SeqPlots - the user guide

Przemyslaw Stempor

July 16, 2014

Contents

Welcome to SeqPlots

An interactive tool for visualizing track signals and sequence motif densities along genomic features using average plots and heatmaps.

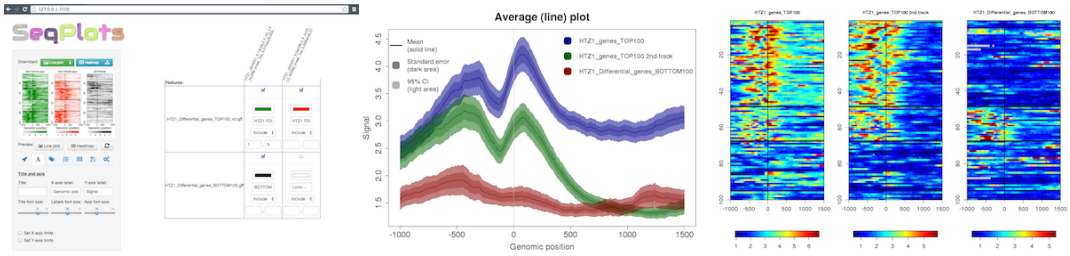

Figure 1: Examples of Seq Plots interfce and outputs

Summary

SeqPlots is a web browser tool for plotting average track signals (e.g. read coverage) and sequence motif densities over user specified genomic features. The data can be visualized in linear plots with error estimates or as series of heatmaps that can be sorted and clustered. The software can be run locally on a desktop or deployed on a server and allows easy data sharing. SeqPlots pre-calculates and stores binary result matrices, allowing rapid plot generation. Plots can also be run in batch.

Availability

Standalone versions of SeqPlots are available as a Mac OS X (10.6 or higher) app bundle combing R, all required packages and scripts([subproject home]) or as an R package ([subproject home]). SeqPlots can also be deployed on a server using free and open sourced (GPL licensed) Shiny Server (https://github.com/rstudio/shiny-server), making it available for multiple clients. For the server version, clone this repo to Shiny Server application directory and set up data location: `git clone https://przemol@bitbucket.org/przemol/seqplots.git`.

Key features

- Easy to use web interface (R or shell expertise not required)
- Web server or desktop versions
- Generates publication quality plots out of the box
- Plots average signals or heatmaps
- Accepts Wiggle, BedGraph, BigWiggle, and GFF and BED formats
- Calculates motif density from reference genome packages
- Tracks and features are searchable and old calculations stored
- Converts tracks to binary BigWiggle format for rapid data extraction and efficient storage
- Implemented using Shiny R framework providing internet browser reactive GUI and session based connectivity (websocets)

Issues and bugs

Please visit **issues tracker** to view currently know issue. To report new issue/bug/feature request please click **here**. If issue is connected to file upload please attach the file in the form.

Source code

If you require code repository access, please request by email. The code will become public upon release. This repository contains only Shiny framework server code [R] and client browser code/libraries [HTML5, CSS3 and JavaScript]. There are two derivative projects for desktop end users:

- SeqPlots OSX App
- SeqPlots Bioconductor R package

References

Quick start guide

1. Start the SeqPlots. Refer to installation guides for platform specific information. After successful initiation the web interface should automatically open in your default web browser. If you are web server user just navigate your browser to server address.

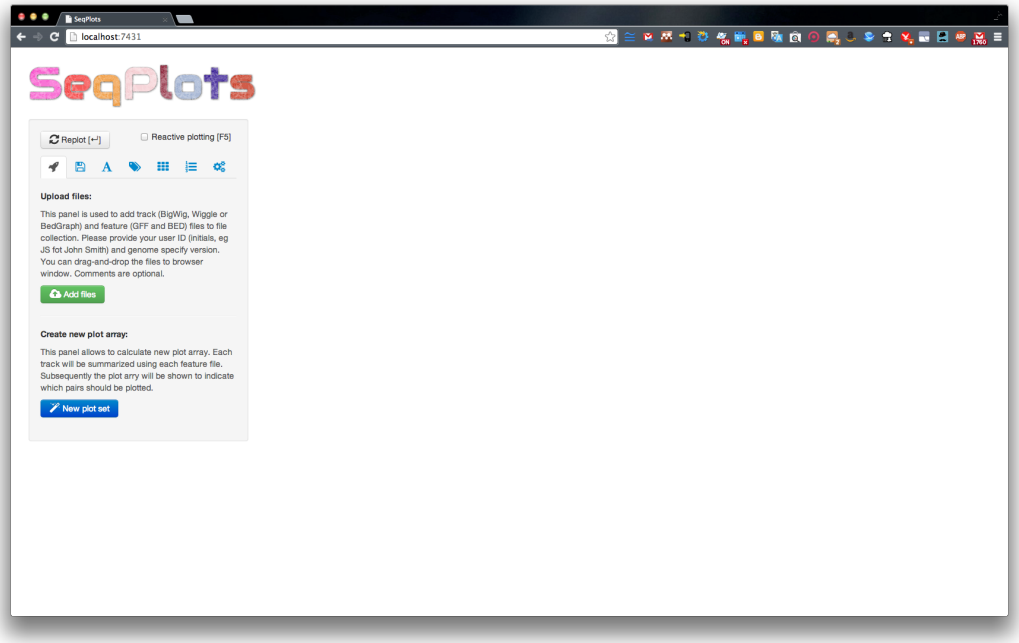

Figure 2: The SeqPlots interface in web browser

2. Upload some feature (BED or GFF) and track (BigWig or WIG) files. They can be gzip compressed (e.g. file1.bed.gz). Press green "Add files..." button or just drag and drop files into the window. The ready to upload files will show up in upload window, where you select user name, reference genome and optionally add some comments.

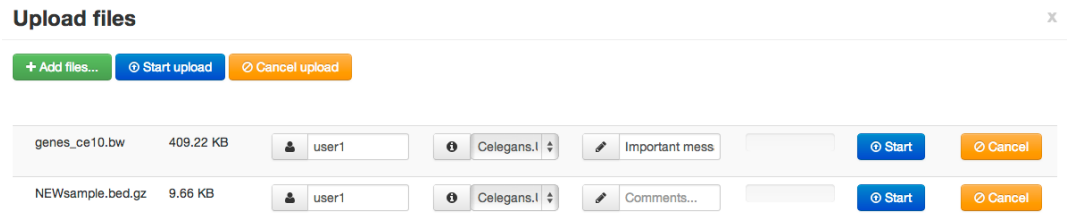

Figure 3: File upload panel

3. When all is done press blue "Start upload" button. After upload and processing is done the green "SUCCESS" label should show. It means that file is on the registered and ready to use. Occasionally the file might be mot formatted properly or chromosome names might not agree with reference genome. In such case a verbose error will window appears and file as labeled as "ERROR". For further information please refer to **errors chapter**.

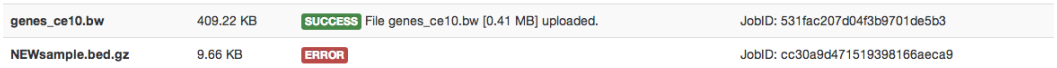

Figure 4: File upload progress infoermation

4. Dismiss upload window and press blue "New plot set" button on side panel. This will bring up file management window. In file management window select ant least one file from "Features" tab and either one or more file from "Tracks" file or sequence motif(s). The sequence motifs and tracks can be processed and plotted together.

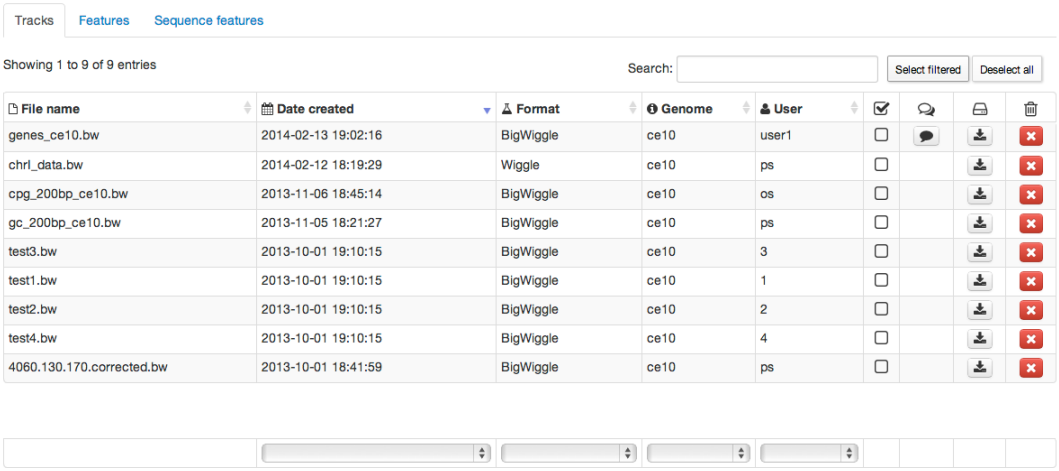

Figure 5: File management panel

5. When you decided which files/motifs to plot it is time to set up the processing options. You can find these in the button of plotting window. For first plot you should do just fine with default options, to learn more check **this** section.

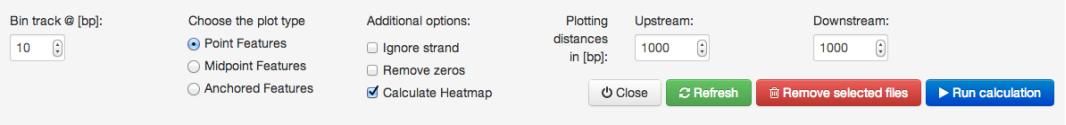

Figure 6: Plot set calculation options

6. After options are set up press blue "Run calculation" button. This will dismiss the file management modal and show processing window. Here you can observe the progress of the task and optionally cancel it if no longer required (or you forgot to add some very important file to the plot-set).

