

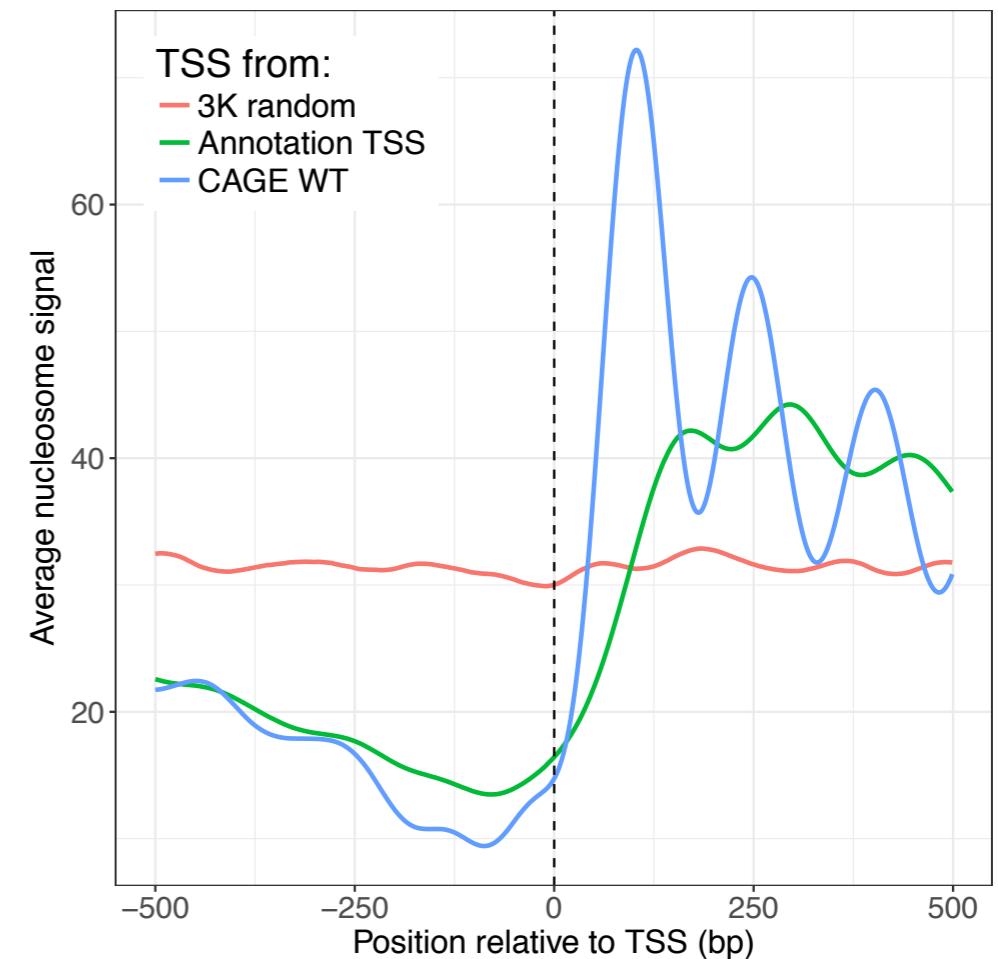
Coverage Footprints & Heatmaps

A quick tutorial using



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

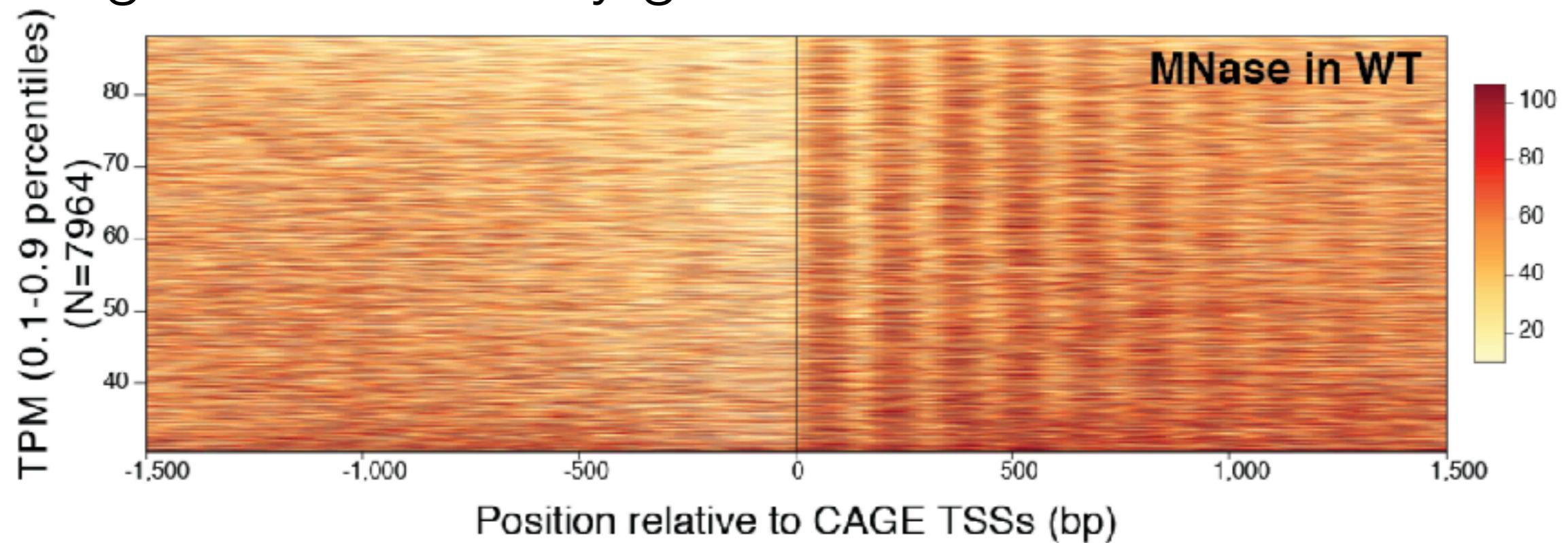
- easy to compare different signal
- summarisation
- readability

