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

Rene Welch

Preliminary Examination

Department of Statistics, University of Wisconsin-Madison

December 1st, 2015

Committee Members:

Professor Sündüz Keleş, Department of Statistics, Department of Biostatistics and Medical Informatics

Professor Karl Broman, Department of Biostatistics and Medical Informatics

Professor Colin Dewey, Department of Computer Sciences, Department of Biostatistics and Medical Informatics

Professor Christina Kendziorski, Department of Biostatistics and Medical Informatics

Professor Ming Yuan, Department of Statistics

Contents

List of Figures

Abstract

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

We provide a quality control pipeline which visually assess ChIP-exo biases, library complexity and enrichment; and calculates a signal-to-noise measure. Also, we updated dPeak (Chung et al., 2012 [4]), which makes a striking balance in sensitivity, specificity and spatial resolution for ChIP-exo data analysis.

1 Introduction

ChIP-exo (Chromatin Immunopecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh, 2011 ([23]) 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.

While the number of produced ChIP-exo data keeps increasing, characteristics of ChIP-exo data and optimal strategies for experimental design and analysis of ChIP-exo data are not fully investigated yet, including issues of sequence biases inherent to ChIP-exo data, choice of optimal statistical methods, and determination of optimal sequencing depth. However, currently the number of available ChIP-exo data is still limited and their sequencing depths are still insufficient for such investigation. To address this limitation we gathered ChIP-exo data from diverse organisms: CTCF factor in human [23]; ER factor in human and FoxA1 factor in mouse (Serandour et al., 2013 [26]); and generated σ^{70} factor in Escherichia Coli (E. Coli) under aerobic ($+O_2$) condition, and treated by rifampicin by 0 and 20 minutes.

DNA libraries generated by the ChIP-exo protocol seem to be less complex than the libraries generated by ChIP-Seq (Mahony et al., 2015 [19]). Hence, most of current QC guidelines (Landt et al., 2012 [16]) may not be applicable on ChIP-exo, additionally to our knowledge there are not established quality control pipelines for ChIP-exo. To address this challenge, we suggest a collection of quality control visualizations to understand which biases are present in ChIP-exo data. Previous ChIP-exo analysis used ChIP-Seq samples to compare the resolution between experiments ([23], [24], [26]); Carroll et al., 2014 [3] studied the use of the Strand-Cross Correlation (SCC) (Kharchenko et al., 2008 [14]) and showed that by filtering blacklisted regions the estimation of the SCC is improved. However, this method requires to know blacklisted regions in advance which may not be available, additionally using SCC may not be useful since the peaks that are attained at the read and fragment length are confused in a typical ChIP-exo SCC curve. In our pipeline we propose two out-the-shelf analysis: an enrichment plot and the local normalized SCC cofficient.

In order to archive 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 [23] discussed that reads in the forward and reverse strand might construct peak pairs around bound protein, of which heights were implicitly assumed to be symmetric. Hence, they used the

"peak pair method" that predicts the midpoint of two modes of peak pairs as potential binding site. Mace (Wang et al., 2014 [28]), CexoR (Madrigal, 2015 [18]) and peakzilla (Bardet et al., 2013 [1]), 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 [30] and Pepke and Wold, 2009 [21]), such as MACS (Zhang et al., 2008 [32]), CisGenome (Hongkai et al., 2008 [13]) 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 [17]), GEM (Guo et al., 2012 [7], an improved version of Guo et al., 2010 [8]) and PICS (Zhang et al., 2010 [31]) have been proposed to identify binding sites in higher resolution using ChIP-Seq data. However, most of them are still not tailored for ChIP-exo and PET and SET ChIP-Seq data in a unified framework and as a result, currently available methods are not appropriate for fair comparison between ChIP-exo and ChIP-Seq. To address these limitations, we developed an improved of dPeak (Chung et al., 2013 [4]), a high resolution binding site identification (deconvolution) algorithm that we previously developed for PET and SET ChIP-Seq data, so that it can also handle ChIP-exo data. The dPeak algorithm implements a probabilistic model that accurately describes the ChIP-exo and ChIP-Seq data generation process.

In this work, we demonstrate that the "peak pair" assumption of Rhee and Pugh [23] 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 method to identify binding events and dPeak outperforms or performs competitively respecto 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 PET ChIP-Seq and both significantly improve

the resolution of protein binding site identification compared to SET-based analysis with any of the available methods.

2 ChIP-exo and ChIP-Seq review

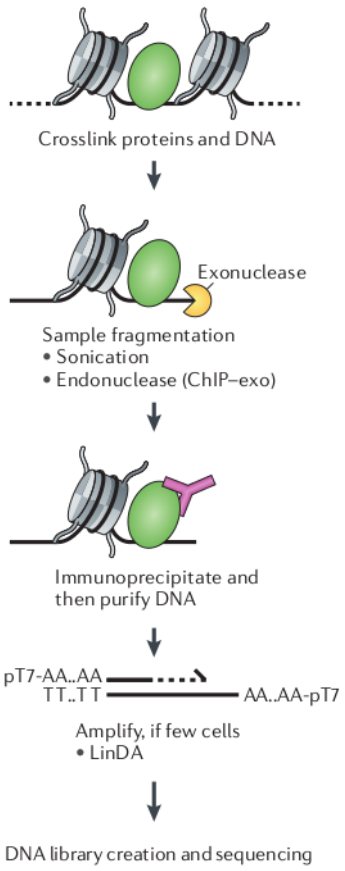


Figure 1: ChIP-exo diagram. The only difference with ChIP-Seq is the second step, where an exonuclease enzyme is added and it digest the DNA fragments starting from the 5' end until it finds a crosslink protein (this figure is taken from Furey, 2012 [6]).

ChIP-exo is based in a modification to the ChIP-seq protocol, where an exonuclease enzyme is added and it digests the fragments starting by the 5' ends and stops until it finds a protein that is crosslinked to the DNA.

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 5' end of each DNA fragment to a fixed distance from the bound protein. As a result, boundaries around the protein of interest constructed with 5' ends of fragments are located much closed to bound protein compared to ChIP-Seq. This 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 5' end of each fragment generates millions of reads or tags.

Since this protocol is a modification to ChIP-Seq, some of its characteristics are still maintained in ChIP-exo while several new are being found. Rhee and Pugh, 2011 [23], showed that ChIP-exo generated fragments are located around the binding events more tightly than ChIP-Seq generated fragments. Hence, it is of utmost importance to understand how this new protocol works, specifically we focused on understanding which biases are mantained from ChIP-Seq to ChIP-exo, which biases are specific to ChIP-exo, how the current ChIP-Seq QC guidelines behave in ChIP-exo and we developed a new QC pipeline specific to ChIP-exo. A typical

example is shown in figure 2, which shows the σ^{70} and β'_f factors (and the β factor only on top) on a E. Coli genome's region: First, both regions show that the σ^{70} sample is enriched; second the σ^{70} sample in the ChIP-exo panel shows that its respective form in the ChIP-Seq panel may be form of more binding events; and third, either of the β factors show that the complexity of the ChIP-exo libraries may be lower in some cases, i.e. the peak is mapped with fewer positions.

