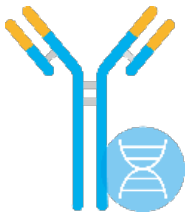
Abalign: a comprehensive multiple sequence alignment platform for B-cell receptor immune repertoires



User's Manual

Sichuan University

Mar. 2023

MANUAL

Introduction

Multiple sequence alignment (MSA) 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 domain 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 and insertion positions by immunology studies, which enable the result to be consistent with the structural and immunological knowledge.

We compared Abalign with three state-of-the-art MSA tools: Clustal Omega[2], MAFFT[3], MUSLCE[4]. Comprehensive benchmark tests showed that Abalign outperformed the existing MSA tools in accuracy, speed and memory consumption significantly. Users can see detailed test information on the homepage of Abalign (http://cao.labshare.cn/abalign/).

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 B-cell lineage tree, assigning VJ gene, 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 public clonotypes, or residue preferences, etc. Abalign will profit immunoinformatic and pharmaceutical communities by analyzing massive B cell receptor immune repertoire data or antibodies sequences and making new discoveries.

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.

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: Once you have extracted the files, navigate to the extracted folder and double-click "**Abalign _setup.exe**" to launch the installation wizard. Follow the prompts to complete

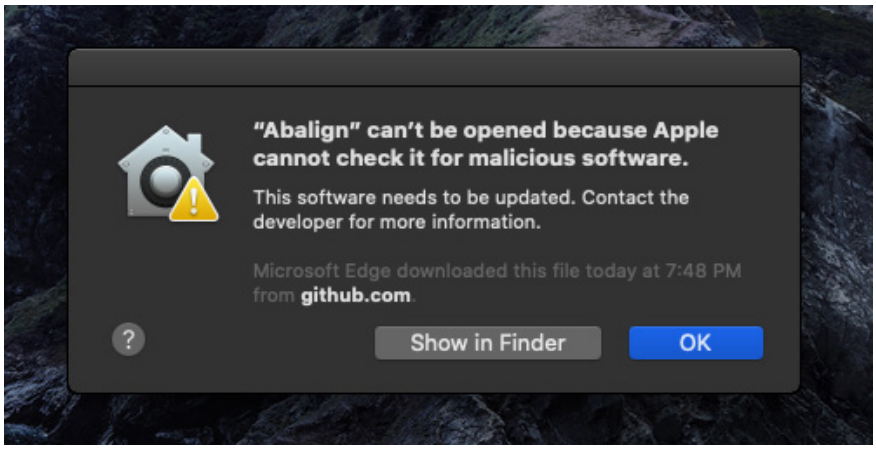
the installation process.

Linux: After extracting the files, navigate to the extracted folder and locate "**Abalign_installer.run**". Open a terminal and execute the following command to launch the installation guide:

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

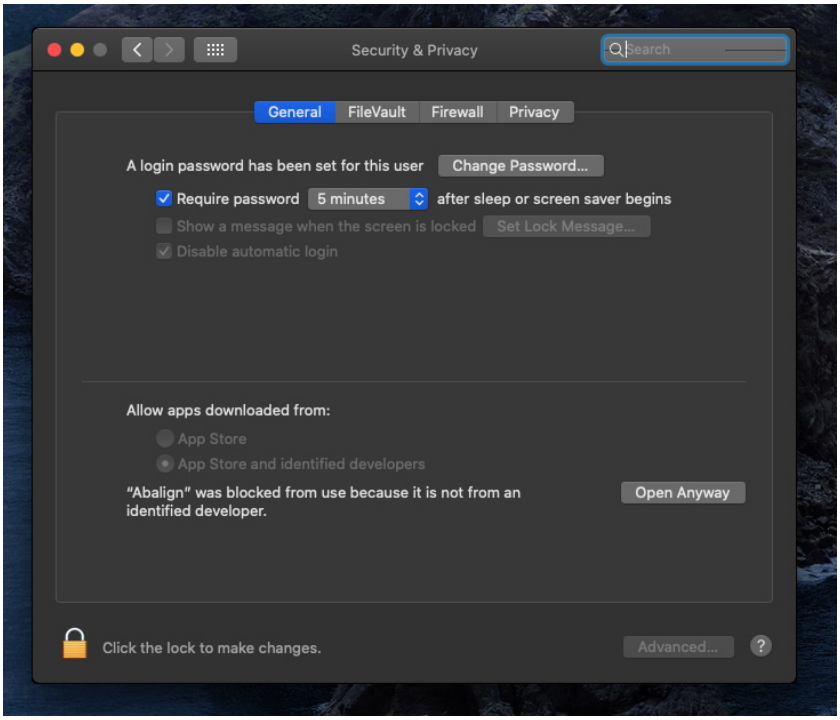
macOS: After extracting the files, navigate to the extracted folder and double-click "**Abalign.app**" to launch Abalign directly. Please move Abalign.app into the Applications folder. Please note that Abalign is currently undergoing compatibility testing for macOS. If you encounter any issues, please contact us at cy_scu@yeah.net.

Tips 1: The first time you run Abalign on macOS, you may see a dialog box. This is normal - simply click "OK" to proceed.

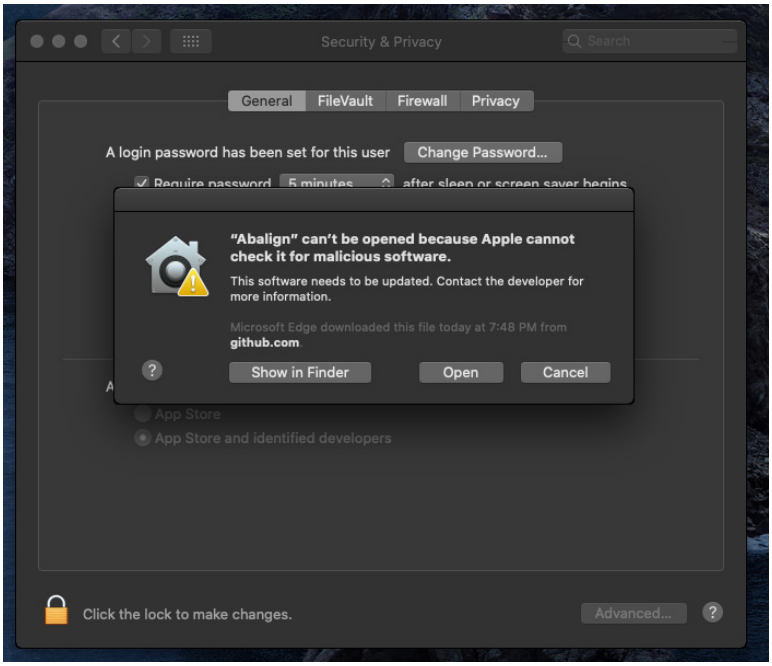


Then please click "System Preferences"->"Security & Privacy", and click "Open Anyway"

in this window.



Once you have completed the above steps, a new dialog box will appear. Click "Open" to launch Abalign.



Tips 2: If you see a dialog box that reads "Abalign is damaged and can't be opened, You should move it to the Trash", execute the following command in Terminal to resolve the issue:

1. sudo xattr -r -d com.apple.quarantine /Applications/Abalign.app

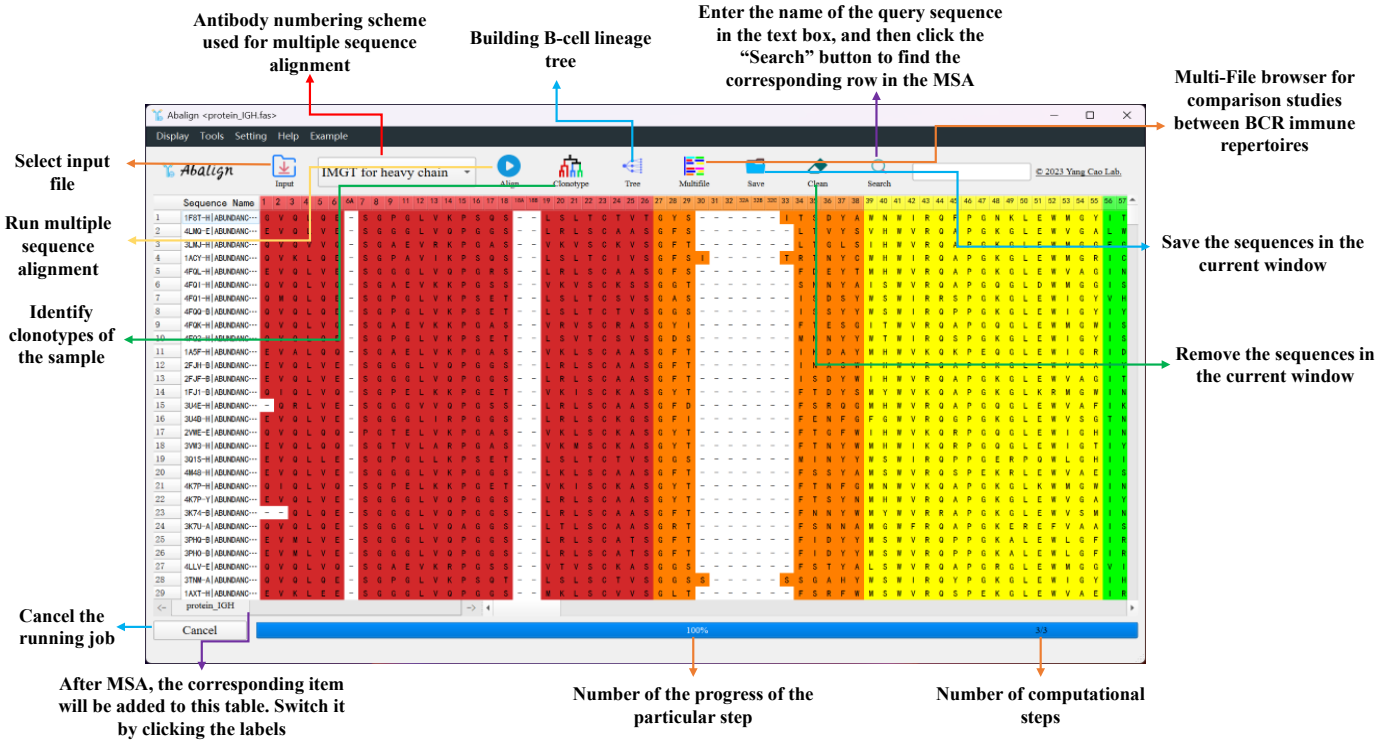
Uninstallation

Windows: To uninstall Abalign, navigate to the installation path and double-click "**unins000.exe**". Click "**Yes**" to confirm.

Linux: To uninstall Abalign, execute "**Abalign_uninstall.sh**" in the terminal if you have added Abalign to the environment variable. Alternatively, delete the Abalign installation folder to uninstall Abalign directly.

macOS: To uninstall Abalign, delete "**Abalign.app**" directly. Additionally, Abalign's configuration information is stored in "**~/Library/Preferences/Abalign_conf.txt**". To completely uninstall Abalign, delete this file as well.

Software Layout



Tool Buttons

Input: Choose the input file in FASTA format.

Scheme: There are eight options available for selecting the antibody numbering schemes and chain types between **"Input"** and **"Align"**. The default selection is **"IMGT for heavy chain"**.

Align: This feature enables users to perform multiple sequence alignment. The software also matches the VJ genes of the input sequences by aligning them with the germline gene database of the selected species. The alignment results are displayed in the main window and rendered

with different colors for FR1, CDR1, CDR2, CDR3, and FR4 regions. Users can customize the rendering method by clicking on the "**Display**" in the "**Sequence Render Mode**" menu.

