

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

1	Overview of ChIP-exo	2
2	ENCODE's quality metrics on ChIP-exo datasets	2
2.1	PCR bottleneck coefficient	3
2.2	Strand cross-correlation	4
3	Methodology description	5
3.1	Strand imbalance	5
3.2	Signal to noise extraction and PCR amplification bias	9
3.3	Signal to noise examples	9
3.4	Signal to noise replicability	11
3.5	Enzyme overdigestion	17
4	Comparison among replicates	18

List of Figures

1	ChIP-exo protocol, the figure is from [1]	2
2	PBC comparison among protocols	4
3	Strand cross-correlation for ChIP-exo	4
4	ChIP-exo vs ChIP-seq forward strand ratio comparison	5
5	Forward strand ratio boxplots	6
6	Strand composition proportion plots	7
7	Fwd strand ratio evaluation	7
8	ER-rep1. Each panel consists of the MA plot as defined above, using all regions with depth higher than it's title.	8
9	depth/width vs npos/depth plots	9

10	High quality regions of ER in MCF-7 human cell lines (rep 1) showing the contrast between average read coverage and unique read coverage rate	10
11	Regions of Fox-A1 in mouse liver tissue (rep 3) showing low unique read coverage rate but high average read coverage	10
12	Regions of ER in human MCF-7 cell lines (rep 1) showing low average read coverage but high unique read coverage rate	11
13	Regions of ER in human MCF-7 cell line (rep 1) showing the very low average read coverage and high unique coverage rate	11
14	Comparison of the arc vs urcr plot among all three replicates of ER tf in MCF-7 cell. In the top panel, we are considering all region in gold standard peak set and in the bottom all region with number of unique 5' positions greater than 25	12
15	Comparison of the arc vs urcr plot among both replicates of H3k4me1 in K562 cell. In the top panel, we are considering all region in gold standard peak set and in the middle and bottom the regions with number of unique 5' positions greater than 5 and 25 respectively	13
16	Number of unique positions vs depth	15
17	depth/width vs npos depth plot, filtered by depth	16
18	Density of difference in summit positions separated by sequencing for "both" label	18

1 Overview of ChIP-exo

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

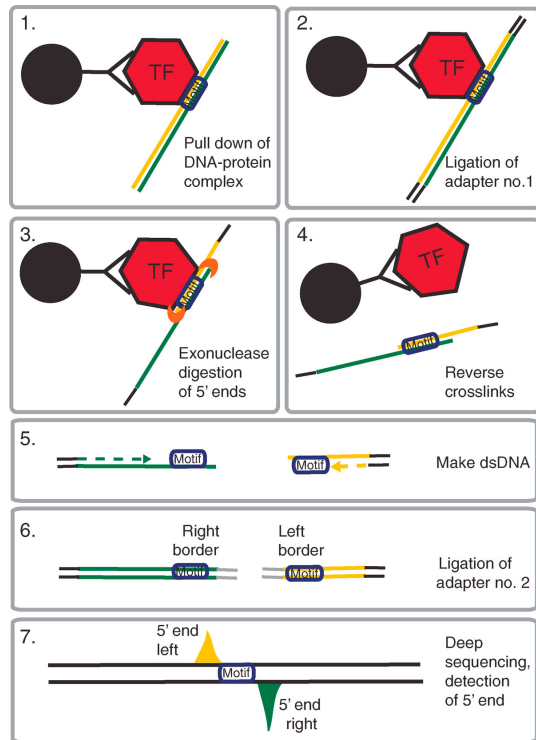


Figure 1: ChIP-exo protocol, the figure is from [1]

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

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

1. The motif is occurring 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 reaches 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 analyzed 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 $2\delta_2$.

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	114.132
Sig70-exp-rep2	864,714	0.247	160.850
Sig70-stat-rep1	1,584,532	0.259	230.152
Sig70-stat-rep2	1,012,936	0.247	325.520
SigmaS-exp-rep1	1,593,964	0.294	-162.618
SigmaS-exp-rep2	3,405,118	0.177	40.318
SigmaS-stat-rep1	1,822,585	0.232	58.953
SigmaS-stat-rep2	9,898,733	0.163	52.232
dataset	depth	pbc	nsc
BetaPf-rif0-rep1	1,875,127	0.251	1.794
BetaPf-rif0-rep2	898,641	0.302	1.896
BetaPf-rif20-rep1	4,900,071	0.234	1.197
BetaPf-rif20-rep2	6,550,805	0.215	15.388
Beta-rif0-rep1	3,909,669	0.283	1.854
Beta-rif0-rep2	5,157,768	0.256	1.859
Beta-rif20-rep1	5,153,689	0.338	1.280
Beta-rif20-rep2	1,509,554	0.498	-81.999
Sig70-rif0-rep1	960,256	0.282	10.294
Sig70-rif0-rep2	2,247,295	0.266	25.092
Sig70-rif20-rep1	1,940,387	0.270	1.835
Sig70-rif20-rep2	4,229,574	0.215	14.106

dataset	depth	pbc	nsc
CTCF	48,478,450	0.458	15.977
dataset	depth	pbc	nsc
FoxA1-rep1	22,210,461	0.656	21.452
FoxA1-rep2	23,307,557	0.800	60.661
FoxA1-rep3	22,421,729	0.107	72.312
dataset	depth	pbc	nsc
ER-rep1	9,289,835	0.808	19.843
ER-rep2	11,041,833	0.802	21.442
ER-rep3	12,464,836	0.820	18.699
dataset	depth	pbc	nsc
H3k27ac	29,599,796	0.305	1.531
H3k4me1-rep1	28,794,319	0.258	1.311
H3k4me1-rep2	31,818,368	0.252	1.291

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 suggest 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-existent

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 immunoprecipitate, and as a result it 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

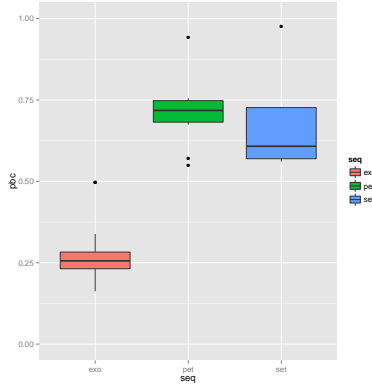


Figure 2: PBC comparison among protocols

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 pairs 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.

