Introduction to NGS analyses

Giorgio L Papadopoulos

Institute of Molecular Biology and Biotechnology

Bioinformatics Support Group

04/12/2015

Overview

1. Introduction
   - DNA sequencing
   - Applications

Overview

1. Introduction
   - DNA sequencing
   - Applications
2. Primary Analysis
   - ChIPseq
   - RNAseq

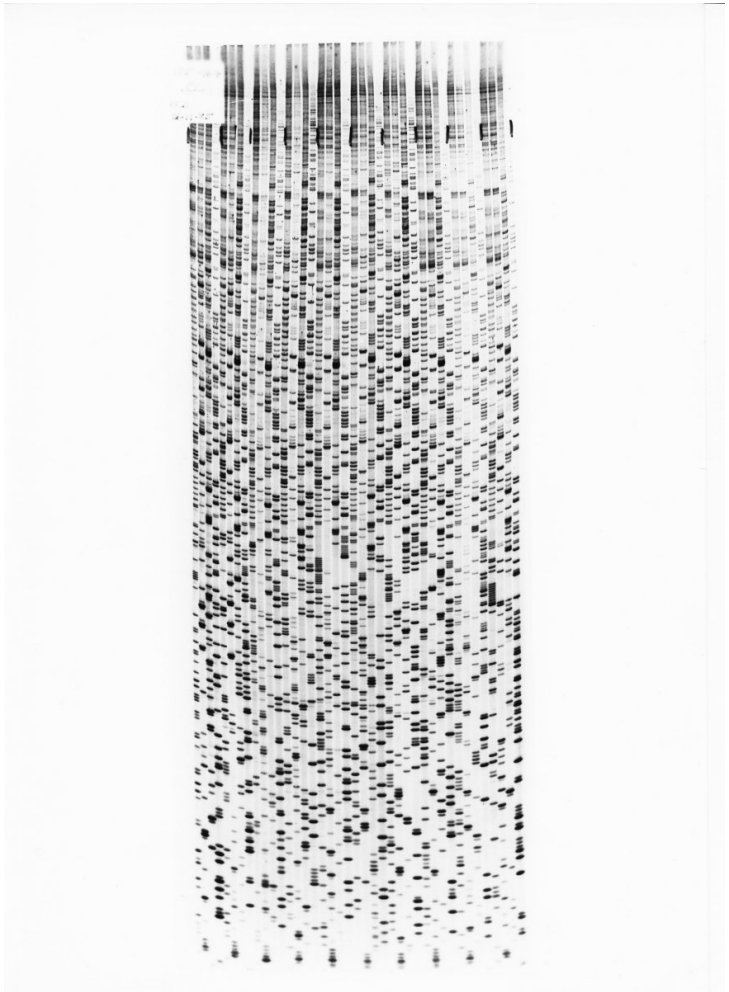
Overview

1. Introduction
   - DNA sequencing
   - Applications
2. Primary Analysis
   - ChIPseq
   - RNAseq
3. Interpretation
   - Visualization
   - Analysis Software
   - Downstream Analysis

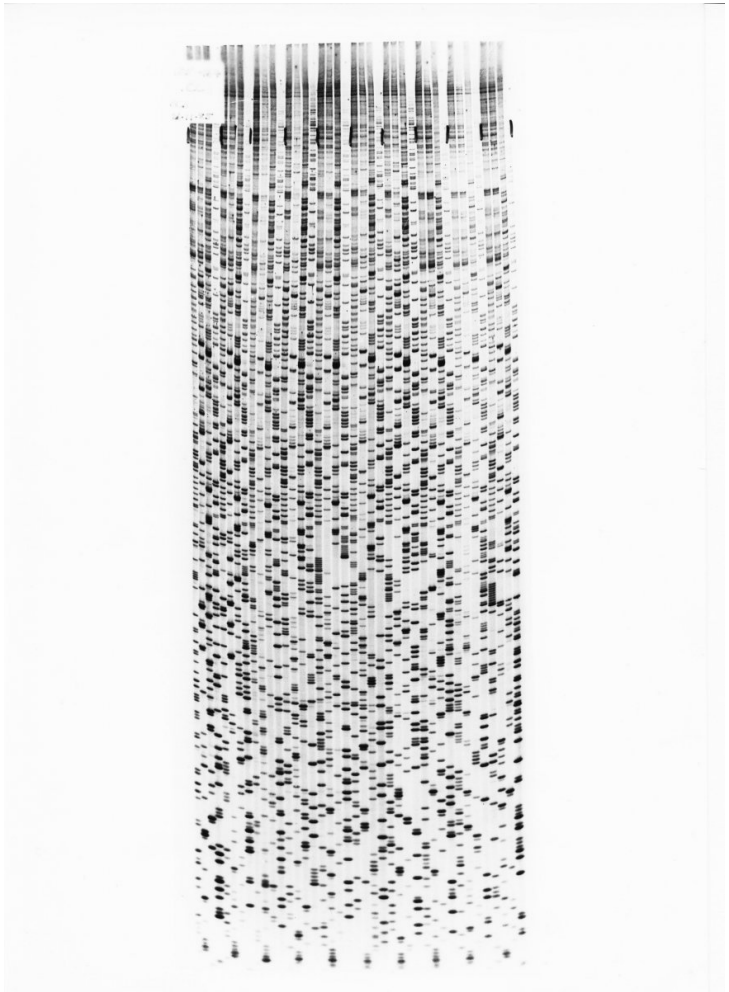
Overview

1. Introduction
   - DNA sequencing
   - Applications
2. Primary Analysis
   - ChIPseq
   - RNAseq
3. Interpretation
   - Visualization
   - Analysis Software
   - Downstream Analysis
4. Online Resources
   - ENCODE
   - GEO
   - ENA

