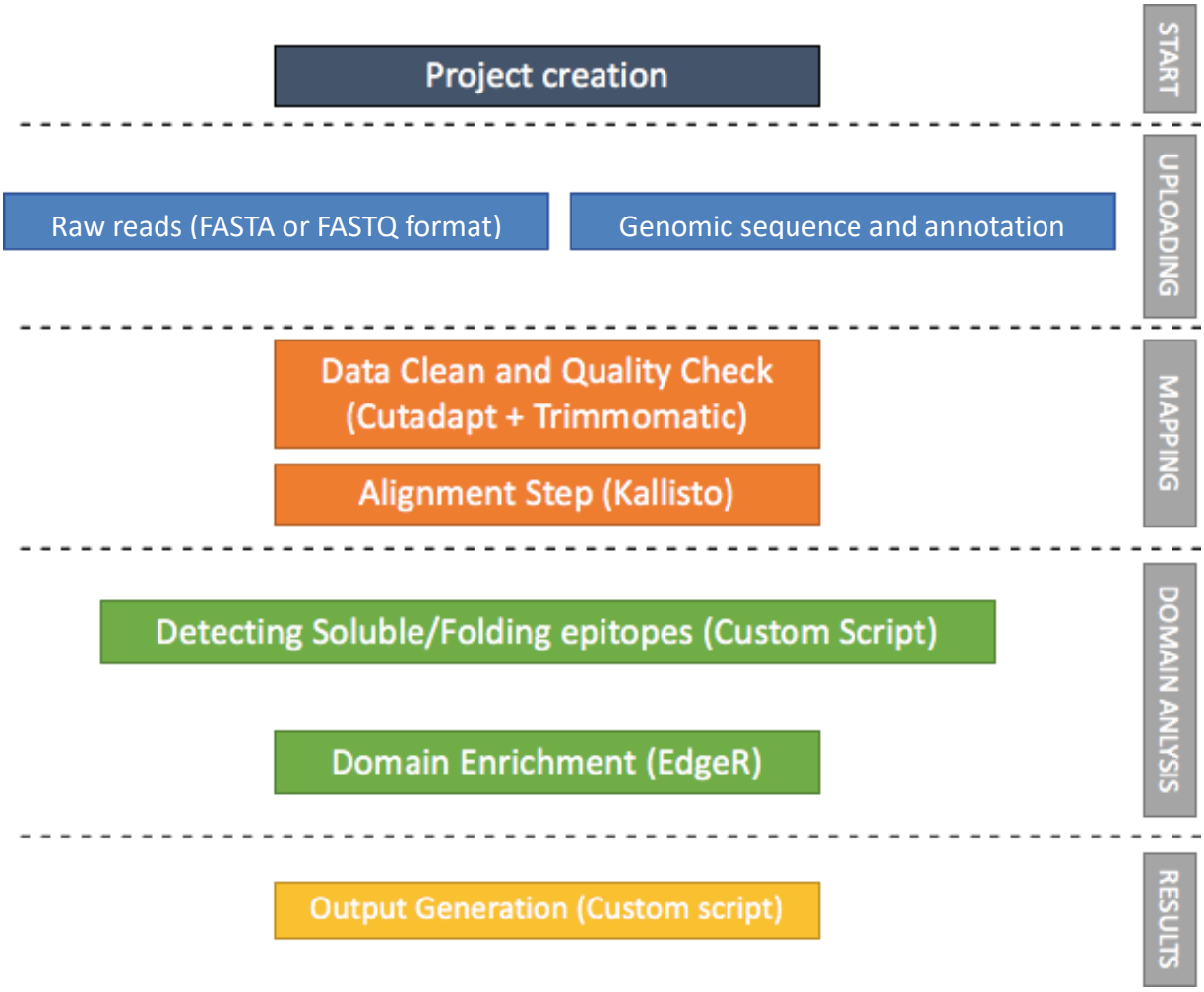
Eukaryote - User Guide

This introductory section provides an overview of **Eukaryote** pipeline drafting and design. The vertical gray rectangles correspond to the website sections.

