A quality control and analysis pipeline for ChIP-exo data

February 2015

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 for data from ChIP-exo experiments. This automated pipeline evaluates the overall signal to noise level of a given experiment and investigates ChIP-exo data for artifacts such as (i) strand imbalance where one strand is digested more than the other; (ii) enzyme over-digestion where the 5 ends of the forward strand reads are located after the 5 ends of the reverse strand reads; and (iii) PCR amplification bias the reads of a candidate binding event are concentrated on an extremely small numbers of positions. Assessment of these biases and artifacts are facilitated through diagnostic plots and summary statistics that compare a large portion of the genome as partitioned into islands with regions of high depth. We also systematically compare multiple ChIP-seq and ChIP-exo datasets to quantify differences between these two protocols. Our analysis indicate that spatial resolution of ChIP-exo is comparable to that of paired-end ChIP-seq and both of them are significantly better than resolution of single-end ChIP-seq. Furthermore, for a given fixed sequencing depth, ChIP-exo provides higher sensitivity, specificity, and spatial resolution than paired-end ChIP-seq.

Contents

List of Figures

1 Overview of ChIP-exo

ChIP-exo is a new techonology which is claimmed to be more precise than ChIP-seq (both SET and PET) to detect the location of protein-DNA interactions.

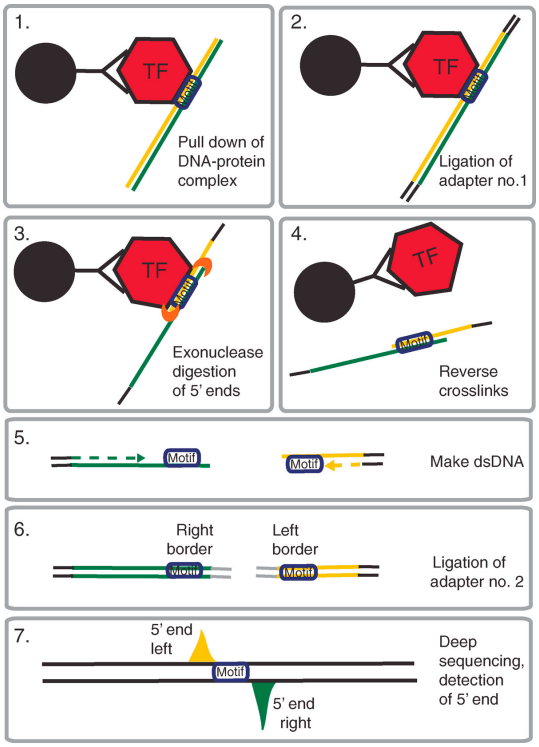


Figure 1: ChIP-exo protocol, the figure is from [1]

If we consider the coordinate system as being ordered from 5' to 3' then, we can think the diagram as:

1. The motif is ocurring at position μ
2. The # 1 adaptors are being added at positions $\mu-\delta_2$ and $\mu+\delta_2$
3. The exonuclease digestion, digest the DNA starting at the 5' ends, and digest until it reachs the positions $\mu-\delta_1$ and $\mu+\delta_1$

4 - 6. This steps separate the transcription factor from the DNA, make the DNA double stranded again and add the #2 adaptors. This adaptors are at the positions $\mu\pm\delta_1$ respectively. From this step we can see that there are two types of reads being sequenced, one for each adaptor.

7. To build this plot, the 5' ends of both sides are sequenced. Thus the sequenced reads are going to occur approximately around μ. The forward strand is going to be in the interval $(\mu-\delta_2, \mu+\delta_1)$ and the backward strand is going to be in $(\mu-\delta_1, \mu+\delta_2)$

Thus, the data will consist of two sets of aligned reads, one for each adaptor. So far, we have analized the one that corresponds to the second adaptor, since is the one related to the digested ends.

A brief remark is that in a SET ChIP-seq experiment, the exonuclease digestion step is ommited, therefore the sequenced reads are going to have limits $\mu\pm\delta_2$. In other words, the fragment length is equal to $2\delta_2$.

for the poster, it would be nice to make a pictorial description of the second column

2 ENCODE's quality metrics on ChIP-exo datasets

the point here is that classical ChIP-seq quality metrics may not work with ChIP-exo datasets

We are considering a diverse collection of datasets, where the genomes range among organisms such as human, mouse and e.coli. To review the current ENCODE metrics efficacy to asses the quality of a dataset, we calculated some of the current metrics:

dataset	depth	pbc	nsc
Sig70-exp-rep1	1,454,566	0.216	12.386
Sig70-exp-rep2	864,714	0.247	12.592
Sig70-stat-rep1	1,584,532	0.259	14.700
Sig70-stat-rep2	1,012,936	0.247	14.499
SigmaS-exp-rep1	1,593,964	0.294	5.568
SigmaS-exp-rep2	3,405,118	0.177	7.251
SigmaS-stat-rep1	1,822,585	0.232	5.479
SigmaS-stat-rep2	9,898,733	0.163	8.156

dataset	depth	pbc	nsc
BetaPf-rif0-rep1	1,875,127	0.251	1.391
BetaPf-rif0-rep2	898,641	0.302	1.297
BetaPf-rif20-rep1	4,900,071	0.234	6.233
BetaPf-rif20-rep2	6,550,805	0.215	5.766
Beta-rif0-rep1	3,909,669	0.283	1.379
Beta-rif0-rep2	5,157,768	0.256	1.403
Beta-rif20-rep1	5,153,689	0.338	2.318
Beta-rif20-rep2	1,509,554	0.498	4.550
Sig70-rif0-rep1	960,256	0.282	4.307
Sig70-rif0-rep2	2,247,295	0.266	4.802
Sig70-rif20-rep1	1,940,387	0.270	6.129
Sig70-rif20-rep2	4,229,574	0.215	6.392

dataset	depth	pbc	nsc
FoxA1-rep1	22,210,461	0.656	4.337
FoxA1-rep2	23,307,557	0.800	5.227
FoxA1-rep3	22,421,729	0.107	6.000

dataset	depth	pbc	nsc
ER-rep1	9,289,835	0.808	3.783
ER-rep2	11,041,833	0.802	3.749
ER-rep3	12,464,836	0.820	3.992

dataset	depth	pbc	nsc
H3k27ac	29,599,796	0.305	1.332
H3k4me1-rep1	28,794,319	0.258	1.127
H3k4me1-rep2	31,818,368	0.252	1.131

