





Analysis of ChIP-seq data (answers to questions)

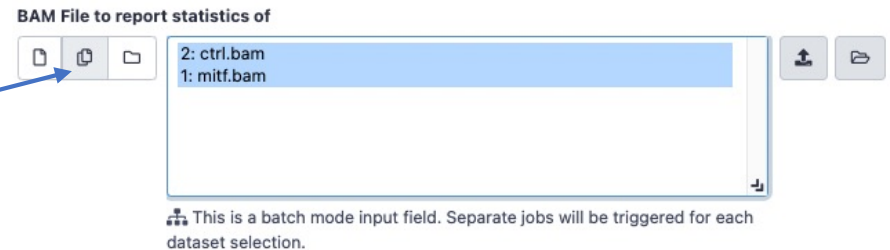
Stéphanie Le Gras
(slegras@igbmc.fr)

Exercise 1: mapping statistics

- 2.
 - Click on the button  to create a new history
 - Click on the history name “Unnamed history”, erase “Unnamed history”, enter “ChIP-seq data analysis” and press enter
- 3.
 - Click on “View all histories” 
 - Drag the two files `22:mitf.bam` and `23:ctrl.bam` from the imported history « NGS data analysis training Strasbourg » and drop them to your current history “ChIP-seq data analysis”.



Exercise 1: mapping statistics

- 4
 - Search for “flagstat” in the search field (tool panel)
 - Click on the name of the tool
 - Click on  to select multiple datasets
 - Select all 2 datasets
 - Click on 



Sample name	No. of raw reads	No. of aligned reads
MITF	31,334,257	23,015,734
Ctrl	29,433,042	19,857,374

Exercise 2: duplicate reads estimate

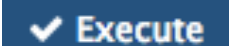
- 1.
 - Search for “markdup” in the search field (tool panel)
 - Click on the name of the tool
 - Click on  to select multiple datasets
 - Select the 2 bam files
 - **Select validation stringency:** Silent
 - Click on 
 - Open the datasets [MarkDuplicates on data * : MarkDuplicate metrics]

Sample name	No. of raw reads	No. of aligned reads	No. of duplicate reads
MITF	31,334,257	23,015,734	16,806,838
Ctrl	29,433,042	19,857,374	15,070,921

Exercise 3: Visualization of the data

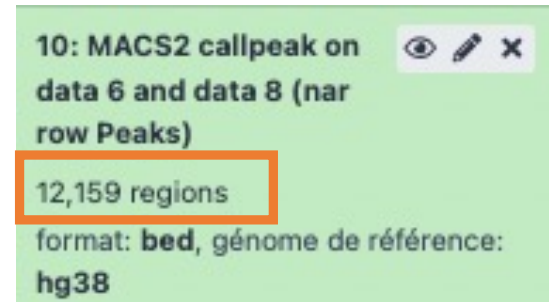
- 1.
 - Idh1 -> No peak
 - NPAS2 -> peak
 - AP1S2 -> Peak,
 - PABPC1I -> No peak
 - Park7 -> No peak
 - Pmel -> Peak
 - Cdk2 -> Peak
 - Actb -> No peak

Exercise 4: peak calling

- 1.
 - Search for “macs2 callpeak” in the search field (tool panel)
 - Click on the name of the tool
 - Set parameters:
 - **Are you pooling Treatment Files?** No
 - CHIP-Seq Treatment File: [mitf bam file marked by MarkDuplicates] (1)
 - **Do you have a Control File?** Yes
 - Are you pooling Control files? No
 - CHIP-Seq Control File: [control bam file marked by MarkDuplicates] (2)
 - **Effective genome size:** H.Sapiens (2.7e9)
 - **Outputs:** select Peaks as tabular file, summits, Summary page (html), Plot in PDF
 - Click on 

Exercise 4: peak calling

- 2.
 - There is 12,159 peaks



10: MACS2 callpeak on data 6 and data 8 (narrow Peaks)

12,159 regions

format: bed, génome de référence: hg38


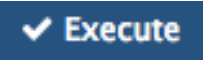
- 3. Look at the tabular file

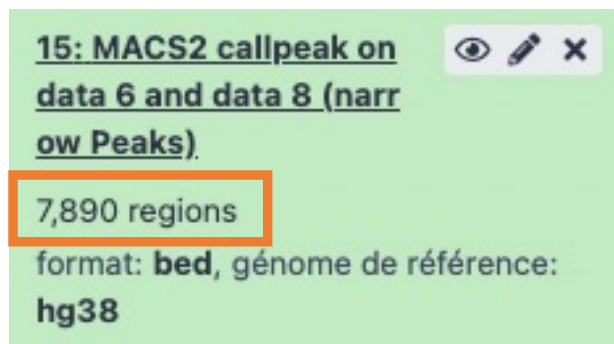
```
# Redundant rate in control: 0.76
# d = 76
# alternative fragment length(s) may be 76 bps
```

chr	start	end	length	abs_summit	pileup
chr1	980687	980817	131	980745	8.48

- The d value estimated by MACS seems a bit small. Let's try to re-run MACS with the expected fragment size : 200

Exercise 5: peak calling

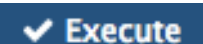
- 1.
 - Click on the name of one of the datasets generated by Macs2.
 - Click on  to display Macs2 form with the same parameters as for the previous run of Macs2
 - In Build Model:
 - select Do not build the shifting model (--nomodel)
 - **Set extension size: 200**
 - Click on 
- 2.
 - 7,890 peaks are now found



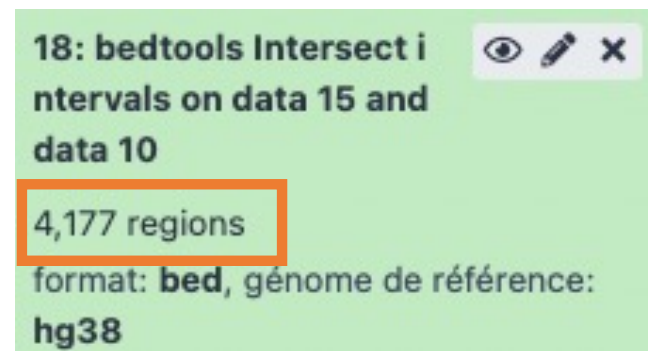
- NOTE: the graphs (showing the estimate of d value) are no longer generated

Exercise 6: compare the two runs of MACS

1.



