# Prokaryote – User Guide

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



# Input Files

Mandatory inputs for InteractomeSeq - Prokaryote execution are:

- genome reference file in FASTA format (either a custom annotation file or one selected from the drop-down menu) (Organism).
- a genome annotation (either a custom annotation file or one selected from the drop-down menu) (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 the user to upload at least two datasets. The input datasets must be generated with the same sequencing platform.

### **Raw Data Files**

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.

E	Raw Data Files
L	File
L	HP_genomic_26695_R1.fastq.gz
	HP_genomic_26695_R2.fastq.gz

### **Genome Sequence and Annotation**

Genome annotation can be provided either as by selecting one of the annotations pre-loaded in the internal **InteractomeSeq** database (derived from bacterial genome annotations by the National Center for Biotechnology Information of the National Institute of Health) (NCBI Genome Annotation) (1) or a custom file (Custom Genome Annotation tab) (2).



### **NCBI** Genome Annotation

Users can select one of the pre-loaded NCBI bacteria genome annotations (NCBI ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/, last updated on November 2019), containing complete information about 15593 bacterial strains (Complete Genome).

In order to select the proper strain from the drop-down menu, the user has to type just 3 or more characters of the strain name to activate the automatic search in the database.

Clicking on the "Preview" button allows quick check of the Genome Annotation selected.

Note: check that the "Gene Name" field contains the same values as the GED file, otherwise **InteractomeSeq** will not process the data.

Annotation	Strain <b>1</b> HELIO	OBACTER PYLOR	l 26695 - NC_000915		• X
			PREV	/IEW @	
# LINES : 1,46	9				
Chromosome	Start	End	Strand Locus Tag	Gene Name	Description
NO 0000151	217	633	- HP0001	nusB	transcription antitermination protein NusB
NC_000915.1		1105	- HP0002	ribH	6,7-dimethyl-8-ribityllumazine synthase
NC_000915.1 NC_000915.1	635	1105			
NC_000915.1 NC_000915.1 NC_000915.1	635 1115	1945	- HP0003	-	2-denydro-3-deoxyphosphooctonate aldolase
NC_000915.1 NC_000915.1 NC_000915.1 NC_000915.1	635 1115 1932	1945	- HP0003 - HP0004	-	2-dehydro-3-deoxyphosphooctonate aldolase carbonic anhydrase IcfA

## **Custom Genome Sequence and Annotation**

Alternatively, users can provide their own preferred genome sequence and annotation, provided the custom annotation file fulfills the specific requirements listed below.

Note: organisms with more than one chromosome or containing plasmids must be analyzed at the same time. Thus, when building custom genome annotations, users have to provide two files, one for nucleotide sequence (multi fasta) and one with gene annotation.

- Users can submit a custom nucleotide sequence file in one of the recognized file formats:
  - o Fasta
  - o Multi-Fasta
- Users can submit a custom gene annotation file in one of the recognized file formats:
  - o BED
  - o GFF
  - CSV/TSV
- Users can indicate (optionally):
  - Header Line and Rows (This file can optionally have a custom number of header lines at the top).
  - Column separator can be set selecting from TAB, SPACE, comma (",") or semicolon (";").

#### **Annotation**

E Annotation	Lustom Ann	otation					
Reference	ce						
	Reference File	SELECT FILE		Drop	File		
■ Annotat	ion						
,	Annotation File 0	SELECT FILE		Drop	File		
	File Format <b>O</b>			Header Line <b>O</b>			
Colu	mn Separator	TAB	CSV/TSV	# Header Rows ①	1 3	-	

### **BED Format**

Gene annotation in Browser Extensible Data format (BED) must be provided according to the UCSC standard, with at least 6 columns (also called "BED6" format) (https://genome.ucsc.edu/FAQ/FAQformat#format1).

BED file fields must contain the following information:

- 1. chrom The name of the chromosome or scaffold.
- 2. chromStart The starting position of the gene in the chromosome or scaffold. The first base in a chromosome is numbered 1.
- 3. chromEnd The ending position of the gene in the chromosome or scaffold.
- 4. name Defines the gene name.
- 5. score not used. Please that the field cannot be empty, but must contain a value (set it to "0" or "." if the BED file has no score associated to the gene).
- 6. strand Defines the transcription strand for each gene. Either "+" or "-".

An example of a valid Genome annotation BED6 file for use in **InteractomeSeq** is reported below.

locus	chr_start	chr_end	GeneName	score	strand
NC_007650	1	1188	BTH_110001	0	+
NC_007650	1281	2324	BTH_110002	0	÷.,
NC_007650	2490	2870	BTH_110003	0	<b>4</b> 0
NC_007650	2950	3558	BTH_110004	0	-)
NC_007650	3726	4925	BTH_110005	0	+
NC_007650	4938	5969	BTH_110006	0	+
NC 007650	6192	6740	<b>BTH II0007</b>	0	-0

### **GFF Format**

Alternatively, annotations can be provided also in the General Feature Format (GFF) format, which has nine required fields that must be tab-separated. Please, refer to:

https://genome.ucsc.edu/FAQ/FAQformat#format3 and http://gmod.org/wiki/GFF3 for the complete description of this format.

