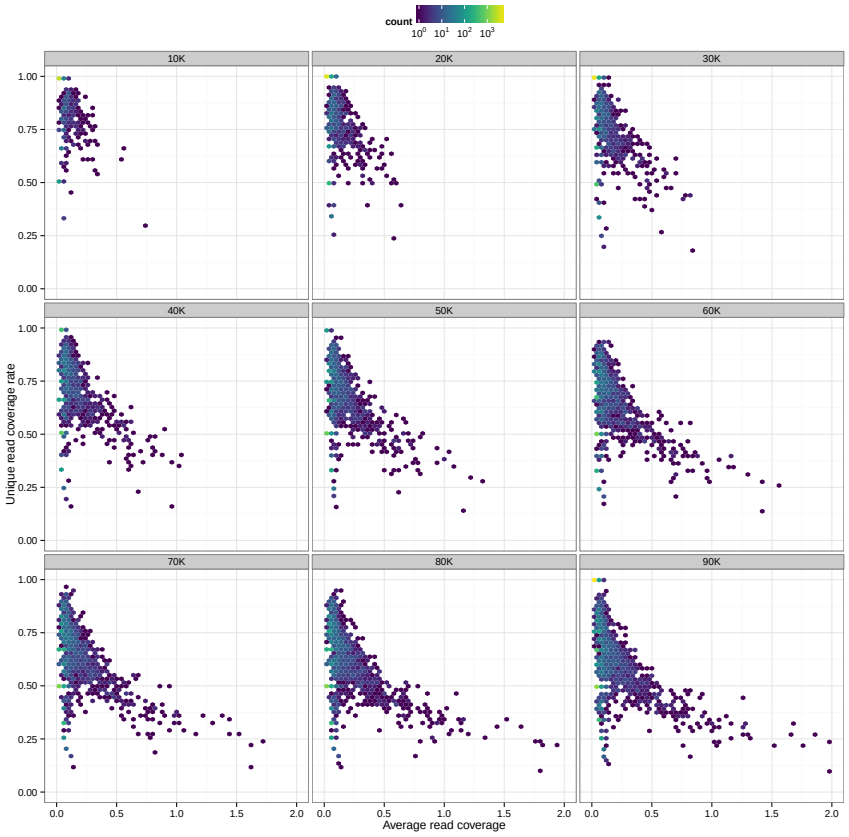


Figure 8 Hexbin plots of ARC vs URCR of the σ^{70} ChIP-exo experiment under aerobic condition when 10K to 90K reads are being sampled.

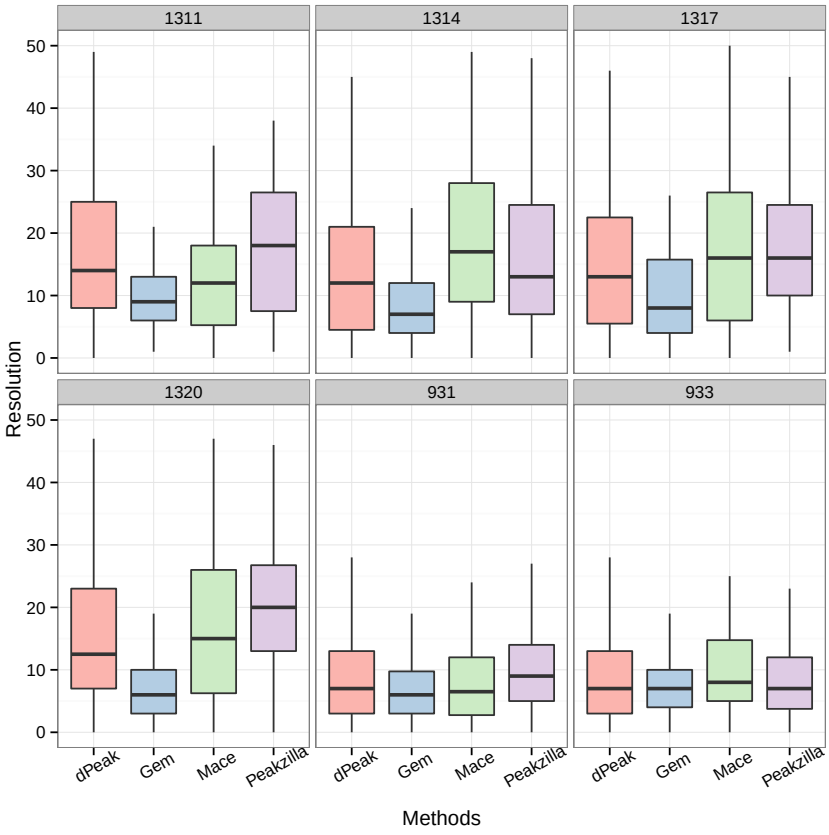


Figure 9 Comparison of the resolution between dPeak, Gem, Mace and Peakzilla methods. Resolution is defined as the minimum distance between a RegulonDB annotation an a predicted binding event.

