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

1. Background
2. ChIP-exo Data Exploration and Qualiy Control
   - ChIP-Seq QC measures
   - Evaluation of ChIP-Seq QC Metrics for ChIP-exo
   - Comparison of ChIP-exo and ChIP-seq
   - Quality Control Pipeline
3. Systematic Evaluation of ChIP-Seq and ChIP-exo
4. Comparison with Other Algorithms
5. Conclusions and future work

ChIP-exo procedure

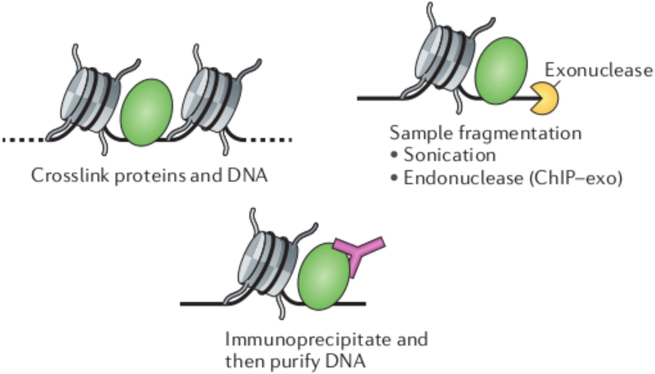

Crosslink proteins and DNA
Exonuclease
Sample fragmentation
• Sonication
• Endonuclease (ChIP–exo)
Immunoprecipitate and then purify DNA

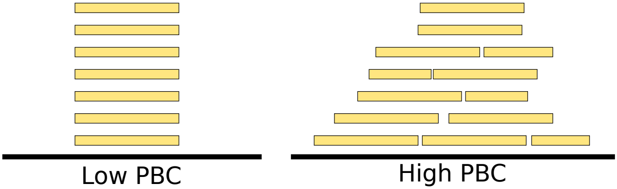
Figure: ChIP-exo procedure, Furey, 2012 [2]

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.

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

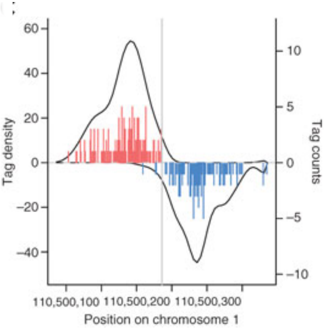


Figure: SCC explanation. Kharchenko et al., 2008 [3]

ChIP-Seq QC measures

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

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

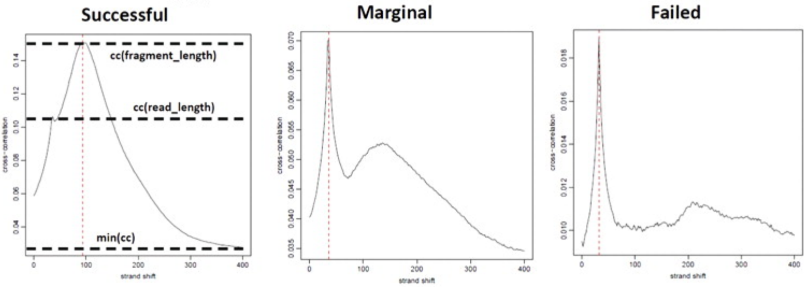


Figure: SCC as QC. Landt et al., 2012 [4]

ChIP-Seq QC measures

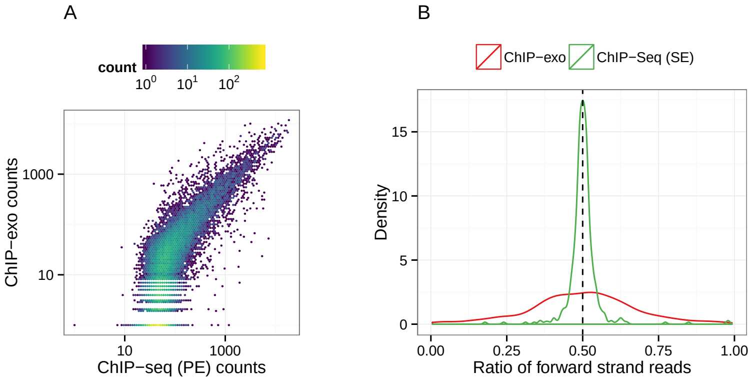
- Number of reads.
- PCR bottleneck Coeff.
- Strand Cross-Correlation.
- Normalized Strand Cross-Correlation. Ratio between the SCC when the shift is the fragment length and min. value of the SCC.

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	
R	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

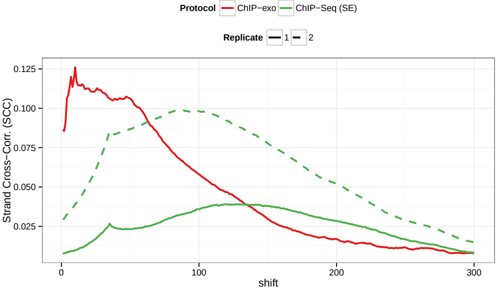
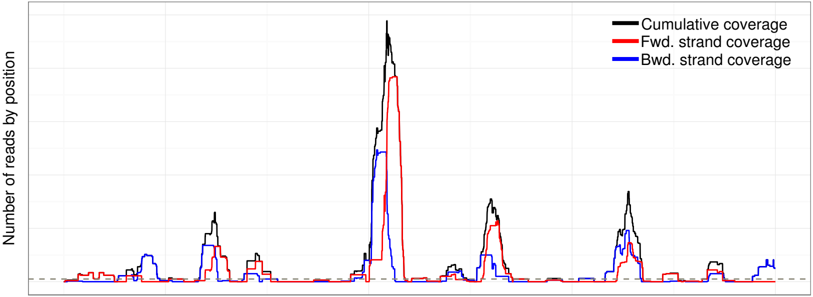


Figure: SCC for CTCF factor in HeLa cell line for ChIP-exo and SET-ChIP-Seq.
- PBC for rep1 is 0.56
- PBC for rep2 is 0.94

- 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.