✗ 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

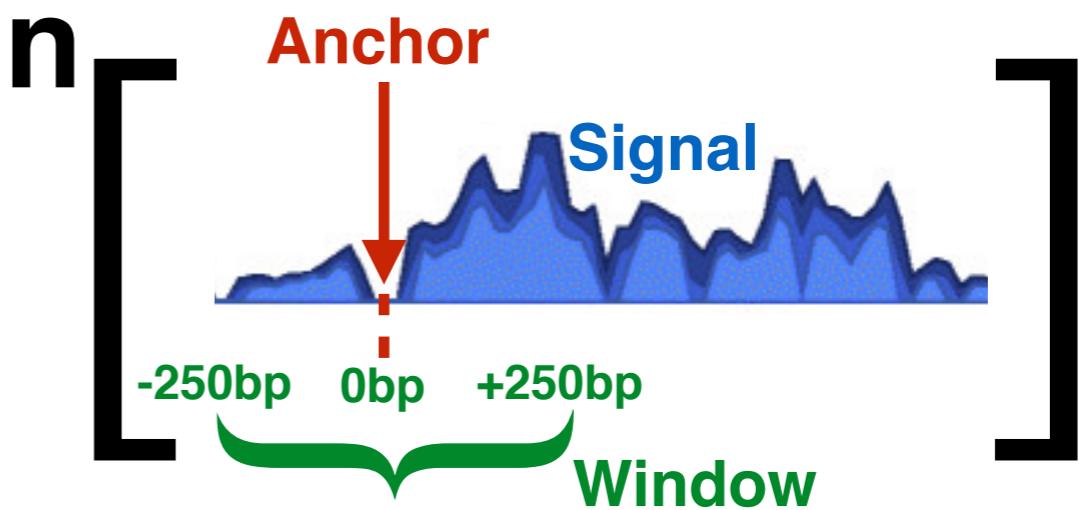
✗ Cons

- 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**
 - > *rows* = genomic feature,
 - > *columns* = 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

The screenshot shows the SeqPlots web application interface. At the top, there's a status bar with "Status: SeqPlots running". Below it is a navigation bar with "Data directory: /Users/axelthieffry/SeqPlots_data" and buttons for "Change data directory", "Quick reload", and "Exit". The main title "SeqPlots" is in large, colorful letters. Below the title is a toolbar with icons for signal tracks, feature files, and other options. A large green button labeled "Add files" is highlighted with a black arrow pointing to it from a circled number 1. To the right of the toolbar, a section titled "where data are stored + directory options, not useful today" is shown. Below the toolbar, there's a "Upload files:" section with instructions for adding signal tracks and feature files, and a "Create new plot array:" section with instructions for choosing signal tracks and feature files. A blue button labeled "New plot set" is highlighted with a black arrow pointing to it from a circled number 2. At the bottom left, there's a "Stop and exit the web interface:" section with an orange "Exit SeqPlots" button. A "Help" link at the bottom left points to documentation or a tutorial. A large black arrow points from a circled number 3 to the "Configure and draw plots" section, which is located near the top right of the interface.

1 *add your own files* (not done here because already loaded)

2 *configure and run an analysis*

3 *configure and draw plots*

Very well written doc, have a look!

1 - Adding files & quick reminder on common formats

Upload files

Info Use "Add files" button or drag and drop files here. Specify genome version & file type for each file. You can upload multiple files.

+ Add files... ⏪ Start upload ⚡ Cancel upload ★ Set defaults...

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,+	.	+

BED 6

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

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

bedGraph

wig

- a variation of bedGraph

SAM

- from a mapping, very detailed,
- very heavy

“binarisation”

makes it faster & lighter
cannot be viewed directly

bigWig
BAM

2 - Running an analysis: Select signal tracks

Status: SeqPlots running Data directory: /Users/axelthieffry/SeqPlots_data Change data directory Quick reload Exit

Info Choose file by clicking on file name. Chosen files will be highlighted. Click file name again to cancel choice. At least one signal track or motif and one feature file must be selected.