GFF file fields must contain the following information:

- 1. seqname The name of the chromosome or scaffold.
- 2. source not used. Please note that the field cannot be empty, but must contain a value (set it to "." if the GFF file has no value associated to this field).
- 3. feature The feature type (e.g.: "gene", "CDS", "tRNA", etc.).
- 4. start The starting position of the feature in the chromosome or scaffold. The first base is numbered 1.
- 5. end The ending position of the feature in the chromosome or scaffold.
- 6. score not used. Please note that the field cannot be empty, but must contain a value (set it to "0" or "." if the GFF file has no score associated to the feature).
- 7. strand Valid entries are "+", "-", or "." (for not available/not relevant).
- 8. frame not used. If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be ".". Please note that the field cannot be empty, but must contain a value.
- 9. attributes A list of feature attributes in the format tag = value. Multiple tag = value pairs are separated by semicolons.

**Note**: Please check that records defining gene features have a pair: "locus\_tag = LOCUS\_NAME" in the "attributes" section (column 9), since InteractomeSeq infers the gene name from this tag-value pair.

**Note**: Valid GFF files are those downloaded from NCBI (https://www.ncbi.nlm.nih.gov/) and Patric (https://www.patricbrc.org/) and custom genome files formatted accordingly. Please, be sure to upload only plain text files to **InteractomeSeq** since it does not accept compressed formats (e.g.: zip, gzip or bzip archives).

### Example of a GFF downloaded from NCBI

the second s				-	
##gfl-version 3					
#igf-spec-version	1.21				
#Iprocessor NCB	l annotwriter				
#Igenome-build A	SM852v1				
#Igenome-build-a	ocession NCBI_As	isembly:GC	F_0000	0852	5.1
##sequence-regi	on NC_000915.11	1667867	2.50.50		
##species http://w	www.ncbi.nlm.nlh.go	ov/Taxonom	y/Brows	erhw	wtax.cgi?id=85962
NC_000915.1	RefSeq gene	217	633	0 -	0 ID=gene0;Dbxref=GeneID:898756;Name=nusB;gbkey=Gene;gene=nusB;gene_biotype=protein_coding;locus_tag=HP0001
NC_000915.1	RefSeq CDS	217	633	0 -	0 ID=cds0;Parent=gene0;Dbxref=Genbank NP_206803.1;GeneID:898756;Name=NP_206803.1;Note=Regulates rRNA biosynthe
NC 000915.1	RefSeq gene	635	1105	0-	0 ID=gene1;Dbxref=GeneID:898768;Name=rbH;gbkey=Gene;gene=ribH;gene_biotype=protein_coding;locus_tag=HP0002
NC_000915.1	RefSeq CDS	635	1105	0-	01D=cds1_Parent=gene1_Dtxref=Genbank NP_206804.1_GeneID.898768_Name=NP_206804.1_Note=RibE%38_6%2C7-dimet
NC 000915.1	RefSeg gene	1115	1945	0 -	0 ID=gene2; Dbxref=GeneID:#98773;Name=HP0003;gbkey=Gene.gene_biotype=protein_coding.tocus_tag=HP0003
NC 000915.1	RefSeq CDS	1115	1945	0 -	01D=cds2;Parent=gene2;Dbxref=Genbank:NP_206805;1;GeneID:896773;Name=NP_206805;1;Note=catalyzes the formation of
NC_000915.1	RefSeq gene	1932	2597	0-	0 ID-gene3 Dbxref-GeneID 898779 Name-HP0004;gbkey-Gene.gene_biotype-protein_coding.locus_tag=HP0004
NC 000915.1	RefSeq CDS	1932	2597	0-	0 ID=cds3.Parent=gene3.Dbxref=Genbank.NP_206806.1.GeneID.898779.Name=NP_206806.1.gbkey=CDS.product=carbonic
NC 000915.1	RefSeq gene	2719	3402	0+	0 ID=gene4;Dbxref=GeneID:898802:Name=HP0005;gbkey=Gene;gene_biotype=protein_coding3ocus_tag=HP0005
NC_000915.1	RefSeq CDS	2719	3402	0+	01D=cds4.Parent=gene4.Dbxref=Genbank.NP_206807.1,GeneID.896802;Name=NP_206807.1,Note=type 1 subfamily%3B inv
NC 000915.1	RefSeq gene	3403	4233	0+	0 ID=gene5;Dbxref=GeneID:898828;Name=panC;gbkey=Gene;gene=panC;gene_biotype=protein_coding;locus_tag=HP0006
NC 000915.1	RefSeq CDS	3403	4233	0+	01D=cds5;Parent=gene5;Dbxref=Genbank:NP_206808;1;GeneID:896828;Name=NP_206808;1;Note=catalyzes the formation of
NC 000915.1	RefSeg gene	4250	4322	0-	01D=gene5;Dbxre1=Gene1D \$98829;Name=tRNA-Glu-1;gbkey=Gene;gene=tRNA-Glu-1;gene biotype=tRNA.jocus tag=HPI01
NC_000915.1	RefSeq IRNA	4250	4322	0-	01D=ma0;Parent=gene6;Dbxref=Gene1D:898829;gbkey=tRNA;gene=tRNA-Glu-1;product=tRNA-Glu
NC 000915.1	RefSeg exon	4250	4322	0-	0 ID:ad1 Parent: ma0 Douret: GeneID 898829 obsev: IRNA gene: IRNA-Gu-1 product: IRNA-Gu