Figure 7: Plot set calculation progress window

7. After some time the calculation will finish (fingers crossed, without the error) and you will be able to see plot set array. In here you can choose which feature-track or feature-motif pairs to plot. Choose one or more checkboxes and press grey "Replot" button (or RETURN from your keyboard). You can also check "Reactive plotting" checkbook just next to it - it will automatically apply changes to plots as soon you make them.

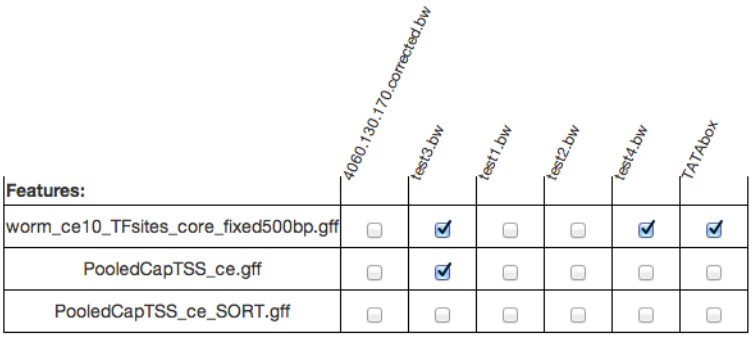

Figure 8: The plot selection grid

8. Congratulation! Your First plot is complete, you can see the preview of it on the side panel.

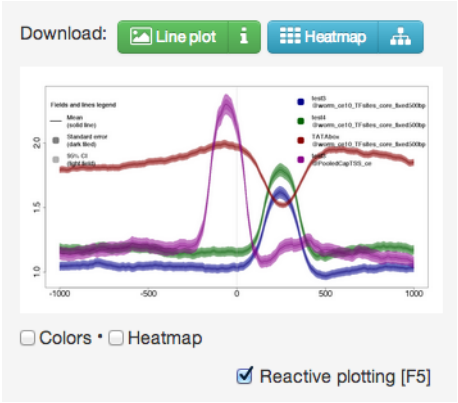

Figure 9: The plot preview panel

9. You are able to set up labels, titles, font sizes, legends and may more on side panel tabs, take a look at **this** chapter.

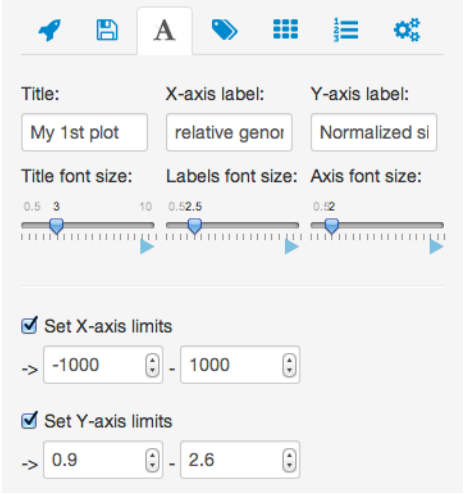

Figure 10: Plot settings tabs

10. By clicking the plot preview you can enlarge it for better view. When everything is ready you can get the plot as PDF bt clicking green "Line plot" button just on the top of side panel.

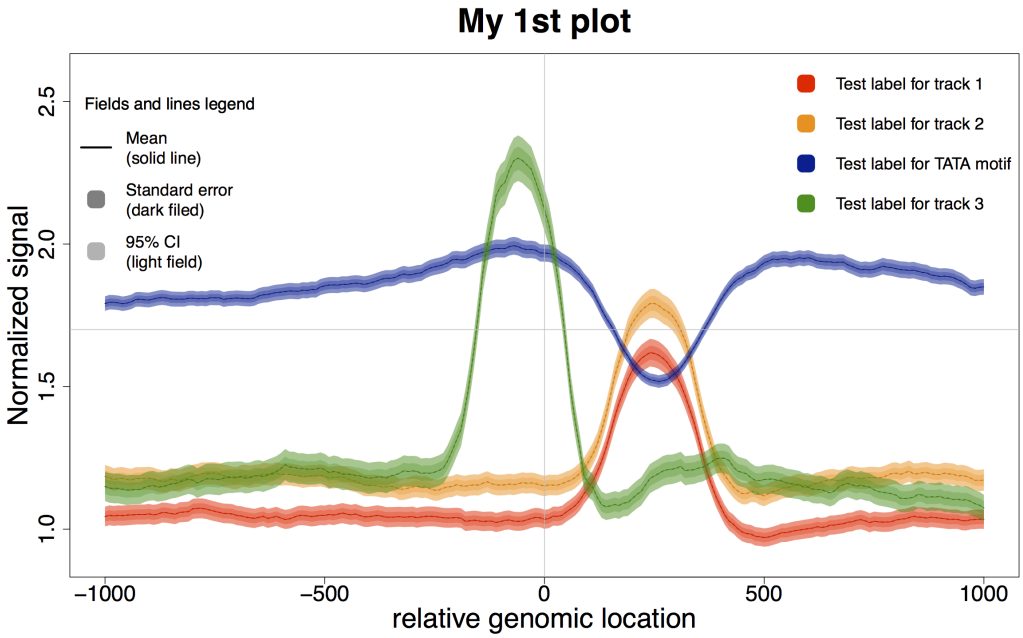

Figure 11: Average plot example

11. You can also visualize the signal as a heatmap. Please note that heatmap plotting is possible only for single feature file or files containing exactly the same number of genomic ranges (which will become the rows of heat map). For heatmap you can choose to sort and/or cluster it using k-means. To learn more about heatmaps click (**here**).

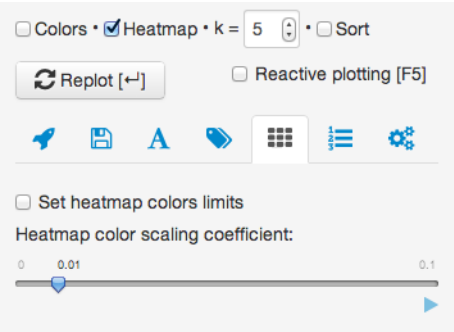

Figure 12: Heatmap settings tab

12. Similarly to line plot you can grab the heatmap PDF using 'Heatmap' button just on the top of side panel. Just attached to it, there is small button allowing you to get cluster definitions.

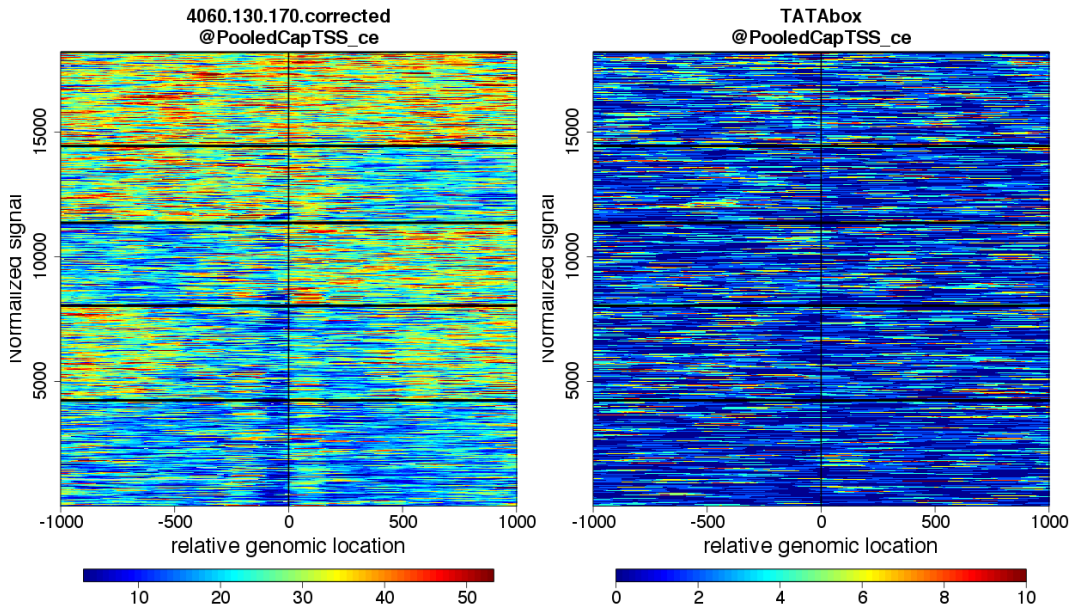

Figure 13: Heatmap example

Installation: R package

R package is the best way to use SeqPlots on desktop computers when operating system is Windows or Linux. It is also the recommended method for R users Mac OS X.

System requirements:

- R 3.1 or higher

Installation

To install SeqPlots package, start R and enter:

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

To install SeqPlots development package, start R and enter:

```
install.packages('devtools')
devtools::install_bitbucket(repo='SeqPlots_Rpackage', username='przemol',
auth_user=readline('User?> '), password=readline('Password?> '), dependencies="Depends")
```

Usage

From R prompt run:

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

After a successful initiation the user interface will be opened in your default web browser. For further usage please refer to quick start guide or specific chapters of documentation.

The `seqplots` function accepts "root" argument, which allows to change the data location folder (by default your home directory will be used), e.g.:

```
seqplots(root='/path/to/data/location')
```

Additional genome packages

Genomic packages can be installed using standard bioconductor installer (Internet connection required). For example, to instal human reference genome (hg19):

```
source("http://bioconductor.org/biocLite.R")
biocLite("BSgenome.Hsapiens.UCSC.hg19")
```

Corresponding genome packages are required before uploading the files for plotting. Full list of supported genomes is avilable here: http://www.bioconductor.org/packages/release/BiocViews.html#__BSgenome

Installation: Deployment on the server

> The server version is designed to be used by advanced users, who would like to provide shared SeqPlots service to a group of people.

