ChIP-seq/Functional Genomics/Epigenomics

CBSU/3CPG/CVG Next-Gen Sequencing Workshop

Josh Waterfall

March 31, 2010
Outline

Introduction to ChIP-seq
  Control data sets
  Peak/enriched region identification
Related functional genomics assays
Useful web resources
Chromatin ImmunoPrecipitation (ChIP-seq)

Advantages over array-based methods

No cross-hybridization background
Lower end of sensitivity largely dependent on just sequencing depth
More linear, quantitative assay
Unmappable portion of genome is distinct from, and much smaller than, repeat masked portion
No limit based on probe locations
Needs less starting material
Higher resolution
Important Issues:

Appropriate Controls
Identifying Enriched Regions
Controls

Sequencing is such high-sensitivity that signals invisible in any other assay are now apparent. Need rigorous controls to be confident of enrichment.

- Input DNA has non-random pattern (open chromatin shears more easily) - Sono-seq is an actual assay.
- Mock-IP controls for more steps in ChIP protocol than input DNA but not antibody cross-reactivity.
- Different antibodies (to different epitopes) in separate experiments, or ChIP after target protein has been depleted (or in cell-line without tagged protein), help control for cross-reactivity.

- To characterize new antibody, IP and mass-spec everything that comes down to verify only expected binding partners are seen.
Background rate is non-uniform

Auerbach R K et al. PNAS 2009;106:14926-14931
Identifying enriched regions

Identifying narrow ChIP peaks is very different from:
identifying broadly enriched regions in ChIP.
identifying narrow peaks from other assays.

• Many well-worked out programs for identifying narrow ChIP peaks (e.g. from sequence specific binding factors).
  The best programs can exploit strand-specific patterns,
  local versus global background levels, mappability, etc.
• Other assays also result in localized peaks but without same strand pattern.
• Most work on broad regions (e.g. particular histone modifications or RNA polymerase) is based on sliding windows (either of fixed length or fixed read count).
Strand specific pattern for localized peaks in ChIP-seq
Is strand information used?
If fragment size is used, is it defined by user or estimated from data?
(How) is control data incorporated?
Is background defined locally or globally?
(How) are unmappable regions treated?

Table 3 | Publicly available ChIP-seq software packages discussed in this review.

<table>
<thead>
<tr>
<th>Software</th>
<th>Profile</th>
<th>Peak criterion</th>
<th>Tag shift</th>
<th>Control data</th>
<th>Rank by</th>
<th>User input</th>
<th>Artifact filtering</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastPeaks v3.2</td>
<td>Aggregation of overlapping tags</td>
<td>Height threshold</td>
<td>Input or estimated</td>
<td>NA</td>
<td>Binary log odds ratio with empirical FDR</td>
<td>2</td>
<td>None</td>
<td>Y/Yes</td>
</tr>
<tr>
<td>F-seq v3.82</td>
<td>Kernel density estimation (KDE)</td>
<td>s.d. above KDE (with background)</td>
<td>Input or estimated</td>
<td>KDE for local background</td>
<td>Peak height</td>
<td>2</td>
<td>None</td>
<td>No/No</td>
</tr>
<tr>
<td>GLISTR</td>
<td>Aggregation of overlapping tags</td>
<td>Classification by height and relative enrichment</td>
<td>User input tag extension</td>
<td>Multiplied sampled to estimate background (ta) values</td>
<td>Peak height and fold enrichment</td>
<td>2</td>
<td>control &amp; CISP</td>
<td>No/No</td>
</tr>
<tr>
<td>MACS v3.5.5</td>
<td>Tags, shifted over window scan</td>
<td>Local region Binomial P value</td>
<td>Estimate from high-quality peaks</td>
<td>Input tag extension length</td>
<td>Used for Rosetta filter when available</td>
<td>2</td>
<td>None</td>
<td>P-value threshold, tag length, signal for strand assignment</td>
</tr>
<tr>
<td>PeakSeq</td>
<td>Extended tag aggregation</td>
<td>Local region Binomial P value</td>
<td>Estimate from high-quality peaks</td>
<td>Input tag extension length</td>
<td>Used for significance of sample enrichment with binomial distribution</td>
<td>P-value</td>
<td>2</td>
<td>None</td>
</tr>
<tr>
<td>QuEST v2.3</td>
<td>Kernel density estimation</td>
<td>2</td>
<td>Height threshold, background ratio</td>
<td>KDE for enrichment and empirical FDR estimation</td>
<td>P-value</td>
<td>2</td>
<td>None</td>
<td>Na, control &amp; CISP</td>
</tr>
<tr>
<td>SICER v1.5</td>
<td>Window scan with gaps allowed</td>
<td>P-value from random background with enrichment relative to control</td>
<td>Input</td>
<td>Linearly scaled for candidate peak ( b )-values</td>
<td>q-value</td>
<td>1</td>
<td>None</td>
<td>1</td>
</tr>
<tr>
<td>SICER v1.4</td>
<td>Window scan</td>
<td>( \beta ), ( \sigma ) and ( \rho ) thresholds in peaks</td>
<td>Average nearest paired tag distance</td>
<td>Used to compute enrichment distribution</td>
<td>P-value</td>
<td>1</td>
<td>None</td>
<td>Na, control &amp; CISP</td>
</tr>
<tr>
<td>sSeq v5.0</td>
<td>Strand specific window scan</td>
<td>( P )-values (paired peaks only)</td>
<td>Maximal strand correlation</td>
<td>Subtracted before peak calling</td>
<td>P-value</td>
<td>2</td>
<td>None</td>
<td>Na, control &amp; CISP</td>
</tr>
<tr>
<td>UltraSeq v4.2</td>
<td>Window scan</td>
<td>Binomial P-value</td>
<td>Estimated or user specified</td>
<td>Subtracted before peak calling</td>
<td>P-value</td>
<td>2</td>
<td>P-value, control &amp; CISP</td>
<td>Target FDR</td>
</tr>
</tbody>
</table>

