What is ChIP-Seq?

- ChIP-Seq is a new technology to analyze protein and DNA interactions.
- ChIP-Seq
  - Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing
  - Allow mapping of protein–DNA interactions in-vivo on a genome scale
ChIP-Seq Workflow

1. Cross-link whole cells with formaldehyde
2. Isolate genomic DNA
3. Sonicate DNA to produce sheared, soluble chromatin
4. Add protein-specific antibody
5. Immunoprecipitate and purify immunocomplexes
6. Reverse cross-links, purify DNA and prepare for sequencing
7. Map to genome

Bioinformatics

Bioinformatics Challenges

- Rapid mapping of short sequence reads to the reference genome
- Visualize mapping results
  - Thousand of enriched regions
- Peak analysis
  - Peak detection
  - Finding exact binding sites
- Compare results of different experiments
  - Normalization
  - Statistical tests
Let’s go over PeakSeq, a software package for Chip-Seq…
Visualization: Genome Browser

Visualization: Custom

300 kb region from mouse ES cells

Visual comparison with Chip-Chip
Results: 1 to 20 of 416

1. Activation of the Glucocorticoid Receptor Is Associated with Poor Prognosis in Estrogen Receptor-Negative Breast Cancer
   Pan D, Kocherginsky M, Conzen SD.
   Cancer Res. 2011 Oct 4. [Epub ahead of print]
   PMID: 21868756 [PubMed - as supplied by publisher]
   Related citations

2. Next-generation insights into regulatory T cells: expression profiling and FoxP3 occupancy in human
   Free full text Related citations

3. Discriminative prediction of mammalian enhancers from DNA sequence.
   Lee D, Karchin R, Beer MA.
   Genome Res. 2011 Oct 3. [Epub ahead of print]
   PMID: 21875935 [PubMed - as supplied by publisher]
   Related citations
RNA-seq
RNA-seq

- high-throughput sequencing technology for sequencing RNA (actually cDNA which is reverse transcript of RNA).
- allows researchers to obtain information like:
  - gene/transcript/exon expressions
  - alternative splicing
  - gene fusions
  - post-transcriptional mutations
  - single nucleotide variations
  - others
RNA-Seq workflow

Steps:
1. Reads filtering (quality, B’s, etc.)
2. Align all reads on genome
3. Aligning against the transcriptome all the reads which
   ◦ map uniquely on genome, or
   ◦ do not map on genome
4. Compute (normalized) transcript expressions (e.g. RPKM)
5. Repeat steps 1-4 for all samples
6. Find relative-changes among their transcript expressions of the same gene across the group of samples
Alternative splicing and isoform

Cassette Exon

Mutually Exclusive Exons

Intron Retention

Alternative 5' or 3' Splice Sites

Alternative Promoters

Alternative Splicing and Polyadenylation
Bioinformatics challenge

Fig. 2 – Reads’ mappings at chromosome and transcript level
Bioinformatics challenge

Fig. 5 – Coverage plot
Bioinformatics challenge

Coverage plot for gene ERBB2 in breast cancer

Coverage plot for gene ERBB2 in normal breast

Fig. 6 – Coverage plots visualization
1. Whole-transcriptome RNAseq analysis from minute amount of total RNA.
   Tariq MA, Kim HJ, Jejelowo O, Pourmand N.
   Free full text Related citations

2. Single-nucleotide resolution analysis of the transcriptome structure of Clostridium beijerinckii NCIMB 8052 using RNA-Seq.
   Wang Y, Li X, Mao Y, Blaschek HP.
   PMID: 21962126 [PubMed - as supplied by publisher] Free Article
   Related citations

3. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data.
   Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM.
Let's start to work first. While program running, you can:

In this tutorial, I will analyze mouse paired-end RNA-Seq data from Illumina GAII to calculate expression level of variants, and identify novel gene and novel alternative splicing variants. If you need to analyze RNA-Seq data, prepare the genome reference and gene annotation file according to your request.

Note: You can copy and paste all the text in blue to your Linux command line to run. Anything with "#" is commented out.

#1) ssh to smp node

# Because some scripts need large memory, also some of them can run multiple threads to speed up the analysis.
# Please you should not use the head nodes to run these analysis.
ssh ssh.ccv.brown.edu
ssh -Y smp007
# Note: option -Y will allow linux to forward graphic output from R, or Java

# Check if the node is busy. If yes, try other smp nodes. The available nodes are from "smp001" to "smp007" list.