System requirements:

- R 3.1 or higher
- Shiny Server 1.0 or higher

Installation

1. Install and configure the Shiny Server by following the instructions on https://github.com/rstudio/shiny-server
2. Install SeqPlots R package and dependences by following the instruction here
3. Copy SeqPlots files to Shiny Server application folder:

```
cp -r $(Rscript -e "cat(system.file('seqplots', package='seqplots'))") /srv/shiny-server/
```

4. Set up SeqPlots **data location** by running from R:

```
seqplots(root='/path/to/data/location')
```

5. Edit first line of `/srv/shiny-server/shiny/server_config.R`, so the environment variable `root` matches the **data location**; for example:

```
Sys.setenv('root'='/var/shiny-server/DATA')
```

Usage

After successful installation the SeqPlost web GUI will be available at `your_server_name:3838/seqplots/`.

For further usage please refer to quick start guide or specific chapters of documentation.

Additional genome packages

Genomic packages can be installed using standard bioconductor installer (Internet connection required). For example, to instal human reference genome (hg19):

```
source("http://bioconductor.org/biocLite.R")
biocLite("BSgenome.Hsapiens.UCSC.hg19")
```

Corresponding genome packages are required before uploading the files for plotting. Full list of supported genomes is avilable here: http://www.bioconductor.org/packages/release/BiocViews.html#__BSgenome

Alternative installation

SeqPlots can be directly cloned into Shiny Server application folder from git repository by using `git clone https://przemol@bitbucket.org/przemol/seqplots.git`. The data location can be set up with following R code:

```
root <- /data/location
dir.create(root)
setwd(root)
require(RSQLite)
sqlite <- dbDriver("SQLite")
con <- dbConnect(sqlite, dbname = "files.sqlite")
dbGetQuery(con, "CREATE TABLE files (id INTEGER PRIMARY KEY ASC, name TEXT UNIQUE,
  ctime TEXT, type TEXT, format TEXT, genome TEXT, user TEXT, comment TEXT)")
if (!dbListTables(con) == "files")
    warning("Database not created!")
dbDisconnect(con)
if (!all(sapply(c("removedFiles", "files", "publicFiles", "tmp"), dir.create)))
    warning("Folders not created!")
message("\nData loaction: ", root)
```

The following dependencies must be installed in R:

```
              package  version
1               shiny   0.10.0
2              fields      7.1
3           parallel    3.1.0
4          multicore      0.2
5           BSgenome   1.32.0
6      GenomicRanges   1.16.3
7             plotrix    3.5-7
8        rtracklayer   1.24.2
9             RJSONIO  1.2-0.2
10            RSQLite   0.11.4
11            kohonen   2.0.14
12              Cairo    1.5-5
13             digest    0.6.4
14            methods    3.1.0
15              tools    3.1.0
16              utils    3.1.0
17              httpuv    1.3.0
18            caTools     1.17
19             xtable    1.7-3
20         htmltools    0.2.4
21            bitops    1.0-6
22              Rcpp   0.11.2
```

```
23              spam   0.41-0
24              maps    2.3-7
25              grid    3.1.0
26         grDevices    3.1.0
27      BiocGenerics   0.10.0
28           IRanges   1.22.9
29        Biostrings   2.32.0
30           XVector    0.4.0
31         Rsamtools   1.16.1
32          graphics    3.1.0
33             stats    3.1.0
34          zlibbioc   1.10.0
35      GenomeInfoDb    1.0.2
36            stats4    3.1.0
37               XML 3.98-1.1
38             RCurl 1.95-4.1
39 GenomicAlignments    1.0.1
40      BiocParallel    0.6.1
41           foreach    1.4.2
42         BatchJobs      1.2
43            BBmisc      1.7
44               DBI    0.2-7
45         sendmailR    1.1-2
46              brew    1.0-6
47              plyr    1.8.1
48           stringr    0.6.2
49              fail      1.2
50         checkmate      1.0
51         codetools    0.2-8
52         iterators    1.0.7
53         base64enc    0.1-1
54             class   7.3-10
55              MASS   7.3-33
```

Adding and managing files

Supported file formats

Tracks:

- BigWig (.bw) - http://genome.ucsc.edu/FAQ/FAQformat.html#format6.1
- Wiggle (.wig) - http://genome.ucsc.edu/goldenPath/help/wiggle.html
- BedGraph (.bdg) - http://genome.ucsc.edu/goldenPath/help/bedgraph.html

Features:

- BED - http://genome.ucsc.edu/FAQ/FAQformat.html#format1
- GFF - http://genome.ucsc.edu/FAQ/FAQformat.html#format3
- GTF (with .gff extension) - http://genome.ucsc.edu/FAQ/FAQformat.html#format4

Files must be formatted according to UCSC guidelines. All widely used chromosome names conventions are accepted, e.g. for human files either 'chr1' or '1' can be used, however these conventions should not be mixed within single files.

Adding files

Pressing the `Add files` button brings up the **file upload panel**.

Figure 14: File upload panel

You can drag and drop files here or press the `Add files...` button to opens a file selection menu. Before starting the upload the following mandatory information must be provided about each file:

- User ID
- Reference genome - drop-down menu containing reference genome package currently installed in R

Comments are optional.

The contents of the a text field can be copied to all files by clicking the icon at the left of the field. The default values can be set using `Set defaults...` button. Default values are stored using the browser cookies, and the settings will be remembered across different sessions as long as the same web browser is used. File extensions that are not supported will raise an error.

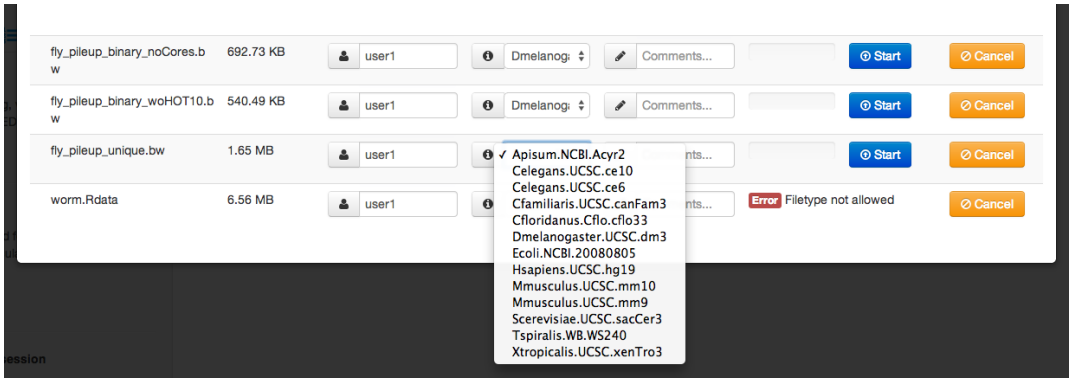

Figure 15: File upload panel with 4 files selected

Individual files can be uploaded by pressing 'start' next to the file name or all files can be uploaded at once by pressing the `Start upload` button at the top of **file upload panel**.

During the upload process a progress bar is displayed. After upload SeqPlots gives a message that upload was successful or or gives an error message. Common errors are misformatted file formats or chromosome names do not matched the reference genome. For more information please refer to errors documantation

Figure 16: A feedback on successfully upload files

To dismiss the upload window, click on `X` or outside the window.

Downloading and removing files

Clicking the `New plot set` button brings up the **file collection window**. The primary function of this window is to choose signal tracks and feature files to use for calculating the plots. However, it also provides basic file management capabilities. Information on files can be reviewed and files can be downloaded or deleted. Fields can be searched, filtered and sorted by any column. The red `x` button on the right site of file table removes a single file from the collection, while `Remove selected files` button will erase all selected files.

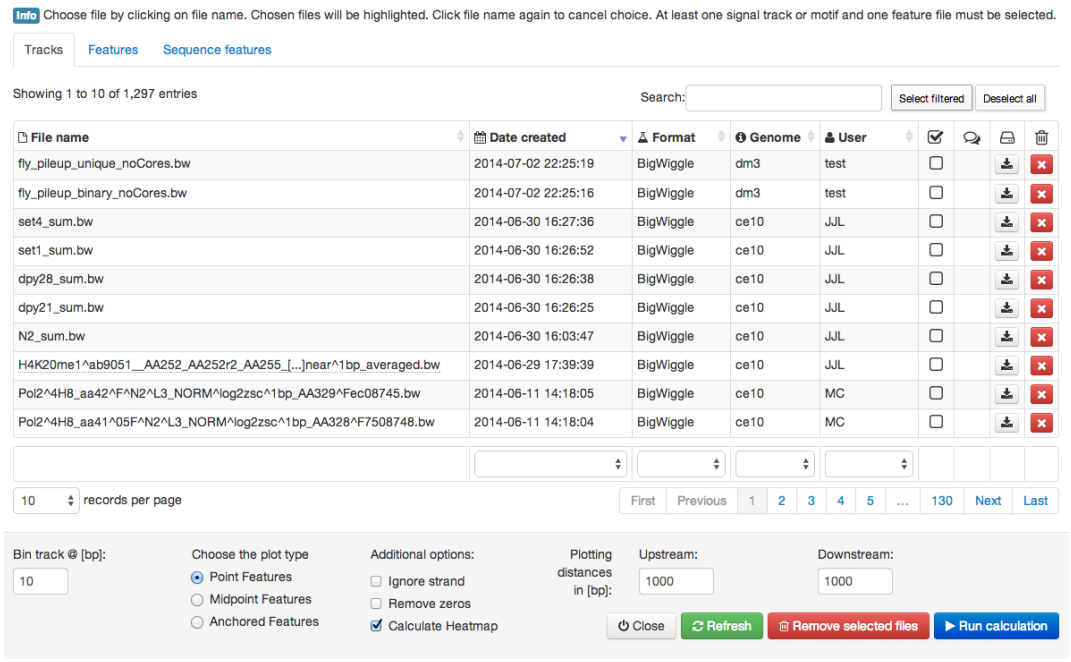

Figure 17: The file collection window

Running the plot-set jobs

The **file collection modal** allows choosing signal tracks and feature files from the collection to calculate average plots and heat maps. Press `New plot set` button to bring it up. If you wish tu upload more files please refer to adding new files documentation.This window have three tabs:

