Analysis of ChIP-seq data (answers to questions)

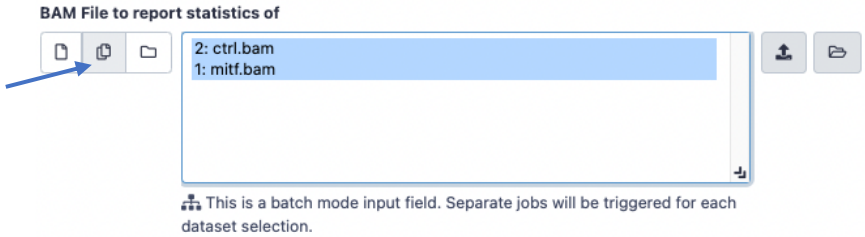
Stéphanie Le Gras
(slegras@igbmc.fr)

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 ✔ Execute



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 Execute
  - 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

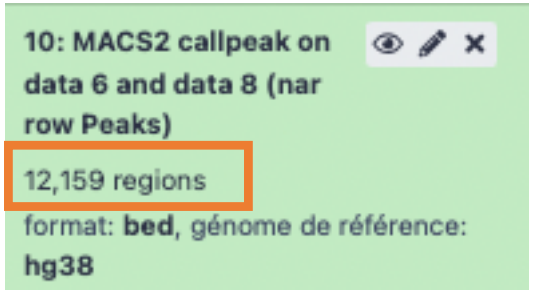
- 1.
  - Idh1 -> No peak
  - NPAS2 -> peak
  - AP1S2 -> Peak,
  - PABPC1l -> No peak
  - Park7 -> No peak
  - Pmel -> Peak
  - Cdk2 -> Peak
  - Actb -> No peak

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



- 3. Look at the tabular file

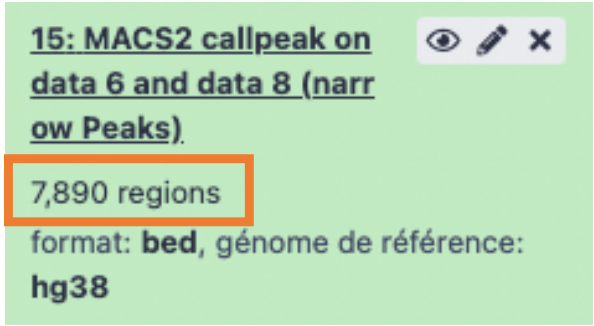
Redundant rate in control: 0.76
d = 76
alternative fragment length(s) may be 76 bps

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 ✔ Execute
- 2.
  - 7,890 peaks are now found