Clonotype: This feature enables the identification of clonotypes for the input sequences. A clonotype is defined as a group of sequences with high CDR3 homology, identical CDR3 length, and V/J gene usage[5]. To use this feature, users must first perform the "Align" function. The identity threshold of CDR3 can be adjusted in the "**Setting**"->"**Align Parameters**" dialog. The default threshold is 100%.

Tree: This feature uses the FastTree[6], which employs the maximum likelihood method, to build a .nwk file and visualize the resulting B-cell lineage tree using Ete3[7]. Clicking the button to perform this function will bring up a dialog box (Fig. 1), where users can adjust the parameters for drawing the tree (Fig. 2).

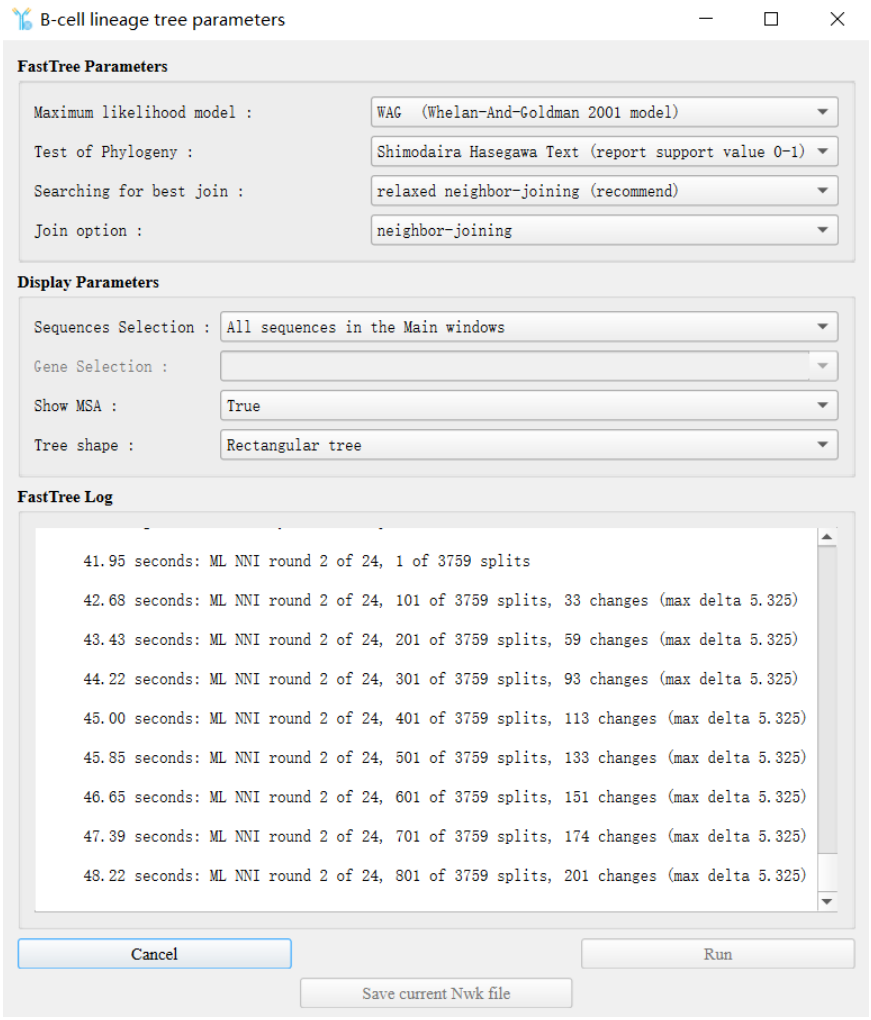


Figure 1. B-cell lineage tree Parameters Dialog.

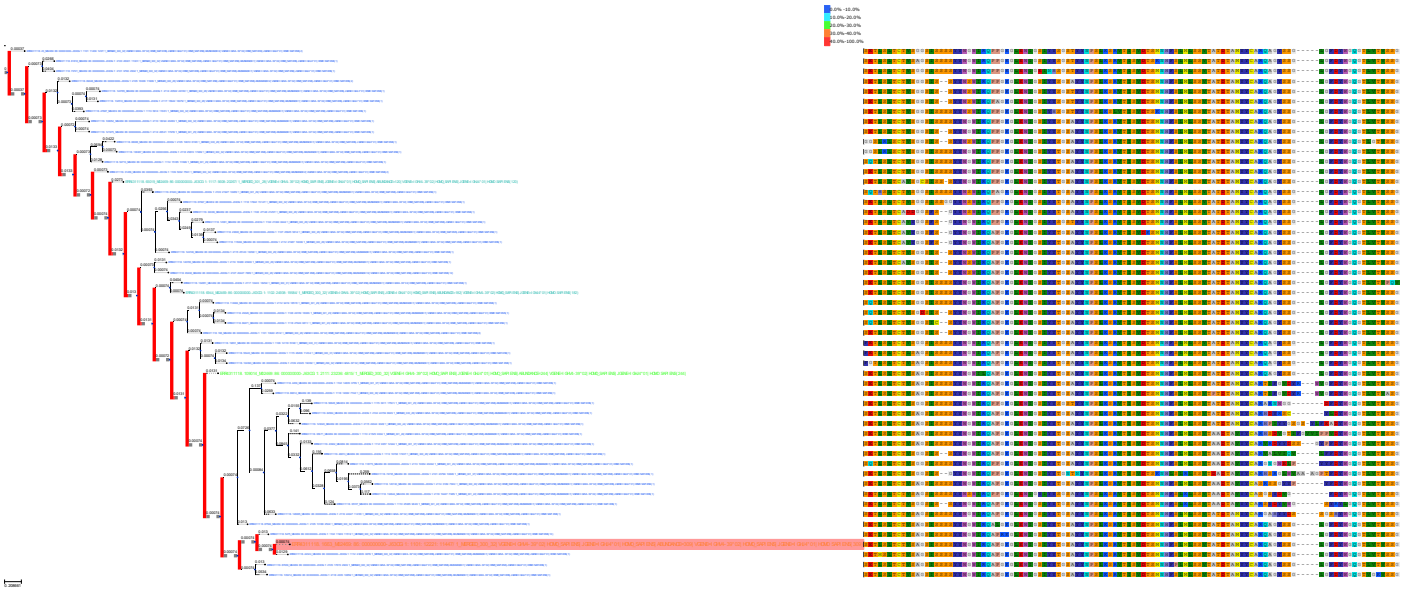


Figure 2. An example of a B-cell lineage tree with corresponding MSA. The left panel displays the B-cell lineage tree of the selected sequences, with highlighted nodes representing high-abundance. The right panel shows the corresponding multiple sequence alignment, with different residues represented by distinct colors.

Multifile: Clicking this button opens a window that allows users to align and analyze multiple BCR repertoires data. Users can load multiple FASTA files for analysis in this window.

Save: Saves the sequences currently displayed in the text box.

Clean: Clears all sequences in the current text box and deletes the sample.

Search: To locate a specific sequence, enter its name and click the search button. The display interface will navigate to the location of the query sequence.

Menu Bar

Display

Remove Duplication: When selecting this option, antibody sequences with identical variable domains are displayed only once, but the duplicates are still used in the abundance analysis. This option is enabled by default and can be adjusted in the "**Align Parameters**" dialog.

Show Statistics: If you select this option, a dialog box will appear each time the sample alignment is completed or when you switch between items, displaying information about the sample. This information will include the number of sequences in the sample, the number of antibody variable domains, and the number of antibody variable domains after deduplication.

Filter Level: This option allows users to filter sequences based on the length of their variable domains. Four levels of length filtering are available: "**Off**", "**Soft**" (default), "**Normal**", and "**Strict**". When "**Off**" is selected, there is no length limit for each region. When "**Soft**" is selected, there must be at least one amino acid in each region. "**Normal**" and "**Strict**" limit the region length based on antibody data with known structures. This option can be adjusted in the "**Align Parameters**" dialog.

Sequence Color Mode: This menu offers two options for adjusting the colors of residue rendering. "**Light mode**" renders the residues in light colors, while "**Soft mode**" renders the residues in darker colors.

Sequence Render Mode: There are two options in this menu for adjusting the sequence rendering mode. **"Color by regions"** divides the antibody sequences into different FRs and CDRs and renders them in different colors. **"Color by amino"** renders different residues in different colors based on their types.

Display Full MSA: This option will display the complete MSA of the input sequences.

Display by Clonotypes: This option will display the complete multiple sequence alignment of the input sequences. However, before using this option, the **"Clonotype"** analysis must be performed first.

Display by Genes: This menu provides three options: "**Display by V genes**", "**Display by J genes**" and "**Display by VJ genes**". Selecting any of these options will cause the software to display the sequences that match the specified condition based on their V and/or J gene usage.

Display by Regions: This menu provides seven options that correspond to the seven regions of the antibody variable domain. Users can select one or more options to display the desired regions.

Tools

V Gene Alignment: This feature provides two options. **"V Gene alignment"** shows the alignment of each sequence with the top 5 scoring V genes (Fig. 3), while **"V Gene Detection in MSA"** identifies the V genes during the multiple sequence alignment process (which is the default setting).

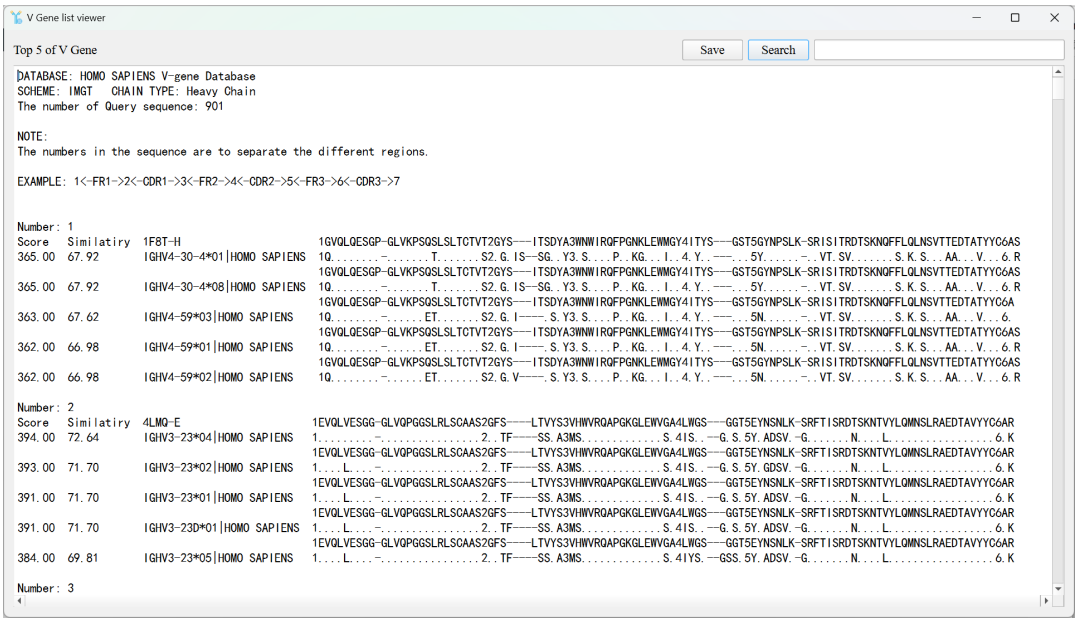


Figure 3. An example of V Gene list viewer.