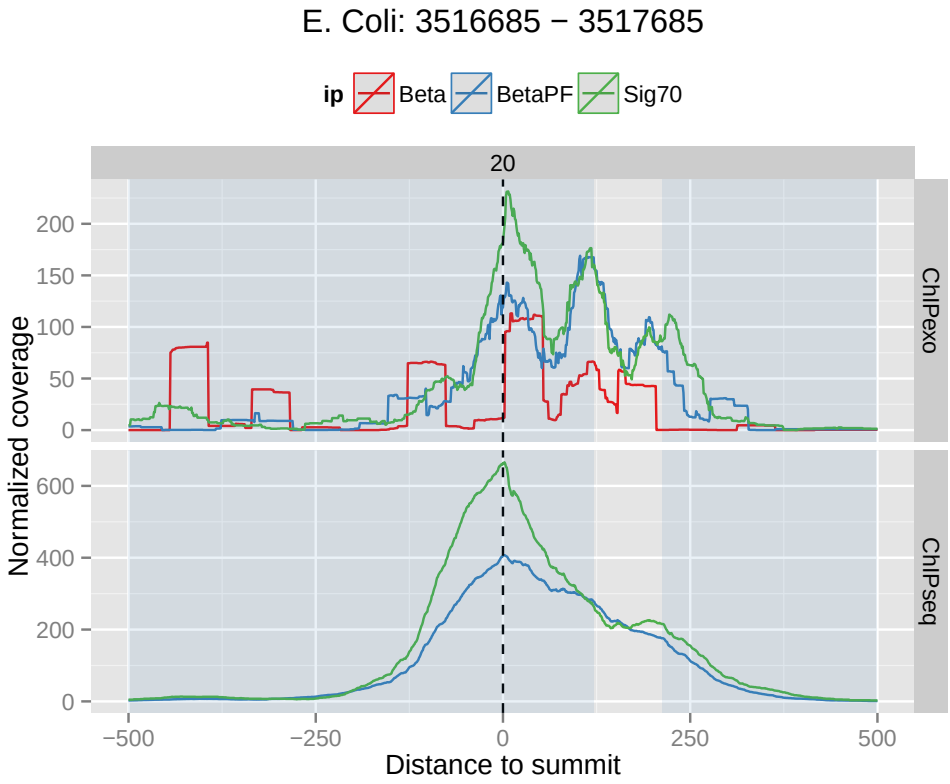


Figure 2: An example of a PET ChIP-Seq peak compared to ChIP-exo. σ^{70}, β and β'_f are subunits of the transcription initiation complex in E. Coli. Gene regions on the reverse strand are highlighted in light blue and the region was centered respect to ChIP-Seq σ^{70} summit. The library complexity is lower in ChIP-exo than in ChIP-Seq but the resolution is increased, this can be observed at the σ^{70} peak that for ChIP-exo was created as the combination of at least three ChIP-exo "sub-peaks".

2.1 Current Quality Control Measures for ChIP-Seq

To our knowledge it doesn't exists a QC pipeline for ChIP-exo. Hence, we analyzed how the current ChIP-Seq QC measures behave on ChIP-exo data.

Table 1 contains measures that are commonly used with ChIP-Seq data. PBC stands for PCR

bottleneck amplification and is defined as the ratio between the number of positions to which exactly one mappping read maps divided by the number of position to which at least one mapping reads maps, the ENCODE project proposed a series of ranges to interpret this quantity (which can be found in https://www.encodeproject.org/data-standards/2012-quality-metrics/), but it doesn't apply to ChIP-exo, if it did all the samples with PBC < 0.5 should be reject due to be showing severe bottlenecking. Planet et al., 2011 proposed the Standarized Standard Deviation which defined as:

$$\text{SSD} = \text{SD}_N/\sqrt{N}$$

where SD_N is the standard deviation of the coverage of the sample and N is the number of reads. This measure is smaller for ChIP-exo since the reads are located more tightly around binding sites, but it may be biased since several fragments are being digested by exonuclease enzyme, hence there are several position in the genome where no fragment is being mapped. Finally the normalized strand cross-correlation is defined as:

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

where $y(\delta)$ is strand cross-correlation defined as in (1).

2.1.1 Strand cross-correlation

The strand cross-correlation (SCC) was proposed by Kharchenko et al., 2008 [14] 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 propotion of tags mapped to chromosome c and n_c^S is the tag count vector for strand S and chromsome c.

For ChIP-Seq data, the SCC will have two peaks: One when δ is equal to reads length and one when δ is equal to the unobserved fragment length, ideally the SCC function is maximized at this point. Landt et al., 2012 [16] explain that a succesful if this happens, it is marginally succesful if its maximized at the read length peak but still shows the fragment length summit and failed if is maximized at the read length peak but the second summit is not present.

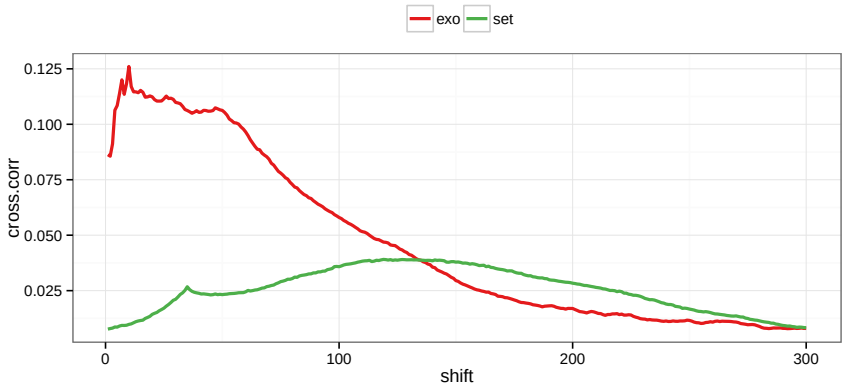


Figure 3: SCC for CTCF factor in HeLa cell line for ChIP-exo (obtained from [23]) and SET ChIP-Seq (obtained from [22])

Figure 3 shows the SCC function for both a ChIP-exo (in red) and a SET ChIP-Seq (in green) for CTCF in HeLa cell line. The ChIP-Seq dataset behaves as the succesful case from [16], but the ChIP-exo doesn't. Additionally, since the exonuclease enzyme digests the DNA fragments, then the shifts where the SCC may reach a locall or a global maximum are confounded. Hence, new QC measures are necessary for ChIP-exo data.

Table 1: Usual quality control indicators applied to the σ^{70} samples. PBC stands PCR-bottleneck coefficient, SSD for standardized standard deviation, NSC for normalized strand cross-correlation coefficient (Another measure related to the strand cross-correlation curve is the RSC was ommited since input sample usually don't exists for ChIP-exo). σ^{70} samples were given by Robert Landick, FoxA1 and ER samples are from Serandour et al., 2013 [26] and CTCF sample is from Rhee and Pugh, 2011 [23].

IP	Organism	Condition	Rep.	Nr. reads	PBC	SSD	NSC
σ^{70}	E.Coli	Rif-0min	1	960,256	0.2823	0.0361	10.29
σ^{70}	E.Coli	Rif-0min	2	2,247,295	0.2656	0.1091	25.08
σ^{70}	E.Coli	Rif-20min	1	1,940,387	0.2698	0.0820	17.69
σ^{70}	E.Coli	Rif-20min	2	4,229,574	0.2153	0.1647	14.11
FoxA1	Mouse	NA	1	22,210,461	0.6562	9.12×10^{-5}	21.452
FoxA1	Mouse	NA	2	22,307,557	0.7996	7.94×10^{-5}	60.661
FoxA1	Mouse	NA	3	22,421,729	0.1068	1.31×10^{-4}	72.312
ER	Human	NA	1	9,289,835	0.8082	3.64×10^{-5}	19.843
ER	Human	NA	2	11,041,833	0.8024	4.6×10^{-5}	21.422
ER	Human	NA	3	12,464,836	0.8203	4.89×10^{-5}	19.699
CTCF	Human	NA	1	48,478,450	0.4579	1.29×10^{-4}	15.977

2.2 The dPeak model

Chung et al., 2013 proposed the dpeak model to identifity transcription factor binding sites in high resolution. Two models are proposed: One to deconvolve the signal of single end tags (SET) data, i.e. when only one end of the DNA fragment is sequenced; and another to deconvolve the signal of paired end tags (PET) data, i.e. when both ends of the DNA fragment are sequenced. For ChIP-exo, both ends are sequenced but only the 5' end is considered, since is the side where the exonuclease enzyme digests the DNA.