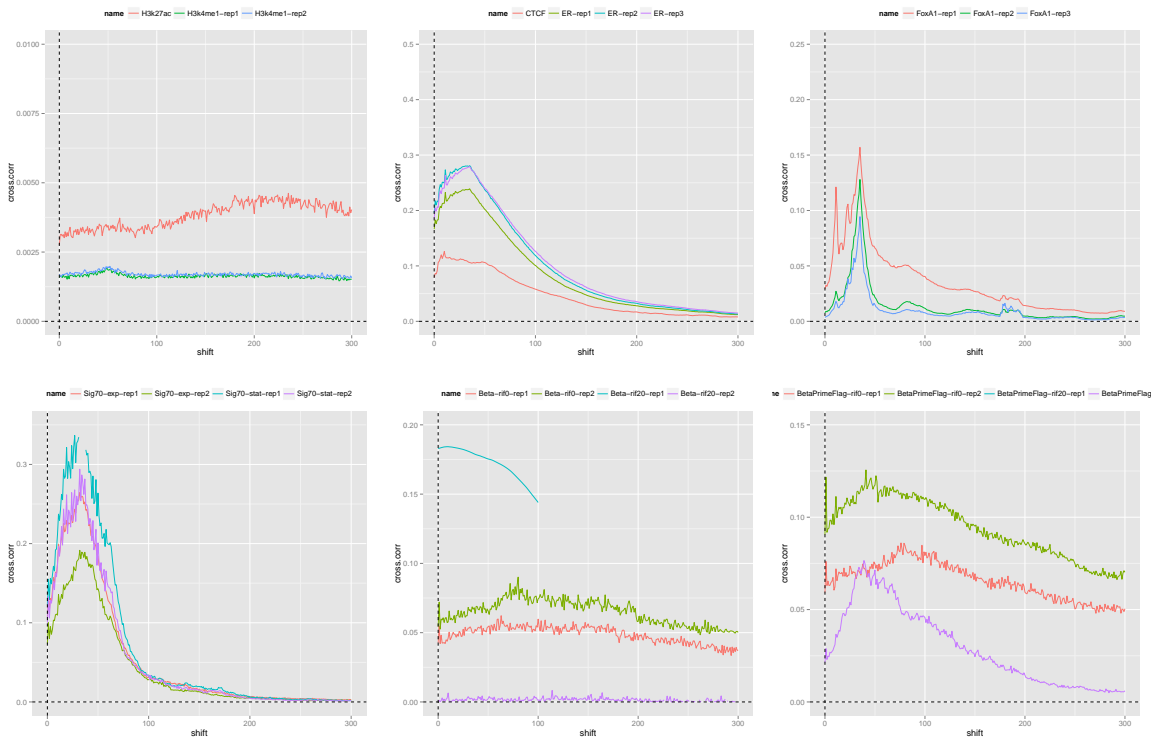


Figure 3: Different examples of ChIP-exo cross-correlations

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 chosen 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

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 digested 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.

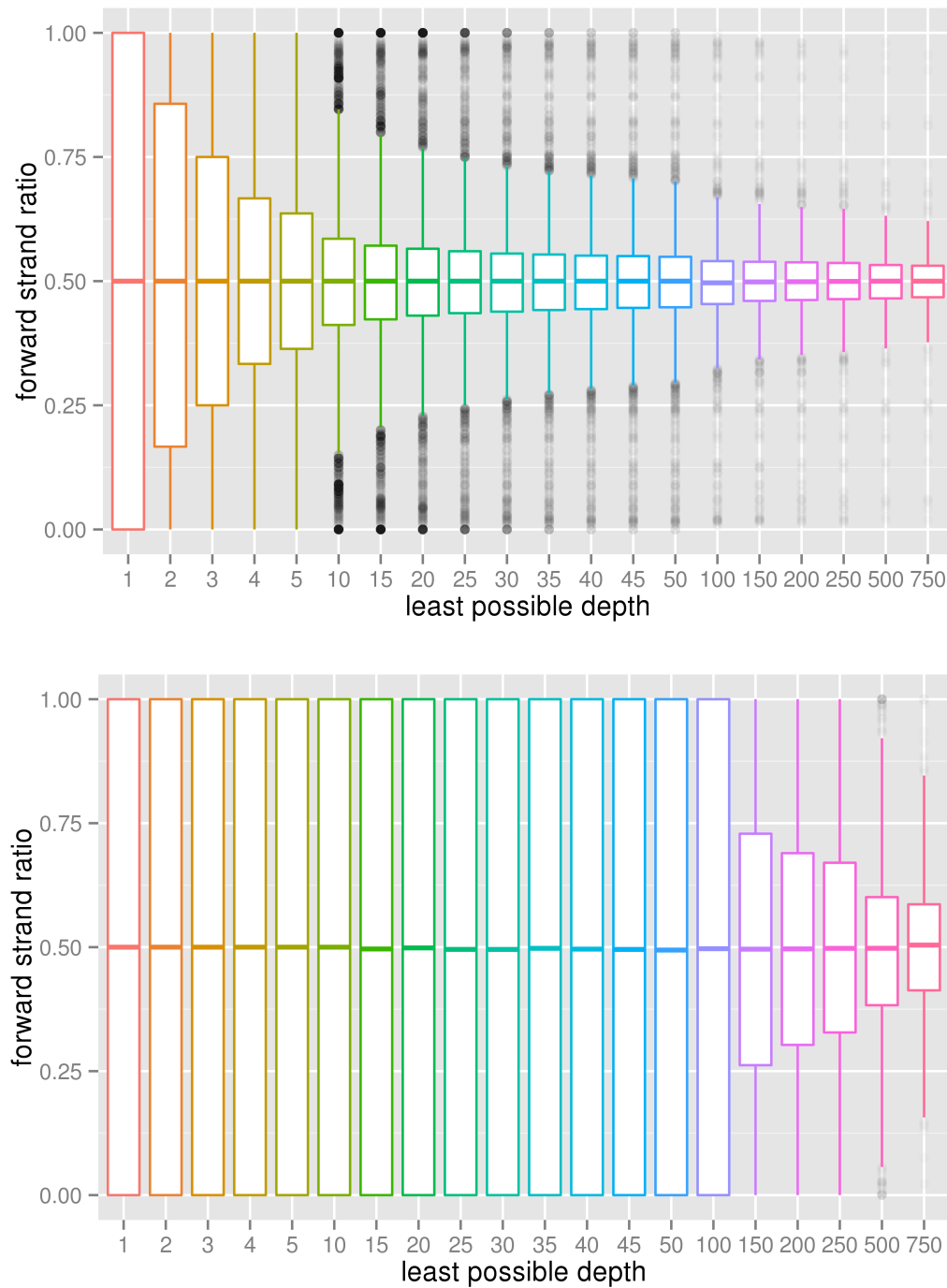


Figure 5: Forward strand ratio boxplot of all regions with depth greater than the x label: On top, ER transcription factor from MCF-7 cell line, generated in [2]; On bottom, H3k27ac histone from K562 cell line, provided by Ren's lab

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.

