

# 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 Kendzierski**, Department of Biostatistics and Medical Informatics

**Professor Ming Yuan**, Department of Statistics

# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>ChIP-exo and ChIP-Seq review</b>	<b>7</b>
2.1	Current Quality Control Measures for ChIP-Seq . . . . .	8
2.1.1	Strand cross-correlation . . . . .	10
2.2	Comparison with ChIP-Seq data . . . . .	11
<b>3</b>	<b>Results</b>	<b>13</b>
3.0.1	Comparison with ChIP-Seq data using dPeak . . . . .	13
3.1	Recommendations for the design of ChIP-exo experiments . . . . .	13
<b>4</b>	<b>Conclusions</b>	<b>16</b>
<b>5</b>	<b>Planned work</b>	<b>17</b>

## List of Figures

- 1 ChIP-exo diagram. The only difference with ChIP-Seq is the second step, where an exonuclease enzyme is added and it digests the DNA fragments starting from the 5' end until it finds a crosslink protein (this figure is taken from Furey, 2012 [4]). . . . . 7
- 2 An example of a PET ChIP-Seq peak compared to ChIP-exo.  $\sigma^{70}$ ,  $\beta$  and  $\beta'_f$  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  $\sigma^{70}$  summit. The library complexity is lower in ChIP-exo than in ChIP-Seq but the resolution is increased, this can be observed at the  $\sigma^{70}$  peak that for ChIP-exo was created as the combination of at least three ChIP-exo “sub-peaks”. . . . . 8
- 3 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. . . . . 11
- 4 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)  $\delta$  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)  $\sigma$  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. . . . 14
- 5 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 . . . 15

## 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 [5]), which makes a striking balance in sensitivity, specificity and spatial resolution for ChIP-exo data analysis.

# 1 Introduction

ChIP-exo (Chromatin Immunoprecipitation followed by exonuclease digestion and next generation sequencing) Rhee and Pugh, 2011 ([18]) 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 [18]; ER factor in human and FoxA1 factor in mouse (Serandour et al., 2013 [20]); and generated  $\sigma^{70}$  factor in *Escherichia coli* (E. Coli) under aerobic (+O<sub>2</sub>) 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 [15]). Hence, most of current QC guidelines (Landt et al., 2012 [12]) 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 ([18], [19], [20]); Carroll et al., 2014 [3] studied the use of the Strand-Cross Correlation (SCC) (Kharchenko et al., 2008 [10]) 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 coefficient.

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 [18] 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 [21]), CexoR (Madrigal, 2015 [14]) 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 [22] and Pepke and Wold, 2009 [17]), such as MACS (Zhang et al., 2008 [24]), CisGenome (Hongkai et al., 2008 [9]) and MOSAiCS (Kuan et al., 2009 [11]), 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 [13]), GEM (Guo et al., 2012 [6], an improved version of Guo et al., 2010 [7] ) and PICS (Zhang et al., 2010 [23]) 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 dPeak (Chung et al., 2013 [5]), 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 [18] 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

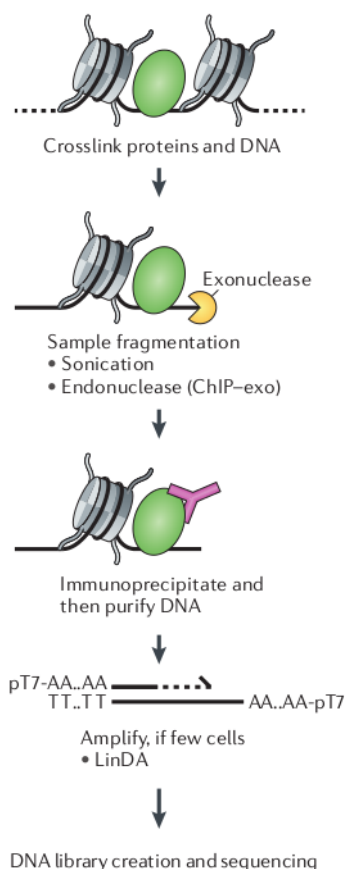


Figure 1: ChIP-exo diagram. The only difference with ChIP-Seq is the second step, where an exonuclease enzyme is added and it digests the DNA fragments starting from the 5' end until it finds a crosslink protein (this figure is taken from Furey, 2012 [4]).

