HOW IT WORKS

HTS-flow is based on the IFOM/IEO/IIT Campus cluster, and therefore a Campus cluster account is needed to log in.

If you do not have a Campus cluster account yet, you have to ask on to the Service Desk system:
http://servicedesk.ieo.it/helpdesk/WebObjects/Helpdesk.woa
You can contact the service desk if you cannot remember your username or password.
Once you have obtained a cluster account, you need to contact the HTS-flow responsible person to be added in the system.

PRIMARY ANALYSIS

Submitting a job

From this page a computational biologist can run a primary analysis on a group of samples. Primary analyses consist of quality control of the raw reads, followed by filtering and alignment to reference genome.

The genomes currently available in HTS-flow are:
mm9 - Mus musculus
mm10 - Mus musculus
hg18 - Homo sapiens
hg19 - Homo sapiens
rn5 - Rattus norvegicus
dm6 - Drosophila melanogaster

Sequencing technologies currently implemented in HTS-flow are RNA-Seq, ChIP-Seq, DNaseI-Seq, BS-Seq.

SECONDARY ANALYSIS

Secondary Analyses are datatype-specific. HTS-flow supports:
Expression Quantification
Differential Genes Expression (DEG calling)
Peak Calling
Footprint Calling

This page is used to launch secondary analyses. In the first panel, the user can choose the appropriate type of secondary analysis, while in the second panel the available aligned samples can be chosen by clicking on them and then pressing the “SELECT” button. A final panel, specific for the chosen secondary analysis, will open automatically.

When a secondary analysis is complete, it will be shown in the COMPLETED ANALYSIS page, under the SECONDARY JOBS table. The output of each analysis is stored in the specific folder within HTS-flow. Secondary analysis results can then be accessed in two ways:

(1) With a web browser, at the URL:
http://www.bioinfo.ieo.eu/BAgroup/HTS-flow/DB/secondary/SECONDARY_ID/
where SECONDARY_ID corresponds to the Secondary ID assigned by HTS-flow.

(2) On the campus cluster (grid.ieo.eu) at the address:
/data/BA/public_html/HTS-flow/DB/secondary/SECONDARY_ID
where SECONDARY_ID corresponds to the Secondary ID assigned by HTS-flow.

The output of a secondary analysis is saved in RDS format, which can be loaded within R with the function readRDS:
results=readRDS(‘path_to_your_result_file’)
Refer to https://stat.ethz.ch/R-manual/R-devel/library/base/html/readRDS.html for more information about this function.

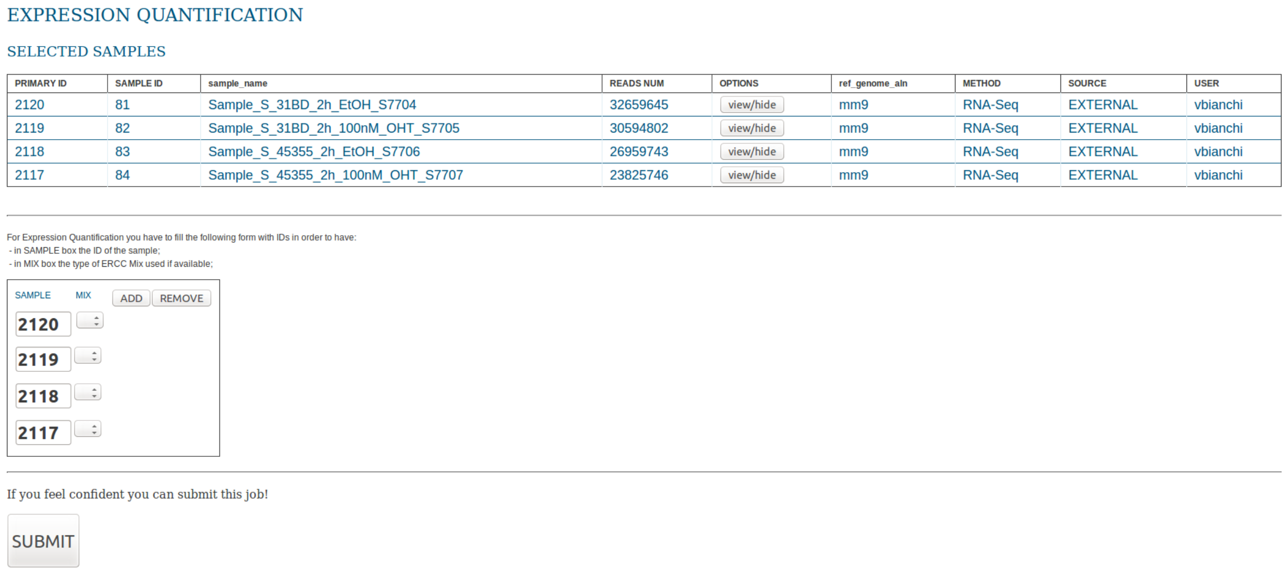
Expression Quantification

Submitting jobs

After choosing the samples, the user must fill a table with two fields:
SAMPLE: the PRIMARY ID of the sample;
MIX: the ERCC spike-in Mix (either 1 or 2) added for normalization. If no mix was added, leave empty. This feature has not yet implemented.

The ADD and REMOVE buttons can be used to create/delete more lines in the table and quantify the expression of multiple samples within the same secondary analysis.

Finally, the job can be submitted by clicking the SUBMIT button (Figure 1). HTS-flow will assign a SECONDARY ID to the job and automatically redirect the web browser to the RUNNING ANALYSES page, which contains the list of running jobs. When the analysis is complete, it will be moved to the COMPLETED ANALYSES webpage, in the SECONDARY JOBS panel.



