Coverage Footprints & Heatmaps

A quick tutorial using

SeqPlots

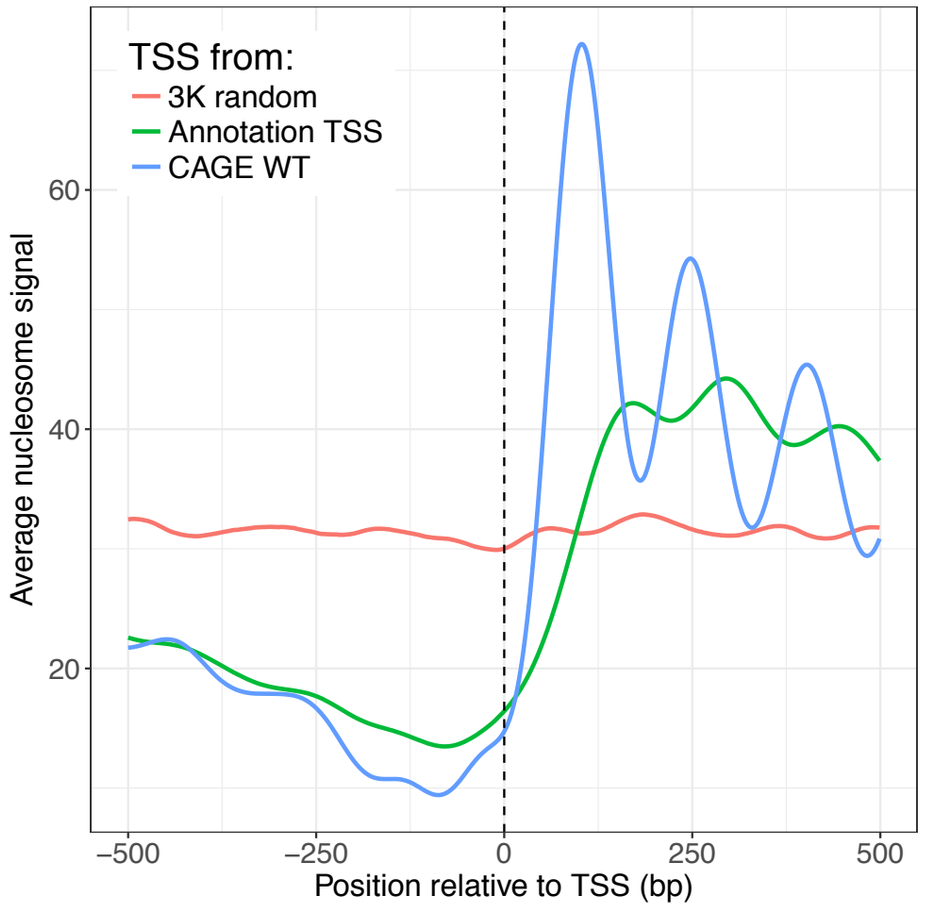
Bioinformatics of High-Throughput Analysis 2017 - Axel Thieffry

Footprint: The Concept

- Summarisation of a signal across many different genomic locations of interest

- Usually ***mean*** or ***median*** is used to summarise the information

- Different names: TSS-plots, meta-gene plots, profile-plots, …

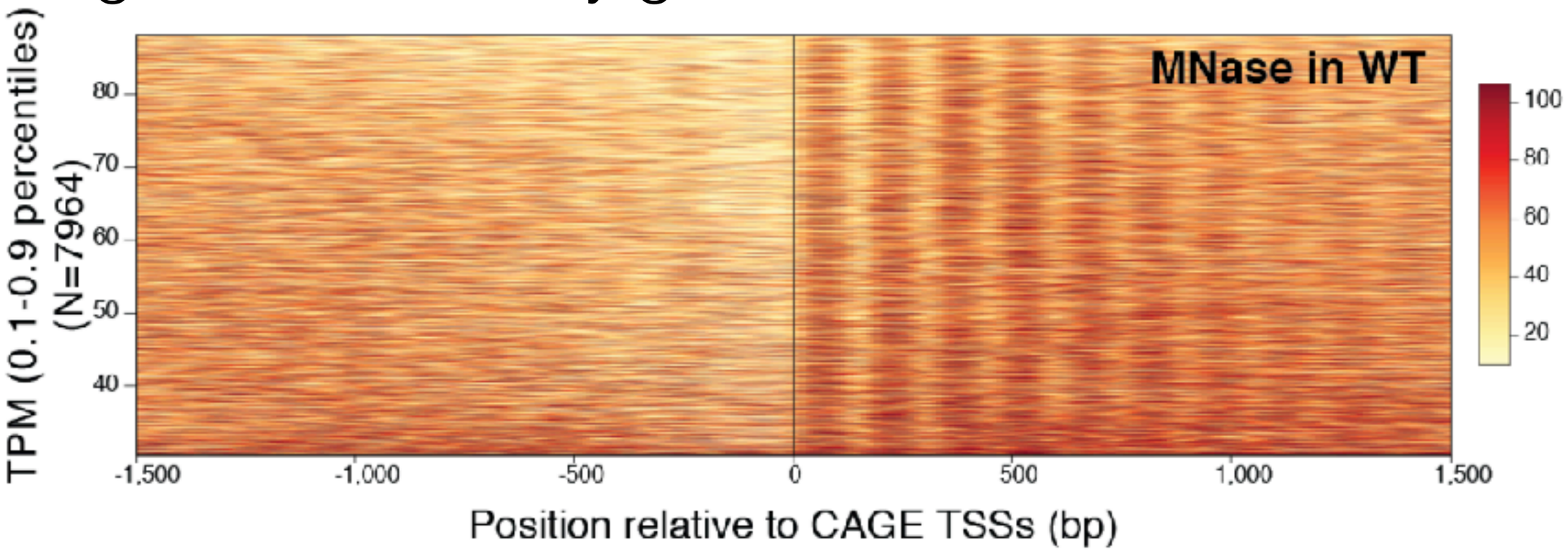


✓ **Pros**
- easy to compare different signal
- summarisation
- readability

✗ **Cons**
- possible stratification masked by statistic
- statistic might be sensitive to outliers (mean)

Heatmap: The Concept

- Non-summarised, colour-based visualisation of signal across many genomic locations of interest



✓ **Pros**
- can reveal stratification/subgroups
- clustering & ordering of signal (overall expression, width of DHS, distance to another reference point…)
- more credibility: non-summarised
- pretty and impressive

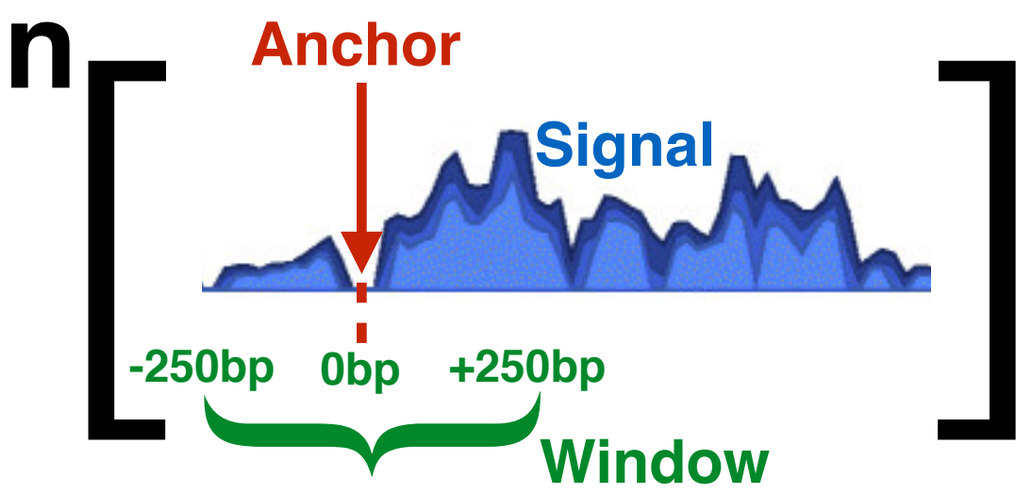
✗ **Cons**
- difficult to compare != signals
- arbitrary mapping to colour scale
- requires some reading habit
- large image processing

