DRAFT

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

Rene Welch [1†], Dongjun Chung [6†], 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
[†]These two authors contributed equally.

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 a ChIP-exo experiment's strand imbalance, enrichment and library complexity. Assessment of these characteristics 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; Quality Control; ChIP-Seq; Spatial Resolution; Transcription Factor; Binding Site Identification on High-Res; Deconvolution

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 (36 to 100 bp) at the 5′ end of each fragment generates millions of reads. Figure 1 illustrates the differences between ChIP-exo, Single End (SE) ChIP-Seq and Paired End (PE) ChIP-Seq: The 5′ ends of a ChIP-exo experiment are located more tightly around the binding proteins than in a ChIP-Seq experiment; in a PE ChIP-Seq experiment both ends are observed while in a SE ChIP-Seq experiment only the 5′ end.

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]). Second, although there are roughly the same amount of reads in both strands, locally there may be more reads in one strand than in the other. Finally, most of current ChIP-Seq quality control (QC) guidelines (Landt et al., [3]) may not be applicable on ChIP-exo, while there are not established QC pipelines for ChIP-exo; previous ChIP-exo analyses used ChIP-Seq samples to compare the resolution between experiments ([1], [4], [5])). To address these challenges, we suggest a collection of quality control visualizations to interrogate these biases 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 [4]); Glucocorticoid receptor (GR) in IMR90, K562 and U20S cell lines (Starick et al., 2015 [6]).Furthermore, we also generated ChIP-exo and ChIP-Seq data for σ^{70} factor in *Escherichia Coli* (*E. Coli*) measured under aerobic ($+O_2$) condition, and treated by rifampicin by 0 and 20 minutes (courtesy of Professor Robert Landick's lab).

In order to obtain the potential benefits of ChIP-exo on protein binding site identification, it is critical 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. Based on this rationale, they used the "peak pair method" that predicts the midpoint of two modes of peak pairs as potential binding sites. Recently developed ChIP-exo data analysis methods, such as Mace (Wang et al, 20114 [7]), CexoR (Madrigal, 2015 [8]) and Peakzilla (Bardet et al., 2013 [9]), 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 [10] and Pepke and Wold, 2009 [11]), such as MACS (Zhang et al., 2008 [12]), CisGenome (Ji et al., 2008 [13]) and MOSAiCS (Kuan et al., 2009 [14]), these approaches might not fully utilize potentials of ChIP-exo data for high resolution identification of protein binding sites. Specifically these approaches reveal protein binding sites only in lower resolution, i.e., at an interval of hundreds to thousands of base pairs. Furthermore, they implicitly assume that there is only one "mode" or "predicted binding location" per this wide genomic interval. More recently, deconvolution algorithms such as Deconvolution (Lun et al., 2009 [15]), GEM (Guo et al., 2012 [16], an improved version of Guo et al., 2010 [17]) and PICS (Zhang et al., 2010 [18]) 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 and utilized an improved version of dPeak (Chung et al., 2013 [19]), 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.

Some of the key findings in this work are as follows. First, we demonstrate that the "peak pair" assumption of Rhee and Pugh, 2013 [5] does not hold well in real ChIP-exo data. Second, 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. Third, we evaluated several methods to identify binding events and dPeak performs competitively respect to GEM and MACE when analyzing ChIP-exo data. Finally, 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 sequenced E. Coli σ^{70} ChIP-exo and ChIP-Seq data

σ^{70} factor is a transcription initiation factor of housekeeping genes in 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 [20]. Therefore, investigation of ChIP-exo's potential for identification and differentiation of closely spaced binding sites is invaluable for elucidating the transcriptional networks of prokaryotic genomes.

2.2 Application of current ChIP-Seq QC guidelines on ChIP-exo data

We started our exploration by investigating whether the current state-of-the-art QC pipelines for ChIP-Seq are suitable 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 [3]. We omitted the Relative Strand Cross-Correlation (RSC) which is another commonly used QC measure because a typical ChIP-exo experiment is not accompanied by an input sample.

```
##     Organism  IP/TF Condition/Cell Rep.       Depth    PBC      NSC
##  1:   E.Coli  Sig70        Aerobic    1  13,961,493 0.1399 103.4239
##  2:   E.Coli  Sig70        Aerobic    2  14,810,838 0.1634 162.8002
##  3:   E.Coli  Sig70      Anaerobic    1  16,108,774 0.1354 153.5088
##  4:   E.Coli  Sig70      Anaerobic    2  13,636,541 0.1532 172.5815
##  5:   E.Coli  Sig70       Rif-0min    1     902,921 0.2690  13.7691
##  6:   E.Coli  Sig70       Rif-0min    1   1,852,124 0.2591  17.9188
##  7:   E.Coli  Sig70      Rif-20min    2   2,104,427 0.2584  29.6083
##  8:   E.Coli  Sig70      Rif-20min    2  11,548,572 0.1511  13.0863
##  9:    Human   CTCF           HeLA    1  48,478,450 0.4580  16.0248
## 10:    Mouse Fox A1    Mouse Liver    1  22,210,461 0.6562  21.2820
## 11:    Mouse Fox A1    Mouse Liver    2  23,307,557 0.7996  60.4219
## 12:    Mouse Fox A1    Mouse Liver    3  22,421,729 0.1068  72.0424
## 13:    Human     GR          IMR90    1  47,443,803 0.2979   8.8602
## 14:    Human     GR           K562    1 116,518,000 0.0505   4.1179
## 15:    Human     GR           U2OS    1   3,255,111 0.7714  10.0588
## 16:    Human     ER          MCF-7    1   9,289,835 0.8083  19.8752
## 17:    Human     ER          MCF-7    2  11,041,833 0.8024  21.4851
## 18:    Human     ER          MCF-7    3  12,464,836 0.8204  18.7281
```

Table 1 Current QC metrics applied to ChIP-exo data. PBC stands for PCR Bottleneck Coefficient and NSC for Normalized Strand Cross-Correlation.

DNA libraries generated by the ChIP-exo protocol seem to be less complex than the libraries generated by ChIP-Seq, since the possible number of positions to which the reads can be aligned is being reduced due to the exonuclease digestion, hence considerable amounts of reads are being mapped to specific positions. This affects the interpretation of the PBC, since for ChIP-Seq low PBC values indicate that the same read has been copied by the amplification process and aligned multiple times to the same position; while for ChIP-exo when several reads are aligned to the same position are not necessarily the same read amplified, but several reads that their $5\prime$ end was digested to the same position before the amplification step. It is of special importance to notice that for several ChIP-exo datasets, the PBC values are quite low. Therefore, by blindly following ChIP-Seq guidelines those experiment could have been considered as not useful and repeated.

The Strand Cross-Correlation (SCC) introduced by Kharchenko et al., 2008 [21] is the most commonly used quality measure in ChIP-Seq. It is calculated as the correlation between both strand coverages, where each one is shifted $\delta/2$ bp towards the $3\prime$ direction. In general it measures how well the reads mapped to each strand are clustered around the locations where the proteins are binding to the DNA, and usually it is expected to observed two local maxima, one when the profiles are shifted by the average read length and another when the profiles are shifted by the unobserved fragment length. In a good ChIP-Seq dataset the last one is also the SCC global maxima. However, in ChIP-exo's case these two peaks are confounded. Hence the Normalized Strand Cross-Correlation (NSC) which is a measure based

on the SCC is harder to interpret. Figure 2 shows 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 (Upper part of Figure 3A). 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 (lower part of Figure 3A). Specifically, 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. Note that currently available ChIP-exo data analysis methods, such as Wang et al., 2014 [7], Madrigal 2015 [8] and Bardet et al., 2013 [9] rely on this assumption. In order to evaluate this assumption, we reviewed the proportion of reads in the forward strand in ChIP-exo peaks such as at least one binding site is predicted in both ChIP-exo and ChIP-Seq 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. 3B) 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 3C and 3D 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. In Figure 3C it is shown that the ChIP-exo tag counts linearly increases with the mappability score and in Figure 3D it is shown that for GC - content below 0.6, the mean ChIP tag count increases and for GC - content greater than 0.6 it shows a decreasing trend. Kuan et al., 2011 [14] studied the presence of the mappability and GC - content biases in ChIP-Seq's background. It is not surprising to see these biases also present in ChIP-exo data, since ChIP-Seq and ChIP-exo signal seems to be linearly correlated for enriched regions (Figure 3A). Rozowsky et al., 2009 [22] and Valouev et al., 2008 [23] provide in depth analysis of the mappability and GC - content biases for ChIP-Seq

respectively. Finally, these results indicate that binding site identification for ChIP-exo data benefits from using methods that take into account of apparent sequence biases such as mappability and GC content.