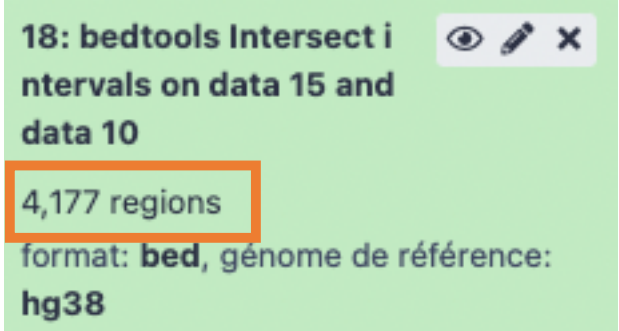


  - NOTE: the graphs (showing the estimate of d value) are no longer generated

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 ✔ Execute

4,177 regions are found



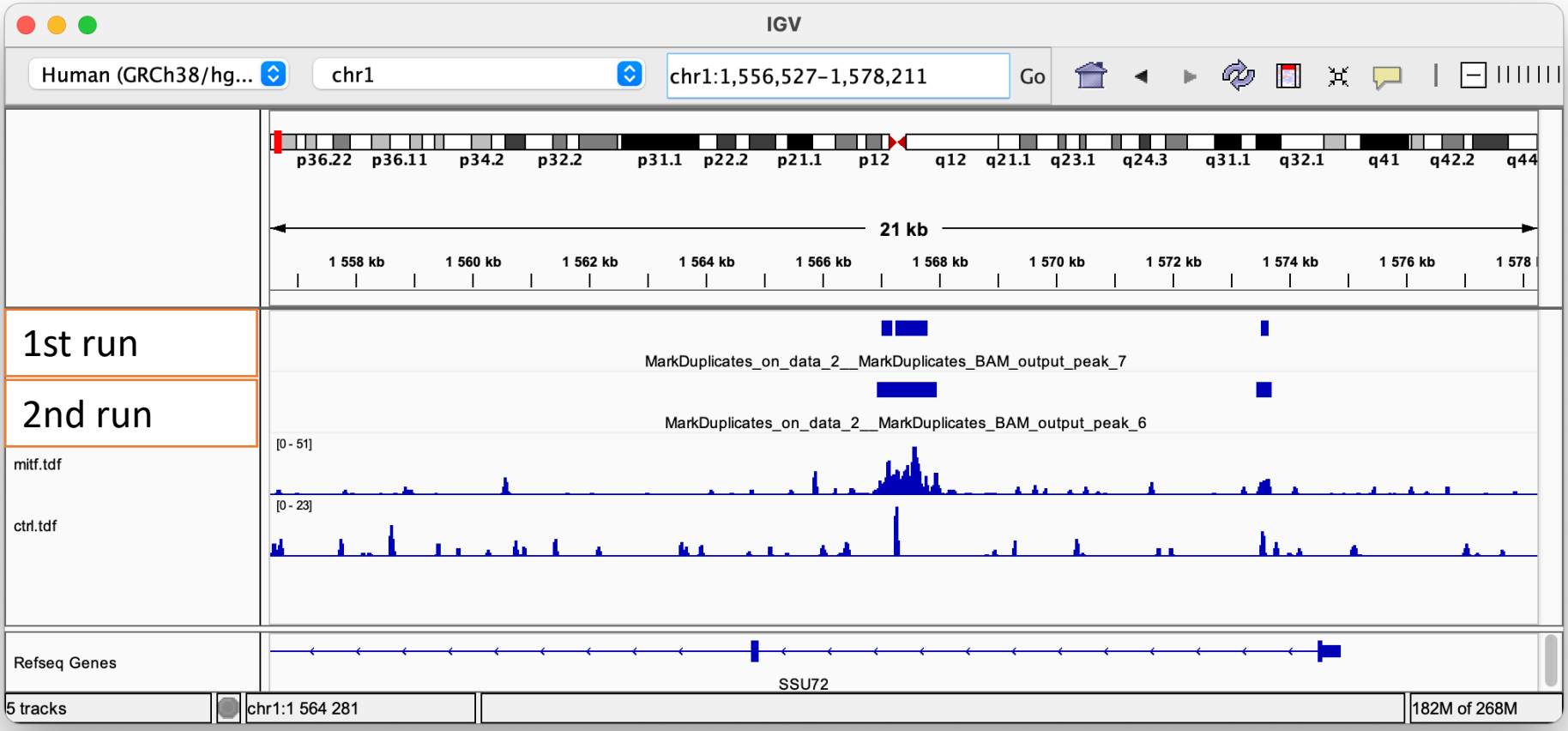
Exercise 6: compare the two runs of MACS

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

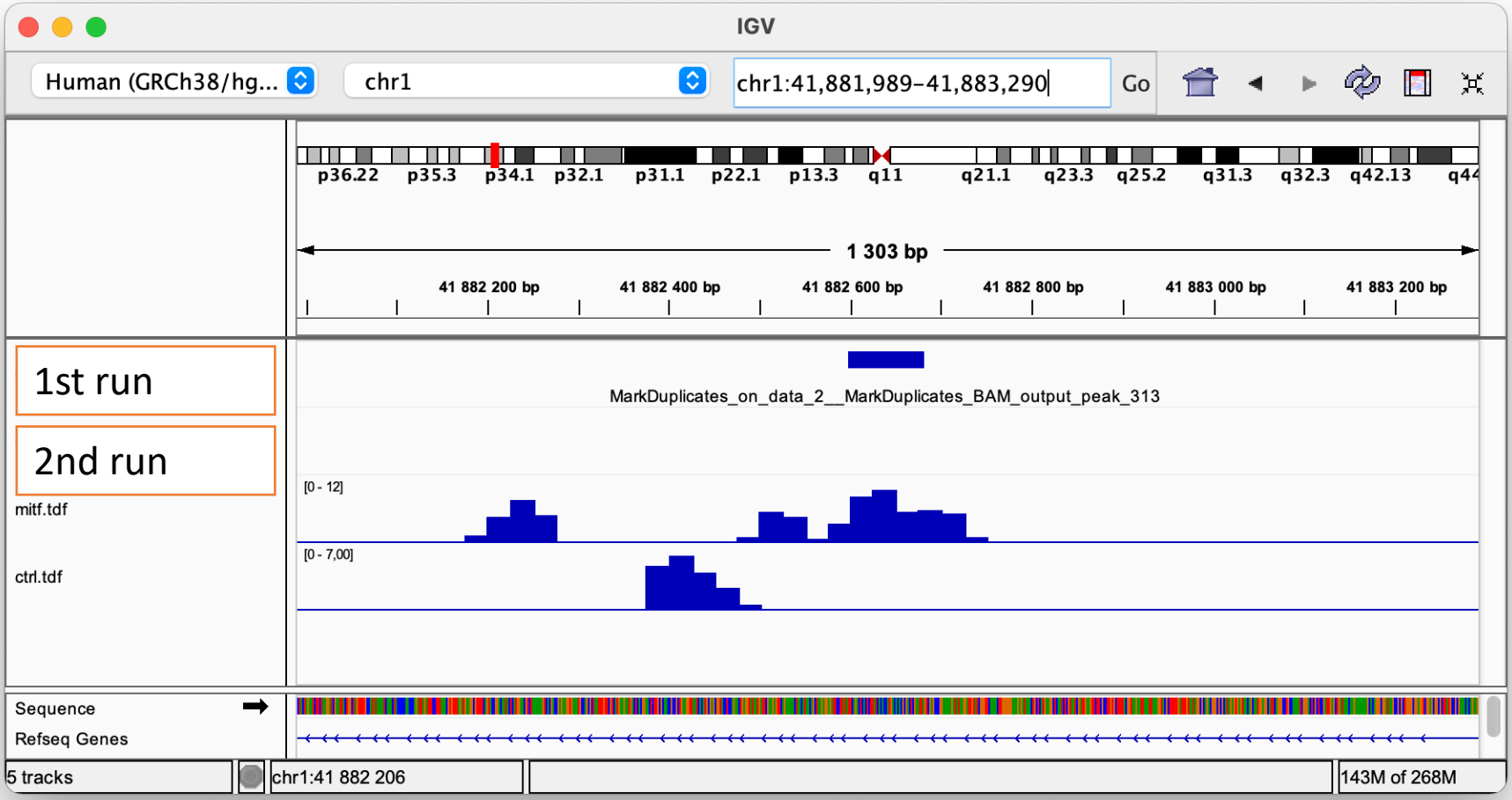
Exercise 6: compare the two runs of MACS