- `Tracks` gather signal files, that is Wiggle, BigWiggle and BedGraph
- `Features` gather genomic feature files, that is BED, GFF and GTF
- `Sequence features` allows to set up the sequence motif density track

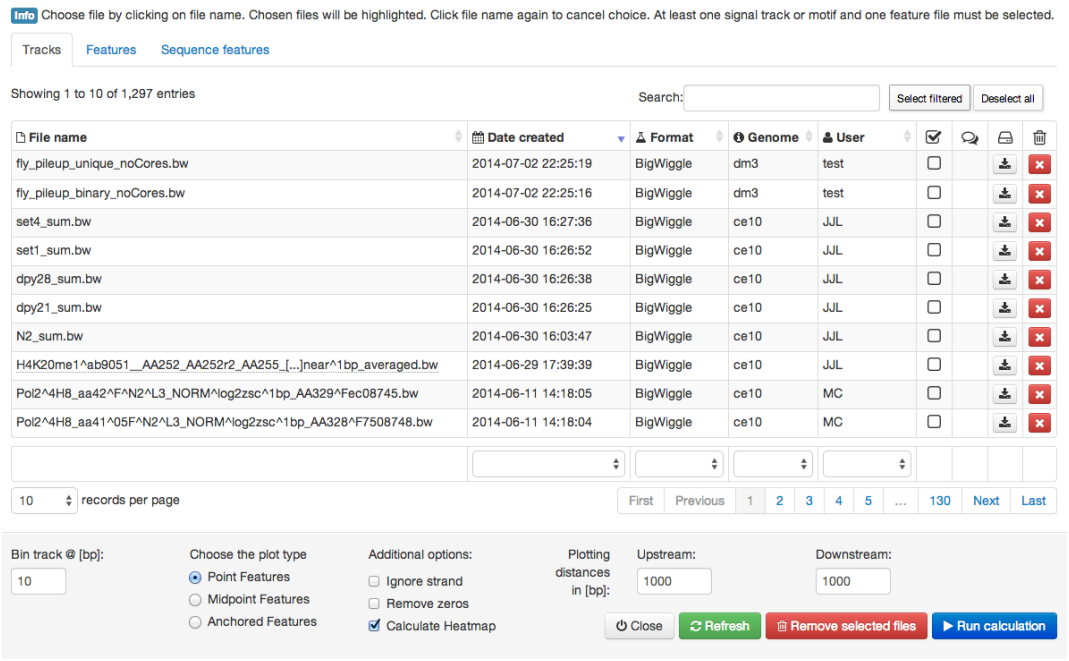

Figure 18: The file collection modal

Selecting files

Both `Tracks` and `Features` tabs allow to review all the information about files, filter them and sort by any column. The "Search:" dialog allows to quickly filter the files by any field, while dropdowns below the file grid allow for more advanced filtering on specific columns.

Files are selected by clicking on file name, or any other part of the row beside `Show comment` and `Download` or `Remove` buttons. Chosen files are highlighted in light blue. Clicking the file name again will cancel the selection. At least one signal track or motif and one feature file must be selected before starting the calculation.

Setting up plot options

The set of options controlling the plot settings is available below the file grid/motif option:

1. `Bin track @ [bp]:` - this numeric input determines the resolution of data acquisition; the default value 10 means that 10bp intervals within the plotting range will be summarized by calculating the mean. Higher values increases the speed of calculation and produces smoother plots. See the explanations.
2. `Choose the plot type` - this radio box determines the mode of plots
   - *`Point Features`* - plot a range around feature start or end depending on it's directionality, see explanations
   - *`Midpoint Features`* - calculates the middle point of the feature and plot a range around it
   - *`Anchored Features`* - scale the features to given pseudo-length and plots the range upstream of the beginning and downstream of the end
3. `Ignore strand` - the directionality (strand) will be ignored, that its +, − and ∗ ranges will be centered on start and plotted in the same direction
4. `Ignore zeros` - the signal values equal to 0 in the track will be ignored, that is will be excluded from mean and errors calculation

5. `Calculate heatmap` - this checkbox determines if heat map matrix should be saved; uncheck it will speed up calculation calculation, but only average plots will be feasible in this plot set.
6. `Plotting distances in [bp]` - the distances in to be plotted:
   - *`Upsteram`* - the plotting distance in base pairs upstream to the feature
   - *`Anchored`* - the pseudo-length, to which the features will be extended or shrunk using linear approximation (only for anchored plots)
   - *`Downstream`* - the plotting distance in base pairs downstream to the feature

Plotting sequence motif density

`Sequence features` tab allows to calculate and plot the motif density around genomic features using the reference sequence package. Motif plots can be mixed with track files' signal plots. The following options can be set here:

1. `DNA motif` - the DNA motif
2. `Sliding window size in base pairs [bp]` - the size of the sliding window for motif calculation. The value (number of matching motifs within the window) is reported in the middle of the window, e.g. if window is set to 200bp, DNA motif is "GC" and there are 8 CpGs in first 200 bp of the chromosome the value 8 will be reported at 100th bp.
3. `Display name` - The name of the motif that will be shown in key and heatmap labels. Leave blank to use `DNA motif` value.
4. `Plot heatmap or error estimates` - this checkbox determines if heatmap matrix and error estimates should be calculated. If unchecked much faster algorithm will be used for motif density calculation, but only the average plot without the error estimates will be available.
5. `Match reverse complement as well` - determined if reverse complement motif should be reported as well. For example the TATA motif will report both TATA and ATAT with this option selected.

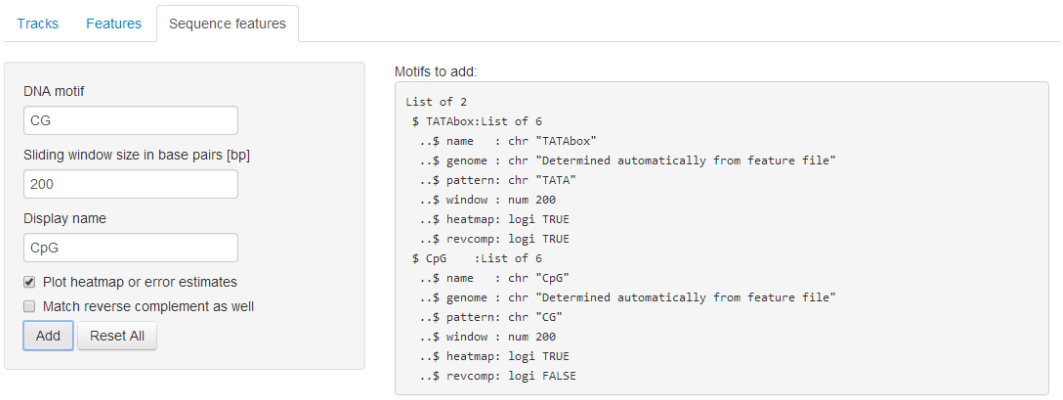

Figure 19: Sequence motifs selection tab

Clicking `Add` button adds the motif to plot set, while `Reset All` clears the motif selection. On the right side from motif setting panel is the list summary of included motifs.

Starting the plot set calculation

The option are executed by pressing `Run calculation` button. This dismisses the **file collection modal** and brings up the calculation dialog, which shows the progress. On Linux and Mac OS X (systems supporting fork based parallelization) the calculation can be stopped using the `Cancel` button - this will bring back all settings in **file collection modal**.

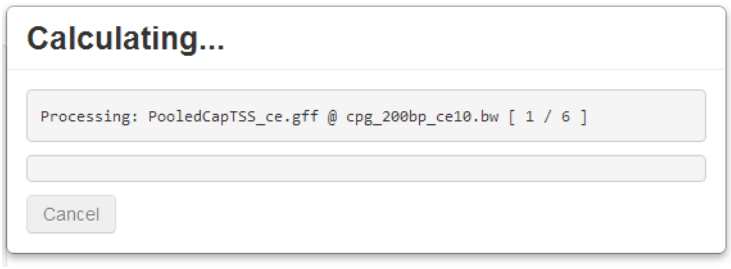

Figure 20: The calculation progress dialog

After the successful execution the **plot array** will appear. In case of error the informative error pop-up will explain the problem. Please reffer to error section for further information.

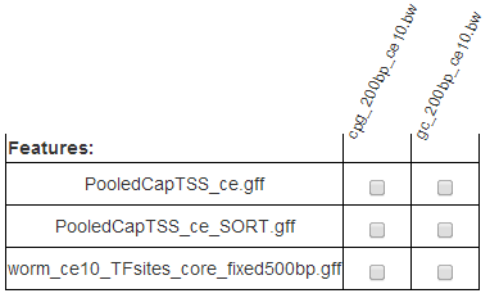

Figure 21: The plot array

Plotting

This section focuses on average (line) plots and options common between these and heatmaps. For heatmap options please refer to heatmps documentation.

Previewing plot

After calculating or loading plot set select the pairs of features and tracks/motifs using **plot array** checkboxes. Clicking on the column name (tracks/motifs) toggles the whole column selection, Similarly clicking on row name (features) toggles the whole row selection. Clicking on top-left most cell of **plot array** toggles the selection of whole array.

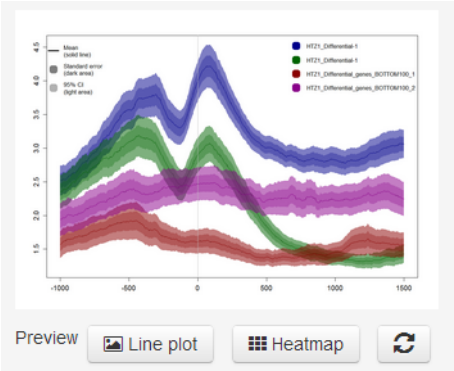

Figure 22: Plot preview plus `Line plot`, `Heatmap` and `refresh` buttons

If at least one pair on **plot array** is selected pressing `Line plot` button produces average plot preview and `Heatmap` button the heatmap preview. Finally, pressing `refresh` button or [RETURN] key from keyboard applies the new selection and options. These operations are done automatically in reactive mode.

