



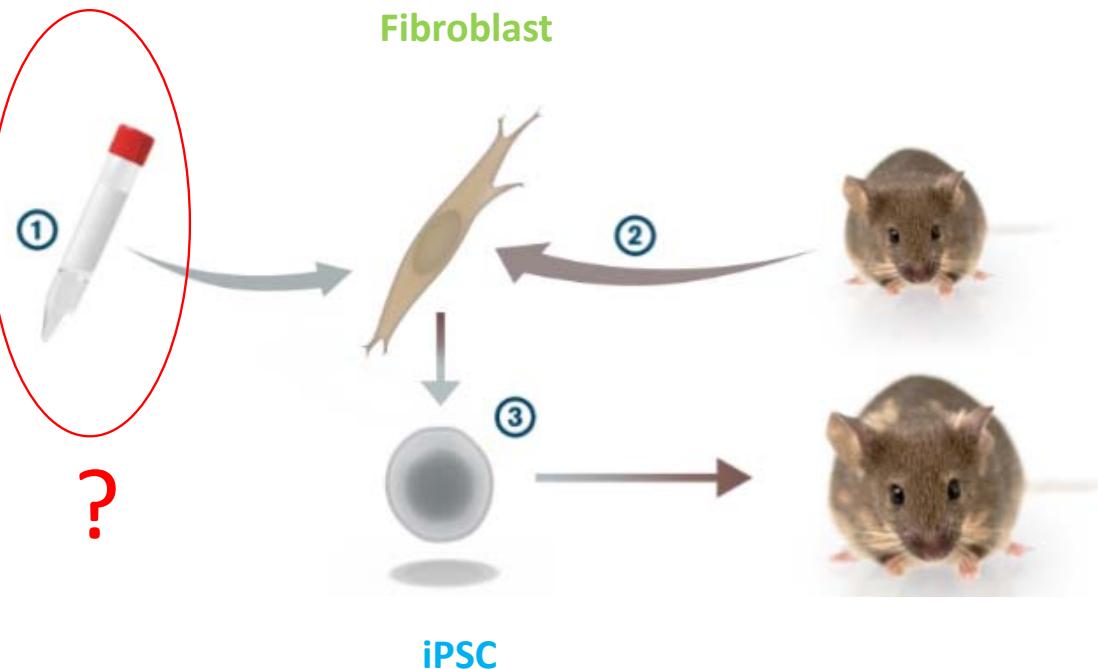
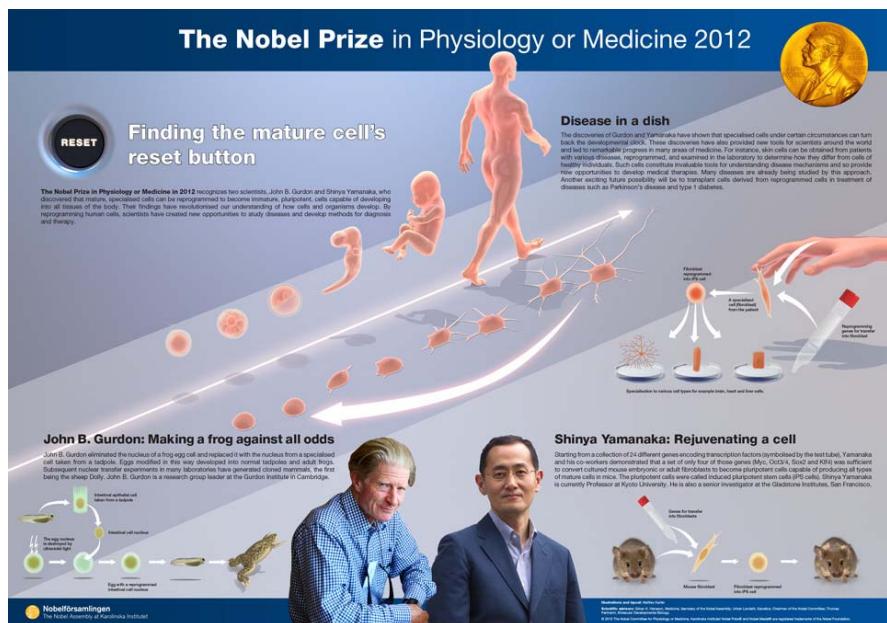
ChIP Sequencing (transcription factors)

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



III, Niklas Elmehed, © Nobel Media.

William G. Kaelin Jr.
Prize share: 1/3



III, Niklas Elmehed, © Nobel Media.

Sir Peter J. Ratcliffe
Prize share: 1/3



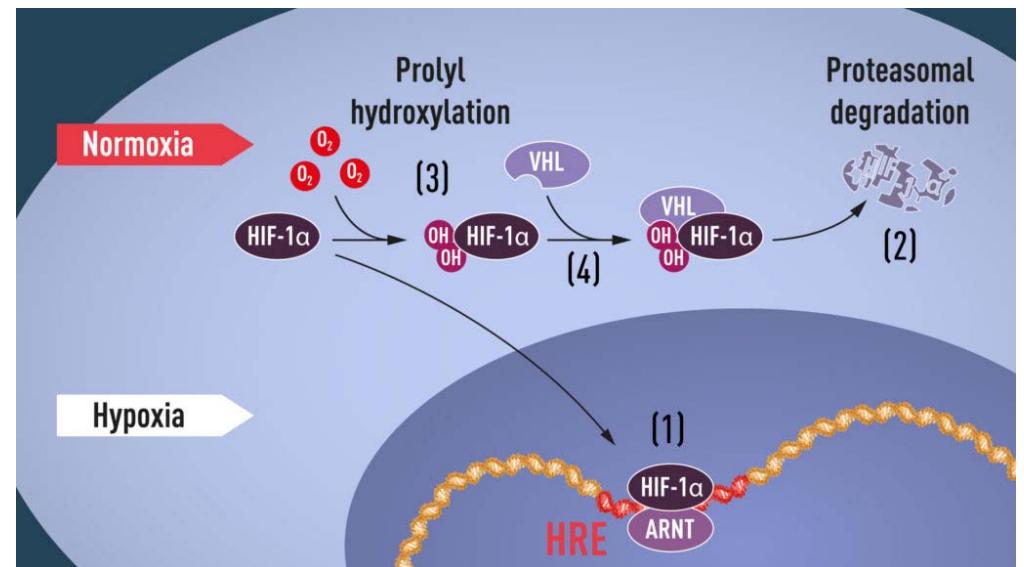
III, Niklas Elmehed, © Nobel Media.

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 alterations 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.



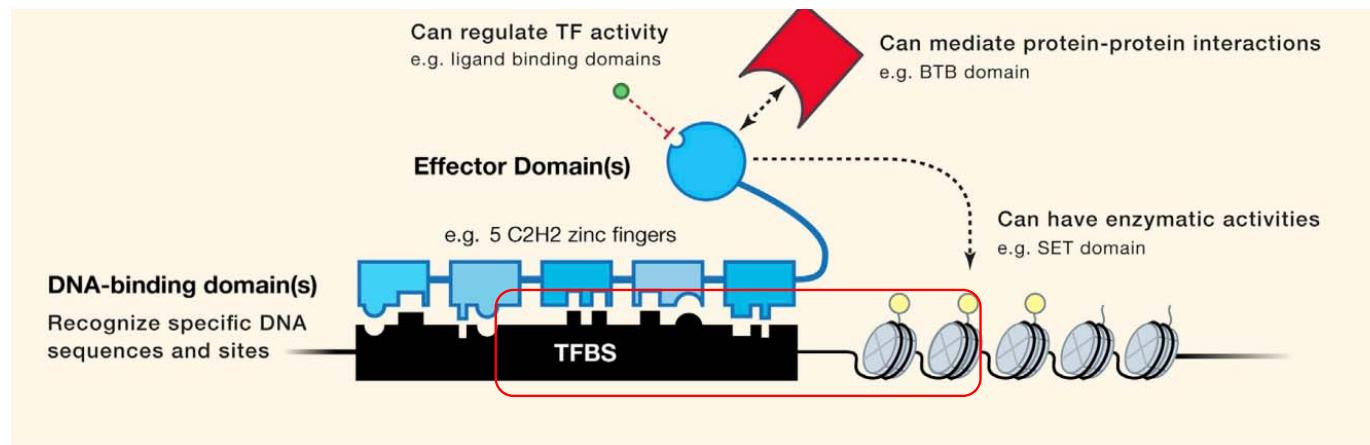
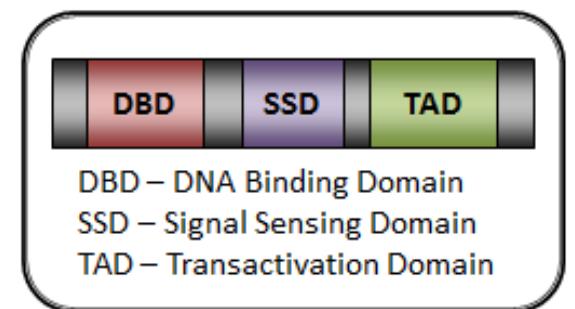
Transcription Factor

- 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)

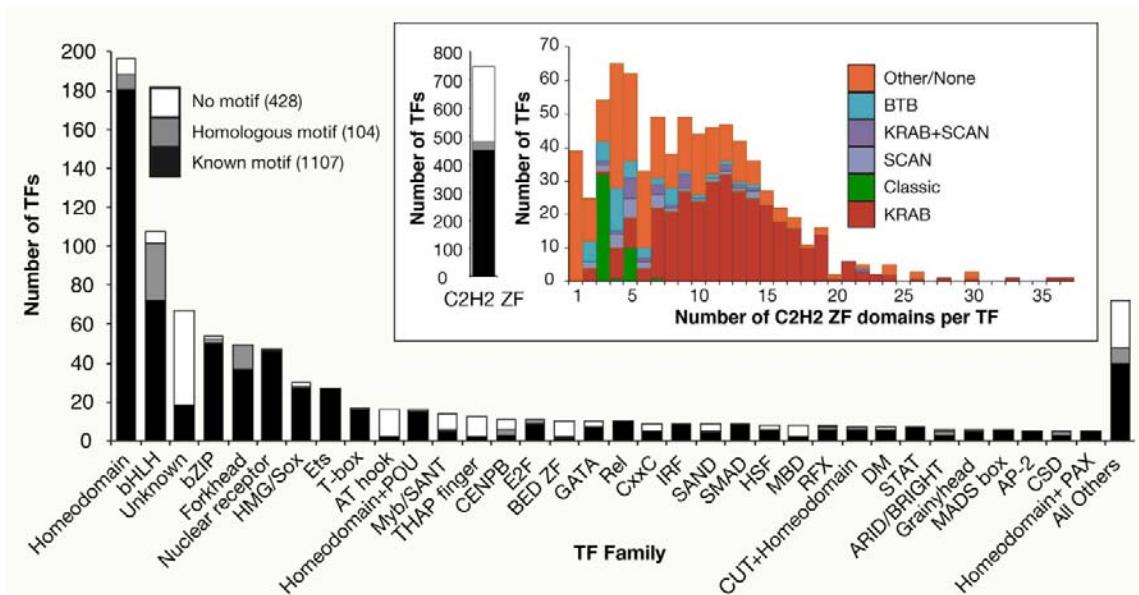
DNA Binding Domain (DBD)

- 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

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

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

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

Position Probability Matrix (PPM)

- Position Probability Matrix (PPM)

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

- 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.