Fred Sanger
Late 1970s

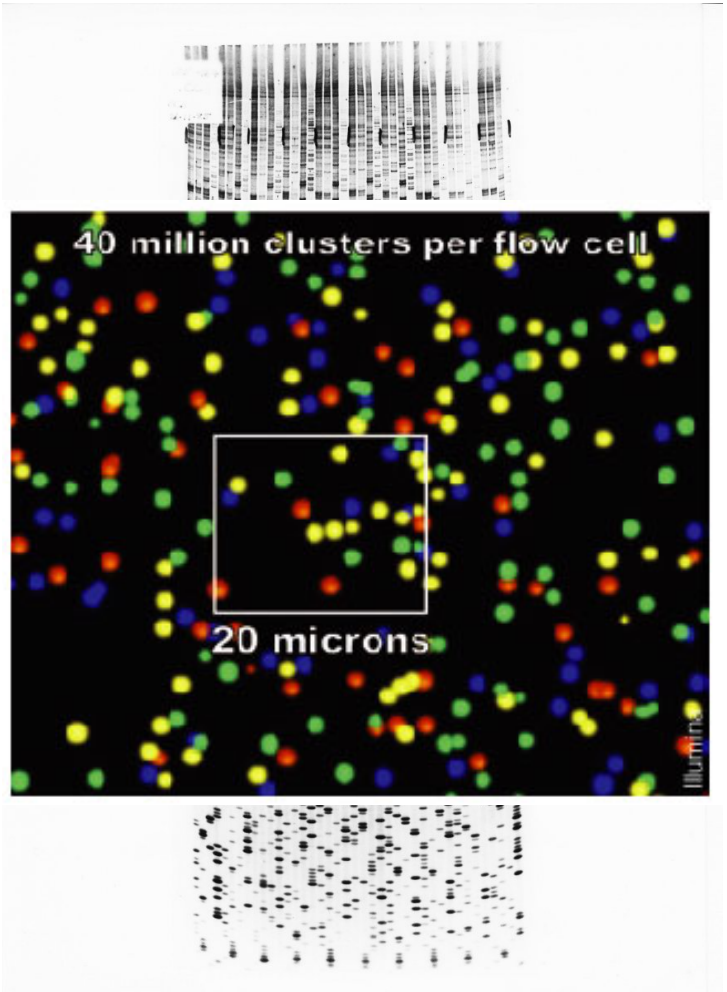


Fred Sanger
Late 1970s



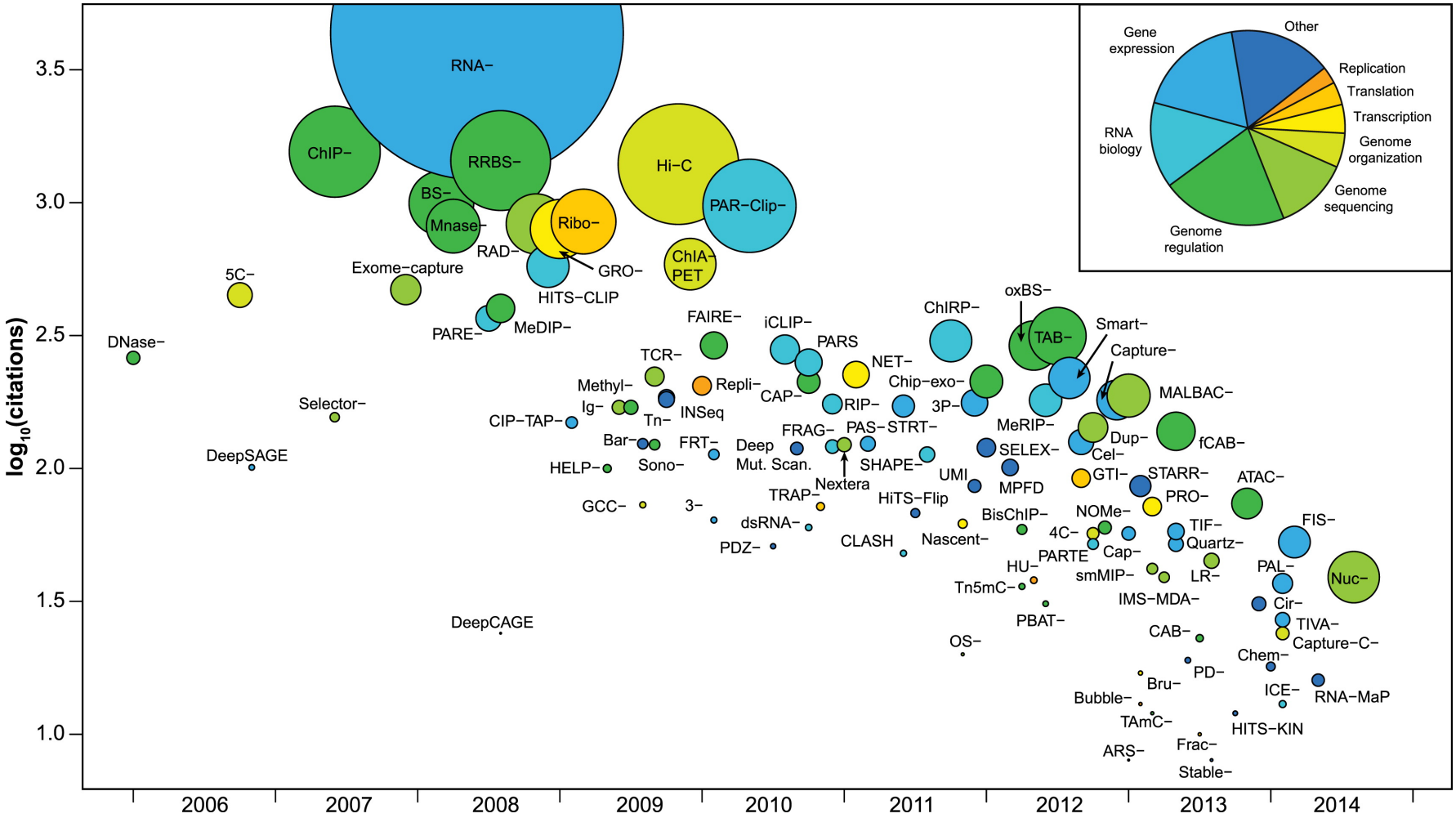
Fred Sanger
Late 1970s

... Several million times ...
Late 2000s

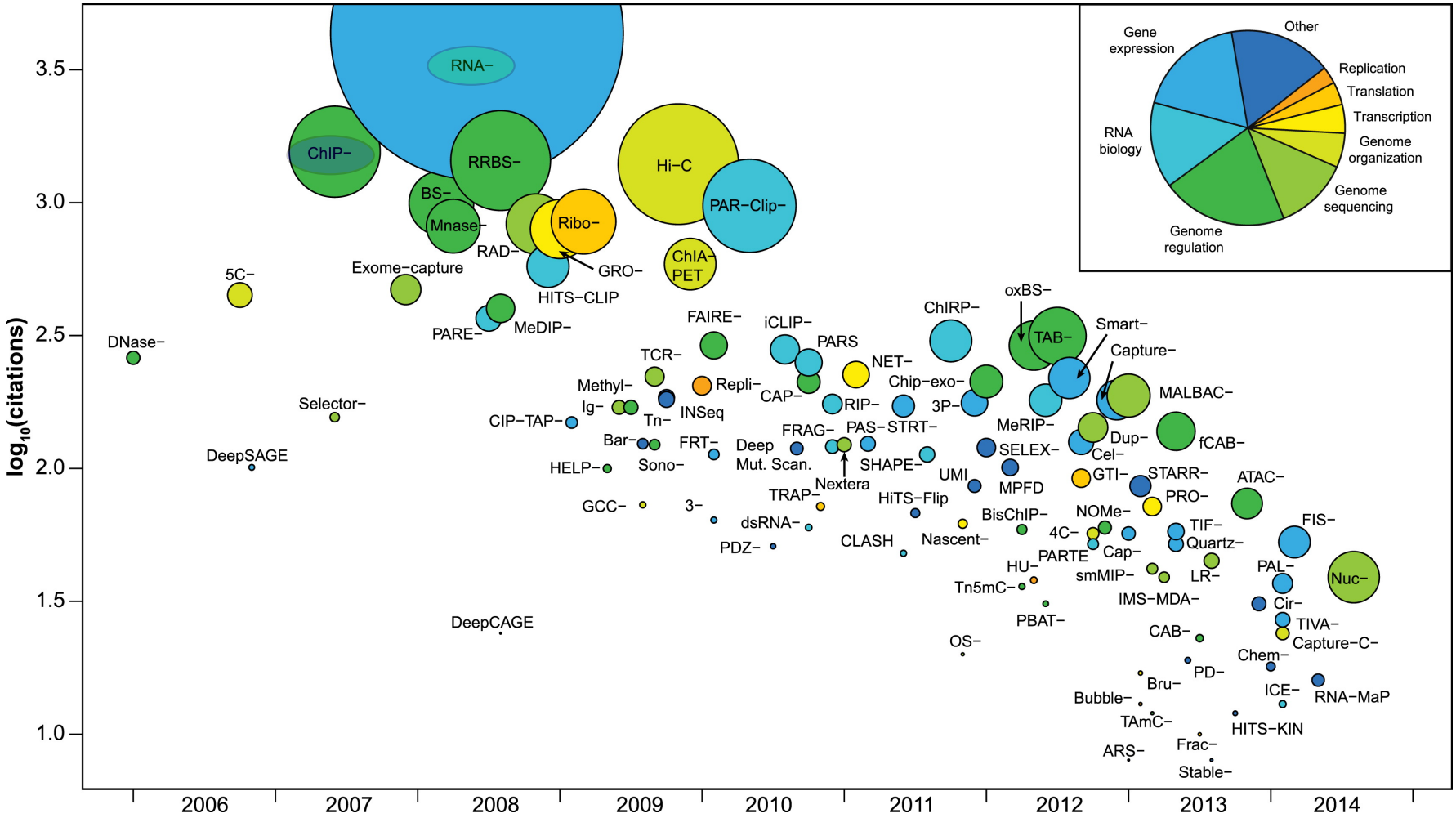


Fred Sanger
Late 1970s

... Several million times ...
Late 2000s

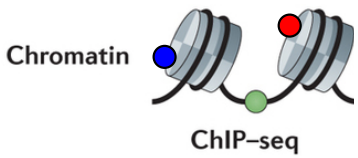


Reuter et al, Mol Cell, 2015

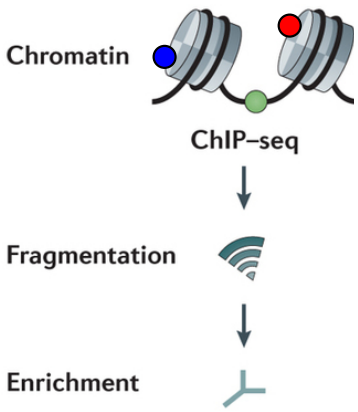


Reuter et al, Mol Cell, 2015

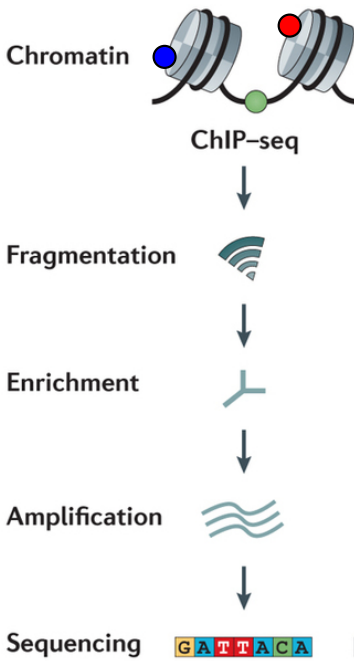
Genome Wide Occupancy Profiling



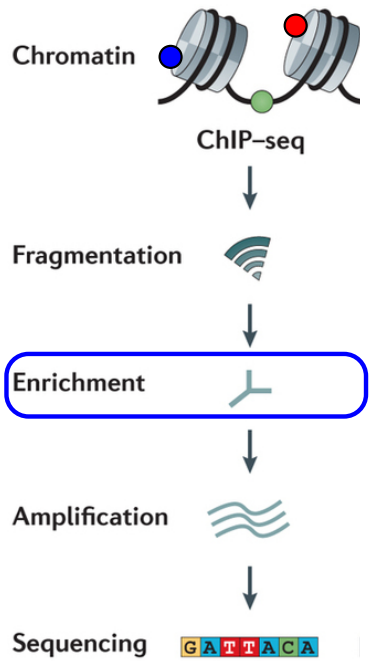
Genome Wide Occupancy Profiling



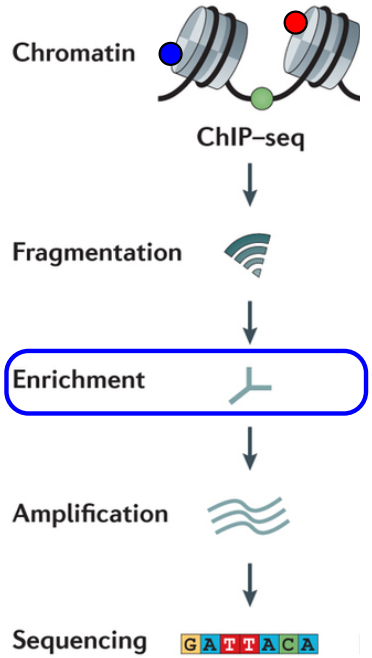
Genome Wide Occupancy Profiling

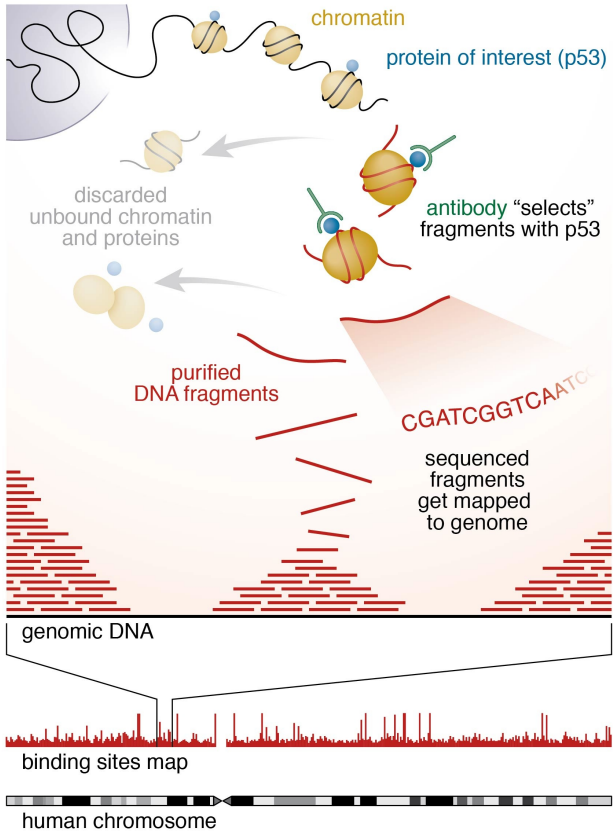


Genome Wide Occupancy Profiling

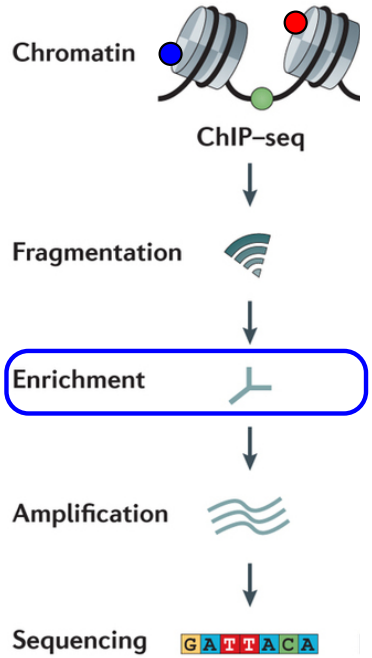


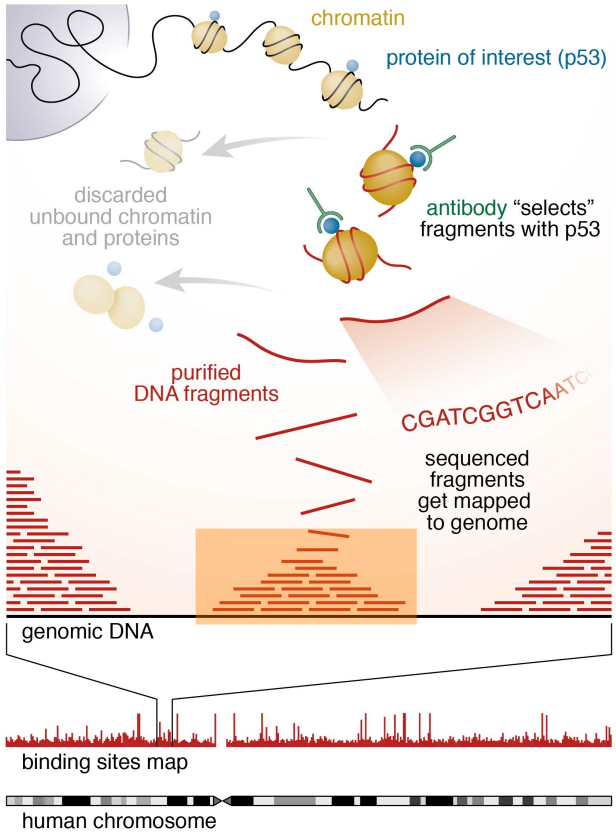
Genome Wide Occupancy Profiling



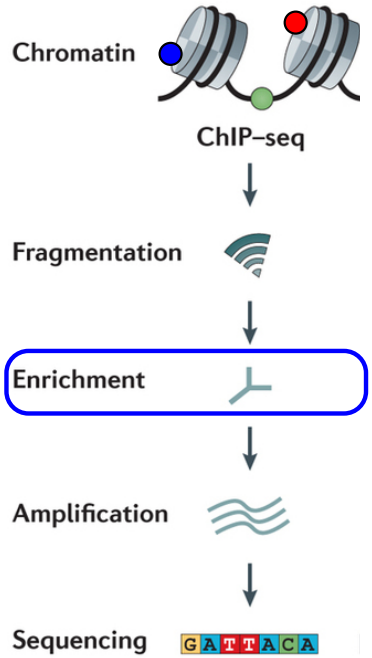


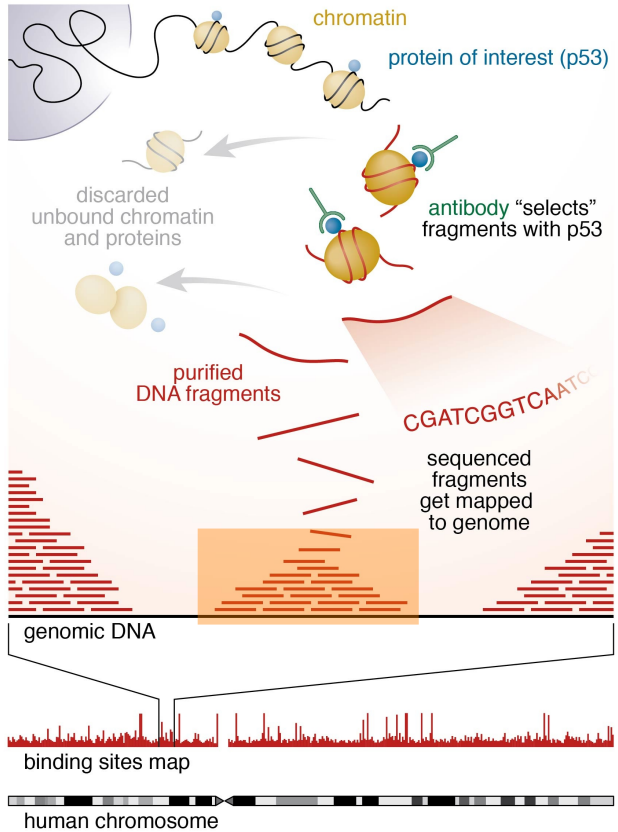
Genome Wide Occupancy Profiling





Genome Wide Occupancy Profiling

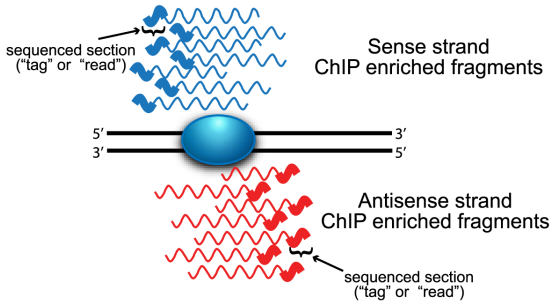




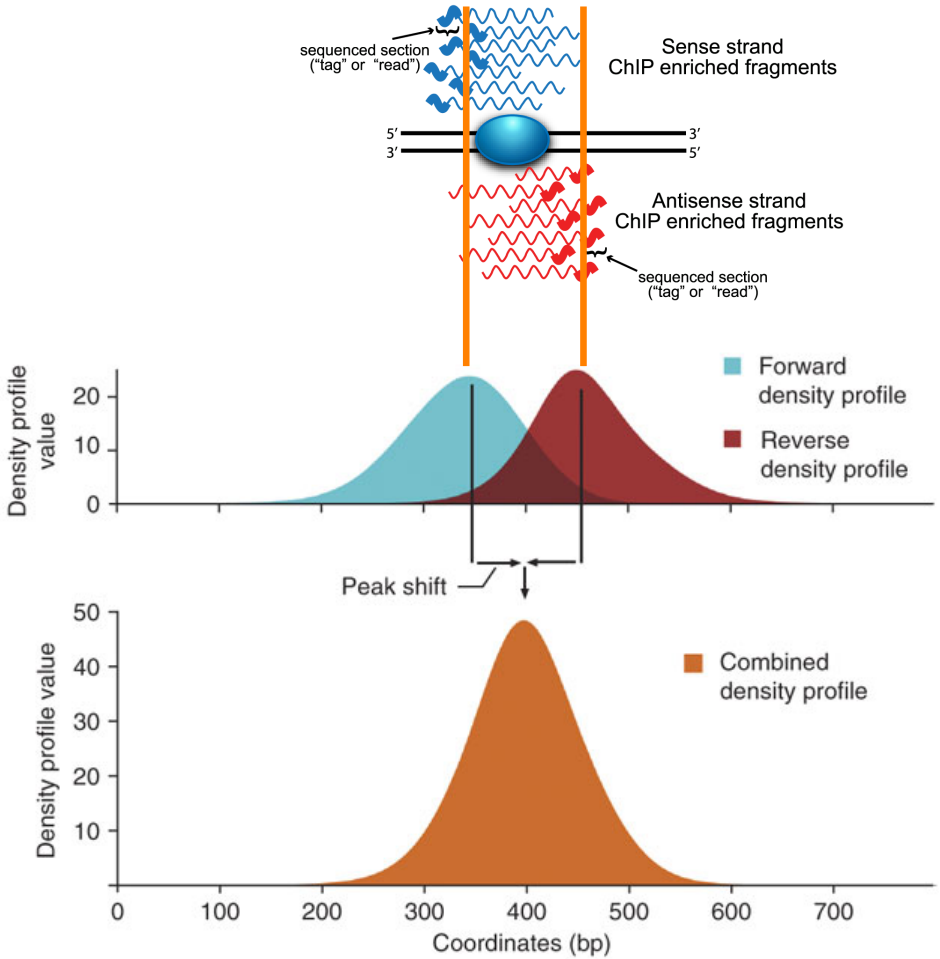
Overlapping reads come from different cells!!!

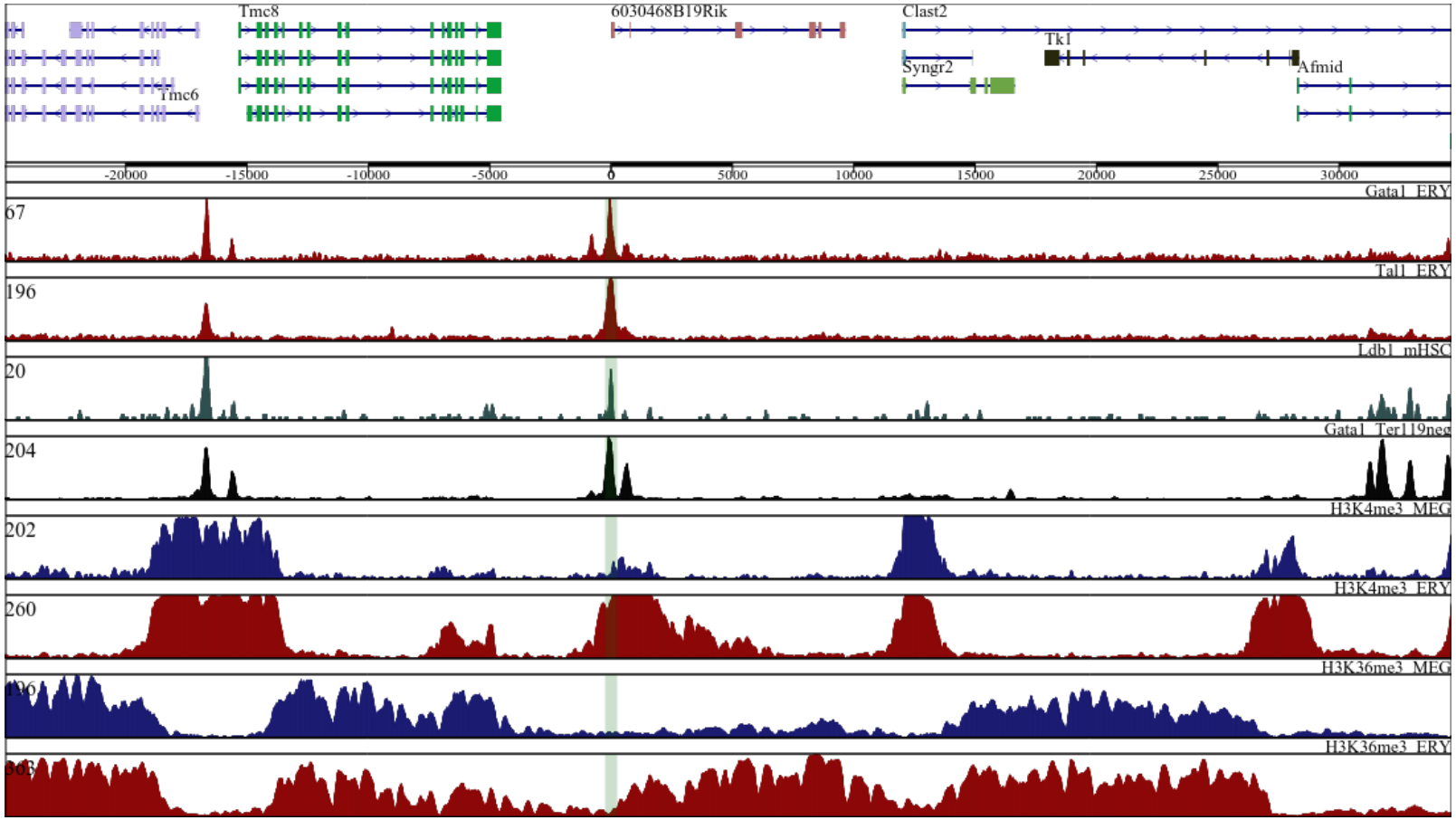
NGS is a cell population readout (50 million cells)

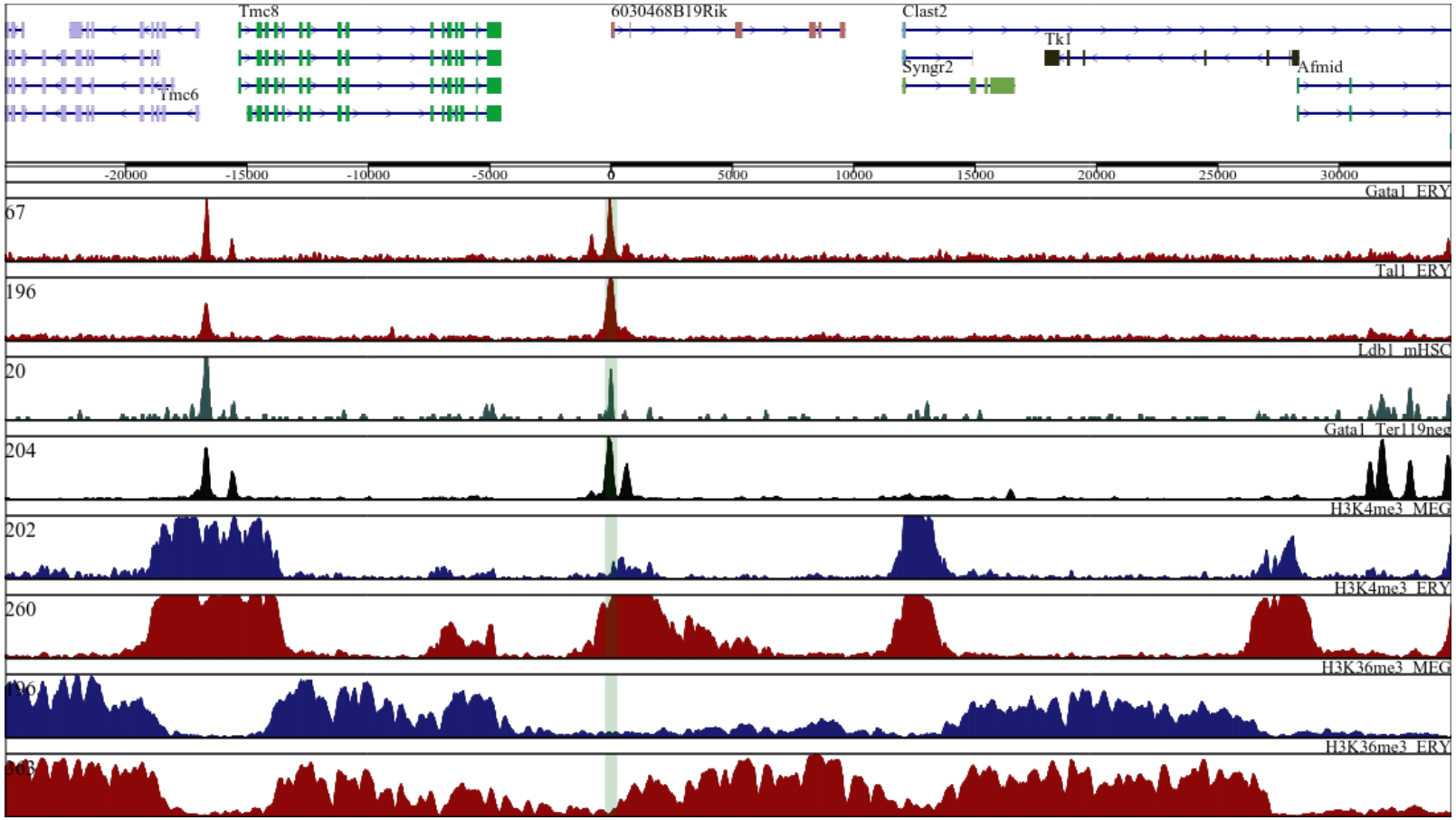
Primary Sequencing Data



Primary Sequencing Data

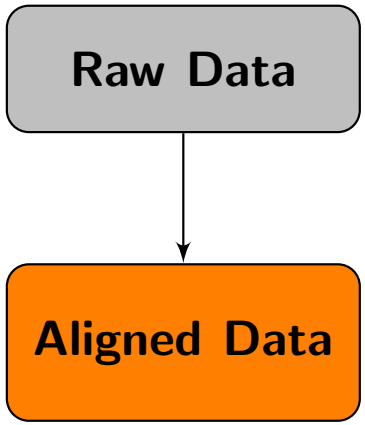


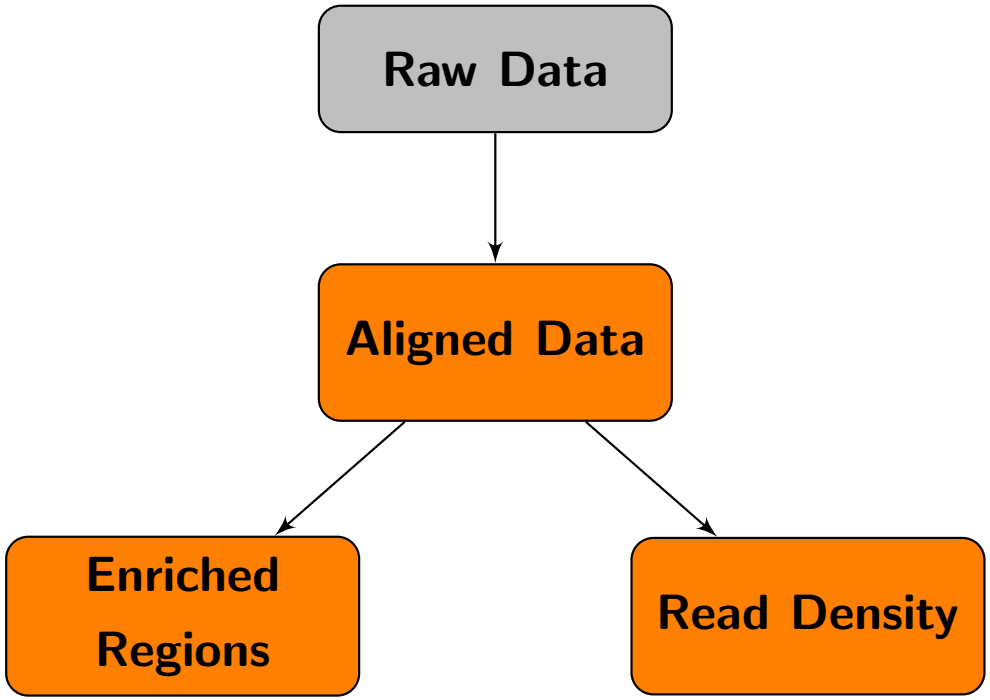

Tmc8
6030468B19Rik
Clast2
Tk1
Syngr2
Afmid
Tmc6
-20000
-15000
-10000
-5000
0
5000
10000
15000
20000
25000
30000
Gata1_ERY
67
Tal1_ERY
196
Ldb1_mHSC
20
Gata1_Ter119neg
204
H3K4me3_MEG
202
H3K4me3_ERY
260
H3K36me3_MEG
H3K36me3_ERY

