

Chromatin Immunoprecipitation-sequencing (ChIP-seq) Data Analysis

Weihong Yan

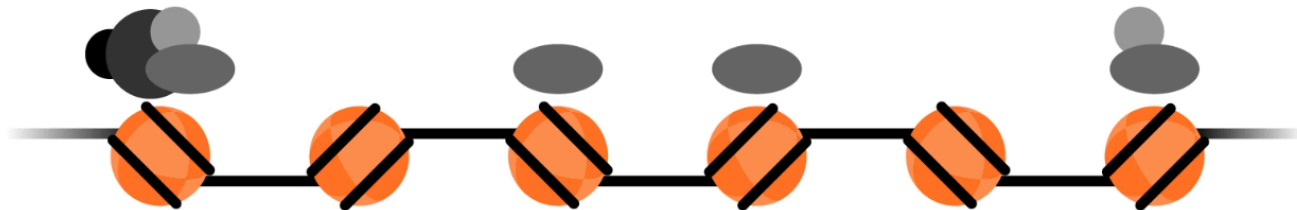
wyan@chem.ucla.edu

ChIP-seq Data Analysis

- ❑ Day 1: ChIP-seq Background and Concept
 - ChIP-seq Protocol
 - Quality Control and Guidelines
 - ChIP-seq data from sequence read archive (SRA)
- ❑ Day 2: ChIP-seq data analysis workflow
 - Bowtie2 Alignment
 - MACS2 Peak calling
 - IGV genome visualization
- ❑ Day 3: Peak Annotation and Functional Analysis
 - CEAS: Cis-regulatory element annotation system
 - BEDTools for genome arithmetic
 - HOMER peak annotation, functional and motif analysis
 - GREAT

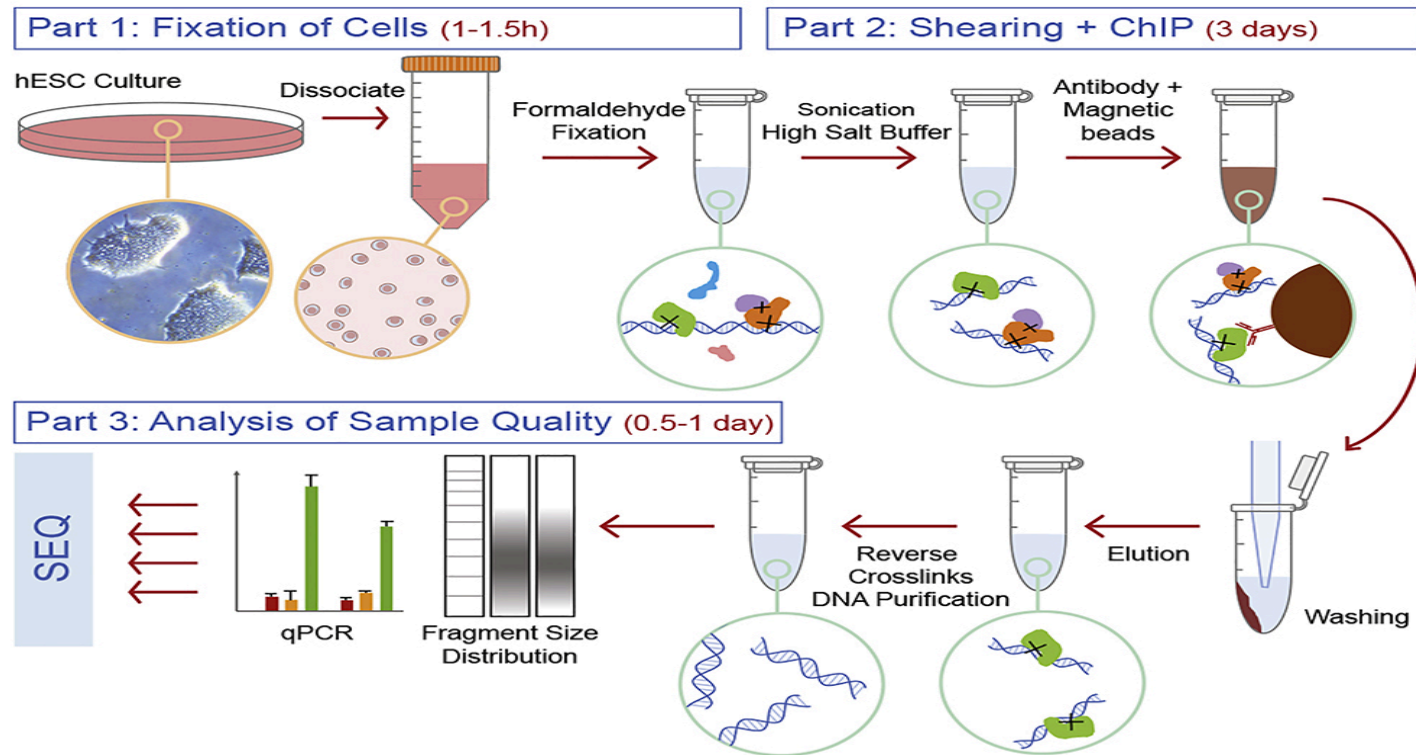
What is ChIP-seq

- Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a powerful method for identifying genome-wide DNA binding sites for histones, transcription factors and other proteins.
- Chromatin is a complex of nucleic acids and proteins (histones, transcription factors and other proteins).
- ChIP-Seq technique makes its feasible to exam the interactions between proteins and nucleic acids on a genome-wide scale and reveals insights into gene regulation events that play critical roles in biological pathways and diseases.



How does ChIP-seq work

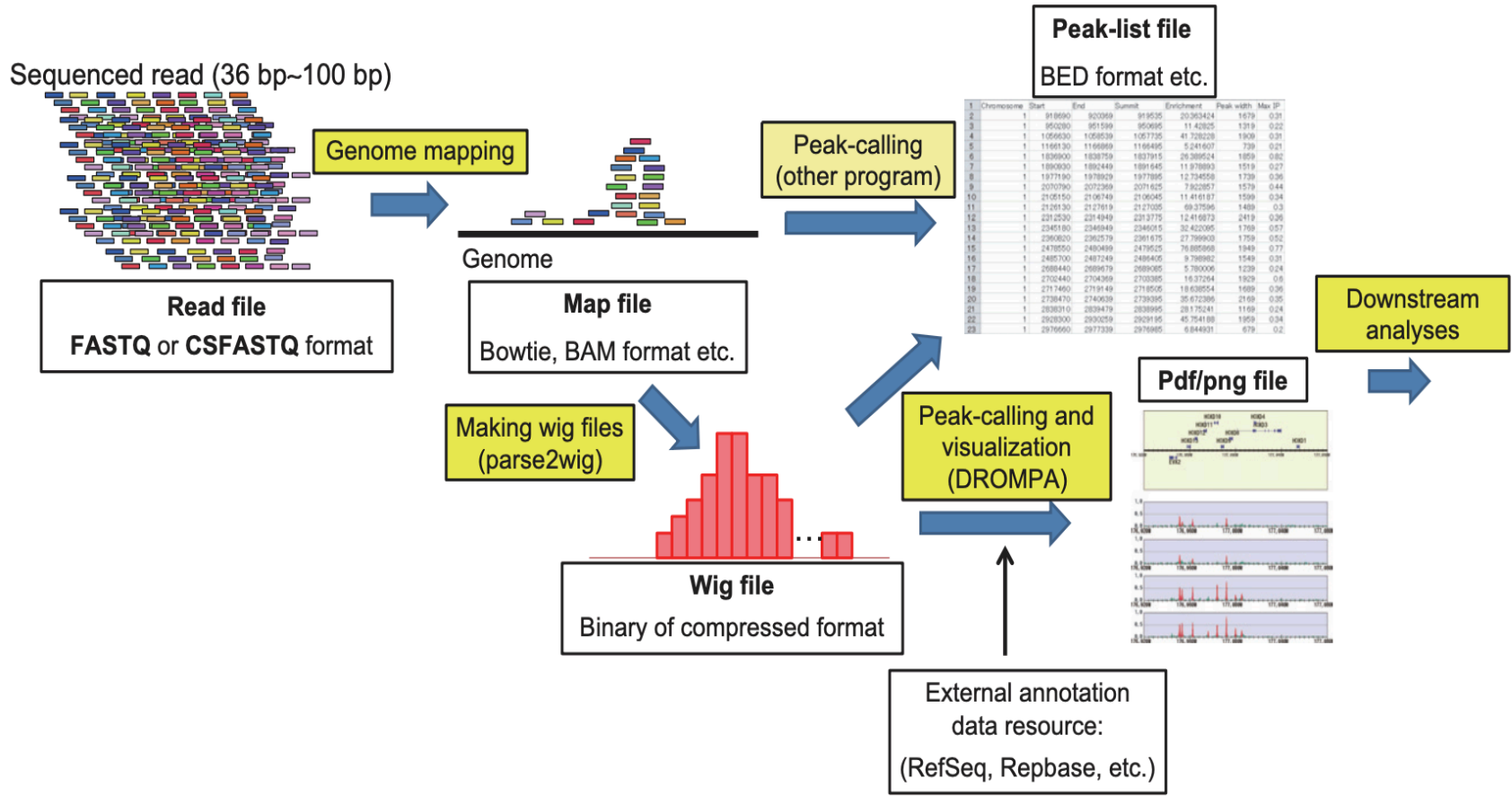
Experimental Protocol



<https://star-protocols.cell.com/protocols/110>

How does ChIP-seq work

Sequencing Data Analysis Pipeline



ChIP-seq vs ChIP-chip

Table1: ChIP-Seq Compared to ChIP-ChIP Analysis

	ChIP-Seq	ChIP-chip	ChIP-Seq Advantage
Starting material	Low: Down to 10 ng	4 µg	Hundreds-fold lower DNA input requirements means fewer IP reactions
Flexibility	Yes: Genome-wide assay of any sequenced organism	Limited: Dependent on available products	Not limited to content available on arrays
Positional resolution	± 50bp	± 500–1000bp	Site mapping can be an order of magnitude more precise
Sensitivity	Widely tunable: Increase counts to increase sensitivity	Poor: Based on hybridization and ratios	Simply increase the number of counts to obtain desired sensitivity
Cross-hybridization	None: Each DNA is individually sequenced	Significant	Higher quality data even in complex genomes

ChIP-seq Experimental Considerations

Biases and Artifacts from Sample Preparation and Data Processing

- Crosslink/fixation: influence efficiency of fragmentation and binding of antibody to its protein target.
- Sonication: open chromatin regions more easily sheared than other regions. Over sonication can lead to loss of protein-DNA interactions and cause protein degradation.
- Antibody: binding specificity to its target and non-specific binding to other proteins.
- PCR biases: GC-rich fragments, adapter dimers
- Library complexity: same DNA fragments sequenced repeatedly in low complexity library
- Sequencing depth
- Read mapping: repetitive elements, duplications of genomic sequences, mapping algorithms
- Peak calling methods

ChIP-seq Quality Control Guidelines

- Antibody Specificity
- Input controls
- Data quality assessment
 - Sequencing quality (fastqc)
 - Library complexity
 - Sequencing depth
 - Peak identification
 - Visual inspection
 - Global ChIP enrichment score
 - Cross-correlation analysis

Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

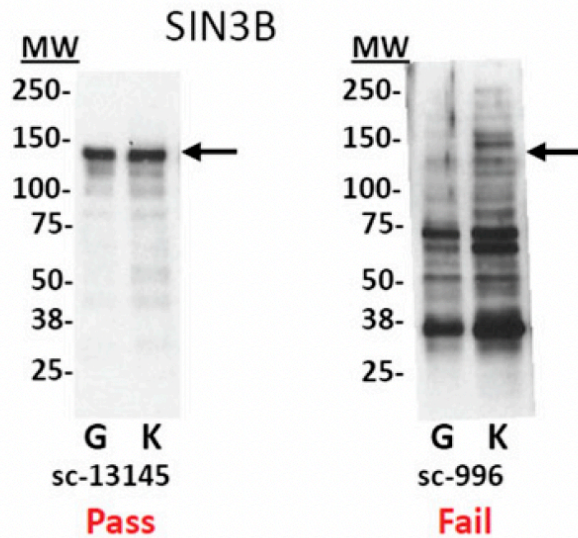
R Nakat, T Sakata, Methods, 2021

C.A Meyer and X Shirley Liu, Nature review genetics, 2014

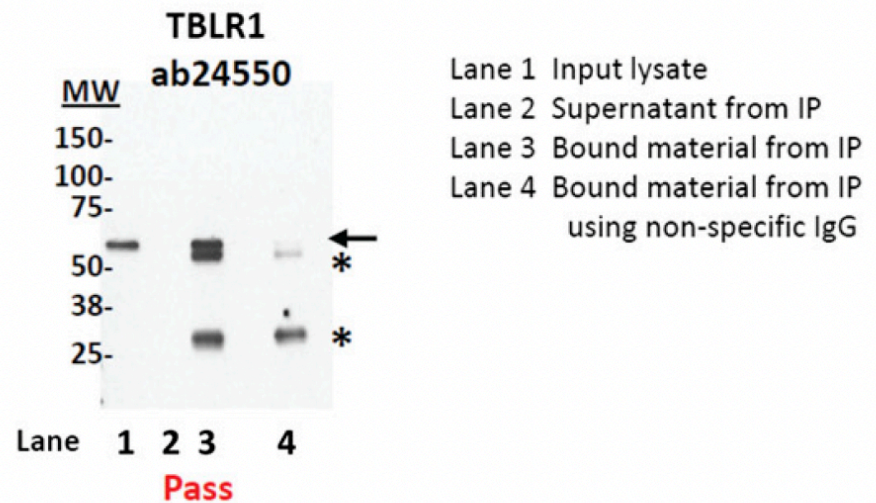
S.G Landt, et al, Genome research, 2012

Antibody Specificity

A Immunoblot assay



B Immunoprecipitation (IP) assay



ChIP-seq Input Controls

- Input DNA

The DNA sample that has been cross-linked and sonicated but not immunoprecipitated

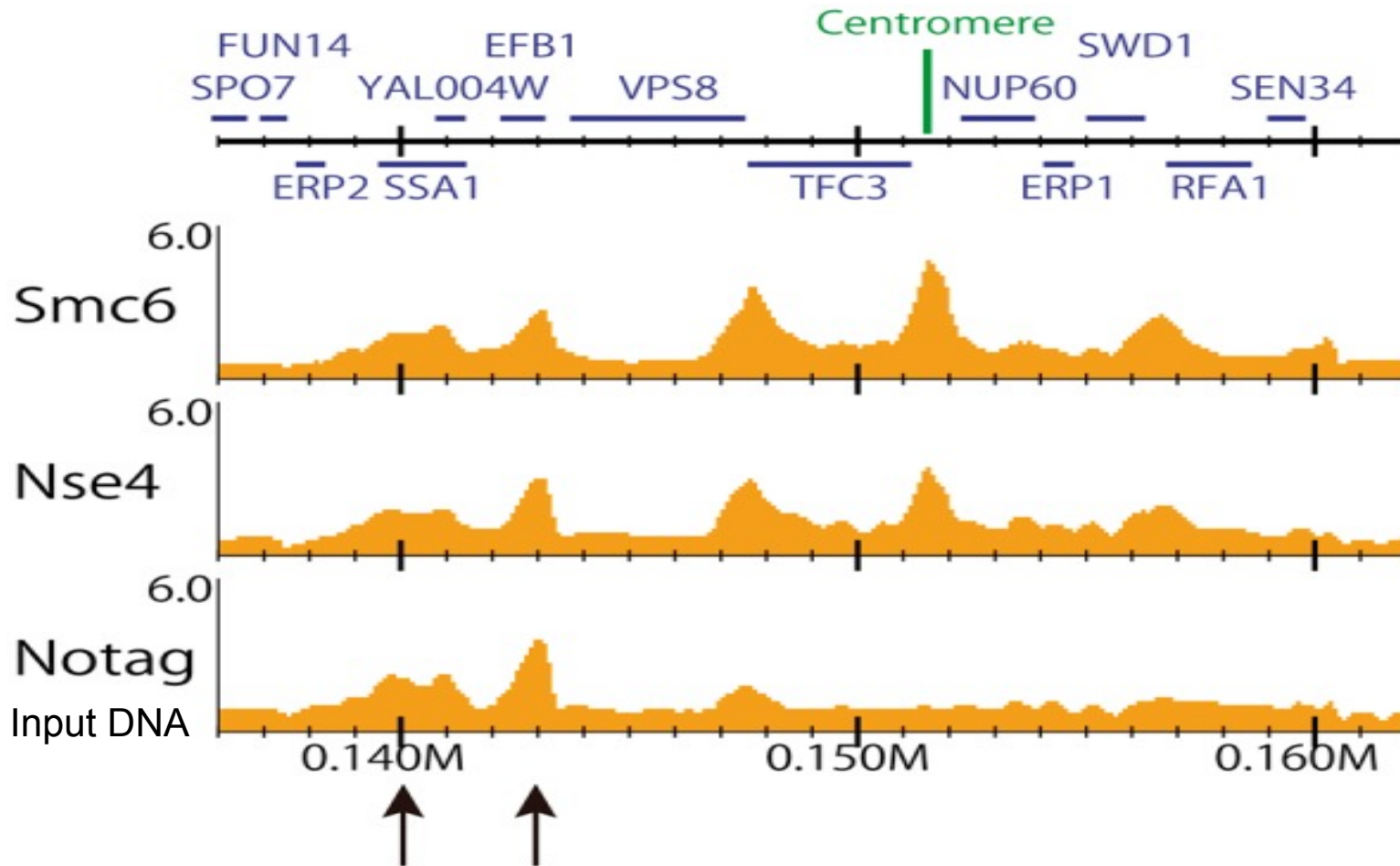
Benefit: Normalize biases introduced in the processes of crosslink, sonication, PCR, Sequencing, read mapping and peak identification.

- Mock IP (IgG)

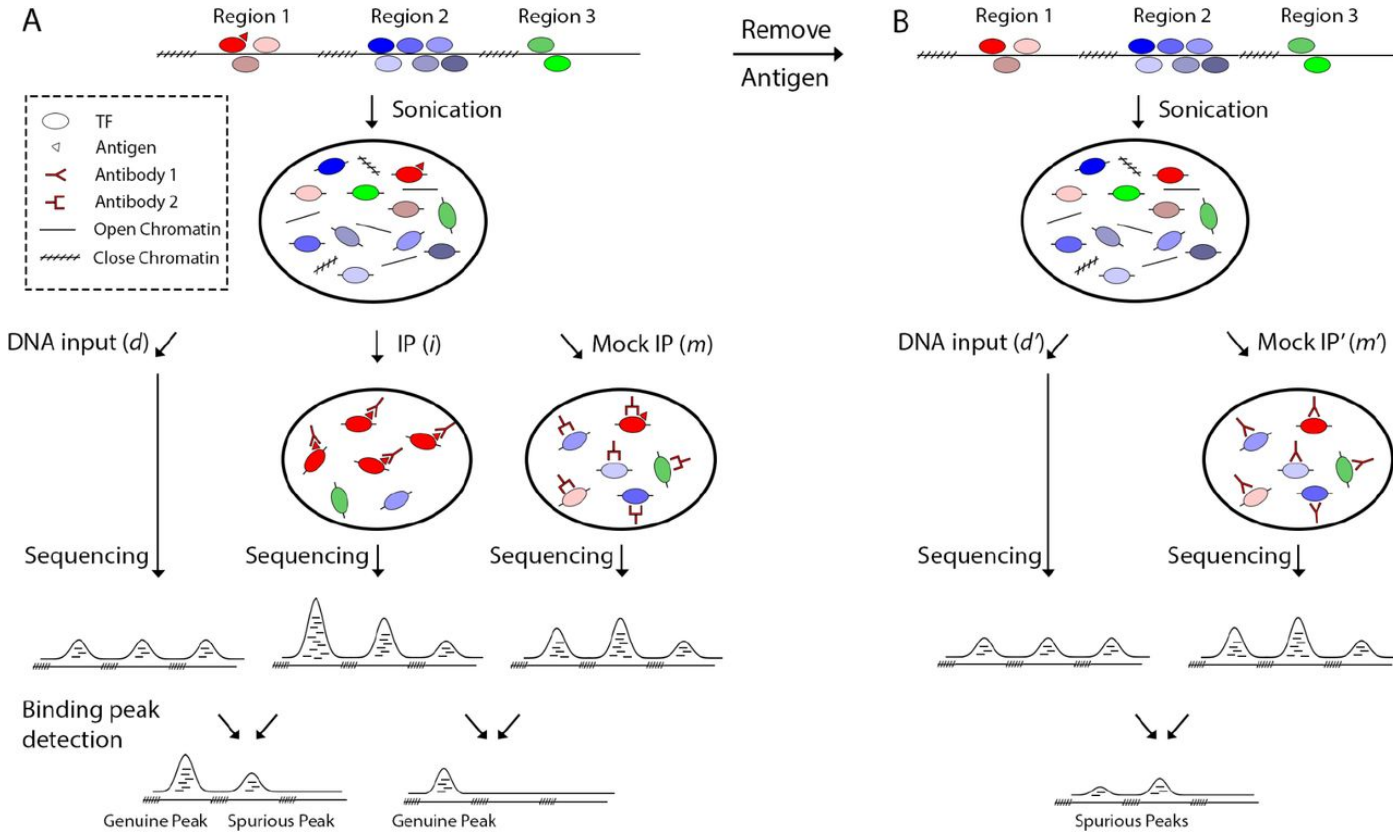
IgG antibody used for immunoprecipitation

Benefit: Normalize crosslinking bias and antibody nonspecificity.