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 closer 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 [18], 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 maintained 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  $\sigma^{70}$  and  $\beta'_f$  factors (and the  $\beta$  factor only on top) on a E. Coli genome's region: First, both regions show that the  $\sigma^{70}$  sample is enriched; second the  $\sigma^{70}$  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  $\beta$  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.

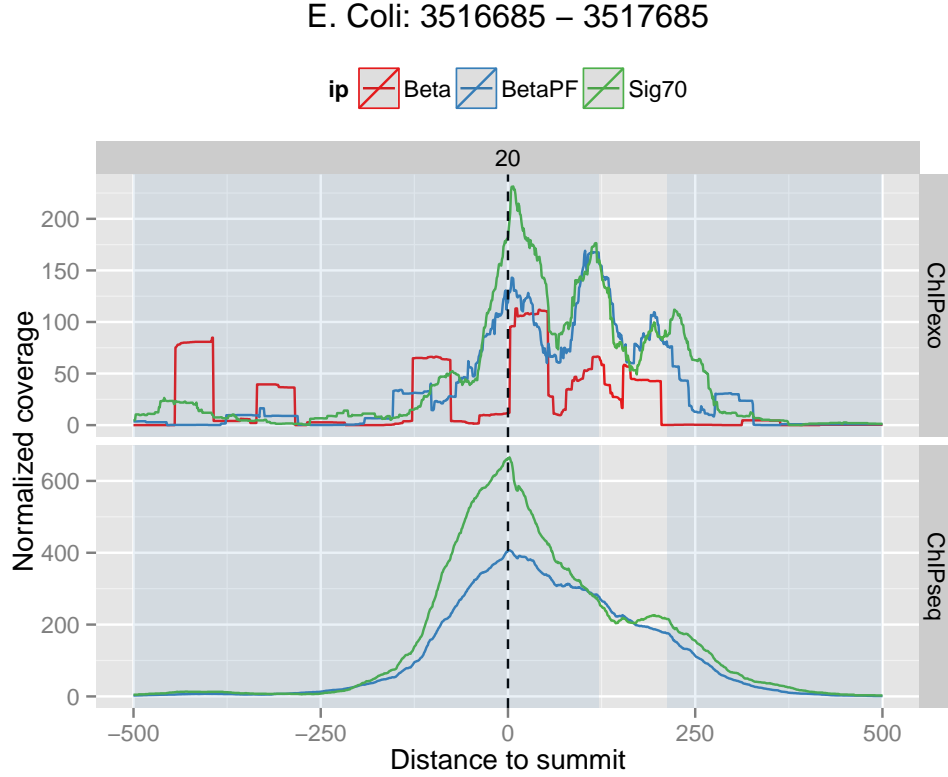


Figure 2: An example of a PET ChIP-Seq peak compared to ChIP-exo.  $\sigma^{70}$ ,  $\beta$  and  $\beta'_f$  are subunits of the transcription initiation complex in E. Coli. Gene regions on the reverse strand are highlighted in light blue and the region was centered respect to ChIP-Seq  $\sigma^{70}$  summit. The library complexity is lower in ChIP-exo than in ChIP-Seq but the resolution is increased, this can be observed at the  $\sigma^{70}$  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: Usual quality control indicators applied to the  $\sigma^{70}$  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 omitted since input sample usually don't exists for ChIP-exo).  $\sigma^{70}$  samples were given by Robert Landick, FoxA1 and ER samples are from Serandour et al., 2013 [20] and CTCF sample is from Rhee and Pugh, 2011 [18].