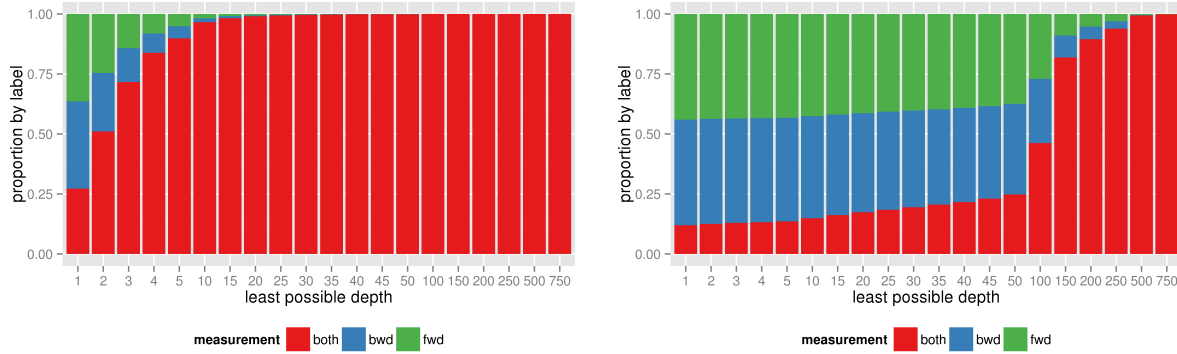


Figure 6: Strand specific proportion of all regions with depth greater than the x label: On left, ER transcription factor from MCF-7 cell line, generated in [2]; On right, H3k27ac histone from K562 cell line, provided by Ren's lab
again, change the x-lab to be more descriptive, perhaps y-axis should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

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

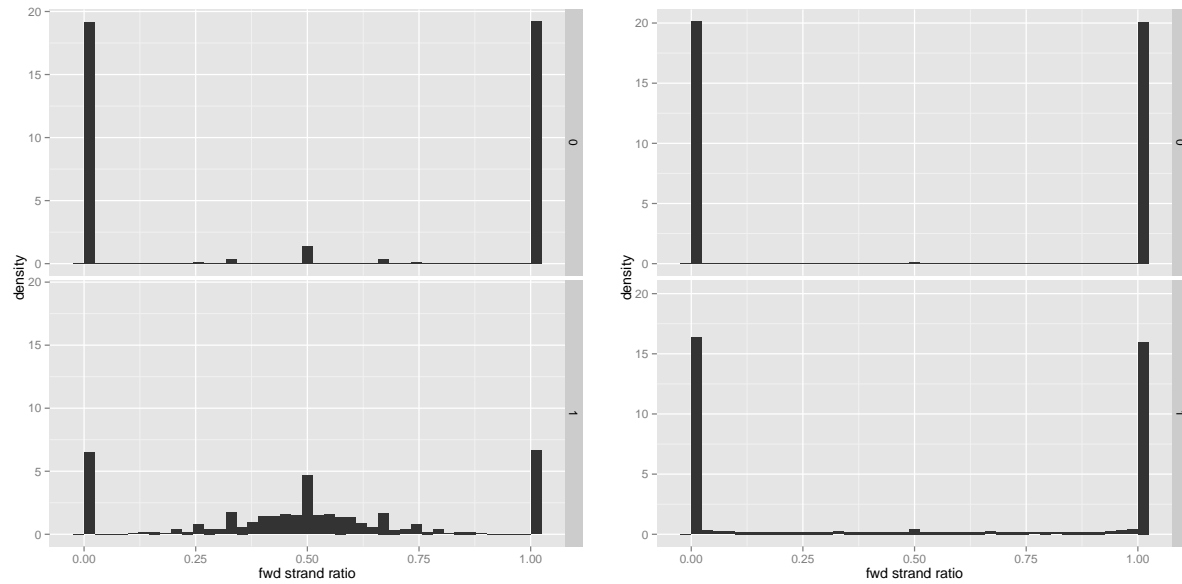


Figure 7: Same datasets as before. For the left case: the densities on top don't overlap any peaks, most of this regions are strand specific, in contrast with the bottom panels where most of the mass is around the 0.5, while for the right case both are very similar

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:

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

$$M := \log_2 f - \log_2 r$$

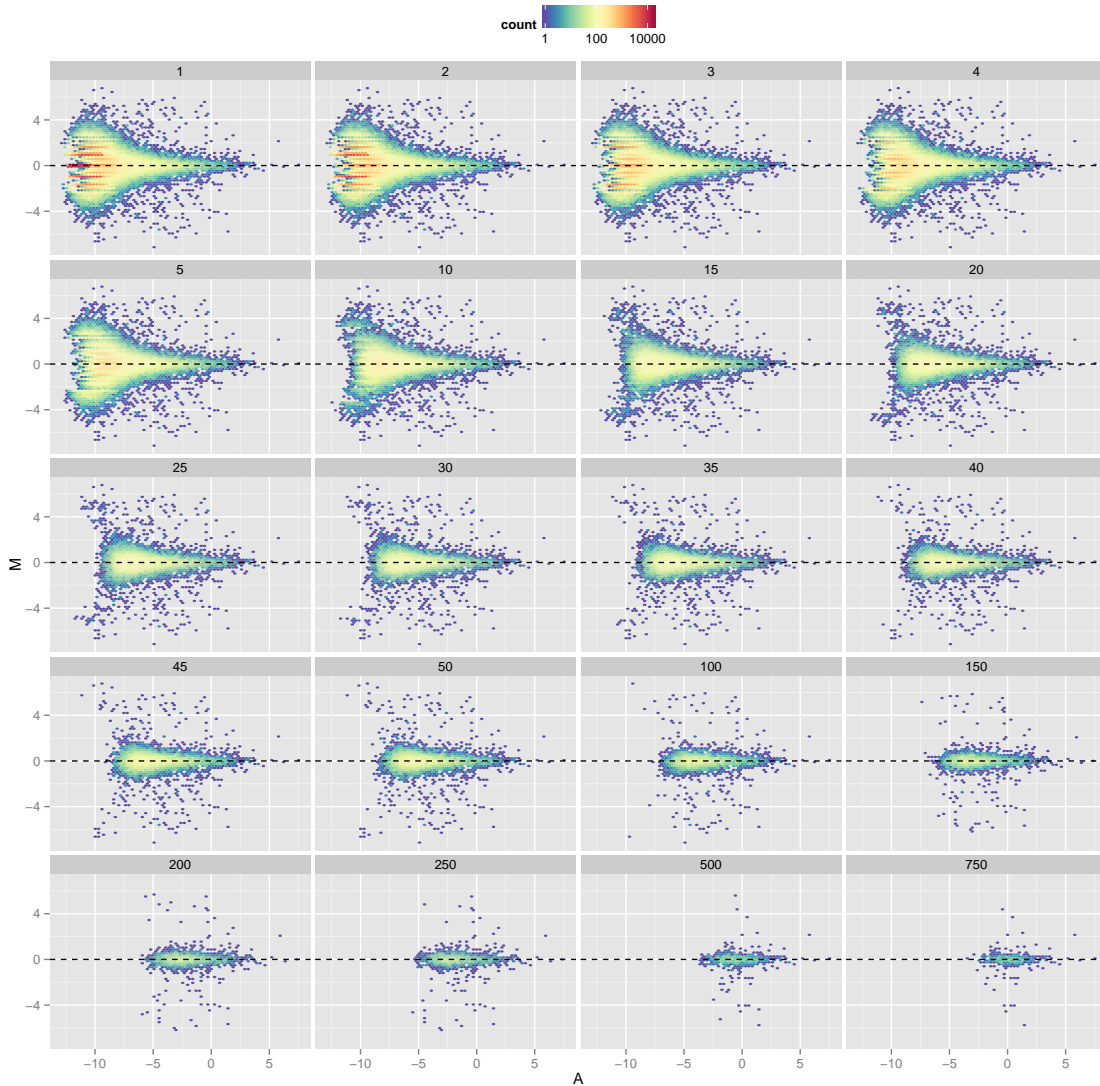


Figure 8: ER-rep1. Each panel consists of the MA plot as defined above, using all regions with depth higher than it's title.

