ChIP-seq peak identification and analysis

Martin Morgan

Fred Hutchinson Cancer Research Center

January 29, 2010

Acknowledgments

- Deepayan Sarkar, Robert Gentleman, Zizhen Zhao, Michael Lawrence, Patrick Aboyoun
- Stephen Tapscott, Yi Cao,
- Hervé Pagès, Marc Carlson, Chao-Jen Wong, Nishant Gopalakrishnan
- NIH / NHGRI P41-HG004059

Classical ChIP-chip

Diverse 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

Overall approach

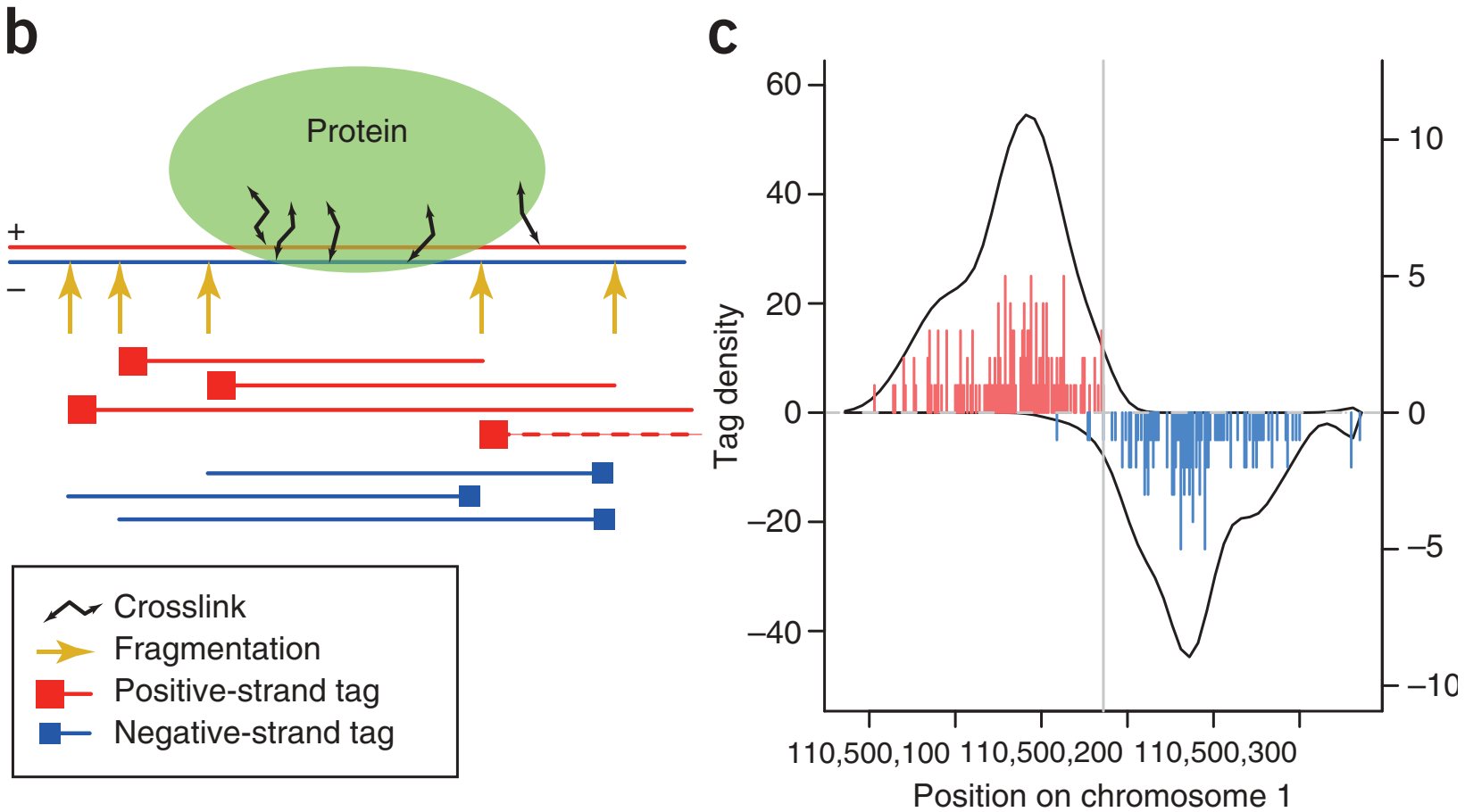
- Cross-link chromatin, e.g., formaldehyde
- Immunoprecipitate with specific antibodies $\longrightarrow$ enriched DNA fragments of desired length, e.g., 500bp
- Quantify enrichment by hybridization to tiling microarrays