SSU72 (chr1:1556527-1578211)



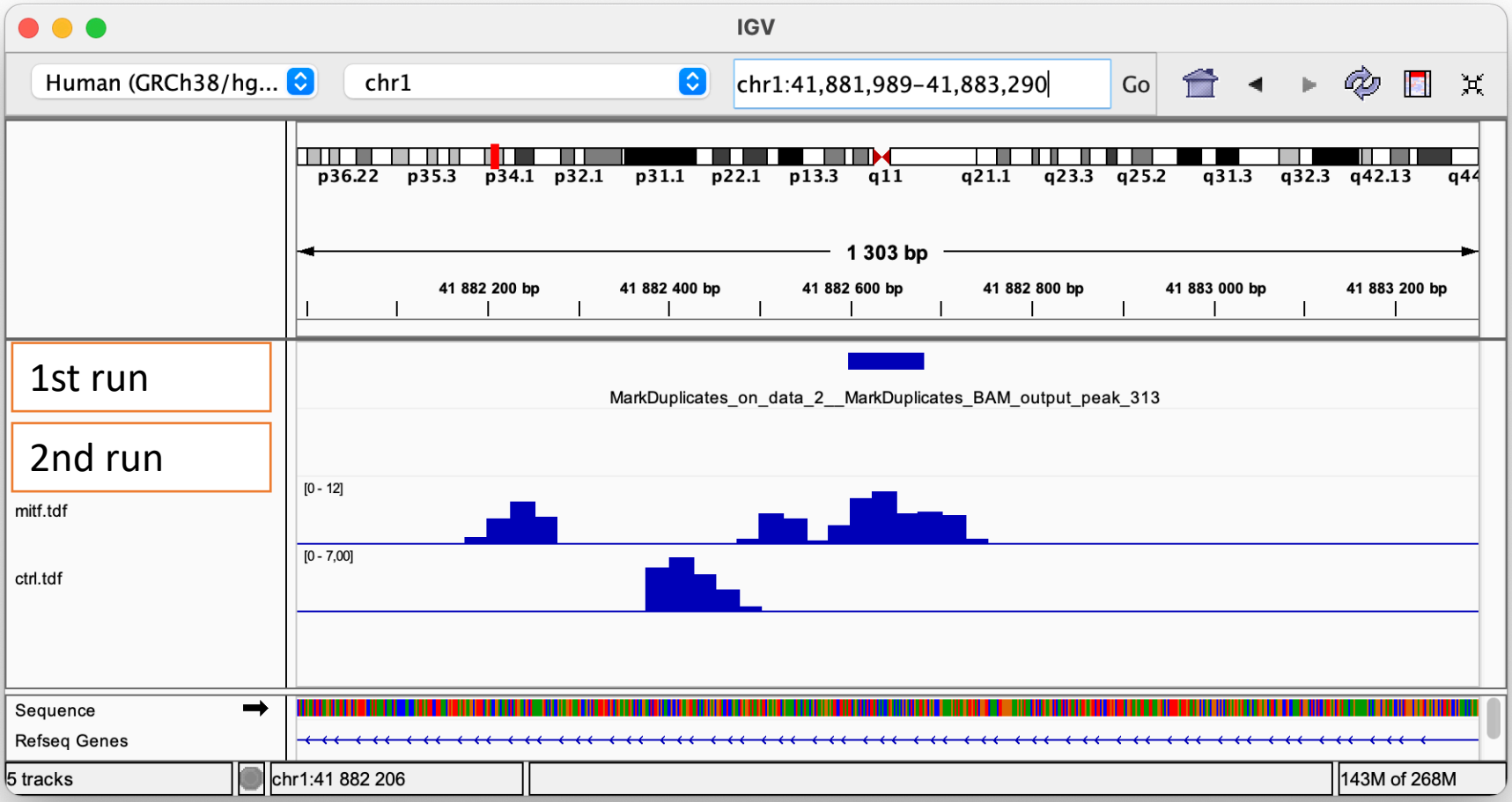
Exercise 6: compare the two MACS runs

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 defintion:**
  - 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 

- 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 ✔ Execute
- 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 ✔ Execute

Exercise 8: *de novo* motif discovery

2.

