

MANUAL

Introduction	4
Installation	5
Windows:.....	5
Linux:	5
macOS:	5
Software layout.....	7
Tool Buttons	7
Input:.....	7
Align:	7
ClonoType:	7
Phylogeny:.....	8
Multifile:.....	9
Save:	9
Clean:.....	9
Search:	9
Menu Bar	9
Display.....	9
Remove duplication.....	9
Filter Level	10
Sequence color mode	10
Sequence render mode	10
Display full msa	10

Display by ClonoType.....	10
Display by Genes	11
Display by region	11
Tools	11
V Gene.....	11
Abundance.....	12
Seqlogo (only for linux)	13
Unusual Residue.....	13
Length Distribution	15
Parameter.....	15
Align parameter	15
Temporary Path	16
Example.....	16
Usage case	17
Multi-File Navigator.....	23
Navigator layout.....	23
Explanation of terms	23
Clonotype	23
Usage	23
Add file.....	23
Delete file	23
Multiple sequence alignment	24
Save file.....	24
Tool buttons	24

Length Distribution	24
Pairwise Density	25
Diversity Heatmap.....	26
Gene Abundance	27
Clonotype Abundance	28
Residue Changes	32
Notice	34
Reference	35

Introduction

Multiple sequence alignment has long been used as a powerful tool to investigate the evolutionary, structural and functional properties of protein families. Compared with ordinary protein families, antibodies or BCR sequences have highly variable regions, which make the existing multiple sequence alignment methods unable to produce precise result on antibodies. Recently, the increasing data of BCR sequencing along with COVID-19's global popularity has stimulated the urgent needs for multiple BCR-sequence alignment and bioinformatics analysis. To address this issue, we developed a free multiple sequence alignment software based on AbRSA^[1], named Abalign, which incorporated the heuristic knowledge of antibody numberings, including IMGT, KABAT, Chothia and Martin. It follows the well-characterized patterns of conserved or insertion positions by immunology studies, which enable the result to be consistent with the structural and immunological knowledge.

Abalign has a user-friendly interactive interface that supports multiple sequence alignment, length statistics for each region, V/J gene matching, clonotype matching, evolutionary analysis, antibody humanization, and other functions. In addition, we have developed a series of functions for the cross-analysis of multiple files to help users analyze valuable information from multiple samples, such as shared clono types, or changes in residue ratios, etc. Compared with traditional multiple sequence alignment software, Abalign requires very little computational resources and can complete the alignment and analysis of 1G of DNA FASTA sequences at a very fast speed on a PC with only 16G of RAM. Abalign will profit immunoinformatic and pharmaceutical communities on analyzing massive BCRs or antibodies and making new discoveries.

Installation

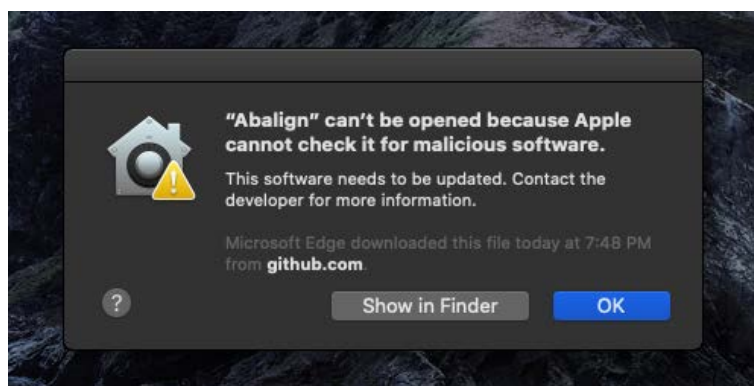
Windows: After extracting the files, go to the extracted folder, Double-click **Align_Setup.exe** to enter the installation, click **Browse** to select the installation path, and click **Next**.

Linux: After extracting the files, go to the extracted folder, find **Abalign_installer.run**, and execute the following command in the terminal to start the installation guide.

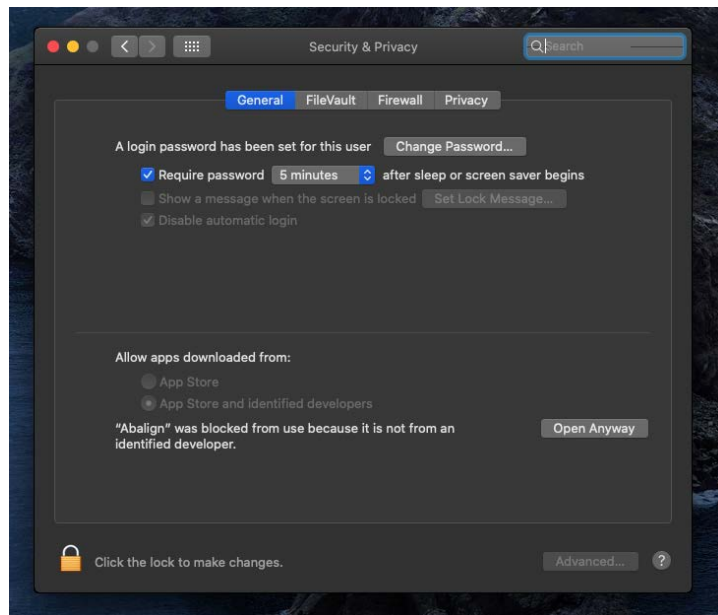
1. `chmod +x Abalign_installer.run`
2. `./Abalign_installer.run`

macOS: After extracting the files, go to the extracted folder, Double-click **Abalign.app** to run Abalign directly. Please move Abalign.app into the Applications folder for easier use.

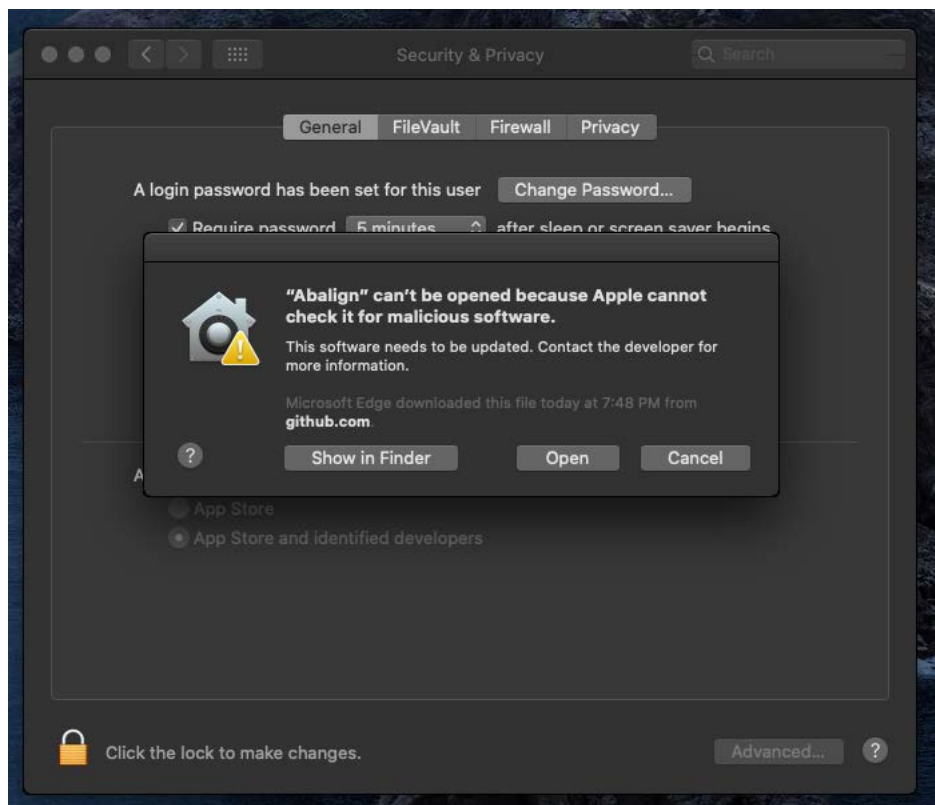
When you run Abalign for the first time on macOS, the following dialog box may pop up. This is normal, please click "OK" in this dialog box. When you run Abalign for the first time, the following dialog box may pop up. This is normal, please click "OK" in this dialog.



Then please click "**System Preferences**"->"**Security & Privacy**", and click "**Open Anyway**" in this window

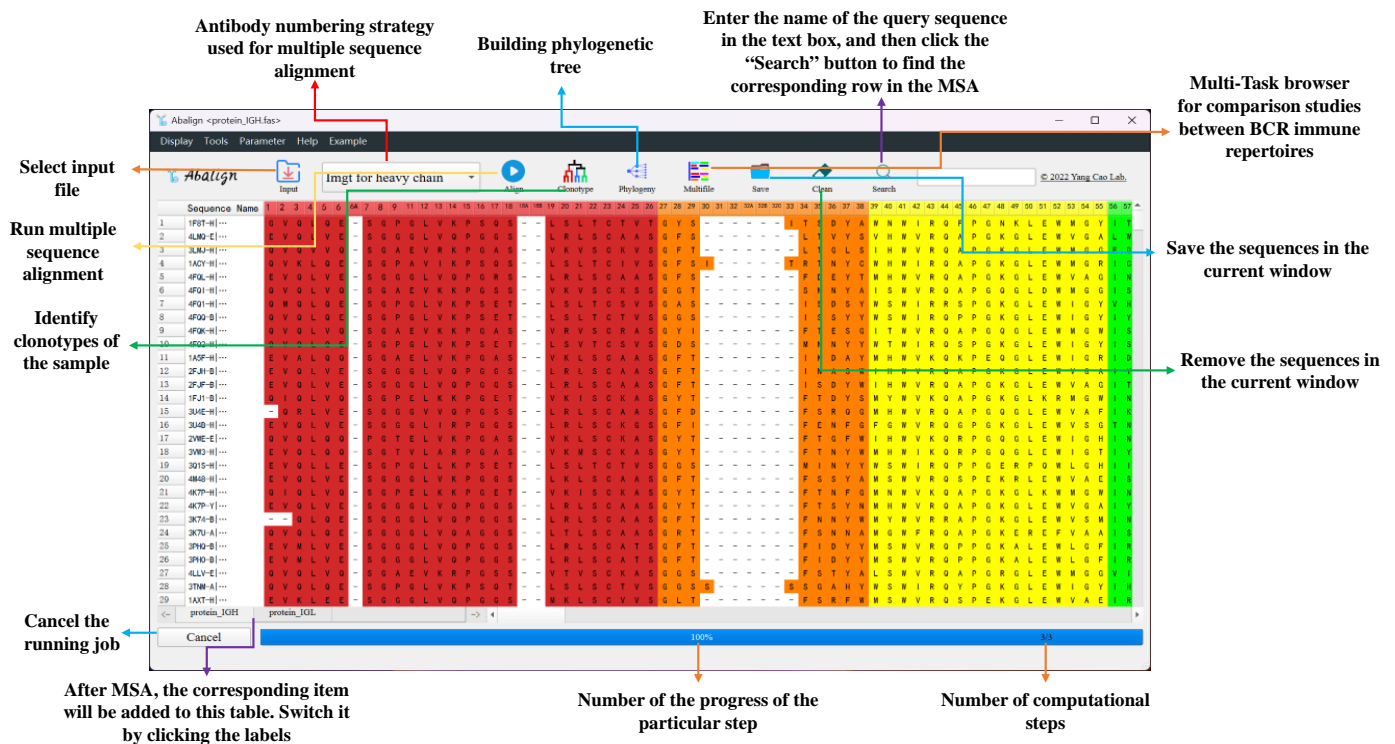


After completing the above steps, a new dialog will appear, click “Open” in the dialog box to open Abalign.



Note: When running Abalign under macOS, if a dialog prompts "Abalign is damaged and can't be opened. You should move it to the Trash". Execute '`sudo xattr -r -d com.apple.quarantine /Applications/Abalign.app`' in the terminal to resolve the issue.

Software layout



Tool Buttons

Input: Select the input file in Fasta format.

Align: Execute multiple sequence alignment. It will also search for V genes and species that are most similar to each sequence. The results are shown in the window and rendered with different colors for FR1, CDR1... CDR3, and FR4. Users can change the rendering method by clicking **Display- > Sequence render mode**.

ClonoType: Identify clonotypes. It should mention that running "Align" needs to be the first step. The clonotype is defined as sequences that share the same V and J genes as well as the same CDR3 [2].

Phylogeny: Perform FastTree software (maximum likelihood method)^[2] to build .nwk file and visualize it with Ete3^[3]. After clicking the button, a dialog will pop up (Fig. 1), in this dialog you can adjust the parameters for building the phylogenetic tree (Fig. 2).