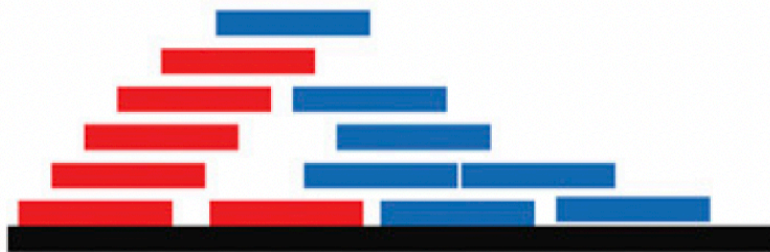
ChIP-seq Input Control



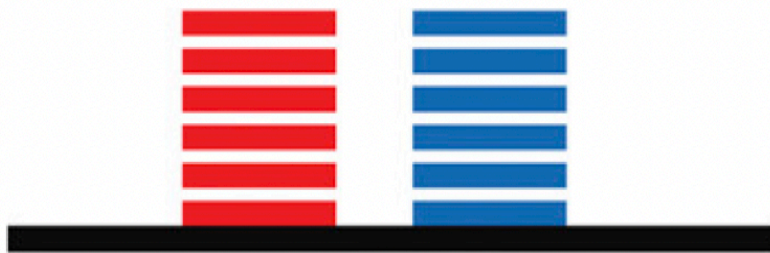
ChIP-seq Input Controls



Library Complexity



Typical ChIP-seq peak



Low-complexity ChIP-seq peak

Sources of low complexity:

- Low amount of DNA library from IP
- PCR bias
- Sequencing bias
- Adapter dimers

Adapter Dimers

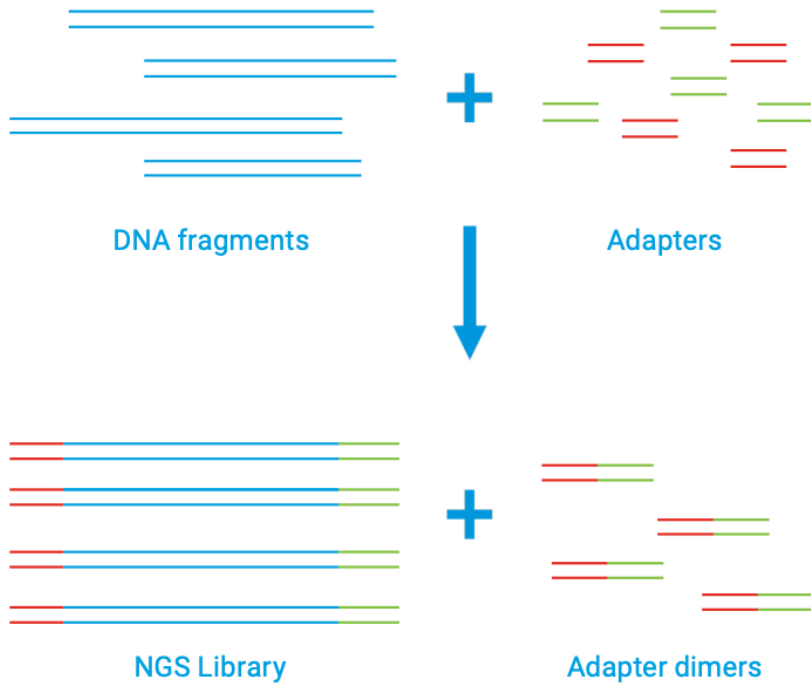
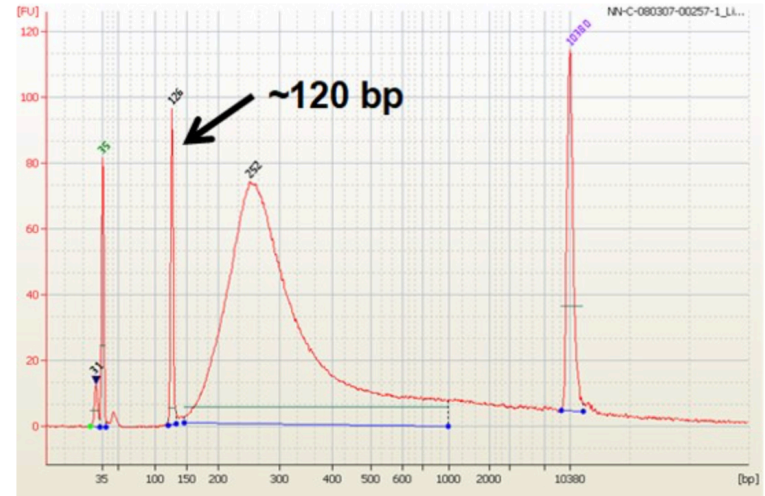


Figure 1. During the process of NGS library preparation, known DNA adapter sequences are ligated to the 5' and 3' ends of the DNA. Generally, one adapter will contain the primer sequence, while the other is used to bind the library to the flow cell for sequencing. Adapter dimers form when the two adapters ligate to each other instead of the target insert.

What do adapter dimers look like?



If adapter dimers present in the library, clean-up the library with solid-phase reversible immobilization (SPRI) beads or gel purification

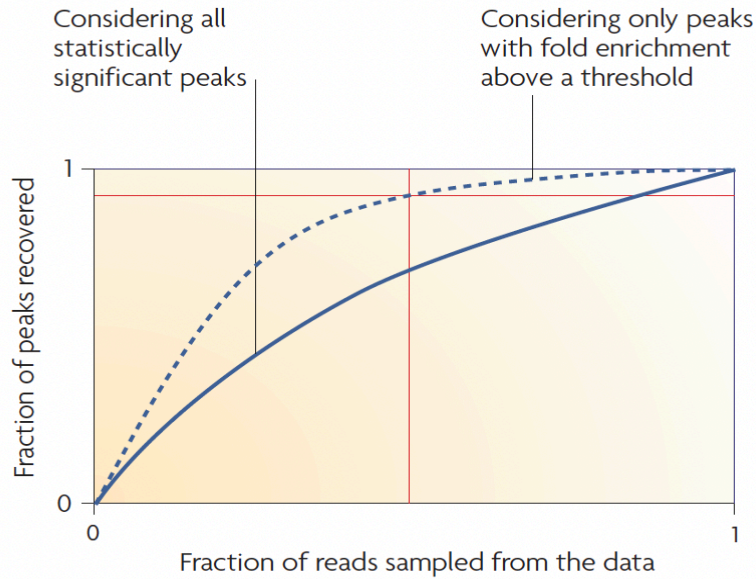
Guideline for the Library Complexity

Non-redundant fraction(NRF): the ratio between the number of non-redundant reads and the total number of unique mapped reads.

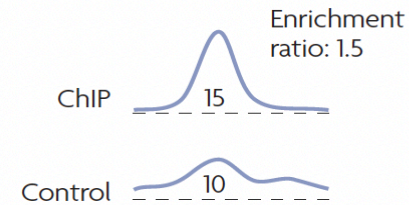
NRF	Complexity
$NRF < 0.5$	concerning
0.5	acceptable
$0.5 \leq NRF \leq 0.9$	compliant
$NRF > 0.9$	Ideal

Sequencing Depth

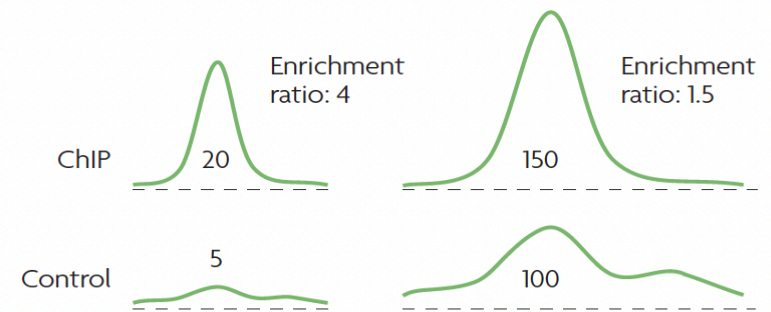
A



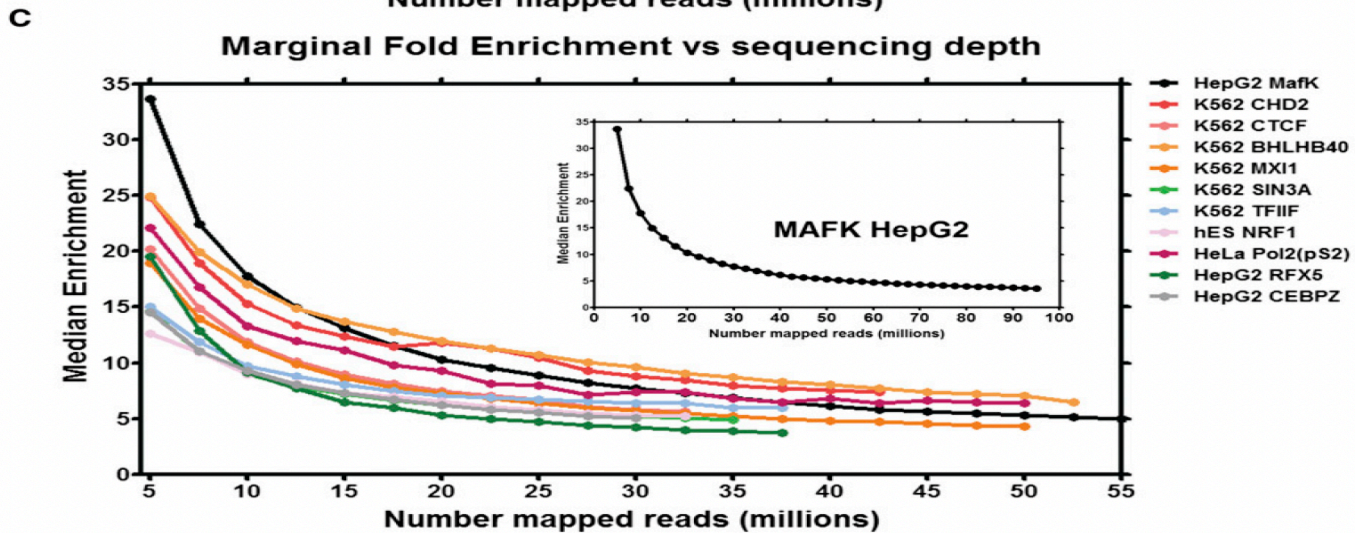
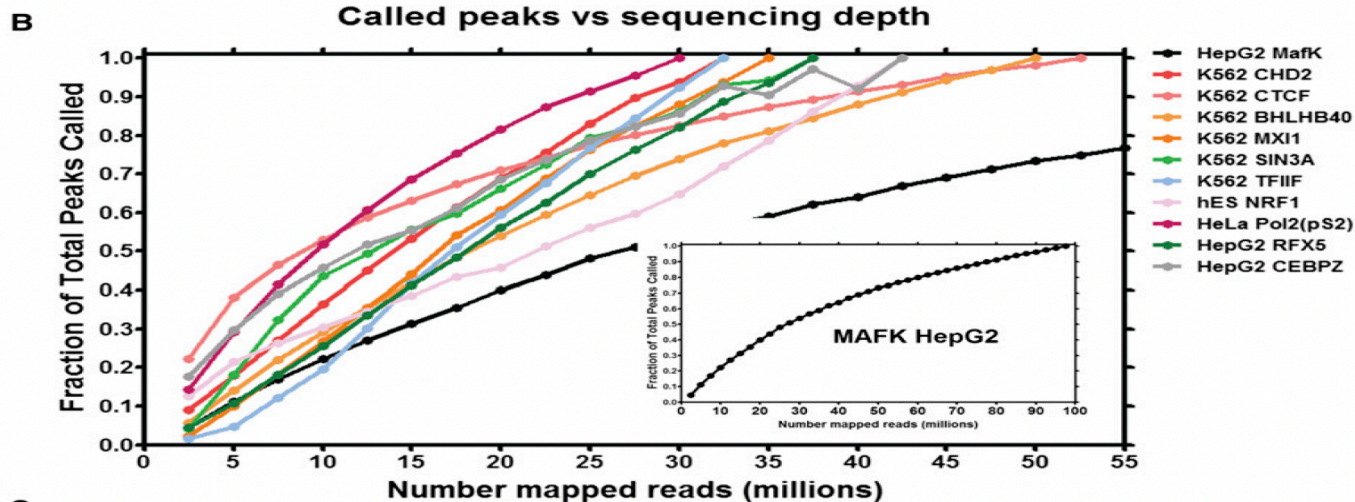
Ba Not statistically significant