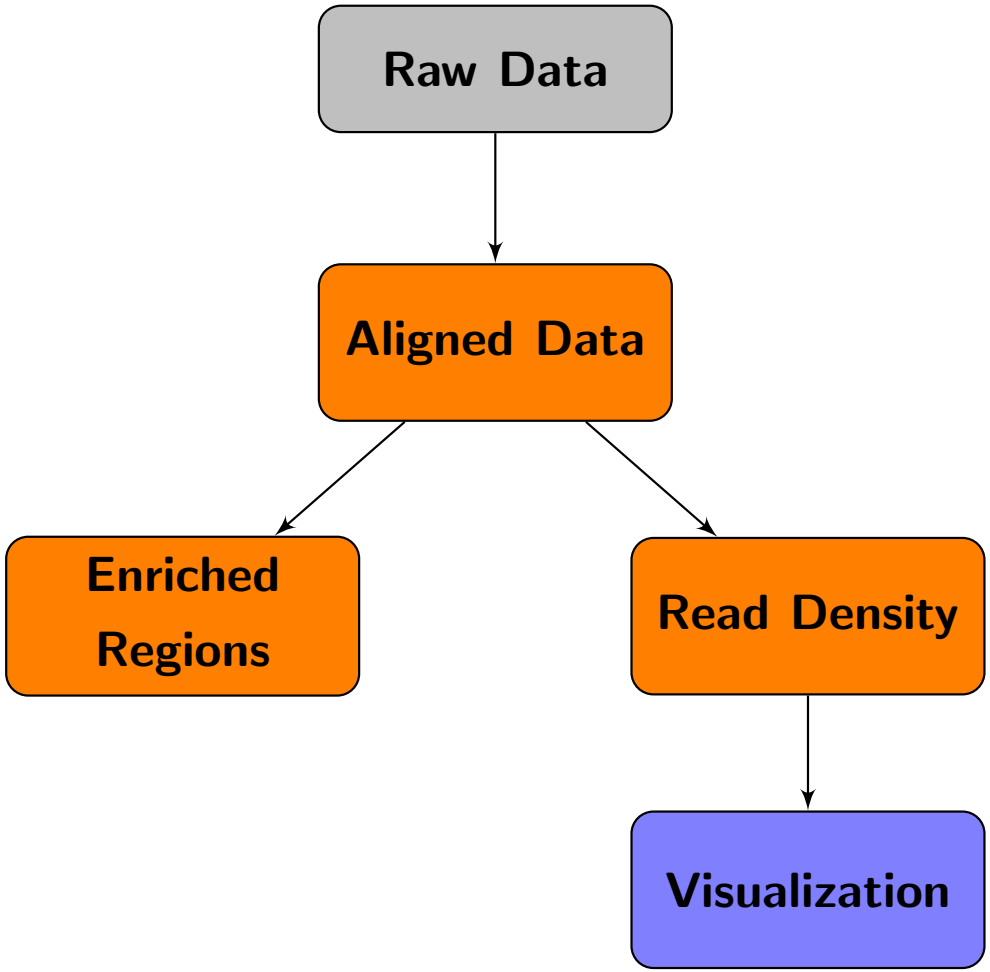


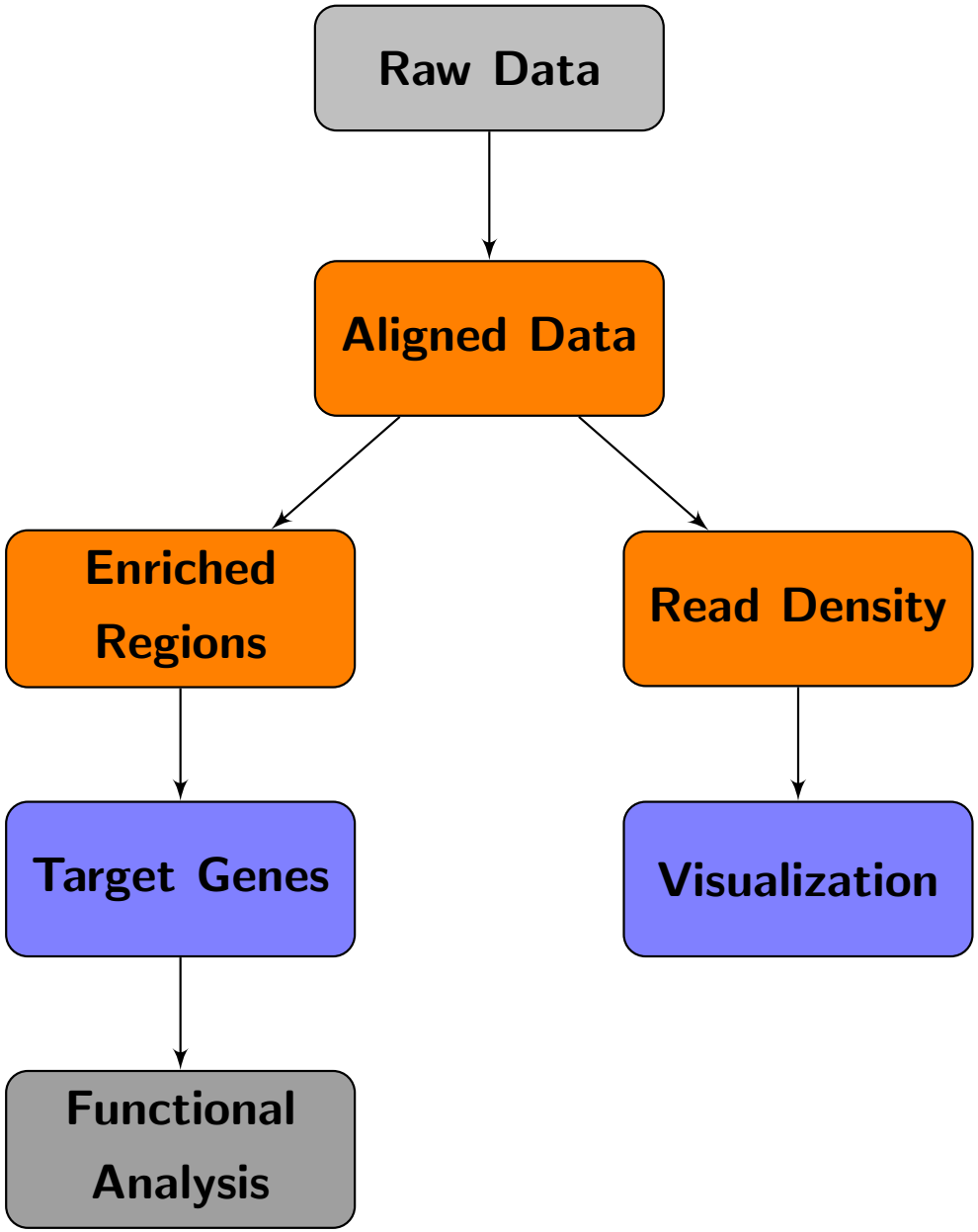
Genome wide, highly accurate representation of genomic states
Fairly unbiased
Complementary and overlapping information

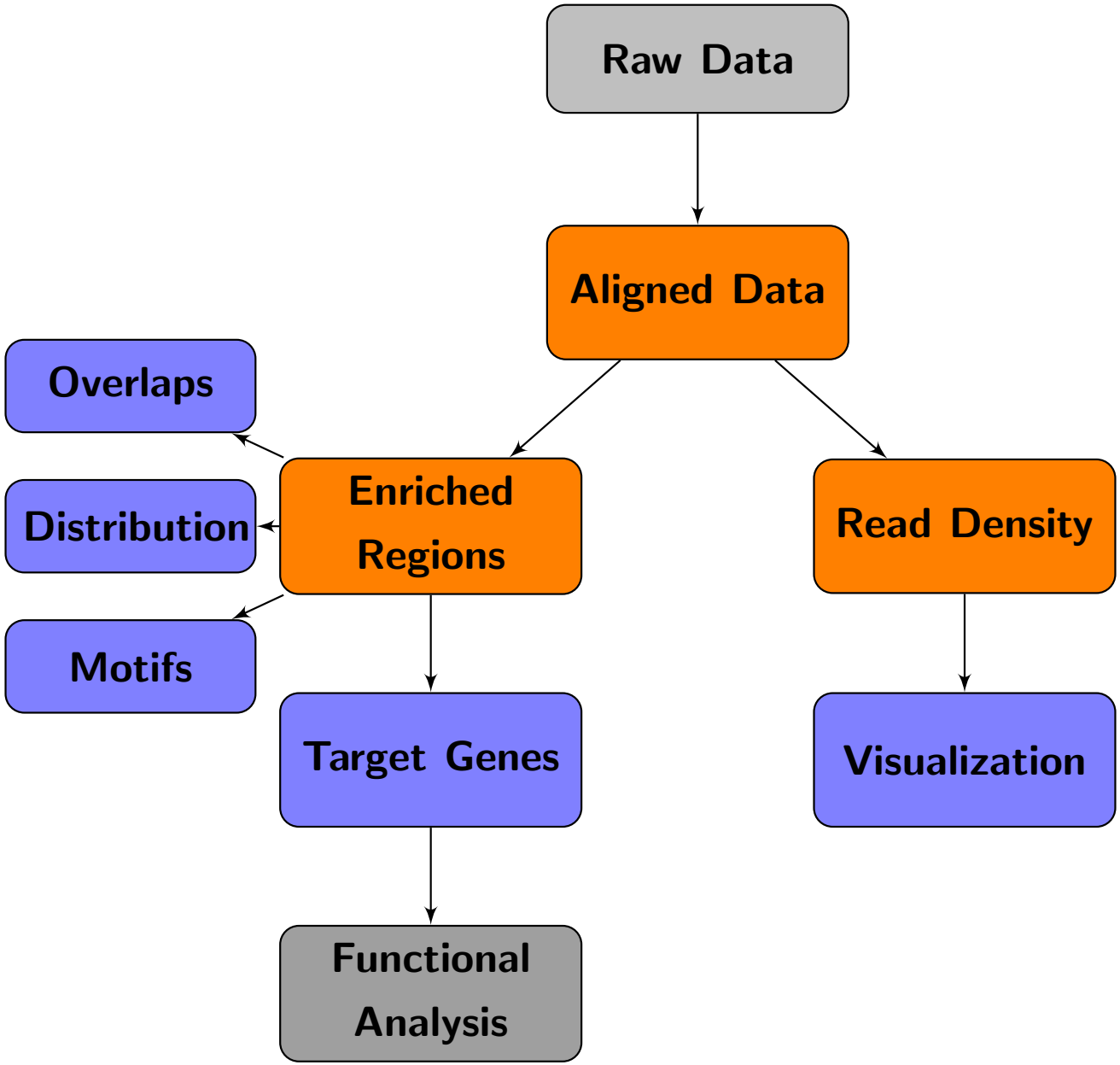
Raw Data


Raw Data
Aligned Data


Raw Data
Aligned Data
Enriched Regions
Read Density


Raw Data
Aligned Data
Enriched Regions
Read Density
Visualization


Raw Data
Aligned Data
Enriched Regions
Read Density
Target Genes
Visualization
Functional Analysis


Raw Data
Aligned Data
Overlaps
Distribution
Enriched Regions
Motifs
Read Density
Target Genes
Visualization
Functional Analysis

Raw Data

Raw Data

What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```

Raw Data

What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```

Operations:

DeMultiplexing

Quality Control

Raw Data

What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```

FastQC

Operations:
DeMultiplexing
Quality Control

Raw Data

What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

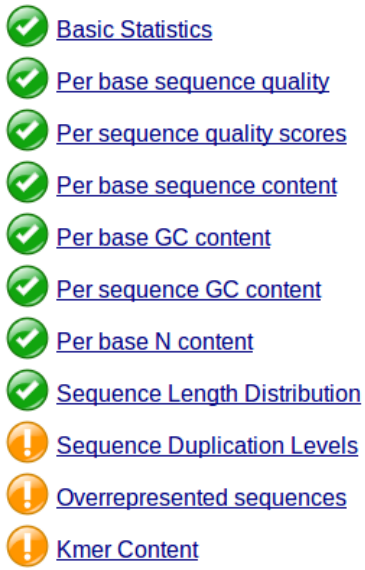
```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```

Operations:

DeMultiplexing

Quality Control

FastQC



Raw Data

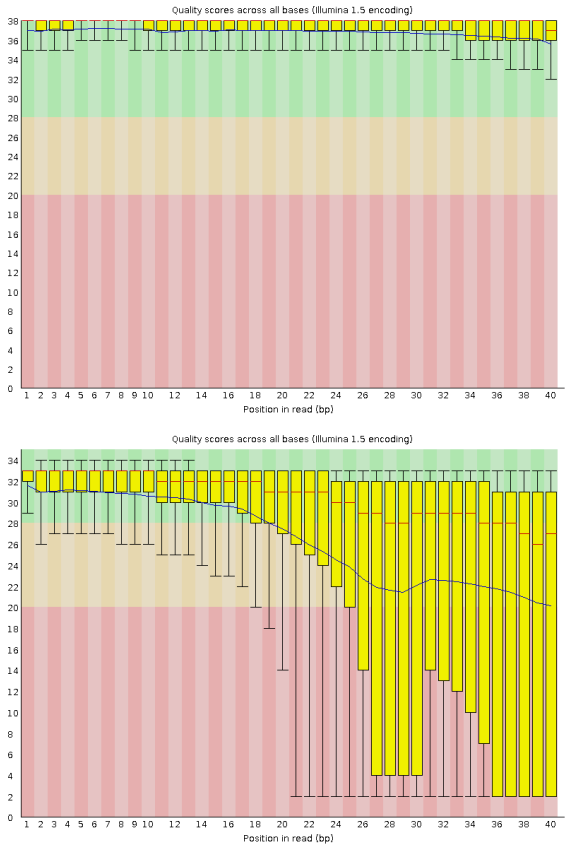
What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```

Operations:
DeMultiplexing
Quality Control

FastQC



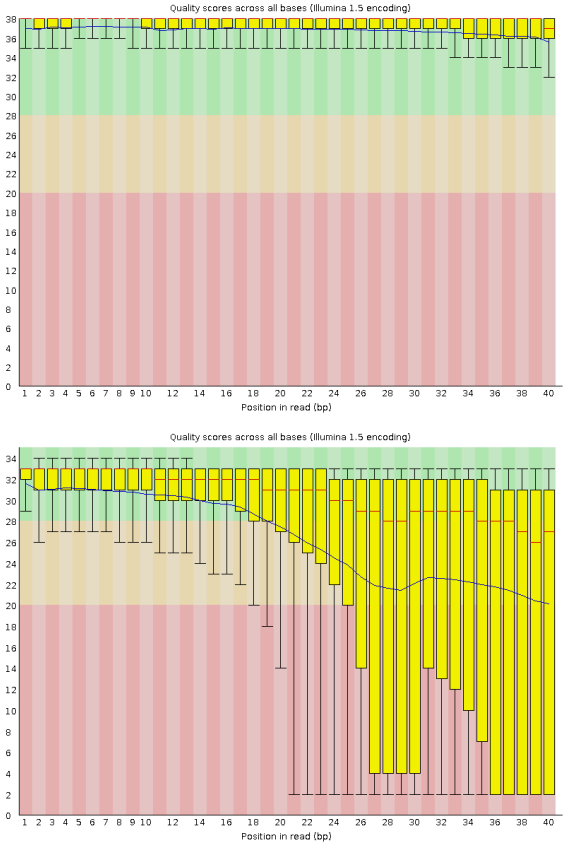
Raw Data

FastQC

What is FASTQ?

- Text-based format for storing both biological sequences and corresponding quality scores.
- FASTQ = FASTA + QUALITY
- A FASTQ file uses four lines per sequence.

