
CyVerse Documentation

CyVerse

Sep 23, 2020

Contents

1	Learning Objectives	3
1.1	Step-by-step Tutorial:	3
2	Manual Maintainer(s)	13
3	Prerequisites	15
3.1	Downloads, access, and services	15
3.2	Platform(s)	15
3.3	Application(s) used	15
3.4	Input and example data	16



 [Learning Center Home](#)

This tutorial introduces the basic principles of analyzing chromatin immunoprecipitation sequencing (ChIP-seq) data and is intended to provide enough information to perform key steps of ChIP-seq analysis: quality control of data, mapping data to a reference genome and peak calling. We will be using Cyverse discovery environment public apps for analysis of a sample set. Command-line expertise is not required to follow most of this tutorial.

Learning Objectives

- Efficiently manage and analyze ChIP-seq data
 - How to do quality assessment using ChIPQC R package
 - Using CyVerse Discovery Environment apps to do ChIP-seq analysis
-

1.1 Step-by-step Tutorial:



 [Learning Center Home](#)

1.1.1 Sample dataset and preprocessing

We are analyzing Fumarate and nitrate reduction (FNR) transcription factor dataset in this tutorial (Myers et al., 2013). FNR transcription factor controls the expression of over 100 target genes in response to anoxia. It facilitates the adaptation to anaerobic growth conditions by regulating the expression of gene products that are involved in anaerobic energy metabolism. We will use the FNR IP ChIP-seq Anaerobic A (GSM1010219) dataset and compare this with the control sample (GSM1010224).

Input Data:

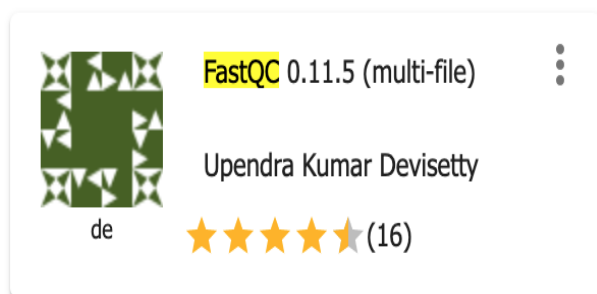
Input	Description	Location
FNR transcription factor data	FNR IP and Input DNA in anaerobic condition	iplantcollaborative > example_data > chipseq_webinar -> fastqfiles

Preprocessing

Evaluate the quality of your sequencing data using FastQC

Preprocessing of ChIP-seq data is similar to that of any other sequencing data and will assess the quality of the raw reads to identify possible sequencing errors or biases. FastQC can be used for an overview of the data quality but this does not assess if your ChIP experiment has worked. We will assess that in Step4- Postprocessing- ChIP quality assessment.

1. Login to the .
2. Click on “Apps” tab in the Discovery Environment and search for “fastqc”.
3. Click on the app icon.



4. Change the name of the analysis and output folder as needed or leave for defaults.
5. Under “Input” click on Add to provide input files for both ChIP and input dataset. Sample dataset location iplantcollaborative > example_data > chipseq_webinar -> fastqfiles. Check both files (SRR576933_IP.fastq, SRR576938_input.fastq) and click ‘OK’.
6. For next section “Resource Requirements” request resources as needed or leave for defaults
7. Click **Launch Analysis**. You will receive a notification that the job has been submitted and running. Click on the Analyses tab to check the status of your job. When the analysis completes, click on the right three dots menu and click on ‘Go to output folder’ to access you output files.

Output/Results

Output	Description	Example
html and zip files	FastqQC report	SRR576933_IP_fastqc.html

Description of output and results

Click on the html report files and check if your sequencing data has any red flags that you should be aware of. There are few red flags in the report. You will notice that “Per base sequence quality” decrease towards the end of the reads which is usual with illumina sequencing. Other useful metrics that should be checked for ChIP-seq data are: sequence duplication levels and over-represented sequences. Check the tutorial on how to .

As this report does not present any major concerns regarding the quality of this dataset, we will proceed for the next step ,i.e., reads alignment. However, for your own data, it is a good practice to rerun fastqc after quality filtering your reads: remove adapter sequences and low-quality bases (Phred quality score< 20) and discard any short reads after

trimming (<20bp reads). Check in CyVerse DE which can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. Access CyVerse trimmomatic app tutorial .

