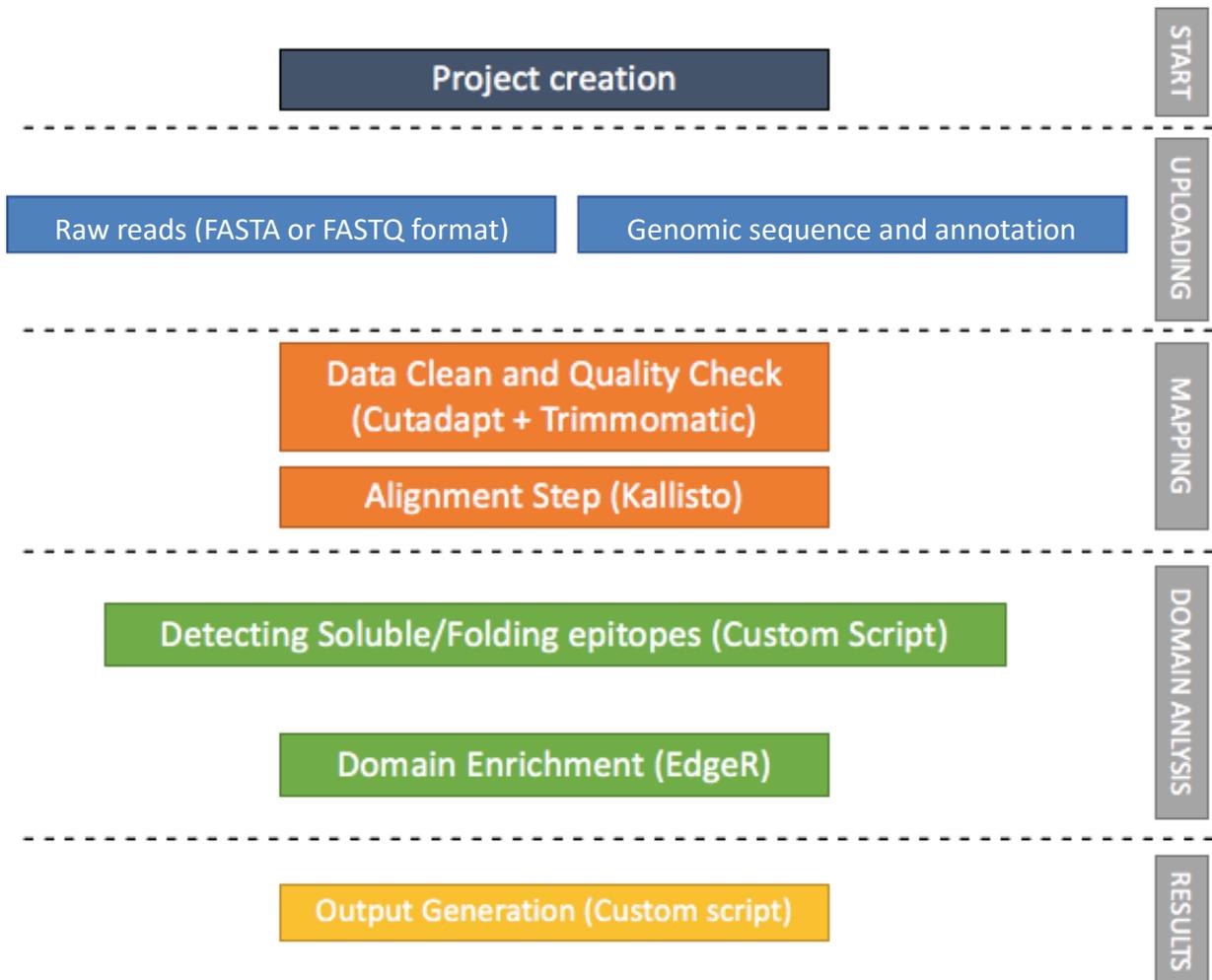


Eukaryote - User Guide

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



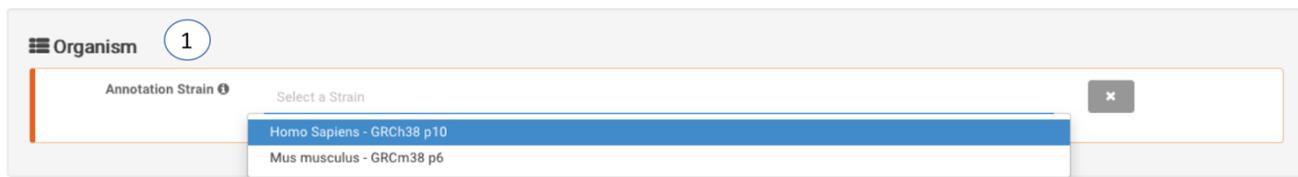
Input Files

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

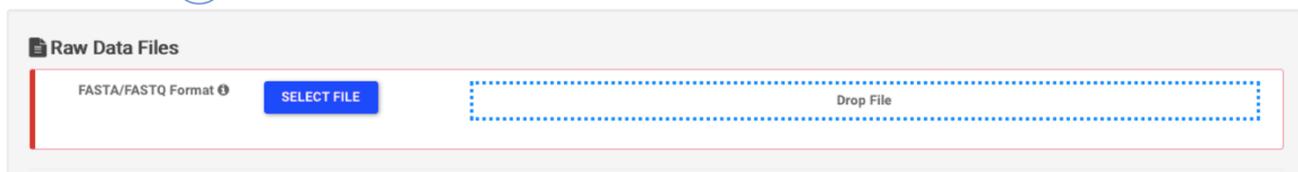
- transcripts and annotation of organism in multiFASTA format (available organisms are listed in selection box) (Organism).
