Analysis of ChIP-seq data with R / Bioconductor

Martin Morgan
Bioconductor / Fred Hutchinson Cancer Research Center
Seattle, WA, USA

19 November, 2009

ChIP-seq

- Chromatin immunopreciptation to enrich sample DNA for sequences of interest
  - Typically: transcription factors bound to chromatin
  - Cross-link protein with DNA; sonicate ($< 1kb$); immunoprecipitate; DNA purification
- Sequencing
  - Process ChIP'ed DNA, e.g., size selection, adapter ligation
  - Perform whole-genome alignment
- Data analysis
  - Identify areas of high coverage – 'peaks'
  - Compare across experimental conditions

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, Cell 133: 1106-17. PMID: 18555785.
- Mouse embryonic stem cells transcription factor binding sites
- GFP: negative control; no peaks anticipated

Bioconductor tools

- Today's lab uses chipseq, ShortRead, Biostrings, IRanges, ...

Starting point: aligned reads I

Issues

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

Decisions

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

Starting point: aligned reads II

Psuedo-code

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

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

Several possible approaches, e.g.,

- Kharchenko et al., 2008, Nature Biotechnology 26: 1351-9
- Jothi et al., 2008, Nucleic Acids Research 36: 5221-31
- Implemented as `estimate.mean.fraglen` in chipseq

Identifying enriched regions: our approach I

Coverage

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

Islands

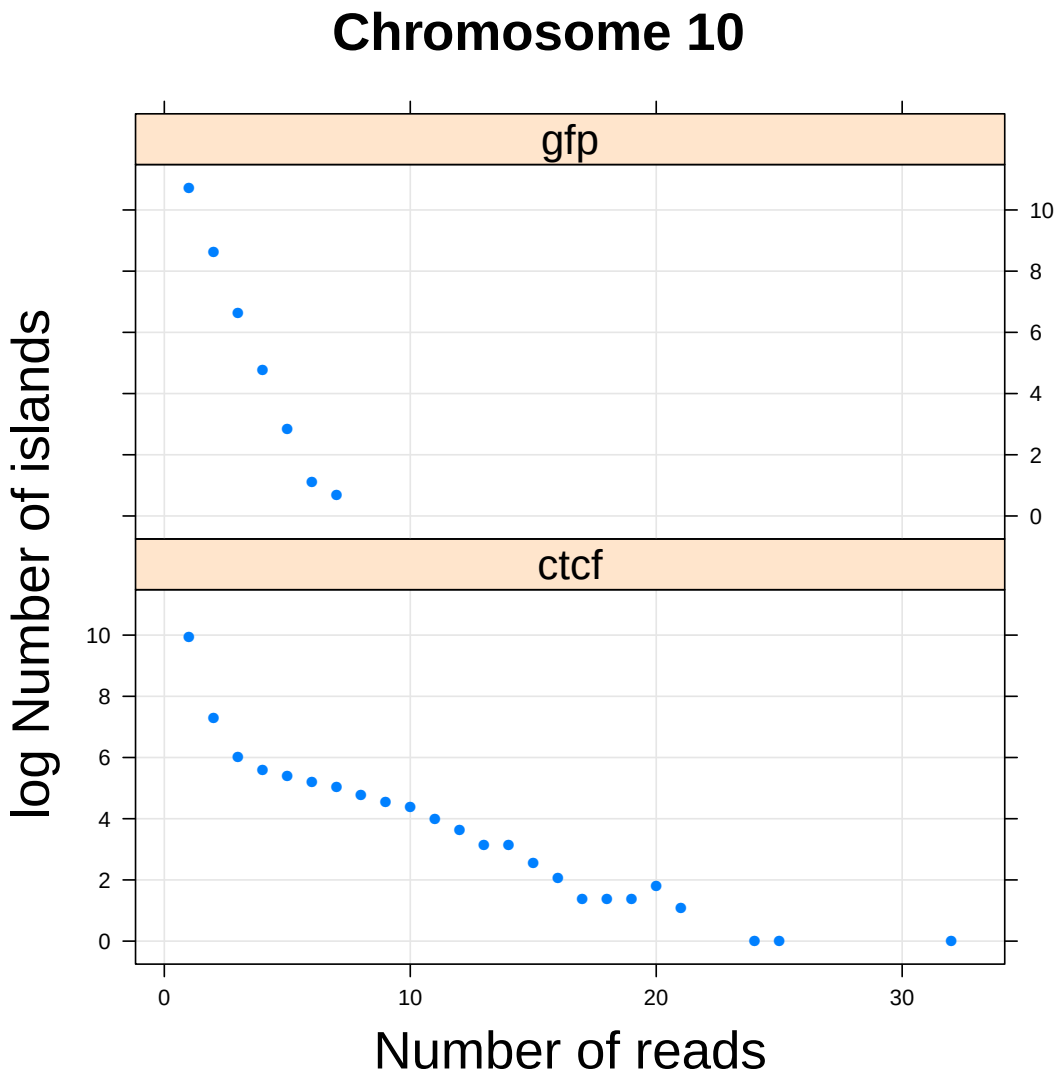
- Contiguous regions of non-zero coverage
- Characterize islands: area under the coverage curve, i.e., number of reads in the island