ChIP-seq

Overall approach

1. Chromatin immunoprecipitation
2. Sequence
   - Process ChIP'ed DNA, e.g., size selection, adapter ligation
   - Perform whole-genome alignment
3. Characterize areas of high coverage – 'peaks'
4. Compare across experimental conditions

Useful reference: Park (2009).



Kharchenko et al. (2008)

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

Analysis using the *chipseq* package

Biological background: CTCF

- Insulator protein, blocking enhancer / promoter interactions (e.g., IGF-2); zinc finger protein
- 15,000 binding sites in human genome

Source: Chen et al. (2008)

- Mouse embryonic stem cells transcription factor binding sites
- GFP: negative control; no peaks anticipated

Aligned reads

Issues

- Reads aligning to multiple genomic locations? Technology sequence bias?
- Genomic coordinates where multiple reads align?

Decisions

- Ignore reads aligned to multiple genomic locations, because alternative not clear; ignore sequence bias.
- Select a maximum of one read starting at each position – concern is that multiple identically aligned reads reflect PCR artifact during sample preparation

Aligned reads

Psuedo-code

```
> filter <- compose(
+   strandFilter(strandLevels=c("-", "+")),
+   chromosomeFilter(regex = "chr[0-9]+$"),
+   alignQualityFilter(1),
+   uniqueFilter(withSread = FALSE))
> aln <- readAligned(aFile, type="MAQMap", filter=filter)
```

Read extension

What is sequenced?

- 5' end of size-selected ChIP-enriched regions
- Upstream of actual binding site on plus strand, downstream on minus strand
- Strand-specific distribution reflects size-selected fragment lengths – e.g., left-skewed on plus strand

Consequence: extend reads in 3' direction

Read extension

Several possibile approaches (e.g., Kharchenko 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