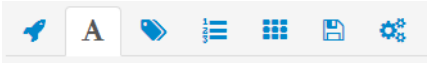

Figure 23: The tab selection area - icons represnts seven panels

Below the plotting buttons are seven panels. On application start the first panel responsible for bringing file upload, management and plot set calculation modals is active. The further three panels hold common plot settings.

Titles and axis panel

This panel groups settings influencing the plot main title, axis labels, various font sizes plus vertical and horizontal plot limits.

- `Title` - The main title of the plot, shown in top-center part of the figure; default empty
- `X-axis label` - Label shown below horizontal axis; default empty
- `Y-axis label` - Label shown below vertical axis; default empty
- `Title font size` - Font size of the title in points (point = ˜1/72 an inch for standard A4 output); default 20 points
- `Labels font size` - Font size of axis labels in points; default 16 points
- `Axis font size` - Controls axis ticks font size, that is size of the numbers indicating position in base pairs on X-axis and means signal value on X-axis; default 14 points
- `Set X-axis limits` - Set hard plotting limits for X-axis; default values are whole range chosen during plot set calculation
- `Set Y-axis limits` - Set hard plotting limits for Y-axis; default values are a range between lowest and highest mean signal extended by error estimate

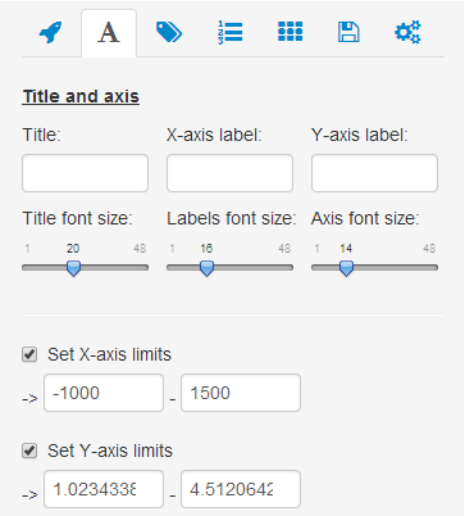

Figure 24: The view on titles and axis panel

Guide lines and data scaling

Controls in this panel controls the display of guide lines and error estimates, and allows to log scale the signal prior to plotting.

- `Transform signal` - if set to *`Log2 transform`* performs log2 transformation of the signal prior to plotting; default setting is *`Do not transform`*
- `Show vertical guide line` - show the vertical line at point 0 - beginning of the feature or midpoint and end of the pseudo-length scaled features (only for anchored plots); turn on by default
- `Show horizontal guide line` - show the horizontal line at user determined height; turn off by default
- `Show error estimates` - show error standard error and 95% confidence interval as fields, if turned off only the line representing the mean signal is shown; turn on by default

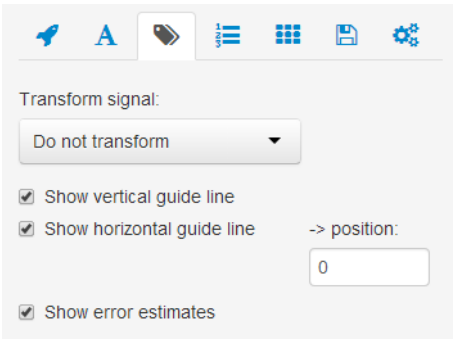

Figure 25: The view on guide lines and data scaling

Keys, labels and colors panel

This panel groups two types of controls. `Colors`, `Label` and `Priority/Order` are a checkboxes revealing further controls on **plot set grid**, specific for a feature-track pair or sub-heatmap. `Show plot key`, `Show error estimate key` and `Legend font size` re global controls specific for average plots. Inputs on **plot set grid** do not have specific labels, but the tooltip explaining their meaning is shown on mouse cursor hover.

- `Colors` - checkboxes revealing a color picker on **plot set grid**. This input allows to control the colors of specific feature-track pair average plots or sub-heatmaps. In browser supporting the color picker 'e.g Chrome' the system dialog will show up. In other browsers (e.g. Firefox) the javaScript color picker will be initialized.
- `Label` - checkboxes revealing a label text input **plot set grid**. This controls the names shown on the **key** with average plots or the heatmap top labels.
- `Priority/Order` - checkboxes revealing numeric input on **plot set grid**. These number determine the order of average plots and hetamaps. Feature-track pair with the highest priority will be listed on the top of **key** for average plots and left-most for heatmaps.
- `Show plot key` - shows the key giving the color to feature-track pair label mapping. If turned on the additional drop-down allows to choose the position on the plot, top-right by default
- `Show error estimate key` - shows the key gexplaining the meaning of error fields. If turnedon the additional drop-down allows to choose the position on the plot, top-left by default
- `Legend font size` - set the size of font used to plot the keys; 12 default

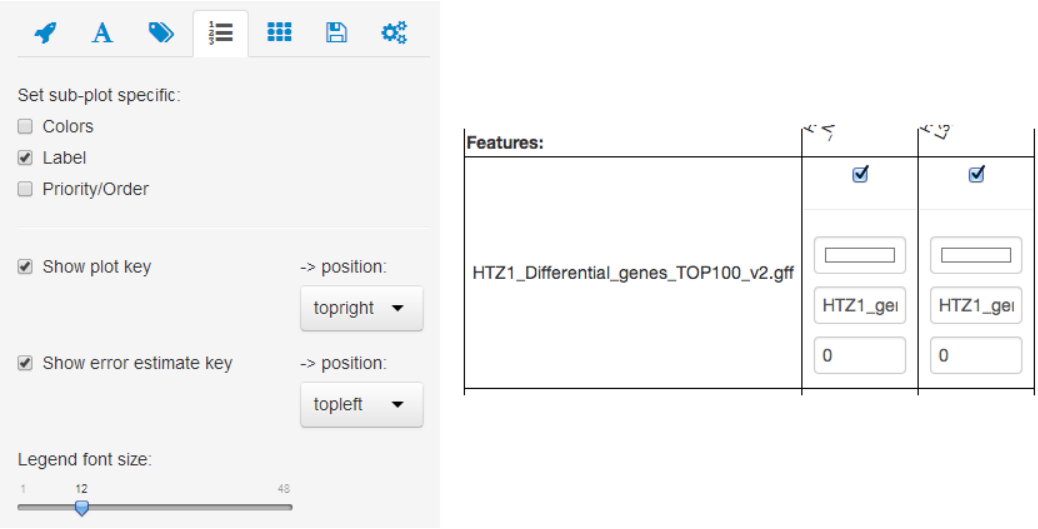

Figure 26: The view on keys, labels and colors panel (left). Color picker, label text input and Priority/Order checkboxes reviled on plot set grid (right).

Plotting and adjusting heatmaps

Heatmaps can be more informative comparing to average plots. If the variability in signal along given genomic feature comes from different biological classes the average plot might not be sufficient for proper examination of the signal or even misleading. SeqPlots implements heatmap plotting in similar way to Galaxy, plotting track-feature pairs as sub-heatmaps horizontally aligned on single figure. All sub-heatmaps must have the same number of data rows, hence in single plot mode simultaneous plotting is possible only on single feature or features containing exact same number of ranges. The heatmaps can be sorted and clustered by k-means, hierarchical clustering and super self organising maps (SupreSOM).

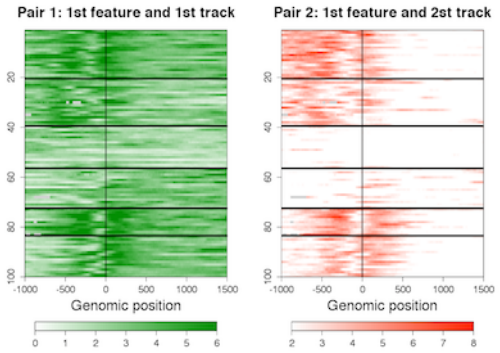

Figure 27: An example of heatmap plot

Heatmap setup tab

This tab groups heatmap specific options, that allows to manipulate various data processing and graphical options.

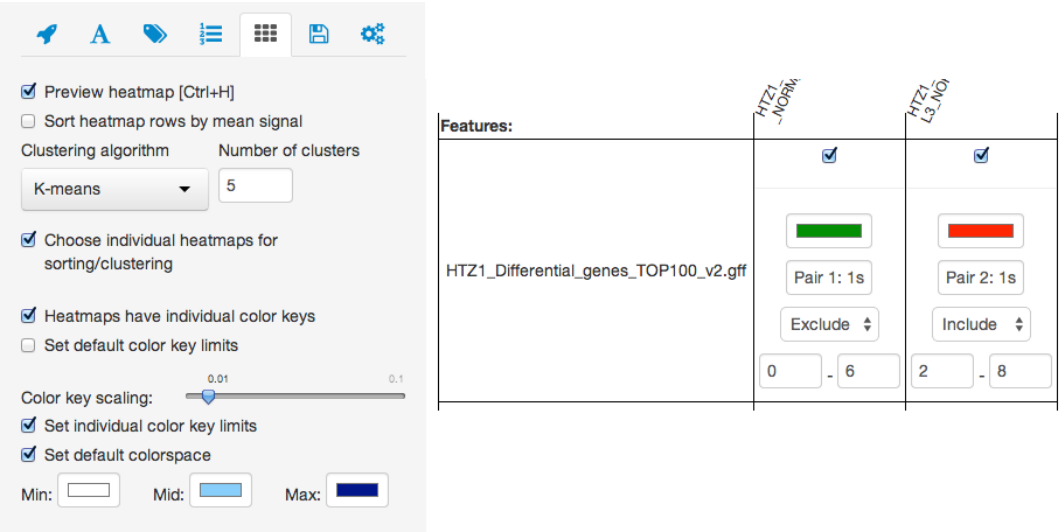

Figure 28: The view on Heatmap setup tab (left). Color picker, Label text input, Priority/Order checkboxes, Choose individual heatmaps for sorting/clustering control and Set individual color key limits numric inputs reviled on plot set grid (right).