Identifying enriched regions: our approach II

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)
```

Island coverage



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)
```

Multiple lanes

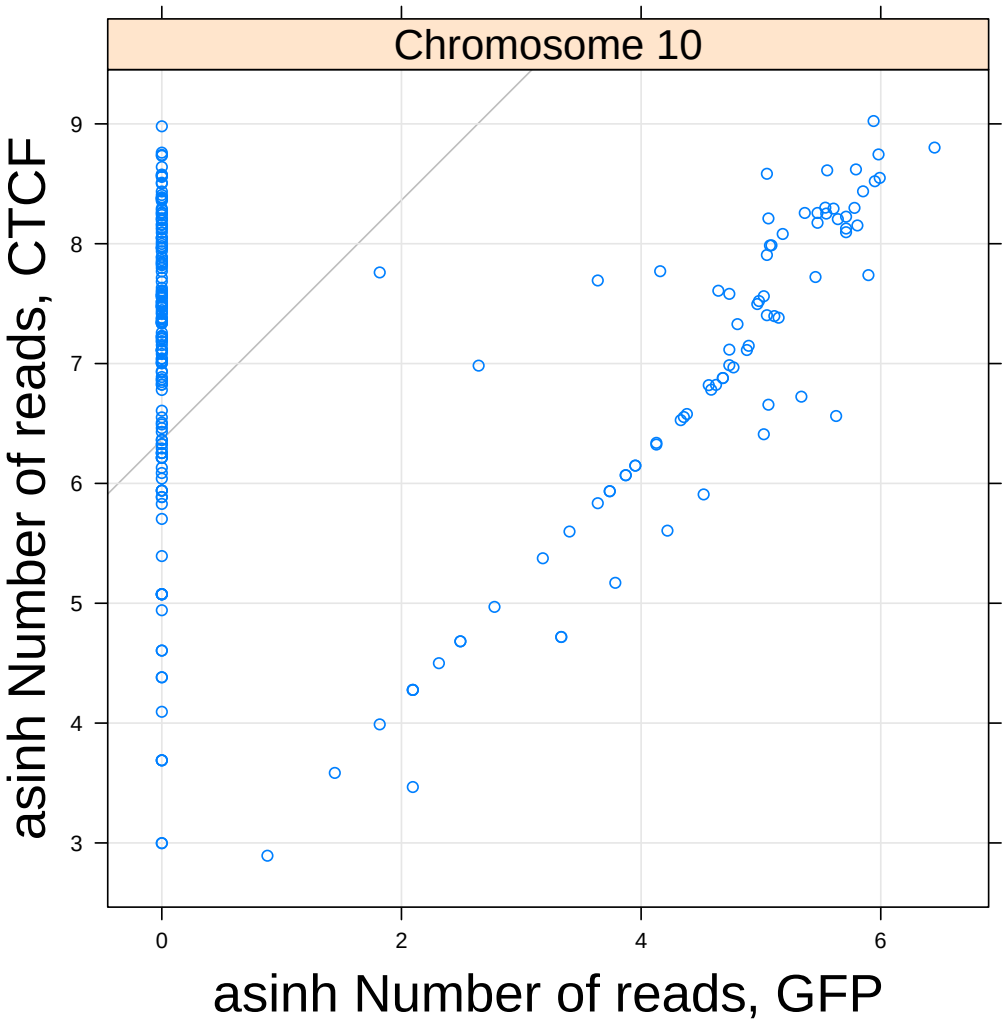
Challenges

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

Possible solutions

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

Island differential coverage



Designed experiments

Summarized read counts

- Matrix with rows being islands, columns be samples, values be read counts

Statistical issues

- 'Peaks' are estimated, not defined *a priori*
- Data is count-based, not continuous; see edgeR for one solution

Additional analysis

`contextDistribution`

- Overlap between discovered peaks and genomic features

Export to genome browsers or otherwise visualize

- Use rtracklayer, hilbertViz, etc., to visualize

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

Summary: an initial 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