A Practical Guide to Integrative Genomics by RNA-seq and ChIP-seq Analysis

Jian Xu, Ph.D.

Children’s Research Institute, UT SW
Outline

- Introduction
- Overview of genomic and next-gen sequencing technologies
- Basic concepts of genomic data analysis
- A practical guide for integrative genomic analysis
  - RNA-seq
  - ChIP-seq
  - Integrated analysis
CRI Sequencing Core Facility

- Equipped with illumina NextSeq500 and BaseSpace Onsite server, mounted to BioHPC storage

- Sequencing types: 1x75bp, 2x75bp, 2x150bp

- Application: RNA-seq, ChIP-seq, exome-seq, targeted resequencing, amplicon sequencing, WGS, etc.

- Managed by Xin Liu & Jian Xu

- More information: http://cri.utsw.edu/sequencing-facility-home/
Introduction to Omics

• BIG data

• Genomics, Proteomics, Metabolomics, …

• Omics studies have been greatly advanced by the high-throughput technologies

• Aim: To characterize and quantify pools of biological molecules that translate into intracellular structure, function and dynamics
Next-Gen Sequencing (NGS)

Single Index Sequencing Utilizes 3 Sequencing Reads

1. Read 1 Seq Primer (HP6)
   - DNA insert

2. Index Seq Primer (HP8)
   - Index
   - Paired End Turnaround
   - DNA insert

3. Read 2 Seq Primer (HP7)
   - DNA insert
Illumina Sequence by Synthesis (SBS) Technology

- Reversible terminator
- Two or four-color removable fluorescent dyes
- Ultra resolution, high-speed scanning optics
Subjects of Omics Study

- **DNA** → **Genome**
- **RNA** → **Transcriptome**
- **Protein** → **Proteome**
- **Metabolites** → **Metabolome**
I have the data, now what?

RNA-seq
ChIP-seq
Proteomics
Epigenomics
Metabolomics
Animal data
Clinical data
Imaging data
Bioinformatics 101 for Bench Scientists

1. Data Generation and QC
   - How to process and QC the data?
   - How to download, organize and store your data?

2. Data Processing
   - What questions do you want to answer?
   - What file formats do you need?

3. Data Analysis and Integration
   - Web-based (e.g. BioHPC, Galaxy/Cistrome)
   - Command lines

4. Data Presentation
   - How to visualize the data? (e.g. Graphing tools)
Overview of Genomics Data Analysis

**Data Generation**

- Sample Collection
- Sample Preparation
  - Microarray
  - NGS
  - MS
- Data Collection
- Data Processing
- Database

**Data Analysis**

- Integrated Analysis
  - Data filtering
  - Annotation
  - Pathway reconstruction
  - Modeling
- Identification/Quantification of Interactions
- Data Visualization
- Biological Validation
- Functional Candidates
- Testable Hypothesis
Why RNA-seq and ChIP-seq?

• **RNA-seq**
  1. Genes differentially expressed between condition A and B
  2. Relative expression levels of genes in condition A and B
  3. Functional groups or pathways

• **ChIP-seq**
  1. Binding sites and densities of factorX across the whole genome
  2. Cooperating factors of factorX

• **Integrative analysis of RNA-seq and ChIP-seq**
  1. Target genes of factorX
  2. How factorX cooperates with co-factors (TFs or epi-factors) and epigenetic landscapes to regulate gene expression
  3. Mechanism of action and testable hypothesis
## Overview of RNA-seq and ChIP-seq Analysis

### RNA-seq

- **BioHPC NGS Pipeline**
- **Cufflinks**

<table>
<thead>
<tr>
<th>Output File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes_count.txt</td>
<td>Read count for genes</td>
</tr>
<tr>
<td>Genes_fpkm.txt</td>
<td>*FPKM of genes</td>
</tr>
<tr>
<td>Gene_exp.diff</td>
<td>Differential gene list</td>
</tr>
<tr>
<td>Isoforms_count.txt</td>
<td>Read count for transcripts</td>
</tr>
<tr>
<td>Isoforms_fpkm.txt</td>
<td>*FPKM of transcripts</td>
</tr>
<tr>
<td>Isoform_exp.diff</td>
<td>Differential transcript list</td>
</tr>
</tbody>
</table>

