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

Outline

ChIP-exo procedure

ChIP-Seq QC measures

Comparison of ChIP-exo and ChIP-seq

The ChIP-exo QC pipeline

Comparison with ChIP-Seq using dPeak

Conclusions and future work

ChIP-exo procedure

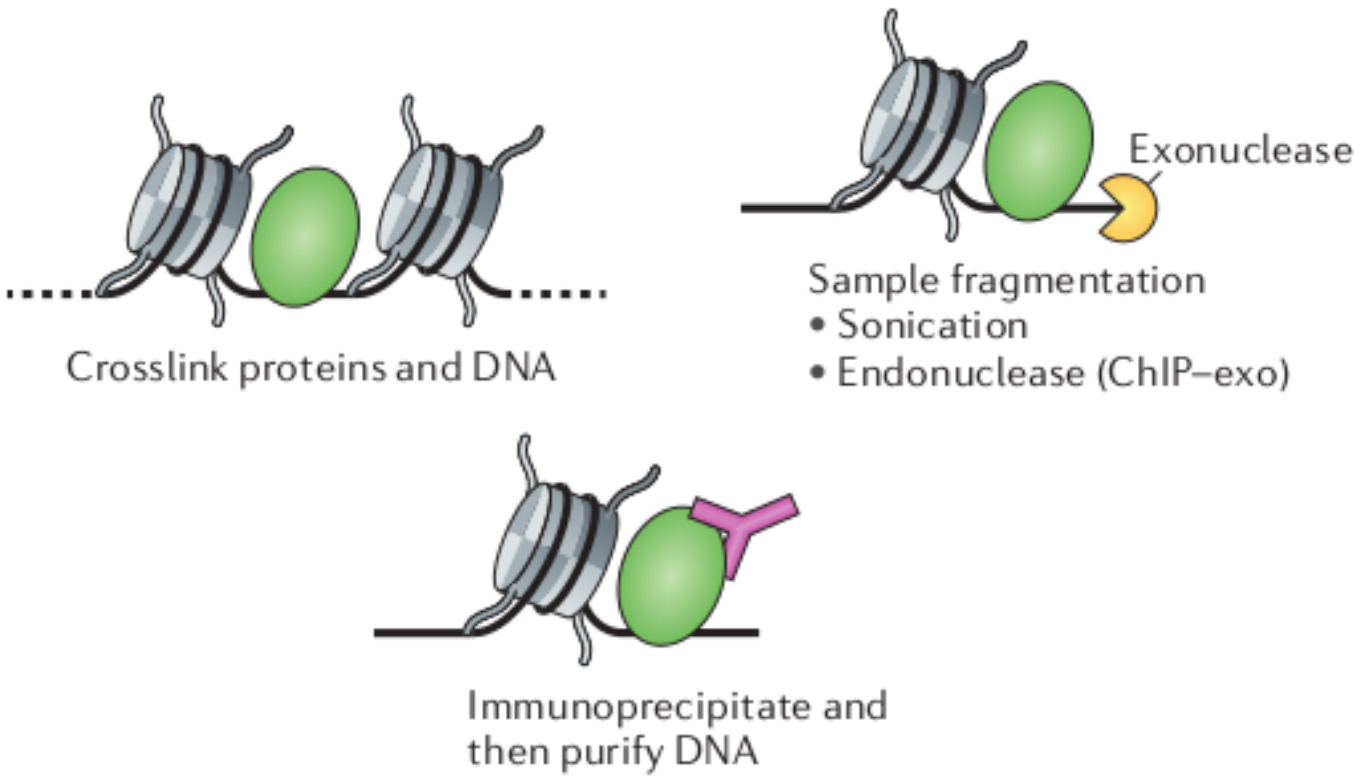


Figure: ChIP-exo procedure, the diagram is from Furey, 2012 [2]

ChIP-Seq QC measures

QC measure	Definition
Nr. reads	Self-explanatory. The higher the better...
PCR bottleneck Coeff.	Ratio of number of pos. to which EXACTLY one read maps and number of pos. to which AT LEAST one read maps
Standarized Std. Dev.	Normalized Std. Deviation of the sequencing coverage
Strand Cross-Corr.	$y(\delta) = \sum_c w_c r \left[n_c^+ \left(x + \frac{\delta}{2} \right), n_c^- \left(x - \frac{\delta}{2} \right) \right]$
Normalized SCC	Ratio of max value of SCC and min value of SCC*

where n_c^S is the coverage for chromosome c and strand S. r is the Pearson correlation and w_c is the proportion of reads in the experiment for chromosome c

