Data exploration, quality control and statistical analysis of ChIP-exo experiments

Rene Welch

Preliminary Examination

Department of Statistics
University of Wisconsin - Madison

December 1st, 2015

Outline

ChIP-exo procedure

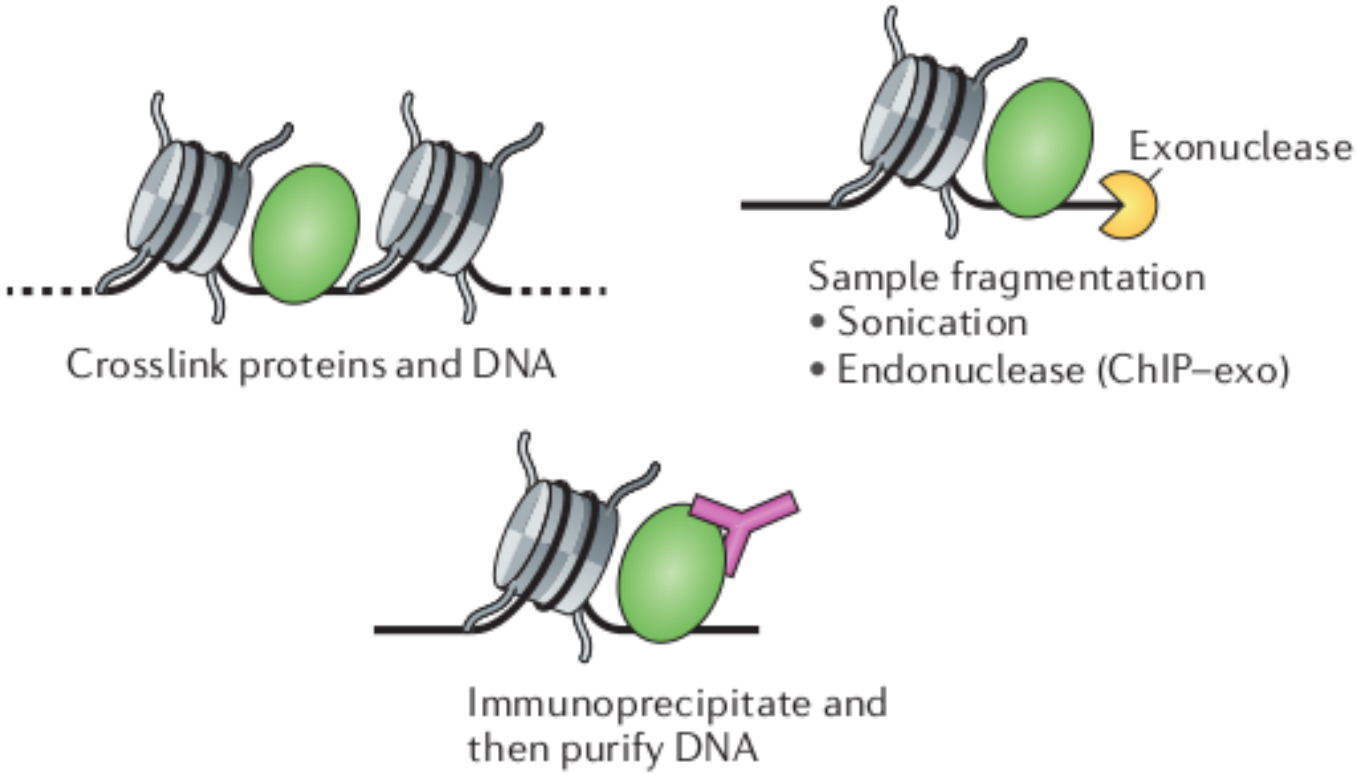


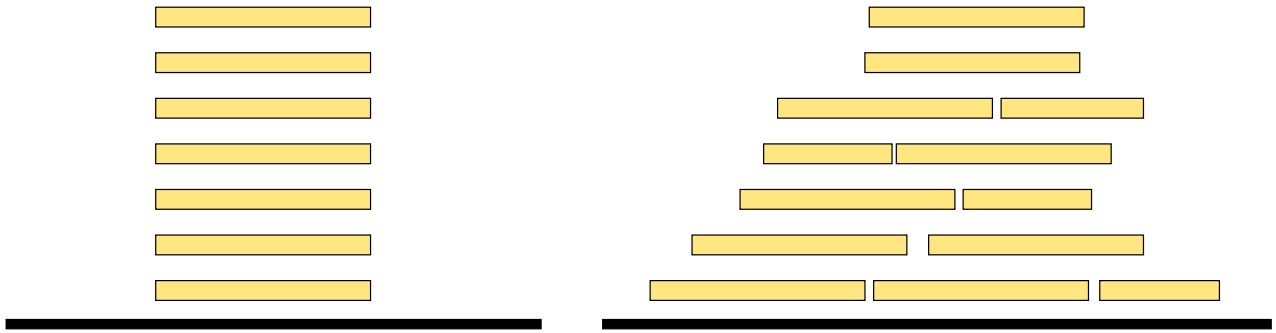
Figure: ChIP-exo procedure, Furey, 2012 [2]

ChIP-Seq QC measures

- Number of reads. Self-explanatory, the higher the better

ChIP-Seq QC measures

- Number of reads.
- PCR bottleneck Coeff. Ratio of number of pos. to which EXACTLY one reads maps and number of pos. to which AT LEAST one reads maps



ChIP-Seq QC measures

- Number of reads.
- PCR bottleneck Coeff.
- Strand Cross-Correlation.

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

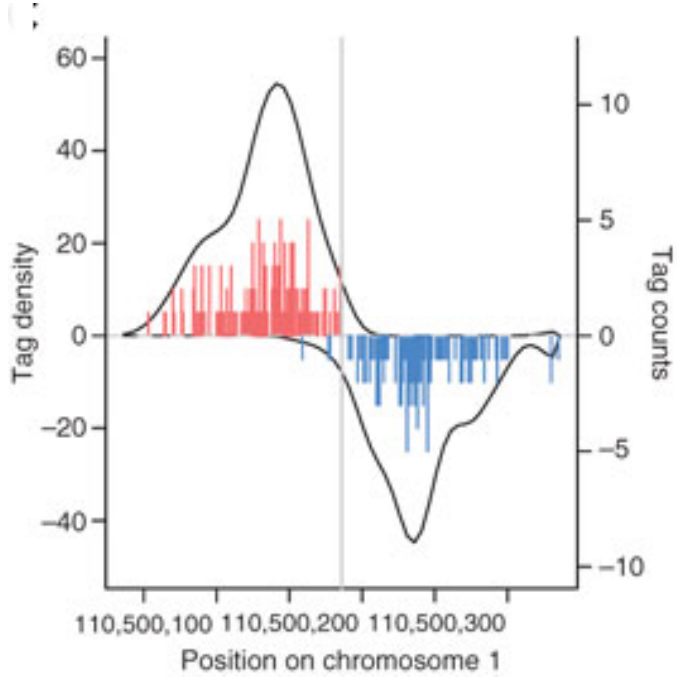


Figure: SCC explanation. Kharchenko et al., 2008 [3]

ChIP-Seq QC measures

- Number of reads.
- PCR bottleneck Coeff.
- Strand Cross-Correlation.

