Contents

[2 Instructions 1](#_Toc444509187)

[2.1 Selecting a dataset 1](#_Toc444509188)

[2.2 Loading a peak file 1](#_Toc444509189)

[2.3 Looking at the Correlation table 2](#_Toc444509190)

[2.4 Static heatmap 2](#_Toc444509191)

[2.5 Responsive heatmap 3](#_Toc444509192)

[2.6 Tree 3](#_Toc444509193)

[2.7 Samples metadata 3](#_Toc444509194)

[2.8 Correlation correction: 3](#_Toc444509195)

[3 FAQ 4](#_Toc444509196)

[3.1 How to cite? 4](#_Toc444509197)

[3.2 Where can I find the source code? 4](#_Toc444509198)

[3.3 Can you add this dataset on HeatChIPseq? One of your dataset does not seemed up to date, can you update it? 4](#_Toc444509199)

[3.4 Could you implement this new feature? 4](#_Toc444509200)

[3.5 I have uploaded a peak file. How do I remove it? 4](#_Toc444509201)

# Instructions

To access the application, click on the **Use Application** button on the **top bar**.

## Selecting a dataset

The first thing you must do is to **select a dataset** to work with. At present, there is 3 datasets available: Transcription Factor (TF) ChIP-seq from [ENCODE](http://www.genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html) in human (hg19), and [CODEX](http://codex.stemcells.cam.ac.uk/about.php) TF ChIP-seq for human (hg19) or mouse (mm10).

The dataset will only be loaded when you will have uploaded a peak file, or once you will have focused on a plot panels.

## Loading a peak file

You can now upload a peak file (a list of genomic coordinates). Note that this is not mandatory, and you can jump to section [2.4](#_Static_heatmap) if you simply want to browse the selected dataset. The application accept tab-delimited text file following a bed format: first column must be chromosome names (chr1, chr2, chrX, etc.), the second must be region start, and the third region ends. Any additional column will be discarded (peak name, score, strand, etc.). Please, be sure to untick the **My peak file contains a header** option if your file do not contains any header, as the first line of the file would be discarded. You can also fill the **Name of your experiment** field, it will modify the label of your experiment in the plots.

The maximum size you can upload is 10 Mb. If your bed file is larger than that, try keeping only the first three columns of it to reduce file size. If after removing the non-essential columns the file is still bigger than 10 Mb and you are confident you want to analyse that many region, [contact us](mailto:guillaume.devailly%20at%20rolsin.ed.ac.uk) (replace at with @) and we will consider increasing the size limit.

Once the upload is complete, a subtle progress bar will appears on top of the page, and quick description of the on-going steps can be found on the top right of the page. It should take less than a minute for 30.000 regions. The **My Peaks** tab should now display a table of your coordinates.

You can download **an example file**. It contains coordinates for human CpG island (hg19) fetch from the [UCSC table browser](http://genome.ucsc.edu/cgi-bin/hgTables), so it will work only with human datasets. The file contains a header, so leave the **My peak file contains a header** option ticked.

Currently, the easiest way to unload a peak file without uploading a new one is to come back to the main page and open a new HeatChIPseq window.

## Looking at the Correlation table

If you uploaded a peak file, the **Correlation table** tab should now display a 3 column table. The first column contains the name of **experiments** in the selected dataset, the second column contains the **correlation** coefficient between the uploaded coordinates and the corresponding experiment in the dataset. The third column contains **scaled correlation** values that we will discuss on section [2.8](#_Correlation_correction:). At the bottom of the page, you will find a **Save as tab delimited .txt** download button, allowing you to download a copy of the correlation table as a tab delimited text file. Those file can be imported into many datasheet software including Microsoft Excel.

You can copy an experiment name from the correlation table and paste it into the Search field of the Samples metadata tab to learn all the metadata we have for that experiment.

## Static heatmap

The heatmap can take up to a minute to be displayed, and a bit longer if you are importing a new peak file or switching datasets.

The **static heatmap** tab will display a clustered heatmap representation of the correlation matrix: each row and column represent a single experiment. The colours code for the correlation value, and a colour legend can be found bellow the heatmap: white is a correlation coefficient equal to 0, red equals to 0.5, and black equals to 1. Several options are available on the side bar panel on the left, **3 - Plot customisation**. In the order of appearance:

* The **Highlight my experiment in the heatmap** tick box. When checked, if you uploaded a peak file, the row and line corresponding to your experiment will be highlighted: a 0 correlation value will be displayed in yellow, a 0.5 value in bright green, and a 1 correlation value still in black. This can help you localised your experiment in crowded heatmaps.
* One or several **subset** options: You can restrict the experiments to be displayed using those fields. They may vary a bit between datasets. You can select experiment concerning one or several **transcription factors** (TFs), as well as select experiment done in one or **several cell lines/tissues**. Leaving the field empty will select all experiments.

If only two or less experiments from the dataset match your criteria, the heatmap will not be displayed. Please be less stringent in your filtering criteria. You can look for available experiment in the Samples metadata tab.

The uploaded peak list will never be filtered out, and will always be displayed in the heatmap.

* The function of the **Uploaded experiment correlation correction** field and the **Maximum expected correlation value for linear scaling correction** slider is detailed in section [2.8](#_Correlation_correction:).
* The **Advance clustering options** button will show and hide the **Distance calculation** and **Clusterisation method** options. The default value are fine, but you can select your favourite one anyway. Values are passed to the R [dist()](https://stat.ethz.ch/R-manual/R-patched/library/stats/html/dist.html) and [hclust()](https://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html) method parameter. A special case is the **1 – correlations** distance method: instead of using the dist() function on correlation values, we use 1 – correlations as a measure of distance between experiments ( a correlation of 0.75 will give a distance of 0.25).
* The **Sample name size** slider will modify label size.
* The **Sample name margin** will modify the size of the margin of the plot.