V Gene Usage: This feature allows users to count the usage frequency of V genes in selected samples and generate a chart displaying the top 20(default) most commonly used V genes (Fig. 4)

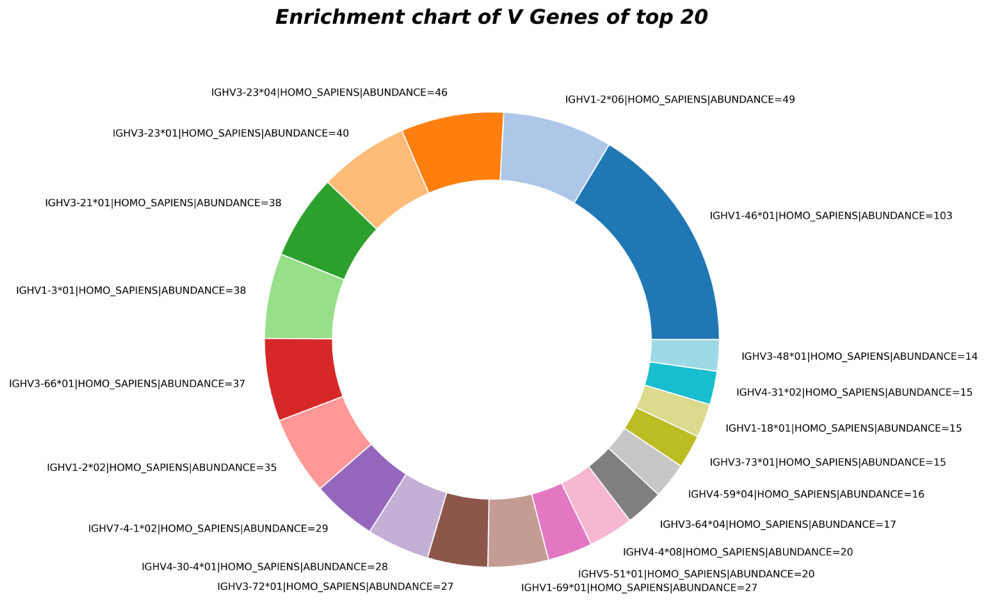


Figure 4. An example of Enrichment chart of V Genes of top 20. The abundance information is represented by a pie chart in which different colors indicate different V genes. The abundance data is displayed alongside the corresponding gene name.

J Gene Usage: This feature allows users to count the usage frequency of J genes in selected samples and generate a chart displaying enriched J genes. (Fig. 5).

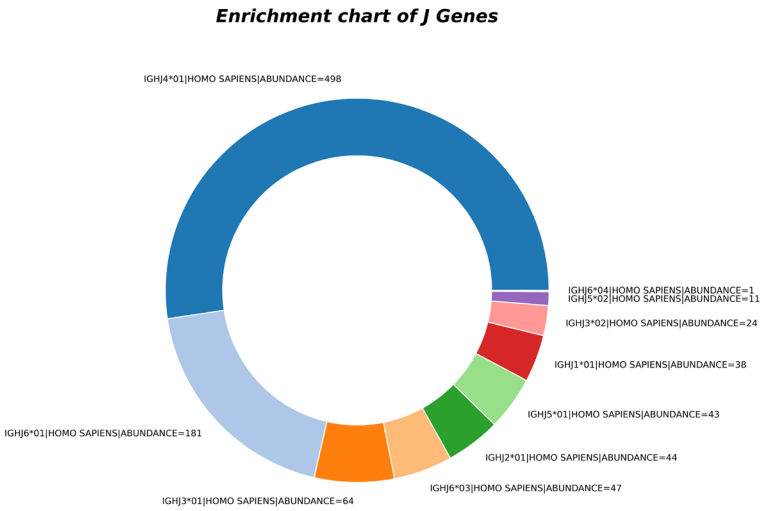


Figure 5. An example of Enrichment chart of J Genes. The abundance information is represented by a pie chart in which different colors indicate different J genes. The abundance data is displayed alongside the corresponding gene name.

VJ Gene Usage: This feature allows users to count the usage frequency of VJ genes in selected samples and generate a chart displaying the top 20(default) enriched VJ genes. (Fig. 6)

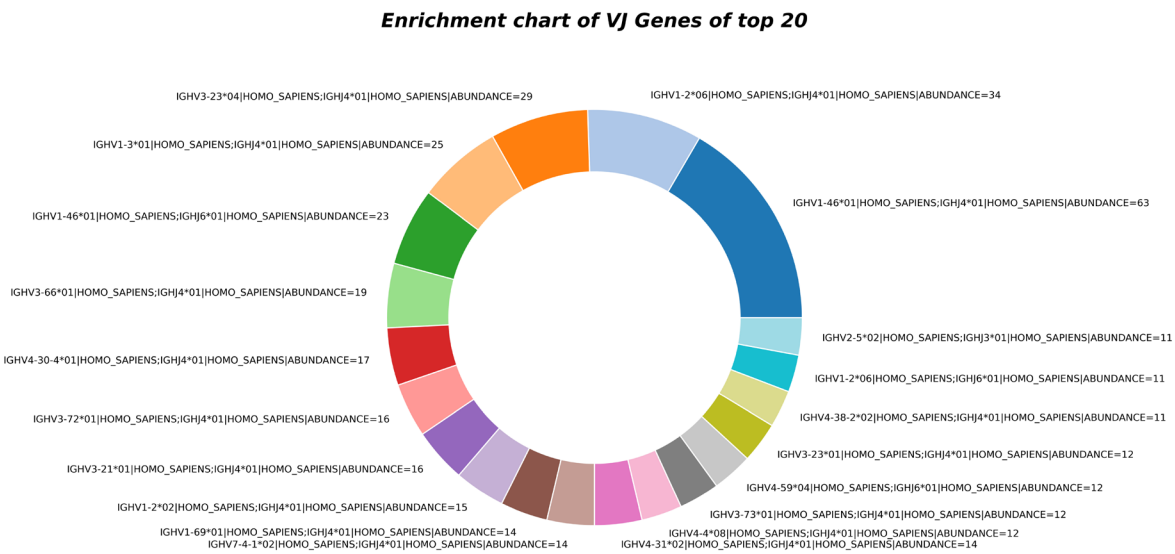


Figure 6. An example of Enrichment chart of VJ Genes of top 20. The abundance information is represented by a pie chart in which different colors indicate different VJ genes. The abundance data is displayed alongside the corresponding gene name.

Sequence Enrichment: This feature shows the top 20 most abundant sequences (Fig. 7).

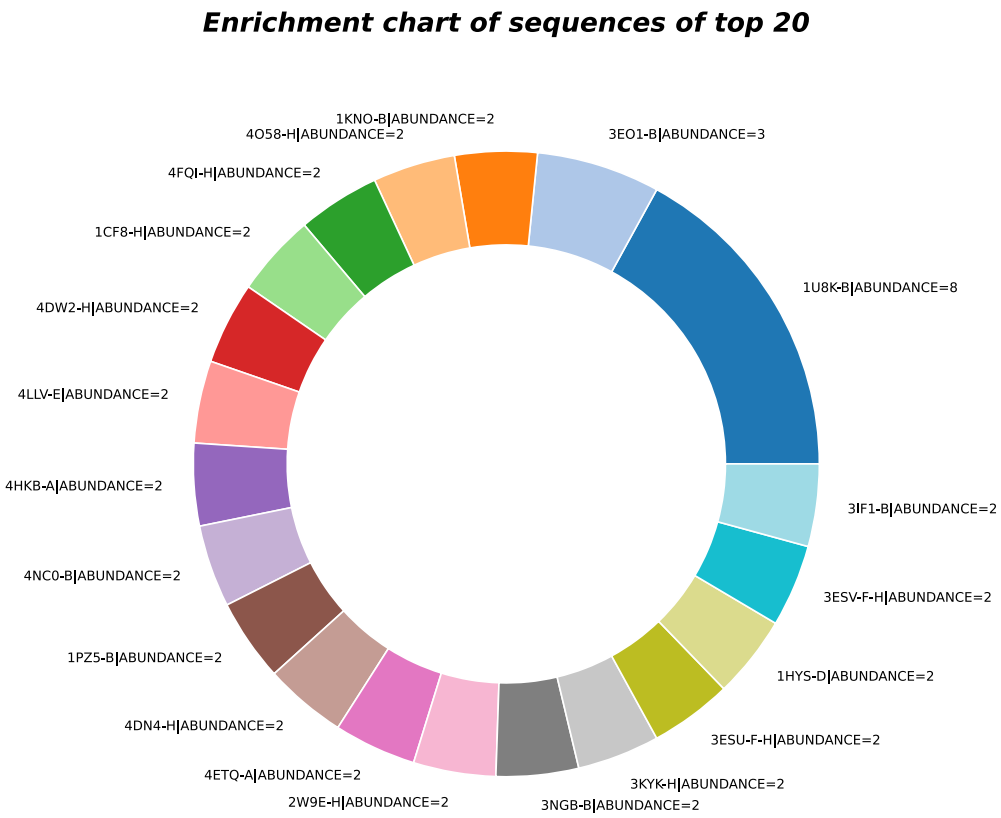


Figure 7. An example of Enrichment chart of sequences of top 20. The abundance information is displayed in a pie chart with different colors indicate different sequences. The abundance data is displayed alongside the corresponding gene name.

Region Enrichment: This feature shows the abundance of the top 20 most abundant sequences for FR1, CDR1, FR2, CDR2, FR3, CDR3, and FR4(Fig. 8).

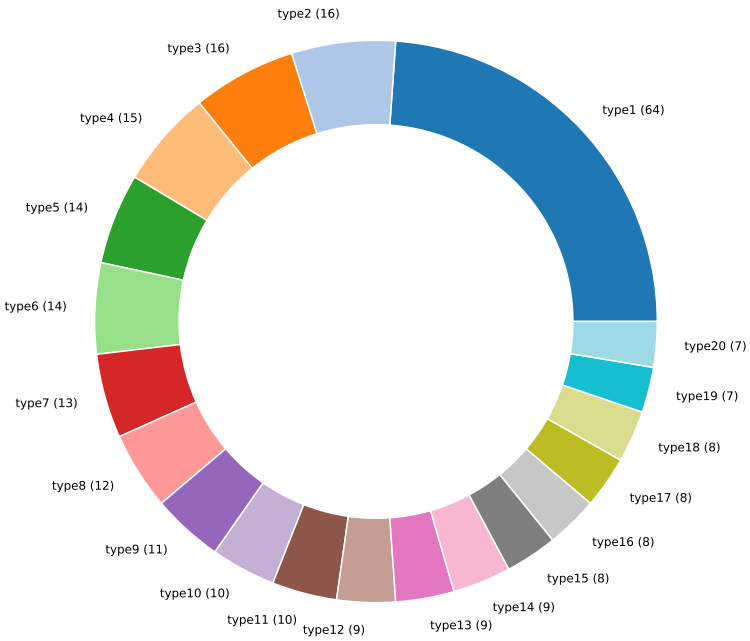


Figure 8. An example of Enrichment chart of FR1 of top 20. The abundance information is displayed in a pie chart with different colors indicate different regions. The abundance data is displayed alongside the corresponding gene name.

Length Distribution: This feature allows users to analyze the length distribution of each region of the antibody for selected samples and draw a kernel density estimation curve (Fig. 18). The regions analyzed include FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4, Variable domain and CDR1-FR4 (note that FR1 may be incomplete if the sequencing data quality is poor) (Fig. 9).

