

ChIP-seq

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

June 25, 2014

ChIP-seq

Chromatin immunoprecipitation, followed by sequencing

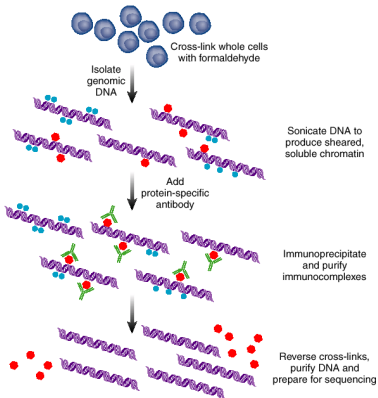
- ▶ Determine location of proteins bound to DNA

Useful for detecting

- ▶ Transcription factor binding sites
- ▶ Histone modification patterns

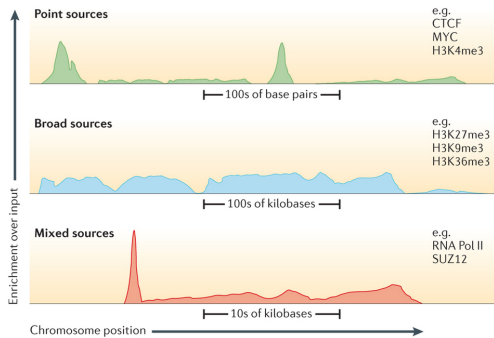
Common questions

- ▶ Which genes is this TF regulating?
- ▶ How do histone modifications affect expression?



Markus Bittig

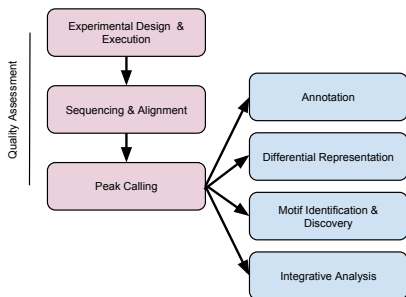
ChIP-seq: peak calling



Nature Reviews | Genetics

- ▶ Peaks and strand cross-correlation, Kharchenko et al. (2008)
- ▶ Broad vs. narrow peaks, Sims et al. (2014)

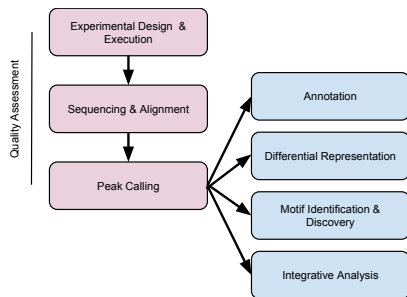
Work flow



Analysis overview

- ▶ Bailey et al. (2013)

Work flow: experimental design & execution



Analysis overview

- ▶ Bailey et al. (2013)

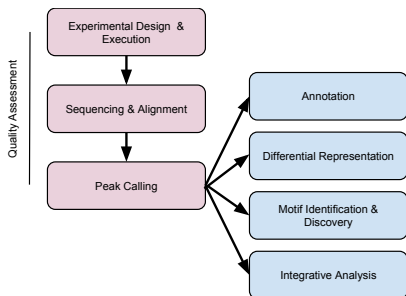
Single sample

- ▶ ChIPed transcription factor and...
- ▶ Input (fragmented genomic DNA) or control (e.g., IP with non-specific antibody such as immunoglobulin G, IgG)

Designed experiments

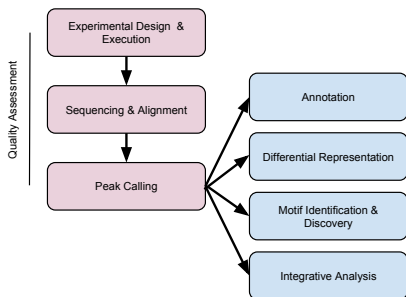
- ▶ Replication of TF / control pairs

Work flow: sequencing & alignment



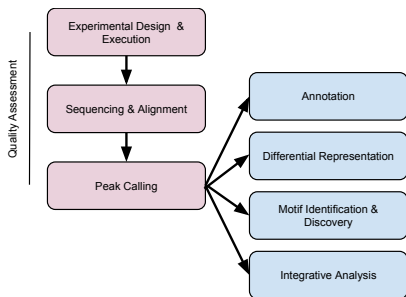
- ▶ Sequencing depth rules of thumb: $> 10M$ reads for narrow peaks, $> 20M$ for broad peaks
- ▶ Long & paired end useful but not essential – alignment in ambiguous regions
- ▶ Basic aligners generally adequate, e.g., no need to align splice junctions
- ▶ Sims et al. (2014)

Work flow: peak calling



- ▶ Very large number of peak calling programs; some specialized for e.g., narrow vs. broad peaks.
- ▶ Commonly used: MACS, PeakSeq, CisGenome, . . .