ChIP-seq peak calling software contest on http://seqanswers.com

Assessing saturation and significance

Saturation is dependent on both ChIP efficiency (globally) and factor:DNA affinity (locally)

Significance depends on both ChIP and control read counts

ChIP-seq read density is a quantitative measure of binding level.

Solutions to GC bias

Improve Illumina protocol (especially gel purification)

Use single molecule sequencing technology


Fragmenting DNA

AFA (Adaptive Focused Acoustics)/Covaris enzyme (MNase, NEB Fragmentase)

Nebulization

Sonication

Quality Evaluation with Agilent Bioanalyzer

Fragmenting DNA

Open chromatin fragments more easily than closed chromatin - ChIP for factors associated with transcriptional repression or heterochromatin can be more difficult than for factors associated with transcriptional activation.

Related Functional Genomic Assays

- Chromatin ImmunoPrecipitation (ChIP)
- DNAse I Hypersensitivity
- Formaldehyde Assisted Isolation of Regulatory Elements
- DNA Methylation (bisulfite, affinity, or restriction enzyme based)
- Cap Analysis of Gene Expression (CAGE)
- Genomic Run On (GRO-seq)
- Many more…

Each can couple to next-gen sequencing and entails analysis/identification of enriched regions
High-throughput methods to identify DNase HS sites.

1) Digest with DNase and blunt end
2) Ligate biotinylated linkers
3) Sonicate to shear DNA
4) Enrich on Streptavidin column
5) Add second linkers, amplify

Label and Hybridize to Tiled Arrays

Raw DNase-chip data

OR

Sequence Using Solexa Platform

DNase-sequences

http://www.genome.duke.edu/people/faculty/crawford/research/
Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE)

Micrococcal Nuclease digestion (MNase-seq)

Weiner A et al. Genome Res. 2010;20:90-100
DNA methylation

3 broad classes of assays:
Enzyme based
  methylation sensitive restriction enzymes
Affinity based
  antibodies or other meDNA binding proteins used in ChIP-like experiment
Bisulfite based
  NaHSO$_3$ deaminates unmethylated cytosines to uracils but does not affect 5-methylcytosine.
Reduced Representation based on enzymatic digest or hybridization enrichment common for cost efficiency in large, sparsely methylated (e.g. mammalian) genomes.
Read alignments done to in silico bisulfite-converted genome.

Direct sequencing of 5th base (future technologies)
Cap Analysis of Gene Expression (CAGE)

http://www.osc.riken.jp/english/activity/cage/basic/
Genomic Run On (GRO-seq)

Web resources for analyzing, viewing, sharing, and collecting genomics data

UCSC Genome Browser ([http://genome.ucsc.edu](http://genome.ucsc.edu))
Galaxy ([http://main.g2.bx.psu.edu](http://main.g2.bx.psu.edu))
The UCSC Genome Browser

http://genome.ucsc.edu

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of in situ mouse and frog images to examine expression patterns. Genome Graphs allows you to upload and display genome-wide data sets.

The UCSC Genome Browser is developed and maintained by the Genome Bioinformatics Group, a cross-departmental team within the Center for Biomolecular Science and Engineering (CBSIE) at the University of California Santa Cruz (UCSC). If you have feedback or questions concerning the tools or data on this website, feel free to contact us on our public mailing list.

If you use ENCODE or modENCODE data, or are interested in exploring it in the future, we invite you to take the 2010 ENCODE/modENCODE Usability Survey. Your input will help us to make this data more accessible to the scientific community. Thank you!

News

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the genome-announce mailing list.

24 Mar. 2010 – African Savannah Elephant Genome Browser Released

We have released a Genome Browser for the African savannah elephant, Loxodonta africana. This assembly (UCSC version loxAfr3, Broad lloxAfr3) was produced by the Broad Institute, Cambridge, MA, USA. The elephant was the first member of Afrotheria to be sequenced. Afrotheria is the deepest node of Eutheria, and the elephant sequence should be useful in reconstructing the ancestral eutharian genome.

This draft of the elephant genome has a size of approximately 3 Gb with 7X coverage. The assembly comprises 2352 scaffolds and chrM (mitochondrial DNA). For more information on the assembly, see the Broad Institute Elephant Genome Project page.