Implemented as `estimate.mean.fraglen` in *chipseq*

Coverage and islands

Coverage

- Number of (extended) reads aligning over each nucleotide position

Islands

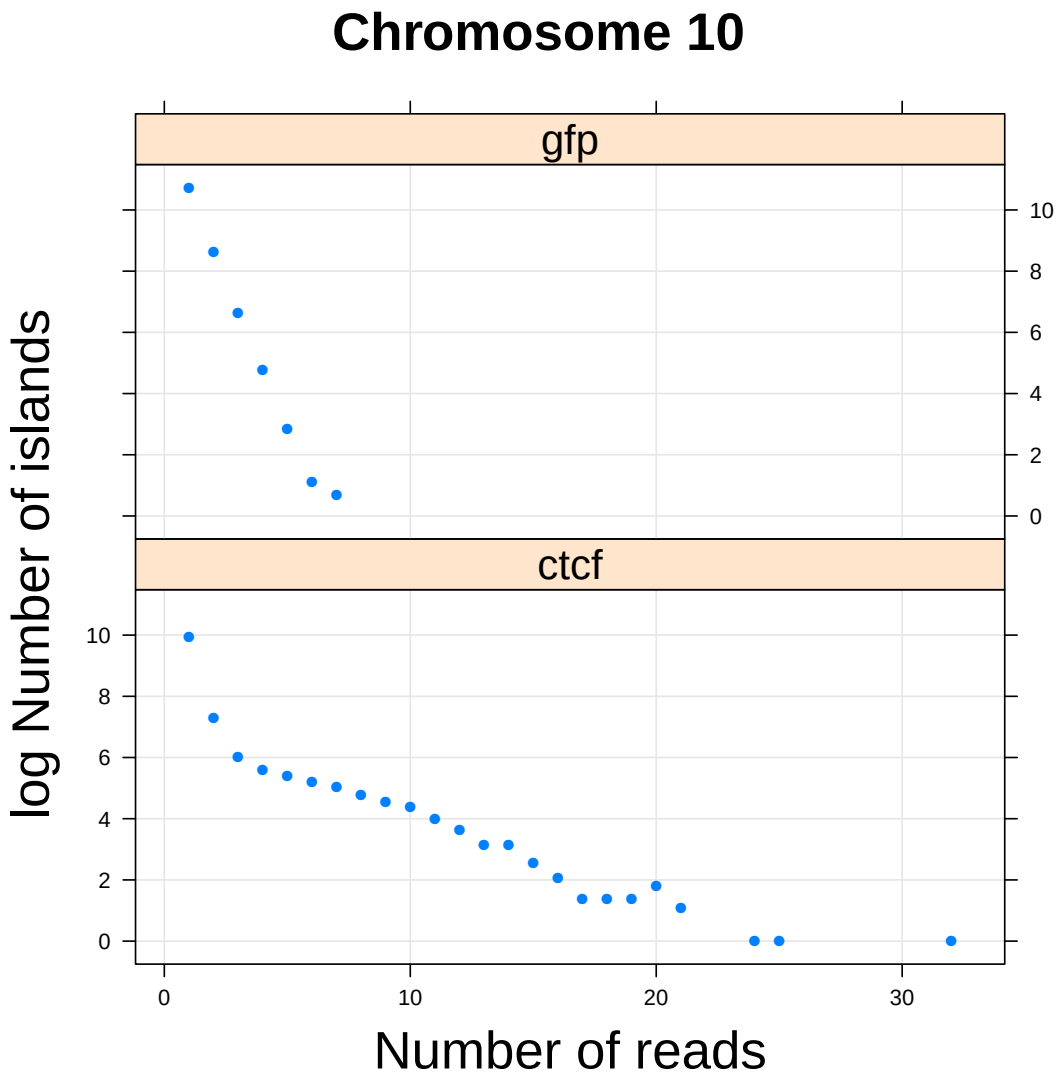
- Contiguous regions of non-zero coverage
- Characterize islands: area under the coverage curve, i.e., number of reads in the island

Coverage and islands

Psuedo-code

```
> cvg <- coverage(aln, extend=150L)
> islandReadSummary <- function(chr, islandDepth)
+ {
+     s <- slice(chr, lower=islandDepth)
+     tab <- table(viewSums(s) / 150L)
+     data.frame(nread=as.numeric(names(tab)),
+                count=as.numeric(tab))
+ }
> islands <- gdapply(cvg, islandReadSummary, islandDepth=1l
```

Coverage and islands



Differential peaks: Background versus signal

Null model $P(K = k) = p^{k-1}(1 - p)$

- Random sample of reads from mappable genome
- Coverage K, with probability p that a read starts at a given position
- Estimate p by assuming islands of depth 1 or 2 derive from the null

Background threshold

- Data usually show strong evidence of departure from null at $k >= 5$; we use $k >= 8$ below
- Model-based and adaptive algorithms areas of active research

```
> islands <- gdapply(cvg, islandReadSummary, islandDepth=8l
```