For the ChIP-exo datasets, some of ENCODE's current quality metrics were evaluated showing that current metrics aren't adequate for ChIP-exo datasets. The metrics evaluated were:

- PCR bottleneck coefficient
- Strand cross-correlation
- Normalized strand cross-correlation

2.1 PCR bottleneck coefficient

ENCODE's sugest the following classification of the PBC:

PBC range	Bottleneck
0 - 0.5	Severe
0.5 - 0.8	Moderate
0.8 - 0.9	Mild
0.9 - 1	Non-existant

In general, the calculated PBC values for ChIP-exo data sets are very low due to the nature of the experiment: An exonuclease enzyme is applied to the inmunoprecitate, and as a result if will trim the 5' end of each DNA fragment. After being aligned, several fragments are going to be mapped to the same 5' starting positions, resulting in an effect that may be confounded with PCR artifacts.

Right now we have this plot for Landick's data sets, but I am considering in make a table with columns: dataset, organism, depth, pbc, that reflect the same information as the figure

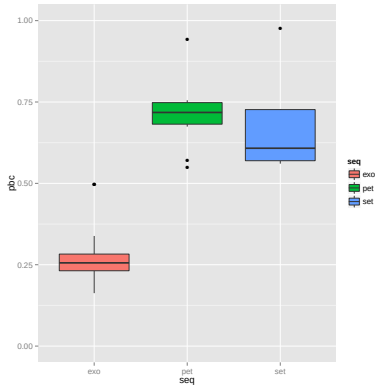


Figure 2: PBC comparison among protocols

2.2 Strand cross-correlation

The strand cross-correlation is calculated by shifting both strands and calculating the correlation between both coverages. For ChIP-seq data, it is shown that the fragment length of the reads can be estimated by the shift where the cross correlation curve is maximized, since the forward and reverse reads might construct peak pais around the protein it is often assumed that both peaks present the same height.

The normalized strand cross-correlation was calculated as the ratio between the maximal and minimal values of the cross-correlation function. Higher values indicates more enrichment, while the minimum possible value is 1.

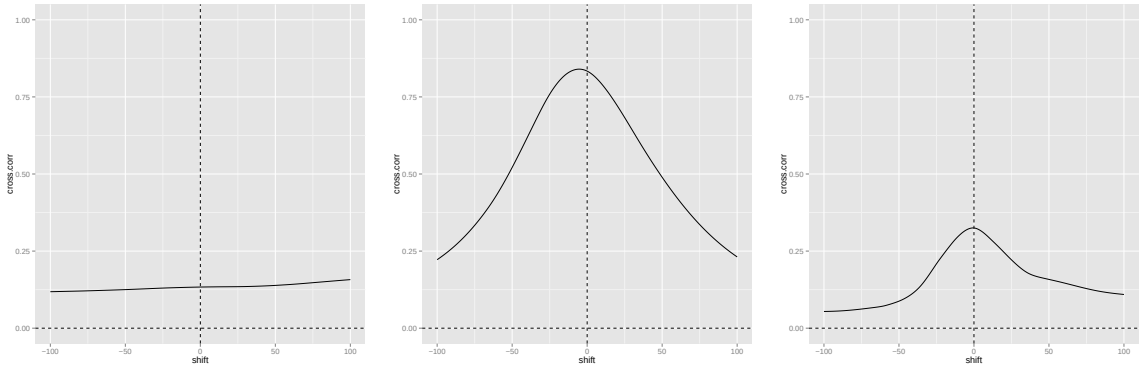


Figure 3: Different examples of ChIP-exo cross-correlations

this is a place holder, until I plot all cross-corr. together by group

After calculating this curves we can see:

- The cc function was calculated for the $[-S, S]$ range where $S = 100$ is the maximum shift . In the extremes this function is still decreasing, thus if a larger S is choosen it is possible to for the cc-function to show non-positive values.
- The maximum shift is obtained at a non-positive shift; therefore the fragment length can't be estimated as the shift where the cc curve is maximized
- Since several reads are digested, the minimal values of the cc-function may be very small and therefore the nsc very high. This may be misleading as a quality indicator.
- Another interpretation may arise from ENCODE's description since it is mentioned that this score is sensitive to technical effects, and in the ChIP-exo case it is not well studied the exonuclease enzyme digestion.
- The nsc calculated values are shown in table 2

use a longer (more typical) range to show how this measure would behave under normal circumstances

have chip-seq vs chip-exo data sets for, perhaps calculate the cross-correlation curves (and measures) for the ChIP-seq data sets.

1. *Ren's histone datasets*
2. *Carroll's human ER datasets*

3. Landick chip-seq set sampled data sets

3 Methodology description

Using *IRanges*, the coverage of the genome is calculated and partitioned into a set of regions. For each region, the following statistics were calculated:

- Forward strand ratio
- Depth
- Depth/width
- Number of unique positions (npos)
- npos/depth (Ratio between number of unique positions in region and depth)

An alternative was to partition the genome into fixed length bins and calculate the same statistics.