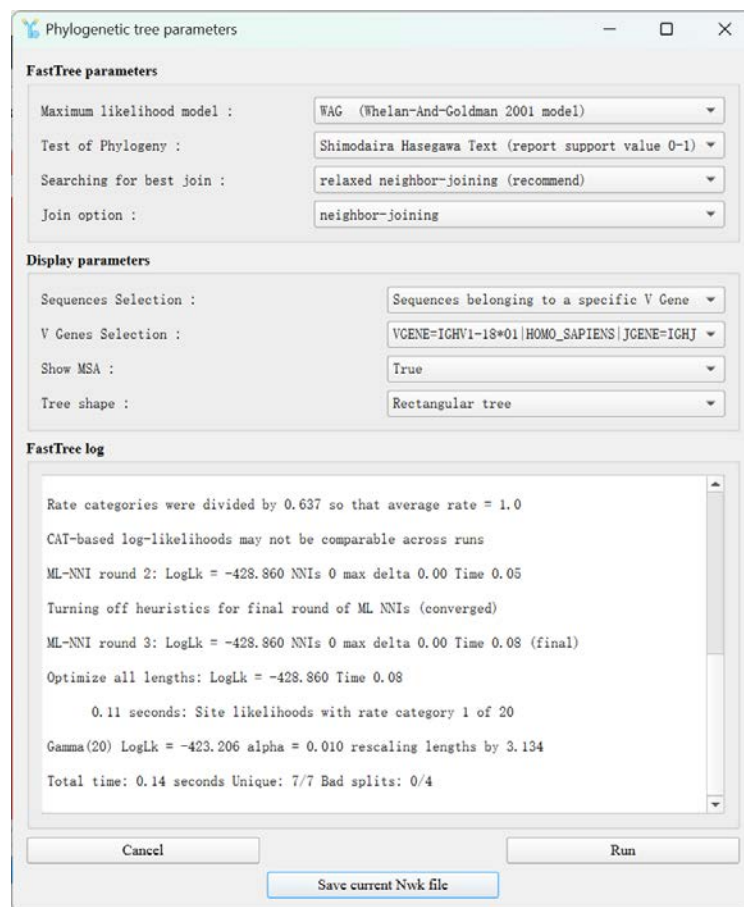


Figure 1. Phylogenetic Tree Parameters Dialog

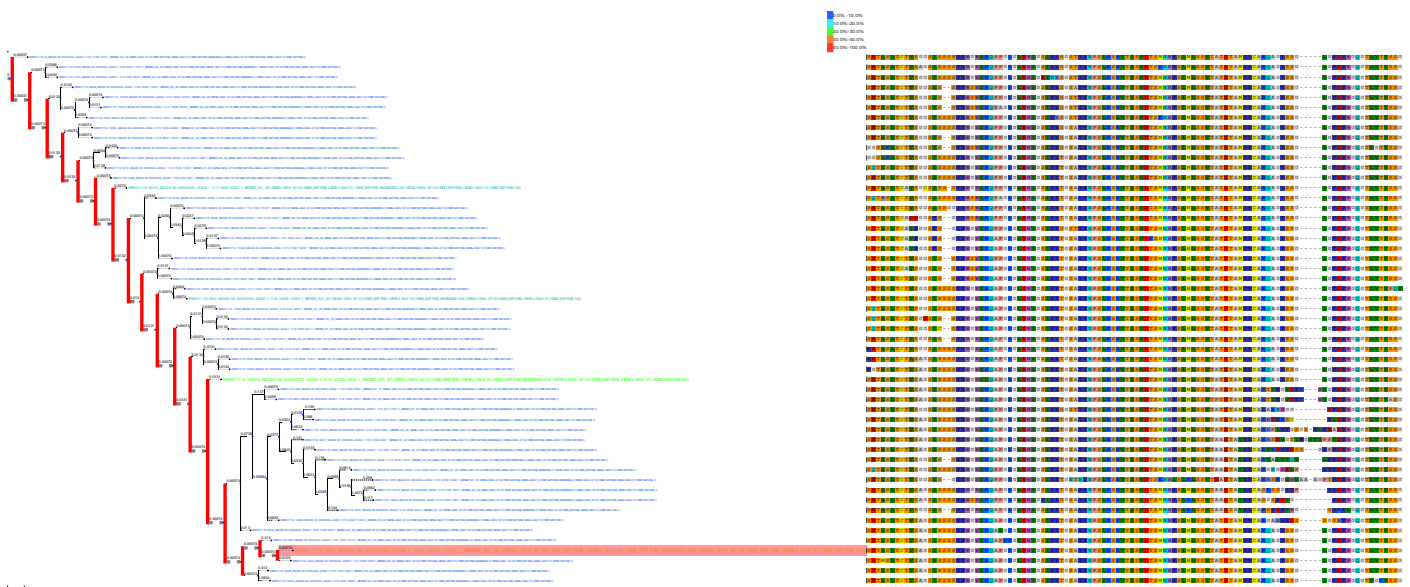


Figure 2. An example of a phylogenetic tree with corresponding MSA

The figure on the left is the phylogenetic tree of the selected sequences, and highlighted with the sequence abundances in the BCR immune repertoire. The multiple sequence alignment corresponding to tree is showed on the right and highlighted with the mutations.

Multifile: After clicking this button, aa navigator will pop up, in which users can perform alignments and comparisons for different BCR immune repertoires. Users need to load multiple Fasta files in the navigator. Please see "Help" for detailed information.

Save: Save the sequences in the current window.

Clean: Clicking this button will clear all sequences in the MSA text box and delete the sample.

Search: Enter the name of the sequence and then click the button, the display region will jump to the query sequence.

Menu Bar

Display

Remove duplication: If you check this option, antibody sequences with the same variable region will not be displayed, nevertheless the **duplicated ones will be accounted for** abundance analysis. This option is enabled by default, and can be adjusted in “Align parameter”.

Filter Level: Filtering sequences based on the length of the antibody variable domain sequence. If the FRs and CDRs of the sequence variable domain do not meet the length condition, then this sequence will be deleted. There are four levels of length filtering: "Off", "Soft"(default), "Normal" and "Strict". "Off" means that there is no length limit for each region, "Soft" requires that the length of each region is not less than 1, "Normal" and "Strict" limit the region length according to antibody data with known structures. This option can also be adjusted in “Align parameter”.

Sequence color mode: There are two options in this menu, which can be toggled to adjust the color of the residue rendering. "Light mode" renders the residue as a light color and "Soft mode" renders the residue as a dark color."Light mode" is used by default, and this feature can be used with "Sequence render mode".

Sequence render mode: There are two options in this menu, and the mode of color rendering can be adjusted by toggling different options." Color by region" will divide the antibody sequence into different FRs and CDRs and render the different regions in different colors." Color by amino" renders different residues in different colors depending on the type of residue."Color by region" is used by default, and this feature can be used with "Sequence color mode".

Display full msa: Selecting this option will display all sequences of the MSA.

Display by ClonoType: Selecting this option will display the specific clonotypes in the sample, which needs to clonotype screening first.

Display by Genes: This menu has three options, by clicking on different options you can display the sequences that match the conditions. "Display by Vgene" brings up the V genes of all sequences when you click on this option, you can select the specified V genes to display the sequences matching the selected criteria. "Display by Jgene" will bring up the J genes of all sequences, and you can select the specified J genes to display the sequences matching the selected criteria. "Display by VJgene" will bring up the VJ gene combinations of all sequences, and you can select the specified VJ gene combinations to display the sequences matching the selected criteria.

Display by region: There are seven options in this menu, corresponding to the seven regions of the variable domain of the antibody, and one or more options can be selected to display the different regions of the antibody.

Tools

V Gene: Here are two editable options in this menu. "Species" provides the V gene species to select. If selecting "Homo sapiens", the V gene identification will only use human V germline gene database. "Display V Gene list" shows the alignment of each sequence with top five alignment scoring V genes (Fig. 3).

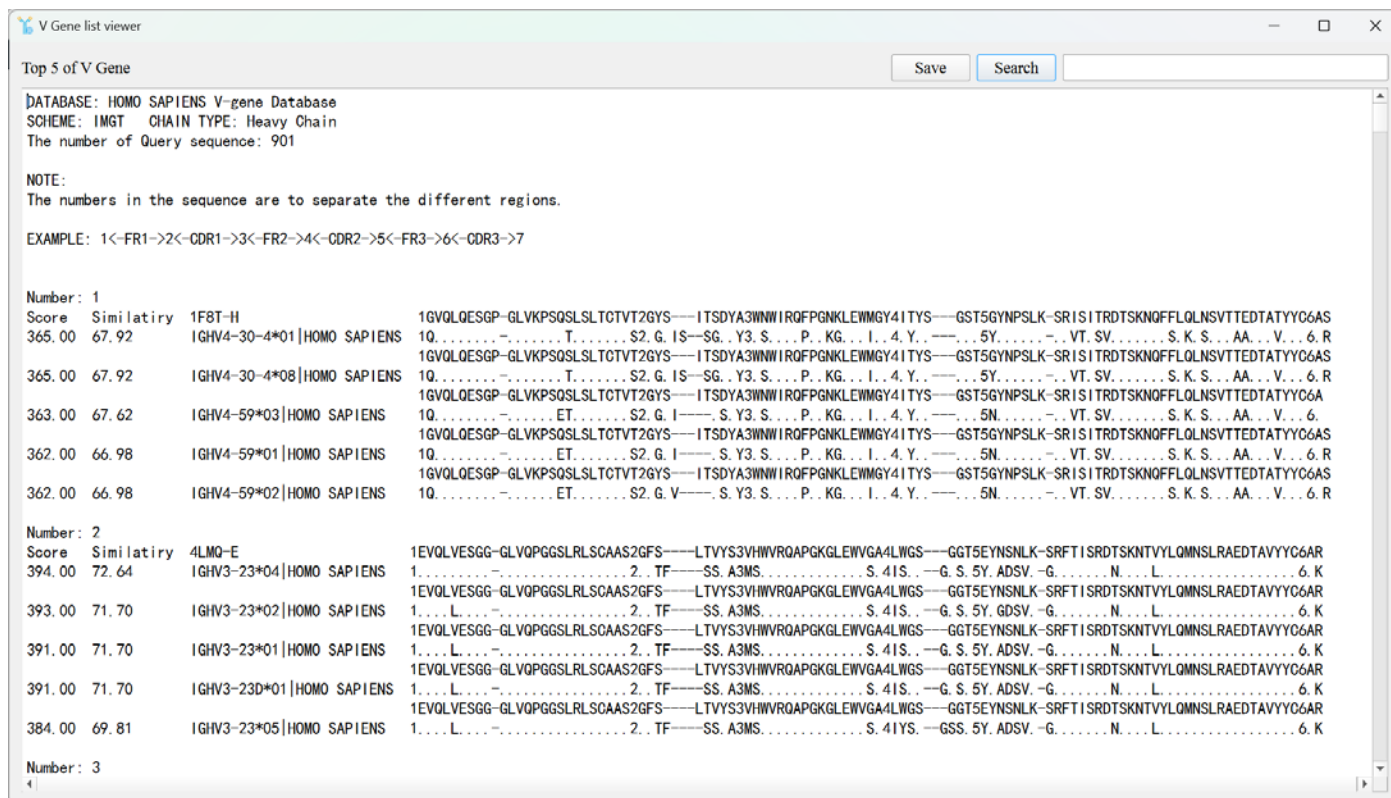


Figure 3. An example of V Gene list viewer.

Abundance: "V Gene abundance" shows the top 20 abundant V genes (Fig. 4). "Sequence abundance" shows the top 20 abundant sequences. "Region abundance" menu shows the abundance of top 20 sequences in a particular region.

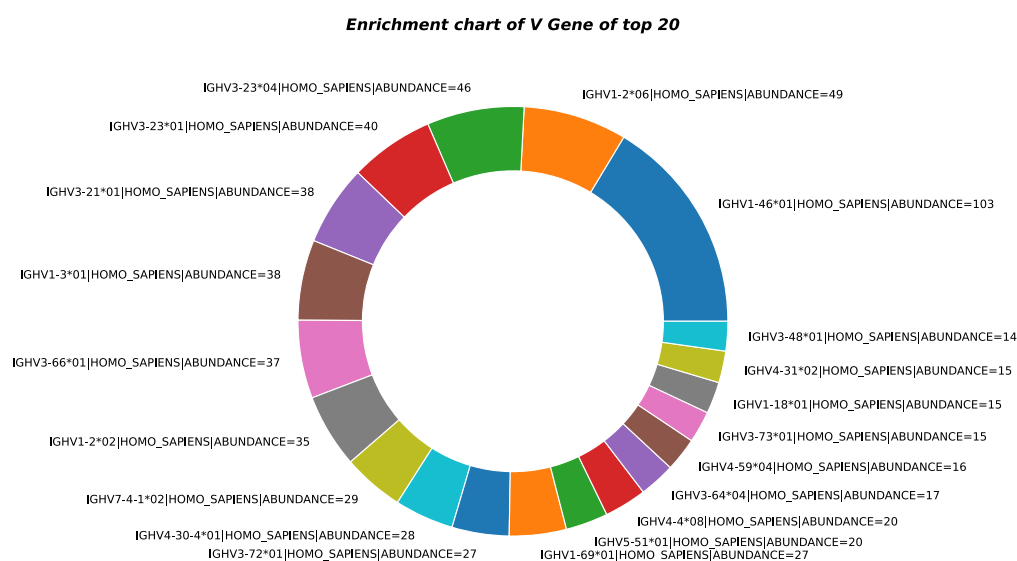


Figure 4. An example of Enrichment chart of V Gene of top 20.

Abundance information is displayed in a pie chart, different colors represent different types, and their corresponding abundance values are in brackets.

Seqlogo (only for linux): "By Entropy" generates a Seqlogo plot with entropy as the y-axis for the currently displayed MSA (Fig. 5). "By Frequency" generates a Seqlogo plot with frequency as the Y-axis for the currently displayed MSA. "Color" changes the rendering mode of the Seqlogo plot.