Tracks Features Sequence features | 0 track(s) selected | Select filtered Add visible on page Select none

Showing 1 to 3 of 3 entries Search:

Select the two signal tracks of histone marks by clicking on them (they turn grey)

File name	Date created	Format	Genome	User	Download	Delete
HTZ1_celegans_N2_L3_chrl.bw	2016-01-29 17:12:00	BigWiggle	ce10	demo		
H3K36me3_celegans_N2_L3_chrl.bw	2016-01-29 17:11:58	BigWiggle	ce10	demo		
H3K4me3_celegans_N2_L3_chrl.bw	2016-01-29 17:11:57	BigWiggle	ce10	demo		

All A A A A 10 records per page First Previous 1 Next Last

Bin track @ [bp]: 10 Choose the plot type Additional options: Plotting distances in [bp]: Upstream: Downstream:

Point Features Ignore strand 1000 1000

Midpoint Features Remove zeros

Endpoint Features Calculate Heatmap

Anchored Features

Close Refresh Remove selected files Run calculation

help on specific controls. To run tutorial click [here](#).

2 - Running an analysis: Select reference points/anchors

Status: SeqPlots running Data directory: /Users/axelthieffry/SeqPlots_data Change data directory Quick reload Exit

Info Choose file by clicking on file name. Chosen files will be highlighted. Click file name again to cancel choice. At least one signal track or motif and one feature file must be selected.

Tracks Features Sequence features | 1 feature(s) selected | Select filtered Add visible on page Select none

Showing 1 to 2 of 2 entries Search:

Select the top 20 most expressed genes in *C. elegans*

File name	Date created	Format	Genome	User	Download	Delete
Genes_celegans_bottom_20pct_expression_chr1.bed	2016-01-29 17:11:56	BED	ce10	demo		
Genes_celegans_top_20pct_expression_chr1.bed	2016-01-29 17:11:56	BED	ce10	demo		

All A A AI A First Previous 1 Next Last

10 records per page

Bin track @ [bp]: 10 Choose the plot type Additional options: Plotting distances in [bp]: Upstream: Downstream:

Statistic: mean Ignore strand Remove zeros Upstream: 1000 Downstream: 1000

median Point Features Calculate Heatmap Refresh Remove selected files Run calculation

Anchored Features

Help Read documentation or press ? button to get help on specific controls. To run tutorial click here.

2 - Running an analysis: Configuration panel (bottom)

Bin track @ [bp]: 10

Statistic:

- mean
- median

Choose the plot type

- Point Features
- Midpoint Features
- Endpoint Features
- Anchored Features

Additional options:

- Ignore strand
- Remove zeros
- Calculate Heatmap

Plotting distances in [bp]:

Upstream: 1000

Downstream: 1000

Signal binning option and related statistic

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

Point Midpoint Endpoint

start end

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

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.

And finally :

3 -Overview of plot configuration:

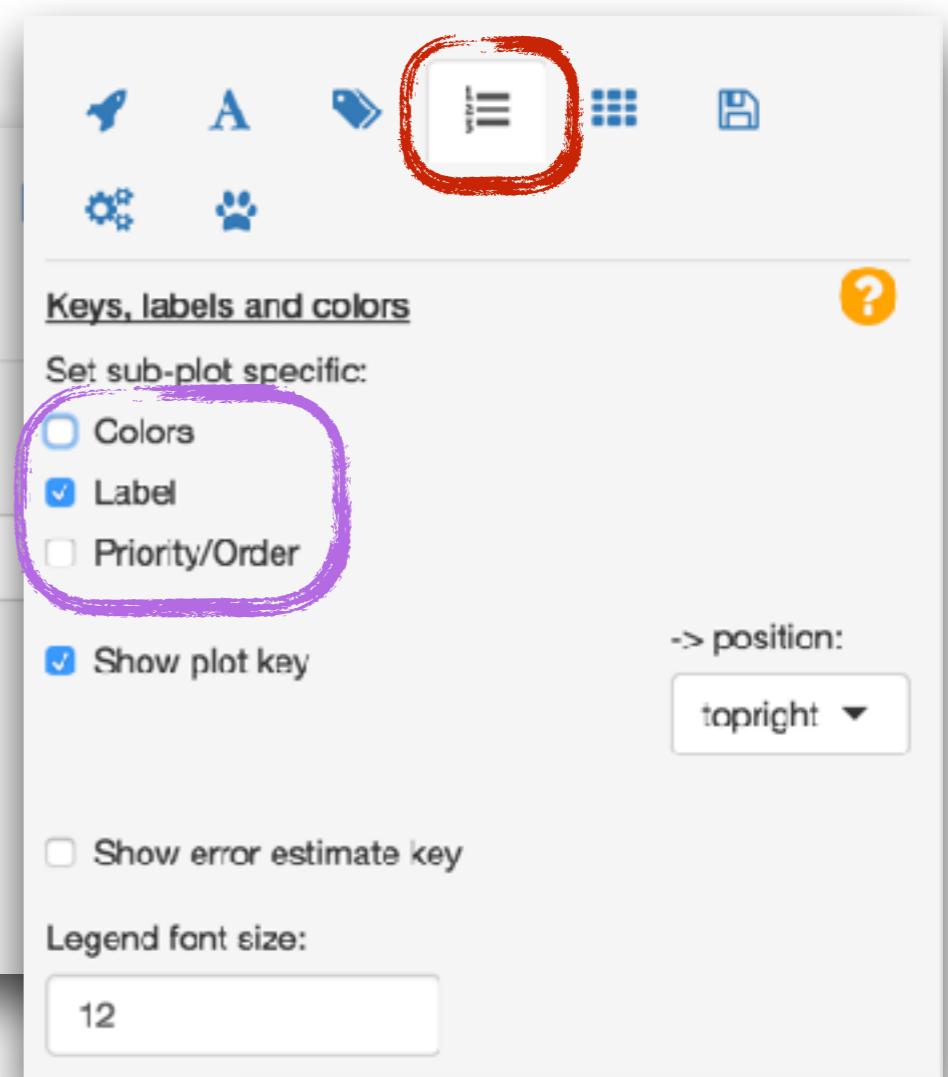
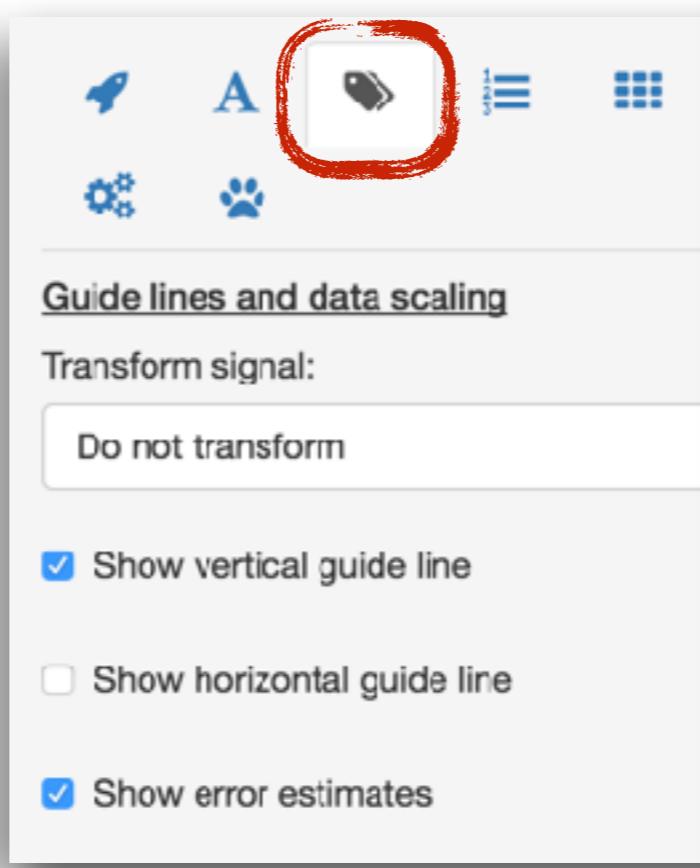
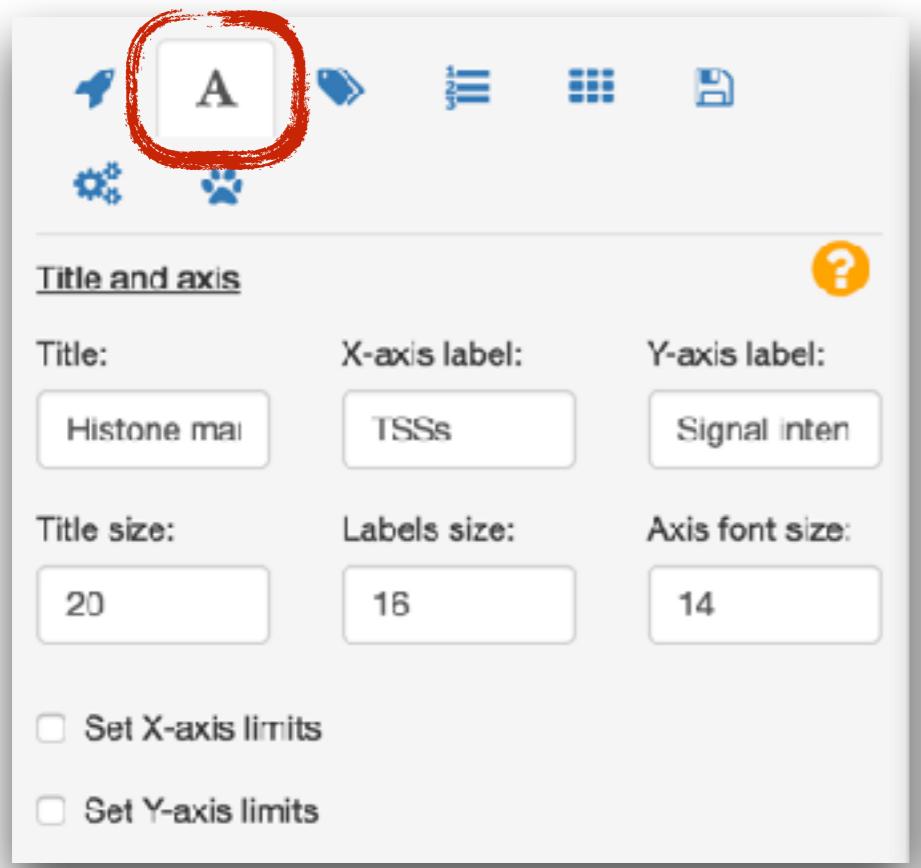
The screenshot shows the SeqPlots software interface. At the top, there's a status bar with "Status: SeqPlots running", a data directory path "/Users/axelthieffry/SeqPlots_data", and buttons for "Change data directory", "Quick reload", and "Exit". The main area has a "SeqPlots" logo and a message: "Select feature/track pair(s) and press "Profile" or "Heatmap" button to activate the preview". Below this are several buttons: "Profile" (green), "Heatmap" (blue, highlighted with a blue circle), "PDF" (white), and a question mark icon. To the right of these are icons for "Upload files", "Add signal tracks", "Create new plot array", and "New plot set". A large orange arrow points from the "Heatmap" button to a detailed view of the data selection table on the right. This table has columns for "Features" and "Signals". It shows a row for "Genes_cellegans_top_20pct_expression_chr1" and two other rows labeled "H3K3me3_cellegans_N2_L3_chp" and "H3K4me3_cellegans_N2_L3_chp". There's also a "Show/hide selection buttons" checkbox. A black arrow points from the bottom right of the table back towards the "Heatmap" button.