```
1 @SEQ_ID
2 GATTTGGGGTTCAAAGCAGTATCGATCAAA
3 +SEQ_ID(Optional)
4 !''*((((***+))%%%++)(%%%%).1**
```



Operations:
DeMultiplexing
Quality Control

Align Data...

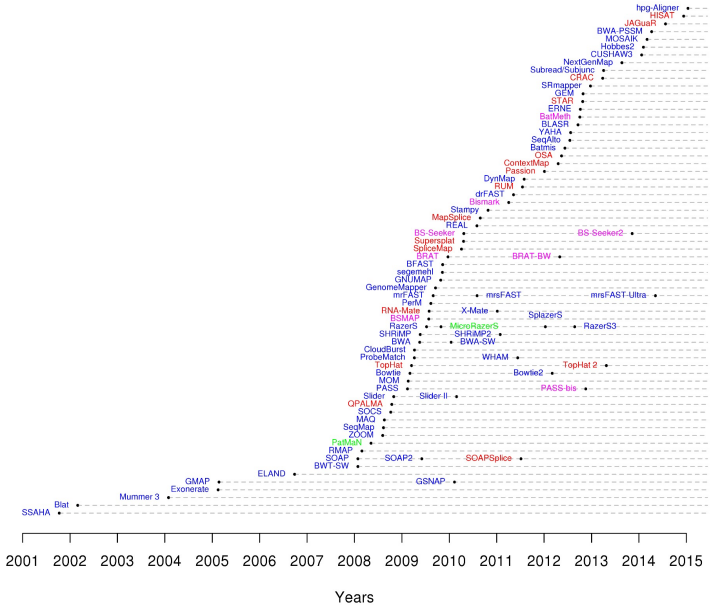
Align Data...

Align Data...

Aligner (Mapper)

Align Data...

Aligner (Mapper)



http://www.ebi.ac.uk/~nf/hts_mappers/

Align Data...

Aligner (Mapper)

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

Align Data...

Aligner (Mapper)

bowtie.2

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

Align Data...

Aligner (Mapper)

bowtie.2

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

Fast...
Versatile...
Documented and supported...

Align Data...

Aligner (Mapper)

bowtie.2

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

% of alignment...

Align Data...

Aligner (Mapper)

bowtie.2

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

% of alignment...

Low percentage?
Contaminations,
adapter/barcode trimming,
wrong reference...

Align Data...

Aligner (Mapper)

bowtie.2

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
(iGenomes, UCSC, NCBI, EBI...)

Different mappers will use different Index Formats...

% of alignment...

Low percentage?
Contaminations,
adapter/barcode trimming,
wrong reference...

Bad Library...

Aligned Data

Aligned Data

SAM $\Leftrightarrow$ BAM $\Rightarrow$ BED

Aligned Data

SAM $\Leftrightarrow$ BAM $\Rightarrow$ BED

$\Downarrow$

Enriched Regions

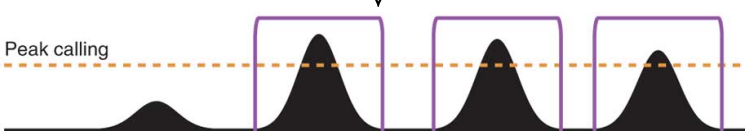
Aligned Data

SAM $\Leftrightarrow$ BAM $\Rightarrow$ BED

$\Downarrow$

Enriched Regions

$\Downarrow$



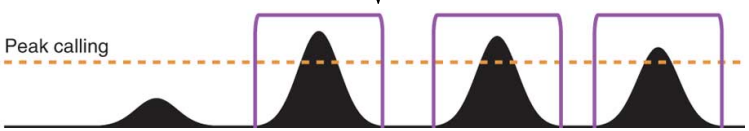
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

SET of GENOMIC COORDINATES (PEAKS...)

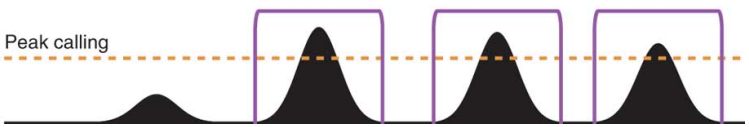
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

How Many???

SET of GENOMIC COORDINATES (PEAKS...)

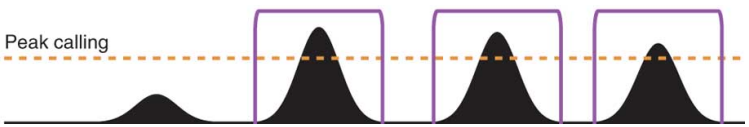
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

How Many???

Where???

SET of GENOMIC COORDINATES (PEAKS...)

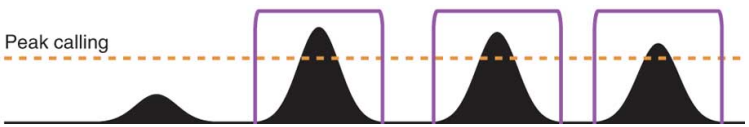
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

How Many???

Where???

With Who???

SET of GENOMIC COORDINATES (PEAKS...)

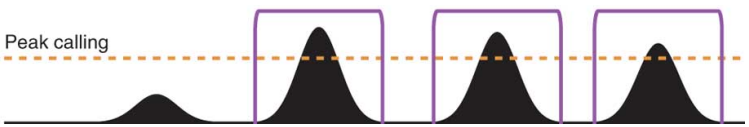
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

How Many???

Where???

With Who???

Function???

SET of GENOMIC COORDINATES (PEAKS...)

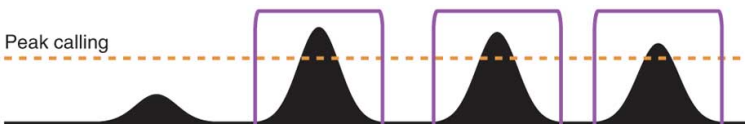
Aligned Data

SAM ⇔ BAM ⇒ BED

⇓

Enriched Regions

⇓



⇓

Chr Start Stop ID Score

SET of GENOMIC COORDINATES (PEAKS...)

How Many???

Where???

With Who???

Function???

HOW???

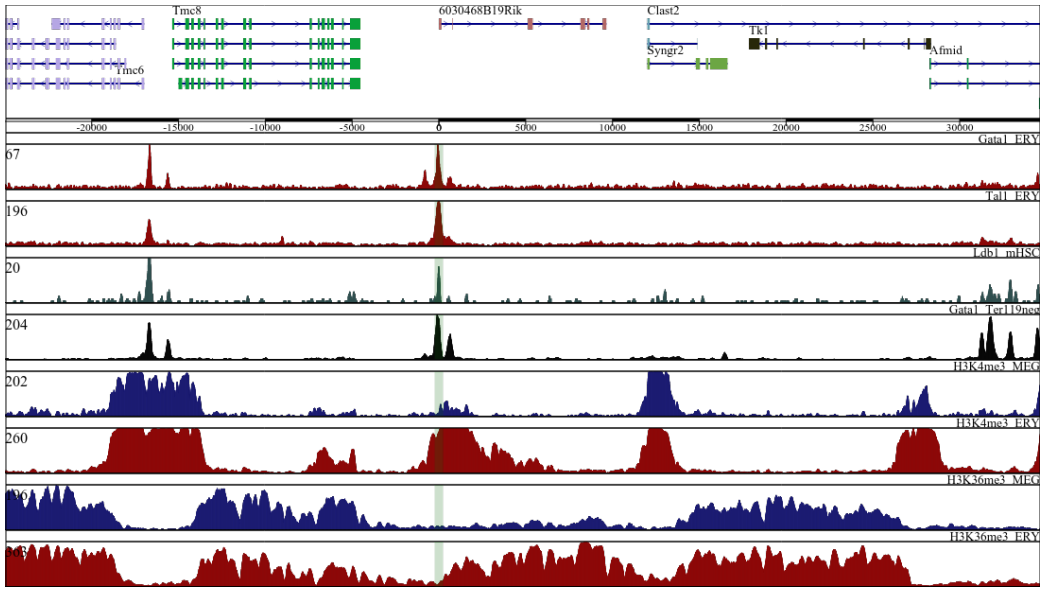
"Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads"

"Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads"

Challenge: Not one single 'peak shape'...

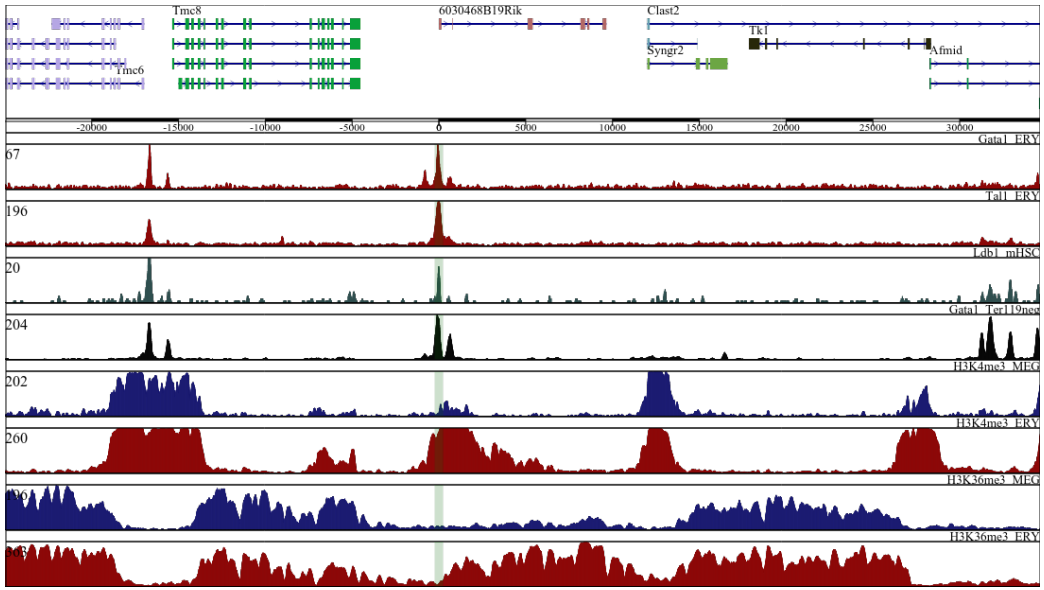
"Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads"

Challenge: Not one single 'peak shape'...



"Peak calling is a computational method used to identify areas in a genome that have been enriched with aligned reads"

Challenge: Not one single 'peak shape'...



Number of peaks
Peak distribution
Peak height
Range of measurements
Experimental variation (IP...)

Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...

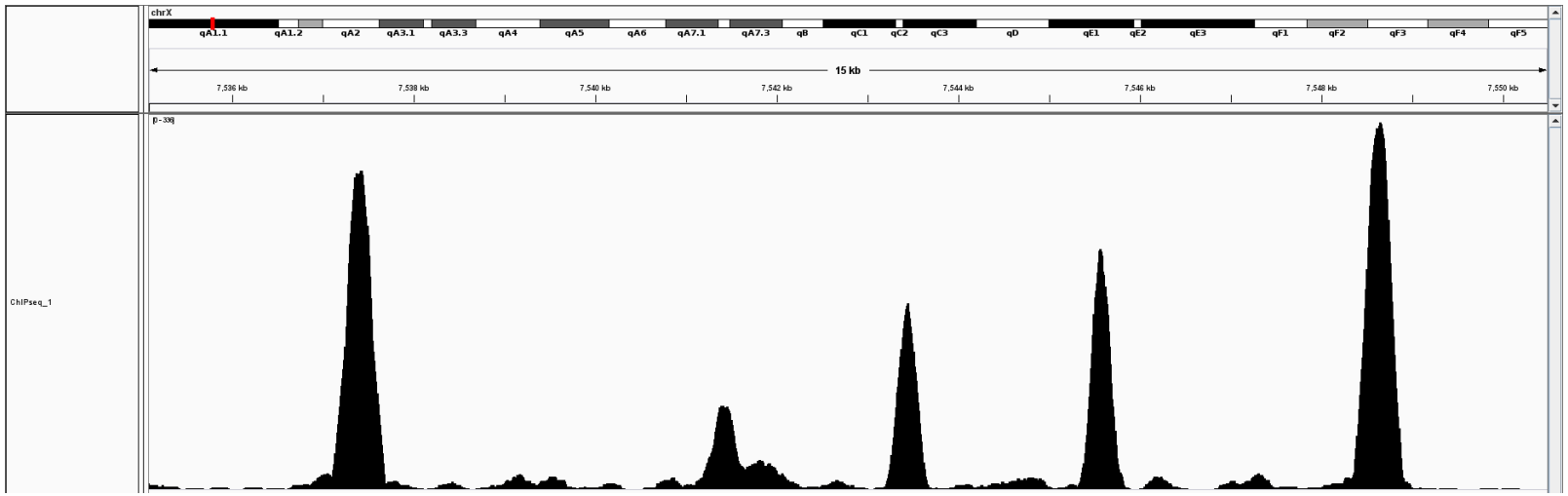
Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...



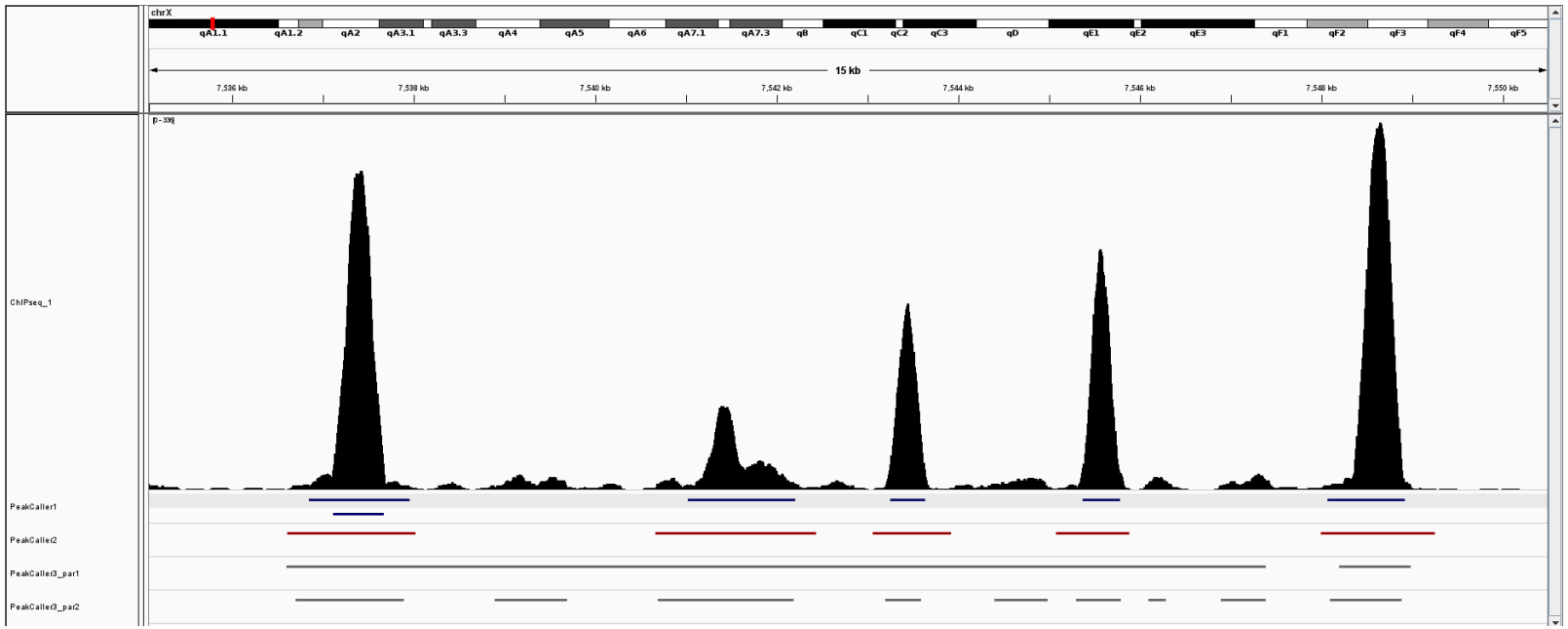
Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...



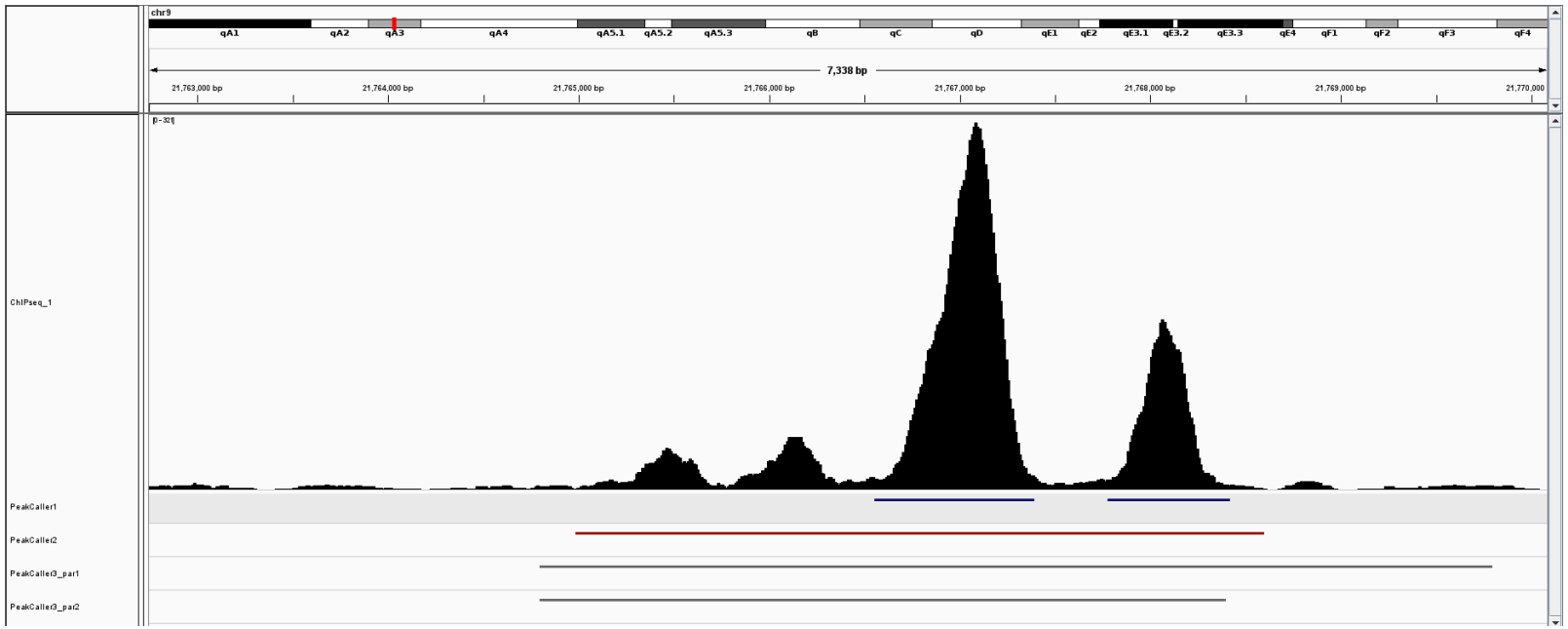
Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...



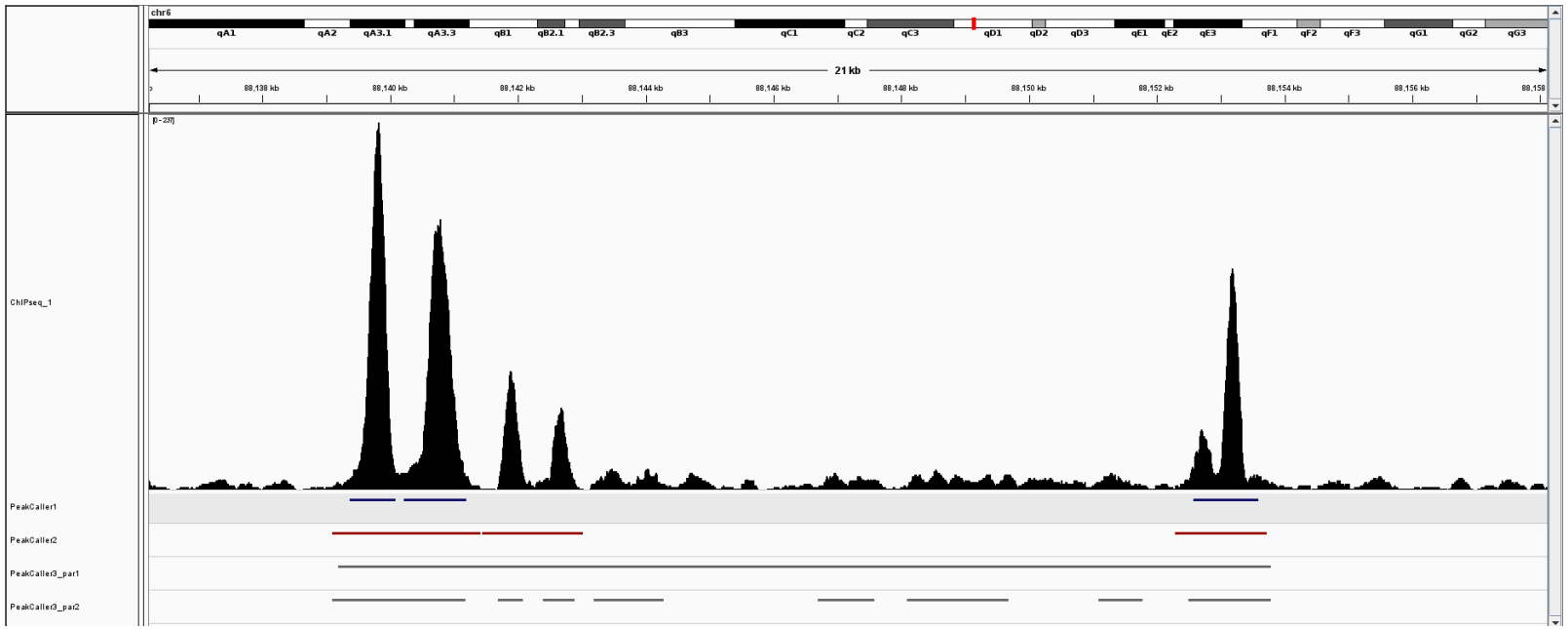
Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...



Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...

Peaks represent only a 'summary' of a ChIPseq experiment...

Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...

Peaks represent only a 'summary' of a ChIPseq experiment...

Useful metrics:

FRiP: Fraction of Reads in Peaks (>1%)

IDR: Irreproducible Discovery Rate

Significance: pValue, ChIPtoINPUT ratio, # of reads...

Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...

Peaks represent only a 'summary' of a ChIPseq experiment...