*FPKM: fragments per kilo base of exon per million reads mapped

### ChIP-seq

- **BWA, MACS**
- **CRI ChIP-seq Pipeline**

<table>
<thead>
<tr>
<th>Output File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peaks.bed</td>
<td>Chr locations for called peaks</td>
</tr>
</tbody>
</table>
| Peaks.xls                        | • Having more information of calling parameters and statistical data for each peak  
|                                  | • Can be used to further filter the peaks         |
| Summits.bed                      | Chr locations for the centers of called peaks     |
| Control.wig                      | • Having the peak intensity information           |
| Treat.wig                        | • Used in further downstream analysis             |
| Control.tdf                      | Used to visualize peak shape and intensity in IGV |
| Treat.tdf                        |                                                   |

### Tools

- Cufflinks
- BWA
- MACS
- BioHPC NGS Pipeline
- CRI ChIP-seq Pipeline

---

Jian Xu Lab @ CRI, UTSW
RNA-seq Analysis using BioHPC NGS Pipeline

Welcome to the BioHPC NGS Web Toolkit

Build Your Own Next Generation Sequencing Pipeline

- Raw Sequencing Reads
- Map reads to genome
- Assemble Transcripts
- Quantify Transcripts
- Find Differential Expression
# Data Analysis: RNA-seq

## RNA-seq Output Files

<table>
<thead>
<tr>
<th>File Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genes_count.txt</td>
<td>Read count for genes</td>
</tr>
<tr>
<td>Genes_fpkm.txt</td>
<td>*FPKM of genes</td>
</tr>
<tr>
<td>Gene_exp.diff</td>
<td>Differential gene list</td>
</tr>
<tr>
<td>Isoforms_count.txt</td>
<td>Read count for transcripts</td>
</tr>
<tr>
<td>Isoforms_fpkm.txt</td>
<td>*FPKM of transcripts</td>
</tr>
<tr>
<td>Isoform_exp.diff</td>
<td>Differential transcript list</td>
</tr>
</tbody>
</table>

*FPKM: fragments per kilo base of exon per million reads mapped

**Note:**
Commonly used gene identities: 1) gene symbol; 2) Entrez gene ID; 3) refseq_mRNA (transcripts)

---

### BioHPC NGS Pipeline

- Cufflinks
- GSEA MSigDB
- DAVID
- Oncomine
- cBioPortal

### A gene list

**Quantitative gene expression profiles**

- javaGSEA

---

14 Jian Xu Lab @ CRI, UTSW
What information do we get from RNA-seq?

- Differential expressed genes between condition A and B

**Functional Annotation:** To look for enriched pathways, biological processes or features, or correlation with other defined gene signatures in different cellular or pathological contexts

---

**A list of genes**
(no quantitative data)

- **GSEA/Molecular.Signature.Databas/Investigate.Gene.Sets**
  - > select gene signatures to compare
  - http://www.broadinstitute.org/gsea/msigdb/annotate.jsp

- **Oncomine cBioPortal**

- **DAVID Bioinformatics Resources**
  - /Start.Analysis -> paste your gene list -> select correct identifier (official gene symbol or Refseq_mRNA) -> gene list as list type -> submit -> Functional annotation tools
  - https://david.ncifcrf.gov/tools.jsp
ChIP-seq Analysis using CRI ChIP-seq Pipeline

Available Workflows

Astrocyte CRI ChiPSeq Workflow
This is a CRI chipseq workflow package for the BioHPC astrocyte workflow system. It implements a simple ChiPSeq analysis workflow.

Parameters

- Project
- Name for this run
- FASTQ files
- design file
- Reference genome for alignment

Run Workflow
To run a workflow against input data you need to upload it into this project. Click the button below to add new files from your web browser or the BioHPC cluster. You can also download or delete existing files from the project in the list below.

### Add Data To This Project

<table>
<thead>
<tr>
<th>Filename</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-seq_K562_DMSO_BRD4.fastq</td>
<td>4.3 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>ChIP-seq_K562_JQ1-6h_BRD4.fastq</td>
<td>3.9 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>ChIP-seq_K562_JQ1-2h_BRD4.fastq</td>
<td>4.3 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>design.txt</td>
<td>18 bytes</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>ChIP-seq_K562_JQ1-2h_RNAPII.fastq</td>
<td>4.6 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>ChIP-seq_K562_DMSO_RNAPII.fastq</td>
<td>4.9 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
<tr>
<td>ChIP-seq_K562_JQ1-6h_RNAPII.fastq</td>
<td>3.7 GB</td>
<td><img src="#" alt="Download" /> <img src="#" alt="Delete" /></td>
</tr>
</tbody>
</table>

Astrocyte provides many workflow created by different groups at UTSW for you to run against your data. To begin, make sure you have added input data into your project and then click the ‘Run a workflow’ button to choose a workflow to run.