- Raw Data files, FASTA or FASTQ format for query reads are allowed in the input, therefore the web interface additionally allows the submission of compressed files (gz format)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



DataSets



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)

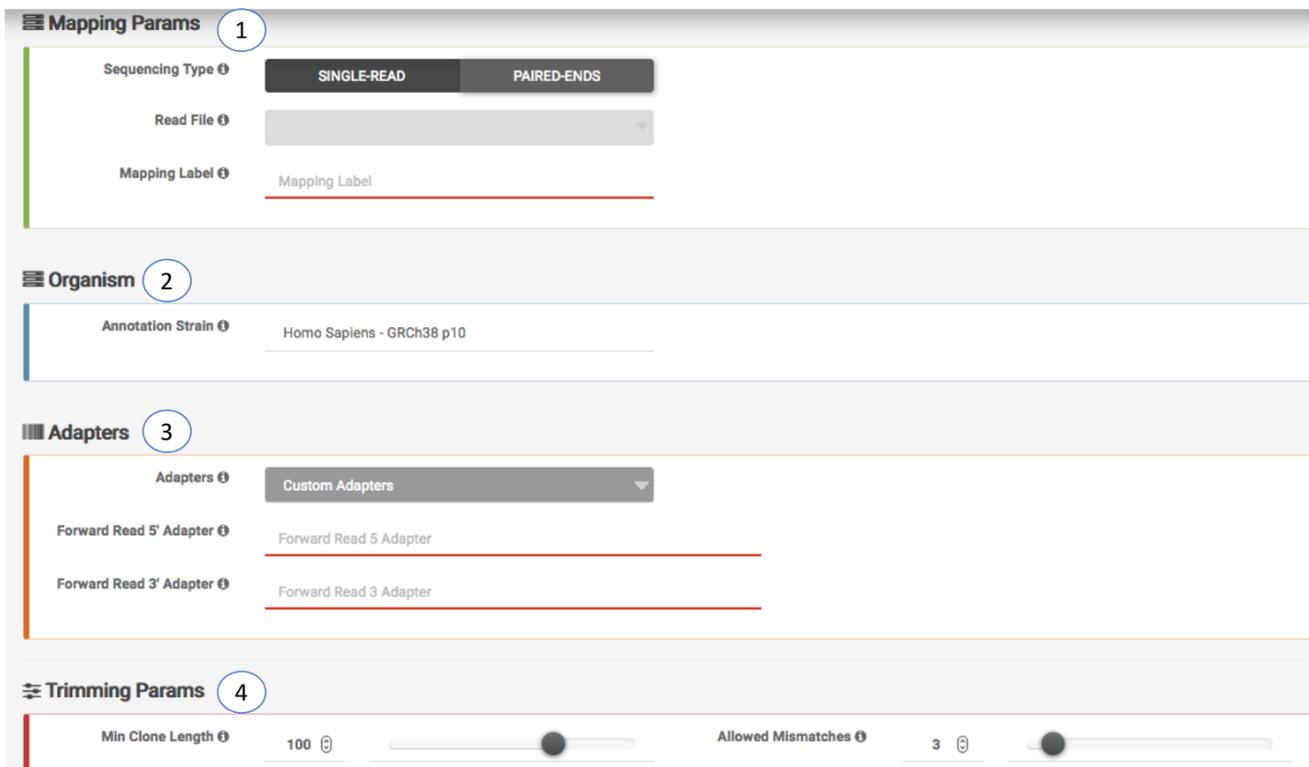
FASTA or FASTQ format are allowed as input, therefore the web interface additionally allows the submission of compressed files (gz format) to reduce the time of data upload (DataSets).