Given a peak region with m positions, with n DNA fragments and that the region is generated by a known g^* amount of binding events. Each fragment is described by the start position where the starting end is aligned and the strand from which the fragment is sequenced. Since only one end is sequenced, then the fragment length is not observed. dPeak assumes that the strand D_i follows a Bernoulli distribution with known parameter p_D, and the start position R_i for the i-th fragment is generated according to the following procedure:

- If the fragment was sequenced from the forward strand ($D_i = 1$):
  - The reads belongs to the background: $R_i|Z_i = 0, D_i = 1 \sim \text{Unif}(1 - \beta + 1, m)$
  - The read belong to the g-th binding event: $R_i|Z_i = g, D_i = 1 \sim \text{N}(\mu_g - \delta, \sigma^2)$
- If the fragment was sequenced from the backward strand ($D_i = 0$):
  - The reads belongs to the background: $R_i|Z_i = 0, D_i = 1 \sim \text{Unif}(1, m + \beta - 1)$
  - The read belong to the g-th binding event: $R_i|Z_i = g, D_i = 1 \sim \text{N}(\mu_g + \delta, \sigma^2)$

  and the binding event to which the i-th fragment belongs $Z_i = g$ is generated from a Multinomial$(\pi_0, \pi_1, \cdots, \pi_{g^*})$.

2.2.1 dPeak's algorithm initialization strategy

dPeak's model likelihood is optimized by the stochastic EM algorithm. Hence, it is of utmost importance how the parameters are initialized in order to obtain a global maximum.

Chung et al., 2013 [4] derived an initialization strategy for the stochastic EM algorithm where each region is initialized by separate independently and the binding sites locations $\mu_g, g = 1, \cdots, g^*$ are initialized by a uniform partition.

For this work, we considered a new strategy where the all the regions are ranked by a statistical measure (as the $-\log(p)$ where p is the empirical p.value of the test $H_0 : \mathcal{R}$ is enriched, where $\mathcal{R}$

is a genomic region) or a empirical quantity (as the number of ChIP tag counts in the region) and use as initial estimates of the δ and σ parameters the average of the estimates from the highest ranked N_{top} peaks.

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 4A). 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 background distribution of ChIP-exo data is less homogeneous than that of ChIP-Seq data.

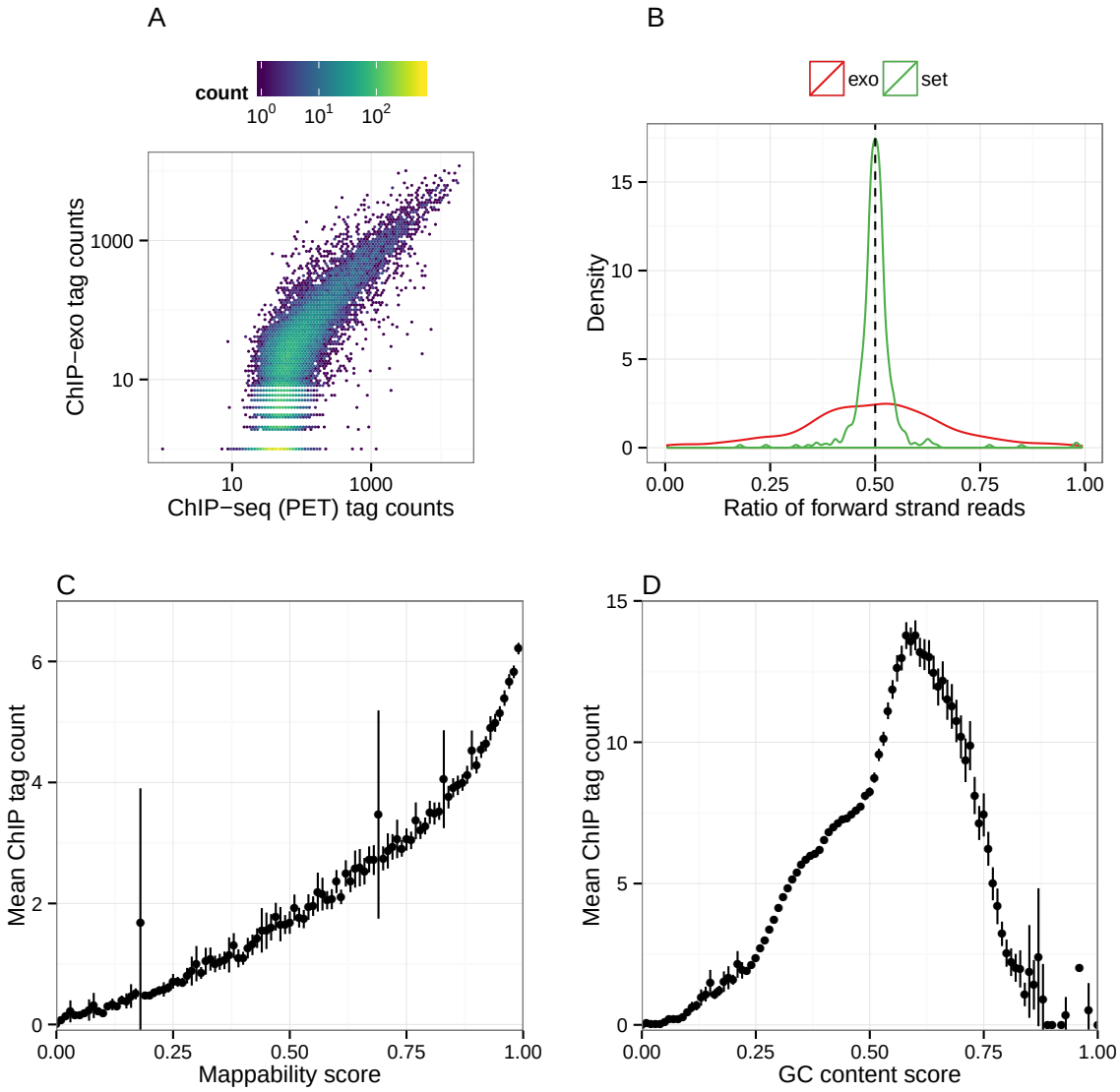


Figure 4: A) ChIP signal in ChIP-exo was linearly related to that in PET ChIP-Seq data in the region with high ChIP tag counts. In contrast, there were clear differences in their background distribution, where several ChIP-exo background regions were empty. B) In ChIP-exo data, strands of reads were significantly less balanced in the regions with potential binding sites compared to SET ChIP-Seq data. 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.

We next evaluated the "peak pair" assumption from Rhee and Pugh, 2011 [23], 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 [28], Madrigal 2015 [18] and Bardet et al., 2013 [1] 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. 4B) 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 [23] to investigate issues specific to eukaryotic genomes for binding sites identification. Figures 4C and 4D 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.

3 Results

3.1 ChIP-exo Quality Control pipeline

Figure 5 shows a flowchart for the ChIPexoQC pipeline. Which basically partitions the genome by keeping the non-digested ChIP-exo regions. Then, for each region it calculates a series of statistics. Finally, it creates several visualizations designed to assess the quality levels of a ChIP-exo sample.