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 becomes more localized around the x-axis. This strong arm across 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 assess 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 three quantities are correlated between each other. Therefore it is necessary to adjust them in order to find more meaningful relationships in the data.

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

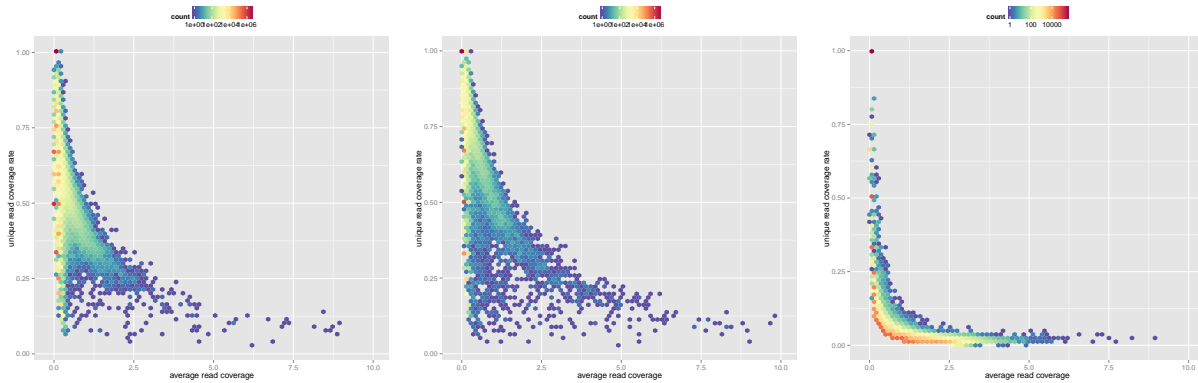


Figure 9: 3 different examples of the signal to noise plot: On left FoxA1 on mouse liver tissue; at the center ER in MCF-7 cell line (both from [2]; At right H3k27ac in K562 provided by Ren's lab ENCODE3

In both the left and center panels of figure 9, it is possible to see two strong arms: One is a vertical arm close to the unique read coverage rate axis, which shows regions with small average read coverage, these regions come not from enriched regions but from regions with low depth and may be part of the background. The other, is an arm which shows the contrast between the average read coverage and the unique read coverage rate. In the right panel of figure 9 we are observing a low quality dataset since several of its regions present either low average read coverage or low unique read coverage rate.

For this figure we are plotting:

$$\text{Average read coverage} = \frac{\text{Region's depth}}{\text{Region's width}}$$

$$\text{Unique read coverage rate} = \frac{\text{Number of unique 5' positions in region}}{\text{Region's depth}}$$

- Average read coverage, which is usually higher in enriched regions than in non-enriched ones.
- Unique read coverage rate (Z). This rate is non-negative and its maximum is obtained at one. The statistic denotes a contrast between two extreme behaviors:
 - When $Z = 1$, in these regions we can see a different read fragment for each 5' position.
 - When $Z \rightarrow 0$, then either the region's depth is very high or the number of unique positions that form the region is very low. In this case, then this region's coverage is a step-wise function with "larger" jumps, which may resemble PCR artifacts.

3.3 Signal to noise examples

re do peak plots. for the annotation need to remove MA values. Denote prob as fsr (forward strand ratio) and pbc as urcr (unique read coverage rate) and dw_Ratio as arc (average read coverage)

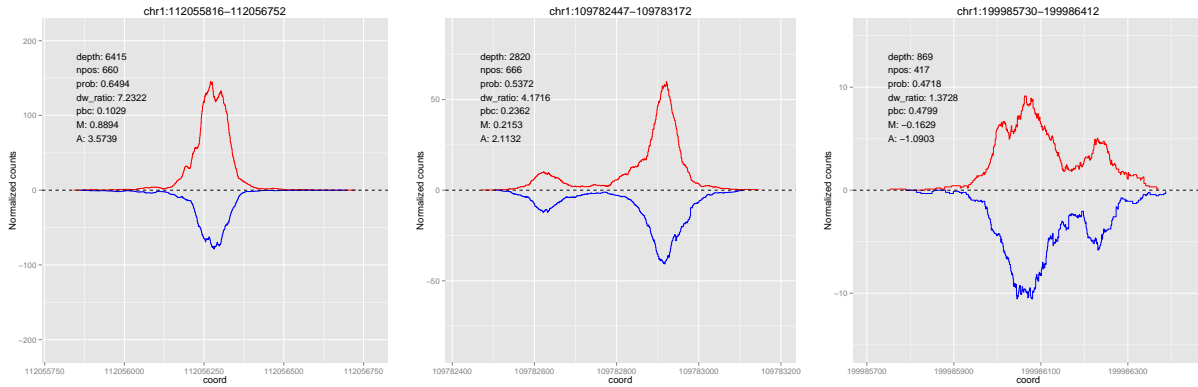


Figure 10: High quality regions of ER in MCF-7 human cell lines (rep 1) showing the contrast between average read coverage and unique read coverage rate

In figure 10 we can see three different cases of good regions, which are sorted decreasingly according to its unique read coverage rate or increasingly according to its average reads coverage. This, shows the existent contrast between both statistics since the values of both are highly different between the three examples but all regions shown possibly enriched regions.

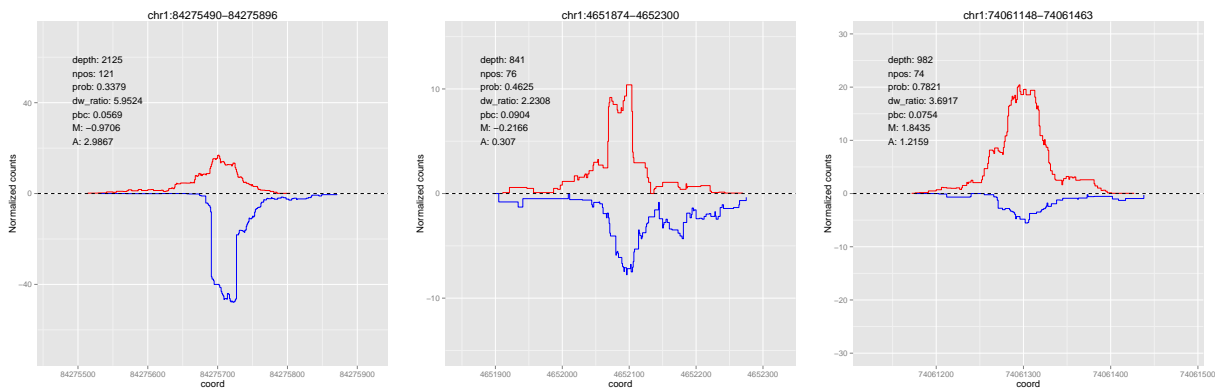


Figure 11: Regions of Fox-A1 in mouse liver tissue (rep 3) showing low unique read coverage rate but high average read coverage

In figure 11 we are observing three examples of regions where the average read coverage is high but the unique read coverage rate is low. This regions are characterized by being constructed with a lower number of unique positions and usually show certain degree of strand imbalance. In a ChIP-seq experiment this patterns may be occurring due to PCR artifact, in which certain position is being sequenced exponentially, but for ChIP-exo this regions may be appearing due to the exonuclease digestion, which trim the fragments starting from the 5' end until it reached the protein binding site.

