Abalign MANUAL

Introduction 4

Installation 5

Windows: 5

Linux: 5

MacOS: 5

Uninstallation 7

Windows: 7

Linux: 7

MacOS: 7

Software Layout 8

Tool Buttons 8

Input: 8

Align: 8

Clonotype: 8

Phylogeny: 9

Multifile: 10

Save: 10

Clean: 10

Search: 10

Menu Bar 10

Display 10

Remove Duplication 10

Filter Level 10
Sequence Color Mode 11
Sequence Render Mode 11
Display Full MSA 11
Display by Clonotype 11
Display by Genes 11
Display by Regions 11
Tools 11
V Gene 11
Abundance 12
Seqlogo (only for linux version) 13
Unusual Residue 13
Length Distribution 15
Parameter 15
Align Parameters 15
Temporary Path 16
Example 16
Example 16
Usage Case 17
Multi-File Navigator 22
Navigator Layout 22
Explanation Of Terms 22
Clonotype 22
Usage 22

Add File 22
Delete File 23
Multiple Sequence Alignment 23
Save File 23
Tool Buttons 23
Length Distribution 23
Pairwise Density 24
Diversity Heatmap 25
Gene Abundance 25
Clonotype Abundance 26
Residue Changes 29
Reference 31

Introduction

Multiple sequence alignment (MAS) has long been used as a powerful method to investigate the evolutionary, structural and functional properties of protein families. It is also a fundamental technique in recent deep-learning based protein 3D structure predictors. Though existing MSA methods have been well-established, they are not suitable for high-throughput computation, and do not fulfill the needs of processing BCRs or antibody sequences, because the highly variable regions cannot be well aligned, without the prior knowledge of gene recombination and hypermutation in antibody maturation. To our knowledge, no MSA tool is particularly designed for BCR alignment up to day. To address this issue, we developed Abalign, which is a high-throughput and accurate MSA tool based on AbRSA[1]. Abalign incorporated the heuristic knowledge of antibody numberings, including IMGT, Kabat, Chothia and Martin, and 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 was implemented in a user-friendly stand-alone program with interactive and visual interfaces, which support the multiple sequence alignment, as well as clustering, antibody numbering, delimiting CDR, constructing phylogenetic tree, VJ gene determination, clonotype analysis, aiding humanization, comparing BCR immune repertoires, etc. by just clicking the buttons. In addition, it supports the cross-analysis of multiple B cell receptor immune repertoire data to investigate information like shared clonotypes, or residue preferences, etc. Abalign can complete the alignment and analysis of 1 GB of DNA FASTA files at a very fast speed on a PC with only 16G of RAM. Abalign will profit immunoinformatic and pharmaceutical communities by analyzing massive BCRs or antibodies and making new discoveries.

Installation

Table 1. Operation System requirements.

OS	Version
Linux	Ubuntu 18.04, Ubuntu 20.04, Ubuntu 22.04
Windows	10, 11
macOS_x86	10.14, 10.15, 11, 12

Table 2. Hardware minimum requirements.

Processors	AMD or Intel Processors
Memory	8 GB RAM
Hard disk	40 GB of available disk space

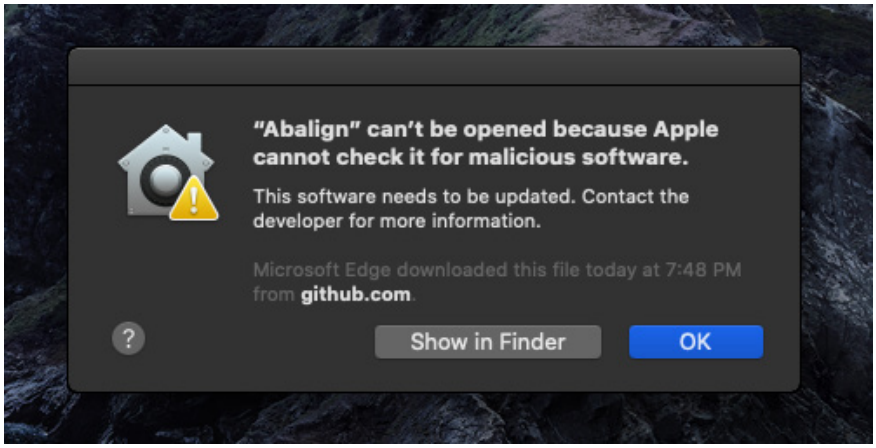
Windows: After extracting the files, go to the extracted folder, and double-click **Abalign_setup.exe** to start the installation guide.

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.

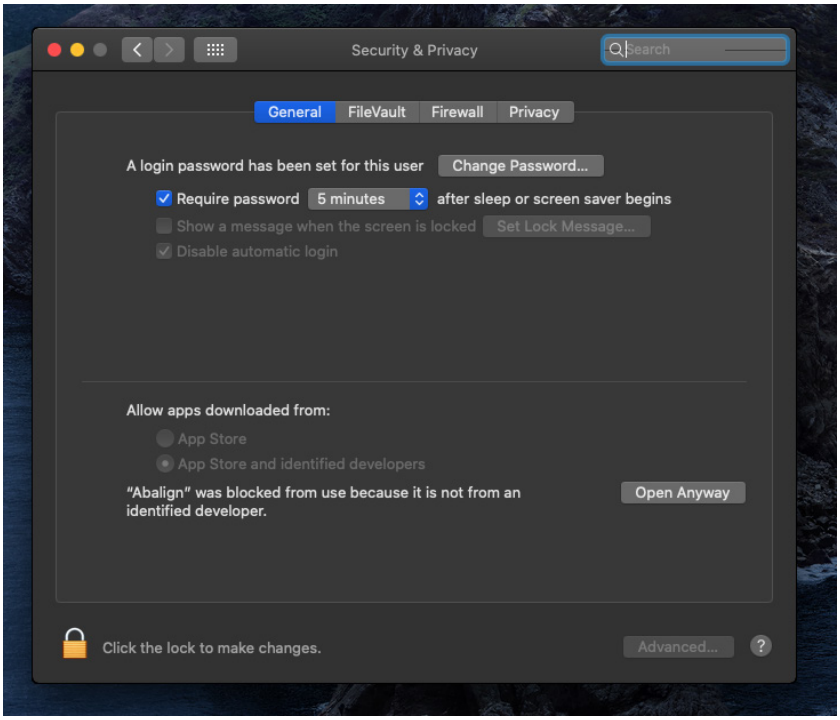
1. chmod +x Abalign_installer.run
2. ./Abalign_installer.run

MacOS: After extracting the files, go to the extracted folder, and double-click **Abalign.app** to run Abalign directly. Please move Abalign.app into the Applications folder. Abalign is currently under compatible testing for MacOS. If you meet any problems, please contact us(cy_scu@yeah.net).

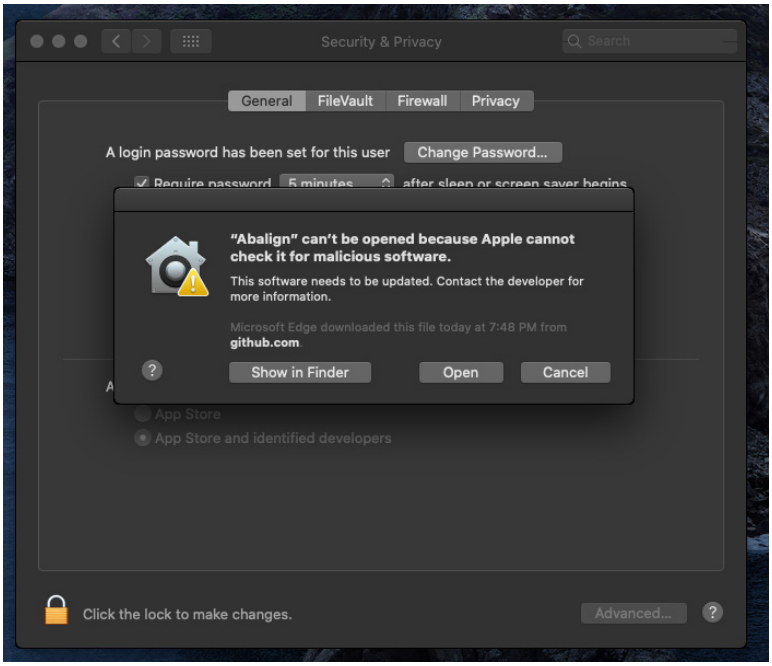
Tips 1: When you run Abalign for the first time on MacOS, the following dialog box may pop up. This is normal, please click "OK" in this dialog box.



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.



Tips 2: If a dialog prompts "Abalign is damaged and can't be opened, You should move it to the Trash" under MacOS, execute "sudo xattr -r -d com.apple.quarantine /Applications/Abalign.app" in the terminal to resolve the issue.

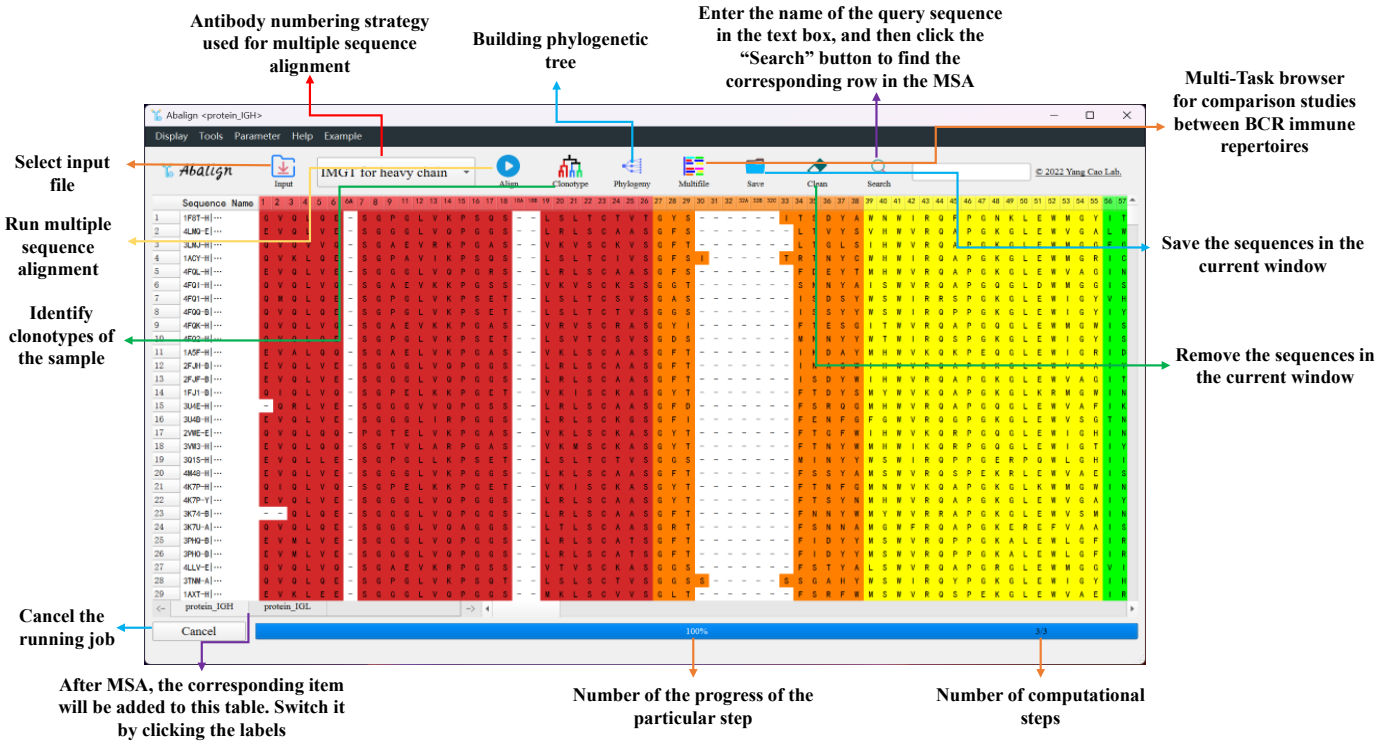
Uninstallation

Windows: Go to the installation path, double-click **unins000.exe** to uninstall Abalign, and click **Yes** to confirm.

