ChIP-seq data analysis

04-05-12
Outlook

- Friday 04-05-12:
  - Next-generation sequencing
  - ChIP-seq
    - experimental design
  - ChIP-seq data analysis:
    - Mapping of sequenced reads to a reference genome
    - Peak calling
    - Peak annotation
    - Discovery of transcription factors sequence motifs

- Friday 11-05-12
  - Practical: ChIP-seq data analysis
Next generation sequencing course, 12th-14th March 2012

Harrold swerdlow, Head of R&D, WTSI
Remco loos and Myrto Kostadima, from EBI
Next-gen Rationale
Capillary Sample Prep

Fragment genome

Clone into bacterial vector

Grow and purify

Harrold swerdlow slide
Capillary Sequencing

Prime

Extend with A, C, G, T terminators

AACGT . . .

Separate by size and detect

Harrold swerdlow slide
Capillary Reactions

1 tube
1 template → 1 capillary
1000 bases

Harrold swerdlow slide
Next-Generation Sample Prep

[Amplify] fragments directly on a surface (bead, chip, etc.)

Harrold swerdlow slide
Sequencing by Synthesis

- Extend by 1 base
- Image
- Reverse termination
- Repeat

Harrold swerdlow slide
Next-Generation Reactions

1 feature
1 template

1 chip
gigabases
The Next-Generation Process

DNA Prep → Library Prep → Chip Prep → Sequencing → Analysis
Illumina Technology
Library Prep

T4 DNA Ligase

Hybridize primers

Limited PCR

Make clusters and sequence

(x2)

Harrold swerdlow slide
Cluster Amplification

Single-molecule array

Cluster ~1000 molecules

1 billion clusters on a single glass chip

Harrold swerdlow slide
Sequencing by Synthesis

Cycle 1

Enzyme

Harrold swerdlow slide
Wash + Detect Fluorescence

Cycle 1

Harrold swerdlow slide
Prepare for Next Cycle

Cycle 1

Removal of fluorescence and reversal of termination

Repeat

Sequence output

Mol a = C

Harrold swerdlow slide
Four Colour Composite

100 MICRONS

20 MICRONS

ACGT

Harrold Swerdlow slide
Base Calling From Raw Data

TGCTACGAT...

TTTTTTTTGT...

Harrold swerdlow slide
Billions of Bases of DNA Sequence (per instrument)

» 8 lanes per chip
» 48 tiles (6 swaths) per lane
» 4,000,000 clusters per tile
» 200 cycles (2 x 100) in 10 days
» $8 \times 48 \times 4,000,000 \times 200 = 300 \text{ Gb}$
» 2 chips = 600 Gb / run = 6 Genomes

Harrold swerdlow slide
• Illumina solexa sequencing video!
Next-generation sequencing applications

• Genome applications:
  • ChIP-seq: TF binding sites, histone modifications, nucleosome positions mapping
  • Dnase-seq: DNA accessibility,
  • Methyl-seq: methylome characterisation
  • Variant discovery: SNPs,
  • De novo genome assembly

• Transcriptome applications:
  • Quantification of gene expression
  • Differential gene expression
  • De novo transcript discovery
  • Detection of aberrant transcripts
<table>
<thead>
<tr>
<th></th>
<th>ChIP-chip</th>
<th>ChIP-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
<td>Array-specific</td>
<td>High - single nucleotide</td>
</tr>
<tr>
<td>Coverage</td>
<td>Limited by sequences on the array</td>
<td>Limited by “alignability” of reads to the genome, increases with read length</td>
</tr>
<tr>
<td>Repeat elements</td>
<td>Masked out</td>
<td>Many can be covered (40% of human genome is repetitive but 80% is uniquely mappable)</td>
</tr>
<tr>
<td>Cost</td>
<td>400-800$ per array (1-6M probes), multiple arrays needed for human genome</td>
<td>Around 1000$ per lane; 20-30M reads</td>
</tr>
<tr>
<td>Source of noise</td>
<td>Cross hybridization</td>
<td>Sequencing bias, GC bias, sequencing error</td>
</tr>
<tr>
<td>Amount of ChIP DNA required</td>
<td>High, few micrograms</td>
<td>Low 10-50ng</td>
</tr>
<tr>
<td>Dynamic range</td>
<td>Lower detection limit and saturation at high signal</td>
<td>Not limited</td>
</tr>
<tr>
<td>Multiplexing</td>
<td>Not possible</td>
<td>Possible</td>
</tr>
</tbody>
</table>
Overview of ChIP-seq experiments

The basic concepts are similar to next-generation sequencing. A ChIP–seq experiment begins with sample preparation, where samples are fragmented and immunoprecipitated using antibodies specific to the protein of interest. After DNA purification, end repair and adaptor ligation are performed. Clustering and amplification steps are carried out using various technologies, such as Illumina Solexa Genome Analyzer, Roche 454 platform, or Helicos by Helicos. The sequencing step involves the enzyme-driven amplification of all templates in parallel. After each extension, the fluorescent labels that are specific to the ChIP–seq experiment are attached to the DNA fragments. The sequencing errors have been reduced substantially as the costs of sequencing per base pair have not affected ChIP–seq.
ChIP-seq experimental design

- Antibody quality
- Control experiment
- Depth of sequencing
- Multiplexing
- Sequencing options:
  - Paired-end or single-end reads
  - 36bp reads or longer
Antibody quality

- A sensitive and specific antibody will give a high level of enrichment
- Limited efficiency of antibody is the main reason for failed ChIP-seq experiments
- Check your antibody ahead if possible. Western blotting to check the cross-reactivity of the antibody
Control experiment

- A ChIP-seq peak should be compared with the same region in a matched control

- Open chromatin regions are fragmented more easily than closed regions

- There is amplification and size selection bias during library preparation

- Repetitive sequences might seem to be enriched (inaccurate repeats copy number in the assembled genome)

Rozowski 2009, nature Biotechnology
Control type

- Input DNA

- Mock IP - DNA obtained from IP without antibody
  - Very little material can be pulled down leading to inconsistent results of multiple mock IPs.