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 button 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.

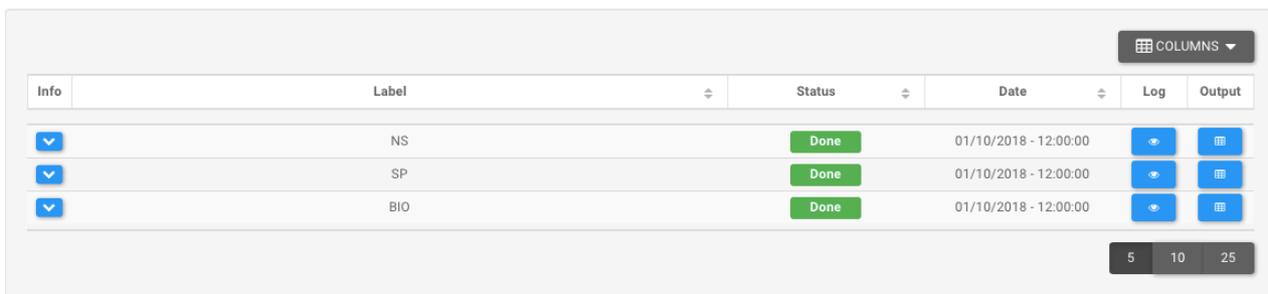


The screenshot displays the Mapping configuration interface, divided into four sections:

- Mapping Params (1):** Sequencing Type (SINGLE-READ, PAIRED-ENDS), Read File (dropdown), Mapping Label (text input).
- Organism (2):** Annotation Strain (Homo Sapiens - GRCh38 p10).
- Adapters (3):** Adapters (Custom Adapters), Forward Read 5' Adapter (text input), Forward Read 3' Adapter (text input).
- Trimming Params (4):** Min Clone Length (100), Allowed Mismatches (3).

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.

Mapping List



Info	Label	Status	Date	Log	Output
	NS	Done	01/10/2018 - 12:00:00		
	SP	Done	01/10/2018 - 12:00:00		
	BIO	Done	01/10/2018 - 12:00:00		

5 10 25

Domain Analysis

Domain Analysis 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 compute 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 defined 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.

The screenshot shows the 'Domain Definition List' interface. At the top, there are four tabs: 'Domain Definition' (selected), 'Domain Enrichment', 'Domain Subtraction', and 'Domain Intersection'. Below the tabs is a table with the following columns: 'Info' (1), 'Label' (2), 'Status' (3), 'Date' (4), 'Log' (5), and 'Output' (6). The table contains four rows of data, each with a dropdown arrow on the left and a 'Done' status button. The data rows are: 26695_S5, HealthyControl, PositiveControl, and AtrophicGastritis, all with a date of 01/10/2018 - 12:00:00. The 'Log' and 'Output' columns contain blue buttons with arrows and document icons respectively. A 'COLUMNS' dropdown menu is visible in the top right corner of the table area.

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.

Domain Definition :: Log :: NS

STATUS
Domain Definition Done
Completed Processing

Eukaryote Domain Definition - Start * Friday November 23, 2018 - 17:00:03
 Computing the depth-of-coverage complete.
 Computing the breadth-of-coverage complete.
 Bam2bed complete.
 Read count complete.
 Max depth coverage computing complete.
 Percentile depth filtering complete.
 Raw definition of domains complete.
 Computing domain start and end complete.
 Parsing output complete
 Eukaryote Domain Definition - End * Friday November 23, 2018 - 17:03:06

6. Output – Hide/show panel with output preview

Domain Definition Output File
DOWNLOAD

TOTAL: 9,377 COLUMNS

Info	Chromosome	Clone Start	Clone End	Clone Length	Gene	Strand	Read Count	Average Depth
▼	chr1	926845	927232	387	ENST00000620200.4	+	64	3.1005
▼	chr1	927130	927517	387	ENST00000618323.4	+	64	3.1005
▼	chr1	927150	927537	387	ENST00000618181.4	+	64	3.1005
▼	chr1	927201	927588	387	ENST00000616125.4	+	64	3.1005
▼	chr1	927285	927672	387	ENST00000617307.4	+	64	3.1005
▼	chr1	927339	927726	387	ENST00000616779.4	+	64	3.1005
▼	chr1	927362	927749	387	ENST00000616016.4	+	64	3.1005
▼	chr1	927525	927912	387	ENST00000342066.7	+	64	3.1005
▼	chr1	927528	927915	387	ENST00000622503.4	+	64	3.1005
▼	chr1	931739	932126	387	ENST00000341065.8	+	10	3.1005

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

Domain Enrichment List

+ DOMAIN ENRICHMENT
REFRESH
COLUMNS

Info	Label	Status	Date	Log	Output	Edit	Delete

Domain Enrichment :: Insert

Control Domain Definition ▼

Selection Domain Definition ▼

Domain Enrichment Label Domain Enrichment Label

Domain Enrichment back-end script uses “bedtools genomcov” 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.