EXPRESSION QUANTIFICATION

SELECTED SAMPLES

PRIMARY ID	SAMPLE ID	sample_name	READS NUM	OPTIONS	ref_genome_aln	METHOD	SOURCE	USER
2120	81	Sample_S_31BD_2h_EtOH_S7704	32659645	view/hide	mm9	RNA-Seq	EXTERNAL	vbianchi
2119	82	Sample_S_31BD_2h_100nM_OHT_S7705	30594802	view/hide	mm9	RNA-Seq	EXTERNAL	vbianchi
2118	83	Sample_S_45355_2h_EtOH_S7706	26959743	view/hide	mm9	RNA-Seq	EXTERNAL	vbianchi
2117	84	Sample_S_45355_2h_100nM_OHT_S7707	23825746	view/hide	mm9	RNA-Seq	EXTERNAL	vbianchi

For Expression Quantification you have to fill the following form with IDs in order to have:
- in SAMPLE box the ID of the sample;
- in MIX box the type of ERCC Mix used if available;

SAMPLE	MIX
2120	
2119	
2118	
2117	

If you feel confident you can submit this job!

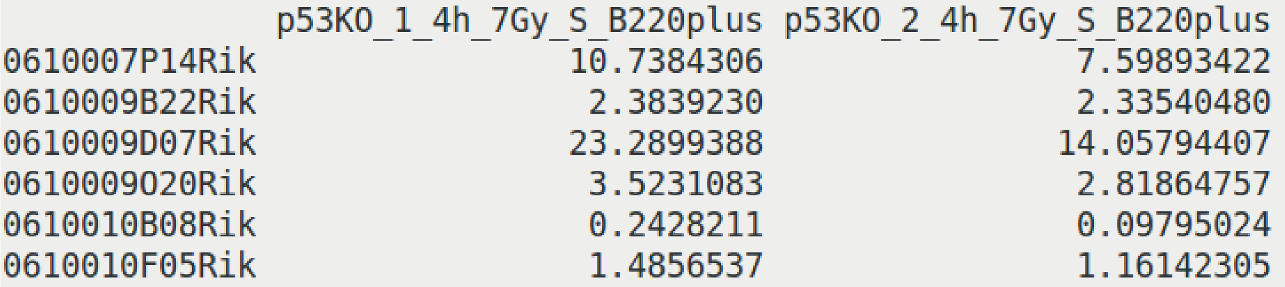
Figure 1: HTS-flow Expression Quantification (Secondary Analysis) web interface. In this example, 4 samples have been selected for expression quantification analysis. The primary IDs were used for filling the table, while no spike-in Mix was selected.

Output

The Expression Quantification analysis will create two different output files: RPKMS.rds and eRPKMS.rds.
(1) RPKMS.rds: absolute quantification of gene expression in terms of Reads per Kilobase per Million of mapped reads (RPKM).

(2) eRPKMS.rds: absolute quantification of gene expression in terms of Reads per Kilobase per Million of mapped exonic reads (eRPKM).

Both files are R data frames where each row is a gene (row names are Gene Symbols) and each column is a sample, and values correspond to gene expressions. In Figure 2, the first 6 rows of the RPKMS.rds file from a two-sample analysis are shown; the column names are the sample names retrieved by the LIMS or defined by users when submitting external data to HTS-flow.



```
              p53KO_1_4h_7Gy_S_B220plus p53KO_2_4h_7Gy_S_B220plus
0610007P14Rik                10.7384306                7.59893422
0610009B22Rik                 2.3839230                2.33540480
0610009D07Rik                23.2899388               14.05794407
0610009O20Rik                 3.5231083                2.81864757
0610010B08Rik                 0.2428211                0.09795024
0610010F05Rik                 1.4856537                1.16142305
```

Figure 2: HTS-flow Expression Quantification output. The first 6 rows of the result table are shown. For each gene, the RPKM value is reported in each sample.

Differential Gene Expression (DEG calling)

