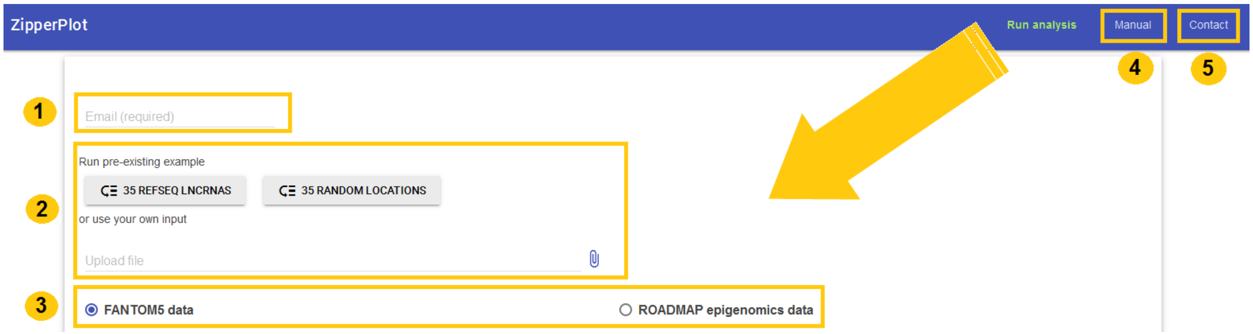
Welcome to Zipper plot webserver's documentation (v.1.2).

Updated last: 29 March 2017



1. Email address (required)

A link containing the detailed report generated from your input will be directly emailed to you at the specified e-mail address.

2. User input

The user has two options: a) Load a pre-existing example (set of 35 RefSeq lncRNAs or 35 random locations); b) Upload a file (in plain text format) containing genomic features of interest (**max: 20,000**).

The Zipper plot application requires three tab-separated fields **as input**: chromosome, genomic coordinate (hg19) of the transcription start site (TSS) and strand (e.g.: `chr1	6653265	+`).

IMPORTANT: Only one genomic feature per line AND each line should contain three tab-separated fields.

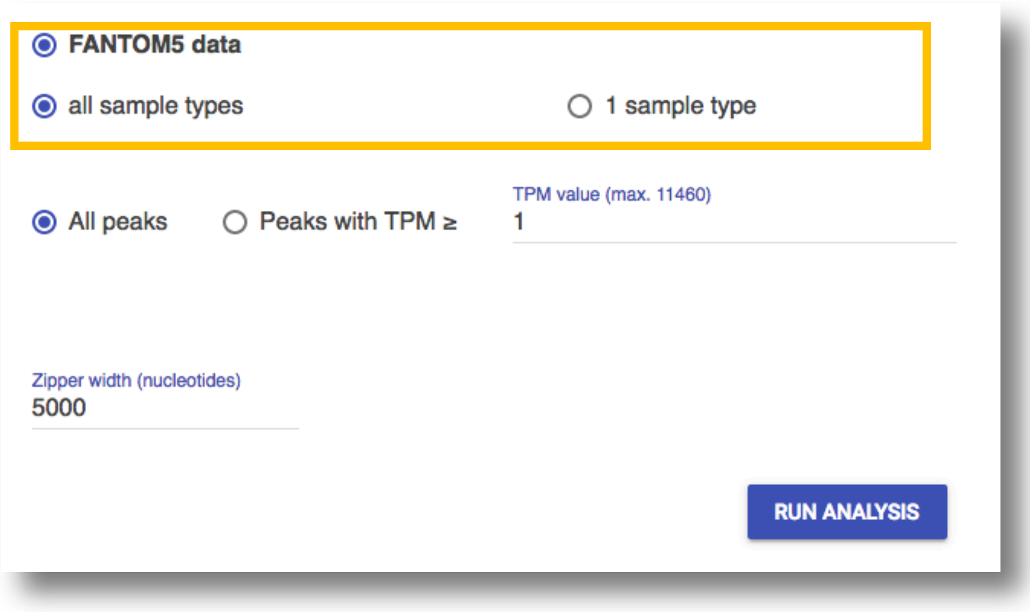
OPTIONAL: The user can include a 4th column if he/she has a label for the genomic feature being studied (e.g.: **`chr12	3884608	+	lnc-PRMT8-4:1`**)

3. Data selection: FANTOM5 data (CAGE-seq) vs Roadmap Epigenomics data (ChIP-Seq, DNase-seq)

After selecting the data type of interest (i.e. CAGE, histone mark or DNase peaks), the user has the option to run the analysis in one sample type of interest or across all available sample types. In the first option, the user knows in advance in which tissue the set of genomic features are more likely to be expressed; with the second option, each individual genomic feature is analyzed across all samples and the sample in which the peak is most closely associated to the genomic feature will be retained for further analysis.