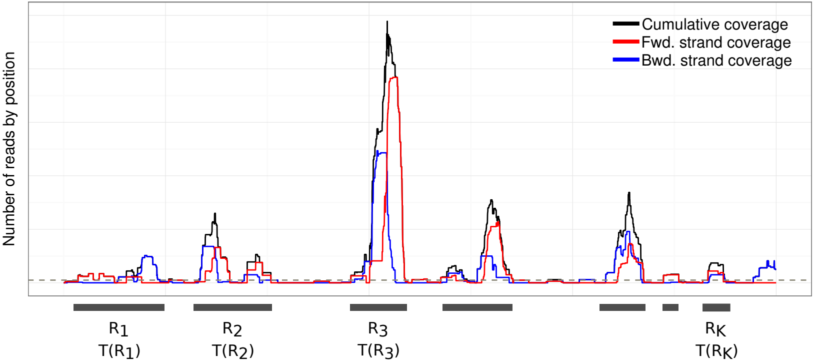
Quality Control Pipeline

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



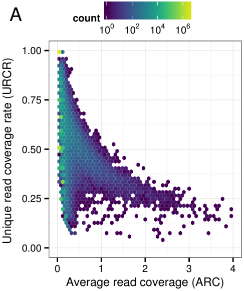
Quality Control Pipeline

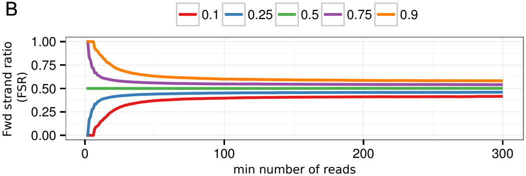
2. Calculate a vector of summary statistics for each islands.

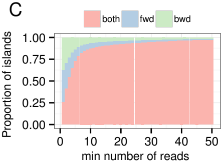


Quality Control Pipeline

3. Visualize all islands together.







3A - ARC vs. URCR - This plot presents 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 where the URCR decreases as the ARC increases.

3B - Min depth vs. FSR - This plot depicts how quickly the distribution of the FSR approximates the median. In a high quality dataset sample, the median is around 0.5, and the other quantiles reach that value quickly.

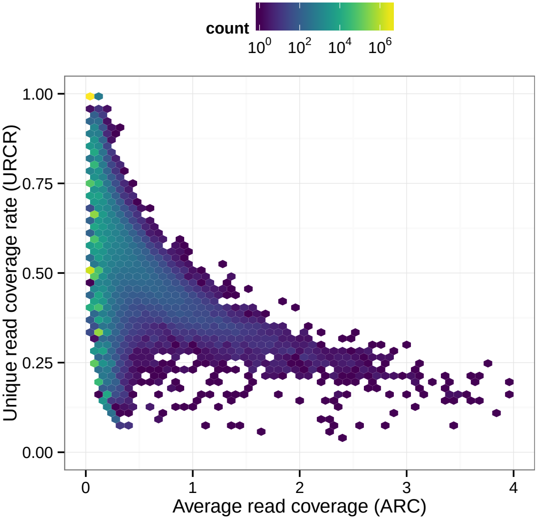
3C - Min depth vs. Proportion of Islands - This plot provides a more detailled view of the FSR. Islands with low depth then to have reads only from one strand. Hence, the plot compares the percentage of islands that contain at least one readof each strand vs. the regions that consist of only reads with only one strand.