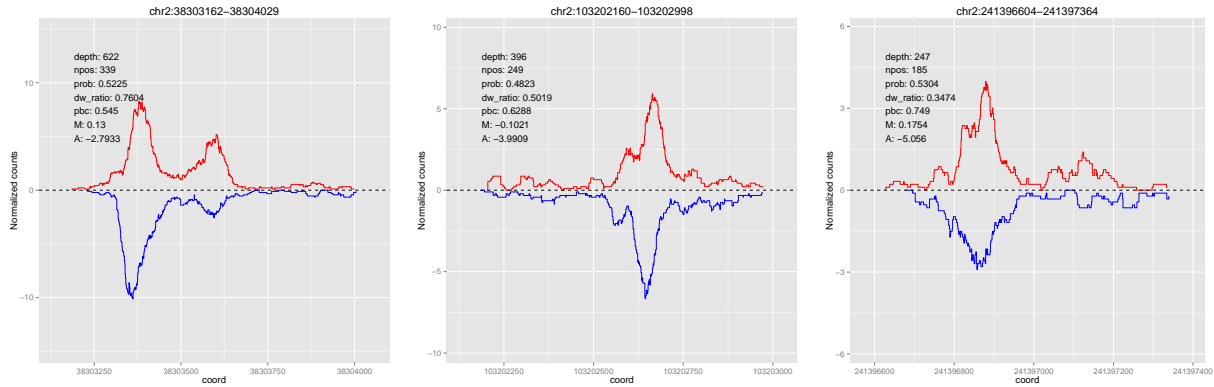


Figure 12: Regions of ER in human MCF-7 cell lines (rep 1) showing low average read coverage but high unique read coverage rate

In figure 12 we are observing examples where the average read coverage is low but the unique read coverage rate is high. In this examples, it is noticeable that the average read coverage is decreasing due to the decreasing depth (from left to right), since all three regions show roughly the same width. It is also possible to notice that the signal is not as defined as in figures 10 or 11.

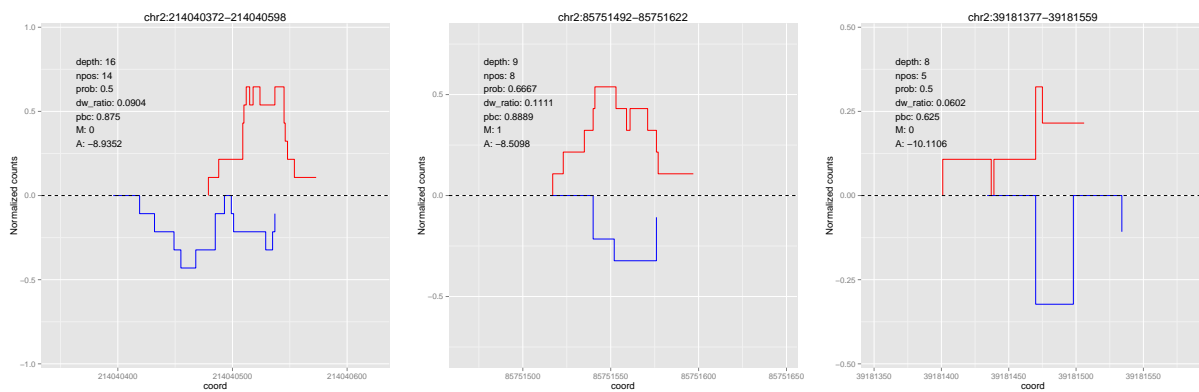


Figure 13: Regions of ER in human MCF-7 cell line (rep 1) showing the very low average read coverage and high unique coverage rate

In figure 13 we are observing several examples of low quality regions. This regions correspond to the strong arms close to the unique read coverage rate axis in the left and middle panels of figure 9. In this case the average read coverage is very low due to very low depth, which may be formed by few background fragment reads that weren't digested by the enzyme.

3.4 Signal to noise replicability

9a) reminds of ChIP vs input plots, which are the key plots I look at before starting to analyze a ChIP-seq dataset.

This plot definitely needs more characterization and documenting typical behavior across multiple datasets. We might want to look at this plot for peaks that are reproducible.

Actually, for any of these metrics, we can define gold standard peak set (reproducible) and justify the metrics that way.

To assess the replicability among samples of the same experiment we built the old standard peak set by considering the intersection among the estimated regions for each data set, and removing peaks with a width smaller than the experiment's average read length.

Then for each region we extracted the reads that overlapped it for all samples, and recalculate the average read coverage and unique read coverage rate for those regions.