Supplement

```
##                          files   nreads       pbc      nsc
## pet.1 edsn1396_Sig70.sort.bam 13445022 0.9426179 8.865244
## pet.2 edsn1398_Sig70.sort.bam 16538920 0.9378843 7.031836
## pet.3 edsn1400_Sig70.sort.bam 11642722 0.8891744 10.77284
## pet.4 edsn1402_Sig70.sort.bam 16854026 0.9407020 7.936239
## set.1 edsn1396_Sig70.sort.bam  6722511 0.6632742  9.01779
## set.2 edsn1398_Sig70.sort.bam  8269460 0.5594449 7.179539
## set.3 edsn1400_Sig70.sort.bam  5821361 0.6472382 10.89898
## set.4 edsn1402_Sig70.sort.bam  8427013 0.5895118 8.124717
```

Table 2 Same QC metrics as in table 1 but applied to Landick's chipseq data of the rif experiment

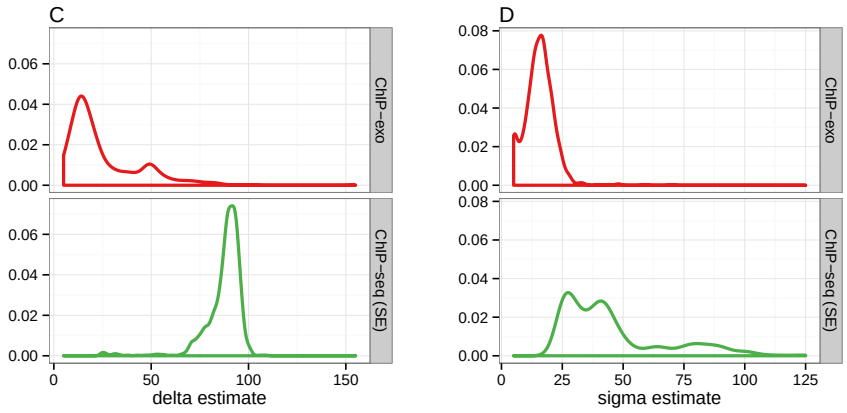


Figure 10 C) δ parameter in dPeak measures average distance of the reads to their respective binding site. In ChIP-exo data, reads were located much closer to the binding site than in SET ChIP-Seq. D) σ parameter measure the dispersion of reads around each binding site. In ChIP-exo data, reads showed less variation around the their respective binding sites compared to SET ChIP-Seq.

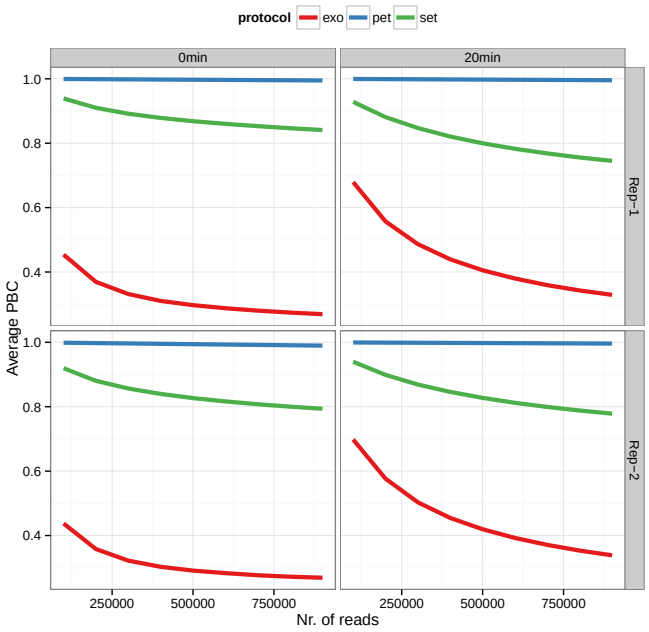
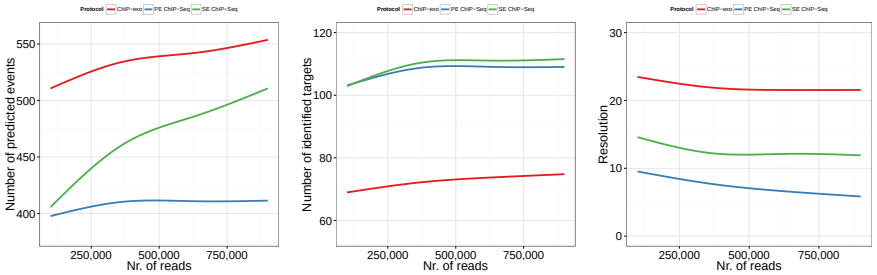
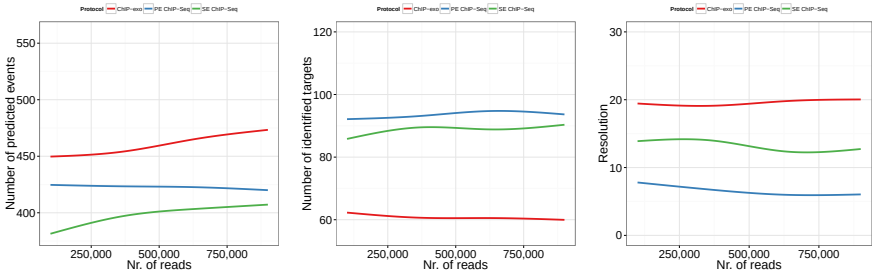