ARC vs. URCR exploration

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

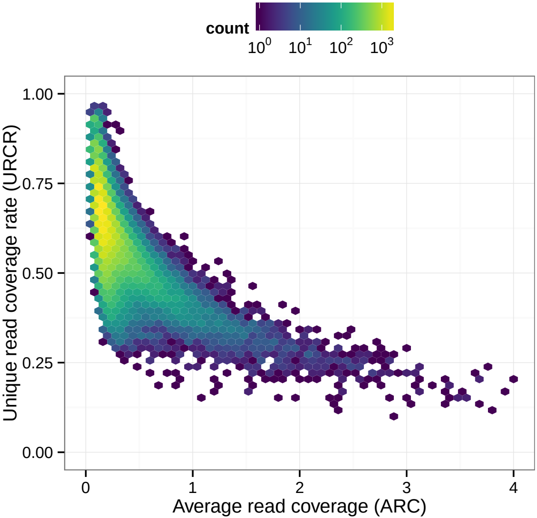
ARC vs. URCR exploration

All regions



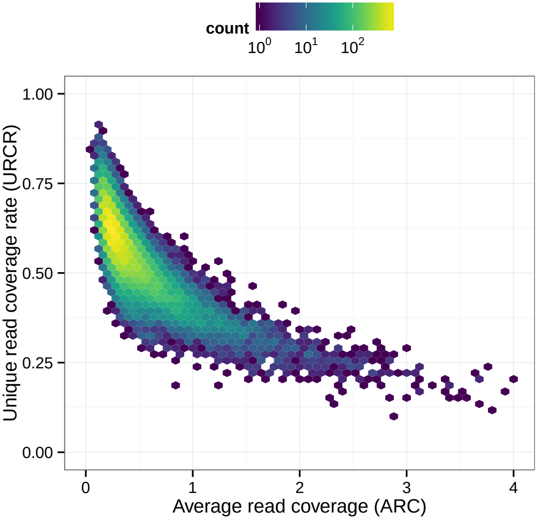
ARC vs. URCR exploration

Mapped to more than 10 positions



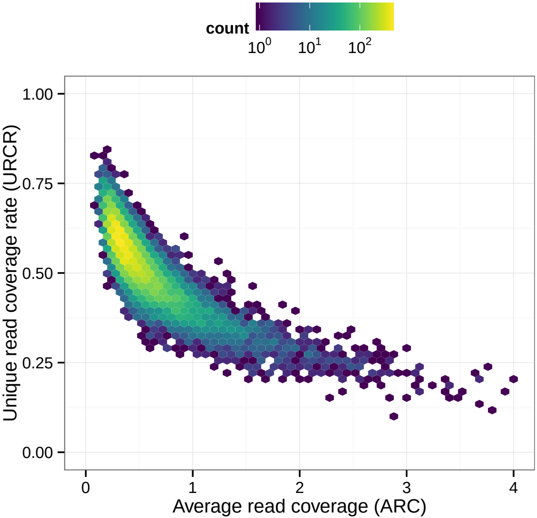
ARC vs. URCR exploration

Mapped to more than 30 positions

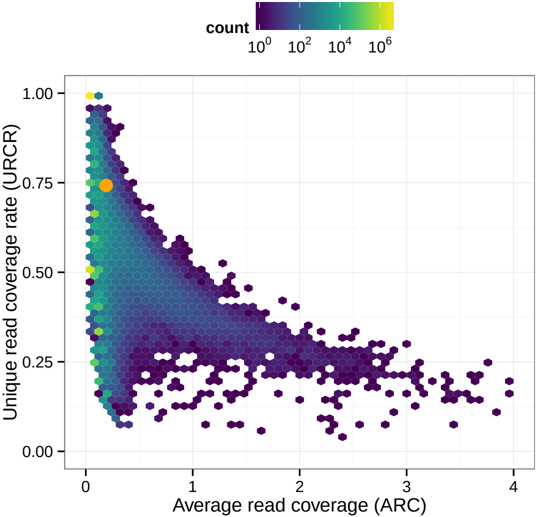


ARC vs. URCR exploration

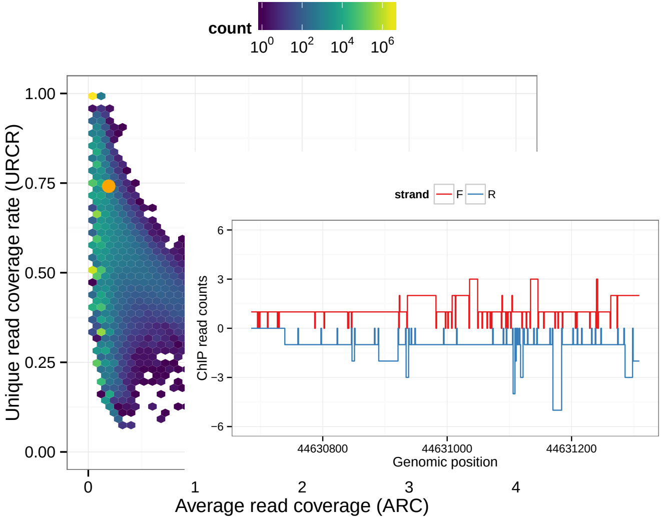
Mapped to more than 50 positions



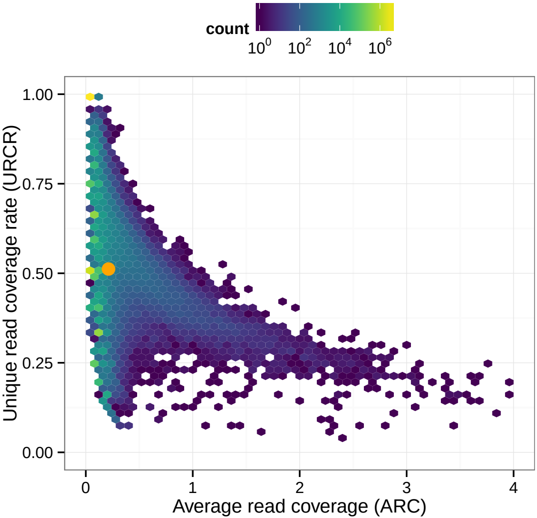
ARC vs. URCR exploration



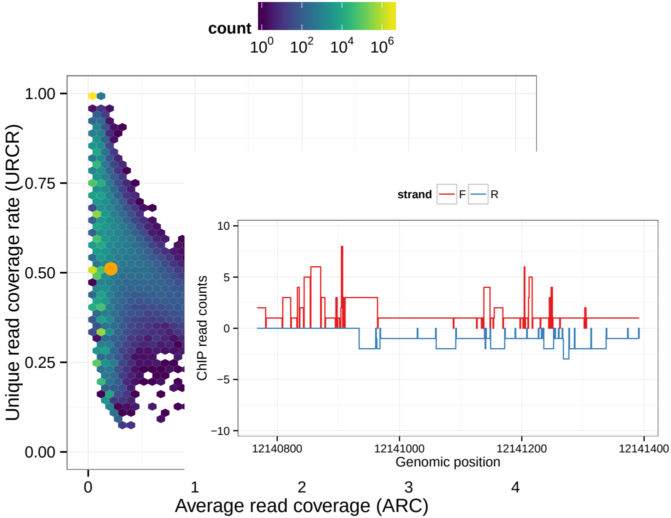