Bb Statistically significant

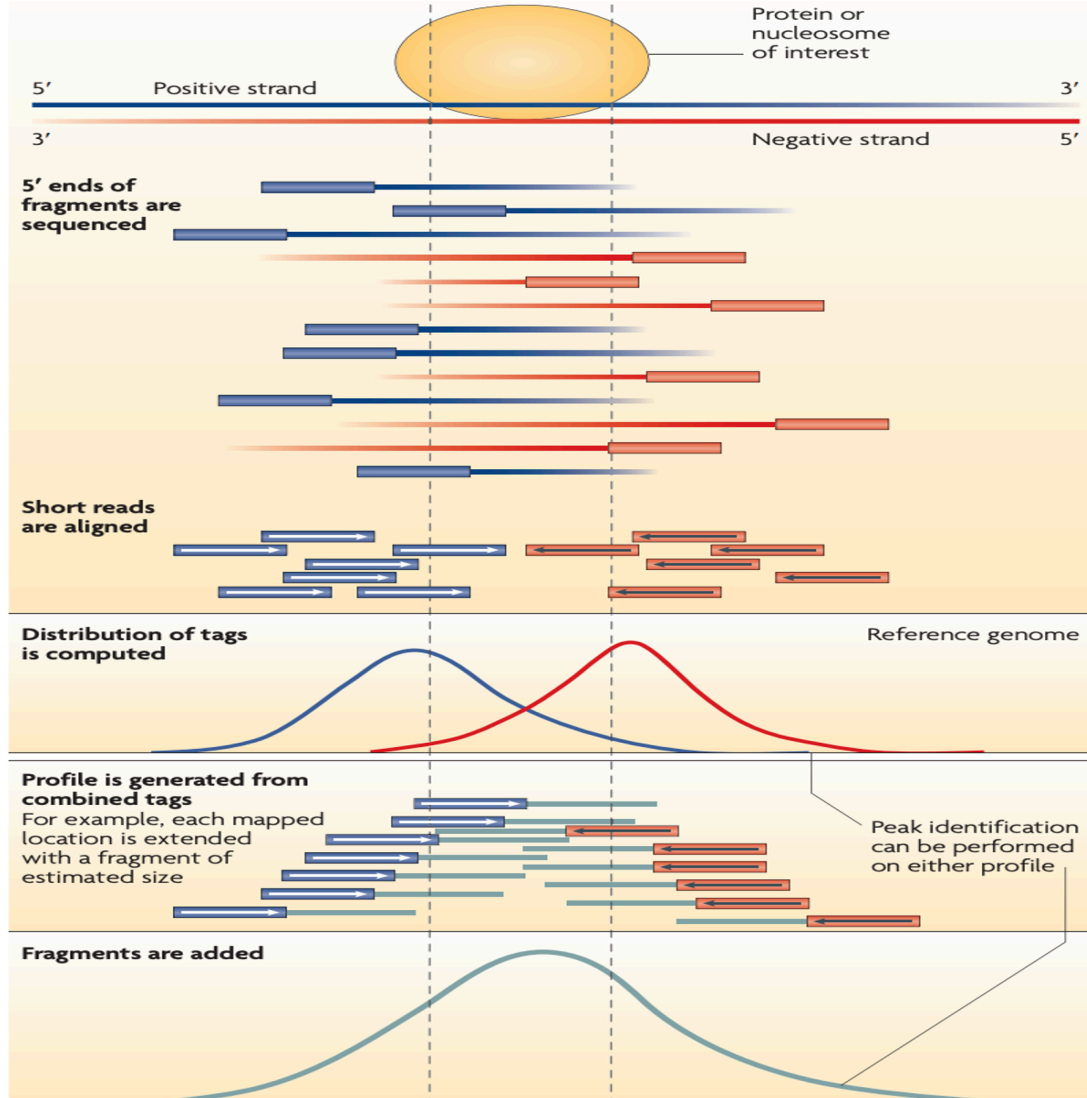


Sequencing Depth



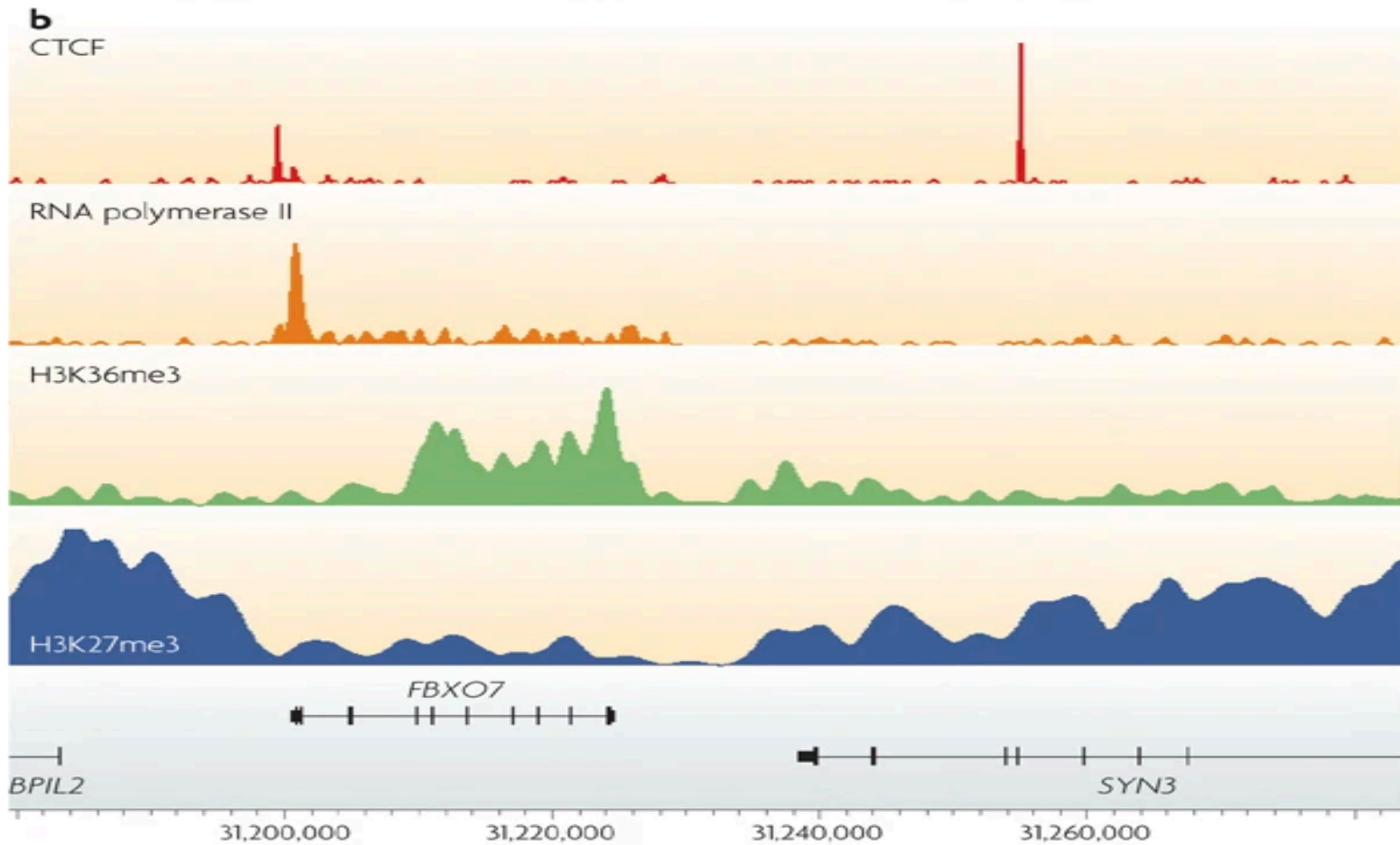
Encode guideline: For point-source library (transcription factor), ≥ 10 million uniquely mapping reads with $\text{NRF} \geq 0.8$. For broad-source library (histone), ≥ 20 million uniquely mapping reads with $\text{NRF} \geq 0.8$.

Peak Identification



Peak calling results are represented in BED format

Types of Enrichment Profiles



Nature Reviews | **Genetics**

CTCF: sharp binding sites

RNA polymerase II: sharp peak plus a broad region

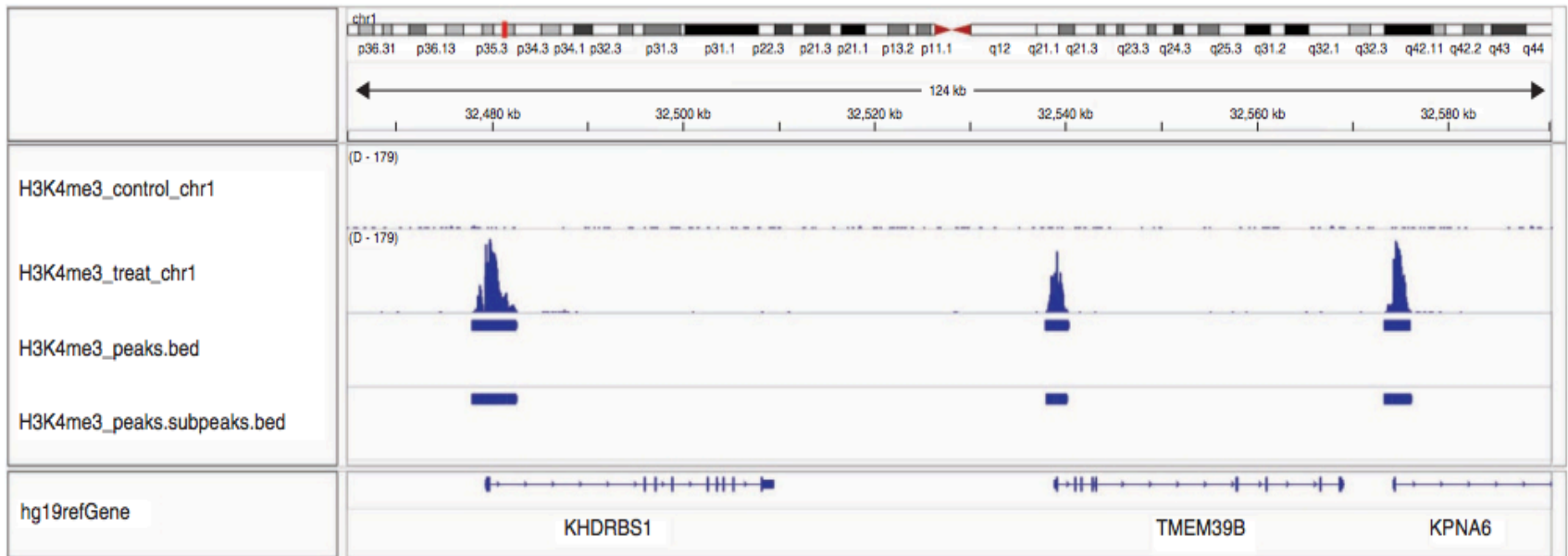
H3K36me3: medium size of peak

PJ Park, 2009

Peak Quality Assessment

- Visualization of ChIP-seq data using genome viewer
- Fraction of reads in peaks (FRiP) measurement
- Cross-correlation analysis
- Reproducibility evaluation