Linux: Execute "**Abalign_uninstall.sh**" in the terminal to uninstall Abalign, if write Abalign to the PATH. Otherwise, delete the Abalign installation folder to uninstall Abalign directly.

MacOS: Delete "**Abalign.app**" directly to uninstall Abalign. In addition, Abalign's configuration is written to "**~/Library/Preferences/Abalign_conf.txt**", delete this file to uninstall Abalign completely.

Software Layout



Tool Buttons

Input: Select the input file in FASTA format.

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

Clonotype: Identify clonotypes for the input sequences. The clonotype is defined as sequences that share the same V and J genes as well as the same CDR3 [2]. It should mention that this needs to run after "Align".

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 which users can adjust the parameters for drawing the phylogenetic tree (Fig. 2).

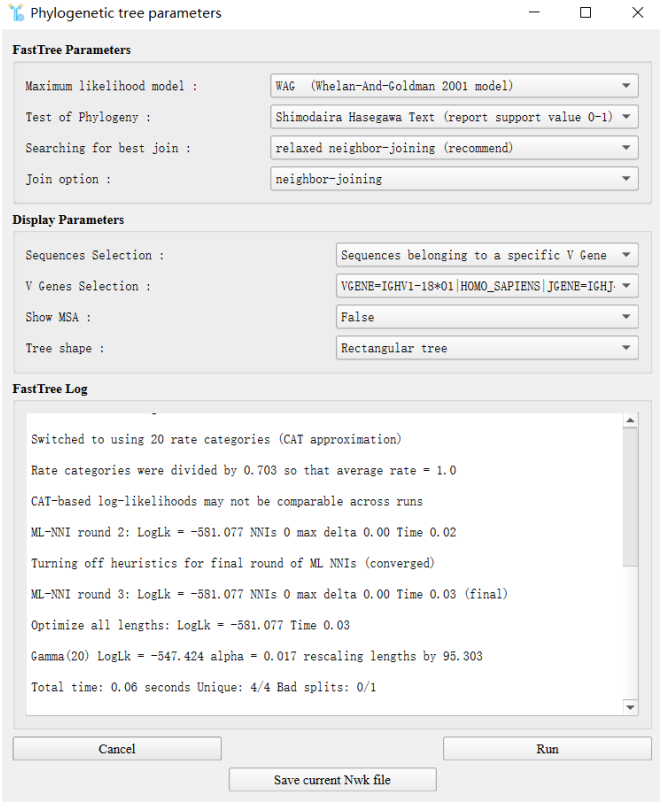


Figure 1. Phylogenetic Tree Parameters Dialog.

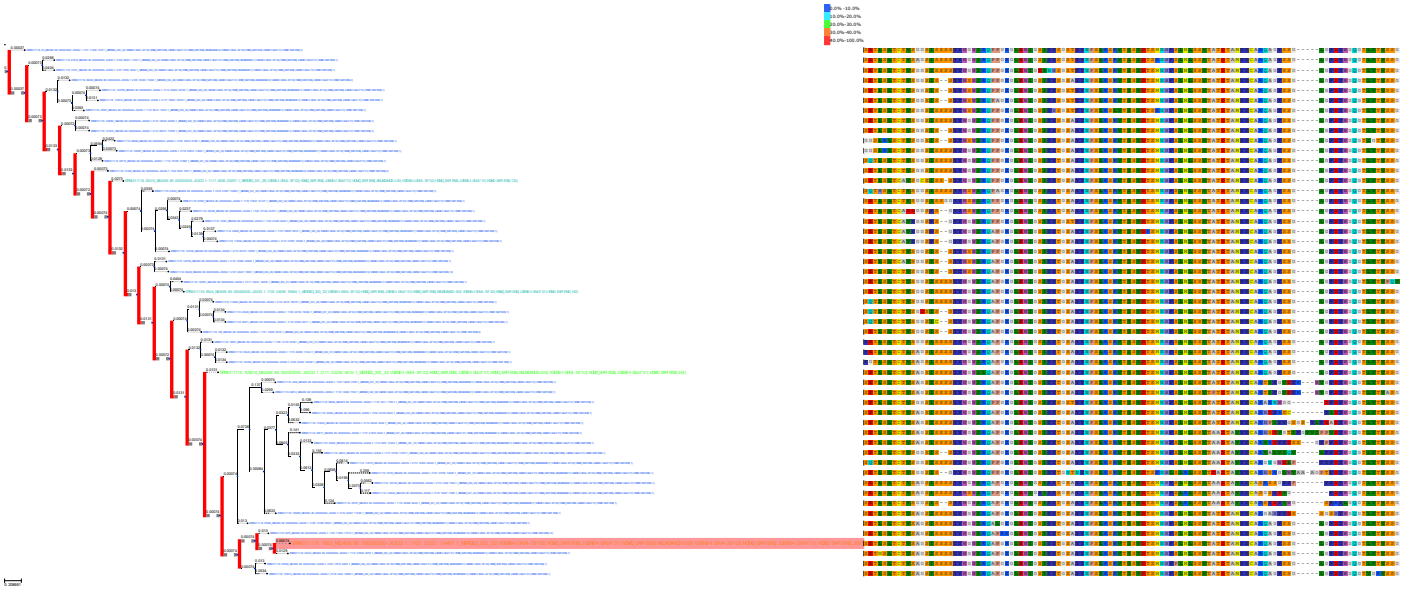


Figure 2. An example of a phylogenetic tree with corresponding MSA. The figure on the left is the phylogenetic tree of the selected sequences, where highlighted sequences represent high-abundance. The multiple sequence alignment corresponding to tree is showed on the right and highlighted with the mutations.

Multifile: After clicking this button, a 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.

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 domain will not be displayed, but the duplicated sequences are still accounted for abundance analysis. This option is enabled by default, and can be adjusted in **"Align Parameters"**.

Filter Level: Filter sequences by the length of the variable domain. There are four levels of length filtering: **"Off"**, **"Soft"(default)**, **"Normal"** and **"Strict"**. **"Off"** indicates no length limit for each region. **"Soft"** requires at least 1 amino acid in each region. **"Normal"** and **"Strict"** limit the region length according to antibody data with known structures. This option can be adjusted in **"Align Parameters"**.

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.

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 regions"** will divide the antibody sequences into different FRs and CDRs and render them in different colors. **" Color by amino"** renders different residues in different colors depending on the types of residues.

Display Full MSA: Select this option will display all the variable domain.

Display by Clonotype: Select this option will display the specific clonotypes in the sample, which needs clicking this button of **" Clonotype"** first.

Display by Genes: This menu has three options (**"Display by V genes"**, **"Display by J genes"** and **"Display by VJ genes"**), by clicking on different options you can display the sequences that match the conditions.

Display by Regions: There are seven options in this menu, corresponding to the seven regions of the variable domain. You can tick one or more options to show the regions.

Tools

V Gene: Here are two editable options. **"Species"** provides the V gene species to select. If selecting **"HOMO SAPIENS"**, the V gene identification will only use human germline gene

database. **"V Gene alignment"** shows the alignment of each sequence with top 5 scoring V genes (Fig. 3).

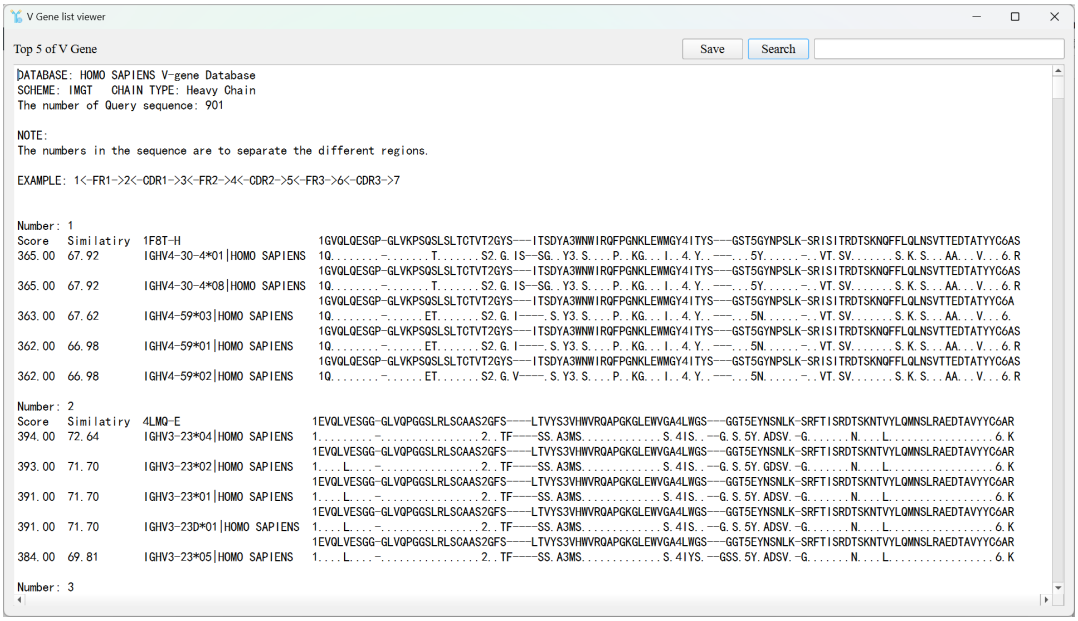


Figure 3. An example of V Gene list viewer.

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

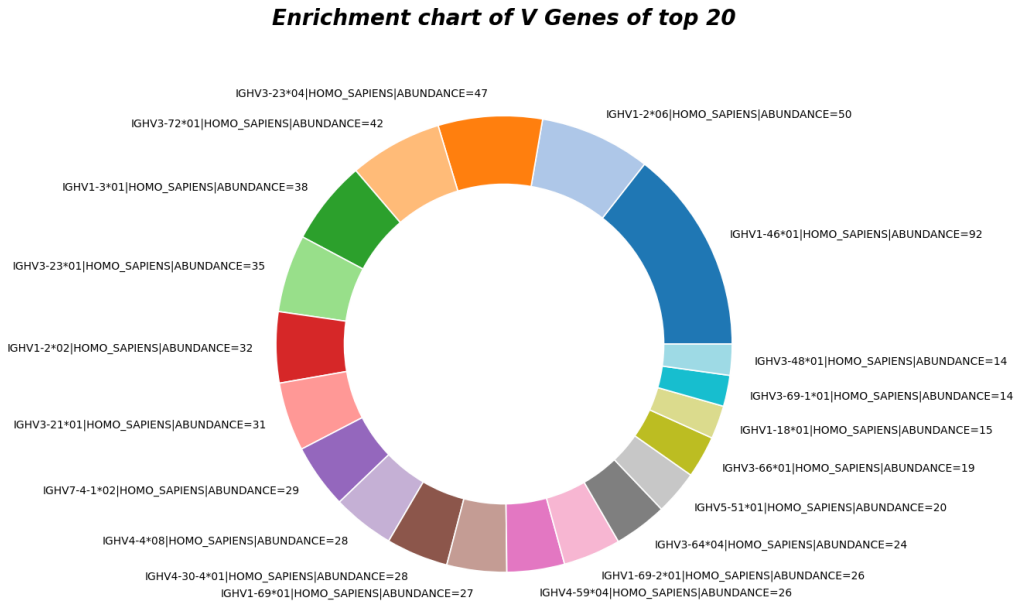


Figure 4. An example of Enrichment chart of V Genes of top 20. The abundance information is displayed in a pie chart with different colors representing different V genes, and the abundance information comes after the corresponding gene name.

Seqlogo (only for linux version): "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"** can change the rendering mode of the Seqlogo plot.