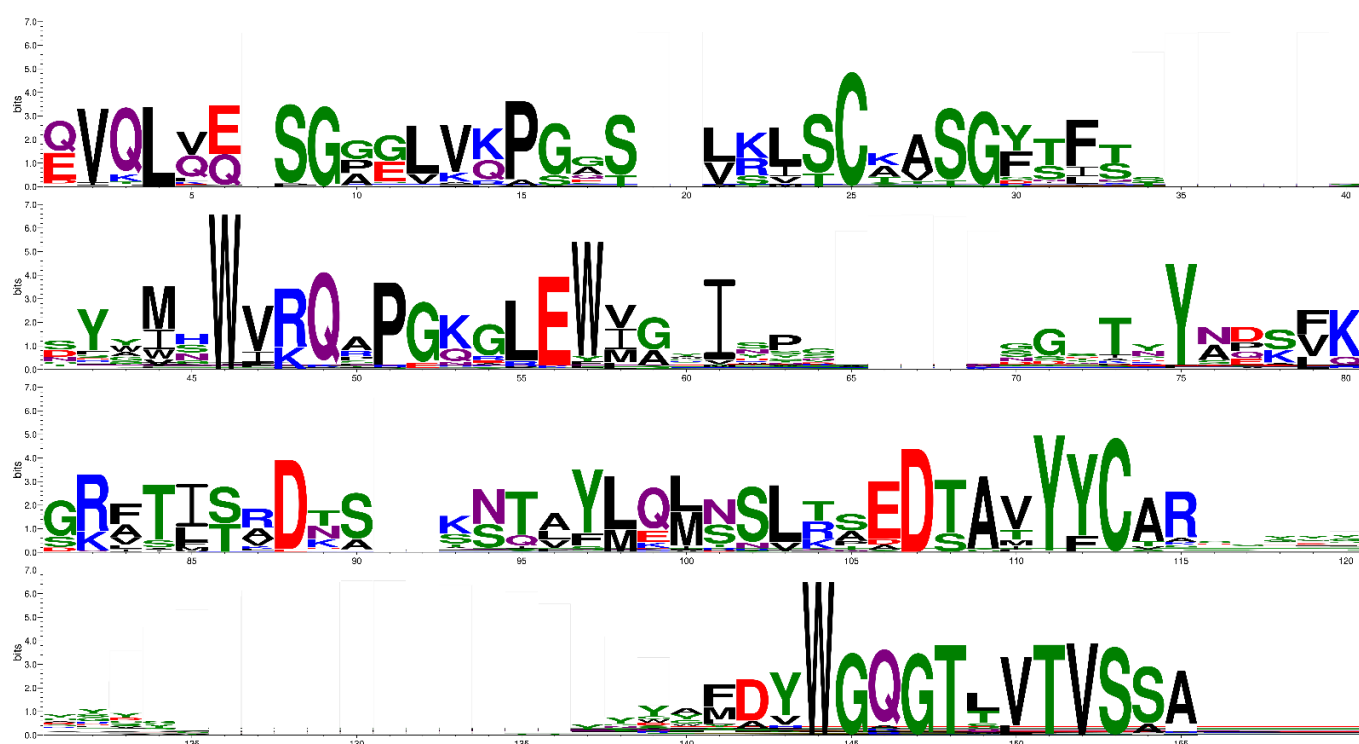


Figure 5. Seqlogo plot by Entropy The ordinate indicates the entropy of amino acids, and the abscissa indicates the position in the variable domain. The larger the letter of the amino acid, the greater the entropy value.

Unusual Residue: This option shows the proportion of residues at each position of the query sequence by comparing with amino acid distribution at each position constructed with tens of millions of human sequences from OAS^[5] (Fig 7). It is used for antibody humanization.

Users are also allowed to build a amino acid distribution graph with their own datasets (Fig. 6).

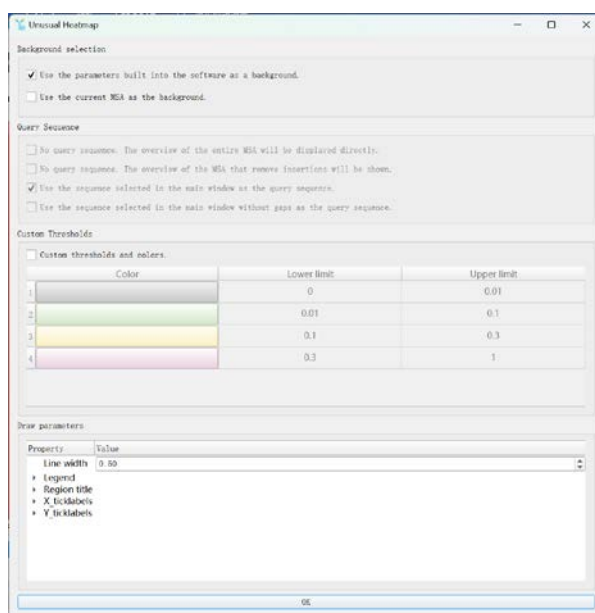


Figure 6. Unusual Residue parameters dialog.

"Background selection" is used to select the background library comparing to the query sequence. "Query Sequence" allows the user to choose whether to use the query sequence and whether to delete the gap in the query sequence. "Custom Thresholds" is used to adjust the thresholds of amino acid distribution and the colors they represent. "Draw parameters" are used to modify drawing parameters.

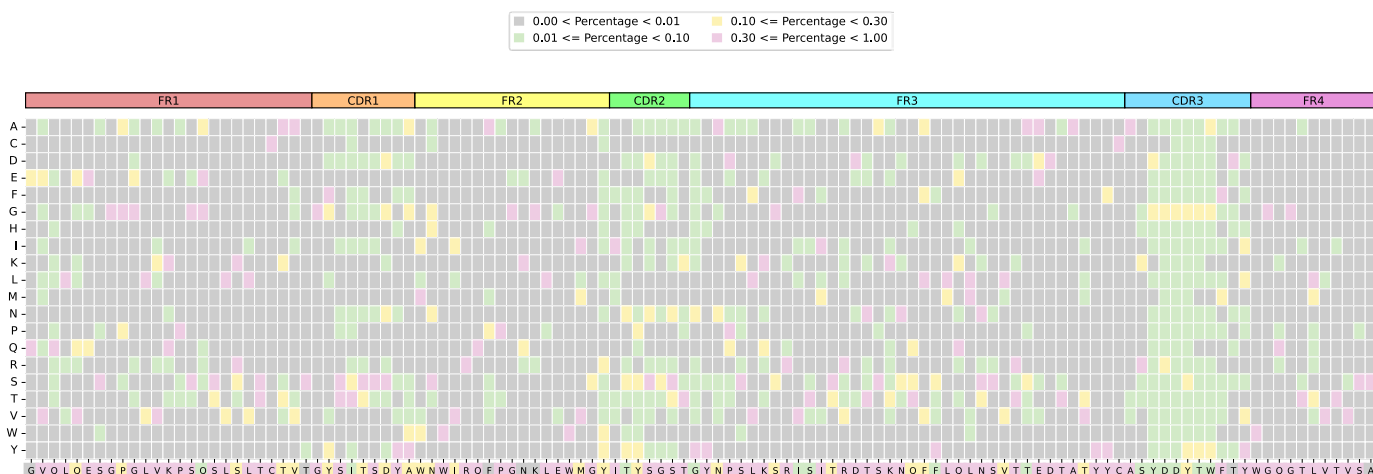


Figure 7. An example of unusual residue map. Different percentage of residues will be rendered as different colors.

The query sequence selected in the Multiple Sequence Alignment window is shown at the bottom of the heatmap, which is used for comparison with the human antibody reference dataset. The FR and CDR regions of the variable domain are marked with different colors on the top. The frequency of each residue in the human dataset is marked with a different color, with frequency below 0.01 (“Unusual Residue”, marked in gray) by default.

Length Distribution: There are eight options in this menu, for each of the FRs, CDRs and the full length of the variable regions. Click on the different options to see the length distribution (Fig 8).

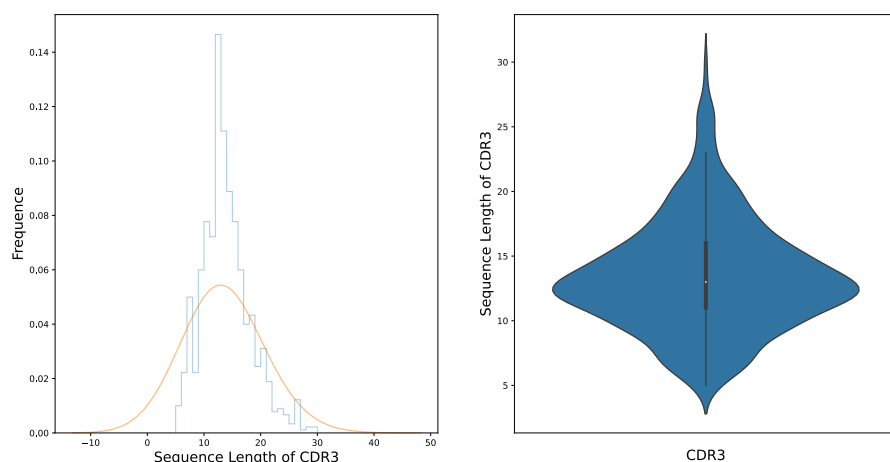


Figure 8. An example of Length Distribution of CDR3

The left figure is a length histogram, the abscissa indicates the length, and the ordinate indicates the proportion. The right figure is a violin plot, the ordinate represents the length, and the larger the width, the greater the number of sequences of the specific length.

Parameter

Align parameter: parameter options for MSA (Fig 9).

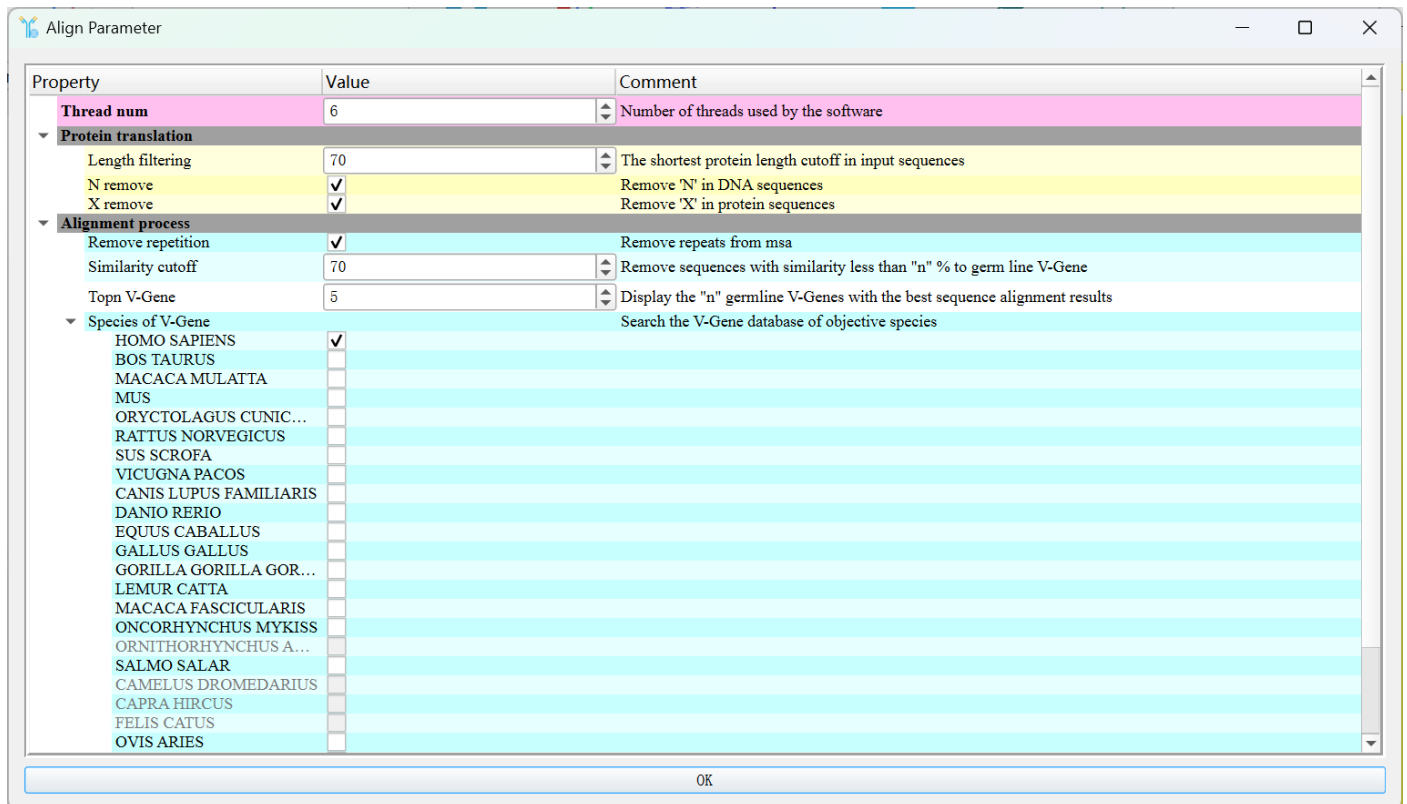


Figure 9. Align parameters dialog.

"Align parameter" is divided into 3 columns, the first column lists the modifiable elements or functions, the second column lists the parameter values corresponding to the elements or whether to enable certain functions, and the third column lists the introduction of the parameters.

Temporary Path: Dialog box for changing the path of the temporary files used by Abalign. Please restart the software after changing it.

