Hands-on introduction to ChIP-Seq analysis

(PHC 6934: Applied Computational Genomics)

Chengqi(Charley) Wang

USF Omics Hub, USF genomics program

The aim is to:

- have an understanding of the nature of ChIP-Seq data
- perform a complete analysis workflow including quality check (QC), read mapping, peak-calling. Use command line and open source software for each step of the workflow and feel the complexity of the task
- have an overview of possible downstream analyses

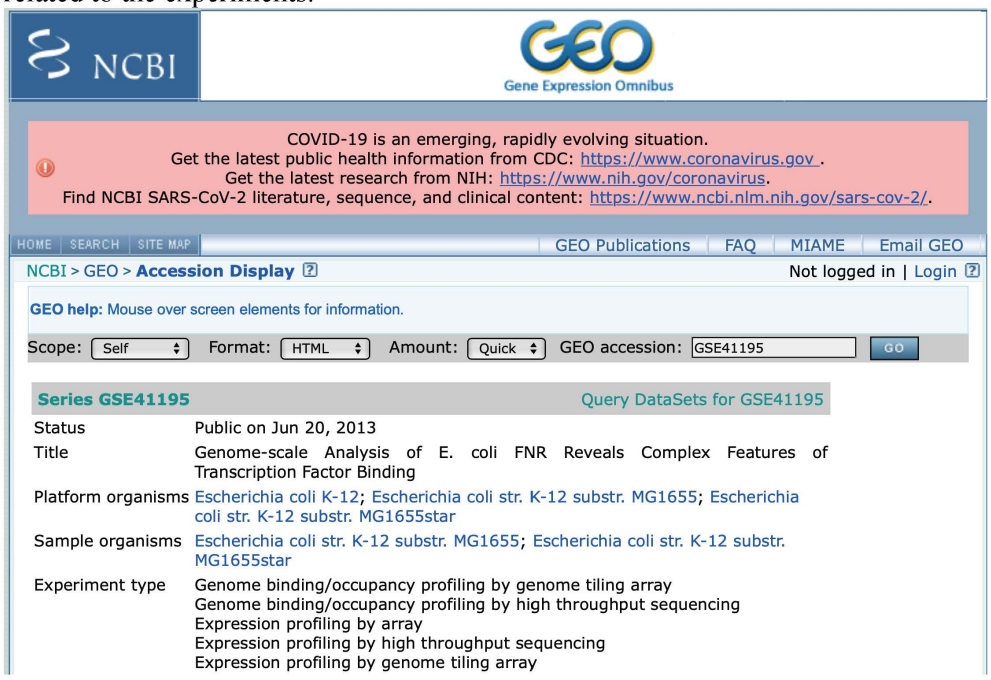
1. Downloading ChIP-Seq reads from NCBI

a. Obtaining an identifier for a chosen dataset

NGS datasets are usually made freely accessible for other scientists, by depositing these datasets into specialized databases. Functional genomic datasets (transcriptomics, genome-wide binding such ChIP-Seq,…) are deposited in the data bases Gene Expression Omnibus (GEO) .

b. Accessing GSE41195 from GEO

1) Search in google GSE41195. Click on the first link to directly access the correct page on the GEO database. The GEO database hosts processed data files and many details related to the experiments.



2) This GEO entry is a mixture of expression analysis and chip-seq. At the bottom of the page, click on the subseries related to the chip-seq datasets. (this subseries has its own identifier:GSE41187)

Submission date	Sep 27, 2012
Last update date	May 15, 2019
Contact name	Kevin Myers
E-mail(s)	kmyers2@wisc.edu
Organization name	University of Wisconsin - Madison
Department	Great Lakes Bioenergy Research Center
Street address	5120 WEI, 1552 University Ave
City	Madison
State/province	WI
ZIP/Postal code	53726
Country	USA

Platforms (3)	GPL8708	UW-Madison Escherichia coli K-12 MG1655 tiling array (YD Design)
	GPL14649	NimbleGen E. coli K12 Gene Expression Array [071112_Ecoli_K12_EXP]
	GPL16109	Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)
Samples (48)	GSM1010199	FNR - Anaerobic - A
More...	GSM1010200	FNR - Anaerobic - B
	GSM1010201	FNR - Anaerobic - C

This SuperSeries is composed of the following SubSeries:
More...

GSE41186 Chip-chip from Escherichia coli MG1655 K-12, WT and Δfnr strains

GSE41187 Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.

GSE41189 Expression analysis of Escherichia coli MG1655 K-12 WT and Δfnr mutant

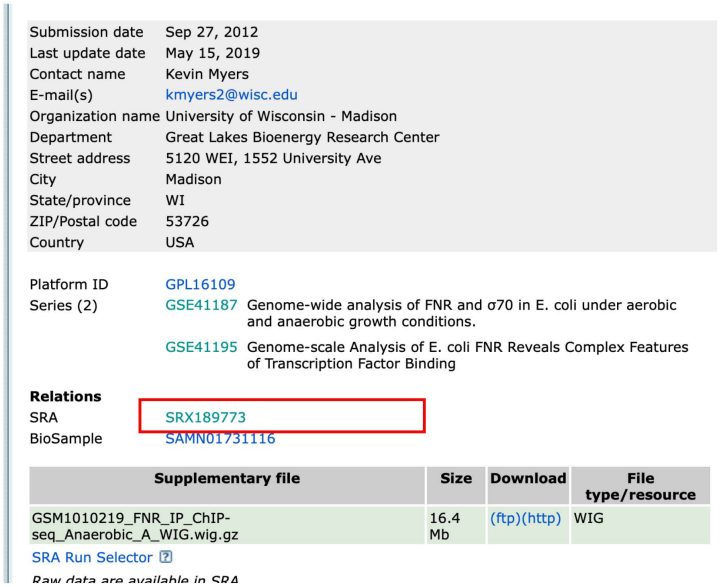
3) From the page, we will focus on the experiment **GSM1010219 FNR IP ChIP-seq Anaerobic A**. At the bottom of the page, click on the link "**GSM1010219 FNR IP ChIP-seq Anaerobic A**".

Overall design	Examination of occupancy of FNR adn σ70 under aerobic and anaerobic growth in conditions.
Contributor(s)	Myers K, Yan H, Ong I, Chung D, Liang K, Tran F, Keles S, Landick R, Kiley P
Citation(s)	Myers KS, Yan H, Ong IM, Chung D et al. Genome-scale analysis of escherichia coli FNR reveals complex features of transcription factor binding. *PLoS Genet* 2013 Jun;9(6):e1003565. PMID: 23818864
	Zuo C, Keleş S. A statistical framework for power calculations in ChIP-seq experiments. *Bioinformatics* 2014 Mar 15;30(6):753-60. PMID: 23665773

Submission date	Sep 27, 2012
Last update date	May 15, 2019
Contact name	Kevin Myers
E-mail(s)	kmyers2@wisc.edu
Organization name	University of Wisconsin - Madison
Department	Great Lakes Bioenergy Research Center
Street address	5120 WEI, 1552 University Ave
City	Madison
State/province	WI
ZIP/Postal code	53726
Country	USA

Platforms (1)	GPL16109	Illumina Genome Analyzer IIx (Escherichia coli str. K-12 substr. MG1655star)
Samples (9)	GSM1010219	FNR IP ChIP-seq Anaerobic A
Less...	GSM1010220	FNR IP ChIP-seq Anaerobic B
	GSM1010221	σ70 IP ChIP-seq Aerobic A
	GSM1010222	σ70 IP ChIP-seq Anaerobic A
	GSM1010223	aerobic INPUT DNA
	GSM1010224	anaerobic INPUT DNA
	GSM1072326	σ70 IP ChIP-seq Aerobic B
	GSM1072327	σ70 IP ChIP-seq Anaerobic B

4) In the new page, go to the bottom to find the SRA identifier. This is the identifier of the raw dataset stored in the SRA database.



Submission date	Sep 27, 2012
Last update date	May 15, 2019
Contact name	Kevin Myers
E-mail(s)	kmyers2@wisc.edu
Organization name	University of Wisconsin - Madison
Department	Great Lakes Bioenergy Research Center
Street address	5120 WEI, 1552 University Ave
City	Madison
State/province	WI
ZIP/Postal code	53726
Country	USA

Platform ID	GPL16109	
Series (2)	GSE41187	Genome-wide analysis of FNR and σ70 in E. coli under aerobic and anaerobic growth conditions.
	GSE41195	Genome-scale Analysis of E. coli FNR Reveals Complex Features of Transcription Factor Binding

Relations

SRA	SRX189773
BioSample	SAMN01731116

Supplementary file	Size	Download	File type/resource
GSM1010219_FNR_IP_ChIP-seq_Anaerobic_A_WIG.wig.gz	16.4 Mb	(ftp)(http)	WIG

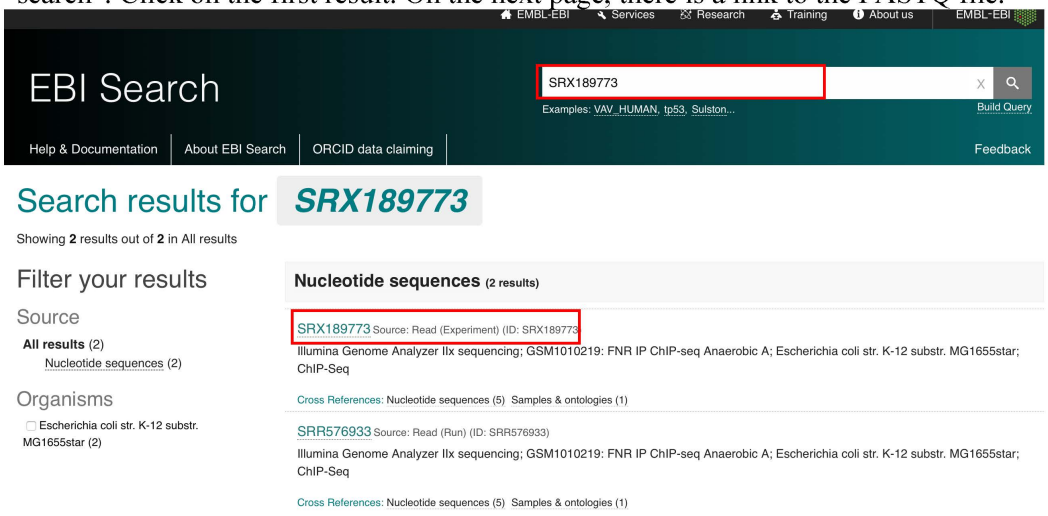
SRA Run Selector

Raw data are available in SRA

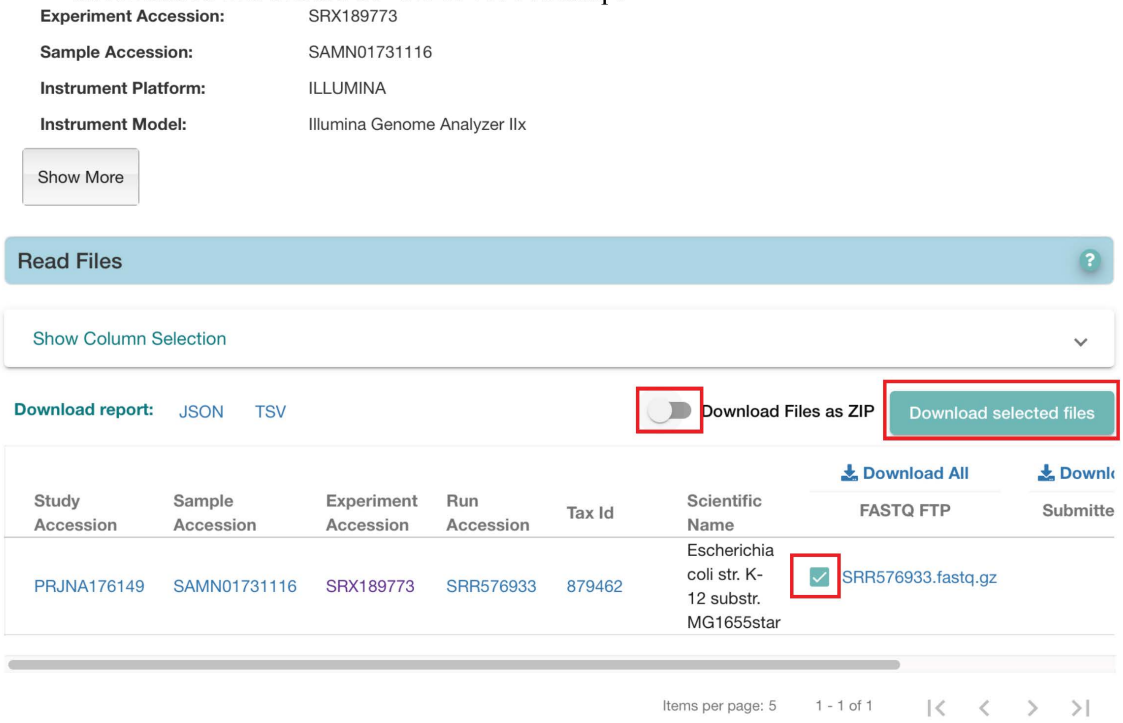
5) Copy the identifier SRX189773 (do not click on the link that would take you to the SRA database. Although direct access to the SRA database at the NCBI is doable, SRA does not store the sequence FASTQ format. In practice, it's simpler and quicker to downloaded datasets from the ENA database(European Nucleotide Archive) hosted by EBI (European Bioinformatics Institute) in UK. ENA encompasses the data from SRA)

c. Downloading FASTQ file from the ENA database

1) Go to the EBI website, paste the SRA identifier (SRX189773) and click on the button "search". Click on the first result. On the next page, there is a link to the FASTQ file.



2) Please select SRR576933.fastq.gz and click 'Download selected files'. The downloaded file should be 'SRR576933.fastq'.



3) Please download the control dataset. We should redo the same steps starting from GEO web page specific to the chip-seq datasets (see step B), and select '**GSE41195**' -> '**GSM1010224 anaerobic INPUT DNA**' -> copy 'SRX189778' -> go to EBI and search 'SRX189778' -> download FASTQ file 'SRR576938.fastq'.

2. Transfer the two downloaded FASTQ file to the student cluster (sc.rc.usf.edu)

a. Using terminal login the student clustering

Opening the terminal and type: 'ssh your_user_name@sc.rc.usf.edu', after typing the password, you should successfully log in to the student clustering.

```
Charleys-MBP-2:~ charleywang$ ssh [redacted]@sc.rc.usf.edu
[redacted]@sc.rc.usf.edu's password:
  Documentation:  https://wiki.rc.usf.edu
        Support:  rc-help@usf.edu
          Phone:  813-974-1222

   To change your password or to recover a lost password, please visit
                        https://netid.usf.edu

   Please report any and all problems to the Helpdesk at rc-help@usf.edu
              so we can track your incident!



[chengqi@scln2 ~]$
```

b. Building the folder for ChIP-Seq analysis

1) Double make sure your working directory is '/home/your_netID', either by checking the '~' inside the bracket or typing 'pwd'.

```
[chengqi@scln2 ~]$ pwd
/home/c/chengqi
[chengqi@scln2 ~]$
```

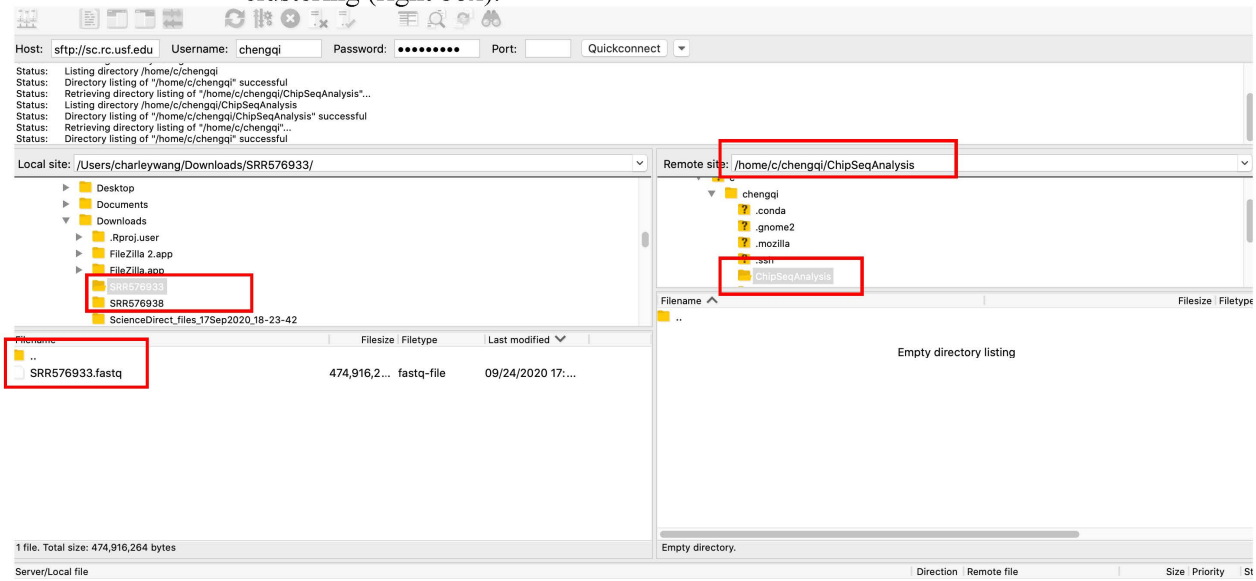
2) If not in your working directory, please move to your home directory by typing 'cd ~'.

3) Building the folder with name 'ChipSeqAnalysis', by typing 'mkdir ChipSeqAnalysis'. And type 'ls' to double check whether the folder has been successfully generated.

```
[chengqi@scln2 ~]$ mkdir ChipSeqAnalysis
[chengqi@scln2 ~]$ ls
ChipSeqAnalysis  Desktop  genomeTrain  test  test.txt  test_GSE77565  test_GSE77565_notSlurm
```

4) Open FileZilla and type 'sc.rc.usf.edu' in Host, Username, Password and '22' in Port. Then click 'Quickconnect' to connect student clustering.

5) After login, opening your folder with the downloaded FASTQ file in your local computer (left box). Also, opening the folder 'ChipSeqAnalysis' in the student clustering (right box).



6) Drag the two FASTQ file to the folder 'ChipSeqAnalysis' and finish the copy. Back to the terminal and type 'ls -lh ~/ChipSeqAnalysis/'

```
[chengqi@scln2 ~]$ ls -lh ~/ChipSeqAnalysis/
total 1.3G
-rw-r----- 1 chengqi usfuser 453M Sep 24 21:32 SRR576933.fastq
-rw-r----- 1 chengqi usfuser 846M Sep 24 21:34 SRR576938.fastq
```

3. Organism length and reference genome FASTA file download

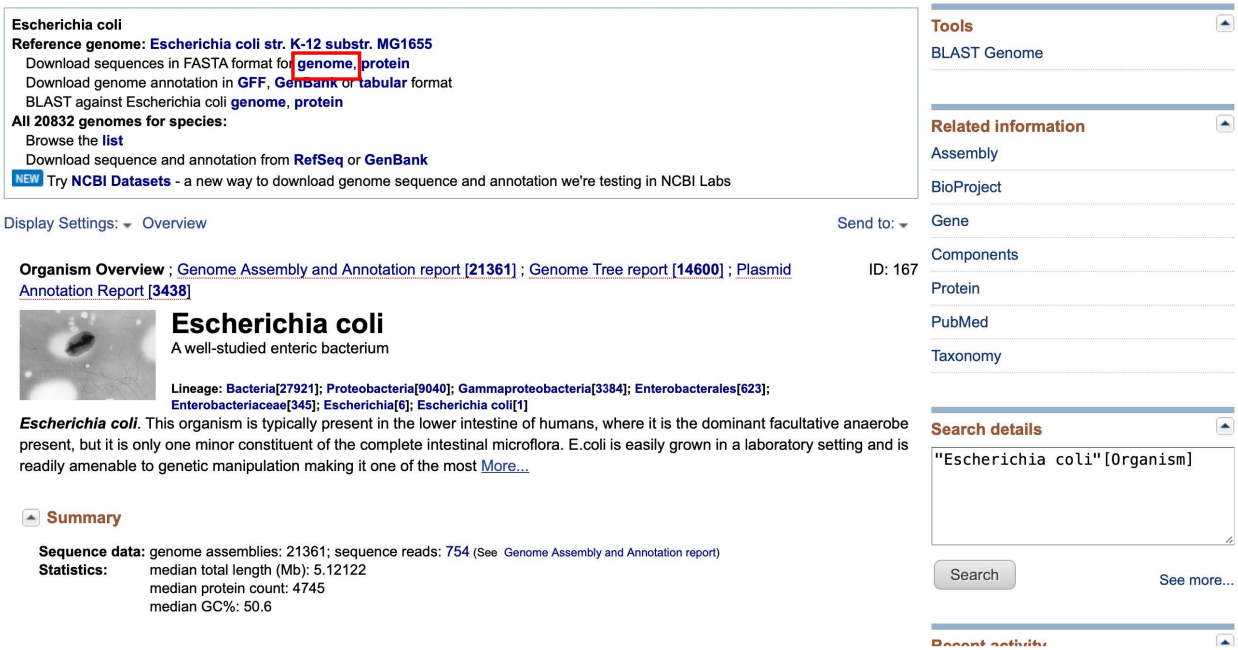
Knowing your organism size is important to evaluate if your dataset has sufficient coverage to continue your analyses. For the human genome (3 Gb), we usually aim at least 10 Million reads.

a. Go to the NCBI Genome website (https://www.ncbi.nlm.nih.gov/genome), and search for the organism Escherichia coli

b. Click on the Escherichia coli str. K-12 substr. MG1655 to access statistics on this genome.

The genome is 4.64 Mbase. The files have respectively 3.6 M reads and 6.7 M reads. As we aim for 10 M for 3 Gb when working with human, the dataset here should enough reads for proper analysis.

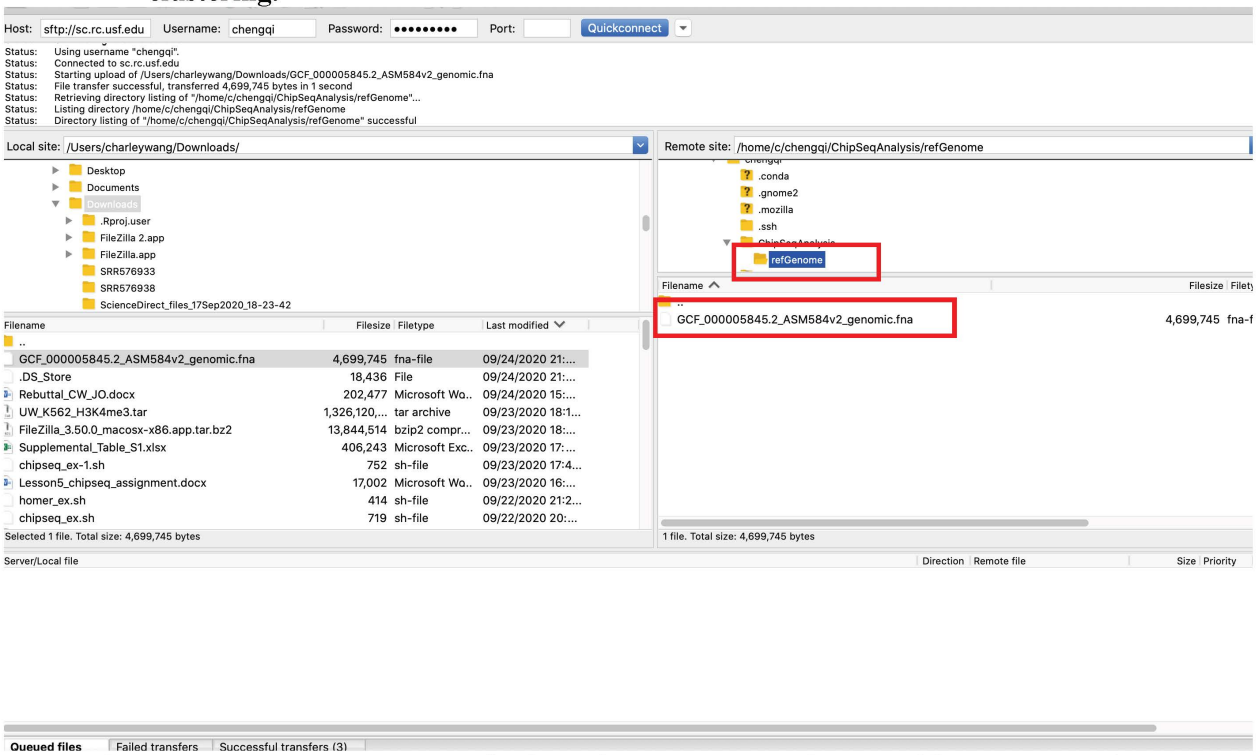
c. Downloading the genome FASTA file by click 'genome'



d. In terminal, building a new folder for reference genome file by typing 'mkdir ~/ChipSeqAnalysis/refGenome'

```
[chengqi@scln2 ~]$ mkdir ~/ChipSeqAnalysis/refGenome
```

e. **By using FileZilla, copy the downloading FAST file 'GCF_000005845.2_ASM584v2_genomic.fna' to the folder 'refGenome' in student clustering.**



f. **Double check whether the file is available under the folder refGenome by typing 'ls -lh ~/ChipSeqAnalysis/refGenome/'**

```
[chengqi@scln2 ~]$ ls -lh ~/ChipSeqAnalysis/refGenome/
total 4.5M
-rw-r----- 1 chengqi usfuser 4.5M Sep 24 21:57 GCF_000005845.2_ASM584v2_genomic.fna
```

4. Prepare the index file

Indexing a genome can be explained similar to indexing a book. If you want to know on which page a certain word appears or a chapter begins, it is much more efficient/faster to look it up in a pre-built index than going through every page of the book until you found it. Same goes for alignments. Indices allow the aligner to narrow down the potential origin of a query sequence within the genome, saving both time and memory.

a. **In terminal (Need to be in student clustering), loading the mapping software by typing 'module purge', then, 'module add apps/bowtie/2.3.5.1'**

```
[chengqi@scln2 ~]$ module purge
[chengqi@scln2 ~]$ module add apps/bowtie/2.3.5.1
[chengqi@scln2 ~]$
```

Research Computing uses a module system to load most software into a user's environment. Most software is not accessible by default and must be loaded in. This allows Research Computing to provide multiple versions of the software concurrently and enables users to easily switch between different versions.

b. **Building the index for bowtie by typing**
'bowtie2-build
~/ChipSeqAnalysis/refGenome/GCF_000005845.2_ASM584v2_genomic.fna
~/ChipSeqAnalysis/refGenome/GCF_000005845.2_ASM584v2_genomic'

```
[chengqi@scln2 ~]$ bowtie2-build ~/ChipSeqAnalysis/refGenome/GCF_000005845.2_ASM584v2_genomic.f
na ~/ChipSeqAnalysis/refGenome/GCF_000005845.2_ASM584v2_genomic
```

c. **Checking whether the index file is available for Bowtie by typing**
'ls -lh ~/ChipSeqAnalysis/refGenome/'
A couple of files ending with '.bt2' should be available under the folder 'ls -lh ~/ChipSeqAnalysis/refGenome'

```
[chengqi@scln2 ~]$ ls -lh ~/ChipSeqAnalysis/refGenome/
total 19M
-rw-r----- 1 chengqi usfuser 5.5M Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.1.bt2
-rw-r----- 1 chengqi usfuser 1.2M Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.2.bt2
-rw-r----- 1 chengqi usfuser   17 Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.3.bt2
-rw-r----- 1 chengqi usfuser 1.2M Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.4.bt2
-rw-r----- 1 chengqi usfuser 4.5M Sep 24 21:57 GCF_000005845.2_ASM584v2_genomic.fna
-rw-r----- 1 chengqi usfuser 5.5M Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.rev.1.bt2
-rw-r----- 1 chengqi usfuser 1.2M Sep 25 10:37 GCF_000005845.2_ASM584v2_genomic.rev.2.bt2
```

5. Quality control of the reads

Now, everything is ready for mapping. But, before mapping the FASTQ files (reads files, SRR576933.fastq SRR576938.fastq) need to be evaluated and make sure enough quality of these file for mapping.

a. **In terminal (Need to be in student clustering), loading the quality control software by typing 'module purge', then, 'module add apps/fastqc/0.11.5'**

```
[chengqi@scln2 ~]$ module purge
[chengqi@scln2 ~]$ module add apps/fastqc/0.11.5
```

b. **Launch the FASTQC program on the experiment file (SRR576933.fastq), by typing 'fastqc ~/ChipSeqAnalysis/SRR576933.fastq'**

```
[chengqi@scln2 ~]$ fastqc ~/ChipSeqAnalysis/SRR576933.fastq
```

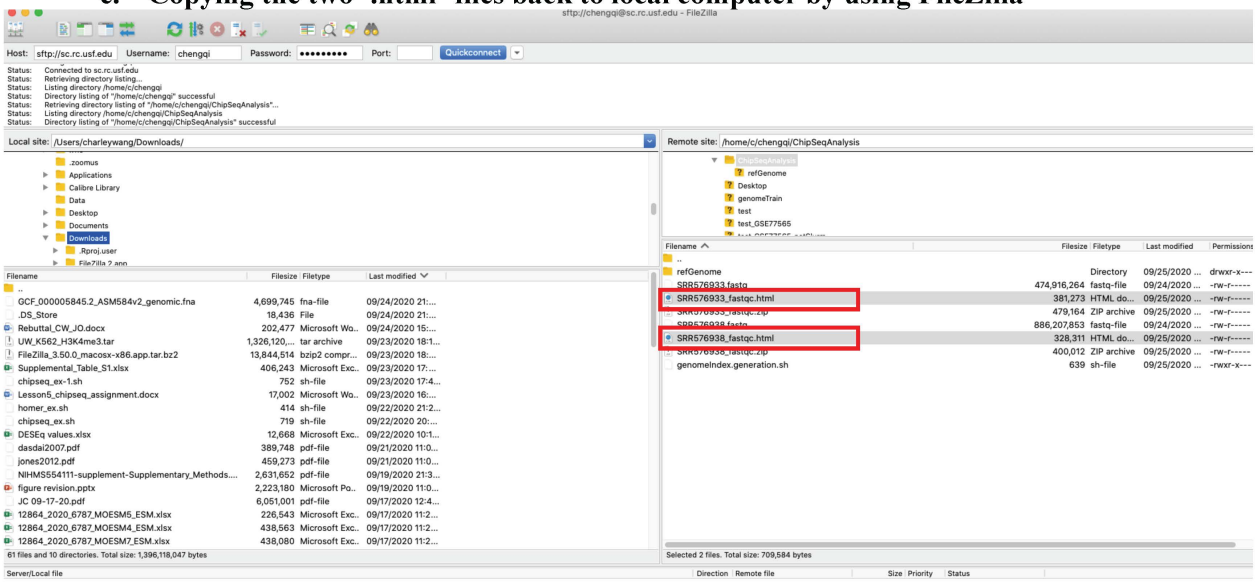
c. **Similarly, performing quality control for control file (SRR576938.fastq)**

```
[chengqi@scln2 ~]$ fastqc ~/ChipSeqAnalysis/SRR576938.fastq
```

d. **The results from quality control (fastqc) are saved in the file ending with '.html'. Copy the two html files 'SRR576933_fastqc.html' and 'SRR576938_fastqc.html' back to your personal local computer from student clustering.**
First, check the two files location by typing
'ls -lh ~/ChipSeqAnalysis/'

```
[chengqi@scln2 ~]$ ls -lh ~/ChipSeqAnalysis/
total 1.3G
-rw-r----- 1 chengqi usfuser 453M Sep 24 21:32 SRR576933.fastq
-rw-r----- 1 chengqi usfuser 373K Sep 25 11:20 SRR576933_fastqc.html
-rw-r----- 1 chengqi usfuser 468K Sep 25 11:20 SRR576933_fastqc.zip
-rw-r----- 1 chengqi usfuser 846M Sep 24 21:34 SRR576938.fastq
-rw-r----- 1 chengqi usfuser 321K Sep 25 11:23 SRR576938_fastqc.html
-rw-r----- 1 chengqi usfuser 391K Sep 25 11:23 SRR576938_fastqc.zip
-rwxr-x--- 1 chengqi usfuser  639 Sep 25 10:18 genomeIndex.generation.sh
drwxr-x--- 2 chengqi usfuser 4.0K Sep 25 10:37 refGenome
```

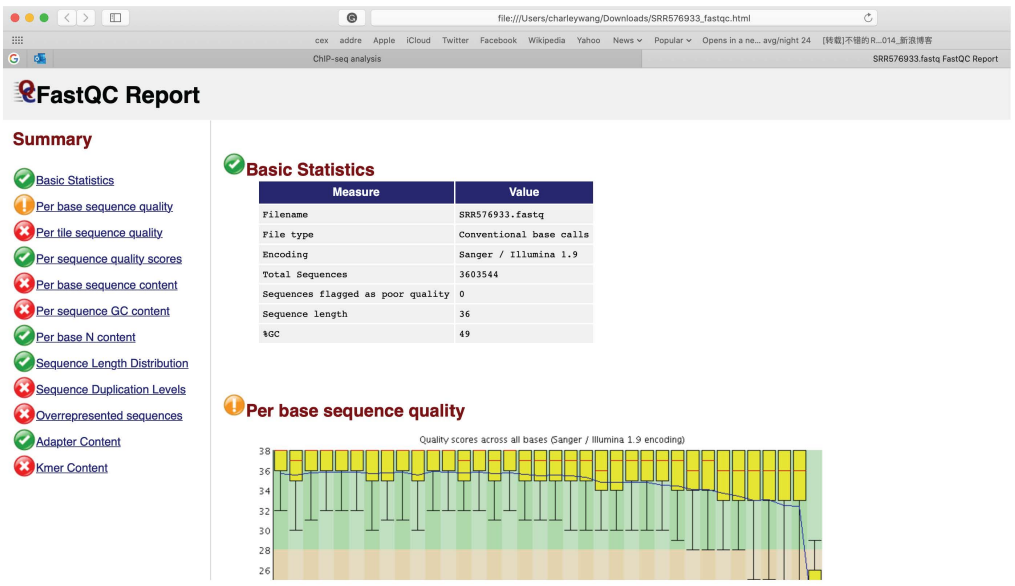
e. Copying the two '.html' files back to local computer by using FileZilla



f. Open the HTML file by double click, or by Firefox, Safari or Chrome

Analyze the result of the FASTQC program: How many reads are present in the file ?
What is the read length ?
Is the overall quality good ?
Are there any concerns raised by the report ?
If so, can you tell where the problem might come from ?

There are 3 603 544 reads of 36bp. The overall quality is good, although it drops at the last position, which is usual with Illumina sequencing, so this feature is not raising hard concerns. There are several "red lights" in the report. In particular, the per sequence GC content and the duplication level are problematic. If you check the "overrepresented sequences", you'll notice a high percentage of adapters (29%!). Ideally, we would remove these adapters (=trim) the reads, and then re-run FASTQC. In practice, we often skip this step, as these reads will anyway not be mapped. Warning: this will affect the future calculated"% of mapped read"!!



6. Mapping the reads with Bowtie

There are multiple programs to perform the mapping step. For reads produced by an Illumina machine for ChIP-seq, the currently "standard" programs are BWA and Bowtie (versions 1 and 2), and STAR is getting popular. We will use Bowtie version 1.1.1 for this exercice, as this program remains effective for short reads (< 50bp).

a. Mapping the experiment and control data by using the shell script 'mapping.sh' in the CANVAS.

Please use 'vim' to change the working directory to your own working directory, and change the email address to your email address. Submitting the mapping.sh by 'sbatch peakCalling.sh'

```
#!/bin/bash

#SBATCH --workdir=/home/c/chengqi/ChipSeqAnalysis/
#SBATCH --job-name=ChIPseqMapping
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --mem=10000
#SBATCH -t 00:10:00
#SBATCH -o run.out
#SBATCH -e run.err
#SBATCH --mail-user=
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL

module purge
module add apps/bowtie/2.3.2
bowtie2 -x refGenome/GCF_000005845.2_ASM584v2_genomic -3 1 -q SRR576933.fastq -S SRR576933.sam
2> SRR576933.out
bowtie2 -x refGenome/GCF_000005845.2_ASM584v2_genomic -3 1 -q SRR576938.fastq -S SRR576938.sam
2> SRR576938.out
```

Please using 'pwd' find out your 'ChipSeqAnalysis' directory and paste here.

Back terminal, you can type 'cd ~/ChipSeqAnalysis', then type 'pwd'

Let's see the parameters of bowtie2 before launching the mapping:

- refGenome/GCF_000005845.2_ASM584v2_genomic is the name of our genome index file
- -q indicates the input file is in FASTQ format
- SRR576933.fastq is the name of our FASTQ file.
- -3 will trim x base from the end of the read. As our last position is of low quality, we'll trim 1 base
- -S will output the result in SAM format
- SRR576938.sam is the output sam file
- 2> SRR576933.out will output some statistics about the mapping in the file SRR576933.out

b. Using 'sbatch mapping.sh' to submit the job. This should take few minutes as we work with a small genome. For the human genome, we would need either more time, or a dedicated server.

Analyze the result of the mapped reads:
Open the file SRR576933.out and SRR576938.out, which contains some statistics about the mapping. How many reads were mapped?
At this point, you should have two SAM files, one for the experiment, one for the control. Check the size of your files, how large are they?

7. Peak calling with MACS2 (Thanks the nice Conda environment setting from Jenna Oberstaller)

For loading MACS2, after login the student clustering, please typing

- module purge
- module add apps/miniconda/3.6.1-intel
- conda activate /[redacted]/.conda/envs/macs2/
- unset PYTHONPATH

```
[chengqi@scln0 ~]$ module purge
[chengqi@scln0 ~]$ module add apps/miniconda/3.6.1-intel
[chengqi@scln0 ~]$ conda activate /[redacted]/.conda/envs/macs2/
(macs2) [chengqi@scln0 ~]$ unset PYTHONPATH
```

You can also use the MACS2 environment generated during Thomas Keller's class

a. Calling the peak by using the shell script 'by using the shell script 'peakCalling.sh' in the CANVAS.

Please use 'vim' to change the working directory to your own working directory, and change the email address to your email address. Submitting the peakCalling.sh by 'sbatch peakCalling.sh'

```
#!/bin/bash

#SBATCH --workdir=/home/c/chengqi/ChipSeqAnalysis/
#SBATCH --job-name=ChIPseqMapping
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --mem=10000
#SBATCH -t 00:10:00
#SBATCH -o run.out
#SBATCH -e run.err
#SBATCH --mail-use[redacted]
#SBATCH --mail-type=BEGIN
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL

module purge
module add apps/miniconda/3.6.1-intel
conda activate /[redacted]/.conda/envs/macs2
unset PYTHONPATH
macs2 callpeak -t SRR576933.sam -c SRR576938.sam -n MACSpeaks -q 0.05 --gsize 4639675
 --keep-dup 1 --nomodel --extsize 400
```

Please using 'pwd' find out your 'ChipSeqAnalysis' directory and paste here.

Back terminal, you can type 'cd ~/ChipSeqAnalysis', then type 'pwd'

Let's see the parameters of MACS before launching the mapping:

- ChIP-seq tag file (-t) is the name of our experiment (treatment) mapped read file SRR576933.sam
- ChIP-seq control file (-c) is the name of our input (control) mapped read file SRR576938.sam
- --gsize Effective genome size: this is the size of the genome considered "usable" for peak calling. This value is given by the MACS developpers on their website. It is smaller than the complete genome because many regions are excluded (telomeres, highly repeated regions...). The default value is for human (2700000000.0), so we need to chnage it. As the value for E. coli is not provided, we will take the complete genome size 4639675.

- -n provides a prefix for the output files. We set this to MACSpeaks, but it could be any name.
- -q indicates the FDR cutoff for picking up peaks
- --keep-dup specifies how MACS should treat the reads that are located at the exact same location (duplicates). The manual specifies that keeping only 1 representative of these "stacks" of reads is giving the best results.
- --nomodel While on, MACS will bypass building the shifting model.
- --extsize While --nomodel is set, MACS uses this parameter to extend reads in 5'->3' direction to fix-sized fragments. For example, if the size of the binding region for your transcription factor is 200 bp, and you want to bypass the model building by MACS, this parameter can be set as 200. This option is only valid when --nomodel is set or when MACS fails to build model and --fix-bimodal is on.

b. Analyzing the MACS results

Look at the files that were created by MACS. Which files contains which information? Please typing 'ls -lh ~/ChipSeqAnalysis/'

```
(macs2) [chengqi@scln1 ChipSeqAnalysis]$ ls -lh ~/ChipSeqAnalysis/
total 3.0G
-rw-r----- 1 chengqi usfuser  15K Sep 26 09:46 MACSpeaks_peaks.narrowPeak
-rw-r----- 1 chengqi usfuser  18K Sep 26 09:46 MACSpeaks_peaks.xls
-rw-r----- 1 chengqi usfuser  11K Sep 26 09:46 MACSpeaks_summits.bed
-rw-r----- 1 chengqi usfuser 453M Sep 24 21:32 SRR576933.fastq
-rw-r----- 1 chengqi usfuser  217 Sep 25 13:26 SRR576933.out
-rw-r----- 1 chengqi usfuser 550M Sep 25 13:26 SRR576933.sam
-rw-r----- 1 chengqi usfuser 373K Sep 25 11:20 SRR576933_fastqc.html
-rw-r----- 1 chengqi usfuser 468K Sep 25 11:20 SRR576933_fastqc.zip
-rw-r----- 1 chengqi usfuser 846M Sep 24 21:34 SRR576938.fastq
-rw-r----- 1 chengqi usfuser  214 Sep 25 13:29 SRR576938.out
-rw-r----- 1 chengqi usfuser 1.2G Sep 25 13:29 SRR576938.sam
-rw-r----- 1 chengqi usfuser 321K Sep 25 11:23 SRR576938_fastqc.html
-rw-r----- 1 chengqi usfuser 391K Sep 25 11:23 SRR576938_fastqc.zip
-rwxr-x--- 1 chengqi usfuser  606 Sep 25 13:25 mapping.sh
-rwxr-x--- 1 chengqi usfuser  580 Sep 26 09:45 peakCalling.sh
drwxr-x--- 2 chengqi usfuser 4.0K Sep 25 10:37 refGenome
-rw-r----- 1 chengqi usfuser 4.1K Sep 26 09:46 run.err
-rw-r----- 1 chengqi usfuser    0 Sep 25 13:18 run.out
```

- MACSpeaks_peaks.narrowPeak: peak coordinates in BED format
- MACSpeaks_peaks.xls: peak coordinates with more information, to be opened with Excel
- MACSpeaks_summits.bed: location of the summit base for each peak (BED format)

How many peaks were detected by MACS?

```
(macs2) [chengqi@scln1 ChipSeqAnalysis]$ wc -l MACSpeaks_peaks.narrowPeak
191 MACSpeaks_peaks.narrowPeak
```