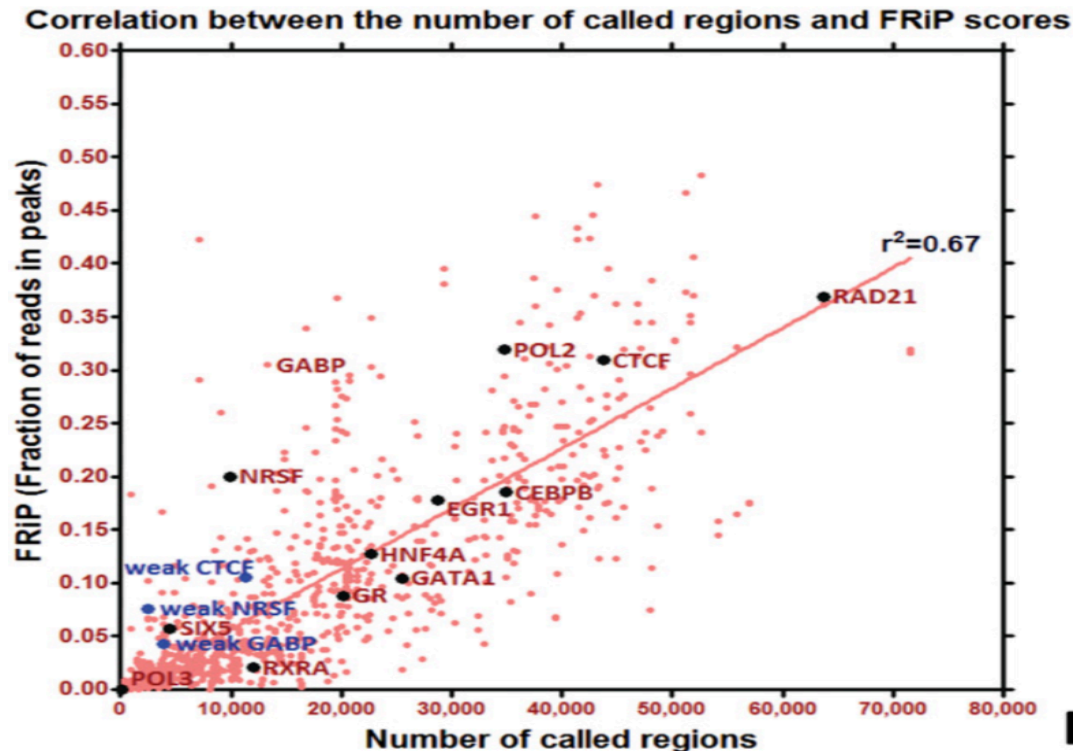
Visualization of Peak Calling Results



Integrative Genomics Viewer (IGV): <https://igv.org/doc/desktop/>

Fraction of Reads in Peaks (FRiP)

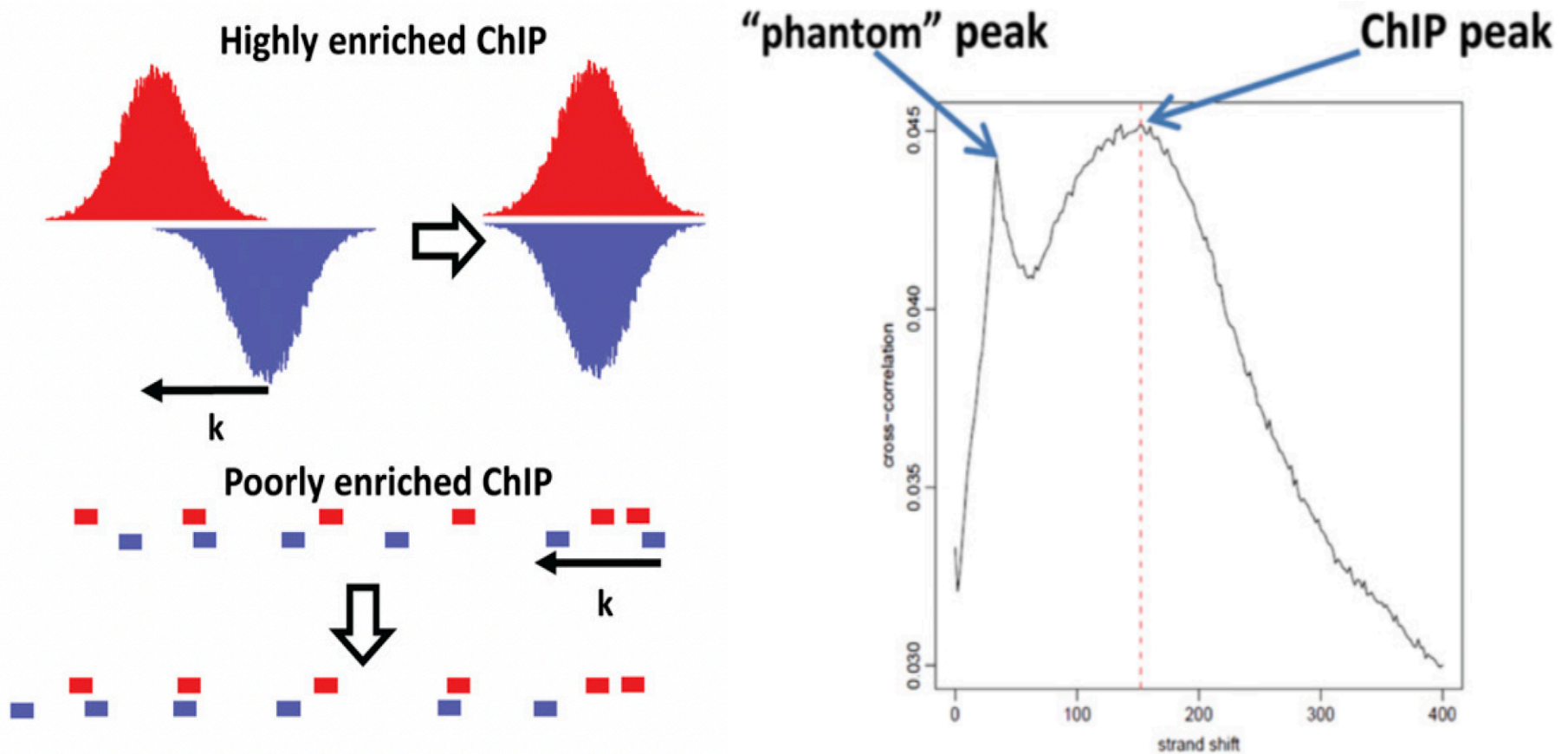
FRiP = number of reads under peaks/total reads



FRiP correlates linearly with the number of called peaks.

FRiP guideline: ≥ 0.01 for the experiments with more than thousands called peaks in large mammalian genome

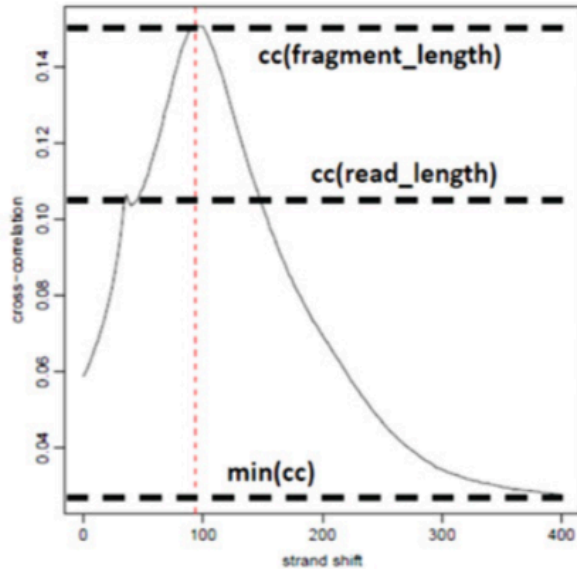
Strand Cross-correlation Analysis



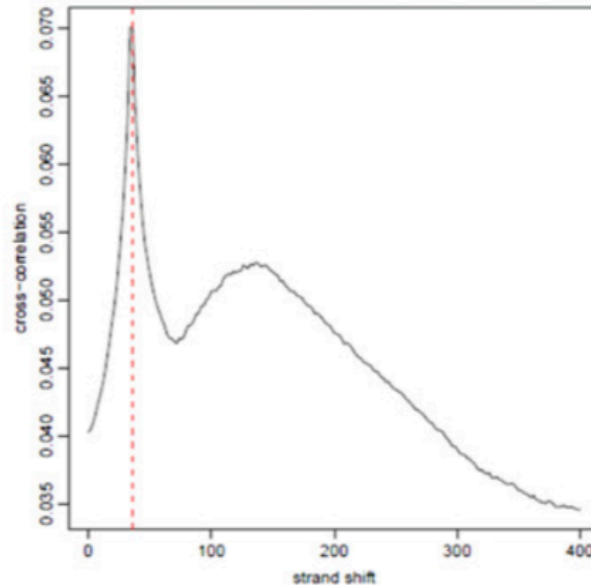
Cross-correlation is computed as the Pearson linear correlation between the Watson strand and the Crick strand after shifting Watson strand by k base pairs. The plot contains two peaks: “phantom” peak at the location matching read length and CHIP peak at the location of matching predominant fragment length.