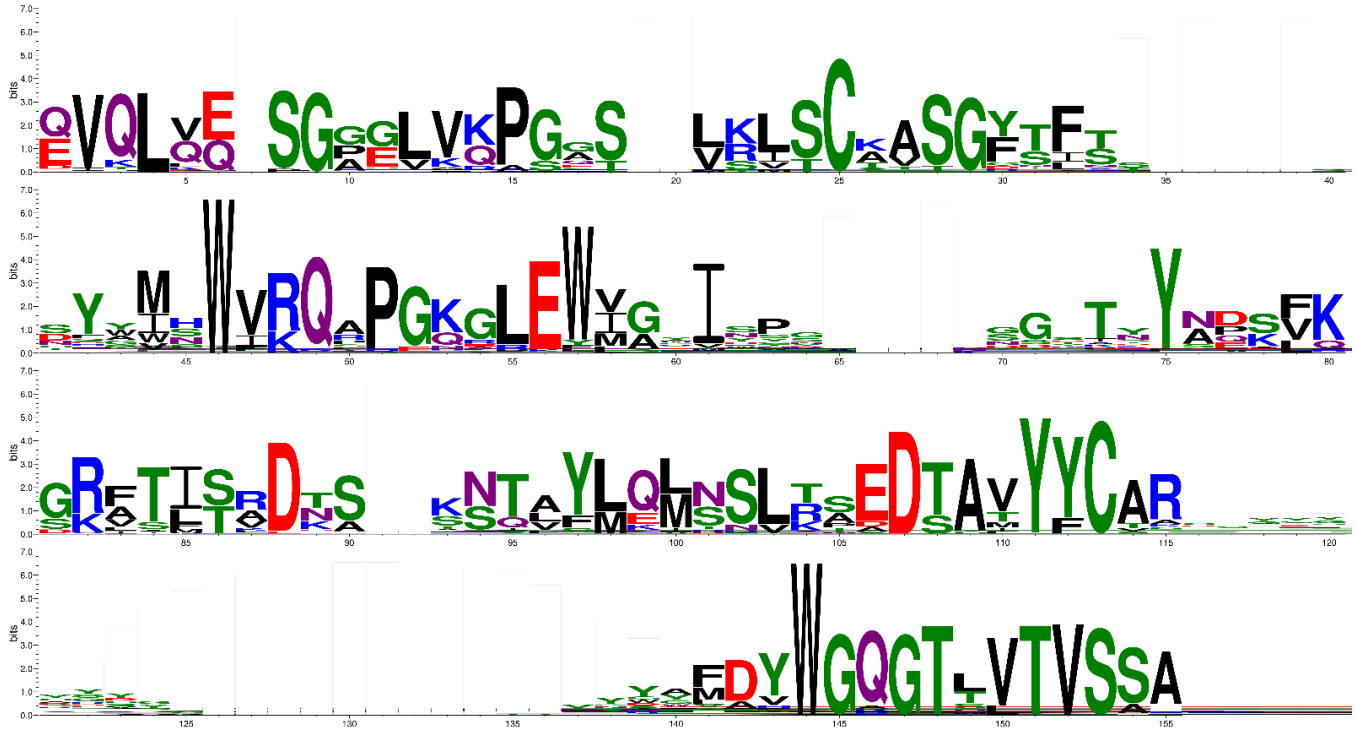


Figure 5. Seqlogo plot by Entropy. The ordinate indicates the entropy of amino acids, and the abscissa indicates the position in the variable domain. The larger the letter of the amino acid, the greater the entropy value.

Unusual Residue: This feature can show the unusual residues at each position of the selected query sequence by the residues proportion information in the dataset, which constructed by tens of millions of human sequences from OAS[5].(Fig. 7). It is used for antibody humanization. Users are also allowed to build an amino acid distribution graph with their own datasets (Fig. 6).

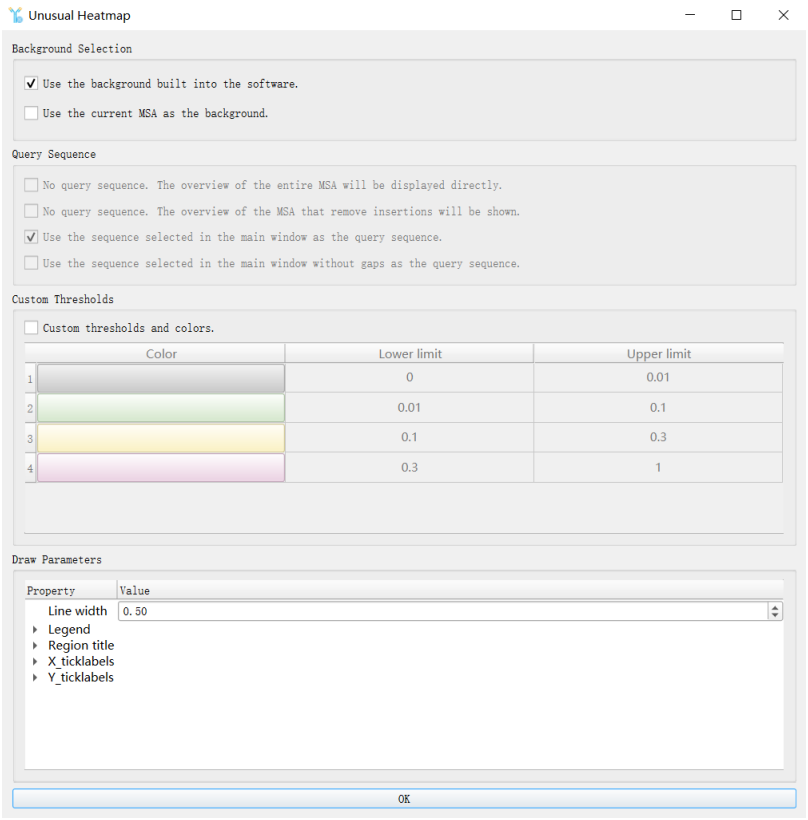


Figure 6. Unusual Residue parameters dialog. **"Background Selection"** is used to select the background comparing to the query sequence. **"Query Sequence"** allows the user to choose whether to use the query sequence and whether to ignore the gap in the query sequence. **"Custom Thresholds"** are used to adjust the thresholds of amino acid distribution and the colors they represent. **"Draw Parameters"** are used to customize the drawing parameters.

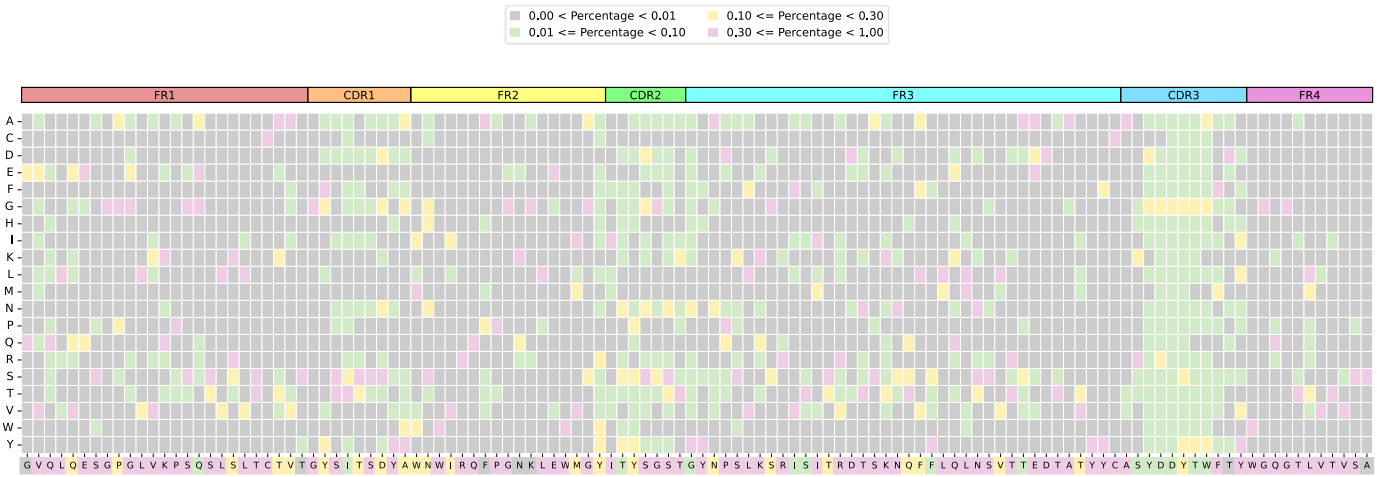


Figure 7. An example of unusual residue map. The residues with different frequencies will be rendered as different colors. The query sequence selected in the Multiple Sequence Alignment window is shown at the bottom of the heatmap, which is used for comparison with the human antibody reference dataset. The FR and CDR regions of the variable domain are marked with different colors on the top. Unusual residues are marked in gray.

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

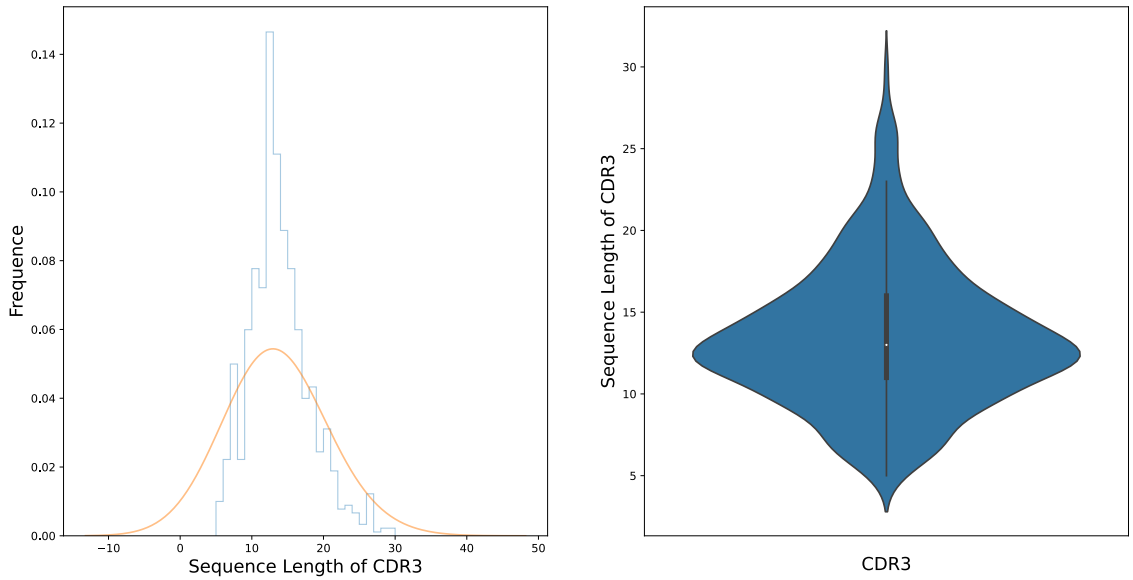


Figure 8. An example of Length Distribution of CDR3. The left figure is a length histogram. The abscissa indicates the length, and the ordinate indicates the proportion. The right figure is a violin plot. The ordinate represents the length, and the wider the width, the greater the number of sequences with the specific length.

Parameter

Align Parameters: Parameter options for MSA (Fig. 9).