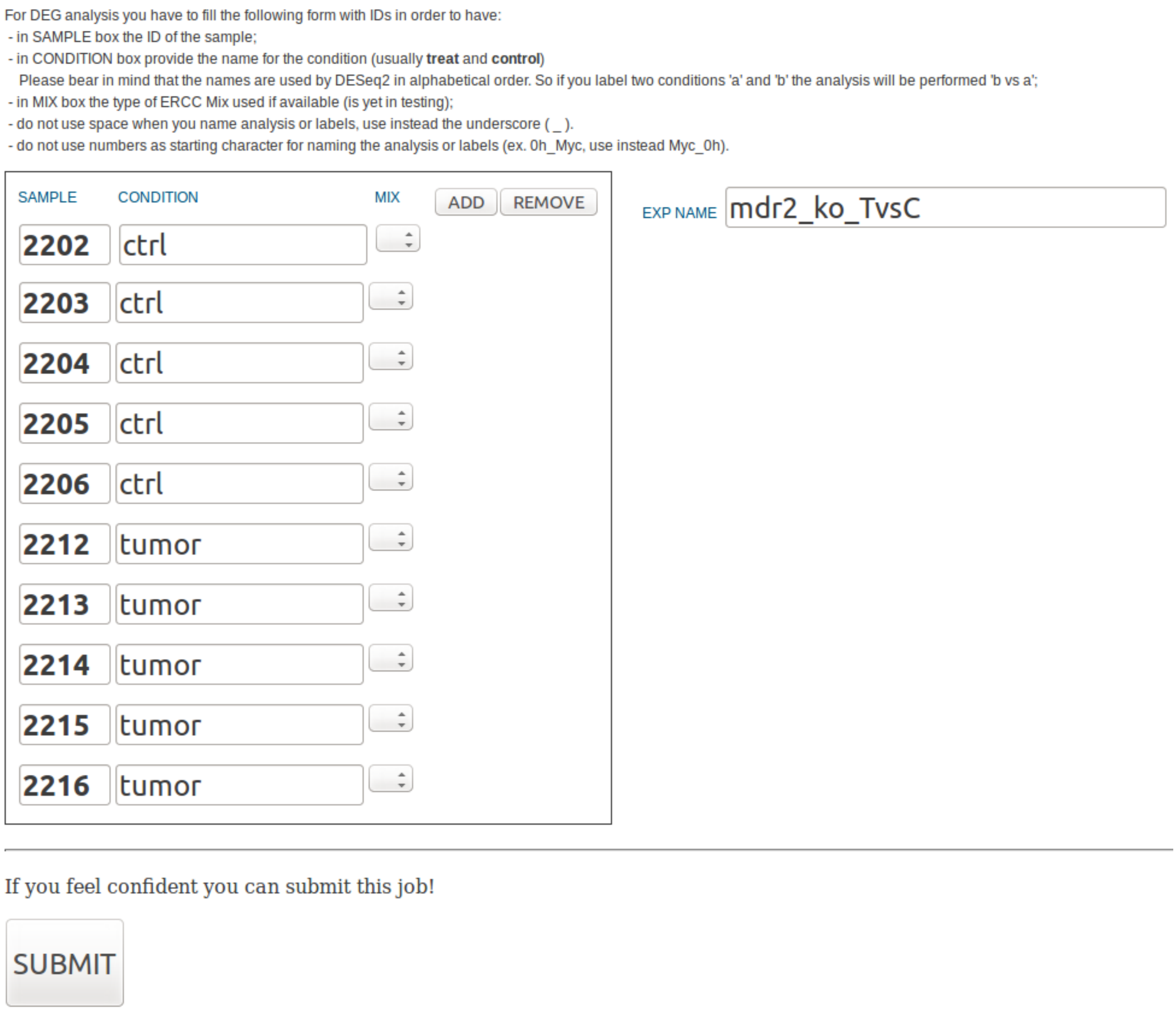
Submitting jobs

Calling Differentially Expressed Genes (DEGs) is performed with DESeq2. The experimental design requires two conditions, typically a treated set of samples and a control set of samples. The presence of replicates in at least one condition is essential.

After choosing the samples, the user must fill a table with three fields:
SAMPLE: the PRIMARY ID of the sample.
CONDITION: one of the two classes of samples (e.g.: treated and control) in the experimental design.
MIX: the ERCC spike-in Mix (either 1 or 2) added for normalization. If no mix was added, leave empty. This feature has not yet implemented.
EXP NAME: a name for the analysis, to be used for the output file. Avoid spaces, use instead underscores.

The ADD and REMOVE buttons can be used to create/delete more lines in the table and add replicates to the conditions.

Finally, the job can be submitted by clicking the SUBMIT button (Figure 3). HTS-flow will assign a SECONDARY ID to the job and automatically redirect the web browser to the RUNNING ANALYSES webpage, showing the list of running jobs. When the analysis is complete, the analysis will be visible on the COMPLETED ANALYSES webpage, under the SECONDARY JOBS panel.



For DEG analysis you have to fill the following form with IDs in order to have:
- in SAMPLE box the ID of the sample;
- in CONDITION box provide the name for the condition (usually **treat** and **control**)
  Please bear in mind that the names are used by DESeq2 in alphabetical order. So if you label two conditions 'a' and 'b' the analysis will be performed 'b vs a';
- in MIX box the type of ERCC Mix used if available (is yet in testing);
- do not use space when you name analysis or labels, use instead the underscore (_).
- do not use numbers as starting character for naming the analysis or labels (ex. 0h_Myc, use instead Myc_0h).

SAMPLE	CONDITION	MIX
2202	ctrl	
2203	ctrl	
2204	ctrl	
2205	ctrl	
2206	ctrl	
2212	tumor	
2213	tumor	
2214	tumor	
2215	tumor	
2216	tumor	

EXP NAME mdr2_ko_TvsC

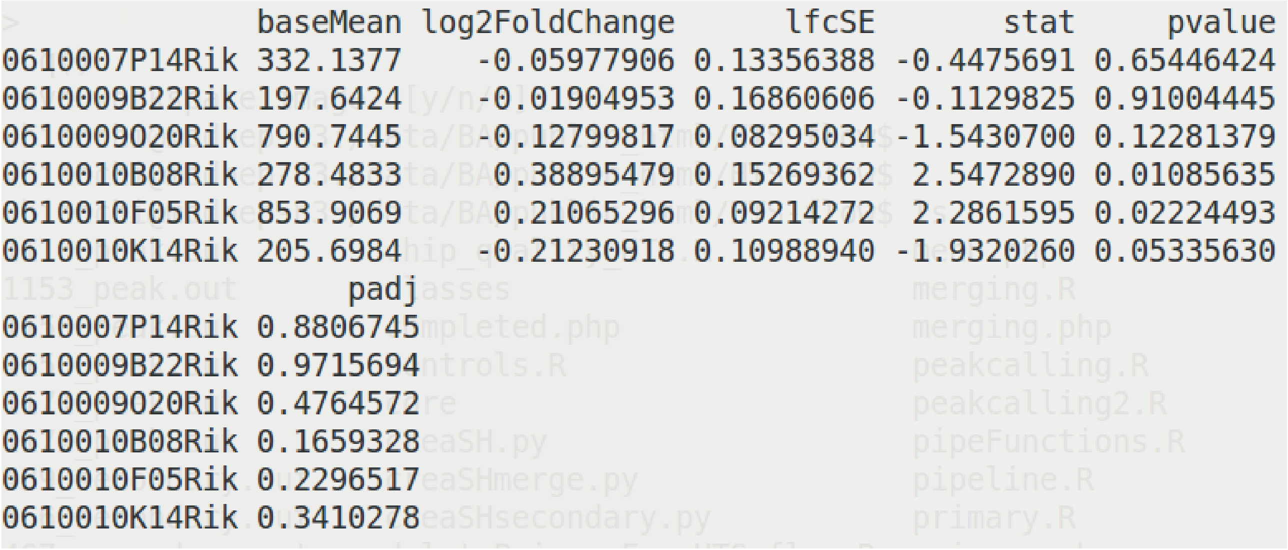
If you feel confident you can submit this job!

