User Manual – VDJ Analysis toolkit

Requirements:

1. Folder structure forked/downloaded from https://github.com/Ghosn-Lab/SKumar/tree/master/VDJ_Analysis
2. Python 3.7.6 or greater installed on the system (Add Python to PATH – instructions follow)
3. Latest version of internet browser, e.g., Google Chrome, Mozilla Firefox, Apple Safari

Installing and Adding Python to PATH:

Python can be downloaded and installed from the website - https://www.python.org/downloads/

To add Python to the PATH variable for execution from any folder in command line:

Windows: https://geek-university.com/python/add-python-to-the-windows-path/

Mac: https://www.educative.io/edpresso/how-to-add-python-to-the-path-variable-in-mac

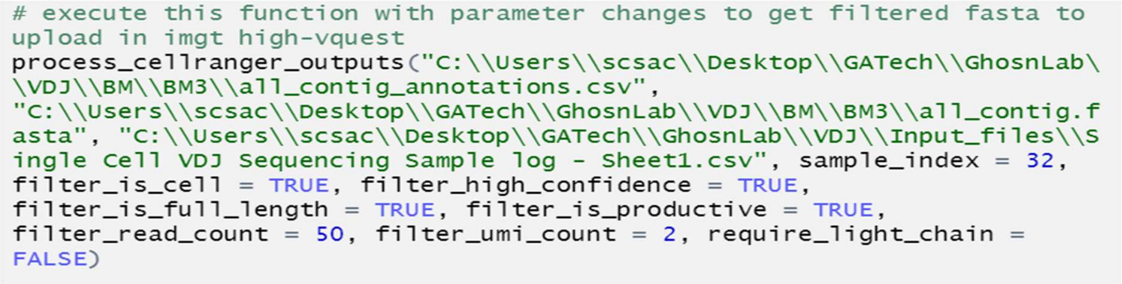
Linux: https://www.tutorialspoint.com/python/python_environment.htm

Once Python has been added to PATH, it can be checked by executing the command:

```
> python/python3 --version
```

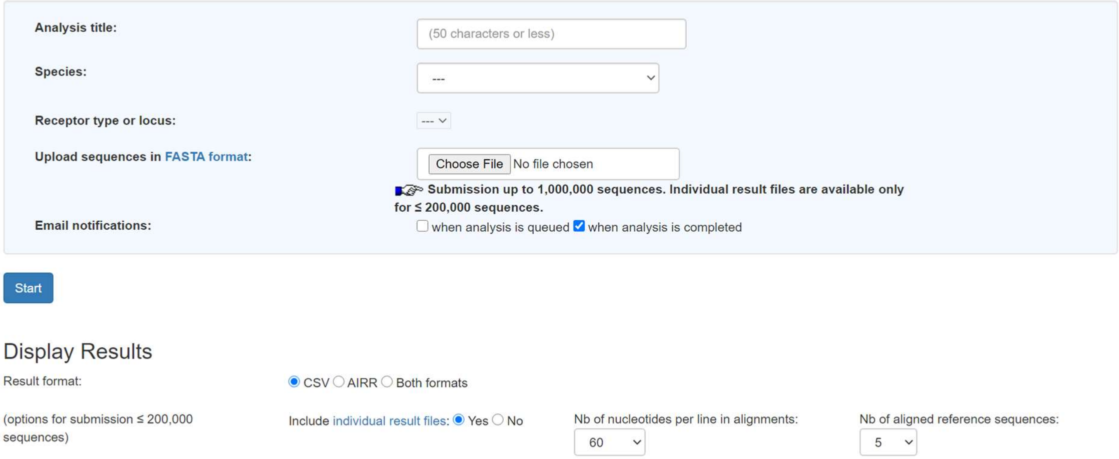
Generation of Input files using IMGT and Matt's pipeline:

1. Open the RMD file in the "Code_files" folder named "running_vdj_single_cell_pipeline" in RStudio.
2. Compile the functions in the environment – "process_cellranger_outputs", "process_imgt_output" and "assign_lineages".
3. Execute the function "process_cellranger_outputs" with input parameters of all contig annotation matrix and fasta from cell ranger. Sample log sheet keeps track of sample and sample_index corresponds to the one in the sheet. The output is filtered annotation csv and high confidence fasta file. Stored in all_contig_annotations file directory.



```
# execute this function with parameter changes to get filtered fasta to
upload in imgt high-vquest
process_cellranger_outputs("C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\
\VDJ\\BM\\BM3\\all_contig_annotations.csv",
"C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\BM\\BM3\\all_contig.f
asta", "C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\Input_files\\S
ingle Cell VDJ Sequencing Sample log - Sheet1.csv", sample_index = 32,
filter_is_cell = TRUE, filter_high_confidence = TRUE,
filter_is_full_length = TRUE, filter_is_productive = TRUE,
filter_read_count = 50, filter_umi_count = 2, require_light_chain =
FALSE)
```

