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

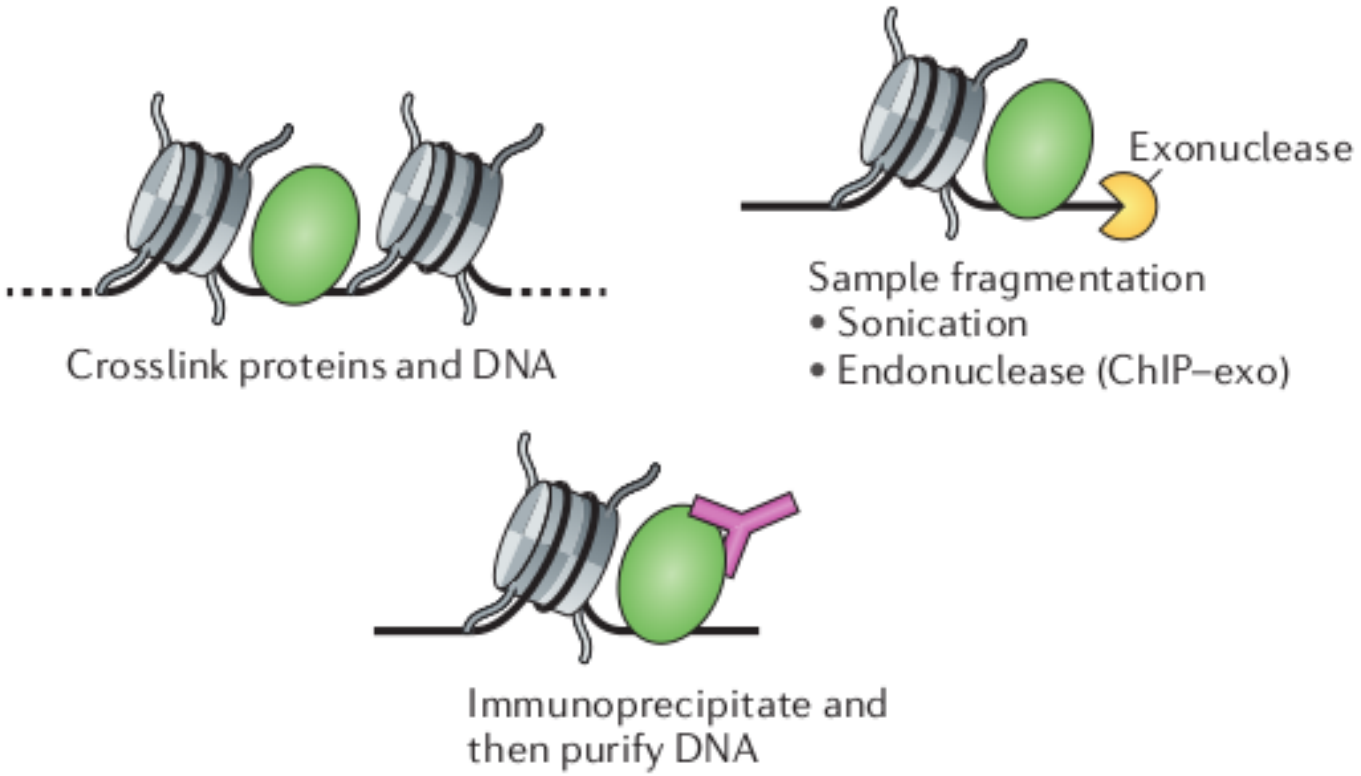


Figure: ChIP-exo procedure, the diagram is from Furey, 2012 [2]