ChIP-Seq QC measures

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

Comparison of ChIP-exo and ChIP-seq I

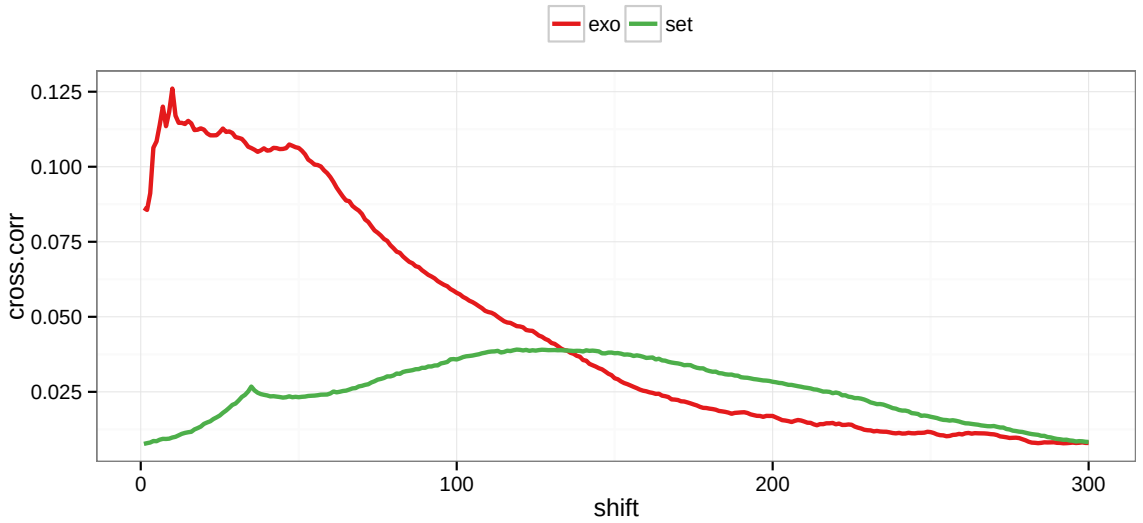
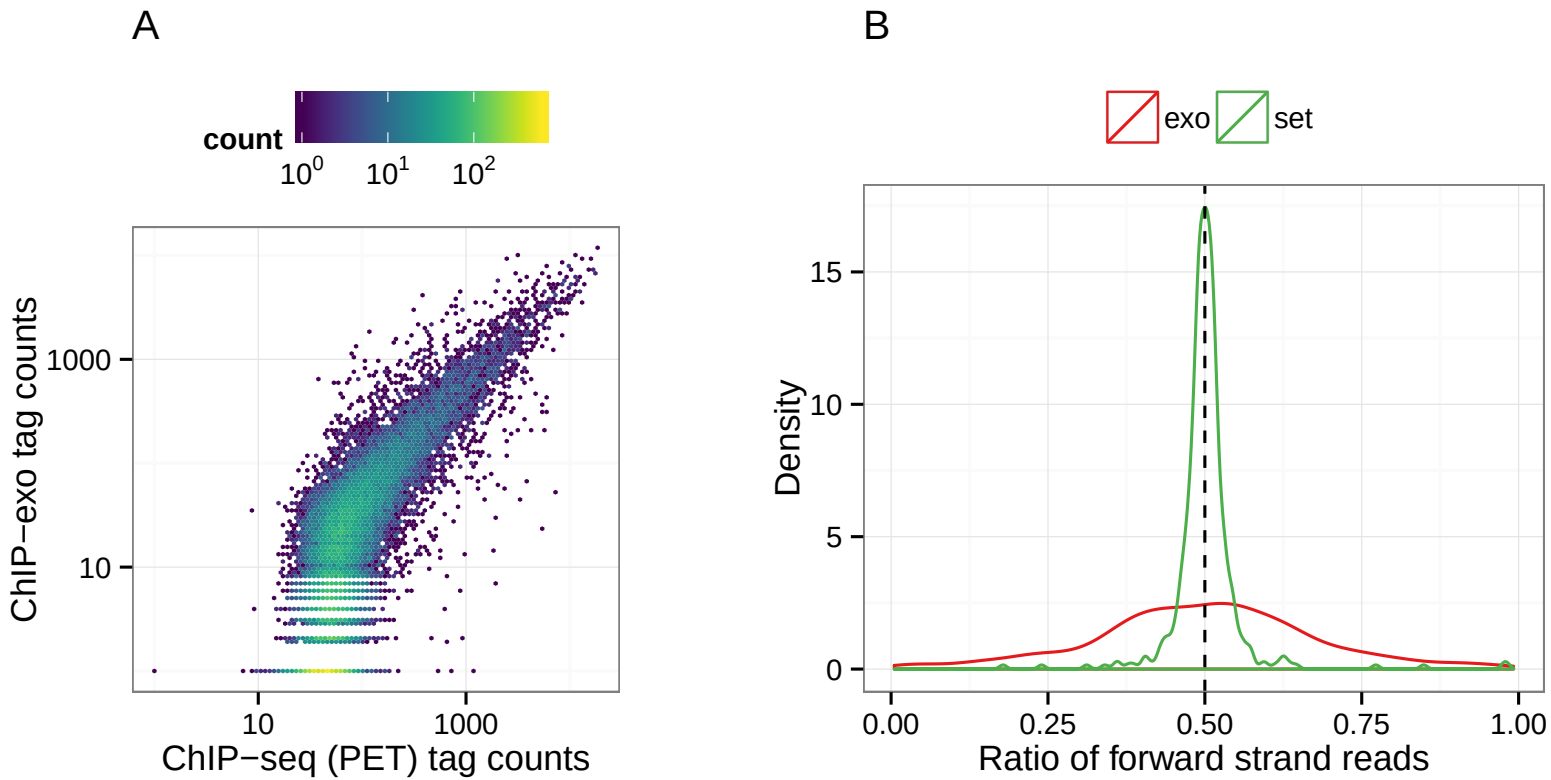


Figure: SCC for CTCF factor in HeLa cell line for ChIP-exo and SET-ChIP-Seq

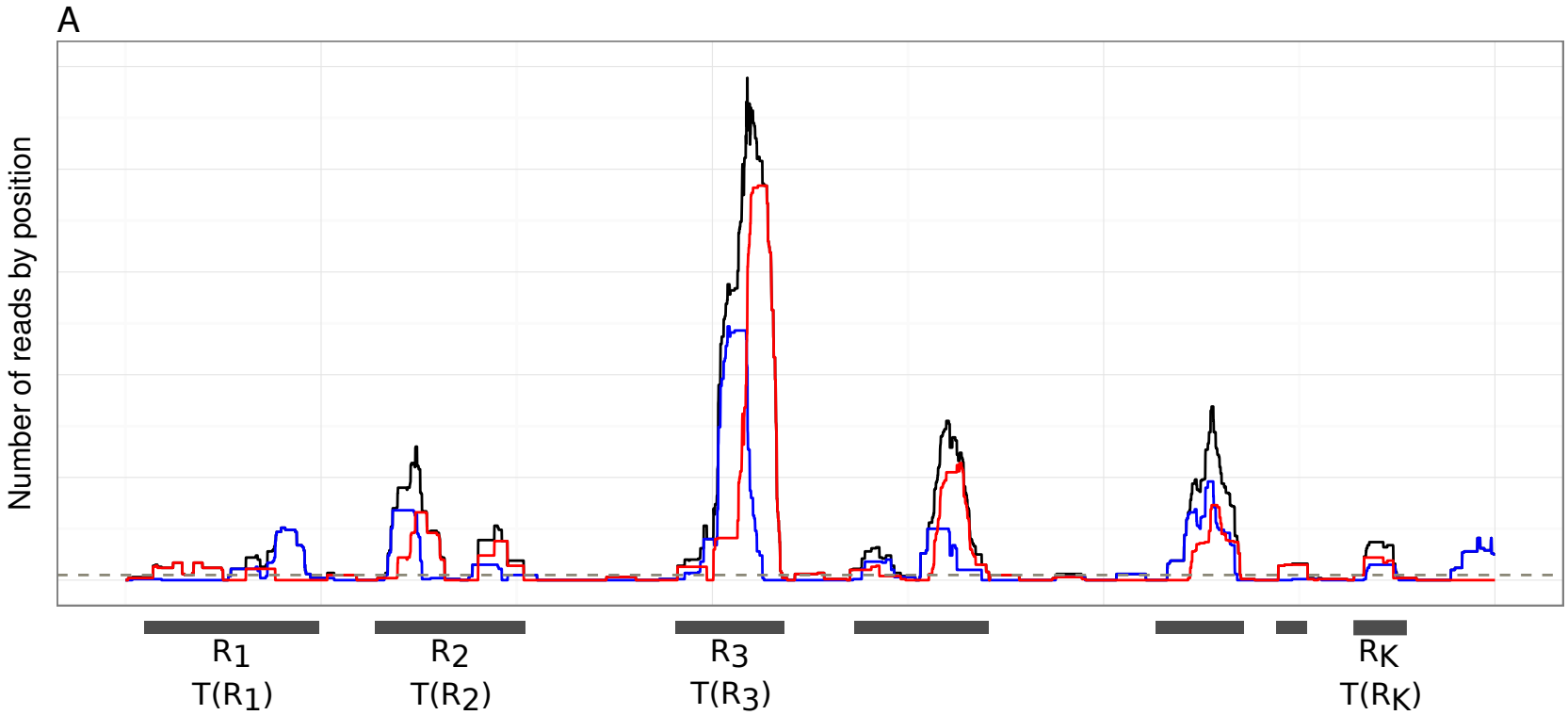
- There is a *"phantom peak"* at read length.
- In ChIP-Seq SCC is maximized at the unobserved fragment length.
- In ChIP-exo, the *"phantom peak"* and the frag. length summit are confounded.

