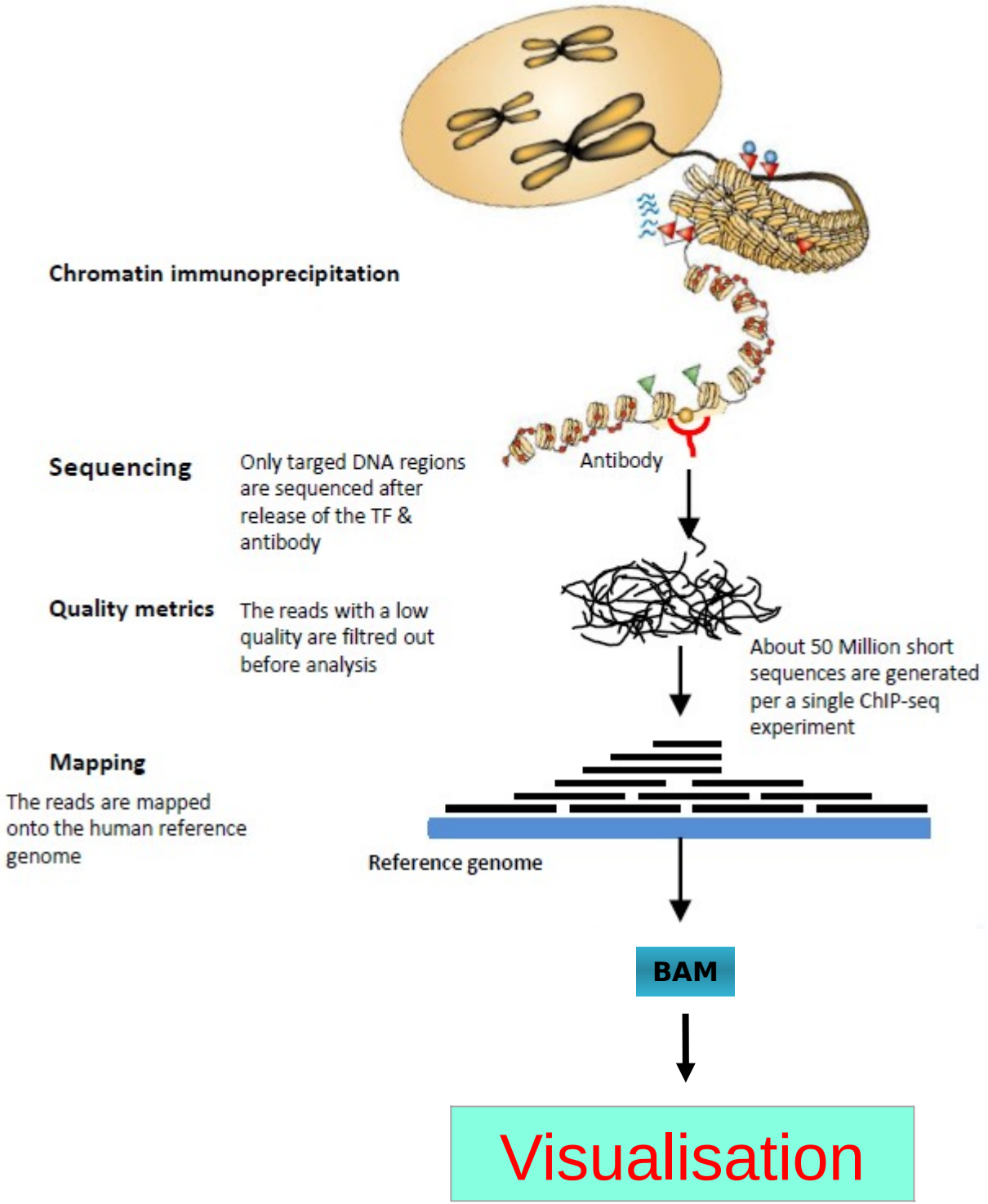
Alignement et visualisation de données NGS