Example: BCR/antibody sequence data (DNA or amino acid) are provided for testing purpose. "Example" contains two options, which are used for single-file mode and multi-file mode respectively, and users can find the example files under the "example" folder in the installation path.

Usage case

Users can click [Example->SingleFile](#) in the menu bar to load the example file, or select the input file through the Input button in the toolbar.

Step 1 Input file: Load the Fasta formatted file after clicking input. (Alternatively, as the example, we click on [Example->SingleFile](#) in the menu bar to load the file.)

Step 2 Search for antibody variable domain and multiple sequence alignment: After loading the sequences, click “Align” to run the program The running progress will be shown in the progress bar. After the computation, a dialog box will be popped, showing the statistics information (Fig 10). duplicated sequences will be detected.

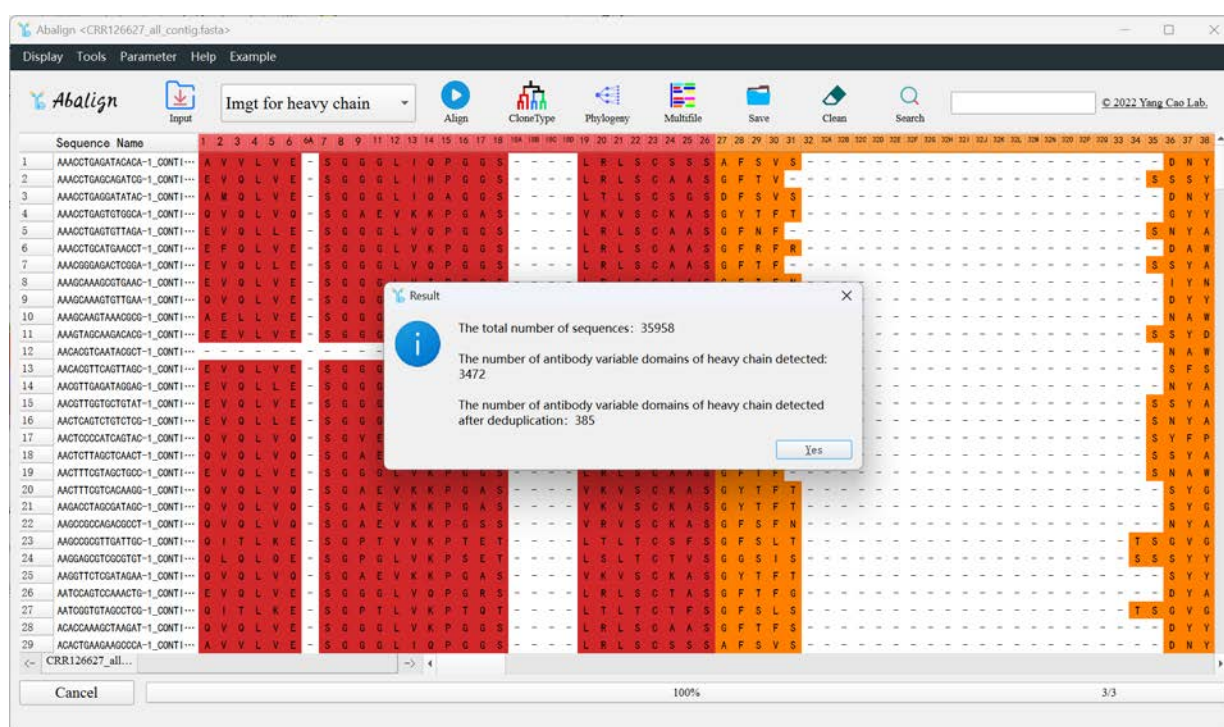


Fig 10. The result of a MSA using the Patient_contig sample in the example folder

Step 3 Analyze multiple sequence alignments: click the “Tools” button in the top menu bar and move the mouse to “Abundance”. Click “V Gene abundance” or “Sequence abundance” to

obtain the abundance map of V Gene and sequence (Fig 11).

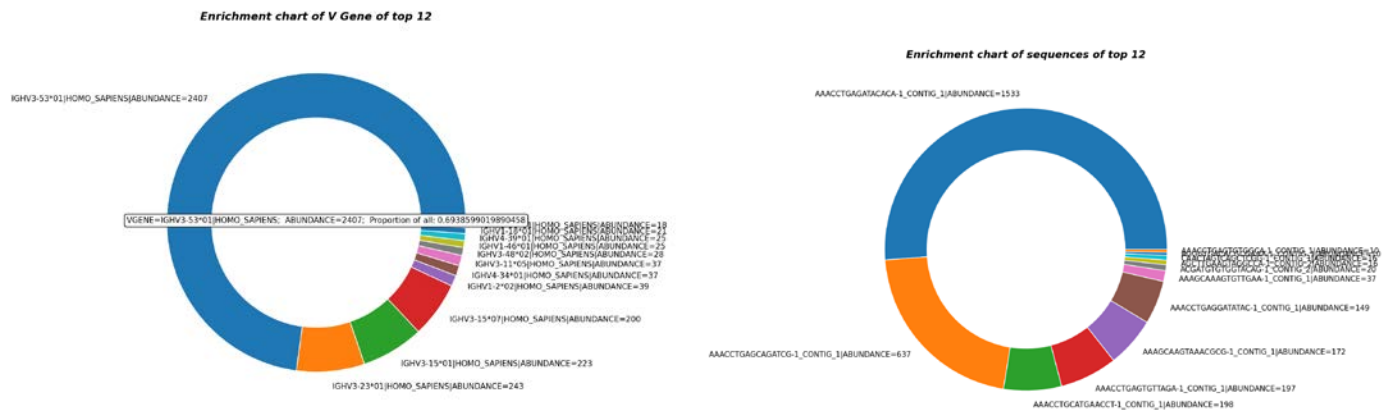


Figure 11. The top12 V gene and sequences abundance of Patient_contig sample

Step 4 Build phylogenetic tree: according to the sequence abundance information, build the tree with the V Gene of the high abundance sequence. In this case, the most abundant sequence belongs to V Gene “IGHVH3-53 * 01”. Users can click [Display->Display by Genes ->Display by Vgene](#) in the menu bar to select the V Gene “IGHVH3-53*01”, and then click the “Run” button to build the phylogenetic tree of sequences belonging to this V Gene (Fig 12). If you are satisfied with the results, click the “[Save the current Nwk file](#)” button in the tree building menu to save the .nwk file.

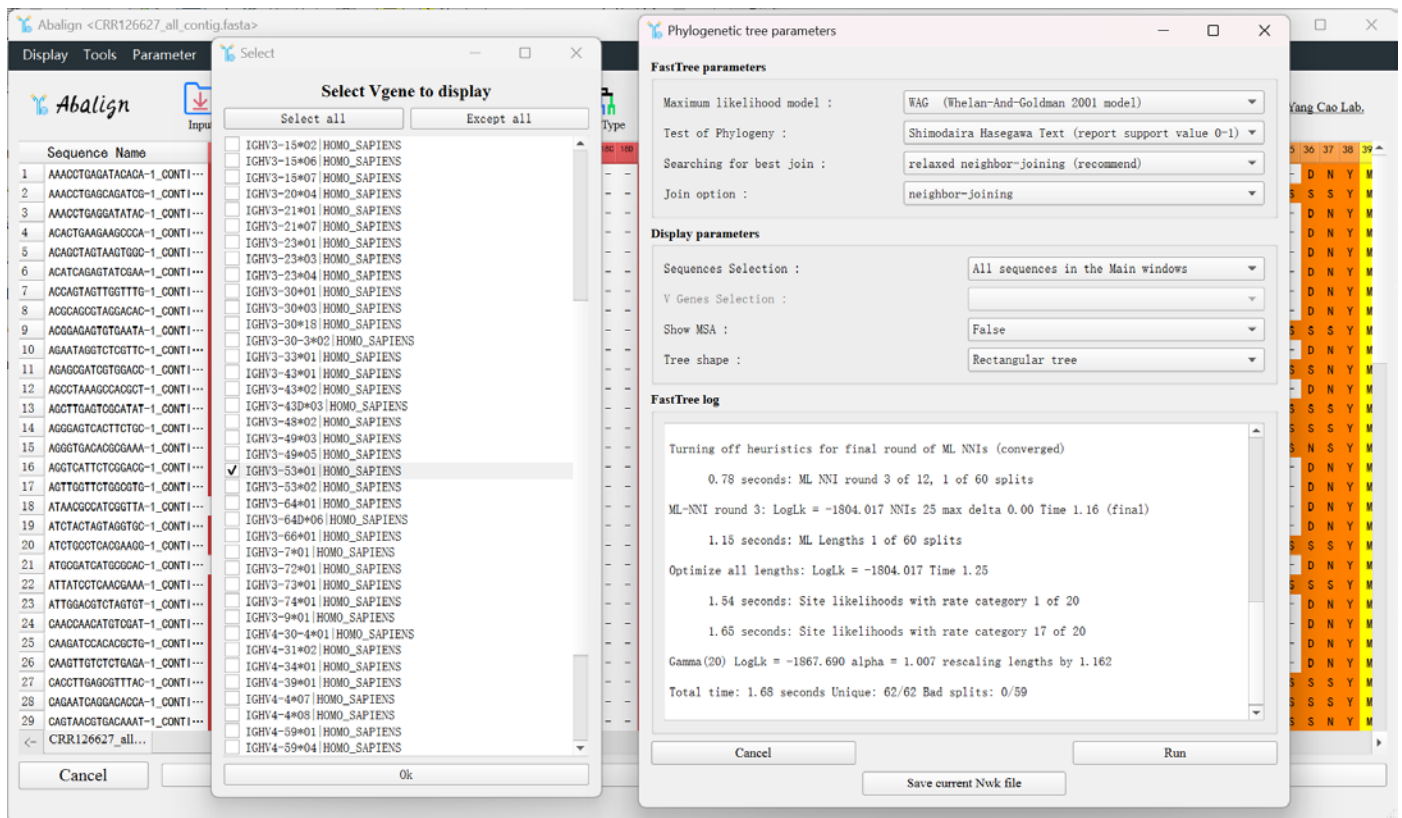


Figure 12. Use sequences belonging to IGHV3-53*01 to build a phylogenetic tree.

Step 5 View the phylogenetic tree: Once the evolution tree is built, a visual window pops up that adjusts the view and displays other information about the tree through the button above the visual window (Fig 13). After selecting the node of the tree, users can also modify the attributes of the node, such as the color of the node, in the right window.

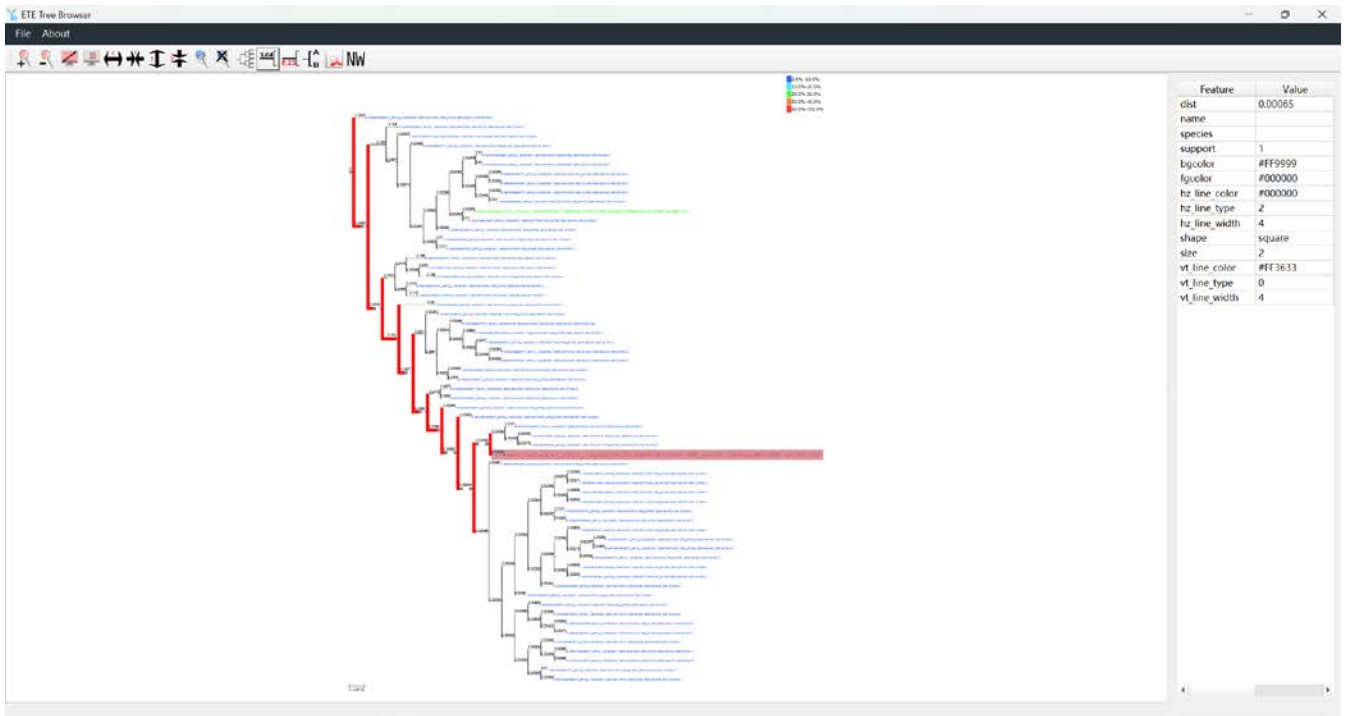


Figure 13. Results of phylogenetic tree construction using sequences belonging to IGHV3-53*01.