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

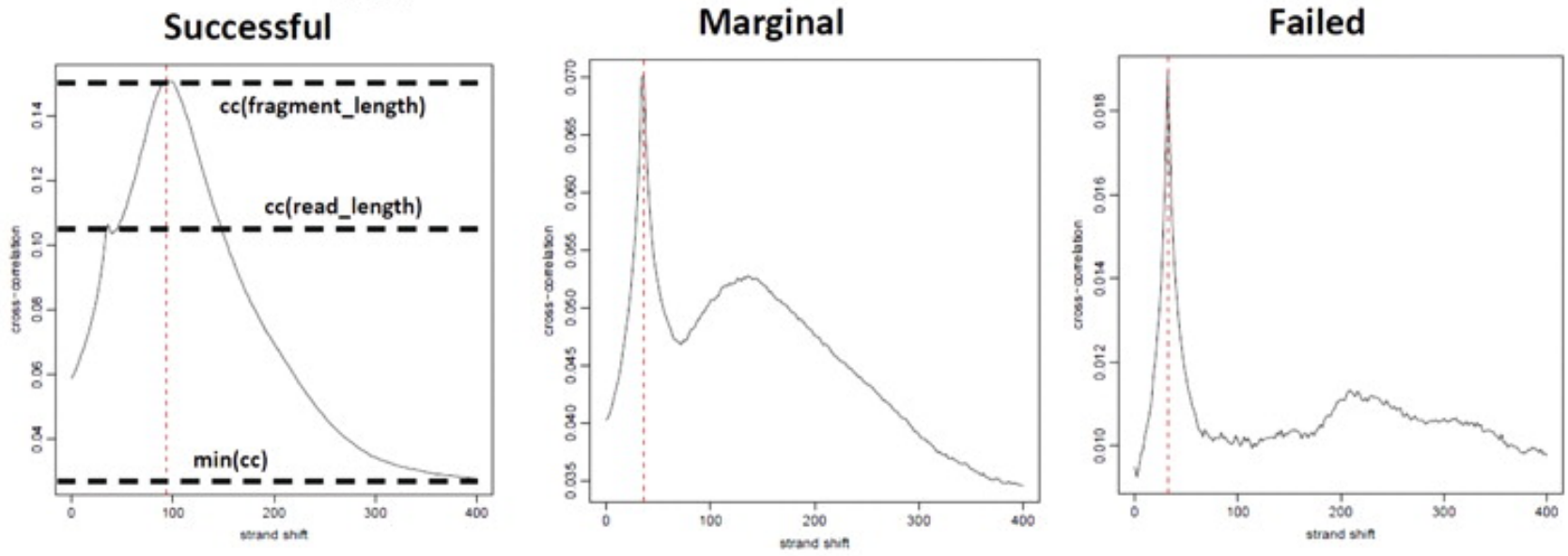


Figure: SCC as QC. Landt et al., 2012 [4]

ChIP-Seq QC measures

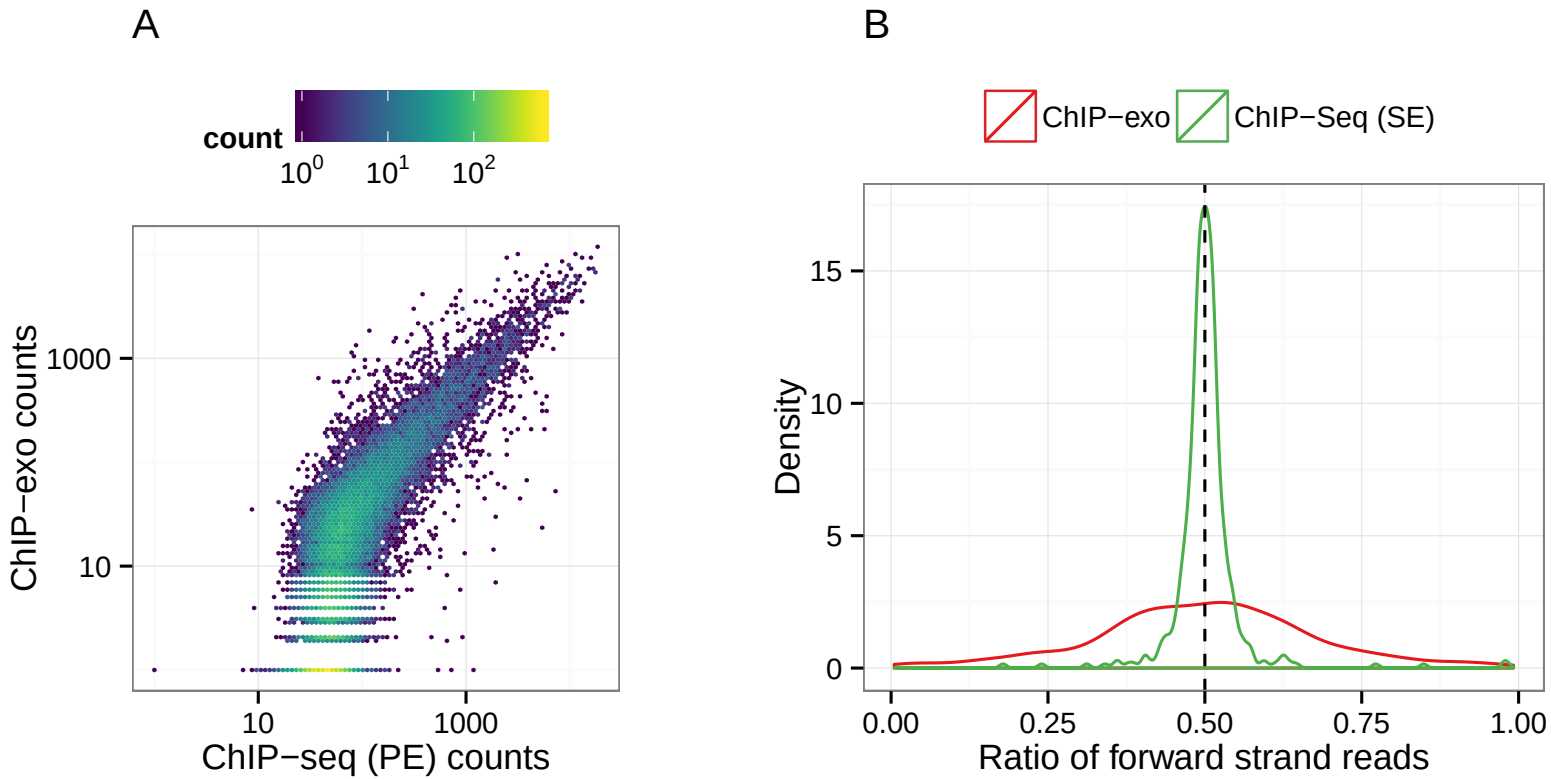
- Number of reads.
- PCR bottleneck Coeff.
- Strand Cross-Correlation.
- Normalized Strand Cross-Correlation. Ratio between the SCC when the shift is the fragment length and min. value of the SCC.