4. Navigate to http://www.imgt.org/HighV-QUEST/search.action . Login using credentials.
5. Specify parameters and toggle “Individual result files” button to “Yes” keeping everything default. Upload the filtered fasta generated in previous step to the portal and press Start.



Analysis title: (50 characters or less)

Species: ---

Receptor type or locus: ---

Upload sequences in FASTA format: Choose File No file chosen

Submission up to 1,000,000 sequences. Individual result files are available only for ≤ 200,000 sequences.

Email notifications: □ when analysis is queued ☑ when analysis is completed

Start

Display Results

Result format: ● CSV ○ AIRR ○ Both formats

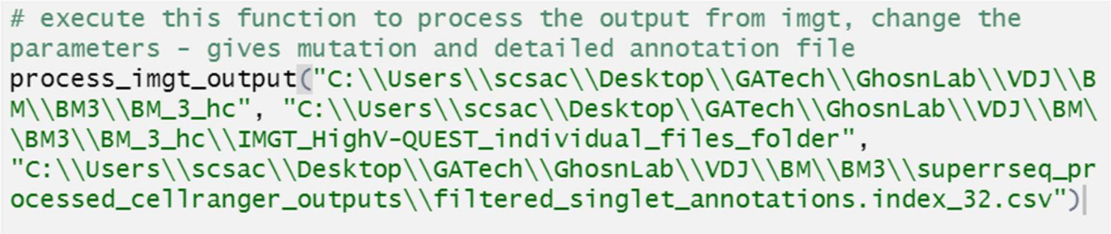
(options for submission ≤ 200,000 sequences)

Include individual result files: ● Yes ○ No

Nb of nucleotides per line in alignments: 60

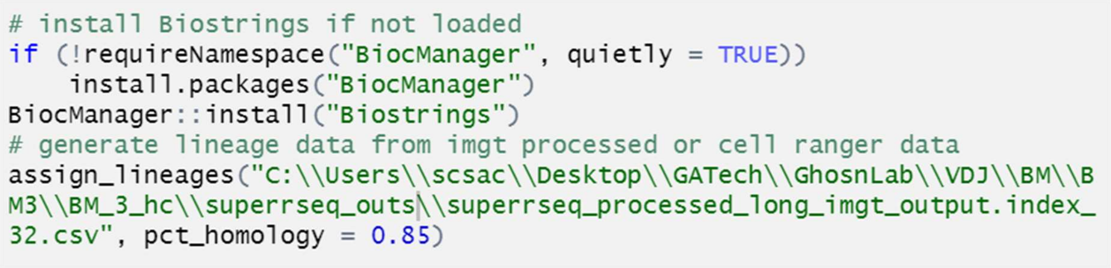
Nb of aligned reference sequences: 5

6. Download the results after analysis is complete. Extract the folder to a location and extract the “IMGT_HighV-QUEST_individual_files_folder” also.
7. Execute the second function in Rmd file – “process_imgt_output”. Specify the individual result folder from IMGT and filtered contig annotation file from last output. The output is “superrseq_processed_long_imgt_output.index_##” csv and “superrseq_processed_detailed_mutation_data.index_##” csv files depicting detailed IMGT processed VDJ matrix and detailed mutation data respectively. Stored in “supperseq_outs” folder in filtered_annotation file directory.



```
# execute this function to process the output from imgt, change the
parameters - gives mutation and detailed annotation file
process_imgt_output("C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\B
M\\BM3\\BM_3_hc", "C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\BM\
\BM3\\BM_3_hc\\IMGT_HighV-QUEST_individual_files_folder",
"C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\BM\\BM3\\superrseq_pr
ocessed_cellranger_outputs\\filtered_singlet_annotations.index_32.csv")
```

8. Execute the third function in Rmd file – “assign_lineages”. Specify the “long_imgt_output” file and homology percentage to calculate bins and consequently lineages. Output is “lineage_processed_data.csv” stored in the parent directory of “long_imgt_output” file.



```
# install Biostrings if not loaded
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("Biostrings")
# generate lineage data from imgt processed or cell ranger data
assign_lineages("C:\\Users\\scsac\\Desktop\\GATech\\GhosnLab\\VDJ\\BM\\B
M3\\BM_3_hc\\superrseq_outs\\superrseq_processed_long_imgt_output.index_
32.csv", pct_homology = 0.85)
```

Usage: Outputs from step 7,8 and cell ranger filtered annotations are used as input for the visualization toolkit.