- `Preview heatmap` - this checkbox indicates whether the preview of average plot or heatmap will be produced, its state is linked to `Line plot` and `Heatmap` buttons above the option tabs. It can be toggled from keyboard by using Ctrl/Cmd+H key combination
- `Sort heatmap rows by mean signal` - sorts the heatmap rows based on the mean value of each row across all sub-heatmps. The highest values on top. Turned off by default.
- `Clustering algorithm` - determines which clustering algorithm (k-means, hierarchical or SupreSOM) will be used to produce the clusters or turns of the clustering while *`do not cluster`* is selected. K-means by default.

- `Choose individual heatmaps for sorting/clustering` - similarly to `Colors`, `Label` and `Priority/Order`, which also works for heatmaps, this checkbox reveals new control on on **plot set grid** that determines if given sub-heatmap should be included in plotting and/or clustering. The excluded sub-plots will be plotted and clustered/ordered along with other sub-heatmaps, but their values would not influence the clustering/sorting. By default all sub-heatmaps are included. Following example shows hierarchical clustering on both heatmaps included (left) and second heatmap excluded (right):

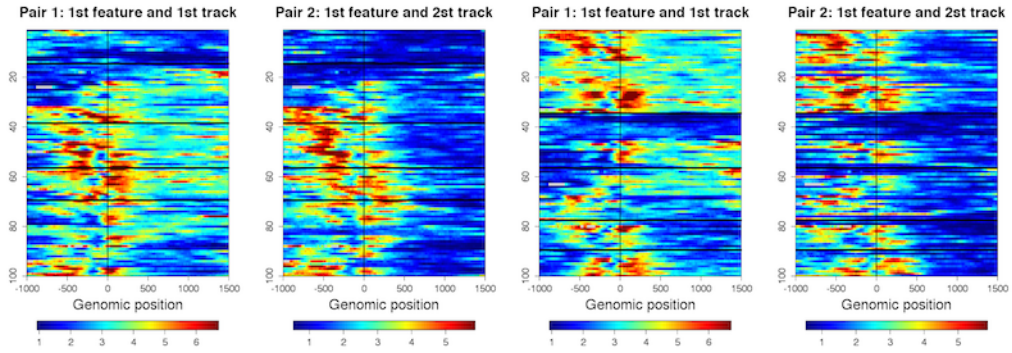

Figure 29: `Choose individual heatmaps for sorting/clustering` usage example: hierarchical clustering on both heatmaps included (left) and second heatmap excluded (right)

- `Heatmaps have individual color keys` - this option determines if each sub-heatmap should have separate color key (plotted below the heatmap) or single, common key should be calculated for all sub-plots (plotted rightmost). By default all sub-heatmap have its own color keys. The example below show the difference between separate (left) and common (right) color keys:

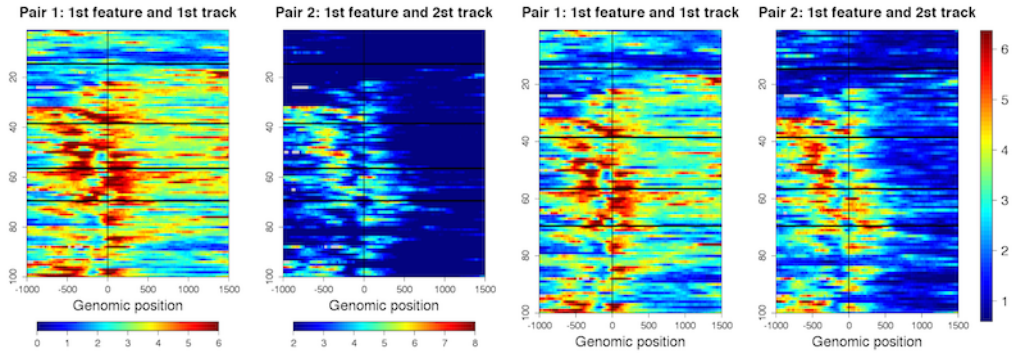

Figure 30: `Heatmaps have individual color keys` usage example: separate (left) and common (right) color keys

- `Set default color key limits` - this option determines the limits in mapping the numerical values to the colors. The range of generated is dependent on these options. Values smaller and lower than given limits will not produce further increase of heatmap color range, but will be plotted in the same color as closest limit value. If this checkbox is not selected, these values are auro-generated using `Color key scaling` parameter. If it is of two numerical fields are shown to hard set the limits.
  - `Color key scaling` - this slider influence how color key limits are generated. For example, 0.01 (default value) calculates limits using data range from 1-99 percentile of available data points. 0.1 uses data range from 10-90 percentile. The general formula for limit is: [quantile(data, `Color key scaling`); quantile(data, 1-`Color key scaling`)]
  - `min` and `max` numeric inputs - in opposite to auto generating color key limits they can be directly given as

numeric values

- `Set individual color key limits` - this option is similar to manual set up of color key limits, but allows to set up different values for individual sub-heatmaps. When this checkbox is selected `min` and `max` numeric inputs are revealed on **plot set grid**
- `Set default colorspace` - When this option is selected three color pickers are being shown. This allows to set up custom color mappings for heatmaps. The following example below shows standard jet colors (left), default blue color mapping after selecting the checkbox (middle) and custom color selection (right):

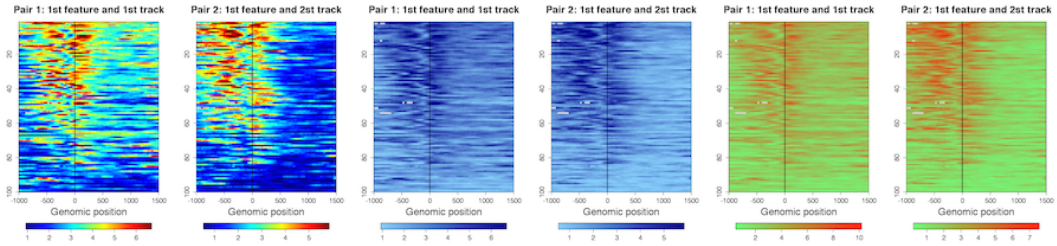

Figure 31: `Set default colorspace` usage example: standard jet colors (left), default blue color mapping after selecting the checkbox (middle) and custom color selection (right)

Other options controlling heatmap appearance

Many options from other tabs influence heatmap output. Here we provide the list of these inputs, please refer to "Viewing and manipulating plots" for further reference.

- **Titles and axis panel**
  - `X-axis label` - Label shown below horizontal axis, drawn separately for each sub-heatmap; default empty
  - `Y-axis label` - Label shown next to vertical axis, drawn separately for each sub-heatmap; default empty
  - `Labels font size` - Font size for axis labels and main labels of sub-heatmaps; default 16 points
  - `Axis font size` - Controls axis ticks font size; default 14 points
  - `Set X-axis limits` - Set hard plotting limits for X-axis; default values are whole range chosen during plot set calculation
- **Guide lines and data scaling panel**
  - `Transform signal` - if set to *`Log2 transform`* performs log2 transformation of the signal prior to plotting; default setting is *`Do not transform`*
  - `Show vertical guide line` - show the vertical line at point 0 - beginning of the feature or midpoint and end of the pseudo-length scaled features (only for anchored plots); turn on by default
- **Keys, labels and colors panel**
  - `Colors` - for hetmaps this input allows to control the color mapping of specific sub-heatmaps. The map allways start with white (for low color key limit) and finishes with selected color (for high color key limit).
  - `Label` - allows to set up custom sub-heatmap top labels
  - `Priority/Order` - The feature-track pairs with the highest priority will be plotted as left-most sub-heatmaps.
  - `Legend font size` - control the font size of common color key, inactive if heatmaps have individual color keys; 12 default

Output files and batch operations

Plots can be downladed as portable document files (PDFs) by clicking `Line plot` or `Heatmap` buttons in "Download:" section of **tool panel** (above the plot preview).

Figure 32: Download:" section of tool panel with `Line plot` and `Heatmap` buttons

Small buttons next to `Line plot` and `Heatmap` produce additional output files:

- `i` button next to `Line plot` downloads the PDF containing average plot keys
- `cluster diagram` button next to `Heatmap` downloads a cluster report giving cluster assignments for each feature as a comma separated value (CSV) spreadsheet.