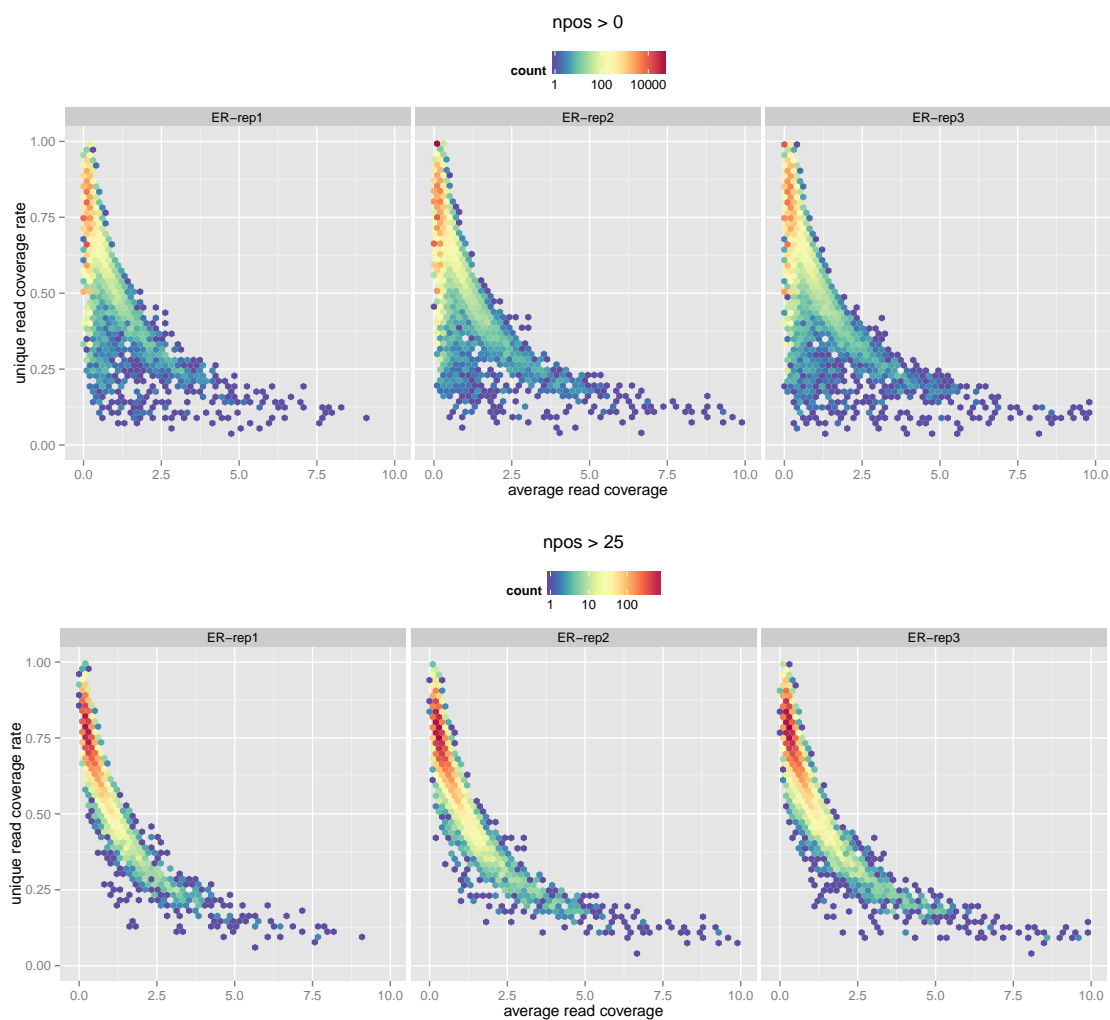


Figure 14: Comparison of the arc vs urcr plot among all three replicates of ER tf in MCF-7 cell. In the top panel, we are considering all region in gold standard peak set and in the bottom all region with number of unique 5' positions greater than 25

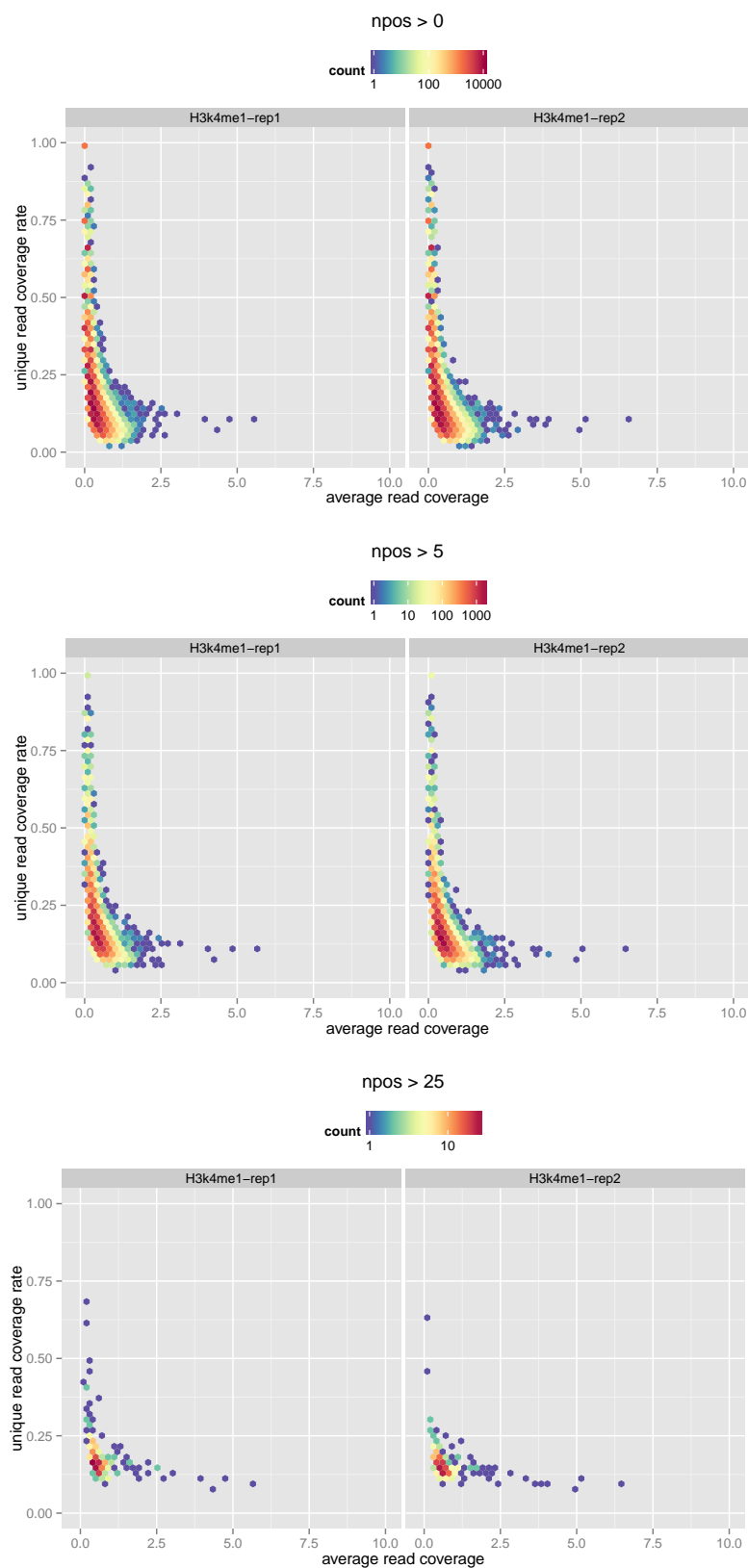


Figure 15: Comparison of the arc vs urcr plot among both replicates of H3k4me1 in K562 cell. In the top panel, we are considering all region in gold standard peak set and in the middle and bottom the regions with number of unique 5' positions greater than 5 and 25 respectively

In figure 14 we can observe that the we are covering the same region in the arc vs urcr plot, which means that this plot is replicable by considering all three samples. On the other hand, by comparing top vs bottom, we are seeing that the region with low arc and varying urcr is filtered out, therefore that region of the plot consist of regions with low number of unique 5' positions. While in figure 15 we are observing i) most of the region have a low arc and low urcr for both replicates; ii) the number of unique position in each region is usually low, since most of the regions seem to be filtered out when the number of unique positions is larger than 25; and iii)