*"NOTE: For visualization purposes, since CAGE-seq peaks are very **narrow (width ranging from 1 to 283nt; with more than 50% < 17 nt)**, we have artificially made all CAGE peaks uniform in width. Even though this may result in peaks that visually seem to be overlapping with the TSS even when they are not **(or few cases where the peak actually overlaps the TSS but not visually)**, the summary report table contains all the correct (unmodified) information and all the derived statistics have been computed with the unmodified data."*

3.1 FANTOM5 + ALL sample types



● FANTOM5 data

● all sample types ○ 1 sample type

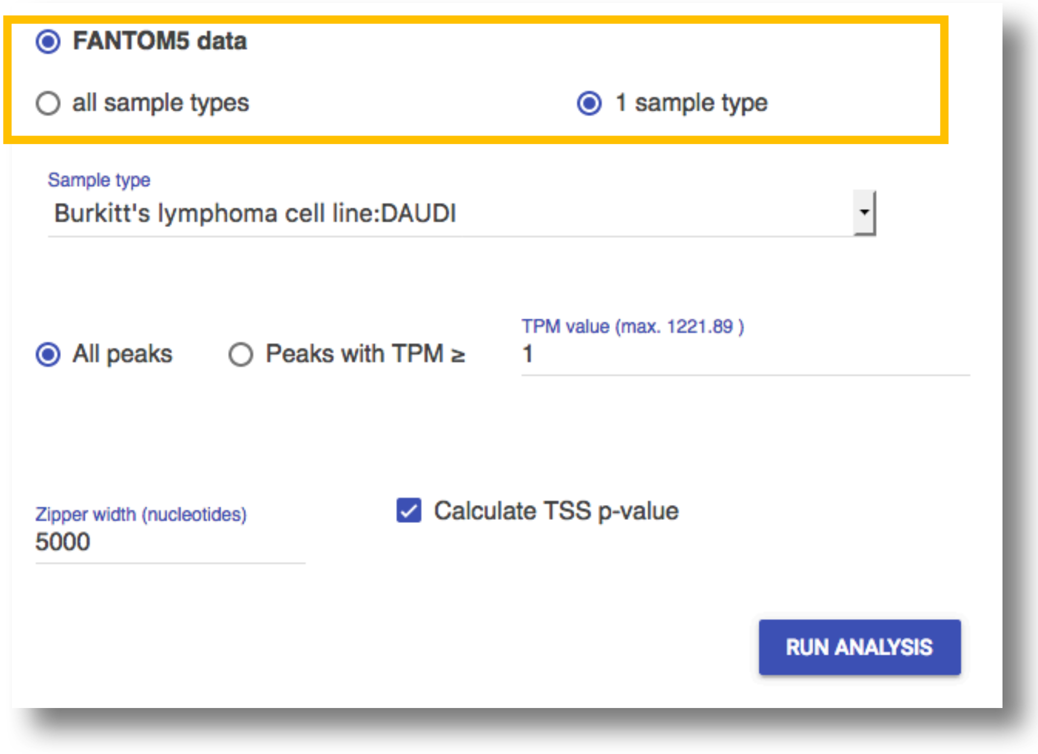
● All peaks ○ Peaks with TPM ≥

TPM value (max. 11460)
1

Zipper width (nucleotides)
5000

1) By default, **"All (CAGE) peaks"** in the database are used (**TPM value** > 0). However, the user can introduce a different value. Depending on which chromosomes were present in the user's input, the maximum TPM value that can be chosen (across the different chromosomes **for all sample types**) will appear. This information will allow the user to choose a reasonable threshold.
2) Zipper width: By default, the Zipper plot is displayed in a +/- 5000-nucleotide window (+/- 5kb from TSS). This value can easily be changed by the user.
3) Finally, by clicking the **button RUN ANALYSIS,** your analysis starts and a detailed report is generated and directly emailed to you.

