# ChIP-Seq Concepts and Applications

Martin Morgan

Fred Hutchinson Cancer Research Center

14-18 June, 2010

# Outline

- Experimental Setup
  - ChIP-Seq
  - Biological Questions
  - Experimental Designs

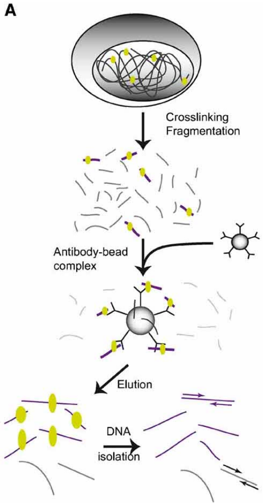
- Pre-Processing & Analysis
  - Alignment and Quality Assessment
  - Peak Identification & Quantification
  - Additional Considerations

- Annotation and Integration

- Resources

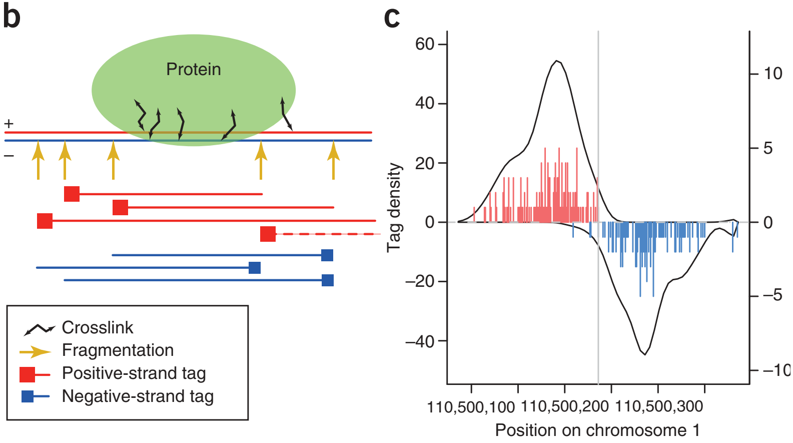
# Chromatin Immunoprecipitation

- Cross-link
- Cell lysis and fragmentation
  - Sonicate (transcription factors)
  - Micrococcal nuclease (nucleosomes)
- Enrichment by immuno-precipitation
  - Antibody + magnetic beads
  - DNA purification
- Adaptor-mediated PCR amplification



Barski and Zhao (2009)

# Experimental Setup: ChIP and Seq

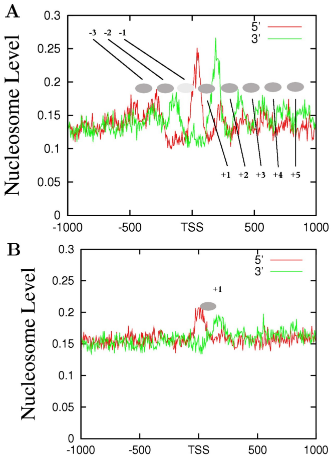


Kharchenko et al. (2008)

# Biological Questions: Nucleosomes

Example: Human CD4$^+$T cells

- Phasing tightly correlated with Pol II binding
- Differential positioning of first nucleosome

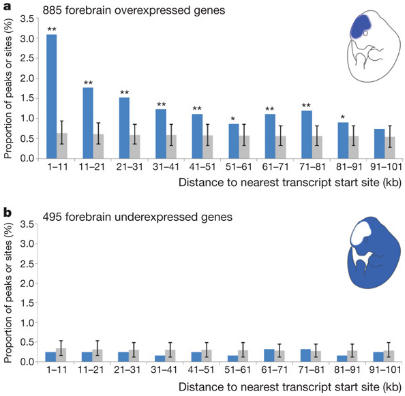


Schones et al. (2008)

# Biological Questions: Transcription Enhancers

Example: tissue-specific enhancers in mouse embryonic forebrain

- ChIP of enhancer associated protein p300 identifies 1000's of binding sites
- *In vivo* effects reproducible in transgenic mice
- Enriched binding near expressed genes



Visel et al. (2009)

# Experimental Designs

- Single sample
- Control lane (Park, 2009)
  - 'Input' control (DNA prior to IP)
  - Mock IP (no antibodies)
  - Non-specific IP
- Designed experiment – factor(s) with 2 or more levels.

"ChIP experiment depends on many intractable parameters, likely including. . . phase of the moon" (Barski and Zhao, 2009)

# Depth of Coverage

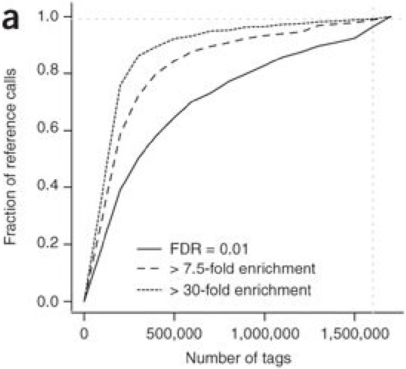
Heuristic, modified from Barski and Zhao (2009)

$$S \propto \frac{G}{W} \times \frac{N}{E}$$

| | |
|---|---|
| S | Amount of sequencing |
| G | Non-repetitive genome size |
| W | Window size (resolution) |
| N | Fraction of genome after ‘ChIP’ |
| E | Enrichment factor (antibody immunoprecipitation) |

- $E$ implies that the input lane is well-characterized, and this implies extensive sequencing of the input, where $N$ is large

# Depth of Coverage



Kharchenko et al. (2008)

- No saturation at specified FDR: more peaks continually found, because larger read counts increase statistical power
- Imposing a fixed fold enrichment criterion established
- Multiplexing increasingly attractive, especially for small genomes / well defined ChIP targets

# Sequencing

Longer reads

- Better mapping in repetitive regions

Paired end reads

- Easier transcription factor binding site identification?
- Better mapping (on the borders of) repetitive regions?