Evaluation of ChIP-Seq QC Metrics for ChIP-exo

IP	Organism	Condition	Rep.	Nr. reads	PBC	NSC	Source
σ^{70}	E.Coli	Rif-0min	1	960,256	0.2823	10.29	Courtesy of Prof. Landick's lab
σ^{70}	E.Coli	Rif-0min	2	2,247,295	0.2656	25.08	
σ^{70}	E.Coli	Rif-20min	1	1,940,387	0.2698	17.69	
σ^{70}	E.Coli	Rif-20min	2	4,229,574	0.2153	14.11	
FoxA1	Mouse	-	1	22,210,461	0.6562	21.452	From Serandour et al., 2013 [7]
FoxA1	Mouse	-	2	22,307,557	0.7996	60.661	
FoxA1	Mouse	-	3	22,421,729	0.1068	72.312	
ER	Human	-	1	9,289,835	0.8082	19.843	From Serandour et al., 2013 [7]
ER	Human	-	2	11,041,833	0.8024	21.422	
ER	Human	-	3	12,464,836	0.8203	19.699	
CTCF	Human	-	1	48,478,450	0.4579	15.977	From Rhee and Pugh 2011, [5]

- For PBC (human and mouse): 0 - 0.5 severe bottlenecking , 0.5 - 0.8 moderate bottlenecking, 0.8 - 0.9 mild bottlenecking, 0.9 - 1 no bottlenecking.
- For NSC (human and mouse): < 1.1 is relatively low.

Comparison of ChIP-exo and ChIP-Seq



- A shows that high density regions are similar between ChIP-Seq and ChIP-exo but background regions are not.
- The peak-pair assumption does hold in ChIP-exo data, but not locally since some regions show strand-imbalance.

Comparison of ChIP-exo and ChIP-seq

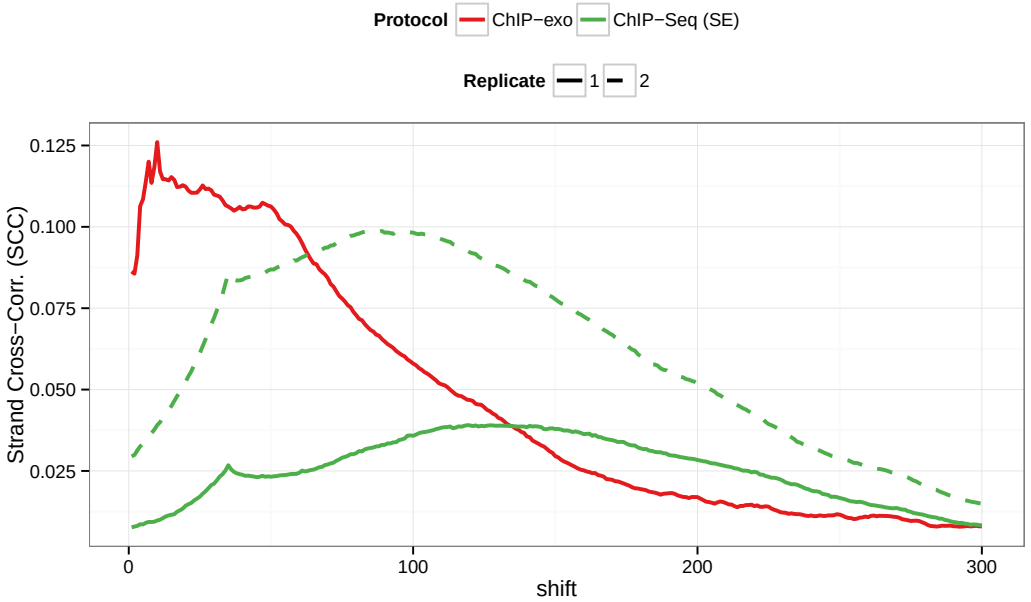
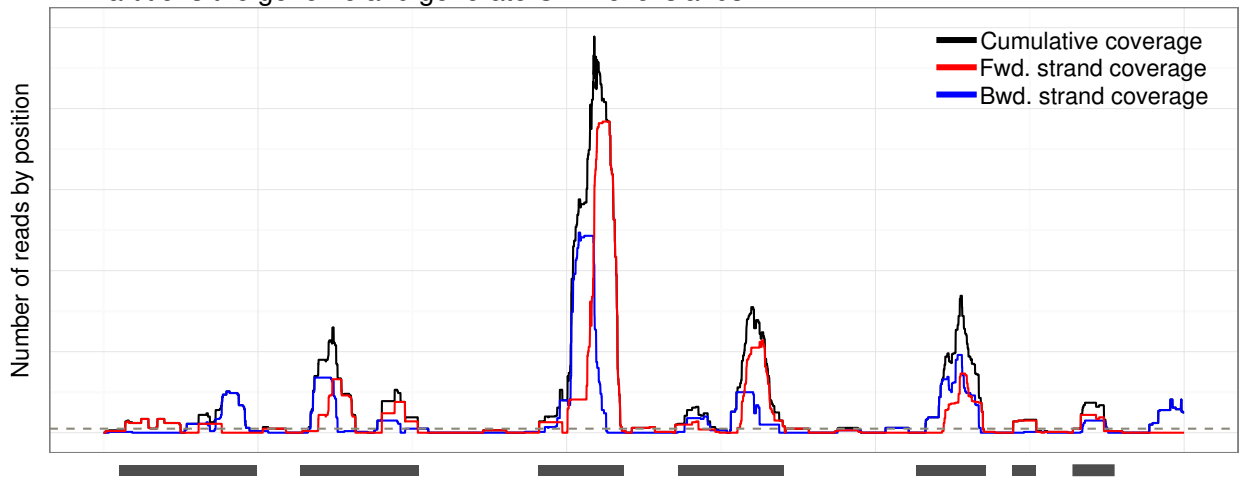


Figure: SCC for CTCF factor in HeLa cell line for ChIP-exo and SET-ChIP-Seq.

- PBC for rep1 is 0.56
- PBC for rep2 is 0.94

- 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 fragment length summit are confounded.

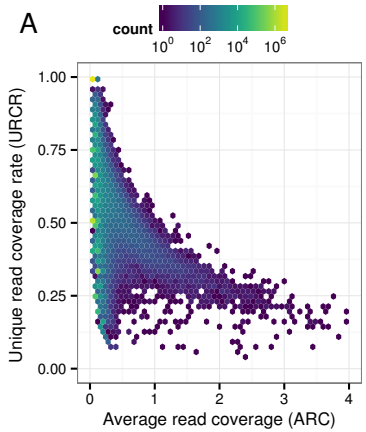
1 - Partitions the genome and generate ChIP-exo islands.