### Example of a GFF downloaded from Patric

##gfl-version 3		1				
#Genome: 400667.7)	Acinetobacter baum	annii ATCC 1	17978			
#Date:02/24/2015						
##sequence-region	accrifNC 009083		1 13408			
accn/NC_009083	RefSeq	CDS	1	957	0+	0 ID=A15_3471 jocus_tag=A15_3471;product=hypothetical protein
accn(NC_009083	RefSeq	CDS	950	1504	0+	0 ID=A15_3461 locus_tag=A15_3461;product=DNA replication protein
accriNC_009083	RefSeq	CDS	2523	3437	0 -	0 ID=A15_3462.locus_tag=A1S_3462:product=hypothetical protein
accn/NC_009083	RefSeq	CDS	3538	4788	0+	0 ID=A15_3463.locus_tag=A15_3463.product=diaminopimelate decarboxylase;ec_number=4.1.1.20
accn/NC_009083	RefSeq	CDS	5039	5629	0+	0 ID=A1S_3464;locus_tag=A1S_3464;product=Cro-like protein
accn/NC_009083	RefSeq	CDS	6340	6906	0 -	0 ID=A15_3465;locus_tag=A1S_3465;product=hypothetical protein
accn/NC_009083	RefSeq	CDS	7074	7685	0 -	0 ID=A15_3466:locus_tag=A15_3466:product=resolvase
accn/NC_009083	RefSeq	CDS	8602	9732	0 -	0 ID=A15_3467;locus_tag=A15_3467;product=hypothetical protein
accn/NC_009083	RefSeq	CDS	10072	10374	0-	0 ID=A15_3468:locus_tag=A1S_3468:product=putative lipoprotein
accnINC_009083	RefSeq	CDS	10367	10723	0-	0 ID=A15_3469;locus_tag=A1S_3469;product=diaminopimelate decarboxylase
accn/NC_009083	RetSeq	CDS	12076	12444	0-	0 ID=A15_3470;locus_tag=A1S_3470;product=regulatory protein LysR

### CSV/TSV Format

Finally, annotations can be provided also as a separated-columns text. Column separator can be chosen among TAB, SPACE, comma (",") or semicolon (";"). This file must have 5 columns.

CSV file fields must contain the following information:

- 1. chrom The name of the chromosome or scaffold.
- 2. chromStart The starting position of the gene in the chromosome or scaffold. The first base in a chromosome is numbered 1.
- 3. chromEnd The ending position of the gene in the chromosome or scaffold.
- 4. strand Defines the transcription strand for each gene. Either "+" or "-".
- 5. name Defines the gene name.

An example of a valid Genome annotation CSV file for use in **InteractomeSeq** is reported below.

NC\_011334.1,485,2020,+,HPG27\_RS07980 NC\_011333.1,899,1729,-,HPG27\_RS00015 NC\_011333.1,1716,2381,-,HPG27\_RS00020 NC\_011334.1,2065,2799,+,HPG27\_RS07985 NC\_011333.1,2503,3186,+,HPG27\_RS07985 NC\_011334.1,2849,3553,+,HPG27\_RS07990 NC\_011334.1,2849,3553,+,HPG27\_RS07990 NC\_011333.1,3187,4017,+,HPG27\_RS07995 NC\_011334.1,3672,4835,-,HPG27\_RS07995 NC\_011333.1,4031,4106,-,HPG27\_RS00035 NC\_011333.1,4179,4255,-,HPG27\_RS00040 NC\_011333.1,4306,4381,-,HPG27\_RS00045 NC\_011333.1,4423,4497,-,HPG27\_RS00050

# 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.

	Mapping Params	
	Sequencing Type	SINGLE-READ PAIRED-ENDS
L	Read File	26695_S5_L001_R1_001.fastq
L	Mapping Label	26695_S5_L001_R1_001
=	Organism	
	organishi	
	Annotation Strain	Helicobacter pylori 26695 - NC_000915
	Adapters	
	Adapters	Autodetect Adapters
ŧ	Trimming Params	
	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.