The required ingredients

- **Signal(s):**
  - coverage from sequencing experiments
  (RNA-seq, CAGE-seq, MNase-Seq, ChIP-seq, …)
  - computed statistics: di-/trinucleotide frequencies,
  logFC, p-values, ratio to a control, motifs, patterns…

- **Reference points (aka *Anchors*):**
  List of genomic coordinates of interest such as:
  TSSs, TTSs, Enhancer midpoints, DHS Summits, …

- **Window:** length of the investigated region
  - spanning the reference points
  - can be asymmetric
  - ex: TSS +/- 250bp



n = as many reference points as required

Similar recipes

1. Compute signal on a genome-wide basis
   - most common format is **BigWig**
2. Get genomic regions of interest
   - most common format is **BED**
3. Establish window length and symmetry status
   - most common format is **BED**, can also be on the fly
4. Recover signal at all windows and stack
   - most common format is a **matrix**
     > *<u>rows</u> = genomic feature,*
     > *<u>columns</u> = signal at each window point*

- **Footprint:**
  - summarise on columns (mean/median)
  - plot

- **Heatmap:**
  - map signal to colour scale
  - plot

Considerations

- Signals on — strand must be flipped to always look in the transcriptional direction: 5’ -> 3’
- Asymmetrical window: flipping will produce a shift, need to correct.
  ex: if you want a -150/+300 window, make a matrix of +/-300, flip when required, and then discard the first 150 columns
- Signal most probably requires normalisation
- Overlapping windows will artificially inflate the signal and lead to mis-interpretations

Tools

- Usually using the command line and dedicated software suite such as **BEDtools**, with custom scripts (**bash, python, perl…**). Laborious, error-prone and difficult.

- **R & Bioconductor:** high-level functions make the task easier but there is a learning curve

- **SeqPlots:** quick and easy GUI based on R/Bioconductor

SeqPlots - installation

- **install SeqPlots from R**:

System requirements:

- R 3.1 or higher

How to install

To install SeqPlots package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
          biocLite("seqplots")
```

How to start

To start SeqPlots web interface, start R and enter:

```
library(seqplots)
          run()
```

SeqPlots - Task

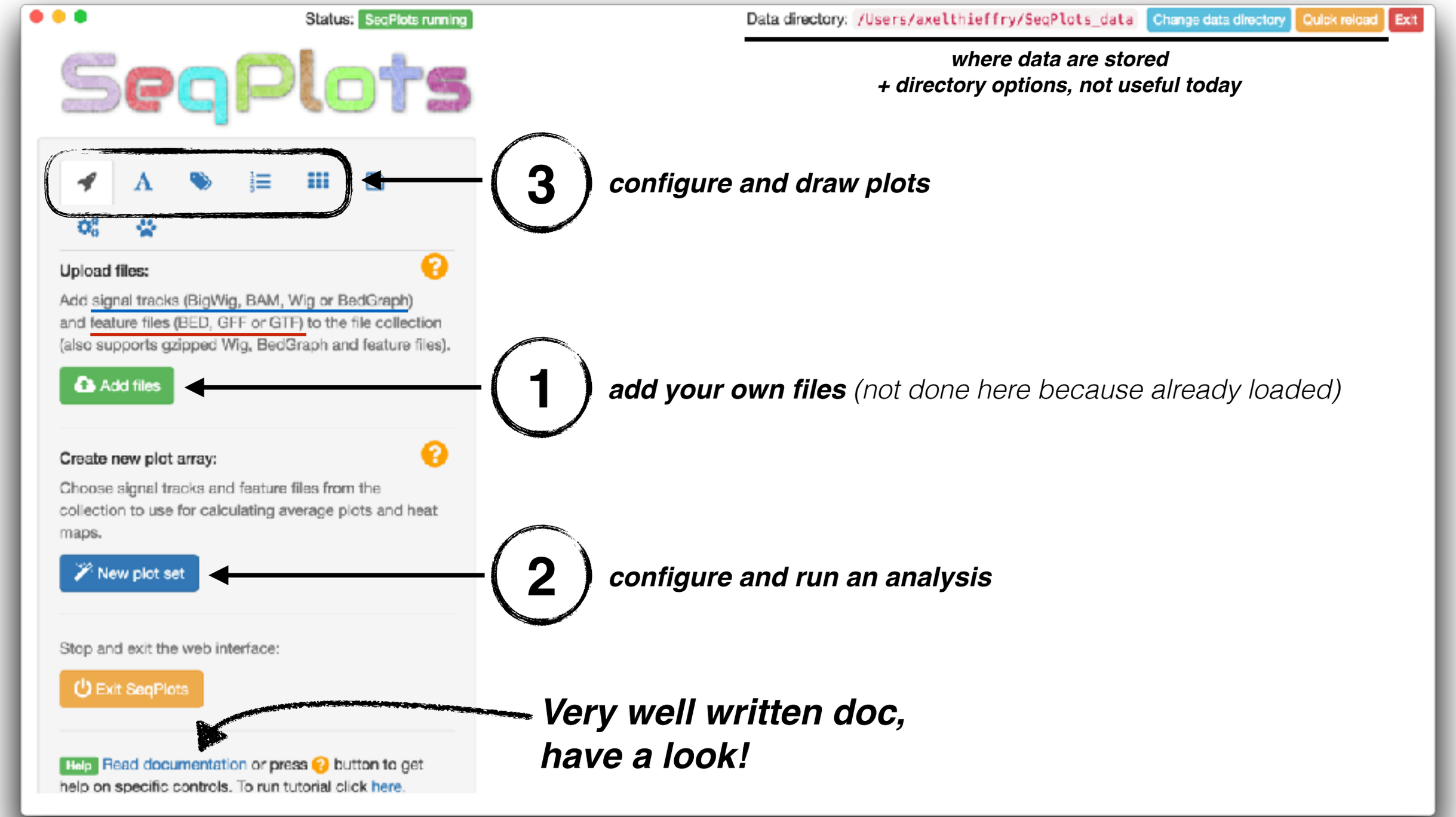
- SeqPlots comes with pre-loaded data on ***C. elegans***, a roundworm model organism

- **Signal data:** histone marks
  - H3K36me3
  - H3K4me3

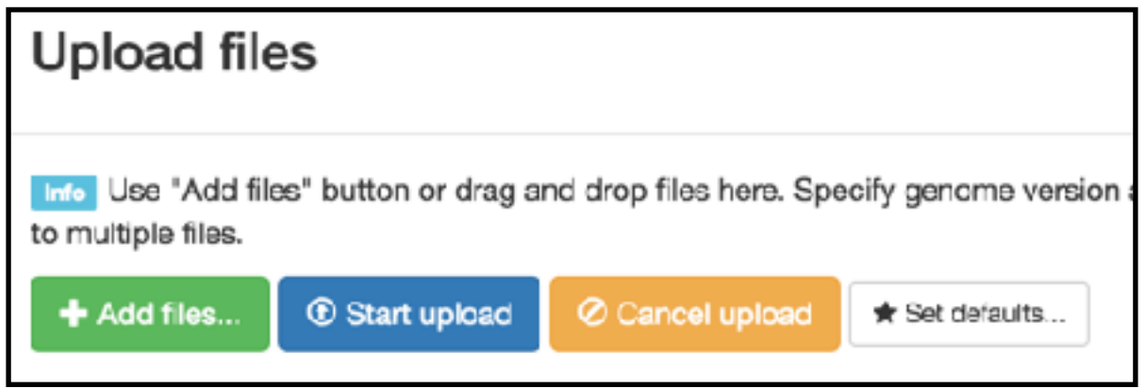
- **Genomic Coordinates:** (subset for chr1 only)
  - Top 20% expressed genes
  - Bottom 20% expressed genes

- **Objective:**
  Look at histone marks around all Top 20 genes starts (TSSs)

SeqPlots - Menu & Workflow



1 - Adding files
& quick reminder on common formats



Anchors

BED 6

chr	start	end	ID	score	strand
III	43480	43844	III:43480-43844,+	.	+
III	50292	50526	III:50292-50526,+	.	+
III	62686	62998	III:62686-62998,+	.	+
III	68088	68530	III:68088-68530,+	.	+
III	72916	73482	III:72916-73482,+	.	+
III	77080	78043	III:77080-78043,+	.	+
III	80382	80546	III:80382-80546,+	.	+

- Score not important here
- Strand can be omitted if non-stranded data
- Omit stuff in BED format with a dot "**.**"