Les mardis de la technologie gourmande de DISC

A. Elkaoutari, CRCM, CIBI

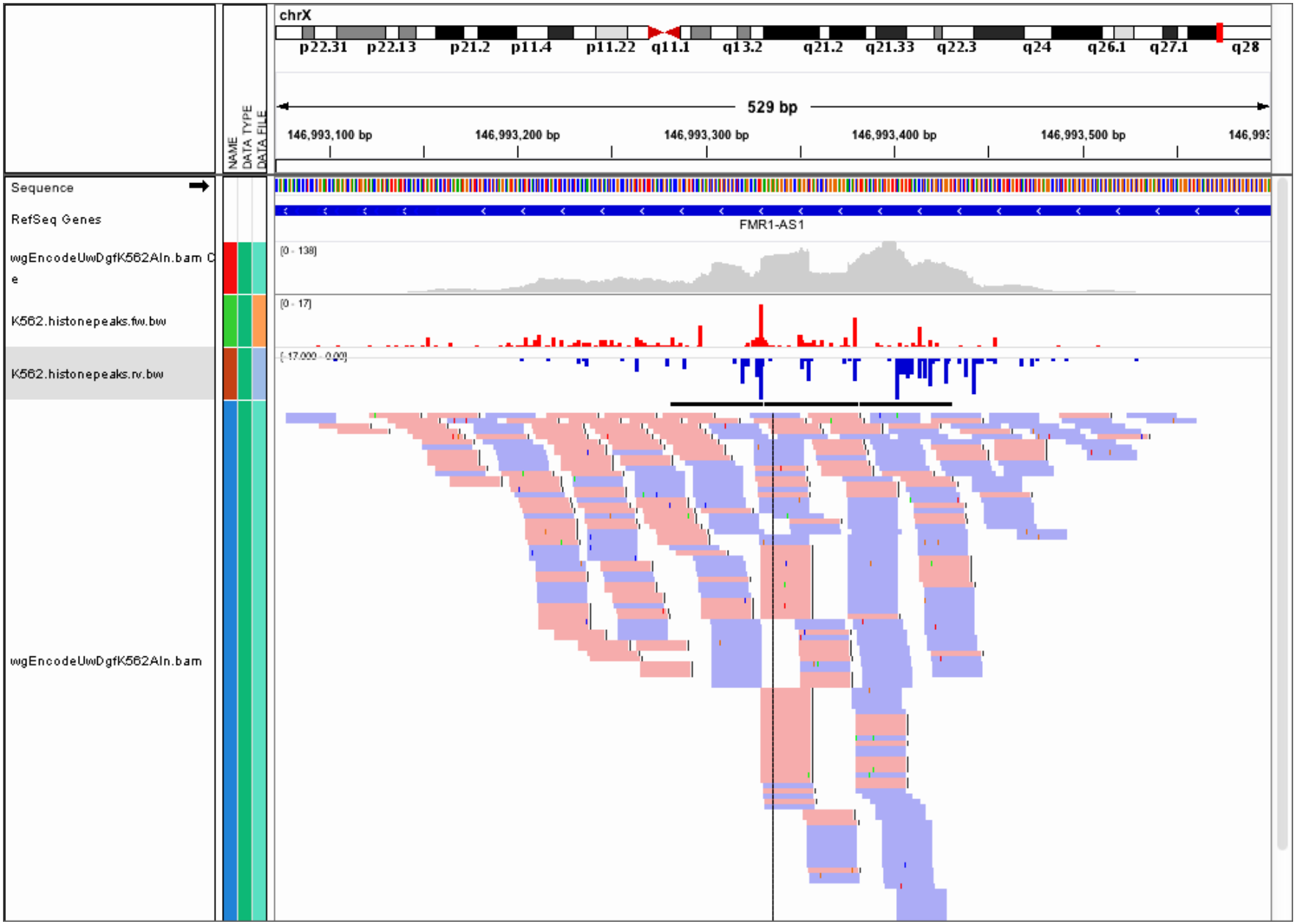
24/05/2016


Chromatin immunoprecipitation
Antibody
Sequencing
Only targed DNA regions are sequenced after release of the TF & antibody
Quality metrics
The reads with a low quality are filtred out before analysis
About 50 Million short sequences are generated per a single ChIP-seq experiment
Mapping
The reads are mapped onto the human reference genome
Reference genome
BAM
Visualisation