- Search for “**Intersect**” in the search field (tool panel)
- Click on the name of the tool **bedtools Intersect intervals**
- Set parameters:
 - **File A to intersect with B:** [MACS2 callpeak on data * and data * (narrow Peaks)] (*1st run of MACS*)
 - **Combined or separate output files**
 - One output per file 'input B' file
 - **File B to intersect with A:** [MACS2 callpeak on data * and data * (narrow Peaks)] (*2nd run of MACS*)
 - **Calculation based on strandedness?** Overlaps on either strand
 - **What should be written to the output file?** Write the original entry in A for each overlap (-wa)
 - **Report only those alignments that ****do not**** overlap the BED file:** Yes
- Click on 

4,177 regions are found



Exercise 6: compare the two runs of MACS

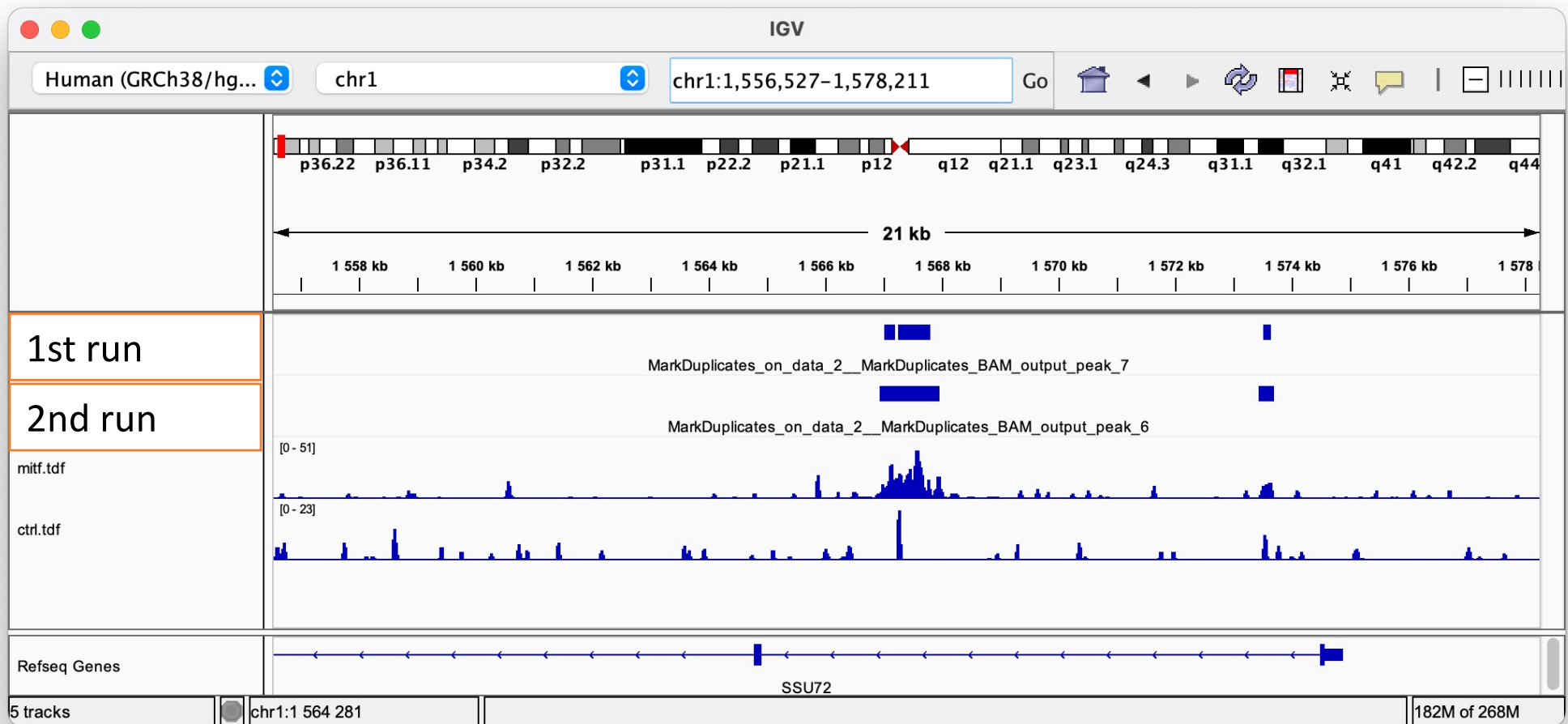
2.