3.2 FANTOM5 + 1 sample type



● **FANTOM5 data**

○ all sample types ● 1 sample type

Sample type
Burkitt's lymphoma cell line:DAUDI

● All peaks ○ Peaks with TPM ≥

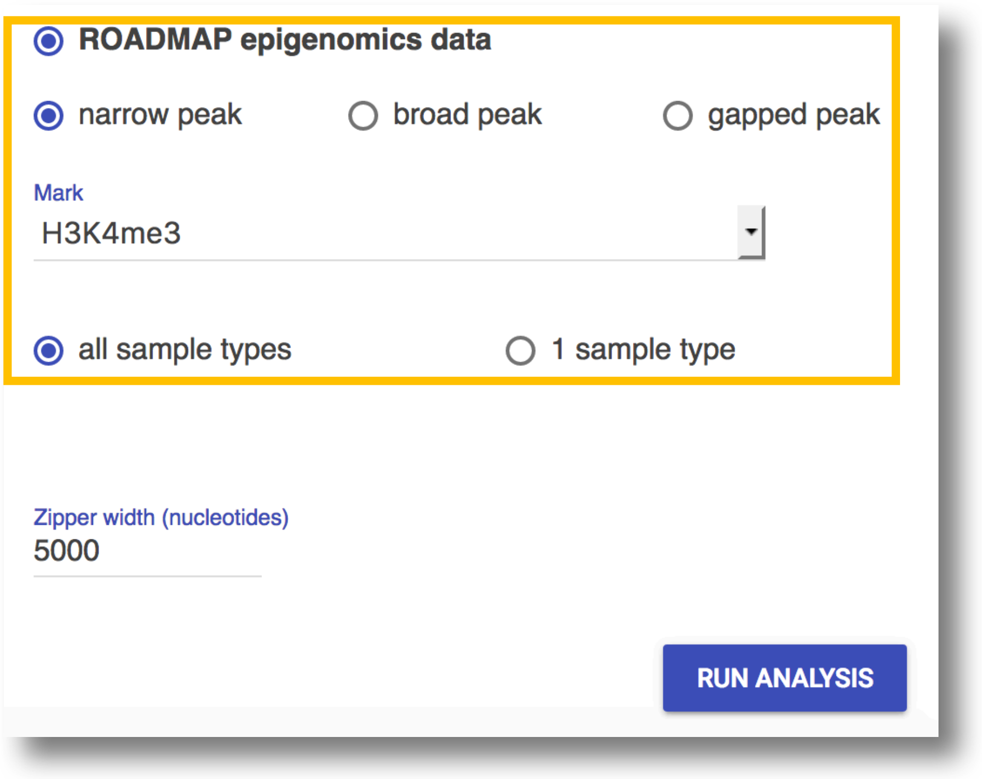
TPM value (max. 1221.89)
1

Zipper width (nucleotides)
5000

☑ Calculate TSS p-value

1) First the user selects one sample type from the **dropdown menu** (e.g.: Burkitt's lymphoma cell line:DAUDI).
2) By default, **"All (CAGE) peaks"** in the database are used (**TPM value** > 0). However, the user can introduce a different value. Depending on which chromosomes were present in the user's input, the maximum TPM value that can be chosen (across the different chromosomes **for the selected sample type**) will appear. This information will allow the user to choose a reasonable threshold.
3) Optionally, the closest peak (for each TSS) can be retrieved for all sample types present in the database. Therefore, for each TSS, a **TSS p-value** is calculated comparing how many sample types have a peak as close (or closer) to the TSS than the one found in the sample type chosen by the user.
4) Zipper width: By default, the Zipper plot is displayed in a +/- 5000-nucleotide window (+/- 5kb from TSS). This value can easily be changed by the user.
5) Finally, by clicking the **button RUN ANALYSIS,** your analysis starts and a detailed report is generated and directly emailed to you.

