# ChIP-seq experimental design and analysis

Martin Morgan (mtmorgan@fhcrc.org)
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

19 November, 2009

# Classical ChIP-chip

Biological context

- ‘Punctuations’, e.g., <200bp; transcription factor finding sites, e.g., associated with CTCF
- Broad, e.g., RNA polymerase II binding to promoters, but also over body of actively transcribed regions
- Histone marks and chromatin domains

Approach

- Cross-link chromatin, e.g., formaldehyde
- Immunopreciptate with specific antibodies → enriched DNA fragments of desired length, e.g., 500bp
- Quantify enrichment by hybridization to tiling microarrays

# From ChIP-chip to ChIP-seq

Limitations

- Probe-specific behavior
- Dye bias
- Tiling resolution

The promise of ChIP-seq

- Greater sensitivity; smaller sample volumes
- Useful early references: Johnson et al. (2007); Robertson et al. (2007)

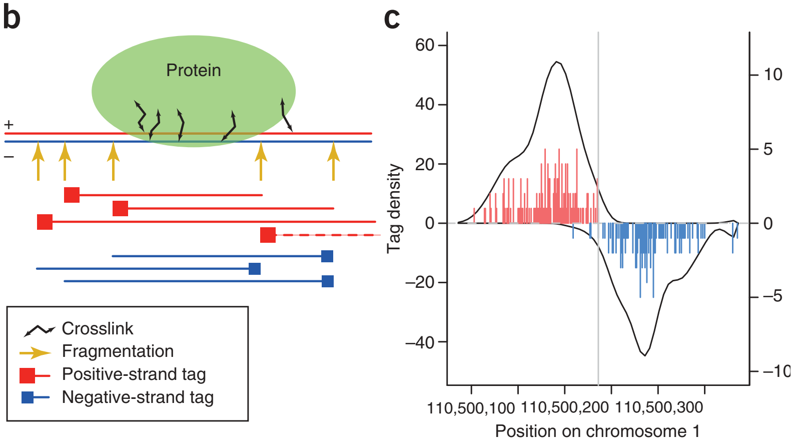
# Sample preparation and mapping

Sample preparation

- Pull-down / enrichment protocols comparable to ChIP-chip
- Sequence preparation: fragmentation (sonication); size selection; primer / adapter ligation

Sequencing and mapping

- Short reads, with characteristic errors
- Mapping with exact or near exact matchingn



Kharchenko et al. (2008)

# ChIP-seq

Criteria for success

- Broad range in number of mapped reads required for ‘success’: 2-20M (Pepke et al., 2009)
- Target properties
  - Number and size of occupied sites
  - Signal intensities
- Library properties
  - Enrichment relative to background
  - Each read from a different founder molecule in the ChIP library
- Trade-offs: specificity (unique reads) vs. sensitivity (multiple reads)

# Sample characteristics

- Majority (60-90%?) are ‘background’ (Pepke et al., 2009)
  - Not as bad as it sounds – 40% of reads distributed over 99.9% of the genome, vs 60% over 0.1%.
- Unmappable genome
  - Repeat regions: reads align to multiple locations; hard to know how to incorporate into read counts
  - Underrepresentation in regions of extreme base composition
- Artifacts of (ChIP) sample preparation
  - E.g., PCR amplification

# Peak identification: major steps

1. Refine signal profile, e.g., smoothing
   - Exercise: implement methods on p. 525 of Pepke et al. (2009)
2. Characterize background
   - Subtract 'input' control
   - Model backgroud, e.g., uniform and strand independent (though several anomalies commonly seen, e.g., excessively large or wide peaks)
3. Determine binding position and strength
   - Aboslute, or relative to background
   - Not always appropriate – e.g., dispersed chromatin marks
4. Filtering
   - *A posteriori* exclusion of discovered peak
   - E.g., Peaks shifted correctly on $+$, $-$ strand
5. Assessment of significance and false discovery rate

# Determining binding position and strength

Several possibilities (e.g., Kharchenko et al., 2008)

- Enrichment relative to ‘input’ (Johnson et al., 2007; Rozowsky et al., 2009) or negative control (Chen et al., 2008)
- XSET
  - Extend reads by expected DNA fragment length
  - Binding regions occur where high numbers of fragments overlap
- Strand-specific shift, e.g., based on fragment length, or estimated from high-quality binding sites
- Strand cross-correlation
  - Shift to maximize correlation between 5’ to 3’ counts on the plus and minus strands

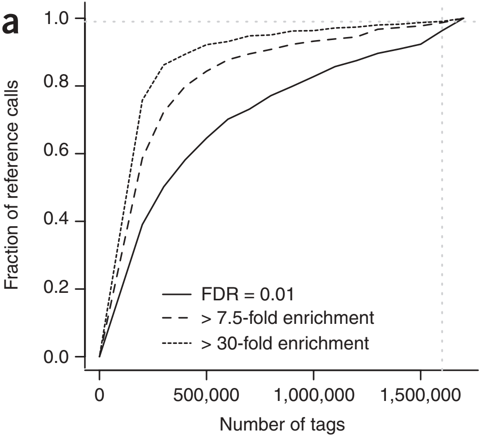
# Statistical characterization

Enrichment, significance, and false discovery

- Parametric assumptions, e.g., background negative binomial
- Empirical
  - Covered binding motifs as a function of binding positions (Kharchenko et al., 2008)
  - False discovery rate as binding regions in control / binding regions in ChIP
- Permutation
  - Maintain spatially proximal tags
- Simulation

# Sufficient sequence depth

Reference binding sites as a function of subsample size (from Kharchenko et al., 2008)



# Annotation and down-stream analysis

- Annotation
- Motif characterization (via position weight matricies)
- Integration with other high-throughput analyses

# R and Bioconductor tools

- chipseq
- ChIPseqR – nucleosome marks
- ChIPsim – simulation
- ChIPpeakAnno – e.g., nearby transcription start sites, enriched GO terms, . . .

# Acknowledgements

- Robert Gentleman, Zizhen Zhao, Deepayan Sarkar, Michael Lawrence, Patrick Aboyoun
- Stephen Tapscott, Yi Cao,
- Hervé Pagès, Marc Carlson, Chao-Jen Wong, Nishant Gopalakrishnan

- NIH / NHGRI 5P41 HG004059-04

# References I

X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133:1106–1117, Jun 2008.

David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007. URL `http://www.sciencemag.org/cgi/content/abstract/316/5830/1497`.

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.

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

# References II

G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4:651–657, Aug 2007.

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.