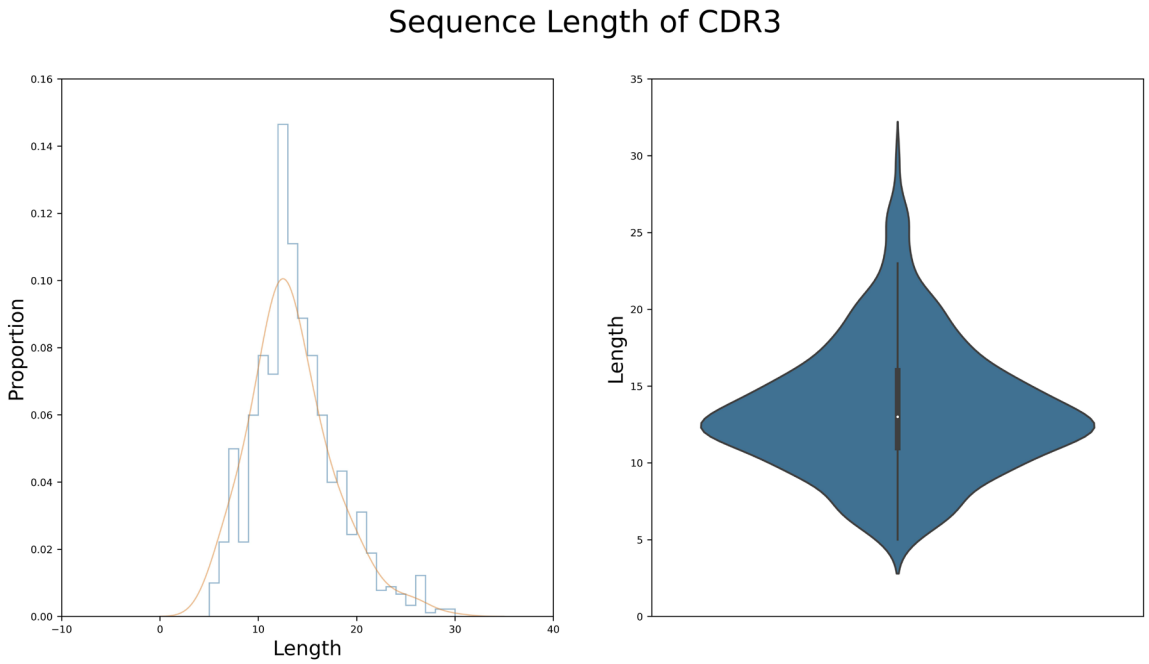


Figure 9. An example of Length Distribution of CDR3. The left figure is a length histogram. The x-axis represents the length, and the y-axis represents the proportion. The right figure is a violin plot. The y-axis represents the length, and the wider the width, the greater the number of sequences with the specific length.

Clonotype Distribution: This feature is used to display the abundance distribution of clonotypes in the selected samples and draw a distribution bar plot. (Fig. 10).

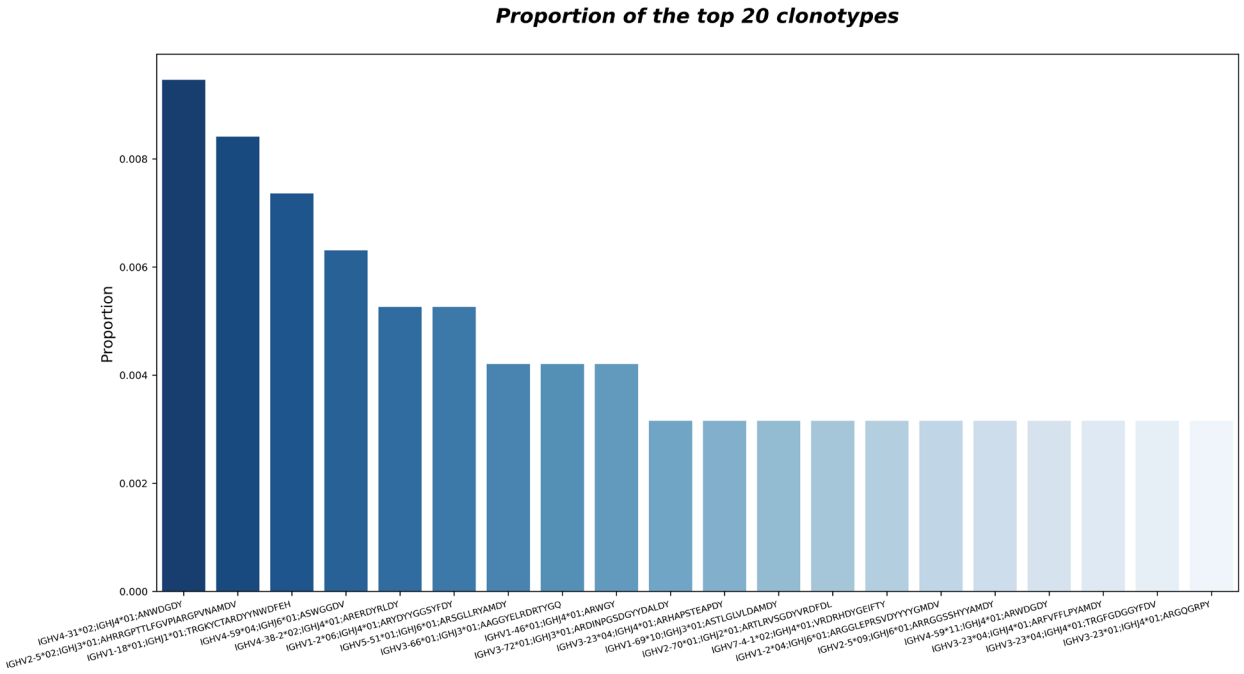


Figure 10. An example of clonotype distribution. The x-axis represents the top 20 clonotypes, and the y-axis represents the proportion of clonotypes.

Clonotype Diversity: This feature calculates various diversity indices for the selected sample and displays them in a radar chart (Fig. 11).

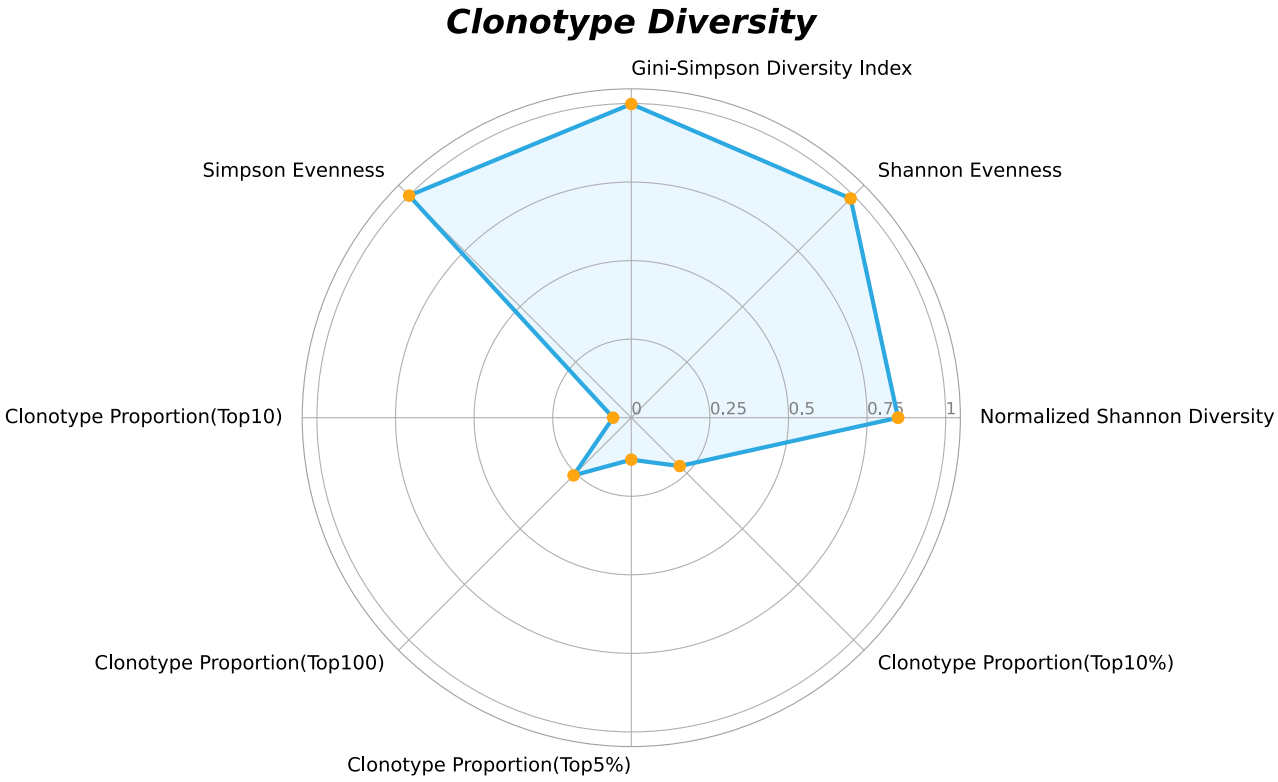


Figure 11. An example of radar chart.

Humanness & Mutation Profile: This feature enables the identification of unusual residues in the selected query sequence based on residues frequency information in a dataset comprising tens of millions of human sequences from OAS[8]. The unusual residues information can aid users in the process of humanizing antibodies. Additionally, users can construct their own residues frequency background by inputting their own datasets, ensuring that the analysis is tailored to their individual needs (Fig. 12,13).

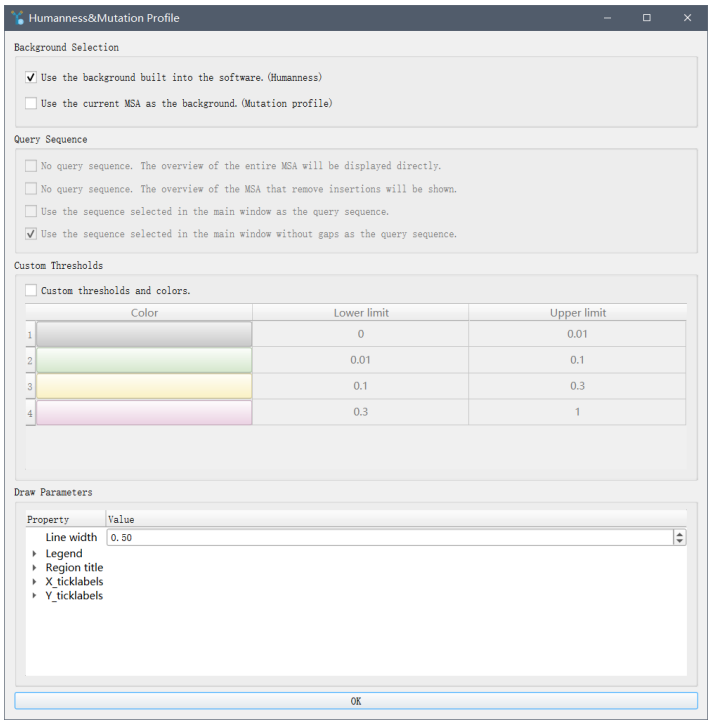


Figure 12. Humanness & Mutation Profile parameters dialog. The **"Background Selection"** allows users to select the background consisting of residues frequency information at each position. The **"Query Sequence"** allows the users to choose whether to use the query sequence and whether to ignore gaps in the query sequence. The **"Custom Thresholds"** can be adjusted to modify the thresholds of amino acid distribution and the corresponding colors they represent. The **"Draw Parameters"** can be customized to adjust the drawing parameters according to individual preferences.