ARC vs. URCR exploration



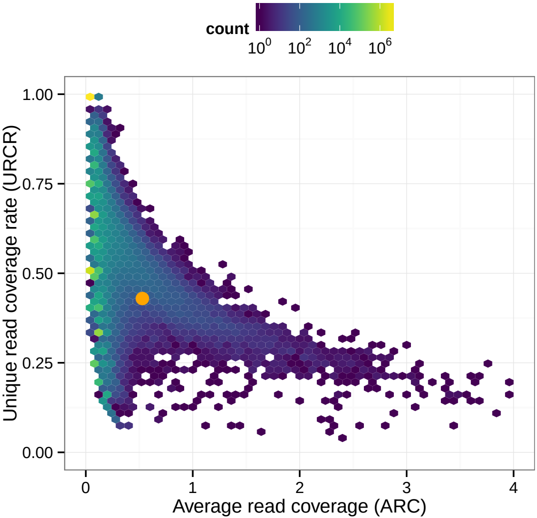
ARC vs. URCR exploration



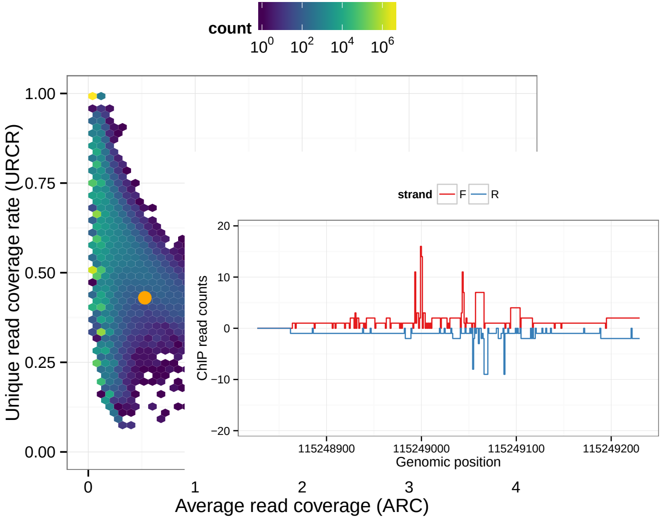
ARC vs. URCR exploration



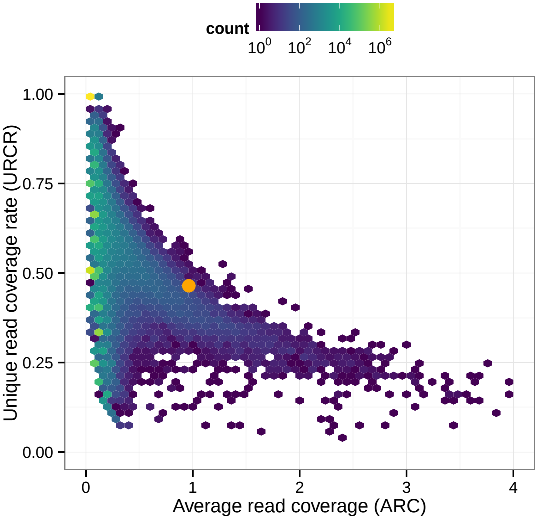
ARC vs. URCR exploration



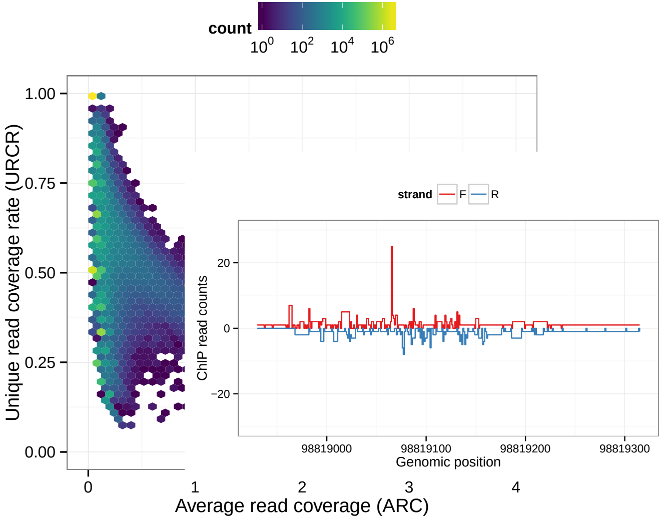
ARC vs. URCR exploration



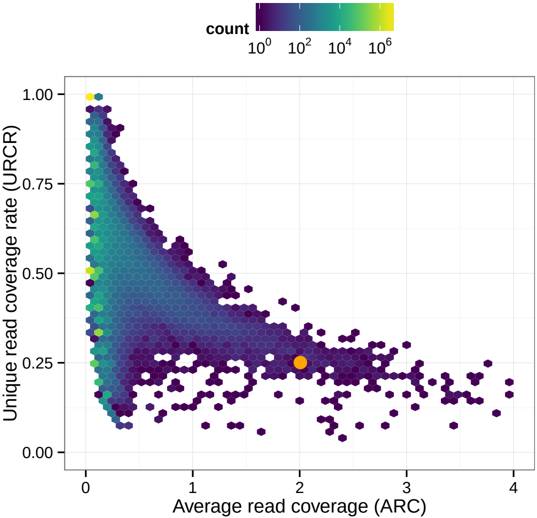
ARC vs. URCR exploration



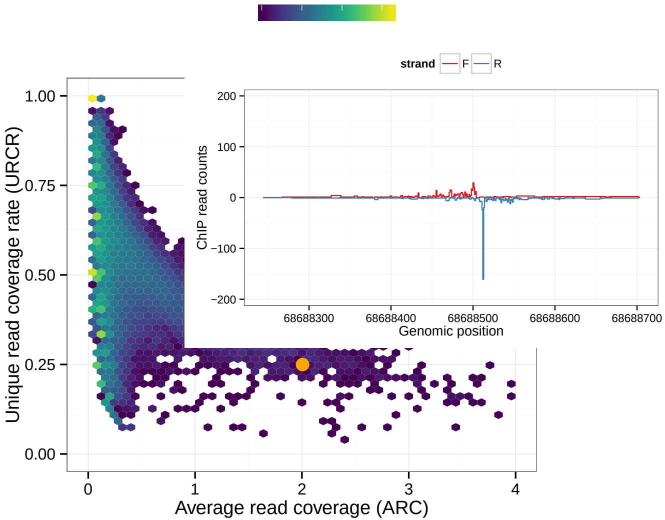
ARC vs. URCR exploration



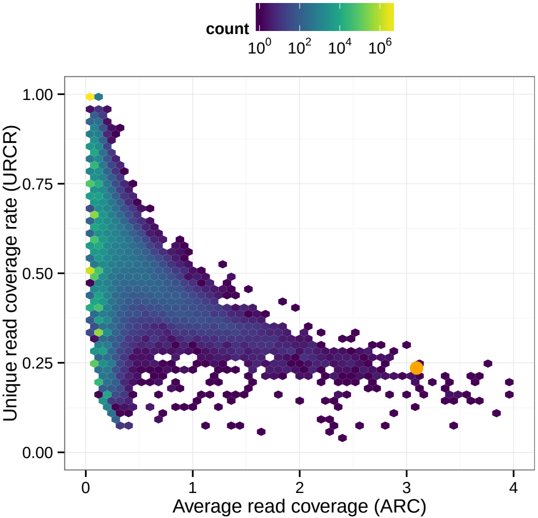
ARC vs. URCR exploration



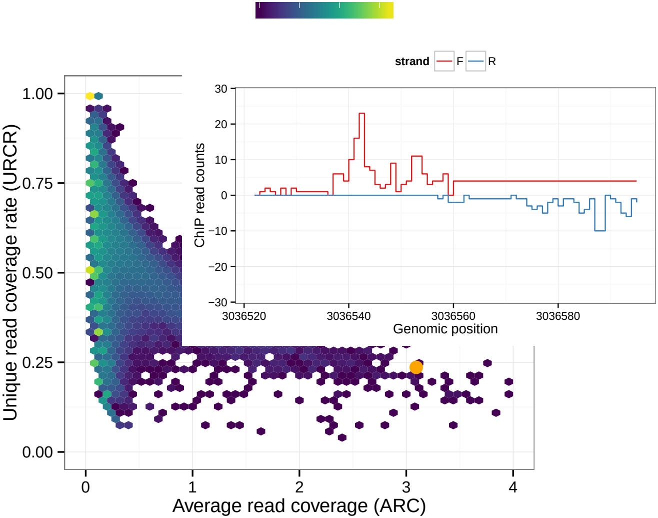
ARC vs. URCR exploration



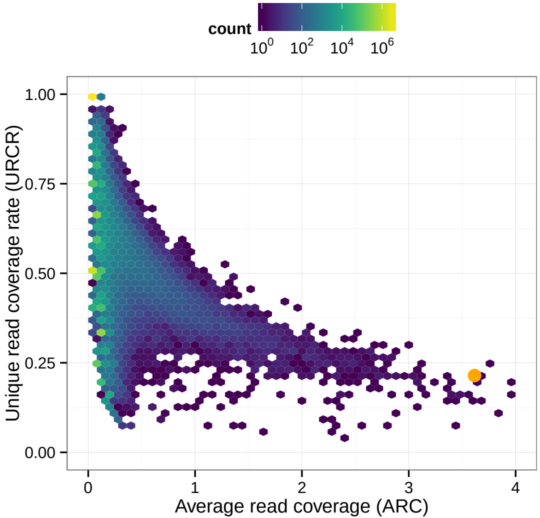
ARC vs. URCR exploration



