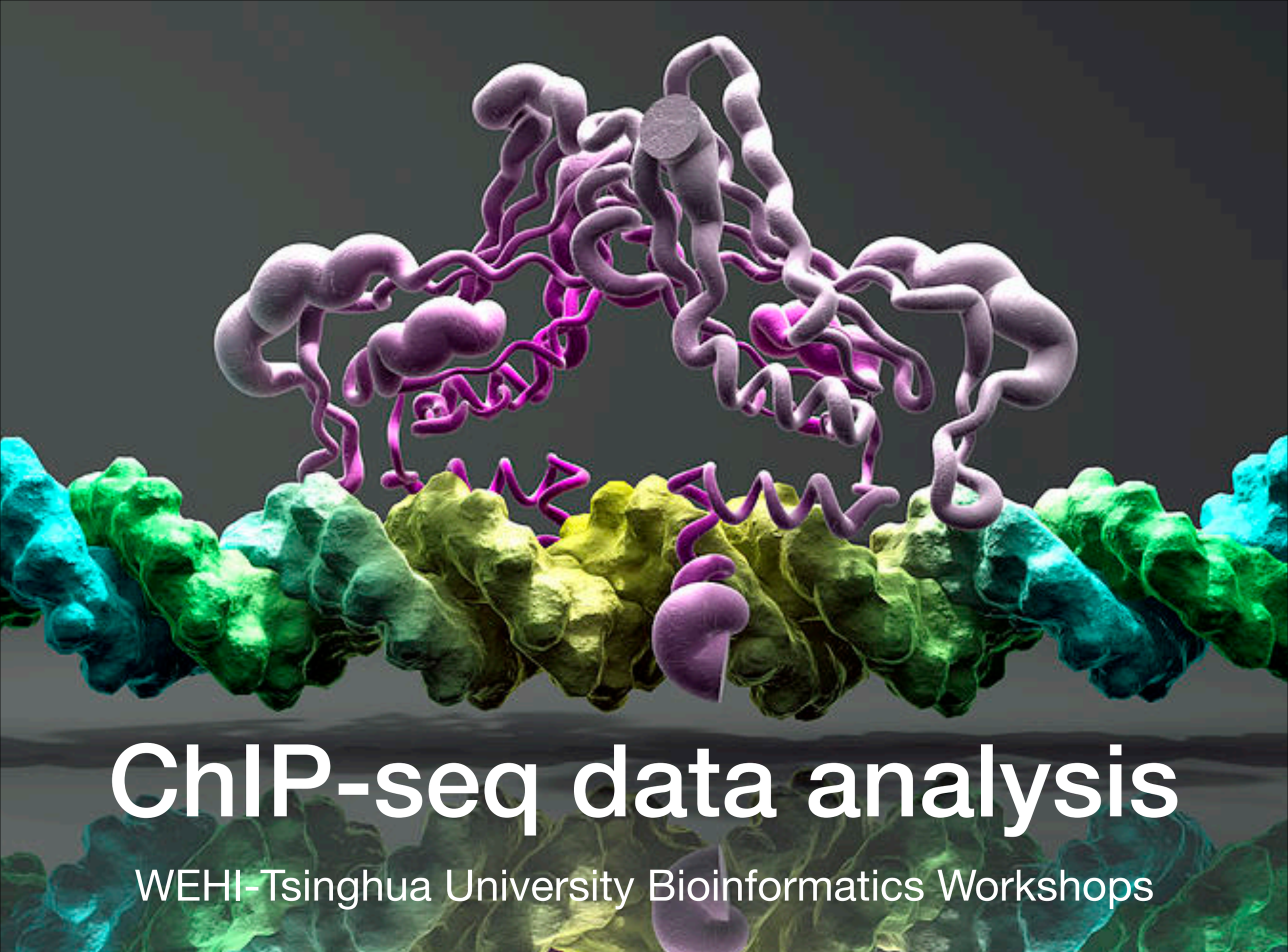


ChIP-seq data analysis

WEHI-Tsinghua University Bioinformatics Workshops

Topics

- What is ChIP-seq data?

- Data pre-processing

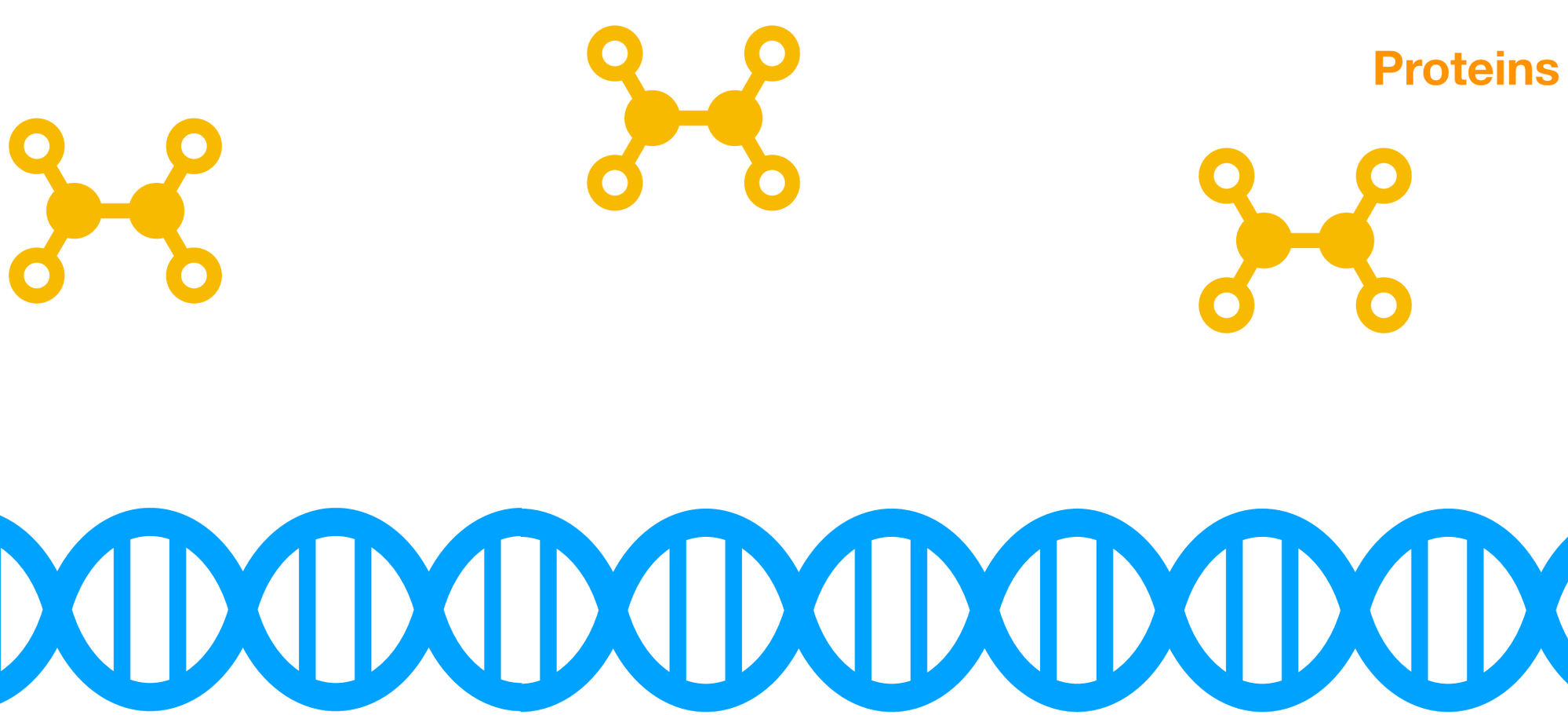
- Analysis pipelines
  - Peak calling
  - Differential binding

What is ChIP-seq data?