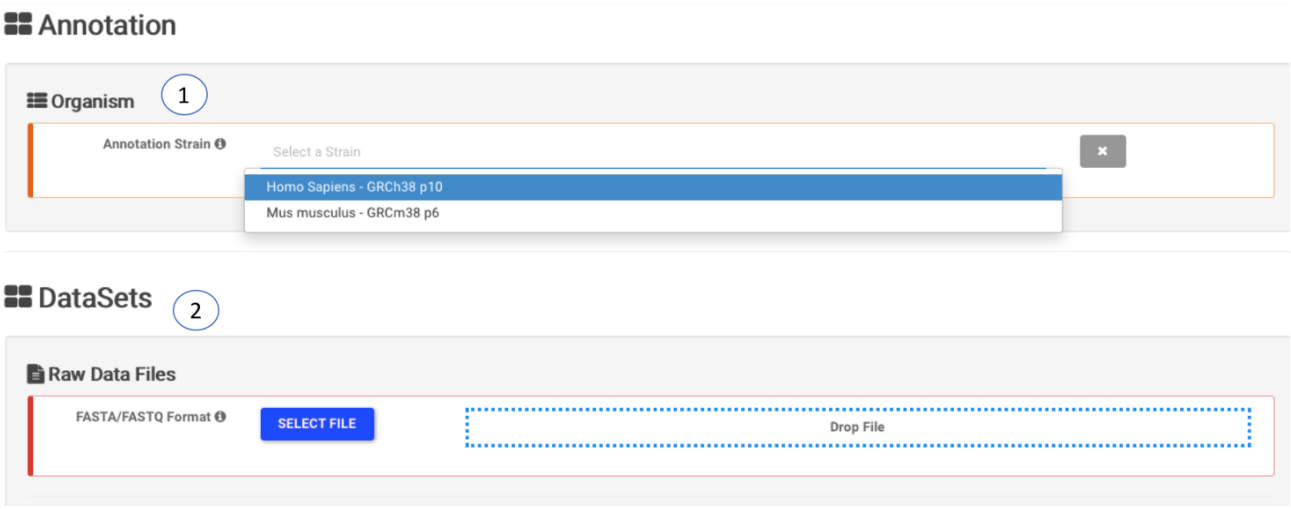


Input Files

Mandatory inputs for **InteractomeSeq - Eukaryote** execution are:

- trascripts and annotation of organism in multiFASTA format (available organisms are listed in selection box) (Organism).
- Raw Data files, FASTA or FASTQ format for query reads are allowed in the input, therefore the web interface additionally allows the submission of compressed files (gz format) to reduce the time of data upload (DataSets).

InteractomeSeq requires at least two datasets. The input datasets must be generated with the same sequencing platform.



Annotation (1)

Pre-loaded nucleotide sequences of all protein-coding transcripts and annotation file are available for Homo Sapiens and Mus Musculus genome. The genome assembly version for Homo Sapiens is the GRCh38 downloaded from NCBI and the annotation was downloaded from GenCode consortium https://www.gencodegenes.org/human/. The genome assembly version for Mus Musculus is the GRCm38 downloaded from NCBI and the annotation was downloaded from GenCode consortium https://www.gencodegenes.org/mouse/.

Raw Data Files (2)

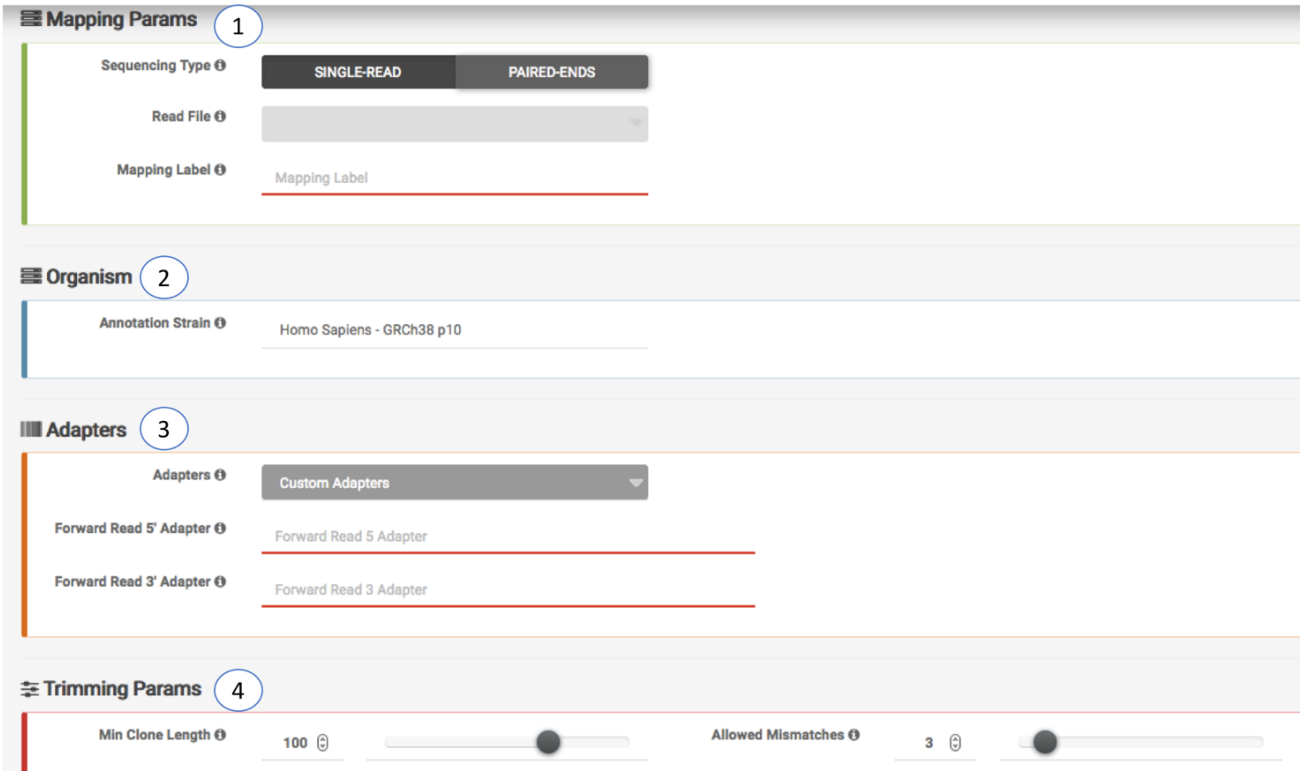
FASTA or FASTQ format are allowed as input, therefore the web interface additionally allows the submission of compressed files (gz format) to reduce the time of data upload (DataSets).