#### **B** Mapping List

							E COLI	umns 🔫
Info	Label	÷	Status	÷	Date	÷	Log	Output
<b>~</b>	26695_S5		Done		01/10/2018 - 12:00:0	0	•	
<b>~</b>	HealthyControl		Done		01/10/2018 - 12:00:0	0	۲	
<b>~</b>	PositiveControl		Done		01/10/2018 - 12:00:0	10	۲	
	AtrophicGastritis		Done		01/10/2018 - 12:00:0	0	۲	
-							5 10	2
							_	

## Domain Analysis

Domain Anlysis is composed by four sheets:

- 1. Domain Definition
- 2. Domain Enrichment
- 3. Domain Subtraction
- 4. Domain Intersection

1 Domain Definition 2 Domain Enrichment 3 Domain Subtraction 4 Domain Intersection	
--	--

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

III Domain Definition :: Insert

E Domain Definition Para	ms
Mapping 🖲	Ψ.
Domain Definition Label	Domain Definition Label
🗃 Organism	
Annotation Strain	Helicobacter pylori 26695 - NC_000915

Domain Definition back-end script launches "bedtool genomomecov" to computes coverage depth at each genome position of the coding regions (CDS). Next with a custom script it calculates the average depth coverage and only the genome 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.

E Domain Definition	Domain Enrichment	E Domain Subtraction	E Domain Intersection		
👪 Domain Definiti	ion List				
Info 1	Label	2	3 Status	÷ 4 Date 5	Log Output 6
	26695_S5		Done	01/10/2018 - 12:00:00	•
	HealthyControl		Done	01/10/2018 - 12:00:00	•
	PositiveControl		Done	01/10/2018 - 12:00:00	•
<b>2</b>	AtrophicGastriti	S	Done	01/10/2018 - 12:00:00	•

- 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.

EDomain Definition :: Log :: 26695\_S5



#### 6. **Output** – Hide/open panel with output preview

Domain Definition :: Output :: 26695\_S5

				_					
TOTAL: 2,986									
Info	Chromosome	÷	Clone Start $\ \ \Leftrightarrow$	Clone End $\Leftrightarrow$	Clone Length \$	Start \$	End \$	Gene	\$ Strand
	NC_000915.1		346	526	180	217	633	HP0001	-
<b>~</b>	NC_000915.1		724	1073	349	635	1105	HP0002	-
<b>~</b>	NC_000915.1		1178	1721	543	1115	1945	HP0003	-
<b>~</b>	NC_000915.1		1775	1955	170	1115	1945	HP0003	-
<b>~</b>	NC_000915.1		1983	2463	480	1932	2597	HP0004	-
<b>~</b>	NC_000915.1		2751	3018	267	2719	3402	HP0005	+
<b>~</b>	NC_000915.1		3060	3168	108	2719	3402	HP0005	+
<b>~</b>	NC_000915.1		3789	3897	108	3403	4233	HP0006	+
<b>~</b>	NC_000915.1		3900	4035	135	3403	4233	HP0006	+
	NC_000915.1		5743	5985	242	5241	7145	HP0009	-

2. **Domain Enrichment** takes as input the Genomic and Target output of Domain Definition step.

**B** Domain Enrichment List

									🖽 COLUN	MNS 🗕
Info	Label	÷	Status	÷	Date	÷	Log	Output	Edit	Delete

III Domain Enrichment :: Insert

Target Domain Definition () Domain Enrichment Label Domain Enrichment Label	Genomic Domain Definition		Ψ.	
Domain Enrichment Label Domain Enrichment Label	Target Domain Definition			
	Domain Enrichment Label 🚯	Domain Enrichment Label	_	

Domain Enrichment back-end script launches "bedtools genomecov" to compute the number of feature (reads) that map inside the epitope regions. 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 background (Genomic Domain Definition). When the computational steps are complete, user can check the status of his analysis.

#### E Domain Enrichment List

				[ ⊞ COL	umns 🔻
Info	2 Label	3 Status	\$ <b>4</b> Date	5 Log	Output
<b>×</b>	26695_S5 + HealthyControl	Done	01/10/2018 - 12:00:00	•	
<b>~</b>	26695_S5 + PositiveControl	Done	01/10/2018 - 12:00:00		
<b>&gt;</b>	26695_S5 + AtrophicGastritis	Done	01/10/2018 - 12:00:00		
				5 10	0 25
Genomic Domain Definition Label	26695_S5				
Target Domain Definition Label	AtrophicGastritis				

- **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
- Log Button that hide/open a box with execution log file.
   Domain Enrichment :: Log :: 26695\_S5 + HealthyControl

≔ status 🛇	Domain Enrichment Done 🗹	Completed Processing
Prokaryote Domain Enrichn Parsing of mapping output	ient - Start * Friday November 16, 2018 - 13:07:49	
Parsing of mapping output	file complete.	
Parsing of domain definition	output file complete.	
Bedtools coverage complet	8.	
Bedtools coverage complet	B.	
Parsing output bedtools co	rerage complete.	
Parsing output bedtools co	rerage complete.	
Differential expression anal	ysis complete.	
Parsing output edgeR comp	lete.	
Prokaryote Domain Enrichn	ent - End * Friday November 16, 2018 - 13:08:07	