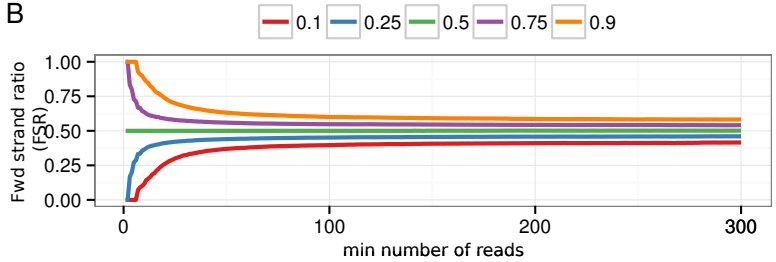


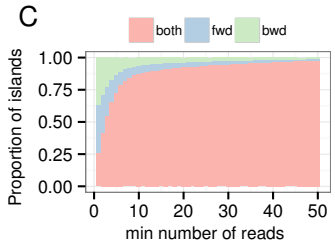
2 - Calculate a vector of summary statistics for each island

R_1 $T(R_1)$ R_2 $T(R_2)$ R_3 $T(R_3)$ R_K $T(R_K)$

3 - Visualize all islands together:





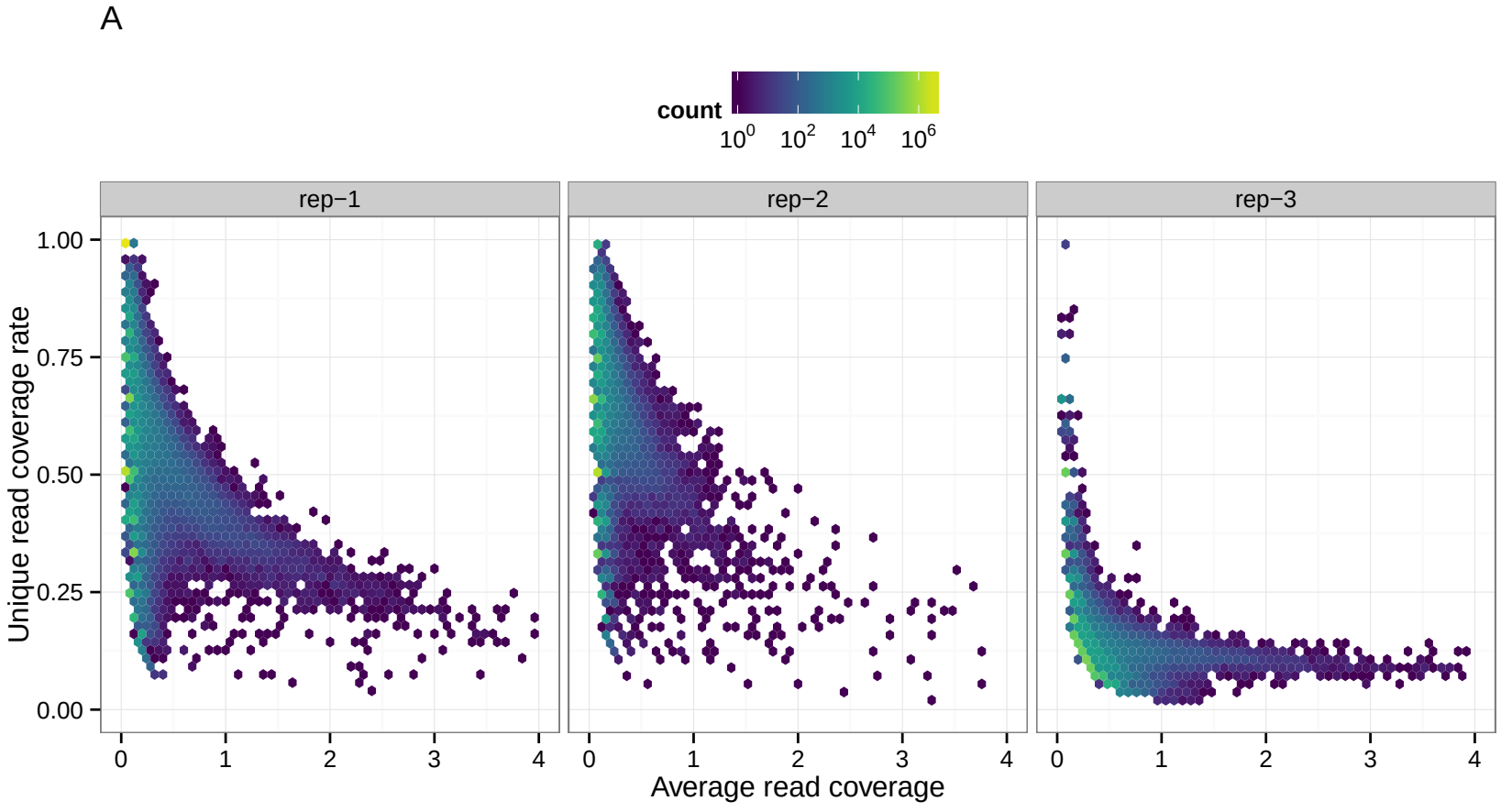


3A - ARC VS URCR - This plot gives a global view of the balance between library complexity and enrichment. There are two arms, one with low ARC, which corresponds to regions formed by few aligned positions, and the other were the URCR decreases as the ARC increases.