Files: “filtered_contig_annotations.csv”,
“superrseq_processed_long_imgt_output.index_##.csv”,
“superrseq_processed_detailed_mutation_data.index_##.csv”,
“lineage_processed_data.csv”

Description of toolkit and usage:

Note: Do not change the folder structure downloaded from GitHub and don’t delete any files within “Code_files” and “Results” folder.

The VDJ Analysis toolkit helps to generate visualization files as well as processed data for multiple scenarios.

There are five main modes of operation in the toolkit, each generating the relevant visualization or files to be utilized by the user.

Mode 1: CDR3 Amino Acid distribution for a sample in a “Treemap” format with relationships across IGH, IGK and IGL barcodes.

Mode 2: Integration of ADT normalized matrix from SuPPER-seq pipeline with mutation data for each of regions CDR 1-3 and FR 1-3 including “Total mutations”, “Number of Hotspots” and “Number of hotspot mutations”.

Mode 3: Generates Circos plot for relationship of lineages of cell barcodes and the isotype information associated with it.

Mode 4: Generates a Hierarchical bar chart for the mutation cumulative count by hotspot for each region CDR 1-3 and FR 1-3.

Mode 5: Generates a VDJ comparison Stacked bar chart for multiple matrix data sets or a single data set.

Navigate to the “Code_files” folder using the command prompt. The general usage of the tool is done using the following command line:

> **python/python3 mainProcess.py [1/2/3/4/5] [input_files] [output_directory]**

Note: The python command can be executed as python/python3 depending on the first execution after installing and adding python to PATH.

Arguments:

[1/2/3/4/5] – Specify the mode of operation to be used for the specific visualization or output.

[input_files] – Specify the full input file paths corresponding to the mode, separated by space.

- **Mode 1:** Filtered VDJ matrix from Cell ranger
- **Mode 2:** ADT normalized matrix, IMGT processed detailed mutation file and IMGT processed VDJ matrix
- **Mode 3:** Lineage data processed VDJ matrix
- **Mode 4:** IMGT processed detailed mutation file and IMGT processed VDJ matrix
- **Mode 5:** Filtered VDJ matrices or matrix

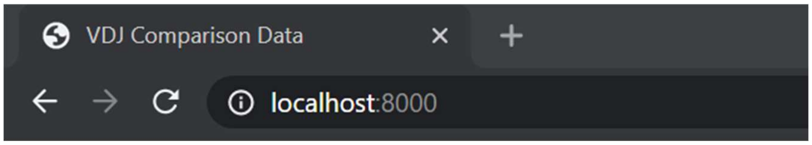
[output_directory] – Specify the output directory for results. It is recommended to specify as the "Results" folder path in the downloaded folder structure as it contains the html files for visualization.

Output and its usage:

For the mode, input file and output directory specified, the outputs will be stored in one of the four folders for visualization or the "Results" folder in case of Mode 2.

Instructions to display the output HTML file:

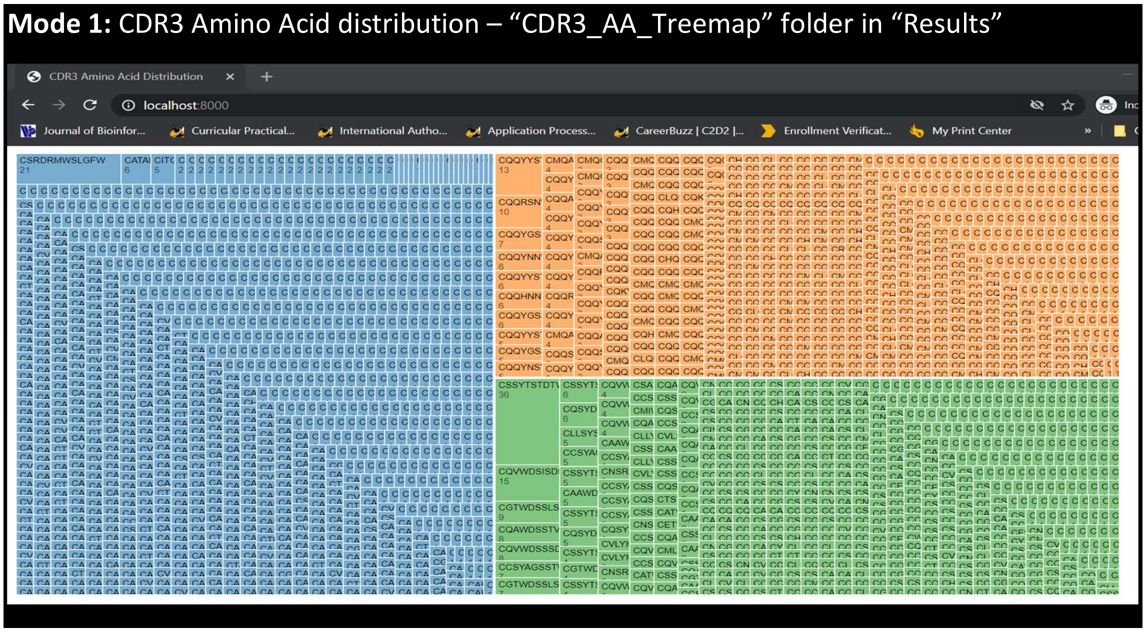
1. Using command line navigate to the required folder inside the "Results" folder after running the tool, e.g.,
   > **cd C:\Users\scsac\VDJ_Analysis\Results\CDR3_AA_Treemap**
2. Run the command
   > **python/python3 -m http.server**
3. Open an **incognito/private window** in your browser (Chrome, Firefox, Safari)
4. In the address bar type, **localhost:8000**



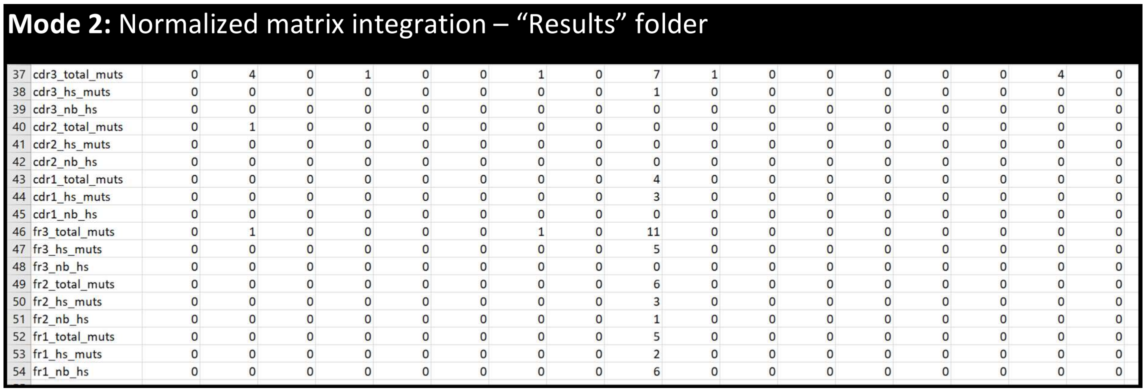
5. Interactive visualization will be displayed in the browser
6. Repeat steps 1-5 for other folders in "Results" to display their visualizations
7. In case of Mode 2, the csv file can be used for visualization in SeqGeq

Below are the expected output visualization file examples in case of each mode.

Mode 1: CDR3 Amino Acid distribution – “CDR3_AA_Treemap” folder in “Results”