1. In Galaxy, click on  for the two datasets named [MACS2 callpeak on data * and data * (narrow Peaks)] and save the files on your computer
2. Go to IGV and load the two files along with the [two tdf](#) files already loaded (mitf.tdf and ctrl.tdf)
3. In Galaxy, click on the  of the dataset named [bedtools intersect intervals on data * and data *]

Chrom	Start	End	Name	Score	Strand	ThickStart	ThickEnd	ItemRGB	BlockCount
chr1	983819	983925	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_2	53	.	6.75700	9.07098	5.31956	50
chr1	1586289	1586365	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_10	14	.	4.10564	4.39929	1.45337	7
chr1	1728644	1728730	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_11	15	.	4.27812	4.90906	1.52693	66
chr1	1807103	1807179	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_12	44	.	5.62660	8.09168	4.44141	33
chr1	2167152	2167228	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_19	38	.	5.44460	7.46705	3.86461	48
chr1	3276552	3276628	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_21	15	.	4.27812	4.90906	1.52693	52
chr1	3444380	3444456	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_22	13	.	3.43160	4.33100	1.39739	40
chr1	3681035	3681111	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_23	13	.	4.05353	4.26549	1.34476	59
chr1	3900155	3900272	MarkDuplicates_on_data_1__MarkDuplicates_BAM_output_peak_24	26	.	4.85117	6.12167	2.65739	64

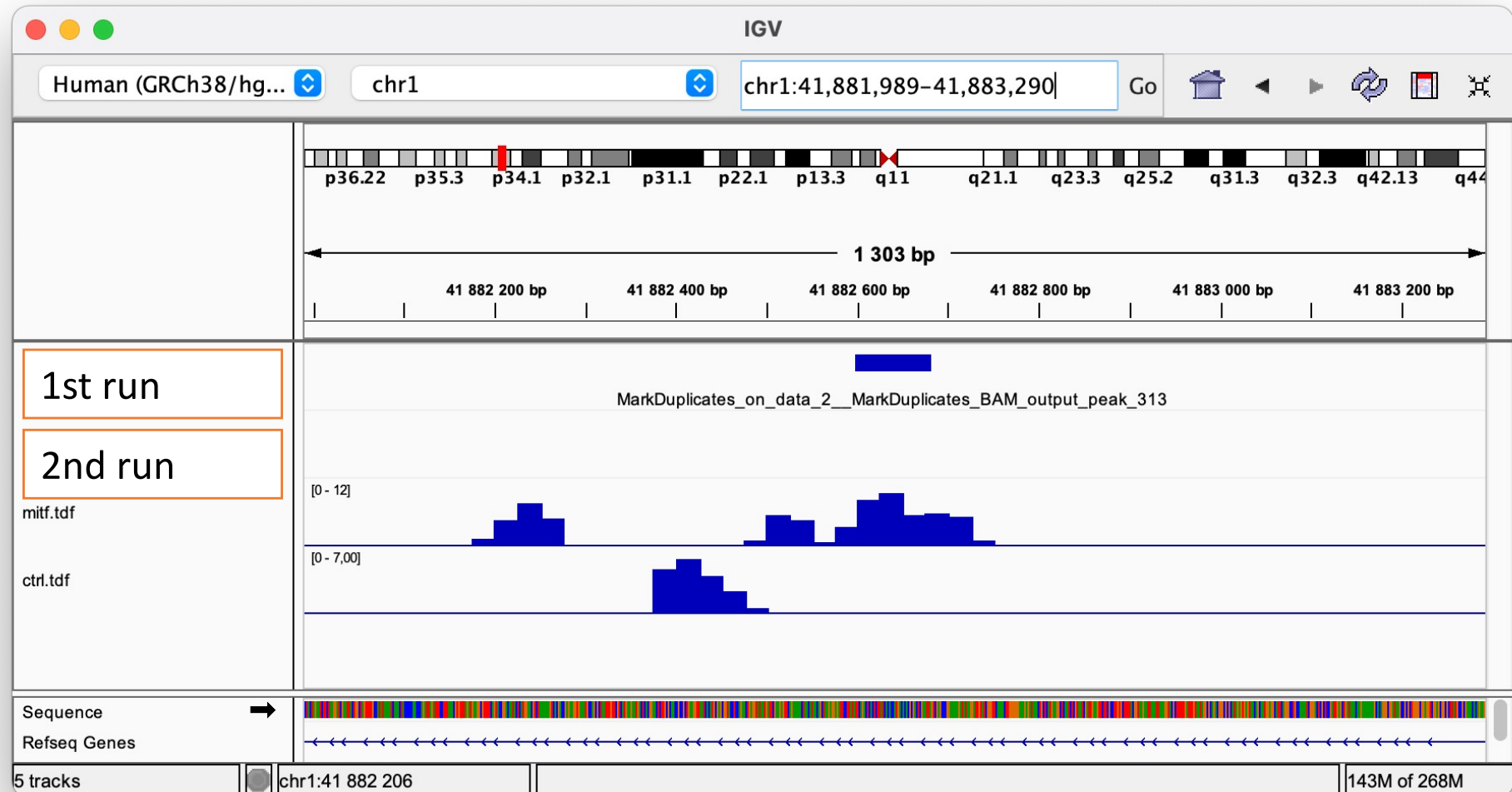
Exercise 6: compare the two runs of MACS

SSU72 (chr1:1556527-1578211)