2.4 ChIP-exo Quality Control Pipeline

Figure 4 shows a flowchart for the ChIP-exo QC pipeline. In the first step, we partition the genome by keeping the non-digested ChIP-exo regions. Then, for each region it counts the number of fragments that compose the region and the number of positions to which the reads are being mapped to in each strand. With these values it calculates the following summary statistics:

$$\begin{aligned}
\text{ARC} &= \frac{\text{Nr. of reads in the region}}{\text{Width of the region}}, \\
\text{URCR} &= \frac{\text{Nr. of reads in the region mapped to exactly one position}}{\text{Nr. of reads in the region}}, \\
\text{FSR} &= \frac{\text{Nr. of fwd. strand reads in the region}}{\text{Number of reads in the region}}.
\end{aligned}$$

Finally it creates several visualizations designed to diagnose the quality level of a ChIP-exo sample. Figure 4A shows the typical behavior of the ARC vs. URCR plot. In general, the plot depicts two strong arms: One on the left with low ARC values and varying URCR which corresponds to ChIP-exo's background, regions that are usually composed by scattered reads that were not digested during the exonuclease step; and another one where the URCR decreases as the ARC increases, which corresponds to regions that are usually enriched and as the URCR decreases the library complexity does it as well, on the other hand high URCR values correspond to regions composed by position with relatively few fragments aligned to. Figures 4B and 4C analyze the strand imbalance bias in a ChIP-exo experiment: The first one depicts how quickly the regions exclusively formed by fragments in one strand are being filtered out as regions with higher depth are observed; and the second one shows how quickly the FSR's distribution approach the median, since in a high quality sample it is expected for the median to be approximately 0.5 and the enriched regions are going to be composed by fragments sequenced from both strands.

The ChIP-exo QC pipeline provides a model free framework to analyze the biases in a ChIP-exo experiment by taking advantage of the exonuclease enzyme that digests the non-enriched regions to partition the genome, calculates common ChIP-Seq QC metrics in ChIP-exo regions locally and allows the interpretation of these metrics by the use of diagnostic plots.

2.4.1 Enrichment and library complexity in ChIP-exo data

In ChIP-exo experiments, background fragments are often digested by the exonuclease enzyme, therefore the balance between the enrichment and library complexity of an experiment is a key factor determining the sample's data quality.

Using the Fox A1 in mouse liver cell lines from [4] and these two quantities, we explored the relationship between library complexity and experiment enrichment. In

Figure 5A we present ARC vs. URCR plots for all three replicates. As a case of study, we compare the three plots to differentiate the quality of the three experiment. Hence, this might imply that 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 statement, we extracted the sequences around high confidence binding events and look for the FoxA1 motif using FIMO [24]. Figure 5B shows the number of candidate regions, which shows that the first replicate is being allocated into more enriched regions than the other ones. Figure 5C shows that for the first and third replicates, the FoxA1 motif is being detected in roughly the same proportion of sites, and finally in 5D 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 is 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, as it evaluates the "peak pair" assumption that the original ChIP-exo paper suggested and multiple ChIP-exo data analysis methods rely on. For every ChIP-exo experiment, 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.

In order to asses the strand imbalance we created the visualization shown in Figure 6: Figure 6A presents the FSR's behavior as the lower depth regions are being filtered out, while Figure 6B) 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 (using the Wilcoxon rank sum test over the imbalance index defined in the methods section) if the strand imbalance's distribution is the same for both classes. For regions composed by a higher amount of reads, it is harder to distinguish their peaks by considering only the strand imbalance, hence in a better quality ChIP-exo experiment it is easier to distinguish enriched regions by the amount of reads in both strands. Similarly, we may consider that the strands of reads for background of a ChIP-exo experiment is more unbalanced than those for the enriched regions. In conclusion, Figure 6 shows that the global FSR does not represent the experiment's local strand imbalance, hence the "peak pair" assumption does not hold locally in ChIP-exo data.

2.5 Comparison with ChIP-Seq data using dPeak

Figure 7 shows comparisons among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. We considered the RegulonDB data as ground truth, since those are the most recent annotation on *Escherichia Coli*. A RegulonDB annotation (Salgado et al, 2012 [20])

was considered to be identified if the distance from the closest dPeak binding site estimate was less that or equal to 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 closest dPeak binding site estimate. Figure 7A shows that the sensitivity increases as the mean distance between binding events increases. When the average distance is greater than 200 bp, dPeak identifies more than the 75% of the binding events in each peak, this is intuitive as the mean ChIP-Seq's fragment length is shorter than 200 bp, hence the read does not contribute to more than one binding event [19]. 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. Figure 7B 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 only at a fixed depth level in the literature, while they did not include PE ChIP-Seq as well either. In order to address this limitation in previous studies, 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 (N reads for both ChIP-exo experiment and $N/2$ or N pairs for PE ChIP-Seq to assume more realistic situation of a fixed cost). For each sampled dataset we applied our lower-to-higher resolution pipeline by calling peaks with MOSAiCS [14] and then deconvolving the binding events by using dPeak [19]. For the ChIP-exo datasets we called peaks by using the GC-content and mappability models with MOSAiCS, since it's background is usually composed of scattered reads across the genome; and for the ChIP-Seq datasets we used their respective Input samples. 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.

Figures 8 shows the behavior of each data type in σ^{70} experiment under aerobic condition when comparable number of reads is used for all of ChIP-exo, SE ChIP-Seq and PE ChIP-Seq. In Figure 8A we show the number of candidate regions defined as the number of regions where a binding event was identified in a collection of high quality peaks; in Figure 8B we depict the number of binding events; in Figure 8C we show the number of identified targets, where a RegulonDB was considered identified if a binding event was identified in a 15 bp vicinity of it; and finally Figure 8D show the resolution defined as the distance from a RegulonDB annotation to the closest dPeak prediction. 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 false 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 earlier than the cases for ChIP-Seq, which may indicate that with ChIP-exo a smaller amount of reads is needed

to identify the same number of targets than ChIP-Seq, but it may be also possible that this is an artifact occurring due to ChIP-exo's lower library complexity.

Figure S3 shows an analogous analysis but using the σ^{70} replicates with and without rif treatment. The left, middle and right columns show 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 8, hence we used the ChIP-exo QC pipeline in the fixed depth ChIP-exo experiments. Figure 9 shows scatter plots 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 S4 to S7, 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 10 compares the resolution defined as the minimum distance between a RegulonDB annotation and a binding site predicted by either Peakzilla [9], MACE [7], GEM [16] or dPeak [19]. In a good dataset such as both of the ChIP-exo experiments under aerobic (panel 931 and 933) conditions, all the methods are comparable in resolution, and dPeak slightly outperforms the rest. On the other hand, for experiments dPeak identified binding sites with higher resolution than Mace and Peakzilla, but with low resolution 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. We provided a list of factors that reflect the quality of a ChIP-exo experiment and we developed 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 quality of a ChIP-exo experiment. While 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 that is computationally more expensive to perform or requires additional inputs that may not be available, such as motif detection in a set of high quality regions or resolution analysis given a set of annotations as gold-standard.

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

We studied the shared biases between ChIP-exo and ChIP-Seq data, and noticed that for eukaryotic genomes the relationship between ChIP-Seq data and either the mappability or the GC content scores are still present in ChIP-exo. We also examined ChIP-exo's background and noticed that is significantly different from the ChIP-Seq one, since it consists of only a small quantity of fragments that was not digested by the exonuclease enzyme. Additionally, we showed that we have unbalanced number of reads in forward and reverse strands, and that in a lower quality ChIP-exo experiment those regions are going to be harder to differentiate from the possibly enriched regions.

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 rigorous 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. The proposed ChIP-exo QC pipeline provides a rigorous, easily interpretable, computationally efficient framework to diagnose if the library complexity of a ChIP-exo experiment is adequate.

4 Methods

Growth conditions.

ChIP-exo experiments.

Definition of current ChIP-Seq QC guidelines.