Info	Label	Status	Date	Log	Output
▼	NS + BIO	Done	01/10/2018 - 12:00:00	🔍	📄
▼	NS + SP	Done	01/10/2018 - 12:00:00	🔍	📄

Control Domain Definition Label: NS

Selection Domain Definition Label: SP

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.

📄 Domain Enrichment :: Log :: NS + BIO

STATUS
Domain Enrichment Done
Completed Processing

Eukaryote Domain Enrichment - Start * Friday November 23, 2018 - 18:15:27
 Parsing input file complete.
 Parsing input file complete.
 Bedtools intersect of common domains complete.
 Bedtools intersect of unique domains complete.
 Parsing files for edgeR analysis complete.
 Differential expression analysis complete.
 Parsing output file with common domains complete.
 Parsing output file with unique domains complete.
 Eukaryote Domain Enrichment - End * Friday November 23, 2018 - 18:15:32

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

Domain Enrichment Output File 📄 DOWNLOAD

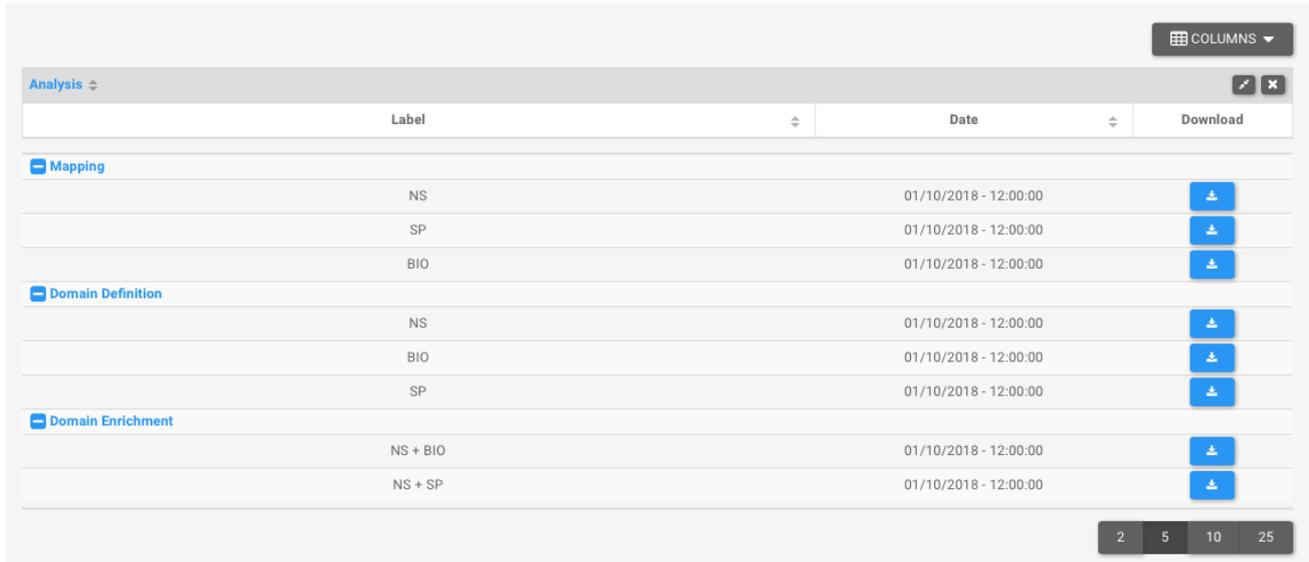
TOTAL: 188 COLUMNS

Info	Chromosome	Clone Start	Clone End	Clone Length	Start	End	Gene	Strand	Log FC	Adjust PValue
▼	chr1	228140761	228141044	283	1269385	1269724	ENST00000435153.5	+	2.3935	4.7426e-3
▼	chr1	228144866	228145236	370	6724537	6724953	ENST00000366718.5	+	2.5732	8.0270e-3
▼	chr1	228145624	228145994	370	6724537	6724953	ENST00000366716.1	+	2.5630	2.5140e-2
▼	chr1	228140570	228140853	283	6724713	6725236	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140570	228140853	283	6724713	6725236	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140570	228140853	283	6724713	6725139	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140570	228140853	283	6724713	6725139	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140570	228140853	283	6724713	6725139	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140570	228140853	283	6724713	6725139	ENST00000366721.5	+	2.3935	4.7426e-3
▼	chr1	228140343	228140847	504	6748373	6749077	ENST00000366723.5	+	2.3935	4.7426e-3

5 10 25 50

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.