Strand Cross-correlation Analysis

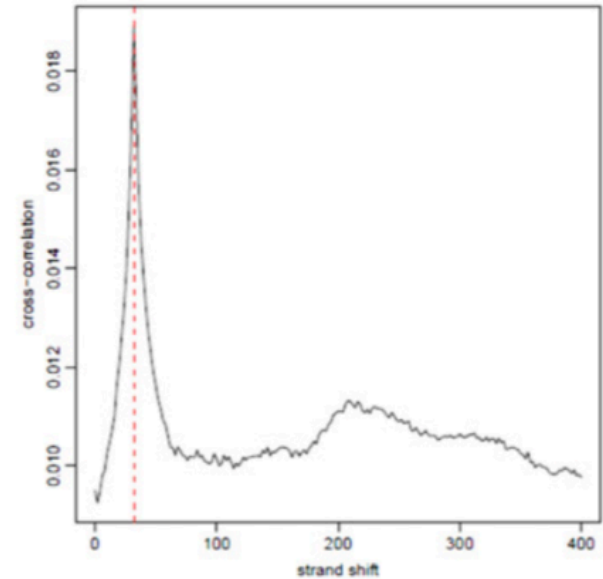
Successful



Marginal



Failed

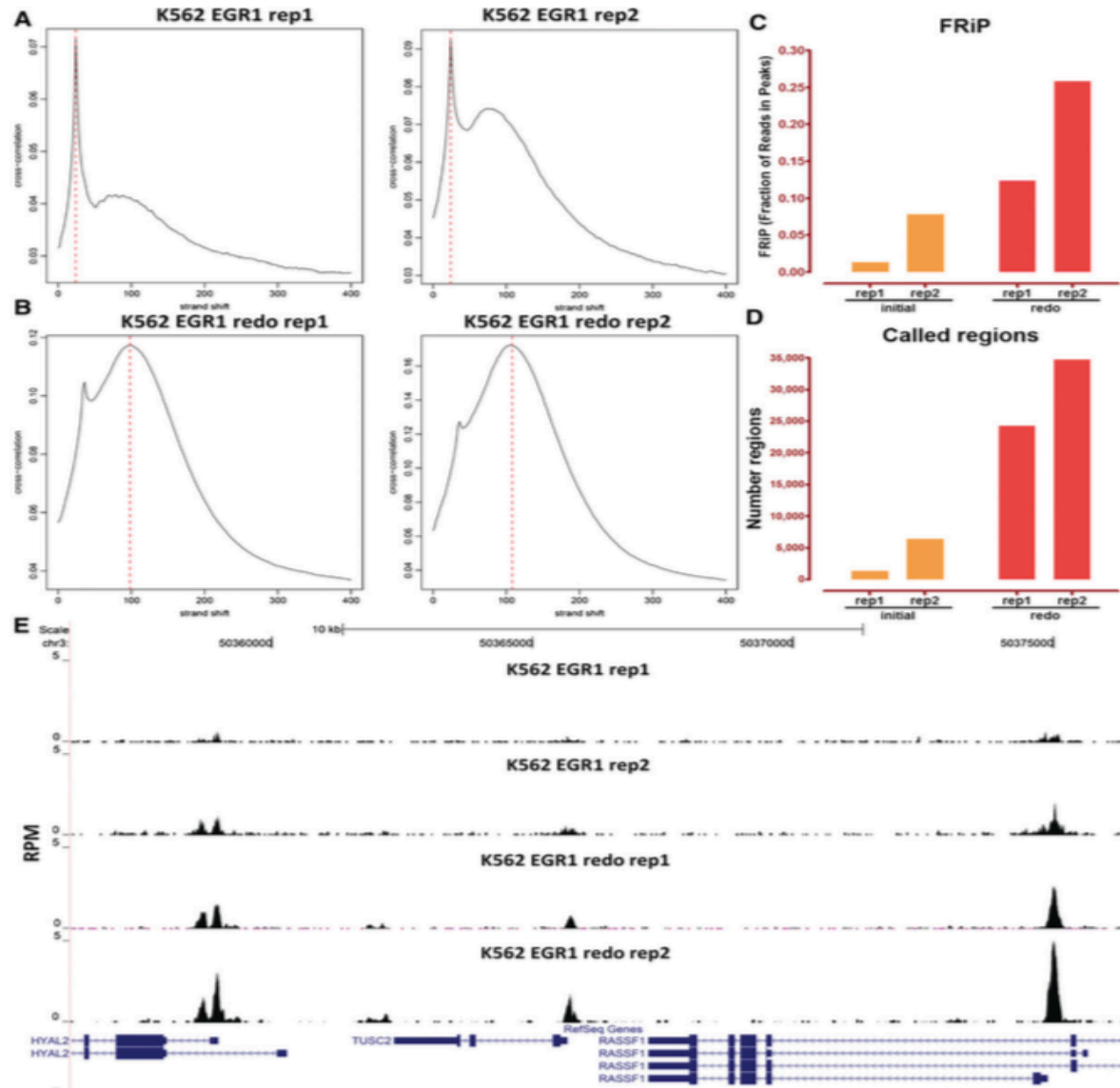


$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

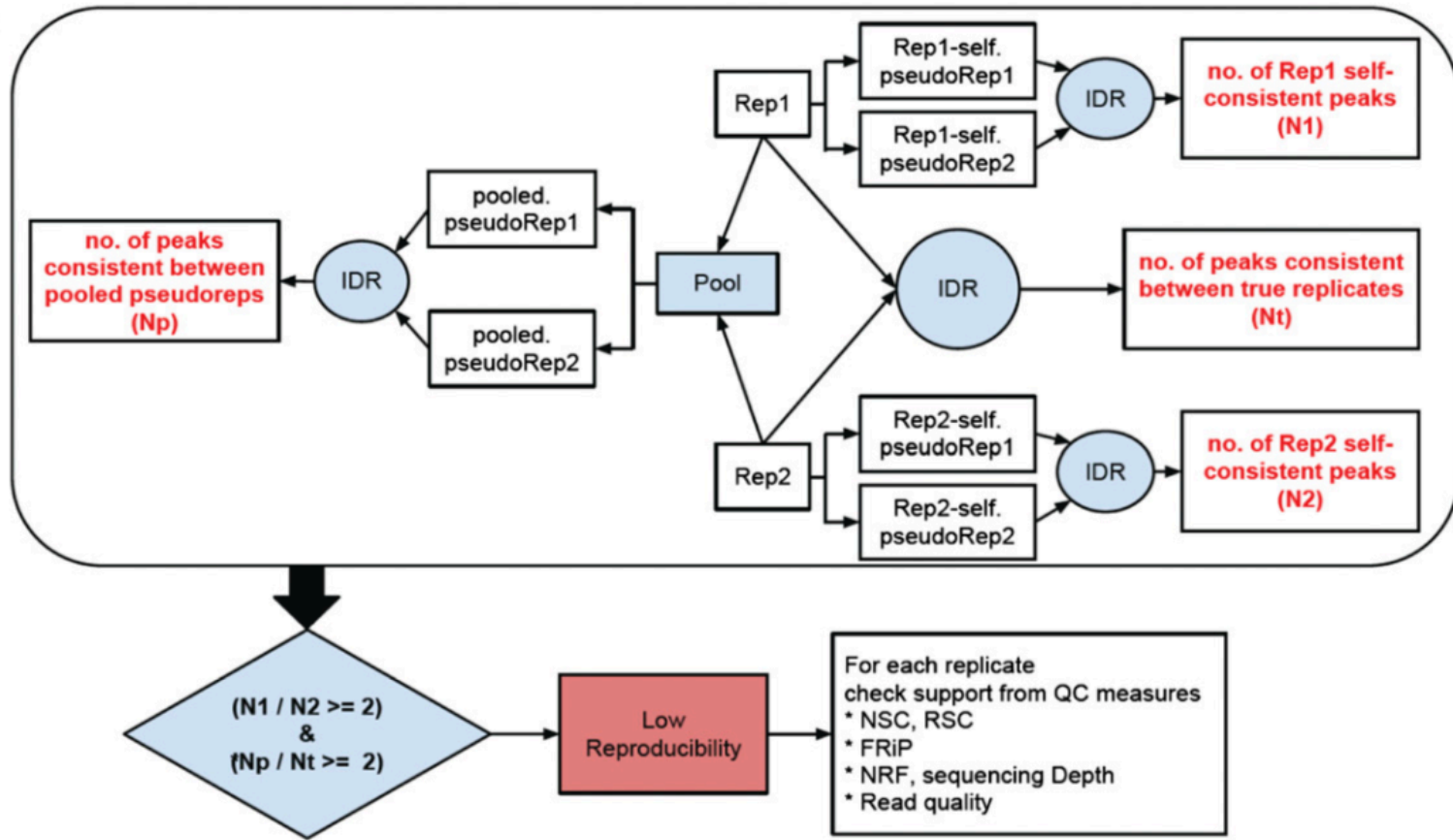
$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Encode Guideline: Repeat experiments with NSC (normalized strand coefficient) less than 1.05 and RSC (relative strand correlation) less than 0.8

An Example of Peaks Assessment



Irreproducible Discovery Rate (IDR)