Mapping
Reads vs. reference genome

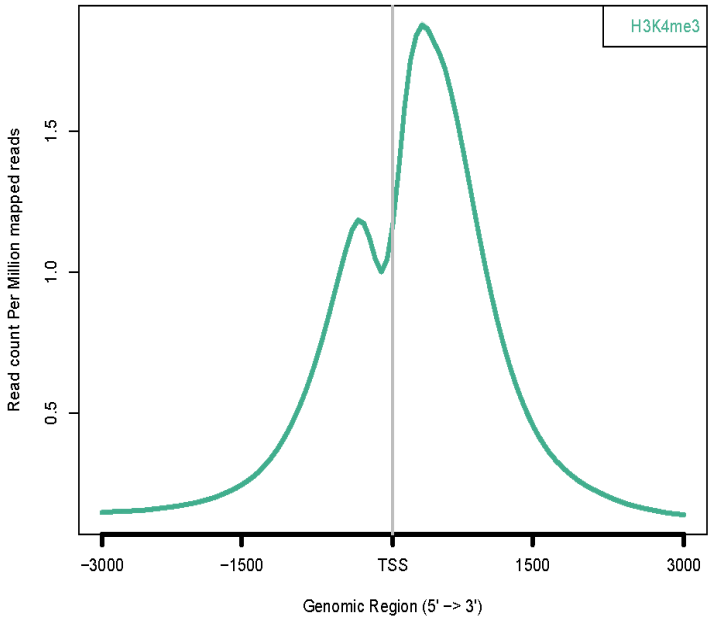
```
#!/bin/bash
# EL KAOUTARI
# 07/07/2015

# faire l'alignement de reads fastq ontre le génome humain hg19
# readsAlign.sh permet de faire les 3 étapes principales d'alignment en utilisant bfast
# en utilisant samtools
        # - convertir les sam aux BAM files,
        # - samtools merge multiple BAM files to single BAM
        # - samtools sort the BAM file
        # - samtools index the BAM file
# ....

echo "launching bfast alignment (readsAlign.sh) and samtools:
        - bfast match
        - bfast localalign
        - bfast postprocess
        - samtools view : from SAM to BAM files
        - samtools merge multiple BAM files to single BAM
        - samtools sort the BAM file
        - samtools index the BAM file"

# Running 1st example, mapping of H3K36me3_shCtrl.fastq against hg19

./scripts/bin/readsAlign.sh -n H3K36_test -f ./Ref_Genome/hg19.fa -r ./seq_data/H3K36me3_shCtrl.fastq

echo "Alignment of 1st dataset done."
################################################################################
# Running 2nd example, mapping of PolII_shCtrl.fastq against hg19

./scripts/bin/readsAlign.sh -n PolII_test -f ./Ref_Genome/hg19.fa -r ./seq_data/PolII_shCtrl.fastq

echo "Alignment of 2nd dataset done."
```

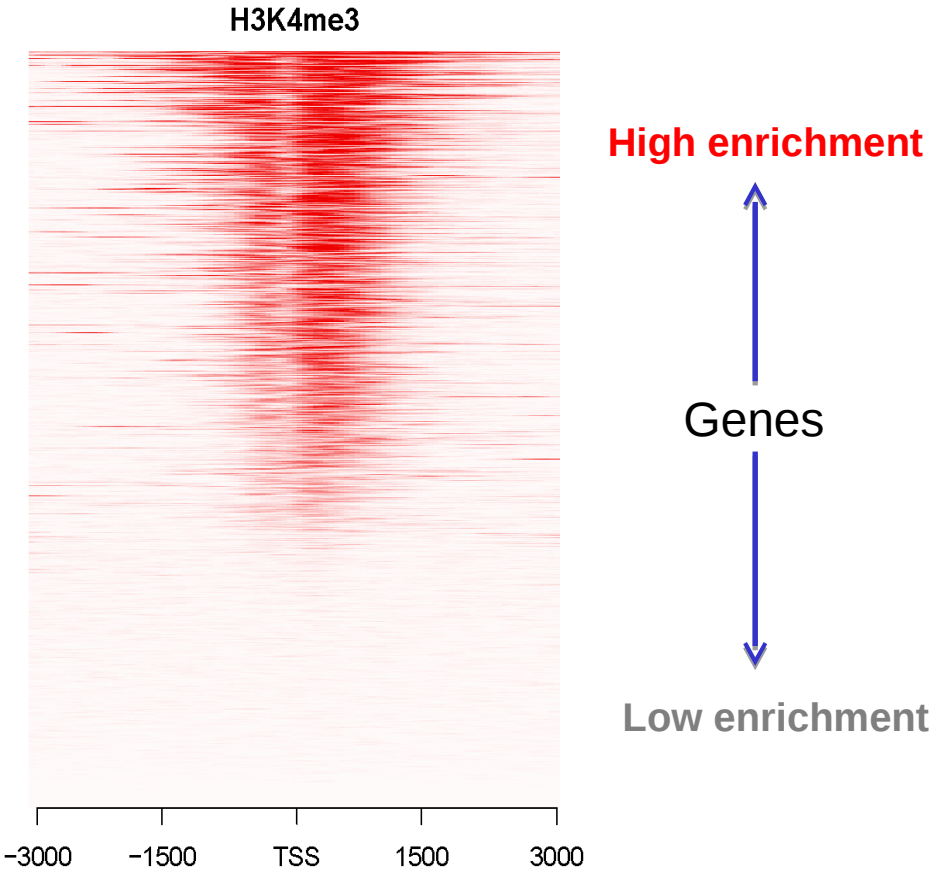
We can open visualize BAMs directly in genome browser



Mining and visualization by integrating genomic DBs
(using ngs.plot.r)

ngs.plot is a program that allows you to easily visualize your next-generation sequencing (NGS) samples at functional genomic regions.