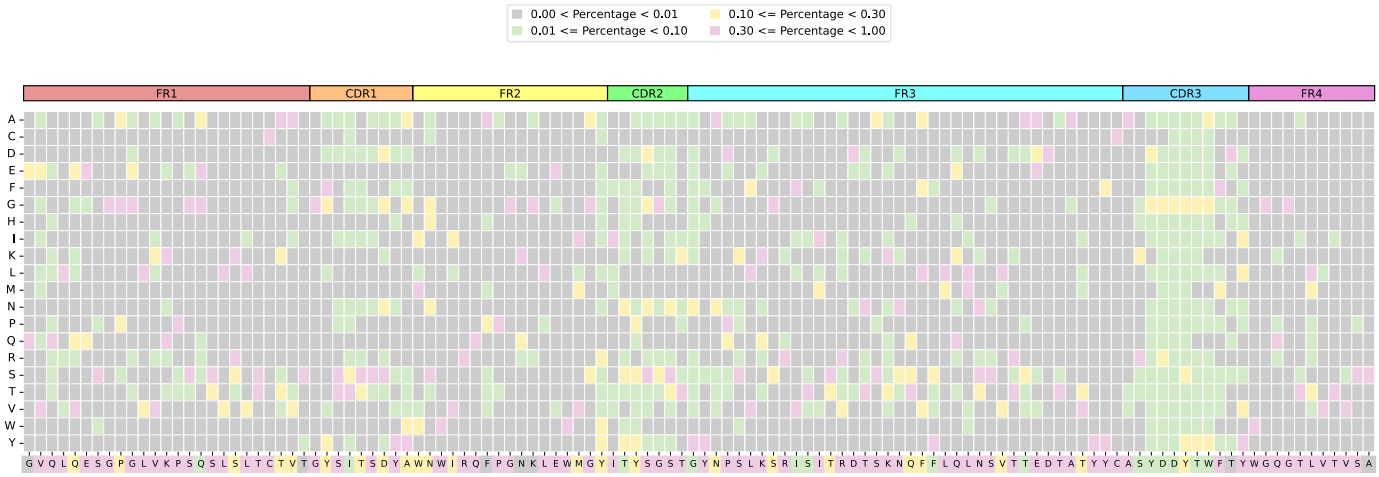


Figure 13. An example of unusual residue map. The residues with different frequencies will be rendered as different colors. The query sequence selected in the Multiple Sequence Alignment window is shown at the

bottom of the heatmap, which is used for comparison with the background. The FR and CDR regions of the variable domain are marked with different colors on the top. Unusual residues are marked in gray.

Seqlogo (only for linux version): This feature allows users to generate a visualization of the aligned sequences, where each position in the alignment is represented by a stack of letters, with the height of each letter representing its frequency or entropy. The **"By Entropy"** option generates a Seqlogo plot with entropy as the Y-axis for the displayed MSA currently (Fig. 14). The **"By Frequency"** option generates a Seqlogo plot with frequency as the Y-axis for the displayed MSA currently. The **"Color"** option be used to change the rendering mode of the Seqlogo plot.

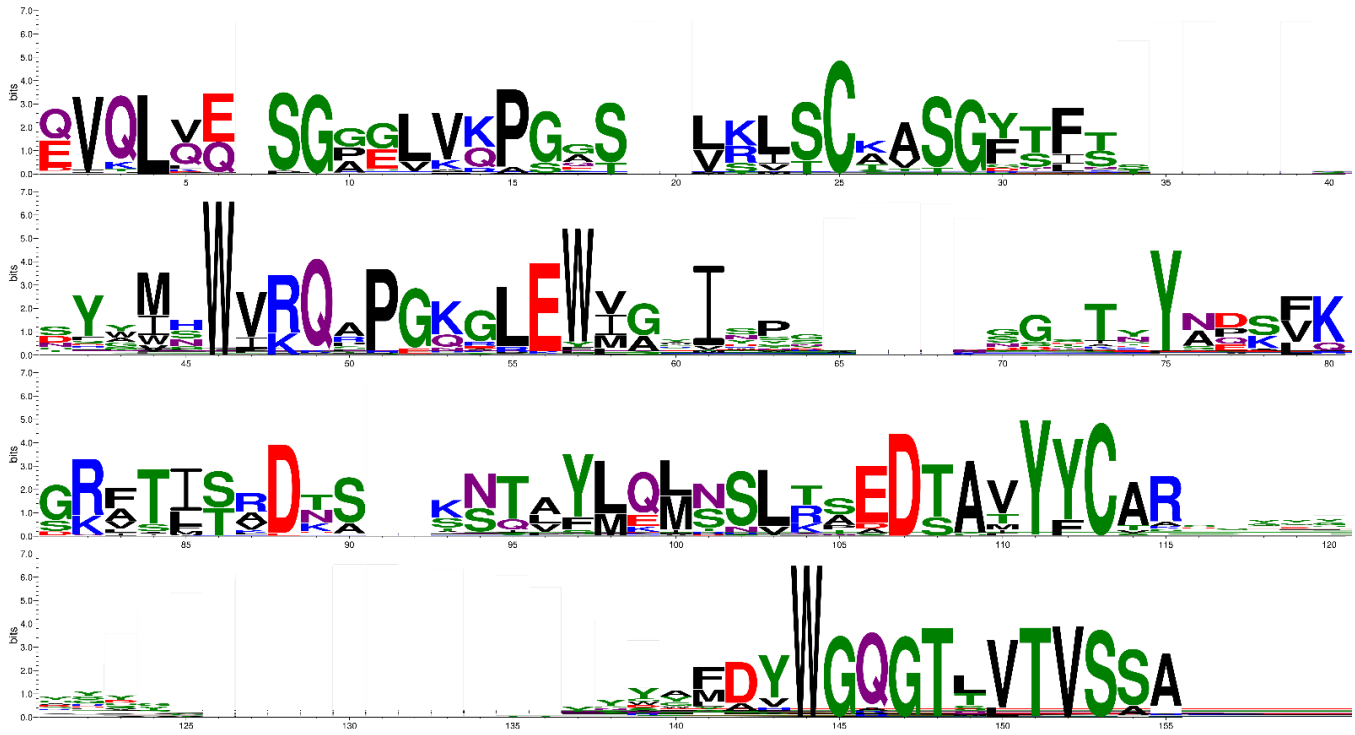


Figure 14. A seqlogo plot by Entropy. The y-axis represents the entropy of amino acids, and the x-axis represents the position in the variable domain. The larger the letter of the amino acid, the greater the entropy value.

Setting

Here are two options available: **"Align Parameters"** and **"Temporary Path"**. The **"Align Parameters"** menu provides options for MSA (Fig. 14). The **"Temporary Path"** menu allows for customization of the path of the temporary files used by Abalign. After making changes to the temporary path, please restart the software.

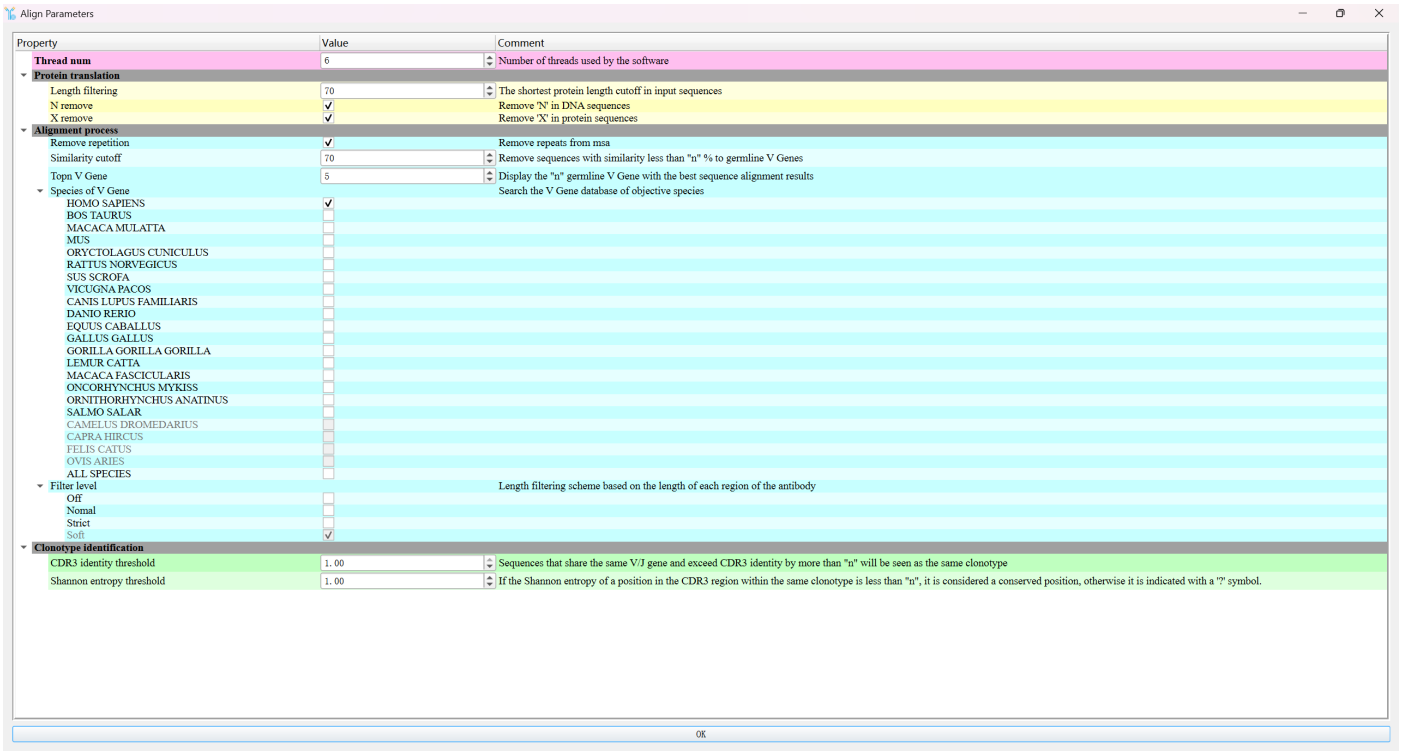


Figure 15. The **"Align Parameters"** contains three columns. The first column lists the functions or elements that can be modified, while the second column shows the corresponding values and whether or not the functions are enabled. The third column provides a description of each parameter.

Help

User can click the **"Help"** button to open the user manual file.

Example

Example: BCR/antibody sequence data (nucleic acid or protein) are provided for testing purposes. The **"Example"** function provides two options: **"SingleFile"** and **"MultiFile"**. Users can find the example files under the "example" folder in the installation path.

Usage Case

Users can choose to load the example file by clicking on **"Example"** in the menu bar, followed by selecting **"SingleFile"**. Alternatively, they can select the input file by clicking on the **"Input"** button in the toolbar.

Step 1. Input file: To begin, load a FASTA format file by clicking on the **"Input"** button in the toolbar or by selecting **"Example"** -> **"SingleFile"** in the menu bar to load the example file.

Step 2. Identify antibody variable domain and run multiple sequence alignment: Once the sequences have been loaded, select the desired scheme and chain type, and then click on **"Align"** to initiate the program. The progress of the program will be displayed in the progress bar located at the bottom. Once the program has completed, a dialog box will appear, which shows the statistical information (Fig. 16). During this process, duplicated sequences will be detected by default.