- Nonspecific IP - using an antibody against a protein that is not known to be involved in DNA binding

- There is no consensus on which is the most appropriate

- Sequencing a control can be avoided when looking at:
  - time points
  - differential binding pattern between conditions
Depth of sequencing

More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth.

Number of putative target regions continues to increase significantly as a function of sequencing depth.

With current sequencing technologies, one lane is usually sufficient.

Park J 2009, Nature Reviews, Genetics
# Saturation-MACS « diag »

## Table

<table>
<thead>
<tr>
<th>FC</th>
<th># peaks</th>
<th>90%</th>
<th>80%</th>
<th>70%</th>
<th>60%</th>
<th>50%</th>
<th>40%</th>
<th>30%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20</td>
<td>31530</td>
<td>75.01</td>
<td>55.98</td>
<td>39.58</td>
<td>26.01</td>
<td>15.35</td>
<td>7.43</td>
<td>2.64</td>
<td>0.51</td>
</tr>
<tr>
<td>20-40</td>
<td>5481</td>
<td>99.62</td>
<td>97.7</td>
<td>92.52</td>
<td>80.46</td>
<td>61.34</td>
<td>36.75</td>
<td>14.61</td>
<td>2.81</td>
</tr>
<tr>
<td>40-60</td>
<td>235</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.57</td>
<td>90.21</td>
<td>68.51</td>
<td>28.09</td>
</tr>
<tr>
<td>60-80</td>
<td>40</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>95</td>
<td>62.5</td>
<td></td>
</tr>
<tr>
<td>80-100</td>
<td>7</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>85.71</td>
</tr>
<tr>
<td>100-120</td>
<td>2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>120-140</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>160-180</td>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Sequencing options

- **Pared-ends vs single-end:**
  - DNA fragments are sequenced from both ends
  - Costs twice as much as single-end sequencing
  - Increase « mappability » of reads specially in repetitive regions
  - For ChIP-seq, usually not worth the extra cost, unless you have a specific interest in repeat regions

- **Short vs long reads:**
  - For ChIP-seq of 36 bp single-end reads are sufficient
Overview of ChIP-seq analysis

1. Check sequencing depth
2. Sequencing platform → Image analysis (base calling) → 35–50 bp sequences → Genome alignment → Peak calling
3. Quality scores → Control sample
4. Enriched regions → Visualization with genome browser → Motif discovery
5. Relationship to gene structure → Integration with gene expression → Gene set analysis → Other advanced analysis → Differential profile analysis
Fasq format

@HWUSI-EAS100R:6:73:941:1973#0/1

GATTTGGGTCTAAGCAGTGATGATGAAATAGTAAATCCATTTGTCTACATCACTAGTT
+
!"*(((((****))**++)(**++).1***+-**'))**55CCF>>><<<CCCCCCCC65

6 - Flowcell lane
73 - Tile number
941,1973 - ‘x’,’y’-coordinates of the cluster within the tile
#0 - index number for a multiplexed sample (0 for no indexing)
/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)
# Phred quality score

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

A Phred score of a base:

\[ Q \text{ phred} = -10 \times \log_{10}(e) \]

where \( e \) is the estimated probability of a base being wrong.

For example: If a base is estimated to have a 0.1% chance of being wrong, it gets a Phred score of 30
Mapping of sequenced reads

- ELAND-provided with Illumina sequencer
  - Limited reads length
  - Allow 2 substitutions

- MAQ
  - Uses quality values
  - Integrate consensus calling

- Bowtie
  - Ultrafast
  - Can work on workstations with < 2 Gb memory

- Many others: BWA, Novoalign, BFAST,...
Mapping challenges

- Enormous amount of short reads against large genomes
- Presence of repetitive regions, pseudogenes

Mismatch:
  - Allow or not
  - SNP or sequencing errors
  - Insertion/deletion

- Multiple reads: reads that map to more than one genomic location

- Software challenges:
  - Balance between speed, precision and memory usage
Strand specific profile at enriched sites

Park J 2009, Nature Reviews, Genetics
Peak calling

- CisGenome:
  - Peak criteria: number of reads in windows and number ChIP read minus control reads

- ERANGE:
  - High quality peak estimate

- MACS:
  - Poisson P value estimate

- Many others: FindPeaks, QuEST...
Peak calling - Challenges

Park J 2009, Nature Reviews, Genetics
MACS tool

- Model the shift size between +/- strand tags:

- Scan the genome to find regions with tags more than mfold enriched relative to random tag distribution.

- Randomly sample 1000 of these (high quality peaks) and calculate the distance between the modes of their +/- peaks.

- Shift all the tags by d/2 toward the 3’ end.

Feng 2011
Current protocols in bioinformatics
Analysis downstream to peak calling

- Visualization - genome browser: Ensembl, UCSC, IGB
- Peak Annotation - finding interesting features surrounding peak regions: PeakAnalyzer
- Correlation with expression data
- Discovery of binding sequence motifs
  - Split peaks
  - Fetch summit sequences
  - Run motif prediction tool
- Gene Ontology analysis on genes that bind the same factor or have the same modification
- Correlation with SNP data to find allele-specific binding
Tools to install for the next session

- Bowtie (http://sourceforge.net/projects/bowtie-bio/files/latest/download)
- MACS (http://liulab.dfci.harvard.edu/MACS/index.html)
- PeakAnalyser (available at http://www.ebi.ac.uk/bertone/software)
- Java (http://www.java.com/fr/)