ChIP-seq

Chromatin **I**mmuno**P**recipitation **seq**uencing

Aim: To study the interaction between proteins and DNA in the cell nucleus



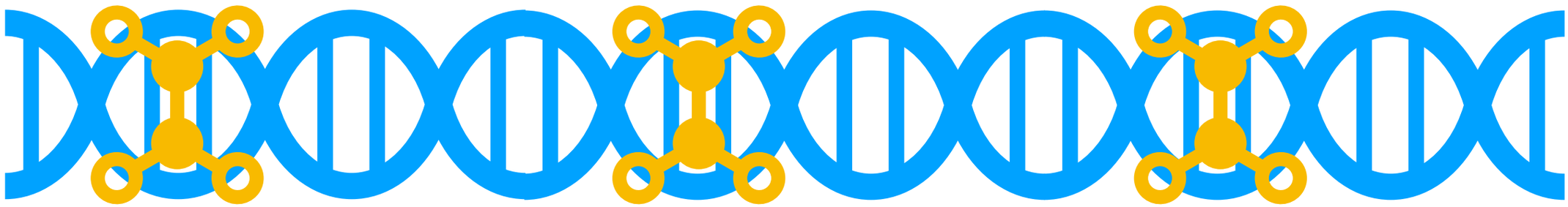
ChIP-seq

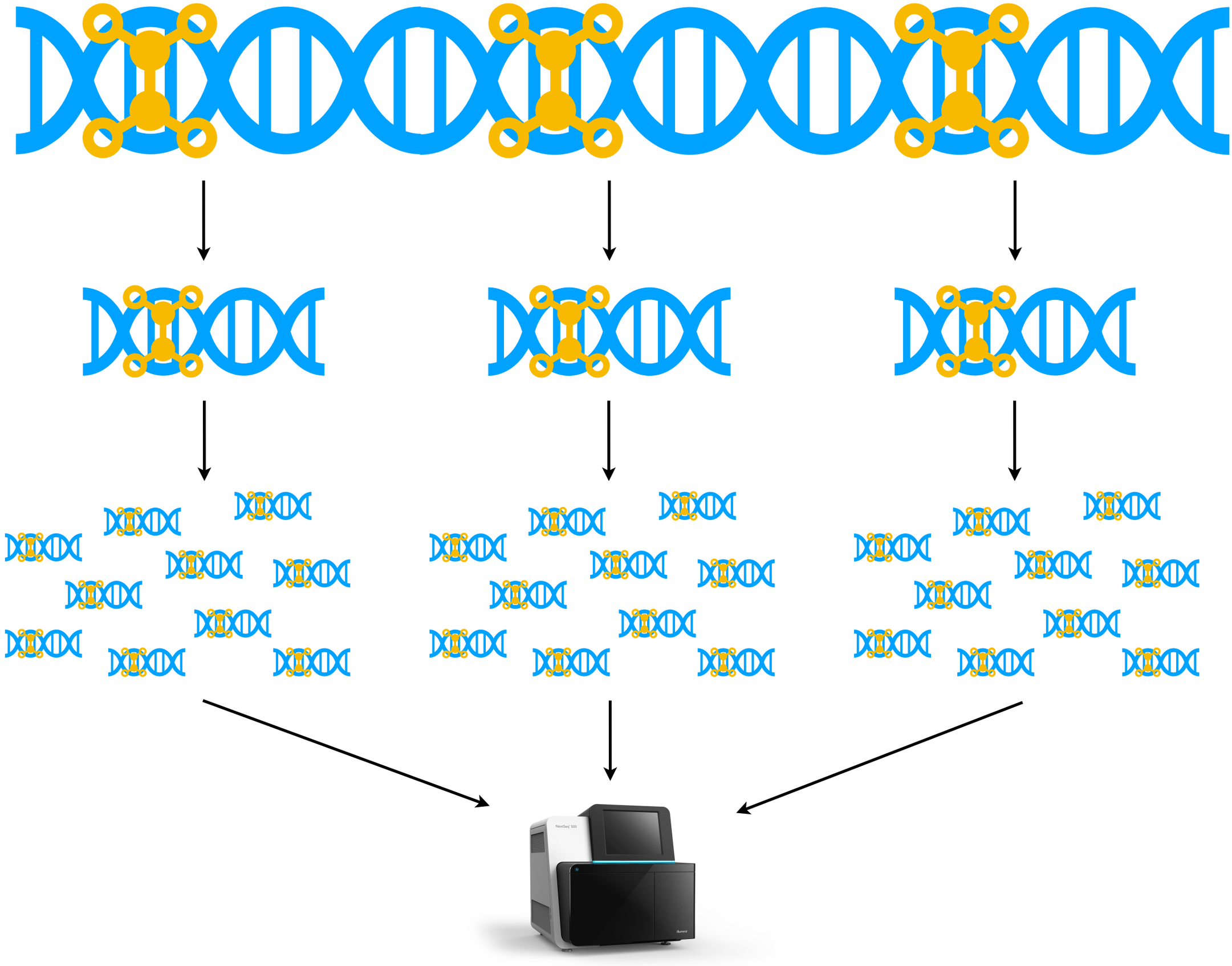
Chromatin **I**mmuno**P**recipitation **seq**uencing

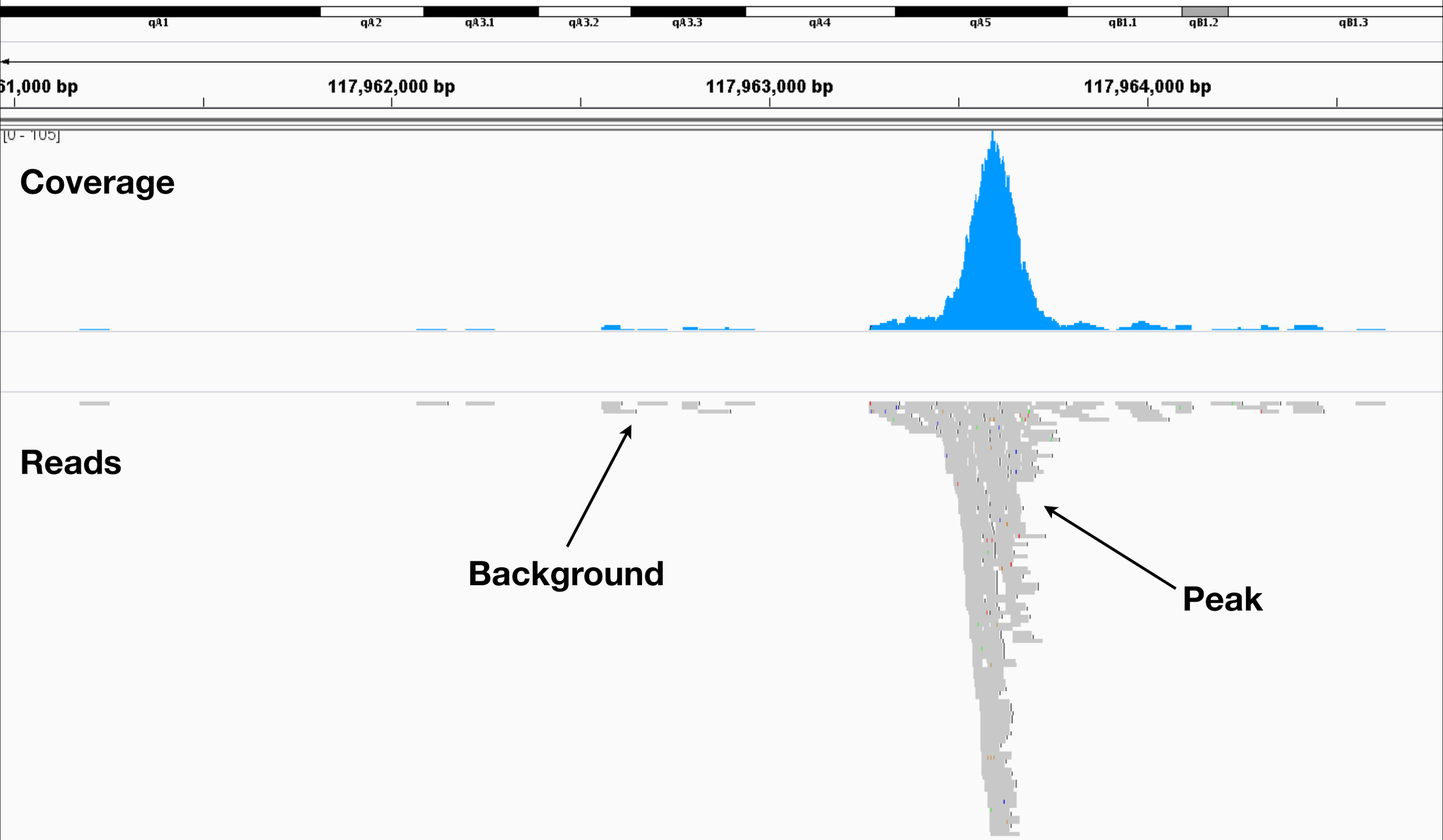
Aim: To study the interaction between proteins and DNA in the cell nucleus