Data selection table (see next slide)
rows = reference points
columns = signals

Generate plots directly with default settings or refresh plot after changing settings
Heatmap-specific options
General plot settings (see next slide)

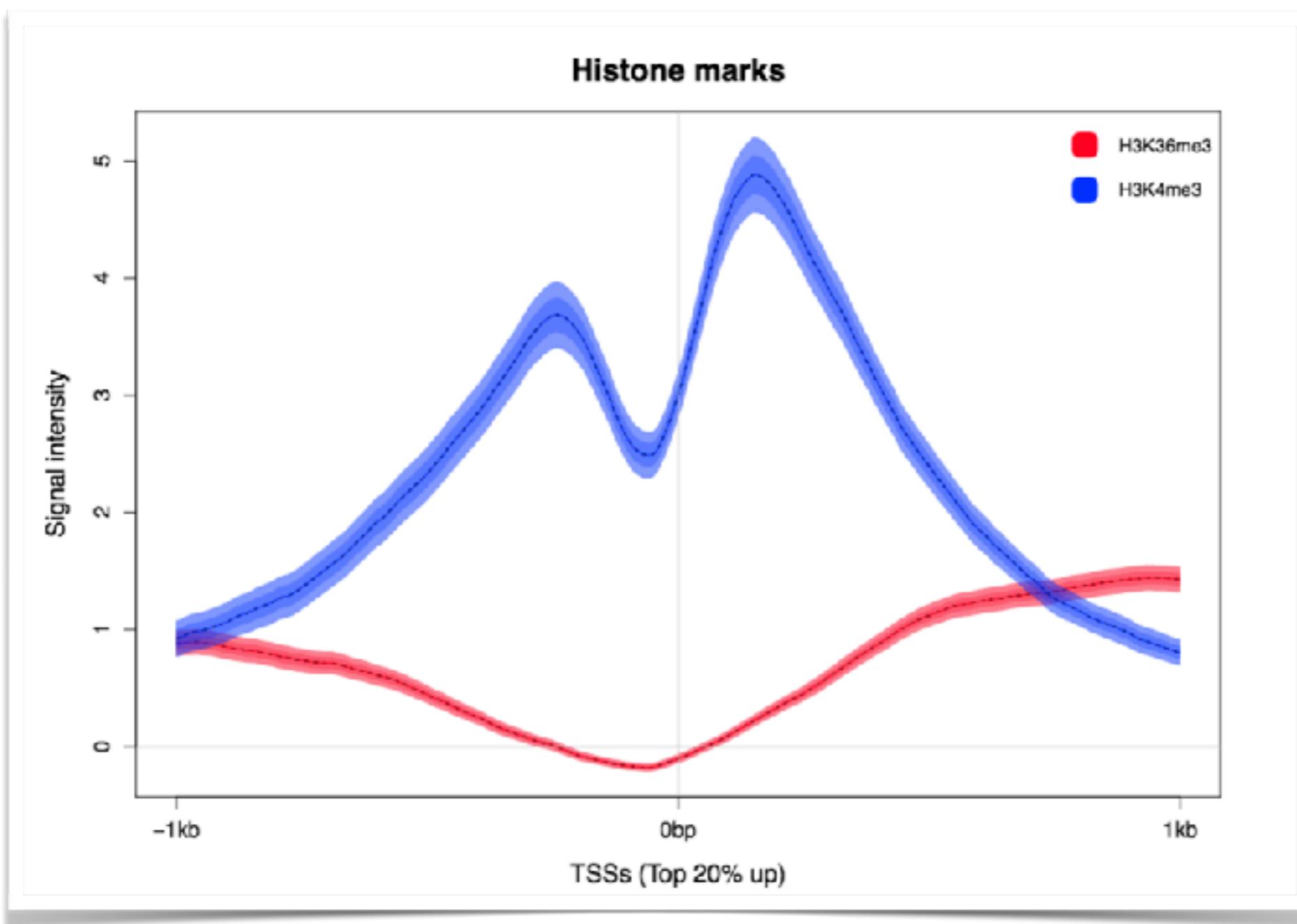
3 - General plot settings

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



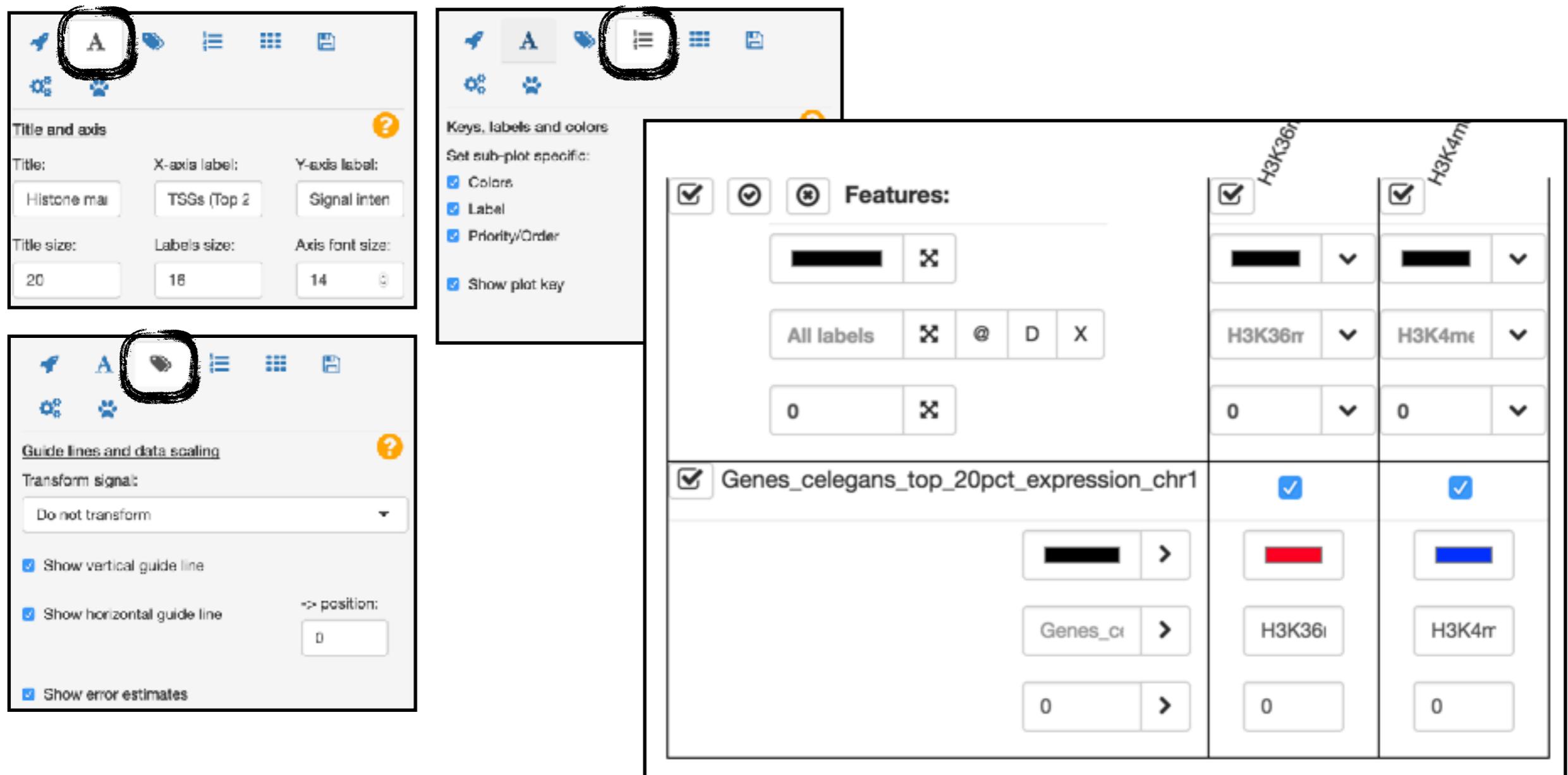
SeqPlots - Footprint

- Try to reproduce the following footprint (aka )



SeqPlots - Footprint answer

- Try to reproduce the following footprint (aka )