Signals

bedGraph

chr	start	end	score
I	6664	6665	0.0861972
I	13391	13392	0.0861972
I	13821	13822	0.0996363
I	15465	15466	0.0860478
I	15505	15506	0.149454
I	15561	15562	0.101189
I	18495	18496	0.0860478
I	36403	36404	0.101189

wig
- a variation of bedGraph

SAM
- from a mapping, very detailed,
- very heavy

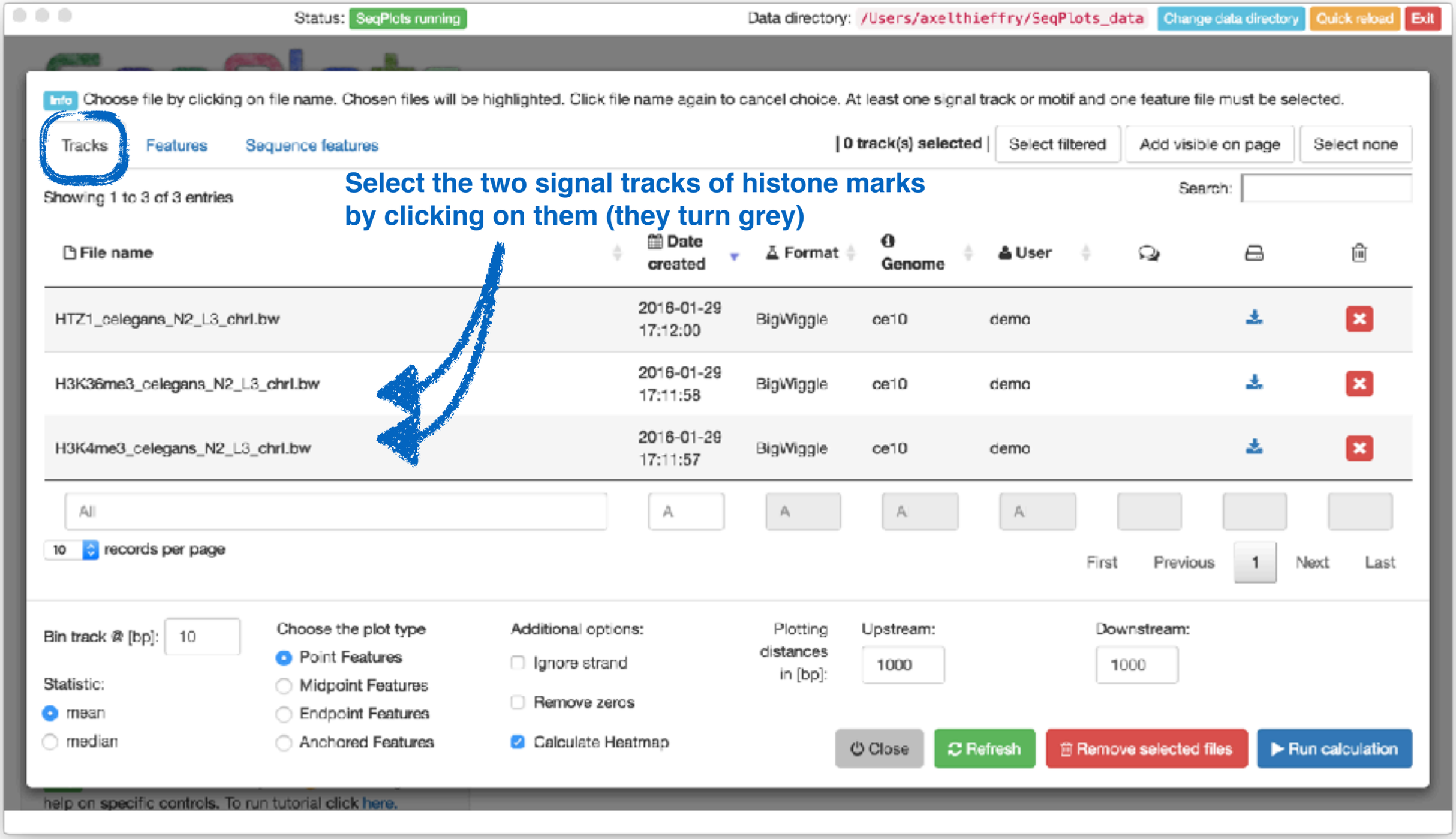
"binarisation"

makes if faster & lighter
cannot be viewed directly

bigWig
BAM

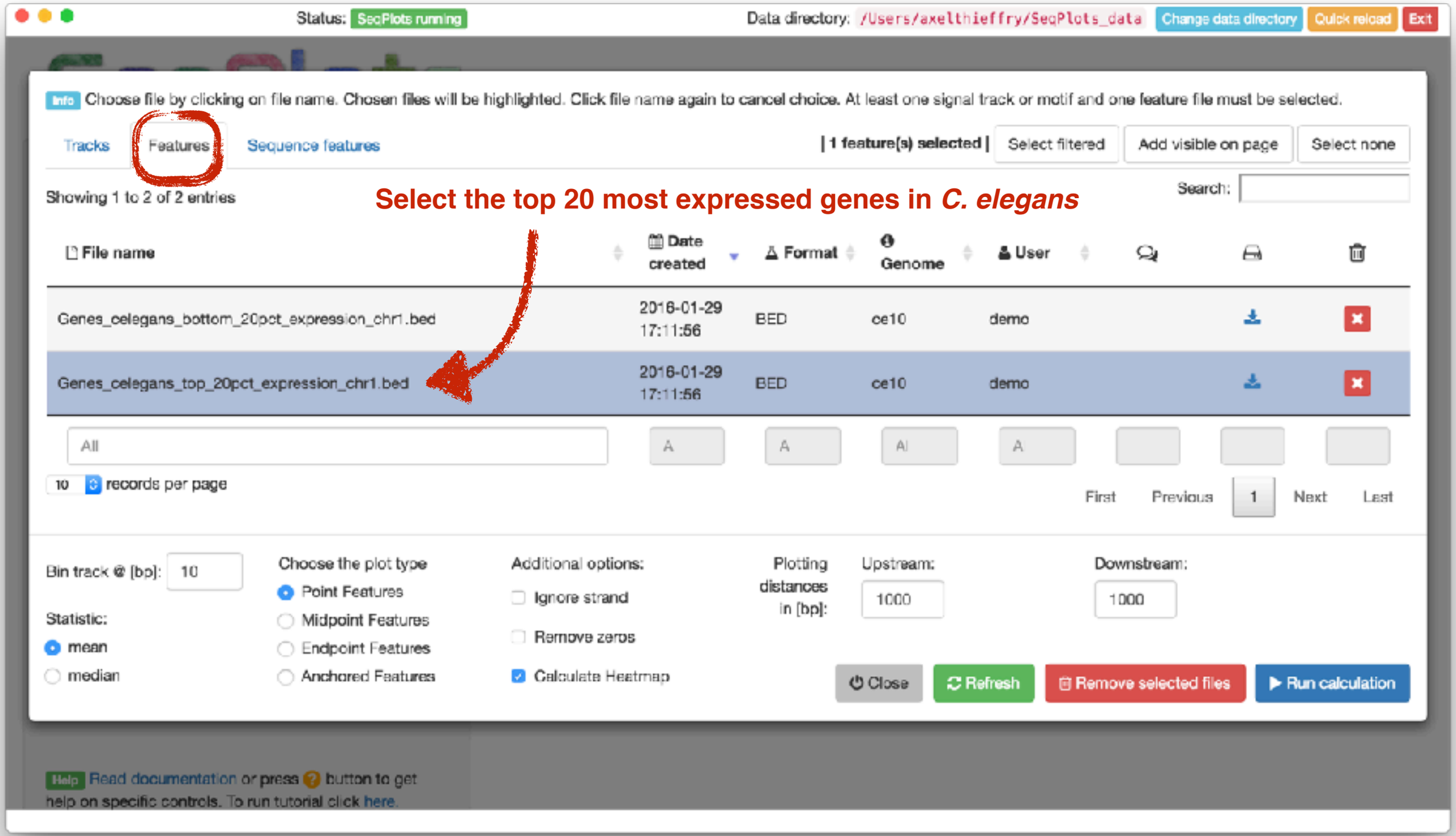
2 - Running an analysis:

Select signal tracks

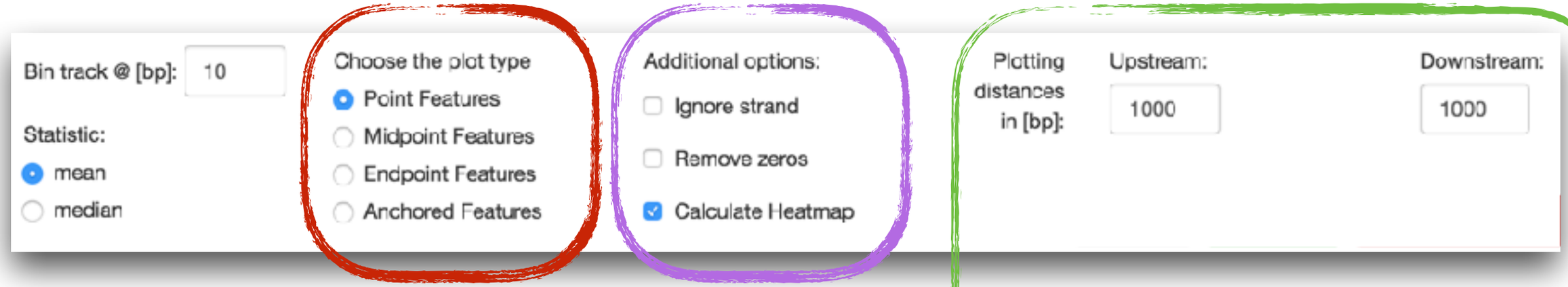


2 - Running an analysis:

Select reference points/anchors

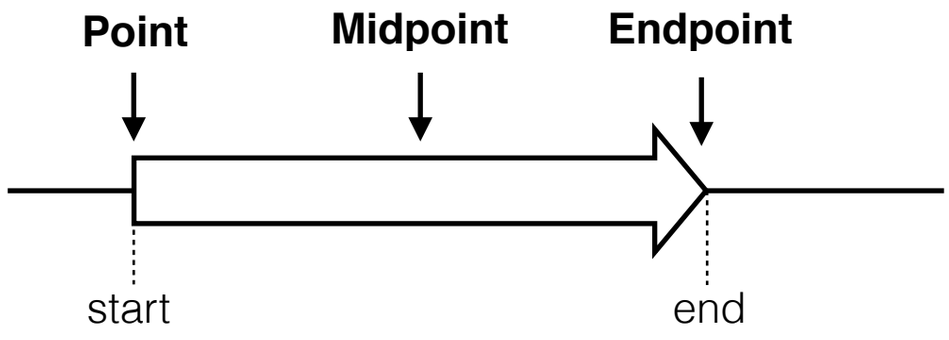


2 - Running an analysis:
Configuration panel (bottom)



Signal binning option and related statistic

What is your reference point, given the genomic region provided in your BED file?



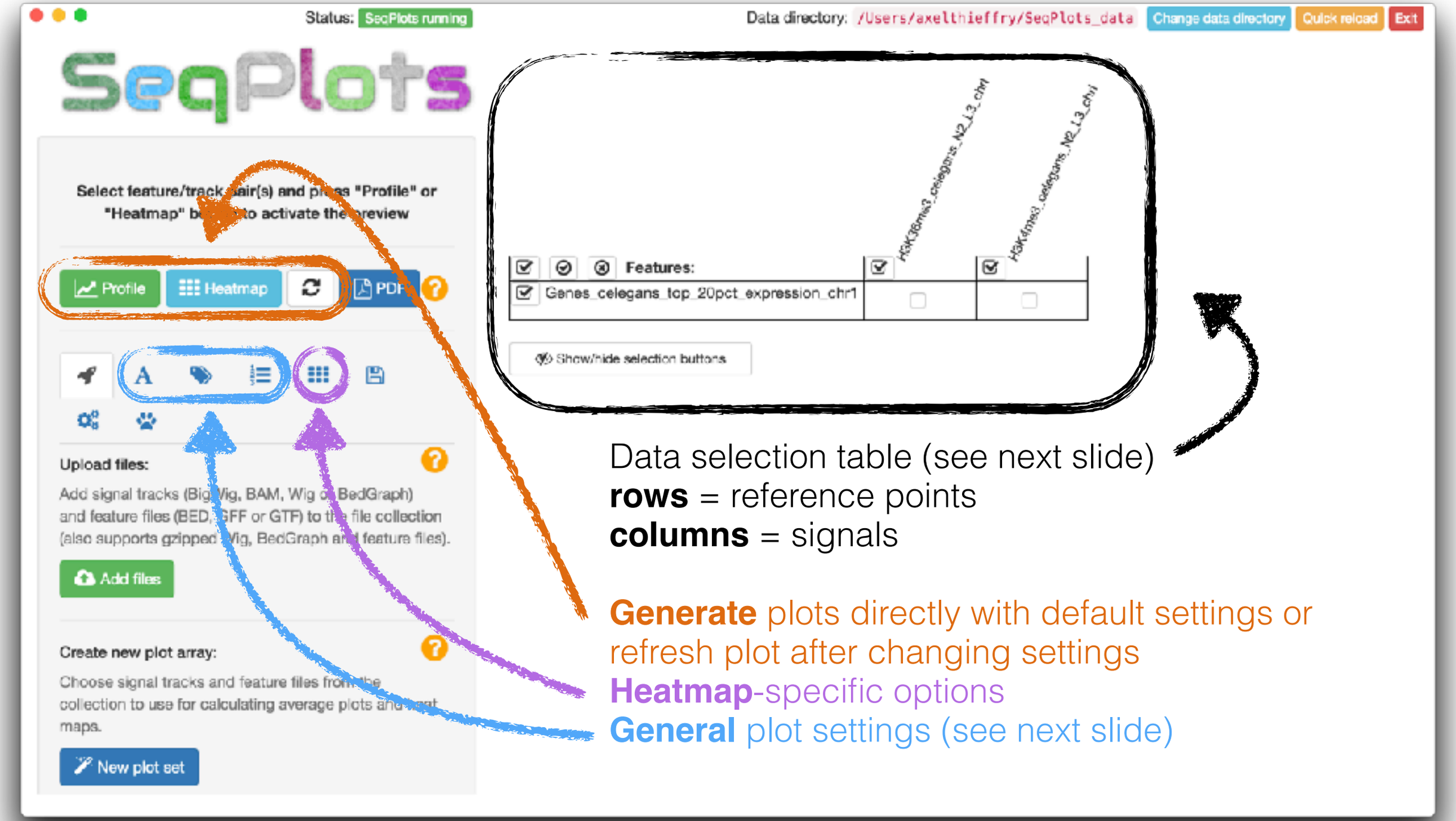
*Special case: '**Anchored**' will make a window from Start to End, and scale all genes to a user-defined pseudo-length*