### 6. Output – Hide/open panel with output preview

Domain Enrichment :: Output :: 26695\_S5 + HealthyControl

omain En	nrichment Output File				± DOWNLO	ND					
TOTAL	275										COLUMNS -
Info	Chromosome	≎ Clone Star	t ÷	Clone End 👙	Clone Length	Start 🍦	End 🔤	Gene 🍦	Strand $\Leftrightarrow$	Log FC 	Adjust PValue <sup>‡</sup>
	NC_000915.1		238	373	135	217	633	HP0001	-	2.4653	3.0548e-2
	NC_000915.1	8	8395	8569	174	7603	9243	HP0010		2.1790	3.4764e-3
	NC_000915.1	10	0861	11046	185	9911	11590	HP0012	+	1.8962	1.3982e-2
~	NC_000915.1	14	4979	15288	309	14248	16611	HP0017	+	3.3727	7.6874e-5
~	NC_000915.1	16	6787	17147	284	16863	18272	HP0018	+	2.0296	6.5923e-3
~	NC_000915.1	17	7966	18052	86	16863	18272	HP0018	+	6.7736	7.9131e-15
<b>~</b>	NC_000915.1	33	3910	33988	78	32680	34905	HP0033	+	2.9902	9.8855e-5
~	NC_000915.1	41	1903	42052	149	40651	42063	HP0043	+	5.7755	1.4830e-4
~	NC_000915.1	43	3269	43360	91	43243	44175	HP0045	+	2.5775	5.2191e-3
~	NC_000915.1	46	5430	46492	62	46042	48351	HP0048	-	2.8890	3.6349e-3

**Note**: The Domain Enrichment step requires that the design of the experiment includes InteractomeSequencing of both the bacterial genome and the selections with patient sera.

3. **Domain Subtraction** takes as input two differentially enriched epitopes/domains lists, one defined as Control Domain Enrichment and one defined as Selection Domain Enrichment.

#### **III** Domain Subtraction :: Insert

Control Domain Enrichment	
Selection Domain Enrichment 🔀	
Domain Subtraction Label	Domain Subtraction Label
⊋ Params	
Overlap 🕄	0,5 ()

Domain Subtraction back-end script launches "bedtools subtract" that searches for domains in Control Enrichment file that overlap with those of Selection Enrichment file. If an overlapping feature is found in Control Enrichment file, the overlapping portion is removed from Selection Enrichment file and the remaining portion of Selection Enrichment domains are reported. When the computational steps are complete, user can check the status of his analysis.

#### Domain Subtraction List

								E COLI	JMNS 🔻	
-	nfo	2 Label	÷	3 Status	\$ <b>(</b>	4 Date	5	Log	Output	6
	<b>У</b> Не	althyControl-PositiveControl		Done		01/10/2018 - 12:	00:00	۲		
	▼ He	althyControl-AtrophicGastritis		Done		01/10/2018 - 12:	00:00	۲		
(1)								5 10	25	
	Control DomainEnrichment Label	26695_S5 + HealthyControl								
	Selection DomainEnrichment Label	26695_S5 + HealthyControl								

- 1. Info Drop-down menu with information of Domain Enrichment 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.
  - I Domain Subtraction :: Log :: HealthyControl-PositiveControl

6. Output – Hide/open panel with output preview

omain Su	ubtraction Output File			2 DOWNLOAD						
TOTAL	251									COLUMNS 🔻
Info	Chromosome 🔶	Clone Start 👙	Clone End 👙	Clone Length $\stackrel{\diamond}{\Rightarrow}$	Start 🔶	End 🔤	Gene 🔶	Strand $\doteqdot$	Log FC  💠	Adjust PValue <sup>‡</sup>
	NC_000915.1	5873	6027	154	5241	7145	HP0009	-	2.7581	1.3126e-3
<b>~</b>	NC_000915.1	6284	6387	103	5241	7145	HP0009		5.1679	3.4435e-3
	NC_000915.1	8695	8837	142	7603	9243	HP0010	-	1.7501	2.7446e-2
~	NC_000915.1	11762	12001	239	11587	12639	HP0013	+	1.7709	4.1700e-2
~	NC_000915.1	13992	14096	104	13983	14246	HP0016	+	1.8357	4.3285e-2
~	NC_000915.1	21880	21964	84	21152	22717	HP0022	-	2.2877	5.3086e-3
<b>~</b>	NC_000915.1	26388	26480	92	26078	27358	HP0026	-	2.1189	8.6137e-3
~	NC_000915.1	36527	36705	149	36556	37611	HP0037	+	3.0844	2.6988e-3
~	NC_000915.1	37340	37464	124	36556	37611	HP0037	+	2.8550	1.2769e-3
<b>~</b>	NC_000915.1	60677	60803	126	57741	61298	HP0056	-	2.3361	1.1336e-2

4. **Domain Intersection** takes as input two differentially enriched epitopes/domains lists output of Domain Definition step.