In general several plots were designed in order to asses a datasets quality.

3.1 Strand imbalance

A unique challenge in ChIP-exo data is to model the strand imbalance, where one strand is more diggested than the other. To asses for this effect we calculated the forward strand ratio, which is defined as:

$$\text{fwd strand ratio} = \frac{f}{f+r}$$

where

- f number of forward reads in region
- r number of reverse reads in region

At first, the genome was partitioned into fixed length bins. And for each bin, both the ChIP-exo and ChIP-seq forward strand ratio were calculated. We found that strands of reads were much less balanced in ChIP-exo data than in ChIP-seq data.

Need to make again this plots. proteins with both chip-exo and chip-seq data sets are:

- *landick, both rif-treatment and stat-vs-exp*
- *carroll, human samples*
- *ren, histones samples, comparing vs encode*

need to re do this figures, this is our version of figure 3b. in that figure only significant regions were used and the fragment reads were extended, therefore the extreme values $\{0, 1\}$ are not a heavy

Figure 4: ChIP-exo vs ChIP-seq forward strand ratio comparison

However, the objective is to asses the quality of a ChIP-exo experiment without a ChIP-seq one. For which, we analyzed the relationship the islands forward strand ratio and the region's depth.

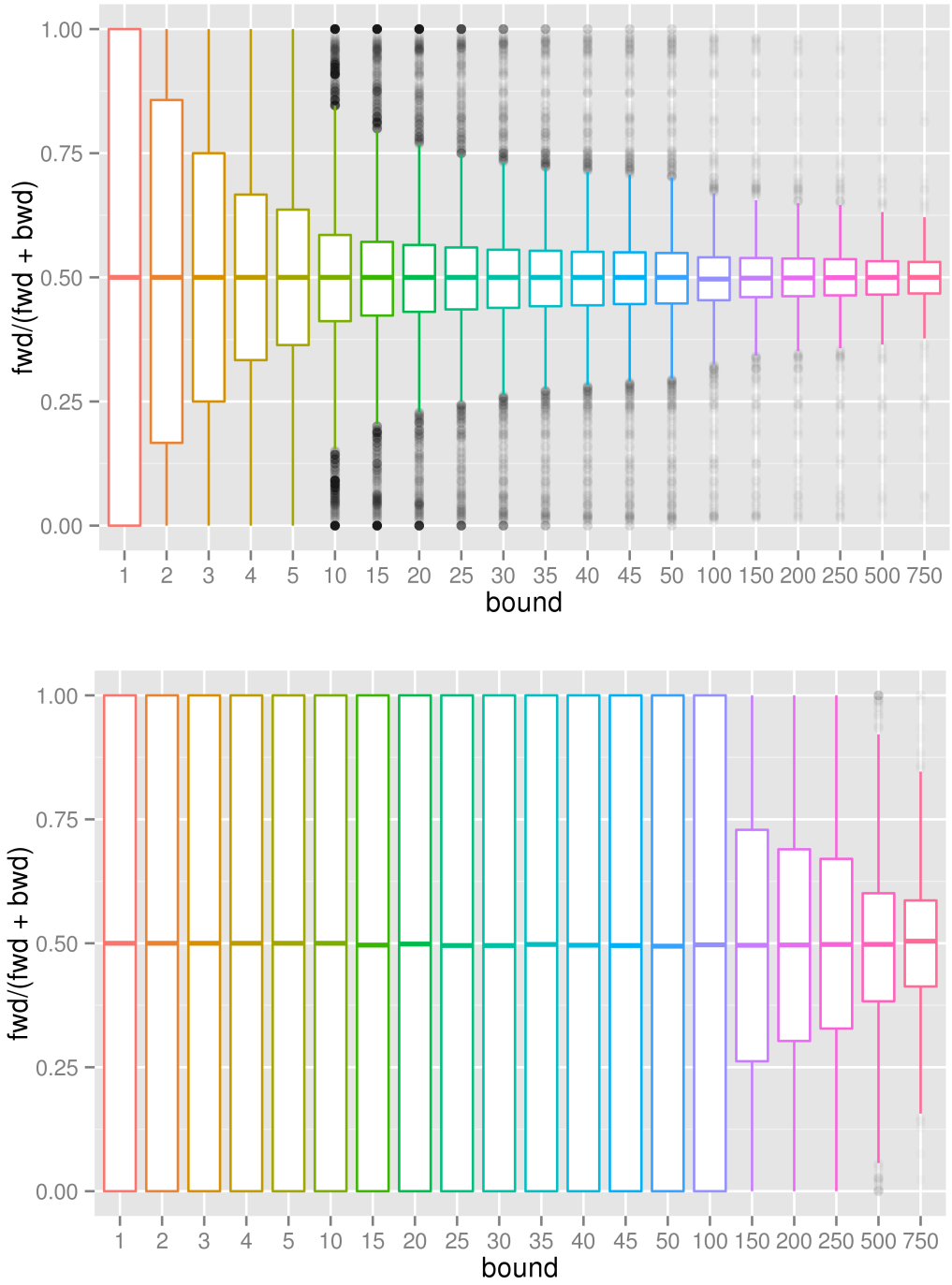


Figure 5: Forward strand ratio boxplot of all regions with depth greater than the x label: On top, ER transcription factor from MCF-7 cell line, generated in [2]; On bottom, H3k27ac histone from K562 cell line, provided by Ren's lab

Re do plot:

- *Change the ylabel is f / (f + r) and being consistent with equation above*
- *also change xlabel to something more descriptive*
- *possibly add the case of depth > 0 (still is the same as > 1)*
- *decide about keeping or removing the outliers in the plot, if I am not going to mention them then remove them*

In both panels of figure 5, it is possible to see that as the number of reads in each region increases then the forward

strand ratio tend to stabilize around 0.5. We observed that usually better quality datasets show a higher stabilization rate. The presence of "full" boxplots in both panels suggests the existence of strand specific regions.