3.3 Roadmap Epigenomics + ALL sample types



● **ROADMAP epigenomics data**

● narrow peak ○ broad peak ○ gapped peak

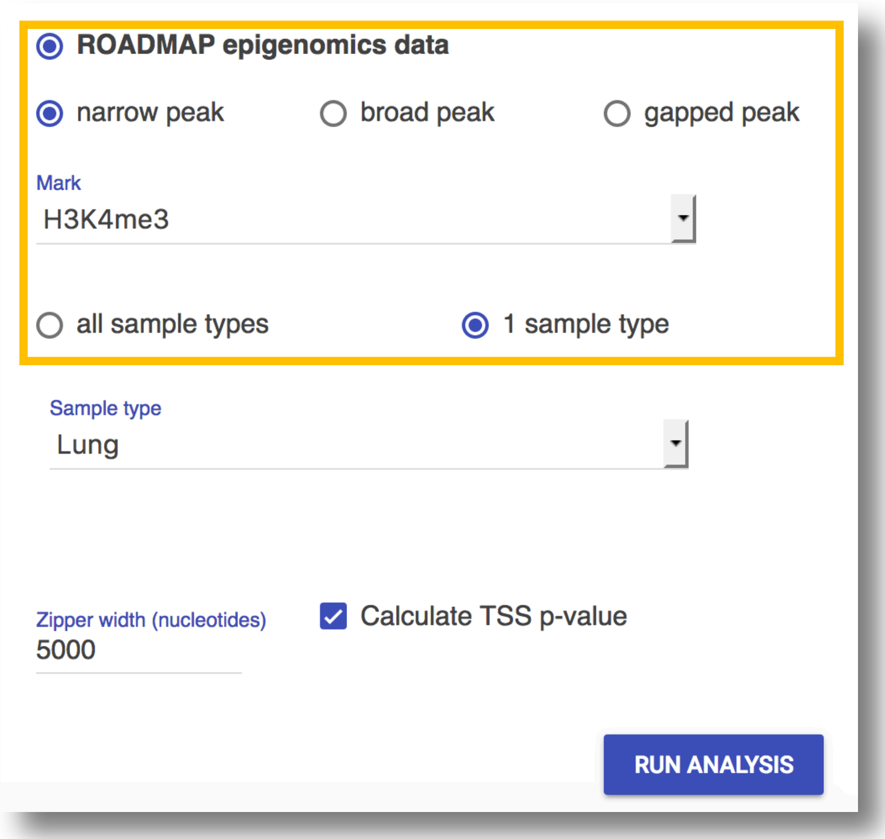
Mark
H3K4me3

● all sample types ○ 1 sample type

Zipper width (nucleotides)
5000

1) First the user selects one type of peaks among narrow, broad and gapped peaks.
2) Next, the user selects one mark indicative of transcriptional activity (e.g.: H3K4me3).
3) Zipper width: By default, the Zipper plot is displayed in a +/- 5000-nucleotide window (+/- 5kb from TSS). This value can easily be changed by the user.
4) Finally, by clicking the button RUN ANALYSIS, your analysis starts and a detailed report is generated and directly emailed to you.

3.4 Roadmap Epigenomics + 1 sample type



● ROADMAP epigenomics data

● narrow peak ○ broad peak ○ gapped peak

Mark
H3K4me3

○ all sample types ● 1 sample type

Sample type
Lung

Zipper width (nucleotides)
5000

☑ Calculate TSS p-value

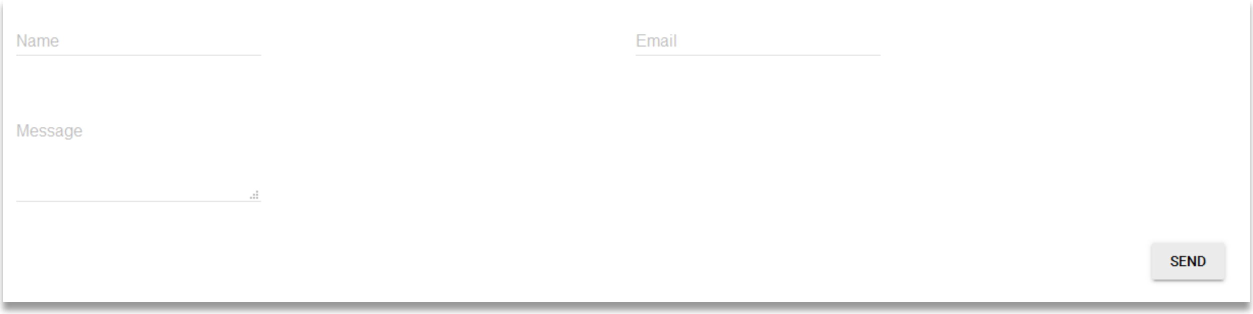
1) First the user selects one type of peaks among narrow, broad and gapped peaks.
2) Next, the user selects one mark indicative of transcriptional activity (e.g.: H3K4me3).
3) Selection of a sample type from the dropdown menu (e.g.: Lung).
4) Optionally, the closest peak (for each TSS) can be retrieved for all sample types present in the database. Therefore, for each TSS, a **TSS p-value** is calculated comparing how many sample types have a peak as close (or closer) to the TSS than the one found in the sample type chosen by the user.
5) Zipper width: By default, the Zipper plot is displayed in a +/- 5000-nucleotide window (+/- 5kb from TSS). This value can easily be changed by the user.
6) Finally, by clicking the **button RUN ANALYSIS,** your analysis starts and a detailed report is generated and directly emailed to you.