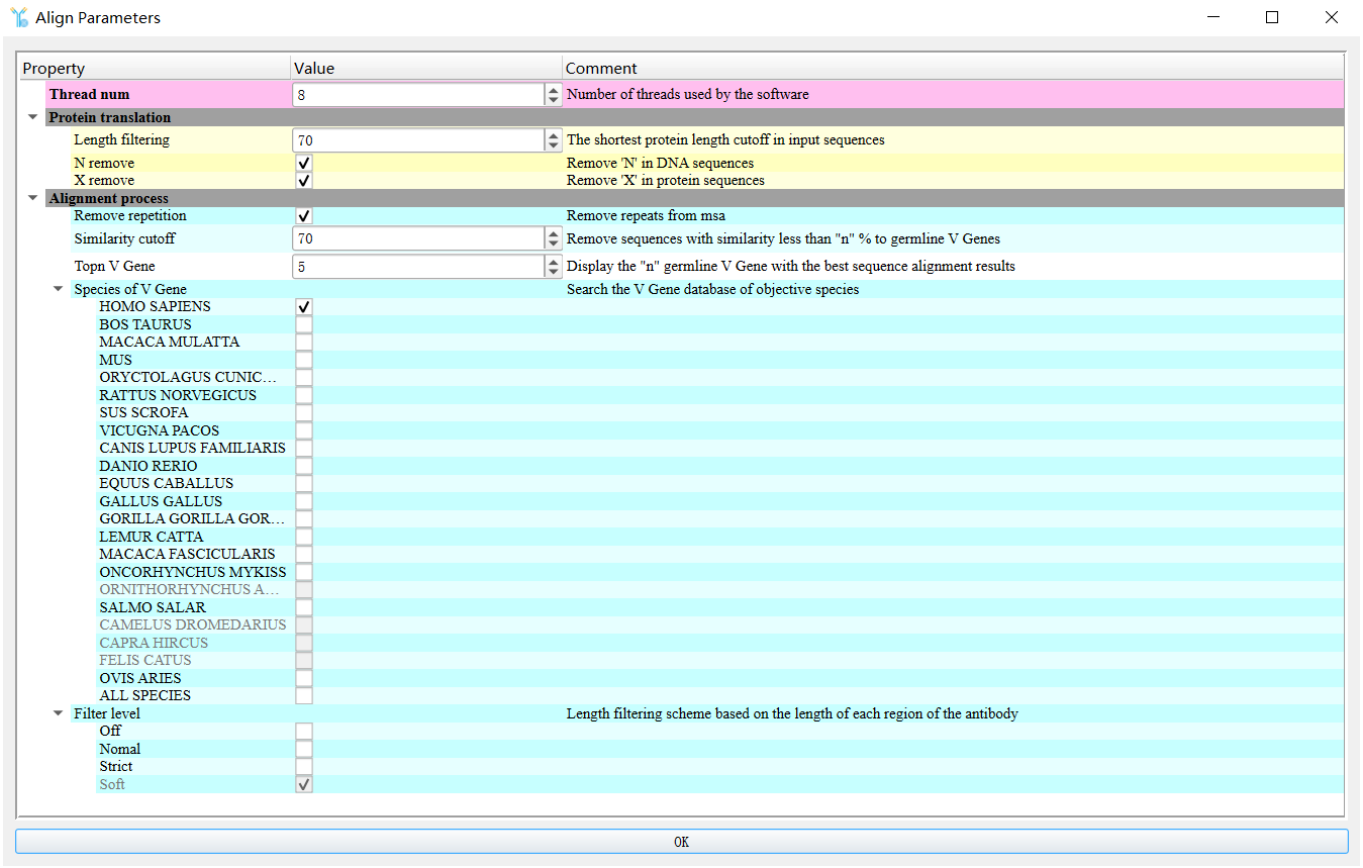


Figure 9. Align Parameters dialog. **"Align Parameters"** is divided into 3 columns, the first column lists the modifiable elements or functions. The second column lists the values and whether to enable certain functions. And the third column lists the introduction of the parameters.

Temporary Path: Change the path of the temporary files used by Abalign. Please restart the software after customizes the path.

Example

Example: BCR/antibody sequence data (DNA or amino acid) are provided for testing. **"Example"** offers two options, which are single mode and multi-file mode respectively, and users can find the 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 format file after clicking **"Input"**. (Alternatively, 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.

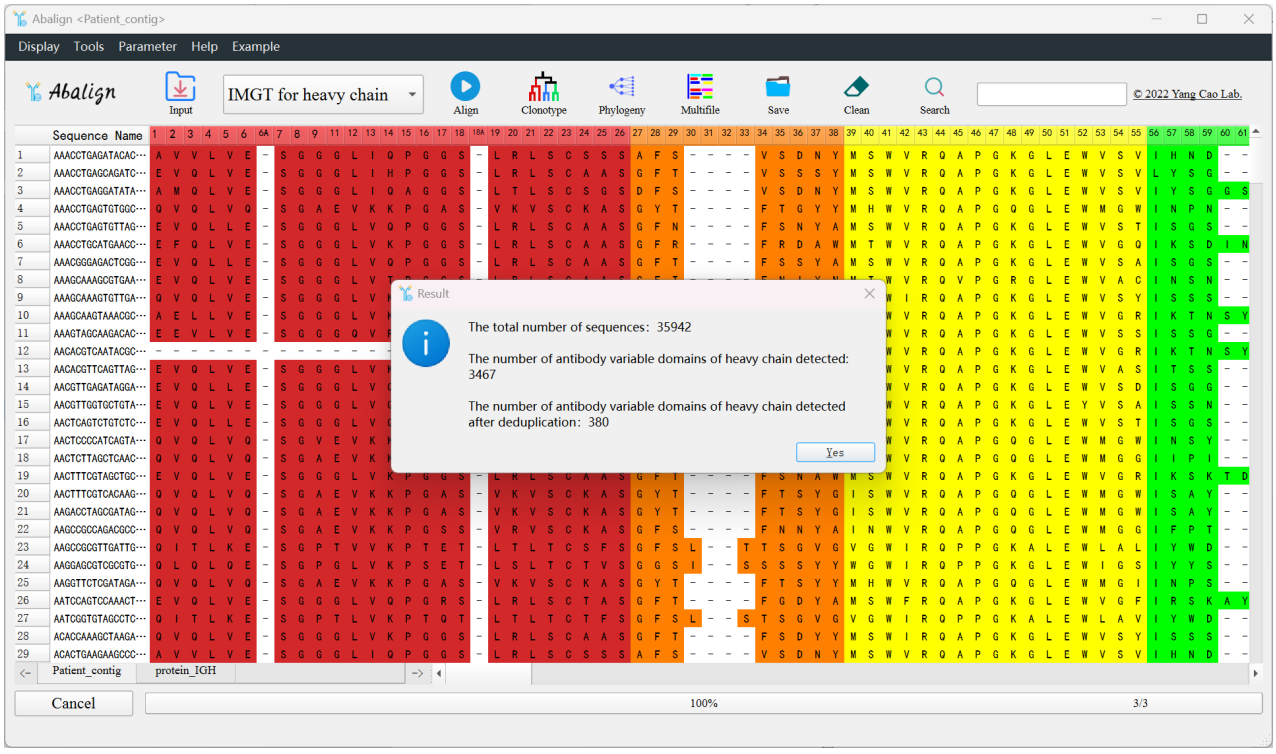


Figure 10. The result of a MSA using the Patient_contig sample in the example folder

Step 3. Analyze multiple sequence alignment: Click the **"Tools"** button in the top menu bar and move the mouse on **"Abundance"**. Click **"V Gene abundance"** or **"Sequence**

abundance" to obtain the abundance map of V Genes or sequences (Fig. 11).

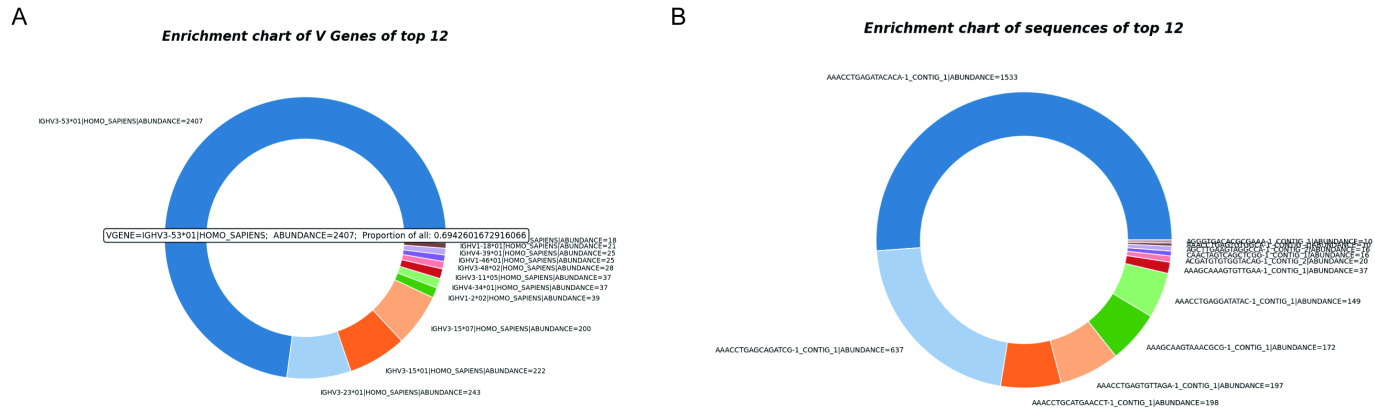


Figure 11. The top12 V genes and sequences abundance of Patient_contig sample

Step 4. Build phylogenetic tree: According to the sequence abundance information, build the tree with the V Gene of the high abundance sequences. In this case, the most abundant sequences belong to V Gene "IGVH3-53 * 01". Users can click **Display->Display by Genes ->Display by V genes** in the menu bar to select the V Gene "IGVH3-53*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 current Nwk file"** button in the tree building menu to save the .nwk file.

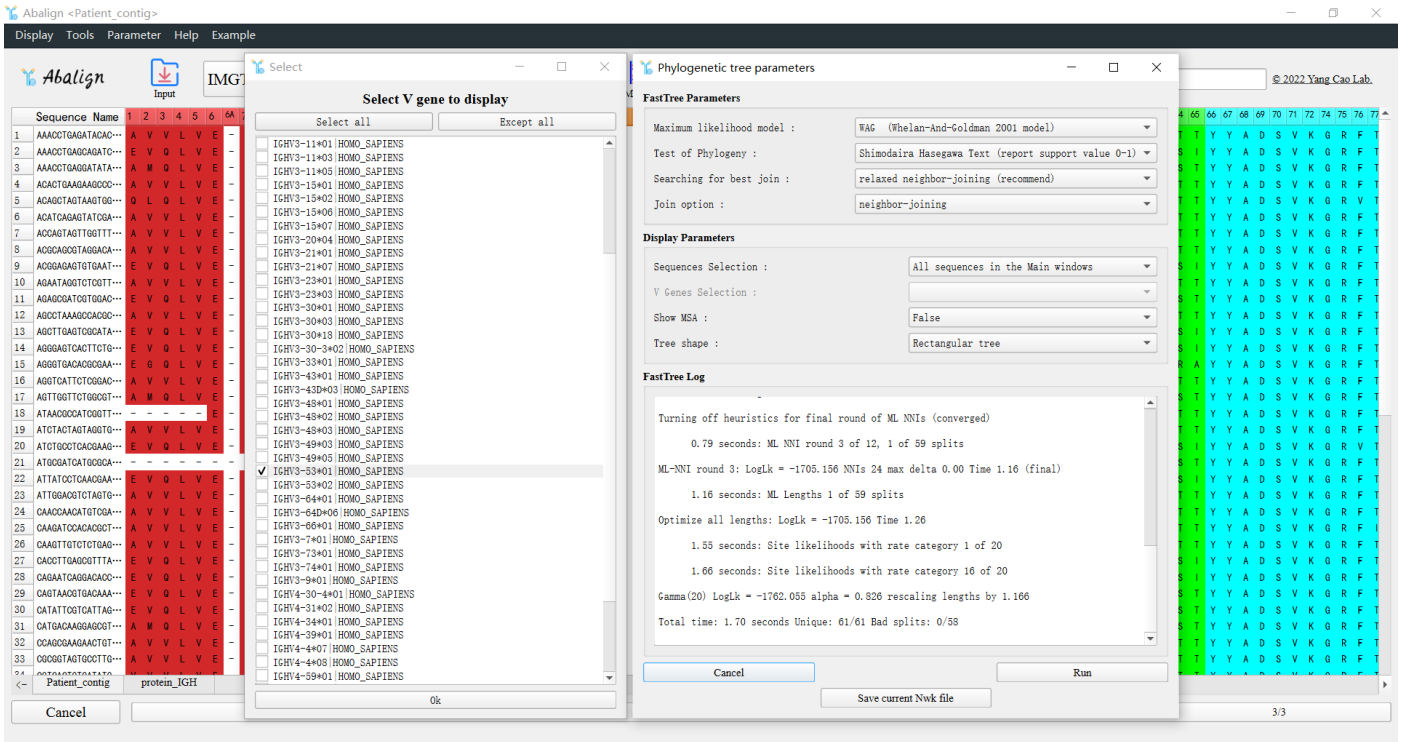


Figure 12. Use sequences belonging to IGHV3-53*01 to build a phylogenetic tree.