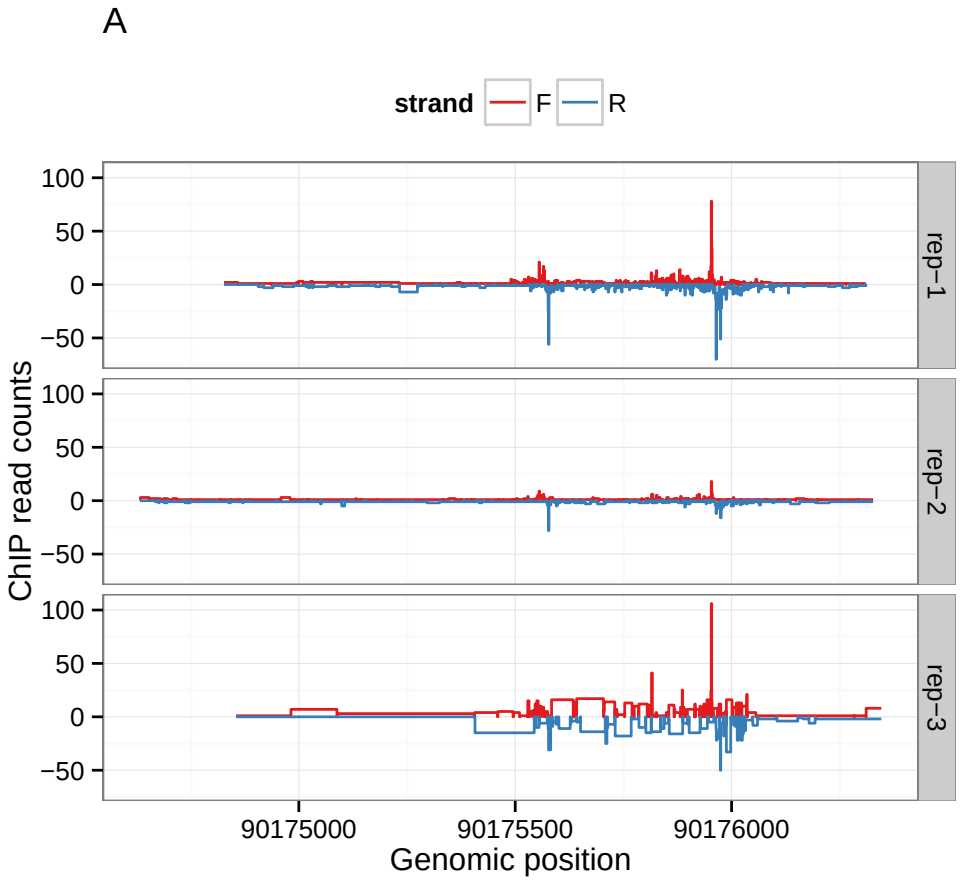
3B - Min depth VS FSR - This plot gives an idea of how quickly the distribution of the FSR approximates the median. In a good sample, the median is approx. 0.5, and the other quantiles reach that value quickly

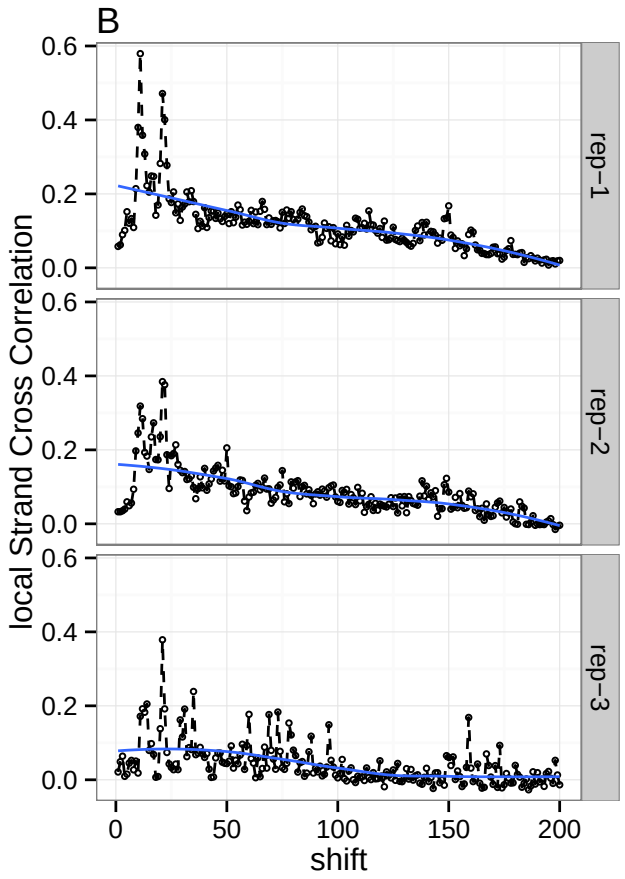
3C - Min depth VS Proportion of Islands - This plot gives a more detailled view of the FSR. For islands with a low amount of reads is common to contain only reads from one strand. Hence, it compares the percentage of islands that contain at least one read of each strand vs the regions that consist of only reads