4. Manual

It will lead you to the most recent version of the Zipper plot webserver's documentation.

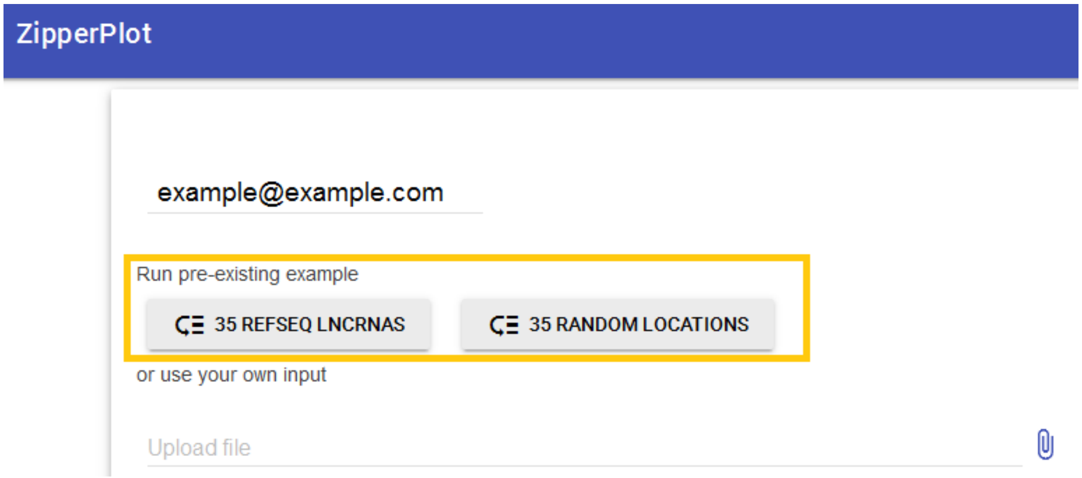
5. Contact

Bug reports, suggestions and contributions are very welcomed. Please contact us via our contact form.



INTERPRETING THE SUMMARY REPORT

We have made available two examples to try out our webtool: 35 random locations across different chromosomes and 35 lncRNAs from Refseq. These can be directly used as input by clicking the corresponding button:



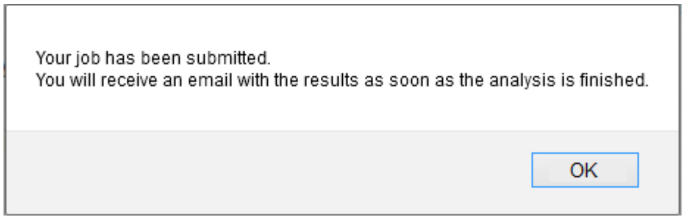
example@example.com

NOTE: All the queries introduced by the users are treated **confidentially** during processing and are **removed from our server one week after its creation**.

To note, the closer the peaks are distributed around the TSSs, the smaller the Area Under the Zipper (AUZ). A "closed zipper" (AUZ=0) indicates perfect overlap between closest peak and TSS for all the genomic features being studied.

As an example, we have selected FANTOM5 (CAGE-seq) data + ALL sample types but we encourage the final user to change this set up and work with their own data.

Once the button "**RUN ANALYSIS**" is clicked, the following pop-up window will appear:

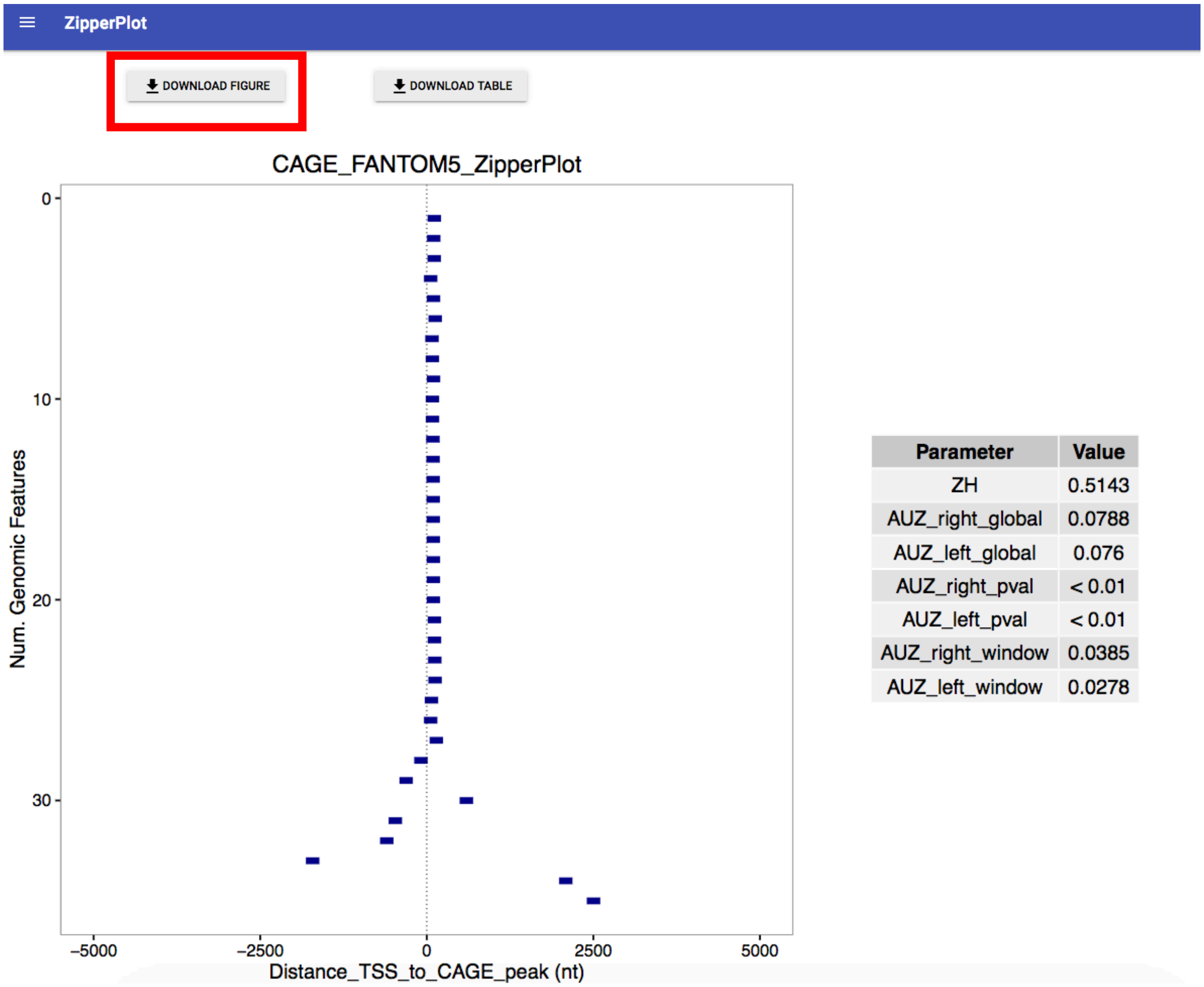


Your job has been submitted.
You will receive an email with the results as soon as the analysis is finished.

Following the link that you receive in your inbox, a **detailed .html report** will appear in your browser.

- Zipper plot and statistics for 35 lncRNAs from Refseq (*downloadable as a .pdf file by clicking “⇓ DOWNLOAD FIGURE”*):