Step 5. View the phylogenetic tree: Once the evolution tree is built, a visual window will pop up, in which users can adjust the view and display other information about the tree through the button (Fig. 13). After selecting the node of the tree, users can also modify the attributes of the nodes, such as the color of the nodes, in the right window.

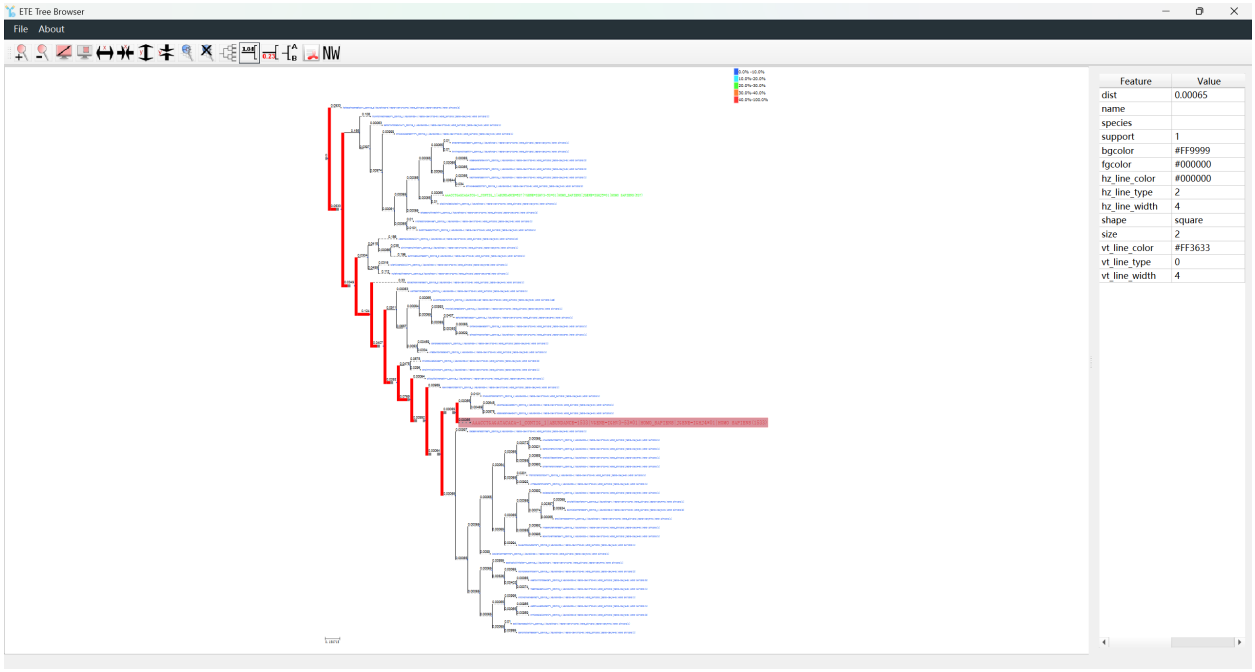


Figure 13. Results of phylogenetic tree construction using sequences belonging to IGHV3-53*01.

Step 6. Aid antibody humanization: If users need to know the frequency of residues used at each position compared with the known human BCRs or antibodies, they can click **Tools->Unusual Residue** to get the information (Fig. 14).

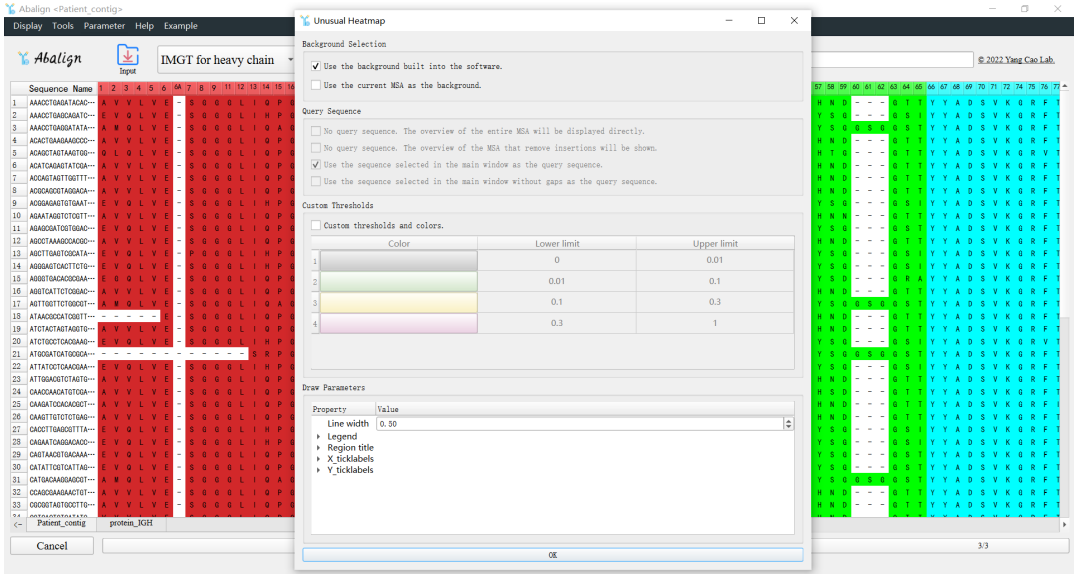


Figure 14. Select the first sequence in MSA to construct the unusual residue map.

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.

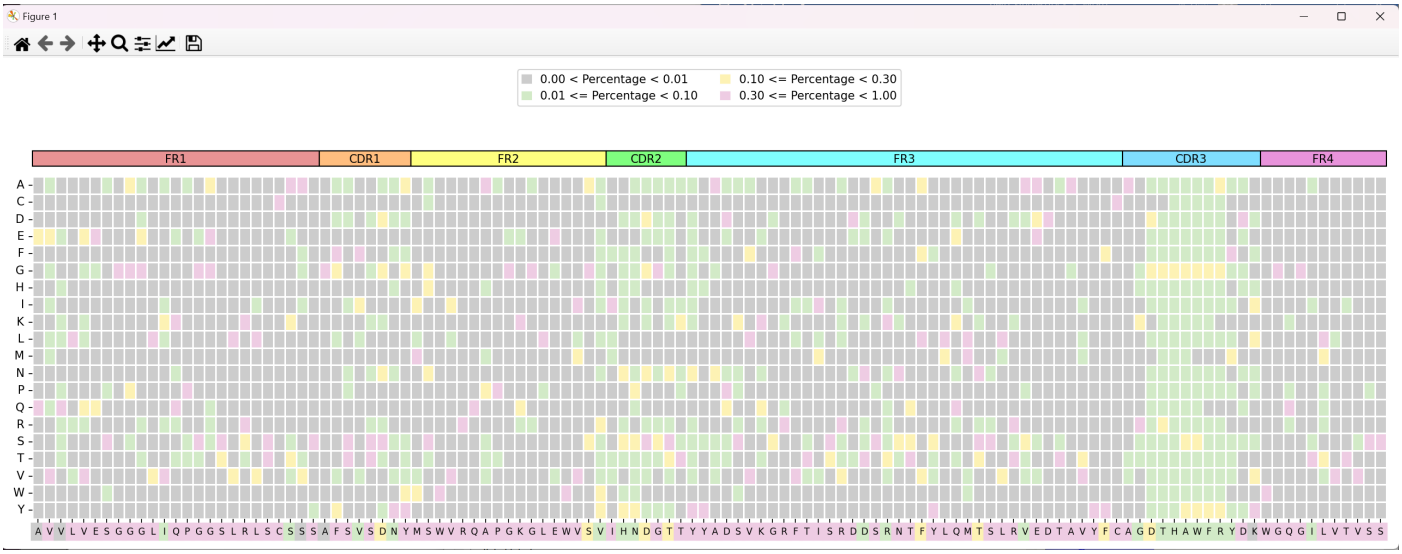


Figure 15. Result graph of unusual residues of the first sequence in MSA.

Step 7. Cluster sequences with Clonotypes: After 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 (Fig. 16).

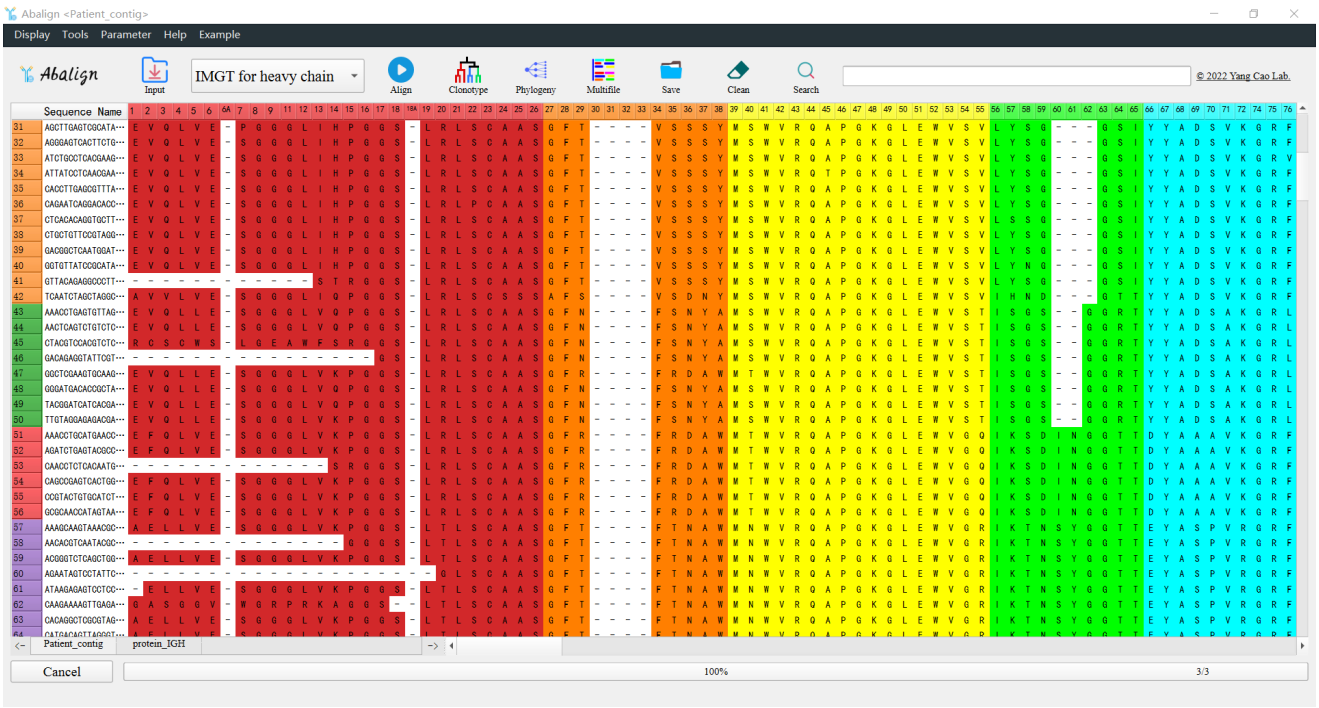


Figure 16. Rearrange sequences belonging to the same clonotype together and render them in the same color in the line label.

Step 8. 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.