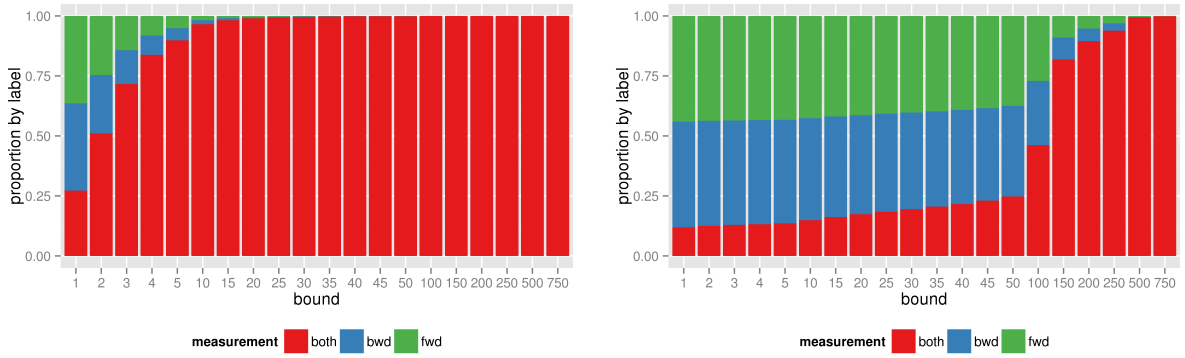


Figure 6: Strand specific proportion of all regions with depth greater than the x label: On left, ER transcription factor from MCF-7 cell line, generated in [2]; On right, H3k27ac histone from K562 cell line, provided by Ren's lab
again, change the x-lab to be more descriptive, perhaps y-axis should be strand specific proportion, it would be nice if we represent the total number of regions by depth

Figure 6 may be used to explain why the stabilization rate in the bottom case of 5 is low, since there are a considerable amount of strand specific regions.

To evaluate that the forward strand ratio is indeed a meaningful statistic to asses the quality of a dataset, we considered a conservative list of peaks called from a ChIP-seq experiment under the same conditions and separate them using the rule (as seen in 7):

$$z_i = \begin{cases} 1 & \text{if overlaps any of the peaks} \\ 0 & \text{otherwise} \end{cases}$$

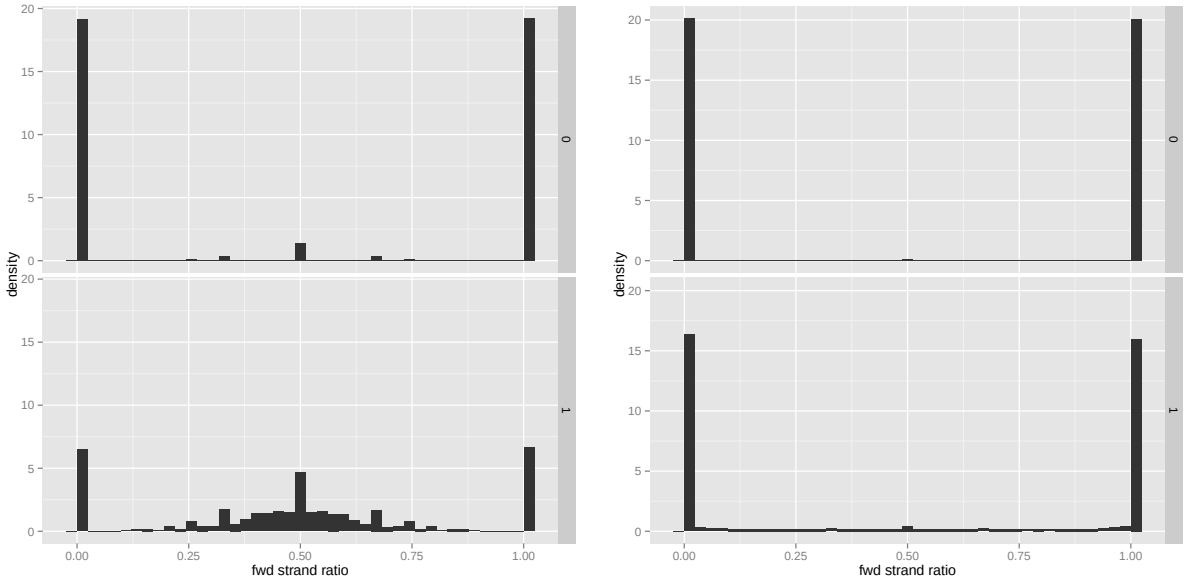


Figure 7: Same datasets as before. For the left case: the densities on top don't overlap any peaks, most of this regions are strand specific, in contrast with the bottom panels where most of the mass is around the 0.5, while for the right case both are very similar

need to call peaks with landick's chip-seq data, this figure is a place holder until I make the analogous plot with the e.coli data the data was generated as in [3]