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

- 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/>

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

DTU Bioinformatics
Department of Bio and Health Informatics

Services are gradually being migrated to <https://services.healthtech.dtu.dk/>.
In the near future, cbs.dtu.dk will be retired. Please try out the new site.

Seq2Logo 2.0

Seq2Logo is a web-based sequence logo generation method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion.

Note that Seq2Logo as default includes a pseudo count correction for lowcounts. This means that the amino acid frequencies displayed in the sequence logos are corrected for low number of observations using a Blosum amino acid similarity matrix. To turn this feature off, the Weight on prior must be set to zero.

View the [version history](#) and the [release notes](#) of this server. All the previous versions are available online, for comparison and reference.

Try out the newest version [Seq2Logo-2.1\(BETA\)](#).

Instructions

SUBMISSION

Provide Input (MDA, [Data and Structure](#), [FASTA](#), [PSMS](#)) For more info [click here](#)

Select Logo type: Kullback-Leibler

Clustering method: Clustering (Hobohm1)

Specify threshold for clustering (Hobohm1) 0.63 Threshold (Hobohm1)

Weight on prior (pseudo counts): 200

Select information content units: Bits Text on y-axis: Bits (the text on the y-axis can be edited at will)

Note: The PSSM of non-weight-matrix inputs will always be calculated in Bits¹ or Halfbits if chosen.

Available Output Formats: (multi)
JPEG ▾
PNG
PDF ▾

Submit Clear fields

Instructions: Simply paste your alignment file in the appropriate box, choose your logo type, clustering method, weight on prior and file format and submit.
Need guidet help? [click here](#).

Advanced Settings show



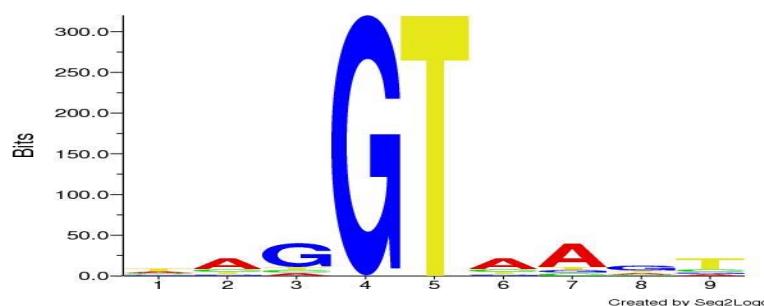
Type in the model in the required formats

<i>A</i>	3	6	1	0	0	6	7	2	1
<i>C</i>	2	2	1	0	0	2	1	1	2
<i>G</i>	1	1	7	10	0	1	1	5	1
<i>T</i>	4	1	1	0	10	1	1	2	6

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/>

JASPAR 2018

≡

Cart JASPAR Blog

Home About Search Browse JASPAR CORE Browse Collections Tools RESTful API Download Data Matrix Clusters Genome Tracks

Search profile(s)

CTCF

Examples: SPI1, P17676, ChIP-seq, Homo sapiens

Search Advanced Options

2 profile(s) found

Display 10 profiles Filter:

ID	Name	Species	Class	Family	Logo
MA0139.1	CTCF	Homo sapiens	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	
MA1102.1	CTCFL	Homo sapiens	C2H2 zinc finger factors	More than 3 adjacent zinc finger factors	

Copy CSV

Showing 2 profiles of page 1 from 1 pages

1

Analyze selected profiles

Please select matrix profiles on the left side to add to your cart or perform the following analysis.

Add to cart

You have 0 profile(s) in your cart. You can add profiles to the cart to download or perform analysis.

Add to cart View cart

Scan

Cluster

Randomize

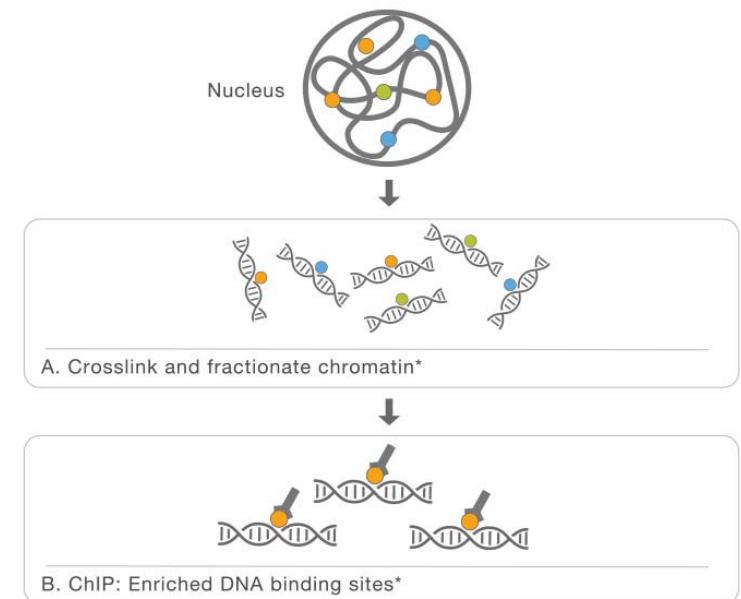
Permute

Download

Experimental Methods to Identify TF Binding

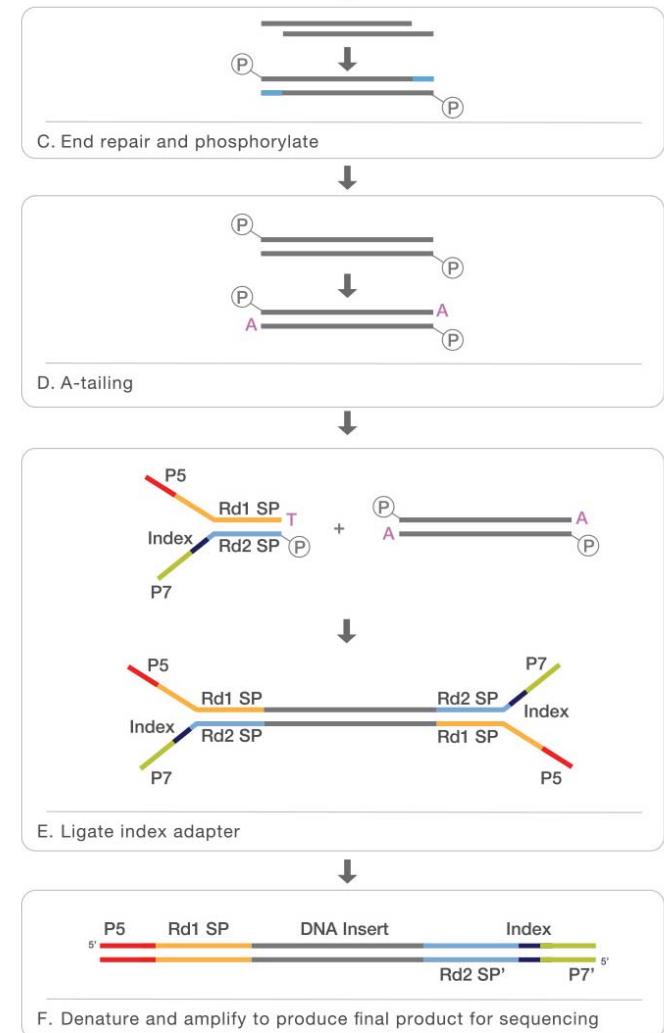
Chromatin Immunoprecipitation (ChIP)