# Outline

- Experimental Setup
  - ChIP-Seq
  - Biological Questions
  - Experimental Designs
- Pre-Processing & Analysis
  - Alignment and Quality Assessment
  - Peak Identification & Quantification
  - Additional Considerations
- Annotation and Integration
- Resources

# Pre-Processing: Alignment & Quality Assessment

Library construction

- Under-representation of AT-rich regions with low melting temperature, GC-rich regions due to PCR bias
- MNase sequence preference (nucleosomes)
- Antibody variability
- Optical and PCR duplicate reads

Alignment

- Micro-repetitive genomic regions (non-alignment)
- Non-specific enrichment, often reads on a single strand

Barski and Zhao (2009), Park (2009)

# Analysis

Simple

- View aligned reads in a browser (or *HilbertVisGUI*!)

Binned

- Divide genome into bins
- Identify bins with greater-than-expected (e.g., Poisson) counts

Model-based

- Exploit $+/-$ strand asymmetry to more narrowly define binding sites

Pepke et al. (2009) and Schmidt et al. (2009) offer comprehensive enumerations of available software

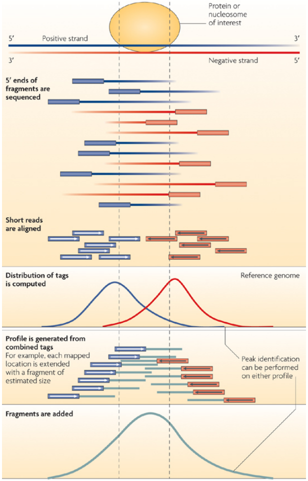
# Peak Identification

Strand asymmetry

- 5’ end of fragments
- Distinct distributions on +, − strands

Smoothed profile

- Shift each distribution toward center, or. . .
- Extend each read by estimated fragment length
- Other algorithms: Kharchenko et al. (2008)


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Nature Reviews | Genetics

Park (2009)

# Peak Quantification

Fold ratio

- Enrichment relative to control
- But: 5-fold change from 10 to 50 has different statistical significance from 100 to 500

Model-based

- Poisson (or other) description of count data, e.g., *MACS*, Zhang et al. (2008)
- Adjust for regional bias in tag density from library construction / micro-repetitive regions, e.g., *PeakSeq*, Rozowsky et al. (2009)

# An Example: MACS, Zhang et al. (2008)

Pre-processing

- Remove duplicate tags if more than expected based on sequence depth
- Scale control lane to same total tag count as experiment

Peak identification

- Use ‘high-quality’ peaks to estimate fragment width $d$
- Shift all peaks $d/2$ toward 3’ end

Peak quantification

- Whole genome Poisson expectation $\lambda_{BG}$, local processes based on control $\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$;
- Local Poisson process $\lambda_{\text{local}} = \max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k})$
- Score is Poisson probability based on $\lambda_{\text{local}}$

# Peak Problems

Three types of peak

- Sharp: protein-DNA binding; histone modification of regulatory elements
- Broad: histone modifications marking domains
- Mixed

Algorithms generally satisfactory for sharp peaks; adopt *ad hoc* approaches for broad / mixed peaks

# Additional Considerations

Assessing algorithm performance

- Validate using quantitative PCR
- Distribution of distances from peaks to nearest known motif

Statistical significance

- Adjusted to reflect multiple comparisons
- Usually reported as false discovery rate
- Requires realistic null model, e.g., capturing local variations in input control

# Designed Experiments

Much like microarray data

- Rectangular data; ‘features’ × ‘samples’
- Easier to compare across samples than features (?)

Important application-specific issues

- Counts: distinct properties require appropriate error model
- Measurement ‘features’ discovered rather than determined *a priori*

# Outline

Experimental Setup

- ChIP-Seq
- Biological Questions
- Experimental Designs

Pre-Processing & Analysis

- Alignment and Quality Assessment
- Peak Identification & Quantification
- Additional Considerations

Annotation and Integration

Resources

# Annotation and Integration

Annotation

- Relate peak locations to known genomic features, e.g., transcription start sites
- Gene set enrichment-style analyses
- Motif discovery from high-scoring peaks

Integration

- Expression levels of genes
- SNPs and allele-specific binding

# Outline

Experimental Setup
- ChIP-Seq
- Biological Questions
- Experimental Designs

Pre-Processing & Analysis
- Alignment and Quality Assessment
- Peak Identification & Quantification
- Additional Considerations

Annotation and Integration

Resources

# Resources

*Bioconductor* packages

- *BayesPeak*, *CSAR*, *chipseq*
- *ChIPpeakAnno*
- *MotIV*, *rGADEM*
- *ChIPsim*

# References I

A. Barski and K. Zhao. Genomic location analysis by ChIP-Seq. *J. Cell. Biochem.*, 107:11–18, May 2009.

P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of chIP experiments for DNA–binding proteins. *Nature Biotechnology*, 26:1351–1359, 2008.

P. J. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10:669–680, Oct 2009.

S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, 6:22–32, Nov 2009.

J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27:66–75, Jan 2009.

# References II

D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom. ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*, 48:240–248, Jul 2009.

D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887 – 898, 2008. ISSN 0092-8674.

A. Visel, M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457: 854–858, Feb 2009.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9:R137, 2008.