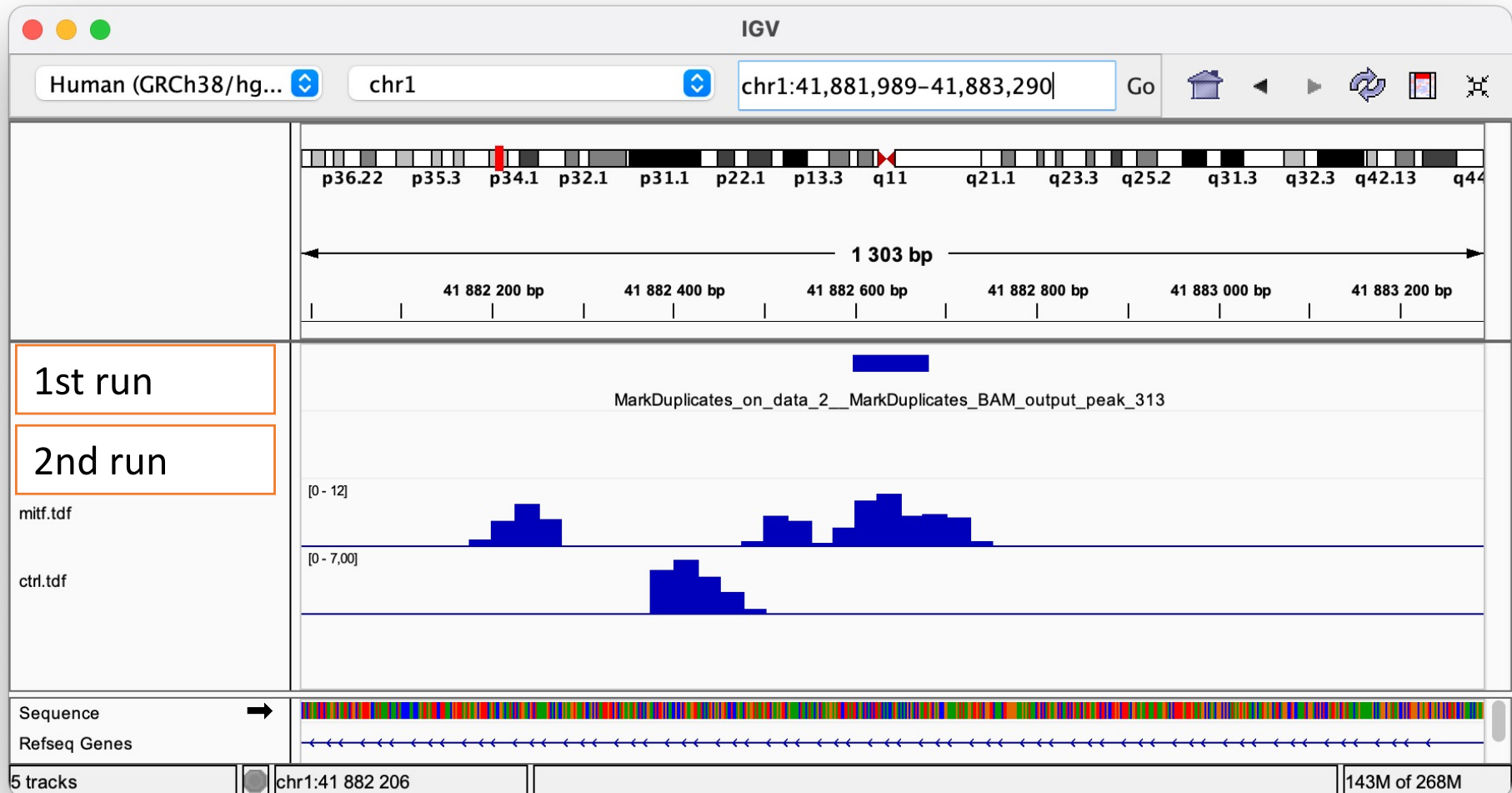
Exercise 6: compare the two MACS runs

chr1:41882599-41882681



Exercise 6: compare the two runs of MACS

chr1:1586290-1586365



Exercise 6: compare the two runs of MACS

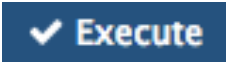
- **We are going to keep the second run of MACS**

For the following dataset of the second run of MACS2, rename the datasets:


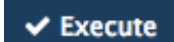
- [MACS2 callpeak on data * and data * (summits in BED)] -> MITF_peak_summits.bed
- [MACS2 callpeak on data * and data * (narrow Peaks)] -> MITF_peaks.narrowPeak

Exercise 7: peak annotation

Search for “closest” in the search field (tool panel)

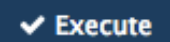
- **BED/bedGraph/GFF/VCF/EncodePeak file:** MITF_peaks.narrowPeak (*second run of MACS2*)
- **Overlap with: will you select a BED/bedGraph/GFF/VCF/EncodePeak file from your history or use a built-in GFF file?**
 - Use a BED/bedGraph/GFF/VCF/EncodePeak file from the history
 - Select a BED/bedGraph/GFF/VCF/EncodePeak file: 25:hg38_ens105_ucsc.bed
- **How ties for closest feature should be handled:** first – Report the first tie that occurred in the B file
- **In addition to the closest feature in B, report its distance to A as an extra column:** Yes
- **Add additional columns to report distance to upstream feature. Distance definition:**
 - Report distance with respect to A. When A is on the – strand, « upstream » means B has a higher (start,stop). (-a)
 - **Choose first from features in B that are upstream of feature in A:** Yes
- Click on  Execute
- Rename the file [Closest regions from data * and data *] -> mitf_peaks.annot.tsv.

Exercise 8: *de novo* motif discovery

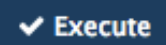

- 1.a
 - Search for “Sort” in the search field (tool panel)
 - Click on the name of the tool
 - Set parameters:
 - Sort Dataset: MITF_peak_summits.bed (*second run of MACS2*)
 - on column: Column: 5
 - with flavor: Numerical sort
 - everything in: Descending order
 - Click on 
- 1.b
 - Search for “select first” in the search field (tool panel)
 - Click on the name of the tool
 - Set parameters:
 - Select first: 800
 - From: [Sort on data *] (*dataset generated in 1.a*)
 - Click on 

Exercise 8: *de novo* motif discovery

2.