Proteins

DNA




NextSeq 500
illumina

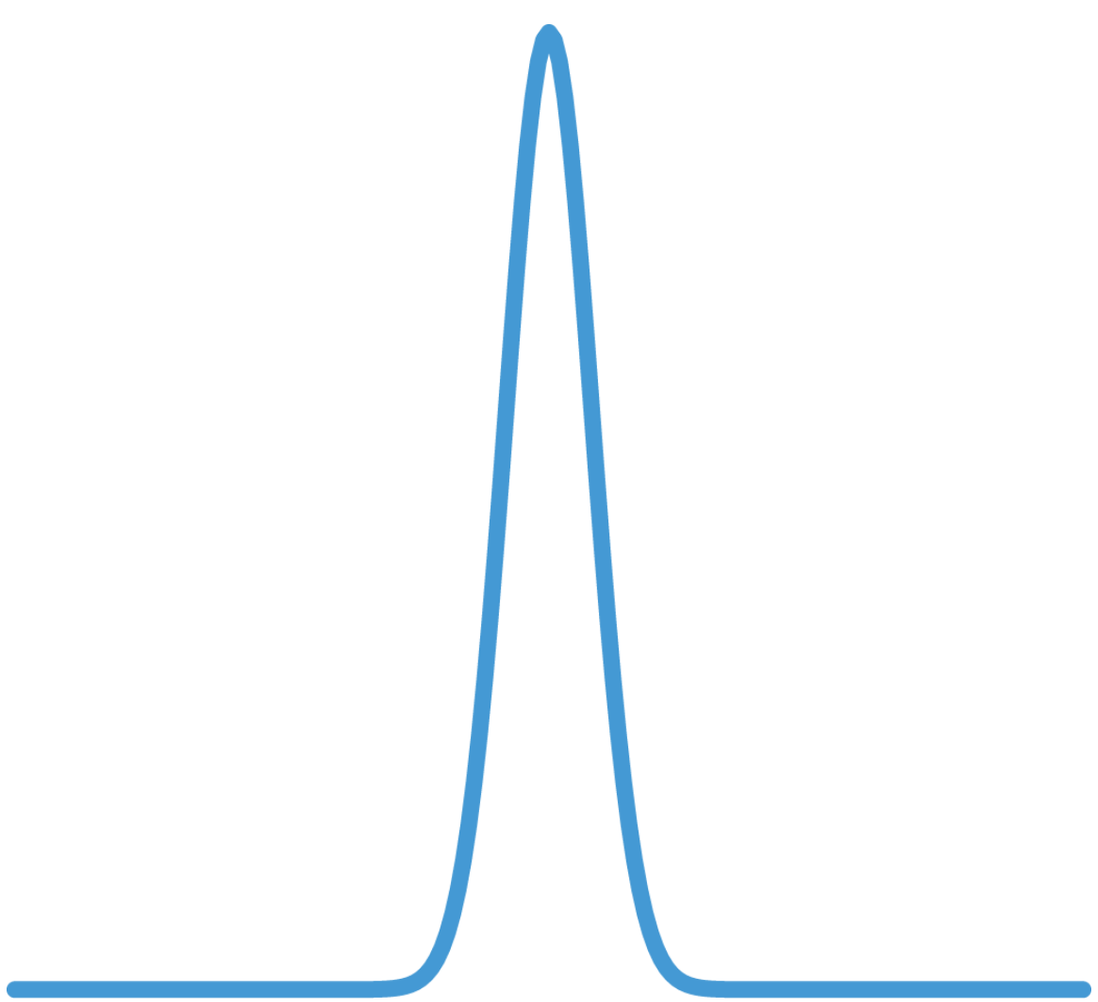


Integrative Genomics Viewer

http://software.broadinstitute.org/software/igv/

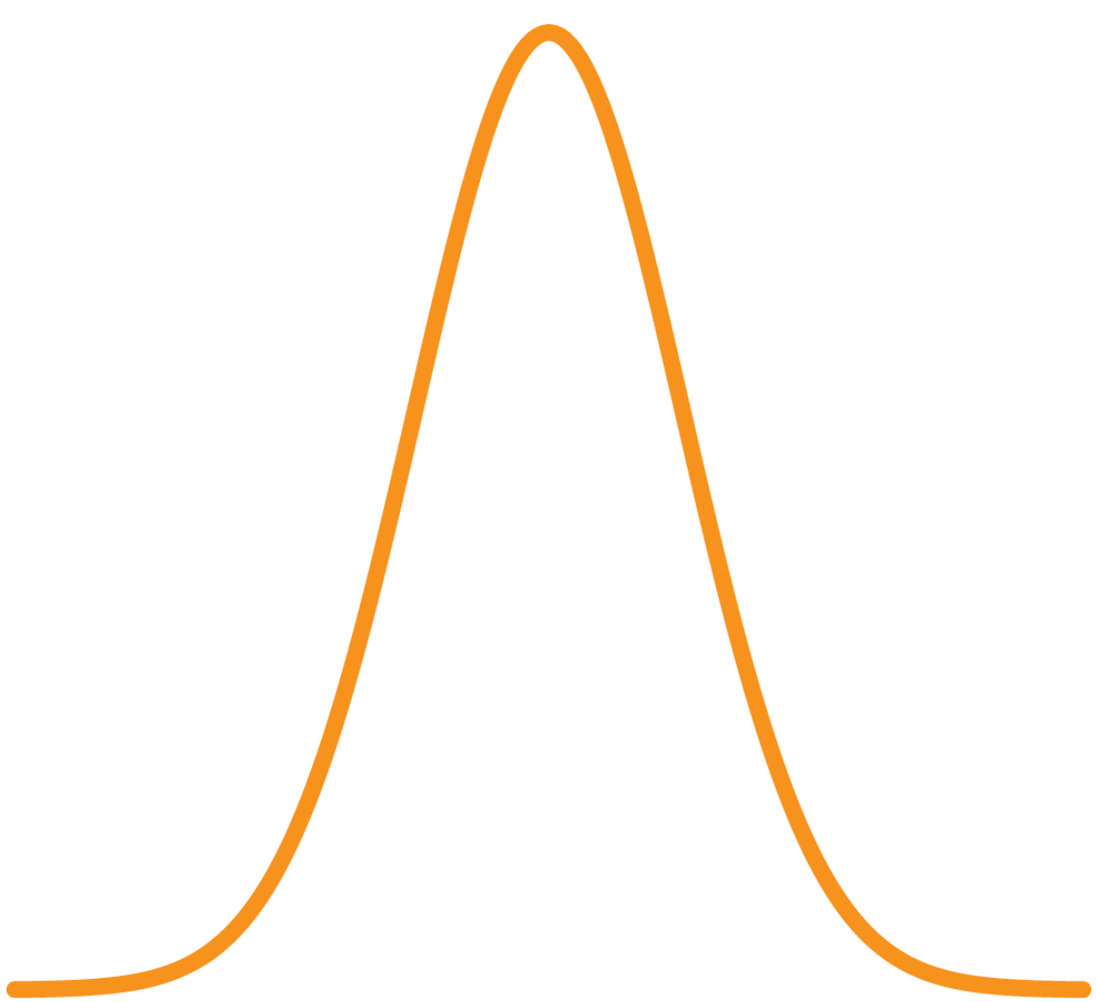
Transcription factor binding

Transcription factors control gene expression by binding with DNA regulatory regions



Histone modification

Histones cause changes within DNA and chromatin confirmation



Programming languages

```
$
```

```
>
```

Example data

NFYA binding

Nuclear transcription **F**actor **Y** subunit **A**lpha

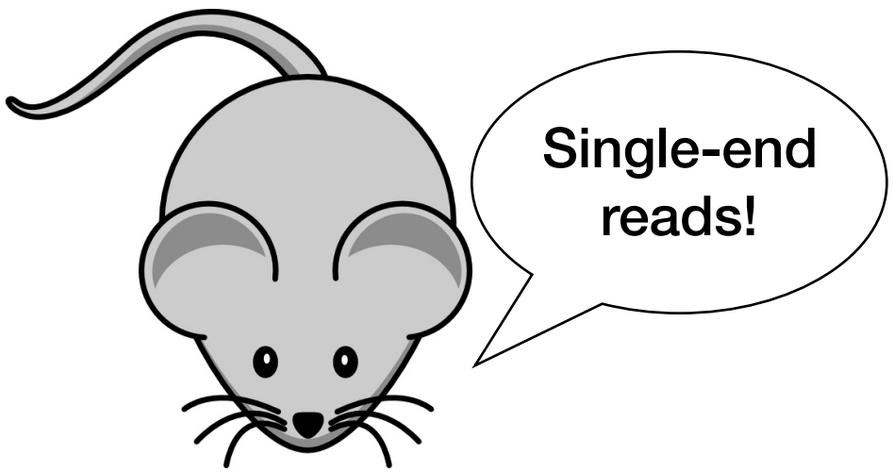
Embryonic Stem Cells (ES)

- ES NFYA biological replicate 1
- ES NFYA biological replicate 2

Terminally differentiated Neurons (TN)

- TN NFYA biological replicate 1
- TN NFYA biological replicate 2

Input



NFYA samples

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25532

NFYA samples

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25532

	Samples	GEO identifier
	ES NFYA biological replicate 1, 2	GSM632038, GSM632039
	TN NFYA biological replicate 1, 2	GSM632057, GSM632058
	Input	GSM632041

NFYA samples

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25532

Samples	GEO identifier	SRA identifier
ES NFYA biological replicate 1, 2	GSM632038, GSM632039	SRR074398, SRR074399
TN NFYA biological replicate 1, 2	GSM632057, GSM632058	SRR074417, SRR074418
Input	GSM632041	SRR074401

SRA files

- SRA toolkit:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

SRA files

- SRA toolkit:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

- FastQ dump:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=fastq-dump

SRA files

- SRA toolkit:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

- FastQ dump:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=fastq-dump

```
$ fastq-dump SRR074398
```

SRA files

- SRA toolkit:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software

- FastQ dump:

https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=fastq-dump

```
$ fastq-dump SRR074398
$ fastq-dump SRR074399
$ fastq-dump SRR074417
$ fastq-dump SRR074418
$ fastq-dump SRR074401
```

Data pre-processing

Step 1: Quality control

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Step 1: Quality control

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/


```
$ fastqc *.fastq
```

Step 1: Quality control

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

MultiQC: http://multiqc.info/docs/


```
$ fastqc *.fastq
```

Step 1: Quality control

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

MultiQC: http://multiqc.info/docs/


```
$ fastqc *.fastq
$ multiqc .
```

Step 2: Read trimming*

- Cutadapt:

https://github.com/marcelm/cutadapt

- Trim Galore:

https://github.com/FelixKrueger/TrimGalore

***only if required**

Step 3: Alignment

http://bioconductor.org/packages/release/bioc/html/Rsubread.html

Alignment

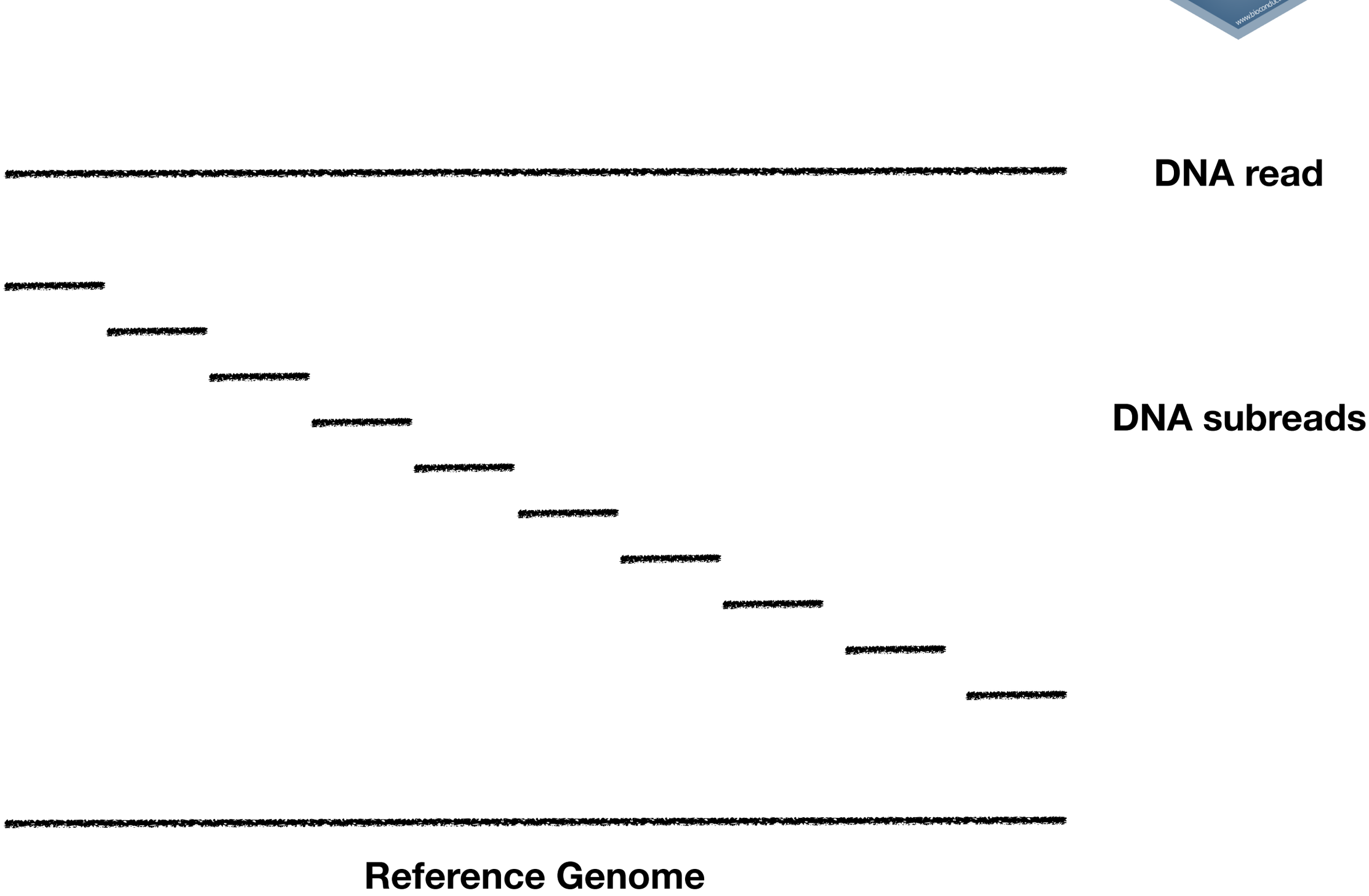
DNA read

Alignment

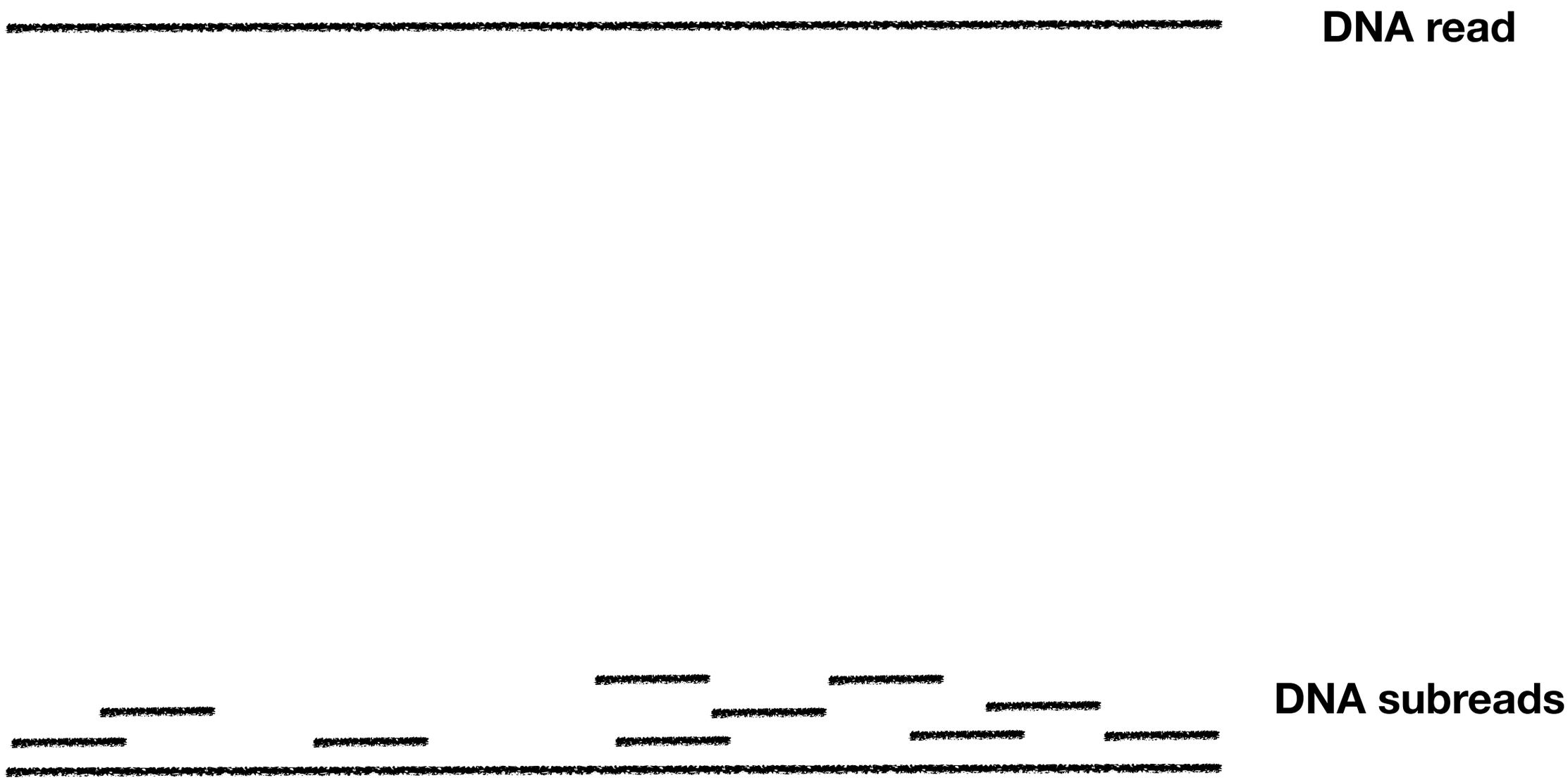
DNA read

DNA subreads

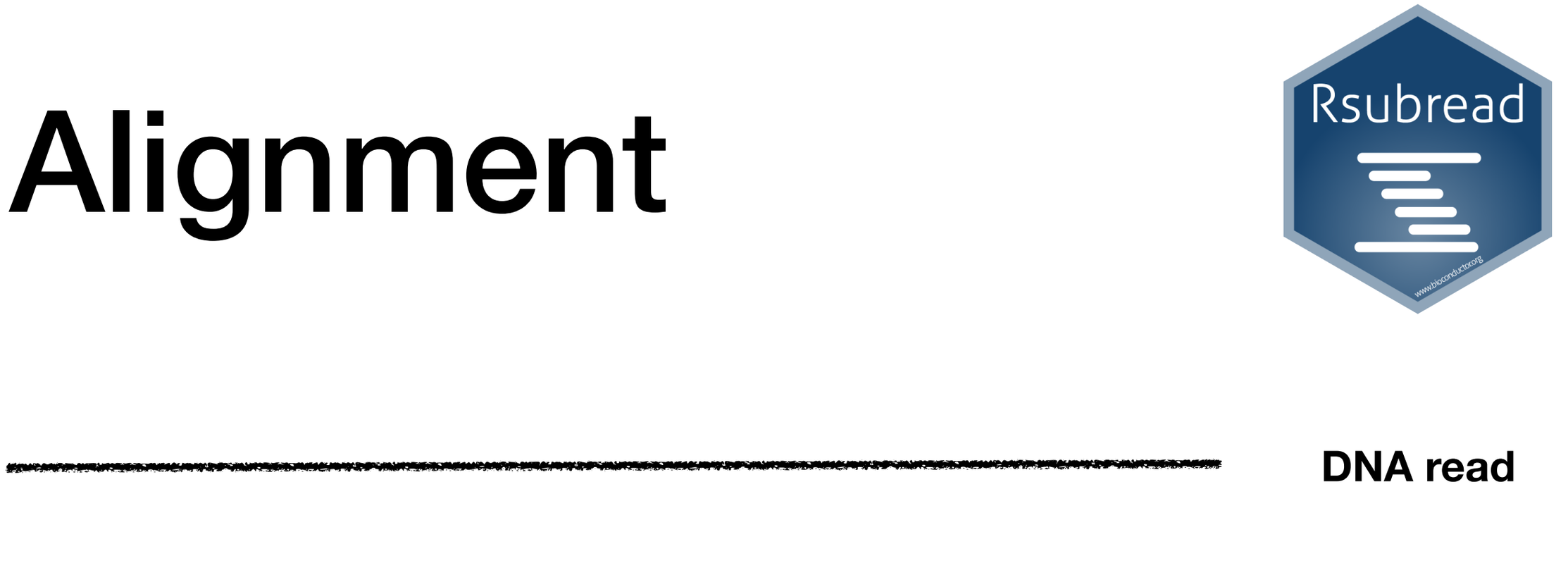
Alignment



Alignment



Alignment


DNA read
Rejected positions
Voted position
DNA subreads
Reference Genome

Build index

Build index


```
> library(Rsubread)
```

Build index


```
> library(Rsubread)
> buildindex(basename="mm10_index",
```

Build index


```
> library(Rsubread)
> buildindex(basename="mm10_index",
             reference="mm10.fa",
```

Build index


```
> library(Rsubread)
> buildindex(basename="mm10_index",
             reference="mm10.fa",
             gappedIndex=FALSE,
             indexSplit=FALSE)
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq        es_1
2 SRR074399.fastq        es_2
3 SRR074417.fastq        tn_1
4 SRR074418.fastq        tn_2
5 SRR074401.fastq       input
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq        es_1
2 SRR074399.fastq        es_2
3 SRR074417.fastq        tn_1
4 SRR074418.fastq        tn_2
5 SRR074401.fastq       input
> align(index="mm10_index",
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq       es_1
2 SRR074399.fastq       es_2
3 SRR074417.fastq       tn_1
4 SRR074418.fastq       tn_2
5 SRR074401.fastq      input
> align(index="mm10_index",
     readfile1=files$FileName,
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq        es_1
2 SRR074399.fastq        es_2
3 SRR074417.fastq        tn_1
4 SRR074418.fastq        tn_2
5 SRR074401.fastq       input
> align(index="mm10_index",
      readfile1=files$FileName,
      type="DNA",
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq        es_1
2 SRR074399.fastq        es_2
3 SRR074417.fastq        tn_1
4 SRR074418.fastq        tn_2
5 SRR074401.fastq       input
> align(index="mm10_index",
      readfile1=files$FileName,
      type="DNA",
      unique=TRUE,
```

Align


```
> files <- data.frame(FileName = c("SRR074398.fastq",
"SRR074399.fastq", "SRR074417.fastq",
"SRR074418.fastq", "SRR074401.fastq"), SampleName =
c("es_1","es_2","tn_1","tn_2","input"))
> files
        FileName SampleName
1 SRR074398.fastq        es_1
2 SRR074399.fastq        es_2
3 SRR074417.fastq        tn_1
4 SRR074418.fastq        tn_2
5 SRR074401.fastq       input
> align(index="mm10_index",
      readfile1=files$FileName,
      type="DNA",
      unique=TRUE,
      nthreads=10)
```

Proportion mapped

Proportion of reads mapped

```
> bam_files <- paste0(files$FileName, ".subread.BAM")
```

Proportion mapped

Proportion of reads mapped

```
> bam_files <- paste0(files$FileName, ".subread.BAM")
> prop.mapped <- propmapped(bam_files)
```

Proportion mapped

Proportion of reads mapped

```
> bam_files <- paste0(files$FileName, ".subread.BAM")
> prop.mapped <- propmapped(bam_files)
> prop.mapped$Samples <- files$FileName
```

Proportion mapped

Proportion of reads mapped

```
> bam_files <- paste0(files$FileName, ".subread.BAM")
> prop.mapped <- propmapped(bam_files)
> prop.mapped$Samples <- files$FileName
> cbind(SampleName=files$SampleName, prop.mapped)
```

Proportion mapped

Proportion of reads mapped

```
> bam_files <- paste0(files$FileName, ".subread.BAM")
> prop.mapped <- propmapped(bam_files)
> prop.mapped$Samples <- files$FileName
> cbind(SampleName=files$SampleName, prop.mapped)
  SampleName         Samples NumTotal NumMapped PropMapped
1       es_1 SRR074398.fastq 32038452  22782623   0.711102
2       es_2 SRR074399.fastq 36749276  26119506   0.710749
3       tn_1 SRR074417.fastq 39283051  28983089   0.737801
4       tn_2 SRR074418.fastq 35423633  27282881   0.770189
5      input SRR074401.fastq 15032584  11492495   0.764506
```

Step 4: Bam file processing

- Sambamba

http://lomereiter.github.io/sambamba/

1. Coordinate sort bam files

2. Mark duplicate reads

Sort bam files


```
$ sambamba sort -t 10 SRR074398.fastq.subread.BAM
```

Sort bam files


```
$ sambamba sort -t 10 SRR074398.fastq.subread.BAM
$ sambamba sort -t 10 SRR074399.fastq.subread.BAM
$ sambamba sort -t 10 SRR074417.fastq.subread.BAM
$ sambamba sort -t 10 SRR074418.fastq.subread.BAM
$ sambamba sort -t 10 SRR074401.fastq.subread.BAM
```

Sort bam files


```
$ sambamba sort -t 10 SRR074398.fastq.subread.BAM
$ sambamba sort -t 10 SRR074399.fastq.subread.BAM
$ sambamba sort -t 10 SRR074417.fastq.subread.BAM
$ sambamba sort -t 10 SRR074418.fastq.subread.BAM
$ sambamba sort -t 10 SRR074401.fastq.subread.BAM
```

