

CCRIFX ChIP-Seq Data Analysis Tutorial for Biologists - DRAFT

Step 1. Setting up your computing environment

Objectives

- Identify computing resources required for data storage and analysis (personal computers, shared lab computers, portable drives, etc.)
- Install stand-alone tools and obtain software licenses for web-based tools
- Identify training opportunities at NIH and online tutorials

Minimum system requirements

- Check if your computer meets the minimum computer specification
 - Computing power – 4-8-16 gigabytes (GB) of RAM
 - 1-2 processors with 2-5 cores
 - Storage capacity - 1-2-4 terabytes (TB) of disk

Online tutorials

- CCR Bioinformatics Training and Education Program (BTEP)
 - <http://bioinformatics.nci.nih.gov/training/>
- Center for Information Technology (CIT) Training
 - <http://training.cit.nih.gov/>
- NIH Library Bioinformatics Training
 - <http://nihlibrary.nih.gov/Services/Bioinformatics>
- FAES Grad School
 - <http://www.faes.org/grad>
- NIAID bioinformatics training seminars (use IE or Safari)
 - <http://collab.niaid.nih.gov/sites/research/SIG/Bioinformatics/Seminars.aspx>
- UCSC Browser - [UCSC Genome Browser](#)
- Online tutorials at software websites
 - Genomatix - http://www.genomatix.de/online_help/help/videos.html
 - [Distance correlation of genomic elements](#)
 - [Introduction to the Genomatix Genome Browser](#)
 - [ChIP-Seq case study](#)

Software tools for ChIP-Seq data analysis

- Some of the tools evaluated and tested by CCRIFX are listed in Table 1 below along with their advantages and limitations.
- Many of these command-line tools are available on [Biowulf](#) (see Step 4). Stand-alone tools need to be installed on your Mac or PC.
- GUI tools are great for initial exploratory data analysis, visualization and analysis of small datasets. If you are new to NGS data analysis, [Genomatix](#)

and [IGV](#) are a good starting point. Click [here](#) for software license information from CCR OSTP.

- Command-line tools offer more flexibility, parameter optimization, and data quality control options. They are often used for analysis of medium and large datasets and in pilot projects that require a lot of parameter optimization.
- This is not a comprehensive list. More tools are listed below (see Step 5).

Application	Input files	GUI	Advantages	Limitations	General Use Suggestions
CEAS	BED, WIG	No	Cis-regulatory Element Annotation System		Peak annotation, gene-centered annotation, R graphics
CLC Genomics Workbench	NGS reads from Chip-Seq, RNA-Seq, miRNA and WGS studies	Yes	One-stop-shop application; a wrapper for many command line tools; good quality <i>de novo</i> assembler	Network connectivity issues	Read alignment, visualization miRNA expression analysis, variant detection, phylogenetic trees
Homer	NGS reads, BAM, BED	No	Comprehensive software suite, ChIP-Seq analysis can be run on a laptop		ChIP-Seq, <i>de novo</i> motif analysis, QC, visualization
Genomatix Software Suite (GSS)	BAM, BED	Yes	Transitions to and from GMS, GMA, genome and pathway annotation tools for 33 organisms	Genome annotations are limited to protein-coding genes and miRNAs	Transcription factor over-representation analysis and <i>de novo</i> motif finding, SNP annotation, microarrays
IGV	BAM, BED, TDF	Yes	Ease of use	Limited numbers of tracks available for viewing	Exploratory data analysis, QC, and visualization
MACS	NGS reads, BAM, BED	No	Gold standard for transcription factor binding site analysis	Not suitable for broad-factor analysis, FPs in MACS1 and FNs in MACSII	Peaks calling and TF binding site analysis
MEME	DNA sequences containing motifs	Yes			<i>De novo</i> motifs discovery, GO annotations
Partek Flow	NGS reads, BAM, BED	Yes	RNA-Seq quantification, variant detection	Network connectivity issues, hidden parameters	Projects can be shared with collaborators

Partek Genomics Suite	BAM, BED, gene lists	Yes	Easy to use, good statistical tools for DEG analysis	Limited beyond-a-gene-list analysis options Limited standard QC workflow	Primary data analysis, Anova T-test, Data integration with ChIP-Seq
SICER	NGS reads, BAM, BED	No	Gold standard for broad-point factor analysis	Non-standard BED output files	Broad-point factor analysis

Step 2. Understand ChIP-Seq terminology and analysis techniques

Objectives

- Understand ChIP-Seq terminology and analysis techniques
- Understand major steps in the ChIP-Seq analysis workflow
- Name common tools and analytical approaches used for ChIP-Seq analysis
- Name ChIP-Seq data mining software and databases

Typical deliverables in a Chip-Seq

- Raw primary data
 - FASTQ (e.g. 36-bp single end reads)
- Intermediate results
 - QC results, BAM, BED, TDF files
- Final results
 - Peak lists, transcription factor (TF) lists, known and de novo motifs, enriched pathways, PCA and clustering diagrams, heat maps, VENN diagrams, data integration results
- Data from public sources
 - [ENCODE](#)
 - [SRA](#)
 - [GEO](#)

Terminology

- **File formats**
 - BAM - Binary Sequence Alignment/Map; contains aligned reads
 - BAI – indexed BAM file
 - BED (Browser Extensible Data) - a tab-delimited text file that defines a feature track
 - 1 line per feature, 3-12 columns, plus track def lines
 - Required fields
 - chrom - name of the chromosome or scaffold
 - chromStart - Start position of the feature
 - chromEnd - End position of the feature
 - BedGraph - to display of continuous-valued data in track format
 - FASTQ – a text file, contains sequence data and quality scores