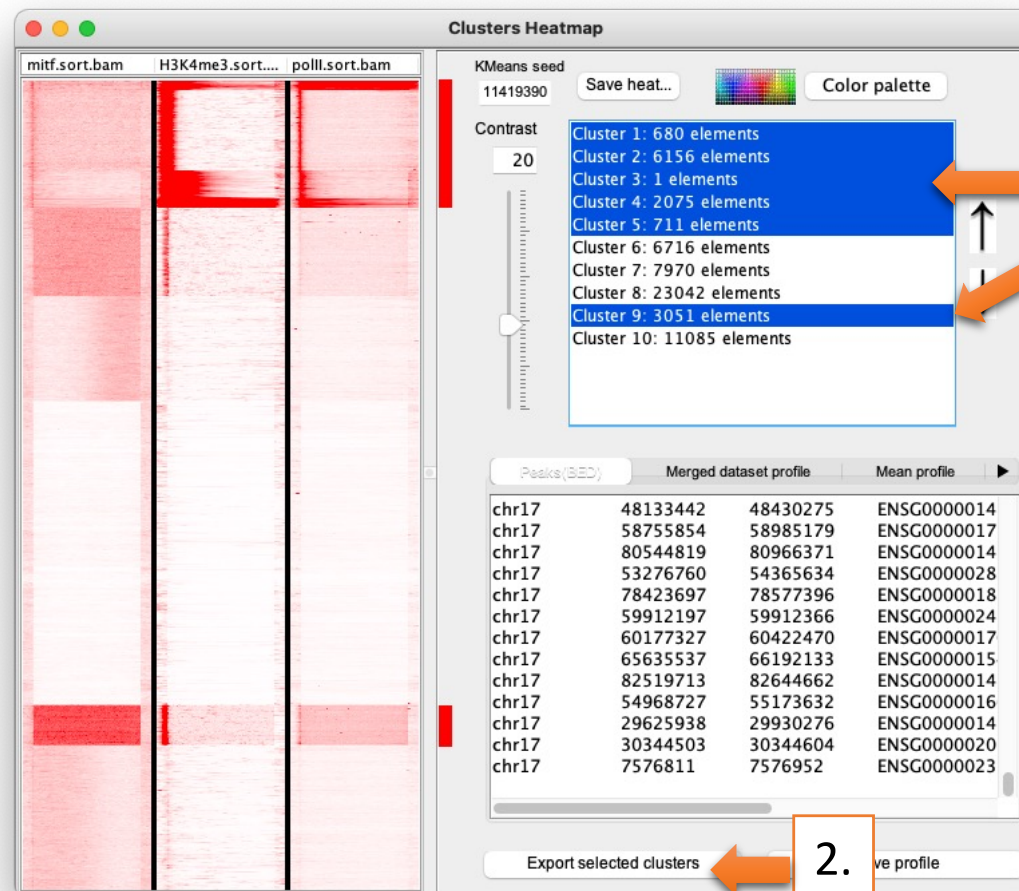
- Search for "slop" in the search field (tool panel)
- Click on the name of the tool **bedtools SlopBed**
- Set the parameters
 - **BED/bedGraph/GFF/VCF/EncodePeak file:** [Select first on data *] (*results of step 1.b.*)
 - **Genome file:**
 - Locally installed Genome file
 - **Genome file:** Human Dec. 2013 (GRCh38/hg38) (hg38)
 - **Choose what you want to do:** Increase the BED/VCF/GFF entry by the same number of base pairs in each direction. (default)
 - Number of base pairs: 50
 - Click on 

Exercise 8: *de novo* motif discovery

- 3.
 - Search for “**getfastabed**” in the search field (tool panel)
 - Click on the name of the tool **bedtools GetFastaBed**
 - Set the parameters:
 - **BED/bedGraph/GFF/VCF/EncodePeak file:** [bedtools SlopBed on data *] (*the dataset generated in 2*)
 - **Choose the source for the FASTA file:** Server indexed files
 - **Fasta_id:** Human (homo sapiens): hg38
 - Click on 
 - Rename the file peakSummits_+/-50nt_top800.fasta
- 4.
 - Expand the box of the dataset peakSummits_+/-50nt_top800.fasta and click on  to download the file
- 5.
 - Go to MEME-chIP website and run the tool with the fasta file peakSummits_+/-50nt_top800.fasta as input file and with default parameters.