Work flow: down-stream analysis



- ▶ Annotation: what genes are my peaks near?
- ▶ Differential representation: which peaks are over- or under-represented in treatment 1, compared to treatment 2?
- ▶ Motif identification (peaks over known motifs?) and discovery
- ▶ Integrative analysis, e.g., association of regulatory elements and expression

Peak calling: MACS

MACS: Model-based Analysis for CHIP-Seq, Zhang et al. (2008)

<http://liulab.dfci.harvard.edu/MACS/>

- ▶ Scale control tag counts to match CHIP counts
- ▶ Center peaks by shifting $d/2$
- ▶ Model occurrence of a tag as a Poisson process
- ▶ Look for fixed width sliding windows with excess number of tag enrichment

Empirical FDR

- ▶ Swap CHIP and control samples; FDR is $\# \text{ control peaks} / \# \text{ CHIP peaks}$

Output: BED file of called peaks

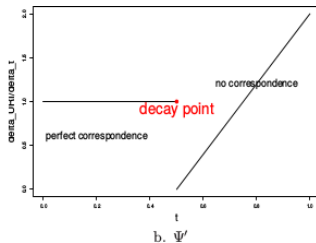
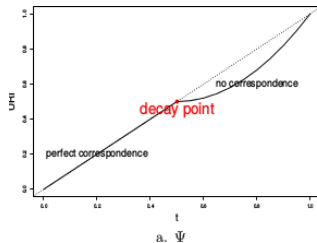
Peak calling: Irreproducible Discovery Rate

When replicates present:

- ▶ Peak callers often consistent on most confidently called peaks, but disagree on more ambiguous peaks
- ▶ When should one stop calling peaks?

Answer: Li et al. (2011) (also IDR101)

- ▶ Ranking of significance coupled with consistency between replicates



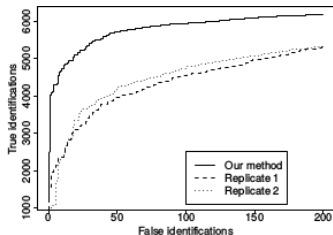
Peak calling: Irreproducible Discovery Rate

When replicates present:

- ▶ Peak callers often consistent on most confidently called peaks, but disagree on more ambiguous peaks
- ▶ When should one stop calling peaks?

Answer: Li et al. (2011) (also IDR101)

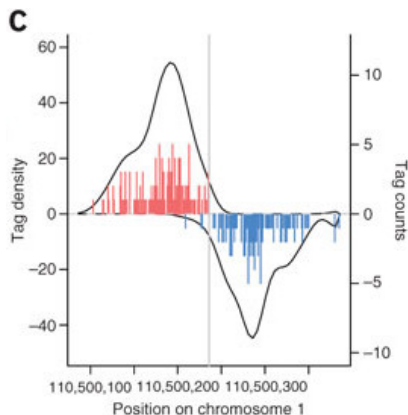
- ▶ Ranking of significance coupled with consistency between replicates



Quality Assessment

ENCODE guidelines: Landt et al. (2012)

- ▶ *Sequencing depth* relevant to TF site occupancy;
> 12M reads
- ▶ *Library complexity* diverse libraries indicate better sample prep, e.g., low complexity if original library contained only a few distinct reads
- ▶ *Cross-correlation* height: quality of ChIP; offset: length of fragments; 'phantom' peak: overlapping singletons

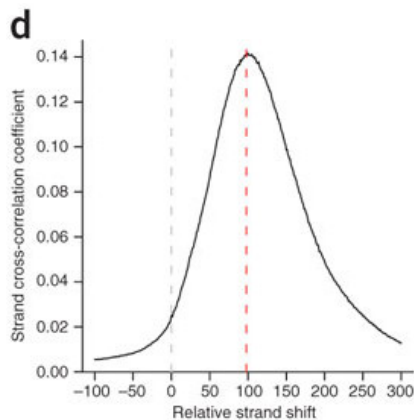


Kharchenko et al. (2008)