- 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

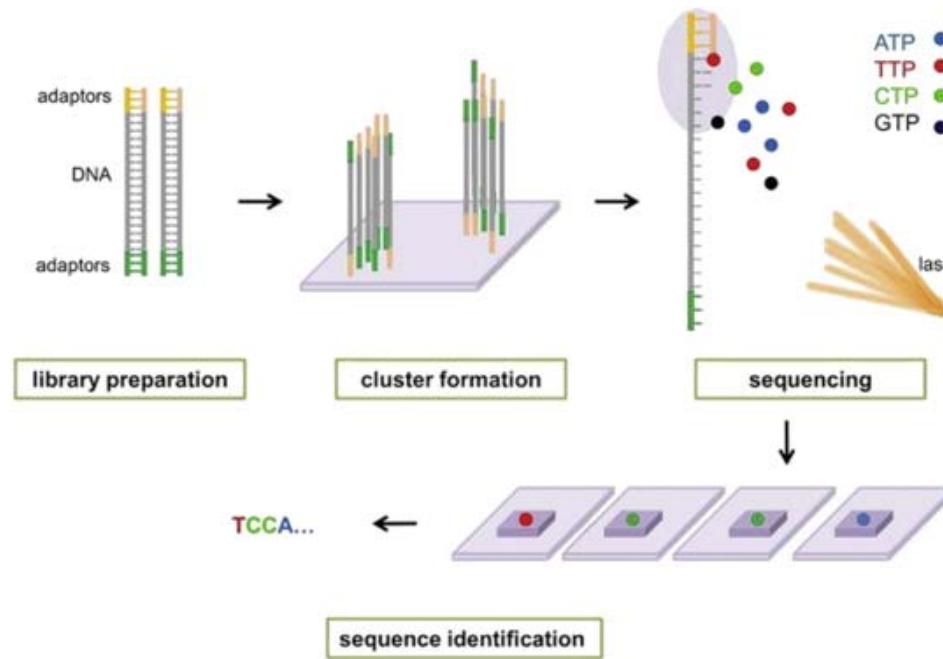


Sequencing Library

- 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



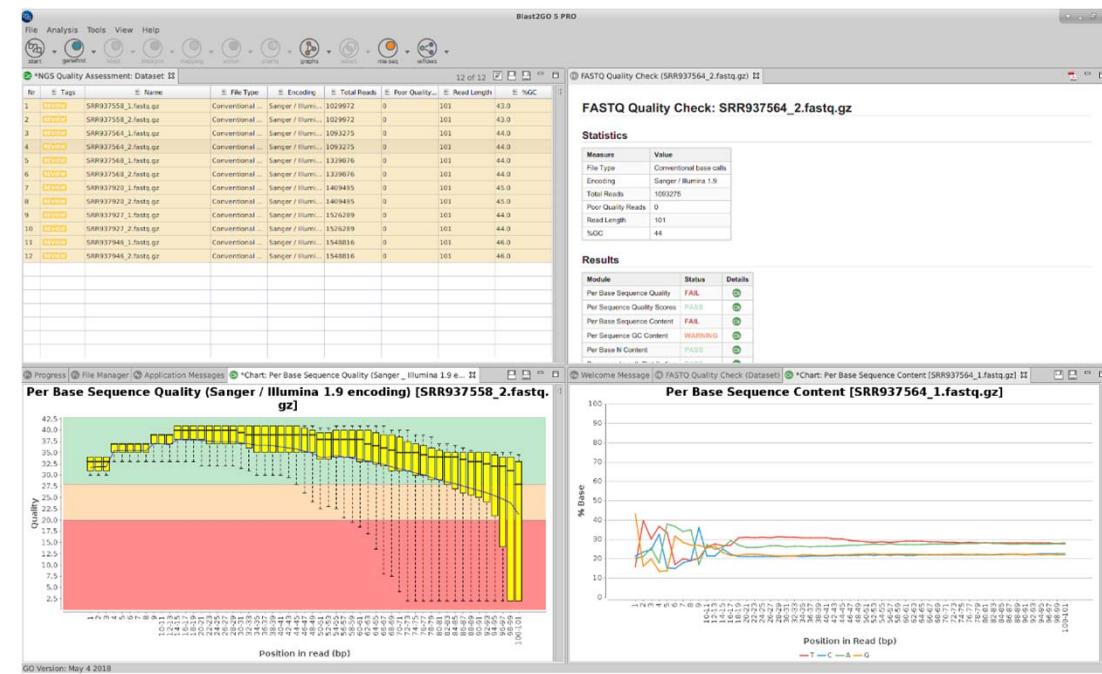
- <https://www.youtube.com/watch?v=fCd6B5HRaZ8>

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.fasta
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

Per base sequence quality

Quality scores across all bases (Illumina 1.5 encoding)

Position in read (bp)

Mapping tools

Read Mapping

bioRxiv preprint doi: <https://doi.org/10.1101/2023.09.04.552321>; this version posted September 4, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY-ND 4.0 International license](https://creativecommons.org/licenses/by-nd/4.0/).

	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

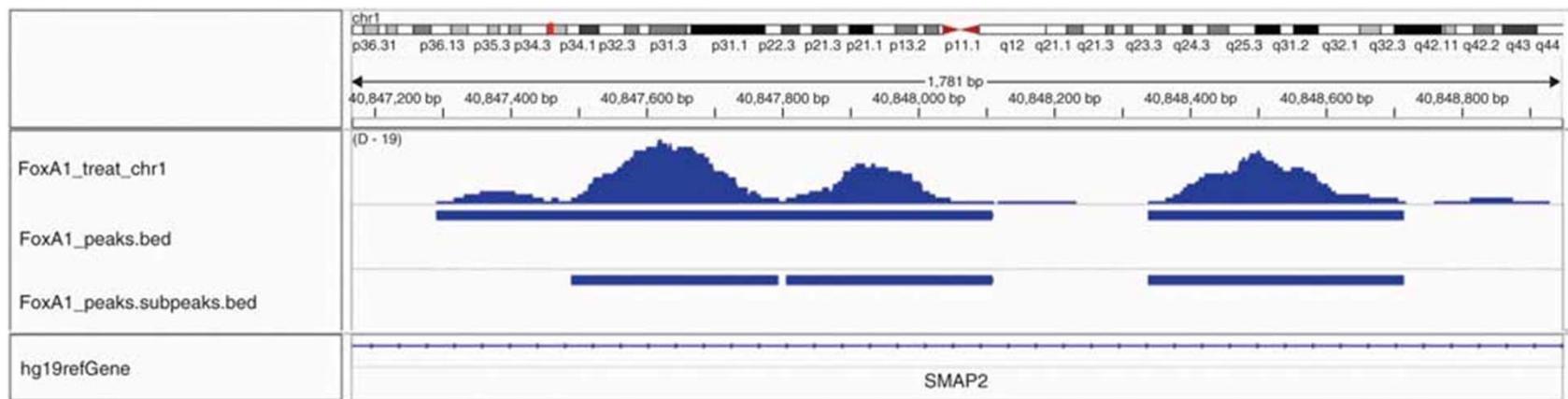
Given 9% duplicates, an additional sequencing run of the same size (from the same library) will give you 1.6x more unique reads. Two additional runs will give you 2.2x more (1.6×1.4).

...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.

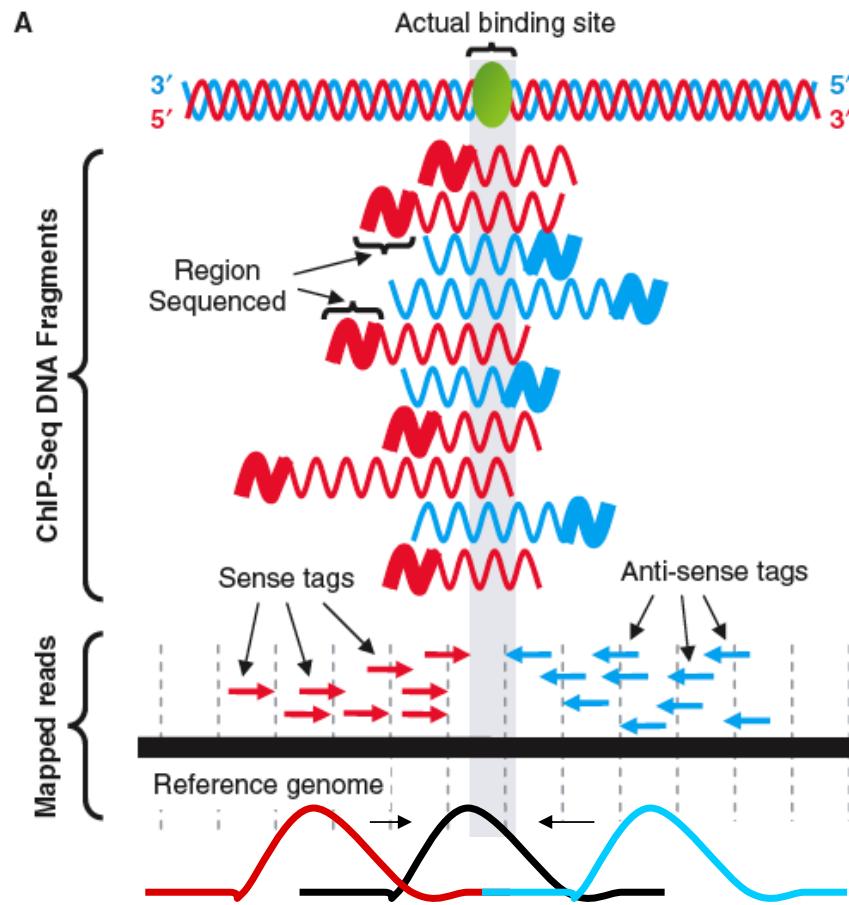
IGV visualization of MACS results



- 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 $\frac{1}{2}$ fragsize & antisense upstream by $\frac{1}{2}$ 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
	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		chromosome scale Poisson dist.
MACS	13	1.3.5		X			X			X		X		local Poisson dist.
QuEST	14	2.3				X	X			X**		X		chromosome 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 excludes putative duplicated regions, no treatment of deletions

Wilbanks and colleagues 2010 Plos One

MACS procedure

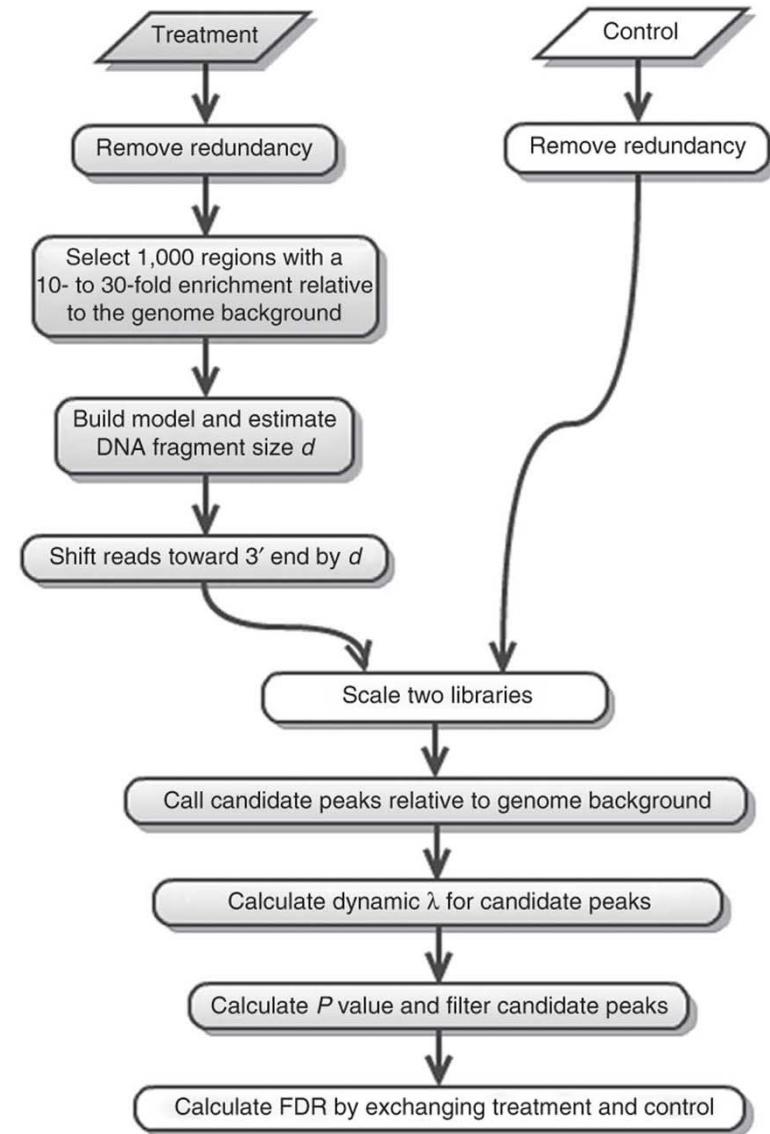
$$P_{\lambda}(X=k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