Step 7 Aid antibody humanization: If users need to know the frequency of residues used at each position in comparing with the known human BCR or antibodies, they can click **Tools->Unusual Residue** to get the information (Fig 14).

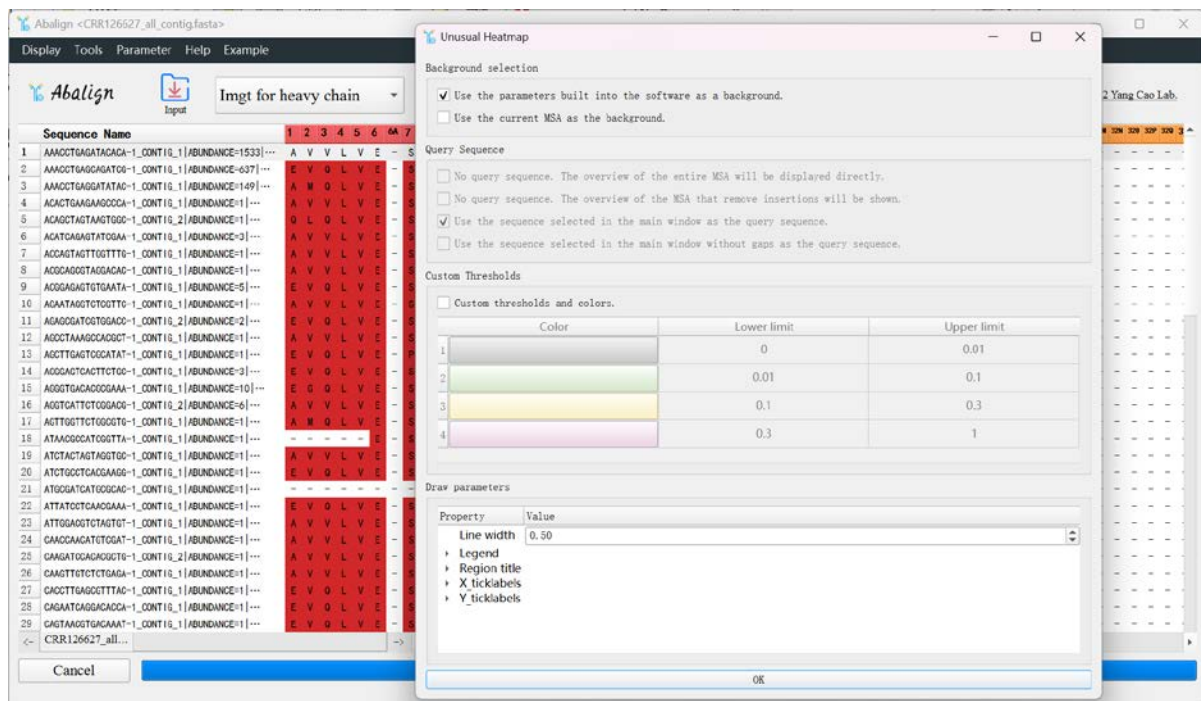


Figure 14. Select the first sequence in MSA to construct the unusual residue map.

By default, low frequency amino acids will be shown in gray, if they exist in the query sequence (Fig 15). It is recommended for finding substitution with the corresponding high frequency amino acids of human BCRs.

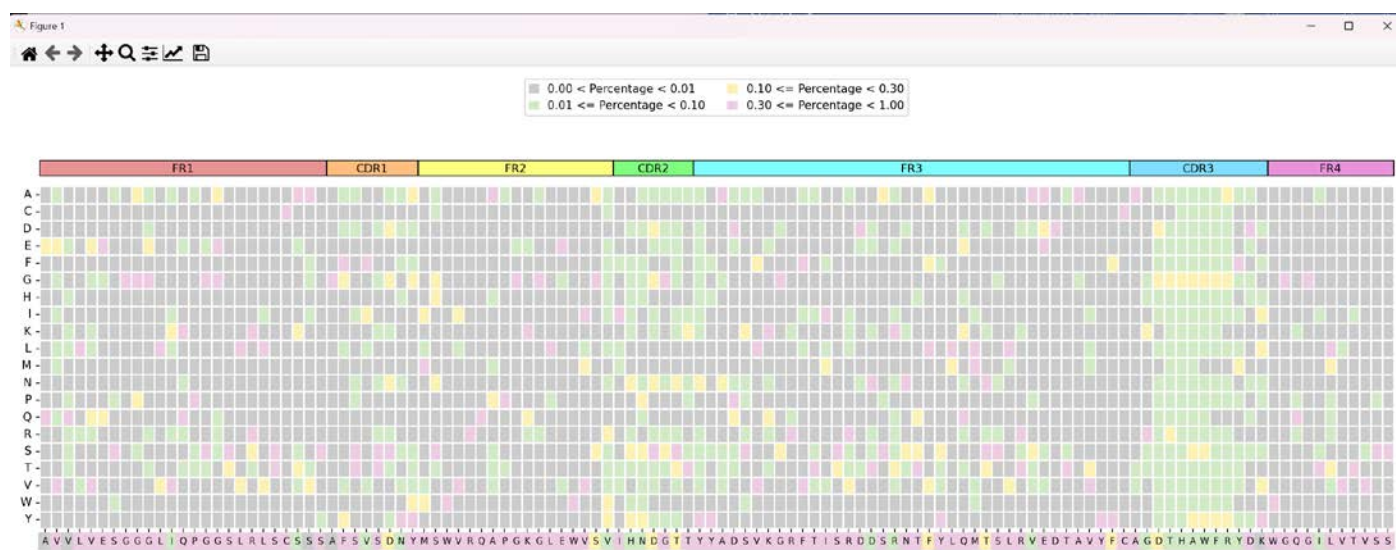


Figure 15. Result graph of unusual residues of the first sequence in MSA.

Step 8 Cluster sequences with Clonotypes: Clicking the “Clonotype” button in the toolbar, Abalign will sort the sequences by clonotype, and sequences with the same clonotype will be lined up together and highlighted by the same color.

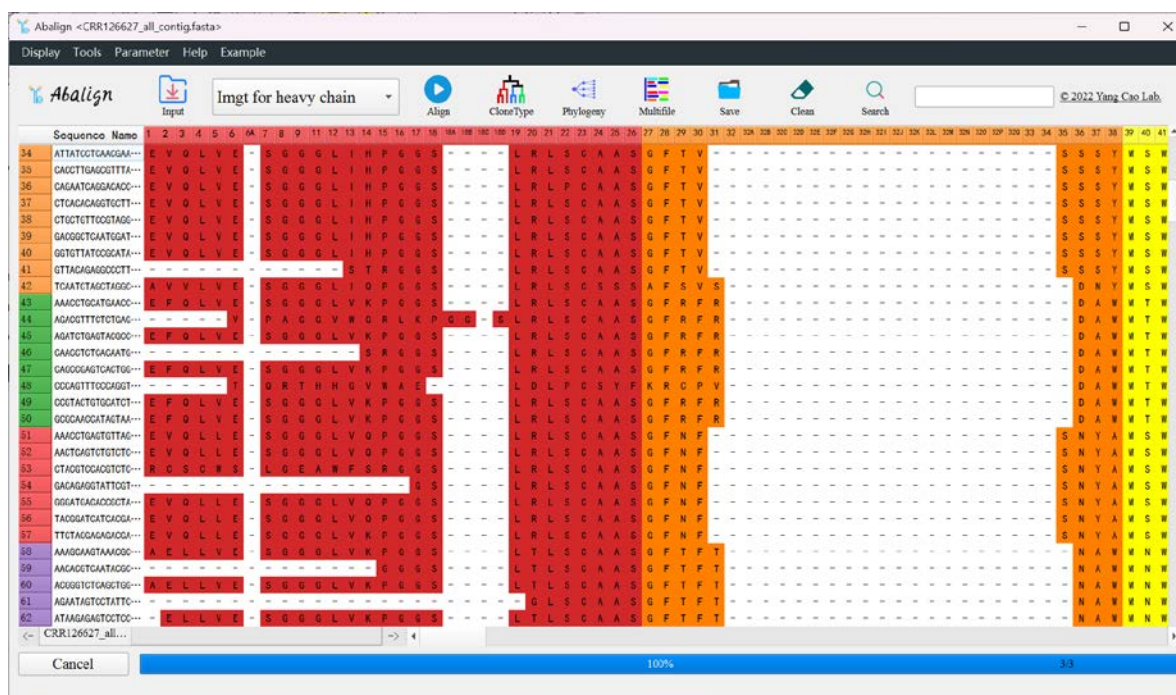


Figure 16. Rearrange sequences belonging to the same clone type together and render them

in the same color in the line label.

Step 9 Build of phylogenetic tree by clonotype: Once the Clonotype button was clicked, users can display the sequences belonging to a specific clonotype individually in the main window (Fig 17). Then, a phylogenetic tree can be constructed with these sequences.