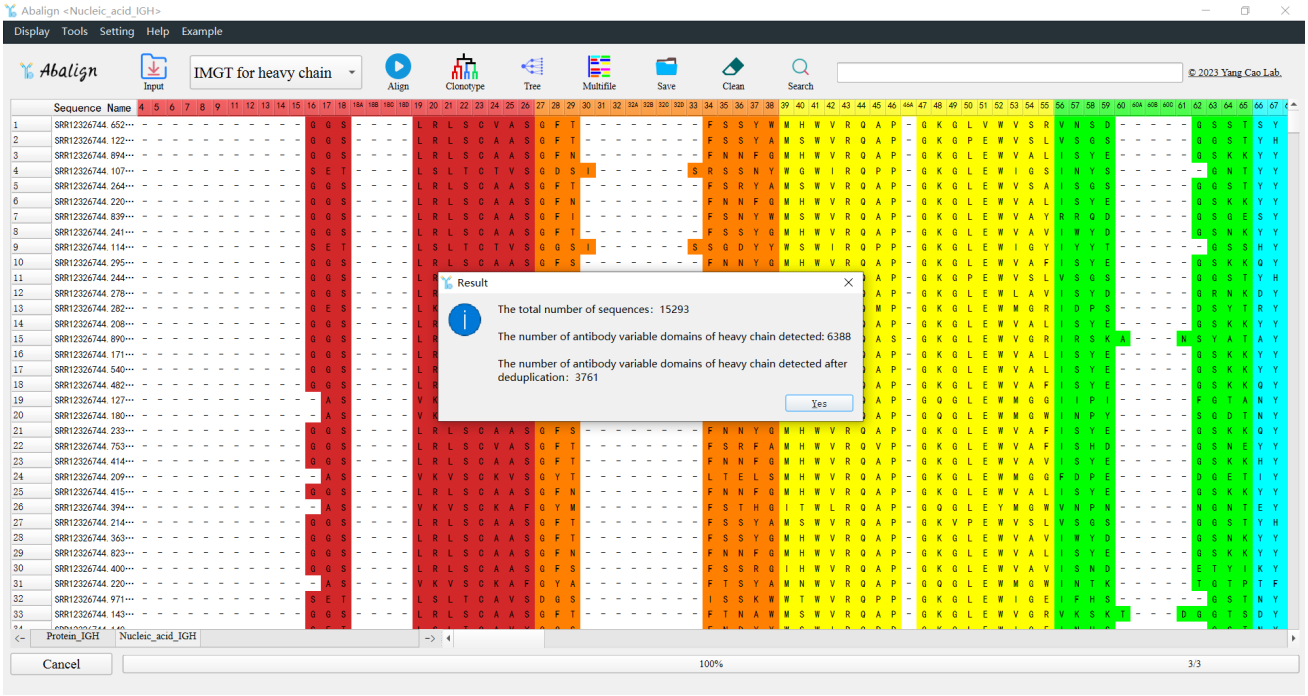


Figure 16. The result of MSA using Nucleic_acid_IGH sample

Step 3. Count sample composition: Users can click the **"Tools"** button in the top menu bar and click either **"Gene Usage"** or **"Sequence Enrichment "** to obtain the abundance map of Genes or sequences (Fig. 17).

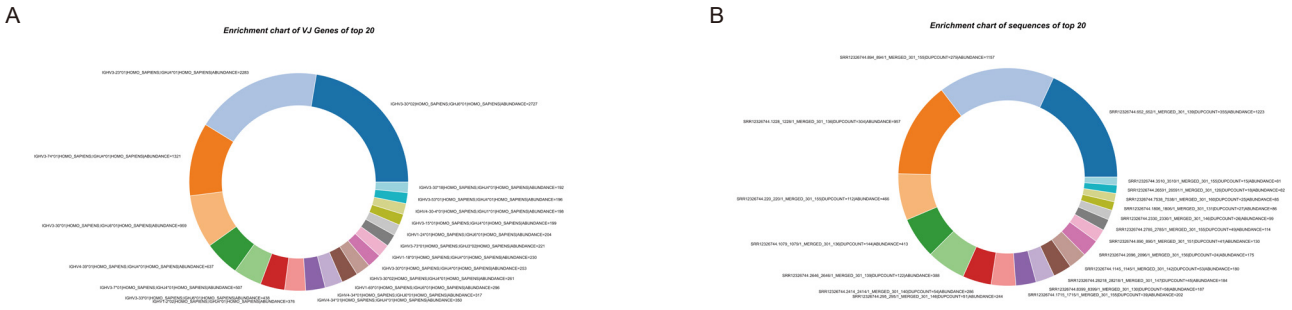


Figure 17. The top20 VJ gene combinations and sequences of Nucleic_acid_IGH sample

Step 4. Build B-cell lineage tree: Users can construct B-cell lineage tree from multiple perspectives, such as different genes or different clonotypes. For example, users are interested

in the most abundant VJ gene combinations "IGHV3-23*01, IGHJ4*01". They can select "Sequences belonging to a specific gene combination" mode and find out the most abundant VJ gene combinations "IGHV3-23*01, IGHJ4*01". Then click on the **"Run"** button to build the B-cell lineage tree(Fig. 18). If users are satisfied with the results, they can save the .nwk file by clicking on the **"Save current Nwk file"** button in the tree building menu.

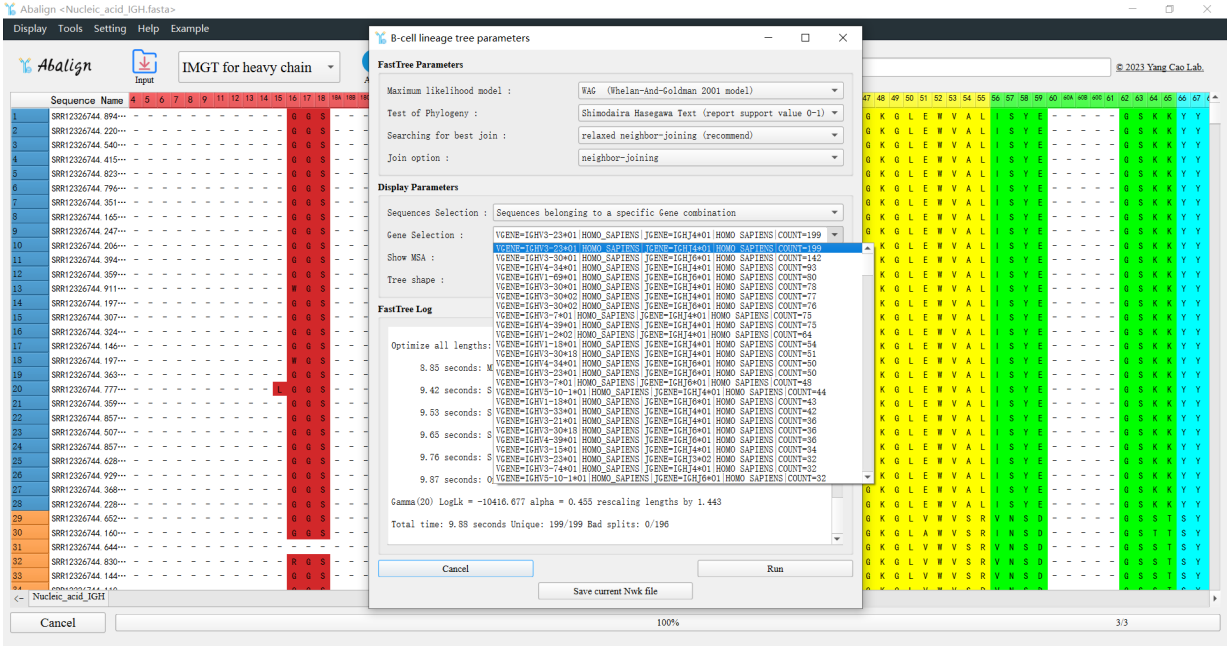


Figure 18. Use sequences belonging to IGHV3-23*01, IGHJ4*01 to build a B-cell lineage tree.

Step 5. View the B-cell lineage tree: Once the B-cell lineage 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. 19). After selecting the node of the tree, users can also modify the attributes of the nodes, such as the color, in the right window.

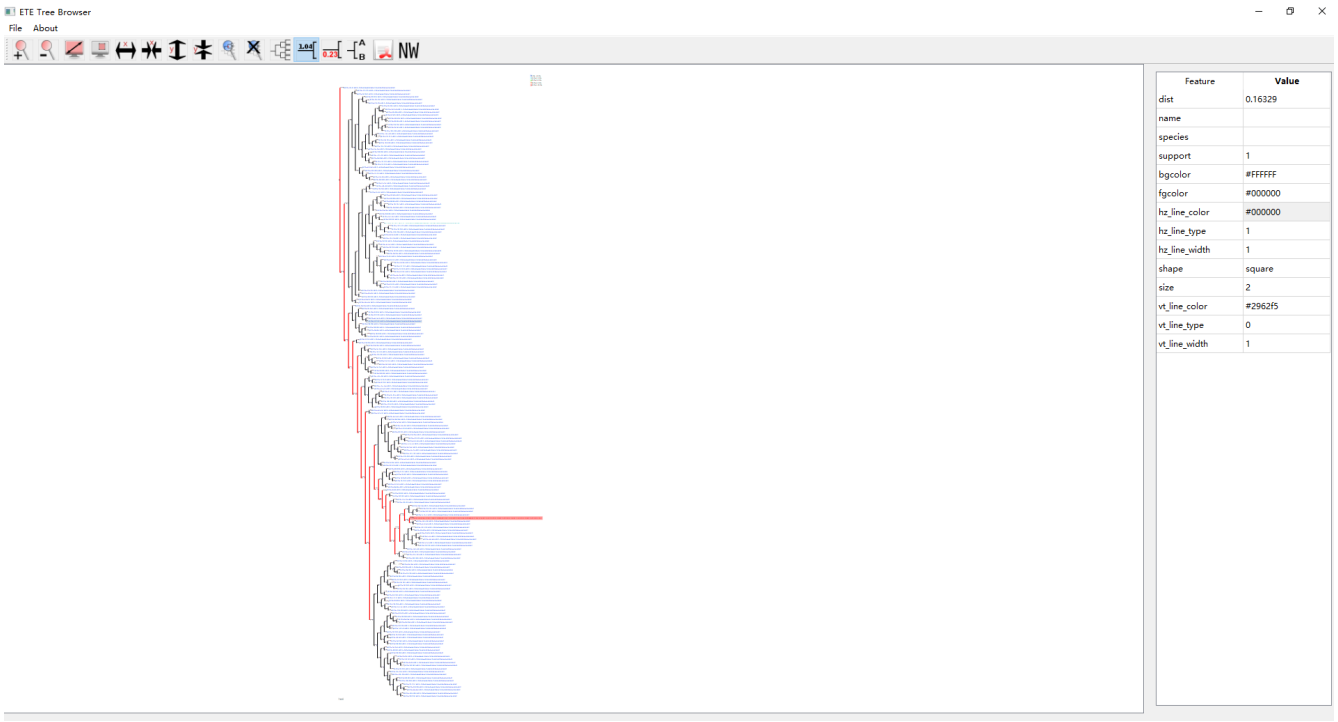


Figure 19. Results of B-cell lineage tree construction using sequences belonging to IGHV3-23*01, IGHJ4*01.

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, users can click on **"Tools"->"Humanness & Mutation Profile"** to view the data (Fig. 20).