Figure 3: HTS-flow DEG web interface. The experiment has two conditions: ctrl and tumor, which were associated to the primary IDs of the samples. An experiment name was provided (mdr2_ko_TvsC).

Output

Differential Gene Expression analysis will create a single output file, corresponding to the EXP NAME field, in the form 'EXP NAME'.rds.
This file contains an R data frame where each row is a gene (Gene Symbol) and columns list the DESeq2 default outputs (see Figure 4):

baseMean: the mean gene expression over all samples in the two conditions
log2FoldChange: log2 Fold Change, treated vs untreated
lfcSE: standard error, treated vs untreated
stat: Wald test statistic
pvalue: Wald test p-value
padj: Benjamini-Hochberg adjusted p-values (False Discovery Rate)



```
              baseMean log2FoldChange      lfcSE       stat     pvalue
0610007P14Rik 332.1377    -0.05977906 0.13356388 -0.4475691 0.65446424
0610009B22Rik 197.6424    -0.01904953 0.16860606 -0.1129825 0.91004445
0610009O20Rik 790.7445    -0.12799817 0.08295034 -1.5430700 0.12281379
0610010B08Rik 278.4833     0.38895479 0.15269362  2.5472890 0.01085635
0610010F05Rik 853.9069     0.21065296 0.09214272  2.2861595 0.02224493
0610010K14Rik 205.6984    -0.21230918 0.10988940 -1.9320260 0.05335630
                   padj
0610007P14Rik 0.8806745
0610009B22Rik 0.9715694
0610009O20Rik 0.4764572
0610010B08Rik 0.1659328
0610010F05Rik 0.2296517
0610010K14Rik 0.3410278
```

Figure 4: DEG calling output in HTS-flow (first 6 rows). Rows correspond to genes, while the different DESeq2 outputs are on the columns.

Peak Calling

Submitting jobs

Peak Calling is performed with MACS2.
After choosing the samples, the user must fill a table with three fields:
SAMPLE1: the PRIMARY ID of the sample used as input in the peak call.
SAMPLE2: the PRIMARY ID of the sample where peaks should be called.
LABEL: label associated to SAMPLE2, which will be used as reference name for SAMPLE2 in this secondary analysis. Avoid spaces, use instead underscores.

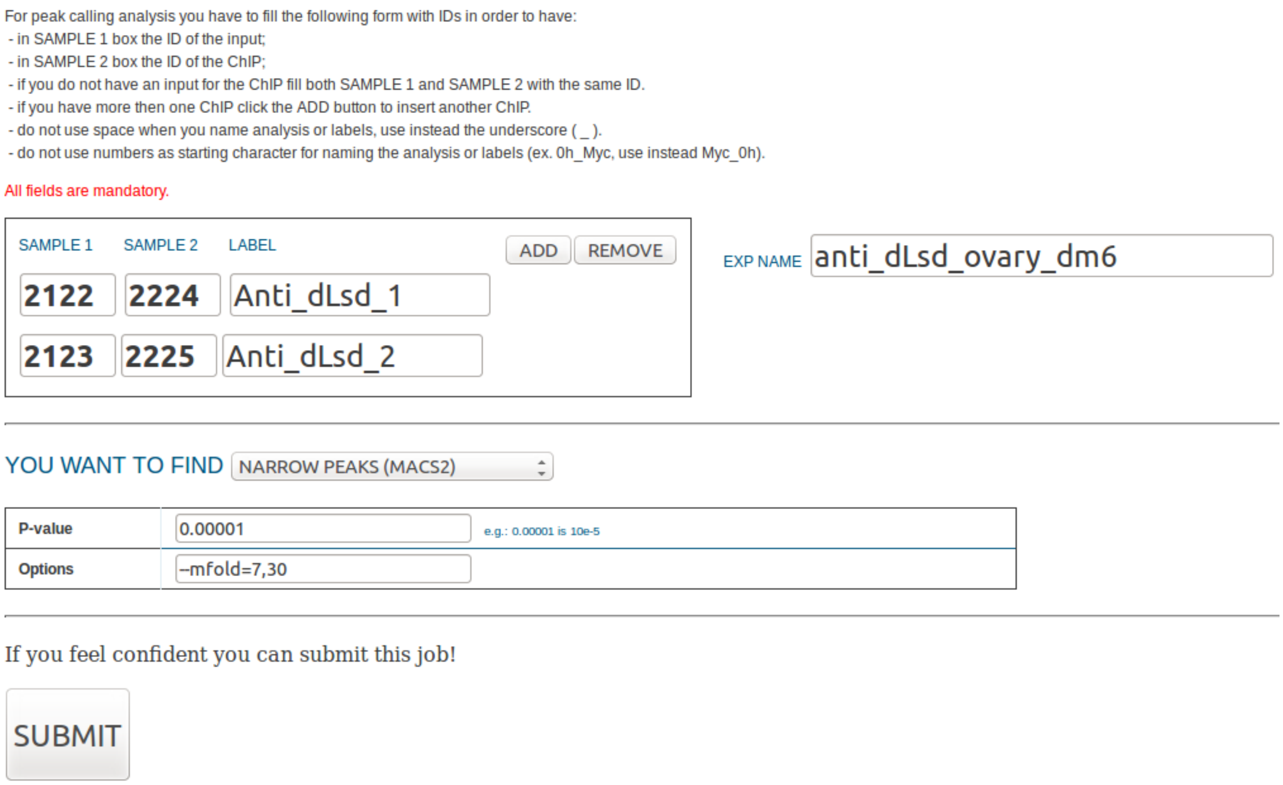
The ADD and REMOVE buttons can be used to create/delete more lines in the table and performing more peak calls within the same secondary job.