If you see an error message: figure margins too large, try reducing the size of the margin as well as the size of the name size.

Bellow the heatmap, you should see 3 buttons: **Save as png**, **Save as pdf** and **Save as svg**, to export the image in those format.

## Responsive heatmap

The **responsive heatmap** tab displays an interactive plot provided through the [plot.ly](https://plot.ly/) API. It can takes tens of seconds to be displayed on powerful computer running later version of Firefox and Chrome, but takes somewhat longer to load on Safari, Internet explorer and Edge (blame plot.ly). It represent the same version of the **Static heatmap** presented in the previous section, without the side dendrograms (trees). Most of the options are the same as for **Static heatmap**, so please check section [2.4](#_Static_heatmap) for more information. Additional options are available on the responsive heatmap itself:

* Mouse over information: when hovering the mouse over the heatmap, a text box will appears, giving you the name of the experiment in the x -axis, the name of the experiment in the y-axis, and there correlation value (z).

If the Highlight my experiment in the heatmap option is enabled, the z value of the highlighted cells will be offset by 2: a correlation value of 0.75 will have a z-value of 2.75.

* You can zoom on the heatmap by drag and dropping the mouse defining a rectangle. To zoom out, double-click on the heatmap or click the **Reset axis** or **Autoscale** buttons on the top left corner of the heatmap.
* By selecting the **Pan** button on the top left corner of the heatmap, drag and dropping the mouse will allow you to pan, which can be useful once you zoomed in a specific part of the heatmap.
* The **Download plot as a png** button is non-functional (blame plot.ly). Please use the one under the **Static heatmap** or look at the next step.
* The **Save and edit plot in the cloud** will send the data to plot.ly where, after the loading, you can play with many tools offered by plot.ly, such as export a JSON version of the data, save the plot as a png, change the theme (and notably the colour scale), etc.

## Tree

The **Tree** tab will display only the dendrogram (or Tree) from the experiment clustering. Options are mostly similar to the one for static heatmap, so please have a look to section [2.4](#_Static_heatmap). At that time, the **Highlight my experiment in the heatmap** option will not highlight your experiment in the dendrogram.

If you see an error message: figure margins too large, try reducing the size of the margin as well as the size of the name size.

## Samples metadata

The sample metadata tab will display metadata information (experience name, cell type, url of the original data, etc.) for the **selected dataset**. The table is sortable and searchable, and can be downloaded as a tab-delimited txt file using the **save** button below the table.

You can copy an experiment name from the correlation table and paste it into the Search field of the Samples metadata tab to learn all the metadata we have for that experiment.

## Correlation correction:

Sometimes, the maximum correlation of a user peak file with any experiment in the dataset can be quite low. In some situations, when the user is confident that the top hits are relevant, this may be evidence of strong “batch effect” that could reflect artefact of library preparation method, peak calling or ChIP efficiency. Those low correlation value might bias the clustering (due to [Long Branch Attraction](https://en.wikipedia.org/wiki/Long_branch_attraction)). The user can manually correct this effect using the **Linear scaling** method on the **Uploaded experiment correlation correction** option. The correlation values will be linearly up-scaled so the maximum correlation value will now be equal to the value of the **Maximum expected correlation value for linear scaling correction** slider (default 0.95). The resulting transformed correlation value can be obtained from the **scaledCorrelation** column of the **Correlation table**.

The linear scaling of correlation value will not change the ordering of the values, only scaled the value (ie he third most correlated experiment will be the same with or without scaling).

# FAQ

## How to cite?

A preprint should be available in a few weeks.

## Where can I find the source code?

The source code can be find on [GitHub](https://github.com/gdevailly/HeatStarSeq_gh).

## Can you add this dataset on HeatChIPseq? One of your dataset does not seemed up to date, can you update it?

Please [contact us](mailto:guillaume.devailly%20at%20rolsin.ed.ac.uk) (replace at with @) and we will have a look. A well curated dataset can be implemented / updated within a working day.

## Could you implement this new feature?

Please [contact us](mailto:guillaume.devailly%20at%20rolsin.ed.ac.uk) (replace at with @). We will be very pleased to consider implementing any feature that will improve the usability of this application.

## I have uploaded a peak file. How do I remove it?

At the moment, there is no option to remove a peak file. You can replace it with any other peak file clicking the **Browse** button again. To remove a peak file without uploading a new one, the simplest method is to open a new HeatChIPseq session by going **Back to the main page**. Refreshing the HeatChIPseq page may result in slightly erratic outcome.

# About

**HeatChIPseq** is a part of **Heat\*Seq**, an attempt to make genome-wide comparison of high throughput sequencing experiments easier. It was developed by Guillaume Devailly, Anna Mantsoki and Anagha Joshi at the [Roslin Institute](http://www.roslin.ed.ac.uk/), and funded by the [Biotechnology and Biological Sciences Research Council](http://www.bbsrc.ac.uk/). It use [R](https://www.r-project.org/) [shiny](https://www.rstudio.com/products/shiny/), [plot.ly](https://plot.ly/), and various [CRAN](https://cran.r-project.org/) and [Bioconductor](http://www.bioconductor.org/) packages, and datasets from [ENCODE](http://www.genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html) and [CODEX](http://codex.stemcells.cam.ac.uk/about.php). Sources are available on [GitHub](https://github.com/gdevailly/HeatStarSeq_gh).

Expression