Figure 11 Average PBC (among all seeds) of the sampled ChIP-exo, PE ChIP-Seq and SE ChIP-Seq experiments under the rif treatment conditions used for saturation analysis.

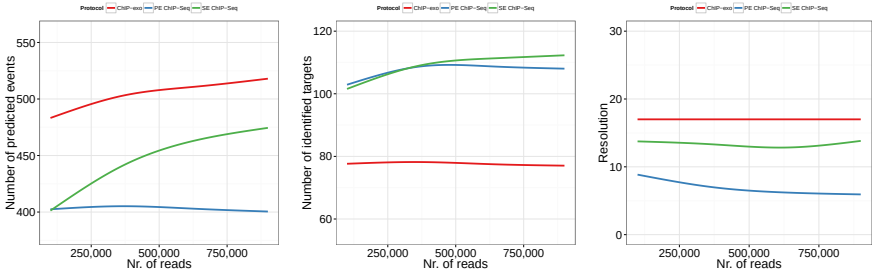
- Rep-1 and rif-0min:



- Rep-1 and rif-20min:



- Rep-2 and rif-0min:



- Rep-2 and rif-20min:

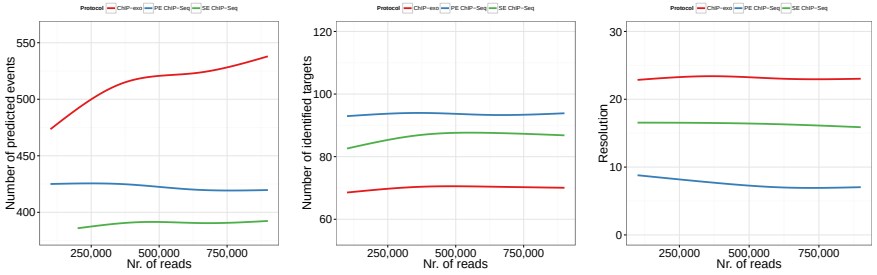


Figure 12 ChIP-exo, PE ChIP-Seq and SE ChIP-Seq comparison at varying sequencing depth Comparison of the number of predicted events (left), identified targets (middle) and resolution (right) among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as gold standard. A RegulonDB binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