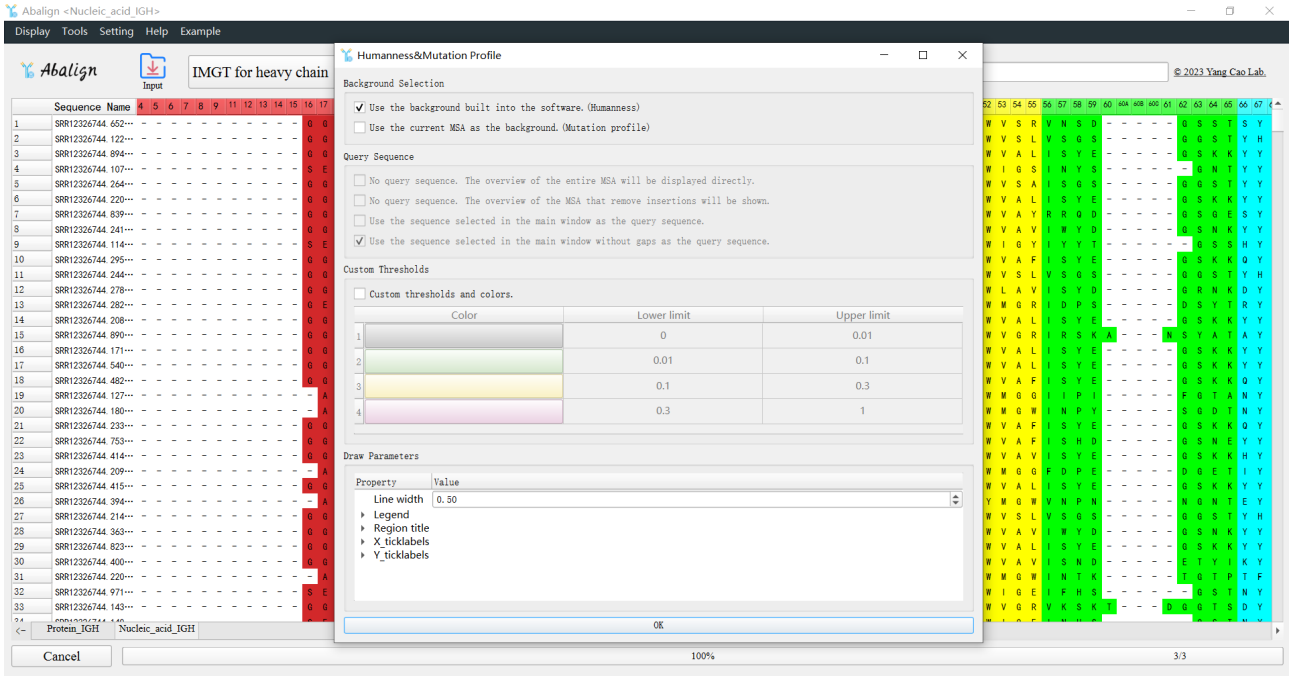


Figure 20. Select the first sequence in MSA to construct the unusual residue map.

The default setting highlights low-frequency residues (unusual residues) in gray if they exist in the query sequence (Fig. 21). This feature is useful in finding substitutions with the corresponding high-frequency residues found in human BCRs.

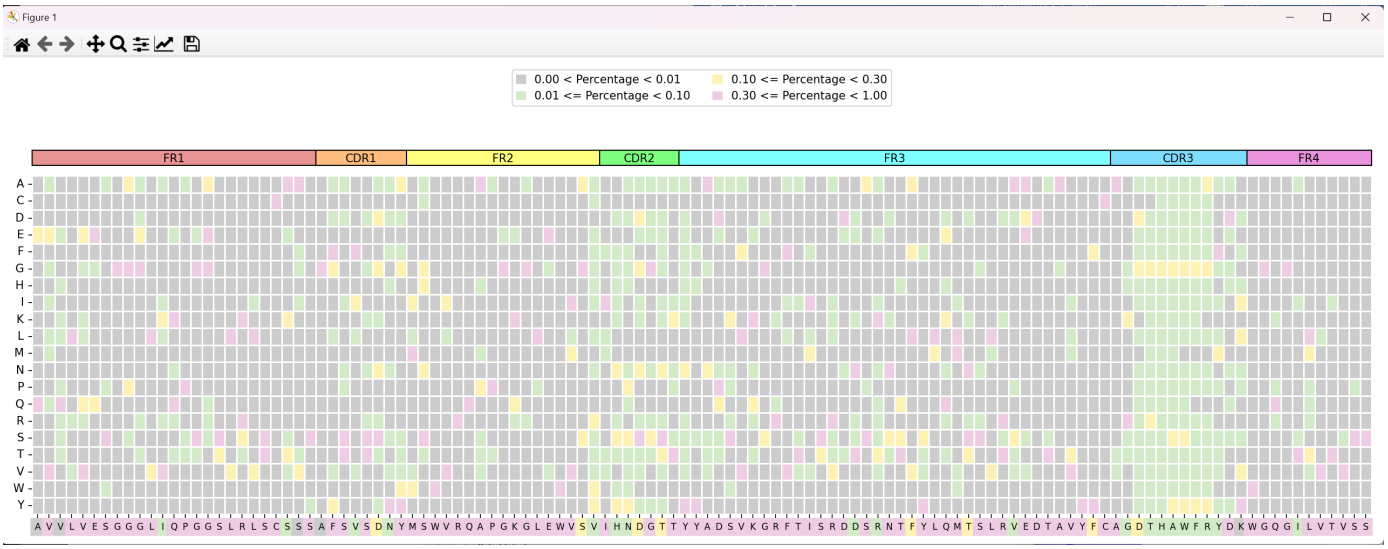


Figure 21. Unusual residue map of the first sequence in MSA.

Step 7. Cluster sequences with clonotypes: By clicking the **"Clonotype"** button in the

toolbar, Abalign will group sequences by clonotype and arrange sequences with the same clonotype together, highlighting them in the same color (Fig. 22).

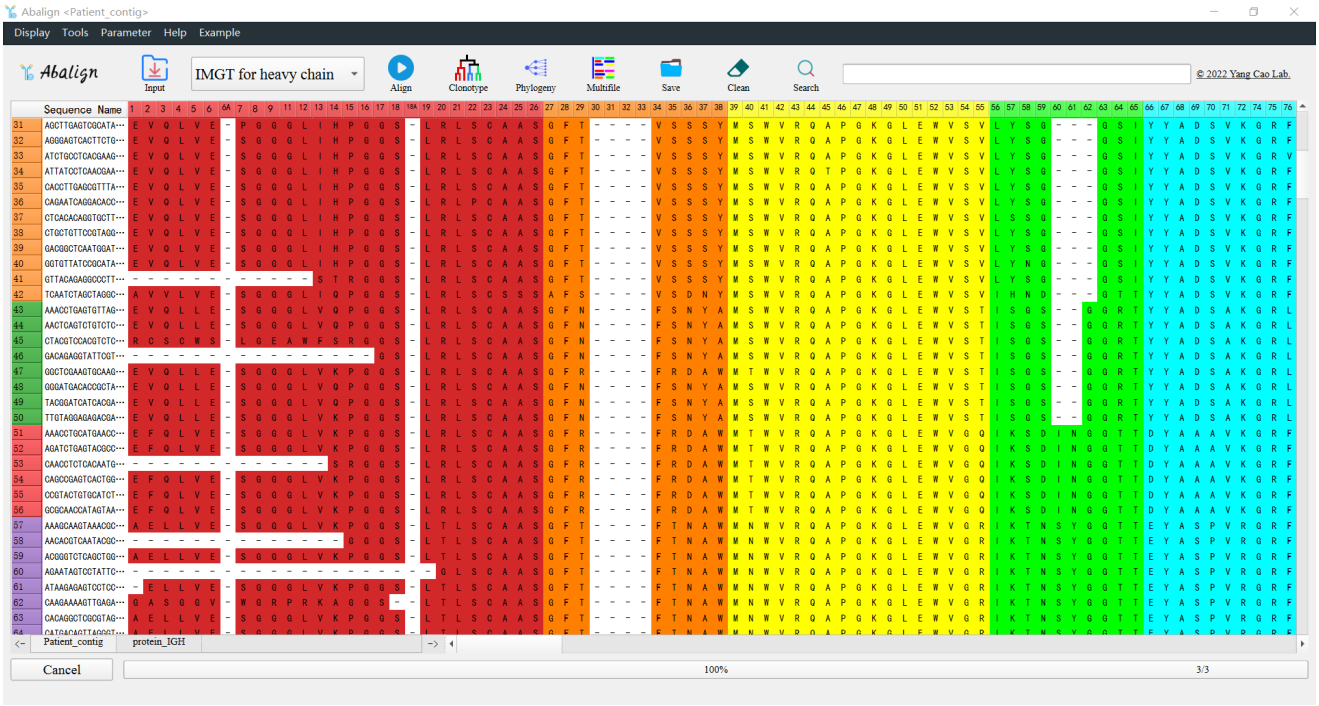


Figure 22. Rearrange sequences belonging to the same clonotype together and render them in the same color in the line label.

Step 8. Build of B-cell lineage tree by clonotype: After clicking the **"Clonotype"** button , users can display the sequences belonging to a specific clonotype individually in the main window (Fig. 23). Then, a B-cell lineage tree can be constructed with these sequences.