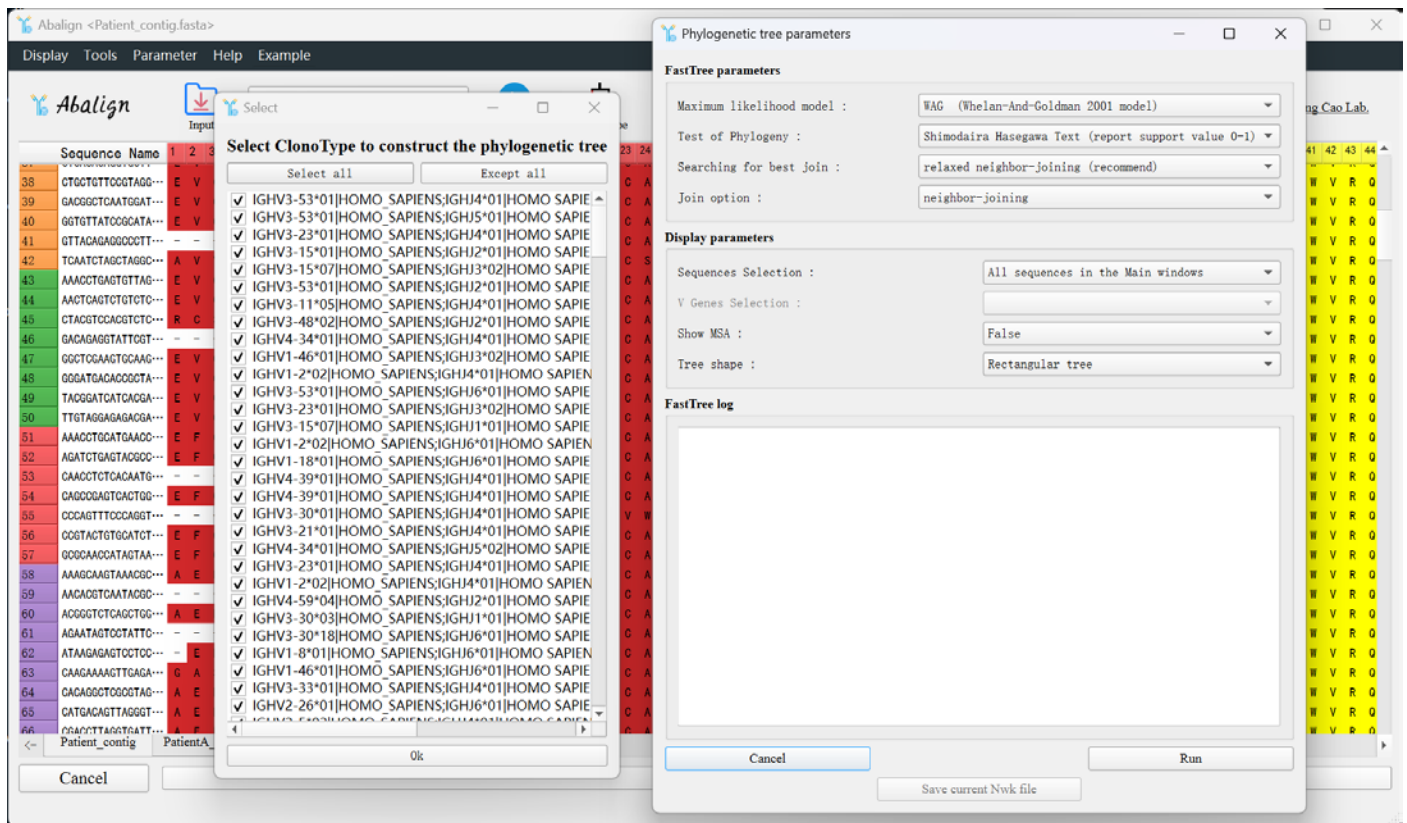


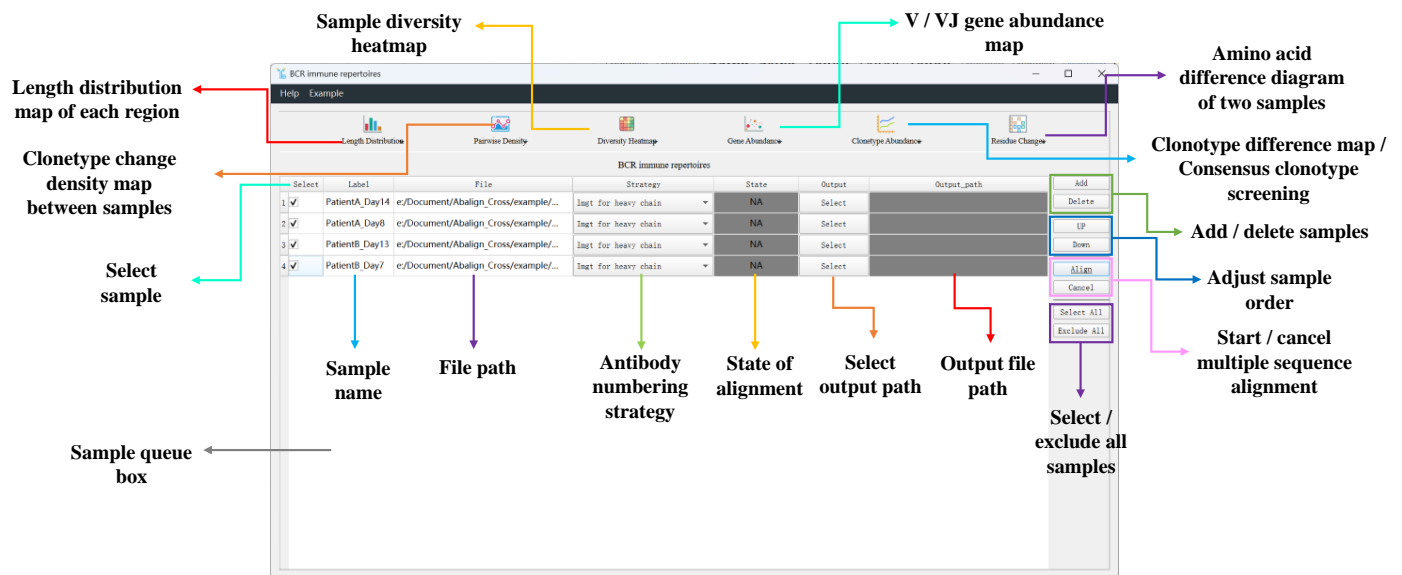
Figure 17. Selecting specific clonotypes to construct phylogenetic trees.

Step 10 Save File: The sequence displayed in the window can be saved by clicking the “Save” button in the toolbar. Thus, in combination with the previous [Display by Genes](#), [Display by Regions](#) and [Display by Clonotype](#), in addition to saving all sequences, we can personalize and save the sequences we need.

Step 11 Multi Task: Click on "[MultiFile](#)" in the main window to use this feature, which supports you to Align and cross-analysis of multiple files.

Multi-File Navigator

Navigator layout



Explanation of terms

Clonotype: Sequences sharing the same VJ genes and the same CDR3 regions are classified as the same clonotype.

Usage

Add file: Click "Add" to select the target files, then the corresponding files will be displayed in "BCR immune repertoires" table. The label of the file can be modified by changing the value of the "Label" column of the table.

Delete file: Check files in "Select" column of the table, then click "Delete" to delete files.

Multiple sequence alignment: Check files in "Select" of the table, then click "Align". Modify the value of "Strategy" to change the numbering strategy, and the alignment parameters can be changed in the window that pops up after clicking "Align". "State" shows the progress of alignment, and "Cancel" is used to cancel alignment in the queue.

Save file: Click buttons in "Output" column of the table to determine the output path. After saving, you will get the following files: **".fas"** and **".temp.txt"** both record the results of multiple sequence alignment of samples, and the difference is that the latter uses "*" to divide FRs and CDRs. **".number.txt"** records the antibody number used for multiple sequence alignment. **".abundance.txt"** and **".vabundance.txt"** record the abundance and proportion of each sequence and V, J gene respectively. **".clonotype.csv"** records the distribution of clonotypes in selected sample, and **".clonotype_seqs.csv"** records the sequence composition of each clonotype.

Tool buttons

Length Distribution: Count the length of each region of antibody for selected samples and draw the kernel density estimation curve (Fig. 18). The regions include FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4, full length and CDR1-FR4 (FR1 will be incomplete when the sequencing data quality is poor).

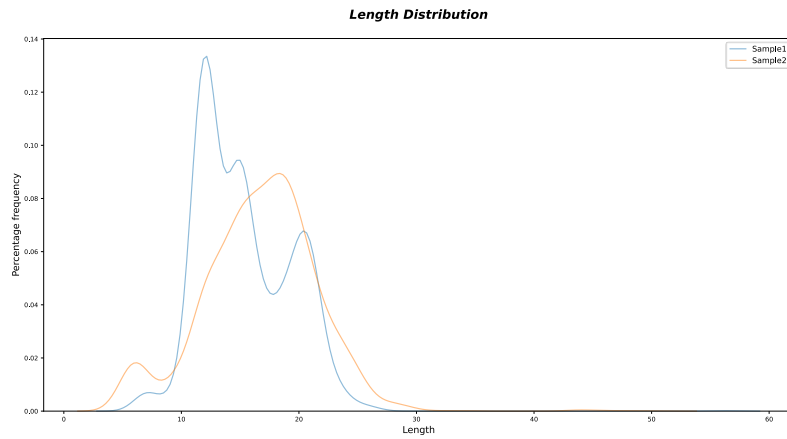


Figure 18. An example of length distribution in CDR3

The abscissa represents the length of the sequence, and the ordinate represents the percentage of the sequences with certain length. The lines with different colors represent different samples. The plot data can be saved by clicking “Save Sources”. After saving, you will get a file named **"length_distribution.csv"**, which records the abundance of each length type.

Pairwise Density: Count the changes of clonotypes abundance/density between two selected samples and draw the density map (Fig. 19).

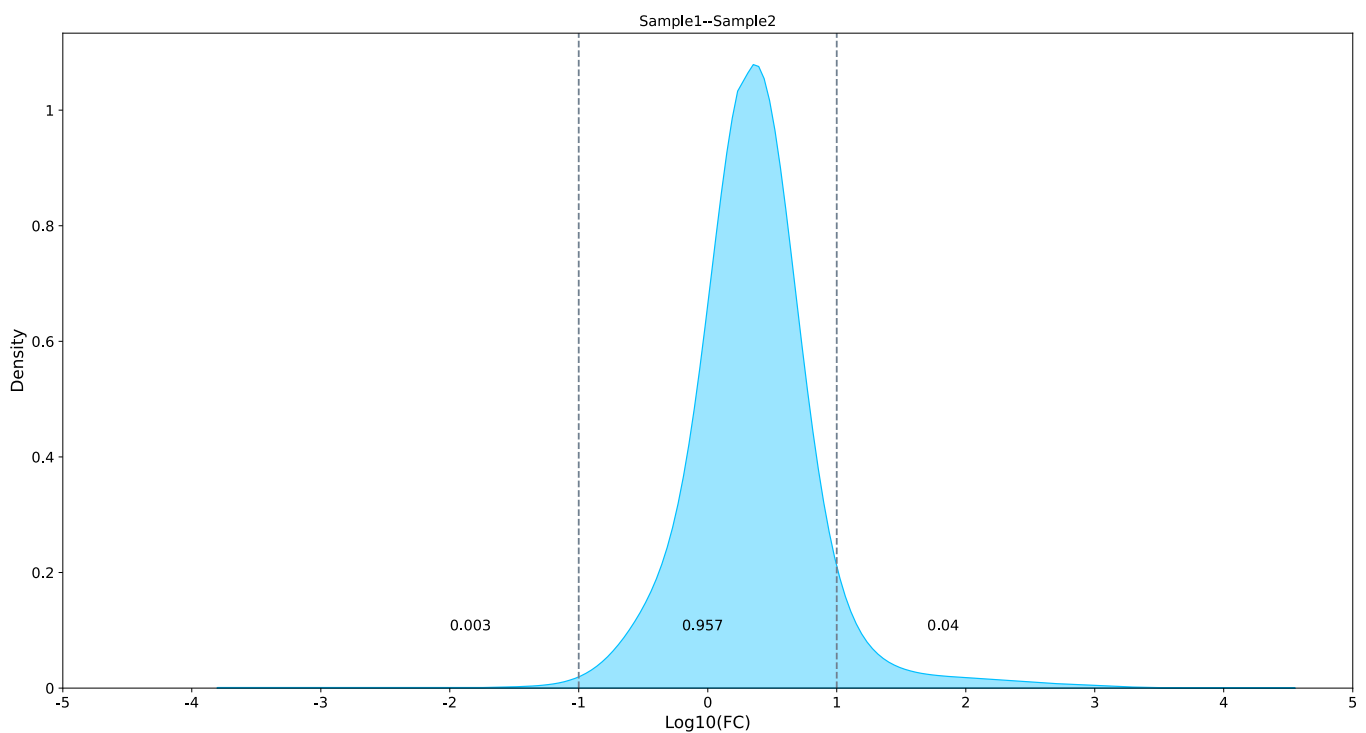


Figure 19. An example of clonotype differential density map between two samples

The abscissa represents the base 10 logarithm of the clonotype fold change, and the ordinate represents the density value. After removing the low abundance clonotypes, clonotypes are divided into 3 groups according to $\log_{10}(\text{FC})$. $\log_{10}(\text{FC}) > 1$ is the expanded group, $\log_{10}(\text{FC}) < -1$ is the reduced group, and $-1 < \log_{10}(\text{FC}) < 1$ is the constant group. The number represents the proportion of the clonotypes contained in each group.

Diversity Heatmap: Calculate the diversity indexes of clonotype among selected samples and draw the heatmap (Fig. 20).

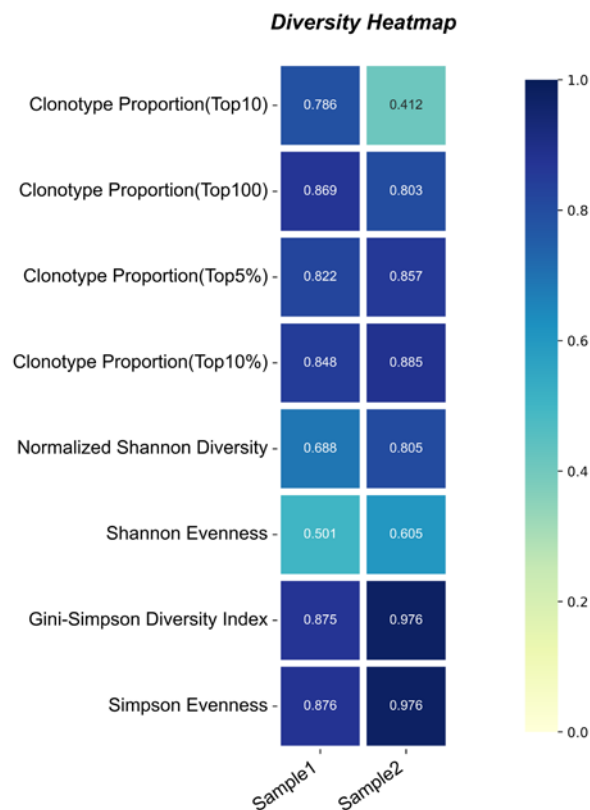


Figure 20. An example of diversity heatmap

The abscissa represents different samples, and the ordinate represents the diversity indexes. The darker the color, the higher the value, and the lighter the color, the lower the value.. The plot data can be saved by clicking “Save Sources”. After saving, you will get a file named **"diversity_heatmap.csv"**, which records the diversity indexes for selected samples.

Gene Abundance: Count the V/VJ gene abundance for selected samples and draw the histogram (Fig. 21) / scatter plot (Fig. 22).

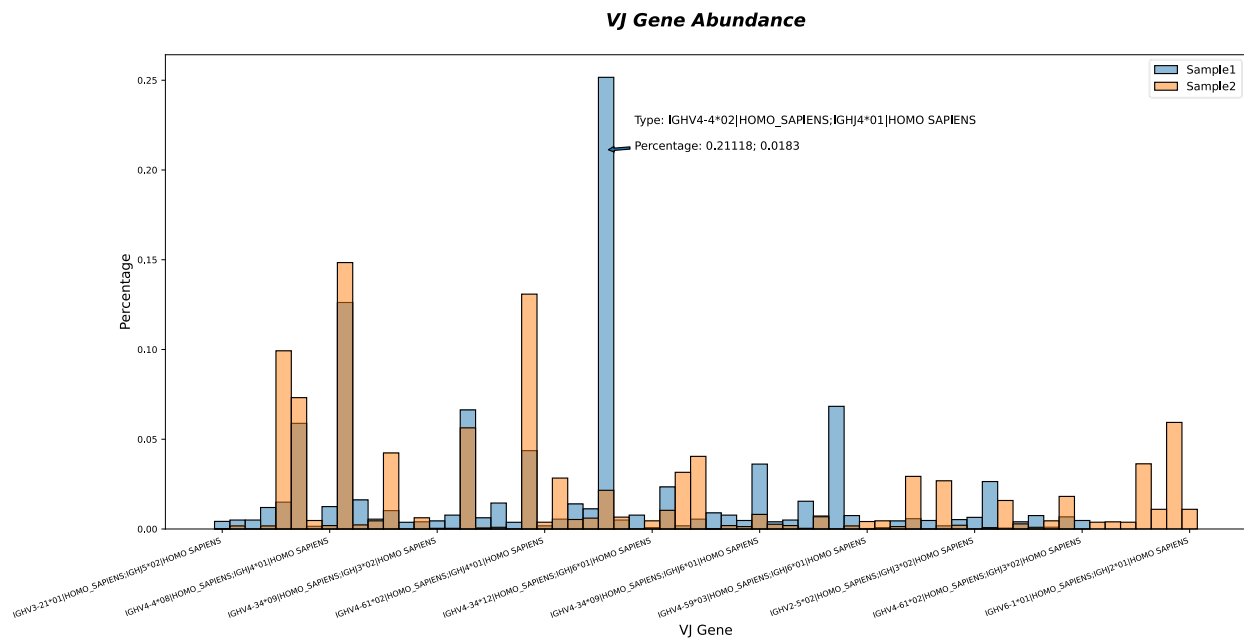


Figure 21. An example of VJ gene abundance histogram

The abscissa represents the type of V/VJ gene, and the ordinate represents the proportion of the specific gene. The different colored bars represent different samples. Details will be displayed when hovering.