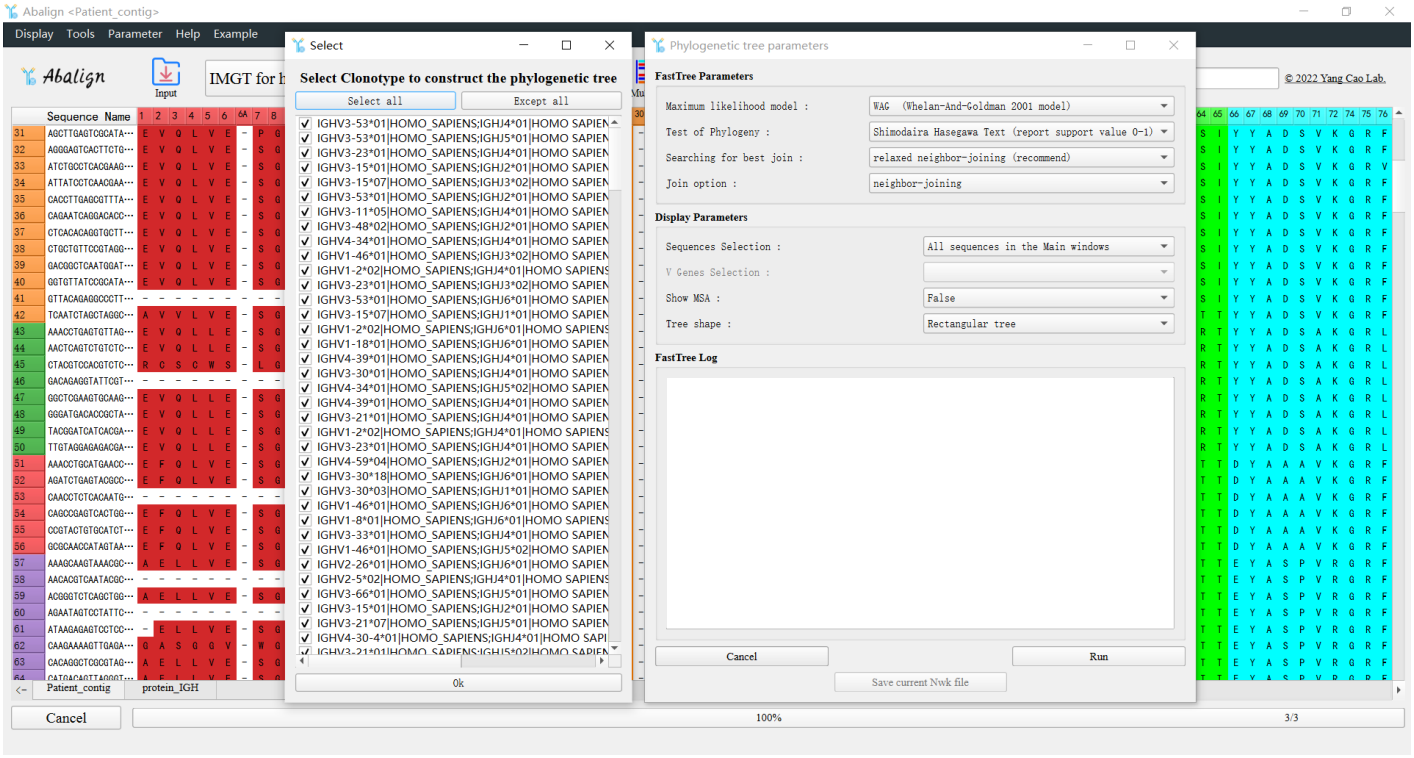


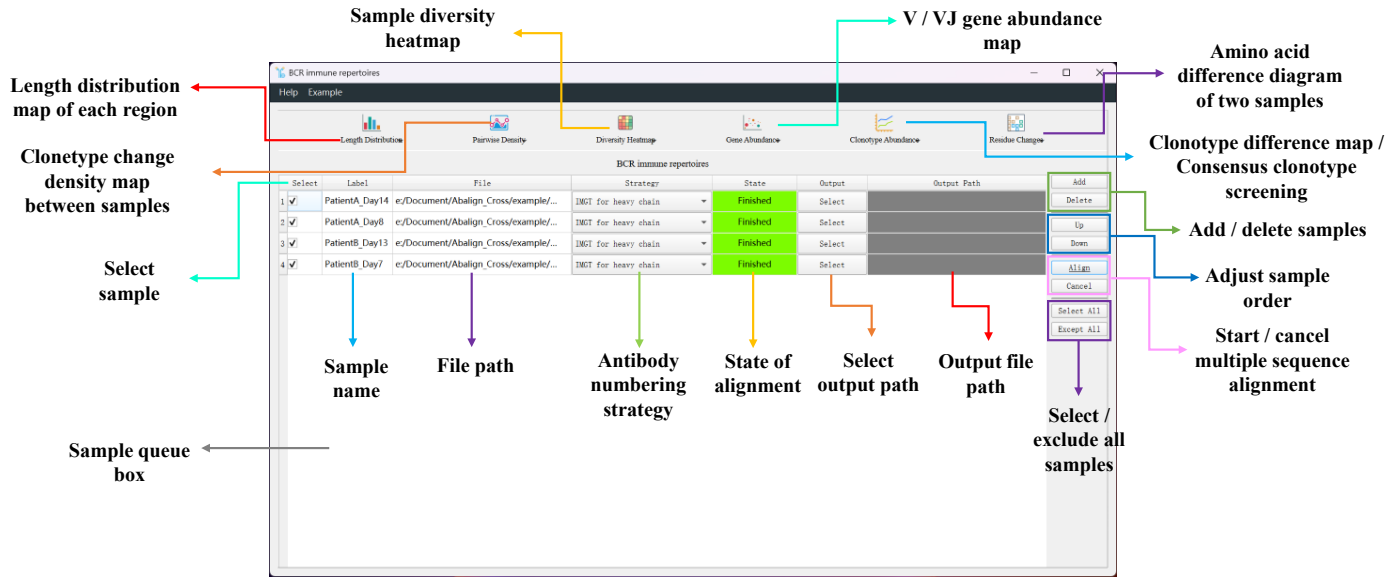
Figure 17. Selecting specific clonotypes to construct phylogenetic trees.

Step 9. Save File: The sequences displayed in the window can be saved by clicking the **"Save"** button in the toolbar. In addition to saving all sequences, we can save the sequences we need by combining **"Display by Genes"**, **"Display by Regions"** and **"Display by Clonotypes"**.

Step 10. Multifile: Click on **"Multifile"** button in the main window to use this feature, which supports users to align and perform cross-analysis on multiple files.

Multi-File Navigator

Navigator Layout



Explanation Of Terms

Clonotype: Sequences sharing the same VJ gene and the same CDR3 region are classified as the same clonotype.

Usage

Add File: Click **"Add"** button to select the target files, then the corresponding files will be displayed in **"BCR immune repertoires"** table. The filename can be customized by changing the value of the **"Label"** column of the table.

Delete File: Tick files in **"Select"** column of the table, then click **"Delete"** button to delete files.

Multiple Sequence Alignment: Tick files in **"Select"** column of the table, then click **"Align"** button. Customize the value of **"Strategy"** column of the table to change the numbering strategy. And the alignment parameters can be changed in the window that pops up after clicking **"Align"** button. **"State"** column of the table shows the progress of alignment, and **"Cancel"** button is used to delete alignment in the queue.

Save File: Click **"Select"** button 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 selected sample, 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 VJ 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 if the sequencing data quality is poor).

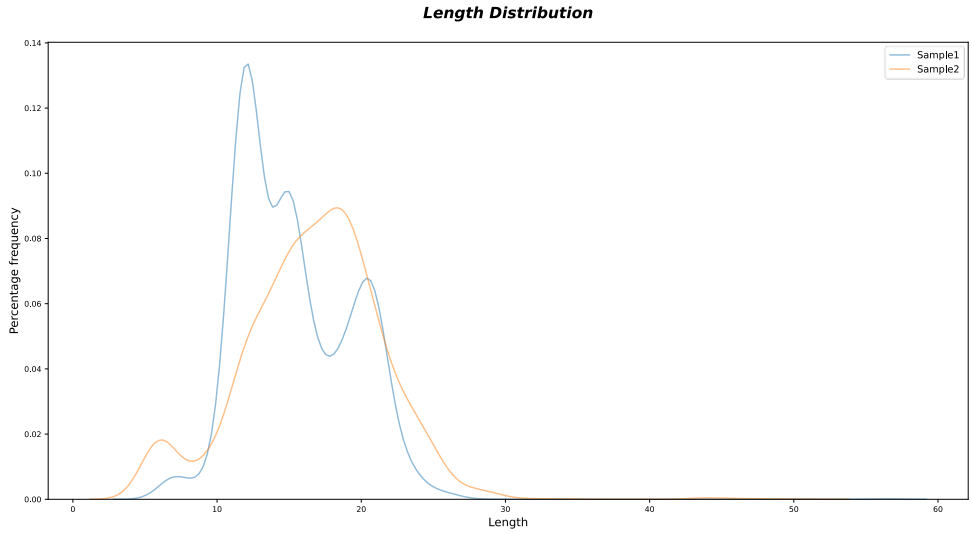


Figure 18. An example of length distribution in CDR3. The abscissa represents the length of the sequences, and the ordinate represents the percentage of the sequences with certain length. The lines with different colors represent different samples. The plot data can be saved by clicking **"Save Sources"**.

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

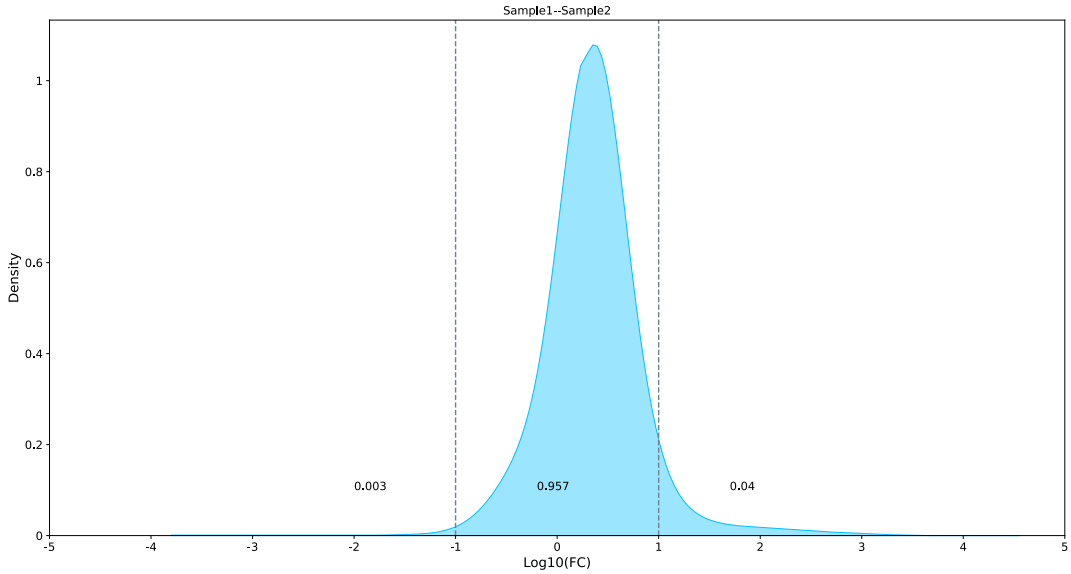


Figure 19. An example of clonotype differential density map between two samples. The abscissa represents the base 10 logarithm of the clonotype fold change, and the ordinate represents the density value. After removing the low abundance clonotypes, clonotypes are divided into 3 groups according to log10(FC). log10(FC)>1 is the expanded group, log10(FC)<-1 is the reduced group, and -1< log10(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).