Sorted bam file	Index file
SRR074398.fastq.subread.sorted.bam	SRR074398.fastq.subread.sorted.bam.bai
SRR074399.fastq.subread.sorted.bam	SRR074399.fastq.subread.sorted.bam.bai
SRR074417.fastq.subread.sorted.bam	SRR074417.fastq.subread.sorted.bam.bai
SRR074418.fastq.subread.sorted.bam	SRR074418.fastq.subread.sorted.bam.bai
SRR074401.fastq.subread.sorted.bam	SRR074401.fastq.subread.sorted.bam.bai

Mark duplicates


```
$ sambamba markdup -t 10 SRR074398.fastq.subread.sorted.bam
SRR074398.fastq.subread.sorted.markdup.bam
```

Mark duplicates

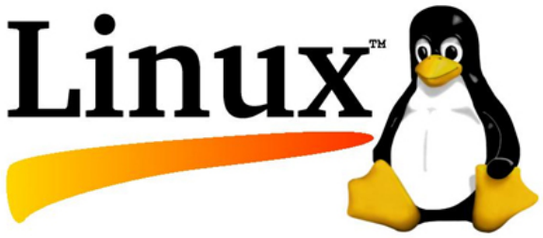

```
$ sambamba markdup -t 10 SRR074398.fastq.subread.sorted.bam
SRR074398.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074399.fastq.subread.sorted.bam
SRR074399.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10  SRR074417.fastq.subread.sorted.bam
SRR074417.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074418.fastq.subread.sorted.bam
SRR074418.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074401.fastq.subread.sorted.bam
SRR074401.fastq.subread.sorted.markdup.bam
```

Mark duplicates


```
$ sambamba markdup -t 10 SRR074398.fastq.subread.sorted.bam
SRR074398.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074399.fastq.subread.sorted.bam
SRR074399.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10  SRR074417.fastq.subread.sorted.bam
SRR074417.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074418.fastq.subread.sorted.bam
SRR074418.fastq.subread.sorted.markdup.bam
$ sambamba markdup -t 10 SRR074401.fastq.subread.sorted.bam
SRR074401.fastq.subread.sorted.markdup.bam
```

Marked duplicated bam file	Index file
SRR074398.fastq.subread.sorted.markdup.bam	SRR074398.fastq.subread.sorted.markdup.bam.bai
SRR074399.fastq.subread.sorted.markdup.bam	SRR074399.fastq.subread.sorted.markdup.bam.bai
SRR074417.fastq.subread.sorted.markdup.bam	SRR074417.fastq.subread.sorted.markdup.bam.bai
SRR074418.fastq.subread.sorted.markdup.bam	SRR074418.fastq.subread.sorted.markdup.bam.bai
SRR074401.fastq.subread.sorted.markdup.bam	SRR074401.fastq.subread.sorted.markdup.bam.bai

Proportion duplicate reads




```
$ sambamba flagstat -t 10
SRR074398.fastq.subread.sorted.markdup.bam
```

Proportion duplicate reads


```
$ sambamba flagstat -t 10
SRR074398.fastq.subread.sorted.markdup.bam >
SRR074398_flagstat.txt
```

Proportion duplicate reads


```
$ sambamba flagstat -t 10
SRR074398.fastq.subread.sorted.markdup.bam >
SRR074398_flagstat.txt

$ sambamba flagstat -t 10
SRR074399.fastq.subread.sorted.markdup.bam >
SRR074399_flagstat.txt

$ sambamba flagstat -t 10
SRR074417.fastq.subread.sorted.markdup.bam >
SRR074417_flagstat.txt

$ sambamba flagstat -t 10
SRR074418.fastq.subread.sorted.markdup.bam >
SRR074418_flagstat.txt

$ sambamba flagstat -t 10
SRR074401.fastq.subread.sorted.markdup.bam >
SRR074401_flagstat.txt
```

Proportion duplicate reads

32038452 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
5011150 + 0 duplicates
22782623 + 0 mapped (71.11%:N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A:N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A:N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Proportion duplicate reads

32038452 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
5011150 + 0 duplicates
22782623 + 0 mapped (71.11%:N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A:N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A:N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Proportion duplicate reads

32038452 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
5011150 + 0 duplicates
22782623 + 0 mapped (71.11%:N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A:N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A:N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Proportion duplicate reads

Sample	Flagstat file	Proportion duplicate reads
es_1	SRR074398_flagstat.txt	0.156
es_2	SRR074399_flagstat.txt	0.242
tn_1	SRR074417_flagstat.txt	0.231
tn_2	SRR074418_flagstat.txt	0.123
input	SRR074401_flagstat.txt	0.138

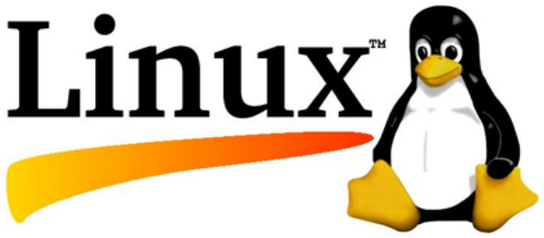
Step 5: Binding visualization

- deepTools

https://deeptools.readthedocs.io/en/develop/

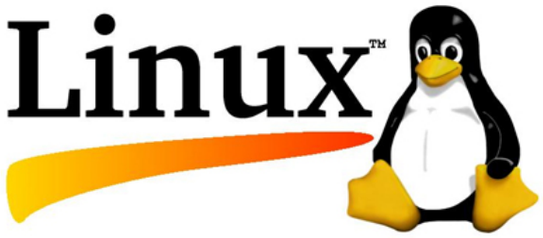
1. Coverage files
2. Scale coverage over gene regions
3. Plot scaled coverage for each sample

Coverage files



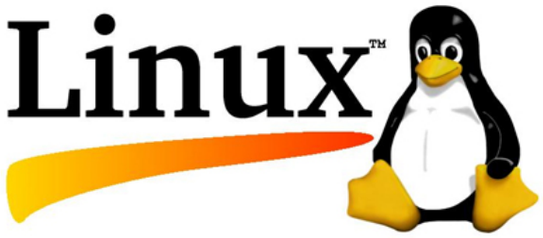

```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
```

Coverage files



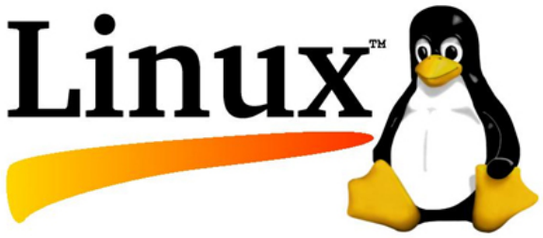

```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
-o SRR074398_coverage.bigWig
```

Coverage files



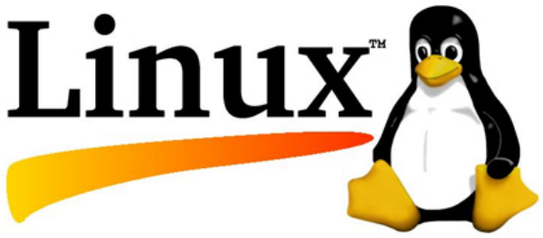

```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
-o SRR074398_coverage.bigWig
--outFileFormat bigwig
```

Coverage files




```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
-o SRR074398_coverage.bigWig
--outFileFormat bigwig
--numberOfProcessors 10
```

Coverage files




```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
-o SRR074398_coverage.bigWig
--outFileFormat bigwig
--numberOfProcessors 10
--ignoreDuplicates
```

Coverage files


```
$ bamCoverage -b SRR074398.fastq.subread.sorted.markdup.bam
-o SRR074398_coverage.bigWig
--outFileFormat bigwig
--numberOfProcessors 10
--ignoreDuplicates

$ bamCoverage -b SRR074399.fastq.subread.sorted.markdup.bam
-o SRR074399_coverage.bigWig
--outFileFormat bigwig --numberOfProcessors 10 --ignoreDuplicates

$bamCoverage -b SRR074417.fastq.subread.sorted.markdup.bam
-o SRR074417_coverage.bigWig
--outFileFormat bigwig --numberOfProcessors 10 --ignoreDuplicates

$ bamCoverage -b SRR074418.fastq.subread.sorted.markdup.bam
-o SRR074418_coverage.bigWig
--outFileFormat bigwig --numberOfProcessors 10 --ignoreDuplicates

$ bamCoverage -b SRR074401.fastq.subread.sorted.markdup.bam
-o SRR074401_coverage.bigWig
--outFileFormat bigwig --numberOfProcessors 10 --ignoreDuplicates
```

Regions

BED file containing gene regions

- Tab delimited file
- No column or row names
- One entry per region

Regions

BED file containing gene regions

- Tab delimited file
- No column or row names
- One entry per region

Chromosome	Start	End	Name	Score	Strand
chr1	0	50	*	*	+/-

Regions

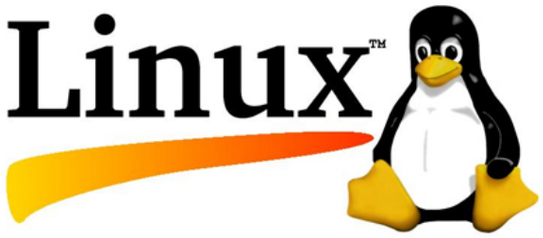
BED file containing gene regions

- Tab delimited file
- No column or row names
- One entry per region

Chromosome	Start	End	Name	Score	Strand
chr1	0	50	*	*	+/-


```
> getInBuiltAnnotation("mm10")
```

Scale region coverage




```
$ computeMatrix scale-regions
```

Scale region coverage


```
$ computeMatrix scale-regions
-S SRR074398_coverage.bigWig
SRR074399_coverage.bigWig SRR074417_coverage.bigWig
SRR074418_coverage.bigWig SRR074401_coverage.bigWig
```

Scale region coverage


```
$ computeMatrix scale-regions
-S SRR074398_coverage.bigWig
SRR074399_coverage.bigWig SRR074417_coverage.bigWig
SRR074418_coverage.bigWig SRR074401_coverage.bigWig
-R mm10_gene_body.bed
```

Scale region coverage


```
$ computeMatrix scale-regions
-S SRR074398_coverage.bigWig
SRR074399_coverage.bigWig SRR074417_coverage.bigWig
SRR074418_coverage.bigWig SRR074401_coverage.bigWig
-R mm10_gene_body.bed
-out ChIP_coverage.gz
```

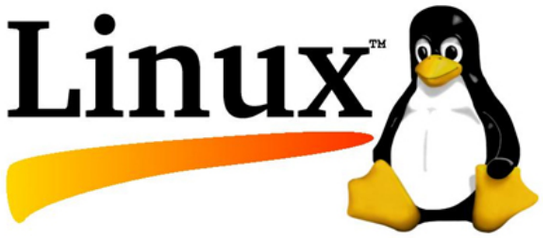
Scale region coverage


```
$ computeMatrix scale-regions
-S SRR074398_coverage.bigWig
SRR074399_coverage.bigWig SRR074417_coverage.bigWig
SRR074418_coverage.bigWig SRR074401_coverage.bigWig
-R mm10_gene_body.bed
-out ChIP_coverage.gz
-a 1000 -b 1000
```