A name for the analysis must to be assigned through the **EXP NAME** input form. Avoid spaces in the name, use instead underscores.

The last part of the form allows to select the parameters for the peak call performed by MACS2.

Peak shapes must chosen accordingly to the specific ChIP-seq experiment: for example, NARROW peaks for transcription factors, BROAD peaks for histone marks.
With the NARROW/BROAD option both calls will be performed and the union of peaks from the NARROW and BROAD analysis will be output. This option will affect the annotation analysis performed by HTS-flow on the identified peaks as explained below.

Finally, the job can be submitted by clicking the SUBMIT button (Figure 5). HTS-flow will assign a SECONDARY ID to the job and automatically redirect the web browser to the RUNNING ANALYSES webpage, showing the list of running jobs. When the analysis is complete, the analysis will be visible on the COMPLETED ANALYSES webpage, under the SECONDARY JOBS panel.



For peak calling analysis you have to fill the following form with IDs in order to have:

- in SAMPLE 1 box the ID of the input;
- in SAMPLE 2 box the ID of the ChIP;
- if you do not have an input for the ChIP fill both SAMPLE 1 and SAMPLE 2 with the same ID.
- if you have more then one ChIP click the ADD button to insert another ChIP.
- do not use space when you name analysis or labels, use instead the underscore (_).
- do not use numbers as starting character for naming the analysis or labels (ex. 0h_Myc, use instead Myc_0h).

All fields are mandatory.

SAMPLE 1	SAMPLE 2	LABEL
2122	2224	Anti_dLsd_1
2123	2225	Anti_dLsd_2

EXP NAME anti_dLsd_ovary_dm6

YOU WANT TO FIND NARROW PEAKS (MACS2)

P-value	0.00001	e.g.: 0.00001 is 10e-5
Options	–mfold=7,30	

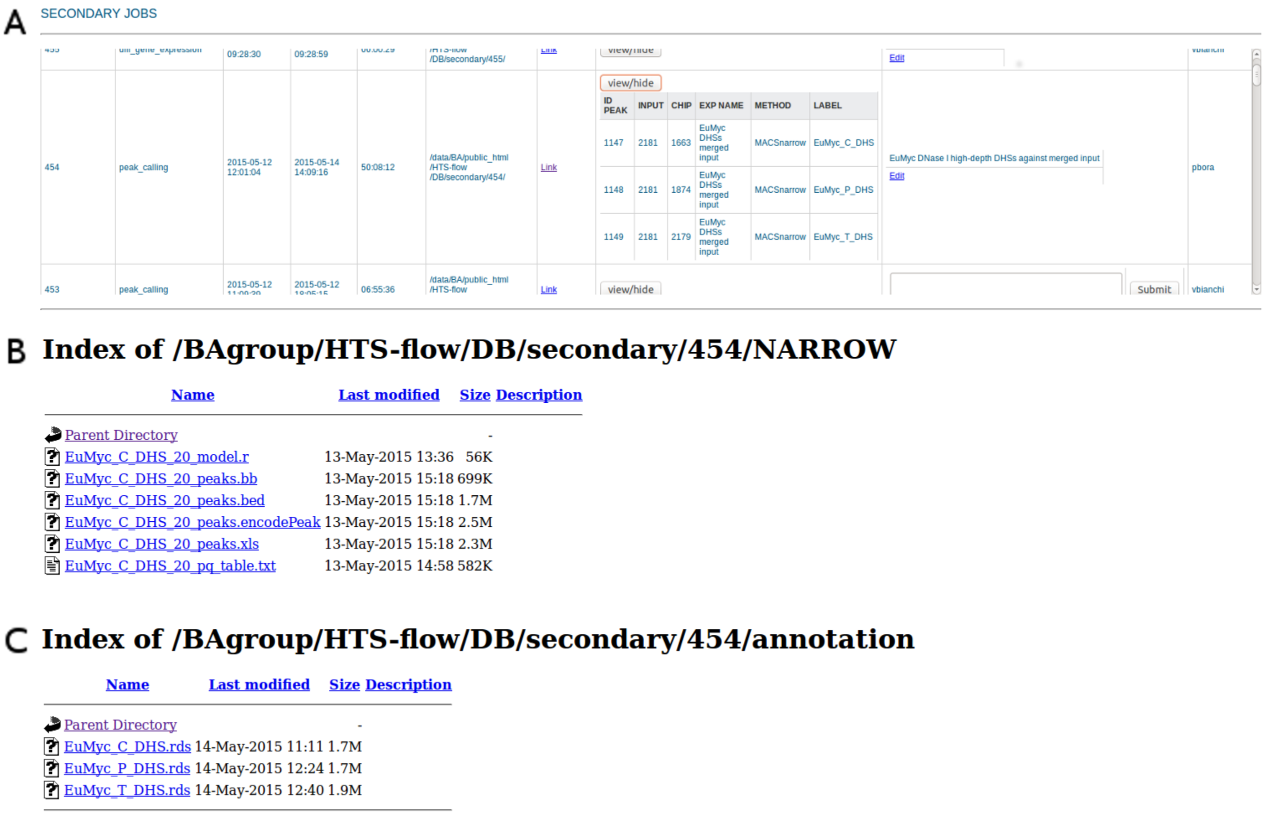
If you feel confident you can submit this job!