λ = mean = expected value = variance

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

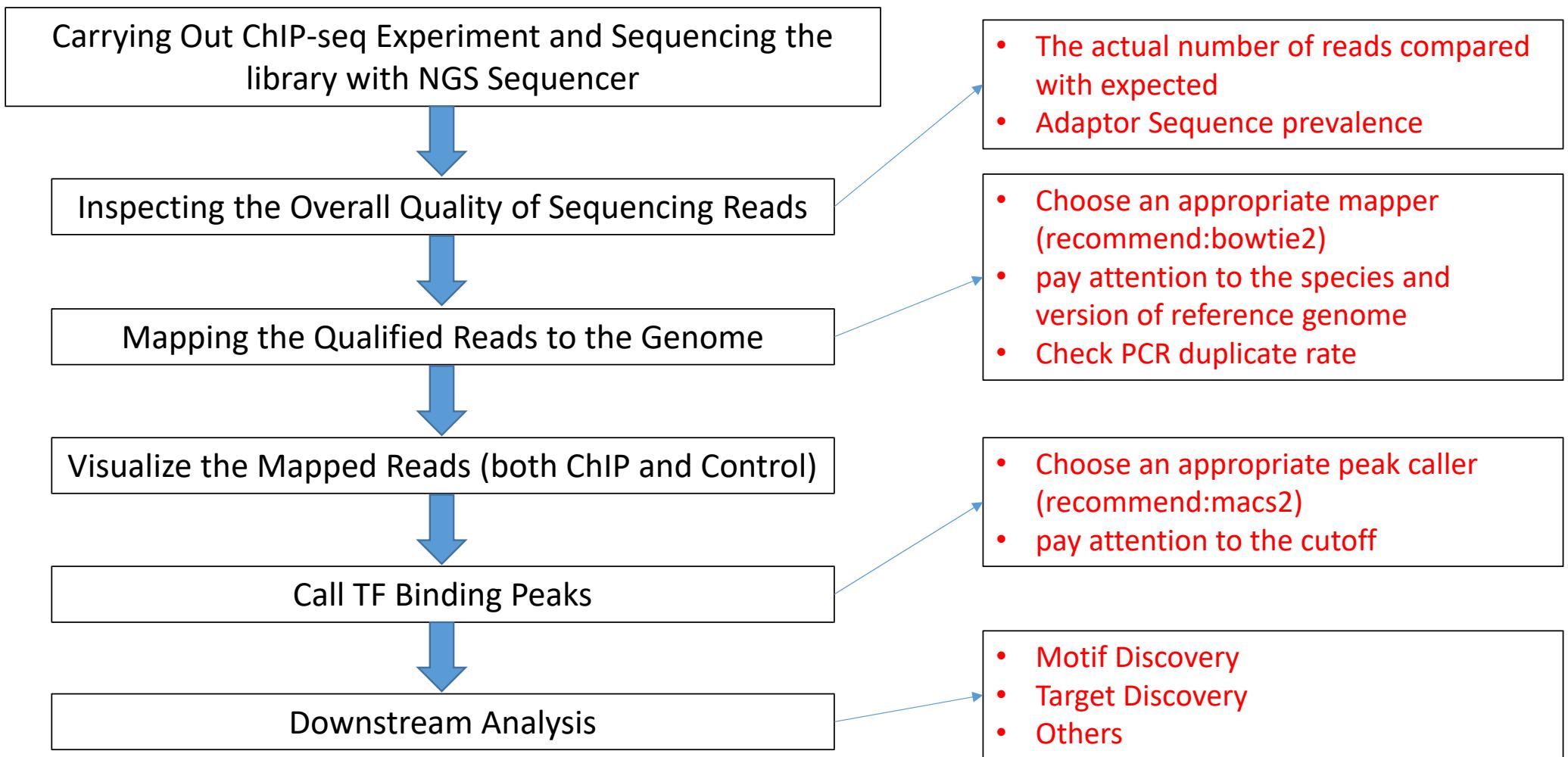
$$= \frac{\text{Read length (nt)} * \text{Total read number}}{\text{Effective genome length (nt)}}$$

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 Rhee¹ and B. Franklin Pugh^{1,*}

¹Center 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

The screenshot shows the Galaxy web interface with the following steps highlighted:

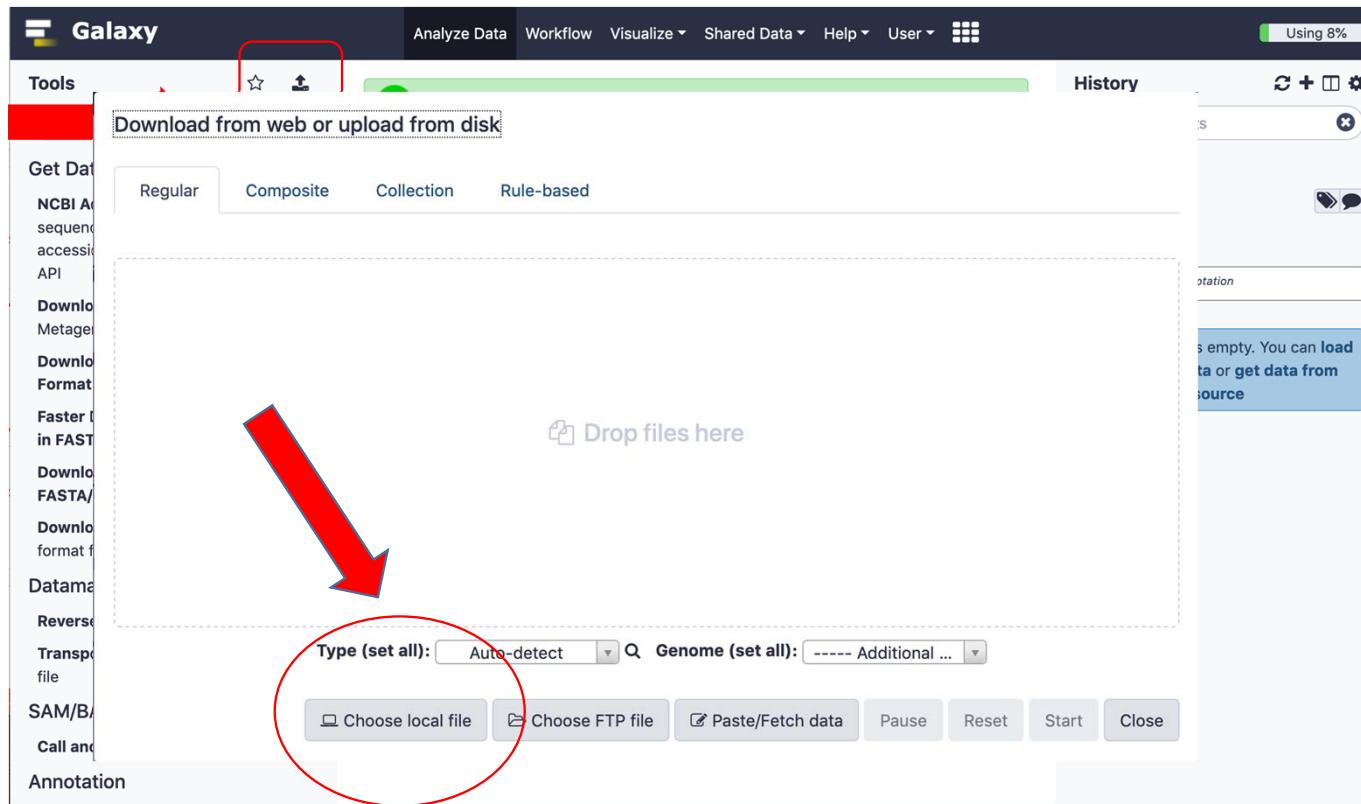
- 1. Find the software**: A red circle highlights the "Faster Download and Extract Reads in FASTQ format from NCBI SRA" tool in the "Tools" sidebar.
- 2. Type the accession number in the text box**: A red arrow points from the text input field to the "SRR346400" accession number.
- 3. Execute...**: A red circle highlights the "Execute" button at the bottom of the tool form.

Galaxy Interface Elements:

- Tools**: A sidebar listing various data retrieval tools like "download and extract".
- Get Data**: A section with links to "NCBI Accession Download", "Download run data from EBI Metagenomics database", "Download and Generate Pileup Format from NCBI SRA", and the circled "Faster Download and Extract Reads in FASTQ format from NCBI SRA".
- Download and Extract Reads in FASTA/Q format from NCBI SRA (Galaxy Version 2.10.4)**: The main tool panel with the following fields:
 - select input type**: A dropdown set to "SRR accession".
 - Accession**: A text input field containing "SRR346400".
 - Select output format**: Radio buttons for "gzip compressed fastq" (selected), "Uncompressed fastq", and "Bzip2 compressed fastq".
 - Email notification**: Buttons for "Yes" and "No".
- History**: A panel showing an "Unnamed history" with 54 deleted datasets, 13.67 GB total size, and a message about loading data.

Or Uploading your own Fastq File

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



1. fastqc

The screenshot shows the Galaxy web interface with the following steps highlighted:

- 2. Find Fastq QC tool box**: A red box highlights the "Tools" sidebar on the left, specifically the "fastqc" entry under the "FASTA/FASTQ" section.
- 3. Choose the dataset you want to analyze**: A red arrow points to the "Short read data from your current history" input field, which contains "2: Reb1_R2.fastqsanger". This field is also highlighted with a red box.
- 4. Click "Execute"**: A red box highlights the "Execute" button at the bottom left of the tool configuration panel.

History panel on the right shows two datasets:

- 2: Reb1_R2.fastqsanger
- 1: Input_R2.fastqsanger

1. Find your uploaded data here.

Galaxy

Analyze Data Workflow Visualize Shared Data Help User Fri 6 Mar 2020 Using 9% Input_R2_fastqsanger

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

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

Top 10 adapter sequences

Sequence	Count	Percentage	Possible Source
GAACTCCAGTCACCTGGCGCA	51709	4.056960163191652	TruSeq Adapter, Index 7 (97% over 36bp)
AACTCCAGTCACCTGGCGCAT	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 Galaxy interface shows the 'Tools' panel on the left with various bioinformatics tools listed. The 'bowtie2' tool is selected and highlighted with a red box. The main workspace on the right shows the configuration for the 'bowtie2' tool. A red box highlights the 'FASTA/Q file' input field, which contains two entries: '2: (unavailable) Reb1_R2.fastqsanger' and '1: (unavailable) Input_R2.fastqsanger'. Below this, a red box highlights the 'Select reference genome' dropdown menu, which is set to 'Yeast (Saccharomyces cerevisiae): sacCer3'. At the bottom, a red box highlights the 'Execute' button.