Empirical:

Known Motif: Enriched in peaks, Central Position

Distribution: Near TSS, Within GeneBody, Enhancers

Target Genes: Gene expression changes, Functional Annotation

Solutions:

Different Peak Callers

Optimal Parameters

Correct Experimental Design

Good ChIP...

Peaks represent only a 'summary' of a ChIPseq experiment...

Empirical:

Known Motif: Enriched in peaks, Central Position

Distribution: Near TSS, Within GeneBody, Enhancers

Target Genes: Gene expression changes, Functional Annotation

Browser inspection and previously known sites

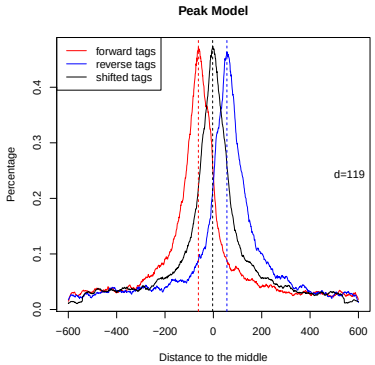
TFs: MACS
HMs: SICER

TFs: MACS
HMs: SICER

Running MACS: **macs14** **-t** ChIP **-c** Input **-g** mm **-n** Name

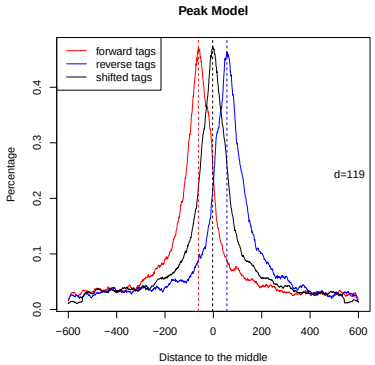
TFs: MACS
HMs: SICER

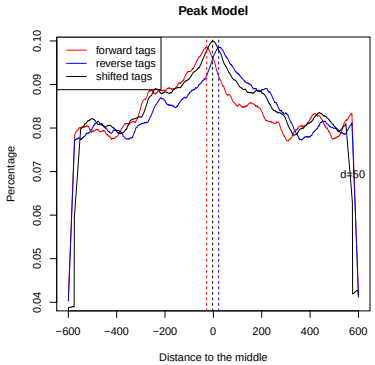
Running MACS: **macs14** -**t** ChIP -**c** Input -**g** mm -**n** Name



TFs: MACS
HMs: SICER

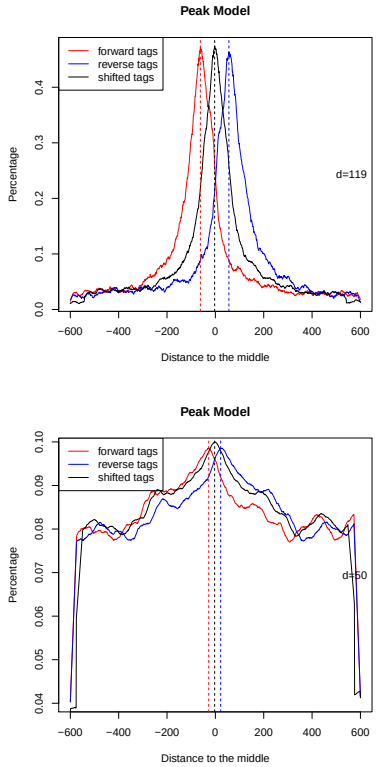
Running MACS: **macs14** -**t** ChIP -**c** Input -**g** mm -**n** Name





TFs: MACS
HMs: SICER

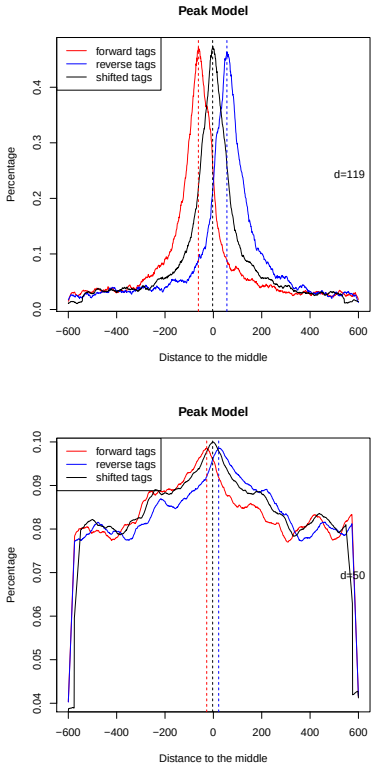
Running MACS: **macs14** -**t** ChIP -**c** Input -**g** mm -**n** Name



# This file is generated by MACS version 1.4.2 20120305								
# ARGUMENTS LIST:								
# name = Gata1_ERYneg								
# format = AUTO								
# ChIP-seq file = Gata1_ERYneg.bam								
# control file = Inputs/ERYter_neg_INPUT.bam								
# effective genome size = 1.87e+09								
# band width = 300								
# model fold = 10	30							
# pvalue cutoff = 1.00e-05								
# Large dataset will be scaled towards smaller dataset.								
# Range for calculating regional lambda is: 1000 bps and 10000 bps								
# tag size is determined as 51 bps								
# total tags in treatment: 22404993								
# tags after filtering in treatment: 19156302								
# maximum duplicate tags at the same position in treatment = 1								
# Redundant rate in treatment: 0.14								
# total tags in control: 10891587								
# tags after filtering in control: 10250192								
# maximum duplicate tags at the same position in control = 1								
# Redundant rate in control: 0.06								
# d = 188								
chr	start	end	length	summit	tags	-10*log10(pvalue)	fold_enrichment	FDR(%)
chr1	3042341	3043008	668	472	38	87.16	30.2	2.67
chr1	3049560	3050013	454	267	47	173.5	41.73	0.57
chr1	3435584	3436610	1027	551	146	414.38	61.7	0.33

TFs: MACS
HMs: SICER

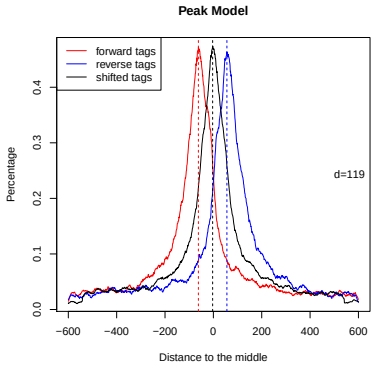
Running MACS: **macs14** -**t** ChIP -**c** Input -**g** mm -**n** Name

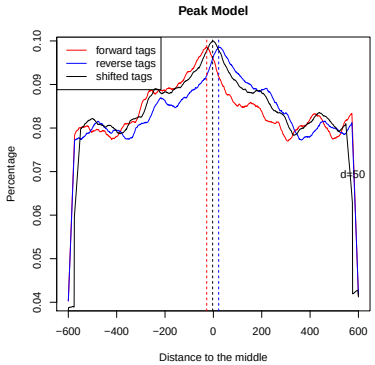


Chromosome	Start	Stop	ID	Score
chr1	3042340	3043008	MACS_peak_1	87.16
chr1	3049559	3050013	MACS_peak_2	173.5
chr1	3435583	3436610	MACS_peak_3	414.38
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
chrX	165596178	165596764	MACS_peak_26347	271.4
chrX	165658009	165658680	MACS_peak_26348	253.88

TFs: MACS
HMs: SICER

Running MACS: **macs14** **-t** ChIP **-c** Input **-g** mm **-n** Name





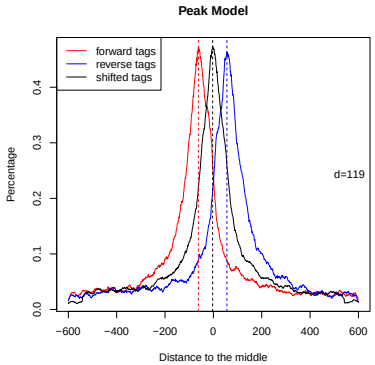
WIG file: Normalized Read Density text format

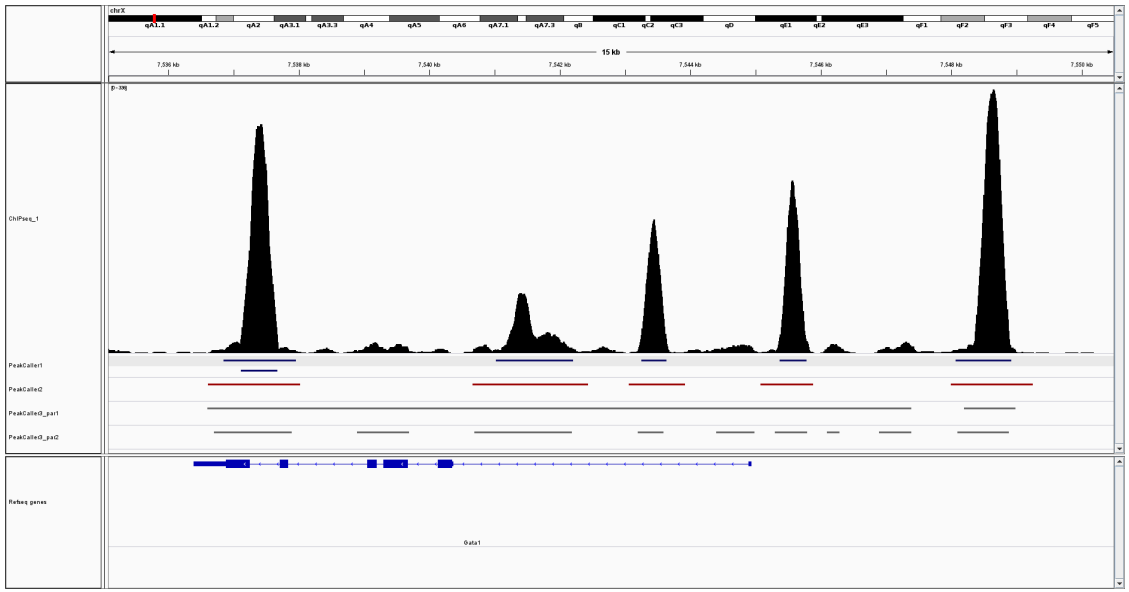
⇓

bigWig file: Binary Read Density format

TFs: MACS
HMs: SICER

Running MACS: **macs14** **-t** ChIP **-c** Input **-g** mm **-n** Name





General Guidelines:

Biological Replicates (2x)

INPUT (noIP sample)

20-25 mln reads per sample

Optimize ChIP conditions before library preparation

Minimize experimental variation

Look at the data

General Guidelines:

Biological Replicates (2x)

INPUT (noIP sample)

20-25 mln reads per sample

Optimize ChIP conditions before library preparation

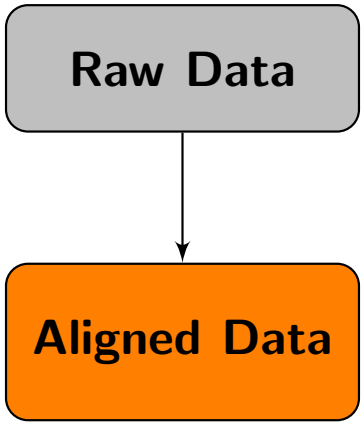
Minimize experimental variation

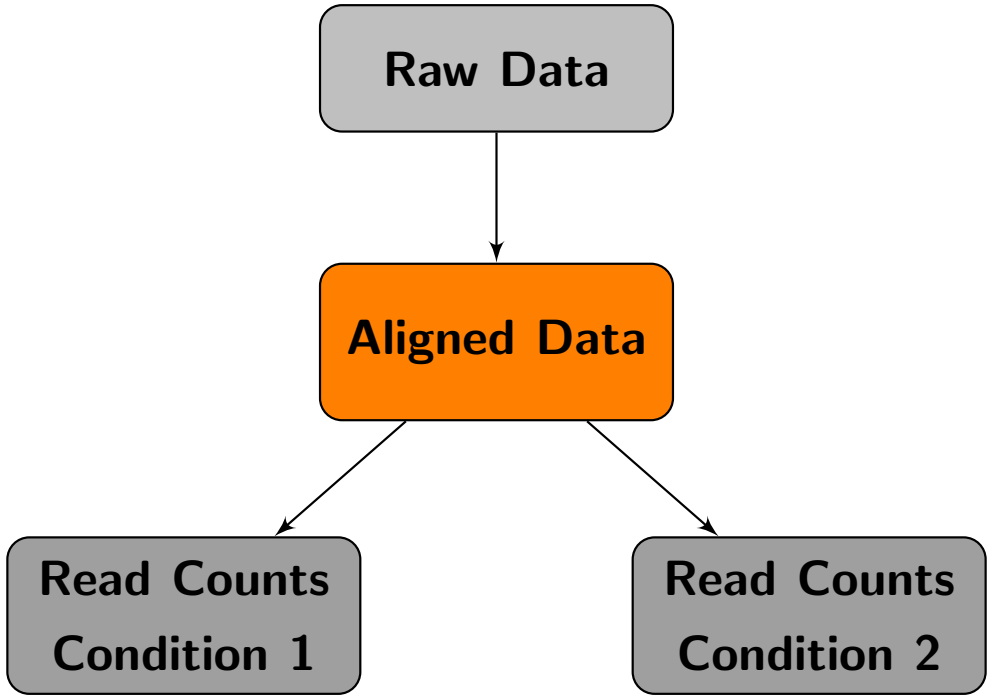
Look at the data

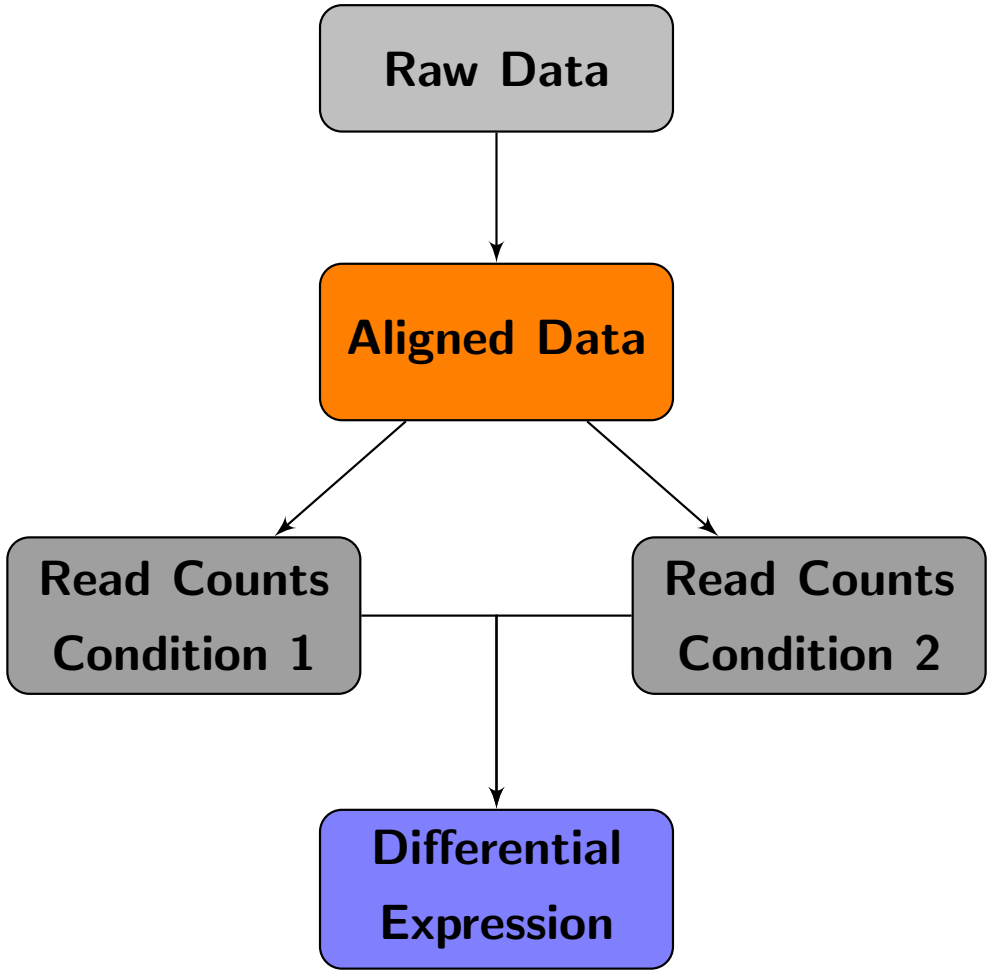
'ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia'
Landt et al, Genome Res, 2012

'Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data'
Bailey et al, PLoS Comp Biol, 2013

Raw Data


Raw Data
Aligned Data


Raw Data
Aligned Data
Read Counts Condition 1
Read Counts Condition 2


Raw Data
Aligned Data
Read Counts Condition 1
Read Counts Condition 2
Differential Expression