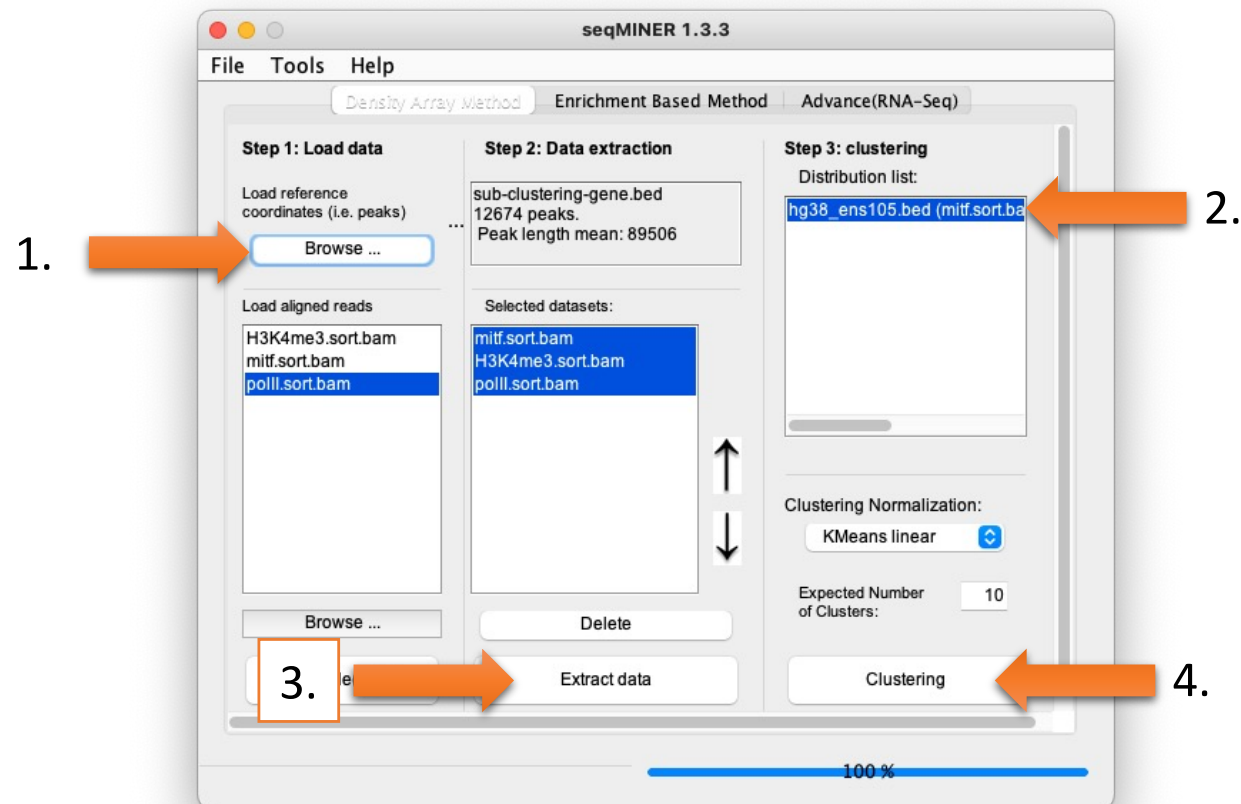
Exercise 9: Clustering

- 1.
 - Select clusters 1, 2, 3, 4, 5, 9 (1)
 - Click on Export Selected clusters (2) and save the file as sub-clustering-gene.bed



Exercise 9: Clustering

- 1.
 - Import the file sub-clustering-gene.bed. (You can use the one provided in [chipseq/seqminer](#)).
 - Click on Browse (1), go to the directory which contains the file and click on open.
 - Remove previous distribution (2)
 - Click on Extract data (3)
 - Click on Clustering (4)



Exercise 9: Clustering

- Go to Annotation (1)
- Click on cluster 4 (2)
- Click on Export selected cluster (3)

The screenshot shows the 'Clusters Heatmap' interface. On the left is a heatmap with three columns labeled 'mitf.sort.bam', 'H3K4me3.sort....', and 'poll.sort.bam'. On the right, there are controls for 'KMeans seed' (11419390), 'Save heat...', and a 'Color palette'. Below these is a 'Contrast' slider set to 20. A list of 10 clusters is shown, with 'Cluster 4: 730 elements' highlighted in blue. An orange arrow labeled '2.' points to this cluster. Below the list are tabs for 'Heatmap', 'Density values', and 'Annotation', with the 'Annotation' tab selected. An orange arrow labeled '1.' points to this tab. The 'Annotation' table is visible, showing columns for 'chr', 'start', 'end', 'name', 'nam...', 'stra...', 'ann...', 'ann...', 'ann...', and 'to T...'. An orange arrow labeled '3.' points to the 'Export selected clusters' button at the bottom.

chr	start	end	name	nam...	stra...	ann...	ann...	ann...	to T...
chr22	22...	22...	ENS...	PRA...	-	ENS...	LL2...	+	-58...
chr22	37...	37...	ENS...	GGA1	+	ENS...	GGA1	+	12...
chr22	23...	23...	ENS...	C2...	+	ENS...	C2...	+	1436
chr22	40...	40...	ENS...		+	ENS...		+	221
chr22	30...	30...	ENS...		-	ENS...		-	177
chr22	30...	30...	ENS...		-	ENS...		-	171
chr22	26...	26...	ENS...	HPS4	-	ENS...	HPS4	-	20...
chr22	41...	41...	ENS...	RA...	-	ENS...	MIR...	-	-12...
chr22	41...	41...	ENS...	POL...	-	ENS...	POL...	-	9404
chr22	19...	19...	ENS...	COMT	+	ENS...	MIR...	+	-80...
chr22	20	20	ENS		-	ENS		-	1406

Exercise 9: Clustering

- Go to DAVID website <https://david.ncifcrf.gov/>
- Click on Start Analysis (1)

1. 



*** Welcome to DAVID 6.8 ***
*** You are welcome to try the newest version of DAVID (2021 Update) on our development site. ***

Analysis Wizard

Tell us how you like the tool
Contact us for questions

← Step 1. Submit your gene list through left panel.

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy_ID" -> List Type as "Gene List" -> Click "Submit" button

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_l_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at

Exercise 9: Clustering

- Fill in the form:
 - **Paste a list (1):** Copy and paste Ensembl Gene IDs from the Cluster4.xls file
 - **Select Identifier (2)** (drop down list): ENSEMBL_GENE_ID
 - **List Type (3):** Gene List
 - Click on **Submit List (4)**

The screenshot shows the DAVID Analysis Wizard interface. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, About DAVID, and About LHRI. A welcome message states: "*** Welcome to DAVID (2021 Update) ***" and "*** If you are looking for DAVID 6.8, please visit our production site. ***". The main heading is "Analysis Wizard".