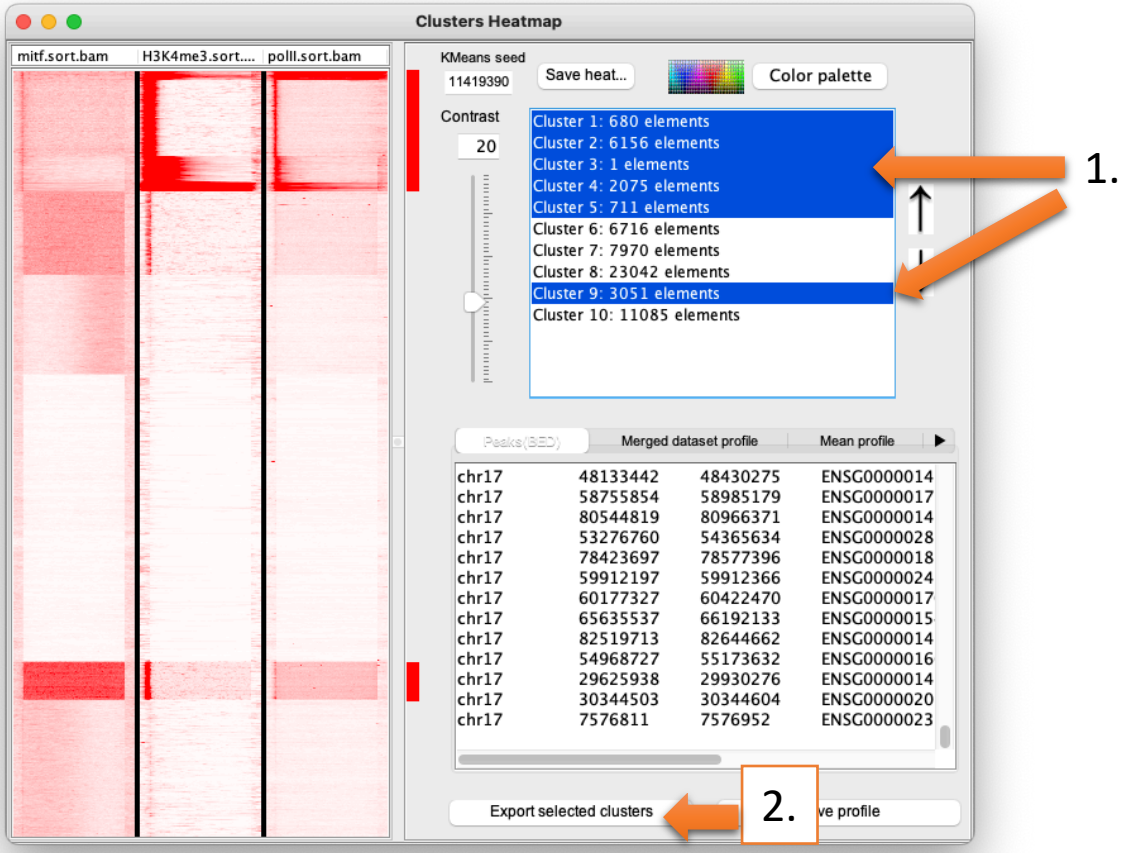
- Search for ”slop” in the search field (tool panel)
- Click on the name of the tool **bedtools SlopBed**
- Set the parameters
  - **BED/bedGraph/GFF/VCF/EncodePeak file**: [Select first on data *] *(results of step 1.b.)*
  - **Genome file:**
    - Locally installed Genome file
    - **Genome file:** 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 ✔ Execute

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 ✔ Execute
  - 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.