Comparison of ChIP-exo and ChIP-Seq II



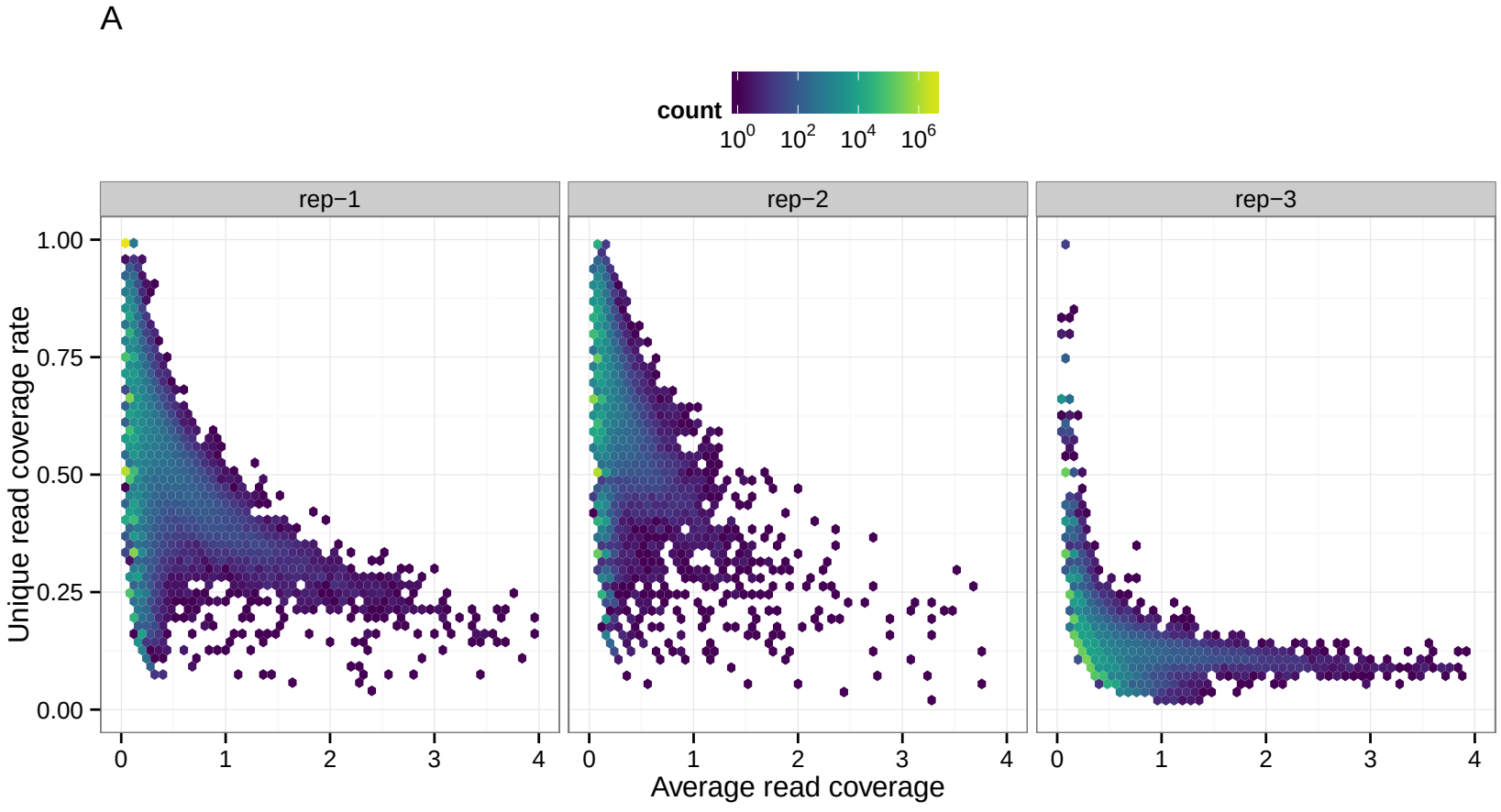
- A shows that high dnesity regions are similar between ChIP-Seq and ChIP-exo but background regions are not.
- The peak-pair assumption doesn't hold in ChIP-exo data, some regions show strand-imbalance

