An analysis of quality control measures for Chip-exo

Rene Welch
Department of Statistics, University of Wisconsin-Madison
Madison, WI

October 2014

Contents

1 Introduction

The idea of this document is to summarize the different analysis made to asses the quality of different ChIP-exo data sets. Several analysis were performed to the ChIP exo data.

2 Summary of the data sets

Right now we have only analized data sets from the E. Coli genome.

Data set	Depth
edsn1310	3909669
edsn1311	960256
edsn1312	1875127
edsn1313	5153689
edsn1314	1940387
edsn1315	4900071
edsn1316	5157768
edsn1317	2247295
edsn1318	898641
edsn1319	1509554
edsn1320	4229574
edsn1321	6550805
edsn930	672217
edsn931	1454566
edsn932	1593964
edsn933	864714
edsn934	3405118
edsn935	1584532
edsn936	1822585
edsn937	1012936
edsn938	9898733

Table 1: Depth for ChIP-exo data sets

Data set	Depth
edsn1369	1294249
edsn1396	1134965
edsn1397	931758
edsn1398	917171
edsn1399	5602290
edsn1400	1188244
edsn1401	1313348
edsn1402	932106
edsn1403	3423076
edsn1416	1295243

Table 2: Depth for ChIP-seq PET data sets

3 PCR bottleneck coefficient

The PCR bottleneck coefficients is defined as a measure of library complexity, i.e. how skewed the distribution of read counts per location is towards 1 read per location. It is defined as:

$$\text{PBC} = \frac{N_1}{N_d} \tag{1}$$

where:

- N_1 is the number of genomic locations to which **exactly** one unique mapping read maps
- N_d is the number of genomic locations to which **at least** one unique mapping read maps, i.e. the number of non-redundant unique mapping reads

Since $N_1 \leq N_d$ then $0 \leq \text{PBC} \leq 1$. Finally ENCODE recommends:

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

Table 3: PBC classification

We grouped the data sets by the protocol it was used to generate it. We considered three protocols: ChIP-exo, ChIP-seq PET and ChIP-seq SET. We wanted to compare if there is an overall level across the same protocols used. Thus, for each data set, we calculated it's PBC and plotted all together by grouping them across protocols:

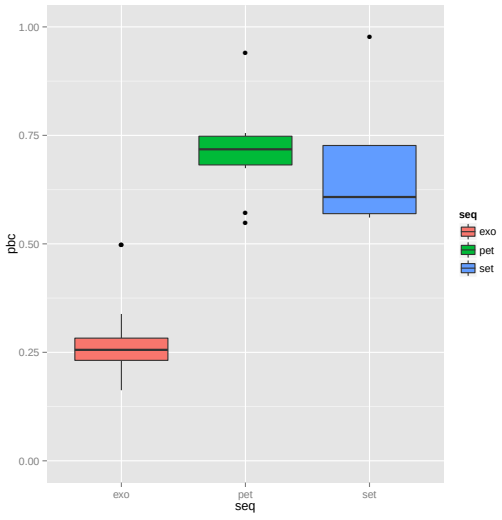

Figure 1: PCR bottleneck coefficient separated by protocol

In figure 1 we can see that PBC is higher for both ChIP-seq protocols. Also, we can see that the range of ChIP-Seq SET PBC is bigger than the other two.

4 Cross-correlation

A problem for ChIP-seq SET data sets is that the read fragment length is unknown. Cross-correlation is a method that gives an estimation of this quantity by the shift where the strand cross - correlation is maximized. For a given shift length δ, the strand cross-correlation is defined as:

$$\rho(\delta) = \sum_{c \in C} \frac{N_c}{N} P\left[n_c^+(x + \delta/2), n_c^-(x - \delta/2)\right] \tag{2}$$

where:

- N_c is the number of reads mapped to chromosome c and N is the sum of all the reads mapped to the genome
- $P(\mathbf{x}, \mathbf{y})$ is the Pearson's correlation between vectors $\mathbf{x}$ and $\mathbf{y}$
- $n_c^S(x)$ is the tag count vector of chromosome c and strand S, centered at position x

For this part of the analysis we calculated the cross - correlation curve using two different pipelines: *spp* and *ChIPQC* and compared both curves visually.