Library complexity and enrichment



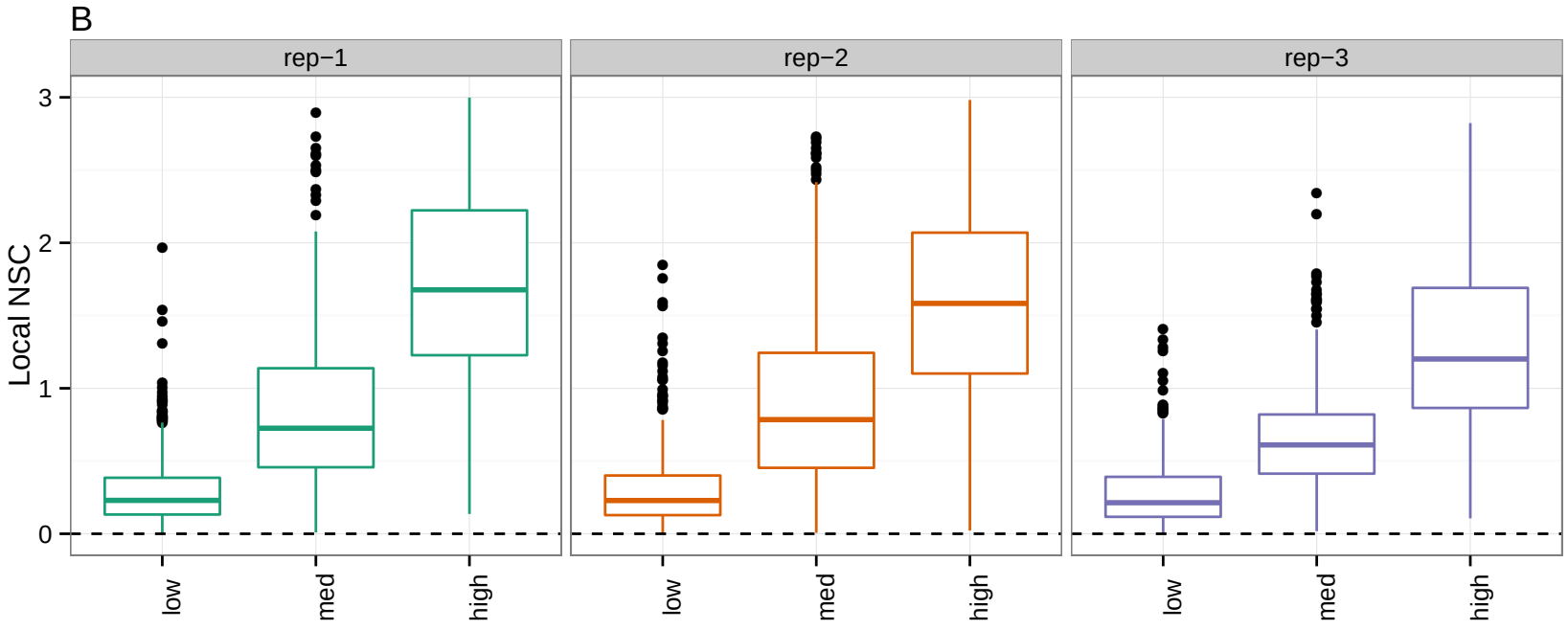
Library complexity and enrichment





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

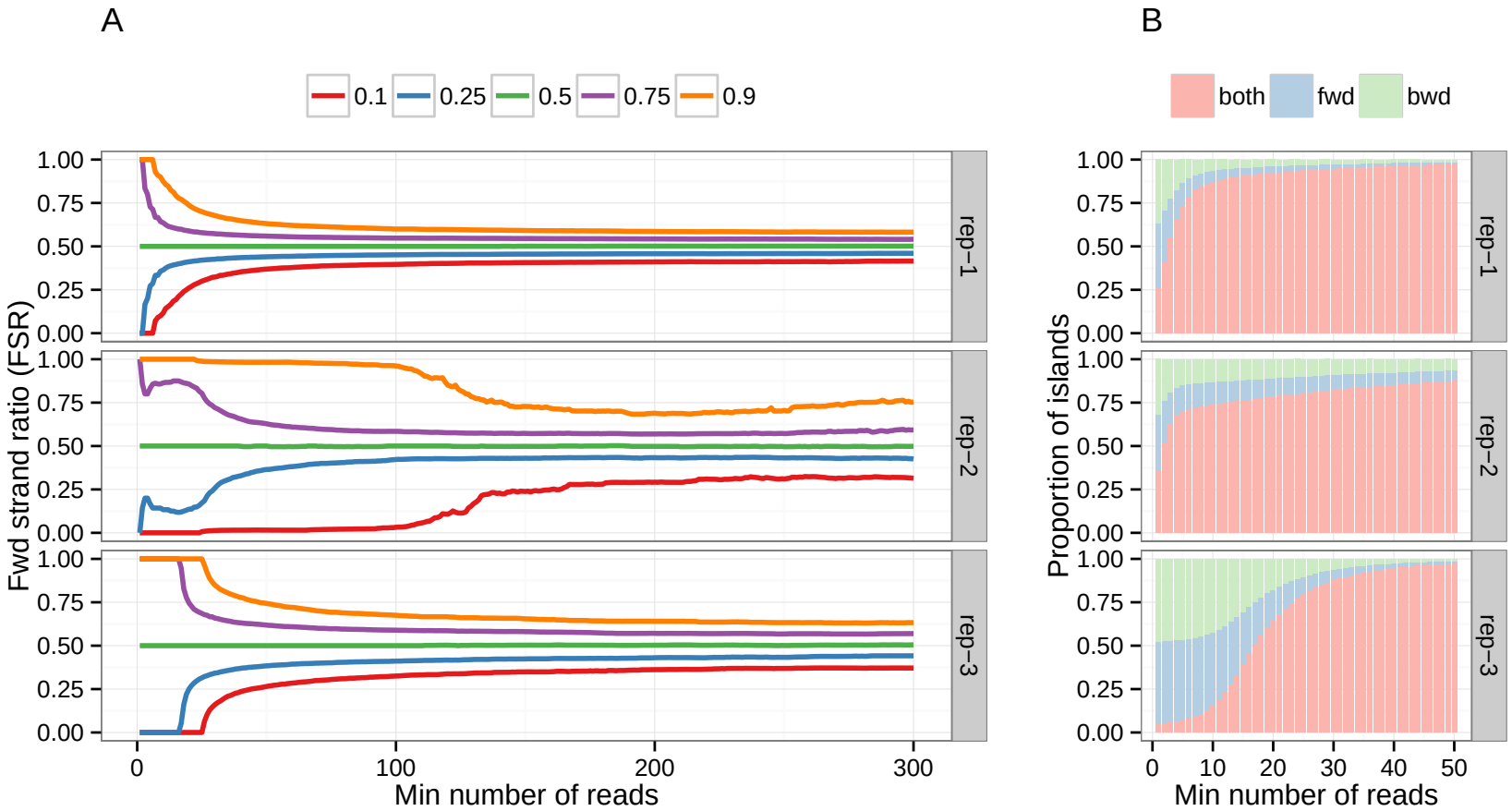
Library complexity and enrichment



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

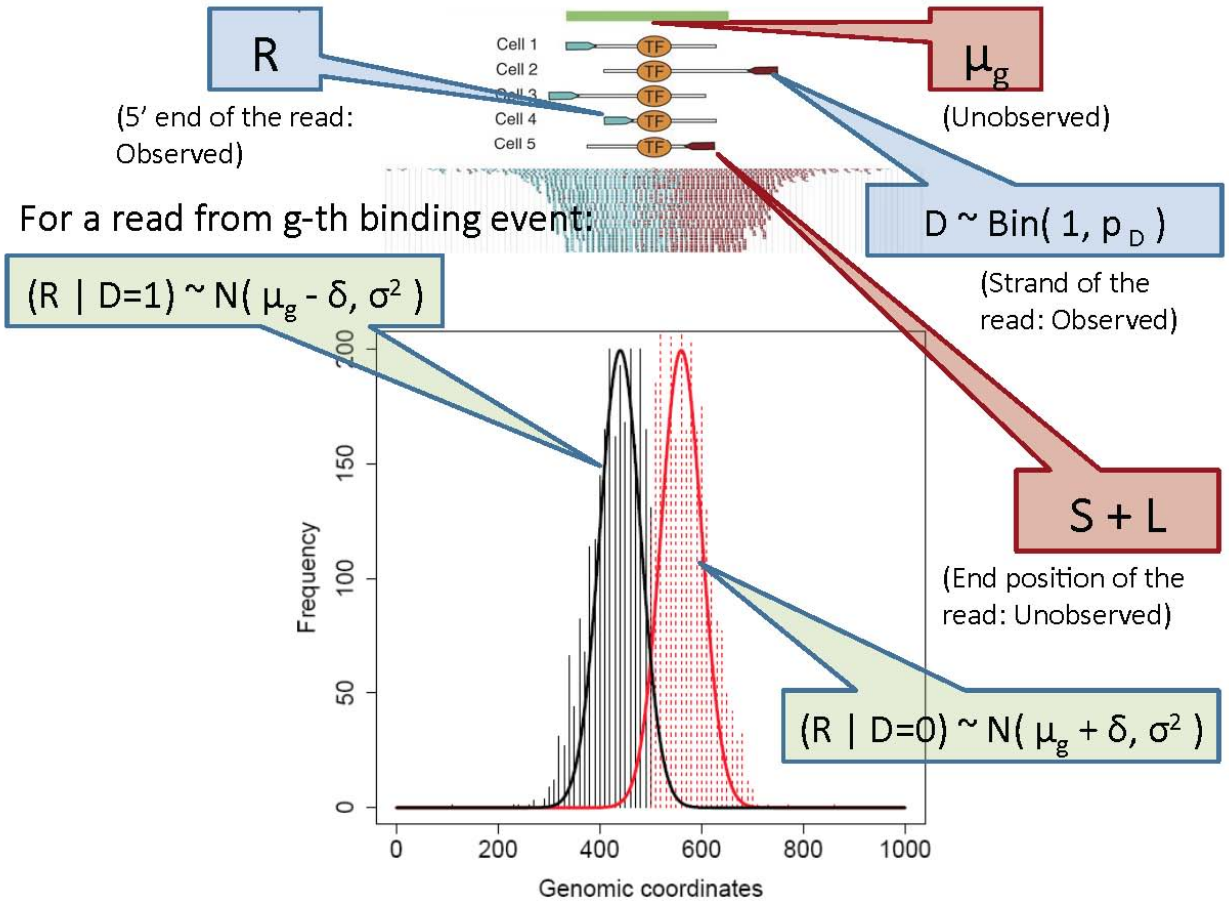


dPeak model for SET case

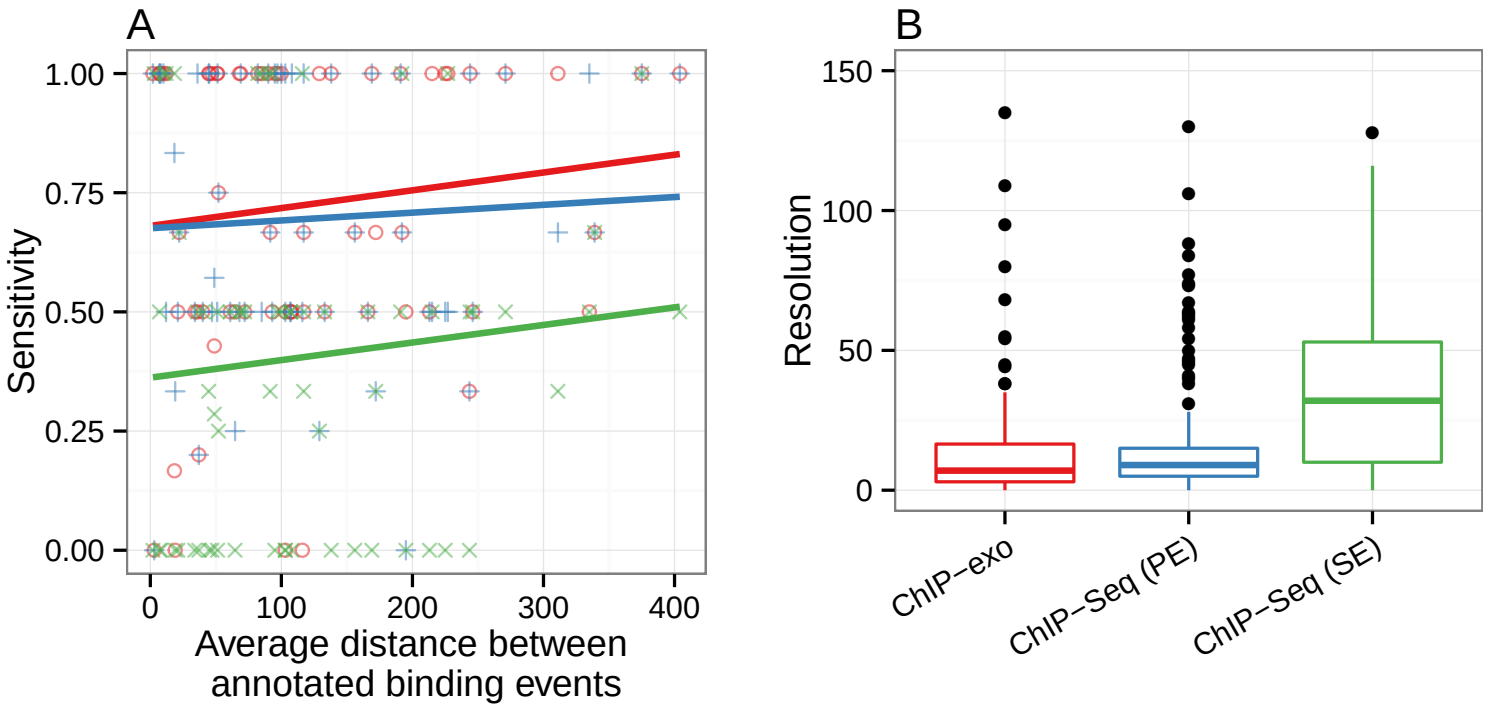
We consider a region with n reads and m positions, for the i-th read:

- $Z_i \sim \text{Multi}(\pi_0, \pi_1, \cdots, \pi_{g^*})$
- $D_i \sim \text{Ber}(p_D)$
  - 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

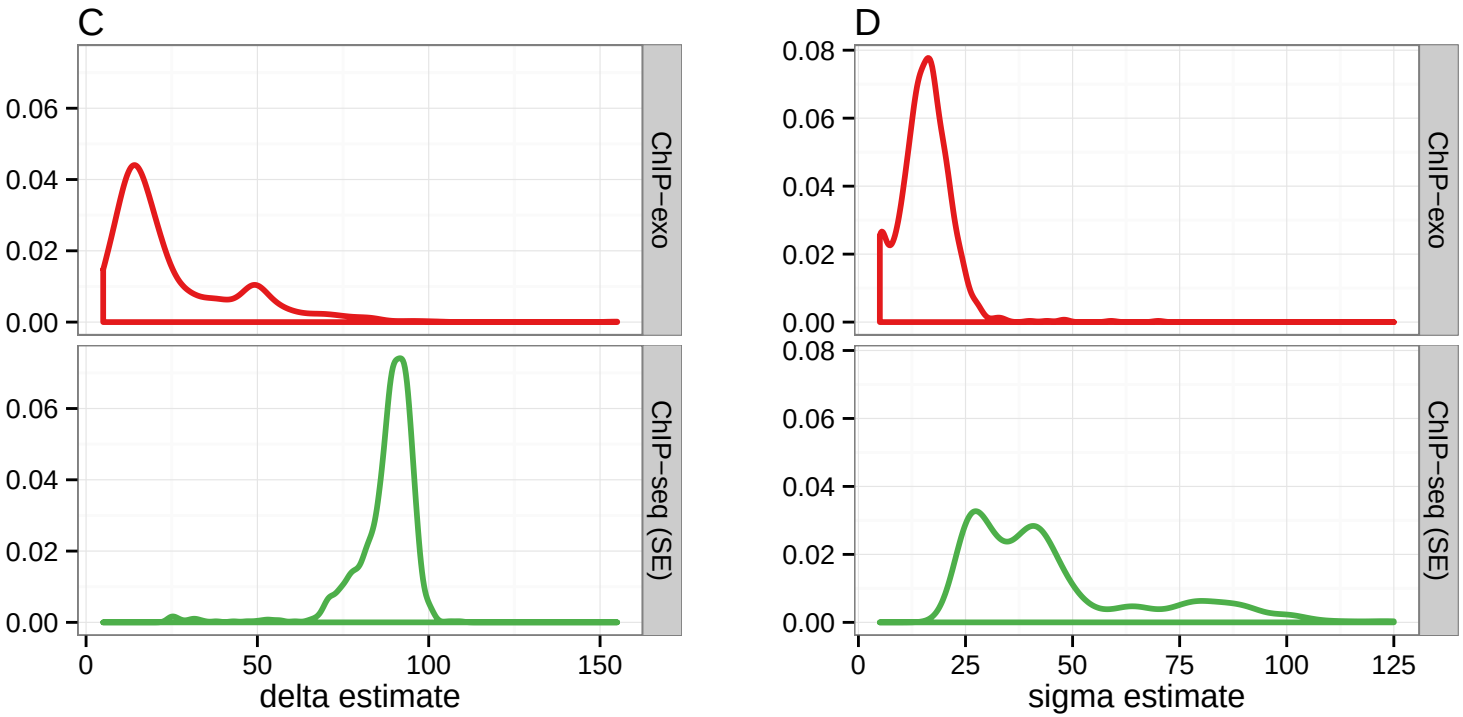


Comparison with ChIP-Seq using dPeak



- Sensitivity is the defined as the proportion of identified peaks (regulonDB [6] is used as gold-standard)
- Resolution is defined as the min. absolute distance of a regulonDB annotation to an est. binding location.