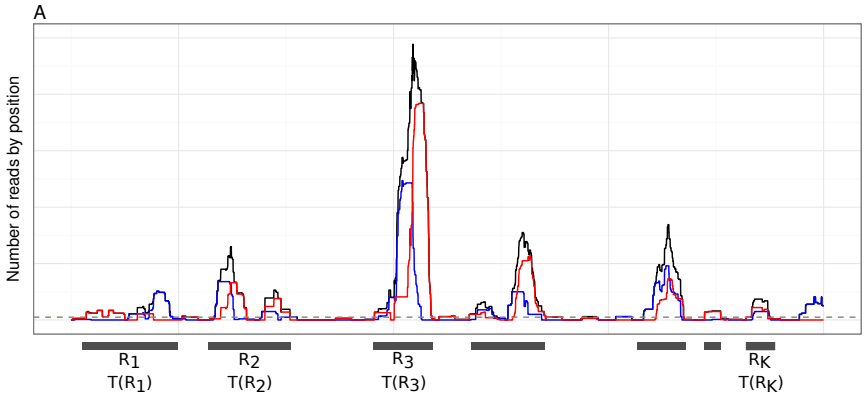


Figure 5: Diagram of the ChIP-exo Quality control pipeline. The genome is partitioned into several regions by removing the empty segments and for each regions we calculate a collection of QC indicators, then for each region R_k a vector of summary statistics $T(R_k)$ is calculated.

3.1.1 Local Normalized Strand cross-correlation

We considered the creation of a novel QC measure for ChIP-exo called local normalized SCC or local-NSC. This measure

$$y(\delta) = r\left[n^{+}\left(x + \frac{\delta}{2}\right), n^{-}\left(x - \frac{\delta}{2}\right)\right]$$

where again r is the Pearson correlation but n^{+} and n^{-} are the coverage vectors of a given region in the genome. Figure 6 shows that as the library complexity of the region is lower, then the local SCC seems to show higher variance. Hence, we fit a nonparametric regression model to smooth the SCC signal:

$$y(\delta) = f(x_{\delta}) + \epsilon_{\delta}$$

where $y(\delta)$ is the observed SCC for shift x_{δ}, $\epsilon_{\delta} \sim N(0, \sigma_f^2)$. Finally the local-NSC is defined as a signal-to-noise ratio by using:

$$\text{local-NSC} = \frac{max_{x_{\delta}}\hat{f}(x_{\delta})}{\hat{\sigma}_f} \tag{2}$$

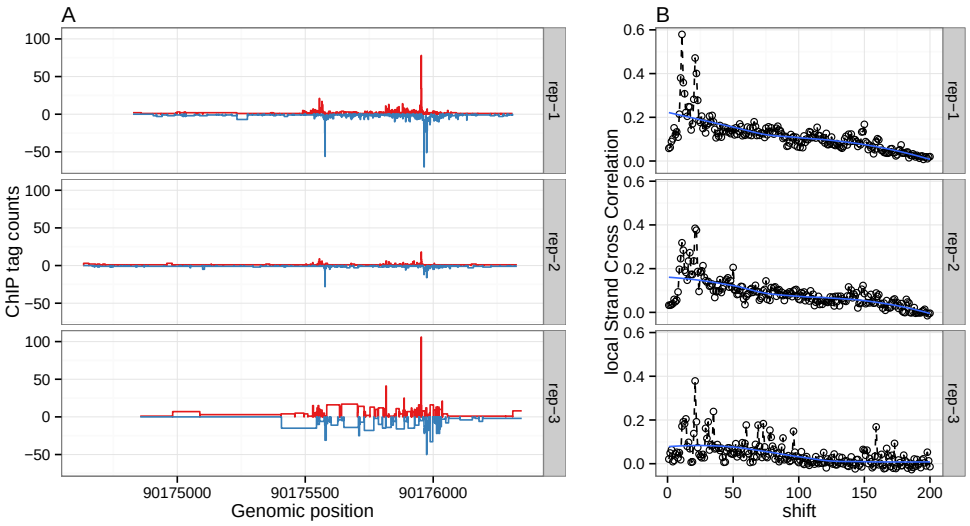


Figure 6: Using the FoxA1 sample from [26]. A) shows an example region for the 3 replicates. The library complexity seem to be comparable between the two top rows and both outperfom the last row. The number of reads for the first replicate is higher than the number of reads for the other replicates. B) Shows the local SCC for the 3 replicates, this measure is comparable among the 3 replicates. The first replicate shows the lowest residual error, while the third one shows the highest. In blue, a "loess" regression model is fitted.

3.1.2 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. To diagnose this, we consider the Average Read Coverage (ARC) and the Unique Read Coverage Rate (URCR) which are defined as:

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

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

Using the mouse-FoxA1 experiment from [26] and the relationship between this two quantities, library complexity and sample enrichment was explored. Figure 7A shows hexbin plots illustrating the interaction between these two quantities. Notice how there are two strong arms in each panel: The first one corresponds to regions with low ARC values and varying URCR values across the

$(0, 1]$ interval, while the second one shows a decreasing trend in URCR as ARC increases. When an experiment shows a higher degree of enrichment, then the separation of this two arms is more noticeable, since the second arm corresponds to possibly enriched regions (12A and 12B).

Figure 7B shows box plots of the local-NSC on regions of the three replicates, stratified by the number of reads mapped exactly to one position for all σ^{70} ChIP-exo experiments; The high stratum is defined by regions consisting of more than 100 unique positions, the medium stratum for regions where the number of unique position is on the (50,100) range and the low stratum in the (20,50) range. For each stratum and biological replicate, we sampled 400 regions whenever possible (in the opposite case, all the regions in the category were considered) and calculated the local-NSC for those regions. For all three replicates, it is shown an increasing trend as the number of unique position increases, which means that the local-NSC is an effective indicator of library complexity. Figure 7C shows the local-NSC coefficient calculated for the same set of regions, stratified by the number of unique positions in the regions of the first replicate using the same criteria as before. In the high panel it is shown that the first replicate sample performs slightly better than the second and both outperform the third one, this agrees with 7A, where only in the left-most panel doesn't seem to be a separation between the low ARC arm and the other.

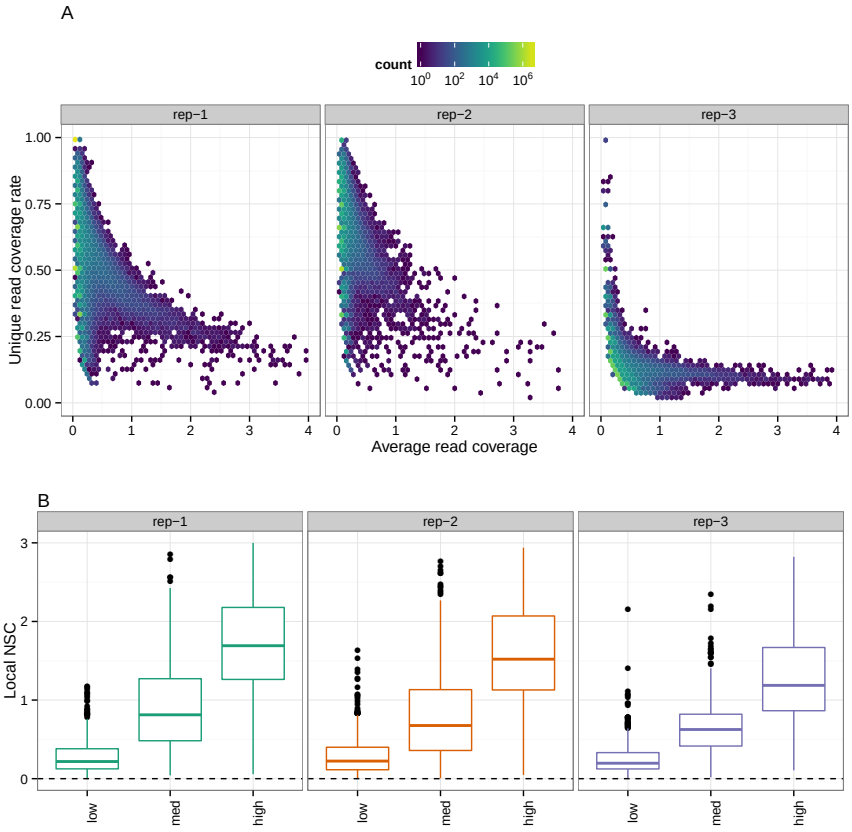


Figure 7: Using the mouse-FoxA1 experiment from [26]: A) Hexbin plots of ARC against URCR, in general we can see a slight separation into two strong arms, one corresponds to low ARC and varying URCR, and for the other URCR decreases as ARC increases B) Boxplots of the local-NSC stratified by nr. of reads mapped to only one position in the regions for the ChIP-exo experiments divided by condition and biological replicate.