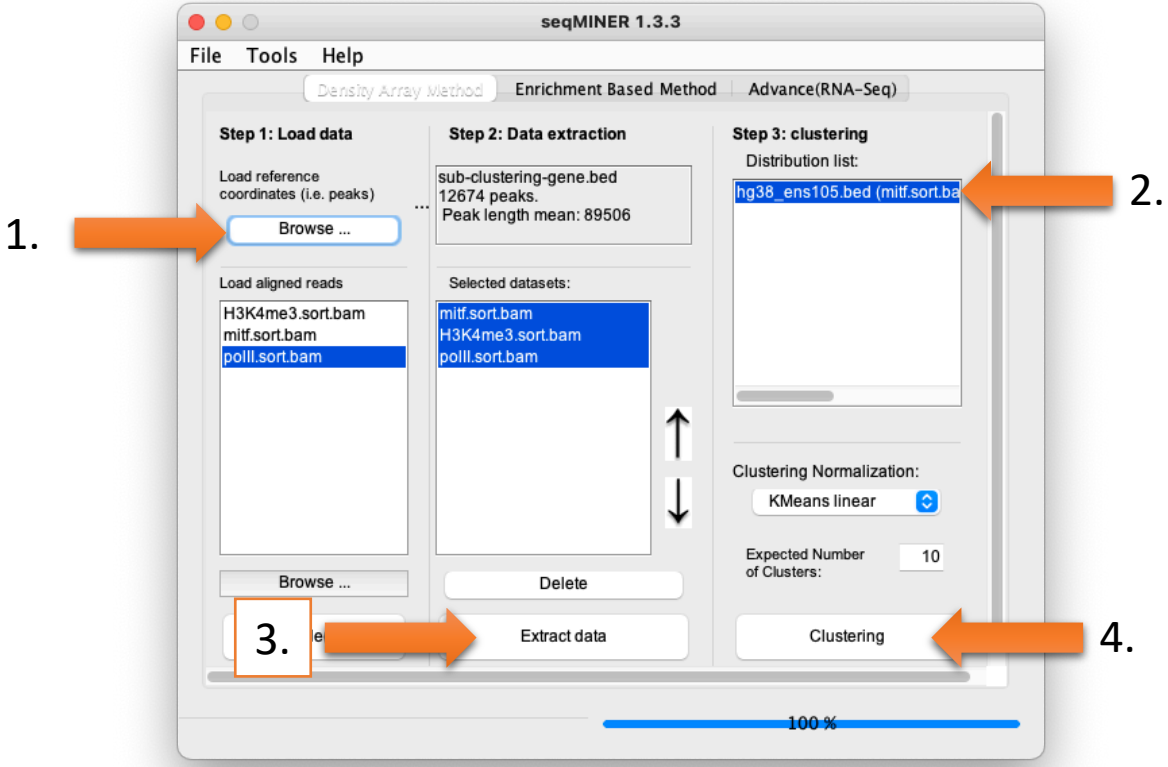
Exercise 9: Clustering

- 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



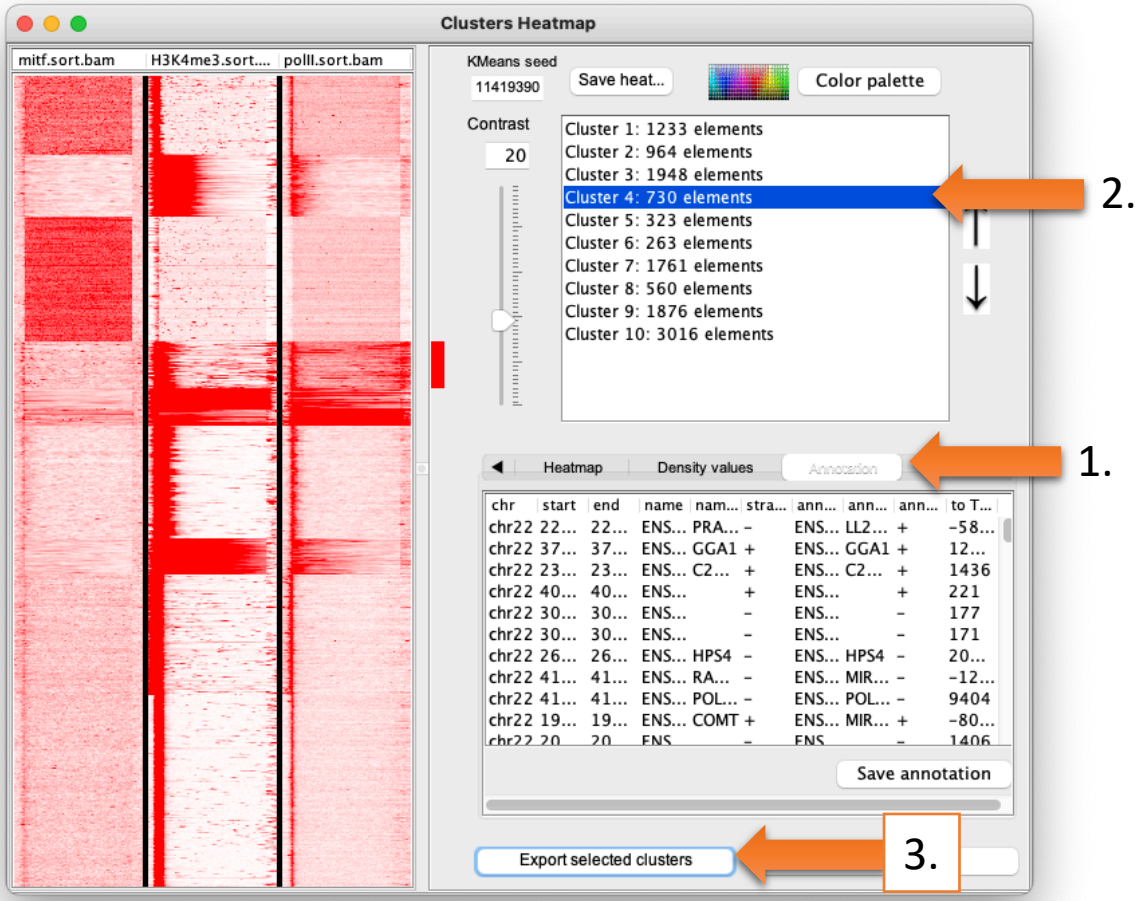
Exercise 9: Clustering