And finally : ▶ Run calculation

window length:
upstream & downstream
of the **reference point**
- here, asymmetric window will be taken care of automatically

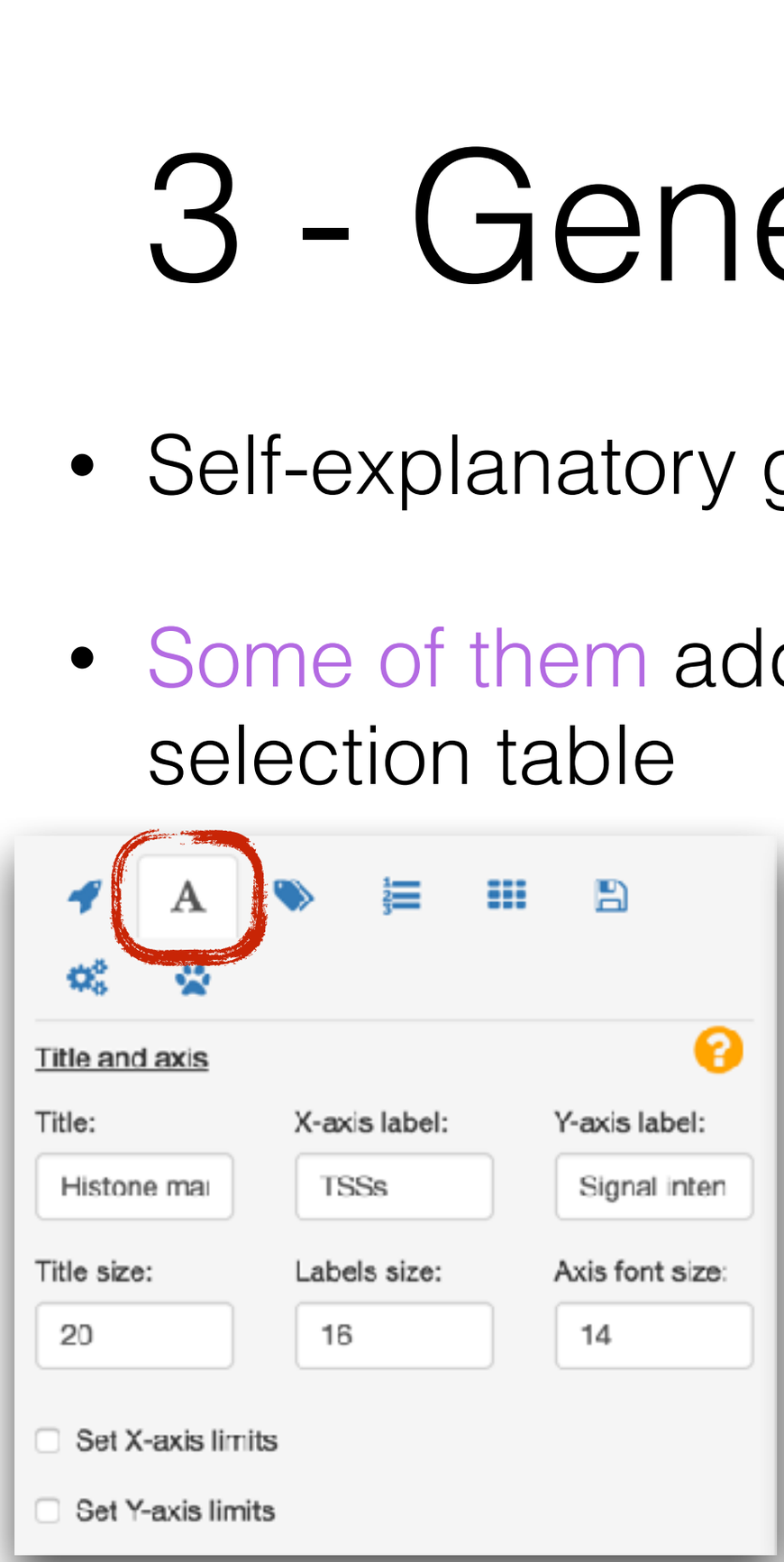
Additional options:
- ignoring strand information:
will not 'flip' the signal according to the strand
- un-check the "calculate heatmap" to decrease calculation time if you are not interested in heatmap plots.

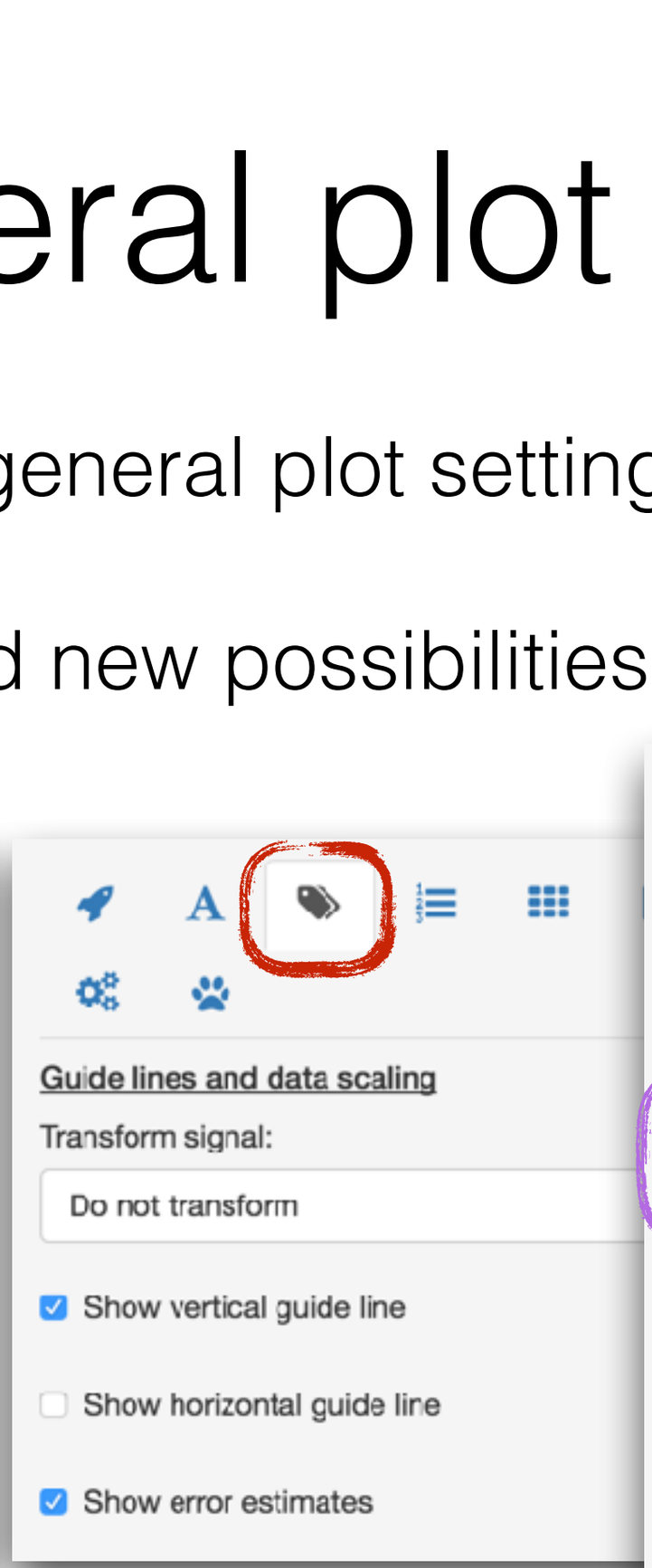
3 -Overview of plot configuration:

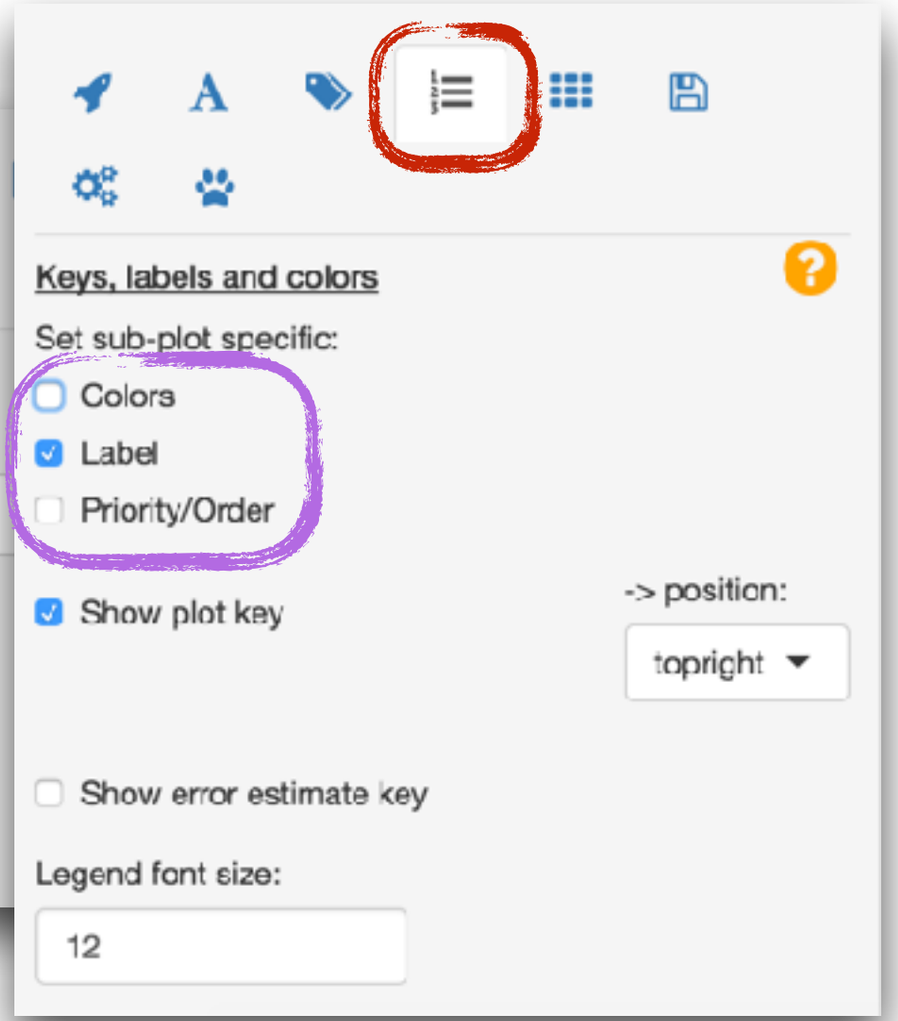


3 - General plot settings

- Self-explanatory general plot settings
- Some of them add new possibilities in the data selection table

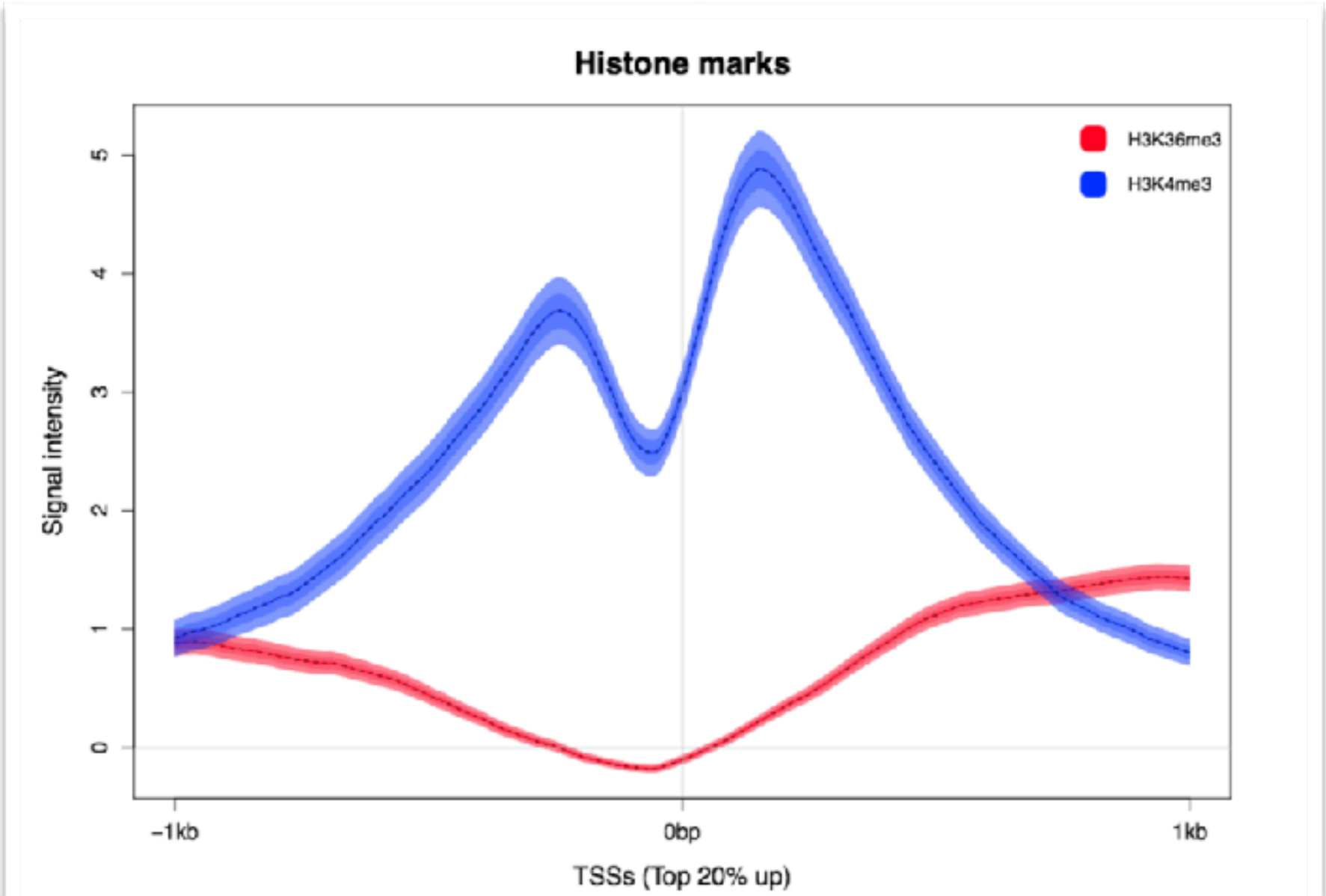






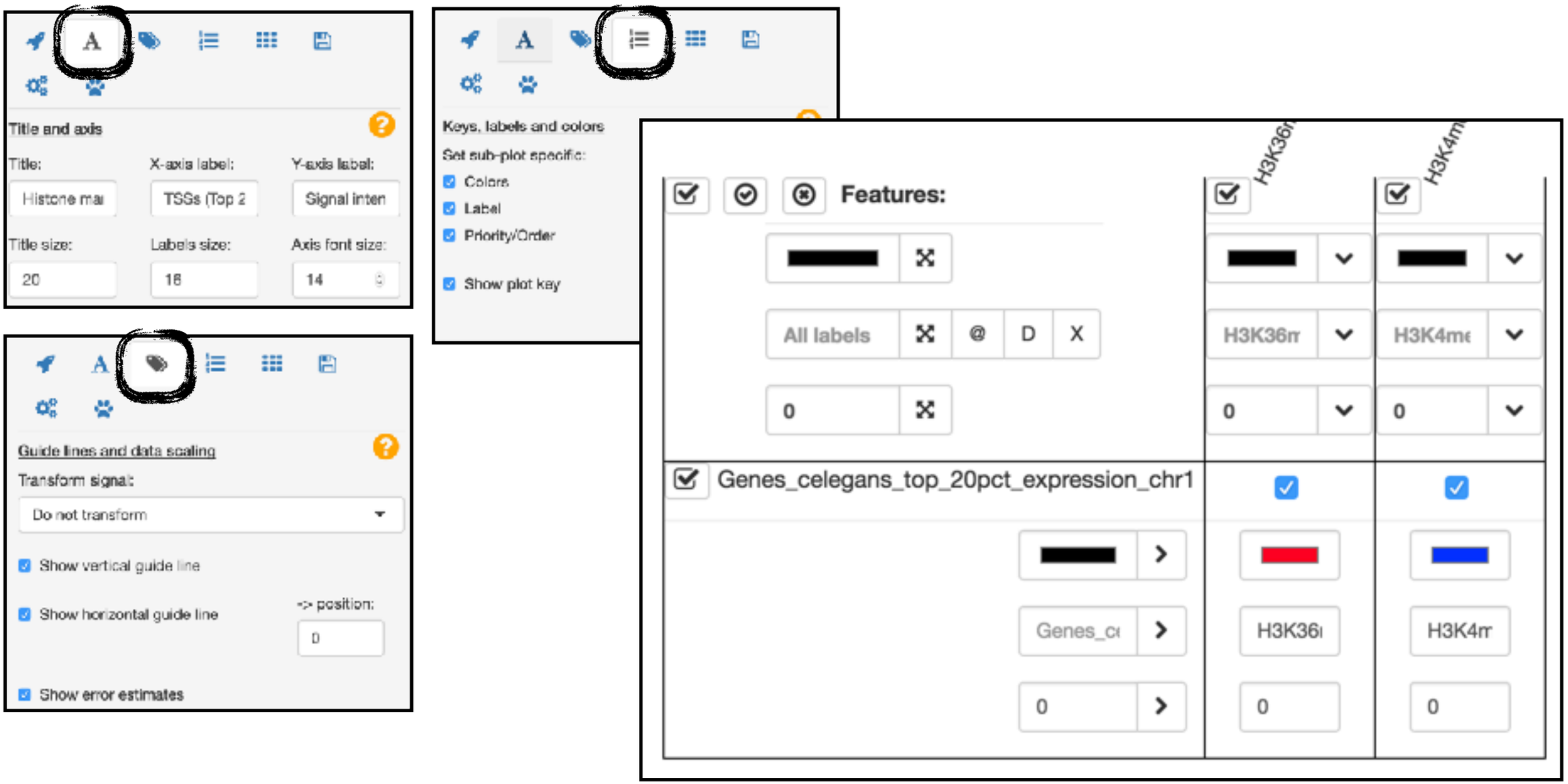
SeqPlots - Footprint

- Try to reproduce the following footprint (aka Profile)



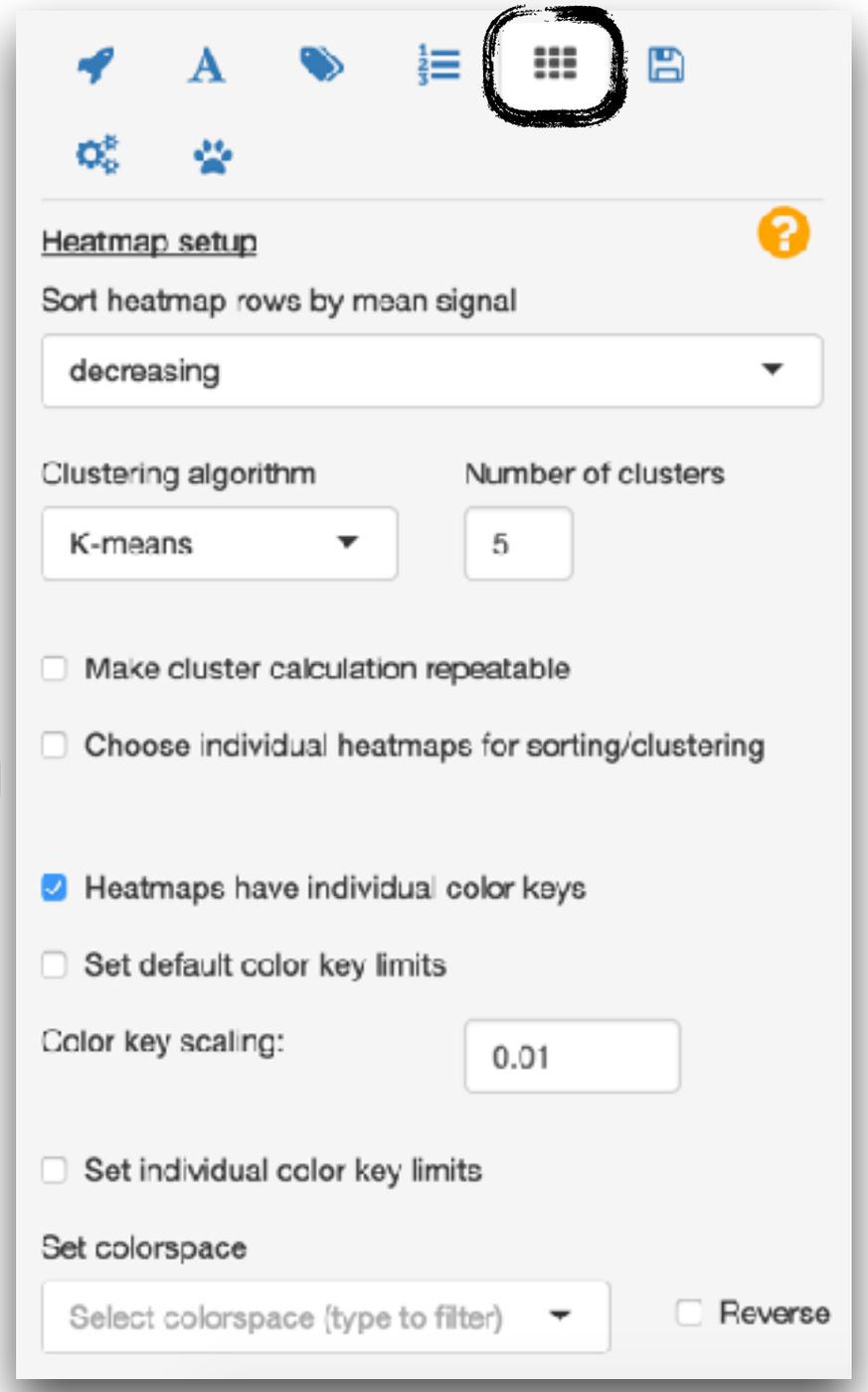
SeqPlots - Footprint answer

- Try to reproduce the following footprint (aka Profile)