### Run a workflow in this project

<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Workflow</th>
<th>Status</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>first_Run</td>
<td>Oct. 21, 2016, 1:32 p.m.</td>
<td>Astrocyte CRI ChIPSeq Workflow Version 0.0.3</td>
<td>Completed successfully</td>
<td>80.5 GB</td>
<td><img src="#" alt="Delete" /></td>
</tr>
</tbody>
</table>
Data Analysis: ChIP-seq

ChIP-seq Output Files

<table>
<thead>
<tr>
<th>File</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peaks.bed</td>
<td>Chr locations for called peaks</td>
</tr>
</tbody>
</table>
| Peaks.xls    | • Having more information of calling parameters and statistical data for each peak  
               • Can be used to further filter the peaks                                |
| Summits.bed  | Chr locations for the centers of called peaks                                |
| Control.wig  | • Having the peak intensity information                                      |
| Treat.wig    | • Used in further downstream analysis                                        |
| Control.tdf  | Used to visualize peak shape and intensity in IGV                            |
| Treat.tdf    |                                                                                   |

1. Load to UCSC genome browser to view peak sites and to obtain the sequence of a specific site
2. Used in analyses of:
   • Motif enrichment
   • Venn diagram
   • Profiling of common and unique peaks

Used in statistical analyses, e.g.
   • Profiling or comparison of binding intensity
   • Correlation of the genomic binding of multiple factors

Load to IGV to view peaks
Cistrome: an integrative platform for transcriptional regulation studies

Tao Liu¹,²†, Jorge A Ortiz³,⁴†, Len Taing¹,², Clifford A Meyer¹, Bernett Lee³,⁵, Yong Zhang⁶, Hyunjin Shin¹,², Swee S Wong³,⁷, Jian Ma⁶, Ying Lei⁸, Utz J Pape¹, Michael Poidinger³,⁵, Yiwen Chen¹, Kevin Yeung³,⁹, Myles Brown²,¹⁰*, Yaron Turpaz³,¹¹* and X Shirley Liu¹,²*

Abstract
The increasing volume of ChIP-chip and ChIP-seq data being generated creates a challenge for standard, integrative and reproducible bioinformatics data analysis platforms. We developed a web-based application called Cistrome, based on the Galaxy open source framework. In addition to the standard Galaxy functions, Cistrome has 29 ChIP-chip- and ChIP-seq-specific tools in three major categories, from preliminary peak calling and correlation analyses to downstream genome feature association, gene expression analyses, and motif discovery. Cistrome is available at http://cistrome.org/ap/.
Integrative Analysis by Galaxy / Cistrome

http://cistrome.org/ap/

Data Preprocessing
- Peak Calling
  - MAT for Affy
  - MA2C for NimbleGen
  - MACS for ChIP-seq
  - MM-ChIP
  - NPS

Integrative Analysis
- Correlation
  - Global Correlation
- Local Correlation
- Venn Diagram
- Association
  - Conservation
  - SitePro
  - Gene Centered Annotation
- Peak Centered Annotation
- DNA Motif
  - Motif enrichment
  - Motif Scan
- Heatmap with clustering

Gene Expression
- Gene Expression Index
- Highly Expressed TFs
- Related Genes
- Differential Expression
- Gene Ontology

Import Data
- Data Upload
- DC Browser
- Auto Retriever from GEO

Peak locations (BED)
Signal profiles (WIGGLE)

Gene lists

Galaxy Tools

Jian Xu Lab @ CRI, UTSW
### Integrative Analysis by Galaxy / Cistrome

http://cistrome.org/ap/

#### Tools

- Genomic Data Preprocessing
- Gene Expression
- Integrative Analysis
- Liftover/Others

#### Input/Output

- Saved Histories

#### View Window

- Datasets: Name, Tags, Sharing, Size on Disk, Created, Last Updated, Status

---

Jian Xu Lab @ CRI, UTSW
**Integrative Analysis by Galaxy / Cistrome**

http://cistrome.org/ap/

### Correlation

- **Multiple wiggle files correlation**
  - Calculate the correlation coefficient on the genome scale using multiple wiggle / bigwig files in fixed-sized windows