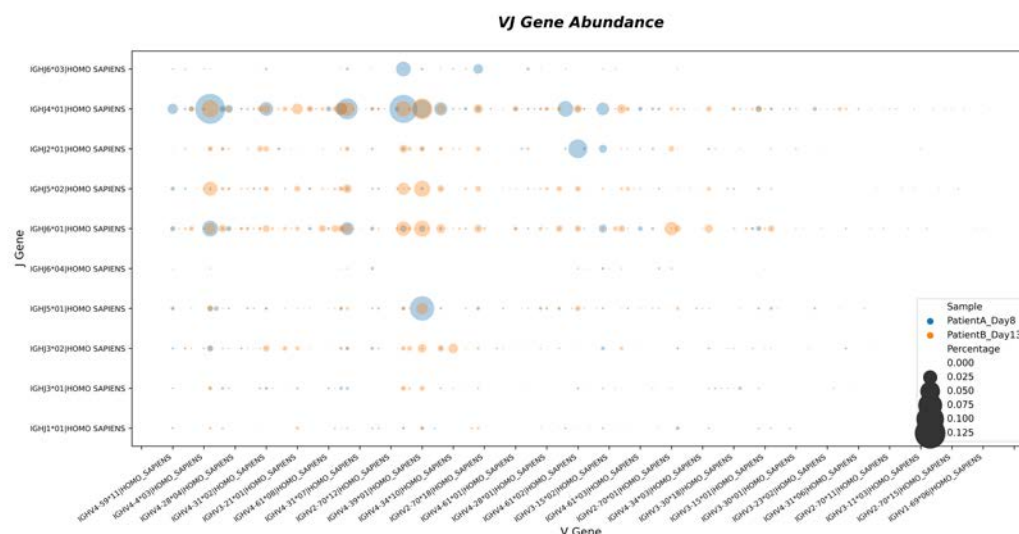


Figure 22. An example of VJ gene abundance scatter plot

The abscissa represents the type of V gene, and the ordinate represents the type of J gene. Points with different colors represent different samples, and the size of points represents the proportion of the specific VJ gene combination. Details will be displayed when hovering.

Clonotype Abundance: There are four options in this menu. “Clonotype Distribution” is used to display the distribution of clonotype abundance in the selected samples and draw the distribution bar plot (Fig. 23). "Clonotype Changes" is used to display the clonotype changes in the selected samples and draw the difference bar plot (Fig. 24). "Consensus Clonotypes" is used to display clonotypes that are shared among different samples and draw the consensus clonotypes bar plot (Fig. 25). "Consensus Clonotypes between groups" is used to display the expanded clonotypes that are shared among different groups. After clicking this button, the "Group Selector" will pop up (Fig. 26), which is used to group different samples and adjust the parameters for screening expanded clonotype . Shared expanded clonotypes results will be presented as an upset plot (Fig. 27).

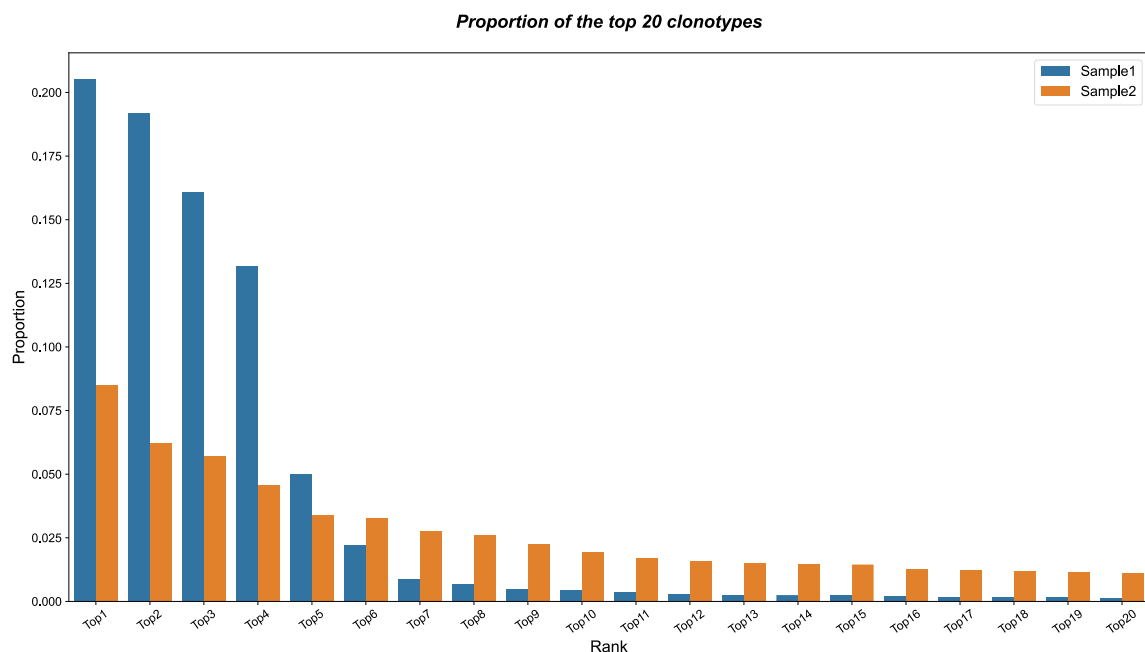


Figure 23. An example of clonotype distribution

The abscissa represents the top 20 clonotypes, and the ordinate represents the proportion of clonotypes, and bars with different colors represent different samples. You can see output file ([clonotype.csv](#) / [clonotype_seqs.csv](#)) for clonotype detail.

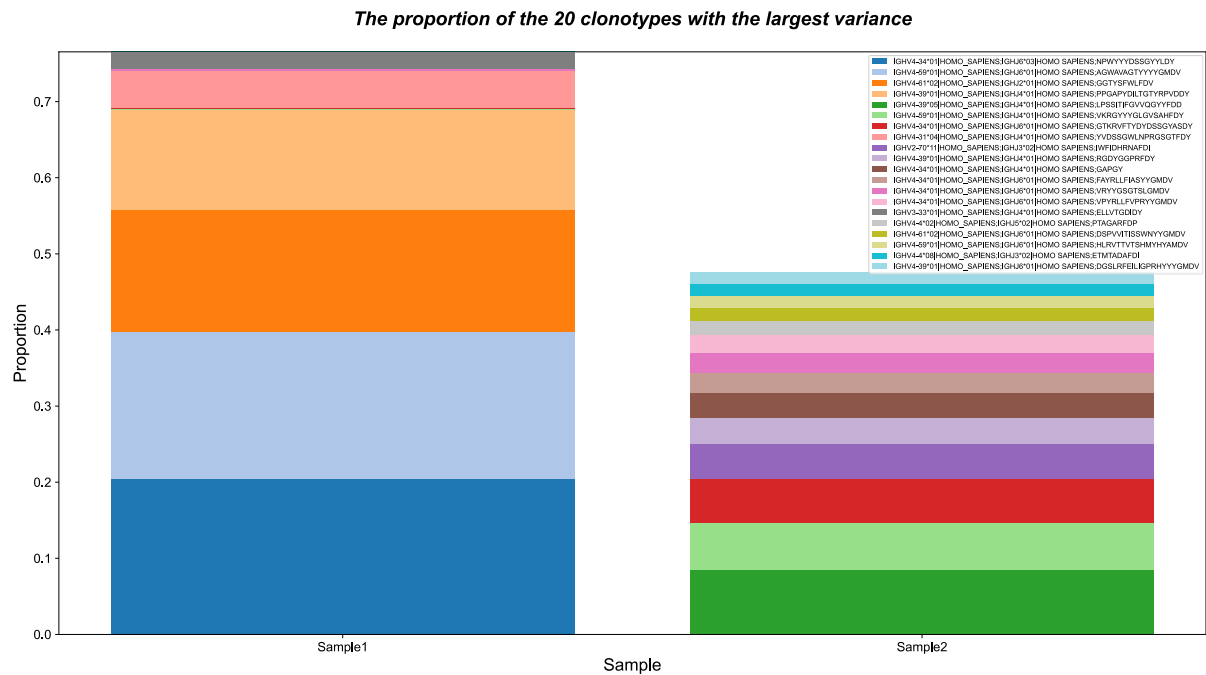


Figure 24. An example of clonotype changes

The abscissa represents different samples, and the ordinate represents the proportion of the top 20 clonotypes that vary across selected samples. The plot data can be saved by clicking “Save Sources”. After saving, you will get a file named "[clonotype_changes.csv](#)", which records the abundance and changes of all clonotypes in selected samples.

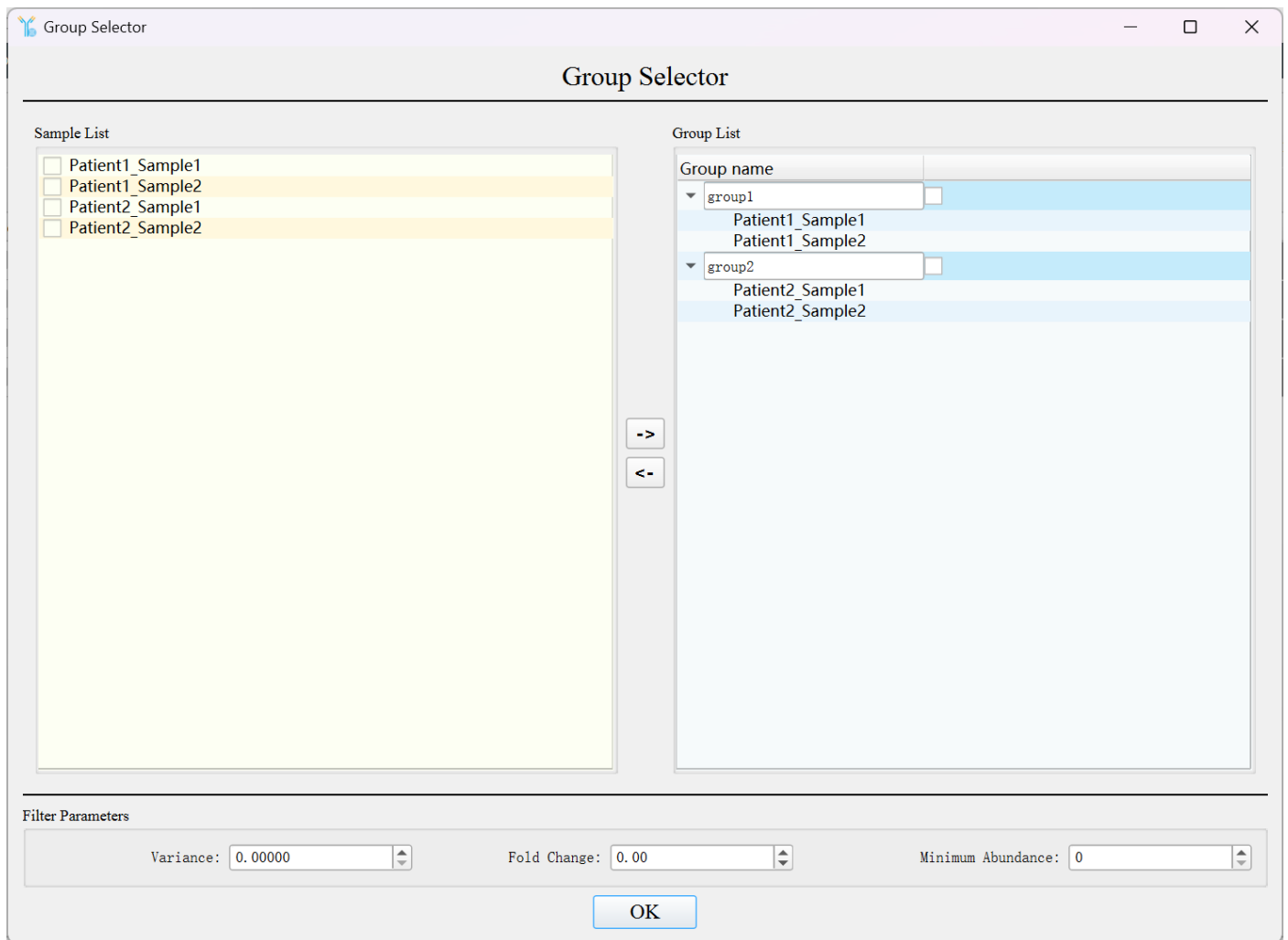


Figure 26. Group Select Parameters Dialog

"Sample List" shows all samples to be analyzed, which selected through the multi-file navigator. "Group List" is used to display the groups constructed by users, which can be generated by selecting the samples on the left and clicking the move button in the middle. Conversely, ticking the right groups and clicking the move button in the middle will remove the selected groups. "Filter parameter" is used to control the conditions for expanded clonotypes. It is worth noting that a group must contains two or more samples, and groups with a single sample will generate an error message.