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

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 mapping 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:

$$SSD = SD_N / \sqrt{N}$$

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 diggested 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 [10] and it may be one of the most used of the ChIP-Seq QC metrics. The SCC curve is defined as:

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

where  $y(\delta)$  is strand cross-correlation for a strand shift  $\delta$ ,  $r$  is the Pearson correlation,  $w_c$  is the proportion of tags mapped to chromosome  $c$  and  $n_c^S$  is the tag count vector for strand  $S$  and chromosome  $c$ .

## 2.2 Comparison with ChIP-Seq data

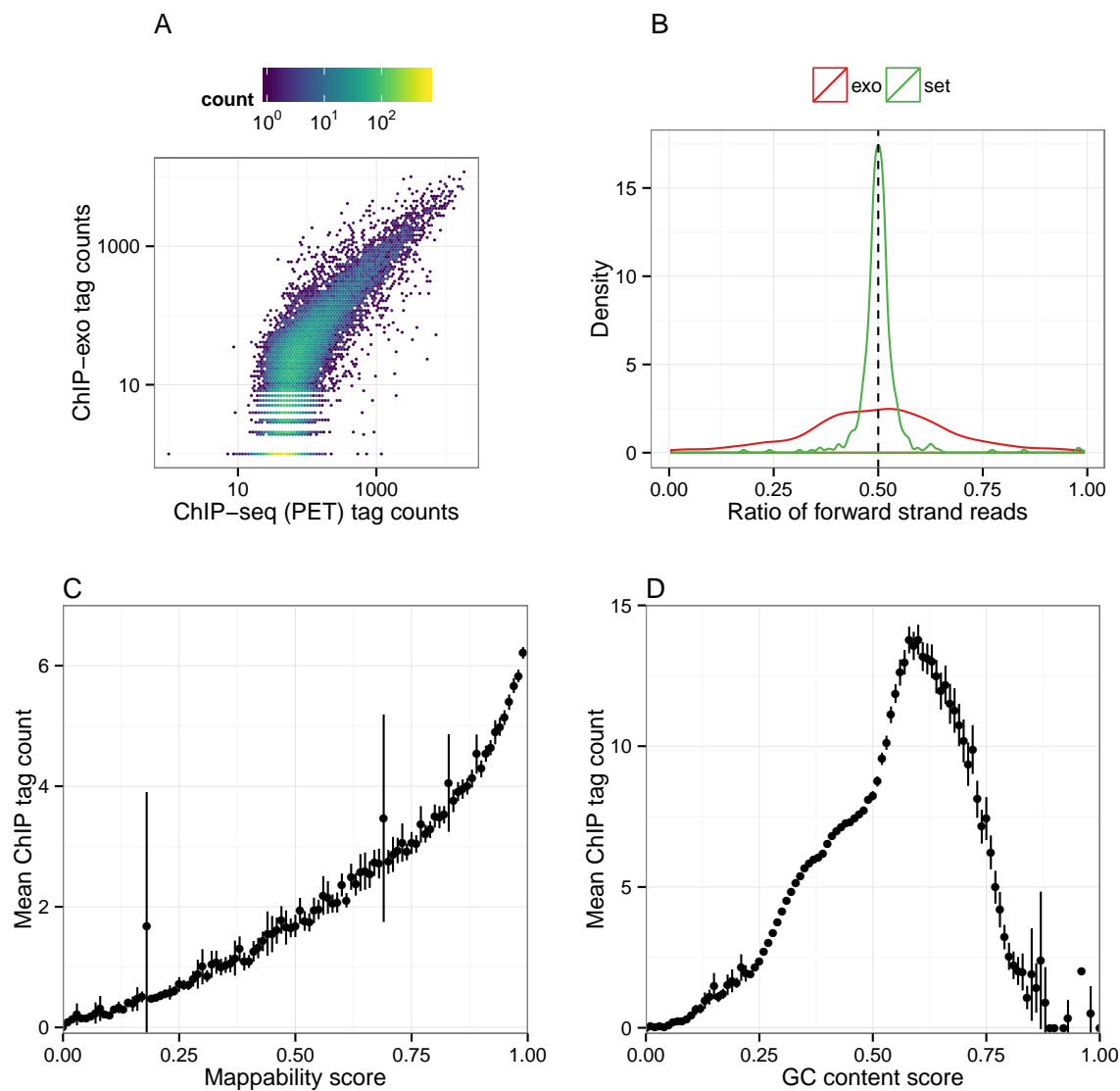


Figure 3: A) ChIP signal in ChIP-exo was linearly related to that in PET ChIP-Seq data in the region with high ChIP tag counts. In contrast, there were clear differences in their background distribution, where several ChIP-exo background regions were empty. B) In ChIP-exo data, strands of reads were 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 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 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. In ChIP-Seq data reads were almost uniformly distributed over background (non-binding) regions and the ChIP tag counts in these regions were significantly larger than zero. In contrast, in ChIP-exo data, there was larger variation in ChIP tag counts among background regions and ChIP tag counts were much lower in these regions compared to ChIP-Seq data. There were also large proportion of regions without any read in ChIP-exo data. These results indicate that background distribution of ChIP-exo data is less homogeneous than that of ChIP-Seq data.

We next evaluated the “peak pair” assumption from Rhee and Pugh, 2011 [18], 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 [21], Madrigal 2015 [14] 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  $\sigma^{70}$  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. 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 [18] 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. 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

#### 3.0.1 Comparison with ChIP-Seq data using dPeak

Figure 4 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 4A 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 4B 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 4C and 4D, we observe the behavior of dPeak’s estimated parameters for single end data (both ChIP-exo and ChIP-Seq). The  $\delta$  parameter measures the distance from the 5’ end of the fragment to its respective binding event and the  $\sigma$  parameter indicates the fragments distribution variability around their respective binding sites (more information in [5]). 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.

#### 3.1 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  $\sigma^{70}$  sample in aerobic conditions. For each sampled dataset we applied our lower-to-higher resolution pipeline by calling peaks with MOSAiCS [11] and then deconvolving the binding events by using dPeak [5]. 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 5 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