The statistics in Table 1 and the SCC curves from Figure 2 were calculated with the **ChIPUtils** package version 0.99.0, available in `https://github.com/welch16/ChIPUtils`. This package provides an easy to use interface to calculate basic quality control metrics and diagnostic plots for ChIP-Seq data.

PCR Bottleneck Coefficient.

The PCR Bottleneck coefficient is a measure of library complexity in ChIP-Seq data:

$$\text{PBC} = \frac{\text{Nr. of positions to which exactly one unique mapping read is aligned}}{\text{Nr. of positions to which at least one unique mapping read is aligned}}$$

For human and mouse genome, the ENCODE project states that a PBC value in the 0 - 0.5 range indicates severe bottlenecking, in the 0.5 - 0.8 range moderate bottlenecking, in the 0.8 - 0.9 range indicates mild bottlenecking and in the 0.9 - 1 range indicates that there is no presence of bottlenecking.

Strand Cross-Correlation.

The strand cross-correlation was proposed by Kharchenko et al., 2008 [21] and it may be one of the most used of the ChIP-Seq QC metrics. The SCC curve is defined as:

$$y(\delta) = \sum_c w_c r \left[n_c^+ \left(x + \frac{\delta}{2} \right), n_c^- \left(x - \frac{\delta}{2} \right) \right], \tag{1}$$

where $y(\delta)$ is the SCC for a strand shift δ, r is the Pearson correlation, w_c is the proportion of reads mapped to chromosome c and n_c^S is the read count vector for strand S and chromosome c. Additionally, two QC metrics are defined:

$$\text{NSC} = \frac{\max_\delta y(\delta)}{\min_\delta y(\delta)}, \tag{2}$$

$$\text{RSC} = \frac{\max_\delta y(\delta) - y_{\text{bgd}}}{y_{\text{rl}} - y_{\text{bgd}}}. \tag{3}$$

where y_{bgd} is the background SCC level and y_{rl} is the SCC "phantom peak" value.

Mappability and GC content scores.

To define the mappability score we follow the definition from Rozowsky et al., 2009 [22]:

$$m_i = \sum_{k=i-L+1}^{i+L-1} \frac{\delta_k}{2L-1}. \tag{4}$$

where δ_i is the indicator if the base at coordinate i can be mapped uniquely by a 32 bp sequence at position i, and L is the expected fragment length. GC - content score is defined analogously, where δ_i represents the occurrence of a G or C at the i-th position in the genome.

The mappability and GC - content scores for a bin are defined as the average of the scores across the nucleotides in the bin.

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 4. We used version 1.0, and it is available in `https://github.com/welch16/ChIPexoqual`.

Motif analysis of Fox A1 enriched regions

For all three replicates we called peaks using the MOSAiCS GC + Mappability model using an FDR level of 5%, filtering out the peaks with average ChIP counts below one hundred fragments and merging peaks gaped by at most 200 bp. Then, we fitted the dPeak model considering at most 5 binding events for each peak, and we searched for the Fox A1 motif over a 10 bp window around the estimated binding events. We used FIMO's command line 4.9.1 version [24].

Imbalance index

For every ChIP-exo experiment, we partitioned the experiment into the non-digested regions using the QC pipeline. For each region we calculated the FSR defined as the ratio between the number of forward stranded reads and the total number of reads in a region.

Then we transformed the FSR into an imbalance index is defined as:

$$\text{Imbalance index} = -\log_{10}(4 \times \text{FSR} \times (1 - \text{FSR}))$$

Construction of a SE ChIP-Seq from a PE 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 with equal probability.

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

For the resolution and sensitivity analysis, we used the MOSAiCS GC content + mappability model to call peaks for ChIP-exo experiments, while for both SE and PE ChIP-Seq experiments we used the MOSAiCS Input model. To avoid false positives we only considered ChIP-exo peaks with average ChIP counts greater than 3000 that overlapped both the SE and PE ChIP-Seq peaks, we considered other cutoff values but still obtained results similar to what we presented in this paper. Then we estimated the binding site events using the dPeak model with a maximum number of 5 binding events. We considered RegulonDB annotation as gold-standard; Resolution is defined as the minimum distance from an annotation to an estimated binding events and Sensitivity is defined as the fraction of annotation within 15 bp from an estimated binding event.

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

To perform the saturation analysis, we sub-sampled N fragments for both ChIP-exo and SE ChIP-Seq protocols. For PE ChIP-Seq we sub-sampled N pairs or $N/2$ fragments. For each seed, we called peaks using MOSAiCS [14] (GC content + mappability for ChIP-exo and Input for SE and PE ChIP-Seq) for the maximum sample size and to avoid false positives we considered only the top 500 peaks for each data protocol. We defined the number of candidate regions as the number of top sample peaks such that a binding events was estimated using the sampled reads and the dPeak's model; the number of predicted events is the total quantity of binding events estimated using the dPeak's model; the number of identified targets are number of gold-standard annotations within 15 bp from an estimated binding events; and the resolution is defined as the minimum distance from a gold-standard annotation to an estimated binding event. We repeated this analysis for ten seeds and reported the median between all those values.

Method comparison for ChIP-exo.

We considered dPeak Chung et al., 2013 [19], GEM Guo et al., 2012 [16], MACE Wang et al., 2014 [7] and Peazilla Bardet et al., 2013 [9] 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 [14] (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. We were unable to use CexoR [8] to estimate ChIP-exo binding sites.

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. 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)
4. 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)
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. Starick, S.R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M.I., Chung, H.-R., Vingron, M., Thomas-Chollier, M., Meijsing, S.H.: ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Research (2015)
7. 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)
8. Madrigal, P.: CexoR: an R/Bioconductor package to uncover high-resolution protein-DNA interactions in ChIP-exo replicates. EMBnet.journal (2015)
9. 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)
10. Wilbanks, E., Facciotti, M.: Evaluation of algorithm performance in ChIP-Seq peak detection. PLOS One (2012)
11. Pepke, S., Wold, B., Ali, M.: Computation for ChIP-seq and RNA-seq studies. Nature (2009)
12. 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)
13. 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)
14. 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)
15. 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)
16. 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)

17. 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)
18. Zhang, X., Robertson, G., Krzwinski, M., Ning, K., Droit, A., Jones, S., Gottardo, R.: PICS: Probabilistic inference for ChIP-Seq. Biometrics (2010)
19. 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)
20. 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
21. Kharchenko, P., Tolstorukov, M., Park, P.: Design and Analysis of ChIP-Seq Experiments for DNA-binding Proteins
22. Rozowsky, J., Euskirchen, G., Auerbach, R., Zhang, Z., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., Gerstein, M.: PeakSeq enables systematic scoring of ChIP-Seq experiments relative to controls. Nature, Biotechnology (2009)
23. Valouev, A., Johnson, D., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R., Sidow, A.: Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature, Methods (2008)
24. Grant, C., Bailey, T., Noble, W.S.: FIMO: Scanning for occurrences of a given motif
25. Mendenhall, E.M., Bernstein, B.E.: DNA-protein interactions in high definition. Genome Biology (2012)
26. 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)
27. 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)
28. Furey, T.S.: ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature Reviews: Genetics (2012)

5 Figures

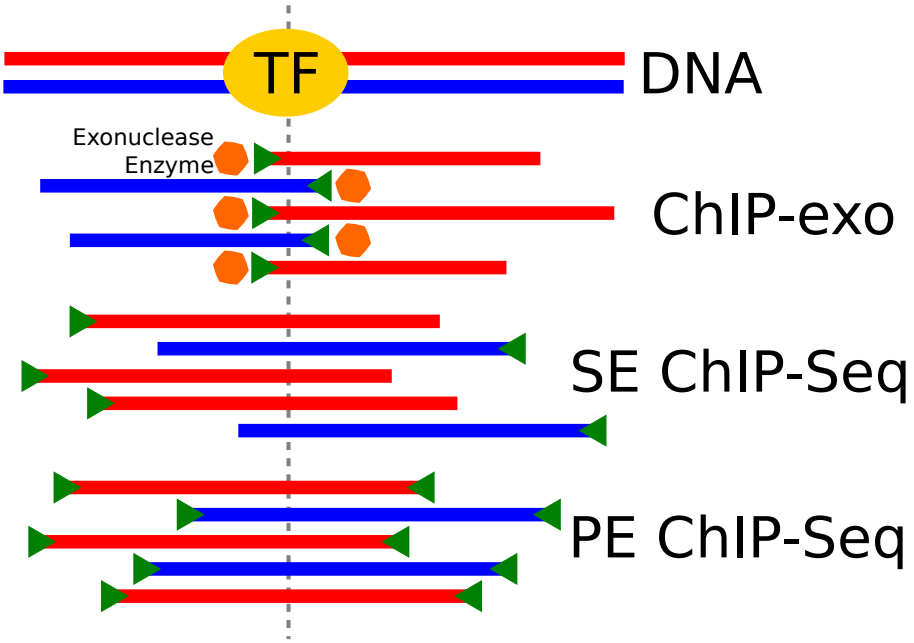


Figure 1 Description of ChIP-exo, SE ChIP-Seq and PE ChIP-Seq: A TF is bound to the forward (red) and backward (blue) strand of the DNA. Then, is sonicated: For ChIP-exo a exonuclease enzyme (orange hexagon) trims the $5\prime$ ends of each DNA fragment to a fixed distance from the bound protein, finally is subjected to Immunoprecipitation and amplification. For both ChIP-exo and SE ChIP-Seq an adapter is ligated (green triangles) at the $5\prime$ ends, while for PE ChIP-Seq is ligated to both ends.