NGSPLOT analysis

https://github.com/shenlab-sinai/ngsplot

https://github.com/shenlab-sinai/ngsplot/wiki/ProgramArguments101

Usage:

ngs.plot.r -G hg19 **-R** genebody **-C** config_file_test.txt **-O** ngsplot_result **-GO** km **-KNC** 10

```
## Mandatory parameters:
  -G   Genome name. Use ngsplotdb.py list to show available genomes.
  -R   Genomic regions to plot: tss, tes, genebody, exon, cgi, enhancer, dhs or bed
  -C   Indexed bam file or a configuration file for multiplot
  -O   Name for output: multiple files will be generated
```

Examples

H3K4 vs. H3K27 trimethylation on all hg19 genes

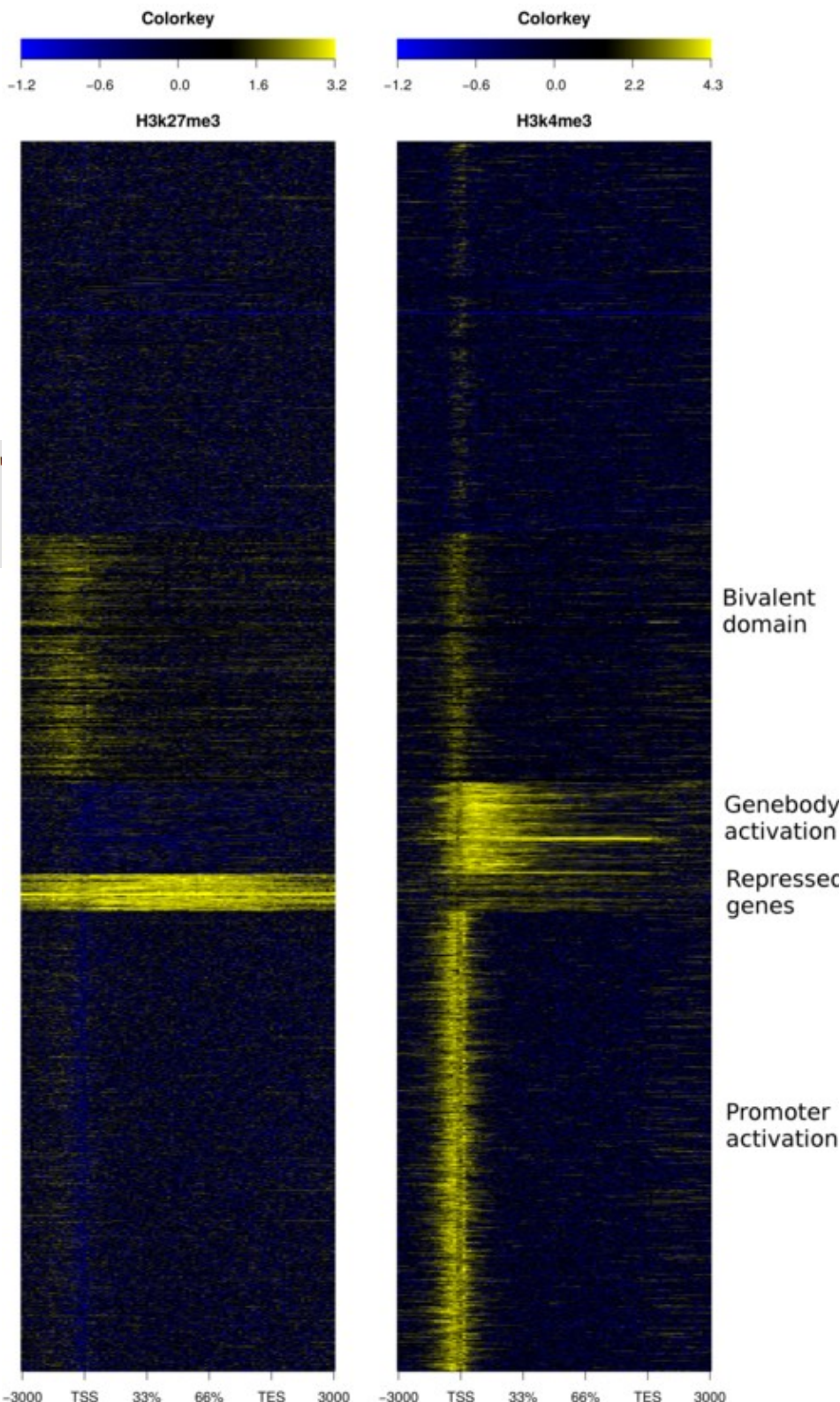
Configure file:

`config.k4k27.txt`

```
hesc.H3k27me3.sort.bam:hesc.Input.sort.bam    -1    "H3k27me3"
hesc.H3k4me3.sort.bam:hesc.Input.sort.bam     -1    "H3k4me3"
```

Command to use:

```
ngs.plot.r -G hg19 -R genebody -L 3000
-C config.k4k27.txt -O k4k27_km -GO km
```



How my genes of interest are enriched on H3K36 ?

The configure file

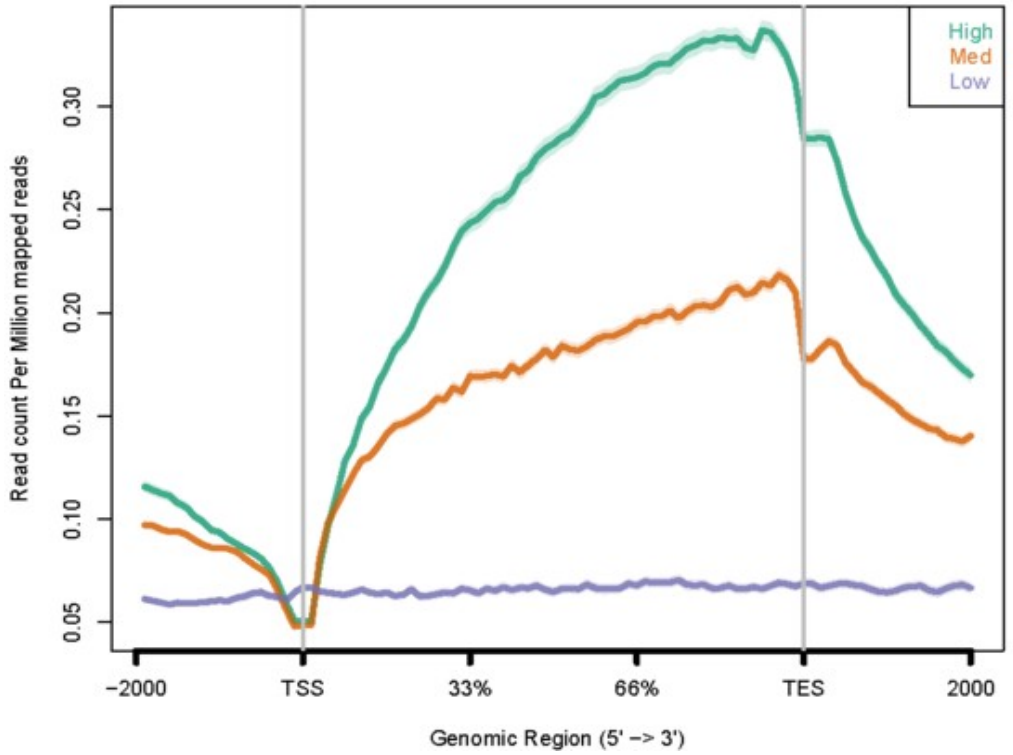
config.hesc.k36.txt