Galaxy

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾

Tools

bowtie2

Annotation

TB-Profiler Profile Infer strain types and drug resistance markers from sequences

Mapping

Map with minimap2 A fast pairwise aligner for genomic and spliced nucleotide sequences

Bowtie2 - map reads against reference genome

RNA-seq

HISAT2 A fast and sensitive alignment program

Metagenomic Analysis

MaxBin2 clusters metagenomic contigs into bins

deepTools

bamPEFragmentSize Estimate the predominant cDNA fragment length from paired-end sequenced BAM/CRAM files

Workflows

All workflows

Is this single or paired library

Single-end

FASTA/Q file

2: (unavailable) Reb1_R2.fastqsanger
1: (unavailable) Input_R2.fastqsanger

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Must be of datatype "fastqsanger" or "fasta"

Write unaligned reads (in fastq format) to separate file(s)

Yes No

--un/--un-conc (possibly with -gz or -bz2); This triggers --un parameter for single reads and --un-conc for paired reads

Write aligned reads (in fastq format) to separate file(s)

Yes No

--al/--al-conc (possibly with -gz or -bz2); This triggers --al parameter for single reads and --al-conc for paired reads

Will you select a reference genome from your history or use a built-in index?

Use a built-in genome index

Built-ins were indexed using default options. See 'Indexes' section of help below

Select reference genome

Yeast (Saccharomyces cerevisiae): sacCer3

Execute

4. click the button

2. Choose the fastq file

3. Choose the reference genome

The mapping takes a while for a large genome

1. You can download the mapped reads

7: Bowtie2 on data 1: alignments
32.8 MB
format: **bam**, database: **sacCer3**

1274575 reads; of these:
1274575 (100.00%) were unpaired; of these:
145967 (11.45%) aligned 0 times
875408 (68.68%) aligned exactly 1 time
253200 (19.87%) aligned >1 times
88.55% overall alignment rate
[bam_sort_core] merging from 0 files
a

Annotation:
[Click here to edit annotation](#)

display at UCSC main
display with IGV local S. cerevisiae (sacCer3)

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

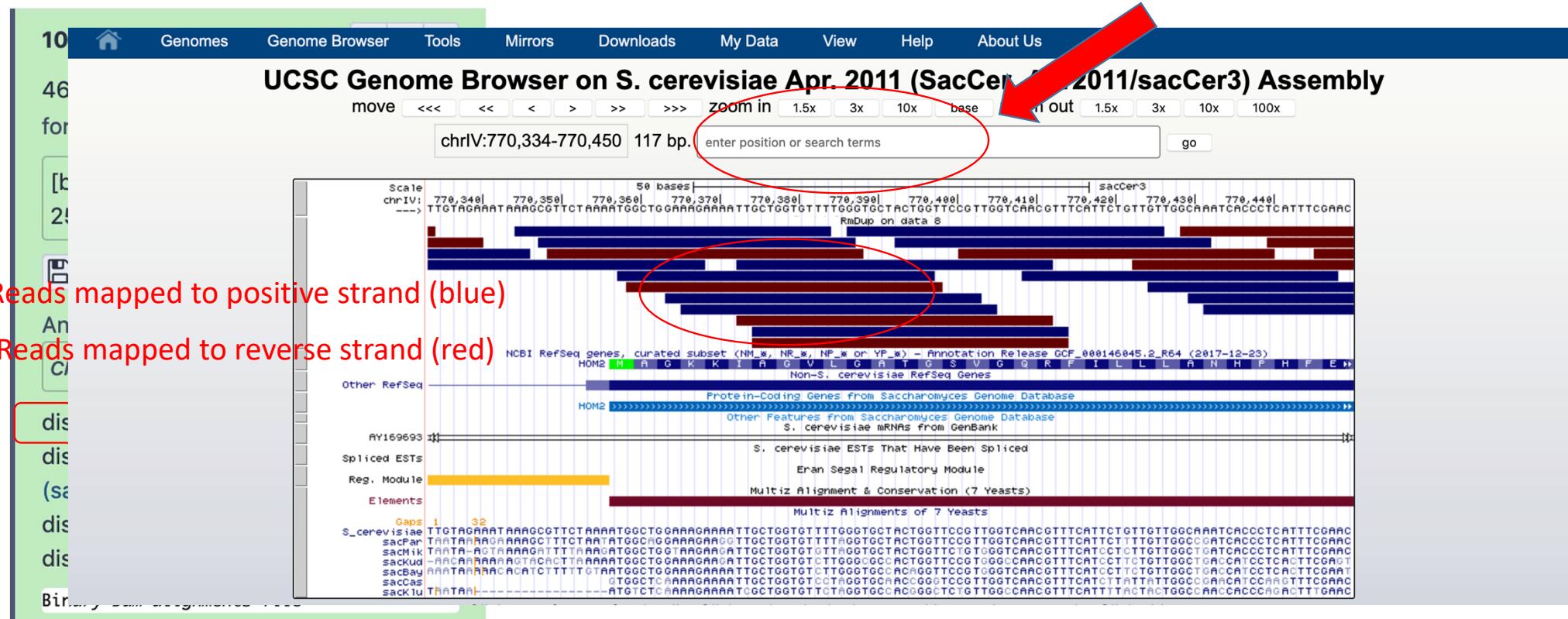
The screenshot shows the Galaxy web interface with the following steps highlighted:

- 1. Find RmDup**: A red box highlights the "Tools" dropdown menu on the left, which contains the "RmDup" entry.
- 2. choose correct dataset**: A red box highlights the "BAM File" input field, which contains two entries: "8: Bowtie2 on data 2: alignments" and "7: Bowtie2 on data 1: alignments".
- 3. choose SR or PE: in our case SR**: A red box highlights the "Is this paired-end or single end data" dropdown menu, which has "BAM is single-end (-s)" selected.
- 4. click "execute"**: A red box highlights the "Execute" button at the bottom left of the tool configuration panel.

The Galaxy interface includes a navigation bar with "Analyze Data", "Workflow", "Visualize", "Shared Data", "Help", and "User" options. The main content area shows the "RmDup remove PCR duplicates (Galaxy Version 2.0.1)" tool details and configuration fields.

3. Visualize the mapping in Genome Browser

Pick up a genomic locus to view



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)

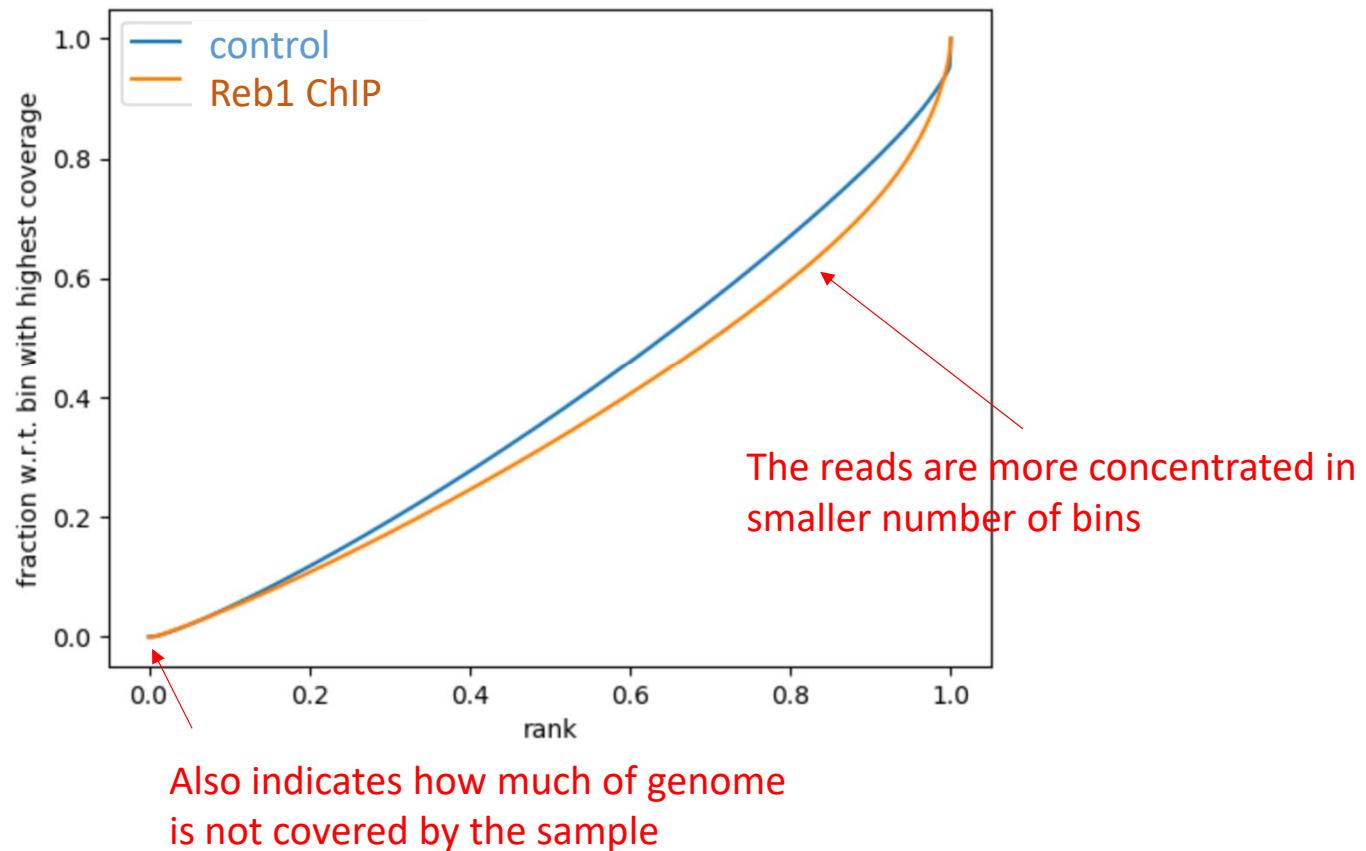
1. Find “plotFingerprint”

The screenshot shows the Galaxy web interface. In the top navigation bar, the 'Tools' tab is selected. Under the 'Tools' tab, a search bar contains the query 'deeptools plotfingers'. Below the search bar, a list of tools is displayed. The 'plotFingerprint' tool is highlighted with a red box and has a tooltip: 'plotFingerprint plots profiles of BAM files; useful for assessing ChIP signal strength (Galaxy Version 3.3.2.0.0)'. Other listed tools include 'plotCorrelation', 'plotCoverage', 'plotEnrichment', 'plotHeatmap', 'plotPCA', 'bamCoverage', 'plotProfile', 'pyGenomeTracks', and 'Workflows'. On the right side of the interface, there are sections for 'Sample order matters' (set to 'No'), 'BAM/CRAM file' (listing '10: RmDup on data 8', '9: RmDup on data 7', '8: Bowtie2 on data 2: alignments', and '7: Bowtie2 on data 1: alignments'), 'Would you like custom sample labels?' (set to 'No, use sample names in the history'), 'Region of the genome to limit the operation to' (empty), 'Show advanced options' (set to 'no'), and 'Show advanced output settings' (empty). At the bottom center is a large blue button with a white checkmark and the text 'Execute'.