The ChIP-exo QC pipeline

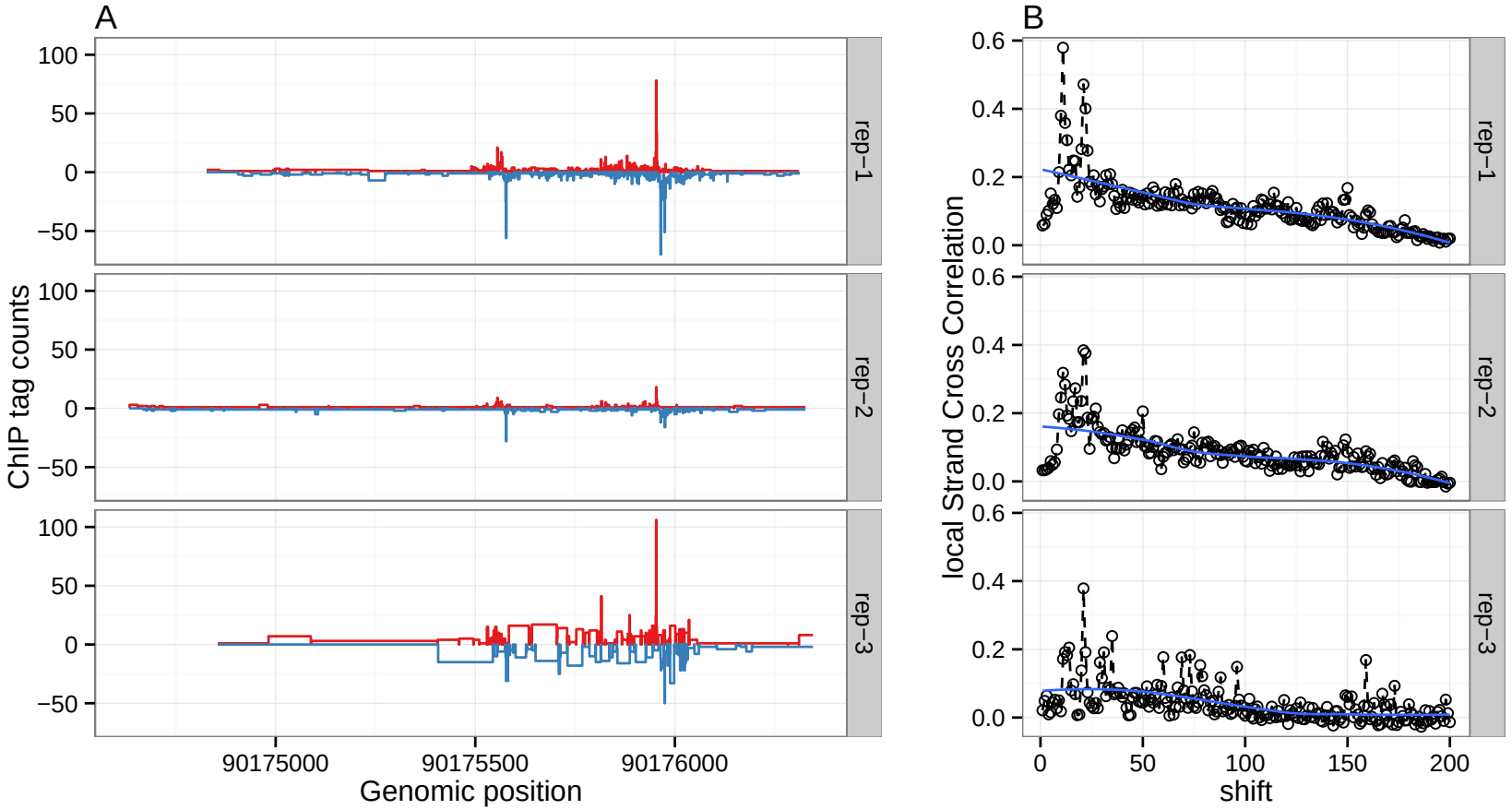


- ARC $= \frac{\text{Nr. of reads in the region}}{\text{Width of the region}}$
- URCR $= \frac{\text{Nr. of reads mapped to only one position in the region}}{\text{Nr. of reads in the region}}$
- local-NSC
- FSR $= \frac{\text{Nr. of fwd. strand reads in region}}{\text{Nr. of reads in region}}$

Library complexity and enrichment

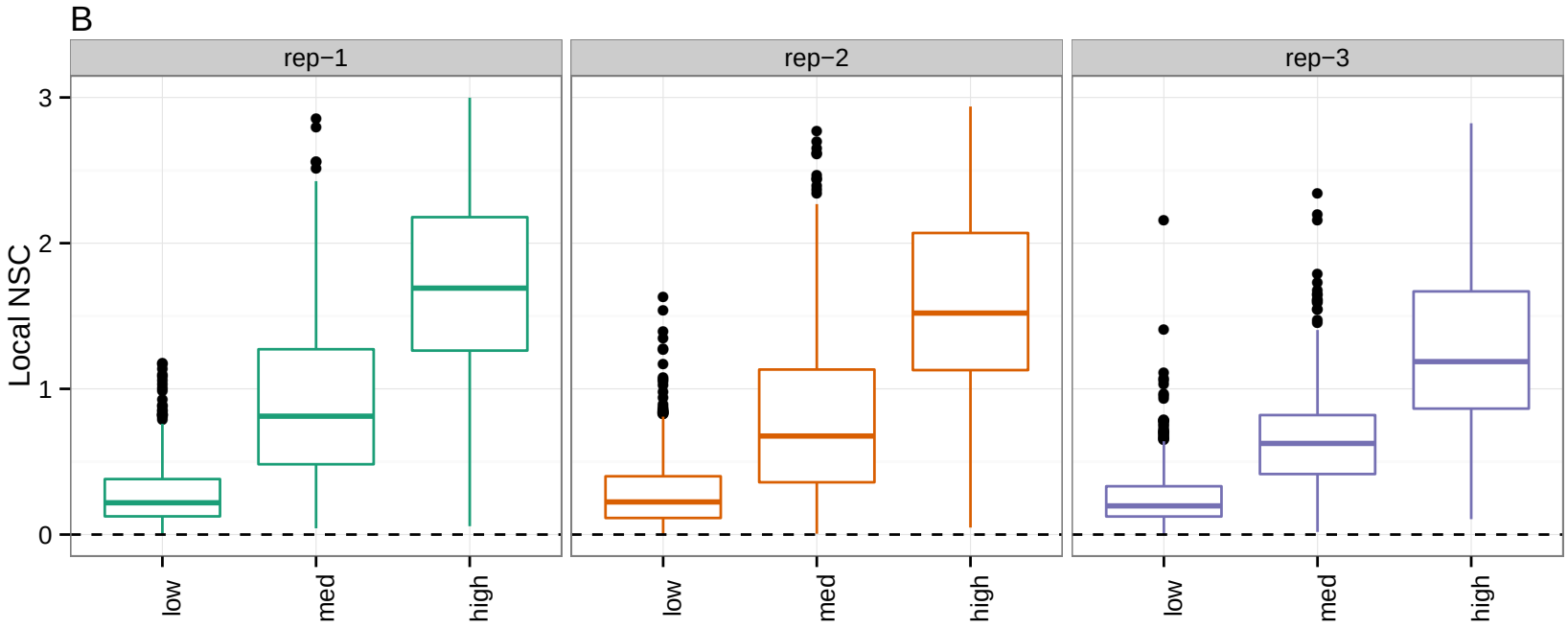


Library complexity and enrichment



$$y(\delta) = f(x_\delta) + \epsilon_\delta \quad \text{local-NSC} = \frac{max_{x_\delta} \hat{f}(x_\delta)}{\hat{\sigma}_f}$$

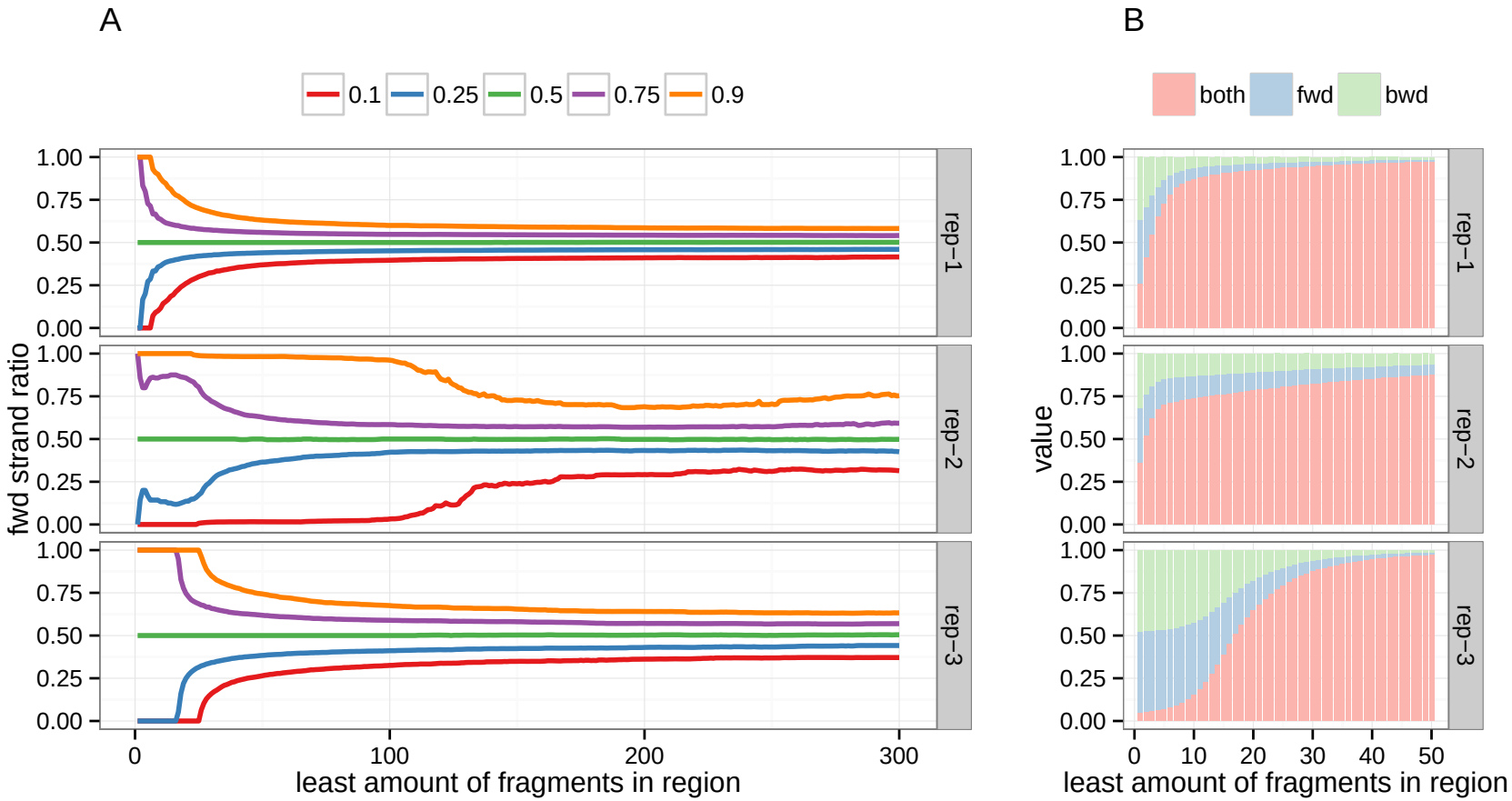
Library complexity and enrichment



where:

- high - regions with nr. of unique positions > 100
- med - regions with nr. of unique positions in $(50, 100)$
- low - regions with nr. of unique positions in $(20, 100)$

Strand imbalance

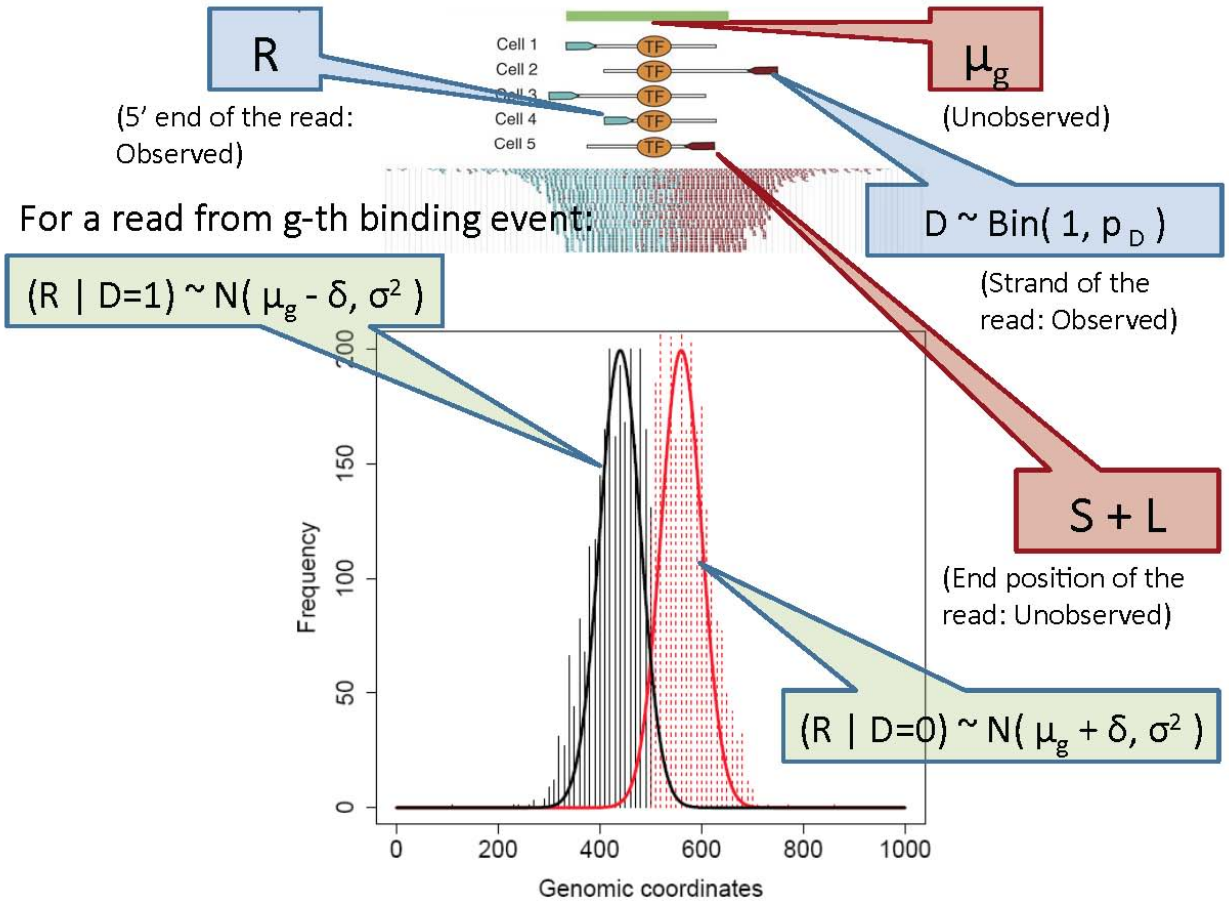


dPeak model for SET case

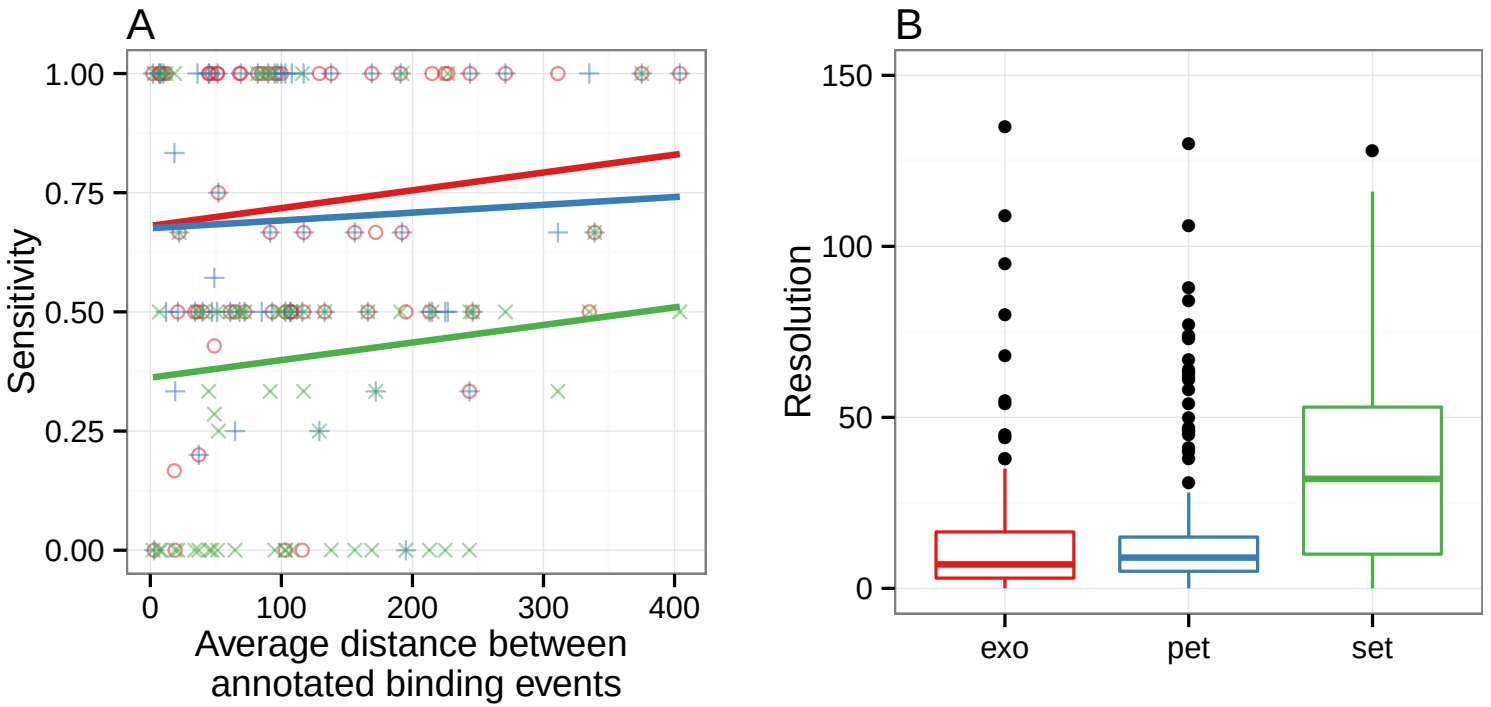
We consider a region with n reads and m positions, for the i-th read:

- $Z_i \sim \text{Multi}(\pi_0, \pi_1, \cdots, \pi_{g^*})$
- $D_i \sim \text{Ber}(p_D)$
  - The read is in the forward strand ($D_i = 1$):
    - The reads belongs to the background: $R_i | Z_i = 0, D_i = 1 \sim \text{Unif}(1 - \beta + 1, m)$
    - The read belong to the g-th binding event: $R_i | Z_i = g, D_i = 1 \sim \text{N}(\mu_g - \delta, \sigma^2)$
  - The read is in the backward strand ($D_i = 0$):
    - The reads belongs to the background: $R_i | Z_i = 0, D_i = 0 \sim \text{Unif}(1, m + \beta - 1)$
    - The read belong to the g-th binding event: $R_i | Z_i = g, D_i = 0 \sim \text{N}(\mu_g + \delta, \sigma^2)$

dPeak model for SET case

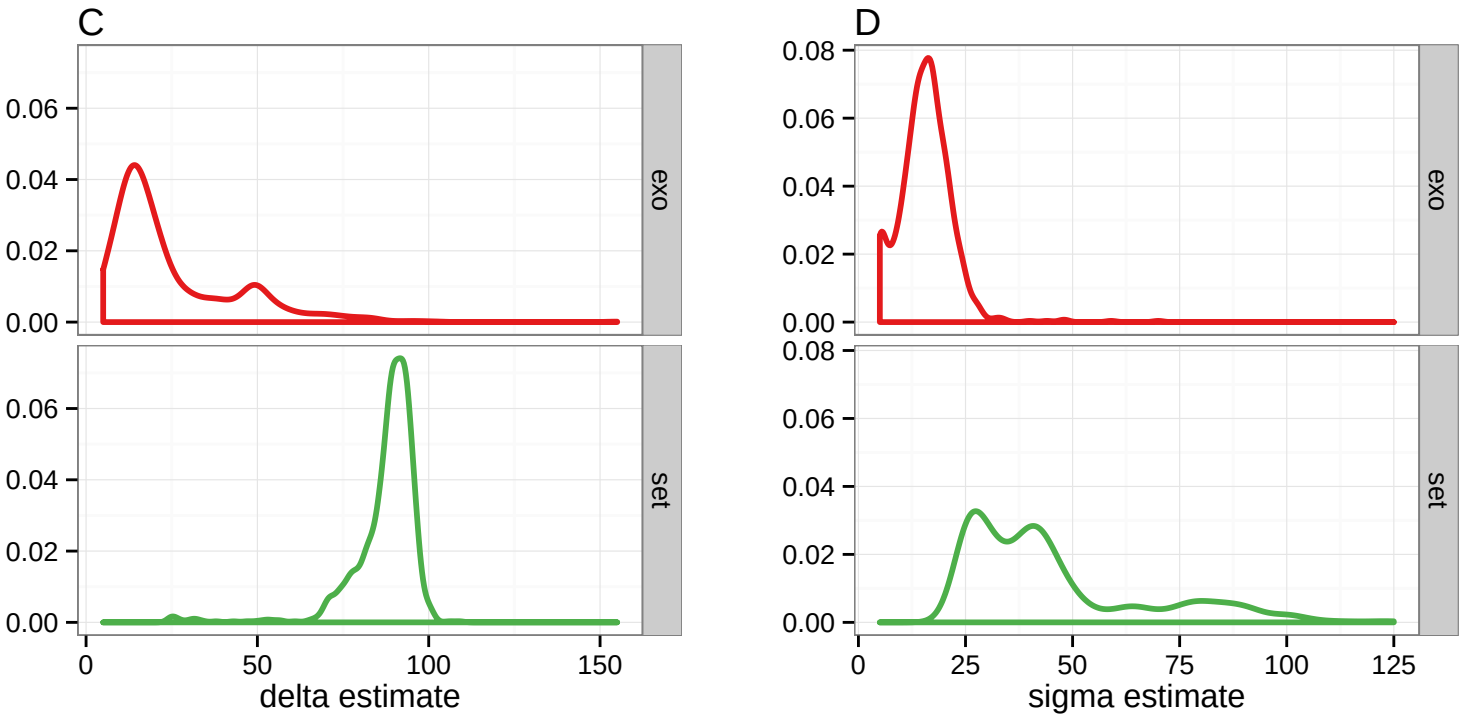


Comparison with ChIP-Seq using dPeak



- Sensitivity is the defined as the proportion of identified peaks (regulonDoB [4] is used as gold-stantard)
- Resolution is defined as the min. absolute distance of a regulonDB annotation to an est. binding location.