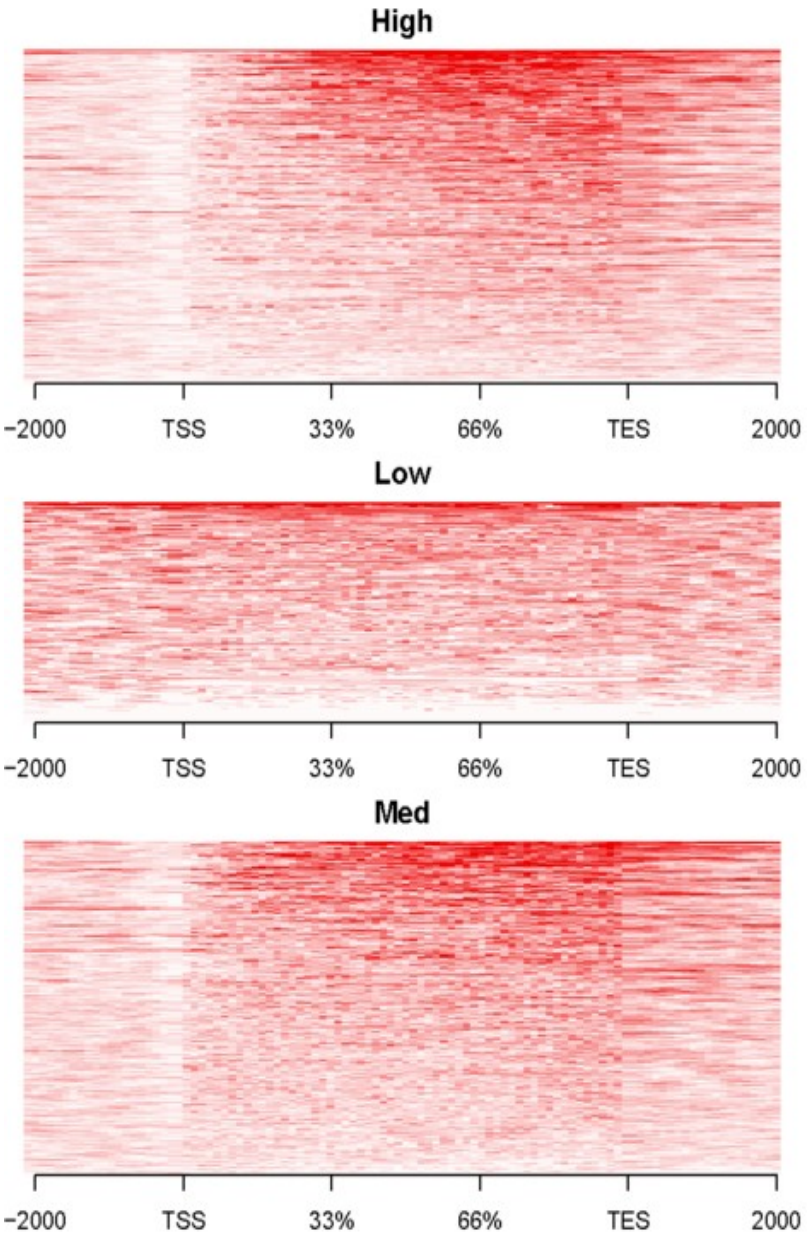
```
hesc.H3k36me3.rmdup.sort.bam     high_expressed_genes.txt          "High"
hesc.H3k36me3.rmdup.sort.bam     medium_expressed_genes.txt        "Med"
hesc.H3k36me3.rmdup.sort.bam     low_expressed_genes.txt           "Low"
```

Command to use

```
ngs.plot.r -G hg19 -R genebody -C config.hesc.k36.txt -O hesc.k36.genebody
-D ensembl -FL 300
```