The cluster report contains following columns:

- `chromosome` - the name of chromosome, contig or scaffold
- `start` - start of the feature (1 based chromosomal coordinate)
- `end` - end of the feature (1 based chromosomal coordinate)
- `width` - width of the feature in base pairs
- `strand` - strand of the feature
- `metadata_...` - the annotation columns driven from original GFF/BED e.g. gene name, score, group
- `originalOrder` - number of feature (row) in GFF/BED, can be used to restore original order after sorting on cluster ID
- `ClusterID` - the numeric ID of the cluster, topmost cluster on heatmap annotated with 1, and the bottom cluster with k, where k equals to number of clusters selected, exported only if clustering is enabled
- `SortingOrder` - the order imposed on heatmap by sorting by mean row(s) values, exported only if sorting is enabled
- `FinalOrder` - the final order of heatmap's rows, this can be influenced by sorting and clustering; 1 indicates topmost row

Sample report:

chromosome	start	end	width	strand	metadata_group	originalOrder	ClusterID	SortingOrder	FinalOrder
chrI	9065087	9070286	5200	+	g1	1	1	3	3
chrI	5171285	5175522	4238	-	g1	2	3	50	43
chrI	9616508	9618109	1602	-	g1	3	3	13	43
chrI	3608395	3611844	3450	+	g1	4	3	11	12

Table view:

chromosome	start	end	width	strand	metadata_group	originalOrder	ClusterID	SortingOrder	FinalOrder
chrI	9065087	9070286	5200	+	g1	1	1	3	3
chrI	5171285	5175522	4238	-	g1	2	3	50	43
chrI	9616508	9618109	1602	-	g1	3	3	13	43
chrI	3608395	3611844	3450	+	g1	4	3	11	12

PDF output size

The last tab (`Batch operation and setup`) on the **tool panel** includes batch operations and various other settings including PDF output size. By default the output PDF will be A4 landscape. This can be changed using the drop-down list to following settings:

- `user defined` - this option reveals two numeric inputs that allows to set output PDF width and height. The values must be given in inches.
- `Legal rotated` - US Legal landscape: 14" by 8.5"
- `A4` - A4 portrait: - 8.27" $\times$ 11.69"

- `Letter` - US Letter portrait: 8.5" × 11"
- `Legal` - US Legal portrait: 8.5" × 14"
- `Executive` - a.k.a Monarch paper: 7.25 × 10.5"

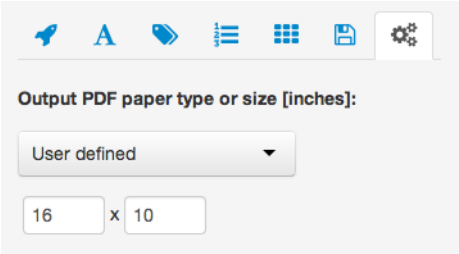

Figure 33: The view on top part of batch operation and setup panel

Batch operations

Controls to plot multiple plots at once are located on the `Batch operation and setup` tab, just below PDF paper options. It is possible to output the plots to multipage PDF, plot an array of plots on a single page (for average plots) or mix these options together.

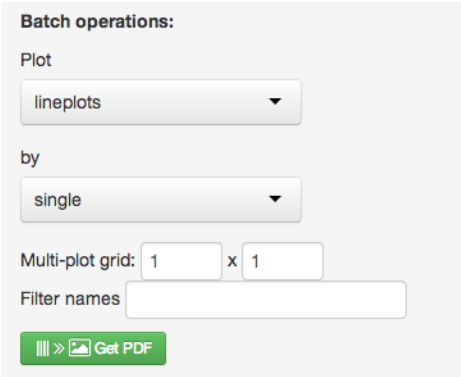

Figure 34: The view on bottom part of batch operation and setup panel

The first drop-down controls the type of the plot - either average or heatmap. The second drop down determines the strategy to traverse the **plot grid**. The options include:

- `single` - every single feature-track pair will be plotted on separate plot
- `rows` - the **plot grid** will be traversed by rows, which means one plot that contains all tracks per feature will be prepared
- `columns` - the **plot grid** will be traversed by columns, which means one plot that contains all features per tracks will be prepared

The `multi plot grid` option controls how many plots will be placed on each page of the PDF output, e.g. 1x1 means one plot per one page, while 3x4 means 3 columns and 4 rows of plots. If number of plots exceeds the number of slots on page the new page will be added to the PDF.

`Filter names` will apply a filter to plot titles, which are based on on uploaded file names. For example, if you uploaded 100 files starting with a prefix of "my_experiment_", you can remove this fragment from each plot title and/or heatmap caption by putting this string in `Filter names`.

Finally, pressing `Get PDF` produces the final output file. Please see example below:

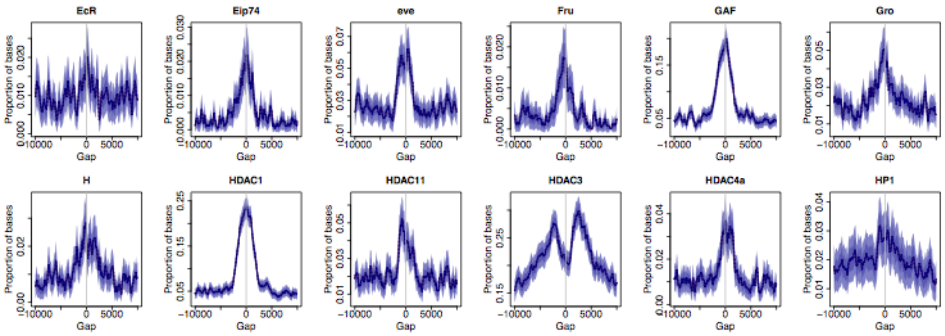

Figure 35: Batch plot usage example - multiple average plots aranged in 6x2 plot grid

Saving and loading plot sets

SeqPlots allows to save the plot sets as binary R files. This allows to quickly load pre-calculated set for replotting. Furthermore, the saved plot sets can be shared with other SePlots users.

Load or save plotset

Following controls are available on "Load or save plotset" panel:

- `Load saved plot set` - this drop-down list allows to select a plotset. Once the Rdata binary file is selected the **plot grid** will be shown instantaneously. Selecting the file reveals two additional buttons:
  - `Remove dataset` - this button deletes the selected saved plot set from user data.
  - `Download plotset` - this button saves a copy of the plotset in selected location.
- `Save current plot set` - this control allows to save the current plot set. Once the desired name of the file is put to the text input the Save button will appear. You can use it after calculating the plot set. It is also possible to save a copy of loaded plot sets. The plot set binary files can be renamed simply by loading them, saving a copy and deleting original source file.

All saved dataset can be found in `data location`/publicFiles. Any SeqPlots Rdata binaries put in the folder will become available for loading in `Load saved plot set` control.

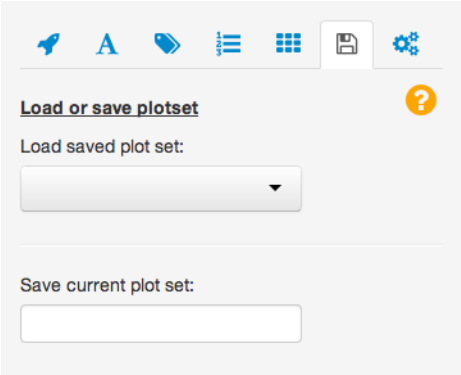

Figure 36: The view on the "Load or save plotset" panel

Plot set files structure

The plot sets files can be also directly loaded in R. This allows further processing and customization of the plots. Data structure is a nested list, which elements be accessed by `[[` R operator. The nesting goes as follow:

- `feature` - R list
  - `track` - R list
    - `means` - numeric vector giving mean signal value for each (binned) genomic position
    - `stderror` - numeric vector giving standard error for each (binned) genomic position
    - `conint` - numeric vector giving 95% confidence interval for each (binned) genomic position
    - `all_ind` - numeric vector giving the genomic position in base pairs
    - `e` - character string giveing numeric vector giving the indicates of anchored distance, NULL for point features plots
    - `desc` - auto generated title of the plot
    - `heatmap` - numeric matrix, (binned) signal values for each genomic position (columns) and each feature (rows)

The example structure:

```
List of 2
 $ HTZ1_Differential_genes_TOP100_v2.gff:List of 2
  ..$ HTZ1_JA00001_IL1andIL2_F_N2_L3_NORM_linear_1bp_IL010andIL009_averaged.bw    :List of 7
  .. ..$ means   : num [1:501] 2.52 2.52 2.52 2.53 2.54 ...
  .. ..$ stderror: num [1:501] 0.114 0.112 0.111 0.11 0.109 ...
  .. ..$ conint  : num [1:501] 0.226 0.223 0.221 0.218 0.217 ...
  .. ..$ all_ind : num [1:501] -1000 -995 -990 -985 -980 -975 -970 -965 -960 -955 ...
  .. ..$ e       : NULL
  .. ..$ desc    : chr "HTZ1_JA00001_IL1andIL2...\n@HTZ1_Differential_genes_TOP100_v2"
  .. ..$ heatmap : num [1:100, 1:501] 2.36 5.25 2.2 3.48 4.32 ...
  ..$ HTZ1_JA00001_IL3andIIL5_F_lin35_L3_NORM_linear_1bp_IL008andIL011_averaged.bw:List of 7
  .. ..$ means   : num [1:501] 2.36 2.35 2.35 2.36 2.38 ...
  .. ..$ stderror: num [1:501] 0.126 0.125 0.125 0.126 0.125 ...
  .. ..$ conint  : num [1:501] 0.249 0.249 0.247 0.251 0.249 ...
  .. ..$ all_ind : num [1:501] -1000 -995 -990 -985 -980 -975 -970 -965 -960 -955 ...
  .. ..$ e       : NULL
  .. ..$ desc    : chr "HTZ1_JA00001_IL3andIIL5...\n@HTZ1_Differential_genes_TOP100_v2"
  .. ..$ heatmap : num [1:100, 1:501] 2.61 3.17 1.42 2.46 4.26 ...
```

```
$ HTZ1_Differential_genes_BOTTOM100.gff:List of 2
 ..$ HTZ1_JA00001_IL1andIL2_F_N2_L3_NORM_linear_1bp_IL010andIL009_averaged.bw    :List of 7
 .. ..$ means   : num [1:501] 1.57 1.57 1.58 1.6 1.62 ...
 .. ..$ stderror: num [1:501] 0.0996 0.0985 0.1003 0.1022 0.1018 ...
 .. ..$ conint  : num [1:501] 0.198 0.195 0.199 0.203 0.202 ...
 .. ..$ all_ind : num [1:501] -1000 -995 -990 -985 -980 -975 -970 -965 -960 -955 ...
 .. ..$ e       : NULL
 .. ..$ desc    : chr "HTZ1_JA00001_IL1andIL2...n@HTZ1_Differential_genes_BOTTOM100"
 .. ..$ heatmap : num [1:100, 1:501] 1.64 1.37 1.61 1.77 1.86 ...
 ..$ HTZ1_JA00001_IL3andIIL5_F_lin35_L3_NORM_linear_1bp_IL008andIL011_averaged.bw:List of 7
 .. ..$ means   : num [1:501] 1.94 1.94 1.95 1.96 1.97 ...
 .. ..$ stderror: num [1:501] 0.123 0.123 0.124 0.126 0.128 ...
 .. ..$ conint  : num [1:501] 0.244 0.245 0.246 0.251 0.253 ...
 .. ..$ all_ind : num [1:501] -1000 -995 -990 -985 -980 -975 -970 -965 -960 -955 ...
 .. ..$ e       : NULL
 .. ..$ desc    : chr "HTZ1_JA00001_IL3andIIL5...\n@HTZ1_Differential_genes_BOTTOM100"
 .. ..$ heatmap : num [1:100, 1:501] 1.61 1.37 1.29 3.04 3.77 ...
```