Input form is designed both for loading Single End or Paired-End sequencing. For Paired-End mode, as shown in the screen-shot below, the loading must be repeated both for the forward and reverse dataset.

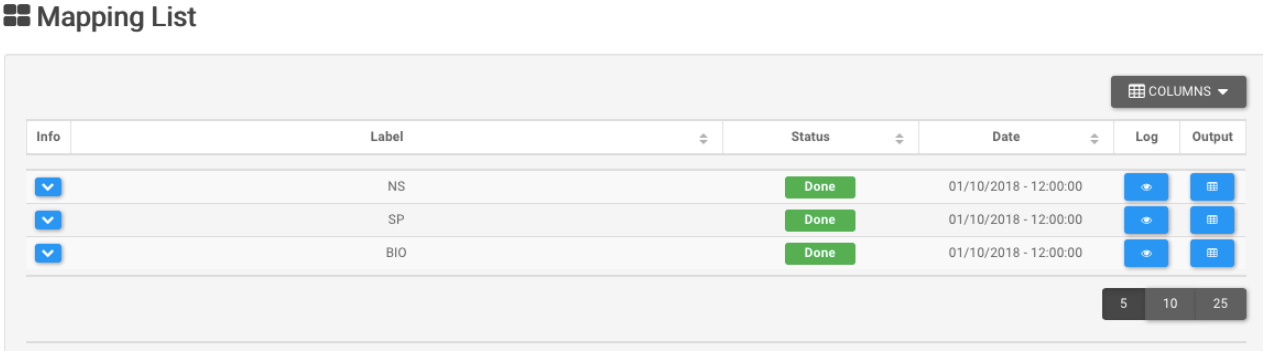
Mapping

By clicking on the botton Mapping  4 sub-sections will appear on the screen.

1. **Mapping Params.** Selection of sequencing type among paired-end reads or single-end.
2. **Organism.** Selected FASTA file that will be used as reference to align the sequences.
3. **Adapters.** Selection of adapters to remove from input sequences. User can select between three options: i) **Autodetec Adapters**; ii) **Custom Adapters**; iii) **Illumina Adapters**
4. **Trimming Params.** Selection of minimum length of sequence and number of mismatch allows; reads below this threshold will be discarded.



Note: Mapping step should be repeated for each input dataset. For each mapping file generated, user can check the log file associated and the Status message.



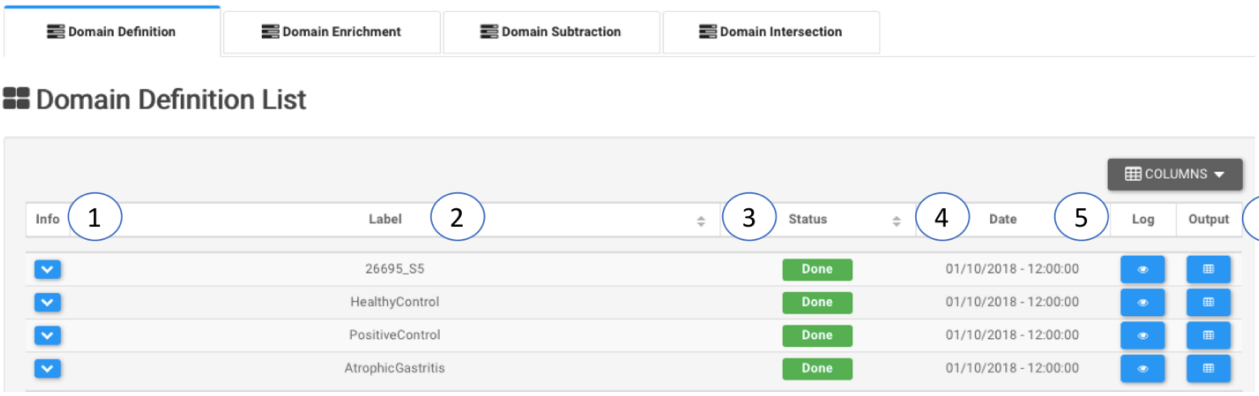
Domain Analysis

Domain Anlysis is composed by four sheets:

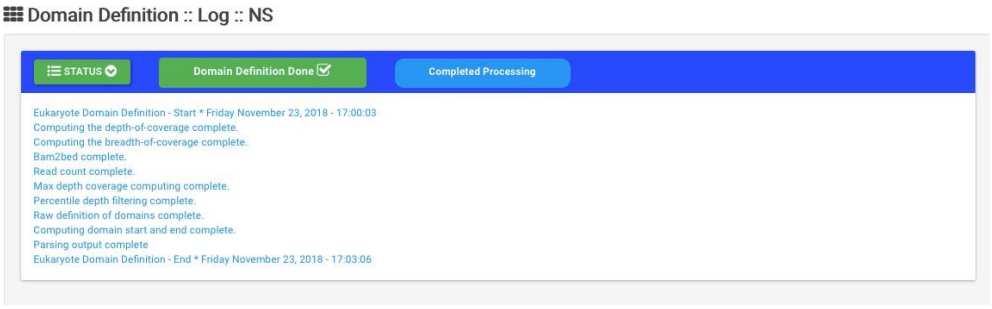
1. **Domain Definition**
2. **Domain Enrichment**

1. **Domain Definition** takes as input the mapping file previously generated.