2. Choose the dataset

3. Click “Execute”

Signal Extraction Scaling Output



4. peak calling (macs2)

1. Find MACS2 callpeak

Galaxy

Analyze Data Workflow Visualize Shared Data Help User

Tools

macs2 callpeak

ChiP-Seq peak data

ChIPseeker for ChIP peak annotation and visualization

MACS2 refinepeak Refine peak summits and give scores measuring balance of forward- backward tags (Experimental)

MACS2 predictd Predict 'd' or fragment size from alignment results

MACS2 randsample Randomly sample number or percentage of total reads

MACS2 filterdup Remove duplicate reads at the same position

MACS2 callpeak Call peaks from alignment results

MACS2 bdgdiff Differential peak detection based on paired four bedgraph files

MACS2 bdgcmp Deduct noise by comparing two signal tracks in bedGraph

MACS2 bdgbroadcast Call broad peaks from bedGraph output

MACS2 bdgpeakcall Call peaks from bedGraph output

MACS2 callpeak Call peaks from alignment results (Galaxy Version 2.1.1.20160309.6)

Are you pooling Treatment Files?

No

For more information, see Help section below

ChIP-Seq Treatment File

10: RmDup on data 8

(-t)

Do you have a Control File?

Yes

Are you pooling Control Files?

No

For more information, see Help section below

ChIP-Seq Control File

9: RmDup on data 7

(-c)

Format of Input Files

Single-end BAM

For Paired-end BAM (BAMPE) the 'Build model step' will be ignored and the real fragments will be used for each template defined by leftmost and rightmost mapping positions. Default: Single-end BAM (--format)

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)

Effective genome size

User defined

The effective genome size is the portion of the genome that is mappable. Large fractions of the genome are stretches of Ns that should be discarded. Also, if repetitive regions were not included in the mapping of reads, the effective genome size needs to be adjusted accordingly. Sizes are from the MACS2 website (--gszie)

Effective genome size

12000000

✓ Execute

4. click the button

Output view

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGE
chr1	14	692	RmDup_on_data_8_peak_1	480	.	11.14585	51.56381	48.0847
chr1	87010	87312	RmDup_on_data_8_peak_2	210	.	6.70586	23.59468	21.0976
chr1	92553	92832	RmDup_on_data_8_peak_3	355	.	9.27723	38.55616	35.5841
chr1	119704	119832	RmDup_on_data_8_peak_4	92	.	4.56829	11.38973	9.2325
chr1	141715	141852	RmDup_on_data_8_peak_5	113	.	4.48007	13.53213	11.3240
chr1	229660	229858	RmDup_on_data_8_peak_6	46	.	3.37237	6.66180	4.6799
chr1	230065	230218	RmDup_on_data_8_peak_7	195	.	6.58641	22.02638	19.5628
chr11	0	196	RmDup_on_data_8_peak_8	365	.	9.35197	39.60216	36.5981
chr11	5516	6459	RmDup_on_data_8_peak_9	328	.	8.08232	35.77129	32.8956
chr11	82372	82536	RmDup_on_data_8_peak_10	170	.	4.68286	19.44128	17.0542
chr11	84367	84512	RmDup_on_data_8_peak_11	87	.	4.30705	10.89353	8.7541
chr11	87932	88044	RmDup_on_data_8_peak_12	34	.	2.78039	5.34426	3.4373
chr11	110976	111477	RmDup_on_data_8_peak_13	407	.	10.09942	43.85774	40.7145
chr11	116888	117024	RmDup_on_data_8_peak_14	65	.	3.82084	8.58513	6.5271
chr11	122078	122333	RmDup_on_data_8_peak_15	170	.	6.13794	19.39235	17.0076
chr11	124846	125105	RmDup_on_data_8_peak_16	178	.	6.28743	20.25953	17.8482
chr11	132232	132377	RmDup_on_data_8_peak_17	58	.	3.67135	7.92574	5.8922
chr11	135724	136171	RmDup_on_data_8_peak_18	396	.	9.87519	42.78334	39.6751
chr11	151182	151560	RmDup_on_data_8_peak_19	386	.	9.72570	41.71587	38.6386
chr11	159863	159969	RmDup_on_data_8_peak_20	114	.	5.01676	13.65277	11.4323
chr11	164913	165024	RmDup_on_data_8_peak_21	52	.	3.59661	7.28438	5.2759
chr11	201706	201990	RmDup_on_data_8_peak_22	334	.	8.46607	36.31299	33.4266
chr11	202426	202554	RmDup_on_data_8_peak_23	53	.	3.51932	7.35632	5.3471
chr11	203186	203348	RmDup_on_data_8_peak_24	46	.	3.44712	6.66180	4.6799
chr11	256594	256684	RmDup_on_data_8_peak_25	21	.	2.19376	3.92038	2.1050
chr11	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								

View the peak result

11: MACS2 callpeak on data 9 and data 10 (narrow Peaks)

856 regions

format: bed, database: sacCer3

/cvmfs/main.galaxyproject.org
/deps/_conda/envs/mulled-v1-9745eaf08708c75d5cb939b005b2
/lib/python2.7/site-packages/MACS2
/OptValidator.py:27: RuntimeWarning:
numpy.dtype size changed, may indicate binary incompatibility

Annotation:

Click here to edit annotation

display in IGB View
display with IGV local S. cerevisiae (sacCer3)
display at UCSC main



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

MEME Suite 5.0.5

Motif Discovery

- MEME
- DREME
- MEME-ChIP
- GLAM2
- MoMo

Motif Enrichment

Motif Scanning

Motif Comparison

Gene Regulation

Manual

Guides & Tutorials

Sample Outputs

File Format Reference

Databases

Download & Install

Help

Alternate Servers

Authors & Citing

Recent Jobs

[Previous version 5.0.4](#)

MEME
Multiple Em for Motif Elicitation

Version 5.0.5

Data Submission Form

Perform motif discovery on DNA, RNA, protein or custom alphabet datasets.

Select the motif discovery mode 
 Classic mode Discriminative mode Differential Enrichment mode 

Select the sequence alphabet
Use sequences with a standard alphabet or specify a custom alphabet. 
 DNA, RNA or Protein Custom No file selected.

Input the primary sequences
Enter sequences in which you want to find motifs. 
 CTCF.LoVo_summits.top500.fa DNA 

Select the site distribution
How do you expect motif sites to be distributed in sequences? 

Select the number of motifs
How many motifs should MEME find? 

Input job details
(Optional) Enter your email address. 

(Optional) Enter a job description. 

Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

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”

The screenshot shows the Galaxy web interface with the 'Tools' panel on the left and the 'Extract Genomic DNA' tool configuration on the right.

Tools Panel:

- Fetch Sequences/Alignments (highlighted with a red box)
- Extract Genomic DNA using coordinates from assembled/unassembled genomes (highlighted with a red box)
- Extract Pairwise MAF blocks given a set of genomic intervals
- Extract MAF blocks given a set of genomic intervals
- Stitch MAF blocks given a set of genomic intervals
- Annotation
- RepeatMasker screen DNA sequences for interspersed repeats and low complexity regions
- Variant Calling
- Mutate Codons with SNPs
- ChIP-seq
- Resize coordinate window of GFF data
- RNA-seq
- pizzly - fast fusion detection using kallisto

Tool Configuration:

- Extract Genomic DNA using coordinates from assembled/unassembled genomes (Galaxy Version 3.0.3+galaxy2):**
 - Fetch sequences for intervals in: 11: MACS2 callpeak on data 9 and data 10 (narrow Peaks) (highlighted with a red box)
 - Interpret features when possible: Yes
 - Choose the source for the reference genome:
 - locally cached (highlighted with a red box)
 - Using reference genome (highlighted with a red box)
 - Select output format: fasta
 - Select fasta header format: bedtools getfasta default
 - Email notification: Yes
- Execute:** A blue button at the bottom left of the configuration area.

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)

The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Analyze Data, Workflow, Visualize, Shared Data, Help, User.
- Tools Panel (Left):**
 - Tools: filter sequence by length (selected)
 - Filter and Sort: Filter data on any column using simple expressions
 - GFF: Filter GFF data by attribute using simple expressions, Filter GFF data by feature count using simple expressions
 - FASTA/FASTQ: Compute sequence length, Filter sequences by length (highlighted with a red box), Filter FASTQ reads by quality score and length
- Tool Configuration (Right):**
 - Tool Name: Filter sequences by length (Galaxy Version 1.2)
 - Fasta file: 16: Extract Genomic DNA on data 15 and data 11
 - Minimal length: 8
 - Maximum length: 0
 - Description: Setting to '0' will return all sequences longer than the 'Minimal length'
 - Email notification: Yes (radio button selected)
 - Description: Send an email notification when the job completes.
 - Execute Button: ✓ Execute

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)

Sequences

16: Extract Geno

Options Configuration

Advanced

Name of sequence set

Galaxy FASTA Input

(-sf)

Sequence Alphabet

DNA

Check reverse complement

Yes No

62: MEME-ChIP (html) on data 56
221.5 bp
format: html, database: sacCer3

Download the MEME report of html file

HTML file

Make sure that you tick the box of “check the reverse complement”

The screenshot shows the Galaxy web interface with the MEME tool selected. A red box highlights the 'MEME-ChIP' tool under the ChIP-seq section. A red arrow points to the 'HTML file' download button, which is circled in red. Another red box highlights the 'Check reverse complement' checkbox. Red text instructions provide guidance for using the tool.



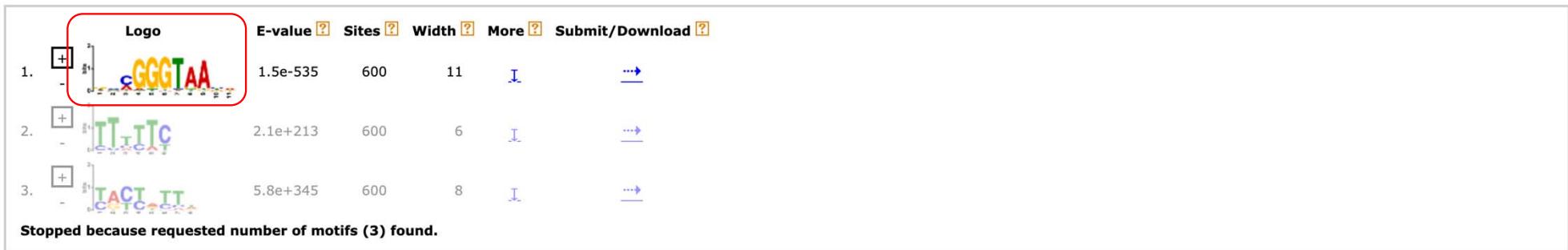
For further information on how to interpret these results or to get a copy of the MEME software please access <http://meme-suite.org>.