- 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

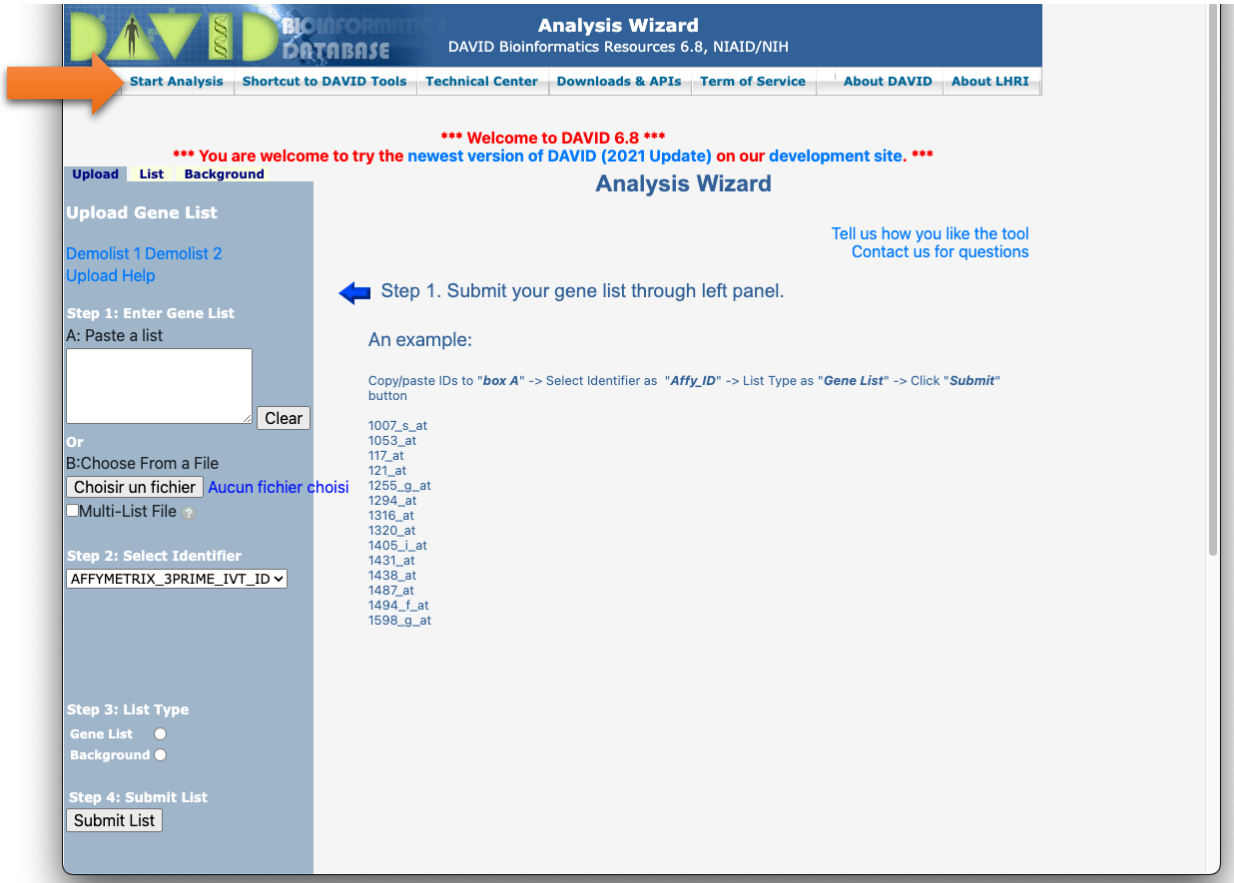
- Go to Annotation (1)
- Click on cluster 4 (2)
- Click on Export selected cluster (3)