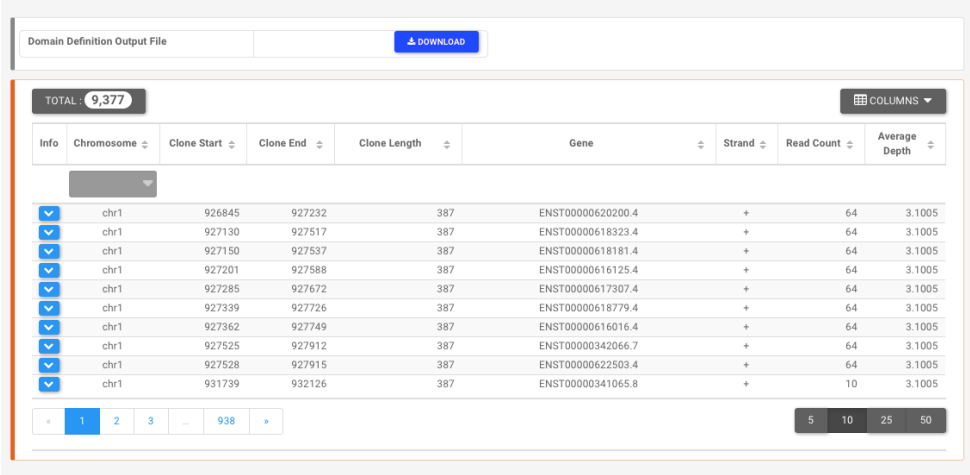
   Domain Definition back-end script uses “bedtool genomomecov” to computes coverage depth at each genome position of the Transcript sequences. Next with a custom script calculates the average depth coverage and only the transcript positions that have a depth coverage greater than the average depth coverage are taken into account. Afterwards the epitopes are defining by combining consecutive bases that have a valid depth coverage. An epitope will be defined by at least 10 consecutive bases. When the computational steps are complete, user can check the status of his analysis.



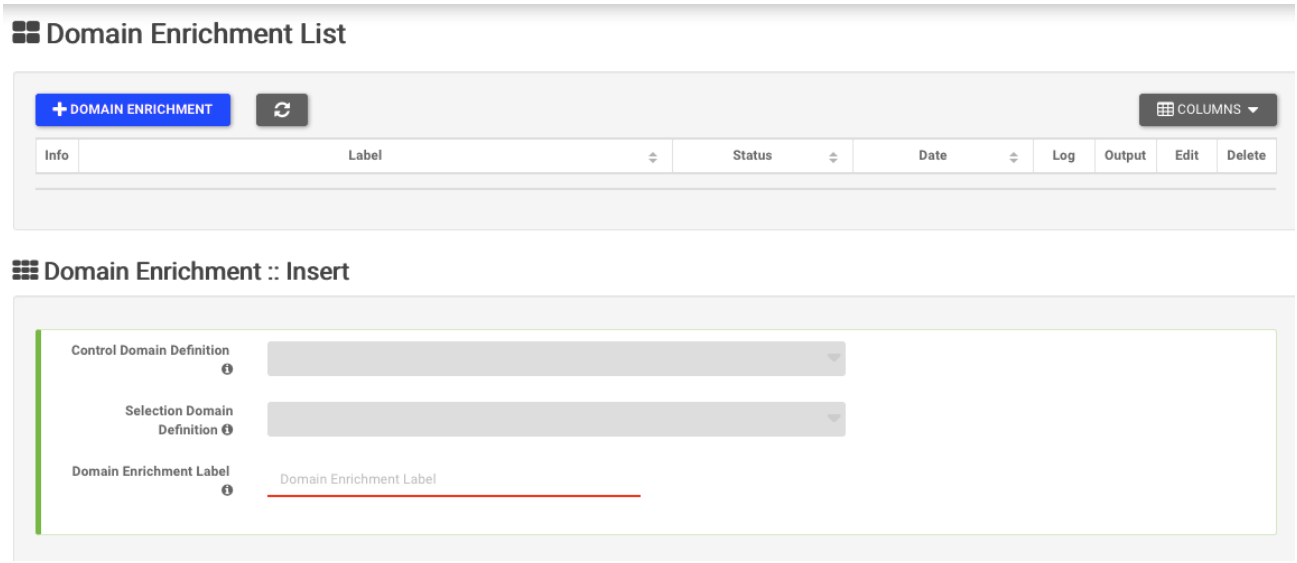
   1. **Info** – Drop-down menu with information of Mapping input file**.**
   2. **Label -** Sample label.
   3. **Status** – When the execution ends successfully, the button turns green, otherwise, it turns red.
   4. **Date** - Day and time of analysis execution
   5. **Log** – Button that hide/open a box with execution log file.