Aligner (Mapper)

Aligner (Mapper)

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
Spliced alignment

Build reference based on known transcripts

Aligner (Mapper)

STAR

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
Spliced alignment

Build reference based on known transcripts

Aligner (Mapper)

STAR

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
Spliced alignment

Build reference based on known transcripts

VERY Fast...
(45 million paired reads
per hour per processor)

Aligner (Mapper)

STAR

Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
Spliced alignment

Build reference based on known transcripts

VERY Fast...
(45 million paired reads per hour per processor)

TopHat2

Aligner (Mapper)

STAR

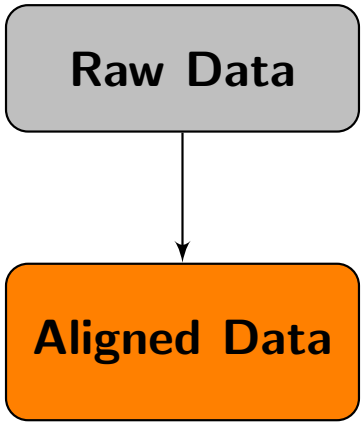
Reference Genome
Human (hg18, hg19)
Mouse (mm9, mm10)
...
Spliced alignment

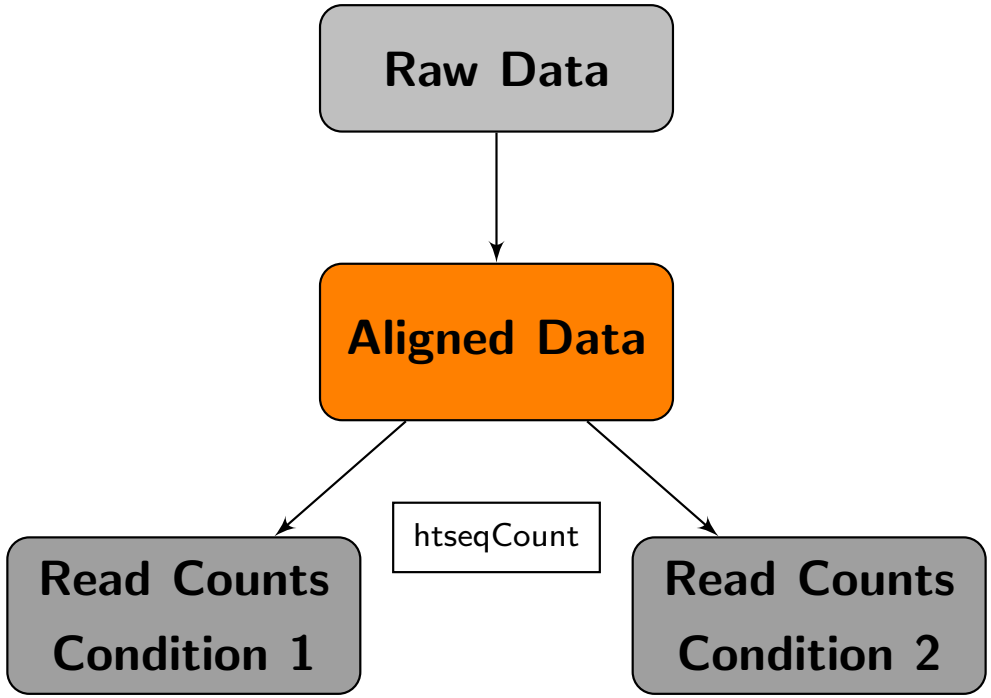
Build reference based on known transcripts

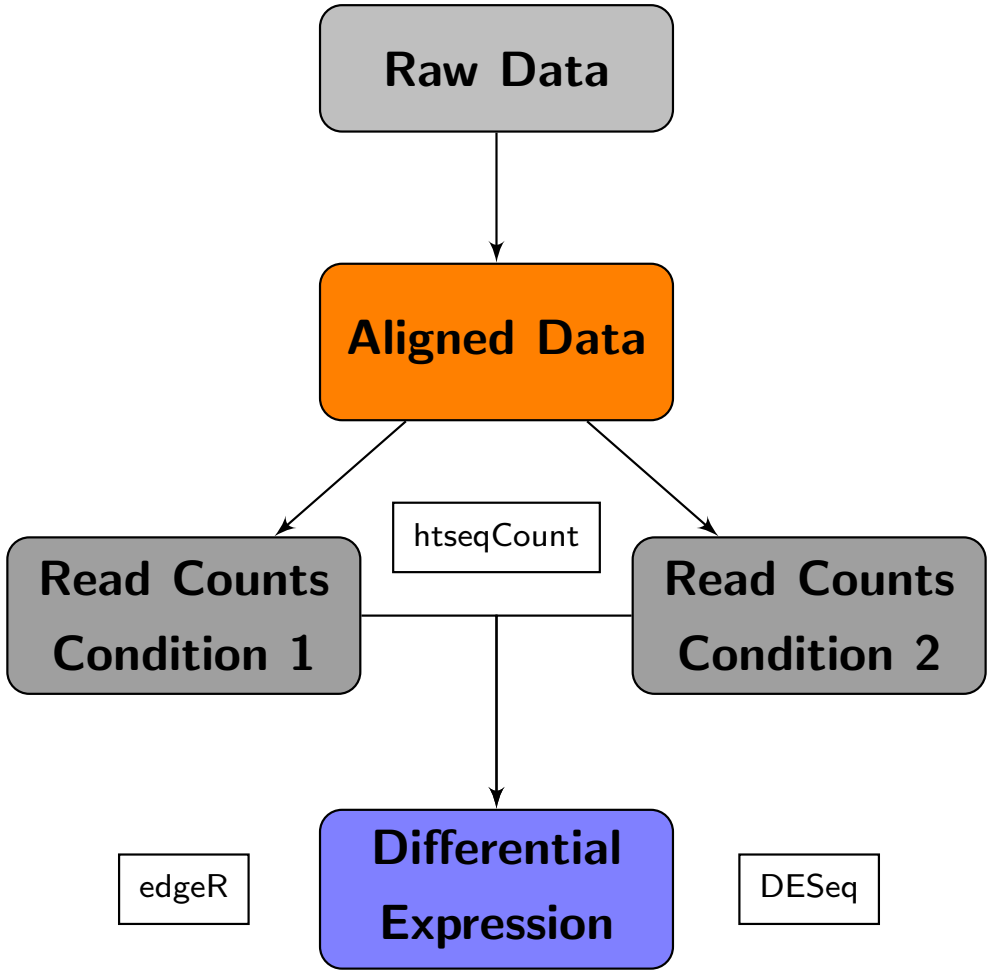
VERY Fast...
(45 million paired reads per hour per processor)

TopHat2

If you want to use CuffLinks...

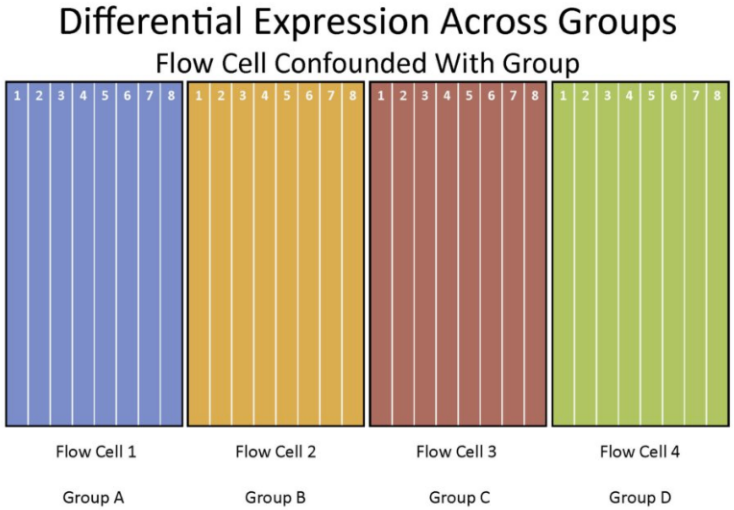

Raw Data
Aligned Data


Raw Data
Aligned Data
htseqCount
Read Counts Condition 1
Read Counts Condition 2

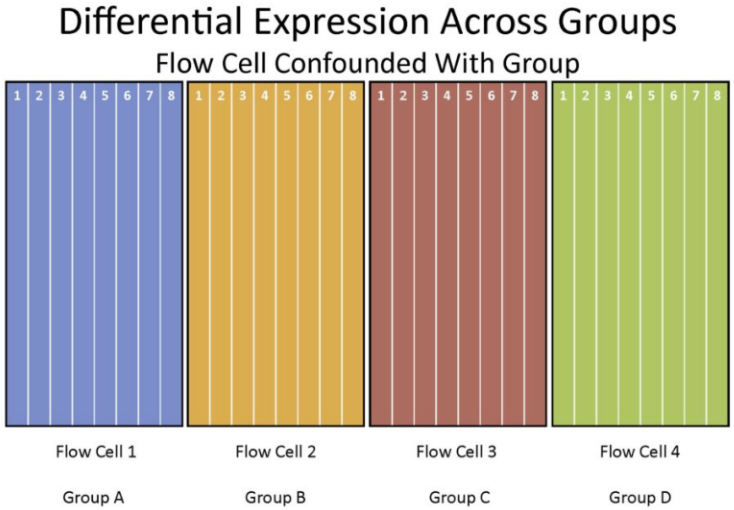

Raw Data
Aligned Data
htseqCount
Read Counts Condition 1
Read Counts Condition 2
edgeR
Differential Expression
DESeq

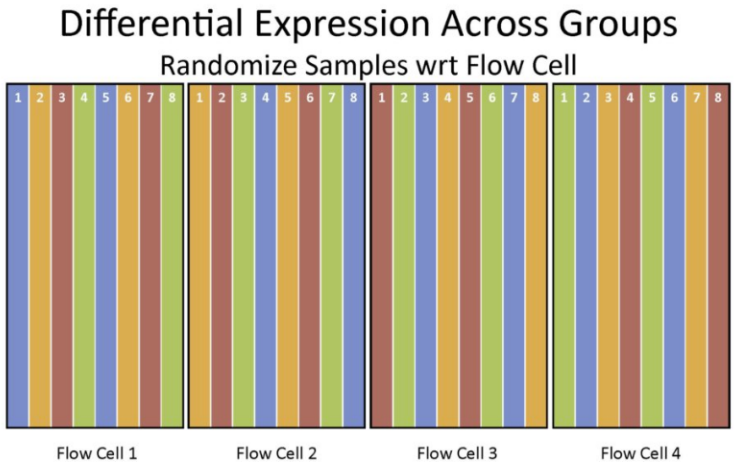
Raw Data

Raw Data

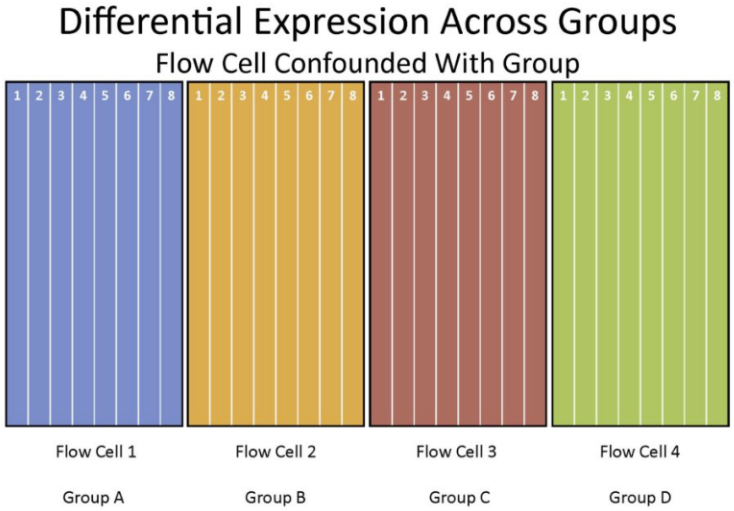

Differential Expression Across Groups
Flow Cell Confounded With Group
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
Flow Cell 1
Flow Cell 2
Flow Cell 3
Flow Cell 4
Group A
Group B
Group C
Group D

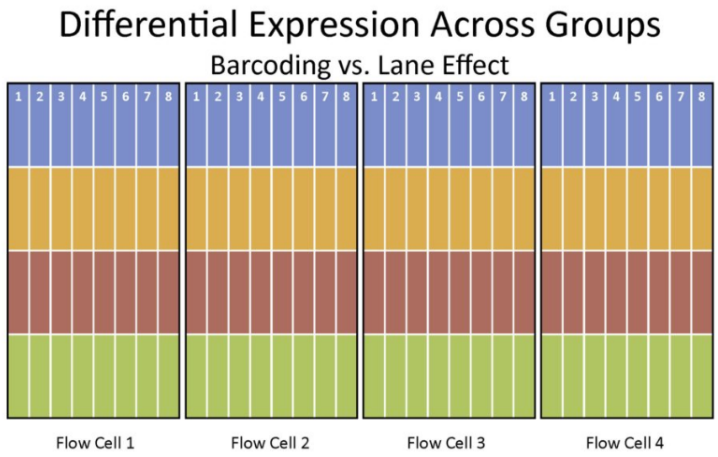
Raw Data


Differential Expression Across Groups
Flow Cell Confounded With Group
1 2 3 4 5 6 7 8
Flow Cell 1
Group A
Flow Cell 2
Group B
Flow Cell 3
Group C
Flow Cell 4
Group D


Differential Expression Across Groups
Randomize Samples wrt Flow Cell
1 2 3 4 5 6 7 8
Flow Cell 1
Flow Cell 2
Flow Cell 3
Flow Cell 4

Raw Data


Differential Expression Across Groups
Flow Cell Confounded With Group
1 2 3 4 5 6 7 8
Flow Cell 1
Flow Cell 2
Flow Cell 3
Flow Cell 4
Group A
Group B
Group C
Group D


Differential Expression Across Groups
Barcoding vs. Lane Effect
1 2 3 4 5 6 7 8
Flow Cell 1
Flow Cell 2
Flow Cell 3
Flow Cell 4

Local

Local

Online

Local

Online

IGV (Integrative Genomics Viewer)

Local

IGV (Integrative Genomics Viewer)

Online

UCSC Genome Browser

Local

IGV (Integrative Genomics Viewer)

Cross Platform (Win, Mac, Linux)
Different file formats
Fast

Online

UCSC Genome Browser

Local

IGV (Integrative Genomics Viewer)

Cross Platform (Win, Mac, Linux)
Different file formats
Fast

Online

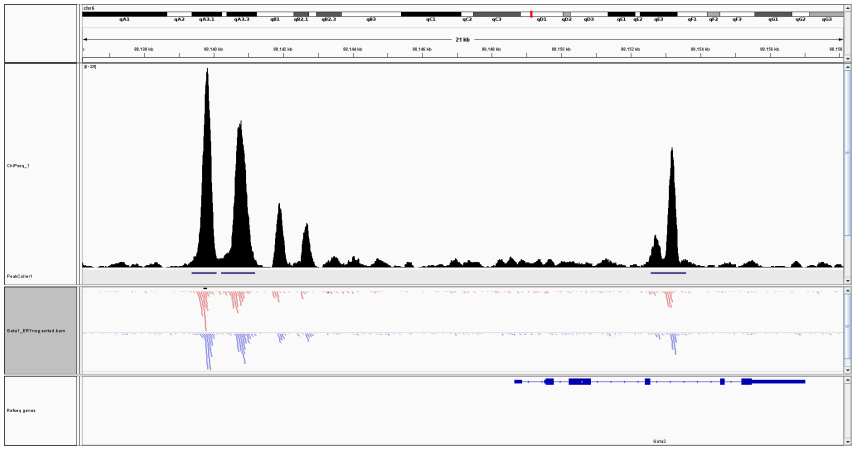
UCSC Genome Browser

Web based
Slow
Comprehensive database

Local

IGV (Integrative Genomics Viewer)

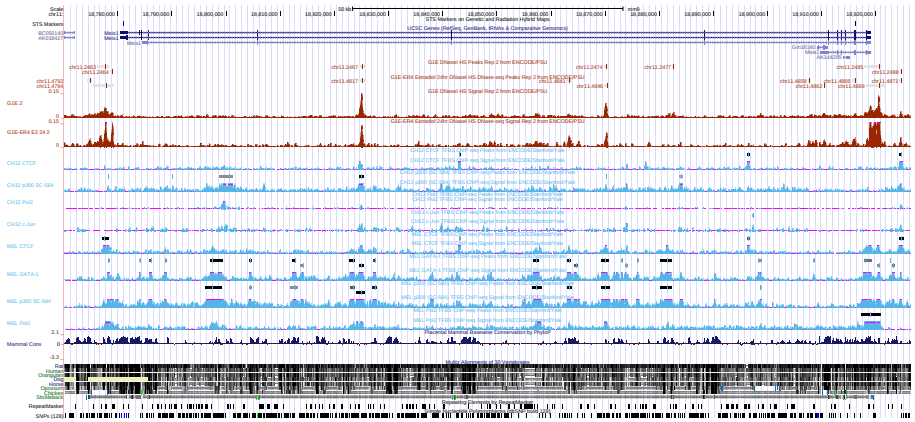
Cross Platform (Win, Mac, Linux)
Different file formats
Fast



Online

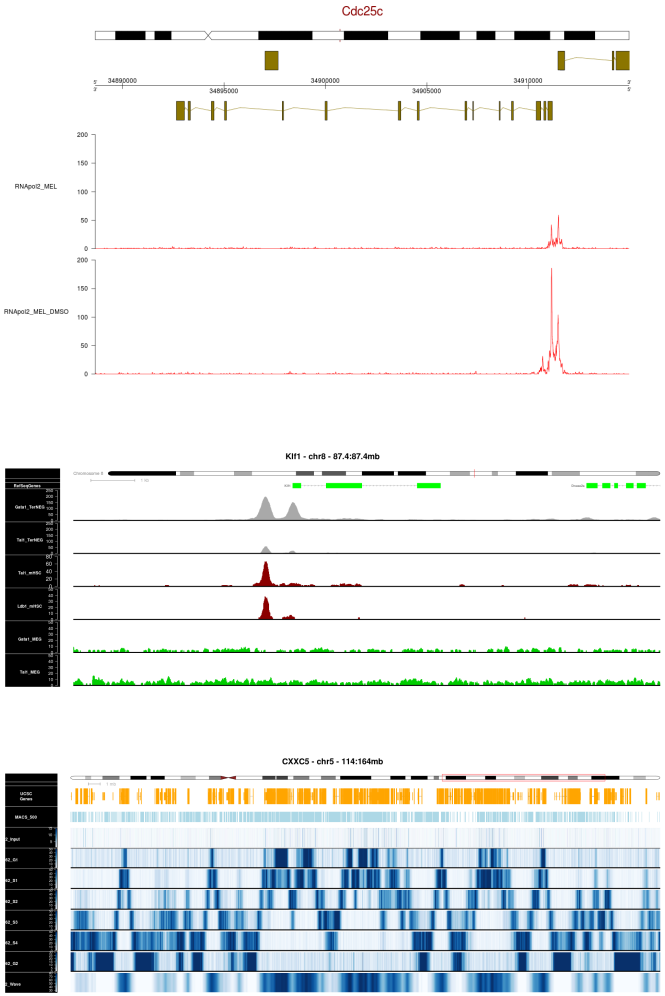
UCSC Genome Browser

Web based
Slow
Comprehensive database



Local

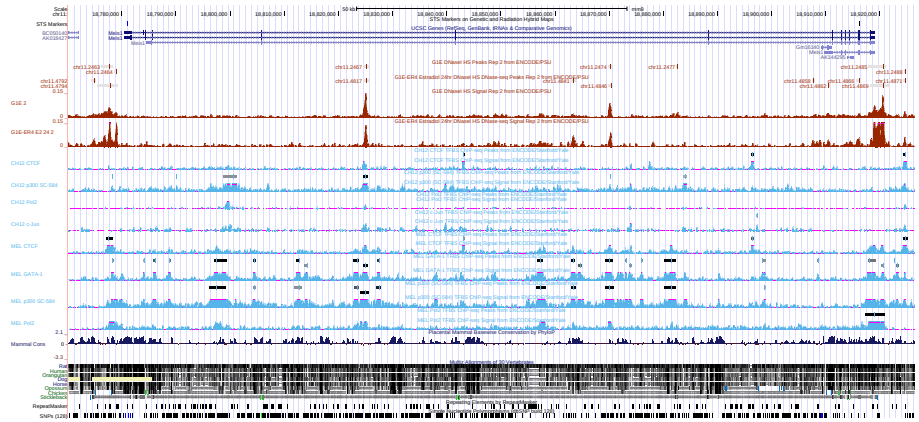
GenomeGraphs (R package)



Online

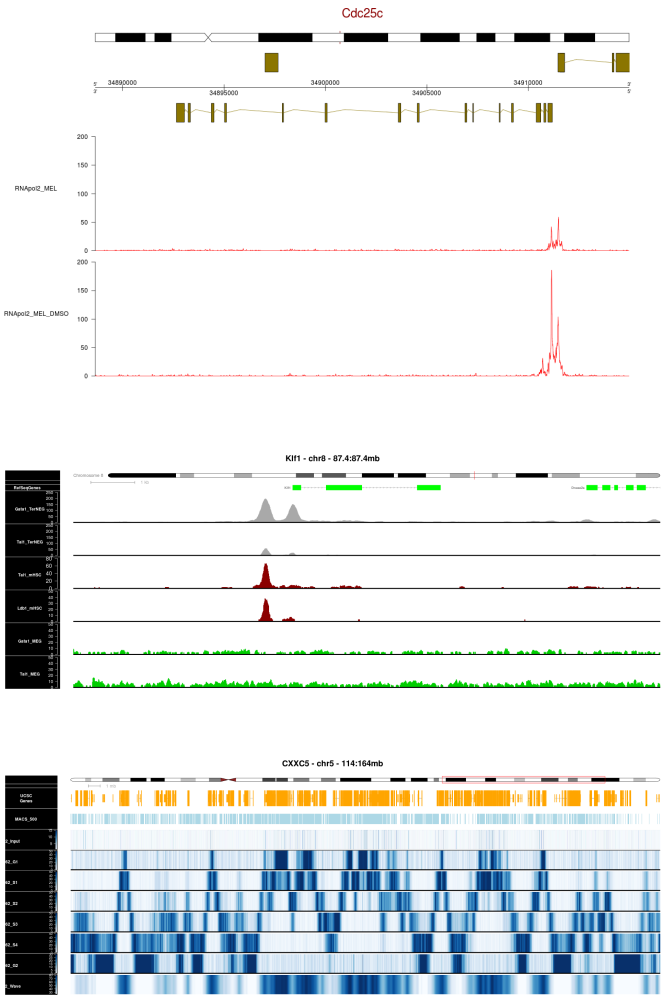
UCSC Genome Browser

Web based
Slow
Comprehensive database



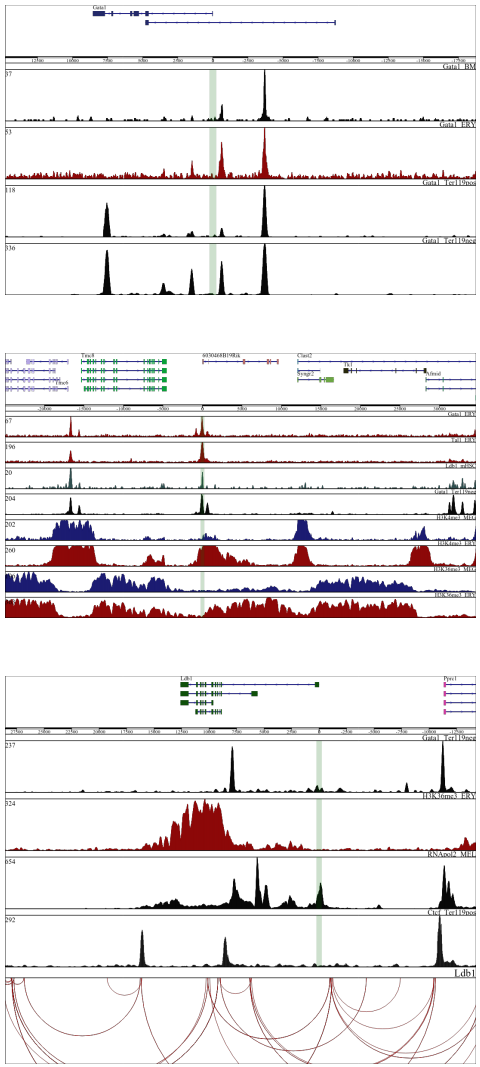
Local

GenomeGraphs (R package)



Online

Ariadne (GeneViewer)



GALAXY server

GALAXY server

Online collection of widely used bioinformatics tools (e.g. FastQC, bowtie.2, CuffLinks, MACS, SICER ...)

GALAXY server

Galaxy

Online collection of widely used bioinformatics tools (e.g. FastQC, bowtie.2, CuffLinks, MACS, SICER ...)

Learn Galaxy...

GALAXY server

Online collection of widely used bioinformatics tools (e.g. FastQC, bowtie.2, CuffLinks, MACS, SICER ...)

Learn Galaxy...

`https://galaxyproject.org/`

GALAXY server

Online collection of widely used bioinformatics tools (e.g. FastQC, bowtie.2, CuffLinks, MACS, SICER ...)

Learn Galaxy...

`https://galaxyproject.org/`

Chipster

GALAXY server

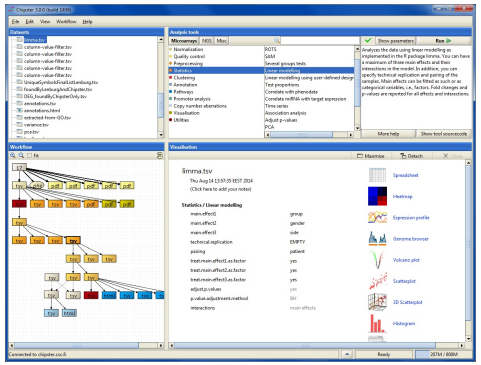
Online collection of widely used bioinformatics tools (e.g. FastQC, bowtie.2, CuffLinks, MACS, SICER ...)