3.1.3 Strand imbalance in ChIP-exo data

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

the sample locally, therefore the assumption doesn't hold in practice.

In order to assess the strand imbalance we created the following visualization presented in figure 8. A) shows the FSR's behavior as lower depth regions are being filtered out, while B) shows which percentage of the regions are composed by fragments of both strands or only one (forward or backward). As the number of reads per region's lower bound increases, the quantiles tend to approach the median. To evaluate this statement, for each replicate of the FoxA1 dataset, we divided the regions formed by at least 100 fragments into into the ones that overlap with any of the sample's peaks. Then we considered a Kolmogorov-Smirnov tests to assess if the distribution of both groups were the same. The p.value for the third replicate's test is 0.1039, for the first replicate's test is 0.0064 and for the second test is 0.0871. This shows that only the middle sample from figure 8 is the least imbalance sample from the the three, while the other two still show a higher imbalance at the 100 level.

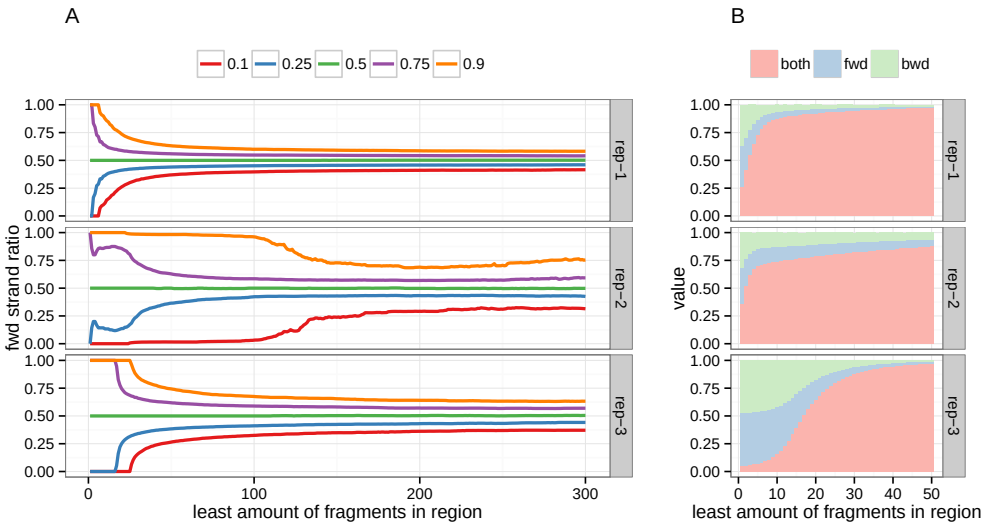
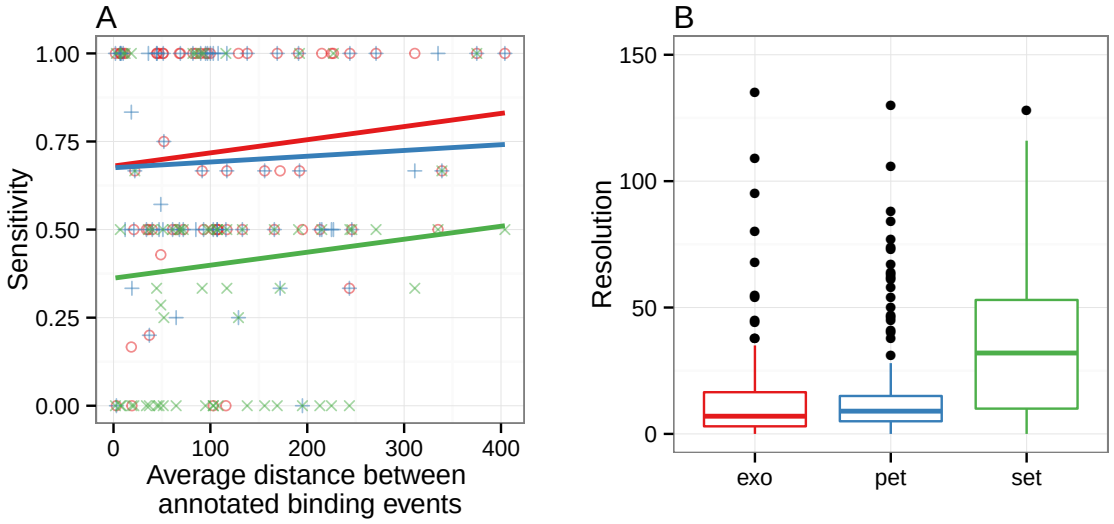


Figure 8: Strand imbalance QC plots for the same data as in figure 7. A) FSR distribution quantiles as the lower depth regions are being filtered out, all quantiles approach to the median as the lower bound increases. B) Stacked histogram with the proportion of regions that are formed by two strands or only one, in a good sample the single-stranded regions are going to be filtered out quickly as in the middle row.

3.2 Comparison with ChIP-Seq data using dPeak

Figure 9 shows different comparisons among ChIP-exo, PET ChIP-Seq and SET ChIP-Seq. A RegulonDB annotation was considered identified if the distance between it and 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. In figure 9A can be seen that the sensitivity of all protocols increases as the average distance between binding sites does. Despite that when the binding events in a peak are closer to each other, both ChIP-exo and PET ChIP-Seq are comparable, as the distance increases ChIP-exo identifies a higher proportion of the RegulonDB annotations; additionally SET ChIP-Seq is significantly less sensible than both ChIP-exo and PET ChIP-Seq. In figure 9B the distance between a RegulonDB annotation to its closest prediction is compared for ChIP-exo, PET and SET ChIP-Seq; while the first two are comparable, both outperform SET ChIP-Seq.

In figures 9C and 9D, we observe the behavior of dPeak's estimated parameters for single end data (both ChIP-exo and ChIP-Seq). The δ parameter measures the distance from the 5' end of the fragment to its respective binding event and the σ parameter indicates the fragments distribution variability around their respective binding sites (more information in [4]). For both parameters, it can be seen that the ChIP-exo estimates are smaller than the SET ChIP-Seq estimates in average, which agrees with the fact that the sample reads are allocated more tightly around the binding events in ChIP-exo data. Hence we can conclude that the peaks shape is quite different between ChIP-exo and SET ChIP-Seq.