If you use MEME in your research, please cite the following paper:

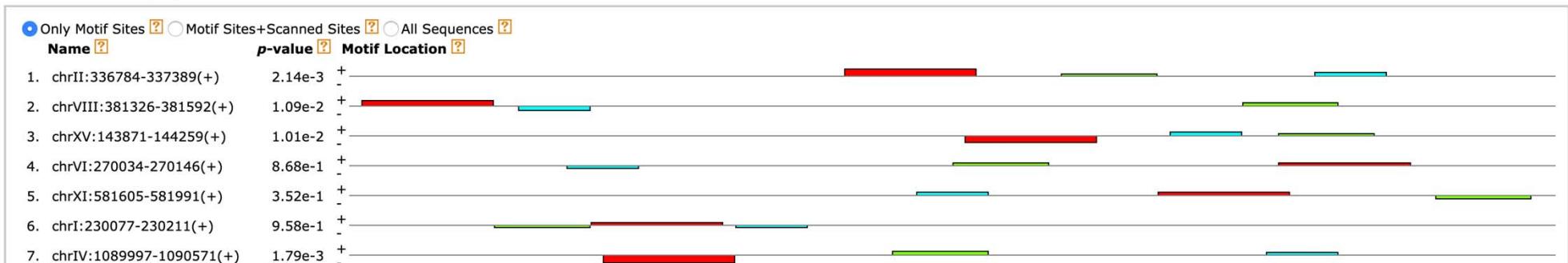
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994. [pdf]

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

DISCOVERED MOTIFS



MOTIF LOCATIONS



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,*}

¹Science for Life Laboratory, Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm 14183, Sweden

²Department of Computer Science, P.O. Box 68

³Genome-Scale Biology Program, P.O. Box 63

00014 University of Helsinki, Helsinki, Finland

⁴These 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

 NCBI

 Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO | Not logged in | Login ?

NCBI > GEO > Accession Display [?](#)

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE49402 GO

Series GSE49402 Query DataSets for GSE49402

Status	Public on Aug 15, 2013
Title	Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites
Organism	Homo sapiens
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	During cell division, transcription factors (TFs) are removed from chromatin twice, during DNA synthesis, and during condensation of chromosomes. How TFs can efficiently find their sites following these stages has been unclear. Here, we have analyzed the binding pattern of expressed TFs in human colorectal cancer cells. We find that binding of TFs is highly clustered, and that the clusters are enriched in binding motifs for several major TF classes. Strikingly, almost all clusters are formed around cohesin, and loss of cohesin decreases both DNA accessibility and binding of TFs to clusters. We show that cohesin remains bound in S phase, holding the nascent sister chromatids together at the TF cluster sites. Furthermore, cohesin remains bound to the cluster sites when TFs are evicted in early M-phase. These results suggest that cohesin binding functions as a cellular memory that promotes re-establishment of TF clusters after DNA replication and chromatin condensation.
Overall design	Examination of TF binding by ChIP-seq in LoVo CRC cell-lines.
Contributor(s)	Enge M, Yan J
Citation(s)	Yan J, Enge M, Whittington T, Dave K et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. <i>Cell</i> 2013 Aug 15;154(4):801-13. PMID: 23953112

Samples (225)			
Less...			
GSM1208590	batch1_chrom1_LoVo_ARNT_Mouse_PassedQC		
GSM1208591	batch1_chrom1_LoVo_ASCL2_Mouse_PassedQC		
GSM1208592	batch1_chrom1_LoVo_ATOH1_Rabbit_PassedQC		
GSM1208593	batch1_chrom1_LoVo_BARX2_Goat_PassedQC		
GSM1208594	batch1_chrom1_LoVo_CAMTA2_Goat_PassedQC		
GSM1208595	batch1_chrom1_LoVo_CASP8AP2_Rabbit_PassedQC		
GSM1208596	batch1_chrom1_LoVo_CASZ1_Rabbit_FailedQC		
GSM1208597	batch1_chrom1_LoVo_CDX2_Rabbit_FailedQC		
GSM1208598	batch1_chrom1_LoVo_CEBPB_Rabbit_PassedQC		
GSM1208599	batch1_chrom1_LoVo_CEBPG_Rabbit_PassedQC		
GSM1208600	batch1_chrom1_LoVo_CEBPZ_Goat_FailedQC		
GSM1208601	batch1_chrom1_LoVo_CREB3L2_Goat_FailedQC		
GSM1208602	batch1_chrom1_LoVo_CSDA_Goat_FailedQC		
GSM1208603	batch1_chrom1_LoVo_CTCF_Goat_PassedQC		
GSM1208604	batch1_chrom1_LoVo_CUX1_Rabbit_FailedQC		
GSM1208605	batch2_chrom1_LoVo_TEAD2_Rabbit_PassedQC		
GSM1208606	batch2_chrom1_LoVo_TEL_Mouse_FailedQC		
GSM1208607	batch2_chrom1_LoVo_TOX4_Goat_FailedQC		
GSM1208608	batch2_chrom1_LoVo_TP73_Rabbit_PassedQC		
GSM1208609	batch2_chrom1_LoVo_TRPS1_Goat_FailedQC		
GSM1208610	batch2_chrom1_LoVo_VEZF1_Rabbit_PassedQC		
GSM1208611	batch2_chrom1_LoVo_ZIC2_Goat_FailedQC		
GSM1208612	batch1_chrom1_LoVo_H3_Rabbit_PassedQC		
GSM1208613	batch1_chrom1_LoVo_H3K4me1_Rabbit_PassedQC		
GSM1208614	batch1_chrom1_LoVo_H3K4me3_Rabbit_PassedQC		
GSM1208615	batch1_chrom1_LoVo_IgG_Mouse		
GSM1208813	batch1_chrom1_LoVo_IgC_Rabbit		
GSM1208814	batch1_chrom1_LoVo_IgG_Goat		
Relations			
BioProject	PRJNA215231		
SRA	SRP028819		
Download family	Format		
SOFT formatted family file(s)	SOFT ?		
MINIML formatted family file(s)	MINIML ?		
Series Matrix File(s)	TXT ?		
Supplementary file	Size	Download	File type/resource
GSE49402_RAW.tar	43.7 Mb	(http://custom)	TAR (of TXT)
SRA Run Selector ?			
Raw data are available in SRA			
Processed data provided as supplementary file			

<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 DNA was added and collected using protein G Sepharose (GE). Cross-links were reversed overnight. DNA was then purified using Qiagen PCR purification kit. Library concentration of the library was determined by Nanodrop spectrophotometer 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 sonicated 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 constructed 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 altern and four to six different samples were multiplexed in one library and sequenced using Illumi

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]$ █
```

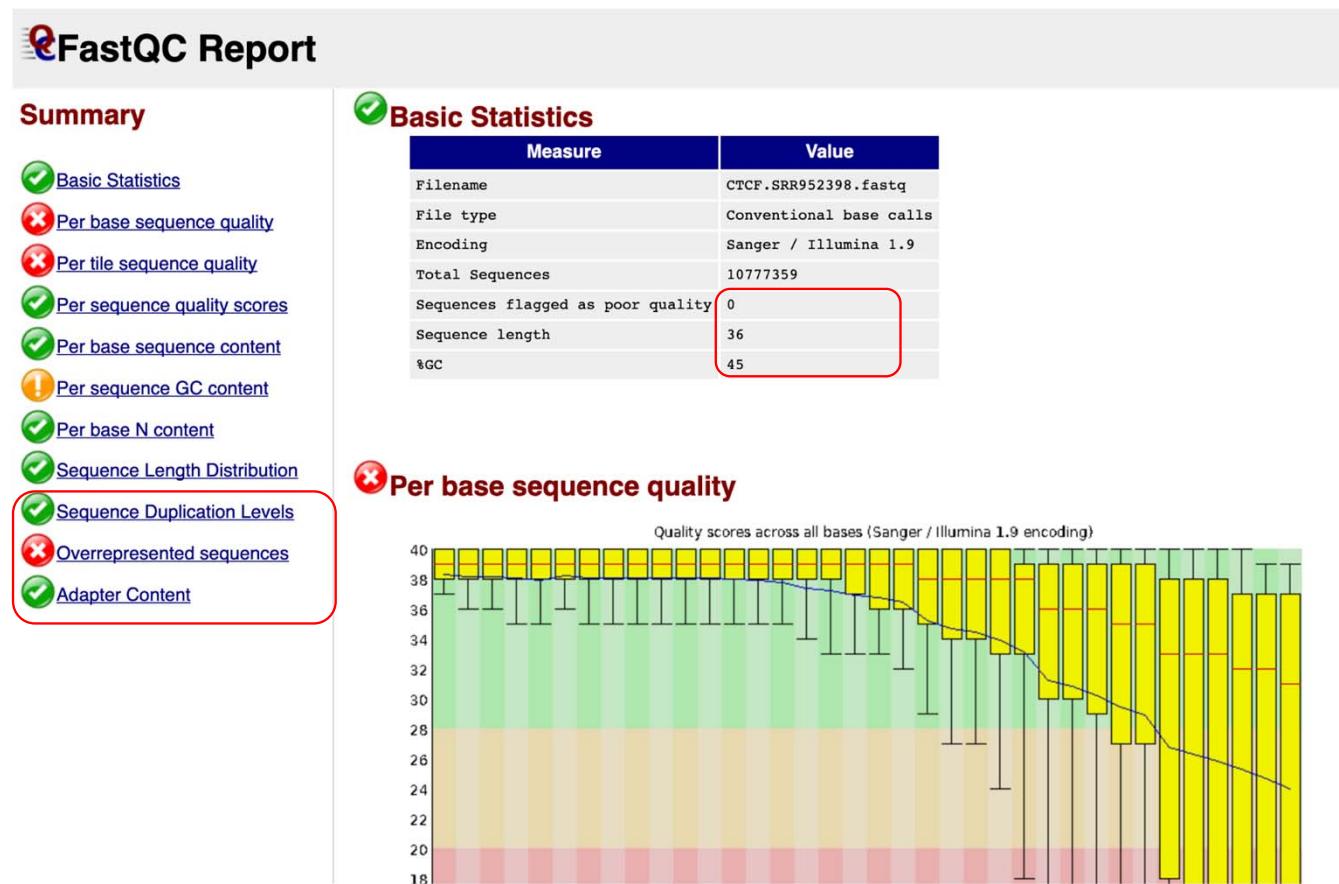
- **43,109,436/4=10,777,359** raw reads (CTCF ChIP-seq)