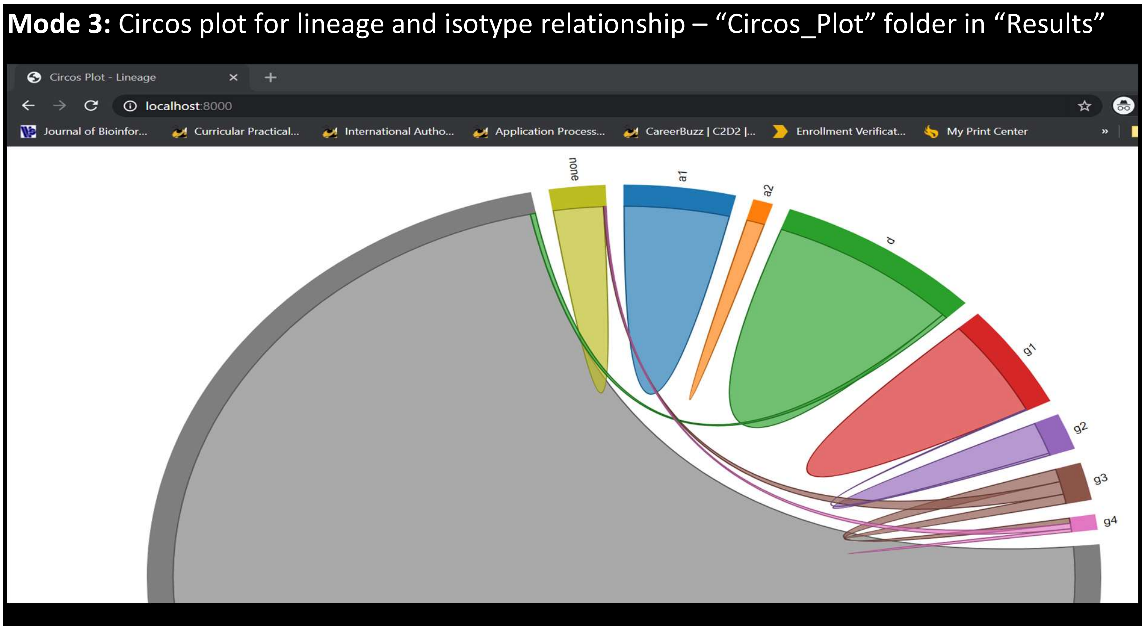


Mode 2: Normalized matrix integration – “Results” folder

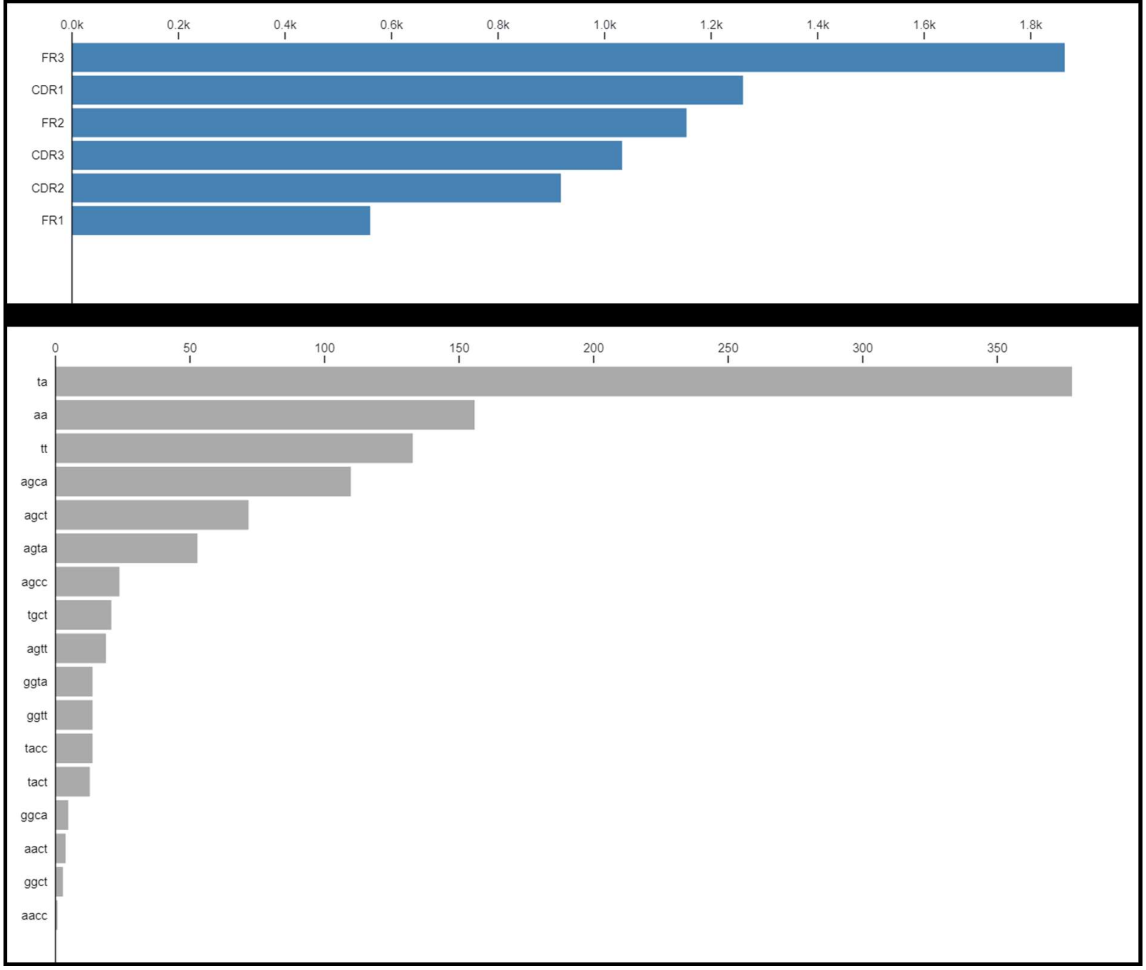


37	cdr3_total_muts	0	4	0	1	0	0	1	0	7	1	0	0	0	0	0	4	0
38	cdr3_hs_muts	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
39	cdr3_nb_hs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	cdr2_total_muts	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	cdr2_hs_muts	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
42	cdr2_nb_hs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	cdr1_total_muts	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
44	cdr1_hs_muts	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
45	cdr1_nb_hs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	fr3_total_muts	0	1	0	0	0	0	1	0	11	0	0	0	0	0	0	0	0
47	fr3_hs_muts	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
48	fr3_nb_hs	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