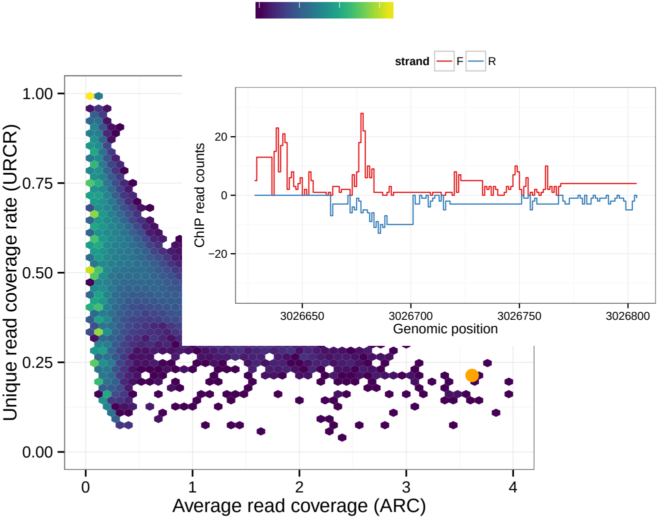
ARC vs. URCR exploration



ARC vs. URCR exploration

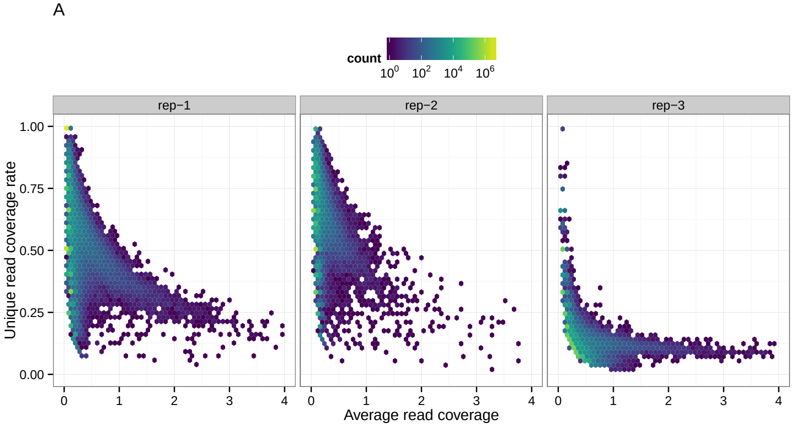


ARC vs. URCR exploration



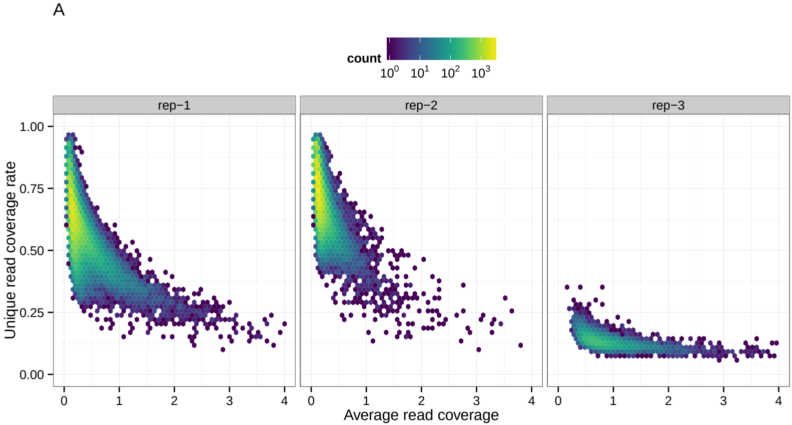
ARC vs. URCR exploration by replicate

All regions



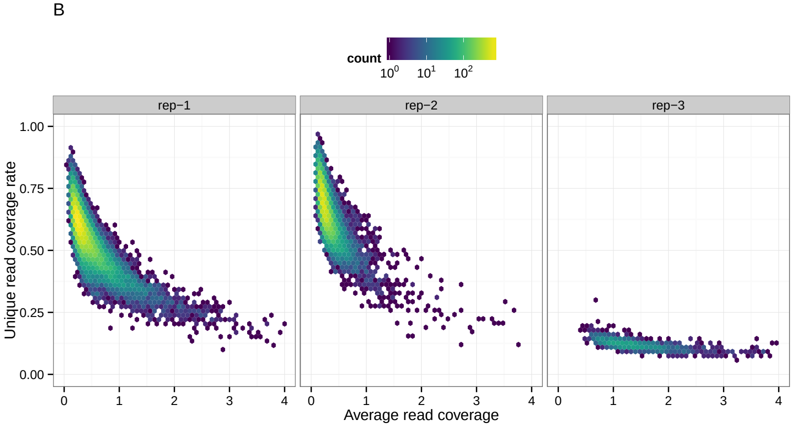
ARC vs. URCR exploration by replicate

Regions formed by reads aligned to more than 10 unique positions



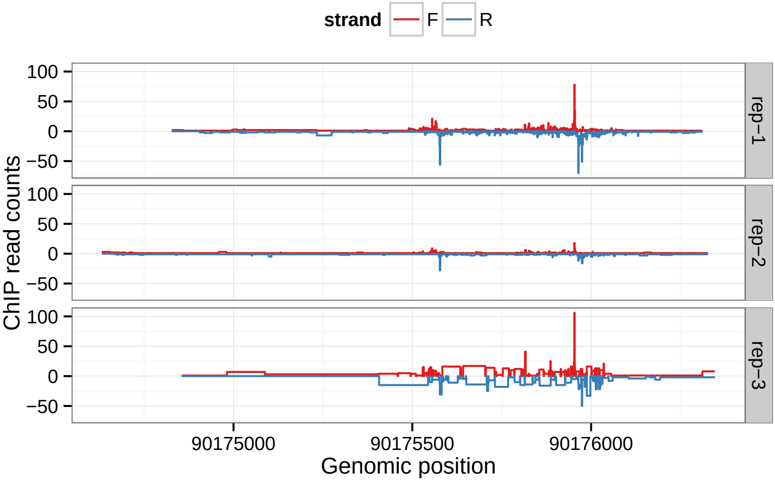
ARC vs. URCR exploration by replicate

Regions formed by reads aligned to more than 30 unique positions



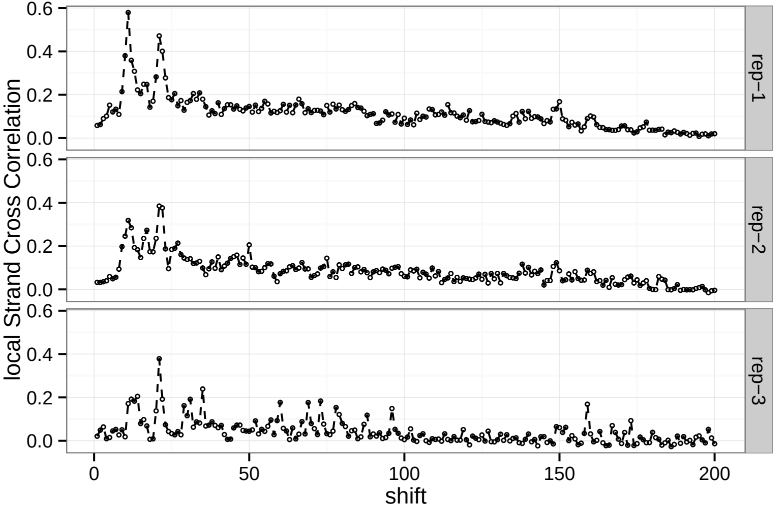
ARC vs. URCR and local-NSC