Differential peaks: case versus control

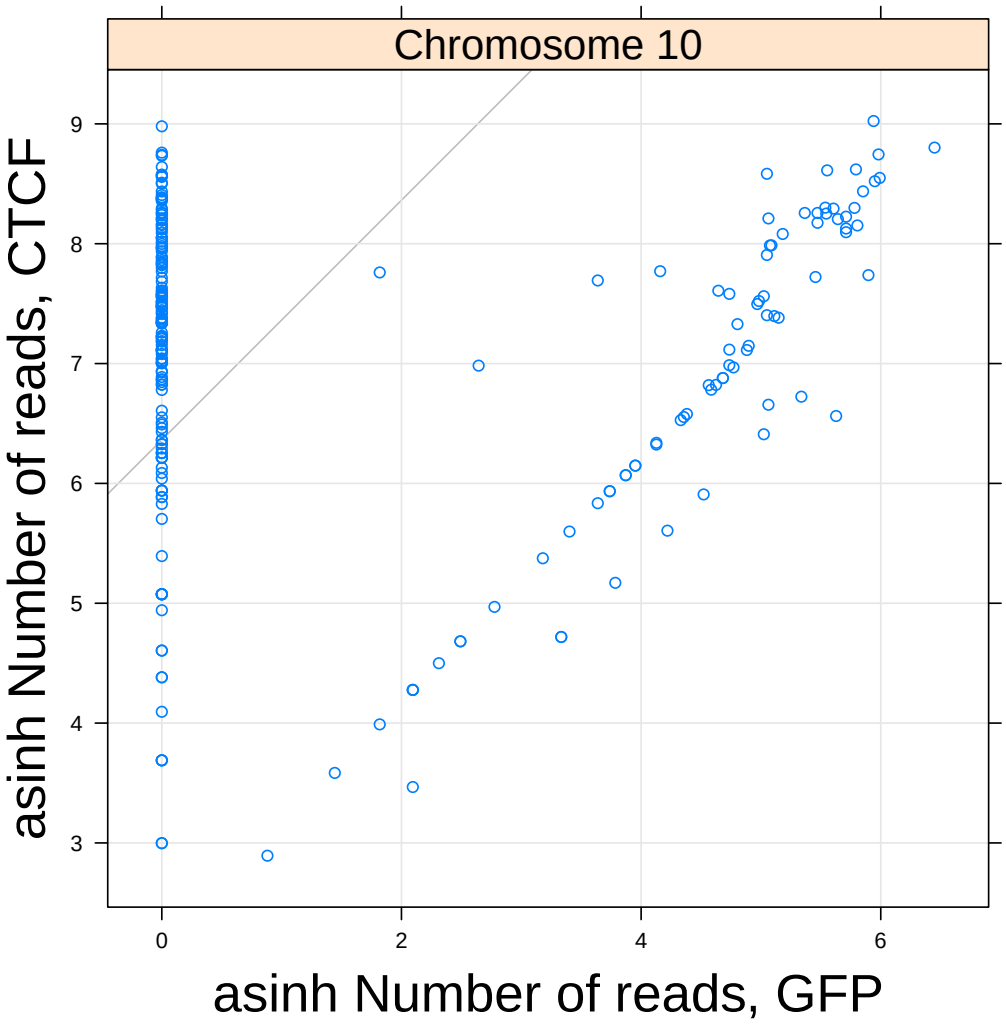
Challenges

- Between-lane variation in number of reads: artifact of sample preparation, or biologically relevant?
- What is a peak – present in one or both samples?

Possible solutions

- Combine lanes and identify peaks
- Compare contributions of each lane, relative to combined lane. `diffPeakSummary` in *chipseq*
- Estimate scaling constant c from robust regression of $y = cx \rightarrow \log y = \log c + \log x$

Differential peaks: case versus control



Differential peaks: designed experiments

Summarized read counts

- Matrix of islands $\times$ samples, values as read counts
- Possible to normalize (e.g., VSN)
- Extend modeling in standard ways, e.g., covariates such as local GC content

Statistical issues

- 'Peaks' are estimated, not defined *a priori*
- Data is count-based, not continuous
- Error model is not simply Poisson; see *edgeR*, *DESeq* for possible solutions

Additional analysis

- Motif exploration with *Biostrings* `matchPWM`
- Record multiple alignments with *Biostrings* `matchPDict`
- `contextDistribution`: overlap between discovered peaks and genomic features
- Export to genome browsers or otherwise visualize, e.g., using *rtracklayer*, *hilbertVis*, etc.,

```
> export(as(cvg[["chr10"]], "RangedData"), "chr10.wig")
```

Summary: a ChIP-seq work flow

- Identify appropriate reads, e.g., uniquely aligned singletons
- Calculate coverage, e.g., with extended reads
- Identify islands
- Restrict to islands above background
- Estimate differential representation
- Analyze designed experiments with linear models appropriate for count-based data

R and *Bioconductor* tools

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

References

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.

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.