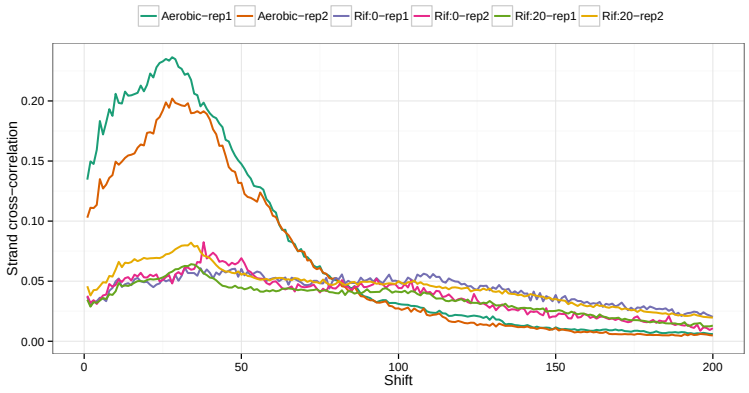


Figure 2 SCC curves for σ^{70} samples. The "phantom peak" and the summit that corresponds to the read and fragment length respectively are confounded due to the exonuclease digestion.

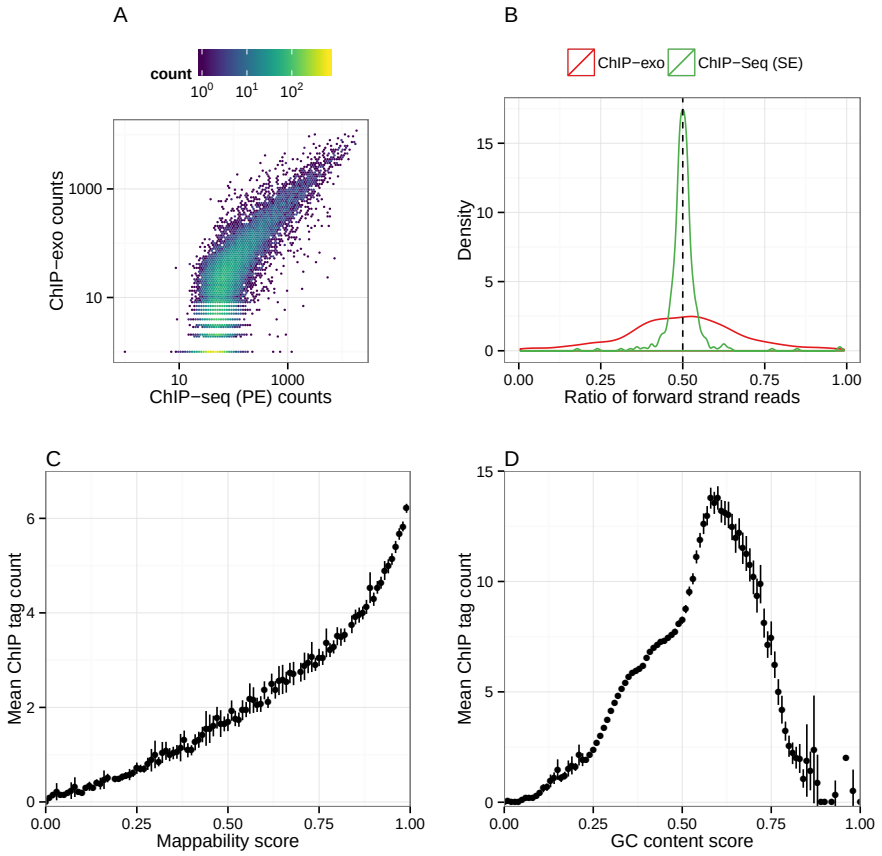


Figure 3 A) Hexbin plot of PE ChIP-Seq bin counts vs ChIP-exo bin counts. B) Forward Strand Ratio densities for SE ChIP-Seq and ChIP-exo peaks. C) Mappability score vs mean ChIP tag counts with 0.95 confidence bands. D) GC - content score vs mean ChIP tag counts with 0.95 confidence bands.