Domain Intersection List

										MNS 👻	
Info		Label	÷	Status	÷	Date	÷	Log	Output	Edit	Delete
Do	main Intersection	:: Insert									
Doi	nain Intersection	:: Insert									
E Doi	main Intersection	:: Insert									
Doi	Selections ()	Domain Intersection Label									

Domain Intersection allows one to screen for overlaps between two sets of epitopes/domains lists.

		E COLUMNS -
	2 Label \$	3 Status
Po	sitiveControl + AtrophicGastritis	Done 01/10/2018 - 12:00:00
Selection Domain Substraction Label	HealthyControl-PositiveControl	5 10 25
Selection Domain Substraction Label	HealthyControl-AtrophicGastritis	
	1. Info – Drop-down menu wit	h information of Domain Enrichment in
	1. Info – Drop-down menu wit file.	h information of Domain Enrichment in
	<ol> <li>Info – Drop-down menu wit file.</li> <li>Label - Sample label.</li> </ol>	h information of Domain Enrichment in
	<ol> <li>Info – Drop-down menu wit file.</li> <li>Label - Sample label.</li> <li>Status – When the execution</li> </ol>	h information of Domain Enrichment in on ends successfully, the button turns
	<ol> <li>Info – Drop-down menu wit file.</li> <li>Label - Sample label.</li> <li>Status – When the execution green, otherwise, it turns re</li> </ol>	h information of Domain Enrichment in on ends successfully, the button turns ed.

5. Log – Button that hide/open a box with execution log file. Domain Intersection :: Log :: PositiveControl + AtrophicGastritis



Domain Intersection List

6. Output – Hide/open 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.

C Mapping		
26695_S5	01/10/2018 - 12:00:00	
HealthyControl	01/10/2018 - 12:00:00	4
PositiveControl	01/10/2018 - 12:00:00	۸.
AtrophicGastritis	01/10/2018 - 12:00:00	<b>A</b>
Domain Definition		
26695_S5	01/10/2018 - 12:00:00	*
HealthyControl	01/10/2018 - 12:00:00	<b>A</b>
PositiveControl	01/10/2018 - 12:00:00	<b>A</b>
AtrophicGastritis	01/10/2018 - 12:00:00	4
Domain Enrichment		
26695_S5 + HealthyControl	01/10/2018 - 12:00:00	*
26695_S5 + PositiveControl	01/10/2018 - 12:00:00	
26695_S5 + AtrophicGastritis	01/10/2018 - 12:00:00	4
Domain Subtraction		
HealthyControl-PositiveControl	01/10/2018 - 12:00:00	<b>A</b>
HealthyControl-AtrophicGastritis	01/10/2018 - 12:00:00	<b>A</b>
Domain Intersection		
PositiveControl + AtrophicGastritis	01/10/2018 - 12:00:00	4

## Outputs

• **Mapping**. Tabular output file is composed by 13 columns and in each row is stored the information about uniquely mapping reads.

seq1:1:15	NC_000915.1	100.0	172	0	0	1	172	273366	273537	3.34e-88	318.0	GAAATACAGAGGCGAGTTTGAAGAGCGCTTGAAAAAG
seq1:1:16	NC_000915.1	100.0	100	0	0	1	100	922010	921911	2.11e-48	185.0	GAATTTAAACGCTGGAAGGCATGCCGATCAAACAGCG
seq1:1:17	NC_000915.1	99.59	244	1	0	1	244	1349358	1349601	2.03e-126	446.0	TAAAGATTCCCCTTTGATCCAAAAAACGCTCAATGTC
seq1:1:19	NC_000915.1	99.465	187	1	0	1	187	941000	941186	7.68e-95	340.0	TCCTTTCGCTCAACCTAGCGCCACTCCTAATTTAGTC
seq1:1:23	NC_000915.1	98.889	90	1	0	1	90	943437	943348	3.23e-41	161.0	CAAAGTTTGAGCTTGGGGATTAACCCGGTGCTGTTGC
seq1:1:24	NC_000915.1	100.0	196	0	0	1	196	1292807	1292612	1.72e-101	363.0	CGCGCGCTATTTTAGGAATTACACCCAATATGTCAAA
1	2	3	4	5	6	(7)	8	9	10	(11)	12	13
			( ')	(-)				- /			( /	

- Tabular data report the following fields:
- 1. Sequence Id
- 2. Reference mapped strand ID used for mapping
- 3. Percentage of sequence aligned
- 4. Sequence length
- 5. Number of mismatch
- 6. Number of gap openings
- 7. Chromosome number
- 8. Alignment length
- 9. Genomic start of alignment
- 10. Genomic end of alignment
- 11. Expected value
- 12. Bit score
- 13. Aligned part of query sequence
- **Domain Definition**. The domain definition step provides output in tabular format. In this file all soluble domains/epitopes detected are listed and the protein coding 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.