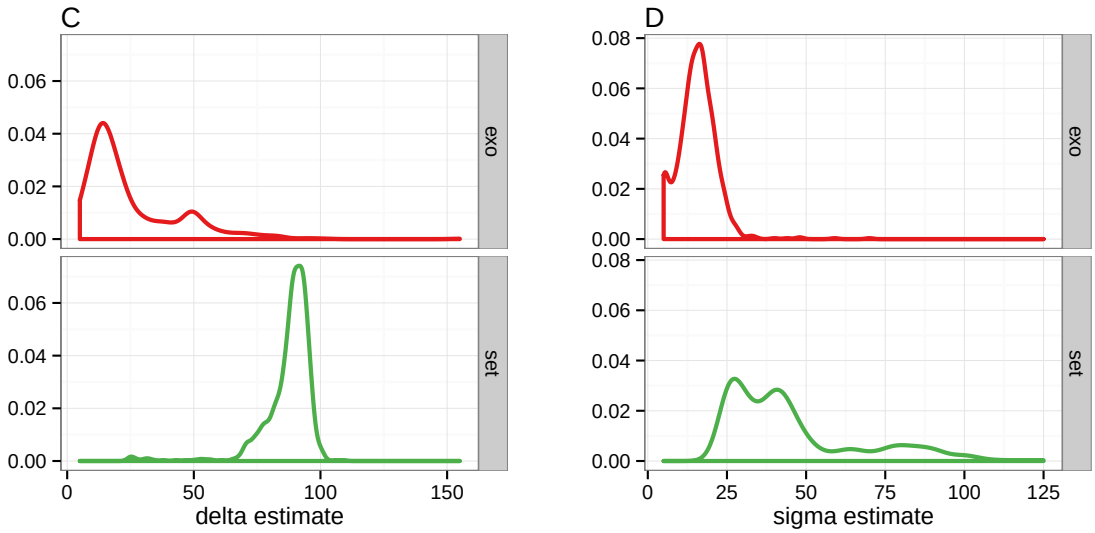


Figure 9: Comparison of (A) sensitivity and (B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction. C) δ parameter in 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.

3.3 Recommendations for the design of ChIP-exo experiments

We sampled a fixed amount of fragments for each of the ChIP-exo, PET ChIP-Seq and SET ChIP-Seq datasets of the σ^{70} sample in aerobic conditions. For each sampled dataset we applied our lower-to-higher resolution pipeline by calling peaks with MOSAiCS [15] and then deconvolving the binding events by using dPeak [4]. For the ChIP-exo datasets we called peaks by using GC-content and mappability with MOSAiCS, and for the ChIP-Seq datasets we used their respective Input samples.

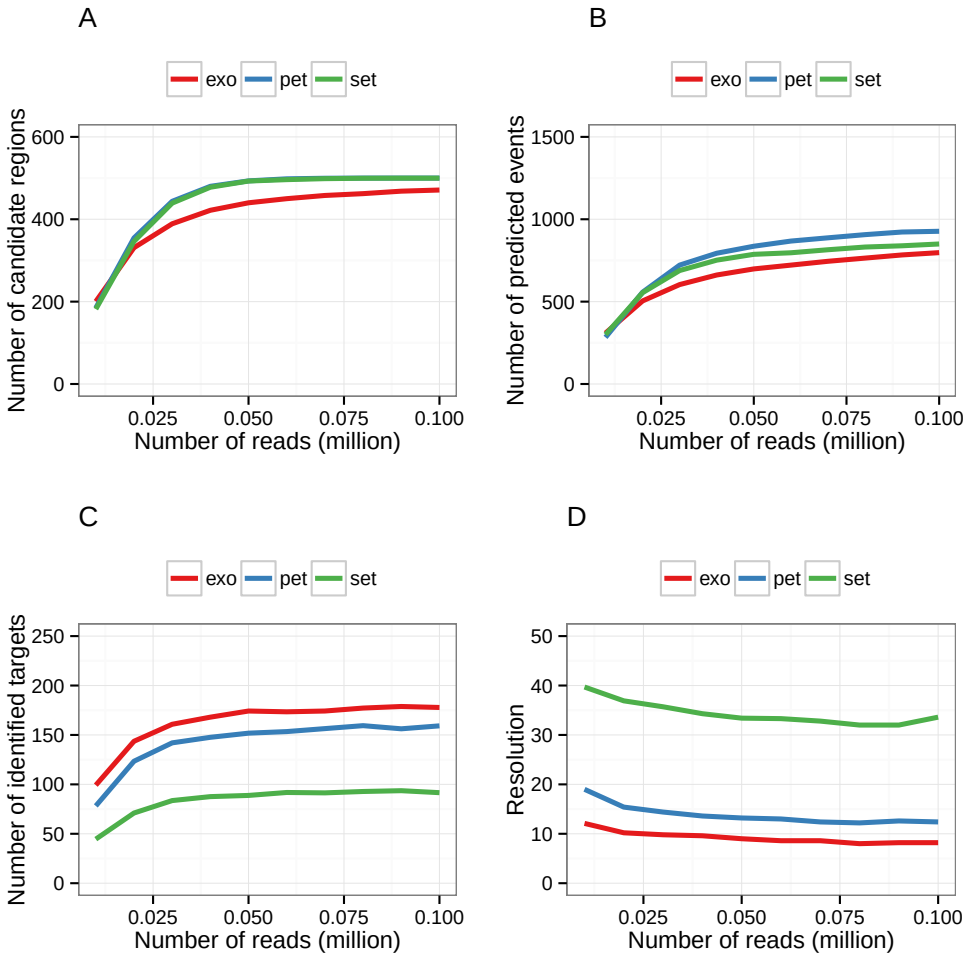


Figure 10: Comparison of the number of candidate regions (A), predicted events (B), identified targets (C) and resolution (D) among ChIP-exo, PET ChIP-Seq and SET ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard was considered identified if a BS was estimated inside a 15 bp window around it

Figure 10 shows the behavior of each data type 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 PET and SET ChIP-Seq in the 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 indicators do so as well, which may indicate that with ChIP-exo a smaller amount of reads is necessary to identify a higher number of targets, but it

may also be possible that this is an artifact occurring due to ChIP-exo's lower library complexity.

4 Software packages

For this project several R packages were created or updated:

- **dPeak**: We updated the initialization strategy. The latest version is currently available from http://dongjunchung.github.io/dpeak/.
- **ChIPexoQual**: This package contains the QC pipeline for ChIP-exo. The last version is available in https://github.com/welch16/ChIPexoQual.
- **Segvis**: The goal of this package is to visualize genomic regions by using aligned reads such as in figures 1 or 5. The latest version is available in https://github.com/keleslab/Segvis.
- **ChIPUtils**: This package attempts to gather the most commonly used ChIP-Seq QC (as in table 1 or figure 3) indicators and additional utilities (as figure 4A) for ChIP-Seq. The latest available version is in https://github.com/welch16/ChIPUtils.

5 Conclusions

We provide a ChIP-exo QC pipeline capable of assess the balance between enriched samples and low complexity regions. It is shown that the "peak-pair" assumption doesn't hold in practice and we provide two out-of-the-box visualization capable to assess the strand imbalance in a ChIP-exo experiment.

We updated the dPeak algorithm from [4] and showed that ChIP-exo is comparable in resolution and sensitivity with PET ChIP-Seq and outperforms SET ChIP-Seq. ChIP-exo compared with PET and SET ChIP-Seq at fixed depth sample is able to identify more targets at a lower resolution. We compared dPeak with another algorithms to estimate binding locations in ChIPexo data, dPeak is comparable to MACE and outperforms GEM in resolution. dPeak provided a striking balance in sensitivity, specificity and spatial resolution for ChIP-exo analysis.

6 Planned work

6.1 Some additional thoughts for this analysis

6.1.1 GC content and mappability quality indicators

Figure 4 shows a relationship between the ChIP tag counts of a ChIP-exo experiment and both mappability and GC content scores. Furthermore, in sections 3.3 and 3.2 we called peaks by using MOSAiCS which considers them as necessary information when there is not input sample. Hence, quality measures we may consider to add indicators to the QC pipeline that considers this biases.

6.1.2 Quality control for ChIP-Nexus

He et al., 2014 [11] proposed a modification to the ChIP-exo protocol that may fix its low library complexity issues by extending the fragments by a random barcode prior to the exonuclease digestion. This last statement is briefly mentioned by Mahony and Pugh, 2015 [19]. We want to apply our QC pipeline to ChIP-Nexus. Depending if ChIP-Nexus data shows background we may need to modify the QC pipeline in the step where it paritions the reads into regions, but the local-NSC may be used to prove this statement.