- FA/FASTA - a text file, contains description and sequence data
- General Feature Format (GFF/GTF) – a tab-delimited text file for describing genomic features
- SAM - Sequence Alignment/Map; contains aligned reads
- [SRA](#) (NCBI) – Short read archive
- TDF - a binary file that contains data preprocessed for faster display
- WIG - a text file that defines dense, continuous data/features such as GC percent, probability scores, and transcriptome data
 - bigWig - to display of dense, continuous data as a graph
- More formats here - <http://genome.ucsc.edu/FAQ/FAQformat>

Major steps in the ChIP-Seq analysis workflow

- Pre-alignment QC (FastQC)
- Read trimming and filtering (Trimmomatic)
- Read alignment to the reference genomic sequence (bowtie)
- Convert BAM files to BED files (SAMTools)
- Estimate fragment length (e.g. Homer tagDir2bed)
- Find significant peaks in sample vs. input (HOMER)
- Downstream analysis
 - Visualization in a genome browser (e.g. IGV)
 - *De novo* motif finding (HOMER)
 - Known motif finding (e.g. HOMER)
 - Analysis of gene lists associated with the peaks (Genomatix, IPA)
 - Analysis of pathways associated with the peaks (IPA)
 - Comparison of ChIP-Seq peak lists to each other (Genomatix)
 - Integration with non-Chip-seq data sets (Partek)
 - Hypothesis generation

Analytical approaches

- Complex biological hypothesis require sophisticated statistical tools
- Different tools use different statistical models or tests
 - E.g. local Poisson dist., HMM, T-test, conditional binomial model

Step 3. Comparing ChIP-Seq data

Objectives

- Use Partek to analyze BAM files
- Use Genomatix to analyze BAM files
- Find significant peaks
- Perform data visualization and downstream analyses
 - Visualization in a genome viewer - IGV
 - Analysis of the gene lists associated with the peaks
- Explore public ChIP-Seq data

- Tracks in the UCSC Genome Browser - can be downloaded as BED files
 - NCBI SRA data sets - can be downloaded as SRA files
-
-

Step 4. Establishing a Helix/Biowulf account (for data storage and access to command line tools and the HTP computing environment)

- Certain steps in ChIP-Seq require extra computing power
 - Short read alignment (BAM)
 - Peak finding
 - *De novo* motif finding
- For these steps, you may need to use HTP computing environment such as Biowulf
- Follow these steps to set up an Helix/Biowulf account
- Identify your 'sponsor' - <http://helix.nih.gov/Documentation/accounts.html>
 - Ask him/her to request a personal Helix account and then a Biowulf account
 - From your mac or PC, login to Helix and Biowulf
 - Review user guidelines
 - <http://helix.nih.gov/Systems/helix.html>
 - http://biowulf.nih.gov/user_guide.html
- If you have a PC, request Putty or another Telnet/SSH client to be installed
 - <http://www.chiark.greenend.org.uk/~sgtatham/putty>
- Unix commands: <http://mally.stanford.edu/~sr/computing/basic-unix.html>
- Data transfer over ssh connections using 'scp'
 - <http://helix.nih.gov/Documentation/transfer.html>
- Request a group account and a shared group directory on Biowulf for
 - Yourself
 - And any other lab member that would be using command-line tools
 - The owner of this group data needs to complete the shared group form at https://helix.nih.gov/nih/shared_data_request.html
- Online tutorials
 - <http://www.linuxtutorialblog.com/post/ssh-and-scp-howto-tips-tricks>
 - <http://helix.nih.gov/Documentation/transfer.html>
 - <http://bioport.lsu.edu/Members/admin/chip-seq-tutorial/chip-seq-in-two-weeks>
 - <http://openhelix.com/ENCODE2>
- Practice executing common Unix commands on Helix and Biowulf
 - Copy your data to the shared group directory with 'scp' to your directory on Biowulf
 - `ssh your_ned_id@helix.nih.gov your_ned_password` - to ssh to HELIX
 - `ssh your_ned_id@biowulf.nih.gov your_ned_password` - to ssh to BIOWULF

- `scp -r your_ned_id@biowulf.nih.gov://path_to_files .` - to pull files from BIOWULF
 - `chgrp -R CCRIFX <ccrifx_dir_name>` - To change the group of the file or directory specified by the File or Directory parameter to the group specified by the Group parameter
 - `md5sum` - to compute *md5* checksum
-

Step 5. Commonly used tools

- BWA (Burrows-Wheeler Aligner) - to map reads to a reference
- Bowtie - another aligner
- MACS1 and MACS2 (Model-based Analysis for ChIP-Seq) - narrow-source peaks
- MEME (Motif-based sequence analysis tools) - to find de novo motifs
- HOMER - to find peaks and motifs
- CEAS (Cis-regulatory Element Annotation System) - to annotate peaks
- SAMtools - to manipulate files
 - E.g. to see the header in BAM file - `samtools view -H <any>.bam`
- SRA toolkit for analysis of NCBI SRA data sets
- CisGenome - GUI*
- E-RANGE - broad source peaks
- HPeak
- PeakSeq
- QuEST
- SICER
- SISR - Windows only
- Sole-Search - GUI*
- And many more - <http://omictools.com/chip-seq>

Step 6. Data and metadata reporting requirements

- Publishing
 - Checklist for articles - adopted by many journals
 - www.nature.com/authors/policies/checklist.pdf
 - Reporting policies
 - <http://www.nature.com/authors/policies/reporting.pdf>
- NGS data submission instructions
 - GEO: <http://www.ncbi.nlm.nih.gov/geo/info/spreadsheet.html>
 - MINSEQE - Minimum Information about a Sequencing Experiment
 - Encode experimental design and submission guidelines
- Recent "Big Data" initiatives in life sciences
 - Global alliance to share genomic data
 - Reproducibility initiative