Scale region coverage


```
$ computeMatrix scale-regions
-S SRR074398_coverage.bigWig
SRR074399_coverage.bigWig SRR074417_coverage.bigWig
SRR074418_coverage.bigWig SRR074401_coverage.bigWig
-R mm10_gene_body.bed
-out ChIP_coverage.gz
-a 1000 -b 1000
--numberOfProcessors 10
```

Plot coverage




```
$ plotHeatmap -m ChIP_coverage.gz
```

Plot coverage

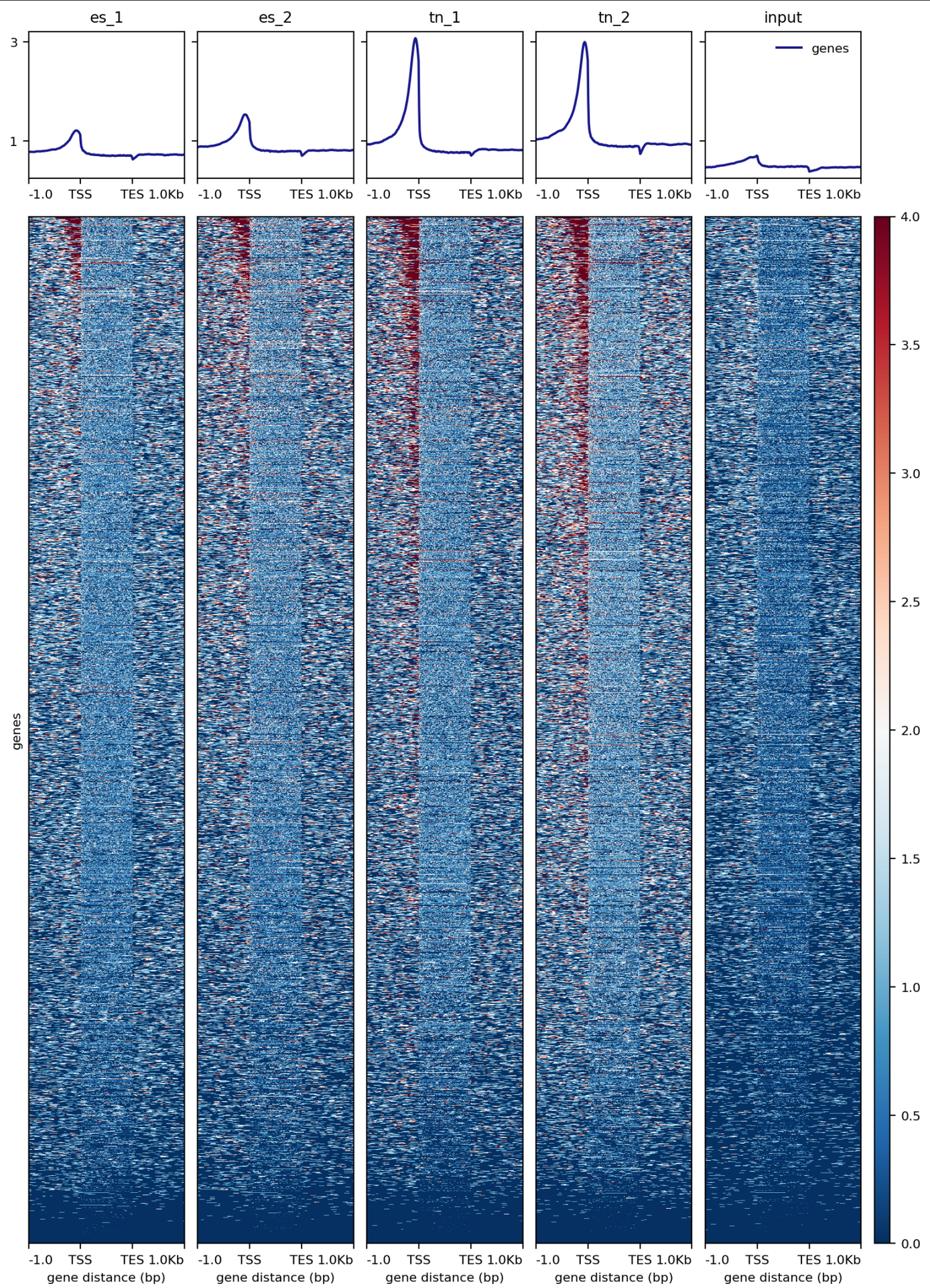

```
$ plotHeatmap -m ChIP_coverage.gz
-out ChIP_coverage_plot.png
```

Plot coverage


```
$ plotHeatmap -m ChIP_coverage.gz
-out ChIP_coverage_plot.png
--colorMap RdBu_r
```

Plot coverage


```
$ plotHeatmap -m ChIP_coverage.gz
-out ChIP_coverage_plot.png
--colorMap RdBu_r
--samplesLabel 'es_1' 'es_2' 'tn_1' 'tn_2' 'input'
```


es_1
es_2
tn_1
tn_2
input
genes
3
1
-1.0
TSS
TES 1.0Kb
genes
gene distance (bp)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

10 minute break

Analysis pipelines

What is your research question?

Where is binding occurring?

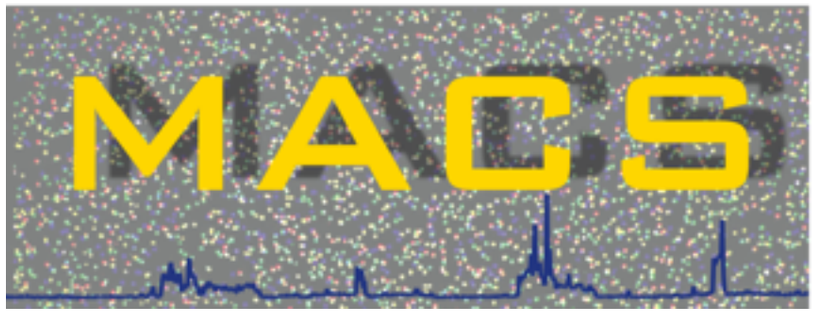
Peak caller

What are the differences in binding between my sample groups?

Differential binding analysis

Peak calling

Peak callers



http://liulab.dfci.harvard.edu/MACS/

http://homer.ucsd.edu/homer/

Peak callers

1. Create tag directories
2. Find peaks
3. Annotate peaks

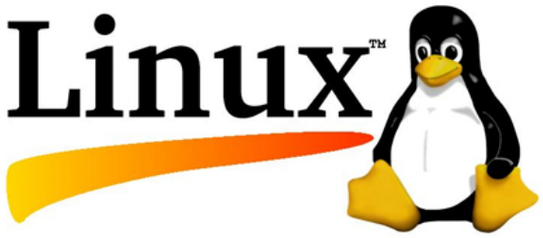
http://homer.ucsd.edu/homer/

Tag directories

- Create a Tag directory for each group

```
$ makeTagDirectory ES/
SRR074398.fastq.subread.sorted.markdup.bam
SRR074399.fastq.subread.sorted.markdup.bam
-tbp 1
```

Tag directories



- Create a Tag directory for each group

```
$ makeTagDirectory ES/
SRR074398.fastq.subread.sorted.markdup.bam
SRR074399.fastq.subread.sorted.markdup.bam
-tbp 1

$ makeTagDirectory TN/
SRR074417.fastq.subread.sorted.markdup.bam
SRR074418.fastq.subread.sorted.markdup.bam  -tbp 1

$ makeTagDirectory INPUT/
SRR074401.fastq.subread.sorted.markdup.bam  -tbp 1
```

Find peaks

Embryonic stem cell peaks relative to input

```
$ findPeaks ES/ -style factor -o ES_peaks.txt
-i INPUT
```

Find peaks

Embryonic stem cell peaks relative to input

```
$ findPeaks ES/ -style factor -o ES_peaks.txt
-i INPUT

0.10% FDR Threshold set at 13.0 (poisson pvalue ~
9.02e-07)
38537 peaks passed threshold
Differential Peaks: 1334 of 38537 (3.46% passed)
Local Background Filtering: 1233 of 1334 (92.43%
passed)
Clonal filtering: 1233 of 1233 (100.00% passed)
Total Peaks identified = 1233
Centering peaks of size 117 using a fragment length
of 117
```

Find peaks

Terminally differentiated neuron peaks relative to input

```
$ findPeaks TN/ -style factor -o TN_peaks.txt -i
INPUT/

0.10% FDR Threshold set at 13.0 (poisson pvalue ~
5.59e-07)
31416 peaks passed threshold
Differential Peaks: 2391 of 31416 (7.61% passed)
Local Background Filtering: 2381 of 2391 (99.58%
passed)
Clonal filtering: 2381 of 2381 (100.00% passed)
Total Peaks identified = 2381
Centering peaks of size 91 using a fragment length of
91
```

Annotate peaks


```
$ annotatePeaks.pl ES_peaks.txt mm10
```

Annotate peaks


```
$ annotatePeaks.pl ES_peaks.txt mm10 >
ES_peaks_annotated.txt
```

Annotate peaks


```
$ annotatePeaks.pl ES_peaks.txt mm10 >
ES_peaks_annotated.txt



$ annotatePeaks.pl TN_peaks.txt mm10 >
TN_peaks_annotated.txt
```

Annotate peaks

1. Peak ID
2. Chromosome
3. Peak start position
4. Peak end position
5. Strand
6. Peak Score
7. FDR/Peak Focus Ratio/Region Size
8. Annotation (i.e. Exon, Intron, ...)
9. Detailed Annotation (Exon, Intron etc. + CpG Islands, repeats, etc.)
10. Distance to nearest RefSeq TSS

Annotate peaks

11. Nearest TSS: Native ID of annotation file

12. Nearest TSS: Entrez Gene ID

13. Nearest TSS: Unigene ID

14. Nearest TSS: RefSeq ID

15. Nearest TSS: Ensembl ID

16. Nearest TSS: Gene Symbol

17. Nearest TSS: Gene Aliases

18. Nearest TSS: Gene description

19. Nearest TSS: Gene type

Differential binding

Do you have replicates?

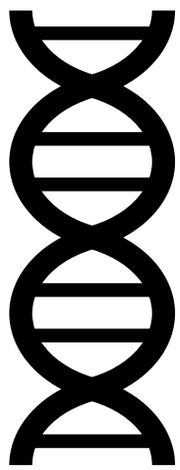
Do you have replicates? No

STOP YOUR ANALYSIS!