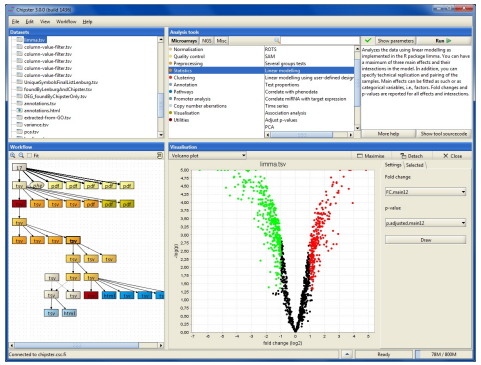
Learn Galaxy...

`https://galaxyproject.org/`

Chipster

User-friendly analysis software for high-throughput data (GUI)





Genomic Regions Enrichment of Annotations Tool (GREAT)

Genomic Regions Enrichment of Annotations Tool (GREAT)

Functional annotation of cis-regulatory regions

Genomic Regions Enrichment of Annotations Tool (GREAT)

Functional annotation of cis-regulatory regions

Input: BED file

Genomic Regions Enrichment of Annotations Tool (GREAT)

Functional annotation of
cis-regulatory regions

Input: BED file

Output:
List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

Genomic Regions Enrichment of Annotations Tool (GREAT)

Functional annotation of
cis-regulatory regions

Input: BED file

Output:
List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

Genomic Regions Enrichment of Annotations Tool (GREAT)

Functional annotation of
cis-regulatory regions

Input: BED file

Output:
List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

MEME-ChIP

Genomic Regions Enrichment of Annotations Tool (GREAT)

GREAT

Functional annotation of cis-regulatory regions

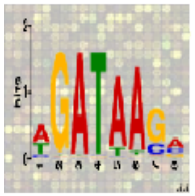
Input: BED file

Output:

List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

MEME-ChIP



Perform motif discovery, motif enrichment analysis and clustering

Genomic Regions Enrichment of Annotations Tool (GREAT)

GREAT

Functional annotation of cis-regulatory regions

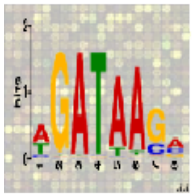
Input: BED file

Output:

List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

MEME-ChIP



Perform motif discovery, motif enrichment analysis and clustering

Input: FASTA file

Genomic Regions Enrichment of Annotations Tool (GREAT)

GREAT

Functional annotation of cis-regulatory regions

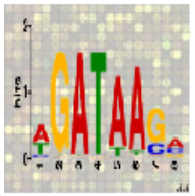
Input: BED file

Output:

List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

MEME-ChIP

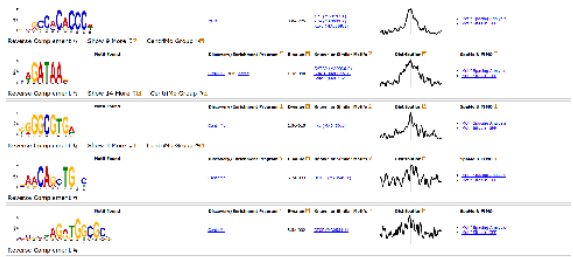


Perform motif discovery, motif enrichment analysis and clustering

Input: FASTA file

Output:

Enriched motifs
Motif distribution plots



Genomic Regions Enrichment of Annotations Tool (GREAT)

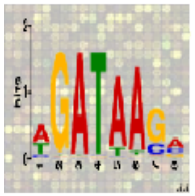
Functional annotation of cis-regulatory regions

Input: BED file

Output:
List of potential target genes
Peak distribution plots
Geneset enrichment analysis
(GO, Pathways, Phenotypes ...)

`http://bejerano.stanford.edu/great/public/html/index.php`

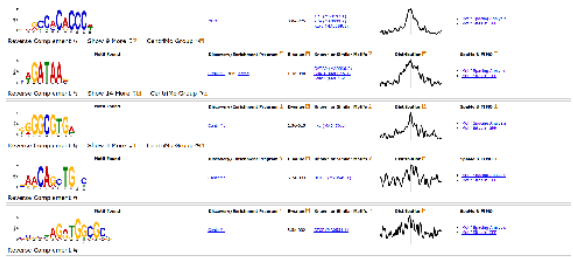
MEME-ChIP



Perform motif discovery, motif enrichment analysis and clustering

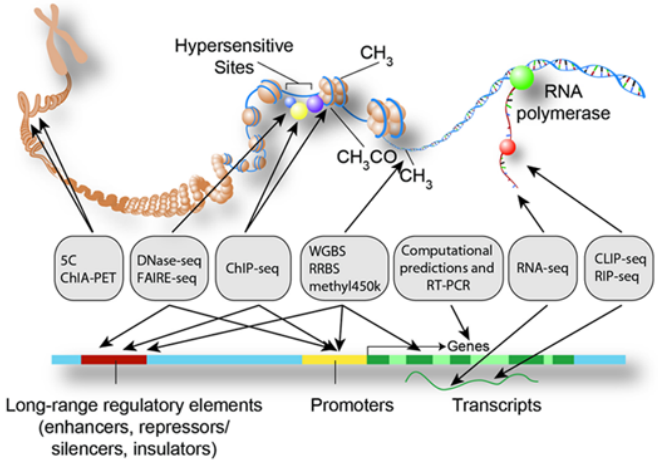
Input: FASTA file

Output:
Enriched motifs
Motif distribution plots



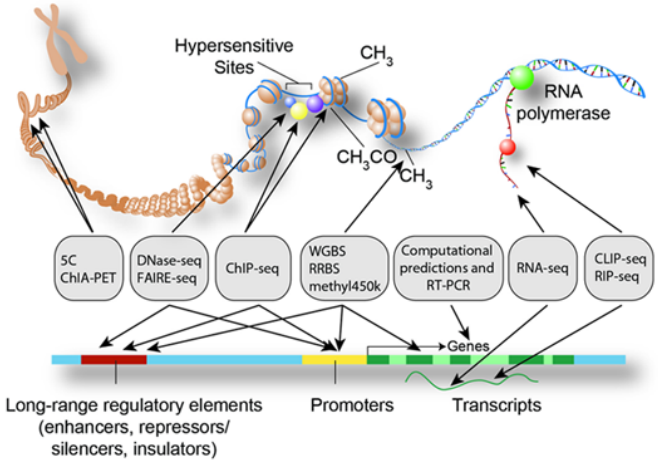
`http://meme-suite.org/tools/meme-chip`

Encyclopedia of DNA Elements



Aim: build a comprehensive parts list of functional elements in the human genome

Encyclopedia of DNA Elements

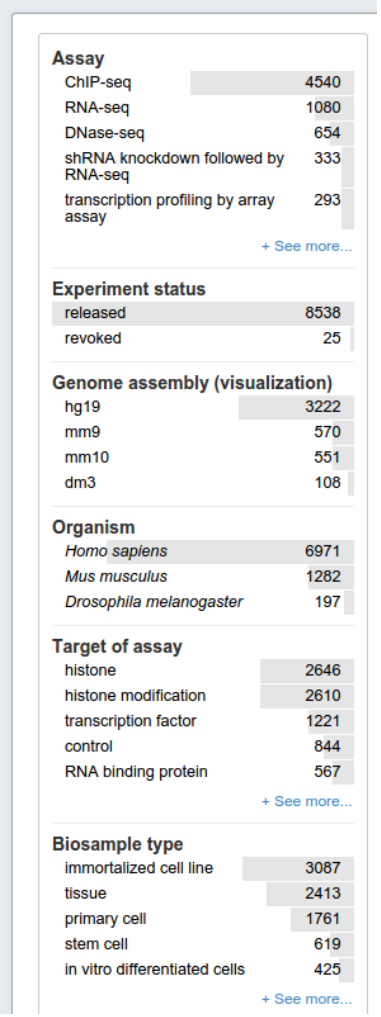


Aim: build a comprehensive parts list of functional elements in the human genome

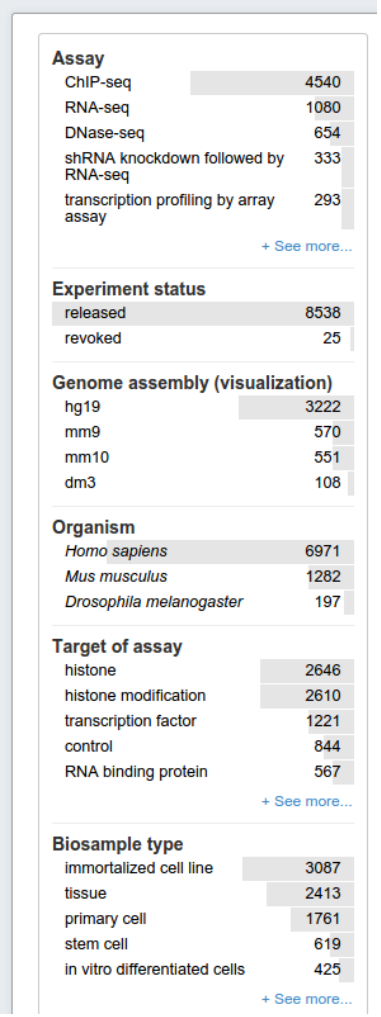
New website is extremely easy to use!

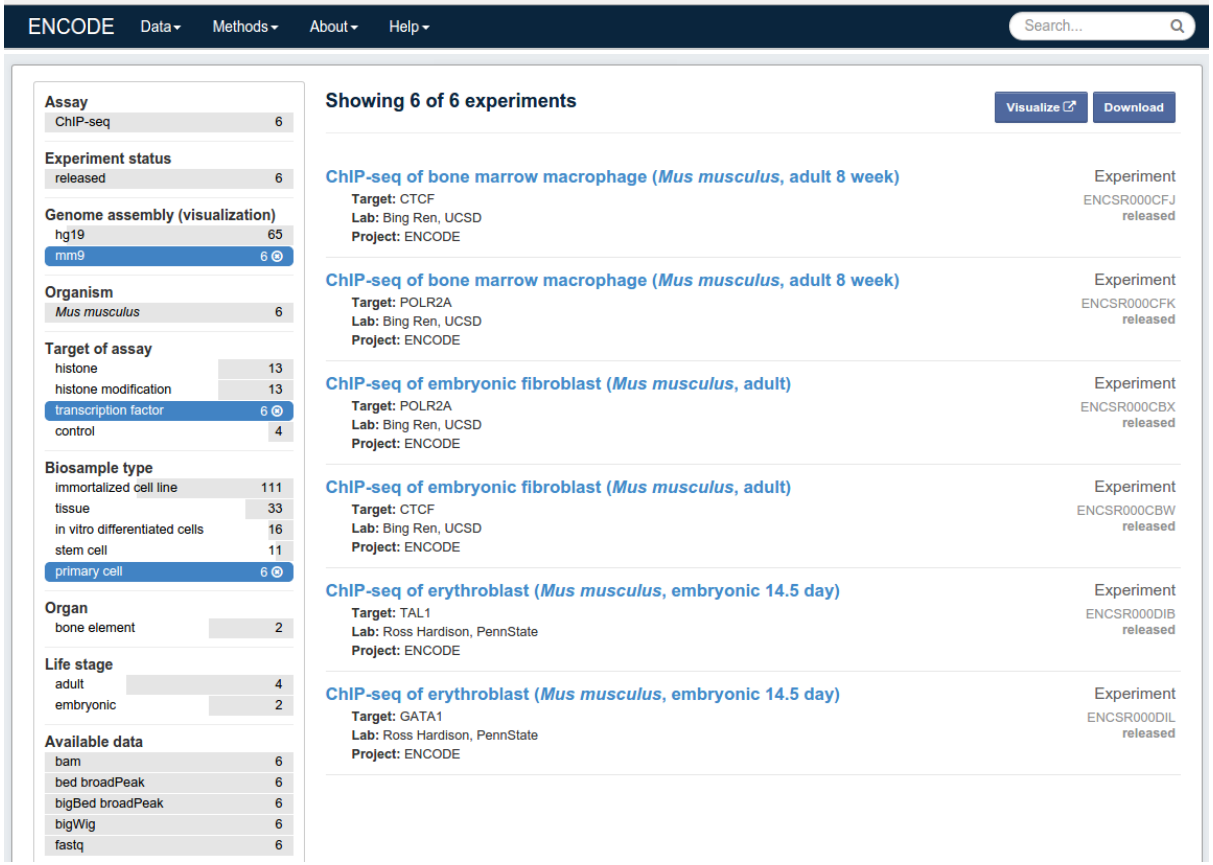
Encyclopedia of DNA Elements

Encyclopedia of DNA Elements

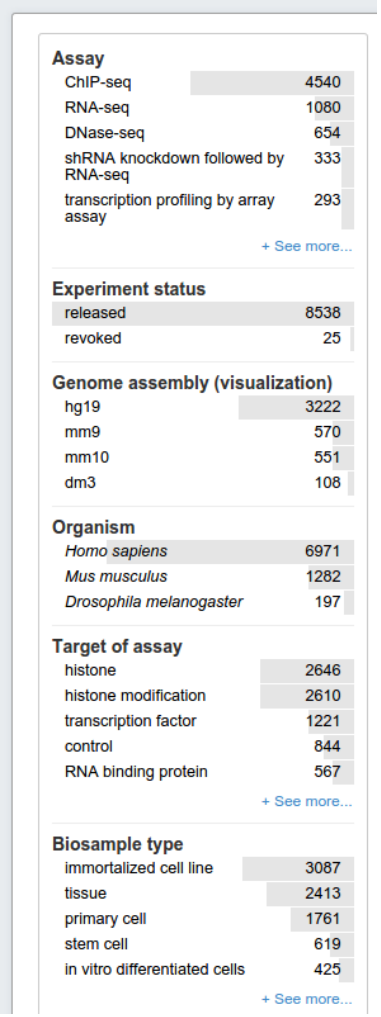


Encyclopedia of DNA Elements

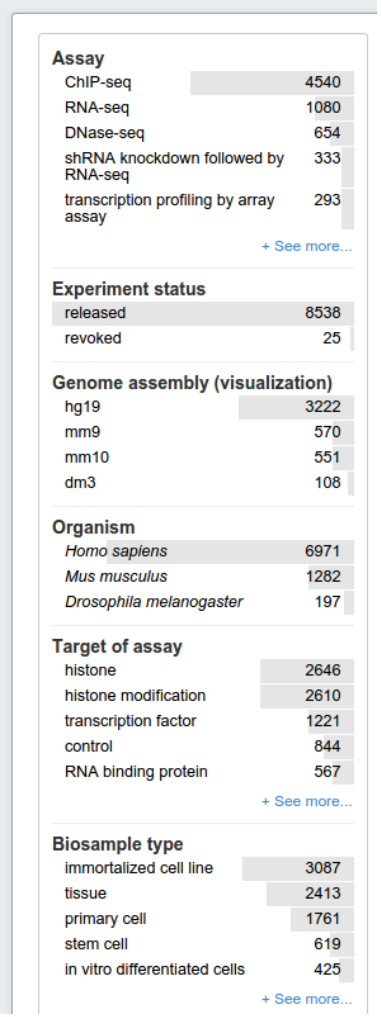


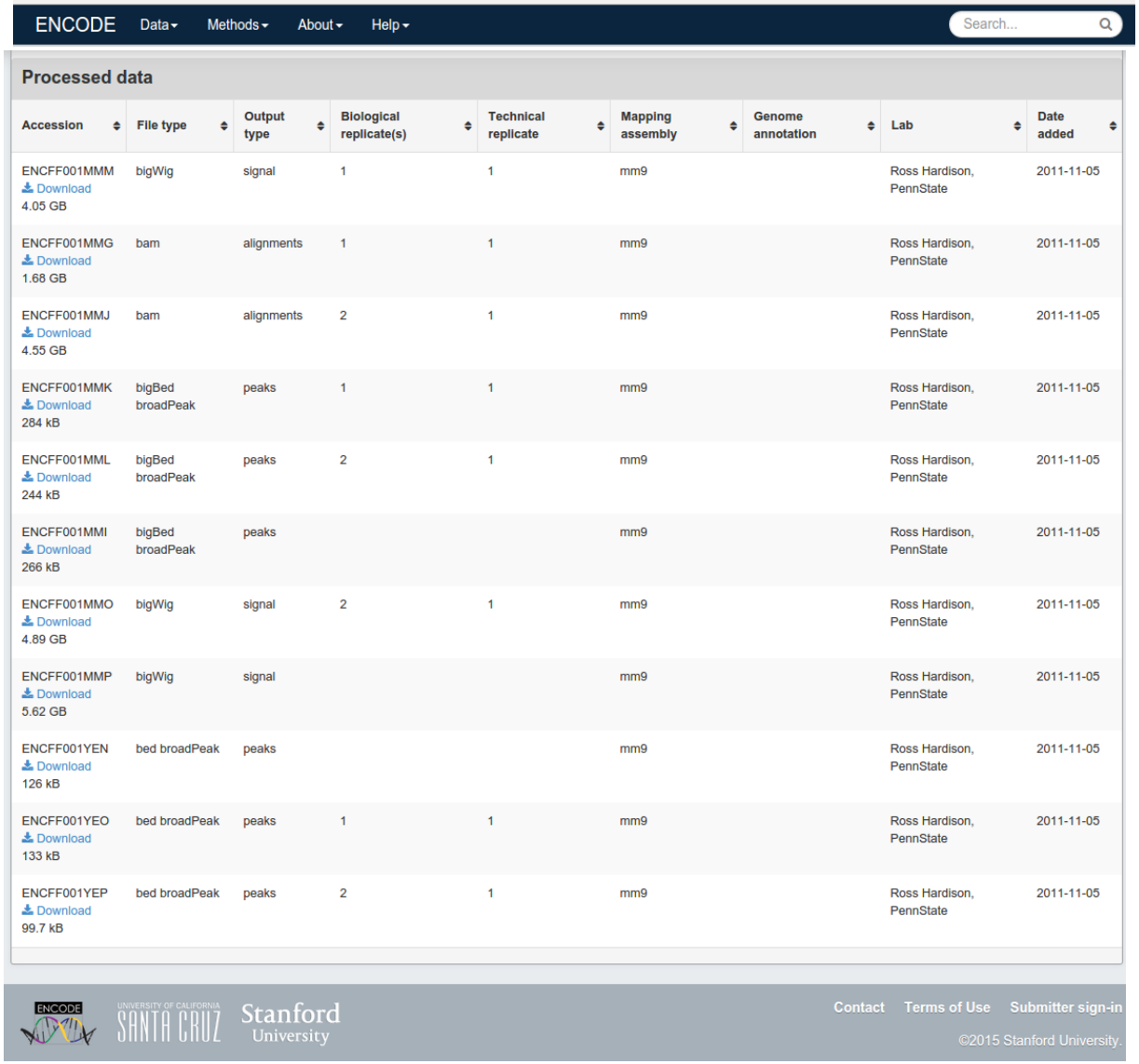


Encyclopedia of DNA Elements

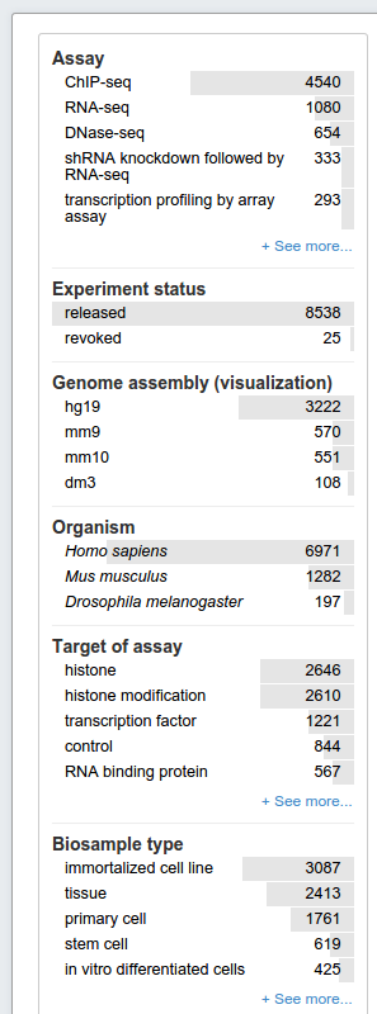


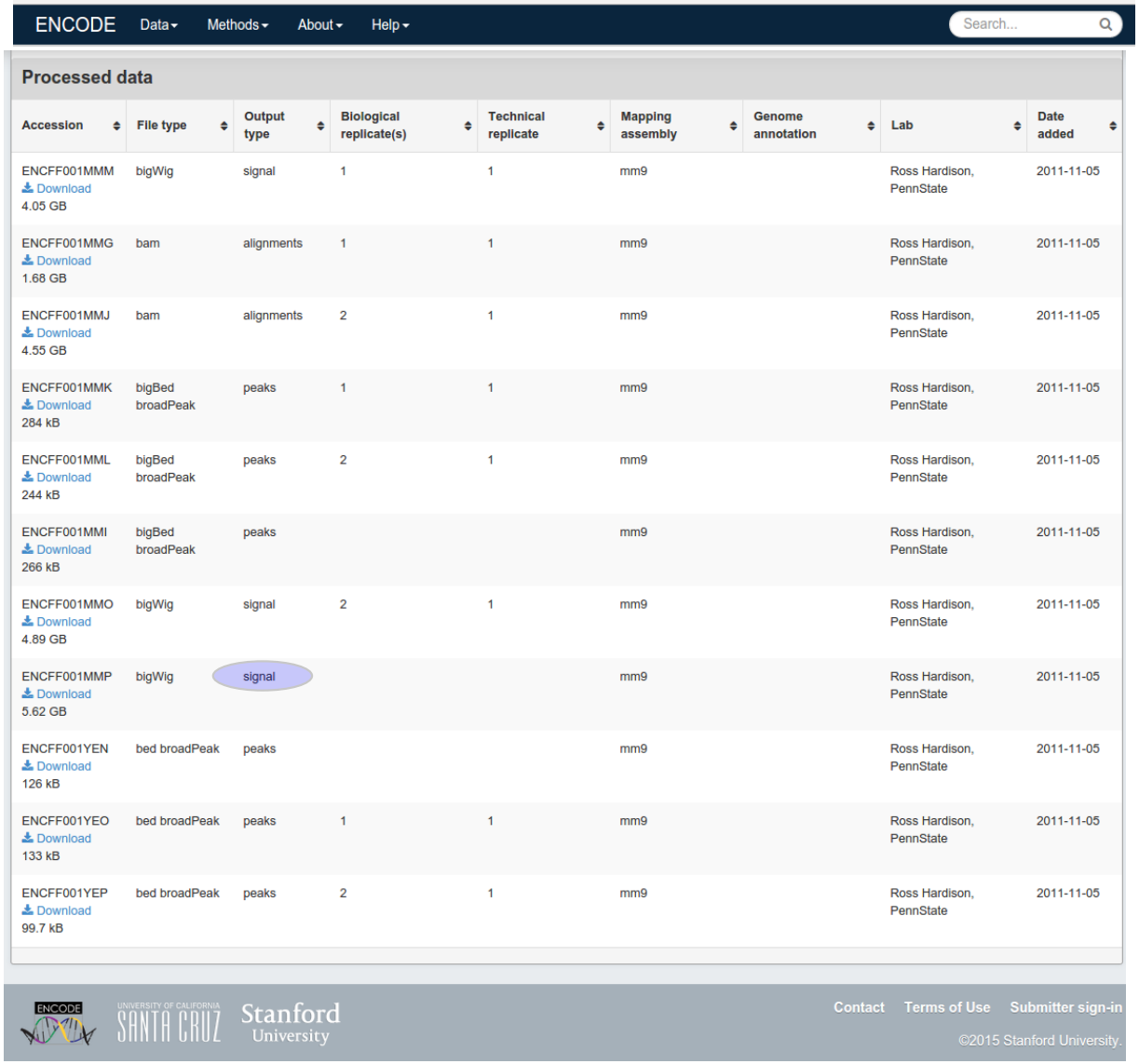
Encyclopedia of DNA Elements



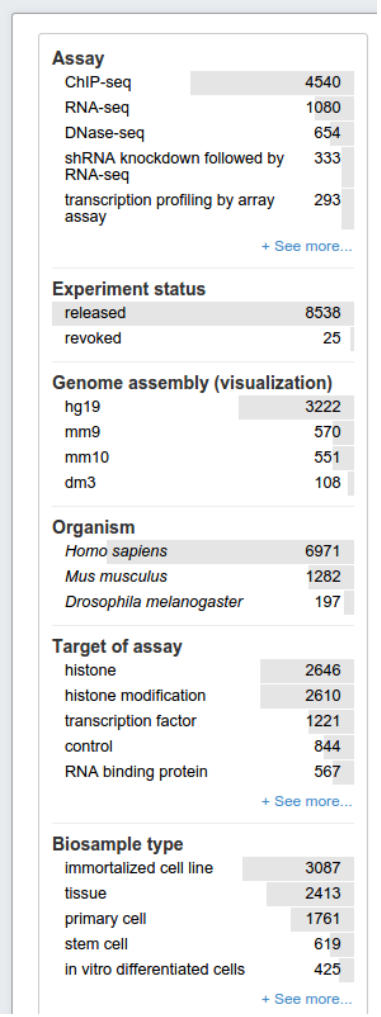


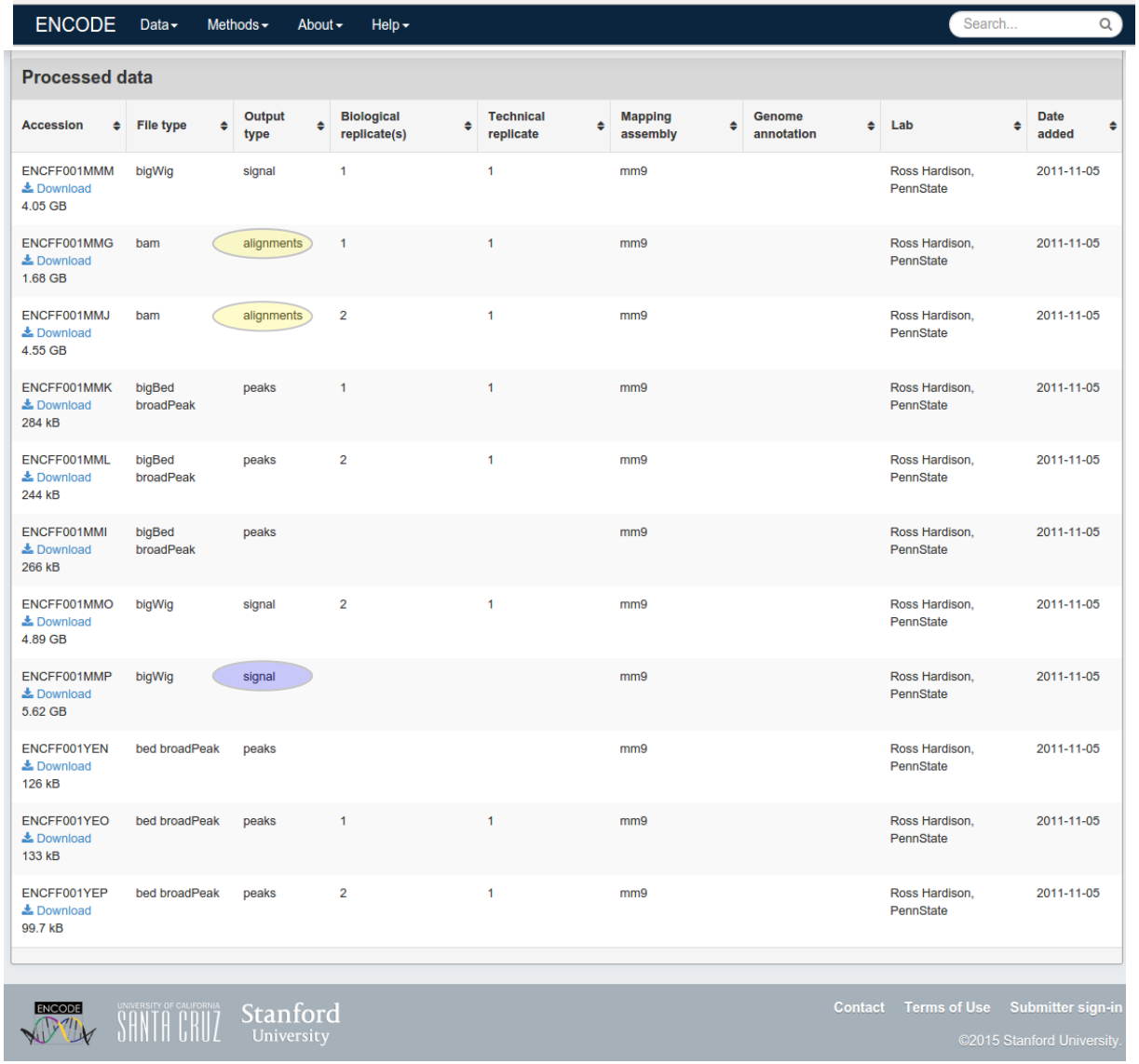
Encyclopedia of DNA Elements





Encyclopedia of DNA Elements





Encyclopedia of DNA Elements

Assay

ChIP-seq	4540
RNA-seq	1080
DNase-seq	654
shRNA knockdown followed by RNA-seq	333
transcription profiling by array assay	293

+ See more...

Experiment status

released	8538
revoked	25

Genome assembly (visualization)

hg19	3222
mm9	570
mm10	551
dm3	108

Organism

Homo sapiens	6971
Mus musculus	1282
Drosophila melanogaster	197

Target of assay

histone	2646
histone modification	2610