<https://github.com/nboley/idr>

Np: Peak consistency between pooled pseudo-replicates

Nt: Peak consistency between true replicates

N1 and N2: Peak self-consistency for each individual replicates

Tools for Quality Assessment of ChIP

- Phantompeakqualtools

<https://github.com/kundajelab/phantompeakqualtools>

- Homer

<http://homer.ucsd.edu/homer/>

Available on hoffman2 cluster

- ChIPQC

<https://bioconductor.org/packages/release/bioc/html/ChIPQC.html>

- IDR

<https://github.com/nboley/idr>

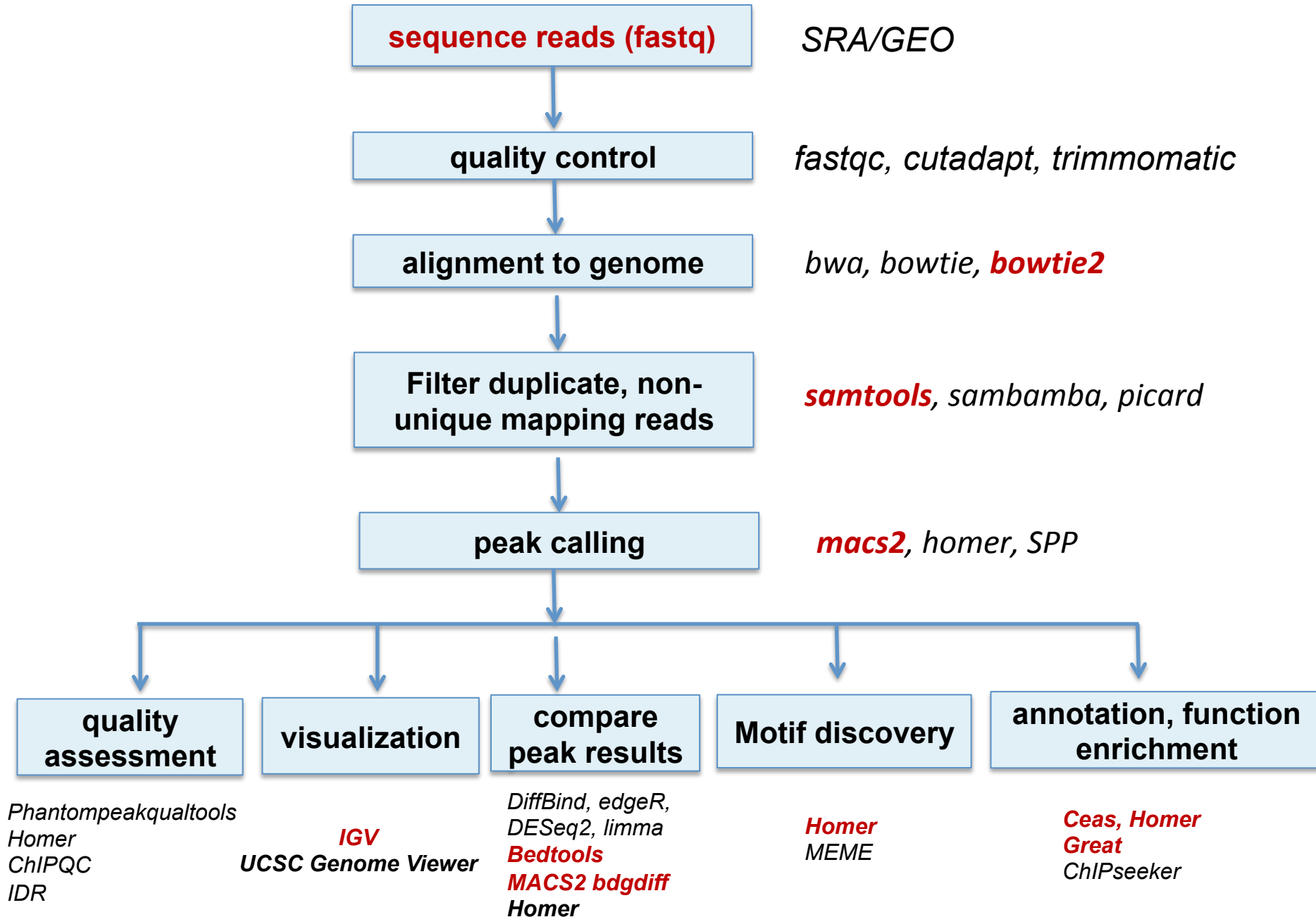
Public ChIP-seq Databases

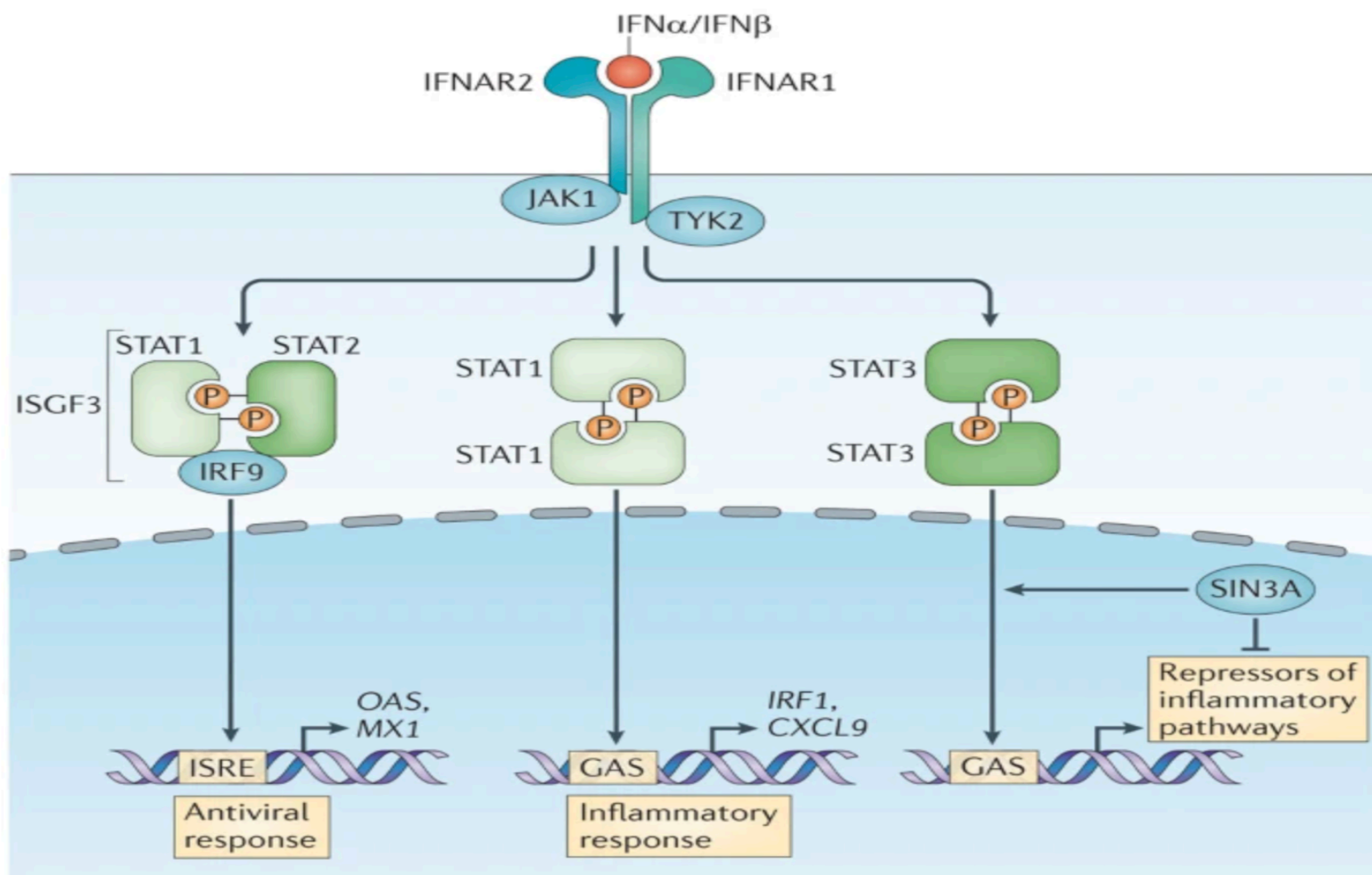
ENCODE portal	https://www.encodeproject.org/
ROADMAP epigenome database	http://www.roadmapepigenomics.org/
IHEC Data Portal	https://epigenomesportal.ca/ihec/
Epigenome database for human endothelial cells	https://rnakato.github.io/HumanEndothelialEpigenome/

<https://www.encodeproject.org/data-standards/chip-seq/>

SG Landt, et al. ChIP-seq guidelines and practices of the Encode and modEncode consortia. Genome research.

ChIP-seq Data Analysis Workflow





- IFN α activates Janus Kinase 1 (JAK1) and tyrosine kinase 2 (TYK2)
- Phosphorylation, dimerization and nuclear translocation of the signal transducer and activator of transcription (STAT) proteins
- STAT1 homodimers bind to gamma-activated sequences (GASs) to induce pro-inflammatory response

Hypothesis: IFN α activates STAT1 and enhances STAT1 bind to ISRE and GAS promoter elements

ChIP-seq Data Analysis of IFN α Induced STAT1 Binding Sites

- Obtain ChIP-seq read files from NCBI sequence read archive (SRA) database
- Bowtie2 alignment to human genome
- MACS2 for peak identification and comparison
- IGV to examine STAT1 peak regions on interferon induced protein with tetratricopeptide repeats (IFIT)
- Annotation and Functional analysis of STAT1 peaks
 - CEAS
 - HOMER
 - GREAT


NCBI Sequence Read Archive

https://www.ncbi.nlm.nih.gov/sra

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information Log in

SRA [Advanced](#) [Help](#)



SRA - Now available on the cloud

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started

- [Documentation](#)
- [How to submit](#)
- [How to search and download](#)
- [How to use SRA in the cloud](#)
- [Submit to SRA](#)

Tools and Software

- [Download SRA Toolkit](#)
- [SRA Toolkit Documentation](#)
- [SRA-BLAST](#)
- [SRA Run Browser](#)
- [SRA Run Selector](#)

Related Resources

- [Submission Portal](#)
- [dbGaP Home](#)
- [BioProject](#)
- [BioSample](#)

SRA [Create alert](#) [Advanced](#) [Help](#)

Summary ▾ 20 per page ▾

Send to: ▾ [Filters: Manage Filters](#)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Search results

Items: 1 to 20 of 425

<< First < Prev Page 1 of 22 Next > Last >>

[GSM1003634: Stanford_ChipSeq_GM12878_JunD_IgG-rab](#)

1. 2 ILLUMINA (Illumina Genome Analyzer) runs: 46.6M spots, 1.7G bases, 806.8Mb downloads
Accession: SRX186639

[GSM1003633: Stanford_ChipSeq_SK-N-SH_CTCF_\(SC-15914\)_IgG-rab](#)

2. 2 ILLUMINA (Illumina Genome Analyzer) runs: 51.6M spots, 1.9G bases, 914Mb downloads

Top Bioprojects

Production ENCODE functional... (425)

Search in related databases

Database	Access		
	public	controlled	all
BioSample			
BioProject			
dbGaP			
GEO Datasets	1		1

NIH National Library of Medicine
National Center for Biotechnology Information

GEO DataSets [Create alert](#) [Advanced](#)

Summary ▾

Send to: ▾

[ENCODE Transcription Factor Binding Sites by ChIP-seq from Stanford/Yale/USC/Harvard](#)

(Submitter supplied) This data was generated by ENCODE. If you have questions about the data, contact the submitting laboratory directly (Philip Cayting <mailto:pcaying@stanford.edu>). If you have questions about the Genome Browser track associated with this data, contact ENCODE (<mailto:genome@soe.ucsc.edu>). This track shows probable binding sites of the specified transcription factors (TFs) in the given cell types as determined by chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq). [more...](#)

Organism: Homo sapiens
Type: Genome binding/occupancy profiling by high throughput sequencing
Platforms: GPL9115 GPL9052 GPL10999 426 Samples
Download data: BIGWIG, NARROWPEAK, TXT
Series Accession: GSE31477 ID: 200031477
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [ENCODE](#) [SRA Run Selector](#)

SRP007993: A study ID. A study contains one or more experiments

NCBI Sequence Read Archive




HOME SEARCH SITE MAP
GEO Publications FAQ MIAME Email GEO

NCBI > GEO > [Accession Display](#) ?
Not logged in | [Login](#) ?

Scope: Format: Amount: GEO accession:

Series GSE31477 [Query DataSets for GSE31477](#)

Status Public on Aug 30, 2011

Title ENCODE Transcription Factor Binding Sites by ChIP-seq from Stanford/Yale/USC/Harvard

Project [ENCODE](#)

Organism [Homo sapiens](#)

Experiment type Genome binding/occupancy profiling by high throughput sequencing

Summary This data was generated by ENCODE. If you have questions about the data, contact the submitting laboratory directly (Philip Cayting <mailto:pcayting@stanford.edu>). If you have questions about the Genome Browser track associated with this data, contact ENCODE (<mailto:genome@soe.ucsc.edu>).

This track shows probable binding sites of the specified transcription factors (TFs) in the given cell types as determined by chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq). Included for each cell type is the input signal, which represents the control condition where no antibody targeting was performed. For each experiment (cell type vs. antibody) this track shows a graph of enrichment for TF binding (Signal), along with sites that have the greatest evidence of transcription factor binding (Peaks).

For data usage terms and conditions, please refer to <http://www.genome.gov/27528022> and <http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf>

Overall design Cells were grown according to the approved ENCODE cell culture protocols. Further preparations were similar to those previously published (Euskirchen et al., 2007) with the exceptions that the cells were unstimulated and sodium orthovanadate was omitted from the buffers. For details on the chromatin immunoprecipitation protocol used, see Euskirchen et al. (2007) and Rozowsky et al. (2009). DNA recovered from the precipitated chromatin was sequenced on the Illumina (Solexa) sequencing platform and mapped to the genome using the Eland alignment program. ChIP-seq data was scored based on sequence reads (length ~30 bps) that align uniquely to the human genome. From the mapped tags a signal map of ChIP DNA fragments (average fragment length ~ 200 bp) was constructed where the signal height is the number of overlapping fragments at each nucleotide position in the genome. For each 1 Mb segment of each chromosome a peak height threshold was determined by requiring a false discovery rate <= 0.05 when comparing the number of peaks above threshold as compared the number obtained from multiple simulations of a random null background with the same number of mapped reads (also accounting for the fraction of mappable bases for sequence tags in that 1 Mb segment). The number of mapped tags in a putative binding region is compared to the normalized (normalized by correlating tag counts in genomic 10 kb windows) number of mapped tags in the same region from an input DNA control. Using a binomial test, only regions that have a p-value <= 0.05 are

Web link <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeSydhTfbs>

Contributor(s) [Snyder M, Gerstein M, Weissman S, Farnham P, Struhl K](#)

Citation(s) ENCODE Project Consortium.. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012 Sep 6;489(7414):57-74. PMID: 22955616

BioProject [PRJNA63447](#) Production ENCODE functional genomics data.