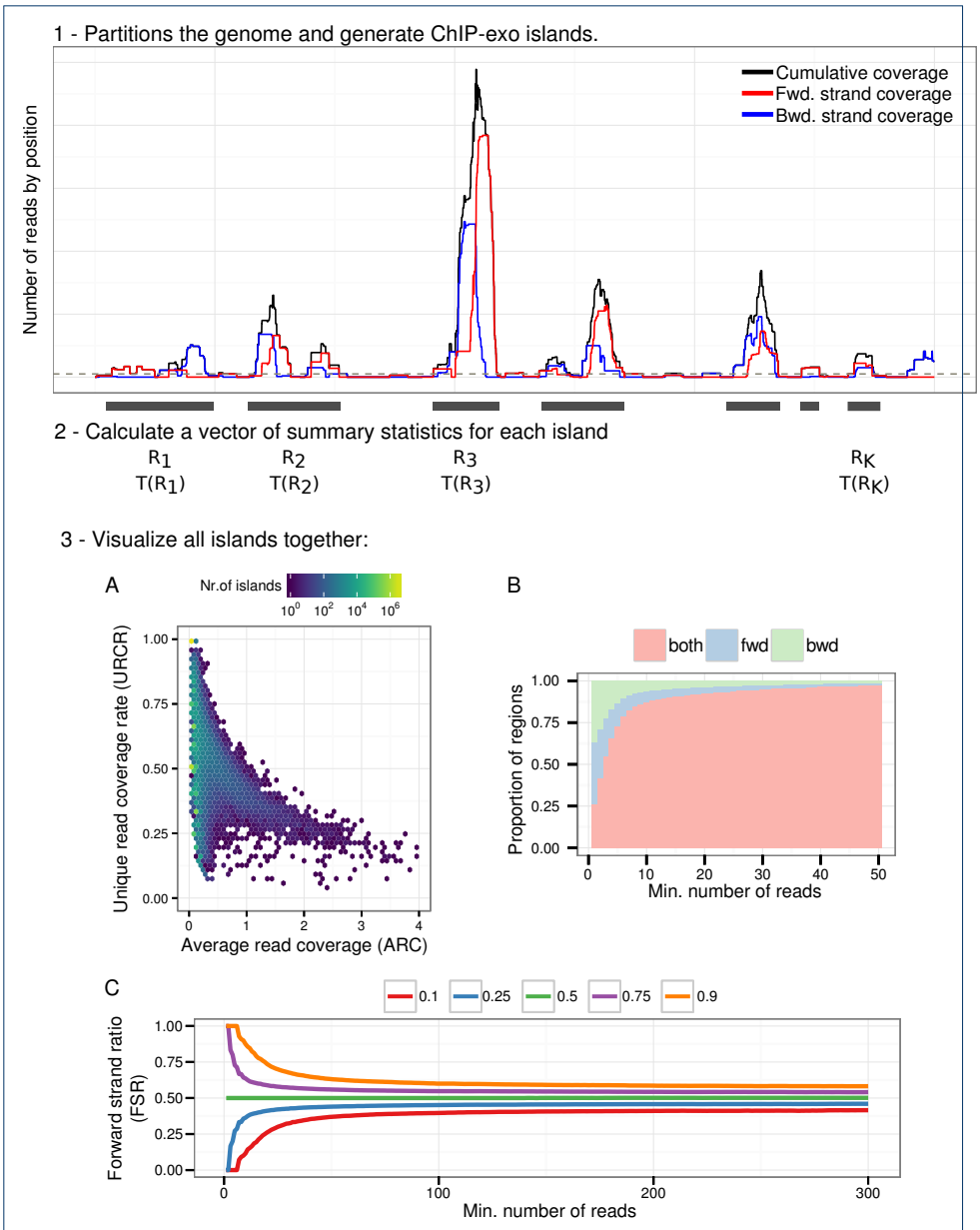


Figure 4 Flowchart of the ChIP-exo quality control pipeline. A) ARC vs. URCR, this plot presents a global view of the balance between library complexity and enrichment. There are two arms: First, one with low ARC, which corresponds to regions formed by few aligned positions; Second, where the URCR decreases as the ARC increases. B) Min. depth vs Proportions of regions, this plot shows the strand composition for all the regions formed by a min. number of reads. Regions with low depth tend to be formed by undigested reads from one unique strand, while regions with higher signal are usually formed by reads in both strands. C) Min depth vs. FSR, this plot depicts how quickly the FSR's distribution approximate the median. In a high quality sample, the median is around 0.5, and the other quantiles reach that value quickly.

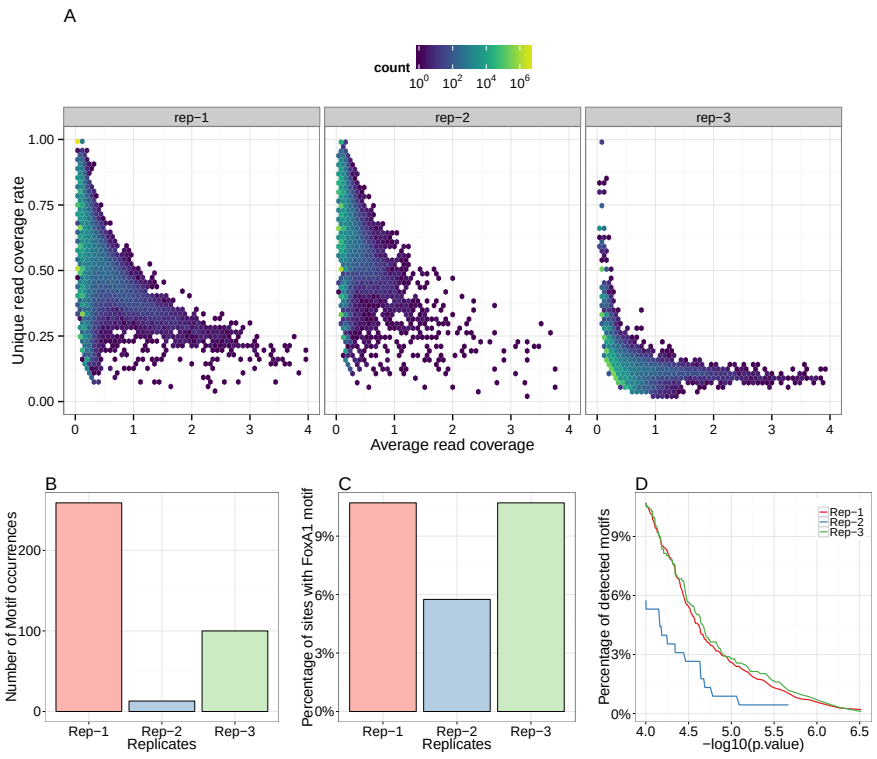


Figure 5 Using the mouse FoxA1 experiment from [4]: A) Hexbin plots of ARC against URCR, there is a slight separation into two strong arms, one corresponds to low ARC and varying URCR, and for the other URCR decreases as ARC increases. B) Number of candidate sites for each replicate. C) Percentage of candidate sites where the FoxA1 motif was detected. D) Cumulative proportion of detected motifs by replicate.

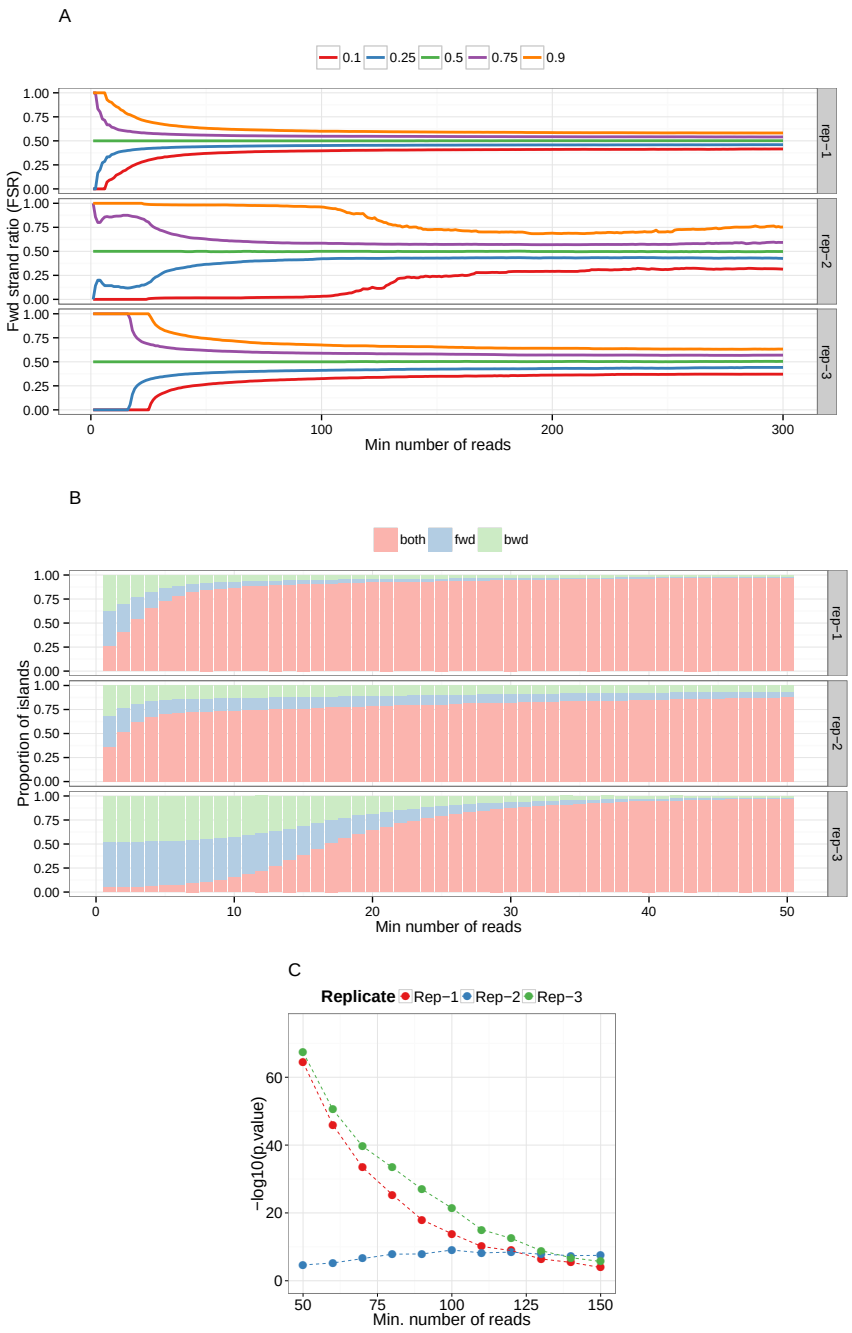


Figure 6 Strand imbalance QC plots for the same data as in Figure 5A. A) FSR distribution quantiles as the lower depth regions are being filtered out, all quantiles approach to the median as the lower bound increases. B) Stacked histogram with the proportion of regions that are formed by two strands or only one, in a good sample the single-stranded regions are going to be filtered out quickly as in the middle row. C) $-\log_{10}(\text{p.value})$ of testing if the imbalance distributions differs when ChIP-exo regions overlap their peaks.