transcription factor	1221
control	844
RNA binding protein	567

+ See more...

Biosample type

immortalized cell line	3087
tissue	2413
primary cell	1761
stem cell	619
in vitro differentiated cells	425

+ See more...

ENCODE Data Methods About Help Search...

Processed data

Accession	File type	Output type	Biological replicate(s)	Technical replicate	Mapping assembly	Genome annotation	Lab	Date added
ENCFF001MMM Download 4.05 GB	bigWig	signal	1	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMG Download 1.68 GB	bam	alignments	1	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMJ Download 4.55 GB	bam	alignments	2	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMK Download 284 kB	bigBed broadPeak	peaks	1	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MML Download 244 kB	bigBed broadPeak	peaks	2	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMI Download 266 kB	bigBed broadPeak	peaks			mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMO Download 4.89 GB	bigWig	signal	2	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001MMP Download 5.62 GB	bigWig	signal			mm9		Ross Hardison, PennState	2011-11-05
ENCFF001YEN Download 126 kB	bed broadPeak	peaks			mm9		Ross Hardison, PennState	2011-11-05
ENCFF001YEO Download 133 kB	bed broadPeak	peaks	1	1	mm9		Ross Hardison, PennState	2011-11-05
ENCFF001YEP Download 99.7 kB	bed broadPeak	peaks	2	1	mm9		Ross Hardison, PennState	2011-11-05

ENCODE UNIVERSITY OF CALIFORNIA SANTA CRUZ Stanford University

Contact Terms of Use Submitter sign-in

Gene Expression Omnibus

Gene Expression Omnibus

GEO is a public functional genomics data repository

Gene Expression Omnibus

GEO is a public functional genomics data repository

All NGS datasets have to be deposited to GEO prior to publication

Gene Expression Omnibus

GEO is a public functional genomics data repository

All NGS datasets have to be deposited to GEO prior to publication

All datasets included in GEO are publicly available for academic use

Gene Expression Omnibus

GEO is a public functional genomics data repository

All NGS datasets have to be deposited to GEO prior to publication

All datasets included in GEO are publicly available for academic use

The search function is pretty close to awful

Gene Expression Omnibus

GEO is a public functional genomics data repository

All NGS datasets have to be deposited to GEO prior to publication

All datasets included in GEO are publicly available for academic use

The search function is pretty close to awful

Best way to query the GEO database is by Accession Number
(declared in the manuscript, GSE30142, GSM746581)

Gene Expression Omnibus

GEO is a public functional genomics data repository

All NGS datasets have to be deposited to GEO prior to publication

All datasets included in GEO are publicly available for academic use

The search function is pretty close to awful

Best way to query the GEO database is by Accession Number
(declared in the manuscript, GSE30142, GSM746581)

FTP access to SRA experiment (Sequence Read Archive)

European Nucleotide Archive

European Nucleotide Archive

ENA provides a comprehensive record of the world's nucleotide sequencing information

European Nucleotide Archive

ENA provides a comprehensive record of the world's nucleotide sequencing information

It covers raw sequencing data, sequence assembly information and functional annotation

European Nucleotide Archive

ENA provides a comprehensive record of the world's nucleotide sequencing information

It covers raw sequencing data, sequence assembly information and functional annotation

The search function is pretty close to awful

European Nucleotide Archive

ENA provides a comprehensive record of the world's nucleotide sequencing information

It covers raw sequencing data, sequence assembly information and functional annotation

The search function is pretty close to awful

Large overlap with GEO database

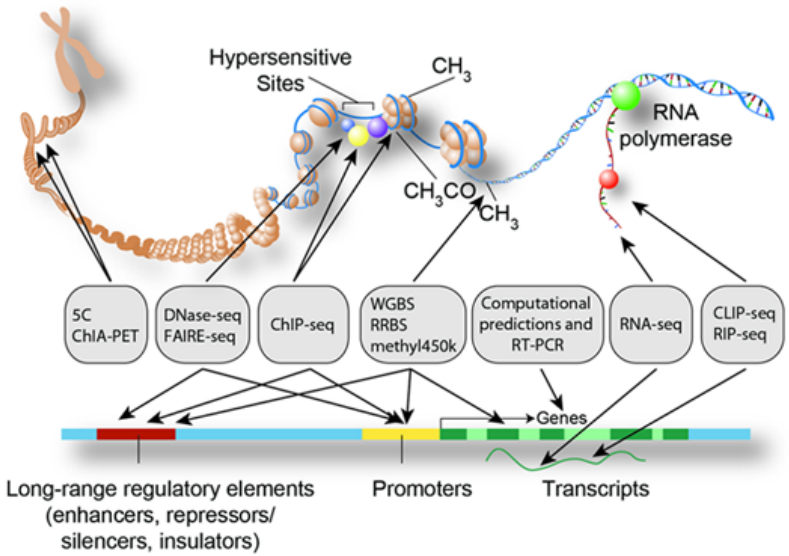
National Human Genome Research Institute

ENCODE

Gene Expression Omnibus

ENA
European Nucleotide Archive

Comprehensive database of genomic features



Systemic understanding
of biological processes

Thank you...