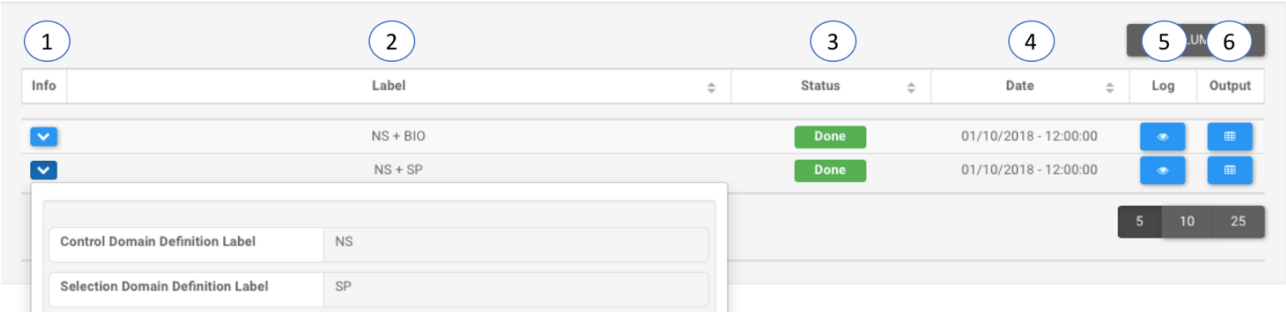
6. **Output** – Hide/show panel with output preview



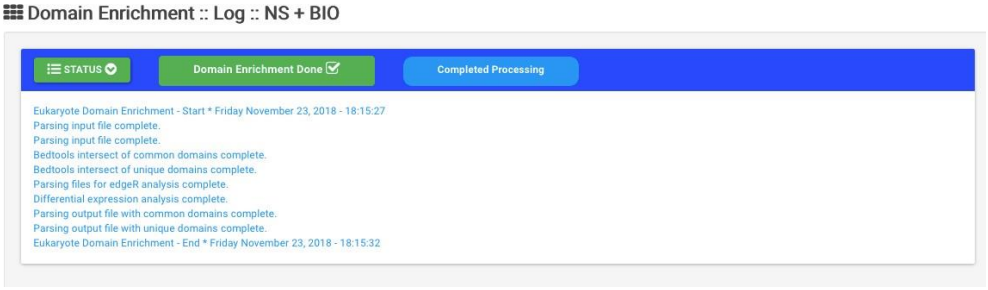
2. **Domain Enrichment** takes as input the Control and Selection output of Domain Definition step.



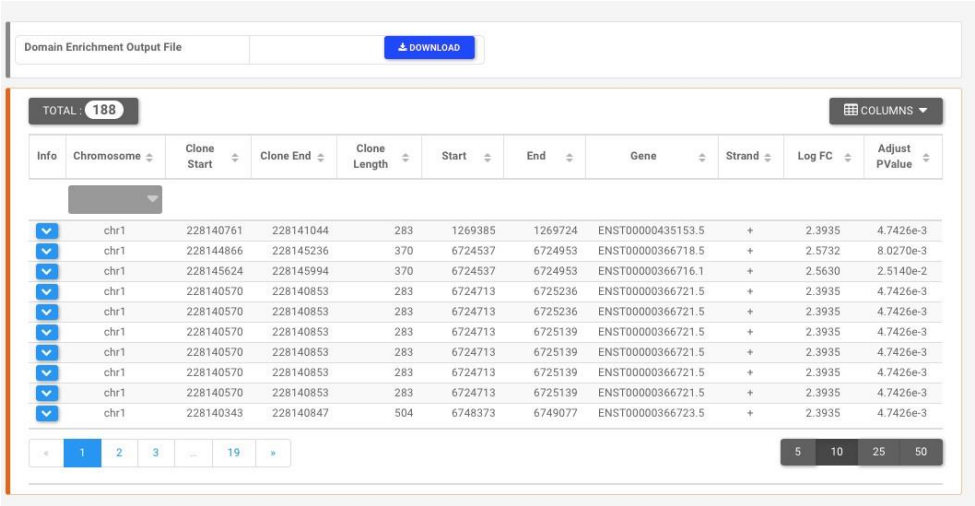
Domain Enrichment back-end script uses “bedtools genomecov” to compute the number of feature (reads) that map inside the epitope regions. Only the common domains between the control selection and the target selection are tested for the statistical analysis. After counting, the epitopes counts are normalized in TPM (transcription per milion) and with R-package EdgeR establish the differentially epitopes of target sample (Target Domain Definition) compare to the Control (Control Domain Definition). When the computational steps are complete, user can check the status of the analysis as shown below.



1. **Info** – Drop-down menu with information of Domain Definition input file**.**
2. **Label -** Sample label.
3. **Status** – When the execution ends successfully, the button turns green, otherwise, it turns red.
4. **Date** - Day and time of analysis execution
5. **Log** – Button that hide/show a box with execution log file.