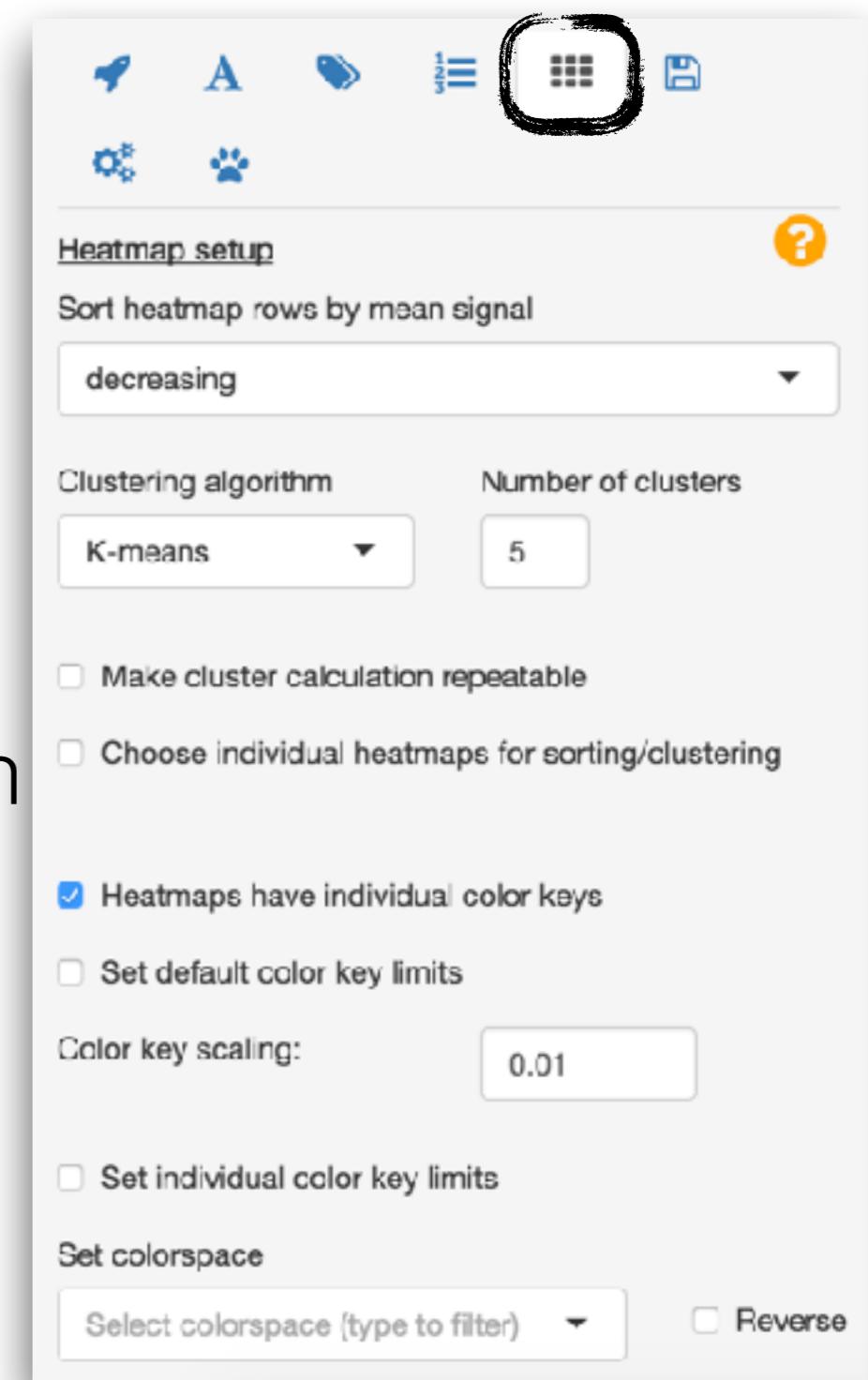


The screenshot shows the SeqPlot software interface with three main panels:

- Title and axis:** Set to "Histone mark" for Title, "TSSs (Top 2)" for X-axis label, and "Signal Inten" for Y-axis label. Title size is 20, Label's size is 18, and Axis font size is 14.
- Keys, labels and colors:** Set sub-plot specific:
 - Colors
 - Label
 - Priority/Order
 - Show plot key
- Features:** A table for adding genomic features. The first row is set up with a black bar and a red 'X' button. The second row has a '0' value and a red 'X' button. The third row contains the text "Genes_celegans_top_20pct_expression_chr1" with a checkmark and a green 'V' button. To the right of the table are dropdown menus for "H3K36r" and "H3K4m".