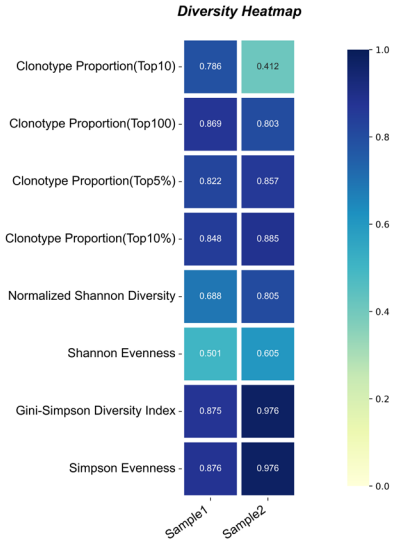


Figure 20. An example of diversity heatmap. The abscissa represents different samples, and the ordinate represents the diversity indexes. The darker the color, the higher the value, and the lighter the color, the lower the value. The plot data can be saved by clicking **"Save Sources"**. After saving, 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 of selected samples, and draw the histogram (Fig. 21) / scatter plot (Fig. 22).

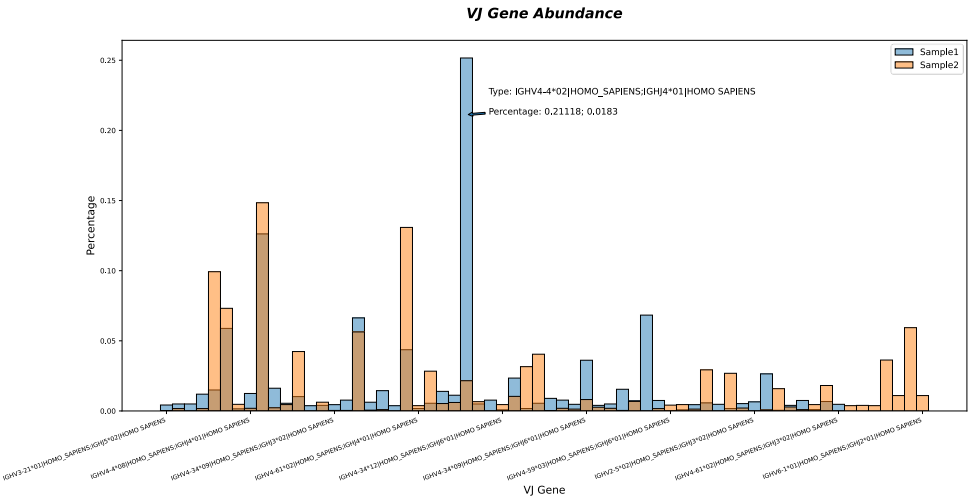


Figure 21. An example of VJ gene abundance histogram. The abscissa represents the type of V/VJ gene, and the ordinate represents the proportion of the specific gene. The different colored bars represent different samples. Details are displayed when hovering.

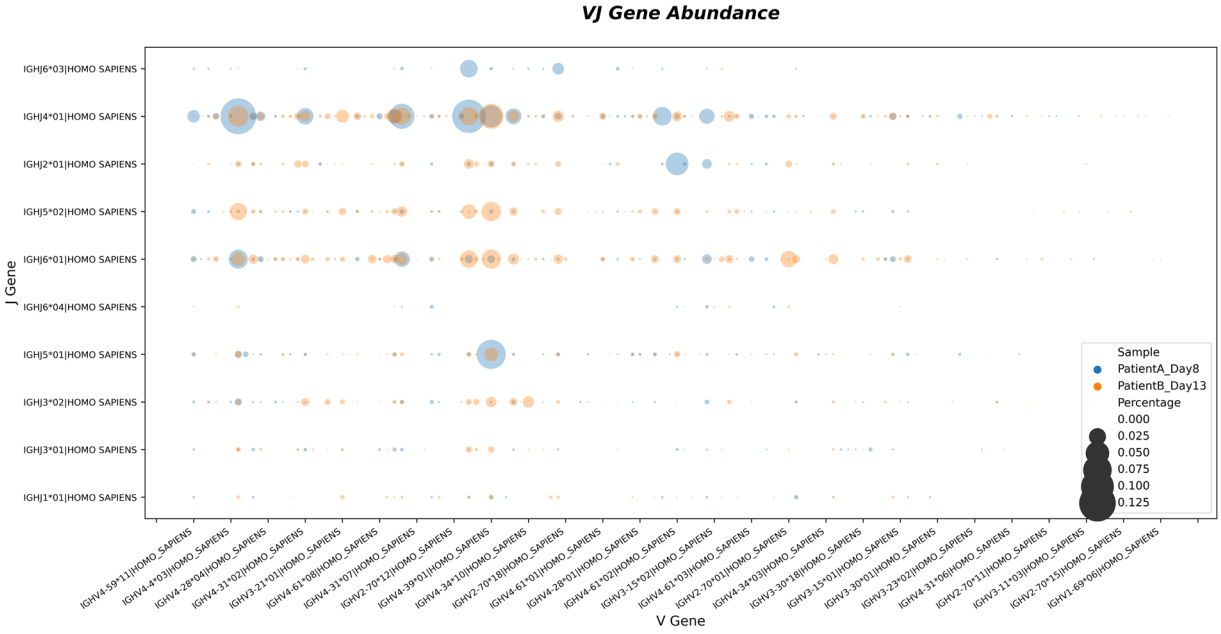


Figure 22. An example of VJ gene abundance scatter plot. The abscissa represents the type of V gene, and the ordinate represents the type of J gene. Points with different colors represent different samples, and the size of points represents the proportion of the specific VJ gene combination.

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).

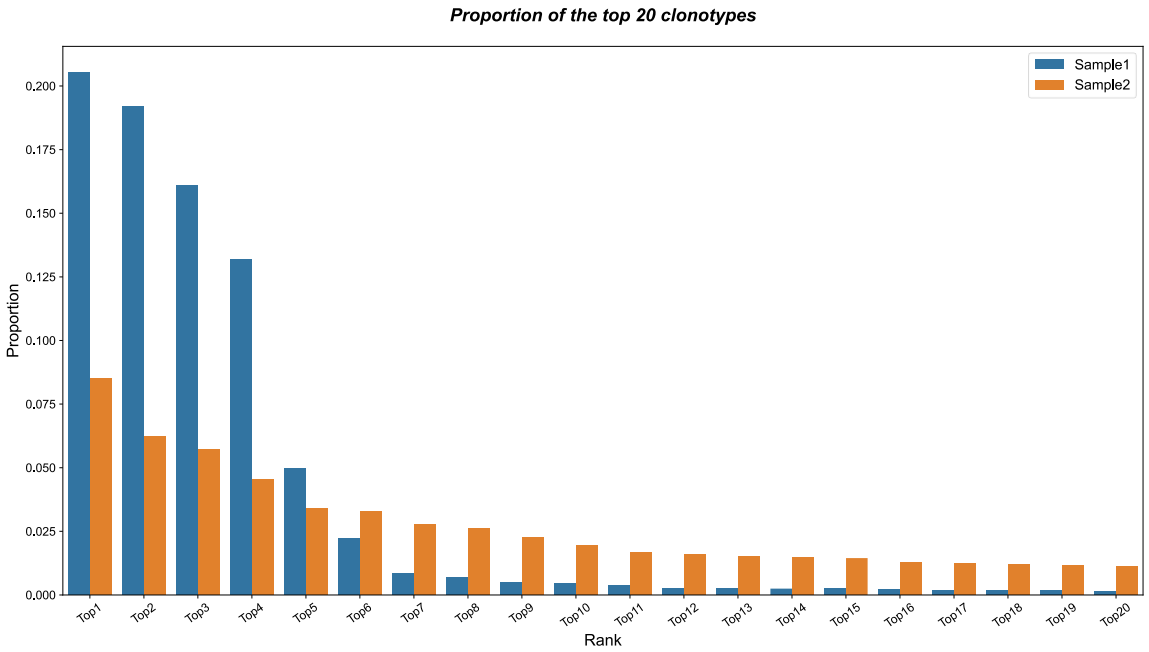


Figure 23. An example of clonotype distribution. The abscissa represents the top 20 clonotypes, and the ordinate represents the proportion of clonotypes, and bars with different colors represent different samples.

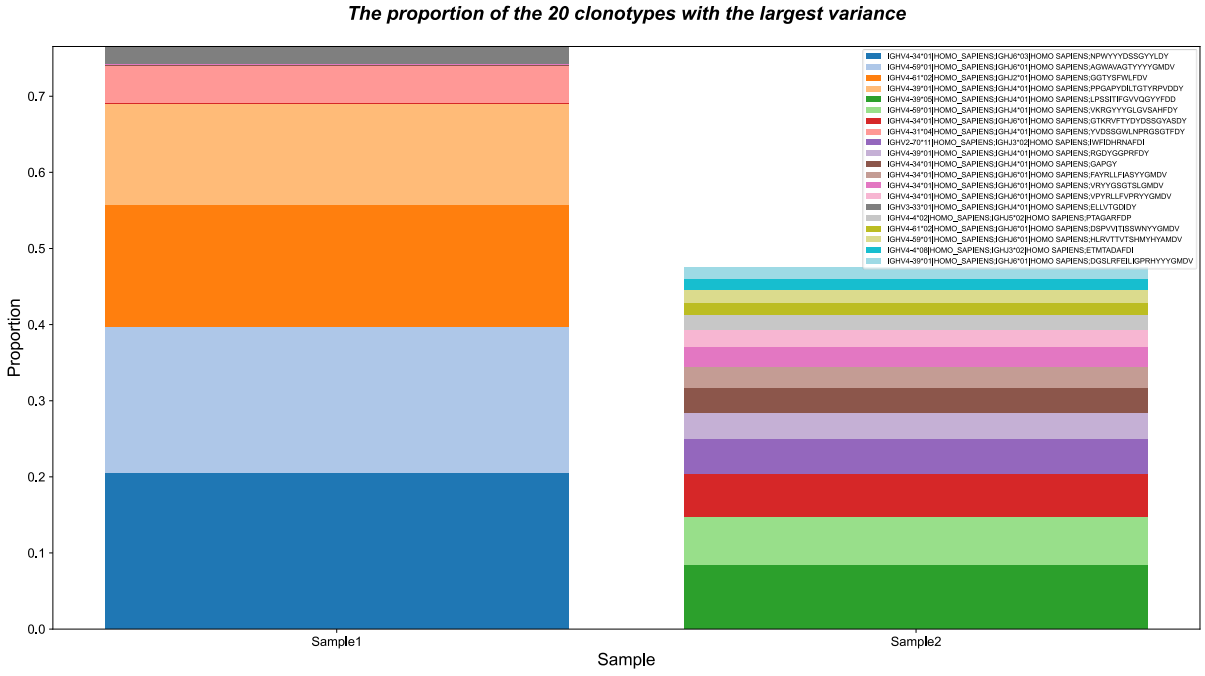


Figure 24. An example of clonotype changes. The abscissa represents different samples, and the ordinate represents the proportion of the top 20 clonotypes that vary across selected samples. The plot data can be saved by clicking **"Save Sources"**. After saving, you will get a file named **"clonotype_changes.csv"**, which records the abundance and changes of all clonotypes of selected samples.