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 - SCC

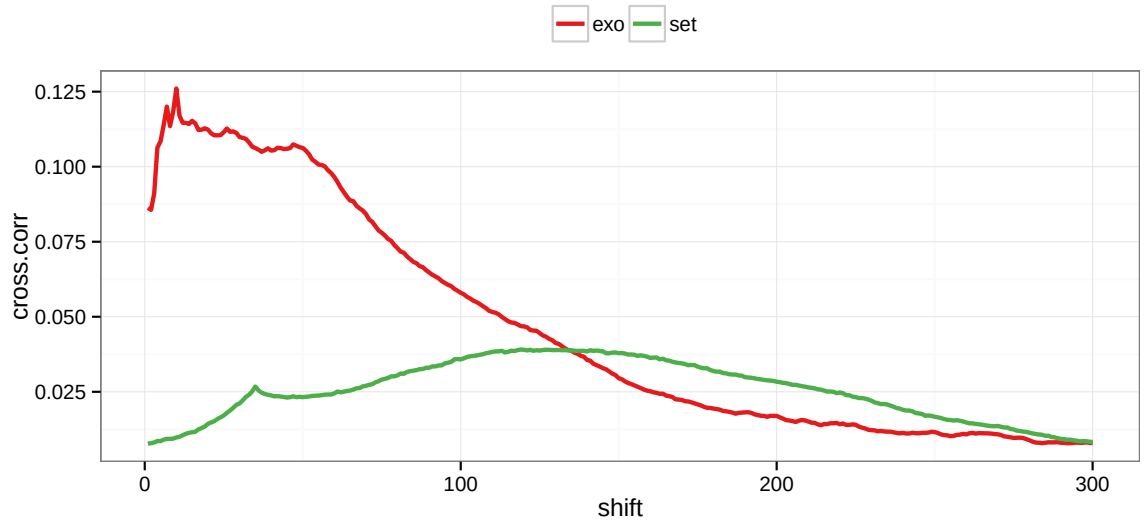
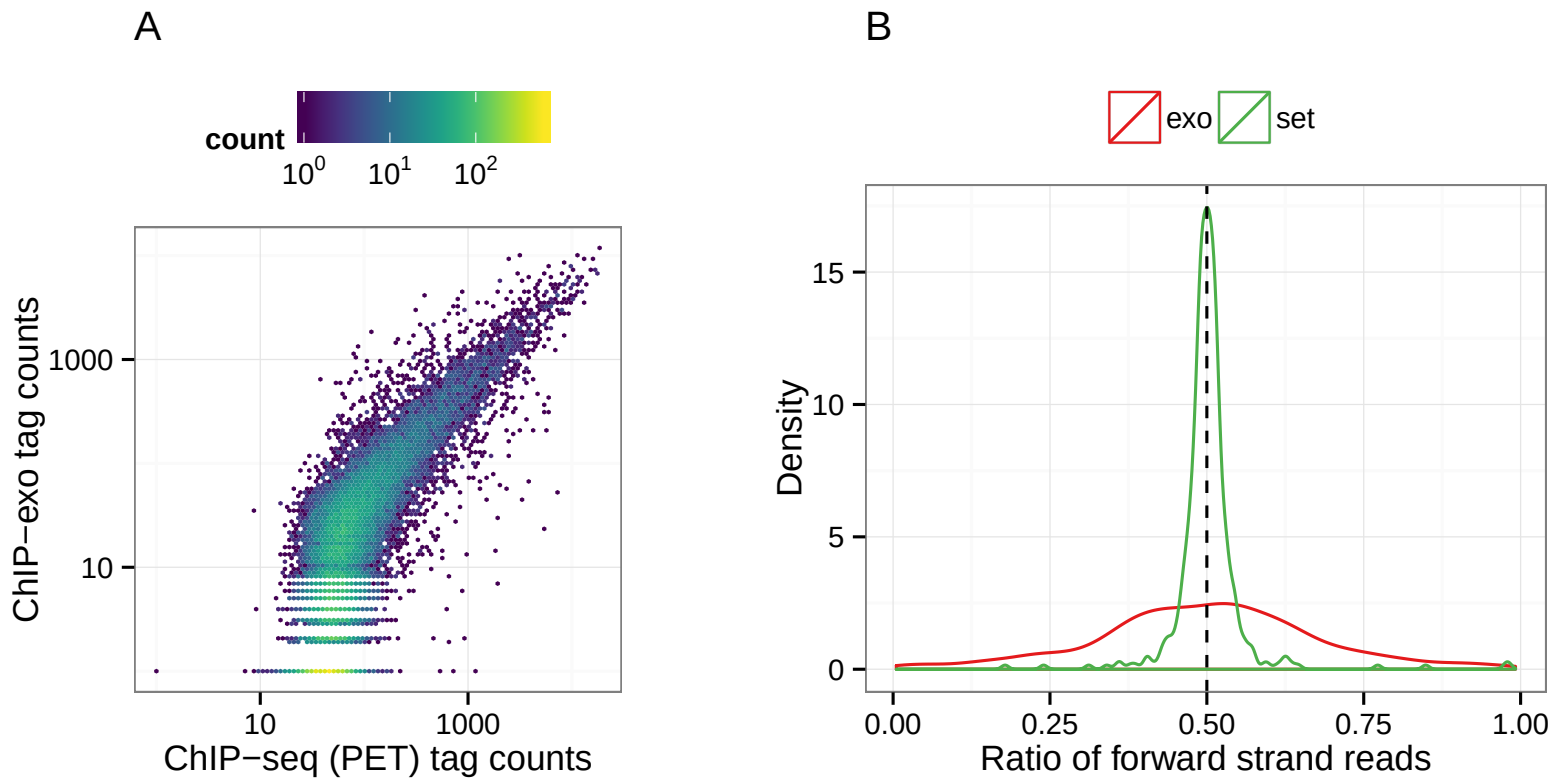


Figure: SCC for CTCF factor in HeLa cell line for ChIP-exo and SET-ChIP-Seq

- 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



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`.

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.