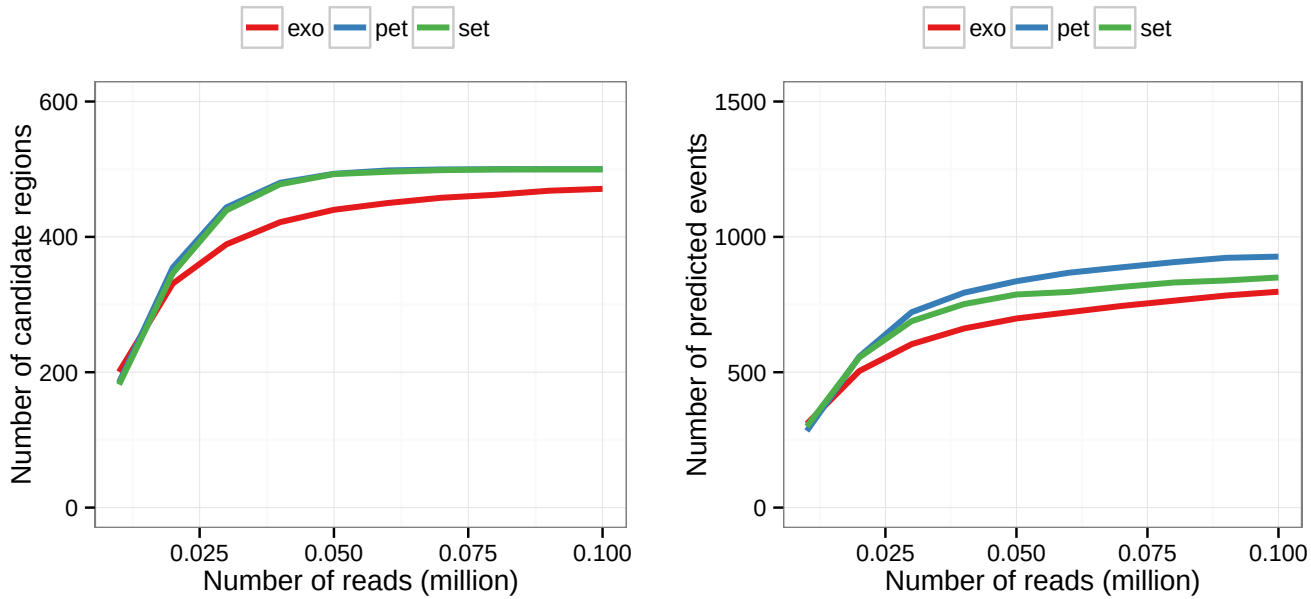
Comparison with ChIP-Seq using dPeak



- δ measures the average distance of reads to their respective binding sites
- σ measures the dispersion of reads around their respective binding sites

ChIP-Seq comparison at fixed depth

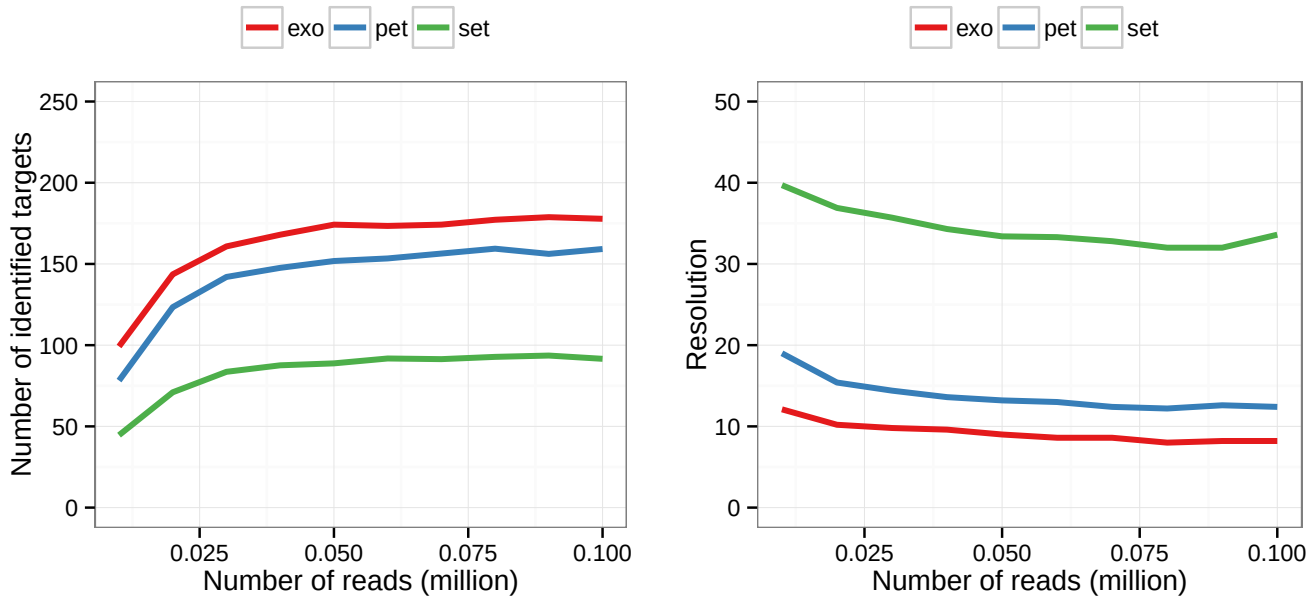
We sampled n fragment reads of each dataset ($2n$ for PET ChIP-Seq), and applied the MOSAiCS / dPeak pipeline:



- ChIP-exo and PET ChIP-Seq are comparable and outperform SET ChIP-Seq

ChIP-Seq comparison at fixed depth

We sampled n fragment reads of each dataset ($2n$ for PET ChIP-Seq), and applied the MOSAiCS / dPeak pipeline:



Conclusions

- Our pipeline is capable of assessing the balance between sample enrichment and library complexity.
- We shown that the "peak-pair" assumption doesn't hold well in practice, and implemented a visualization capable of detecting strand imbalance.
- We updated dPeak, which makes a striking balance in sensitivity, specificity and spatial resolution.
- ChIP-exo and PET ChIP-Seq are comparable in resolution and sensitivity, and both outperform SET ChIP-Seq.
- We showed that with a fixed number of reads, ChIP-exo outperforms PET and SET ChIP-Seq.

Future work

- In the paper, we showed that there is a relationship between ChIP-exo tag counts and both mappability and GC content scores. We want to add a QC measure to the pipeline based on them.
- We want to assess if ChIP-Nexus library complexity is actually higher than ChIP-exo's by using the local-NSC.
- We have been studying E. Coli's transcription initiation complexes with PET ChIP-Seq, being able to label regions as open or closed complexes. We want to improve this analysis by using ChIP-exo data, and hopefully detecting intermediate step between this two states.
- Find a optimal strategy for labelling enhancer out of a predetermined list of regions in the genome by the use of active learning techniques.

Software

- **dPeak**: We updated the initialization strategy. The latest version is currently available from `http://dongjunchung.github.io/dpeak/`.
- **ChIPexoQual**: This package contains the QC pipeline for ChIP-exo. The last version is available in `https://github.com/welch16/ChIPexoQual`.
- **Segvis**: The goal of this package is to visualize genomic regions by using aligned reads. The latest version is available in `https://github.com/keleslab/Segvis`.
- **ChIPUtils**: This package attempts to gather the most commonly used ChIP-Seq QC.The latest available version is in `https://github.com/welch16/ChIPUtils`.

Thank you very much!

References

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.

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

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

Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz-Rascado, Jair s. García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, Gerardo Salgado-Osorio, Shirley Alquicira-Hernández, Kevin Alquicira-Hernández, Porrón-Sotelo Liliana López-Fuentes, Alejandra, Araceli M. Huerta, César Bonavides-Martínez, Yalbi I. Balderas-Martínez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Verónica Jiménez-Jacinto, Leticia Vega-Alvarado, Victor del Moral-Chávez, Alfredo Hernández-Alvarez, Enrique Morett, and Julio Collado-Vides.
Regulondb v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more.