Bulk downloads of the sequence and annotation data are available via the Genome Browser FTP server or the Downloads page. These data have specific conditions for use.

Many thanks to the Broad Institute for the elephant assembly data. The lloxAfr3 annotation tracks were generated by UCSC and collaborators.
Uploading data to UCSC

### UCSC Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz. Software Copyright © The Regents of the University of California. All rights reserved.

<table>
<thead>
<tr>
<th>clade</th>
<th>genome</th>
<th>assembly</th>
<th>position or search term</th>
<th>gene</th>
<th>image width</th>
</tr>
</thead>
</table>

Click here to reset the browser user interface settings to their defaults. Take ENCODE Survey

### About the Human Mar. 2006 (NCBI36/hg18) assembly (sequences)

The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium.

### Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

**Request:**

| chr7 | Displays all of chromosome 7 |
| 20p13 | Displays region for band p13 on chr 20 |
| chr3:1-1000000 | Displays first million bases of chr 3, counting from p-arm telomere |
| chr3:1000000+2000 | Displays a region of chr3 that spans 2000 bases, starting with position 1000000 |
| RH18061:RH80175 15q11:15q13 | Displays region between STS markers RH18061 and RH80175 or chromosome bands 15q11 to 15q13. This syntax may also be used for other range queries, such as between uniquely-determined ESTs, mRNAs, refSeqs, etc. |
Uploading data to UCSC

Add Custom Tracks

Display your own data as custom annotation tracks in the browser. Data must be formatted in BED, bigBed, BEDGRAPH, GFF, GTF, WIG, bigWig, MAF, BAM or PSL formats. To configure the display, set track and browser line attributes as described in the User's Guide. URLs for data in the bigBed and bigWig formats must be embedded in a track line in the box below. Publicly available custom tracks are listed here. Examples are here.

Optional track documentation: Or upload: Choose File no file selected

Click here for an HTML document template that may be used for Genome Browser track descriptions.

Loading Custom Tracks

An annotation data file in one of the supported custom track formats may be uploaded by any of the following methods:

- (Preferred) Enter one or more URLs for custom tracks (one per line) in the data text box. The Genome Browser supports both the HTTP and FTP (passive-only) protocols.
- Click the "Browse" button directly above the URL/data text box, then choose a custom track file from your local computer, or type the pathname of the file into the "upload" text box adjacent to the "Browse" button. The custom track data may be compressed by any of the following programs: gzip (.gz), compress (.Z), or bzip2 (.bz2). Files containing compressed data must include the appropriate suffix in their names.

---
Uploading data to UCSC

Manage Custom Tracks

Name | Description | Type | Doc | Items | Pos | delete | add custom tracks | go to genome browser
-- | -- | -- | -- | -- | -- | -- | -- | --
MACS peaks for HeLa_STAT | MACS peaks for HeLa_STAT | bed | 5902 | chr1 | | | | |

Managing Custom Tracks

This section provides a brief description of the columns in custom track management table. For more details about managing custom tracks, see the Genome Browser User's Guide.

- **Name** - a hyperlink to the update page where you can edit your track data.
- **Description** - the value of the "description" attribute from the track line, if present. If no description is included in the input file, this field contains the track name.
- **Type** - the track type, determined by the Browser based on the format of the data.
- **Doc** - displays "Y" (Yes) if a description page has been uploaded for the track; otherwise the field is blank.
- **Items** - the number of data items in the custom track file. An item count is not displayed for tracks lacking individual items (e.g. wiggle format data).
- **Pos** - the default chromosomal position defined by the track file in either the browser line "position" attribute or the first data line. Clicking this link opens the Genome Browser or Table Browser at the specified position (note: only the chromosome name is shown in this column). The Pos column remains blank if the track lacks individual items (e.g. wiggle format data) and the browser line "position" attribute hasn't been set.
Modifying display settings
Modifying display settings

HeLa_STAT1_control_chr1 Track Settings

Shifted Merged MACS tag counts for every 10 bp

- **Display mode:**
  - Full
  - Submit
  - Remove custom track
  - Update custom track

- **Type of graph:** Bar

- **Track height:** 128 pixels (range: 11 to 128)

- **Vertical viewing range:**
  - min: 1
  - max: 122 (range: 1 to 122)

- **Data view scaling:**
  - auto-scale to data view
  - Always include zero: Off

- **Transform function:**
  - Transform data points by: NONE

- **Windowing function:**
  - Mean ± whiskers

- **Smoothing window:**
  - Off
  - pixels

- **Draw y indicator lines:**
  - at y = 0: Off
  - at y = 0: Off

*Graph configuration help*
Using publicly available tracks
Using publicly available tracks
Galaxy
http://main.g2.mx.psu.edu
NCBI Gene Expression Omnibus (GEO)
ChIP-seq exercise

Use MACS software to call peaks in STAT1 ChIP-seq experiment in human HeLa S3 cells after interferon-γ stimulation. Analyze diagnostics of run and upload data to UCSC genome browser to look at results.

Office hour:
Friday, April 2
1pm - 2pm
102 Weill conference room