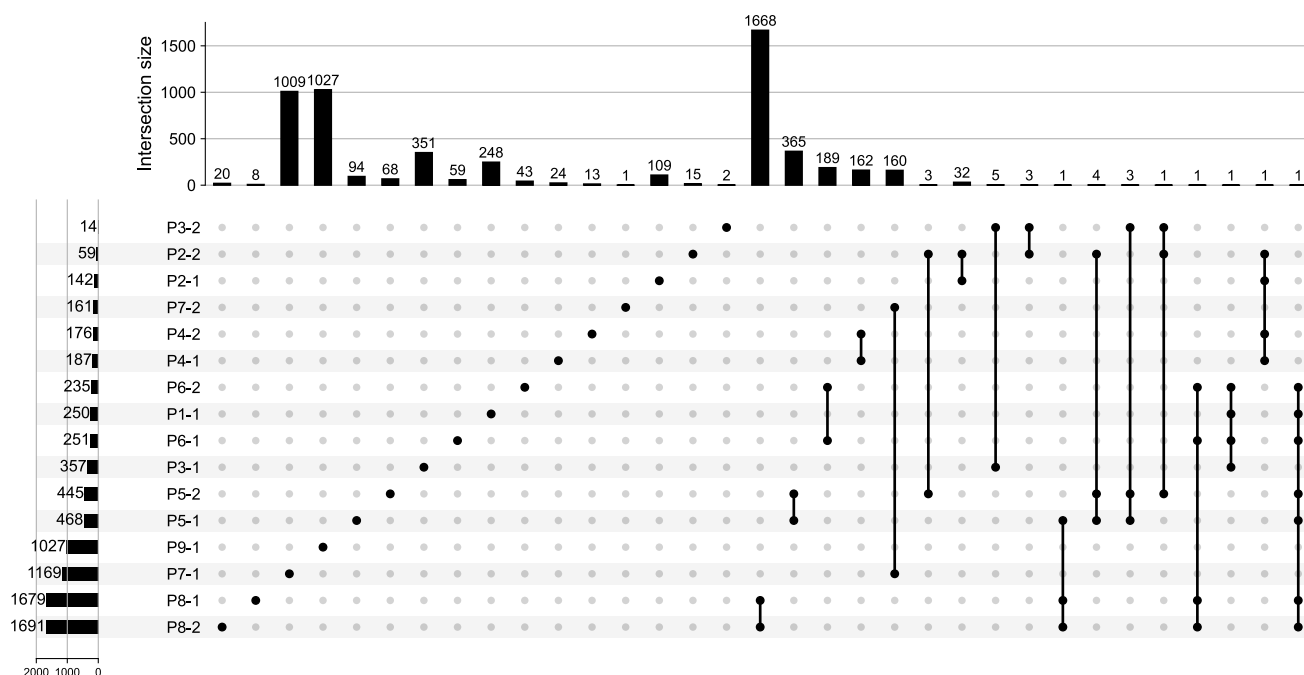
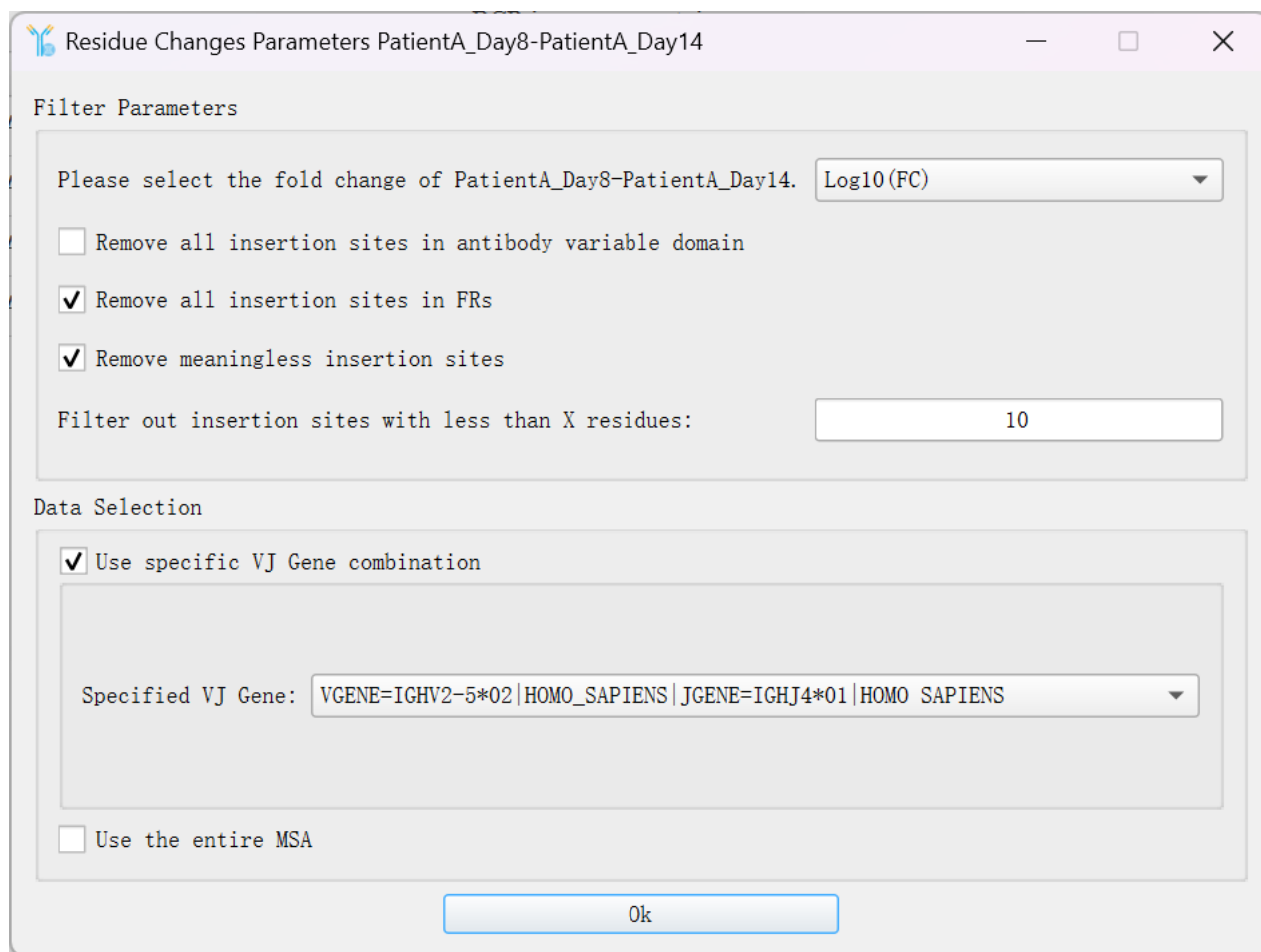


Figure 27. An example of consensus clonotype between groups

The left bar represents the number of clonotypes expanded in each group (for example, one patient represents one group), and the upper bar represents the number of shared clonotypes among different groups. The set information corresponding to the shared clonotypes is shown as the line, and the dot on the line represent the group within the set. The plot data can be saved by clicking “Save Sources”. After saving, you will get a series of files, including multiple files named **"group.csv"** and one file named **"between_groups.csv"**. The former records the changes of clonotypes in each group, and the latter records the occurrences of each clonotype among all groups.

Residue Changes: Count the preferences of amino acids between the two samples. After clicking this button, a dialog will pop up (Fig. 28), which is used to adjust the parameters for amino acid preference (Fig. 29).



Residue Changes Parameters PatientA_Day8-PatientA_Day14

Filter Parameters

Please select the fold change of PatientA_Day8-PatientA_Day14. Log10(FC)

☐ Remove all insertion sites in antibody variable domain

☒ Remove all insertion sites in FRs

☒ Remove meaningless insertion sites

Filter out insertion sites with less than X residues: 10

Data Selection

☒ Use specific VJ Gene combination

Specified VJ Gene: VGENE=IGHV2-5*02|HOMO_SAPIENS|JGENE=IGHJ4*01|HOMO_SAPIENS

☐ Use the entire MSA

Ok

Figure 28. Residue Preference Parameters Dialog

"Filter parameters" are used to control which amino acid difference positions to be displayed. Users can choose the fold change, whether to retain insertion positions or nonsense positions (the positions with residue occurrences lower than X). "Data selection" is used to control the sequences for analysis, allowing the users to select the entire sample or a specific VJ gene combination (only combinations that appear in the selected sample will be displayed). If you get an error message, make sure you select the same sample numbering strategy and chain type.

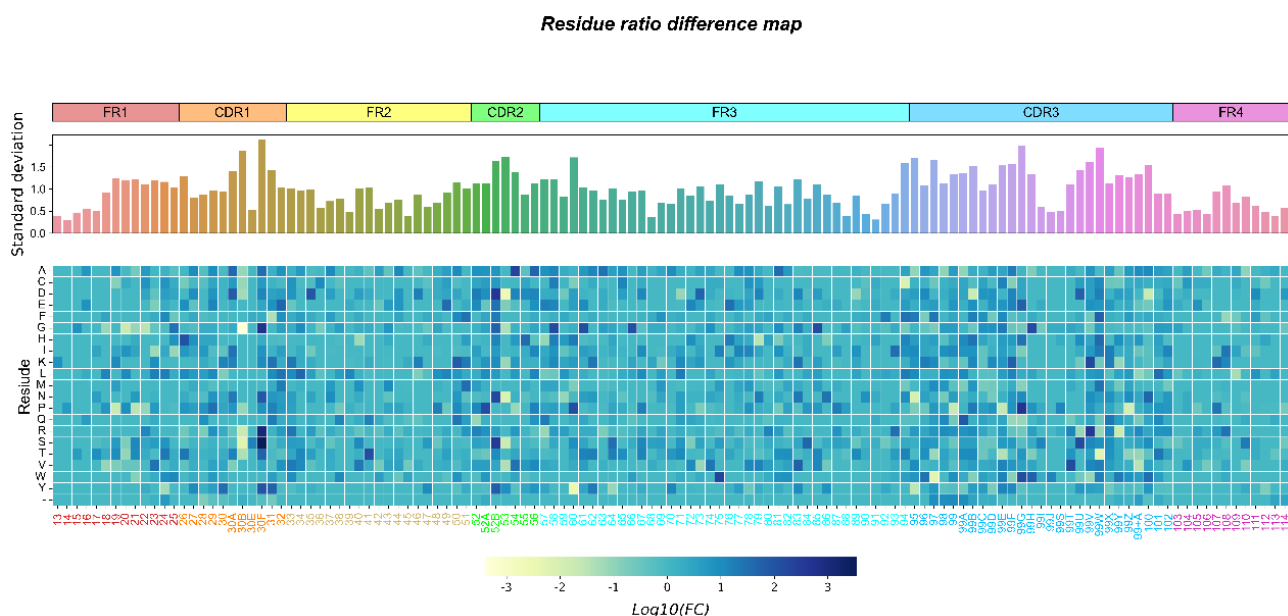


Figure 29. An example of residue preference map

The figure is divided into upper and lower parts. The ribbon at the top uses different colors to distinguish FRs and CDRs. The abscissa at the bottom represents each position in the variable domain. The ordinate of the upper figure represents the standard deviation, and the ordinate of the lower figure represents 20 amino acids and gaps. The histogram above represents the standard deviation of residues differences at each position. If the standard deviation at a certain position is larger, it means that the selection preferences of residues at this position is more significant. The lower figure calculates the fold change in the ratio of residues between the two samples and performs logarithmic processing to the base of 10. The results are presented by the shades of colors, with dark colors indicating positive selection of residues and light colors indicating negative selection of residues. The plot data can be saved by clicking “Save Sources”. After saving, you will get a **csv** file, which records the amino acid preferences of each position in the variable domain.

Notice

1. If there is no response during the operation of the program, wait a little and the program is

still running.

2. Multiple sequence alignment is for input files. If multiple sequence alignment is carried out, then multiple sequence alignment is carried out again, or multiple sequence alignment is carried out for input files, rather than after alignment.
3. Clicking the Cancel button can only terminate the alignment process and cannot be canceled by Cancel when the alignment sequence is finally loaded.
4. If you encounter an unusual situation where the button turns gray and cannot be clicked (for example, when the software is not performing a task), you can click the Cancel button to restore it.

This software is developed by Yang Cao Laboratory, College of Life Sciences, Sichuan University. The main developers are Yang Cao, Fanjie Zong, Chenyu Long, Wanxin Hu and Zhixiong Xiao. If you have any opinions or suggestions, please contact cy_scu@yeah.net.

Reference

- [1] Li L, Chen S, Miao Z, et al. AbRSA: a robust tool for antibody numbering[J]. Protein Science, 2019, 28(8): 1524-1531.
- [2] Młokosiewicz J, Deszyński P, Wilman W, et al. AbDiver: a tool to explore the natural antibody landscape to aid therapeutic design[J]. Bioinformatics, 2022, 38(9): 2628-2630.
- [3] Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix[J]. Mol Biol Evol. 2009 Jul;26(7):1641-50.
- [4] Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data[J]. Molecular biology and evolution, 2016, 33(6): 1635-1638.
- [5] Olsen T H, Boyles F, Deane C M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences[J]. Protein Science, 2022, 31(1): 141-146.