Calculate coverage for all regions



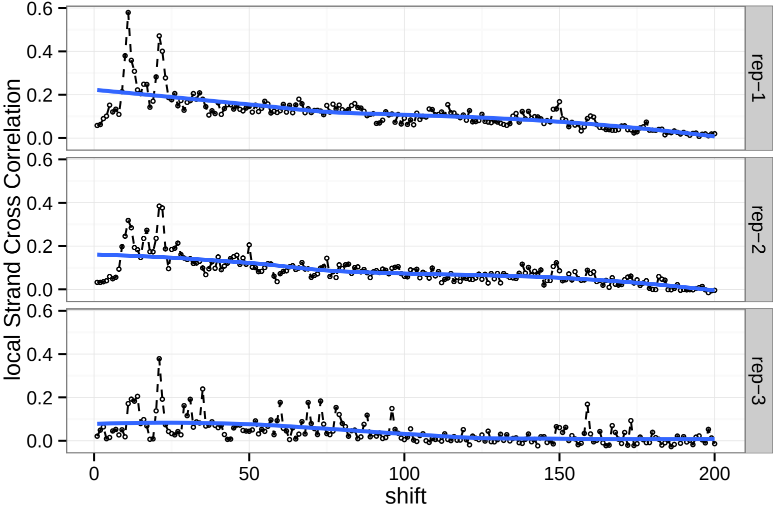
ARC vs. URCR and local-NSC

Calculate local-SCC for regions



ARC vs. URCR and local-NSC

Fit local polynomial regression for the local-SCC



ARC vs. URCR and local-NSC

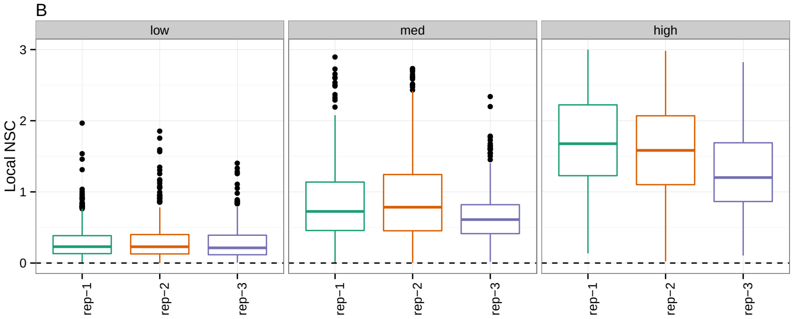
$$y(\delta) = f(x_\delta) + \epsilon_\delta, \quad \text{local-NSC} = \frac{max_{x_\delta} \hat{f}(x_\delta)}{\hat{\sigma}_f},$$

where:

- x_δ is the shift.
- y_δ is the local - SCC at shift x_δ
- $\epsilon \sim N(0, \sigma^2)$ i.i.d

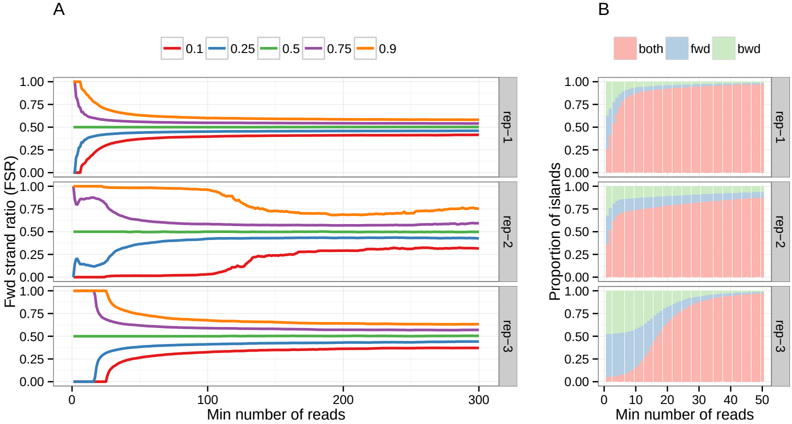
Replicate	$\hat{\sigma}$	$\max_{x_\delta} \hat{f}(x_\delta)$	local-NSC