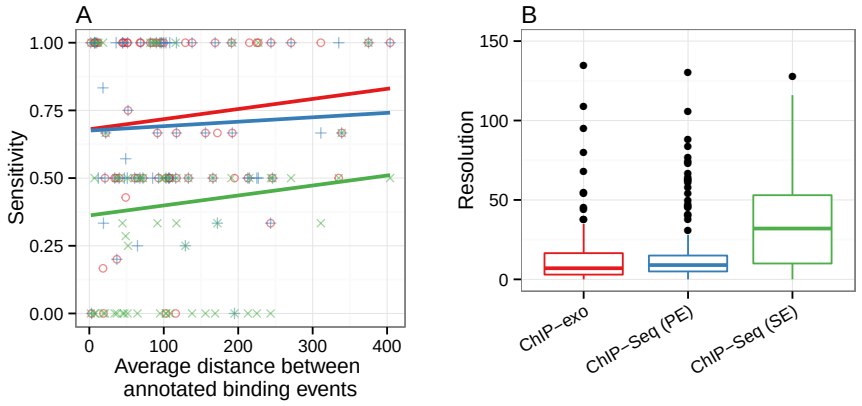


Figure 7 Comparison of (A) sensitivity and (B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction.

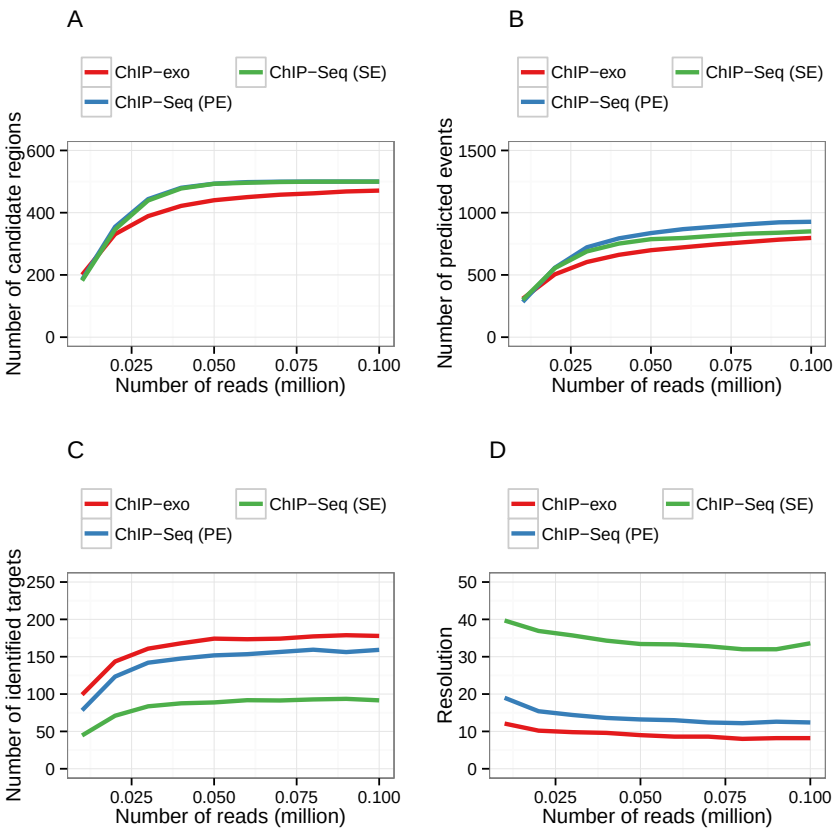


Figure 8 Comparison of the number of candidate regions (A), predicted events (B), identified targets (C) and resolution (D) among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

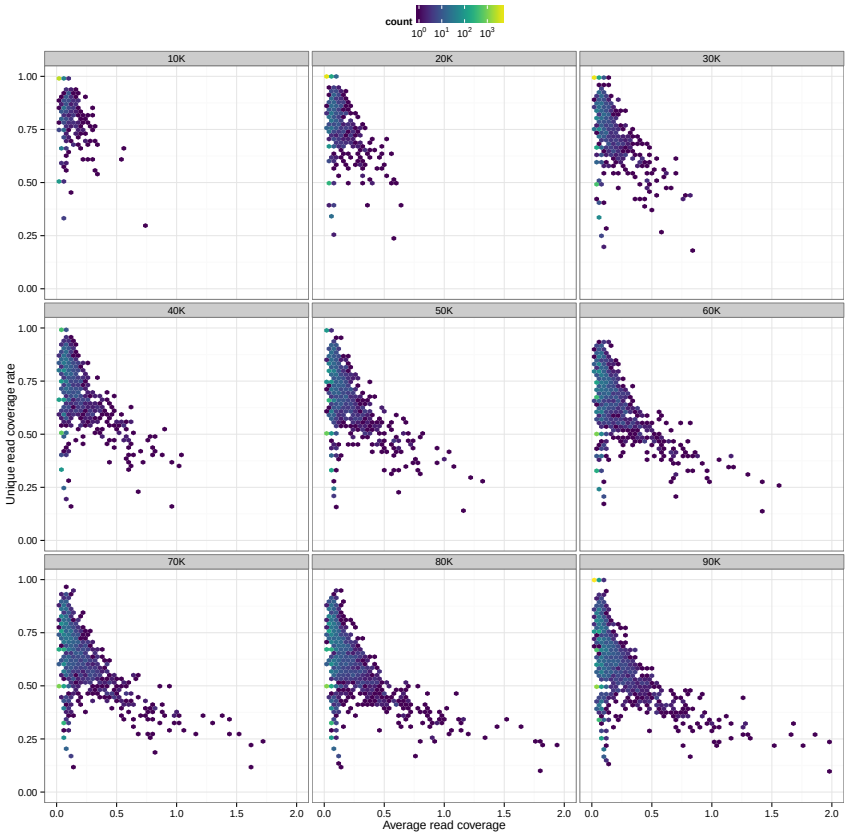


Figure 9 Hexbin plots of ARC vs URCR of the σ^{70} ChIP-exo experiment under aerobic condition when 10K to 90K reads are being sampled.

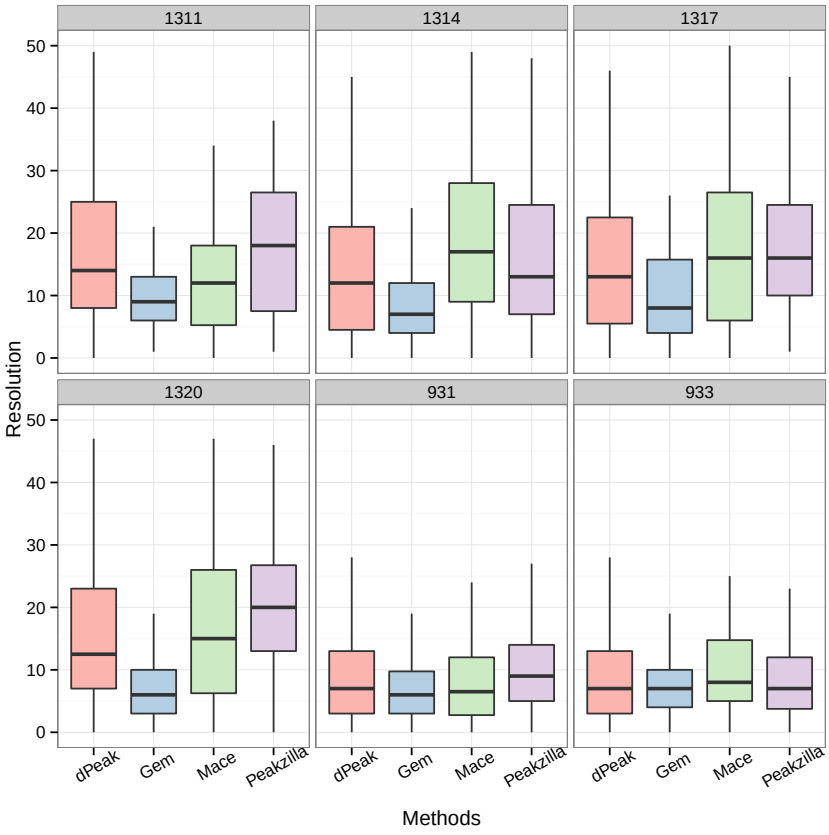


Figure 10 Comparison of the resolution between dPeak, Gem, Mace and Peakzilla methods. Resolution is defined as the minimum distance between a RegulonDB annotation and a predicted binding event.