6.2 Analysis of E. Coli Transcription Initiation Complexes

The data showed in figure 2 is actually part of a more interesting experiment. We have 2 biological replicates of the σ^{70}, β'_f and β transcription factors under two conditions, one where rifampicin was applied by 20 minutes and another where it wasn't applied. As seen in the figure, there is ChIP-exo data for the 3 factors and PET ChIP-Seq for σ^{70} and β'_f (in the figure both replicates were pooled).

σ^{70} factor is a transcription initiation factor of housekeeping genes in E. Coli, and both β factors transcribe the DNA into mRNA. Hence, we want to use this data to have a better understanding of the transcription in E. Coli.

The steps in formation of productive transcription complexes are:

1. Binding of RNAP to rpomoter DNA to form a closed complex.
2. Rearrangement of RNAP-DNA contacts to form an open complex.
3. Binding of nucleoside triphosphates in the active site and synthesis of a initial RNA to yield an initial transcription complex (ITC) - this is sometimes called a "scrunched complex".

4. Promoter escape to form a elongated complex (EC_1) when the promoter contacts are broken, and the σ^{70} factor is realeased.

5. Conversion of EC_1 into EC_2 by loading of the ribosome and elongation factors to enable efficient transcript elongation.

As a proof of concept we were able to identify open and closed regions by centering the binding the peaks respecto to the highest σ^{70} summit (among the rif0 and rif20 samples), and then using hierachical clustering on the β'_f signal of the rif0 samples to produce:

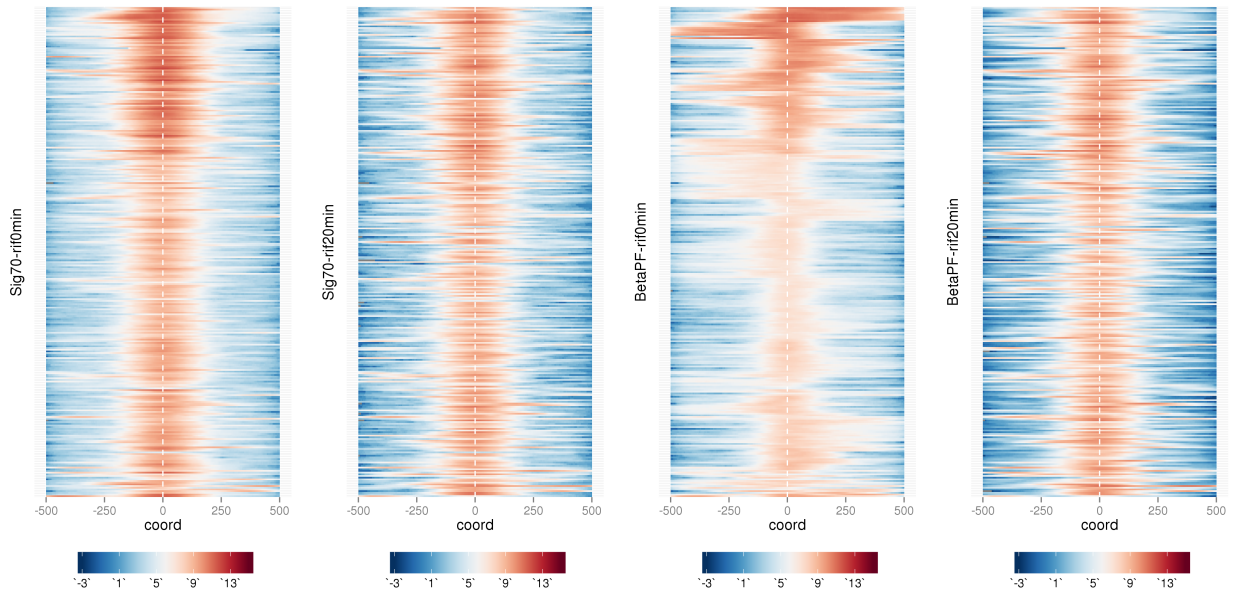


Figure 11: Centered and normalized heatmaps for PET ChIP-Seq peaks of σ^{70} and β'_f under rif0 and rif20. The regions are organized by the centroids clustering of the β'_f-rif0 profiles.

Figure 11 shows heatmaps of the normalized coverages of PET ChIP-Seq peaks for β'_f and $\sigma^{70}70$ under rif0 and rif20 treatments. Each row represents the same region, under the four different samples. As expected the signal for both σ^{70} samples is centered around the middle, for β'_f-rif20, the majority of he signal seems to be centered, hence the trascription was interrupted and for β'_f-rif0 we can see clusters formed by open regions.

Then, using a set of high quality PET ChIP-Seq peaks and clustering with the mclust model by Fraley and Raftery, 2007 [5] we were able to label the a set of regions as closed or open com-

plexes. Hence, we expect that by using ChIP-exo data we may find different patterns that represent transcription units whose output was limited at differemt steps in the formation of EC_2.

Additionally if we find several cases of those patterns, then we may be able to determine which sequences may be causing this behaviour at the different steps.

6.3 Enhancer prediction and active learning

Last summer, we entered to a contest from ENCODE's functional characterization group to fit a model to classify enhancers. Roughly speaks, the data set consided of a set of regions $\mathcal{R}_1, \cdots, \mathcal{R}_n$ and a set of labels $Y_1, \cdots, Y_n \in \{0, 1, \}$. Hence, we built a series of features $X_1, \cdots, X_p$ and fitted a model by dividing the n observations into a training set and a test set, etc. However, there are in total another $m >> n$ regions in the experiment, and the cost of labelling those m regions may be high (in either the actual monetary cost of the experiment or the time the scientists spends labelling those regions). Those $n + m$ regions are H3K27ac peaks, hence they were ranked and the first n were selected randomly from groups with high, medium and low signal. This raises the following questions:

- Is there a better way of ranking this $n + m$ regions ?
- What is the best way to select the regions that are going to be labelled, i.e. we would want to label the regions that gives us the best model.

Hence, it may be worth to investigate which active learning strategies may be the best to apply in this particular problem. Settles, 2012 [27] is a survey of acive learning techniques that we want to explore in order to solve this problems.

References

[1] Anaïs F. Bardet, Jonas Steinmann, Sangeeta Bafna, Juergen A. Knoblich, Julia Zeitlinger, and Alexander Stark. Identification of transcription factor binding sites from chip-seq data at high resolution. *Bioinformatics*, 2013.

[2] Benjamin Bolstad, Rafael Irizarry, Magnus Âstrand, and Terence Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003.

[3] Thomas Carroll, Ziwei Liang, Rafik Salama, Rory Stark, and Ines de Santiago. Impact of artifact removal on chip quality metrics in chip-seq and chip-exo data. *Frontiers in Genetics, Bioinformatics and Computational Biology*, 2014.

[4] Dongjun Chung, Dan Park, Kevin Myers, Jeffrey Grass, Patricia Kiley, Robert Landick, and Sündüz Keleş. dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. *PlOS, Computational Biology*, 2013.

[5] Chris Fraley and Adrian Raftery. Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, 2007.

[6] Terrence S. Furey. Chip-seq and beyond: new and improved methodologies to detect and characterize protein-dna interactions. *Nature Reviews: Genetics*, 2012.

[7] Yuchun Guo, Shaun Mahony, and David K. Gifford. High resolution genome wide binding event finding and motif discovery reals transcription factor spatial bindings constraints. *PlOS, Computational Biology*, 2012.

[8] Yuchun Guo, Georgios Papachristoudis, Robert C. Altshuler, Georg K. Gerber, TOmmi S. Jaakkola, David K. Gifford, and Shaun Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 2010.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference and Prediction.* Springer, New York, 2009.