Figure 2: Strand cross-correlation for a typical ChIP-seq SET data set

Figure 3: Strand cross-correlation for a typical ChIP-seq PET data set

In both figures 2 and 3 we can see typical cross correlation curves for both ChIP-seq SET and PET data sets. In this case, there are no phantom peaks or any weird effect in the curves. So, when estimating the fragment length using the shift where each curve is maximized we can see that the position doesn't variate a lot between methods (it may be a matter of some bp). Also, it's is worth noticing that the range of both curves is on a decimal scale. However, when we plot one of the ChIP-exo datasets, we can see that its curves are both in a centesimal scale:

Figure 4: Strand cross-correlation for a typical ChIP-exo data set

That means that both strand are approximately uncorrelated. Thus, this method may not work to estimate the read fragment length for a ChIP-exo data set. Also, we can think that this low values may be an effect of having sparse tag count vectors (which may be happening because of the enzyme diggesting all positions from the 5' end until reaching the TF.

5 Forward strand ratio density

For this analysis, we divided the e.coli genome into bins of a fixed length b and for each bin we counted the number of reads that overlap with it. Finally we calculated the forward strand ratio as:

$$\text{ratio} = \frac{f+1}{f+r+2} \tag{3}$$

where f (r) is the number of reads in the forward (backward) strand that overlap with the bin. Ideally, we want to observe a uni-model density with small values around the extremes (0 and 1). A typical ChIP-seq density looks like:

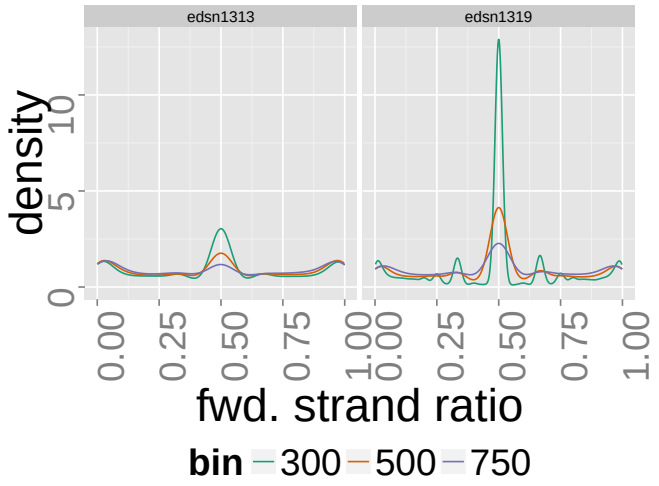

Figure 5: Densities of forward strand ratio, the ip is β and the condition is rif-20 min (both replicates)

Clealy for both ChIP-seq cases the densities are well behaved. Almost all mass from the densities is concentrated around 0.5. However, for the ChIP-exo case the tails seem to be heavier, which means that there are a lot of bins for which reads of only one strand are being mapped. We are assuming that this is due to some sort of diggestion bias.

In figure 5, we can see two replicates of the same data set (ip=β and rif=20 min), where the forward strand density behaves in two different ways, which suggests that this bias is sample dependent.

For comparison purposes, a common case of a ChIP-seq forward strand density is shown (the same effect is observed in the rest of the samples):

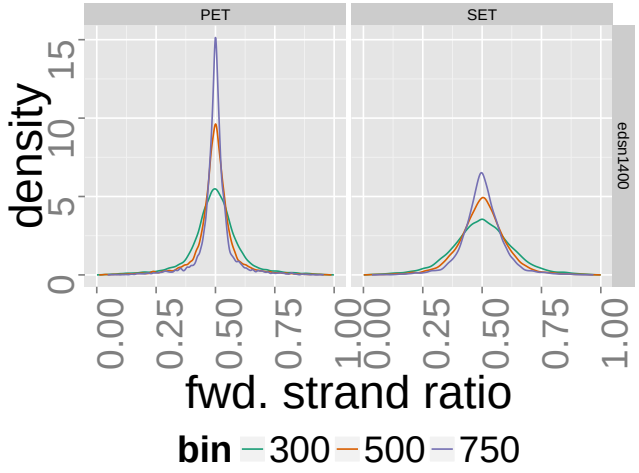

Figure 6: Forward strand ratio for σ_{70} with rif=0 min (both ChIP-seq case)

In figure 6 we can see that there is more probability mass around 0.5 for the PET sample than for the SET sample.

5.1 Forward strand ratio density estimation conditional to higher count bins

After comparing figures 5 versus 6, it is clear that the tails from the 5 are more heavier than the tails of any ChIP-seq case.

Usually, we are more interested in the densities with the highest counts, since those are the regions where some signal could be detected. Thus, we estimated the densities considering only the bins with counts greater than certain quantile.

6 Examples

A Sampling SET reads from PET reads