Supplement

Note: I am not sure whether to include several figures on this supplement

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

Table S1 Same QC metrics as in table 1 but applied to Landick's chipseq data of the rif experiment

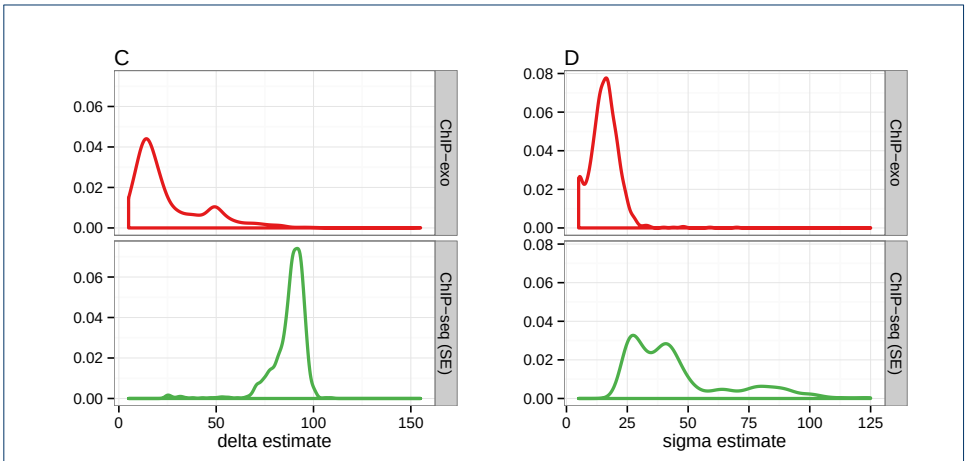


Figure S1 C) δ parameter in dPeak measures average distance of the reads to their respective binding site. In ChIP-exo data, reads were located much closer to the binding site than in SET ChIP-Seq. D) σ parameter measure the dispersion of reads around each binding site. In ChIP-exo data, reads showed less variation around the their respective binding sites compared to SET ChIP-Seq.

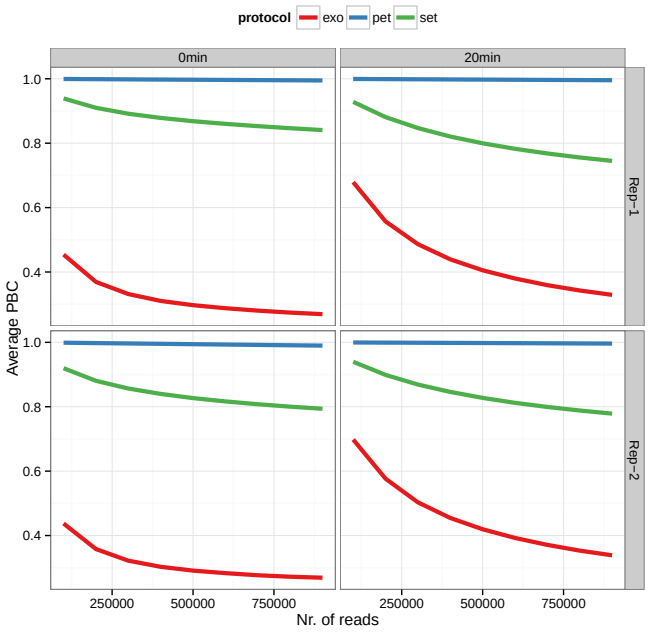
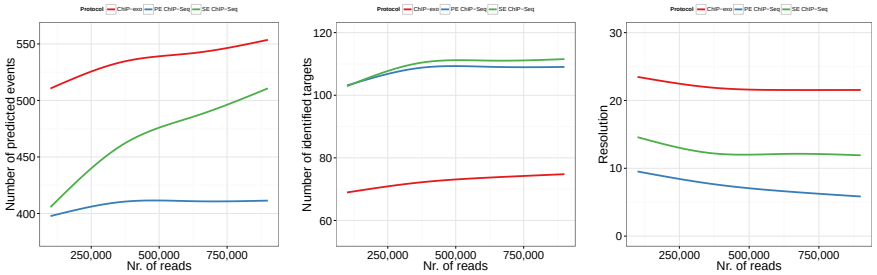
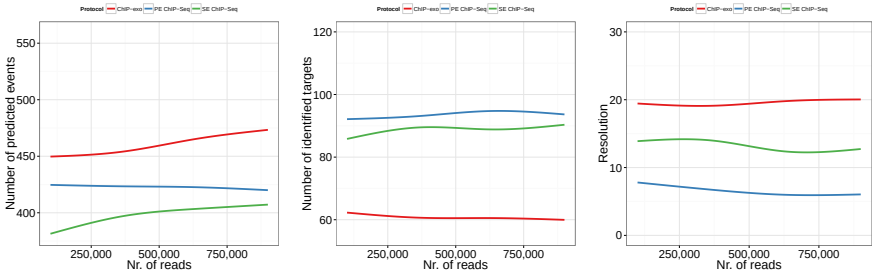


Figure S2 Average PBC (among all seeds) of the sampled ChIP-exo, PE ChIP-Seq and SE ChIP-Seq experiments under the rif treatment conditions used for saturation analysis.

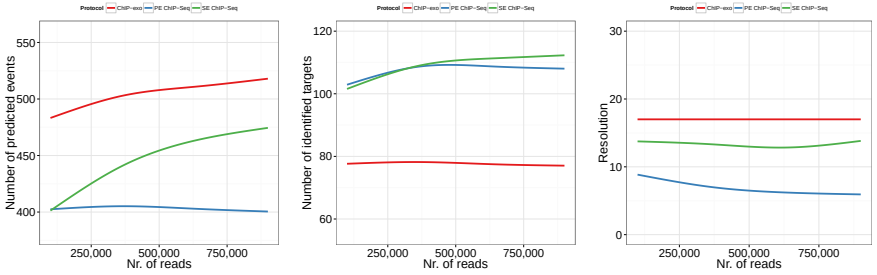
- Rep-1 and rif-0min:



- Rep-1 and rif-20min:



- Rep-2 and rif-0min:



- Rep-2 and rif-20min:

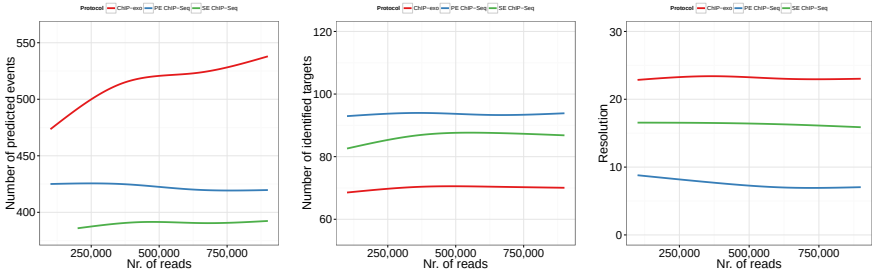


Figure S3 Comparison of the number of predicted events (left), identified targets (middle) and resolution (right) among ChIP-exo, PE ChIP-Seq and SE ChIP-Seq. RegulonDB annotations are considered as gold standard. A RegulonDB binding events was deemed identified if a binding event was estimated at a $\pm$ 15 vicinity of it.

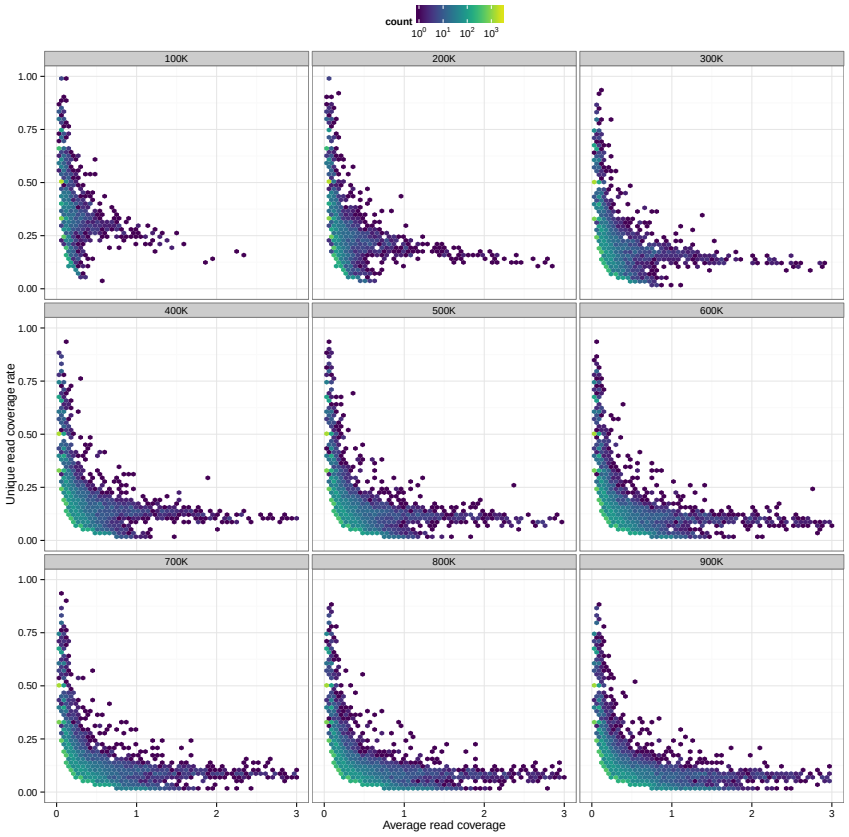


Figure S4 A) ARC vs URCR hexbin plots of Rep-1 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