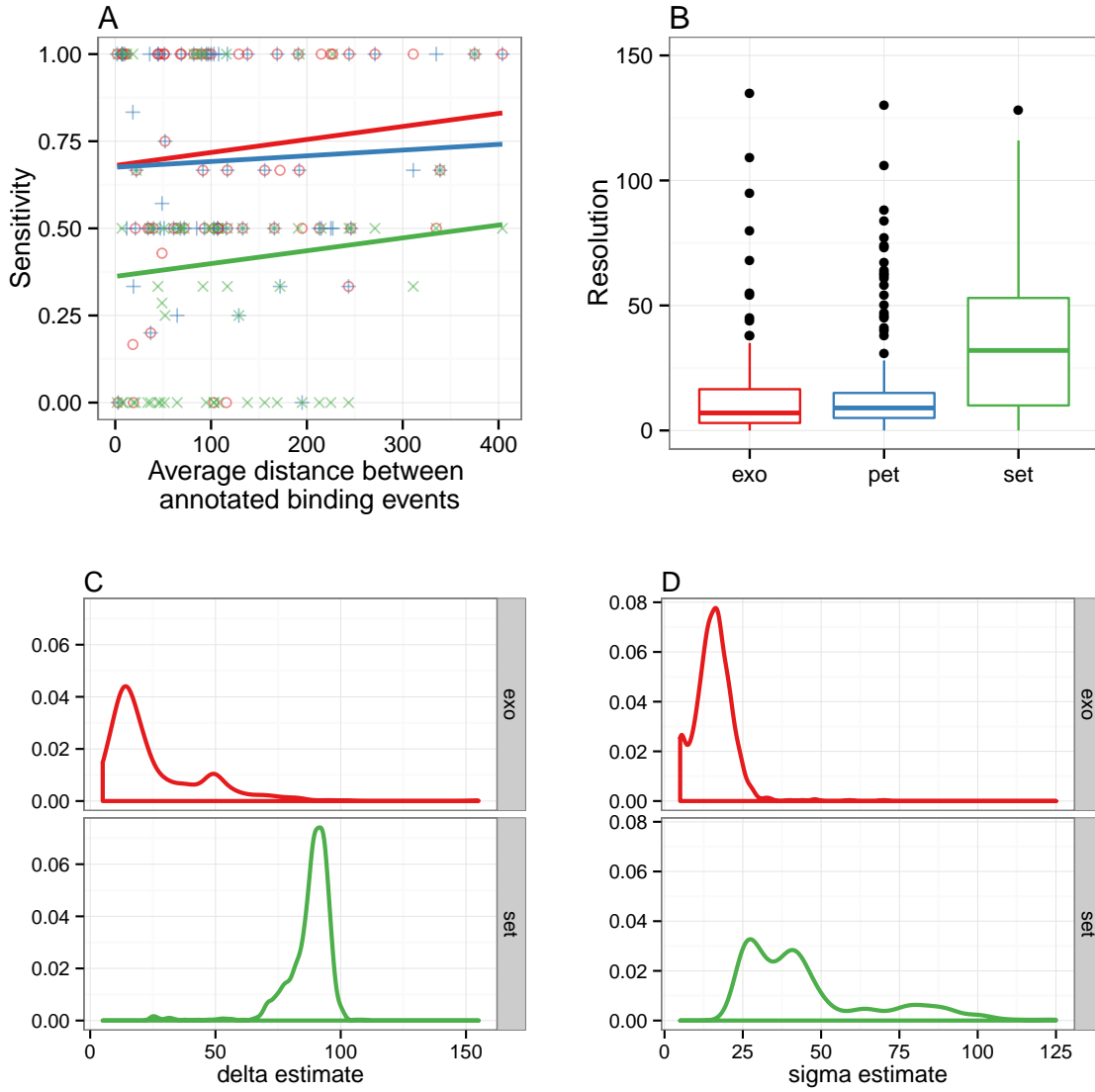


Figure 4: Comparison of (A) sensitivity and (B) resolution between ChIP-exo and ChIP-Seq data. Sensitivity is defined as the proportion of RegulonDB annotations identified using each data. Resolution is defined as the distance between RegulonDB annotation and its closest prediction. C)  $\delta$  parameter 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)  $\sigma$  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.