For more details on each module of the fastqc report, check

Fix or improve this documentation

- Search for an answer:
 - Ask us for help: click  on the lower right-hand side of the page
 - Report an issue or submit a change:
 - Send feedback: Tutorials@CyVerse.org
-

 [Learning Center Home](#)



 [Learning Center Home](#)

1.1.2 Read mapping

In this step, we will align our reads to the E. coli reference genome. Any standard short-read alignment program such as BWA, Bowtie can be used for this step. We will use Bowtie 1 for aligning our reads. For short reads (< 50bp), Bowtie 1 is faster and sensitive than Bowtie2.

Input Data:

Input	Description	Location
Sequence reads	Raw or quality filtered reads	iplantcollaborative > example_data > chipseq_webinar -> fastq-files

Run Bowtie1 in the CyVerse Discovery Environment

1. Click on “Apps” tab in the Discovery Environment and search for “bowtie”.
2. Click on the app icon. Bowtie build and map app builds a bowtie index and then maps reads.



3. Change the name of the analysis and output folder as needed or leave for defaults.
4. Under bowtie build Input section provide Reference genome file in Fasta format. Browse through the datastore and provide GCF_000005845.2_ASM584v2_genomic.fna Ecoli reference genome. This file is provided with the sample dataset- iplantcollaborative > example_data > chipseq_webinar -> ecoli_refgenome.
5. Under bowtie build Output section provide name of the basename for index 'ecoli'. Under bowtie map- reference inputs section provide the same name for the index base name 'ecoli'.
6. Under bowtie map read Input section provide the ChIP sequence reads file SRR576933_IP.fastq. Input data location- iplantcollaborative > example_data > chipseq_webinar -> fastqfiles.
7. Provide an output file name and click on the Analyses to check the status of your job. When the analysis completes, click on the right three dots menu and click on 'Go to output folder' to access you output files. Repeat the same steps for control dataset SRR576933_control.fastq

Sequencing depth

Effective analysis of ChIP-seq data requires sufficient coverage by sequence reads (sequencing depth). The required depth depends mainly on the size of the genome and the number and size of the binding sites of the protein. ENCODE's guidelines is to obtain minimum 10 million uniquely mapping reads per replicate experiment for mammalian genomes (Landt et al, 2009).

Note: If atleast 10 million uniquely mapping reads are required for human genome. How many minimum reads are required for E. coli dataset to have sufficient coverage for further analysis?

Output/Results

Output	Description	Example
Alignment files	Alignment files in SAM format	bowtieout_control.sam

Description of output and results

Bowtie build and map app by default provides output files in SAM format which stands for Sequence Alignment/Map format. For more details on SAM format

Fix or improve this documentation

- Search for an answer:
- Ask us for help: click  on the lower right-hand side of the page
- Report an issue or submit a change:
- Send feedback: Tutorials@CyVerse.org



1.1.3 Peak calling

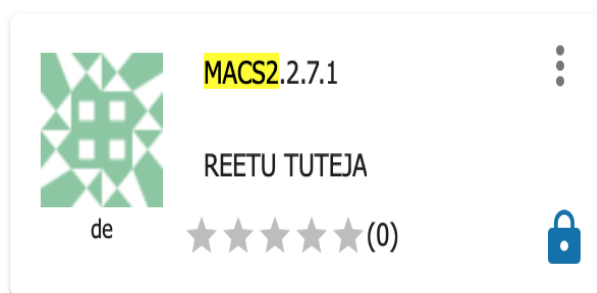
In this step, we will use bowtie alignment files to perform peak calling. Peak calling programs identify set of read enriched regions in the genome that represent binding sites from your protein of interest. We will be using MACS2 (Model-based Analysis for ChIP-seq) app in CyVerse DE for peak identification.

Input Data:

Input	Description	Location
Bowtie output files	Alignment files	iplantcollaborative > example_data > chipseq_webinar -> bowtie_output

Run MACS2 in the CyVerse Discovery Environment

1. Click on “Apps” tab in the Discovery Environment and search for “macs2”.
2. Click on the app icon.



3. Change the name of the analysis and output folder as needed or leave for defaults.
4. Under Callpeaks input section, browse the treatment and control files from the datastore. Provide experiment name ‘ecoli’.

Example treatment file- iplantcollaborative > example_data > chipseq_webinar -> bowtie_output -> bowtieout_chipIP.sam

Example control file- iplantcollaborative > example_data > chipseq_webinar -> bowtie_output -> bowtieout_input.sam
5. Provide mappable genome size- ‘gsize 4639675’ for E. coli genome. Leave the rest of the parameters to defaults. For next section “Resource Requirements” request resources as needed or leave for defaults. Click on Launch Analysis.

6. Click on the Analyses to check the status of your job. When the analysis completes, click on the right three dots menu and click on 'Go to output folder' to access you output files.

Should you discard duplicate reads before peak identification?

Best practice is to remove duplicates prior to peak calling. MACS2 default is to keep a single read at each location but provides different options to deal with duplicates. Bona fide peaks will have multiple overlapping reads with offsets, while samples with only PCR duplicates will stack up perfectly without offsets. These duplicates can arise from experimental artifacts, but can also contribute to true ChIP-signals.

Note: The bad kind of duplicates: If initial starting material is low this can lead to overamplification of this material before sequencing. Any biases in PCR will compound this problem and can lead to artificially enriched regions. Also blacklisted (repeat) regions with ultra high signal will also be high in duplicates. Masking these regions prior to analysis can help remove this problem.

The good kind of duplicates: You can expect some biological duplicates with ChIP-seq since you are only sequencing a small part of the genome. This number can increase if your depth of coverage is excessive or if your protein only binds to few sites. If there are a good proportion of biological duplicates, removal can lead to an underestimation of the ChIP signal (Credits: HBC ChIP-seq workshop for summarizing this info)

Output/Results

Output	Description	Example
NAME_peaks.narrowPeak	Contains the peak locations and other information	ecoli_peaks.narrowPeak
NAME_peaks.xls	Tabular file which contains information about called peaks	ecoli_peaks.xls
NAME_summits.bed	Peak summits locations for every peak	ecoli_summits.bed

Description of output and results

We will be using ecoli_peaks.narrowPeak output file for further analysis. For more information on MACS2 parameters and output files, check the github read me for MACS2 <https://github.com/taoliu/MACS>.

Brief description of narrowpeak output file format (BED6+4 format):

Col1- name of the chromosome

Col2- Peak start position

Col3- Peak end position

Col4- Peak name

Col5- Peak score

Col6- Strand

Col7- Fold enrichment

Col8- log10.pvalue

Col9- log10.qvalue

Col10- peak

For more information about BED format, check

Note: MACS2 mfold parameter specifies an interval of high-confidence enrichment ratio against the background on which to build the model. The default value 10, 30 means that a model will be built on the basis of regions having

read counts that are 10- to 30-fold of the background. Check the effect of changing mfold range to 5,30 on number of resulting peaks.

Fix or improve this documentation

- Search for an answer:
- Ask us for help: click  on the lower right-hand side of the page
- Report an issue or submit a change:
- Send feedback: Tutorials@CyVerse.org

 [Learning Center Home](#)



CYVERSE™

 [Learning Center Home](#)

1.1.4 Postprocessing- ChIP quality assessment

The quality of a ChIP experiment largely depends on the specificity of the antibody and the degree of enrichment achieved in the affinity precipitation step. In this section, we will use ChIPQC R package to calculate cross-correlation and FRiP score to assess quality of our ChIP data. FRiP stands for Fraction of Reads in Peaks. As per ENCODE guidelines, FRiP% values around 5% or higher generally reflect successful enrichment (Landt et al., 2012).

Input Data:

Input	Description	Location
Alignment file	alignment file in bam format	iplantcollaborative > example_data > chipseq_webinar -> bowtie_output
Identified peaks	MACS2 output	iplantcollaborative > example_data > chipseq_webinar -> macs2_output

Run Rstudio-chipqc app in CyVerse DE

1. Click on “Apps” tab in the Discovery Environment and search for “rstudio-chipqc”.
2. Click on the app icon.



3. Change the name of the analysis and output folder as needed or leave for defaults.
4. Under “Input” provide the path for the folder for your input files. This will make your input dataset available in rstudio workspace.
5. Under “Resource Requirements” request resources as needed or leave for defaults
6. Click **Launch Analysis**. You will receive a notification that the job has been submitted and running. Click on ‘Access your analysis here’ link.
7. Once the analysis is launched, provide username (rstudio) and password (rstudio1) for rstudio. Analysis may take few minutes to launch, depends on the size of your input data.
8. Use the following script to calculate FRiP percentage and cross-correlation for one sample using ChIPQCsample function. Check ChIPQC package documentation for more details.


```
library(ChIPQC)

bamFiles <- 'bowtie_chip_sorted.bam'
mypeaks <- read.delim("ecoli_peaks.narrowPeak", header=F)
exampleExp <- ChIPQCsample(bamFiles, peaks=mypeaks)

QCmetrics(exampleExp)
plotFriP(exampleExp)
friP(exampleExp)
plotCC(exampleExp)
```

Note: ChIPQC package accepts sorted BAM files as input. A sorted BAM file is a compressed binary version of a SAM file that has reads sorted by coordinates. Reads from the beginning of the first chromosome are found first in the coordinate sorted alignment file. You can use Samtools SAM to sorted BAM in the DE to convert SAM to sorted BAM files.

Fix or improve this documentation

- Search for an answer:
 - Ask us for help: click  on the lower right-hand side of the page
 - Report an issue or submit a change:
 - Send feedback: Tutorials@CyVerse.org
-



1.1.5 Further reading

Wasserman, W., Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* 5, 276–287 (2004).

Park, P. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10, 669–680 (2009).

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, et al. (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLOS Computational Biology* 9(11): e1003326.

Zhang, Y., Liu, T., Meyer, C.A. et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008).

Carroll TS, Liang Z, Salama R, Stark R and Santiago Id (2014). “Impact of artefact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data.” *Frontiers in Genetics* 10, 5:75.

Ma, W., Noble, W. & Bailey, T. Motif-based analysis of large nucleotide data sets using MEME-ChIP. *Nat Protoc* 9, 1428–1450 (2014).

Fix or improve this documentation

- Search for an answer:
- Ask us for help: click  on the lower right-hand side of the page
- Report an issue or submit a change:
- Send feedback: Tutorials@CyVerse.org



CHAPTER 2

Manual Maintainer(s)

Who to contact if this manual needs fixing. You can also email Tutorials@CyVerse.org

Maintainer	Institution	Contact
Reetu Tuteja	CyVerse / UA	reetututeja@cyverse.org

Note: Acknowledgments: Thanks to VIB and Harvard Chan Bioinformatics training for making wonderful teaching material available for ChIP-seq.

CHAPTER 3

Prerequisites

3.1 Downloads, access, and services

In order to complete this tutorial you will need access to the following services/software

Prerequisite	Preparation/Notes	Link/Download
CyVerse account	You will need a CyVerse account to complete this exercise	
Cyberduck	Standalone software for upload/download to Data Store	

3.2 Platform(s)

We will use the following CyVerse platform(s):

Platform	Interface	Link	Platform Documentation	Quick Start
Data Store	GUI/Command line			
Discovery Environment	Web/Point-and-click			

3.3 Application(s) used

Discovery Environment App(s):

App name	Version	Description	App link	Notes/other links
Bowtie	1.2.2	Short-read sequence aligner		
FastQC	0.11.5	Quality Control tool for HTS data		
MACS2	2.7.1	Model-based analysis of ChIP-seq		
ChIPQC	1.22	Quality assessment tool for ChIP-seq		

3.4 Input and example data

In order to complete this tutorial you will need to have the following inputs prepared

Input File(s)	Format	Preparation/Notes	Example Data
ChIP-seq experiment data	fastq or sra files	Data downloaded from GEO or ENA database	Myers et al., 2013 FNR dataset (iplantcollaborative > example_data > chipseq_webinar)

Fix or improve this documentation

- Search for an answer:
- Ask us for help: click  on the lower right-hand side of the page
- Report an issue or submit a change:
- Send feedback: Tutorials@CyVerse.org

 [Learning Center Home](#)