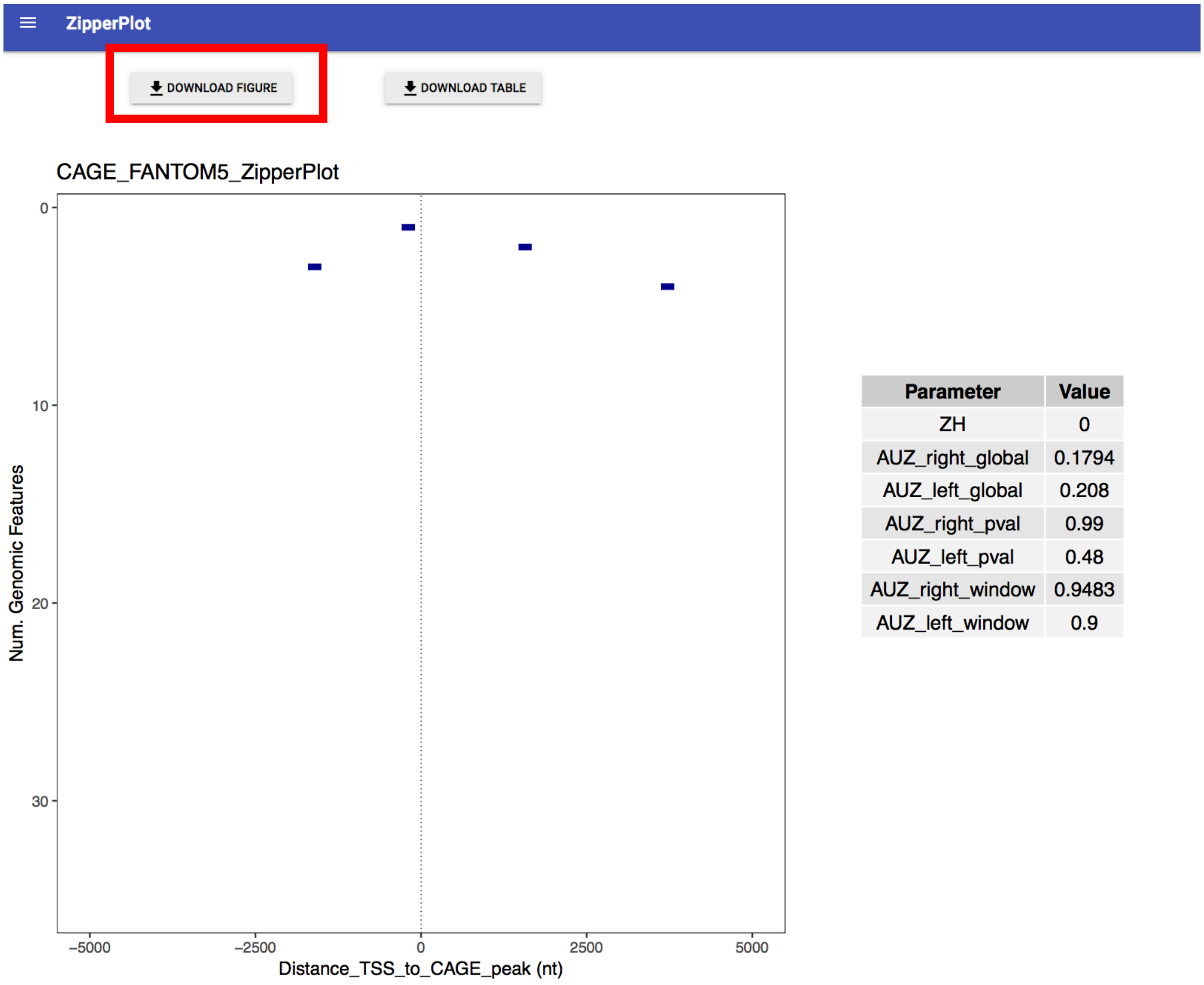
  ZH = 0.5143, AUZ_right_pval < 0.01 & AUZ_left_pval < 0.01. A Zipper Height (ZH) of 0.5143 means that 51.43% of the 35 random TSSs (that is 18 TSSs) overlap with a CAGE-seq peak in at least one sample type. The one-sided AUZ_p-value represents the chance of finding a random Zipper plot with an AUZ_global smaller than or equal to the AUZ_global of the actual use case. As expected, since we have used 35 known lncRNAs from Refseq as input, the p-value (< 0.01) is statistically significant at a significance level of 5%.



Parameter	Value
ZH	0.5143
AUZ_right_global	0.0788
AUZ_left_global	0.076
AUZ_right_pval	< 0.01
AUZ_left_pval	< 0.01
AUZ_right_window	0.0385
AUZ_left_window	0.0278

- Zipper plot and statistics for 35 random locations *(downloadable as a .pdf file by clicking “⇓ DOWNLOAD FIGURE”)*:

  ZH = 0, AUZ_right_pval = 0.99 & AUZ_left_pval = 0.48. A Zipper Height (ZH) of 0 means that none of the 35 random TSSs overlap with a CAGE-seq peak in any sample type (“open zipper”). The one-sided AUZ_p-value represents the chance of finding a random Zipper plot with an AUZ_global smaller than or equal to the AUZ_global of the actual use case. As expected, since we have used random location as input, the p-value is not statistically significant at a significance level of 5%.