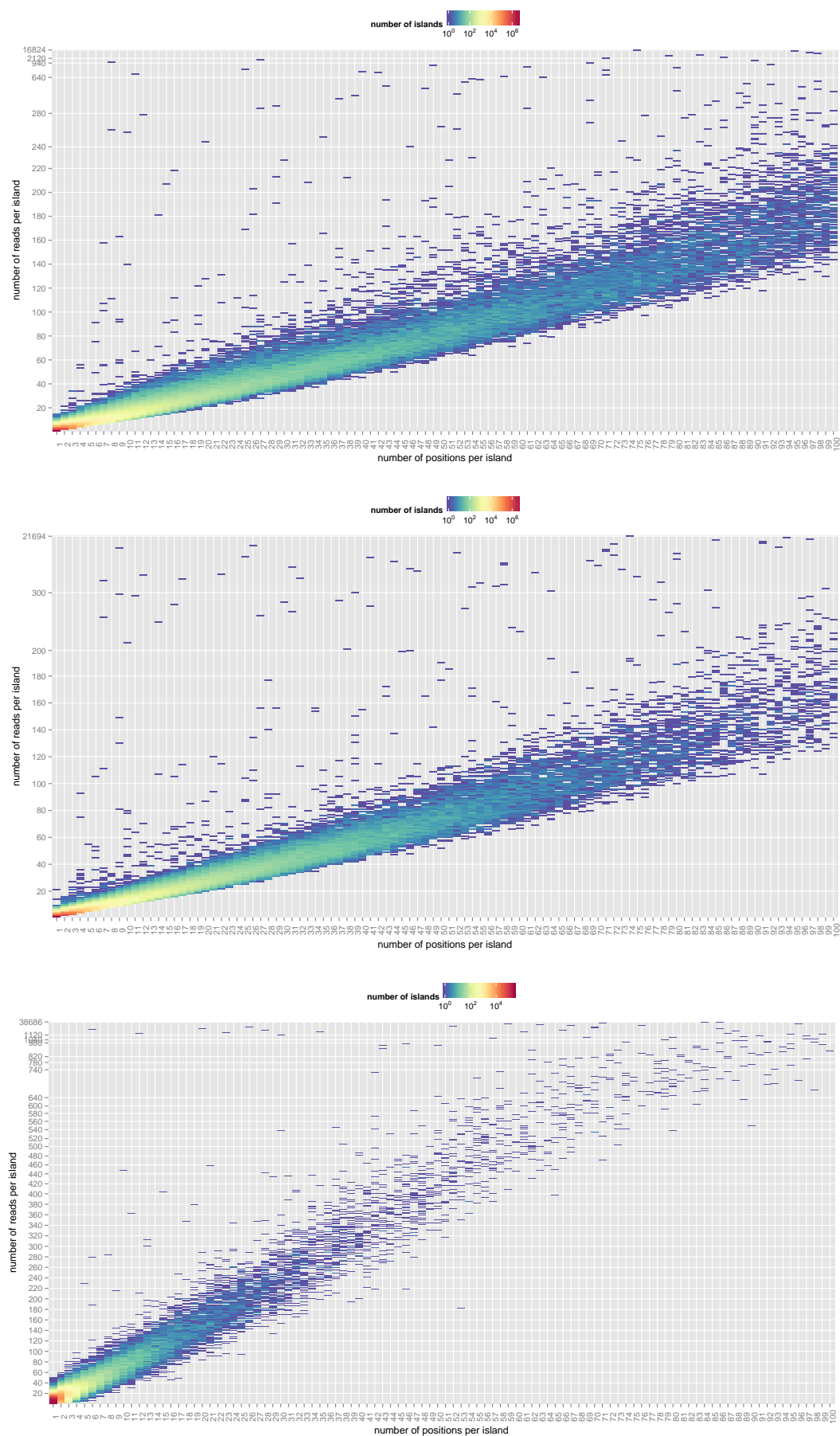


Figure 16: The three plots are biological replicates from FoxA1 transcription factor in mouse liver tissue (from [2])

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

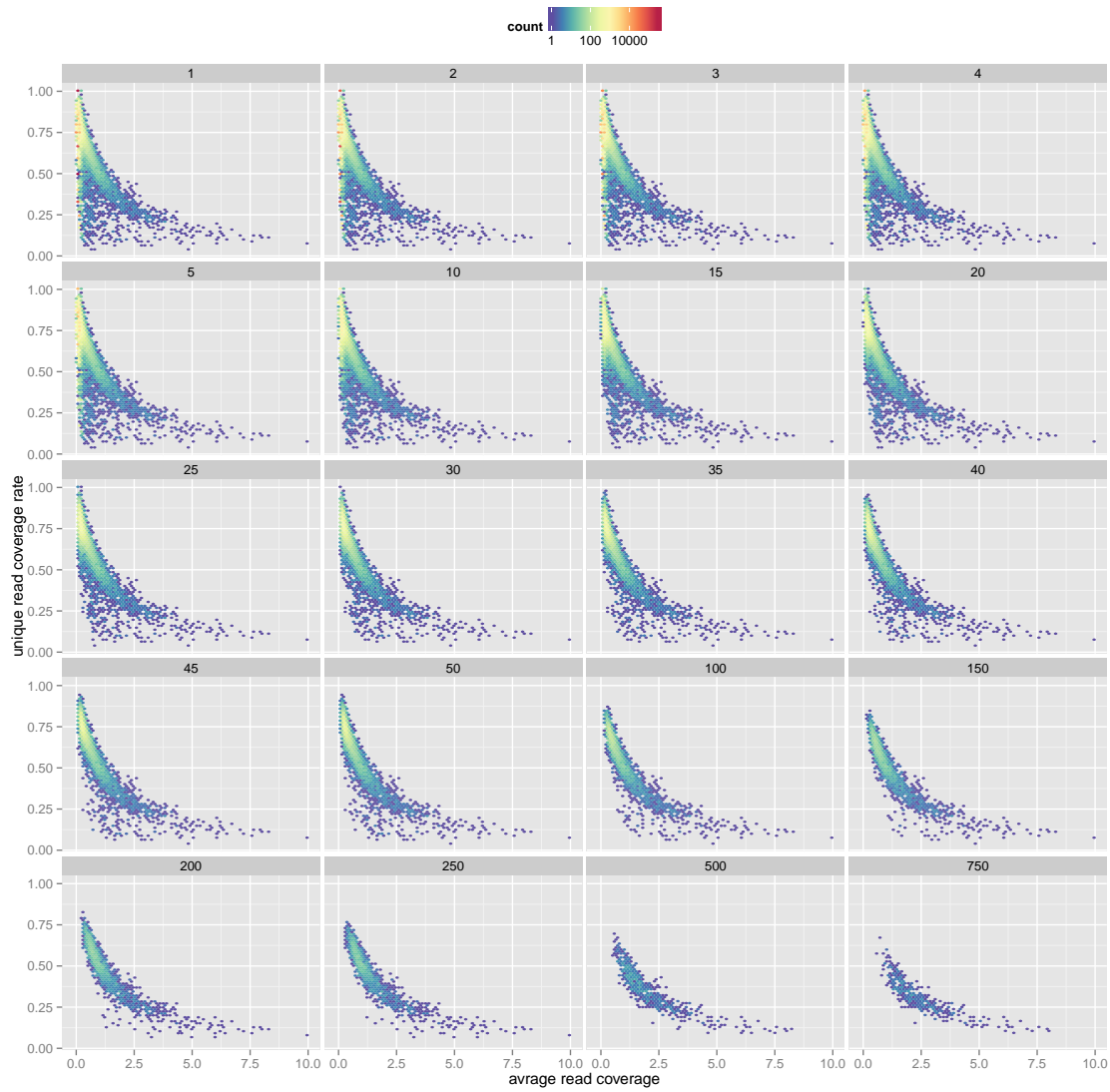
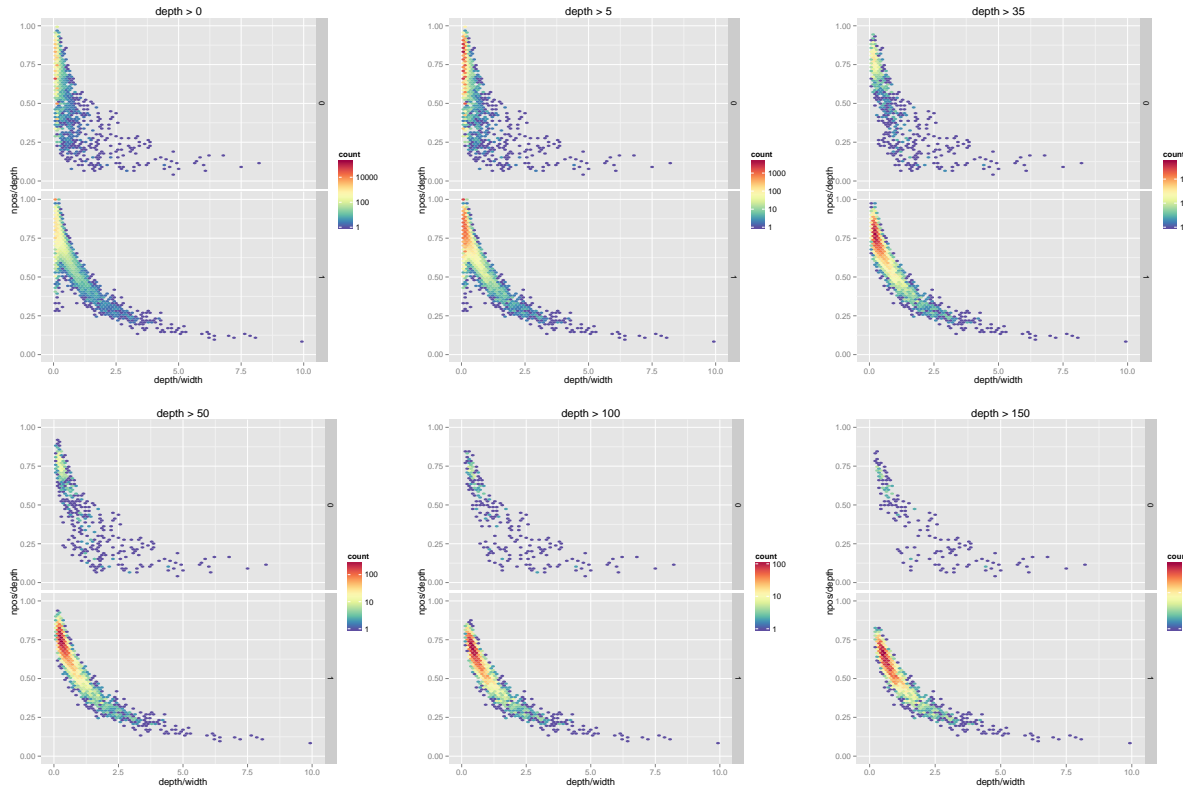