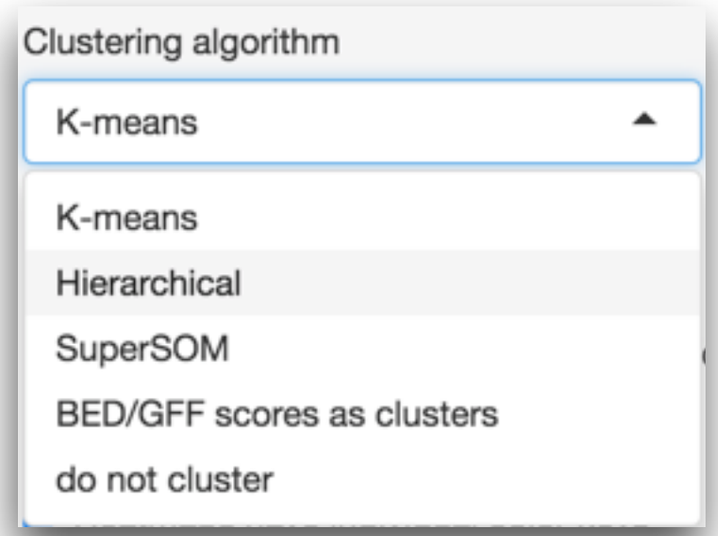
SeqPlots - Heatmap settings

- Most important options:
  - sorting
  - clustering algorithm
  - which signal is used to make the clusters and the sorting? The other signals will be relative to that
  - individual colour keys useful when need to compare != signals
  - colour key scaling
  - colourspace = which palette



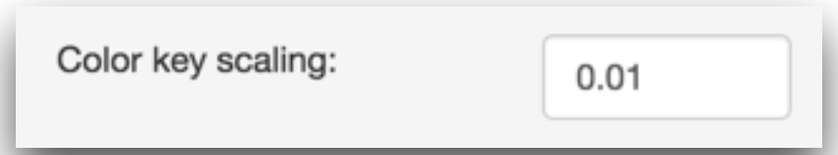
SeqPlots - Heatmap clustering & colour scaling

- Hierarchical clustering can take a significant amount of time on big matrices (binning becomes useful)



- BED scores (column #5) can be used for clustering

- **Colour key scaling:** how colour key limits are generated. For example, 0.01 (default value) calculates limits using data ranging from 1-99 percentile of available data points. 0.1 uses data ranging from 10-90 percentile. From experience, **0.1 works well**, it gets rid of extreme values which would mask all other signal.

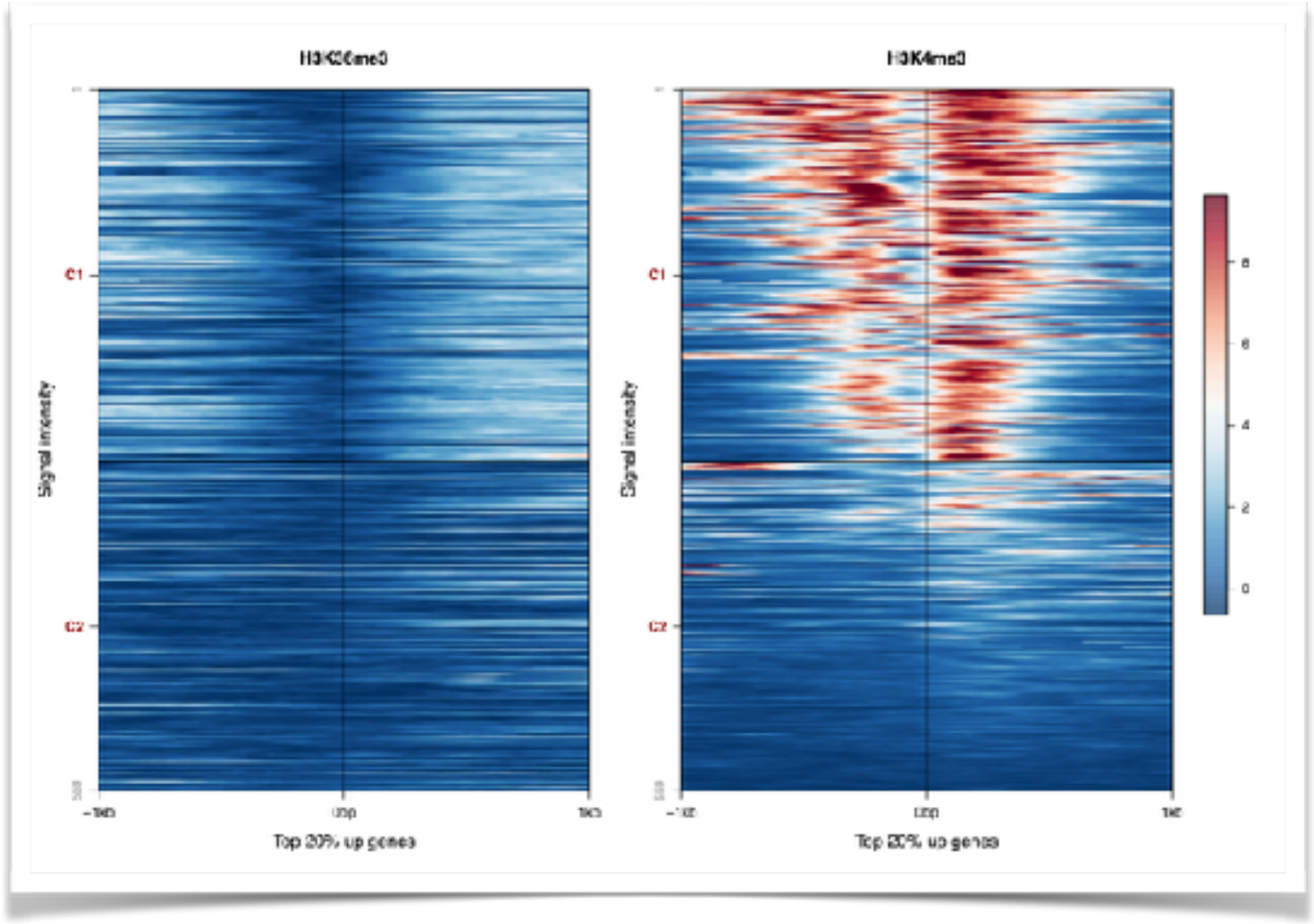


SeqPlots - Heatmap