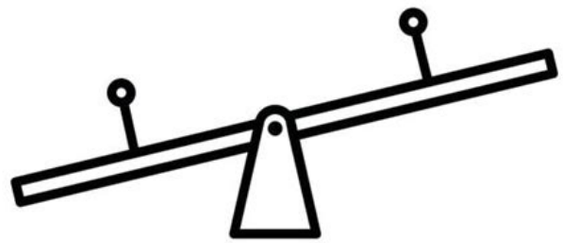
Pipelines

Peaks

Gene oriented

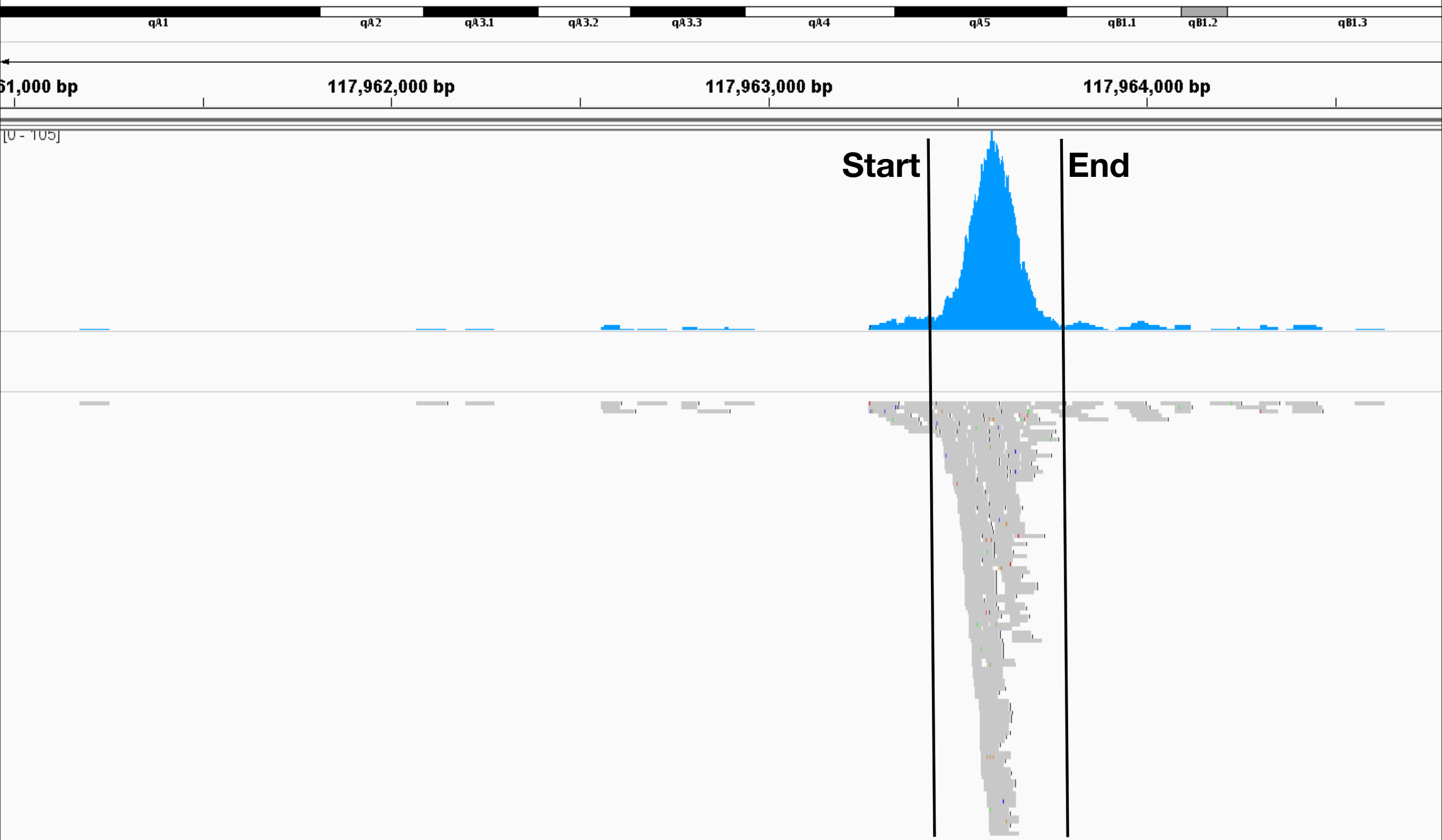


CSAW



Peaks

Peaks for DB


qA1
qA2
qA3.1
qA3.2
qA3.3
qA4
qA5
qB1.1
qB1.2
qB1.3
61,000 bp
117,962,000 bp
117,963,000 bp
117,964,000 bp
[0 - 105]
Start
End

Using peaks

Peak callers do not control error rates correctly

"De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly"

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066778/

Peak calling to control error rates

Initial analysis

Embryonic stem cell peaks relative to input

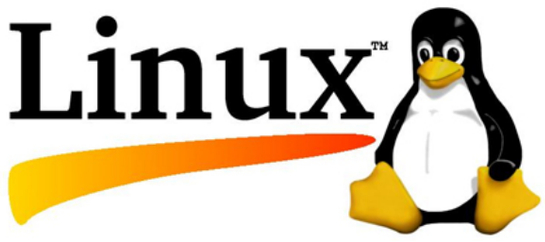
Terminally differentiated neuron peaks relative to input

Does NOT control the error rates!

Peak calling to control error rates

Embryonic stem cell and terminally differentiated neuron peaks relative to input

Calling peaks for DB analysis



Make a Tag directory using all ChIP samples

```
$ makeTagDirectory ES_TN/
SRR074398.fastq.subread.sorted.markdup.bam
SRR074399.fastq.subread.sorted.markdup.bam
SRR074417.fastq.subread.sorted.markdup.bam
SRR074418.fastq.subread.sorted.markdup.bam -tbp 1
```

Calling peaks for DB analysis

Find peaks comparing the combined ChIP files to the input

```
$ findPeaks ES_TN/ -style factor -o ES_TN_peaks.txt -
i INPUT/
```

Calling peaks for DB analysis

Find peaks comparing the combined ChIP files to the input

```
$ findPeaks ES_TN/ -style factor -o ES_TN_peaks.txt -
i INPUT/

0.10% FDR Threshold set at 19.0 (poisson pvalue ~
8.33e-07)
49835 peaks passed threshold
Differential Peaks: 1707 of 49835 (3.43% passed)
Local Background Filtering: 1675 of 1707 (98.13%
passed)
Clonal filtering: 1675 of 1675 (100.00% passed)
Total Peaks identified = 1675
Centering peaks of size 111 using a fragment length
of 111
```

Calling peaks for DB analysis

Annotate peaks

```
$ annotatePeaks.pl ES_TN_peaks.txt mm10 >
ES_TN_peaks_annotated.txt
```

Summarize data


```
> library(Rsubread)
```

Summarize data


```
> library(Rsubread)
> peaks <- read.delim("ES_TN_peaks_annotated.txt",
stringsAsFactors = FALSE)
```

Summarize data


```
> library(Rsubread)
> peaks <- read.delim("ES_TN_peaks_annotated.txt",
stringsAsFactors = FALSE)
> peaks_subread <- peaks[,1:5]
```

Summarize data


```
> library(Rsubread)
> peaks <- read.delim("ES_TN_peaks_annotated.txt",
stringsAsFactors = FALSE)
> peaks_subread <- peaks[,1:5]
> colnames(peaks_subread)[1] <- "GeneID"
```

Summarize data


```
> files <- data.frame(FileName =
c("SRR074398.fastq.subread.sorted.markdup.bam",
"SRR074399.fastq.subread.sorted.markdup.bam",
"SRR074417.fastq.subread.sorted.markdup.bam",
"SRR074418.fastq.subread.sorted.markdup.bam"),
SampleName=c("es_1","es_2","tn_1","tn_2"),
stringsAsFactors = FALSE)
```

Summarize data


```
> peak_counts <- featureCounts(files=files$FileName,
annot.ext=peaks_subread, nthreads=10)
```

Summarize data


```
> peak_counts <- featureCounts(files=files$FileName,
annot.ext=peaks_subread, nthreads=10)
> names(peak_counts)
[1] "counts"     "annotation" "targets"    "stat"
```

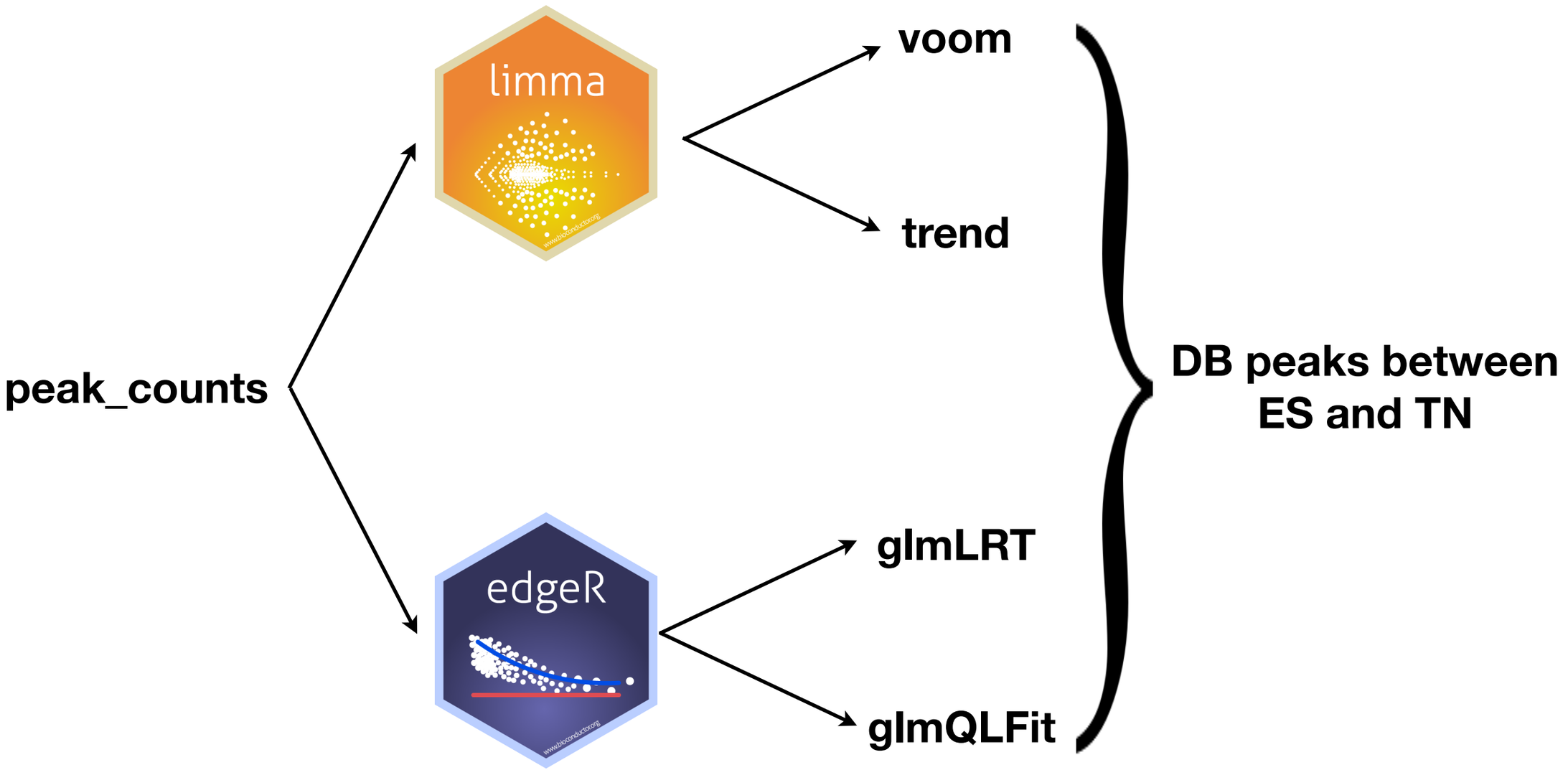
Summarize data


```
> peak_counts <- featureCounts(files=files$FileName,
annot.ext=peaks_subread, nthreads=10)
> names(peak_counts)
[1] "counts"     "annotation" "targets"    "stat"
> peak_counts <- peak_counts$counts
```