Quality Assessment

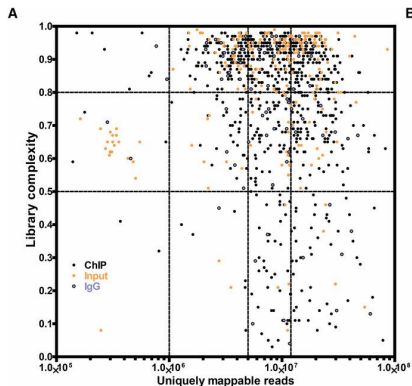
ENCODE guidelines: Landt et al. (2012)

- ▶ *Sequencing depth* relevant to TF site occupancy;
> 12M reads
- ▶ *Library complexity* diverse libraries indicate better sample prep, e.g., low complexity if original library contained only a few distinct reads
- ▶ *Cross-correlation* height: quality of ChIP; offset: length of fragments; 'phantom' peak: overlapping singletons



Kharchenko et al. (2008)

Quality Assessment



Marinov et al. (2014)

- ▶ Large-scale assessment of published ChIP-seq experiments
- ▶ 191 GEO experiments
- ▶ 55% highly successful; 20% poor

Quality Assessment: *ChIPQC*

Inputs: BAM files (raw data) and BED files (called peaks)

```
experiment <- ChIPQC(samples)
ChIPQCreport(experiment)
```

Output: HTML report — <http://starkhome.com/ChIPQC/Reports/tamoxifen/ChIPQC.html>

Annotation: *ChIPpeakAnno*

Inputs

- ▶ Peaks: *RangedData* (*GRanges*-like) peaks, e.g., from `rtracklayer::import()` BED files
- ▶ Annotation: *RangedData* representing gene boundaries, or query to *biomaRt*

```
library(ChIPpeakAnno)
## ...
annotated <- annotatePeakInBatch(peaks,
  AnnotationData=annotation)
```

Output: *RangedData* with annotations about near-by peaks.

Differential Representation: *DiffBind*

Inputs: called peaks and raw BED or BAM files

```
library(DiffBind)
tamoxifen = dba(sampleSheet="tamoxifen.csv")
tamoxifen = dba.count(tamoxifen)
tamoxifen = dba.contrast(tamoxifen,
  categories=DBA_CONDITION)
tamoxifen = dba.analyze(tamoxifen)
tamoxifen.DB = dba.report(tamoxifen)
```

Outputs: diagnostics, visualizations, and 'top table' of differentially expressed regions.

Motifs

Identification

- ▶ JASPAR and other motif catalogs
- ▶ Position Weight Matrix describing probability of nucleotide(s) at each position
- ▶ Scan genome / under peaks for known motifs
- ▶ *MotifDb*, `matchPWM` (*Biostrings*);
- ▶ FIMO, etc

Discovery

- ▶ Collate sequences under peaks, search for recurrent sequences
- ▶ e.g., DREME / MEME-ChIP

Also: enrichment, regulatory modules (2+ motifs co-occurring), function, ...

ChIP-seq in *Bioconductor*: resources

- ▶ EdX MOOC 'Data Analysis for Genomics', chapter on ChIP-seq analysis
- ▶ biocViews terms: ChIPSeq, MotifAnnotation, MotifDiscovery
- ▶ Work flows: Candidate Binding Sites for Known Transcription Factors

ChIP-seq in *Bioconductor*: packages

Sample packages

- ▶ Quality assessment – *ChIPQC*;
- ▶ (Peak calling) – *chipseq*, *PICS*, *triform*, *ChIPseqR*, *iSeq*, ...
- ▶ Single sample summary / exploration – *ChIPpeekAnno*, *chIPseeker*
- ▶ Differential representation – *DiffBind*, *MMDiff*, ...
- ▶ Motifs – *MotifDb*, *TFBSTools* (matching known motifs), *motifRG*, *MotIV*, *rGADEM* *BCRANK* (motif discovery)
- ▶ Integration with expression data – *Rcade*, *epigenomix*

References I

- T. Bailey et al. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS Comput Biol*, 9(11):e1003326, 11 2013. doi: 10.1371/journal.pcbi.1003326.
- P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, 26(12):1351–1359, Dec 2008.
- S. G. Landt et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9): 1813–1831, Sep 2012. doi: 10.1101/gr.136184.111.
- Q. Li, J. B. Brown, H. Huang, and P. J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 09 2011. doi: 10.1214/11-AOAS466.
- G. K. Marinov, A. Kundaje, P. J. Park, and B. J. Wold. Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, 4(2):209–223, Feb 2014.

References II

- D. Sims, I. Sudbery, N. E. Illott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–132, Feb 2014. doi: 10.1038/nrg3642.
- Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.