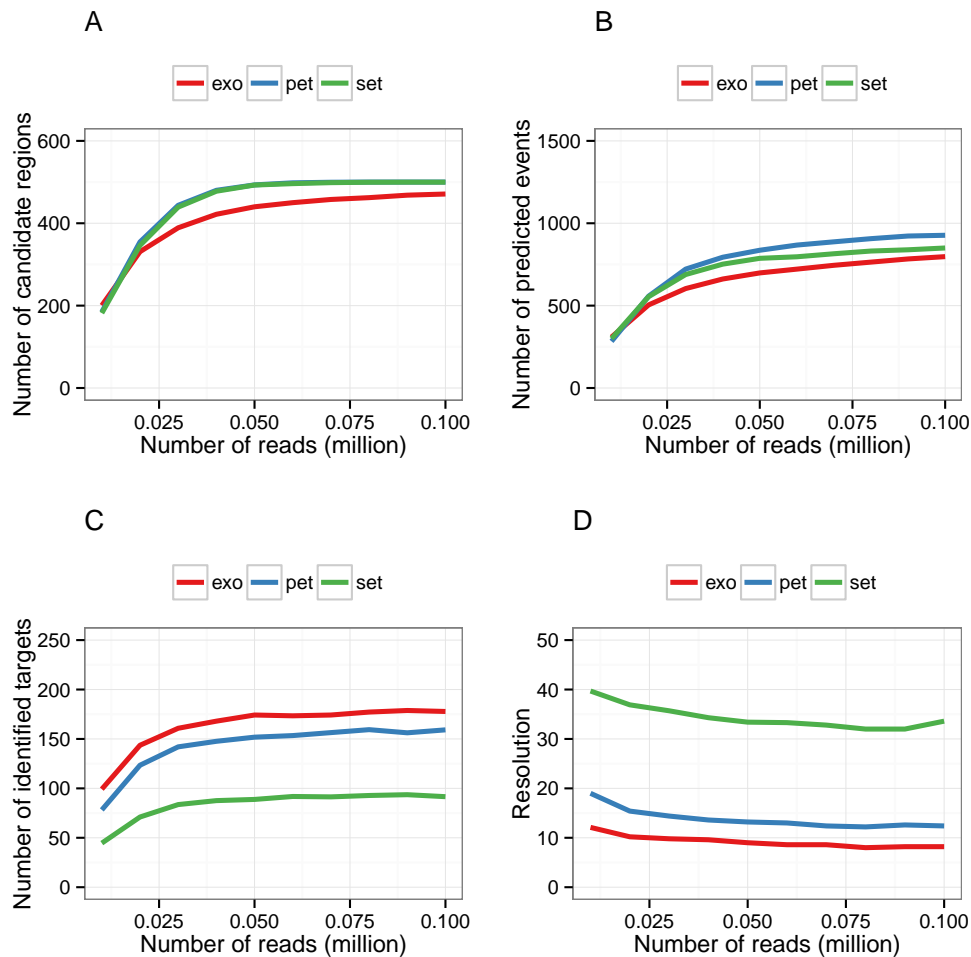


Figure 5: Comparison of the number of candidate regions (A), predicted events (B), identified targets (C) and resolution (D) among ChIP-exo, PET ChIP-Seq and SET ChIP-Seq. RegulonDB annotations are considered as a gold standard. A gold standard was considered identified if a BS was estimated inside a 15 bp window around it

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 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 [5] 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.



## 5 Planned work

## 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 Åstrand, 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] ChIP-seq, beyond: new, improved methodologies to detect, and characterize protein DNA interactions. Furey, terrence s. *Nature Reviews: Genetics*, 2012.
- [5] 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.
- [6] Yuchun Guo, Shaun Mahony, and David K. Gifford. High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial bindings constraints. *PLoS, Computational Biology*, 2012.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] Peter Kharchenko, Michael Tolstorukov, and Peter Park. Design and analysis of chip-seq experiments for dna-binding proteins, 2008.

- [11] 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.
- [12] 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 Shores, Arend Sidow, Matthew Slaterry, John Stamatoyannopoulos, Michael Tolstorukov, Kevin White, Simon Xi, Peggy Farnham, Jason Lieb, Barbara Wold, and Michael Snyder. *Genome Research*, 2012.
- [13] 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.
- [14] Pedro Madrigal. *EMBNet.journal*, 2015.
- [15] Shaun Mahony and Pugh B. Franklin. Protein-dna binding in high-resolution. *Critical Reviews in Biochemistry and Molecular Biology*, 2015.
- [16] Eric M. Mendenhall and Bradley E. Bernstein. Dna-protein interactions in high definition. *Genome Biology*, 2012.
- [17] Shirley Pepke, Barbara Wold, and Mortazavi Ali. Computation for chip-seq nad rna-seq studies. *Nature*, 2009.
- [18] Ho Sung Rhee and Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 2011.
- [19] 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.

- [20] 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.
- [21] Ligu 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.
- [22] Elizabeth Wilbanks and Marc Facciotti. Evaluation of algorithm performance in chip-seq peak detection. *PLoS One*, 2012.
- [23] Xuekui Zhang, Gordon Robertson, Martin Krzwiniski, Kaida Ning, Arnaud Droit, Steven Jones, and Raphael Gottardo. Pics: Probabilistic inference for chip-seq. *Biometrics*, 2010.
- [24] Yong Zhang, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoutte, 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.