Average profiles



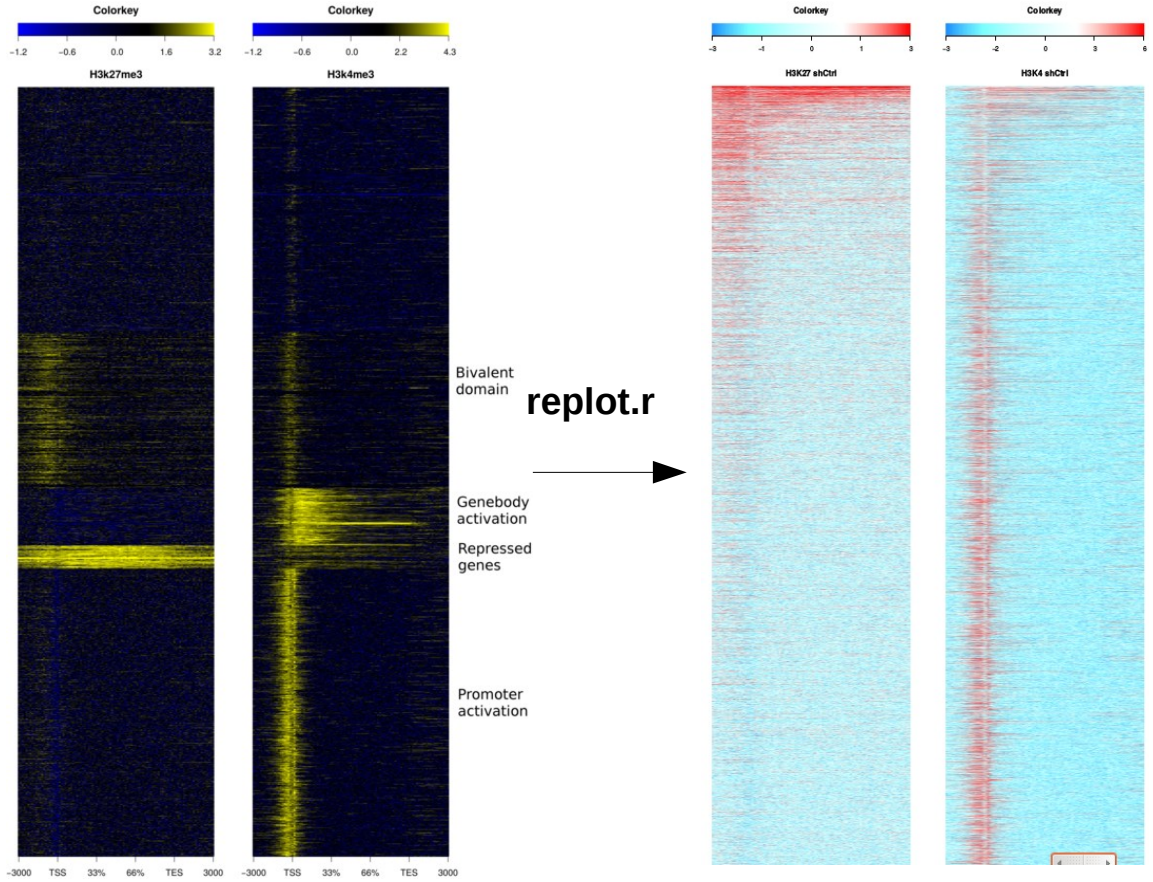


Demo

replot.r

Use replot.r to re-create plots with already generated data using different parameters. There are a couple of options for you to finetune the figures.

```
Usage: replot.r prof/heatmap -I input.zip -O name

## Mandatory parameters:
    -I  Result zip file created by ngs.plot
    -O  Output name
```



Gene order algorithms

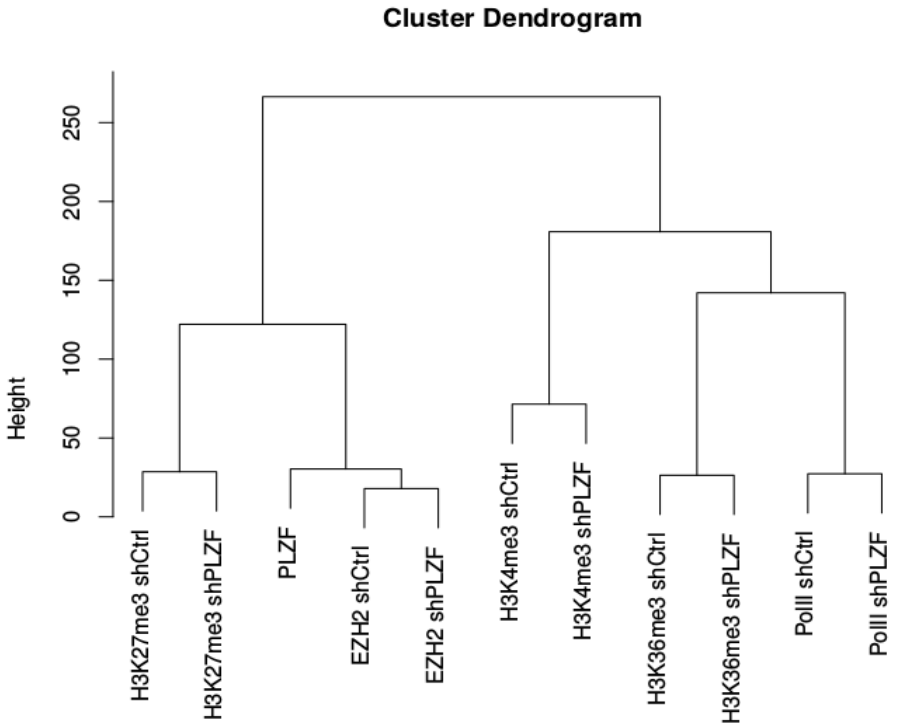
Argument	Explanation	Accepted value and notes
-GO	Gene order algorithm	The algorithm is used to rank genes in heatmaps. Available options:
	total(default)	Overall enrichment of the 1st profile
	hc	Hierarchical clustering
	max	Peak value of the 1st profile. This option makes more sense if the epigenomic mark tends to generate sharper peak.
	prod	Product of all profiles on the same region.
	diff	Difference between the 1st and the 2nd profiles.
	km	K-means clustering. The default number of clusters is 5.
	none	No ranking algorithm applied. Use order provided in the gene list. This can be used to your advantage. For example, you can rank genes by their expression levels and give the list to ngs.plot and see how the epigenomes change with expression.