Figure 5: HTS-flow Peak Calling web interface.

Output

The output of the peak calling analysis is distributed in two folders: the NARROW/ or BROAD/ folders (depending on the type of call) contain the MACS2 output, i.e. a bed file containing the genomic locations for each peak identified for each sample and a saturation table file for each sample. Saturation reports at different fold-enrichments, the proportion of peaks that could still be detected when using 80% to 20% of the sequence reads. Saturation file name is in the form 'LABEL'_saturation.txt.

Besides, in the annotation/ folder the Peak Calling analysis will create an output file per sample submitted, whose name will be in the form ‘LABEL.rds’. Each file contains a GRanges object where each element is a genomic interval (a peak), complemented by information obtained with the GRannotate and GRenrichment functions from compEpiTools



SECONDARY JOBS

455	diff_gene_expression	09:28:30	09:28:59	00:00:29	/HTS-flow /DB/secondary/455/	Link	view/hide	Edit	vbianchi
454	peak_calling	2015-05-12 12:01:04	2015-05-14 14:09:16	50:08:12	/data/BA/public_html /HTS-flow /DB/secondary/454/	Link	view/hide	EuMyc DNase I high-depth DHSs against merged input Edit	pbora
453	peak_calling	2015-05-12 11:00:20	2015-05-12 18:05:15	06:55:36	/data/BA/public_html /HTS-flow	Link	view/hide	Submit	vbianchi

ID PEAK	INPUT	CHIP	EXP NAME	METHOD	LABEL
1147	2181	1663	EuMyc DHSs merged input	MACSnarrow	EuMyc_C_DHS
1148	2181	1874	EuMyc DHSs merged input	MACSnarrow	EuMyc_P_DHS
1149	2181	2179	EuMyc DHSs merged input	MACSnarrow	EuMyc_T_DHS

Index of /BAgroup/HTS-flow/DB/secondary/454/NARROW

Name	Last modified	Size	Description
Parent Directory		-	
EuMyc_C_DHS_20_model.r	13-May-2015 13:36	56K	
EuMyc_C_DHS_20_peaks.bb	13-May-2015 15:18	699K	
EuMyc_C_DHS_20_peaks.bed	13-May-2015 15:18	1.7M	
EuMyc_C_DHS_20_peaks.encodePeak	13-May-2015 15:18	2.5M	
EuMyc_C_DHS_20_peaks.xls	13-May-2015 15:18	2.3M	
EuMyc_C_DHS_20_pq_table.txt	13-May-2015 14:58	582K	

Index of /BAgroup/HTS-flow/DB/secondary/454/annotation

Name	Last modified	Size	Description
Parent Directory		-	
EuMyc_C_DHS.rds	14-May-2015 11:11	1.7M	
EuMyc_P_DHS.rds	14-May-2015 12:24	1.7M	
EuMyc_T_DHS.rds	14-May-2015 12:40	1.9M	

Figure 6: Peak calling output in HTS-flows. (A) Output for secondary ID 454 in the completed analyses page. Clicking on the "Link" button opens the folders with the results: MACS2 output (B) and annotated rds files (C).

.

For each peak, the following fields are available:
enrichment: log2(ChIP/N1-input/N2), where ChIP is the number of reads falling in the interval in the sample, N1 is the library size of the sample, input is the number of reads falling in the interval in the input and N2 is the library size of the input. Computed with the GRenrichment function from compEpiTools.
summit: position of maximum coverage of the peak. Computed with the GRcoverageSummit function from compEpiTools.
midpoint: the midpoint of the peak. Computed with the GRmidpoint function from compEpiTools.

Annotation is computed in two distinct ways, depending on the type of peak calling requested:

- NARROW calls computes annotation from the summit of the peak

- BROAD calls computes annotation from the midpoint of the peak.