To further study the relationship between depth and strand imbalance, we used MA-plots where the "green" and "red" are the forward and reverse reads in each region normalized by the region's width (w). That way we are plotting in figure 8:

$$A := \log_2 f + \log_2 r - 2\log_2 w$$
$$M := \log_2 f - \log_2 r$$

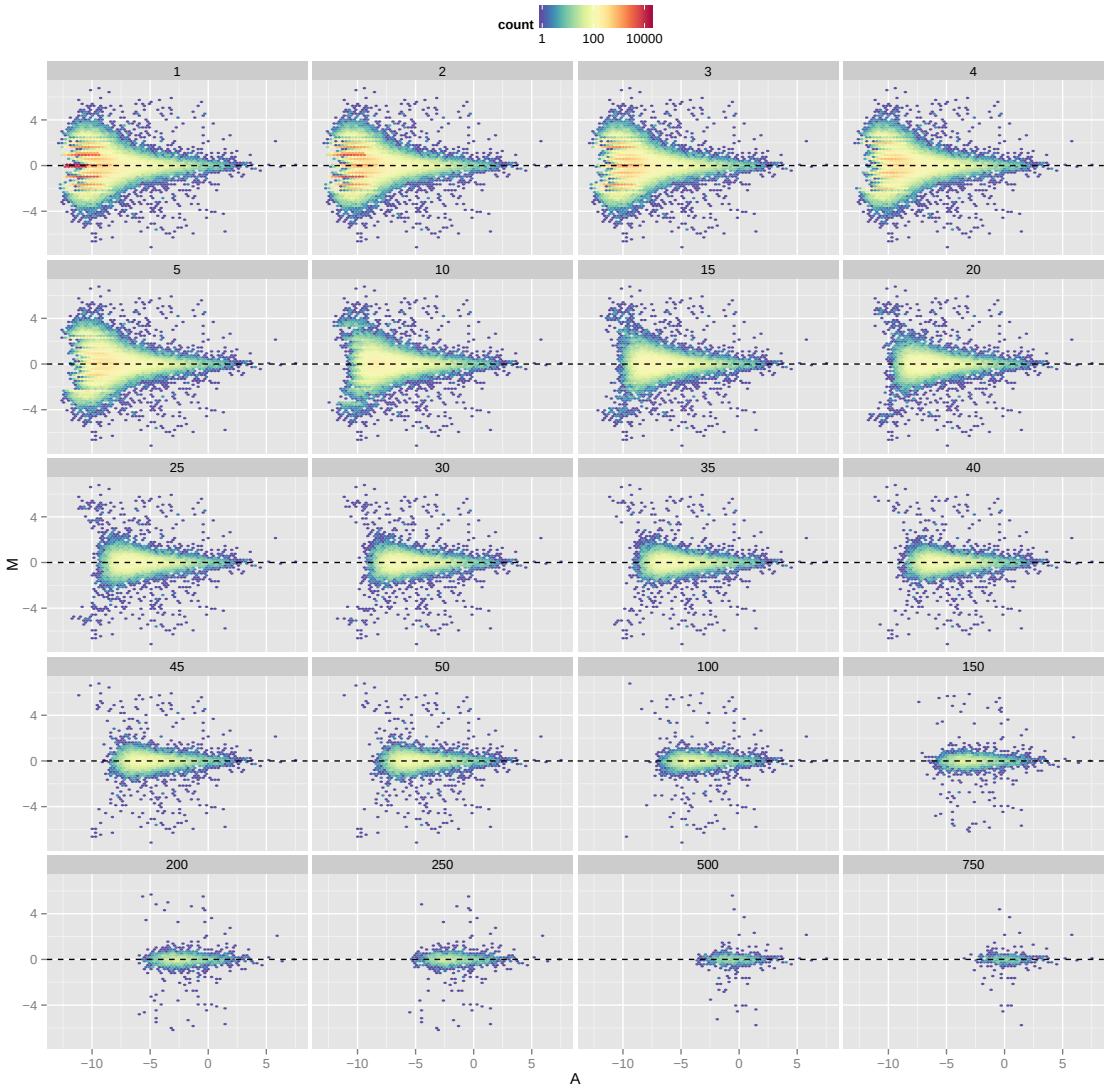


Figure 8: ER-rep1. Each panel consists of the MA plot as defined above, using all regions with depth higher than it's title.

perhaps remove bottom line ... not sure if top

this last plot is a bit of an overkill since we are still showing the same information

From figure 8 we can consider that as the depth increases the signal becames more localized around the x-axis. This strong arm accross the horizontal axis indicates that there are several regions such that they show both: high depth and similar number of forward and reverse fragments.

3.2 Signal to noise extraction and PCR amplification bias

Usually for this topic, the problem would be to find the enriched regions. For this case, we are not attempting to find the enriched regions but to asses the possibility of finding enriched regions in a certain dataset. To do so, we are considering for a given region, the following characteristics:

- Width
- Depth
- Number of unique positions

Not surprisingly all threee quantities are correlated between each other. Therefore is is necessary to adjust them in order to find more meaningful relationships in the data.

In order to understand the relationship among this quantities, we plotted the ratio of depth and width against the ratio of the number of unique positions and the depth of a region:

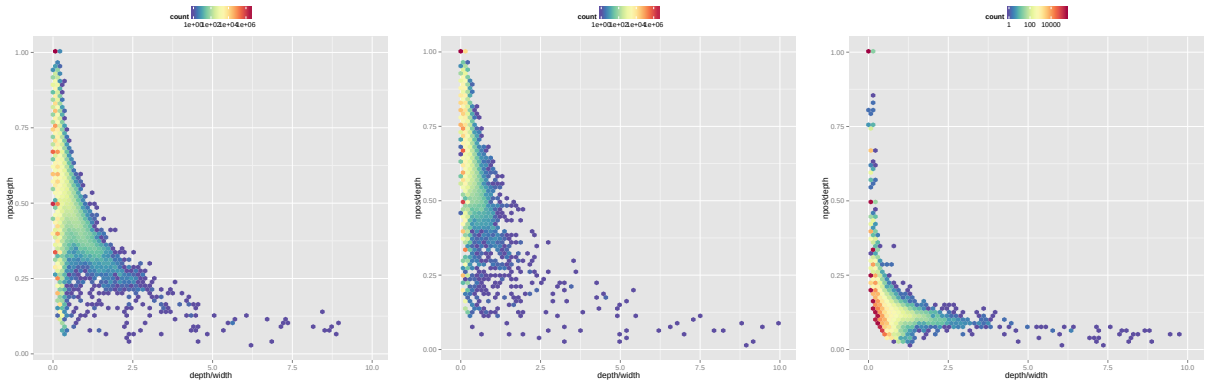


Figure 9: The three plots are biological replicates from FoxA1 transcription factor in mouse liver tissue (from [2])