Rep. 1	0.0533	0.2215	4.14
Rep. 2	0.0453	0.1606	3.54
Rep. 3	0.0475	0.0832	1.75

ARC vs. URCR and local-NSC



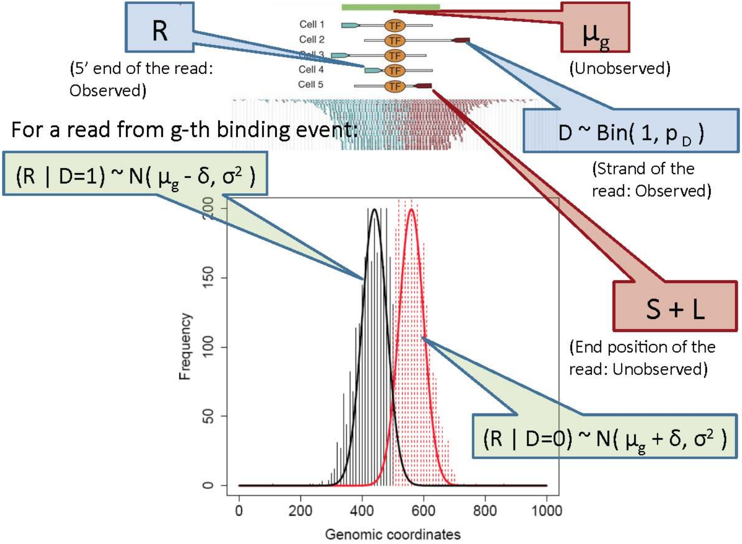
- high: Number of unique positions > 100
- med: Number of unique positions in $(50, 100)$
- low: Number of unique positions in $(20, 50)$

Strand imbalance



-

dPeak model for SET case



R

(5′ end of the read: Observed)

μ_g

(Unobserved)

For a read from g-th binding event:

$(R \mid D=1) \sim N(\mu_g - \delta, \sigma^2)$

$D \sim \text{Bin}(1, p_D)$

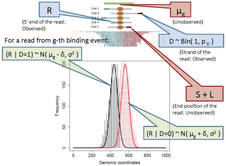
(Strand of the read: Observed)

S + L

(End position of the read: Unobserved)