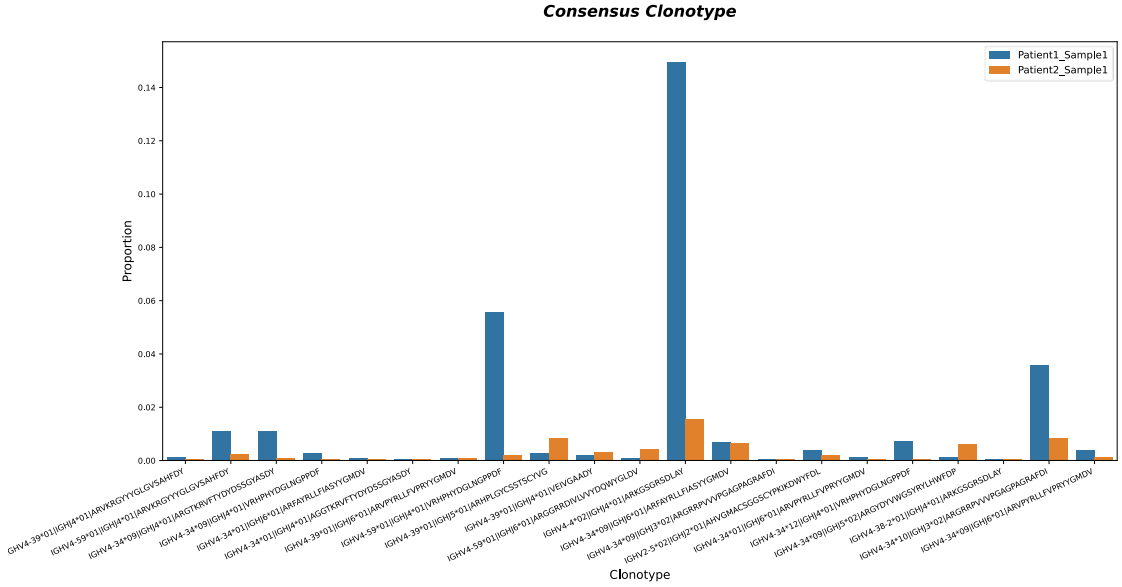


Figure 25. An example of consensus clonotypes between samples. The abscissa represents the specific clonotype, and the ordinate represents the proportion of clonotypes. The bars with different colors represent different samples. The plot data can be saved by clicking **"Save Sources"**. After saving, you will get a file named **"Consensus_clonotype.csv"**, which records the top N clonotypes shared in all samples.

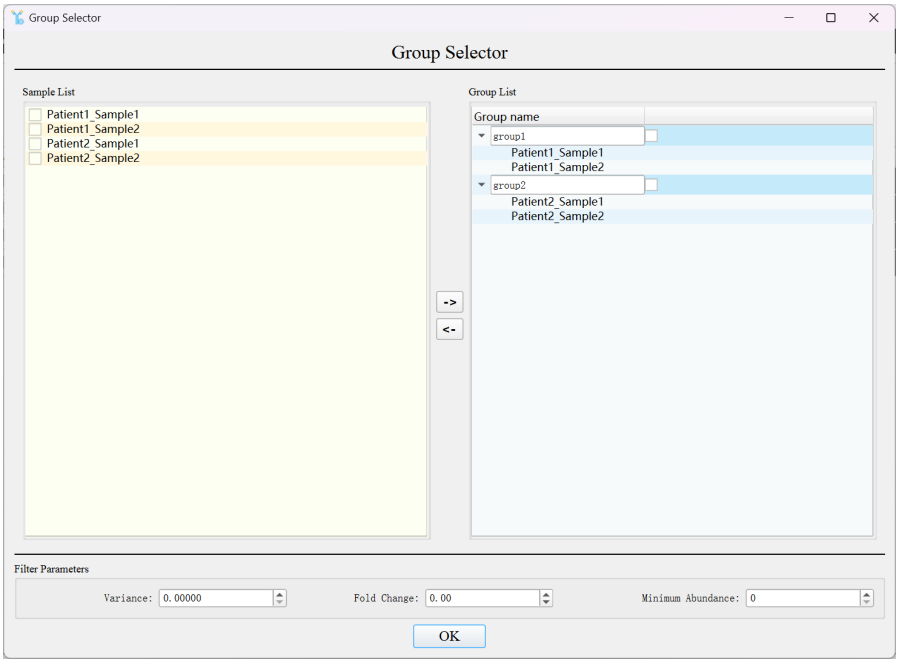


Figure 26. Group Select Parameters Dialog. **"Sample List"** shows all samples to be analyzed, which selected through the multi-file navigator. **"Group List"** is used to display the groups constructed by users, which can be generated by selecting the samples on the left and clicking the move button in the middle. Conversely, ticking the right groups and clicking the move button in the middle will remove the selected groups. **"Filter Parameters"** are used to control the conditions for expanded clonotypes.

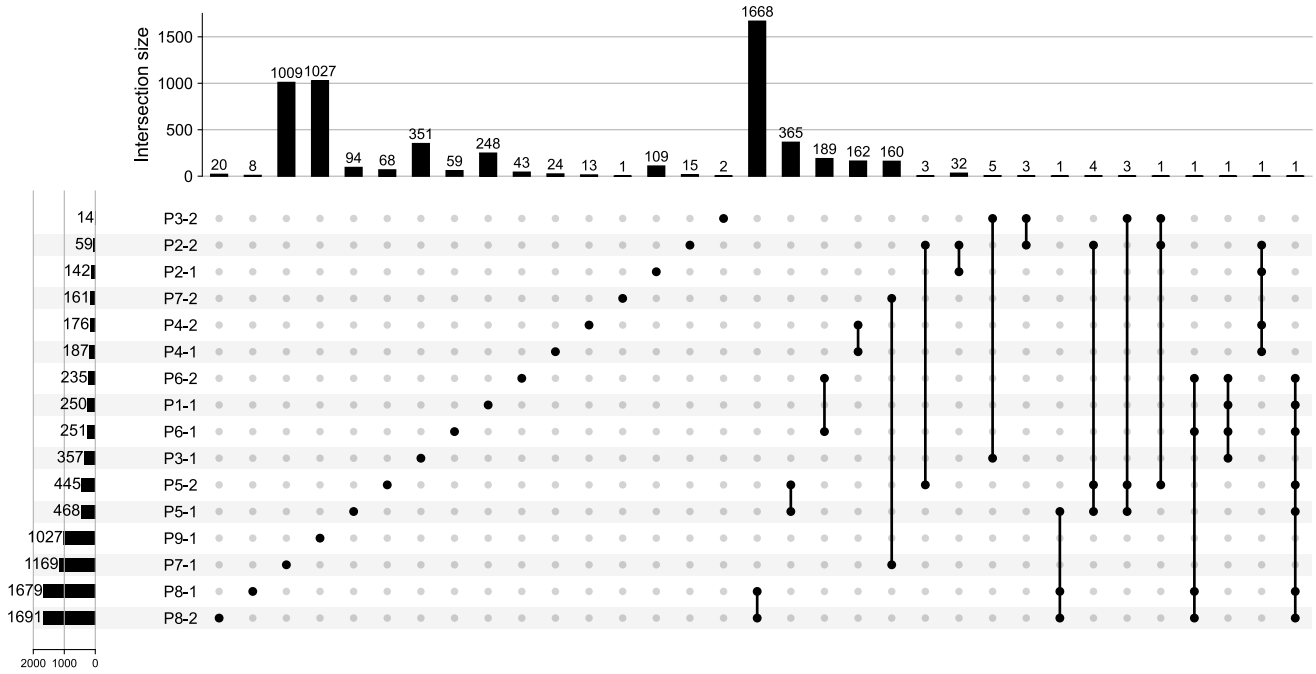


Figure 27. An example of consensus clonotypes between groups. The left bar represents the number of clonotypes expanded in each group (for example, one patient represents one group), and the upper bar represents the number of shared clonotypes among different groups. The set information corresponding 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).

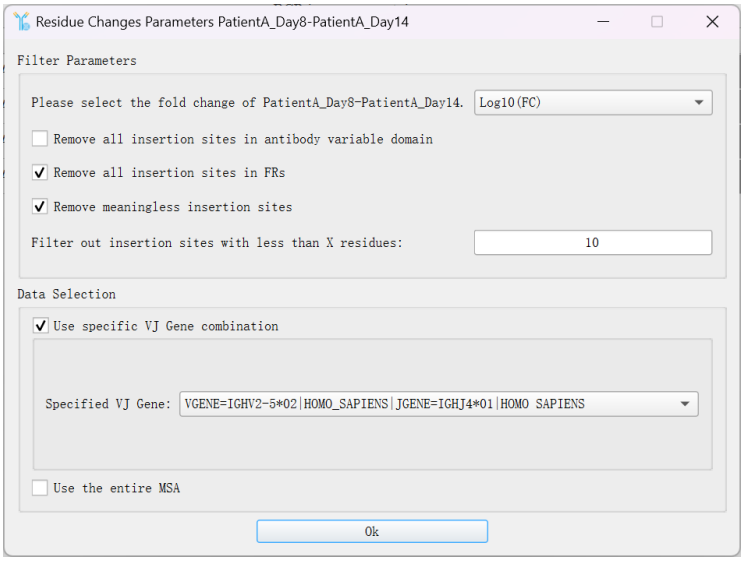


Figure 28. Residue Preference Parameters Dialog. **"Filter Parameters"** are used to control which amino acid difference positions to be displayed. Users can choose the fold change, whether to retain insertion positions or nonsense positions (the positions with residue occurrences lower than X). **"Data Selection"** is used to control the sequences for analysis, allowing the users to select the entire MSA or a specific VJ gene combination.

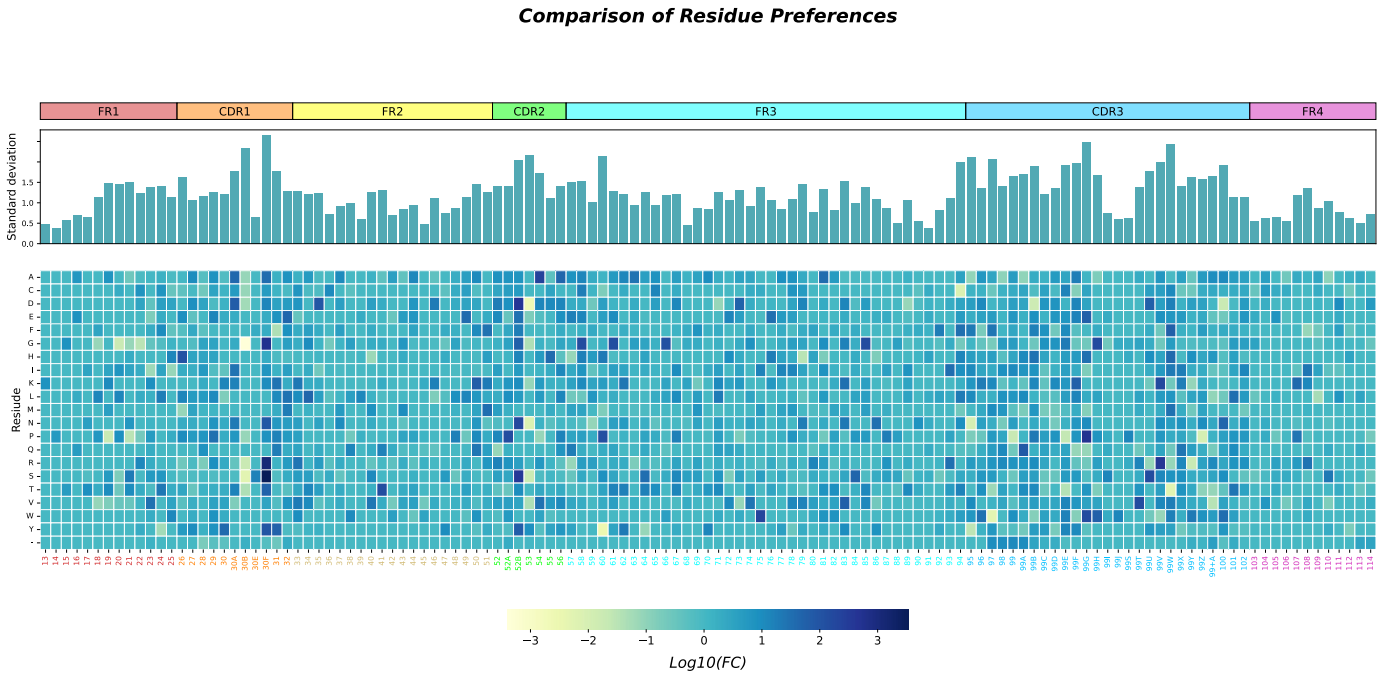


Figure 29. An example of residue preference map. The ribbon at the top uses different colors to distinguish FRs and CDRs. The abscissa at the bottom represents each position in the variable domain. The ordinate of the upper figure represents the standard deviation, and the ordinate of the lower figure represents 20 amino acids and gaps. The histogram above represents the standard deviation of residues differences at each

position, and the larger the value of a certain position, the more significant its residue preference. 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.

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.