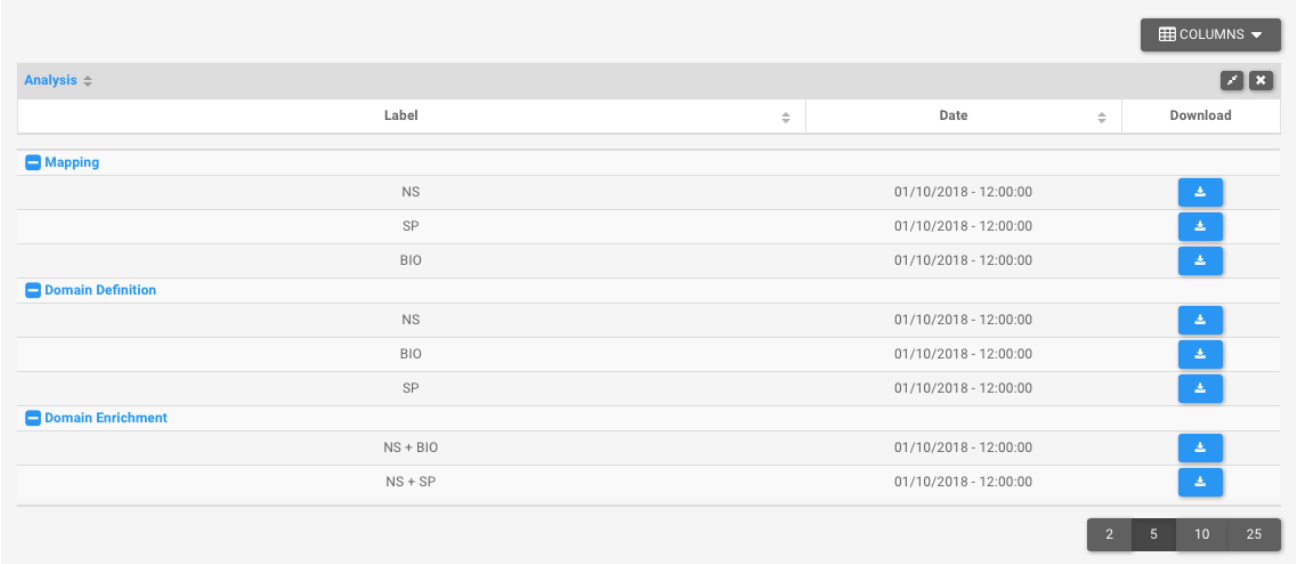


6. **Output** – Hide/show panel with output preview



Results

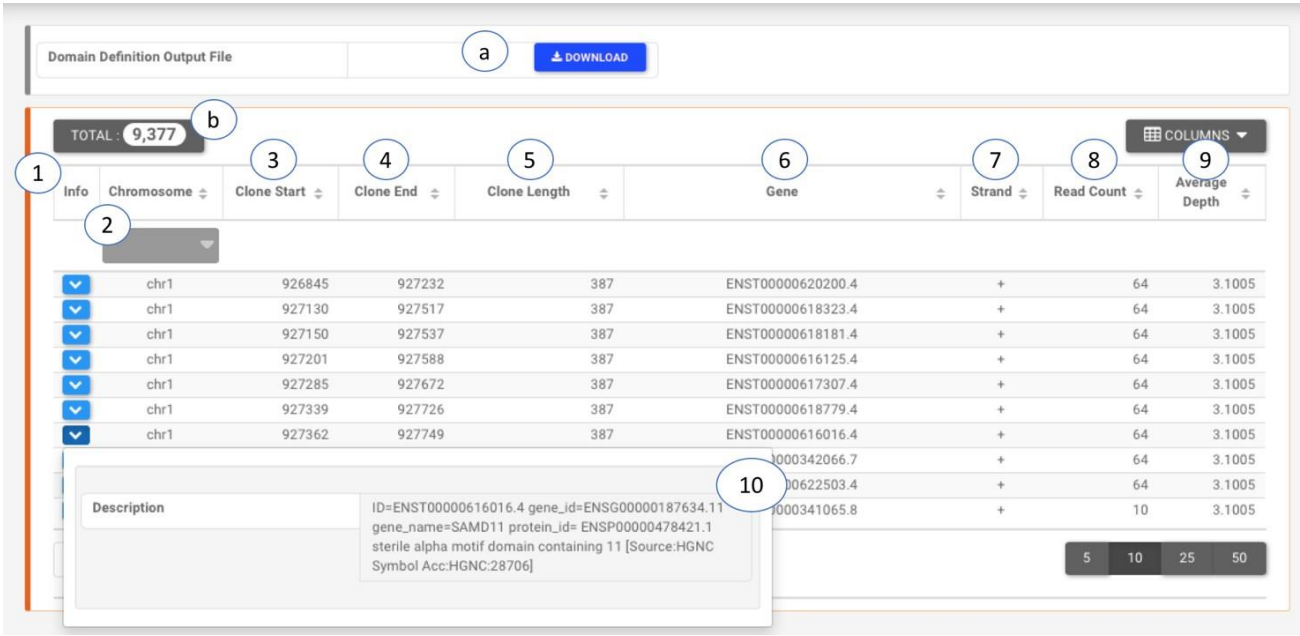
In this section user will find a summary of the output files generated by the execution. Download of the data can be activated clicking on the corresponding button. Data are in zip-compressed archives and can be opened and edited as tab-separated files.