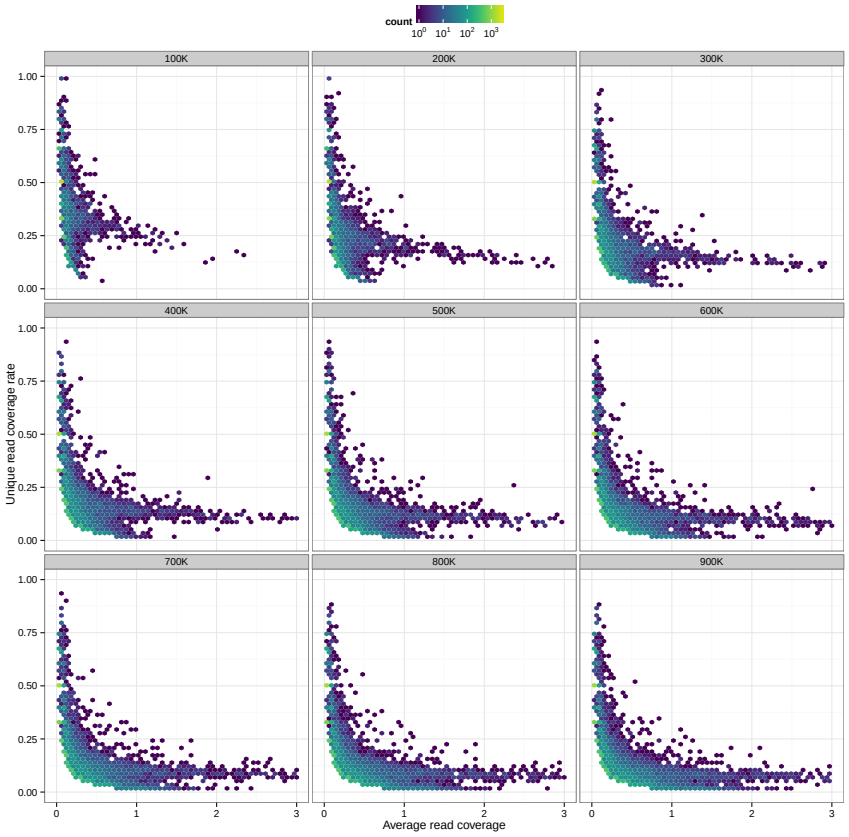


Figure S5 ARC vs URCR hexbin plots of Rep-1 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

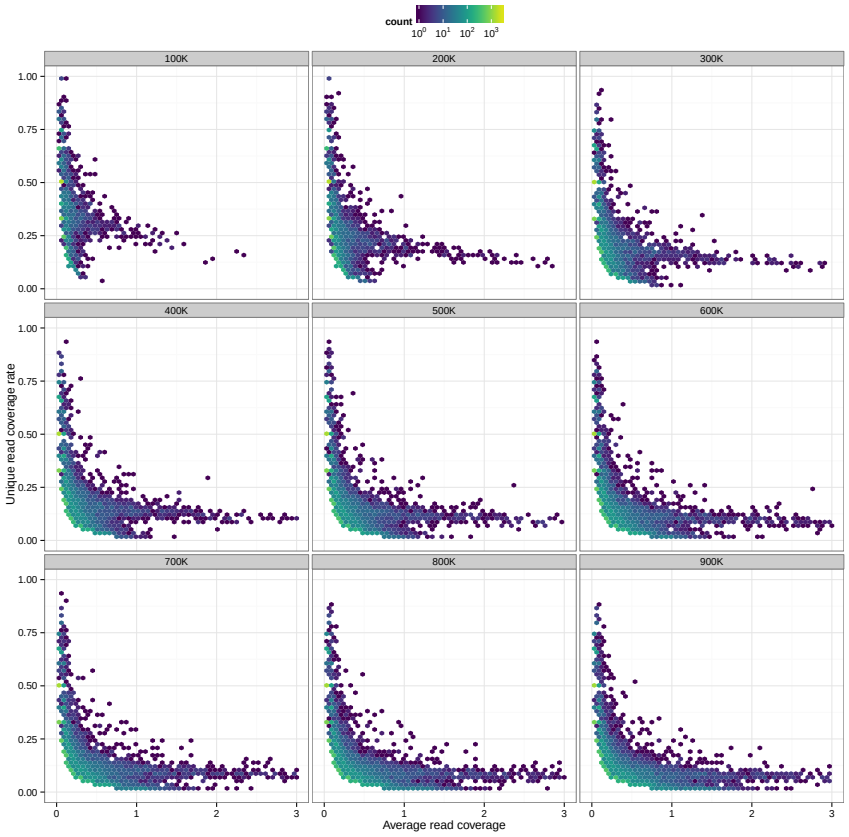


Figure S6 ARC vs URCR hexbin plots of Rep-2 and rif-0min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.

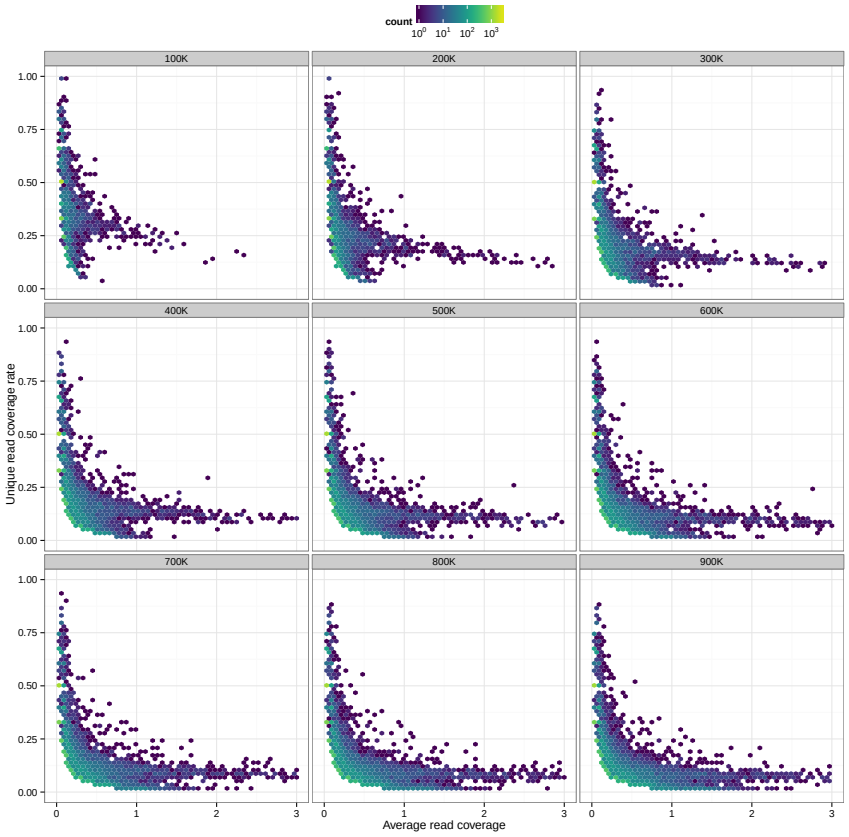


Figure S7 ARC vs URCR hexbin plots of Rep-2 and rif-20min from σ^{70} experiment when 100K to 900K reads are being sampled for each panel.