Submission date Aug 18, 2011

Last update date Nov 10, 2021

Contact name ENCODE DCC

E-mail(s) encode-help@lists.stanford.edu

Organization name ENCODE DCC

Street address 300 Pasteur Dr

City Stanford

State/province CA

ZIP/Postal code 94305-5120

Country USA

Platforms (3) [GPL9052](#) Illumina Genome Analyzer (Homo sapiens)
[GPL9115](#) Illumina Genome Analyzer II (Homo sapiens)
[GPL10999](#) Illumina Genome Analyzer IIX (Homo sapiens)

Samples (426) [GSM782122](#) USC_ChipSeq_HepG2_TCF7L2_UCDavis
[GSM782123](#) USC_ChipSeq_HCT-116_TCF7L2_UCDavis
[GSM782124](#) USC_ChipSeq_HEK293_TCF7L2_UCDavis
[More...](#)

Relations

SRA [SRP007993](#)

[See on Genome Data Viewer](#)

- [GSM935469](#) Yale_ChipSeq_K562_IFNa6h_STAT2_std
- [GSM935470](#) Yale_ChipSeq_K562_IFNa30_STAT2_std
- ➔ [GSM935471](#) Yale_ChipSeq_K562_IFNa6h_STAT1_std
- ➔ [GSM935472](#) Yale_ChipSeq_K562_IFNa30_STAT1_std
- ➔ [GSM935419](#) Yale_ChipSeq_K562_IFNa6h_Input_std
- [GSM935420](#) Yale_ChipSeq_K562_IFNg30_Input_std
- [GSM935421](#) USC_ChipSeq_NT2-D1_Input_UCDavis
- ➔ [GSM935422](#) Yale_ChipSeq_K562_IFNa30_Input_std

NCBI Sequence Read Archive

NCBI GEO > Accession Display [?](#) Not logged in | [Login](#) [?](#)

Scope: Format: Amount: GEO accession:

Sample GSM935471 [Query DataSets for GSM935471](#)

Status: Public on May 22, 2012
 Title: Yale_ChipSeq_K562_IFNa6h_STAT1_std
 Sample type: SRA

Source name: K562
 Organism: [Homo sapiens](#)
 Characteristics: lab: Yale
 lab description: Snyder - Yale University
 datatype: ChipSeq
 datatype description: Chromatin IP Sequencing
 cell: K562
 cell organism: human
 cell description: leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC
 cell karyotype: cancer
 cell lineage: mesoderm
 cell sex: F
 treatment: IFNa6h
 treatment description: Interferon alpha treatment - 6 hours (Snyder)
 antibody: STAT1
 antibody antibodydescription: rabbit polyclonal to STAT1 (alpha) p-91 (C-24).
 Antibody Target: STAT1
 antibody targetdescription: transcription factor, activated by interferon signalling
 antibody vendorname: Santa Cruz Biotech
 antibody vendorid: sc-345
 control: std
 control description: Standard input signal for most experiments.
 control: std
 control description: Standard input signal for most experiments.
 controlid: wgEncodeEH000656
 replicate: 1

Biomaterial provider: ATCC
 Treatment protocol: IFNa6h
 Growth protocol: K562_protocol.pdf
 Extracted molecule: genomic DNA
 Extraction protocol: Instrument model unknown. ("Illumina Genome Analyzer" specified by default). For more information, see <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeSydhTfbs>

GEO DataSet: Curated dataset from submitter
 SRX: experiment ID
 SRR: sequencing run ID

Library strategy: **ChIP-Seq**
 Library source: **genomic**
 Library selection: **ChIP**
 Instrument model: **Illumina Genome Analyzer**

Data processing: <http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeSydhTfbs>

Submission date: May 22, 2012
 Last update date: May 15, 2019
 Contact name: ENCODE DCC
 E-mail(s): encode-help@lists.stanford.edu
 Organization name: ENCODE DCC
 Street address: 300 Pasteur Dr
 City: Stanford
 State/province: CA
 ZIP/Postal code: 94305-5120
 Country: USA

Platform ID: **GPL9052**
 Series (1): **GSE31477** ENCODE Transcription Factor Binding Sites by ChIP-seq from Stanford/Yale/USC/Harvard

Relations
 SRA: [SRX150550](#)
 BioSample: [SAMN01001010](#)

[SRX150550: GSM935471: Yale_ChipSeq_K562_IFNa6h_STAT1_std](#)
 2 ILLUMINA (Illumina Genome Analyzer) runs: 38.9M spots, 1G bases, 780Mb downloads

Submitted by: Gene Expression Omnibus (GEO)

Study: GSE31477: ENCODE Transcription Factor Binding Sites by ChIP-seq from Stanford/Yale/USC/Harvard
 • [SRP007993](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample: Yale_ChipSeq_K562_IFNa6h_STAT1_std
[SAMN01001010](#) • [SRS335942](#) • [All experiments](#) • [All runs](#)
 Organism: [Homo sapiens](#)

Library:
Name: GSM935471: Yale_ChipSeq_K562_IFNa6h_STAT1_std
Instrument: Illumina Genome Analyzer
Strategy: ChIP-Seq
Source: GENOMIC
Selection: ChIP
Layout: SINGLE

Spot descriptor:
 1 forward

Experiment attributes:
 GEO Accession: GSM935471

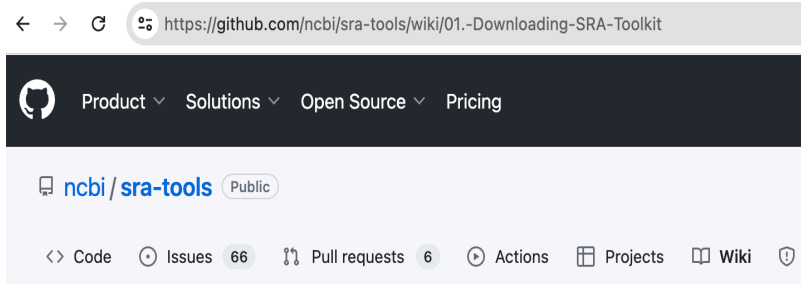
Links:
 External link: [GEO Web Link](#)
 NCBI link: [NCBI Entrez \(gds\)](#)

Runs: 2 runs, 38.9M spots, 1G bases, [780Mb](#)

Run	# of Spots	# of Bases	Size	Published
SRR502327	19,429,794	524.6M	390.4Mb	2012-05-30
SRR502328	19,444,603	525M	389.6Mb	2012-05-30

SRA Dataset Download Toolkit

<https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit>



01. Downloading SRA Toolkit

Andrew Klymenko edited this page on Aug 29 · 31 revisions

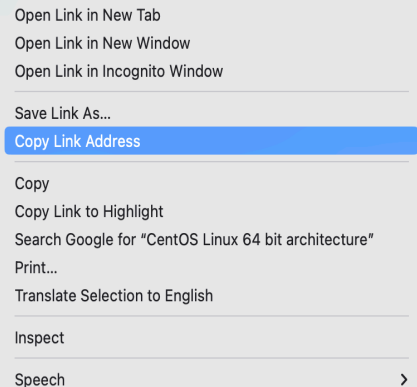
NCBI SRA Toolkit

Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries/install scripts of August 29, 2023, version 3.0.7:

- [CentOS Linux 64 bit architecture](#)
- [Ubuntu Linux 64 bit architecture](#)
- [Cloud - apt-get install script](#)
- [Cloud - yum install script](#)
- [MacOS 64 bit architecture](#)
- [MS Windows 64 bit architecture](#)
- [Docker image repository](#)
- [md5 checksums](#)



Magic-BLAST

login to hoffman:

```
% ssh your_username@hoffman2.idre.ucla.edu
```

Create symbolic link to your scratch directory

```
% ln -s /u/scratch/w/wyan myscratch  
% cd ~/myscratch  
% mkdir workshop  
% cd workshop
```

Download sra toolkit

```
% wget https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/3.0.7/sratoolkit.3.0.7-centos_linux64.tar.gz
```

Extract sra toolkit

```
% tar -xvf sratoolkit.3.0.7-centos_linux64.tar.gz
```

Replace text in red with your account information

SRA CHIP-seq STAT1 Data Download

SRP007993	GSM935422	SRR502228	Yale_ChipSeq_K562_IFNa30_Input_std
SRP007993	GSM935419	SRR502225	Yale_ChipSeq+K562_IFNa6h_Input_std
SRP007993	GSM935472	SRR502329 SRR502330	Yale_ChipSeq_K562_IFNa30_STAT1_std
SRP00793	GSM935471	SRR502327 SRR502328	Yale_ChipSeq_K562_IFNa6h_STAT1_std

request interactive session:

```
% qssh -l h_data=4G,h_rt=2:00:00 -pe shared 4
```

add sratoolkit programs into the system path:

```
% export PATH=~/.myscratch/workshop/sratoolkit.3.0.7-centos_linux64/bin:$PATH
```

run sratoolkit program "fastq-dump"

```
% cd ~/.myscratch/workshop
```

```
% fastq-dump
```

```
[wyan@login2 workshop]$ qssh -l h_data=4G,h_rt=2:00:00 -pe shared 4
[wyan@n6046 ~]$ export PATH=~/.myscratch/workshop/sratoolkit.3.0.7-centos_linux64/bin:$PATH
[wyan@n6046 ~]$ cd ~/.myscratch/workshop
[wyan@n6046 workshop]$ fastq-dump
```

Usage:

```
fastq-dump [options] <path> [<path>...]
fastq-dump [options] <accession>
```

Use option --help for more information

```
fastq-dump : 3.0.7
```

SRA STAT1 and Input Data Download

go to data directory and run sratoolkit program “fastq-dump”

```
% cd /myscratch/workshop
% fastq-dump -Z SRR502228 >INP_30m_IFNa.fastq
% fastq-dump -Z SRR502225 >INP_6h_IFNa.fastq
% fastq-dump -Z SRR502329 >STAT1_30m_IFNa.fastq
% fastq-dump -Z SRR502327 >STAT1_6h_IFNa.fastq
```

```
[wyan@n6046 workshop]$ fastq-dump -Z SRR502228 >INP_30m_IFNa.fastq
Read 26699669 spots for SRR502228
Written 26699669 spots for SRR502228
[wyan@n6046 workshop]$ fastq-dump -Z SRR502225 >INP_6h_IFNa.fasta
Read 31983231 spots for SRR502225
Written 31983231 spots for SRR502225
[wyan@n6046 workshop]$ fastq-dump -Z SRR502329 >STAT1_30m_IFNa.fastq
Read 21192112 spots for SRR502329
Written 21192112 spots for SRR502329
[wyan@n6046 workshop]$ fastq-dump -Z SRR502327 >STAT1_6h_IFNa.fastq
Read 19429794 spots for SRR502327
Written 19429794 spots for SRR502327
[wyan@n6046 workshop]$ mv INP_6h_IFNa.fasta INP_6h_IFNa.fastq
```