Some observations about figure 9:

- The x-axis is the ratio of depth and width. This statistic is going to be large in enriched regions, since in those regions the signal it is likely to be more localized. However, it can be small when either the region is wide or the depth is low
- The y-axis is the ratio between the number of unique positions in the region and it's depth. Clealy the number of positions is less or equal than it's depth, thus it is always positive and ≤ 1.
- When npos is low and the depth is high, this are the regions where there is some sort of PCR-artifact like behaviour.
- When npos is high and depth is low then the region is going to be a dispersed region, such as the input in a ChIP-seq experiment
- Since usually depth $>>$ npos, then it is possible for the dataset to show and adequate quality and the npos / depth ratio to be low.
- In the left figure we can see two strong arms: One is a vertical arm close to the y-axis, usually those regions come from not enriched regions and another arm which decreases as depth increases.

for this datasets, the order of max number of position is rep1 $>$ rep2 $>$ rep3. Therefore an idea to separate this two arms, I can stratify by npos

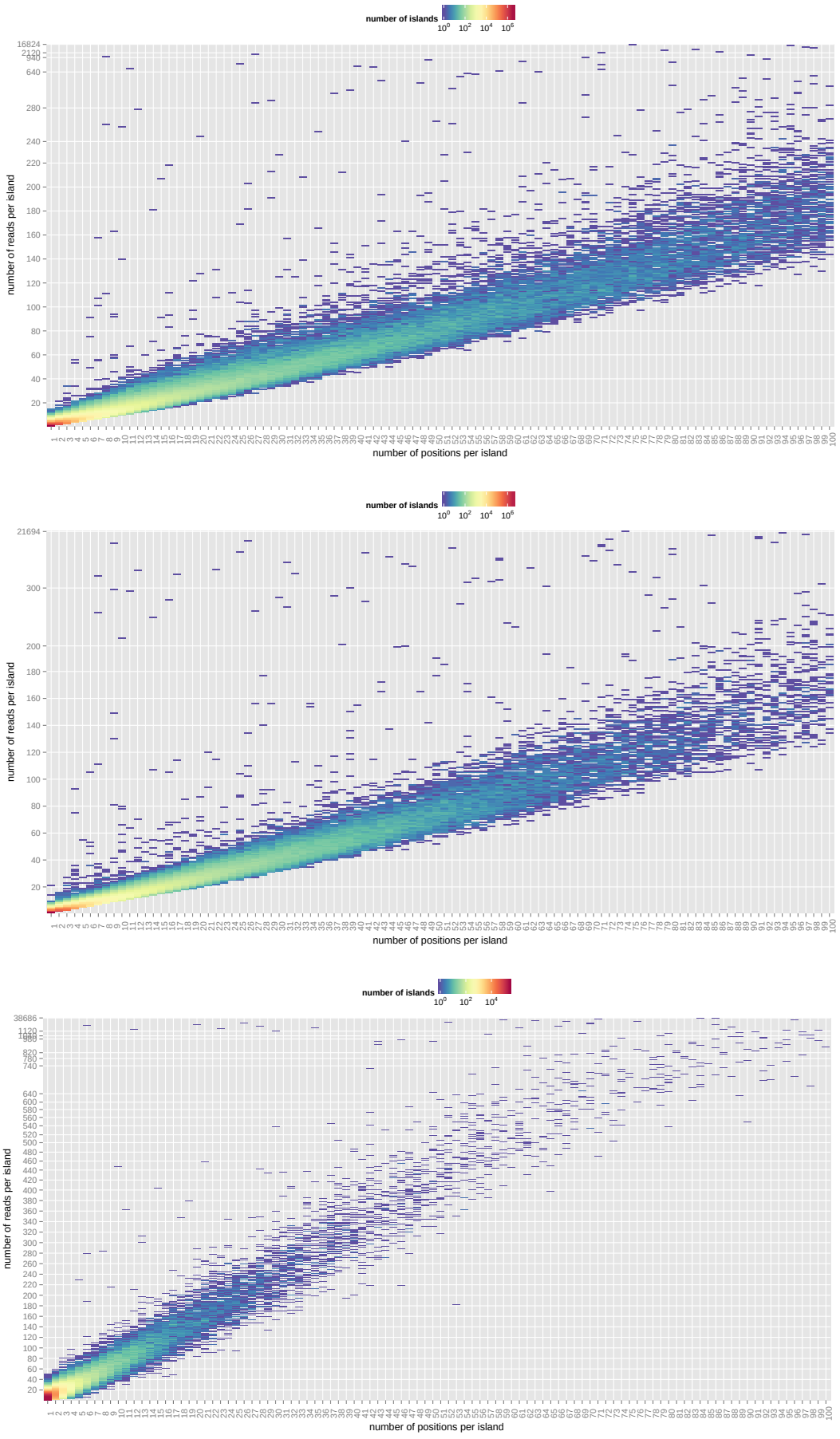


Figure 10: The three plots are biological replicates from FoxA1 transcription factor in mouse liver tissue (from [2])