49	fr2_total_muts	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0
50	fr2_hs_muts	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
51	fr2_nb_hs	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
52	fr1_total_muts	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
53	fr1_hs_muts	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
54	fr1_nb_hs	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0

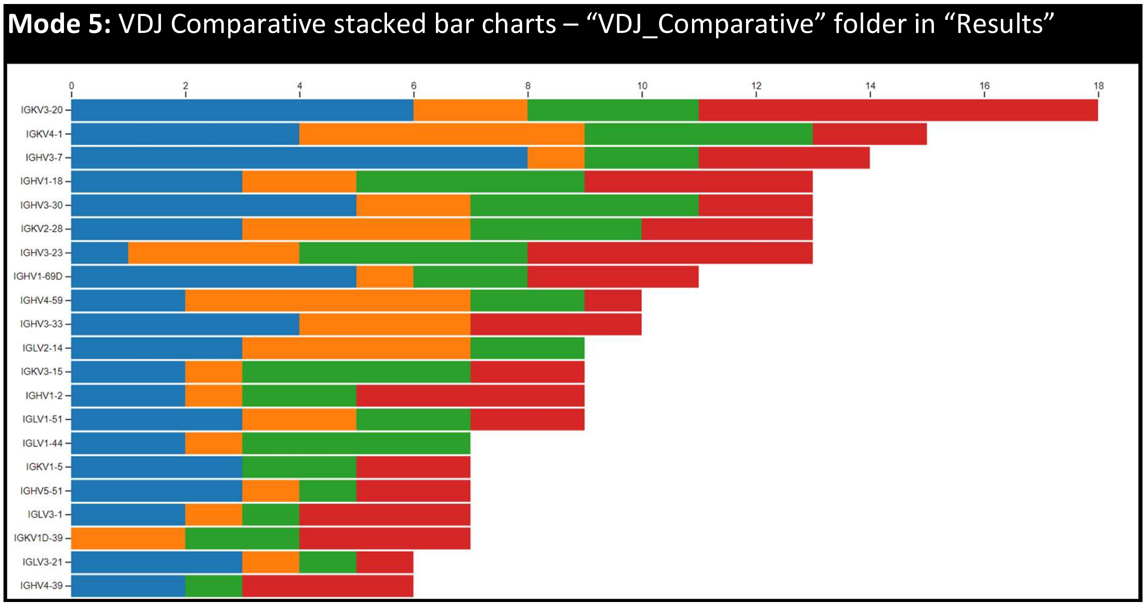
Mode 3: Circos plot for lineage and isotype relationship – “Circos_Plot” folder in “Results”



Mode 4: Hierarchical bar chart for hotspot mutation by region – “Mutation_Data” folder in “Results”



Mode 5: VDJ Comparative stacked bar charts – “VDJ_Comparative” folder in “Results”


IGKV3-20
IGKV4-1
IGHV3-7
IGHV1-18
IGHV3-30
IGKV2-28
IGHV3-23
IGHV1-69D
IGHV4-59
IGHV3-33
IGLV2-14
IGKV3-15
IGHV1-2
IGLV1-51
IGLV1-44
IGKV1-5
IGHV5-51
IGLV3-1
IGKV1D-39
IGLV3-21
IGHV4-39