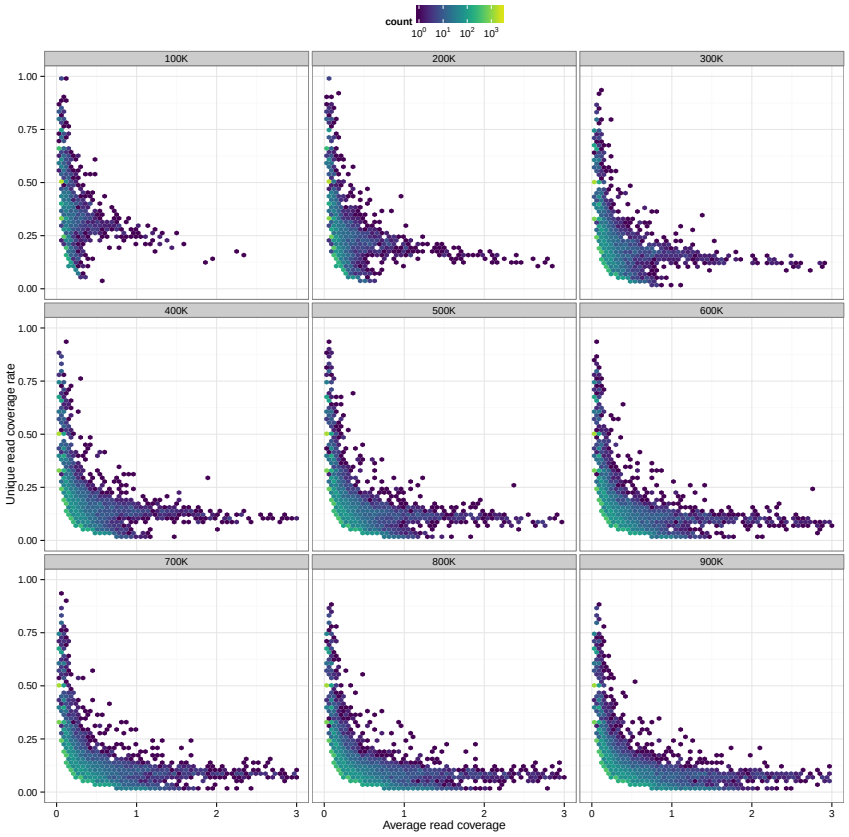


Figure 13 A) ARC vs URCR hexbin plots of Rep-1 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

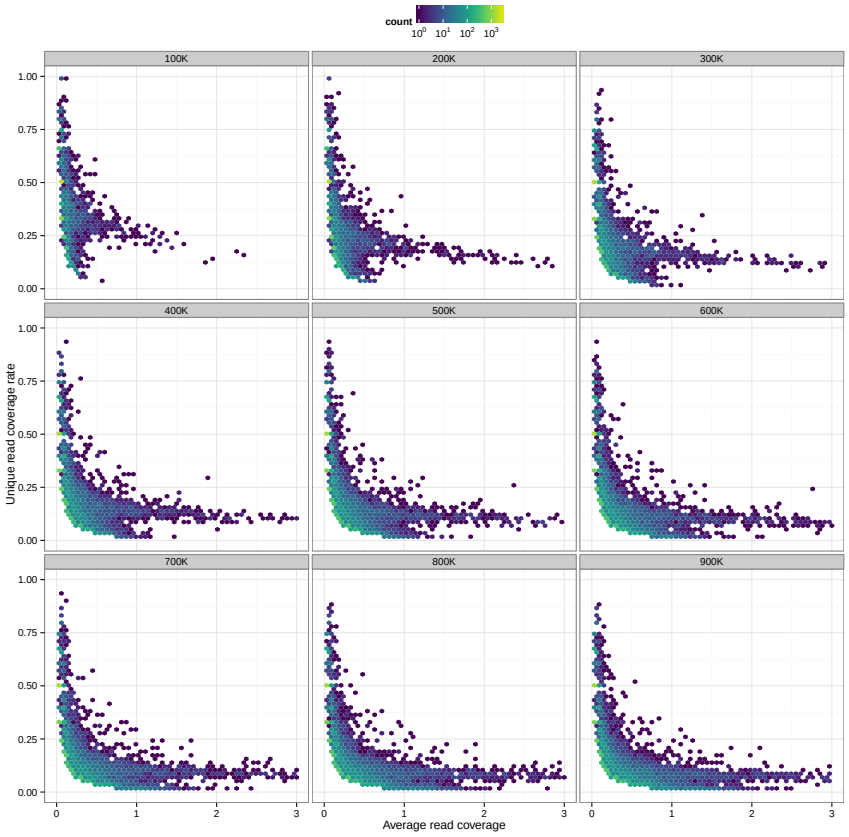


Figure 14 ARC vs URCR hexbin plots of Rep-1 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