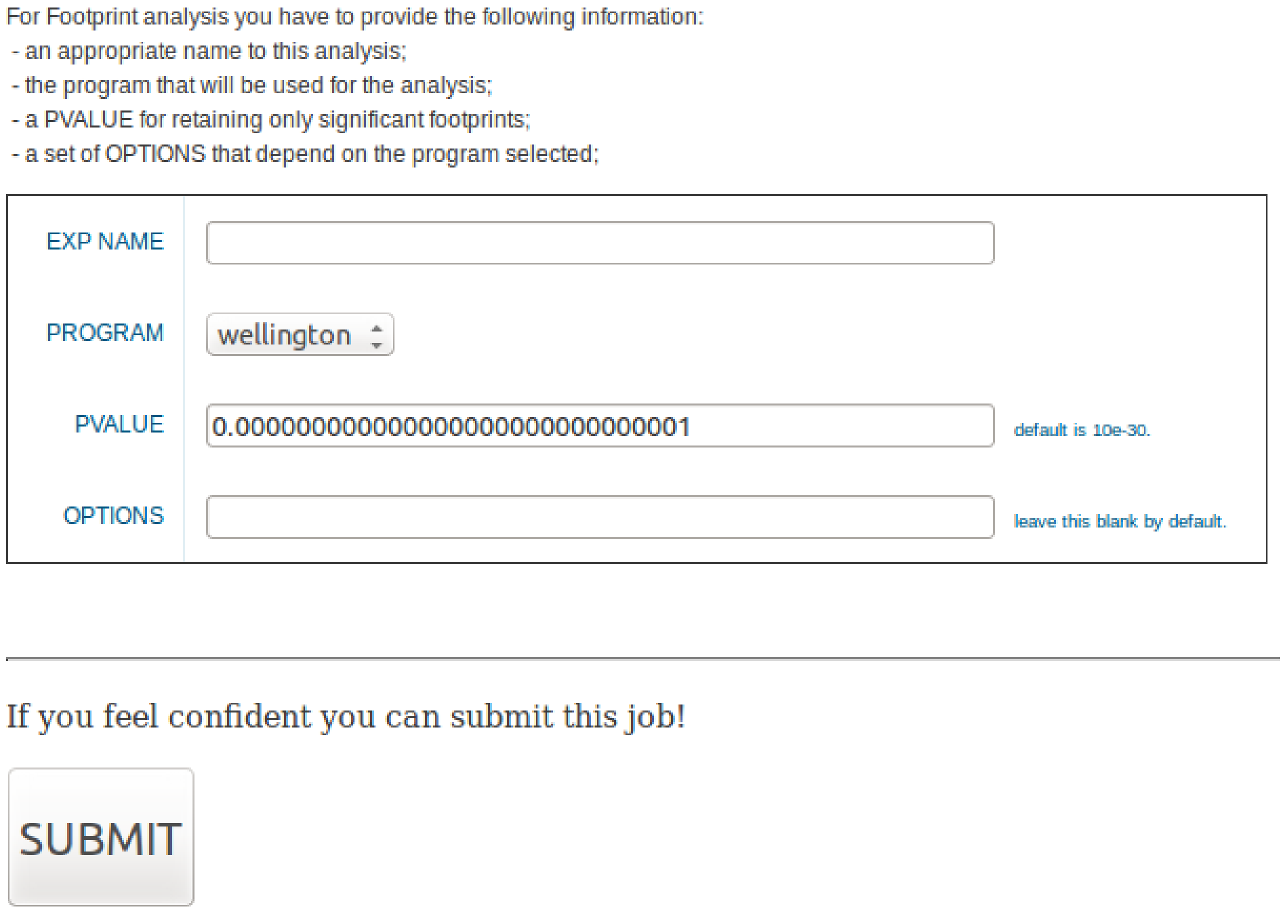
- NARROW/BROAD calls computes annotation from the midpoint of the peak.

Footprint Calling

Submitting jobs

Footprint calls are performed with Wellington.
After choosing the samples, the user must fill a table with four fields:
EXP NAME: A name for this experiment analysis.
PROGRAM: The tool used to call footprint. Currently, only Wellington is available.
PVALUE: The p-value to use as a threshold for statistical significance. By default this is set to **10^{-30}**.
OPTIONS: a set of options that can be used by the selected tool.

Finally, the job can be submitted by clicking the SUBMIT button (Figure 7). HTS-flow will assign a SECONDARY ID to the job and automatically redirect the web browser to the RUNNING ANALYSES webpage, showing the list of running jobs. When the analysis is complete, the analysis will be visible on the COMPLETED ANALYSES webpage, under the SECONDARY JOBS panel.



For Footprint analysis you have to provide the following information:
- an appropriate name to this analysis;
- the program that will be used for the analysis;
- a PVALUE for retaining only significant footprints;
- a set of OPTIONS that depend on the program selected;

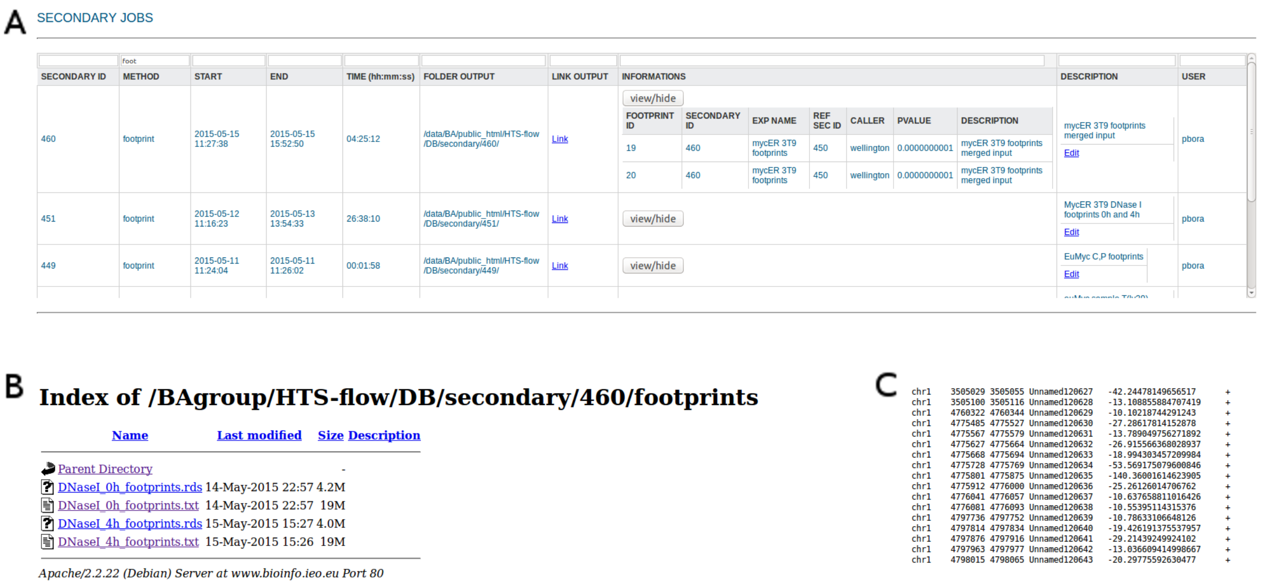
EXP NAME		
PROGRAM	wellington	
PVALUE	0.000000000000000000000000000001	default is 10e-30.
OPTIONS		leave this blank by default.