1. *need to make a better scale of this plot*
2. *the scale of y-axis for the 3rd plot is little bit missleading*
3. *perhaps to bound the number of positions by 70 and the depth by 200*

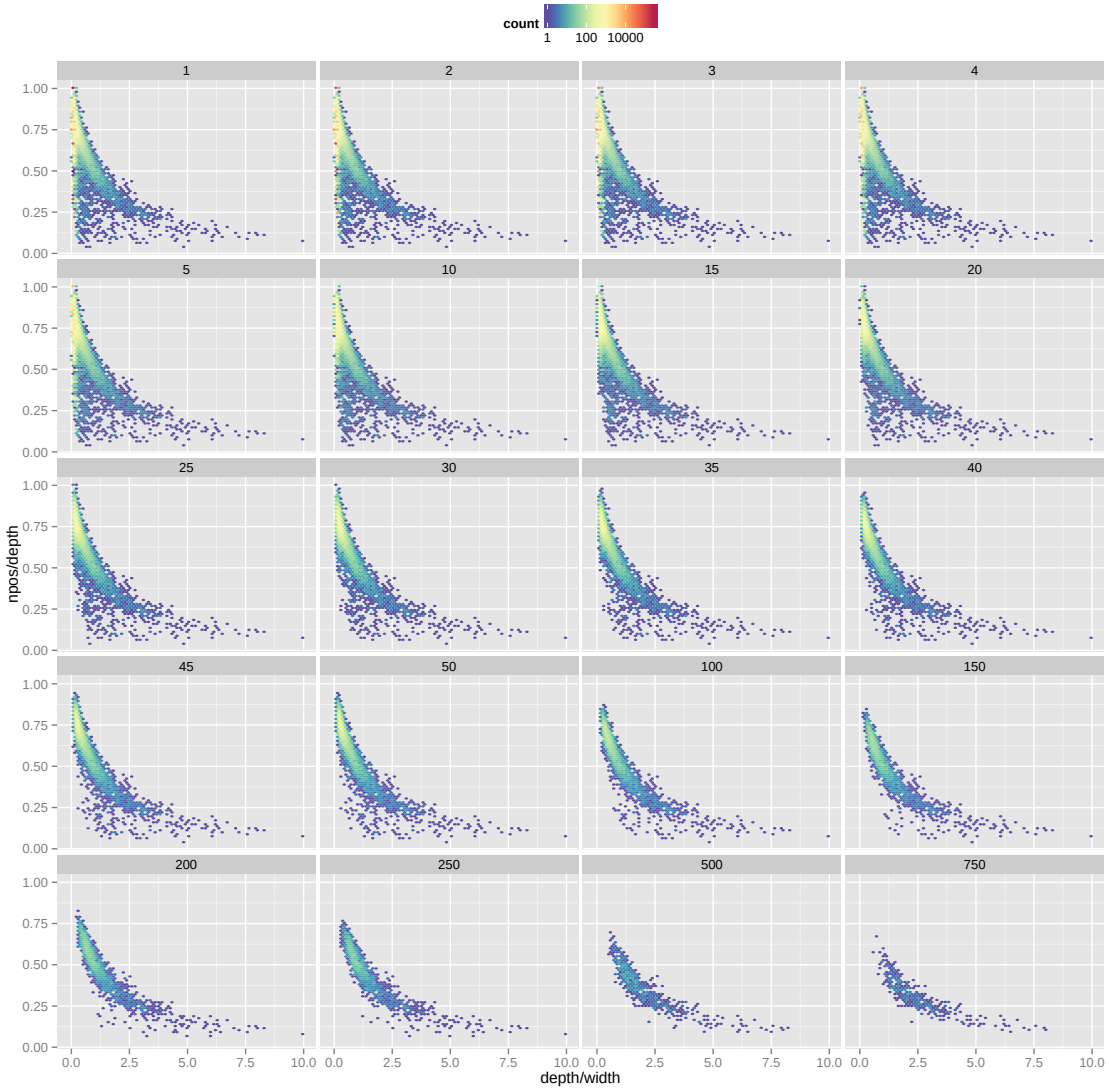
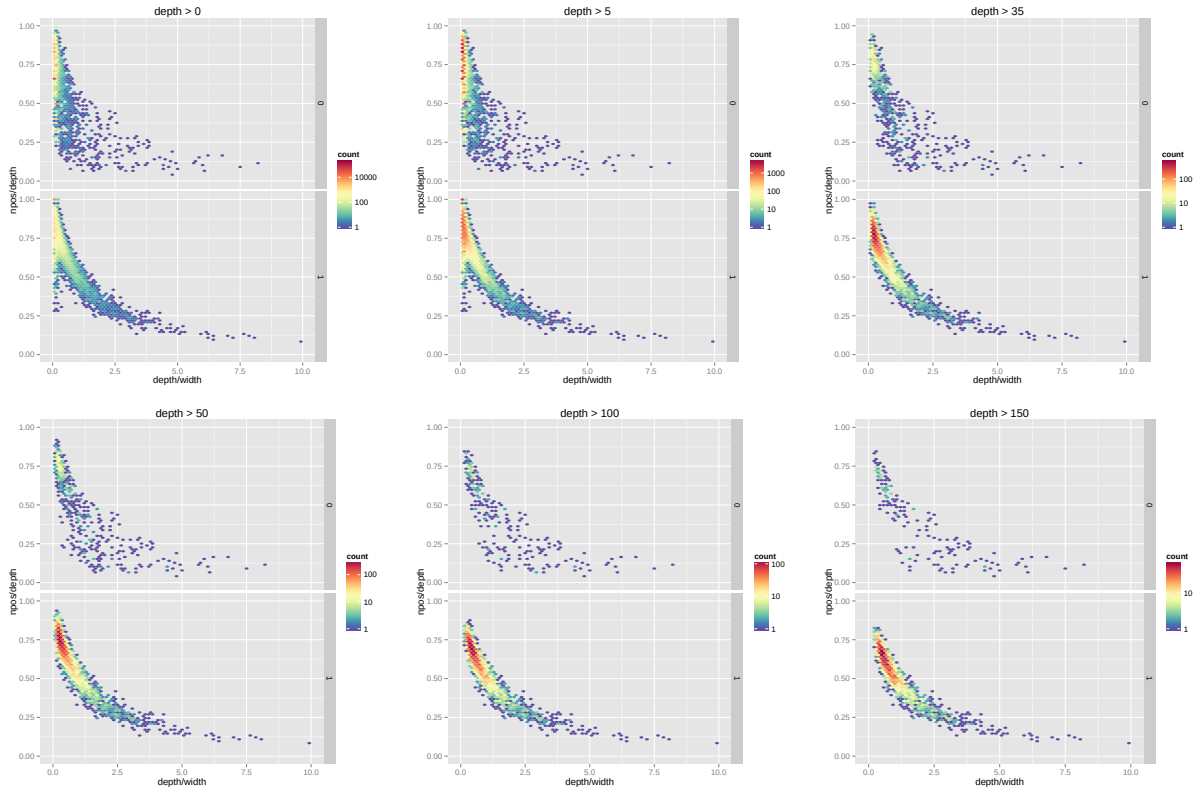