- Try to reproduce the following heatmap (Heatmap)

- What does the heatmap show that we cannot see in the profile plot?

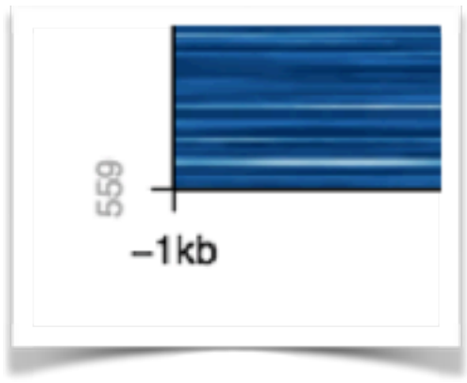
- How many reference points are showed?



SeqPlots - Heatmap answer

- Plotting parameters (Variants of this are of couse valid):
  Colour key scaling of 0.01 works well here
  Clustering: K-means (2 clusters)
  Colourspace: RdBu (reversed)
  Individual heatmap for sorting/clustering: H3K4me3

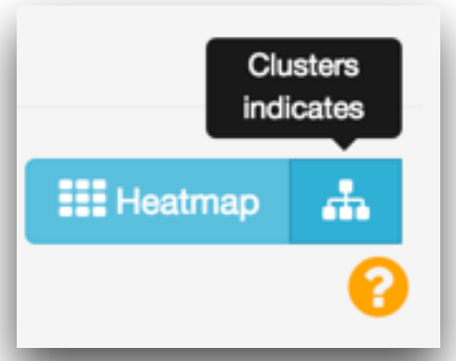
- There are **559** reference points showed



- Heatmap shows that roughly 50% of the
  Top 20% expressed genes (reference points)
  have a strong H3K4me3 presence at TSSs
  (i.e. cluster1, C1, right panel). The profile plot let us suggest
  that all of them have, error estimates are misleading here.

SeqPlots - Exporting matrix

- Exporting the heatmap matrix after parameters for further processing (R / bash/ …)
- CSV format: comma separated values



- Has header to explain each field