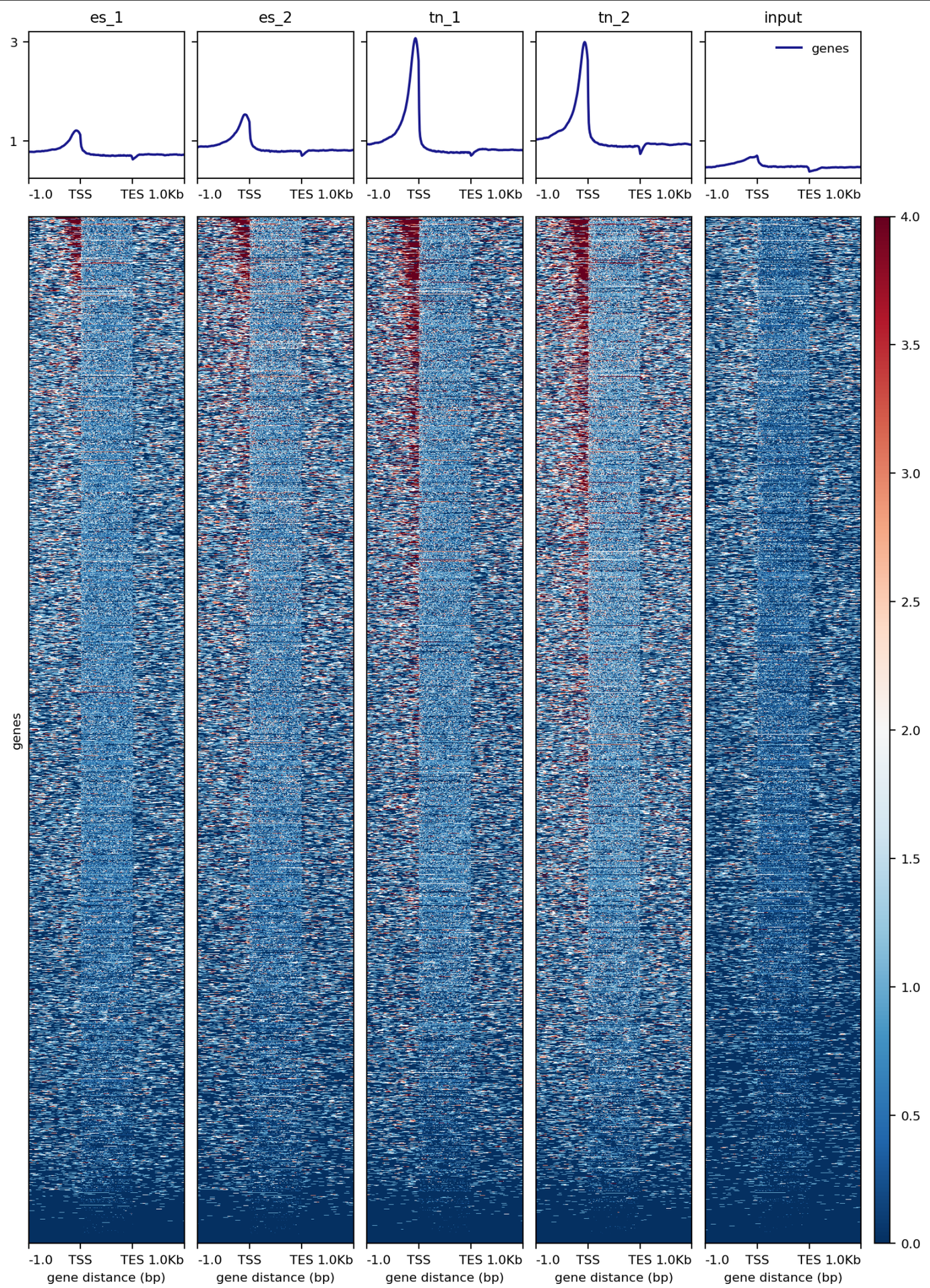
Summarize data


```
> peak_counts <- featureCounts(files=files$FileName,
annot.ext=peaks_subread, nthreads=10)
> names(peak_counts)
[1] "counts"     "annotation" "targets"    "stat"
> peak_counts <- peak_counts$counts
> colnames(peak_counts) <- files$SampleName
```

DB analysis


limma
voom
trend
peak_counts
edgeR
glmLRT
glmQLFit
DB peaks between
ES and TN

Gene oriented


es_1
es_2
tn_1
tn_2
input
genes
3
1
-1.0
TSS
TES 1.0Kb
genes
gene distance (bp)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0

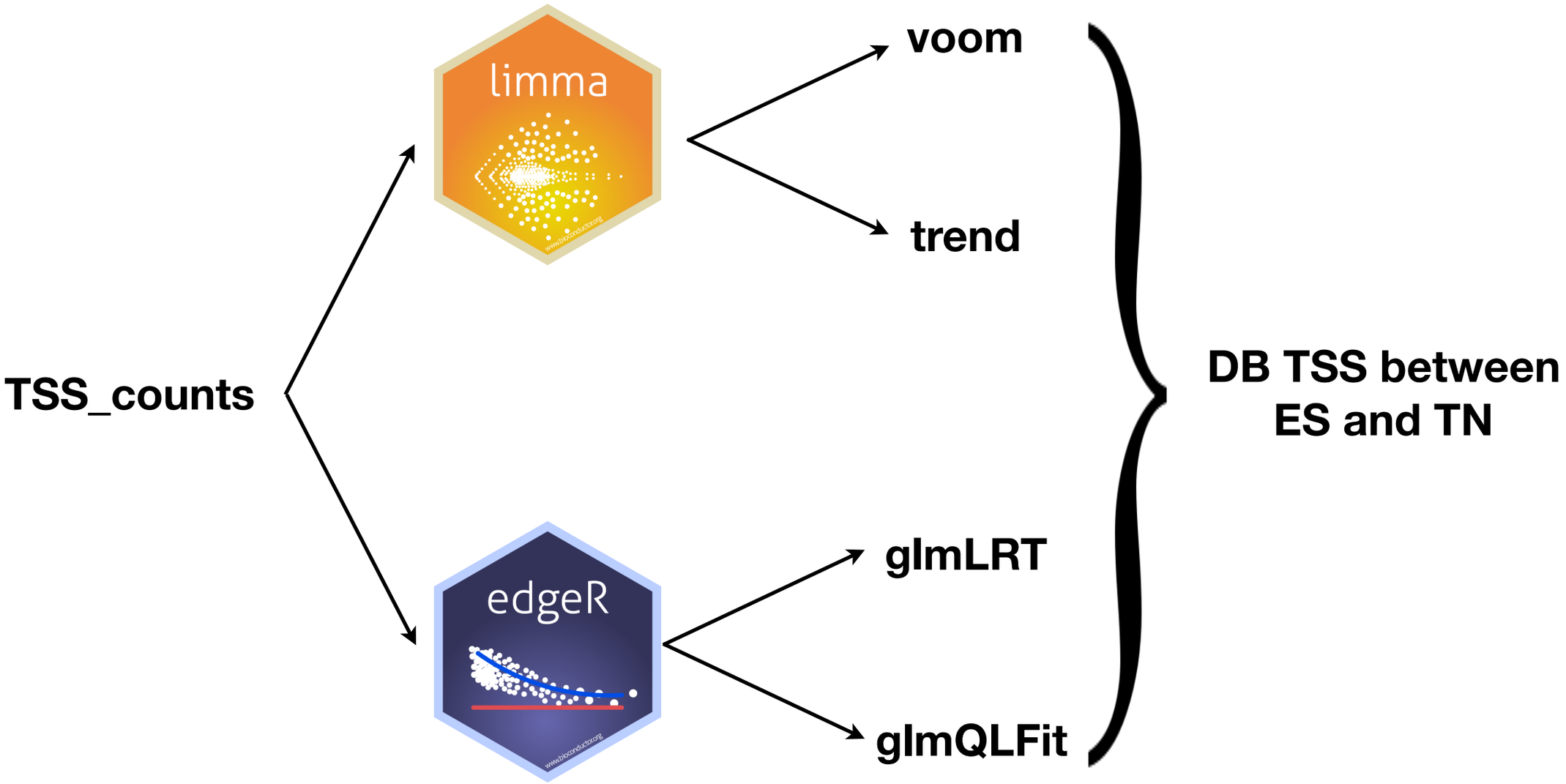
Summarize data around TSS

- Identify TSS of each gene

- Create a region around the TSS - eg. 2000bp upstream and 1000bp downstream of TSS

- Use featureCounts to summarize reads that fall within each region into counts - TSS_counts

DB analysis


limma
voom
trend
TSS_counts
edgeR
glmLRT
glmQLFit
DB TSS between
ES and TN

CSAW

CSAW

CSAW uses sliding windows to identify significant changes in binding patterns across the entire genome

https://bioconductor.org/packages/release/bioc/html/csaw.html

1. Count reads into windows
2. Filter windows
3. Normalize
4. Test for DB using edgeR

CSAW

1. Count read into windows
- Estimate fragment length
- Choose appropriate window size
- Count reads into windows

2. Filter windows (choose 1)
- By count size
- By proportion
- By prior information
- By global enrichment
- By local enrichment
  -Mimicking peak callers
  -Identifying local maxima
  -With negative controls

3. Normalize (choose 1)
- For composition biases
- For efficiency biases
- For trended biases

4. Test for DB
- Either glmLRT or glmQLFit pipeline
- Cluster windows into regions
- Annotate regions

For the experienced user only

References

Data:
Tiwari VK et al. *A chromatin-modifying function of JNK during stem cell differentiation*. Nature Genetics 2011 Dec 18;44(1):94-100.
PMID: 22179133

Subread:
The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3664803/

File formats:
https://genome.ucsc.edu/FAQ/FAQformat.html