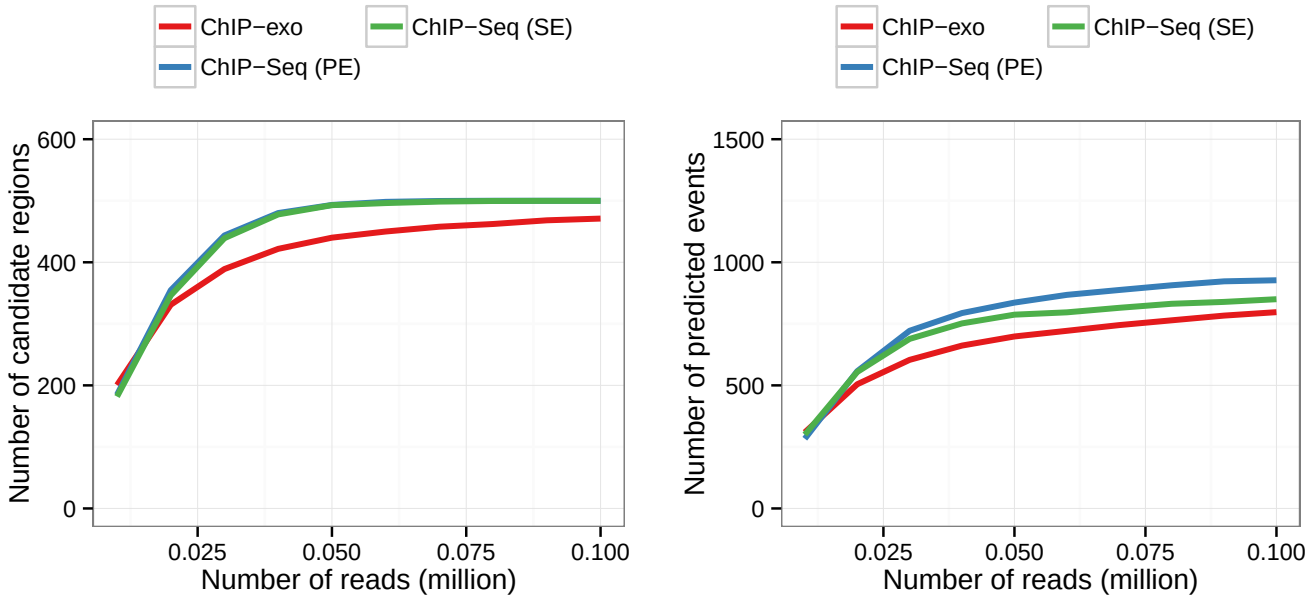
Comparison with ChIP-Seq using dPeak



- 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

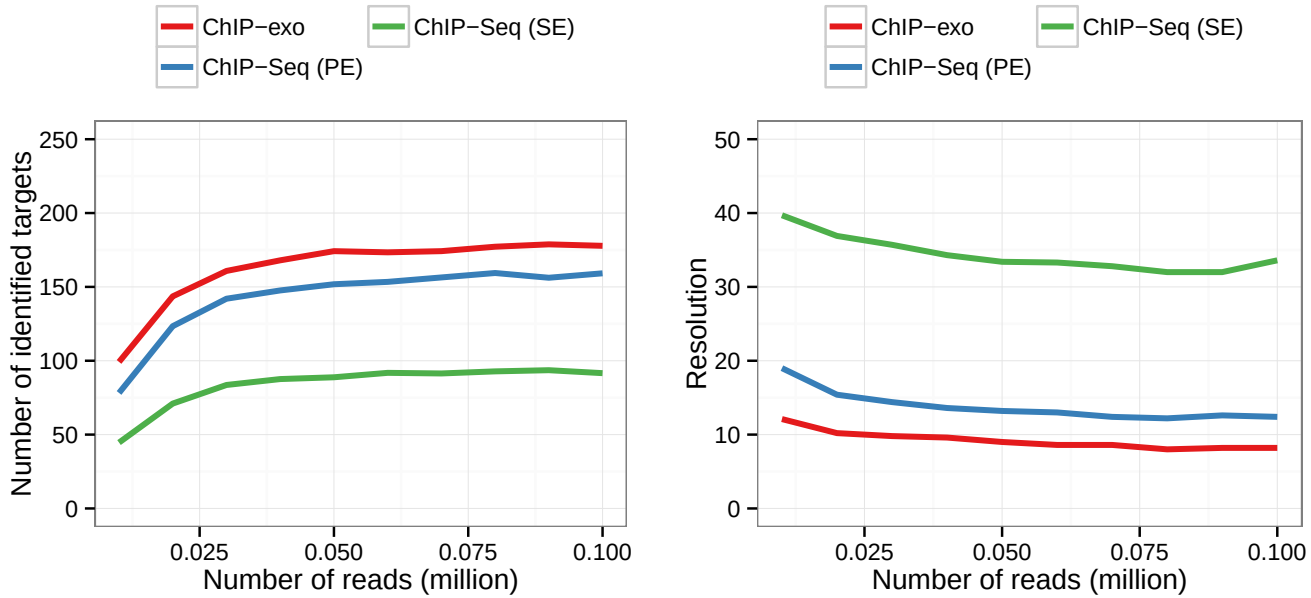
We sampled n fragment reads of each dataset ($2n$ for PET ChIP-Seq), and applied the MOSAiCS / dPeak pipeline:



- 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 labeling 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.

Peter Kharchenko, Michael Tolstorukov, and Peter Park.
Design and analysis of ChIP-Seq experiments for DNA-binding proteins, 2008.

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 Shoresh, Arend Sidow, Matthew Slattery, John Stamatoyannopoulos, Michael Tolstorukov, Kevin White, Simon Xi, Peggy Farnham, Jason Lieb, Barbara Wold, and Michael Snyder.
ChIP-Seq guidelines and practices of the ENCODE and modENCODE consortia.
Genome Research, 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
RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold