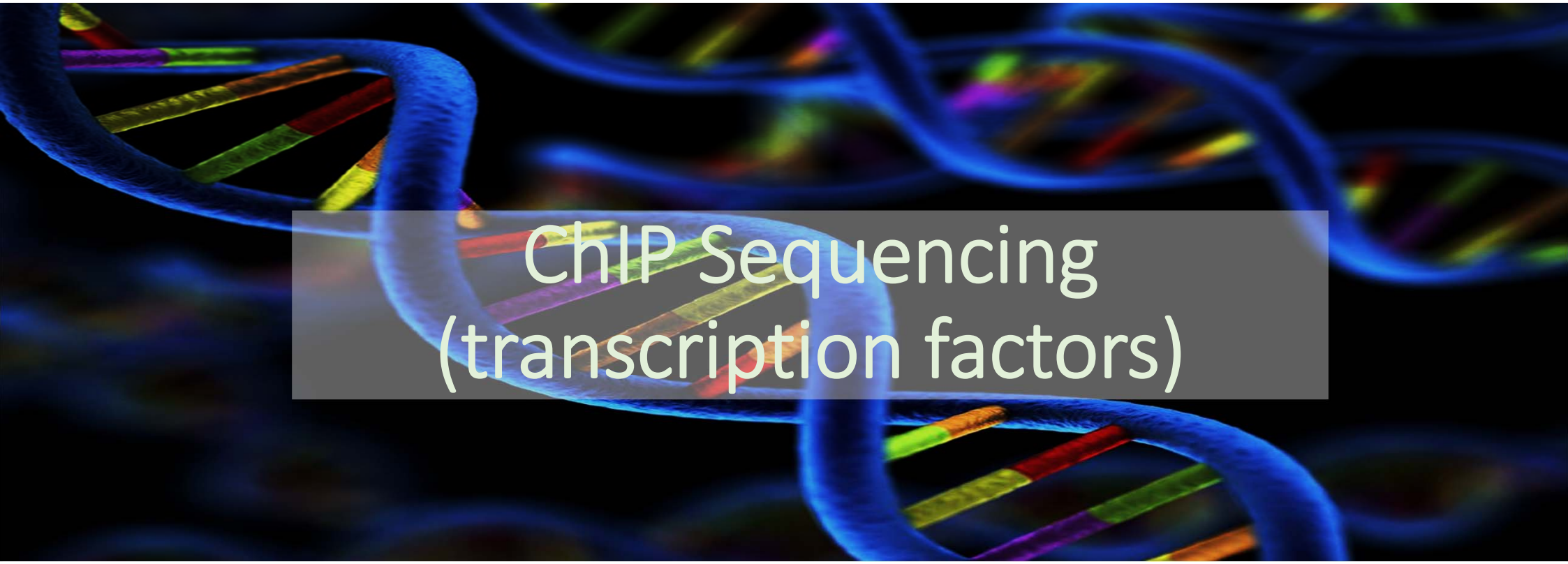

ChIP Sequencing
(transcription factors)

Outline

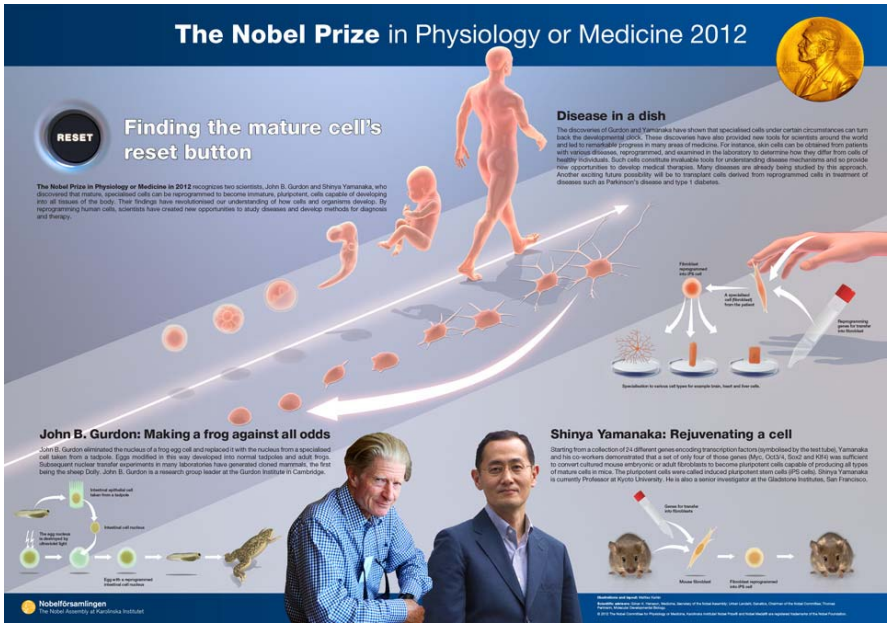
- Introduction to transcription factor
- ChIP-seq Experimental Procedure
- Data Analysis Pipeline

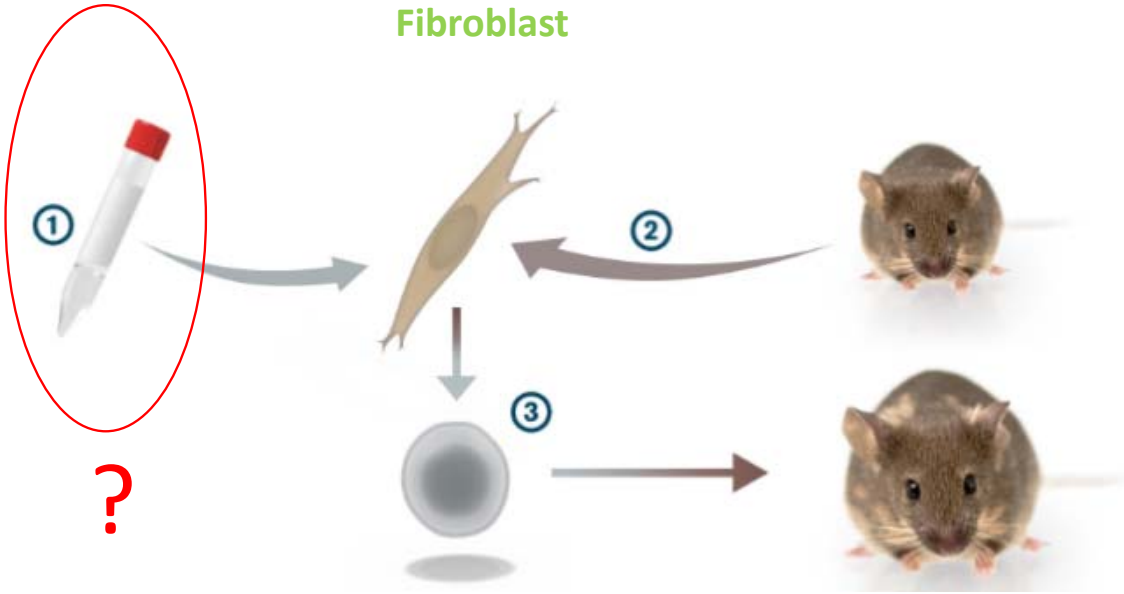
- Case Study I
- Case Study II (Optional)

- Suggested reading
  - **A census of human transcription factors: function, expression and evolution** Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann & Nicholas M. Luscombe *Nature Review Genetics* 2009
  - **The Human Transcription Factors** Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, Matthew T. Weirauch *Cell* 2018

2012 Nobel Prize in Physiology or Medicine

Yamanaka factors (Oct3/4, Sox2, Klf4, c-Myc) → **Transcription Factors (TF)**





2019 Nobel Prize in Physiology or Medicine

The Nobel Prize in Physiology or Medicine 2019

Ill. Niklas Elmedhed. © Nobel Media.
William G. Kaelin Jr
Prize share: 1/3

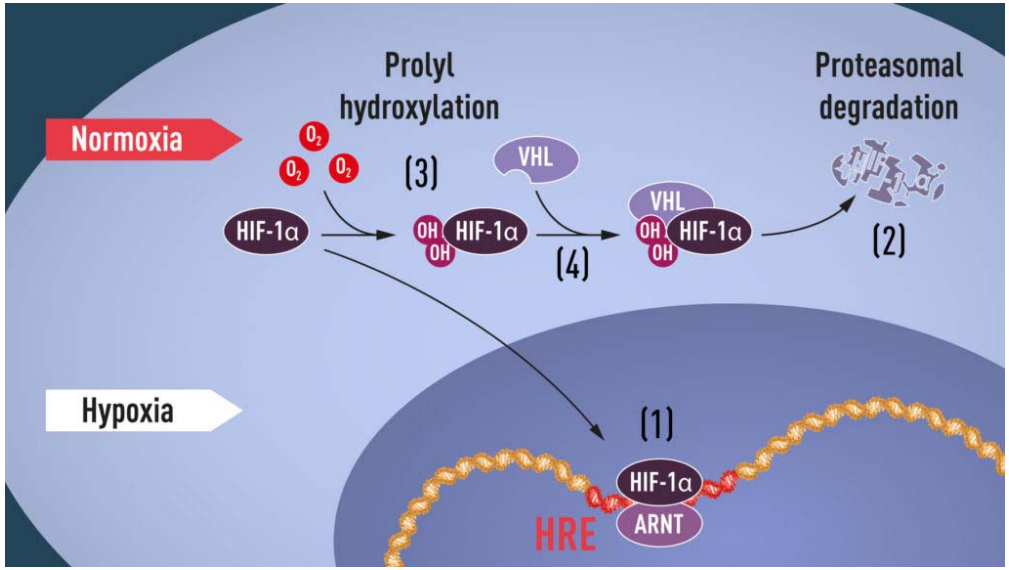
Ill. Niklas Elmedhed. © Nobel Media.
Sir Peter J. Ratcliffe
Prize share: 1/3

Ill. Niklas Elmedhed. © Nobel Media.
Gregg L. Semenza
Prize share: 1/3

The Nobel Prize in Physiology or Medicine 2019 was awarded jointly to William G. Kaelin Jr, Sir Peter J. Ratcliffe and Gregg L. Semenza "for their discoveries of how cells sense and adapt to oxygen availability."

HIF-1-alpha → **Transcription Factors (TF)**

- The dysregulation and overexpression of HIF1A by either hypoxia or genetic alternations have been heavily implicated in cancer biology, as well as a number of other pathophysiology, specifically in areas of vascularization and angiogenesis, energy metabolism, cell survival, and tumour invasion.



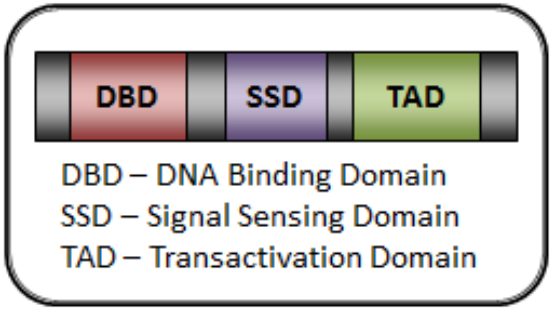
Transcription Factor

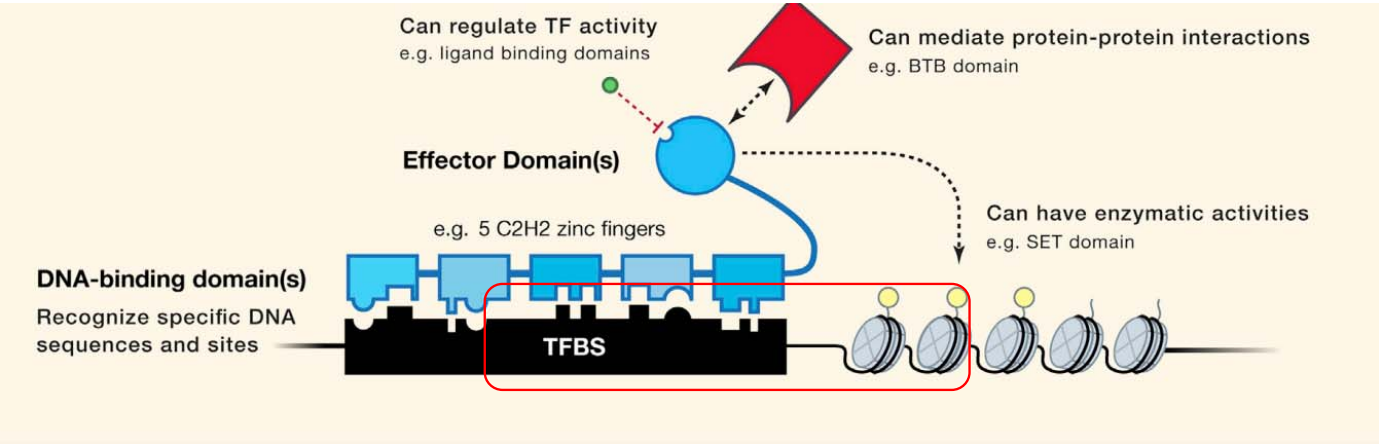
- General transcription factor and sequence specific transcription factor
  - Mediator, transcription machinery, structural proteins
- TF is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.
- The function of TFs is to regulate—turn on and off—genes in order to make sure that they are expressed in the right cell at the right time and in the right amount throughout the life of the cell and the organism.

Sequence Specific Transcription Factor Binding → Tissue Specific Gene Expression (epigenetics)

DNA Binding Domain (DBD)

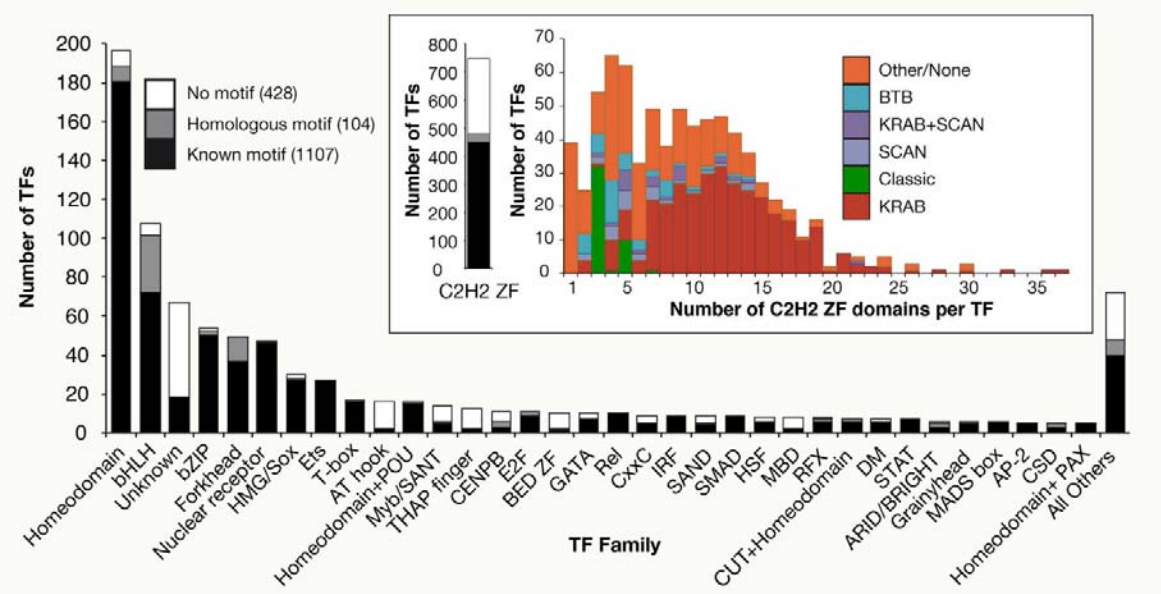
- Most TFs are composed of three parts
  - DBD – DNA binding (sequence specificities)
  - SSD – signal sensing (ligand binding or modifications)
  - TAD – transactivation (releasing transcription initiation)





TF Family

- TFs can be classified according to DBD structures
- In human, there are ~1,600 TFs (out of ~21,000 genes)
- DBD alone determines sequence specificities of the TF



- [A census of human transcription factors: function, expression and evolution] Juan M. Vaquerizas, Sarah K. Kummerfeld, Sarah A. Teichmann & Nicholas M. Luscombe *Nature Review Genetics* 2009
- [The Human Transcription Factors] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, Matthew T. Weirauch *Cell* 2018

DNA Binding Specificities

- Databases that store the information of DNA binding specificities of TFs:
  - TRANSFAC (Matys et al., 2006)
    - TRANSFAC® is the database of eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles.
    - It was developed by a private company. Therefore, one needs to pay and use
    - http://genexplain.com/transfac/
  - JASPAR (Mathelier et al., 2016)
    - Free database
    - **JASPAR** is an open-access database of curated, non-redundant transcription factor (TF) binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups.
    - http://jaspar.genereg.net/

Position Weight Matrix (PWM)

- Consensus sequences
  - Most preferred individual sequences
  - A[CT]N{A}YR (A means that an A is always found in that position; [CT] stands for either C or T; N stands for any base; and {A} means any base except A. Y represents any pyrimidine, and R indicates any purine.)

- PWM is a commonly used representation of motifs (patterns) in biological sequences.
- PWMs are often derived from a set of aligned sequences that are thought to be functionally related and have become an important part of many software tools for computational motif discovery.

Position Weight Matrix (PWM)

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

- Position Frequency Matrix (PFM)
  - Counting frequency for each nucleotide at each position

$$\begin{matrix} A \\ C \\ G \\ T \end{matrix}\begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

- Position Probability Matrix (PPM)
  - Convert frequency to probability (0-1)

$$\begin{matrix} A \\ C \\ G \\ T \end{matrix}\begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

Position Probability Matrix (PPM)

- Position Probability Matrix (PPM)

$$\begin{matrix} A \\ C \\ G \\ T \end{matrix}\begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

- Each column can therefore be regarded as an independent multinomial distribution. This makes it easy to calculate the probability of a sequence given a PPM, by multiplying the relevant probabilities at each position.
- For example, the probability of the sequence S = GAGGTAAAC given the above PPM M can be calculated:

$$p(S|M) = 0.1 \times 0.6 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 0.7 \times 0.2 \times 0.2 = 0.0007056.$$

Position Weight Matrix (PWM)

- Most often the elements in PWMs are calculated as log likelihoods.

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix} \quad M_{k,j} = \log_2 \left(M_{k,j}/b_k\right) \quad \begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 0.26 & 1.26 & -1.32 & -\infty & -\infty & 1.26 & 1.49 & -0.32 & -1.32 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & -0.32 & -1.32 & -1.32 & -0.32 \\ -1.32 & -1.32 & 1.49 & 2.0 & -\infty & -1.32 & -1.32 & 1.0 & -1.32 \\ 0.68 & -1.32 & -1.32 & -\infty & 2.0 & -1.32 & -1.32 & -0.32 & 1.26 \end{bmatrix}$$

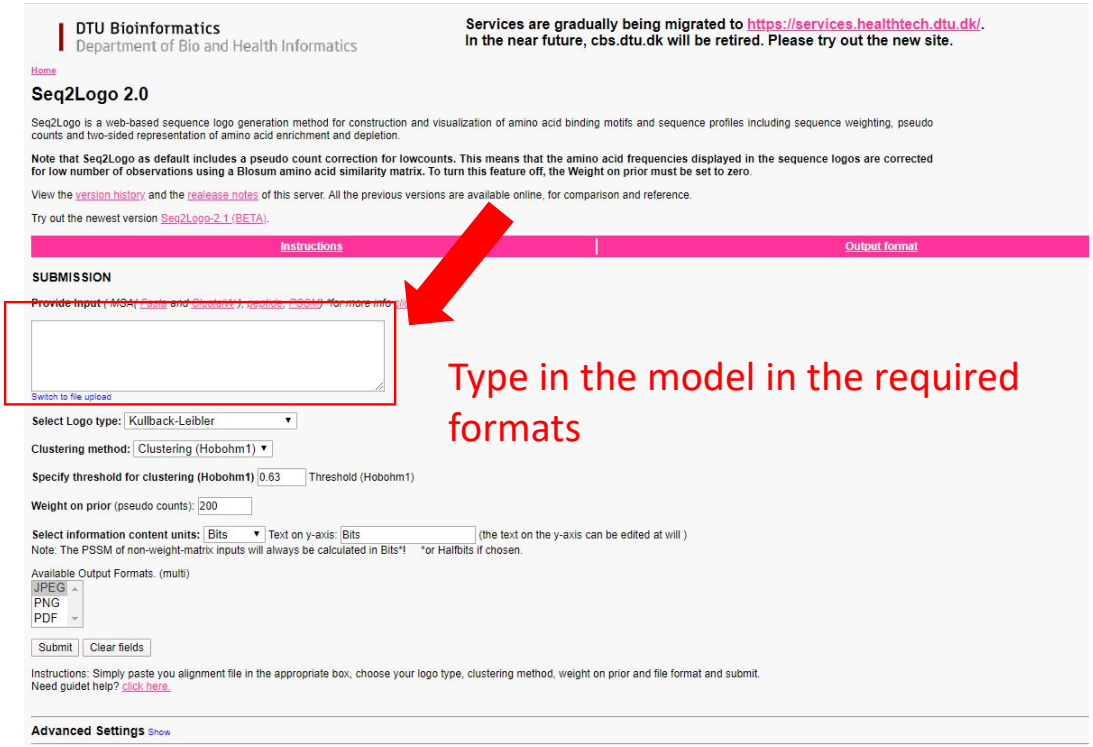
  - bk indicates the background: the simplest would be **0.25**
  - The score is 0 if the sequence has the same probability of being a functional site and of being a random site.
  - The information content (IC) of a PWM is sometimes of interest, as it says something about how different a given PWM is from a **uniform distribution**.

- There are various algorithms to scan for hits of PWMs in sequences. One example is the MATCH algorithm.

Generating a Consensus Logo to illustrate PWM

- A consensus logo is a simplified variation that can be embedded in text format. A consensus logo is created from a collection of DNA sequences and conveys information about the conservation of each position of a sequence motif or sequence alignment
- There are various web-based tools to generate the consensus logos from PWM/PPM/PFM,
  - http://www.cbs.dtu.dk/biotools/Seq2Logo/

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

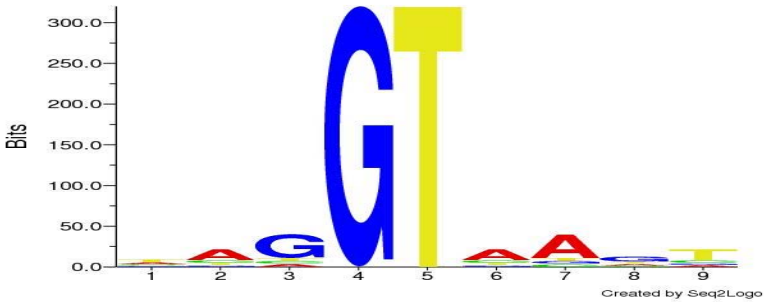


$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

formatting

	A	C	G	T
1	3	2	1	4
2	6	2	1	1
3	1	1	7	1
4	0	0	10	0
5	0	0	0	10
6	6	2	1	1
7	7	1	1	1
8	2	1	5	2
9	1	2	1	6

http://www.cbs.dtu.dk/biotools/Seq2Logo/

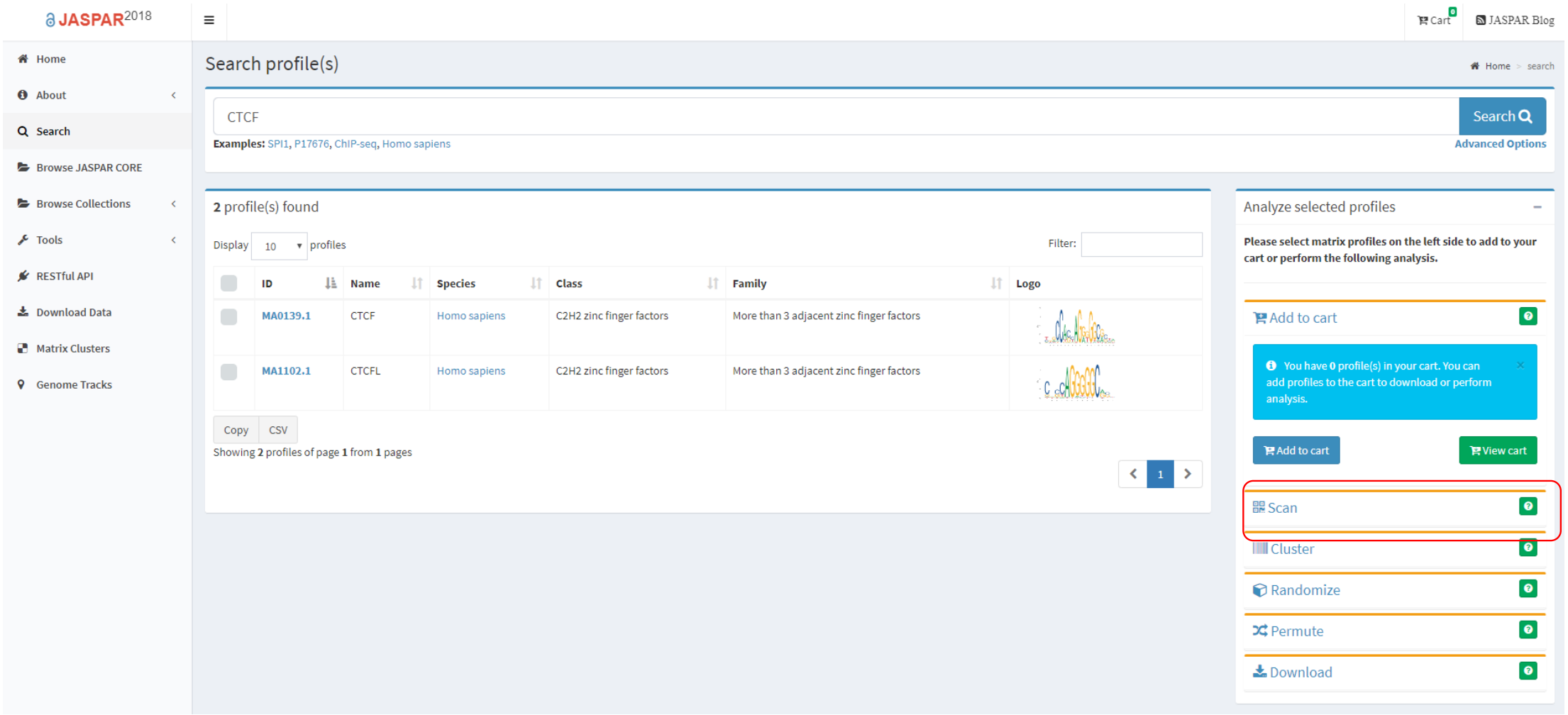


Provide Input *(MSA(Fasta and ClustalW), peptide, PSSM) *for more info click here.*

```
     A     C     G     T
1    3     2     1     4
2    6     2     1     1
3    1     1     7     1
4    0     0    10     0
5    0     0     0    10
6    6     2     1     1
7    7     1     1     1
8    2     1     5     2
9    1     2     1     6
```

Switch to file upload

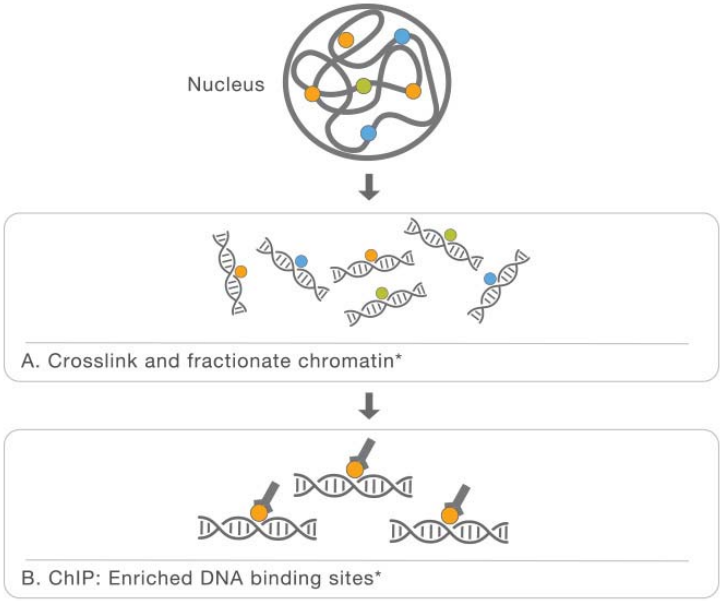
http://jaspar.genereg.net/



Experimental Methods to Identify TF Binding

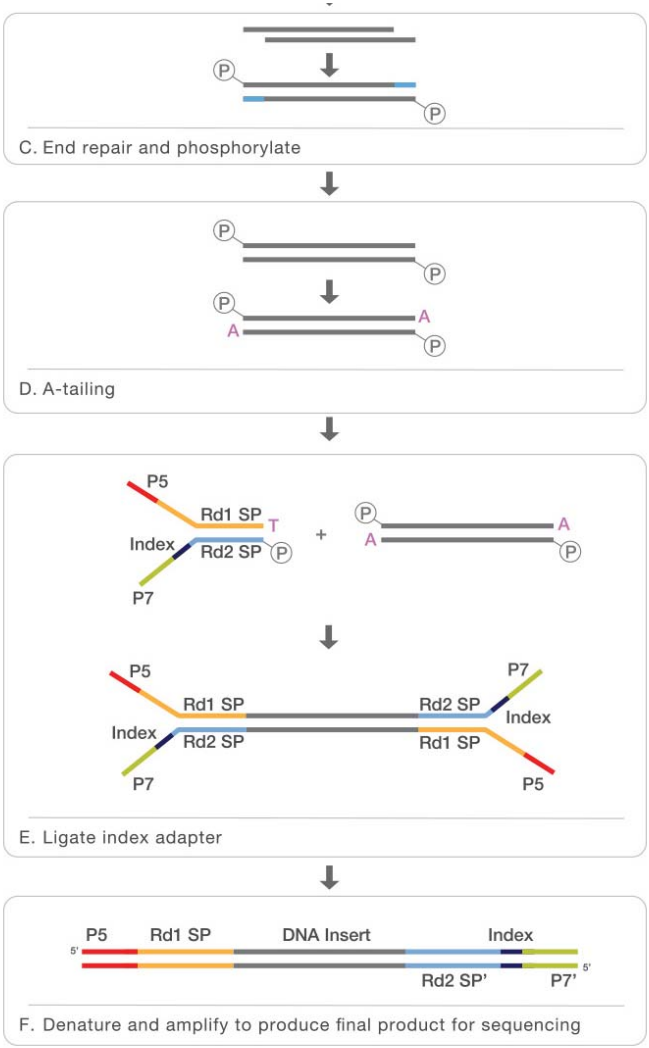
Chromatin Immunoprecipitation (ChIP)

- Cross-linking (1% formaldehyde)
- Shearing chromatin to small fragments (300-500 bp)
- Immunoprecipitation (Ab: a specific TF)
- Wash to remove non-specific binding
- De-crosslinking (heat)
- DNA extraction and purification

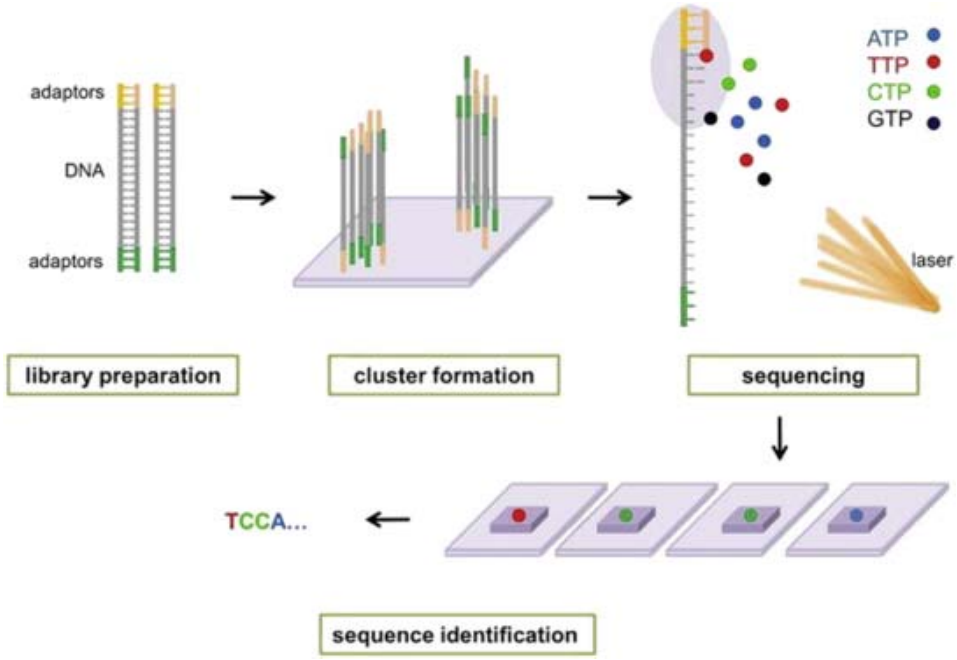


Sequencing Library

- End polishing (generate blunt end for adaptor)
- Adenine nucleotide addition (facilitate TA ligation for adaptor)
- Adaptor ligation (Y shape to generate two different ends for sequencing bridging)
- PCR amplification
- DNA purification and quantification (Qubit)



Sequencing by Synthesis



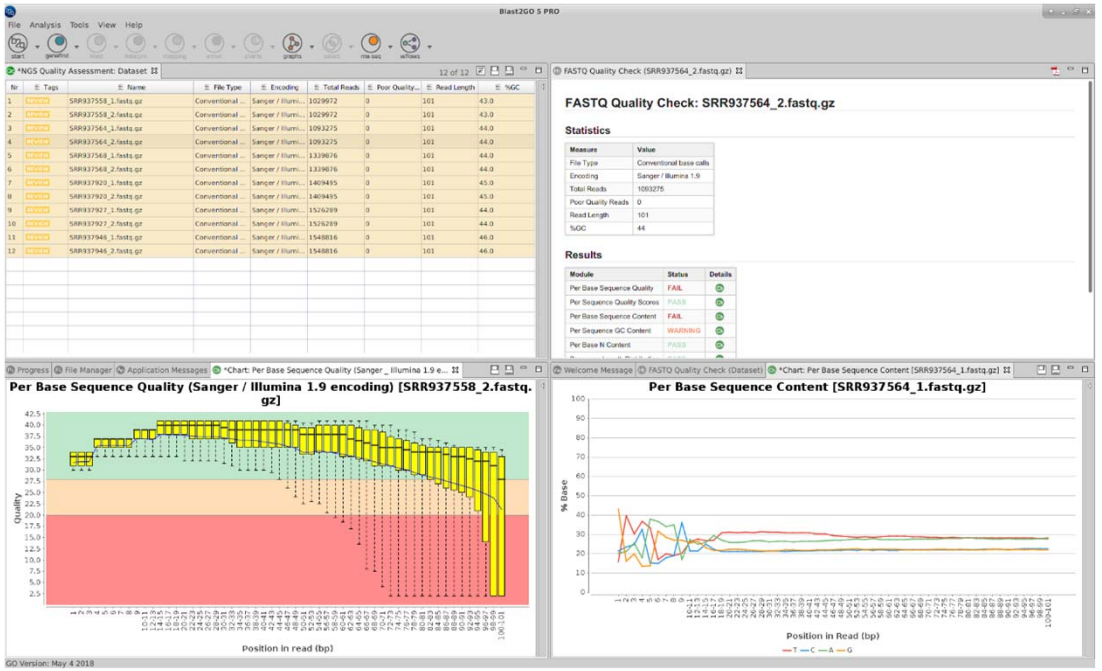
- https://www.youtube.com/watch?v=fCd6B5HRaZ8

Data Analysis

- FastQC: Preliminary Quality Control of Library
- Mapping: Identify the genomic location of each individual read
- Remove redundant reads
- Peak Calling
- Further Analysis
  - Motif Enrichment
  - Gene Ontology Analysis
  - Others

FastQC

- The "FASTQ Quality Check" tool provides an easy way to perform a quality control check on sequence data coming from high throughput sequencing pipelines. The analysis is performed by nine modules which provide a quick overview of whether the data looks good and there are no problems or biases which may affect downstream analysis. Results and evaluations are returned in the form of charts and tables.
- Low Number of Reads
  - Check your fastq file size
  - In general, 5 million reads are required
- Library Contamination
  - GC content
  - K-mer content (over representative)
  - High content of adaptor sequence
- High PCR duplicate



FastQC

FastQC Report

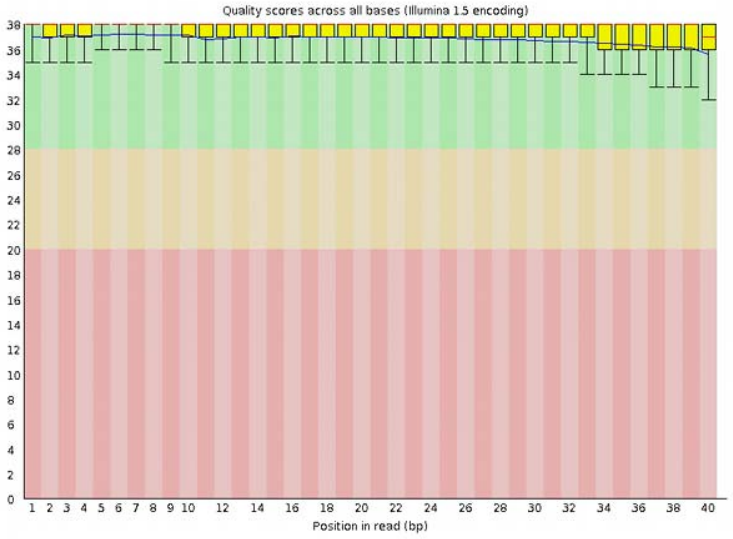
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

Per base sequence quality



Mapping tools

Read Mapping

	Novoalign (3.0)	SOAP3 (version 91)	BWA (0.7.4)	Bowtie2 (2.1.0)	Tophat2 (2.0.8b)	STAR (2.3.0e)
License	Commercial	GPL v3	GPL v3	Artistic	Artistic	GPL v3
Mismatch allowed	up to 8	up to 3	user specified. max is function of read length and error rate	user specified	uses Bowtie2	user specified
Alignments reported per read	random/all/none	random/all/none	user selected	user selected	uses Bowtie2	user selected
Gapped alignment	up to 7bp	1-3bp gap	yes	yes	yes splice junctions introns	yes splice junctions introns
Pair-end reads	yes	yes	yes	yes	yes	yes
Best alignment	highest alignment score	minimal number of mismatches	minimal number of mismatches	highest alignment score	uses Bowtie2	highest alignment score
Trim bases	3' end	3' end	3' and 5' end	3' and 5' end	uses Bowtie2	3' and 5' end
Comments	At one time, best performance and alignment quality		Element of Broad's "best practices" genotyping workflow	Smith-Waterman quality alignments, currently fastest	Currently most popular RNA-seq aligner	Very fast; uses memory to achieve performance

A note on using duplication levels to estimate your library size (complexity)

Assuming you have 100 initial fragments in your library (before amplification) & which fragment gets read is random:

# reads :	25	50	75	100	150	200
# unique reads:	23	37	52	63	78	87
% duplicated:	**8%**	**26%**	**31%**	**43%**	**55%**	**69%**
x-more left in lib:	4.3	2.7	1.9	1.6	1.3	1.15
x-more than prev:		1.6	1.4	1.2	1.24	1.11

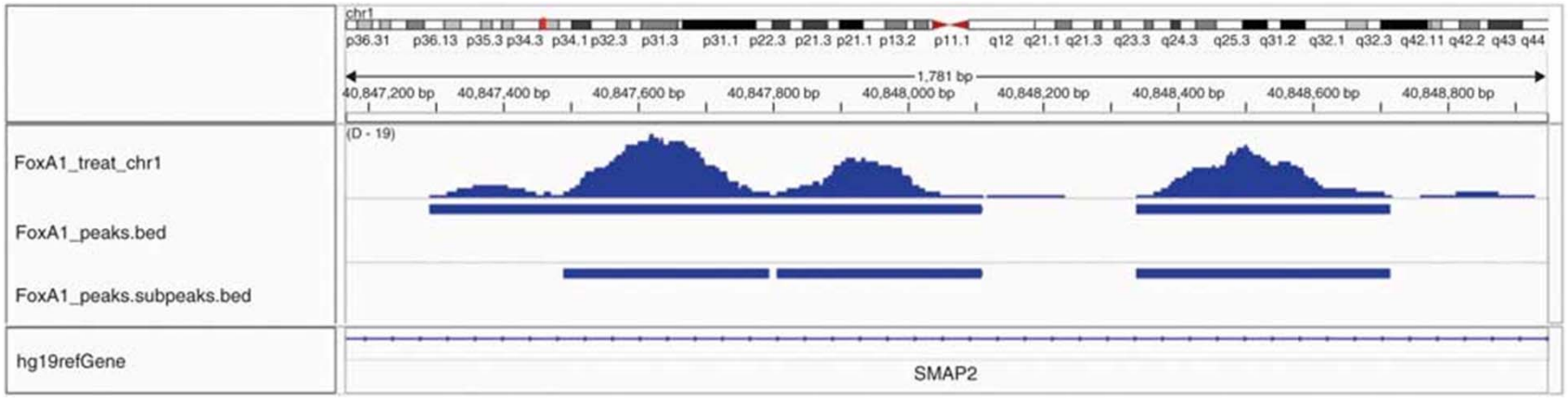
Given 9% duplicates, an additional sequencing run of the same size (from the same library) will give you 1.6x more unique reads. Two additional runs will give you 2.2x more (1.6*1.4).

…but if you have a high % duplicates (e.g. 43%) adding one more lane will only give you 1.37x more unique reads than you had initially. Depending on sequencing depth, this could indicate that your library has low complexity – either because too few fragments from your ChIP survived to the library amplification step, or because the protein binds few sites.

Alternatively: sub-sample your data and check saturation of peak calling

! There are some software (MACS, Picard, etc…) that can help you remove the duplicative reads. You can also simply remove them by deleting the identical sequence.

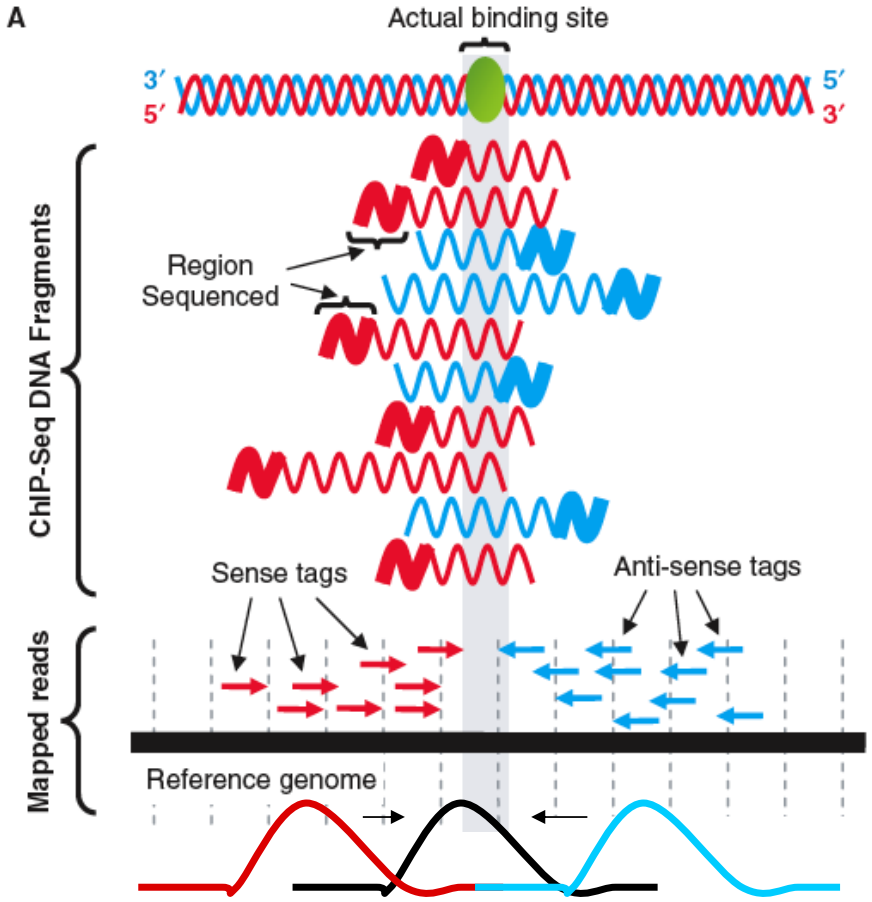
IGV visualization of MACS results



- The peak profile can be visualized using genome browsers
  - IGV (local, java)
  - UCSC Genome Browser

How do peak-finders map binding sites?

- Fragments contain the TF binding site at a (mostly) random position within them.
- Reads are (randomly) from left or right edges (sense or antisense) of fragments.
- Thus peak for sense tags will be 1/2 the fragment length upstream…
- Binding site position = mid-way between sense tag peak & antisense tag peak.
- To get binding site peak, shift sense downstream by ½ fragsize & antisense upstream by ½ fragsize.



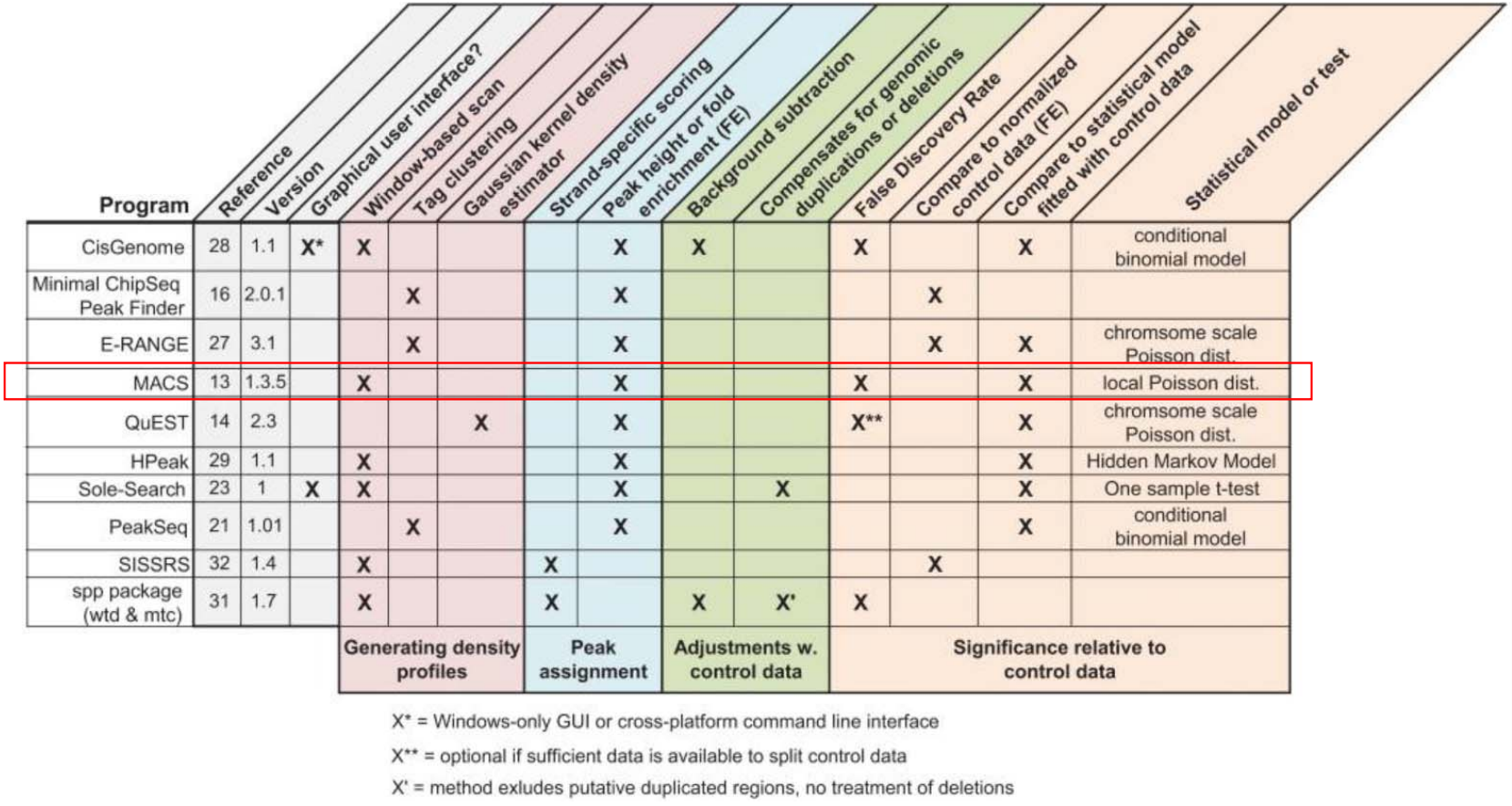
- Adapted from slide set by: Stuart M. Brown, Ph.D., Center for Health Informatics & Bioinformatics, NYU School of Medicine & from Jothi, et al. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. NAR (2008), 36: 5221-31

Peak Calling

- Identify TF binding sites:
  - Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads as a consequence of performing a ChIP-sequencing

- There are lots of softwares: Hpeak, PeakSeq, **MACS**, etc.

Comparison of Peak Callers



Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Wilbanks and colleagues 2010 Plos One

MACS procedure

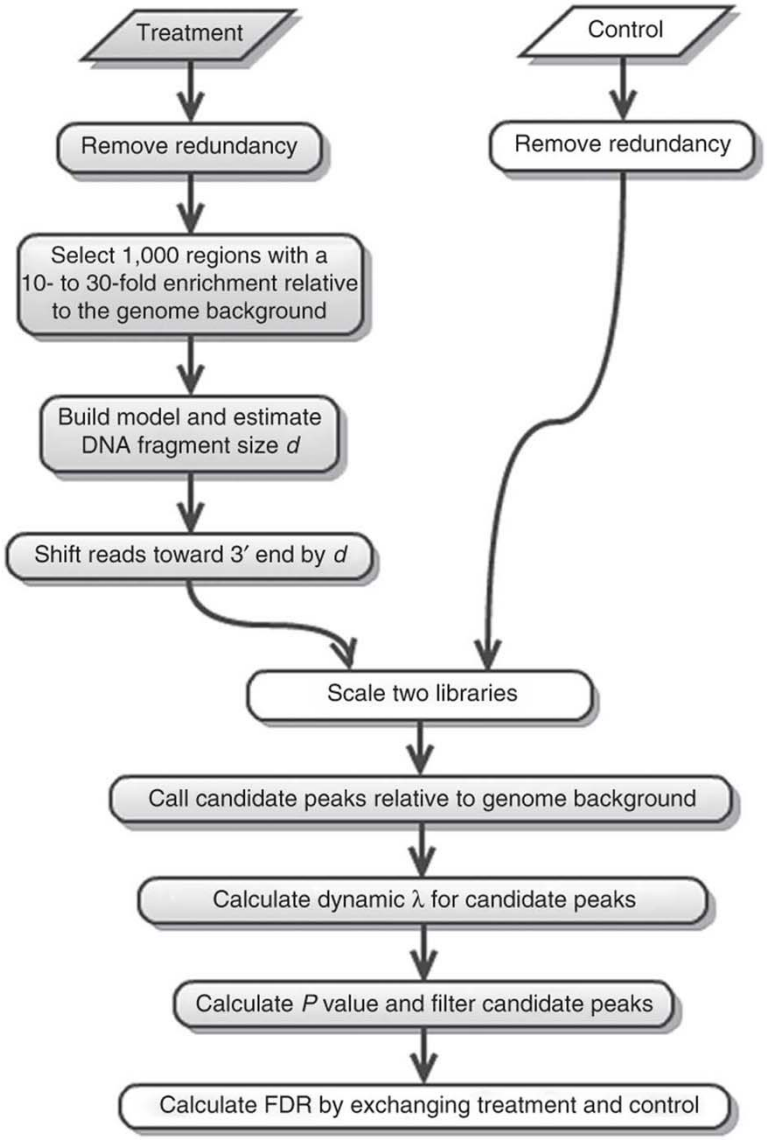
$$P_\lambda(X=k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

λ = mean = expectated value = variance

$$\lambda = \frac{\text{total number of events (k)}}{\text{number of units (n) in the data}}$$

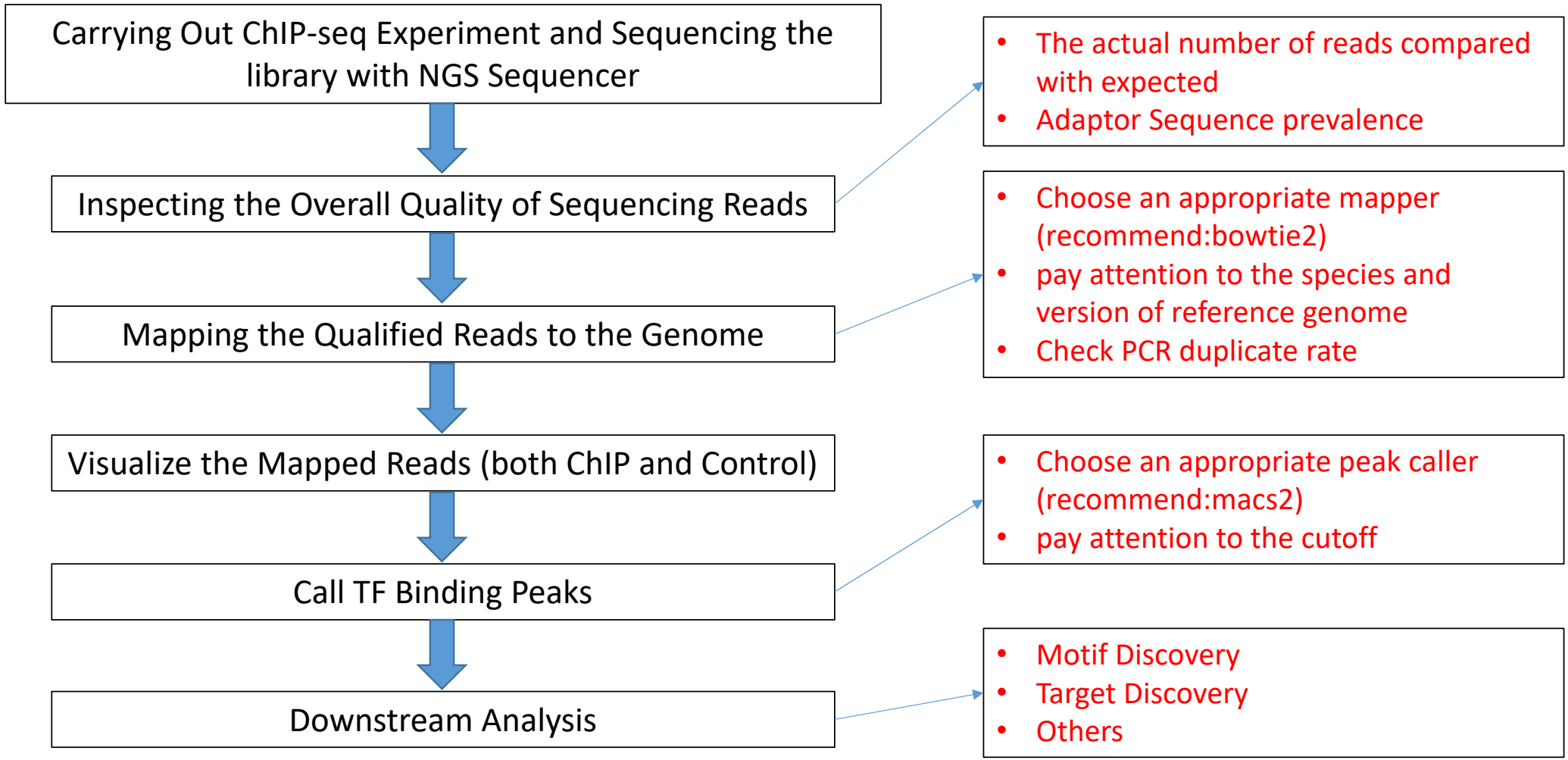
$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$

The tag distribution along the genome can be modeled by a **Poisson distribution**. The Poisson is a one parameter model, where the parameter **λ is the expected number of reads in that window**.



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3120977/

Summary of Analytic Pipeline of ChIP-seq (TF)



Case Study 1 (Galaxy)

- Most softwares are already installed for you
- Space is limited and also slow depends on utility
- Transcription Factor **Reb1** binding in **yeast cell**

Resource

Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution

Ho Sung Rhee[1] and B. Franklin Pugh[1,*]
[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA
*Correspondence: bfp2@psu.edu
DOI 10.1016/j.cell.2011.11.013

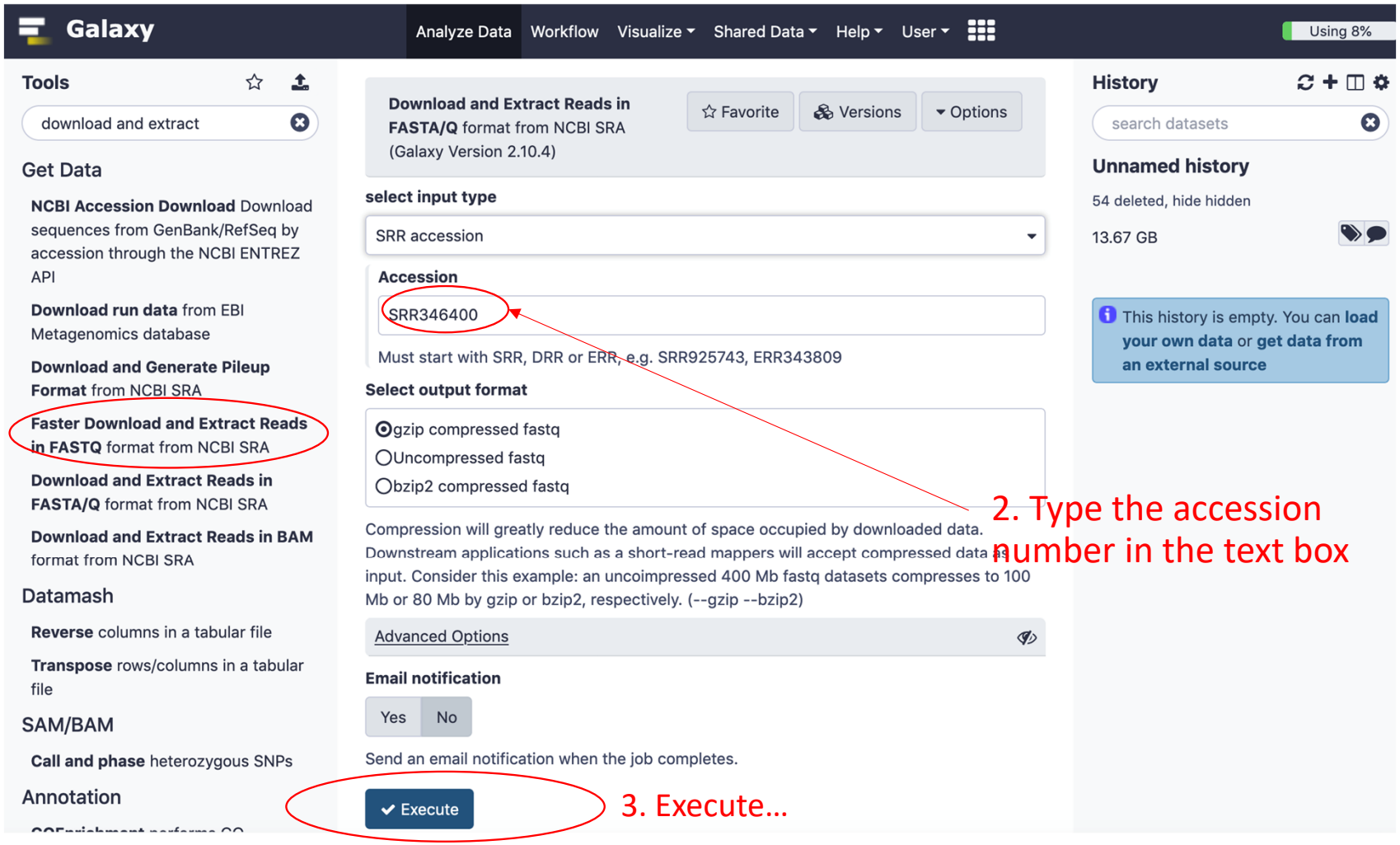
Dataset is stored at NCBI SRA: SRR346400

Runs: 1 run, 42.7M spots, 1.5G bases, 1.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR346400	42,681,641	1.5G	1.3Gb	2011-12-08

0. Obtaining Fastq

1. Find the software

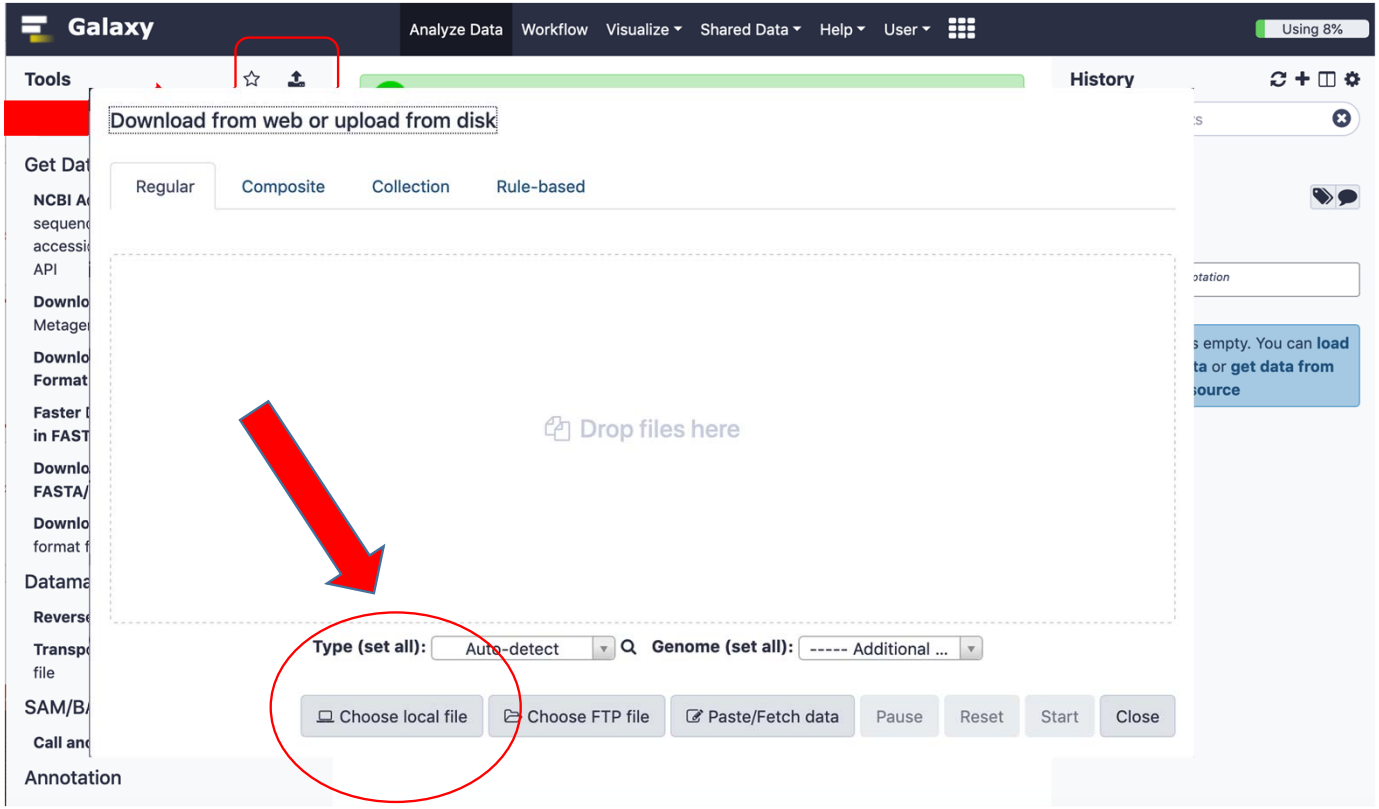


2. Type the accession number in the text box

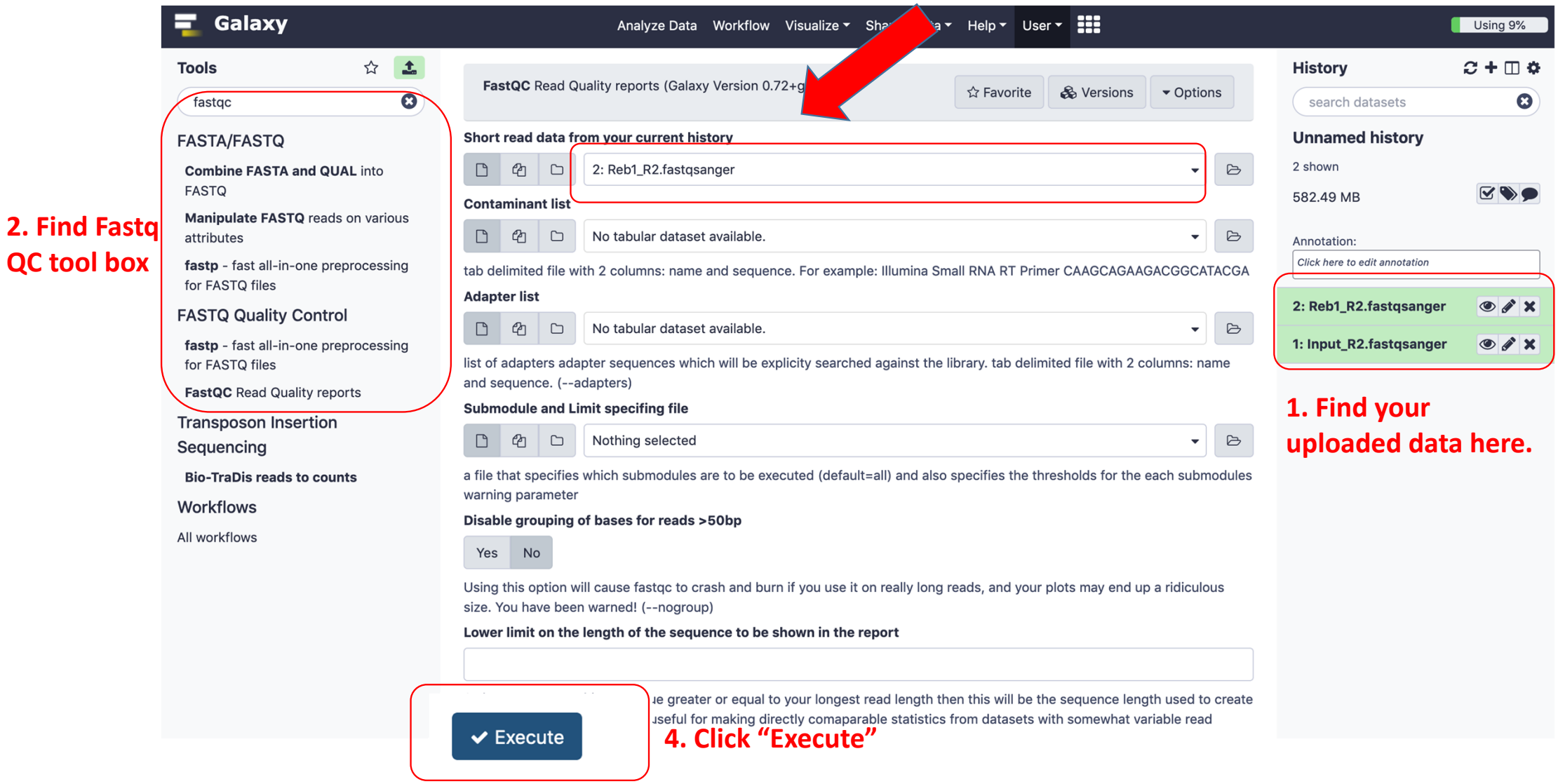
3. Execute…

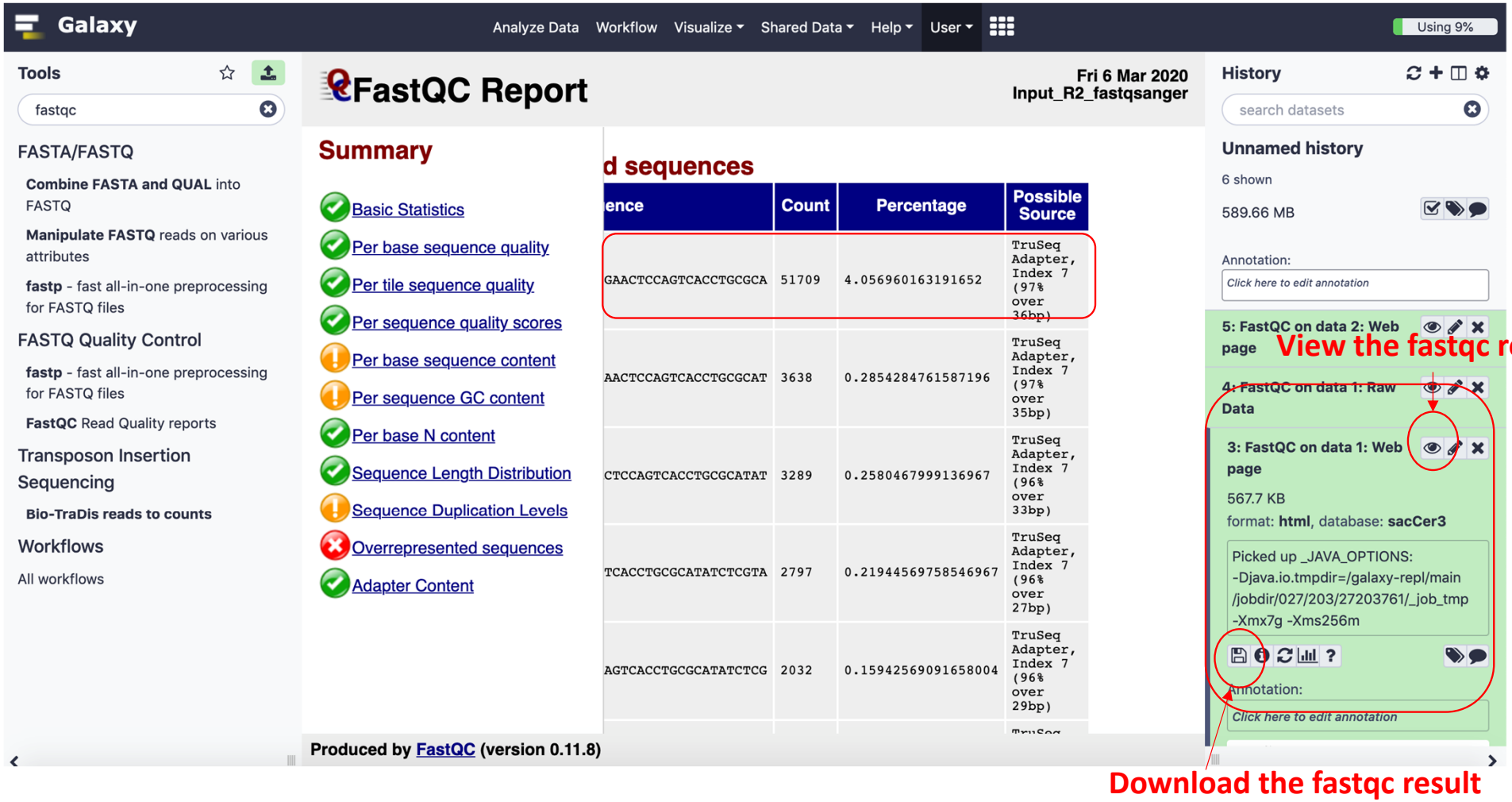
Or Uploading your own Fastq File

1. you can click this button to upload your own reads from local computer



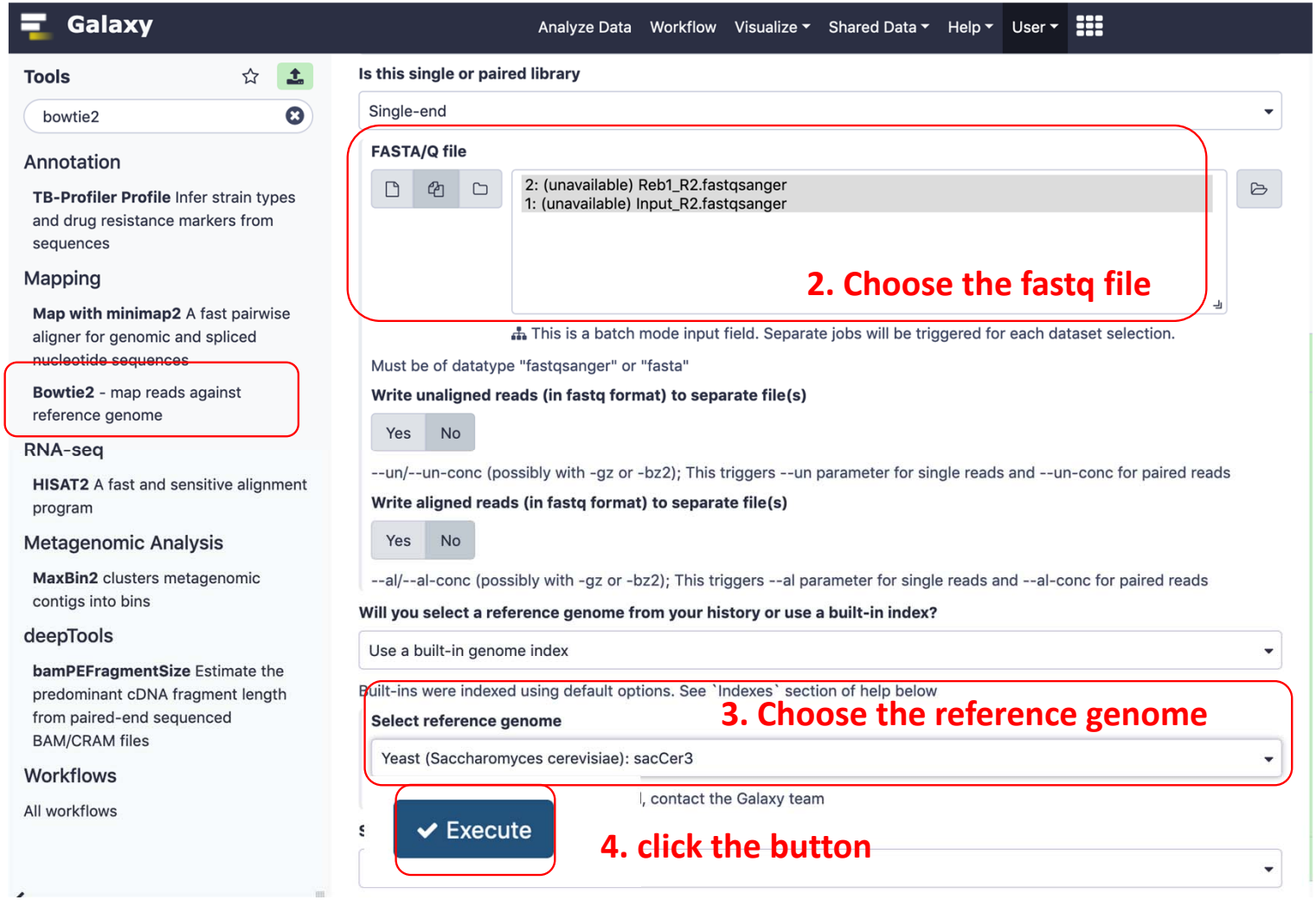
1. fastqc


3. Choose the dataset you want to analyze
Galaxy
Analyze Data
Workflow
Visualize
Help
User
Using 9%
Tools
fastqc
FASTA/FASTQ
Combine FASTA and QUAL into FASTQ
Manipulate FASTQ reads on various attributes
fastp - fast all-in-one preprocessing for FASTQ files
FASTQ Quality Control
fastp - fast all-in-one preprocessing for FASTQ files
FastQC Read Quality reports
Transposon Insertion Sequencing
Bio-TraDis reads to counts
Workflows
All workflows
2. Find Fastq QC tool box
FastQC Read Quality reports (Galaxy Version 0.72+g
Favorite
Versions
Options
Short read data from your current history
2: Reb1_R2.fastqsanger
Contaminant list
No tabular dataset available.
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA
Adapter list
No tabular dataset available.
list of adapters adapter sequences which will be explicity searched against the library. tab delimited file with 2 columns: name and sequence. (--adapters)
Submodule and Limit specifing file
Nothing selected
a file that specifies which submodules are to be executed (default=all) and also specifies the thresholds for the each submodules warning parameter
Disable grouping of bases for reads >50bp
Yes
No
Using this option will cause fastqc to crash and burn if you use it on really long reads, and your plots may end up a ridiculous size. You have been warned! (--nogroup)
Lower limit on the length of the sequence to be shown in the report
ue greater or equal to your longest read length then this will be the sequence length used to create
useful for making directly comaparable statistics from datasets with somewhat variable read
Execute
4. Click "Execute"
History
search datasets
Unnamed history
2 shown
582.49 MB
Annotation:
Click here to edit annotation
2: Reb1_R2.fastqsanger
1: Input_R2.fastqsanger
1. Find your uploaded data here.


Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 9%
Tools
fastqc
FASTA/FASTQ
Combine FASTA and QUAL into FASTQ
Manipulate FASTQ reads on various attributes
fastp - fast all-in-one preprocessing for FASTQ files
FASTQ Quality Control
fastp - fast all-in-one preprocessing for FASTQ files
FastQC Read Quality reports
Transposon Insertion Sequencing
Bio-TraDis reads to counts
Workflows
All workflows
FastQC Report
Fri 6 Mar 2020
Input_R2_fastqsanger
Summary
Basic Statistics
Per base sequence quality
Per tile sequence quality
Per sequence quality scores
Per base sequence content
Per sequence GC content
Per base N content
Sequence Length Distribution
Sequence Duplication Levels
Overrepresented sequences
Adapter Content
d sequences
ence
Count
Percentage
Possible Source
GAACTCCAGTCACCTGCGCA 51709 4.056960163191652 TruSeq Adapter, Index 7 (97% over 36bp)
AACTCCAGTCACCTGCGCAT 3638 0.2854284761587196 TruSeq Adapter, Index 7 (97% over 35bp)
CTCCAGTCACCTGCGCATAT 3289 0.2580467999136967 TruSeq Adapter, Index 7 (96% over 33bp)
TCACCTGCGCATATCTCGTA 2797 0.21944569758546967 TruSeq Adapter, Index 7 (96% over 27bp)
AGTCACCTGCGCATATCTCG 2032 0.15942569091658004 TruSeq Adapter, Index 7 (96% over 29bp)
Produced by FastQC (version 0.11.8)
History
search datasets
Unnamed history
6 shown
589.66 MB
Annotation:
Click here to edit annotation
5: FastQC on data 2: Web page
View the fastqc result
4: FastQC on data 1: Raw Data
3: FastQC on data 1: Web page
567.7 KB
format: html, database: sacCer3
Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/galaxy-repl/main/jobdir/027/203/27203761/_job_tmp -Xmx7g -Xms256m
Annotation:
Click here to edit annotation
Download the fastqc result

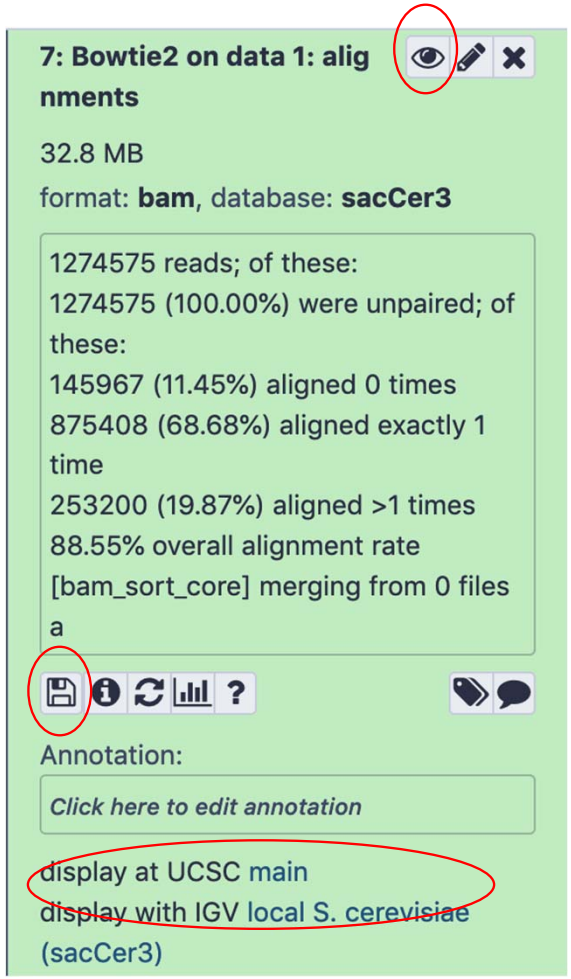
2. Mapping to the yeast genome

1. **Find Bowtie2**



The mapping takes a while for a large genome

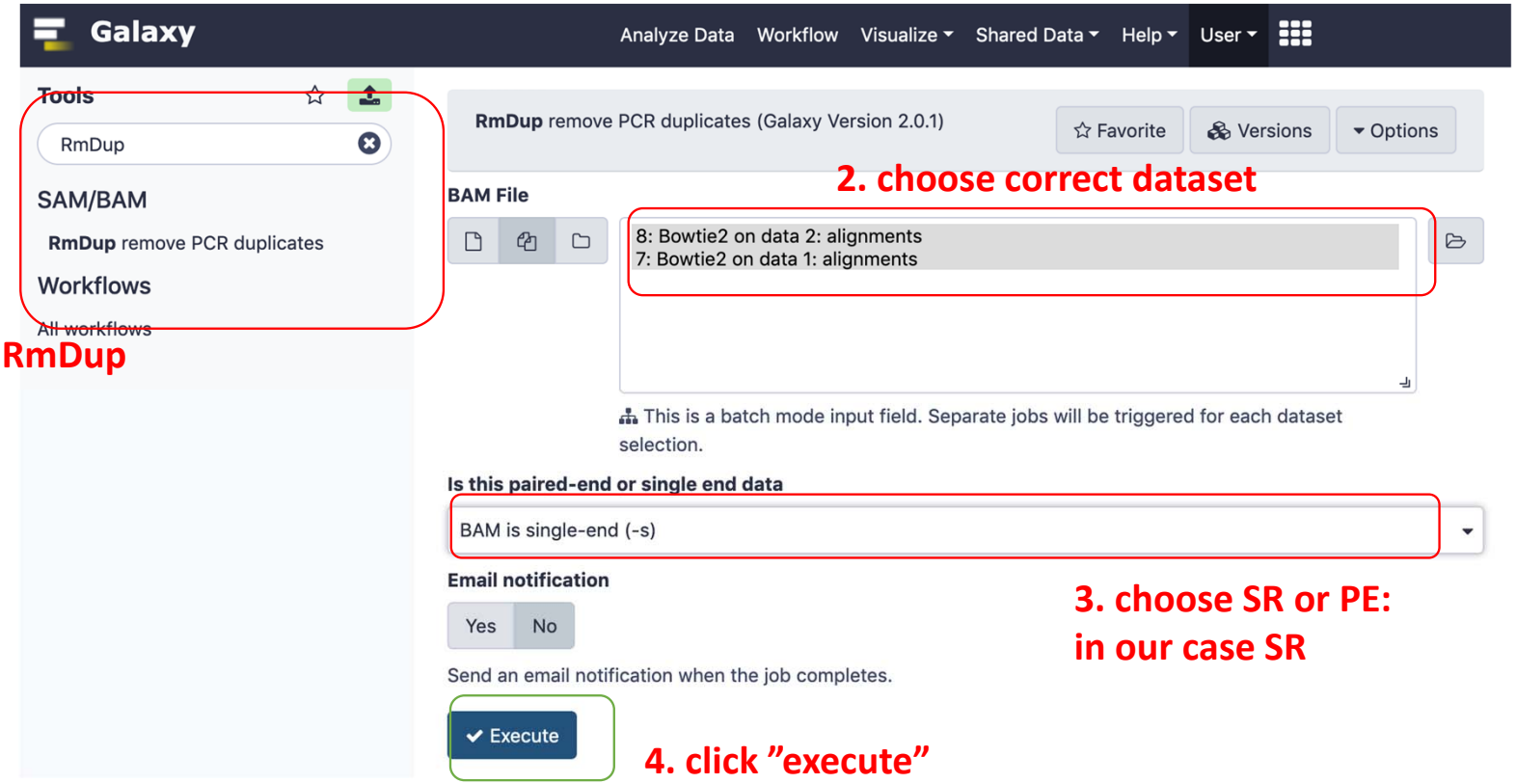
1. You can download the mapped reads



1. You can view the mapped reads in a table format

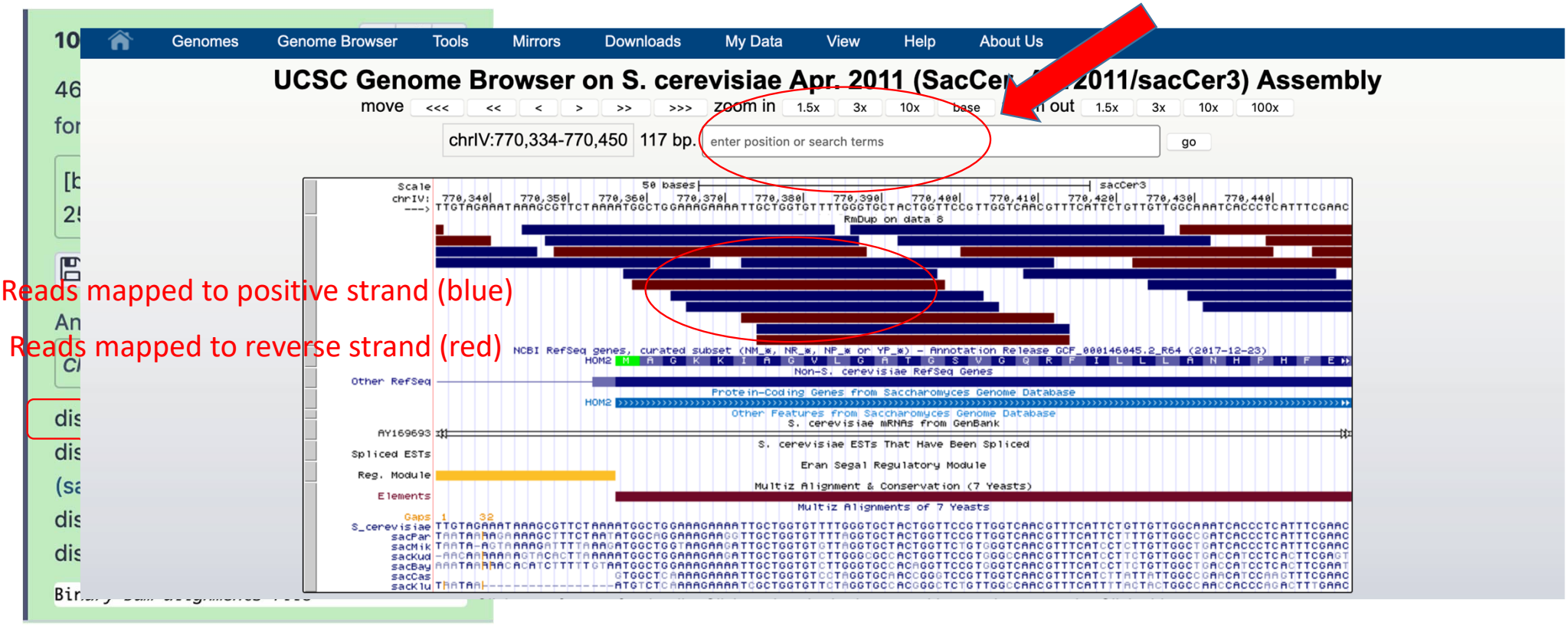
1. You can directly view the mapped reads in genome browser

Remove PCR duplicates



3. Visualize the mapping in Genome Browser

Pick up a genomic locus to view



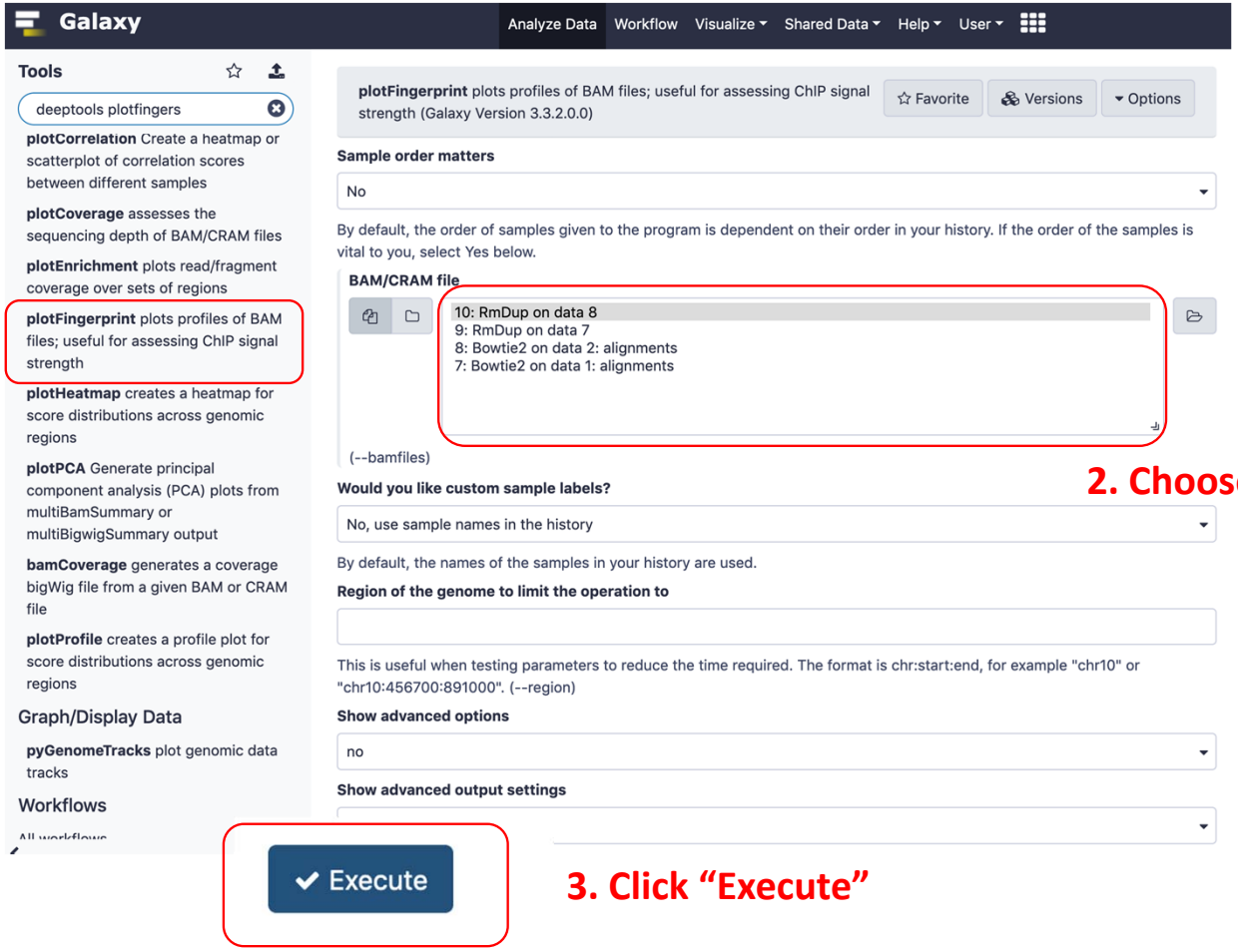
Reads mapped to positive strand (blue)

Reads mapped to reverse strand (red)

Preliminarily Check the Quality of ChIP (Signal Extraction Scaling)

- Suppose that we have two datasets (ChIP and control)
- We divide the genome into *N* non-overlapping windows and for each window we count the number of reads
- Then we sort the ChIP list in ascending order and move elements from the control list to match this order (converting the number of reads to accumulative percentage in its own sample)
- A successful experiment shows that a large portion of reads are concentrated in a few bins close to the bottom while a failed experiment shows linear ascending of the curve (even distribution of reads in bins)

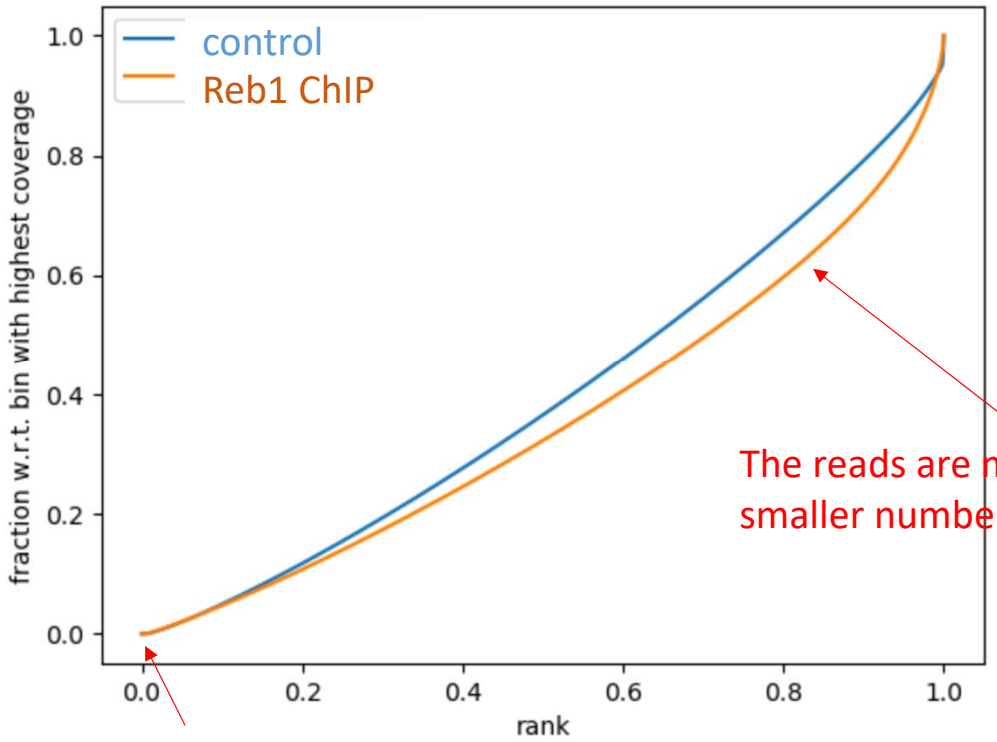
1. Find “plotFingerprint”



2. Choose the dataset

3. Click “Execute”

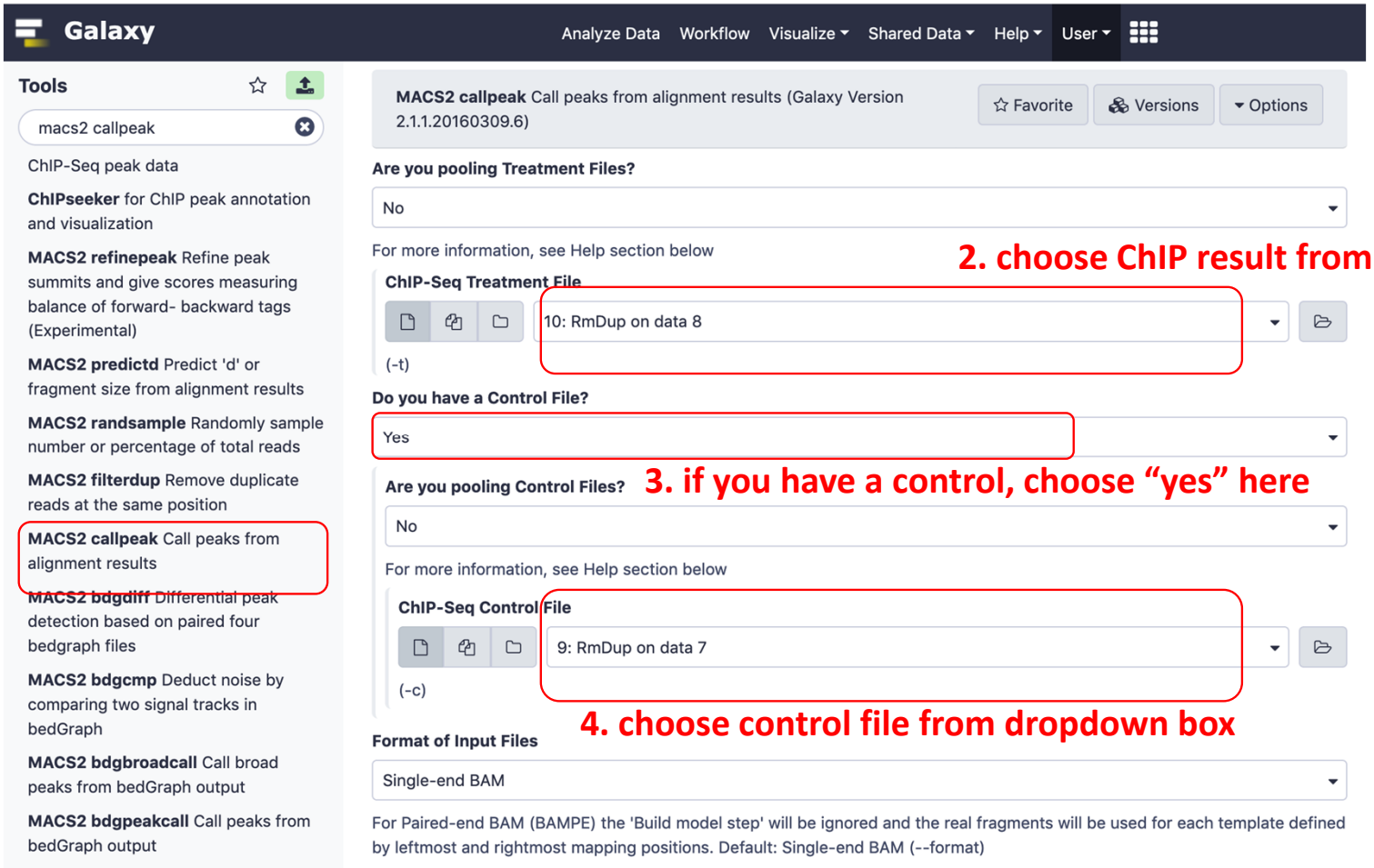
Signal Extraction Scaling Output



The reads are more concentrated in smaller number of bins

Also indicates how much of genome is not covered by the sample

4. peak calling (macs2)



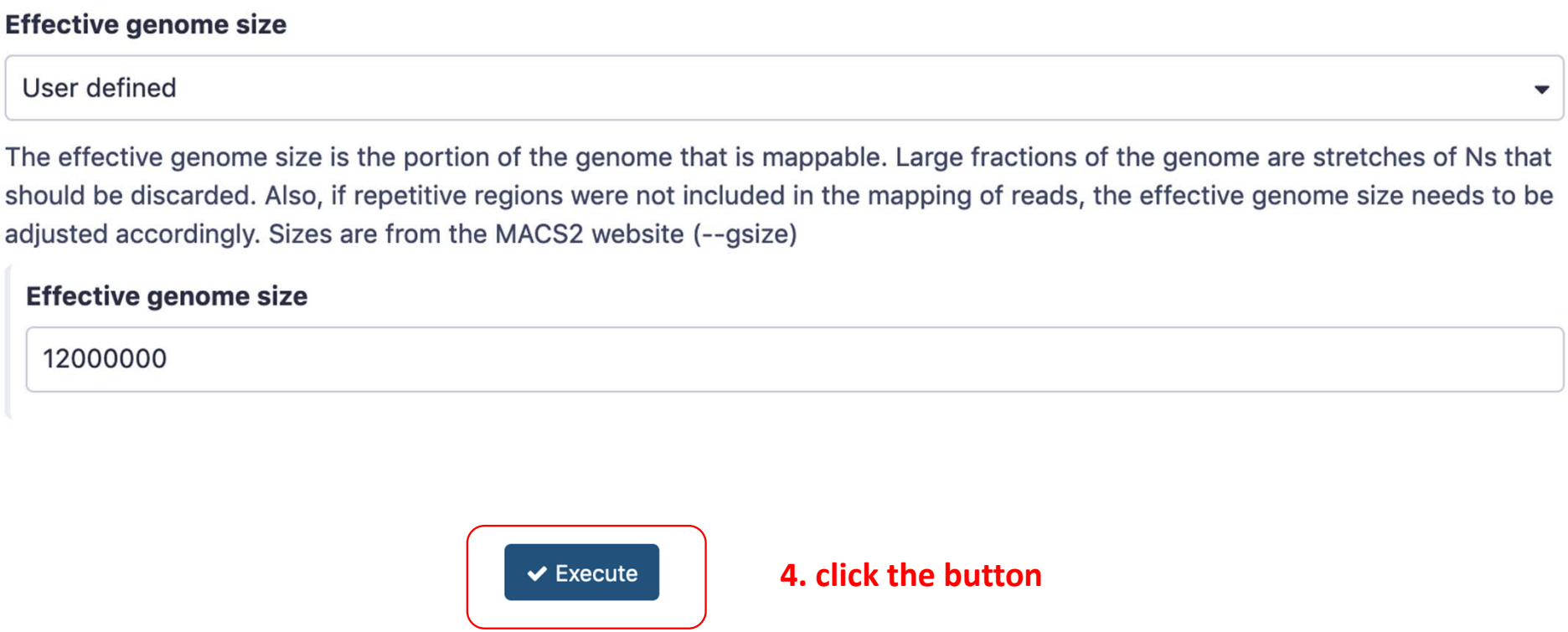
1. Find MACS2 callpeak

2. choose ChIP result from dropdown box

3. if you have a control, choose “yes” here

4. choose control file from dropdown box

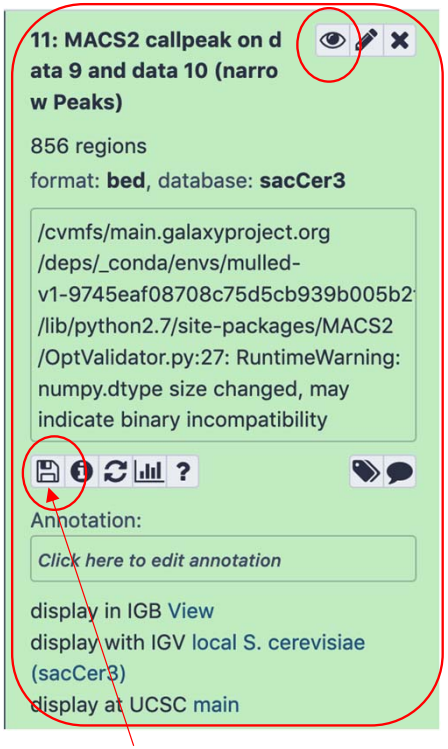
Choose Right Genome Size (Yeast, 12 Mb)



Output view

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB
chrI	14	692	RmDup_on_data_8_peak_1	480	.	11.14585	51.56381	48.0847
chrI	87010	87312	RmDup_on_data_8_peak_2	210	.	6.70586	23.59468	21.0976
chrI	92553	92832	RmDup_on_data_8_peak_3	355	.	9.27723	38.55616	35.5841
chrI	119704	119832	RmDup_on_data_8_peak_4	92	.	4.56829	11.38973	9.2325
chrI	141715	141852	RmDup_on_data_8_peak_5	113	.	4.48007	13.53213	11.3240
chrI	229660	229858	RmDup_on_data_8_peak_6	46	.	3.37237	6.66180	4.6799
chrI	230065	230218	RmDup_on_data_8_peak_7	195	.	6.58641	22.02638	19.5628
chrII	0	196	RmDup_on_data_8_peak_8	365	.	9.35197	39.60216	36.5981
chrII	5516	6459	RmDup_on_data_8_peak_9	328	.	8.08232	35.77129	32.8956
chrII	82372	82536	RmDup_on_data_8_peak_10	170	.	4.68286	19.44128	17.0542
chrII	84367	84512	RmDup_on_data_8_peak_11	87	.	4.30705	10.89353	8.7541
chrII	87932	88044	RmDup_on_data_8_peak_12	34	.	2.78039	5.34426	3.4373
chrII	110976	111477	RmDup_on_data_8_peak_13	407	.	10.09942	43.85774	40.7145
chrII	116888	117024	RmDup_on_data_8_peak_14	65	.	3.82084	8.58513	6.5271
chrII	122078	122333	RmDup_on_data_8_peak_15	170	.	6.13794	19.39235	17.0076
chrII	124846	125105	RmDup_on_data_8_peak_16	178	.	6.28743	20.25953	17.8482
chrII	132232	132377	RmDup_on_data_8_peak_17	58	.	3.67135	7.92574	5.8922
chrII	135724	136171	RmDup_on_data_8_peak_18	396	.	9.87519	42.78334	39.6751
chrII	151182	151560	RmDup_on_data_8_peak_19	386	.	9.72570	41.71587	38.6386
chrII	159863	159969	RmDup_on_data_8_peak_20	114	.	5.01676	13.65277	11.4323
chrII	164913	165024	RmDup_on_data_8_peak_21	52	.	3.59661	7.28438	5.2759
chrII	201706	201990	RmDup_on_data_8_peak_22	334	.	8.46607	36.31299	33.4266
chrII	202426	202554	RmDup_on_data_8_peak_23	53	.	3.51932	7.35632	5.3471
chrII	203186	203348	RmDup_on_data_8_peak_24	46	.	3.44712	6.66180	4.6799
chrII	256594	256684	RmDup_on_data_8_peak_25	21	.	2.19376	3.92038	2.1050
chrII	273787	273997	RmDup_on_data_8_peak_26	52	.	3.52186	7.28438	5.2759
x.psu.edu/repos/rnateam/chipseeker/chipseeker/1.18.0+galaxy1			8_peak_27	210	.	5.72349	23.53512	21.0389

View the peak result



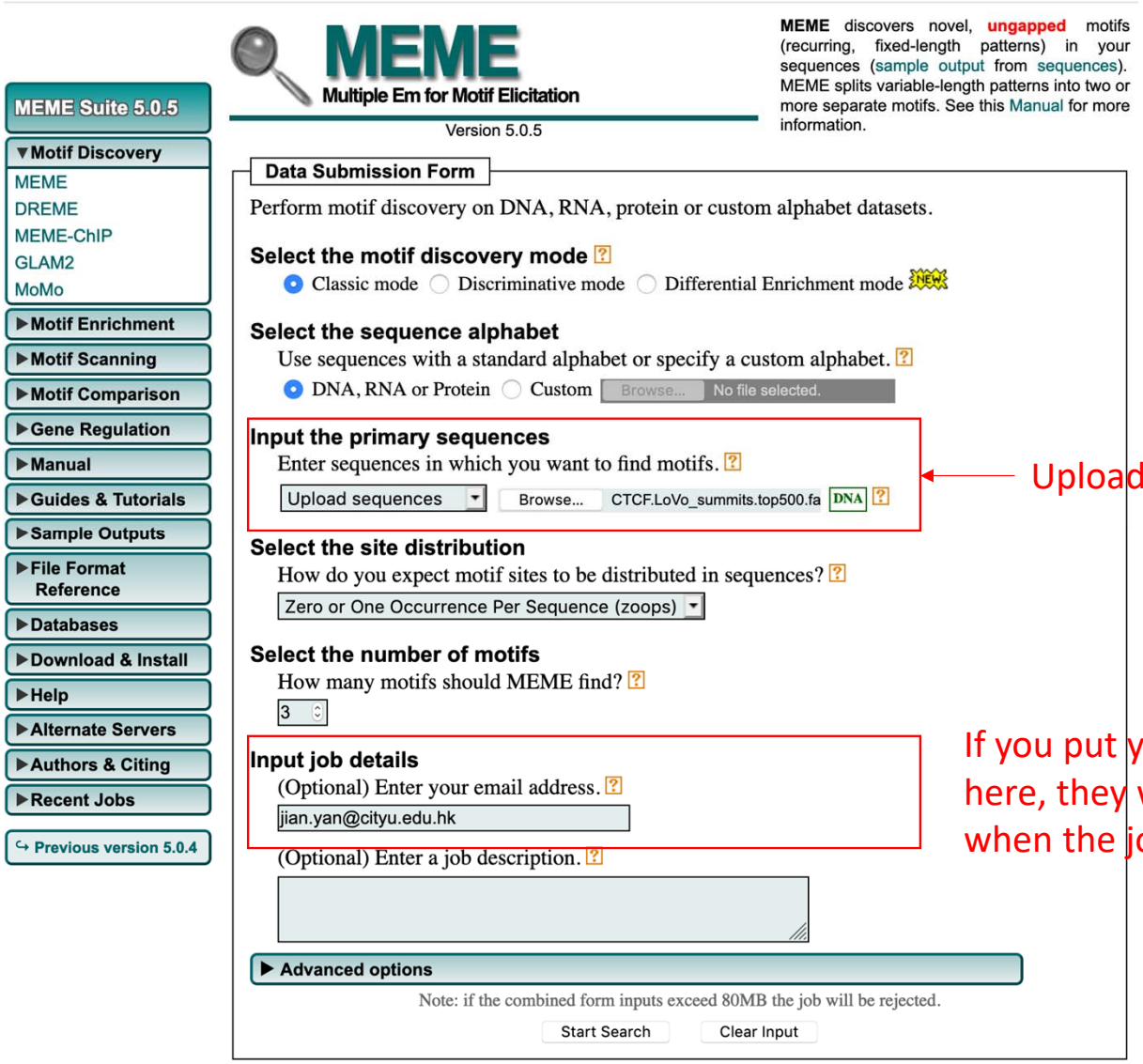
Download the peak result

Downstream Analysis

- TF DNA binding motif discovery (MEME)
- Gene Ontology
- Others….

MEME

- http://meme-suite.org/
- **MEME:** de novo motif discovery
- Note: The input sequences should be in fasta format



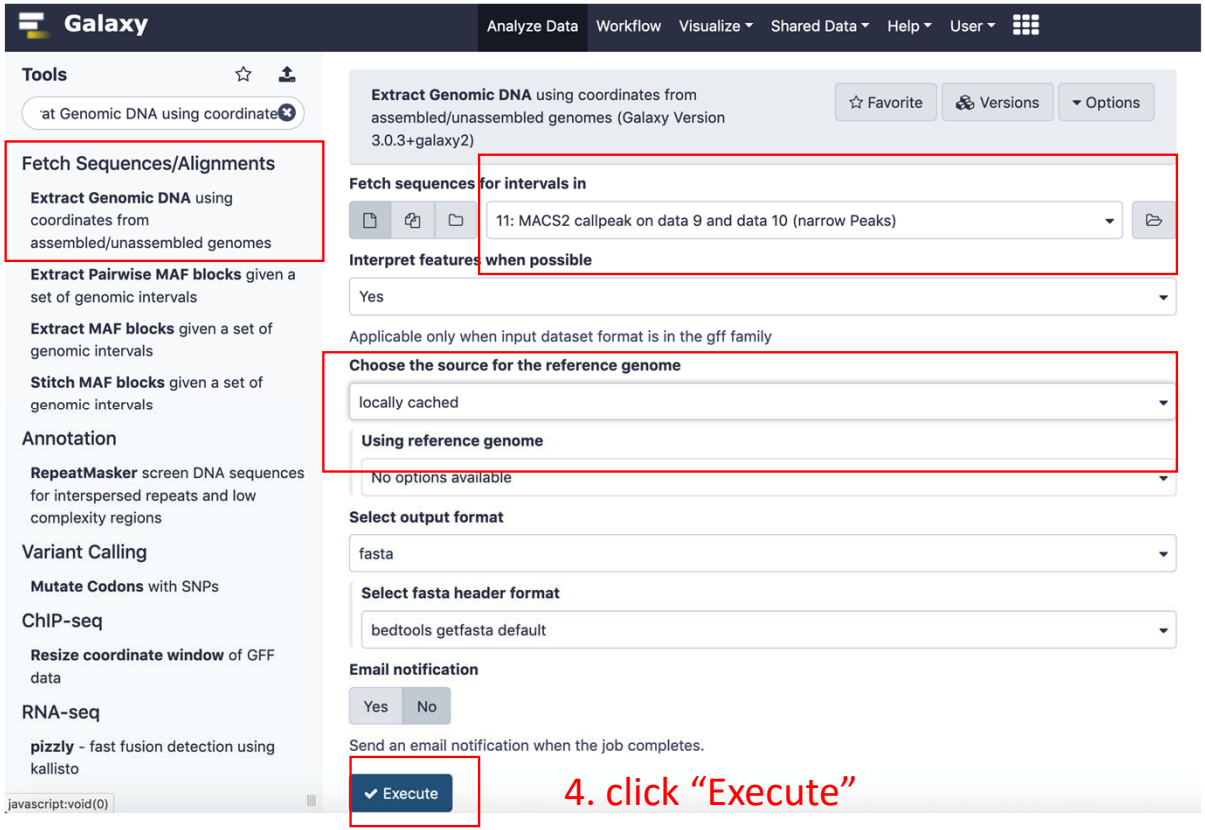
Upload your fasta file here

If you put your email address here, they will inform you when the job is finished.

MEME in Galaxy

- First we need to extract DNA sequences around the peak regions

1. Find appropriate Tools. Here, we choose “Fetch Sequences/Alignments”

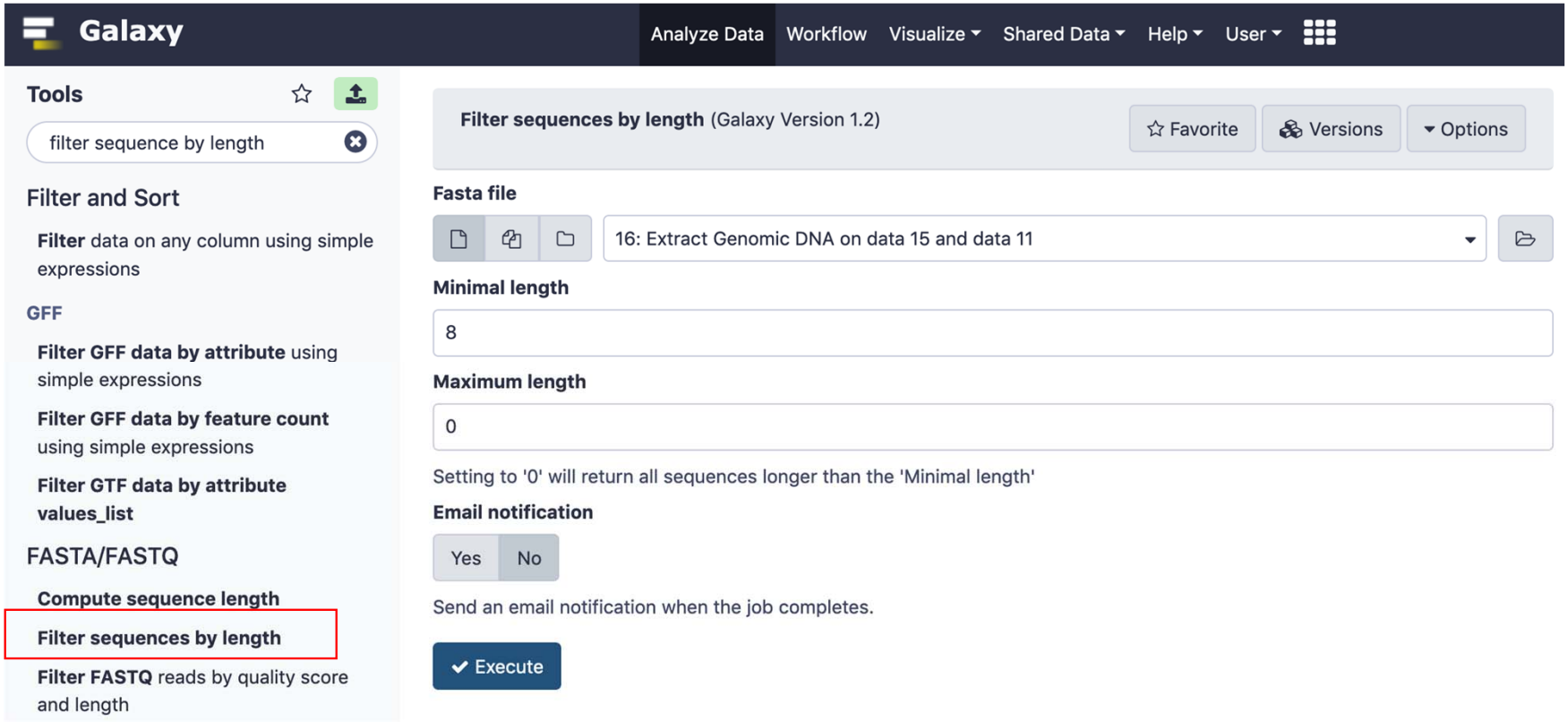


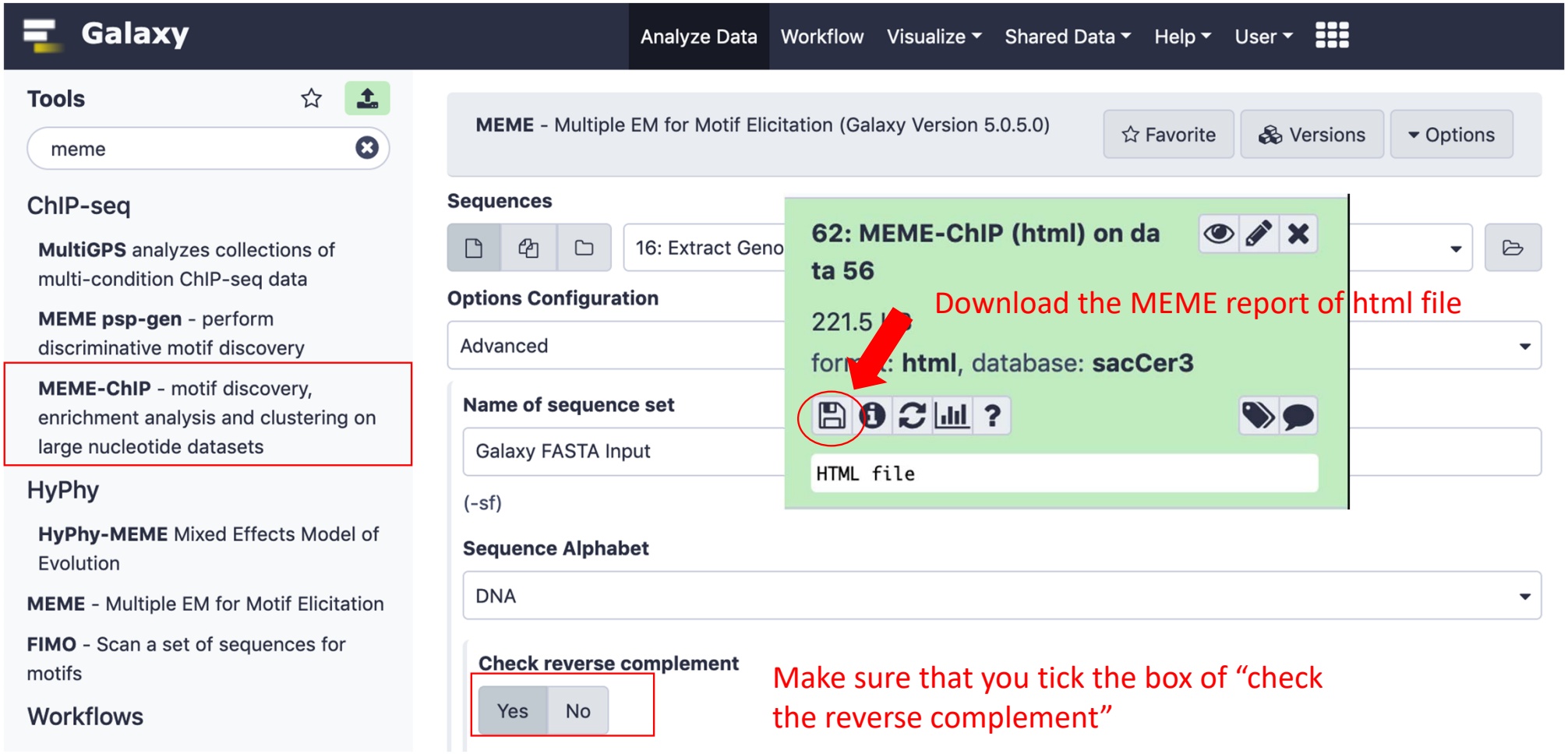
2. Choose the appropriate Peak file

3. You may need to upload your own genomic fasta file

4. click “Execute”

Remove the too short reads (>8bp)




Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Tools
meme
ChIP-seq
MultiGPS analyzes collections of multi-condition ChIP-seq data
MEME psp-gen - perform discriminative motif discovery
MEME-ChIP - motif discovery, enrichment analysis and clustering on large nucleotide datasets
HyPhy
HyPhy-MEME Mixed Effects Model of Evolution
MEME - Multiple EM for Motif Elicitation
FIMO - Scan a set of sequences for motifs
Workflows
MEME - Multiple EM for Motif Elicitation (Galaxy Version 5.0.5.0)
Favorite
Versions
Options
Sequences
16: Extract Geno
62: MEME-ChIP (html) on da
ta 56
221.5
form
: html, database: sacCer3
Download the MEME report of html file
HTML file
Options Configuration
Advanced
Name of sequence set
Galaxy FASTA Input
(-sf)
Sequence Alphabet
DNA
Check reverse complement
Yes
No
Make sure that you tick the box of “check the reverse complement”

MEME

Multiple Em for Motif Elicitation

For further information on how to interpret these results or to get a copy of the MEME software please access http://meme-suite.org.

If you use MEME in your research, please cite the following paper:
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994. [pdf]

DISCOVERED MOTIFS | MOTIF LOCATIONS | INPUTS & SETTINGS | PROGRAM INFORMATION

DISCOVERED MOTIFS

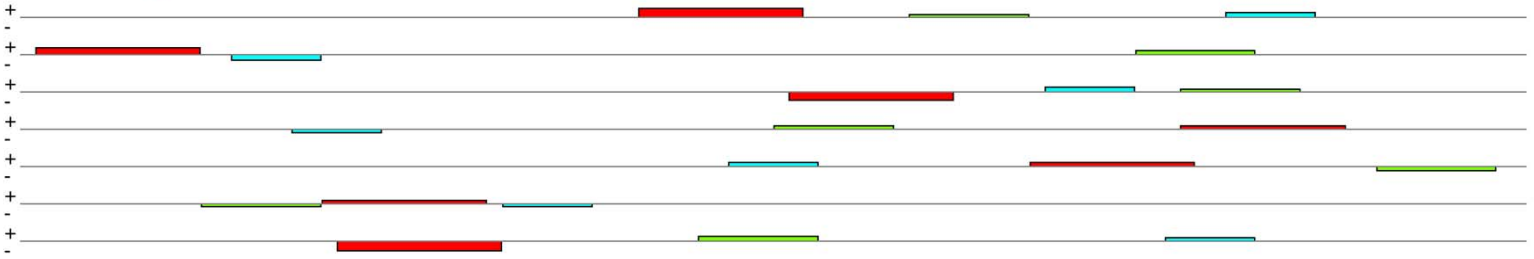
	Logo	E-value	Sites	Width	More	Submit/Download
1.		1.5e-535	600	11		
2.		2.1e+213	600	6		
3.		5.8e+345	600	8		

Stopped because requested number of motifs (3) found.

MOTIF LOCATIONS

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences

	Name	*p*-value	Motif Location
1.	chrII:336784-337389(+)	2.14e-3	
2.	chrVIII:381326-381592(+)	1.09e-2	
3.	chrXV:143871-144259(+)	1.01e-2	
4.	chrVI:270034-270146(+)	8.68e-1	
5.	chrXI:581605-581991(+)	3.52e-1	
6.	chrI:230077-230211(+)	9.58e-1	
7.	chrIV:1089997-1090571(+)	1.79e-3	



Reference: https://galaxyproject.org/tutorials/chip/

Case Study 2 (Command Line) --- Optional

Cell

Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites

Jian Yan,[1,4] Martin Enge,[1,4] Thomas Whitington,[1] Kashyap Dave,[1] Jianping Liu,[1] Inderpreet Sur,[1] Bernhard Schmierer,[1] Arttu Jolma,[1] Teemu Kivioja,[1,2,3] Minna Taipale,[1,3] and Jussi Taipale[1,3,*]
[1]Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm 14183, Sweden
[2]Department of Computer Science, P.O. Box 68
[3]Genome-Scale Biology Program, P.O. Box 63
00014 University of Helsinki, Helsinki, Finland
[4]These authors contributed equally to this work
*Correspondence: jussi.taipale@ki.se
http://dx.doi.org/10.1016/j.cell.2013.07.034

- Investigate the genome-wide binding of transcription factors in colorectal cancer cells.
- Over 100 TF ChIP-seq dataset
- Also include ChIP-seq for H3K4me1, RNA polymerase, etc…
- We focus on an important transcription factor called **CTCF**

Aim of the analysis

- Find the significant binding sites of CTCF in the human genome
- De novo discovery the CTCF binding motif
- Gene Ontology Analysis of CTCF Binding

0. Data access

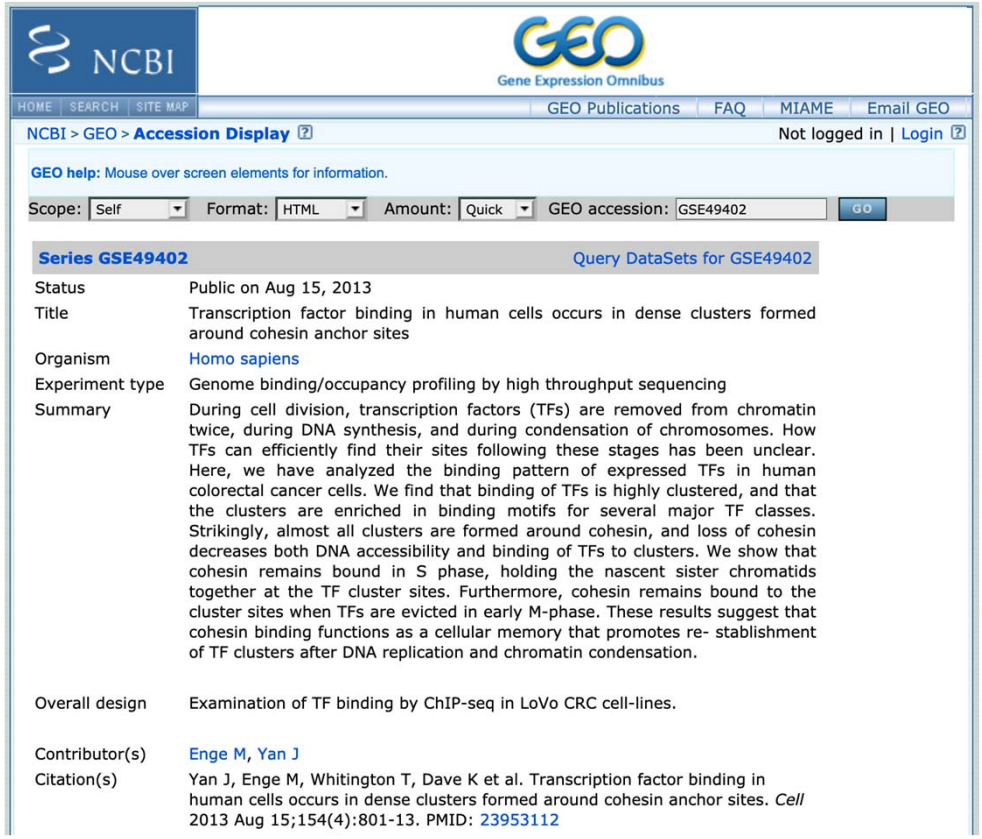
ACCESSION NUMBERS

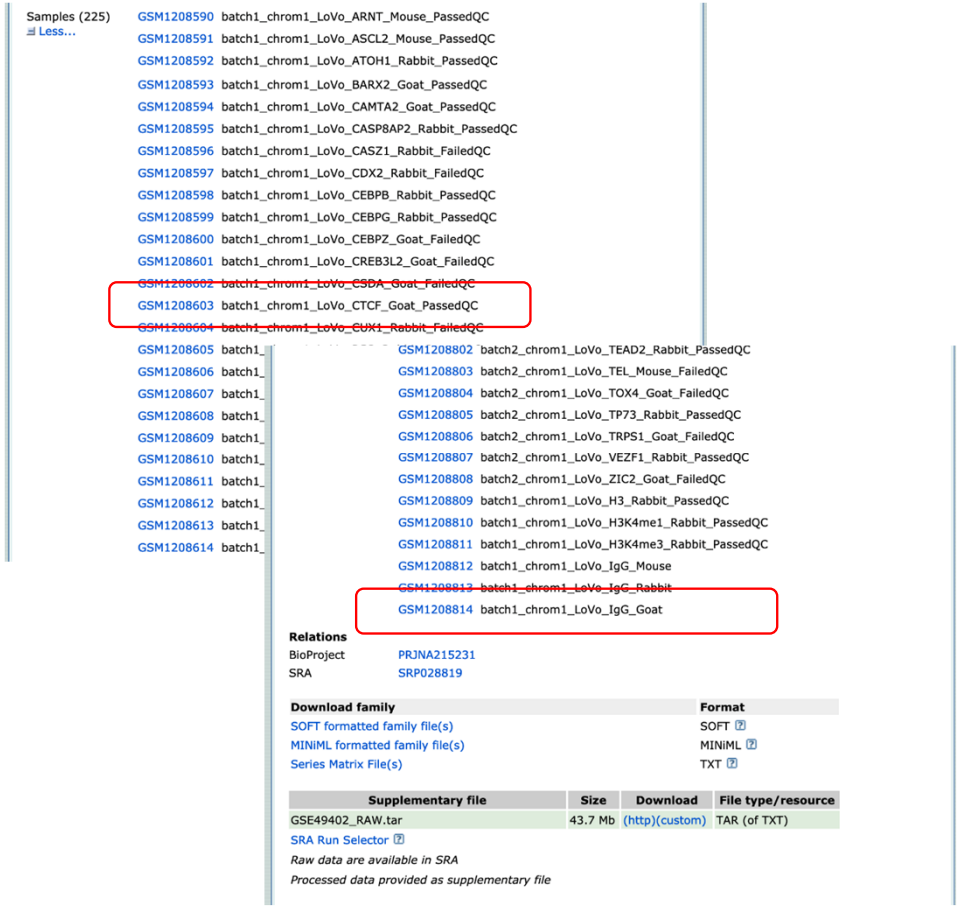
European Nucleotide Archive accession number for genomic sequencing data is ERP002229. Gene Expression Omnibus accession number is GSE48448 for the microarray data and GSE49402 for the ChIP-seq data.

- ChIP-seq data from this study have been deposited in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession no GSE49402
- Gene Expression Omnibus (GEO) is one of the largest public database to store the NGS data
- Find your interesting dataset and download them with SRA tool kit to your computer or the server where you plan to analyze them
  - SRR952398, ChIP-seq CTCF in LoVo cells, 36-bp single read sequencing
  - SRR952609, ChIP-seq IgG control in LoVo cells, 36-bp SR sequencing

```
[@local]$ ./sratoolkit.2.10.0-centos_linux64/bin/fastq-dump-orig.2.10.0 -Z SRR952398 > CTCF.SRR952398.fastq
[@local]$ ./sratoolkit.2.10.0-centos_linux64/bin/fastq-dump-orig.2.10.0 -Z SRR952609 > Control.SRR952609.fastq
```

Find CTCF and IgG control ChIP-seq

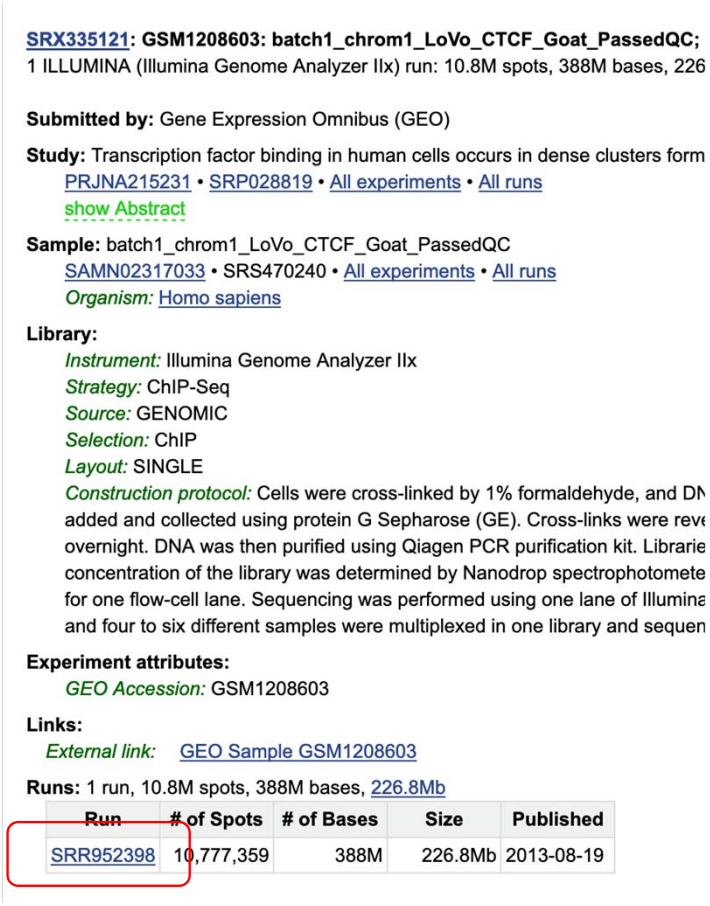




https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49402

Through the link to its SRA (ID no. of sequencing file)

https://www.ncbi.nlm.nih.gov/sra?term=SRX335121



SRX335121: GSM1208603: batch1_chrom1_LoVo_CTCF_Goat_PassedQC;
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 10.8M spots, 388M bases, 226

Submitted by: Gene Expression Omnibus (GEO)

Study: Transcription factor binding in human cells occurs in dense clusters form
PRJNA215231 • SRP028819 • All experiments • All runs
show Abstract

Sample: batch1_chrom1_LoVo_CTCF_Goat_PassedQC
SAMN02317033 • SRS470240 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina Genome Analyzer IIx
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: SINGLE
Construction protocol: Cells were cross-linked by 1% formaldehyde, and DN
added and collected using protein G Sepharose (GE). Cross-links were reve
overnight. DNA was then purified using Qiagen PCR purification kit. Librarie
concentration of the library was determined by Nanodrop spectrophotomete
for one flow-cell lane. Sequencing was performed using one lane of Illumina
and four to six different samples were multiplexed in one library and sequen

Experiment attributes:
GEO Accession: GSM1208603

Links:
External link: GEO Sample GSM1208603

Runs: 1 run, 10.8M spots, 388M bases, 226.8Mb

Run	# of Spots	# of Bases	Size	Published
SRR952398	10,777,359	388M	226.8Mb	2013-08-19

https://www.ncbi.nlm.nih.gov/sra?term=SRX335332

SRX335332: GSM1208814: batch1_chrom1_LoVo_IgG_Goat; Homo sapiens; ChIP-Seq
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 69M spots, 2.5G bases, 1.3Gb downloads

Submitted by: Gene Expression Omnibus (GEO)

Study: Transcription factor binding in human cells occurs in dense clusters formed around cohes
PRJNA215231 • SRP028819 • All experiments • All runs
show Abstract

Sample: batch1_chrom1_LoVo_IgG_Goat
SAMN02316868 • SRS470451 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina Genome Analyzer IIx
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: SINGLE
Construction protocol: Cells were cross-linked by 1% formaldehyde, and DNA was sonicatec
added and collected using protein G Sepharose (GE). Cross-links were reversed and protein
overnight. DNA was then purified using Qiagen PCR purification kit. Libraries were construct
concentration of the library was determined by Nanodrop spectrophotometer (Thermo Fisher
for one flow-cell lane. Sequencing was performed using one lane of Illumina GAIIx, or alterna
and four to six different samples were multiplexed in one library and sequenced using Illumin

Experiment attributes:
GEO Accession: GSM1208814

Links:
External link: GEO Sample GSM1208814

Runs: 1 run, 69M spots, 2.5G bases, 1.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR952609	68,973,382	2.5G	1.3Gb	2013-08-19

Check your file first

- First, check the number of reads in each sample (wc –l)

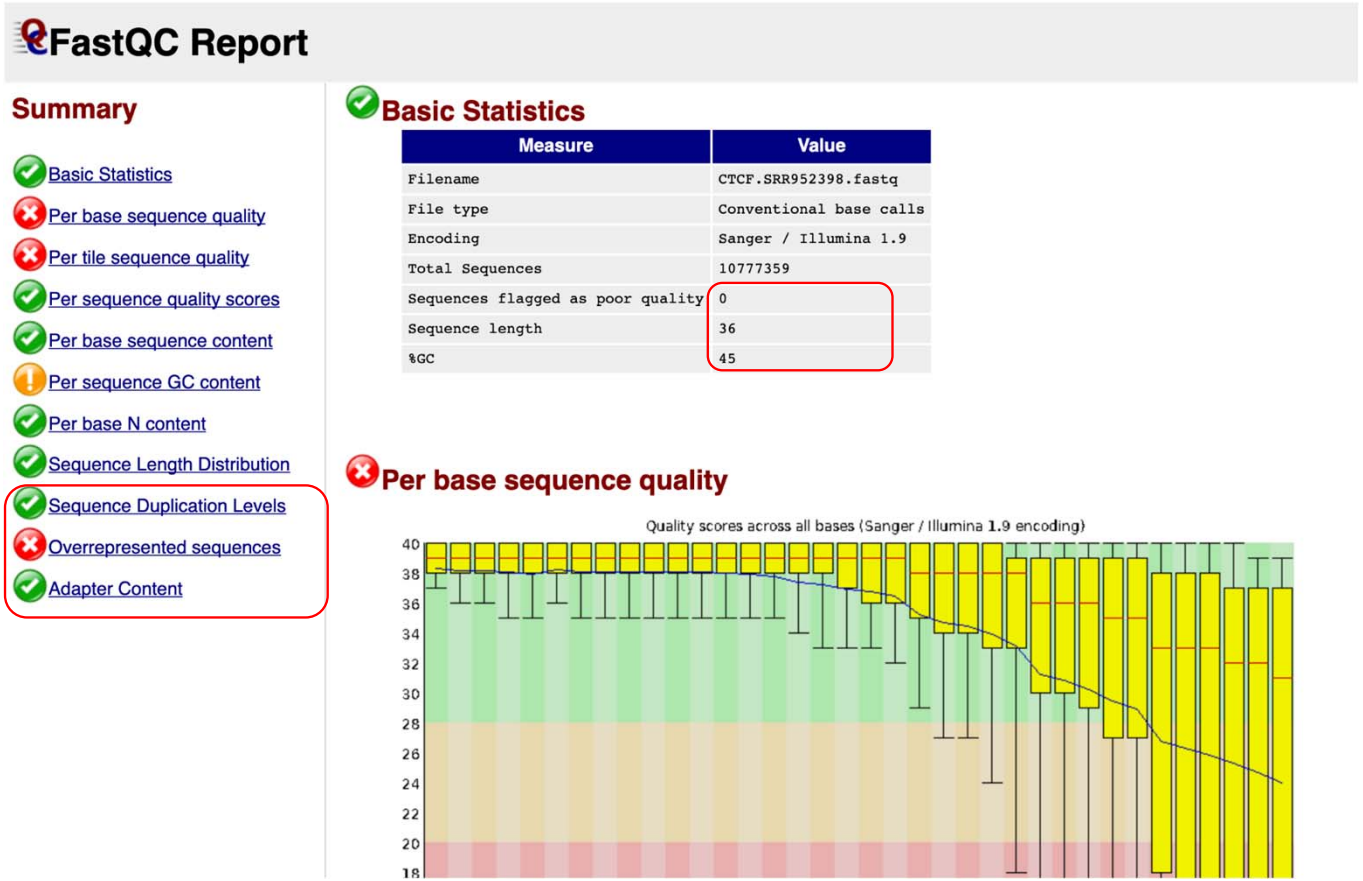
```
[yanjian@localhost fastq]$ wc -l CTCF.SRR952398.fastq
43109436 CTCF.SRR952398.fastq
[yanjian@localhost fastq]$
```

- **43,109,436**/4=10,777,359 raw reads (CTCF ChIP-seq)

```
@SRR952398.1 HWUSI-EAS1683_0015_FC:8:1:1715:921 length=36
NAGAGCCTCCAGTTTAAAAATTAATAAATAACAAAA
+SRR952398.1 HWUSI-EAS1683_0015_FC:8:1:1715:921 length=36
####################################
@SRR952398.2 HWUSI-EAS1683_0015_FC:8:1:3311:928 length=36
NGTAGAATATTATGTGCAATGCTCTGATCCCTTTTT
+SRR952398.2 HWUSI-EAS1683_0015_FC:8:1:3311:928 length=36
####################################
@SRR952398.3 HWUSI-EAS1683_0015_FC:8:1:3469:923 length=36
NTTGTCGGAAAAAGTTCTTATTTTTTTTTCCAAGTT
```

1. Fastqc (quality control of sequencing)

Usage:
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam] [-c contaminant file] seqfile1 .. seqfileN



Fastqc

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAAAA	267245	2.4796891334880837	Illumina Single End Adapter 1 (100% over 28bp)
GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAA	58142	0.5394828176364915	Illumina Single End Adapter 1 (100% over 33bp)
GAAGAGCTCGTATGCCGTCTTCTGCTTGGAAAAAAA	36527	0.33892347837721654	Illumina Single End Adapter 1 (100% over 28bp)
AGCTATATCGGAAGAGCTCGTATGCCGTCTTCTGCT	14286	0.13255566600314603	Illumina Single End Adapter 2 (96% over 32bp)
CGGAAGAGCTCGTATGCCGTCTTCTGCTTGAAAAAA	13390	0.12424194090593066	Illumina Single End Adapter 1 (100% over 30bp)

It suggests that more adaptor clearance step should be taken care of:

a. Run agarose gel
b. SPRI beads
c. others

2. Mapping to the human genome

- Be aware of reference genome
  - Species (human, mouse, ?)
  - version of the genome (hg18, hg19, hg38 or mm8, mm9, mm10)

- bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]
  - Check the option by typing “bowtie”
  - ebwt: specify the path to bowtie index (human reference)
  - s: fastq (single read)
  - hit: output file name (default: stdout)

mapping

1. run bowtie to map fastq reads to the human genome

$ bowtie **-S -q -m 1** /Genome/Human/hg19/hg19 ./CTCF.SRR952398.fastq ./CTCF.SRR952398.sam

Reference genome index | ChIP fastq | mapped output

Important Options:

-S/--sam write hits in SAM format # For further analytic steps

-m <int> suppress all alignments if > <int> exist # we don't need one read to be mapped to multiple loci (1 is preferred)

-q query input files are FASTQ .fq/.fastq # we used fastq as input file

Note that we can feed multiple fastq files to bowtie at the same time by separating them with ","

bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [<hit>]

```
Time loading forward index: 00:00:00
Time loading mirror index: 00:00:01
Seeded quality full-index search: 00:14:39
# reads processed: 10777359
# reads with at least one reported alignment: 7461459 (69.23%)
# reads that failed to align: 797640 (7.40%)
# reads with alignments suppressed due to -m: 2518260 (23.37%)
Reported 7461459 alignments
Time searching: 00:14:40
Overall time: 00:14:40
```

Sort the reads

samtools # convert the output format and sort the reads

samtools view [options] <in.bam>|<in.sam>|<in.sam> [region ...] # Convert the SAM format to BAM

Important Options

S: ignored (input format is auto-detected)

b: output BAM

o: output (otherwise stadut)

Output to a file

$ samtools view /S -b ./CTCF.SRR952398.sam -> ./CTCF.SRR952398.bam

SAM format input

BAM format output

2. run samtools to convert the mapping output to binary format

Sort the reads

samtools sort [options…] <in.bam> # sort the reads to facilitate subsequent analysis

Important option

o: specify the output file

$ samtools sort -o ./CTCF.SRR952398.sort.bam ./CTCF.SRR952398.bam

BAM format input

3. run samtools to sort the mapped reads

Remove PCR duplicates

samtools rmdup [-sS] <input.sort.bam> <output.bam> # other software: picard, macs
-s rmdup for SR reads
-S rmdup for PE reads

$ samtools rmdup /s ./CTCF.SRR952398.sort.bam ./CTCF.SRR952398.nodup.bam
Cleaned output

4. run samtools to remove PCR duplicates

- The results can be viewed in the genome browser or used to call peaks

3. Visualize the mapping in Genome Browser

- First of all, you need to download the files to local computer where genome browser is installed or upload the file to public genome browser such as UCSC genome browser.
- For mapped bam file to be viewed in genome browser, an index file has to be generated using samtools **index**
- samtools index [options…]

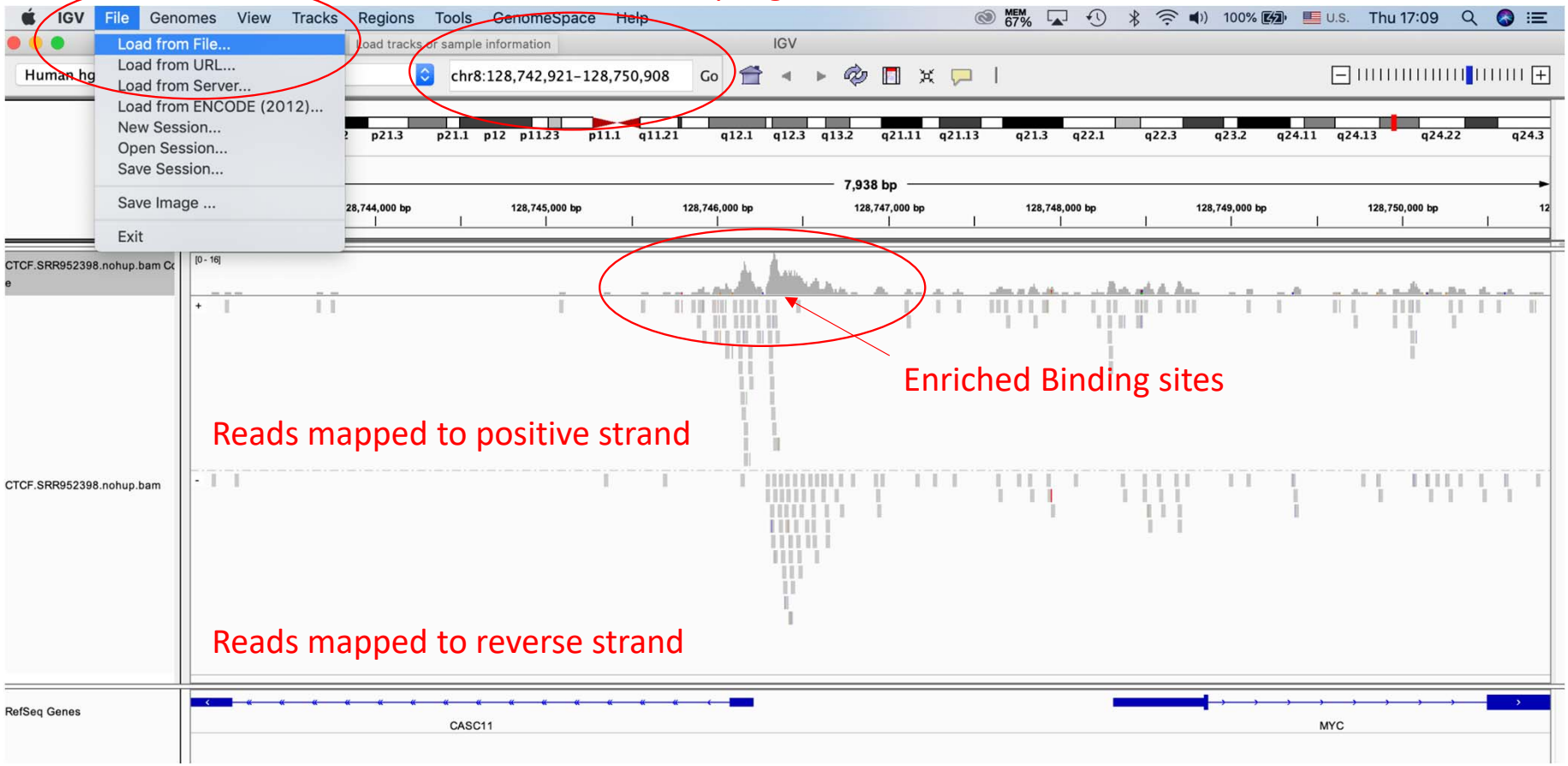
$ samtools index ./CTCF.SRR952398.nodup.bam

- The output will be CTCF.SRR952398.nodup.bam.bai
- Both .bam and .bai files are needed **in the same location** for genome browser!!!

IGV Genome Browser

Load .bam file

Pick up a genomic locus to view



4. MACS2: peak calling

usage: macs2 callpeak -t <TFILE ...> -c <CFILE...>

Important Options:

-f: format (we use BAM)
-g: genome (mm, hs, …)
-n: name of output
--outdir: output directory
/m: fold change cutoff (two values, minimum and maximum)
-q: q value cutoff
-p: p value cutoff

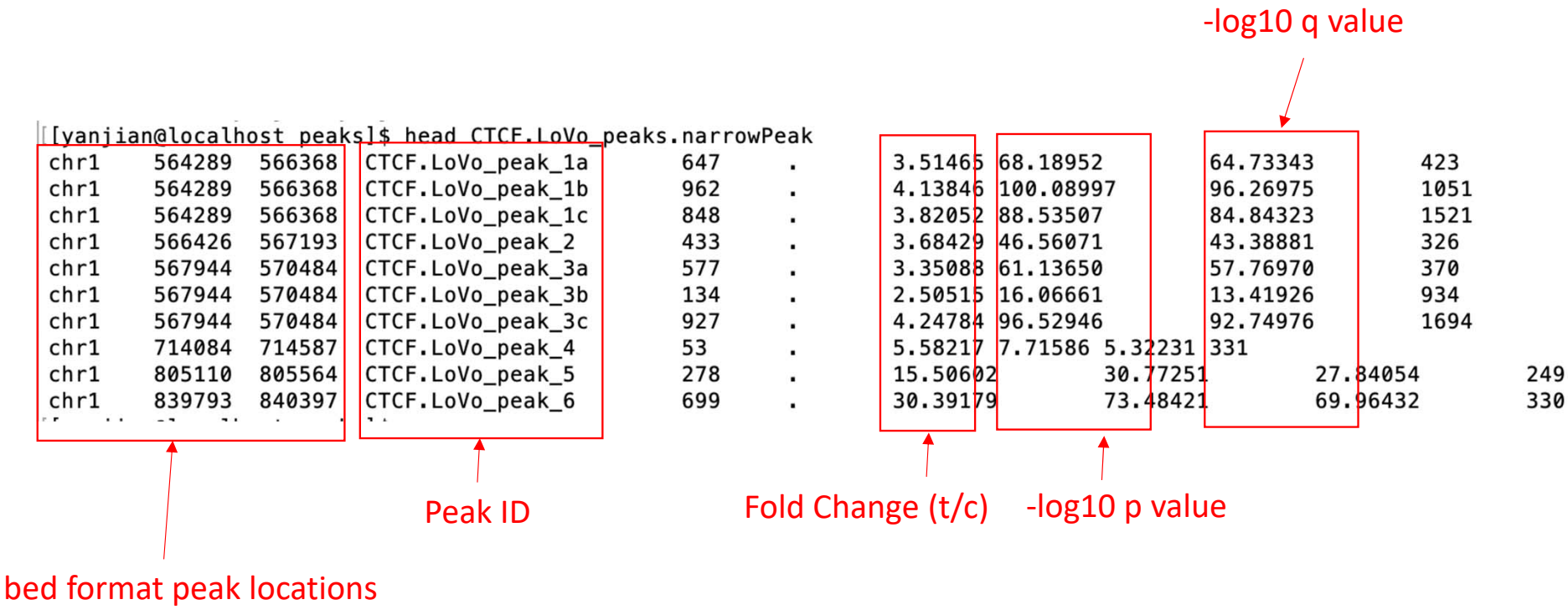
macs2

macs2 callpeak -t ./CTCF.SRR952398.nodup.bam -c ./CTCF.SRR952398.nodup.bam /m 5 50 -g hs /q 0.05 /f BAM /n CTCF.CRC //outdir ./

- The control could be ignored. It assumes a flattened background.
- In total, there are 38,972 narrow peaks are called
- It is difficult to judge whether the experiment is successful based on the number of peaks called

```
     17 CTCF.LoVo_model.r
  38972 CTCF.LoVo_peaks.narrowPeak
  39003 CTCF.LoVo_peaks.xls
  38972 CTCF.LoVo_summits.bed
```

Narrow Peak File (macs2 callpeak output)



Peak File (macs2 callpeak output)

```
# tag size is determined as 36 bps
# total tags in treatment: 6261926
# tags after filtering in treatment: 6261926
# maximum duplicate tags at the same position in treatment = 1
# Redundant rate in treatment: 0.00
# total tags in control: 33873284
# tags after filtering in control: 33873284
# maximum duplicate tags at the same position in control = 1
# Redundant rate in control: 0.00
# d = 264
# alternative fragment length(s) may be 264 bps
chr     start   end     length  abs_summit      pileup  -log10(pvalue)  fold_enrichment -log10(qvalue)  name
chr1    564290  566368  2079    564713  278.00  68.18952        3.51465 64.73343        CTCF.LoVo_peak_1a
chr1    564290  566368  2079    565341  337.00  100.08997       4.13846 96.26975        CTCF.LoVo_peak_1b
chr1    564290  566368  2079    565811  327.00  88.53507        3.82052 84.84323        CTCF.LoVo_peak_1c
chr1    566427  567193  767     566753  175.00  46.56071        3.68429 43.38881        CTCF.LoVo_peak_2
chr1    567945  570484  2540    568315  265.00  61.13650        3.35088 57.76970        CTCF.LoVo_peak_3a
chr1    567945  570484  2540    568879  102.00  16.06661        2.50515 13.41926        CTCF.LoVo_peak_3b
chr1    567945  570484  2540    569639  315.00  96.52946        4.24784 92.74976        CTCF.LoVo_peak_3c
chr1    714085  714587  503     714416  8.00    7.71586 5.58217 5.32231 CTCF.LoVo_peak_4
chr1    805111  805564  454     805360  24.00   30.77251        15.50602        27.84054        CTCF.LoVo_peak_5
```

no. of useful reads in ChIP

no. of useful reads in control

estimated fragment size

peak lengths

peak summit locations

Obtain fasta file from peak file

- Generate the fasta file from .bed file using bedtools getfasta

```
Usage:   bedtools getfasta [OPTIONS] -fi <fasta> -bed <bed/gff/vcf> -fo <fasta>

Options:
        -fi     Input FASTA file
        -bed    BED/GFF/VCF file of ranges to extract from -fi
        -fo     Output file (can be FASTA or TAB-delimited)
        -name   Use the name field for the FASTA header
```

$ bedtools getfasta ⁄fi /Genome/Human/hg19/hg19.fa –bed ./CTCF.LoVo_summits.top500.bed ⁄fo ./CTCF.LoVo_summits.top500.fa

```
>chr12:132991645-132991744
TGGAGGACGGGACTGACCCTCTGGCCACTAGGGGTCTCTCTTGCTCCGCTCTGTGCCAGAGCCGTCGCCAGGAGCCCTGGGATCCGGAACTGGGGCAGA
>chr20:4153197-4153296
AGGACAGAGGCGGGGAGGAGGAAAGGGCCGGAGCCGCCCTGCTCCCGGCGTGGCCTCATCCCCAGCGAAAGCGCGACACCCGCTGTCTGGGAAGGGGAG
>chr8:143820975-143821074
GCGGGTTCCGGGACTCGGGCGCGCGCCCTCTGCTGGCCGTGCGGGCTCAGCATCGGGGGCCCCTCGCAGGTGCCCTCCCAGGGATAGGGGCACCTGCTG
>chr9:104211606-104211705
GGGGTGCAGGGCTAGGGCGGGGACCCTGTCCCGACCGGGTCACTCTCGTGGGAAGAGGGCGGGCGCTGTGTCCGGCAGTCAGAACAAGCGGCTGCGAAG
>chr4:15958000-15958099
ATCATTAGCTCTGGGGGTGGGGCCTGACTTTGTCTGGCAAATACAATACGTTGGTCCTCTGCTGTCACCTGGTGGCCATTACTGGTACGCAGTGGCAAT
```

meme-suite@uw.edu

Inbox - CityU HK 18:17

MEME Submission Information (job appMEME_5.0.51570702630261-471136349)

To: Dr. YAN Jian,

Reply-To: meme-suite@uw.edu

MEME
Multiple Em for Motif Elicitation

This is an auto-generated response to your job submission.

Your job ID is: **appMEME_5.0.51570702630261-471136349**

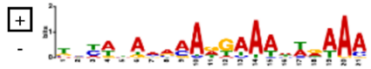
You can view your job results at: http://meme-suite.org/info/status?service=MEME&id=appMEME_5.0.51570702630261-471136349

Job Details

Submitted	Thu Oct 10 10:17:10 UTC 2019
Expires	Mon Oct 14 10:17:10 UTC 2019
(Primary) Sequences	A set of 500 DNA sequences, all 99 in length, from the file `CTCF.LoVo_summits.top500.fa`.
Background	A order-0 background generated from the supplied sequences.
Discovery Mode	Classic: optimizes the E-value of the motif information content
Site Distribution	Zero or one occurrence (of a contributing motif site) per sequence.
Motif Count	Searching for 3 motifs.
Motif Width	Between 6 wide and 50 wide (inclusive).

MEME Result

DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		5.0e-575	375	15		
2.		1.3e-016	20	21		
3.		1.8e+002	18	29		

Stopped because requested number of motifs (3) found.

MOTIF LOCATIONS

Discovered sites of corresponding motifs

Only Motif Sites | Motif Sites+Scanned Sites | All Sequences

	Name	*p*-value	Motif Location
1.	chr12:132991645-132991744	6.19e-3	
2.	chr20:4153197-4153296	2.36e-1	
3.	chr8:143820975-143821074	6.09e-4	
5.	chr4:15958000-15958099	8.98e-8	
6.	chr5:153990059-153990158	2.35e-2	
7.	chr15:73089292-73089391	5.10e-2	
8.	chr2:45870927-45871026	8.57e-4	
9.	chr15:69687040-69687139	2.71e-1	
10.	chr5:98104878-98104977	2.75e-2	
12.	chr22:39930899-39930998	7.67e-3	
13.	chr14:23322403-23322502	7.43e-3	
14.	chr1:211688898-211688997	4.73e-2	
16.	chr12:120524826-120524925	3.10e-2	
19.	chr8:23145533-23145632	1.90e-1