- **Multiple wiggle files correlation in given regions**
  - Calculate the correlation coefficient on the genome scale using one bed file and multiple wiggle / bigwig files

- **Two wiggle file correlation in union regions**
  - Calculate the correlation coefficient of two wiggle / bigwig files in the union regions from two bed files

- **Venn Diagram**
  - Given 2 or 3 intervals, generate a venn diagram of their intersections

### Association

- **CEAS**: Enrichment on chromosome and annotation
  - Annotate the given intervals and scores with genome features such as gene body

- **SitePro**: Aggregation plot tool for signal profiling
  - Draw the score profile near a given interval

- **GCA**: Gene centered annotation
  - Find the nearest interval in the given intervals set for every annotated coding gene

- **peak2gene**: Peak Center Annotation
  - This tool is abolished, please use BETA-minus instead.

### Binding/Expression

- **BETA**
  - **BETA-basic**: Binding and Expression Target Analysis
    - Predict the factors (TFs or CRs) direct target genes by combining the binding and expression data
  - **BETA-plus**: Binding and Expression Target prediction and motif analysis
    - Predict the factors (TFs or CRs) direct target genes by combining the binding and expression data, then do motif analysis on target regions
  - **BETA-minus**: Targets prediction with binding only
    - Predict the factors (TFs or CRs) direct target genes by only binding data

### Motif

- **SeqPos motif tool**
  - Find motifs from given regions enriched near the centers

- **Screen Motif**
  - Given a motif, find all regions that match the motif

- **MISP**: Motif-based Interval Screener with PSSM
  - Input one or more motifs, find all hits in peak regions

---

Jian Xu Lab @ CRI, UTSW
Integrative Analysis of ChIP-seq and RNA-seq Data

**Step 1:** Identify and characterize TF binding sites across the genome

**SeqPose:** find motifs

**Tips:**
1) Less than 5000 lines allowed, so need to filter the peaks in peak.xls file using p-value or FDR as a cutoff;
2) Providing clues for cofactors.

**.wig**

**.bed**

**CEAS:** Enrichment on chr and annotation

**Output:** .pdf or .png

**BETA-minus:** find genes* (among all genes) in a given range (e.g. +/- 50kb of TSS) of the binding sites

* **w/o consideration of differentially regulated genes**

**Conservation plot**
Step 2: Identify candidate target genes of the TF

Direct target genes:
1) the genes differentially regulated between condition A and B; 
   **AND**
2) with TF binding sites near the given region (e.g. +/- 50kb) from TSS of the genes.

Galaxy Toolbox -> Operate on Genomic Intervals

Compare the peak bed file (peaks.bed) and diff_genes_TSS_50kb.bed (genes.bed) file
Step 3: Compare TF binding between conditions or the binding of multiple TFs

- **A.bed**
- **A.wig**
- **B.bed**
- **B.wig**

Venn diagram: A ∩ B

Differential peaks:
1. Common_peaks.bed
2. A_unique_peaks.bed
3. B_unique_peaks.bed

**Search for direct target genes in each condition**
Galaxy tool -> Operate on Genomic intervals

**Sitepro**: Draw the score of binding signals near given peak regions

**Heatmap**: Draw heatmap near given peak regions and also cluster peaks

**SeqPos motif scan**

**Tip**: Different motif enrichment would suggest the use of different co-factors.
What to do with the candidate genes?

Screen the gene list to identify ones with potential link to your biology story

1. Pathway enrichment (DAVID, GSEA)
2. Gene function and pathological correlations (Genecards, BioGPS, NCBI/Gene, Ensembl)
3. Interaction partner proteins (NCBI/Gene, Ensembl)
4. Expression levels in normal tissues and diseased tissues (Genecards, BioGPS, NCBI/GEO, Ensembl)
5. Mutations, frequency and functional impacts (cBioPortal, COSMIC, Ensembl)
6. Testable hypothesis
7. Experimental validation
Questions?

Acknowledgment:
Min Ni, Ph.D., GMDP @ CRI
Yuannyu Zhang, Ph.D.
Yi Du, David Trudgian, Liqiang Wang @ BioHPC
Members of Jian Xu Lab @ CRI