Outputs

- **Mapping**. The alignment results generated by Kallisto is stored in pseudobam file format, user can download this file by clicking for a detailed explanation of this format please refer to this link : https://pachterlab.github.io/kallisto/pseudobam.html.

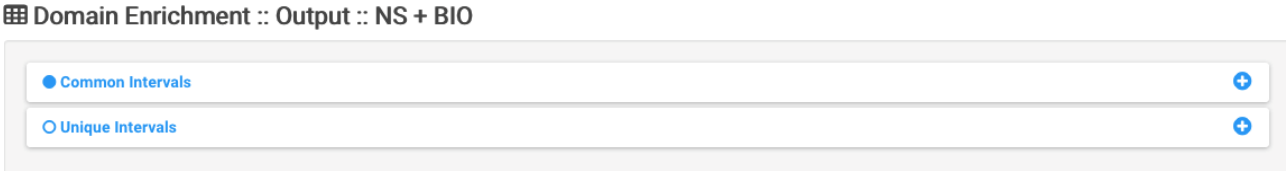
- **Domain Definition**. The domain definition step provides output in tabular format. In this file all soluble domains/epitopes detected are listed and the transcript associated information are provided. As shown in the figure below, download of the data can be activated by clicking on the corresponding button (a) and in box (b) user can quickly check how many domains are detected.



Tabular data report the following fields:

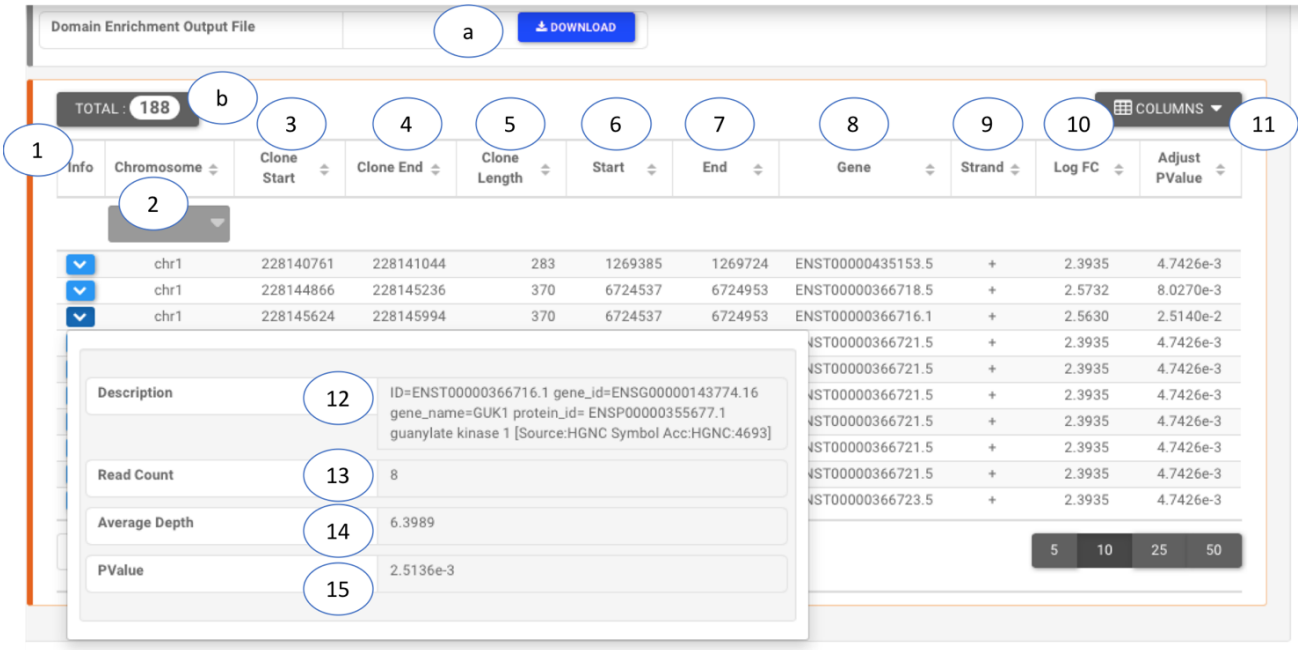
1. Info – Show/hide Description information.
2. Chromosome – Chromosome number, by clicking on the drop-down menu user can select the chromosome of interest.
3. Clone Start – Transcript start of soluble folding domain/epitope
4. Clone End - Transcript end of soluble folding domain/epitope
5. Clone Length - Domain/epitope detected length
6. Gene – Transcript Ensembl ID
7. Strand - Strand associated with transcript annotation
8. Read Count – Number of reads that align inside the domain/epitopes region
9. Average Depth – Indicate th sum of the mapped read depths at each domains base position, divided by the number of known bases in the domains.
10. Description – Transcript annotation

- **Domain Enrichment**. The domain enrichment step provides two output in tabular format.