Figure 17: depth/width vs npos/depth plot. The panel title indicates that this plot shows only the regions with depth greater than the title

the idea here was to study $\text{npos} / \text{depth}$ $\text{npos} \leq \text{depth}$, thus $0 \leq \text{npos} / \text{depth} \leq 1$ 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

3.5 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 18 which show the density of all the sequencing procedures for the “both” label:

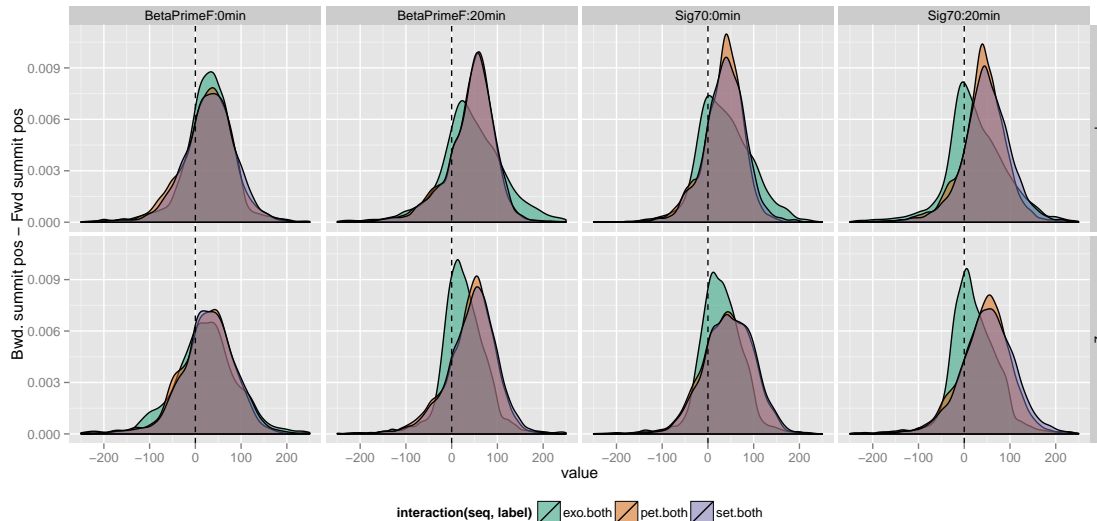


Figure 18: Density of difference in summit positions separated by sequencing for “both” label
it would be nice to update this plots with the new partitions, also there are chip-seq and chip-exo data for human samples (not only e.coli). right now this works as a placeholder

Figure 18 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 consuming 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 $-\log(\text{npos})$ (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

Rene,

I went over your document and have the following comments.

- 1. We need to generate raw data example raw data plots of “perfect”, “typical”, “noisy” ChIP-exo and their corresponding ChIP-seq (possibly from multiple organisms) and at the replicate level. This will set the stage for what we are trying to accomplish.*

human and mouse data can also be uploaded to genome browser - once we identify which regions to show through R plot, we may want browser shots.

- 2. When discussing ENCODE metrics, we should first make the case whether or not they are applicable. Maybe illustrate what we expect cross-correlation to reveal by a cartoon.*

I am a bit confused why we have cross-correlations maximized at zero. Shouldn't we expect them to be maximized at the read length? Does the fact that they are maximized at zero indicate something about the quality? Is this an artifact of smoothing on the cross-correlation plot? Would it improve if we got rid of the islands with severe strand imbalance?

- 3. For the npos/depth vs depth/width plots, once we have these once could have example peaks across different regions of this plot. This plot in the middle and then we highlight peaks with this metric at different coordinates*

- of this plot.*
4. *9a) reminds of ChIP vs input plots, which are the key plots I look at before starting to analyze a ChIP-seq dataset.
This plot definitely needs more characterization and documenting typical behavior across multiple datasets. We might want to look at this plot for peaks that are reproducible.
Actually, for any of these metrics, we can define gold standard peak set (reproducible) and justify the metrics that way.*
 5. *It would also be useful to see an IDR run of ChIP-exo datasets. This question will definitely come up but not sure if we have time for this.*
 6. *You would probably need a pipeline diagram from reads to peaks.*

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 Keleş. dpeak, high resolution identification of transcription factor binding sites from pet and set chip-seq data. *PIOS, Computational Biology*, 2013.