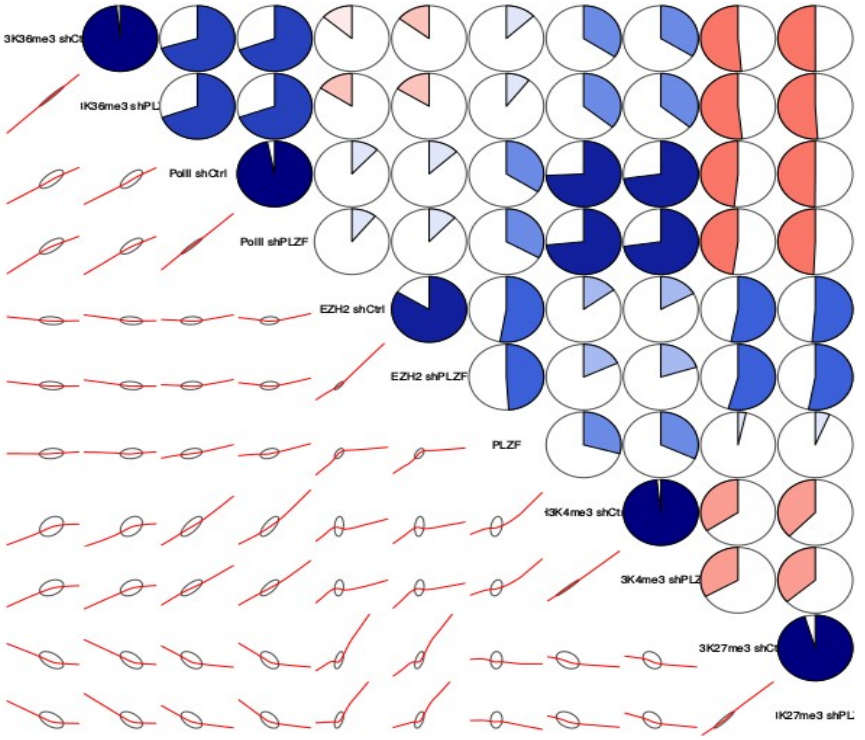
plotCorrGram.r

Demo

Create a corrgram from ngs.plot output files.

```
Usage: plotCorrGram.r -I ngsplot_output.zip -O output_name [Options]
## Mandatory parameters:
  -I   Result zip file created by ngs.plot.
  -O   Output name
## Optional parameters:
  -M   Method used to calculate row stat.
       mean(default): mean of each row.
       max: max of each row.
       window: mean on center region.
  -P   Options for -M method.
       mean: [0,0.5) - trim value for robust estimation, default is 0.
       window: [0,0.5),(0.5,1] - window borders, default:0.33,0.66.
  -D   Options for distance calculation in hierarchical cluster.
       This must be one of 'euclidean'(default), 'maximum', 'manhattan', 'canberra', 'binary' or 'minkowski'.
  -H   Options for agglomeration method in hierarchical cluster.
       This must be one of 'ward'(default), 'single', 'complete', 'average', 'mcquitty', 'median' or 'centroid'.
```





Demo

Sur **Alambic**

```
cd /home/elkaoutari/DemoTechGourm

oarsub -I -n ngsreplot -l host=1/core=10,walltime=2

source /home/progs/ngs/ngs_source

# replot.r

replot.r heatmap -I profiles_H3K36_PolII.zip -O profiles_test -GO total -SC -3,3 -CO blue:white:red

replot.r heatmap -I profiles_H3K36_PolII.zip -O profiles_test -GO total -SC -2,2

replot.r prof -I profiles_H3K36_PolII.zip -O prof_H3K36_PolII

replot.r prof -I prof_allGenes_sans_sh.zip -O avgProf_allGenes

# plotCorrGram.r

plotCorrGram.r -I prof_allGenes_sans_sh.zip -O correlogram_all

plotCorrGram.r -I profiles_H3K36_PolII.zip -O correlogram_H3K36_PolII
```