[10] Shanil Haugen, Wilma Ross, and Richard Gourse. Advances in bacterial promoter recognition and its control by factors that do not bind dna. *Nature Reviews Microbiology.*

[11] Qiye He, Jeff Johnston, and Julia Zeitlinger. Chip-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 2014.

[12] Rafael Irizarry, Bridget Hobbs, Francois Collin, Yasmin Beazer-Barclay, Kristen Antonellis, Uwe Scherf, and Terence Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003.

[13] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S. Johnson, Richard M. Myers, and Wing H. Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature biotechnology*, 2008.

[14] Peter Kharchenko, Michael Tolstorukov, and Peter Park. Design and analysis of chip-seq experiments for dna-binding proteins, 2008.

[15] Pei Fei Kuan, Dongjun Chung, Guangjin Pan, James A. Thomson, Ron Stewart, and Sündüz Keleş. A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 2009.

[16] Stephen Landt, Georgi Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley Bernstein, Peter Bickel, James Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Catherine Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander Hartemink, Michael Hoffman, Vishwanath Iyer, Youngsook Jung, Subhradip Karmakar, Manolis Kellis, Peter Kharchenko, Qunhua Li, Tao Liu, Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard Myers, Peter Park, Michael Pazin, Marc Perry, Debasish Raha, Timothy Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John Stamatoyannopoulos, Michael Tolstorukov, Kevin White, Simon Xi, Peggy Farnham, Jason Lieb, Barbara Wold, and Michael Snyder. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome Research*, 2012.

[17] Desmond S. Lun, Ashley Sherrid, Brian Weined, David R. Sherman, and James E. Galagan. A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from chip-seq data. *Genome Biology*, 2009.

[18] Pedro Madrigal. *EMBnet.journal*, 2015.

[19] Shaun Mahony and Pugh B. Franklin. Protein-dna binding in high-resolution. *Critical Reviews in Biochemistry and Molecular Biology*, 2015.

[20] Eric M. Mendenhall and Bradley E. Bernstein. Dna-protein interactions in high definition. *Genome Biology*, 2012.

[21] Shirley Pepke, Barbara Wold, and Mortazavi Ali. Computation for chip-seq nad rna-seq studies. *Nature*, 2009.

[22] The ENCODE project consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 2011.

[23] Ho Sung Rhee and Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 2011.

[24] Ho Sung Rhee and Franklin Pugh. Chip-exo a method to identify genomic location of dna-binding proteins at near single nucleotide accuracy. *Current Protocols in Molecular Biology*, 2012.

[25] Ho Sung Rhee and Franklin Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 2012.

[26] Aurelien Serandour, Brown Gordon, Joshua Cohen, and Jason Carroll. Development of and illumina-based chip-exonuclease method provides insight into foxa1-dna binding properties. *Genome Biology*, 2013.

[27] Burr Settles. *Active Learning*. Morgan and Claypool Publishers, 2012.

[28] Liguo Wang, Junshend Chen, Chen Wang, Liis Uusküla-Reimand, Kaifu Chen, Alejandra Medina-Rivera, Edwin J. Young, Michael T. Zimmermann, Huihuang Yan, Zhifu Sun, Yuji Zhang, Stephen T. Wu, Haojie Huang, Michael D. Wilson, Jean-Pierre A. Kocher, and Wei Li. Mace: model based analysis of chip-exo. *Nucleic Acids Research*, 2014.

[29] Larry Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2010.

[30] Elizabeth Wilbanks and Marc Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PlOS One*, 2012.

[31] Xuekui Zhang, Gordon Robertson, Martin Krzwinski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo. Pics: Probabilistic inference for chip-seq. *Biometrics*, 2010.

[32] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérˆome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nausbam, Richard M. Myers, Myles Brown, Wei Li, and X. Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biology*, 2008.

Supplement

Additional Enrichment plots for σ^{70}

In figure 12 it is shown the same relationship as in figure 7, but considering only regions formed where the reads are being allocated in more than 10 (A) and 30 (B) unique positions respectively. Both plots show that the vertical arms formed by regions with low ARC is formed by low complexity regions, that way suggesting that this segment correspond to the background of a ChIP-exo experiment.

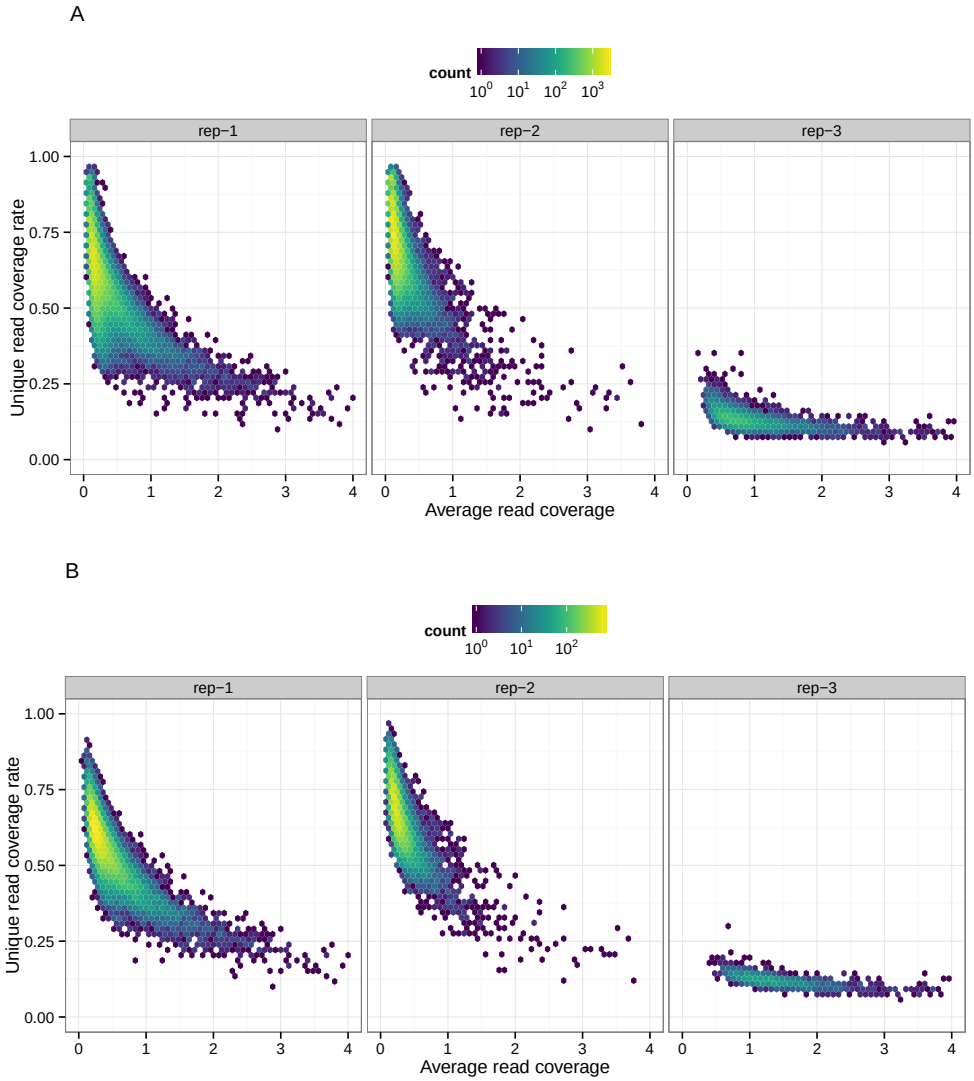


Figure 12: Hexbin plots of ARC vs URCR for each region after partitioning the genome. In A and B the regions with reads mapped to at most 10 and 30 positions respectively where not considered. ARC is defined as the ratio of the nr. of reads and the width of a region and URCR is the ratio of the number of unique position where the reads are being allocated and the number of reads in a region.