Exercise 9: Clustering

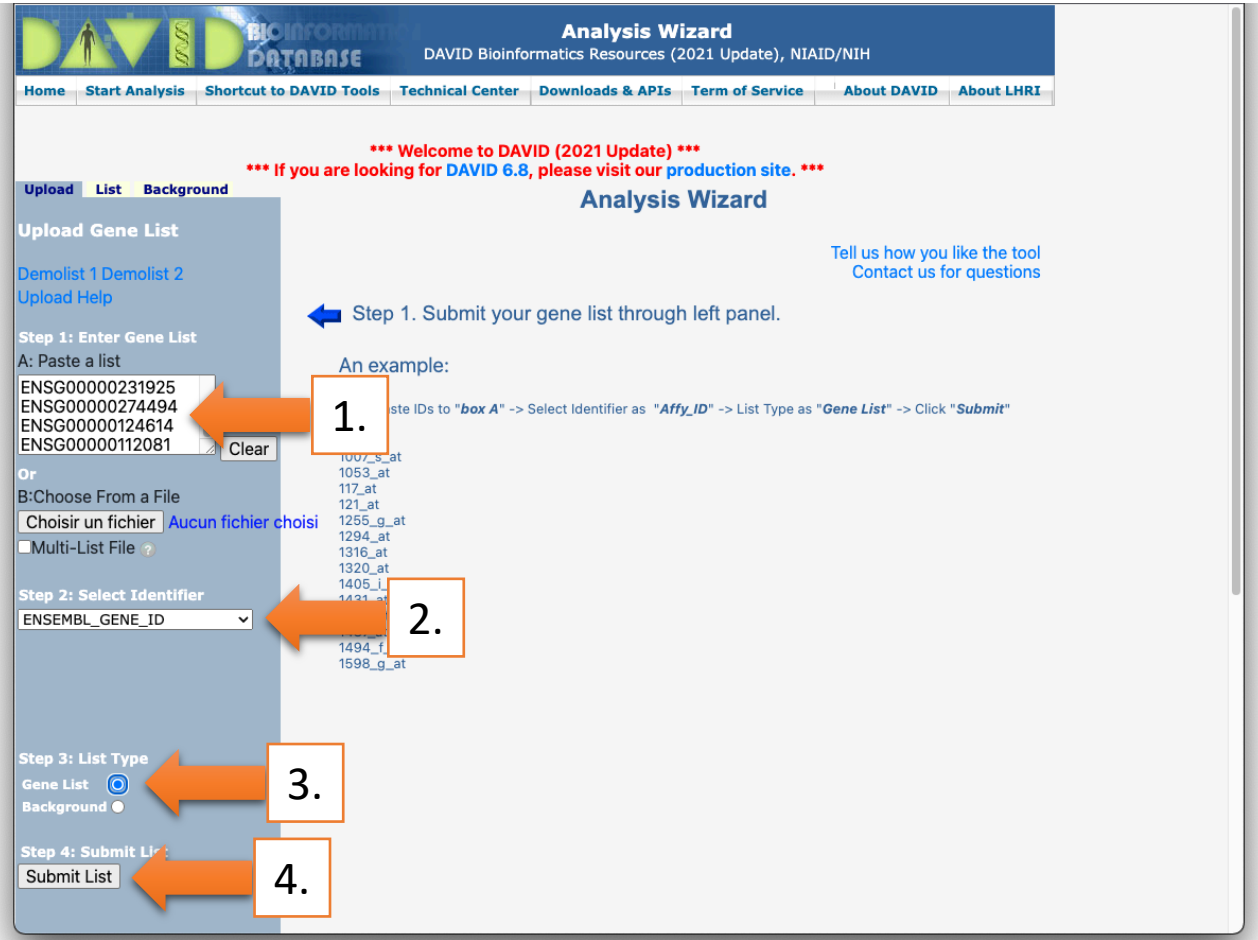
- Go to DAVID website https://david.ncifcrf.gov/
- Click on Start Analysis (1)

1.