The interface is divided into several sections:

- Upload Gene List:** Includes "Demolist 1", "Demolist 2", and "Upload Help".
- Step 1: Enter Gene List:** Contains a text input field "A: Paste a list" with a "Clear" button. Below it is a "B: Choose From a File" section with a file selection button "Choisir un fichier" and a "Multi-List File" checkbox.
- Step 2: Select Identifier:** A dropdown menu currently showing "ENSEMBL_GENE_ID".
- Step 3: List Type:** Radio buttons for "Gene List" (selected) and "Background".
- Step 4: Submit List:** A "Submit List" button.

Four orange arrows with numbers 1, 2, 3, and 4 point to the following elements:

1. The text input field for pasting the gene list.
2. The dropdown menu for selecting the identifier.
3. The "Gene List" radio button.
4. The "Submit List" button.

An example instruction is provided: "Paste IDs to 'box A' -> Select Identifier as 'Affy_ID' -> List Type as 'Gene List' -> Click 'Submit'". A list of example gene IDs is shown on the right side of the interface.

Exercise 9: Clustering

- Click on **Continue to Submit IDs That DAVID Could Map** (1)

DAVID Bioinformatics Resources (2021 Update), NIAID/NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service About DAVID About LHRI

*** Welcome to DAVID (2021 Update) ***
*** If you are looking for DAVID 6.8, please visit our production site. ***

Gene ID Conversion Tool

Help and Tool Manual

You are either not sure which identifier type your list contains, or less than 80% of your list has mapped to your chosen identifier type. Please use the Gene Conversion Tool to determine the identifier type.

Option 1 (Recommended): 1.

Option 2:
Convert the gene list to

For species:

Option 3:

Upload List Background

Upload Gene List

Demolist 1 Demolist 2
Upload Help

Step 1: Enter Gene List
A: Paste a list

Or
B: Choose From a File

 Multi-List File ?

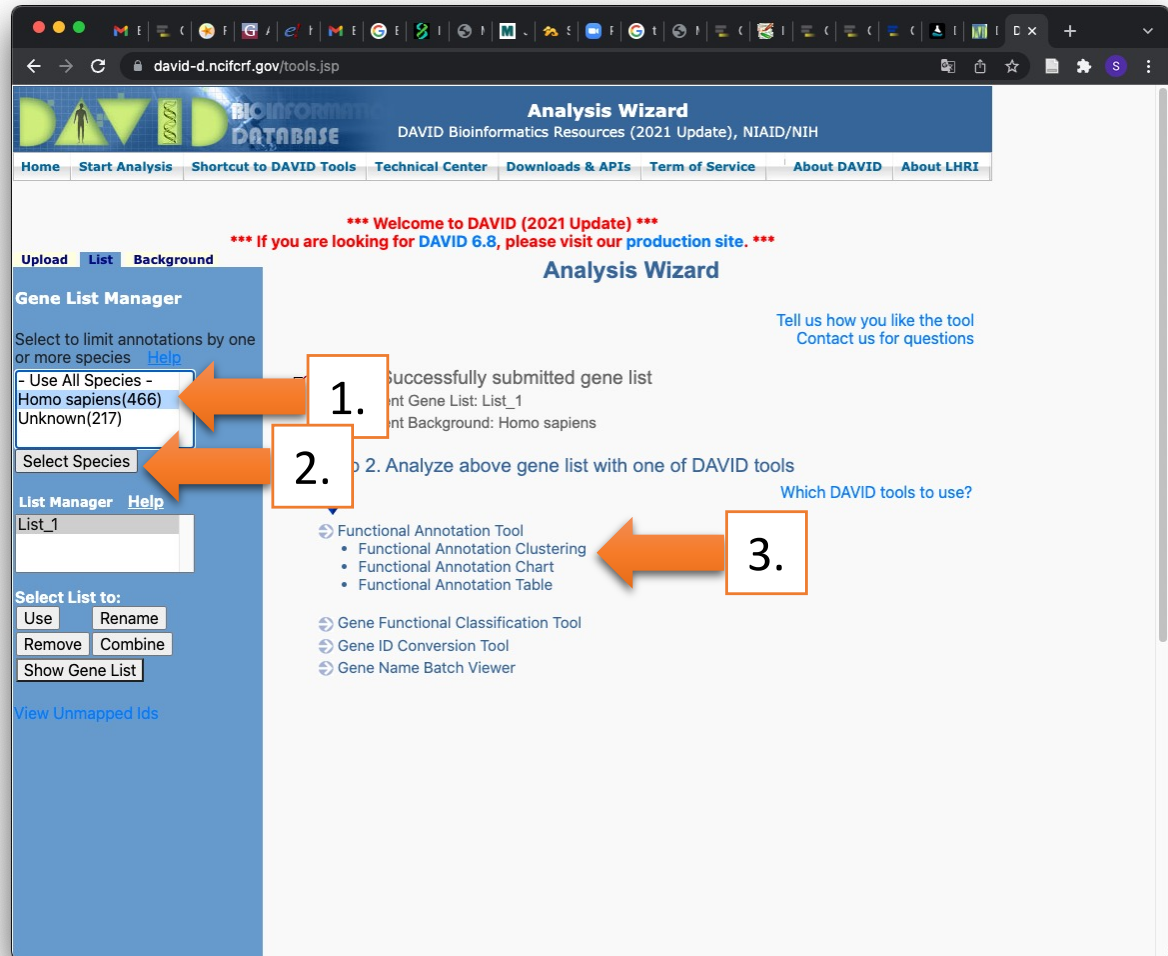
Step 2: Select Identifier

Step 3: List Type
Gene List
Background

Step 4: Submit List

Exercise 9: Clustering

- **Select to limit annotations by one or more species (left panel)**
 - Select Homo sapiens (466) (1)
 - Click on **Select Species** (2)
- Click on **Functional Annotation Tool** (3)