ain Definition Output File		a 400	WNLOAD				
OTAL : 2,986 b	2	3	4	5	6	7	
o 1 Chromosome		lone End 💠	Clone Length 🛛 🌩	Start 💠	End 🌩	Gene	Strand 4
NC_000915.1	346	526	180	217	633	HP0001	-
NC_000915.1	724	1073	349	635	1105	HP0002	-
NC_000915.1	1178	1721	543	1115	1945	HP0003	-
NC_000915.1	1775	1955	170	1115	1945	HP0003	-
NC_000915.1	1983	2463	480	1932	2597	HP0004	-
NC_000915.1	2751	3018	267	2719	3402	HP0005	+
NC_000915.1	3060	3168	108	2719	3402	HP0005	+
NC_000915.1	3789	3897	108	3403	4233	HP0006	+
NC_000915.1	3900	4035	135	3403	4233	HP0006	+
NC_000915.1	5743	5985	242	5241	7145	HP0009	
						5 1	0 25 50
Description	membrane prot	ein 9		)		_	
Nucleotide Sequence	TTTGCACGCCG AGAGCTGATTA AATTCTTGCGT GGTTGACTTGA AGGGTTAGTGG GAATTGGGCGT AA	ATCCCATTCATCG AGCCAACGCGTCT GGCGCTTAATAAT TAAACGGCCTGTA ATTGGCTTATCAA GTTTGAAGCGGT	CACCGTTGTTGGTTTG GAAAGGGTTATGCCCT TGCGAATAAGCGCTTT AGCCCCCGGGGTTATT ATTTTTAAGGAATGGG CGTGGTGATGCTGTTAT	10			

- 1. Chromosome Reference mapped strand ID used for mapping
- 2. Clone Start Genomic start of soluble folding domain/epitope
- 3. Clone End Genomic end of soluble folding domain/epitope
- 4. Clone Length Domain/epitope length
- 5. Start Starting coordinate of the gene associated with the domain/epitope on the chromosome.
- 6. End Ending coordinate of the gene associated with the domain/epitope on the chromosome.
- 7. Gene Gene ID
- 8. Strand Strand associated with gene annotation
- 9. Description Gene associated with the domain/epitope annotation
- 10. Nucleotide Sequence Domain/epitope nucleotide sequence.
- **Domain Enrichment**. The domain enrichment step provides output in tabular format. In this file the expression of all soluble domains/epitopes detected from the Selection sample are statistically tested (edgeR) against Genomic domains/epitopes. 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.

		a	Z DOWNLOAD							
fOTAL: 275 b	2 Clone Start ≑	3 Clone End 💠	(4) Clone Length ≑	5 Start \$	Er	5) 1d ÷	7 Gene ‡	8 Strand ¢	9 Log FC \$	COLUMNS Adjust PValue
NC_000915.1	238	373	135	217		633	HP0001	-	2.4653	3.0548e-2
NC_000915.1	8395	8569	174	7603		9243	HP0010	-	2.1790	3.4764e-3
						11590	HP0012	+	1.8962	1.3982e-2
					_	16611	HP0017	+	3.3727	7.6874e-
Description	mo	lecular chaperone Groß	≅∟ (11			18272	HP0018	+	2.0296	6.5923e-3
		$\bigcirc$			315	18272	HP0018	+	6.7736	7.9131e-1
PValue	8.6	150e-4 (12)				34905	HP0033	+	2.9902	9.8855e-5
And and American	07				-	42063	HP0043	+	5.7755	1.4830e-4
Nucleotide Sequence	CT AC	ACGCCTCTTAATTTAT	GUITTAACCGC	TCGTTAAAGC	13	4175	HP0045	+	2.5775	5.2191e-3
	TT	GCCCTCAATGTCTTC	AGCGATGATTAA	AAGCGGTTTG		48351	HP0048	-	2.8890	3.6349e-3
	CC TC	CTCTTTCATGGTTTTT	TCTAGTAGCGGG	GAGAATGTCTT					5 10	25 5

- 1. Chromosome Reference mapped strand ID used for mapping
- 2. Clone Start Genomic start of soluble folding domain/epitope
- 3. Clone End Genomic end of soluble folding domain/epitope
- 4. Clone Length Domain/epitope length
- 5. Start Starting coordinate of the gene associated with the domain/epitope on the chromosome.
- 6. End Ending coordinate of the gene associated with the domain/epitope on the chromosome.
- 7. Gene Gene ID
- 8. Strand Strand associated with gene annotation
- 9. LogFC log2 fold change estimation
- 10. Adjust Pvalue pvalue adjusted for FDR
- 11. Description Gene associated with the domain/epitope annotation
- 12. Pvalue
- 13. Nucleotide Sequence Domain/epitope nucleotide sequence.
- **Domain Subtraction**. The domain subtraction step provides output in tabular format. In this file are listed the soluble domains/epitopes that result from the subtraction between two differentially enriched Selections. 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 remain after the subtraction.