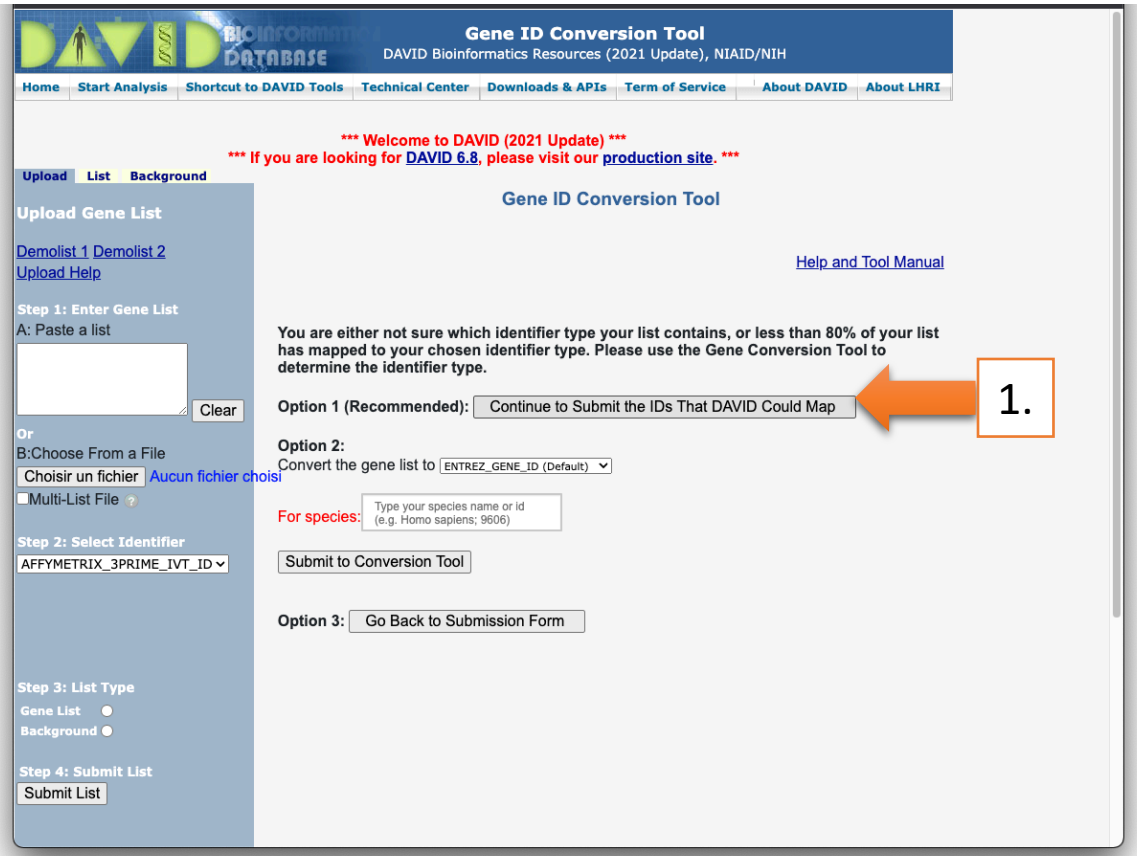
Exercise 9: Clustering

- 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)



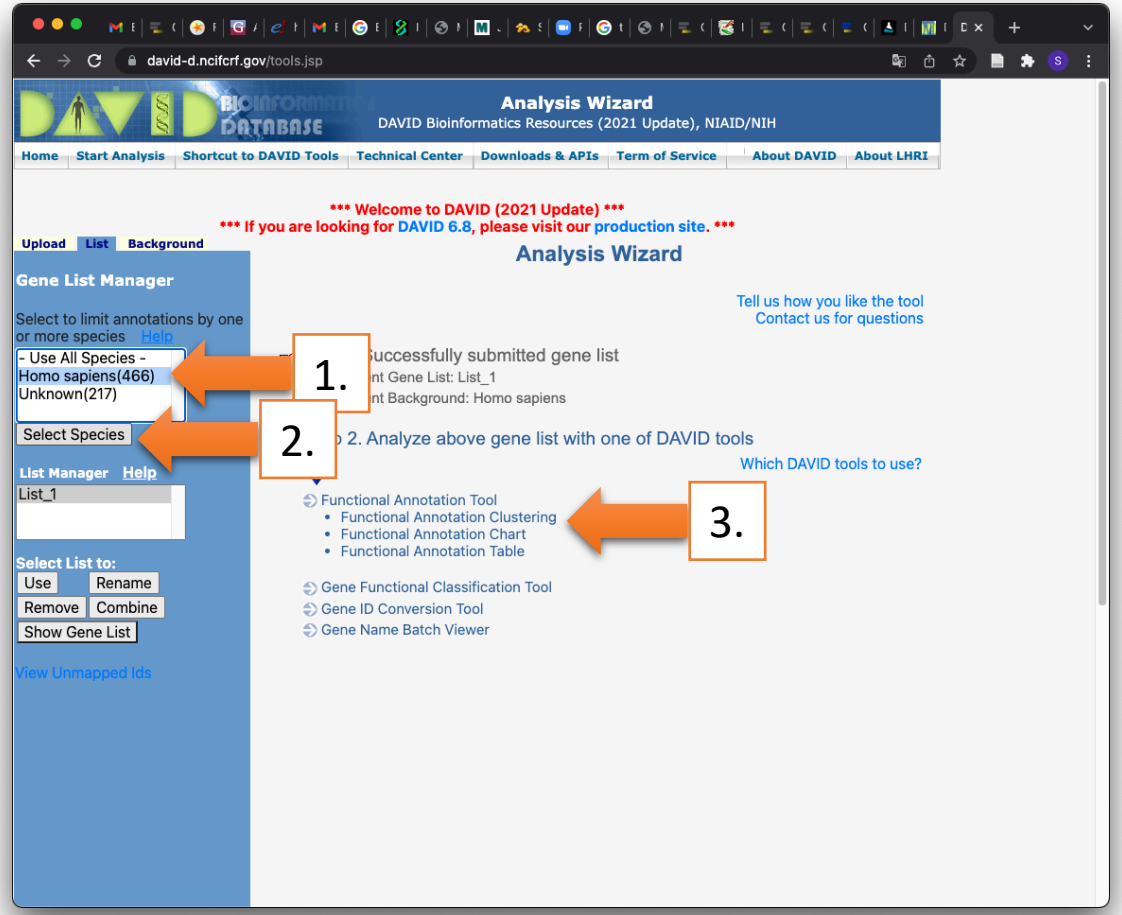
Exercise 9: Clustering

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



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)

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

Rerun using options | Create Sublist

60 Cluster(s) **Download File**

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