Common Intervals are the domains/epitopes that have same transcript position both for Control Selection and Target Selection, these domains/epitopes are statistically tested. As

shown in the figure below, download of the data can be activated by clicking on the corresponding button (a) and in box (b) user can quickly check how many domains are detected.

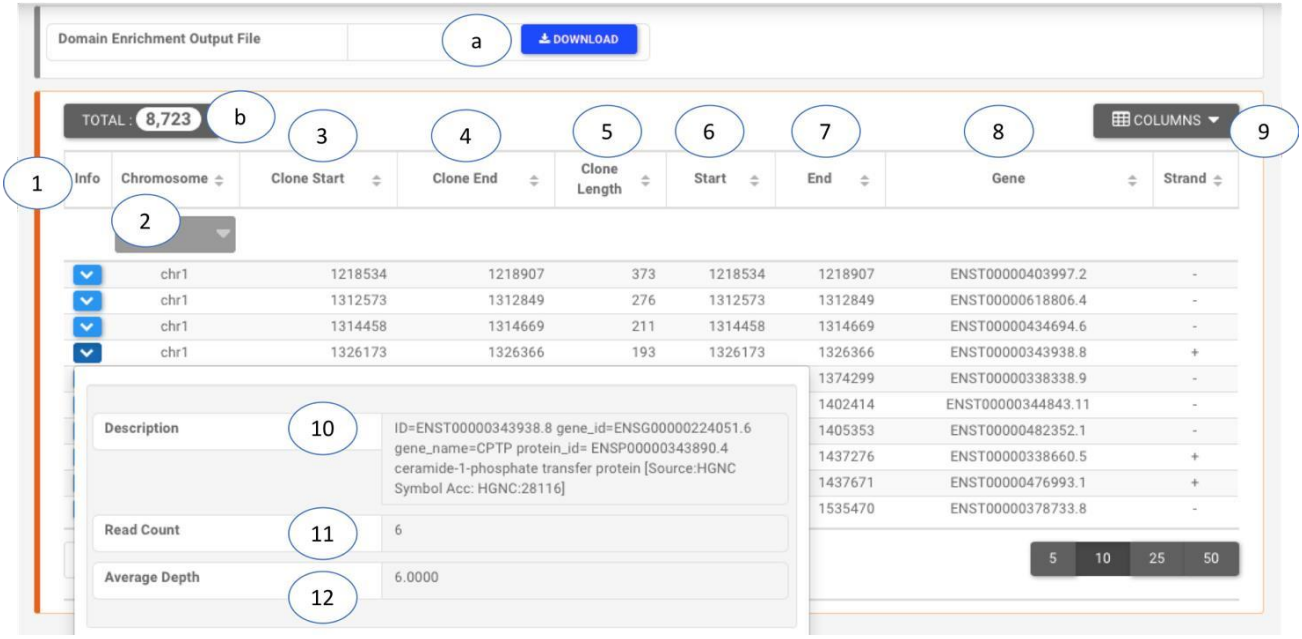


Tabular data report the following fields:

1. Info - Show/hide Description-Read Count-Average Depth-Pvalue information.
2. Chromosome - Chromosome number, by clicking on the drop-down menu user can select the chromosome of interest.
3. Clone Start - Transcript start of soluble folding domain/epitope
4. Clone End - Transcript end of soluble folding domain/epitope
5. Clone Length - Domain/epitope detected length
6. Start – Starting coordinate of the gene associated with the domain/epitope on the chromosome.
7. End - Ending coordinate of the gene associated with the domain/epitope on the chromosome.
8. Gene – Transcript Ensembl ID
9. Strand – Strand associated with transcript annotation
10. LogFC – log2 fold change estimation
11. Adjust Pvalue – pvalue adjusted for FDR
12. Description – Gene associated with the domain/epitope annotation
13. Read Count - Number of reads that align inside the domain/epitopes region
14. Average Depth - Indicate th sum of the mapped read depths at each domains base position, divided by the number of known bases in the domains.

15. Pvalue – probability value

Unique Intervals are the domains/epitopes that results specific of Target Selection. As shown in the figure below, download of the data can be activated by clicking on the corresponding button (a) and in box (b) user can quickly check how many domains are detected.



Tabular data report the following fields:

1. Info - Show/hide Description-Read Count-Average Depth-Pvalue information.
2. Chromosome - Chromosome number, by clicking on the drop-down menu user can select the chromosome of interest.
3. Clone Start - Transcript start of soluble folding domain/epitope
4. Clone End - Transcript end of soluble folding domain/epitope
5. Clone Length - Domain/epitope detected length
6. Start – Starting coordinate of the gene associated with the domain/epitope on the chromosome.
7. End - Ending coordinate of the gene associated with the domain/epitope on the chromosome.
8. Gene – Transcript Ensembl ID
9. Strand – Strand associated with transcript annotation
10. Description – Gene associated with the domain/epitope annotation
11. Read Count - Number of reads that align inside the domain/epitopes region
12. Average Depth - Indicate th sum of the mapped read depths at each domains base position, divided by the number of known bases in the domains.