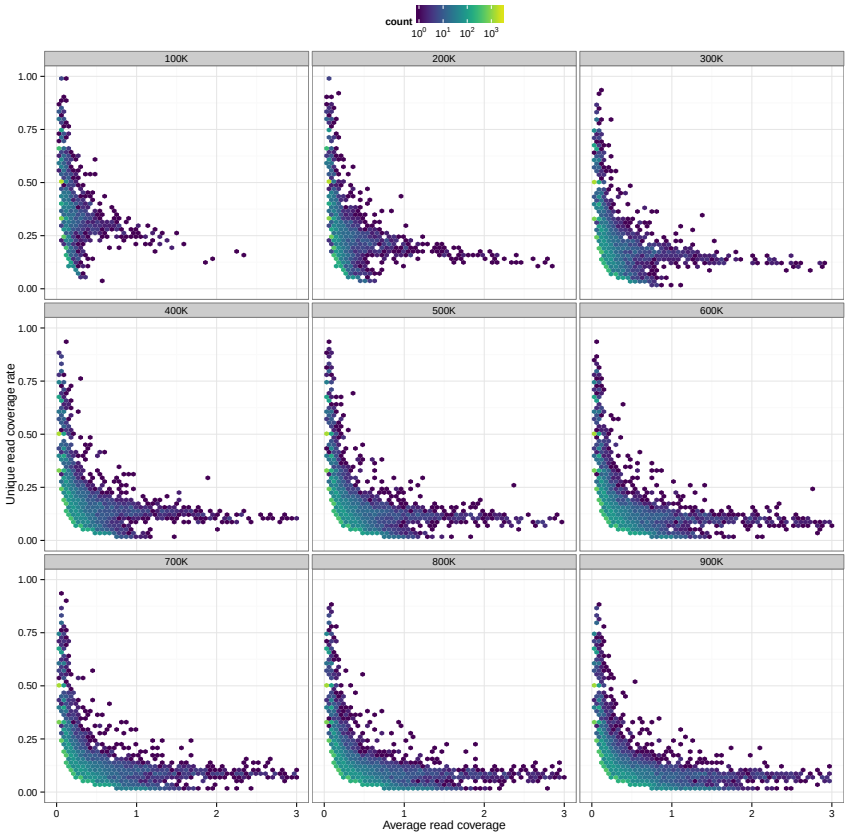


Figure 15 ARC vs URCR hexbin plots of Rep-2 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

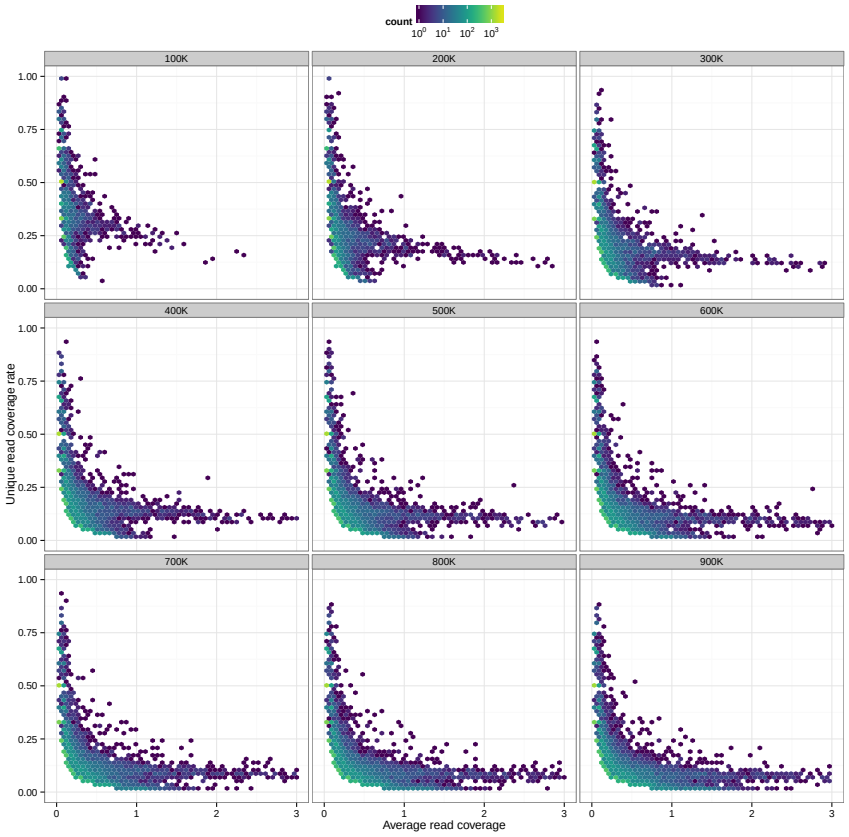


Figure 16 ARC vs URCR hexbin plots of Rep-2 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.