$(R \mid D=0) \sim N(\mu_g + \delta, \sigma^2)$

dPeak model for SET case

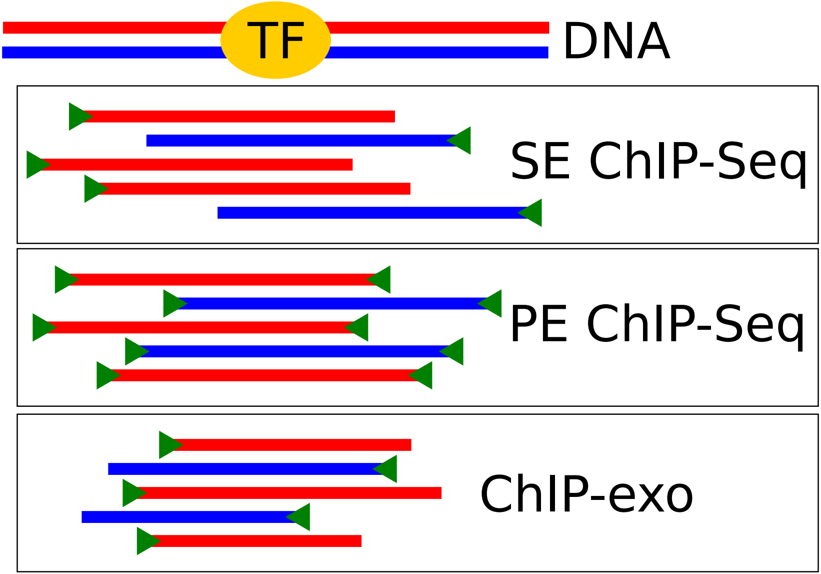

R
(5′ end of the read: Observed)
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
TF
μ_g
(Unobserved)
For a read from g-th binding event:
D ~ Bin(1, p_D)
(Strand of the read: Observed)
(R | D=1) ~ N(μ_g - δ, σ²)
S + L
(End position of the read: Unobserved)
(R | D=0) ~ N(μ_g + δ, σ²)
Frequency
0
50
100
150
200
0
200
400
600
800
1000
Genomic coordinates

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

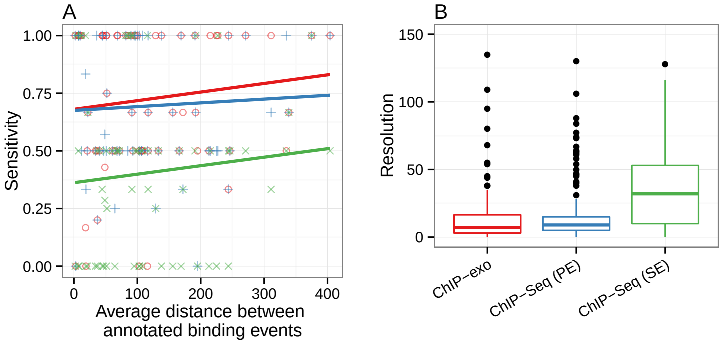
- $Z_i \sim \text{Multi}(\pi_0, \pi_1, \cdots, \pi_{g^*})$
- $D_i \sim \text{Ber}(p_D)$
  - The read is in the forward strand $(D_i = 1)$:
    - The reads belongs to the background: $R_i | Z_i = 0, D_i = 1 \sim \text{Unif}(1, m)$
    - The read belong to the g-th binding event: $R_i | Z_i = g, D_i =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

Data structure

Data available for σ^{70}:

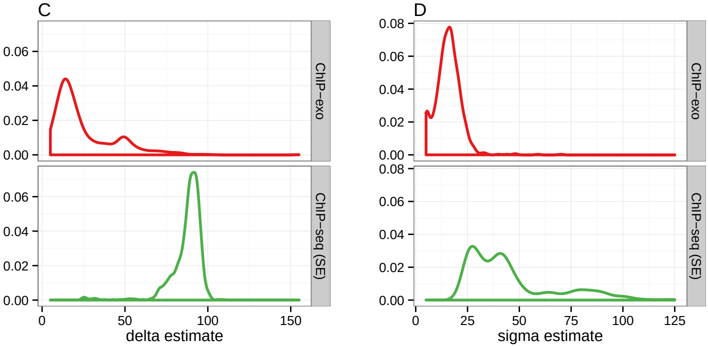


Comparison with ChIP-Seq using dPeak



- 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.

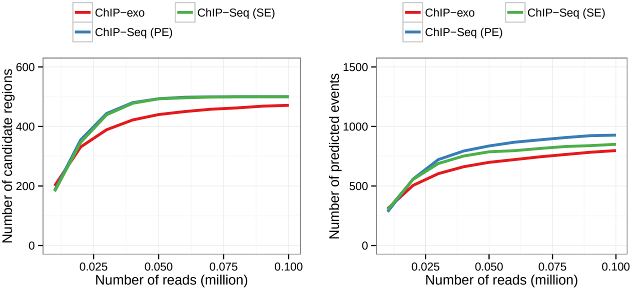
Comparison with ChIP-Seq using dPeak



- δ measures the average distance of reads to their respective binding sites.
- σ measures the dispersion of reads around their respective binding sites.
- Good news !! the model is reflecting the truth.

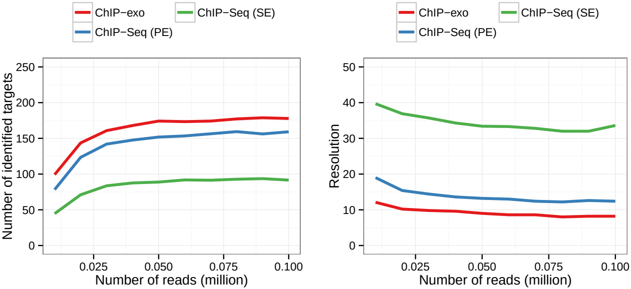
ChIP-Seq comparison at fixed depth

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



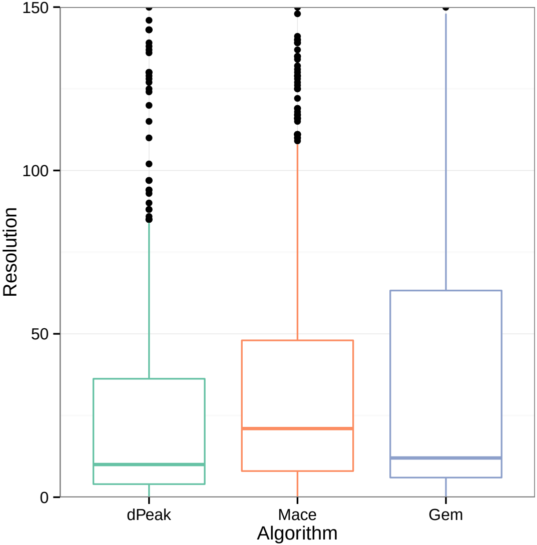
ChIP-Seq comparison at fixed depth

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



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

Comparison with Other Algorithms



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.
- dPeak outperforms other algorithms in resolution.

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. We want to improve this analysis by using ChIP-exo data.
- 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 I

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.

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.

References II

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.

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.