Figure 11: depth/width vs npos/depth plot. The panel title indicates that this plot shows only the regions with depth greater than the title

the idea here was to study npos / depth npos $\leq$ depth , thus $0 \leq$ npos / depth ≤ 1 by itself can't find it by joining with depth / width can find those regions



also, for pcr amplification there is the concept of mesa and we have at least 6 chip-exo datasets with a set of mesas. It would be interesting to see the comparison of mesas vs not- mesas for the human samples since their collection of mesas are much bigger than the case for mouse

3.3 Enzyme overdigestion

For ChIP-exo, we are expecting to see the mode (or modes) of the distribution to have a small positive values. Also, we expect it to have only non-negative values (which is not the case, however this plot considers all the regions and not only the ones where there is a binding event). A promising feature is that the density for the "both" label seem to be lower than the densities of the other labels. Perhaps by considering peaks, we can trim this density.

Finally we can see figure 12 which show the density of all the sequencing procedures for the "both" label:

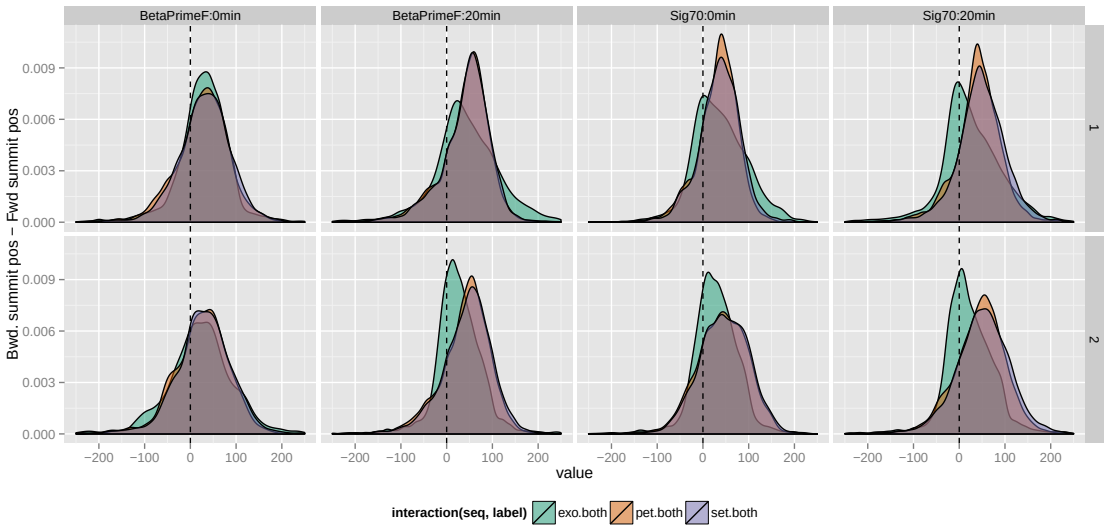


Figure 12: Density of difference in summit positions separated by sequencing for "both" label
it would be nice to update this plots with the new partitions, also there are chip-seq and chip-exo data for human samples (not only e.coli). right now this works as a placeholder

Figure 12 looks very promising since, by considering a very simple classification we are seeing that the difference for the ChIP-exo data set is lower than the difference for the ChIP-seq cases. There are a large amount of cases where the difference is negative, which may be filtered by considering them in an case by case basis.

to do this I tried to calculate local cross - correlation for all islands, we have examples of local islands where is more or less can estimate the difference

for this part I may use the densities of summit_diff. perhaps can re do it using localMaxima + smoothing ... the issue with this indicator is gonna be that calculating the coverage for each region is very time consumming

make an histogram comparison of npos, perhaps strand specific vs both stranded regions

4 Comparison among replicates

this section may be useful but perhaps for future work. so far, we have seen cases where islands with high number of positions are repeated in all replicates of the experiment

this suggests to use number of unique positions as a ranking system, maybe -log(npos) (this to flip the order) and apply idr

need to build this section, one possible idea would be to rank by the use of number of positions and then apply the algorithm

example of raw data, to show the quality of good vs bad data sets

References

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

[2] 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.

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