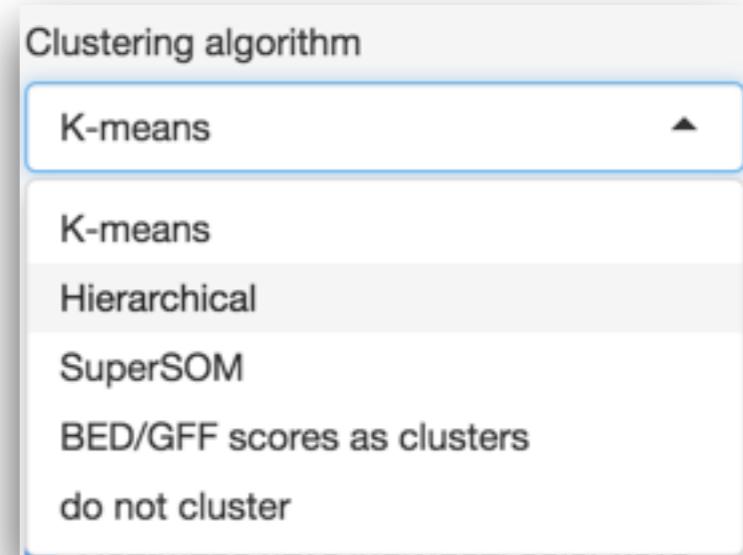
SeqPlots - Heatmap settings

- 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

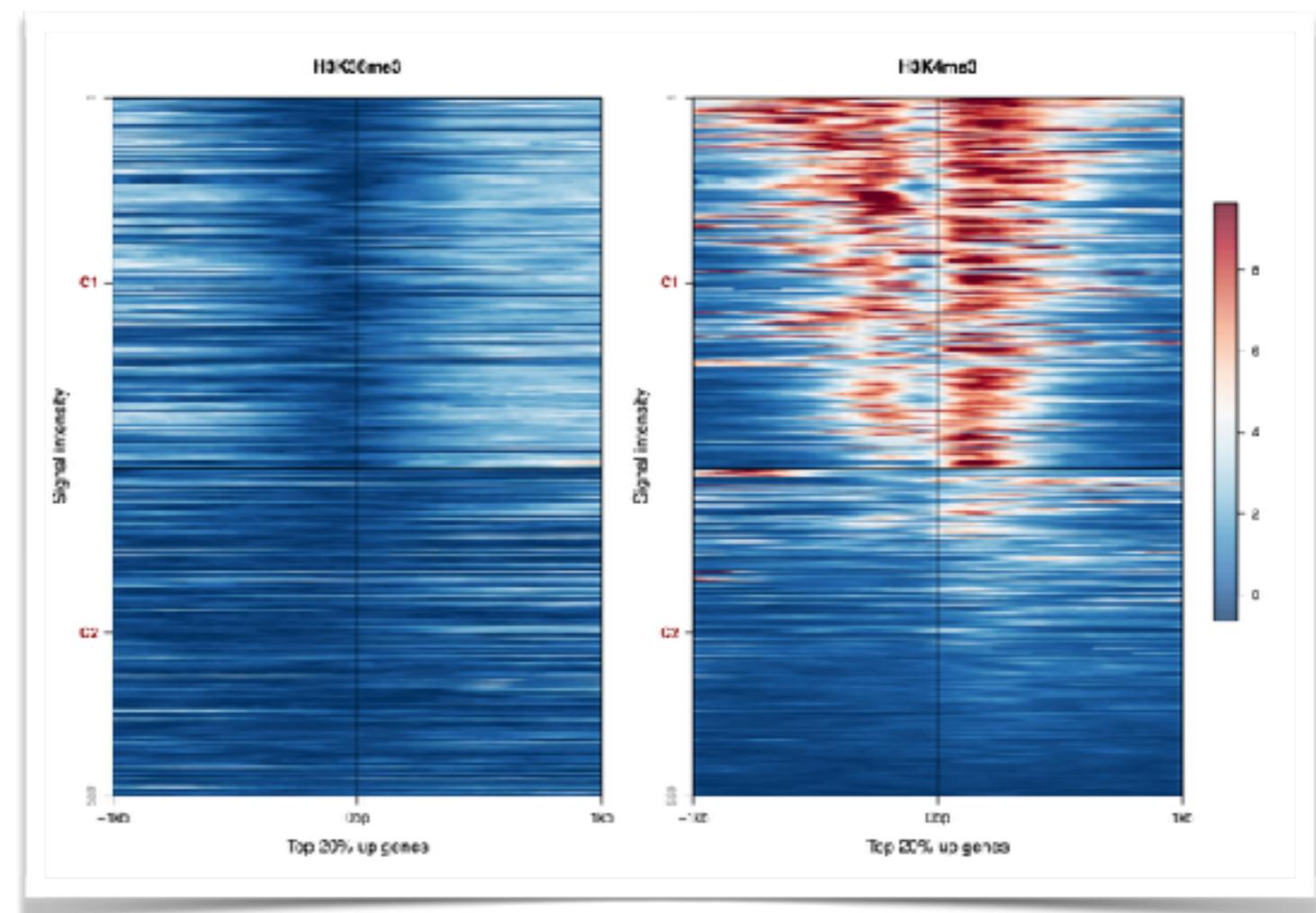
- Hierarchical clustering can take a significant amount of time on big matrices (binning becomes useful)
- BED scores (column #5) can be used for clustering
- **Colour key scaling:** how colour key limits are generated. For example, 0.01 (default value) calculates limits using data ranging from 1-99 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.



Color key scaling:

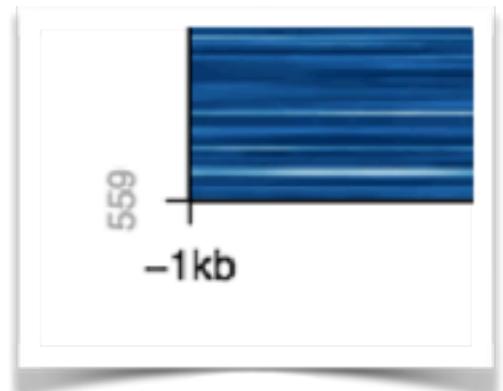
SeqPlots - Heatmap

- Try to reproduce the following heatmap ()
- What does the heatmap show that we cannot see in the profile plot?
- How many reference points are showed?



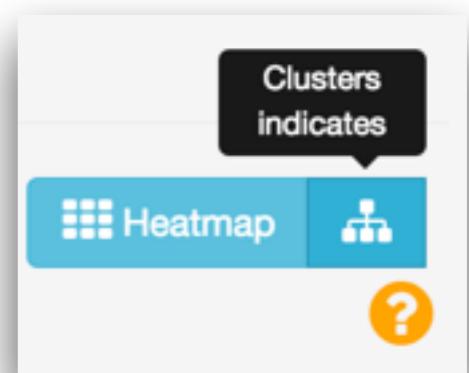
SeqPlots - Heatmap answer

- Plotting parameters (Variants of this are of course 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