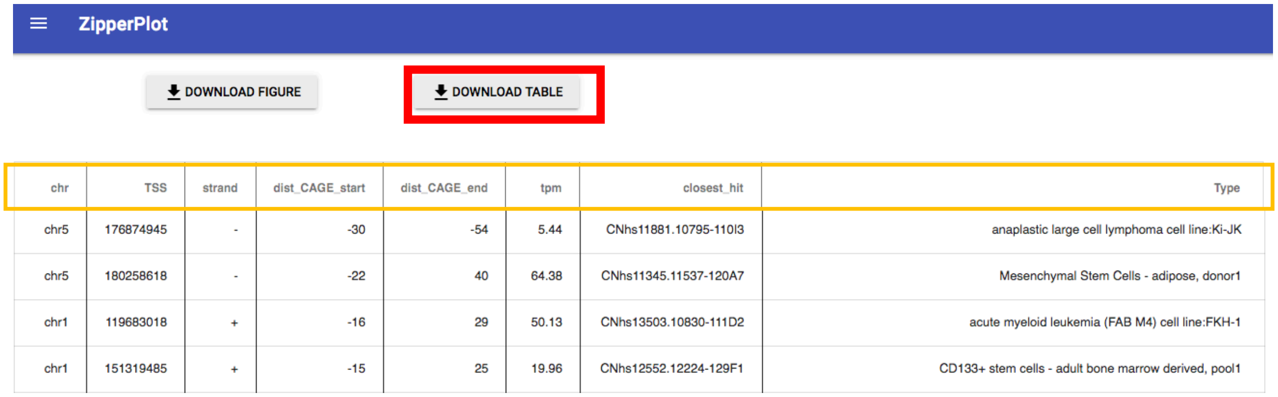
Parameter	Value
ZH	0
AUZ_right_global	0.1794
AUZ_left_global	0.208
AUZ_right_pval	0.99
AUZ_left_pval	0.48
AUZ_right_window	0.9483
AUZ_left_window	0.9

NOTE: The width of the grid for each Zipper plot is determined by the (closest) peak furthest away from the TSS (upstream or downstream) among all retrieved ones. Therefore, the left (upstream of the TSS) and right (downstream of the TSS) grids vary with each different input and thus, AUZ_right_global and AUZ_left_global between different Zipper plots are not directly comparable (see our article for more details).

Nevertheless, AUZ_right_window and AUZ_left_window from two Zipper plots built using the same Zipper width can be directly compared. This is possible because they only depend on the window size choice and are computed using only the retrieved peaks that lie inside the window.

We can observe that AUZ_right_window and AUZ_left_window of the 35 lncRNAs from Refseq are smaller than those of 35 random locations, meaning that the peaks are distributed more closely around the TSSs for the former case.

- The summary table (*downloadable as a plain text file by clicking "⇓ DOWNLOAD TABLE"*) for this combination (FANTOM5 + ALL) contains 8 columns:



chr	TSS	strand	dist_CAGE_start	dist_CAGE_end	tpm	closest_hit	Type
chr5	176874945	-	-30	-54	5.44	CNhs11881.10795-110I3	anaplastic large cell lymphoma cell line:Ki-JK
chr5	180258618	-	-22	40	64.38	CNhs11345.11537-120A7	Mesenchymal Stem Cells - adipose, donor1
chr1	119683018	+	-16	29	50.13	CNhs13503.10830-111D2	acute myeloid leukemia (FAB M4) cell line:FKH-1
chr1	151319485	+	-15	25	19.96	CNhs12552.12224-129F1	CD133+ stem cells - adult bone marrow derived, pool1

i. The first three come directly from the user's input: **chr**, **TSS** & **strand**.
ii. **dist_CAGE_start** and **dist_CAGE_end** are the distances (in nucleotides) from the TSS to the beginning and end of the closest CAGE peak, respectively.
iii. **tpm** (tags per million) is a normalization of the CAGE tag counts defined as the expected count if one million of raw CAGE tags had been extracted.
iv. **closest_hit** and **Type** represent the sample type where the closest CAGE peak has been retrieved (with the original identifier from FANTOM5 data and a brief description, respectively).