```
axelthieffry@UCPH:~/Desktop$ head Clusters_2017-04-27_11-42-20.csv
"chromosome","start","end","width","strand","metadata_name","metadata_score","originalOrder",
"ClusterID","SortingOrder","FinalOrder","RowMeans"
"chrI",8329054,8329677,624,"-","WBGene00004439",36.7803,77,1,226,1,1.91196853992642
"chrI",8329054,8329738,685,"-","WBGene00004439",36.7803,75,1,230,2,1.92430960841687
"chrI",14720576,14721550,975,"-","WBGene00010579",8.10661,211,1,231,3,1.93475108896068
"chrI",12728162,12729861,1700,"+","WBGene00006889",3.78969,459,1,238,4,2.00630422568946
"chrI",12728162,12729861,1700,"+","WBGene00006889",3.78969,460,1,239,5,2.00630422568946
"chrI",12728147,12729861,1715,"+","WBGene00006889",3.78969,458,1,240,6,2.00674199865994
"chrI",8434423,8435808,1386,"+","WBGene00004931",6.41112,300,1,242,7,2.03684042918281
"chrI",9609494,9611662,2169,"-","WBGene00002070",4.55995,399,1,243,8,2.09902043902753
"chrI",8288952,8291131,2180,"-","WBGene00002196",7.95301,212,1,247,9,2.17701954641717
```

Conclusion

- **Footprint/Profile** plots for summarisation of signal
- **Heatmaps** for (visually) detecting otherwise hidden stratification
- **SeqPlots:** easy way to get fast results, which reasonable amount of parameters