```
@SRR952398.1 HWUSI-EAS1683_0015_FC:8:1:1715:921 length=36
NAGAGCCTCCAGTTAAAAATTAATAAACAAAAA
+SRR952398.1 HWUSI-EAS1683_0015_FC:8:1:1715:921 length=36
#####
@SRR952398.2 HWUSI-EAS1683_0015_FC:8:1:3311:928 length=36
NGTACAATATTATGTGCAATGCTCTGATCCCTTTT
+SRR952398.2 HWUSI-EAS1683_0015_FC:8:1:3311:928 length=36
#####
@SRR952398.3 HWUSI-EAS1683_0015_FC:8:1:3469:923 length=36
NTTGTGGAAAAGTTCTTATTTTTTCCAAGTT
```

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
GAAGAGCTCGTATGCCGTCTCTGCTTGAAAAAAA	267245	2.4796891334880837	Illumina Single End Adapter 1 (100% over 28bp)
GATCGGAAGAGCTCGTATGCCGTCTCTGCTTGAAA	58142	0.5394828176364915	Illumina Single End Adapter 1 (100% over 33bp)
GAAGAGCTCGTATGCCGTCTCTGCTTGAAAAAAA	36527	0.33892347837721654	Illumina Single End Adapter 1 (100% over 28bp)
AGCTATATCGGAAGAGCTCGTATGCCGTCTCTGCT	14286	0.13255566600314603	Illumina Single End Adapter 2 (96% over 32bp)
CGGAAGAGCTCGTATGCCGTCTCTGCTTGAAAAAA	13390	0.12424194090593066	Illumina Single End Adapter 1 (100% over 30bp)

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

- Run agarose gel
- SPRI beads
- 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 standard)

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

SAM format input

BAM format output

Output to a file

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

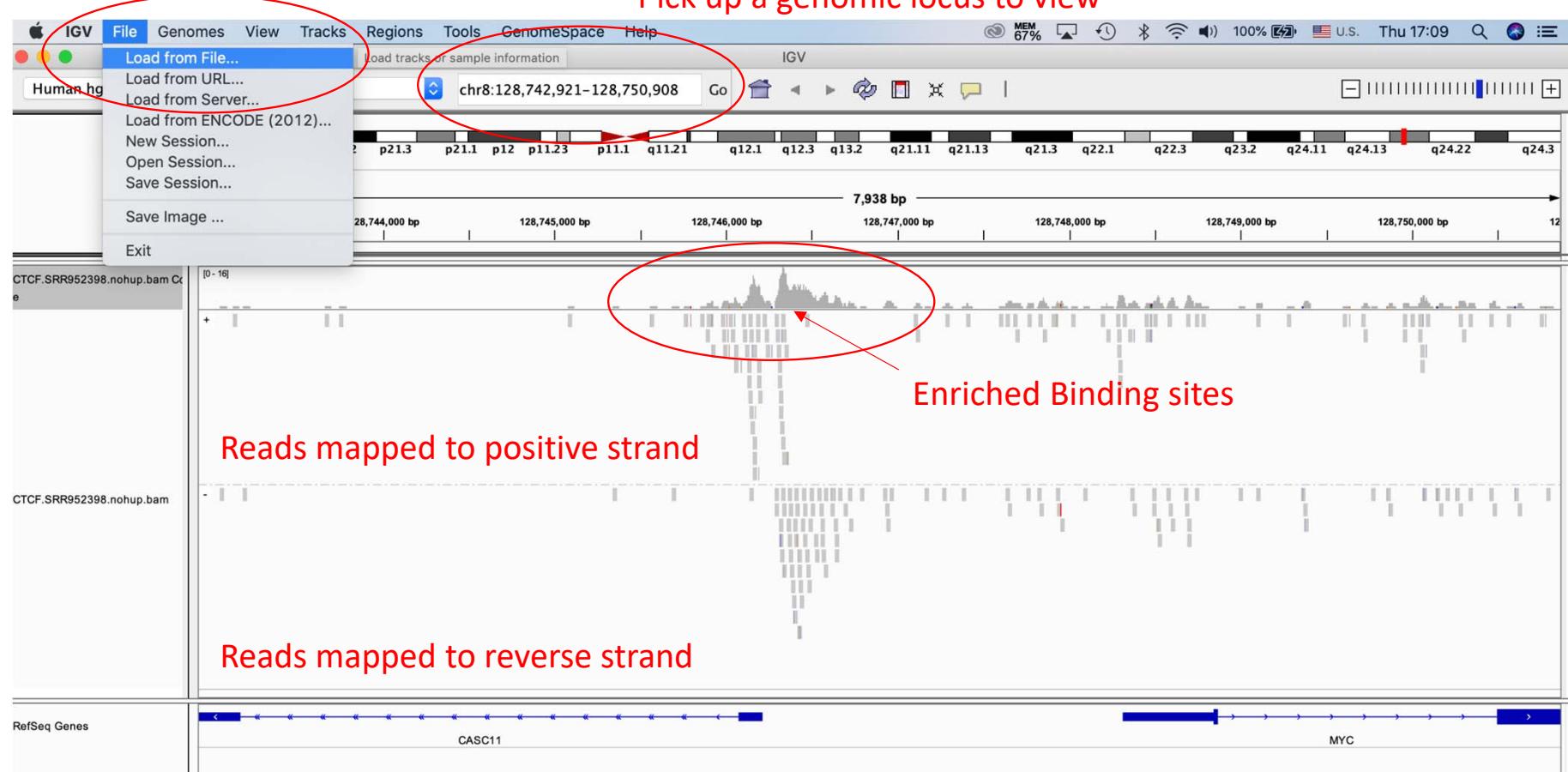
- First of all, you need to download the files to local computer where genome browser is installed or upload the file to public genome browser such as UCSC genome browser.
- For mapped bam file to be viewed in genome browser, an index file has to be generated using samtools **index**
- samtools index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

Enriched Binding sites

Reads mapped to positive strand

Reads mapped to reverse strand

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 /h 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)

```
[yanjian@localhost peaks]$ head CTCF.LoVo_peaks.narrowPeak
chr1 564289 566368 CTCF.LoVo_peak_1a 647 .
chr1 564289 566368 CTCF.LoVo_peak_1b 962 .
chr1 564289 566368 CTCF.LoVo_peak_1c 848 .
chr1 566426 567193 CTCF.LoVo_peak_2 433 .
chr1 567944 570484 CTCF.LoVo_peak_3a 577 .
chr1 567944 570484 CTCF.LoVo_peak_3b 134 .
chr1 567944 570484 CTCF.LoVo_peak_3c 927 .
chr1 714084 714587 CTCF.LoVo_peak_4 53 .
chr1 805110 805564 CTCF.LoVo_peak_5 278 .
chr1 839793 840397 CTCF.LoVo_peak_6 699 .
```

			-log10 q value	
3.51465	68.18952	64.73343	423	
4.13846	100.08997	96.26975	1051	
3.82052	88.53507	84.84323	1521	
3.68429	46.56071	43.38881	326	
3.35088	61.13650	57.76970	370	
2.50515	16.06661	13.41926	934	
4.24784	96.52946	92.74976	1694	
5.58217	7.71586 5.32231	331		
15.50602	30.77251	27.34054	249	
30.39179	73.48421	69.96432	330	

bed format peak locations

Peak ID

Fold Change (t/c) -log10 p value

Peak File (macs2 callpeak output)

```
# tag size is determined as 36 bps
# total tags in treatment: 6261926
# tags after filtering in treatment: 6261926 no. of useful reads in ChIP
# 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 no. of useful reads in control
# 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 estimated fragment size
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
```

↑ ↑
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
TGGAGGACGGGACTGACCCCTCTGGCCACTAGGGGTCTCTCTTGCTCCGCTCTGTGCCAGAGCGTCGCCAGGAGCCCTGGATCCGGAACTGGGGCAGA
>chr20:4153197-4153296
AGGACAGAGGCGGGGAGGAGGAAAGGGCCGGAGCCGCCCTGCTCCGGCGTGGCCTCATCCCCAGCGAAAGCGCGACACCCGCTGTCTGGGAAGGGGAG
>chr8:143820975-143821074
GCGGGTTCCGGGACTCGGGCGCGCCCTCTGCTGGCCGTGGGCTCAGCATGGGGGCCCTCGCAGGTGCCCTCCAGGGATAGGGGCACCTGCTG
>chr9:104211606-104211705
GGGGTGCAGGGCTAGGGCGGGGACCCGTCCCACCGGGTCACTCTCGTGGGAAGAGGGCGGGCGCTGTCCGGCAGTCAGAACACAAGCGGCTGCGAAG
>chr4:15958000-15958099
ATCATTAGCTGGGGTGGGCCTGACTTGTCTGGCAAATACAATACGTTGGCCTCTGCTGTACCTGGTGGCATTACTGGTACGCAGTGGCAAT
```

meme-suite@uw.edu

Inbox - CityU HK 18:17

M

MEME Submission Information (job appMEME_5.0.51570702630261-471136349)

To: Dr. YAN Jian,

Reply-To: meme-suite@uw.edu



Multiple Em for Motif Elicitation

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

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

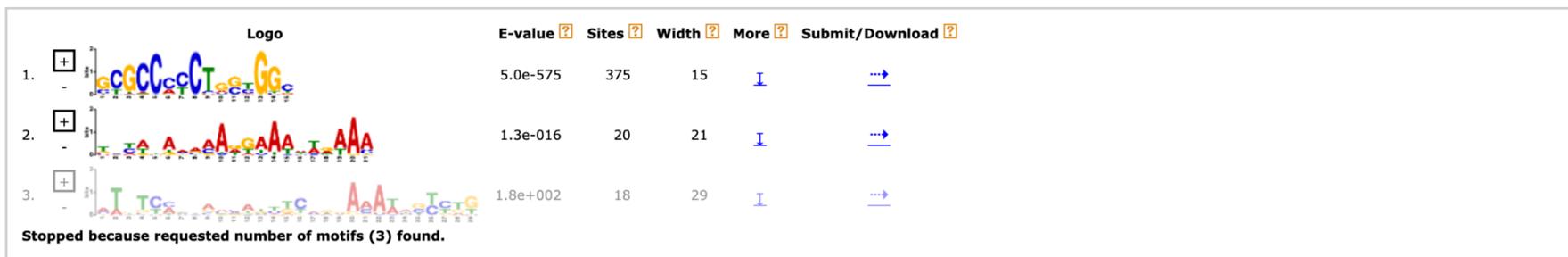
You can view your job results at: http://meme-suite.org/info/status?service=MEME&id=appMEME_5.0.51570702630261-471136349

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 <code>CTCF.LoVo_summits.top500.fa</code> .
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



MOTIF LOCATIONS