If you feel confident you can submit this job!

Figure 7: HTS-flow footprint web interface.

Output

Wellington outputs a bed file containing the genomic locations for each footprint for each sample. These file are located in the output folders footprints/. The bed file contains the the genomic locations for each genomic footprint, followed by a score associated by the footprint caller assessing its statistical significance.
Besides reporting the original wellington output (a txt file), HTS-flow converts the bed files in R GRanges objects in the form 'EXP_NAME'.rds in the output folder footprints/.



SECONDARY JOBS

SECONDARY ID	METHOD	START	END	TIME (hh:mm:ss)	FOLDER OUTPUT	LINK OUTPUT	INFORMATIONS	DESCRIPTION	USER
	foot								
460	footprint	2015-05-15 11:27:38	2015-05-15 15:52:50	04:25:12	/data/BA/public_html/HTS-flow /DB/secondary/460/	Link	view/hide	mycER 3T9 footprints merged input Edit	pbora
451	footprint	2015-05-12 11:16:23	2015-05-13 13:54:33	26:38:10	/data/BA/public_html/HTS-flow /DB/secondary/451/	Link	view/hide	MycER 3T9 DNase I footprints 0h and 4h Edit	pbora
449	footprint	2015-05-11 11:24:04	2015-05-11 11:26:02	00:01:58	/data/BA/public_html/HTS-flow /DB/secondary/449/	Link	view/hide	EuMyc C,P footprints Edit	pbora

FOOTPRINT ID	SECONDARY ID	EXP NAME	REF SEC ID	CALLER	PVALUE	DESCRIPTION
19	460	mycER 3T9 footprints	450	wellington	0.0000000001	mycER 3T9 footprints merged input
20	460	mycER 3T9 footprints	450	wellington	0.0000000001	mycER 3T9 footprints merged input

Index of /BAgroup/HTS-flow/DB/secondary/460/footprints

Name	Last modified	Size	Description
Parent Directory		-	
DNaseI_0h_footprints.rds	14-May-2015 22:57	4.2M	
DNaseI_0h_footprints.txt	14-May-2015 22:57	19M	
DNaseI_4h_footprints.rds	15-May-2015 15:27	4.0M	
DNaseI_4h_footprints.txt	15-May-2015 15:26	19M	

Apache/2.2.22 (Debian) Server at www.bioinfo.ieo.eu Port 80

```
chr1    3505029 3505055 Unnamed120627   -42.24478149656517      +
chr1    3505100 3505116 Unnamed120628   -13.108855884707419     +
chr1    4760322 4760344 Unnamed120629   -10.10218744291243      +
chr1    4775485 4775527 Unnamed120630   -27.28617814152878      +
chr1    4775567 4775579 Unnamed120631   -13.789049756271892     +
chr1    4775627 4775664 Unnamed120632   -26.915566368028937     +
chr1    4775668 4775694 Unnamed120633   -18.994303457209984     +
chr1    4775728 4775769 Unnamed120634   -53.569175079600846     +
chr1    4775801 4775875 Unnamed120635   -140.36001614623905     +
chr1    4775912 4776000 Unnamed120636   -25.26126014706762      +
chr1    4776041 4776057 Unnamed120637   -10.637658811016426     +
chr1    4776081 4776093 Unnamed120638   -10.55395114315376      +
chr1    4797736 4797752 Unnamed120639   -10.78633106648126      +
chr1    4797814 4797834 Unnamed120640   -19.426191375537957     +
chr1    4797876 4797916 Unnamed120641   -29.21439249924102      +
chr1    4797963 4797977 Unnamed120642   -13.036609414998667     +
chr1    4798015 4798065 Unnamed120643   -20.29775592630477      +
```

Figure 8: Footprint results on HTS-flow. (A) Output for footprint analysis with secondary ID 460 in the completed analyses page. Clicking on the "Link" button opens the folders with the results: bed file with footprints identified from the analysis is available both in rds and txt format. In (C) is shown the header of the bed file containing footprints genomic locations.

MERGING ANALYSIS

In this page a user can select a group of aligned samples and pool their reads to obtain a merged alignment file. To be merged, samples need to be aligned to the same reference genome.

Querying completed jobs

The top table (MERGED SAMPLES) shows the completed merging jobs. Each line of this table consists in a sample used in a merging analysis. To identifiy all the samples used in a merging job, gather together all the rows with the same ‘MERGE ID’ (for example, by filling the desired ‘MERGE ID’ on the top of the table.

Submitting a job

The bottom panel (MERGING) can be used to select samples to be merged. Clicking on the SELECT button opens a new panel with two fields:
SAMPLE NAME: the name for the merged sample.
REMOVAL OF DUPLICATES: default TRUE: removal of duplicates in the merged file (suspected to be PCR duplicates). For DNaseI-Seq at high depth (footprint calling), change it to FALSE.

Finally, the job can be submitted by clicking the SUBMIT button. HTS-flow will assign a MERGE ID to the job and automatically redirect the web browser to the RUNNING ANALYSES webpage, showing the list of running jobs. When the analysis is complete, the analysis will be visible on the COMPLETED ANALYSES webpage, under the SECONDARY JOBS panel.