WebHERV

This tool allows the analysis of human genome coordinates regarding the presence of neighboring sequences with similarity to human endogenous retroviruses (HERV). Genome coordinates can be derived from gene expression profiling experiments, e.g. DNA micro-array analyses or RNAseq experiments.

Data must be submitted to the server either as files with genome coordinates or as files with micro-array probe set IDs.

A typical analysis will include the following steps:

1. **Data upload and parameter setting**
   a. **Set the search area**
   You can choose between three possibilities. If the orientation of your genome coordinates with respect to the HERV sequence is not relevant at this stage, you can search in 5′ and 3′ direction from the genome coordinate. In this case, you should use the Area setting "overlap". If you are only interested in HERV sequences downstream or upstream form your genome coordinates, you can use the Area setting "upstream" or "downstream", respectively.
   b. **Set the distance**
   This value defines the distance from the genome coordinates to the HERV sequence. You can start the analysis with the default value of 1000 base pairs but you can also change this value to other values. You can analyze multiple distances at once if you input a comma-separated list (e.g. 100,200,300 or 100, 1000, 10000). Negative values are not allowed.
   c. **Set the minimal sequence length**
   You can set the minimal length that an identified HERV-like sequence should have. You can start the analysis with the default value 1 but you can also change this value to other values. Negative values are not allowed.
   d. **Set the maximal HERV e-value**
   You can set the maximal e-value that an identified HERV-like sequence should have. You can start the analysis with the default value 1.0E-1 but you can also change this value. Negative values are not allowed. Note that the current implementation has the internal limit of 1.0E-10, *i.e.* no HERV-like sequences with higher e-values will be considered as hits.

e. **Upload your data**

You can upload multiple files at once. Here you can choose one of two data formats:

i. If you have already genome coordinates, you can upload a file with these coordinates. In this case you have to select the appropriate genome version of your coordinates. Currently, the versions hg18 ("hervs_hg18") and hg19 ("hervs_hg19") are available. Be sure that your genome coordinates are from the same genome version that you have selected. An example file (based on hg19 coordinates) is available on the server.

ii. If you have probe set IDs from micro-array experiments you can use the server to transform the probe set IDs into genome coordinates. In this case you have to select the appropriate platform. Currently, only Affymetrix Human Exon 1.0 arrays can be processed and you can select one of two different genome versions (hg18 and hg19) for the analysis.

2. **Running the analysis**

After uploading of your files, press the submit button and the program will start the analysis. Depending on the size of the data that you uploaded the analysis might take between few seconds and several hours. After finishing of the analysis the results will be displayed.

3. **Reading the results**

The following information will be displayed:

a. **Total elements analyzed**:

This is the number of items that were analyzed. Usually this will be the number of lines (probe set IDs or genome coordinates) in the uploaded document. If lines were not correctly formatted, the number of total elements will be smaller.

b. **Elements with coordinates**:

This is the number of analyzed item that were correctly assigned to genome coordinates. Especially when you use micro-array data, some probe-sets might have no known coordinates in the used reference genome. Therefore, the number of elements with coordinates can be smaller than the number of up-loaded elements (probe set IDs).

c. **With results** and **Distance:**

Here, the number of genome coordinates or probe sets for which HERV-like sequences were detected are displayed. For each distance the corresponding values are displayed.

d. **Hide inoperable elements** and **Hide elements without hits**:

Here you can select whether all uploaded elements (genome coordinates or probe set IDs) or only elements with an identified neighboring HERV-like sequence should be displayed in the table.

e. **Table:** The table contains for all elements the number (num) the ID (probe set ID or genome coordinates) and the number of found HERV-like sequences for the individual differences. You can download the table as CSV file.
f. **Clicking on an individual hit number** will open a new window with more information about the hits. In the head of the new window you find information about the analyzed item. In the displayed table you will find for each hit the genome coordinates and the e-value.

If you use WebHERV please cite:

Kruse K, Nettling M, Wappler N, Emmer A, Kornhuber M, Staege MS, Grosse I (2018). WebHERV: A web-server for the computational investigation of gene expression associated with human endogenous retroviruses. Submitted.

For further information and the code visit:

https://github.com/etsnok/WebHERV