0TAL : 251 b	$\overline{)}$			5	(6)	7	(9)		10
o 1 Chromosome \$	Clone Start $\Rightarrow$	Clone End 😄	Clone Length $\Rightarrow$	Start ≑	End 💠	Gene 🌲	Strand $\Rightarrow$	Log FC 💠	Adjust PValue
NC_000915.1	5873	6027	154	5241	7145	HP0009		2.7581	1.3126e-3
NC_000915.1	6284	6387	103	5241	7145	HP0009	-	5.1679	3.4435e-
					9243	HP0010	-	1.7501	2.7446e-
			$\frown$		12639	HP0013	+	1.7709	4.1700e-2
Description	me	mbrane protein	(11)		14246	HP0016	+	1.8357	4.3285e-
					22717	HP0022	-	2.2877	5.3086e-
PValue	5.7	327e-4 (12	2)		27358	HP0026	-	2.1189	8.6137e-3
					37611	HP0037	+	3.0844	2.6988e-3
Nucleotide Sequence	CG	CGTGATTGAGCATG		GCCCTATCAGCA	37611	HP0037	+	2.8550	1.2769e-3
	TT	IGCGGGTTGTTAAG	GCTTTGA	CONTROATIO	61298	HP0056	-	2.3361	1.1336e-2

- 1. Chromosome Reference mapped strand ID used for mapping
- 2. Clone Start Genomic start of soluble folding domain/epitope
- 3. Clone End Genomic end of soluble folding domain/epitope
- 4. Clone Length Domain/epitope length
- 5. Start Starting coordinate of the gene associated with the domain/epitope on the chromosome.
- 6. End Ending coordinate of the gene associated with the domain/epitope on the chromosome.
- 7. Gene Gene ID
- 8. Strand Strand associated with gene annotation
- 9. LogFC log2 fold change estimation
- 10. Adjust Pvalue pvalue adjusted for FDR
- 11. Description Gene associated with the domain/epitope annotation
- 12. Pvalue
- 13. Nucleotide Sequence Domain/epitope nucleotide sequence.
- Domain Intersection. The domain subtraction step provides two outputs, on in tabular format and the second is a Venn plot. This outputs represents the unique e common soluble domains/epitopes that result from the intersection of two or three differentially enriched Selections. As shown in the figure below, the unique and common domains lists have different name (1) and the circle of the plot is proportionally to the dimensions of the lists (2). The download of the data can be activated by clicking on the corresponding button (3) that will activate the download of all lists generated. The Venn plot can be saved in png or svg format by clicking the desired format button (4).

Domain Intersection	:: Output :: PositiveCor	ntrol + AtrophicGastritis
---------------------	--------------------------	---------------------------



The preview of unique and common domains/epitopes, can be activated by clicking on the button +

A - B	0
<b>B</b> - A	0
● A ∩ B	0

As shown in the image below, download of the data can be activated by clicking on the corresponding button (a) and in box (b) user can quickly check the number of domains that are common or unique between Selections.

TOTAL: 45 b	2 3	4	5	6	7	8	9	COLUMNS
Inf 1 Chromosome 💠	Start   Clone End	Length ÷	Start ≑	End 💠	Gene 🌻	Strand $\diamondsuit$	Log FC ≑	PValue
NC_000915.1	78752 79019	0	78769	80106	HP0075	-	2.0271	1.1259e-2
				36980	HP0123	+	3.6703	1.8226e-5
				10042	HP0129	+	2.3648	3.9465e-3
Description	phosphoglucosa	mine mutase		\$7916	HP0137	-	1.9050	1.2939e-2
	(	(11)				+	2.8818	1.1592e-2
PValue	2.7800e-3	12)		)5637	HP0199	+	11.9146	3.1948e-31
				29502	HP0220	+	5.9376	1.0422e-9
Nucleotide Sequence	CTTTTTTAGTG	GTTTTTAGCACAAA	TGCCCTTCAAAAA	A 53182	HP0252	+	8.9336	3.1830e-5
	TCTTTAGCTTCT	AAAAGGATTCGCAA	TTTGTTTTCAGTG	c 59342	HP0259	+	7.2241	2.5762e-2
	CGCTATAACGGA	TCAAATGGCGGATT AGCGCTATAACCTT	TTCTAGCTTGTCTA TCAGGCTTTCTAA	A 19106 A	HP0300	+	2.7227	1.6699e-2
	GGGGGCTTTTT GGTATAATTCAA	AGGGGTTTAACGCA		3			5 10	25 50

ANB

- 1. Chromosome Reference mapped strand ID used for mapping
- 2. Clone Start Genomic start of soluble folding domain/epitope
- 3. Clone End Genomic end of soluble folding domain/epitope
- 4. Clone Length Domain/epitope length
- 5. Start Starting coordinate of the gene associated with the domain/epitope on the chromosome.
- 6. End Ending coordinate of the gene associated with the domain/epitope on the chromosome.
- 7. Gene Gene ID
- 8. Strand Strand associated with gene annotation
- 9. LogFC log2 fold change estimation
- 10. Adjust Pvalue pvalue adjusted for FDR
- 11. Description Gene associated with the domain/epitope annotation
- 12. Pvalue
- 13. Nucleotide Sequence Domain/epitope nucleotide sequence.