Exercise 9: Clustering

- Keep all default
- Click on **Functional Annotation Clustering** (1)


The screenshot displays the DAVID Functional Annotation Tool interface. The top navigation bar includes links for Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, About DAVID, and About LHRI. A welcome message states: "*** Welcome to DAVID (2021 Update) ***" and "*** If you are looking for DAVID 6.8, please visit our production site. ***".

The main content area is divided into two columns. The left column, titled "Gene List Manager", allows users to select species (currently set to "Homo sapiens(466)") and manage lists (currently "List_1"). The right column, titled "Annotation Summary Results", shows the current gene list and background, and a list of selected annotation categories:

- Disease** (2 selected)
- Functional_Categories** (6 selected)
- Gene_Ontology** (3 selected)
- General_Annotations** (0 selected)
- Interactions** (1 selected)
- Literature** (0 selected)
- Pathways** (3 selected)
- Protein_Domains** (4 selected)
- Tissue_Expression** (0 selected)

Below the categories, there is a note: "***Red annotation categories denote DAVID defined defaults***". Under the "Combined View for Selected Annotation" section, there are three buttons: "Functional Annotation Clustering", "Functional Annotation Chart", and "Functional Annotation Table". An orange arrow points to the "Functional Annotation Clustering" button, which is labeled with a "1." in a box.

Exercise 9: Clustering


DAVID Bioinformatics Resources (2021 Update)
 Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

*** Welcome to DAVID (2021 Update) ***
 *** If you are looking for DAVID 6.8, please visit our [production site](#). ***

Functional Annotation Clustering

[Help and Manual](#)

Current Gene List: List_1
Current Background: Homo sapiens
466 DAVID IDs

Options **Classification Stringency** Medium

60 Cluster(s) [Download File](#)

Annotation Cluster 1		Enrichment Score: 6.01			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_MF_DIRECT	RNA binding	RT		55	1.1E-7	3.2E-5
<input type="checkbox"/>	UP_KW_PTM	Isopeptide bond	RT		60	6.5E-7	5.4E-6
<input type="checkbox"/>	UP_SEQ_FEATURE	CROSSLNK:Glycyl lysine isopeptide (Lys-Gly) (interchain with G-Cter in SUMO2)	RT		46	8.0E-7	1.1E-3
<input type="checkbox"/>	UP_KW_PTM	Ubl conjugation	RT		72	1.5E-5	9.6E-5
Annotation Cluster 2		Enrichment Score: 5.56			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleoplasm	RT		119	2.5E-9	4.6E-7
<input type="checkbox"/>	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		130	5.5E-7	2.2E-5
<input type="checkbox"/>	GOTERM_CC_DIRECT	nucleus	RT		128	1.6E-2	6.5E-1
Annotation Cluster 3		Enrichment Score: 2.37			Count	P_Value	Benjamini
<input type="checkbox"/>	UP_KW_MOLECULAR_FUNCTION	Ribonucleoprotein	RT		20	1.8E-6	1.1E-4
<input type="checkbox"/>	UP_KW_MOLECULAR_FUNCTION	Ribosomal protein	RT		14	3.9E-5	8.1E-4
<input type="checkbox"/>	GOTERM_CC_DIRECT	ribosome	RT		13	1.5E-4	1.4E-2
<input type="checkbox"/>	GOTERM_BP_DIRECT	translational initiation	RT		11	1.7E-4	9.4E-2
<input type="checkbox"/>	GOTERM_MF_DIRECT	structural constituent of ribosome	RT		12	7.4E-4	1.4E-1
<input type="checkbox"/>	GOTERM_BP_DIRECT	translation	RT		12	2.9E-3	6.6E-1
<input type="checkbox"/>	GOTERM_BP_DIRECT	cytoplasmic translation	RT		7	5.9E-3	9.9E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytosolic small ribosomal subunit	RT		5	9.3E-3	5.0E-1
<input type="checkbox"/>	GOTERM_CC_DIRECT	cytosolic ribosome	RT		6	1.4E-2	6.5E-1
<input type="checkbox"/>	KEGG_PATHWAY	Ribosome	RT		8	2.4E-2	1.0E0
<input type="checkbox"/>	GOTERM_BP_DIRECT	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	RT		7	2.5E-2	1.0E0
<input type="checkbox"/>	GOTERM_BP_DIRECT	SRP-dependent cotranslational protein targeting to membrane	RT		6	2.6E-2	1.0E0
<input type="checkbox"/>	UP_KW_DISEASE	Diamond-Blackfan anemia	RT		3	4.0E-2	5.3E-1
<input type="checkbox"/>	GOTERM_BP_DIRECT	viral transcription	RT		6	5.6E-2	1.0E0
<input type="checkbox"/>	KEGG_PATHWAY	Coronavirus disease - COVID-19	RT		7	2.4E-1	1.0E0
<input type="checkbox"/>	GOTERM_BP_DIRECT	rRNA processing	RT		5	4.8E-1	1.0E0