Advanced options

Some additional SeqPlots options can be located at very bottom of `Bach operation and setup` tab:

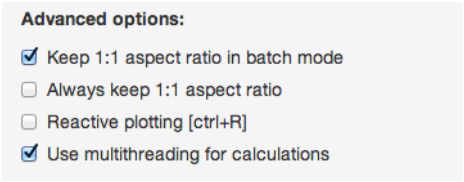

Figure 37: The view on 'Advanced options' section of the batch operation and setup panel

- `Keep 1:1 aspect ratio in batch mode` - This option guarantee that the ratio between X- and Y-axis height will be 1, hence the produced plots will be rectangular in batch mode. This prevent stretching the plots while fitting the single row or column to one page. Turned on by default.
- `Always keep 1:1 aspect ratio` - This checkbox extends previous behaviours on single plots - the figures always will be rectangular, no matter the paper size. Turned off by default.
- `Reactive plotting` - While this checkbox is selected, all plotting operation ar executed on fly. That means changing the font size, title caption, etc. will execute the plotting routine and changes will be visible on preview. `Reactive plotting` might be useful for exploratory data analysis using plots. However, it is not recommended while working with big heatmap plots. Touleble from keyboard by pressing [ctrl/cmd+R]. Turned off by default.
- `Use multithreading for calculations` - This option is available only on desktop instances of SeqPlots under Mac OS X and Linux. While turned off R will not fork the child processes for plotting and plot set calculations. It is useful for debugging, since in single process mode all warning/errors will be directly printed to R console. Also might increase the performance for plotting small average plots. Turned off by default.

Error messages

Adding the files:

`Problem with line N: "line_text" [internal_error]`

The import of feature file (GFF or BED) was not successful due to mis-formatted file.

```
Chromosome names provided in the file does not match ones defined in reference genome.
INPUT: [chr3R, chr2L, chr2R, chr3L]
GENOME: [chrI, chrII, chrIII, chrIV, chrV, ...]
```

There are unexpected chromosome names in input file. Following genomes: *Arabidopsis thaliana, Caenorhabditis elegans, Cyanidioschyzon_merolae, Drosophila melanogaster, Homo sapiens, Oryza sativa, Populus trichocarpa, Saccharomyces cerevisiae and Zea mays* support chromosome names remapping between different naming conventions, including: AGPv2, ASM9120v1, Ensembl, JGI2_0, MSU6, NCBI, TAIR10 and UCSC. If you see above error in one of these genomes there are still unexpected names after the correction. The problematic chromosome names are given in the error message. Remove GFF/BED lines corresponding to them or upgrade the genome to one containing proper naming. Alternatively set genome to NA.

`File already exists, change the name or remove old one.`

File named like this already exists in the database, it is impossible to have two files sharing same filename.

`ERROR: solving row 300: negative widths are not allowed`

The the row 300 have end coordinate smaller than beginning, hence the width in negative. To fix it the start and stop indicates should be swapped. This error often happens when negative strand (-) ranges are misformatted.

Version History

SeqPlots v0.9.1 RC

GENERAL:

- inputs and features sorted alphabetically
- DataTables v1.10.0 with pagination, selection number indicator and infinite row selections
- buttons for heatmap and lineplots, PDF default sizes,
- changes in GUI layout
- default PDF output paper size set to A4 horizontal,
- Font sizes are in points
- preview is compatible with A4 PDF output (at 100 DPI)
- color key for heatmap are always generated using image.plot function that provides better labeling
- batch plots do not override individual labels if set
- option to keep 1:1 aspect ratio (default for batch plots)
- miscellaneous options renamed for clarity

SeqPlots v0.9.0 RC

FEATURES:

- Multi-plot grid option in batch mode - many line plots on single page

GENERAL:

- R 3.1 and BioC 2.14 compatibility
- faster BigWig signal retrieval, no need for modified rtracklayer C code in the package
- warning message if JS File API is not supported (old browsers)
- improved the performance of heatmap plotting by using list of matrices instead concatenated matrix

BUGFIX:

- application start properly without any BSGenome genomic packages installed
- cluster report - the final order agrees with cluster indicates

SeqPlots v0.8.2b

FEATURES:

- Hierarchical and super self-organizing network clustering added for heatmaps
- Anchored motif plots
- The row order of the heatmap is exported along with cluster report

GENERAL:

- JS color picker added for browsers, that do not support i.e. Firefox (checked with modernizr.js library)
- Single process mode and Microsoft Windows compatibility (running without fork parallelization)
- Shiny 0.9.1 compatibility
- Saved datasets can be downloaded for local usage
- Clicking row or column name in plot grid toggles the checkboxes
- Minor GUI changes

SeqPlots v0.8.1b

GENERAL:

- GUI redesign: plot matrix incorporates sub-plot/heatmap specific controls, all heatmap options gathered in single tab
- warning before closing/refreshing a webpage with active session
- cookie based default options: user, genome and deactivate page exit warning
- heat-map clusters provided as cluster report - a CSV file containing original features, annotations and cluster information, see more: https://bitbucket.org/przemol/seqplots/wiki/Heatmaps#markdown-header-cluster-report
- Wiggle files processing: correct for multiple header definitions and roman/arabic chromosome names correction
- Optimised keyboard shortcuts: plot - RETUTRN or ctrl/cmd+SPACE, switch heatmap - ctrl/cmd+H, switch reactive plotting - ctrl/cmd+R
- minor speed improvement

BUGFIX:

- Motif density plots and heatmaps: flip rows on (-) strand

SeqPlots v0.8.0b

GENERAL:

- GUI redesign, option partitioned to more tabs
- preview plot is zoomed on click rather than on mouse hover
- possibility to remove multiple files
- comments visible as popup in file managmed window
- all chromosome naming conventions (most notably chrX/X and variants of chrM/M/MtDNA/MT etc.) are accepted (http://www.bioconductor.org/packages/release/data/annotation/html/seqnames.db.html)
- incoming featurefiles (GFF and BED) are not processed, just chacked for errors
- explicit error handling for incoming flies, the line with problem or unexpected chromosome(s) are indentified to the user
- motif density tracks can be binned (defoult at 10bp)
- tracks amd motif densities cna be mixed together in plots

SERVER:

- server_config.R added - a configuration file that allows to set up server varaiables, e.g. the user data location

MAC OS X APP:

- interface to insrall new genomes from Bioconductor and local resources (R BSgemome format: http://www.bioconductor.org/packag
- option to set up data location

SeqPlots v0.7.0a

- SeqPlots for Mac OS X relesed - an user frienddly wrapper app containig R, packages and SeqPlots coede
- heatmap plotting added
- motif denstty plotting added for lineplot and heatmap
- minior interface redesign
- reactive interface can be turn off for plottting, user plots on demend
- adding files from jQuery File Upload (http://blueimp.github.io/jQuery-File-Upload/) is handled directly by R eliminating additional node.js server application and making proper file handling for desktop version
- computationally expensive operations (calculating plot matrix and plotting) are handeled by new R process (parallel R library) - many proces can run simmutainously in same Shiny instance, user can get fedback from the calcualtion can be cancelled

SeqPlots v0.6.0a

- Shiny (https://github.com/rstudio/shiny) used as R web fraimwork, support for Rserver/EXT JS version dropped
- support for 145 genomes from UCSC database (via user provideing valid genome symbol)
- new reactive user interface
- new plot type: midpoint features - it calculates the middle of given features and centres the summary on it
- the option to ignore the strand (plot always in the same direction)
- the option to remove the zeros (0 value of score in Wiggle track) from mean and error estimate calculations
- the support for BED feature files (in addition to GFF, Wiggle (all variants), BigWiggle)
- automatic chromosome name correction for C. elegans genomes (I => chrI, MtDNA => chrM, etc.)
- accepts wiggle with overlapping ranges (e.g. microarray experiments processed using MA2C)
- basic user management for uploaded files
- option to download the features and track files directly from application

SeqPlots v0.1.0a - v0.5.0a

- Initial test and alpha releases

Explanations

- "**feature**" - a genomic interval defined by **chromosome** name, **start** and **end** positions and the **directionality** (strand). The end must always be a bigger number than start, so the width of the range is not negative. Start and end means here the numeric start of the interval and should not be confused with TSS and TTS.
  For example, in BED format this information is stored in following text tab delimited format: `chr7 127471196 127472363 . . +`

- "**directionality**" - the strand of genomic feature, determining if the plotting range should be anchored around the star or and, and the direction in which signal is being processed to create the average track or heatmap. Unknown directionality is marked by `*` and treated as + for calculations.

- "**track**" - the file assigning the continuous signal (score) to genomic locations across the chromosomes. The signal usually comes from sequencing experiments, like ChIP-seq, RNA-seq, DNase-seq, MNase-seq, or from computational tools, for example nucleosome occupancy prediction, CpG density.
  For example, in BedGraph format this information is stored in following text tab delimited format: `chr19 49302300 49302600 -0.75`

- "**reference genome package**" - the R BSgemome package containing the full reference sequence for given species. It is also used to provide universal chromosome names and chromosome lengths taht are used as plotting boundaries.

- "**reads coverage**" - The basic way to calculate the signal from sequencing based assays. The numeric representation shows how much reads was aligned to given genomic location. This can be a proxy to protein-DNA binding (ChIP-seq) or the expression (RNA-seq). Can be calculated using BedTools: http://bedtools.readthedocs.org/en/latest/content/tools/genomecov.html Also known as `pileups`.