Analysis			COLUMNS
Label	Date	Download	
Mapping			
NS	01/10/2018 - 12:00:00	Download	
SP	01/10/2018 - 12:00:00	Download	
BIO	01/10/2018 - 12:00:00	Download	
Domain Definition			
NS	01/10/2018 - 12:00:00	Download	
BIO	01/10/2018 - 12:00:00	Download	
SP	01/10/2018 - 12:00:00	Download	
Domain Enrichment			
NS + BIO	01/10/2018 - 12:00:00	Download	
NS + SP	01/10/2018 - 12:00:00	Download	

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.

Domain Definition Output File a ↓ DOWNLOAD

TOTAL: 9,377 b COLUMNS ▾

1	2	3	4	5	6	7	8	9
Info	Chromosome	Clone Start	Clone End	Clone Length	Gene	Strand	Read Count	Average Depth
▼	chr1	926845	927232	387	ENST00000620200.4	+	64	3.1005
▼	chr1	927130	927517	387	ENST00000618323.4	+	64	3.1005
▼	chr1	927150	927537	387	ENST00000618181.4	+	64	3.1005
▼	chr1	927201	927588	387	ENST00000616125.4	+	64	3.1005
▼	chr1	927285	927672	387	ENST00000617307.4	+	64	3.1005
▼	chr1	927339	927726	387	ENST00000618779.4	+	64	3.1005
▼	chr1	927362	927749	387	ENST00000616016.4	+	64	3.1005
					000342066.7	+	64	3.1005
					0622503.4	+	64	3.1005
					000341065.8	+	10	3.1005

10

Description: ID=ENST00000616016.4 gene_id=ENSG00000187634.11 gene_name=SAMD11 protein_id= ENSP00000478421.1 sterile alpha motif domain containing 11 [Source:HGNC Symbol Acc:HGNC:28706]

5 10 25 50

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.

Domain Enrichment :: Output :: NS + BIO

Common Intervals	+
Unique Intervals	+

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.

The screenshot displays a web interface for domain enrichment analysis. At the top, there is a 'Domain Enrichment Output File' section with a 'DOWNLOAD' button labeled 'a'. Below this, a table shows the results of the analysis. The table has columns for 'Info', 'Chromosome', 'Clone Start', 'Clone End', 'Clone Length', 'Start', 'End', 'Gene', 'Strand', 'Log FC', and 'Adjust PValue'. A 'TOTAL: 188' indicator is present. A 'COLUMNS' dropdown menu is visible on the right. A detailed view of a row is shown in a pop-up box labeled 'b', displaying fields like 'Description', 'Read Count', 'Average Depth', and 'PValue'. The pop-up box also shows a 'COLUMNS' dropdown menu.

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.

The screenshot displays a web interface for domain enrichment analysis. At the top, there is a 'Domain Enrichment Output File' section with a 'DOWNLOAD' button (labeled 'a'). Below this, a summary bar shows 'TOTAL: 8,723' (labeled 'b'). The main table has columns: Info (labeled '1'), Chromosome (labeled '2'), Clone Start (labeled '3'), Clone End (labeled '4'), Clone Length (labeled '5'), Start (labeled '6'), End (labeled '7'), Gene (labeled '8'), and Strand (labeled '9'). A detailed view of a domain is shown below the table, with fields for Description (labeled '10'), Read Count (labeled '11'), and Average Depth (labeled '12').

Info	Chromosome	Clone Start	Clone End	Clone Length	Start	End	Gene	Strand
▼	chr1	1218534	1218907	373	1218534	1218907	ENST00000403997.2	-
▼	chr1	1312573	1312849	276	1312573	1312849	ENST00000618806.4	-
▼	chr1	1314458	1314669	211	1314458	1314669	ENST00000434694.6	-
▼	chr1	1326173	1326366	193	1326173	1326366	ENST00000343938.8	+
							ENST00000338338.9	-
							ENST00000344843.11	-
							ENST00000482352.1	-
							ENST00000338660.5	+
							ENST00000476993.1	+
							ENST00000378733.8	-

Read Count: 6
Average Depth: 6.0000

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.