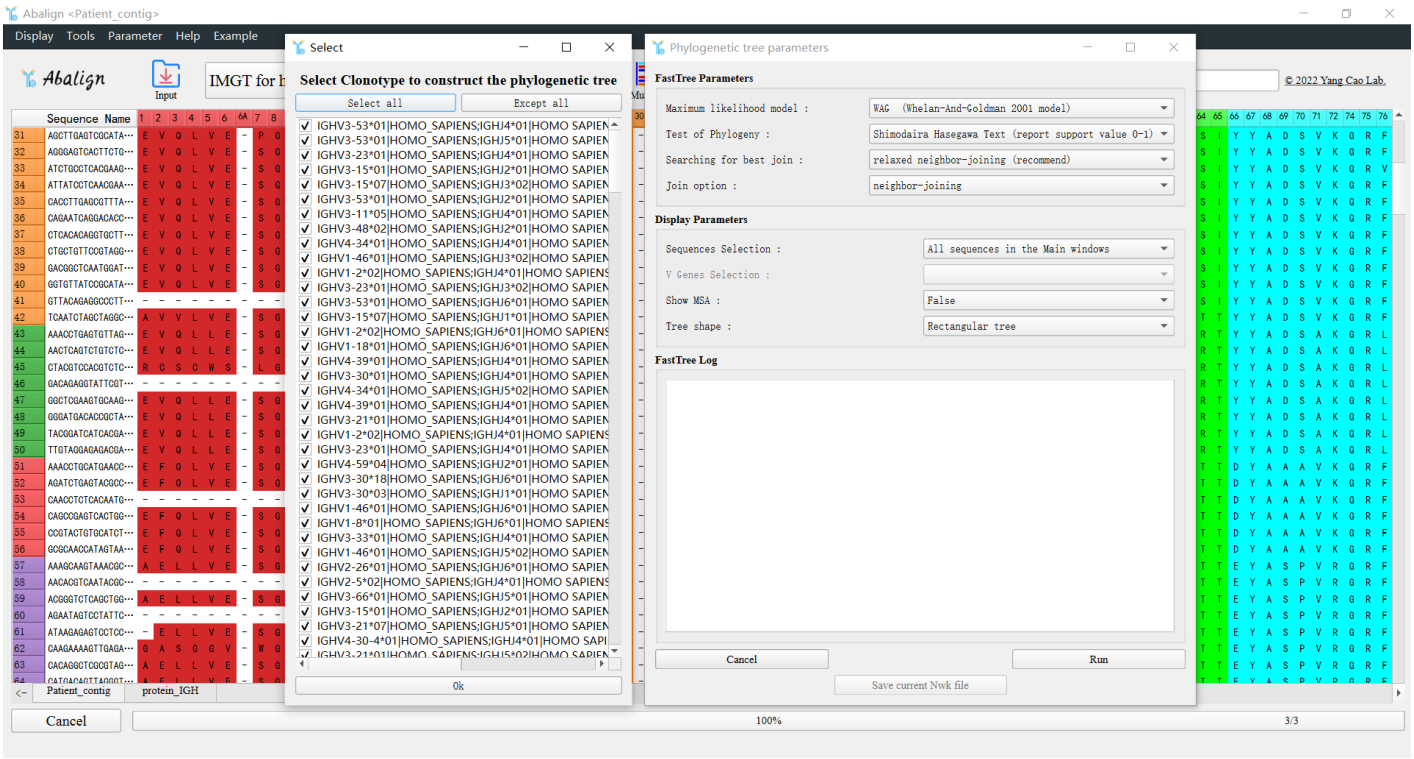


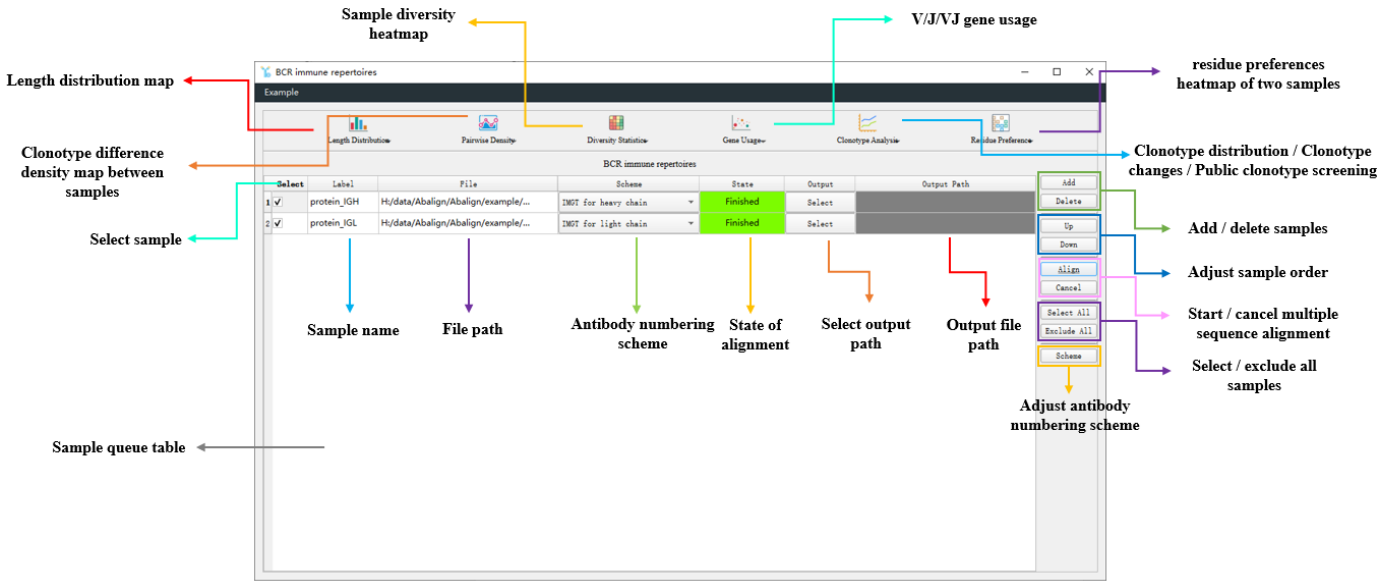
Figure 23. Selecting specific clonotypes to construct B-cell lineage trees.

Step 9. Save File: To save the sequences displayed in the window, click the **"Save"** button located in the toolbar. In addition to saving all sequences, users can selectively save the sequences they need by utilizing the **"Display by Genes"**, **"Display by Regions"** and **"Display by Clonotypes"**.

Step 10. Multifile: To align and analyze multiple B cell receptor immune repertoires data, users can click the **"Multifile"** button located in the main window. This feature allows for cross-analysis of multiple datasets.

BCR repertoires management window

Window Layout



Usage

Add File: Click on **"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 in the table.

Delete File: Select the checkbox in the **"Select"** column of the table corresponding to the files users wish to delete, then click on **"Delete"** button to remove them.

Multiple Sequence Alignment: Select the checkbox in the **"Select"** column of the table corresponding to the files users want to align, then click on **"Align"** button. Select a value in

the **"Scheme"** column of the table to adjust the numbering scheme. In addition, it can also be modified in the window that pops up after clicking on **"Scheme"** button. The alignment parameters can be adjusted in the window that pops up after clicking **"Align"** button. **"State"** column of the table shows the progress of alignment. **"Cancel"** button is used to terminate alignment in the queue.

Save File: Click on **"Select"** button in **"Output"** column of the table to determine the output path. After selecting the output path, Abalign generates several files. **".fas"** and **".temp.txt"** both record the results of multiple sequence alignment of selected sample, with the difference being that the latter uses "*" to divide the FRs and CDRs. **".number.txt"** records the antibody numbering 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. **".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. 24). The options include FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4, variable domain and CDR1-FR4 (FR1 will be incomplete if the sequencing data quality is poor).

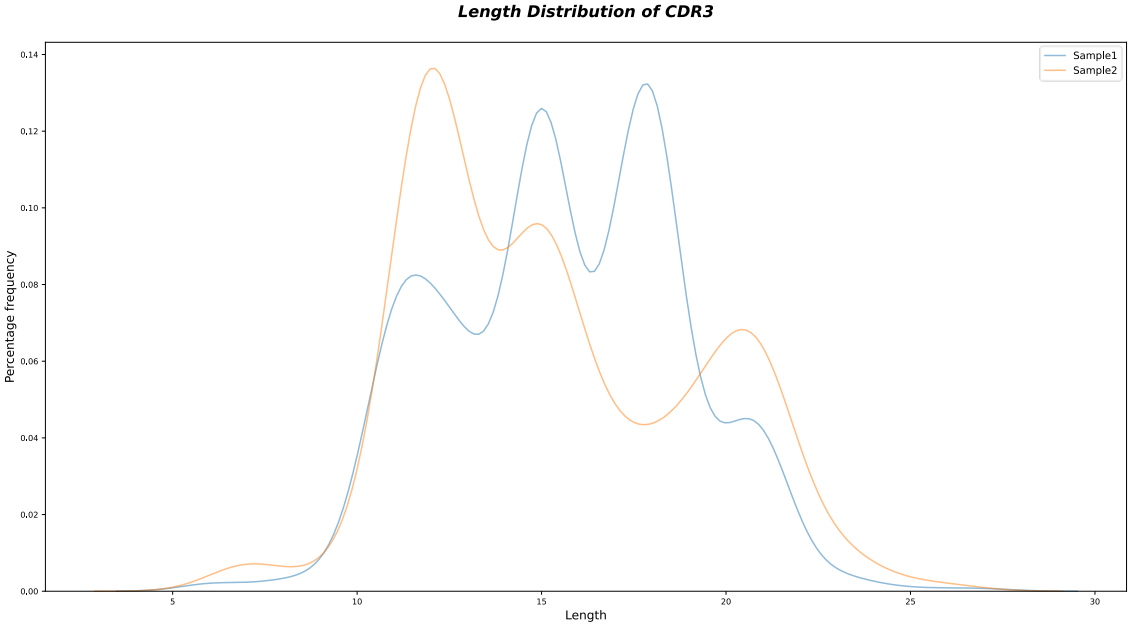


Figure 24. An example of CDR3 length distribution. The x-axis represents the length of the sequences, while the y-axis represents the percentage of sequences with a certain length. The lines with different colors represent different samples. The plotted data can be saved by clicking on **"Save Sources"** button.

Pairwise Density: Count the clonotype difference between two selected samples and draw the density map (Fig. 25).

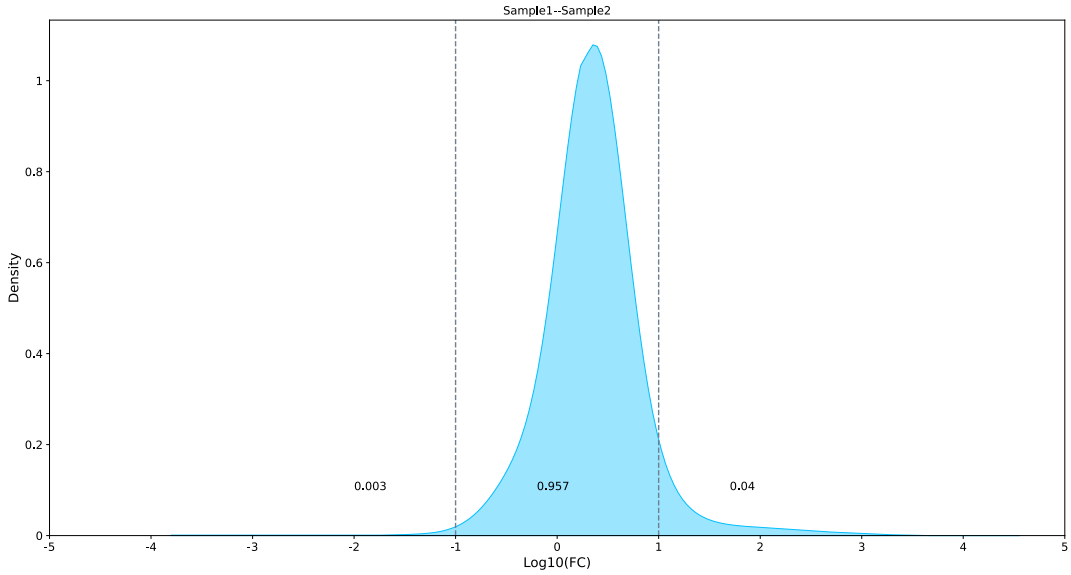


Figure 25. An example of clonotype differential density map between two samples. The x-axis represents the base 10 logarithm of the clonotype fold change, while the y-axis represents the density value. After removing low abundance clonotypes, the remaining clonotypes are divided into three groups based on their log10(FC) values. Clonotypes with log10(FC)>1 are classified as expanded, those with log10(FC)<-1 are reduced, and those with -1<log10(FC)<1 are constant. The numbers shown in the figure represent the proportion of the clonotypes contained in each group.

Diversity Statistics: Assess the clonotype diversity of selected samples and draw the heatmap (Fig. 26).

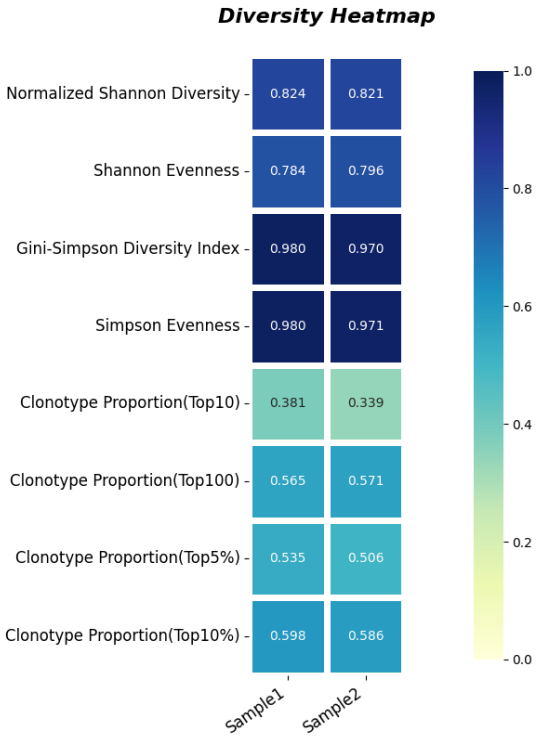


Figure 26. An example of diversity statistics. The x-axis represents different samples, and the y-axis represents the diversity assessment indexes. The darker the color, the higher the value, and the lighter the color, the lower the value. The plotted data can be saved by clicking on **"Save Sources"** button. After saving, Abalign generates a file named **"diversity_heatmap.csv"**, which records the diversity assessments for selected samples.

Gene Usage: Count gene usage of selected samples, and draw the histogram (Fig. 27)/scatter plot (Fig. 28).

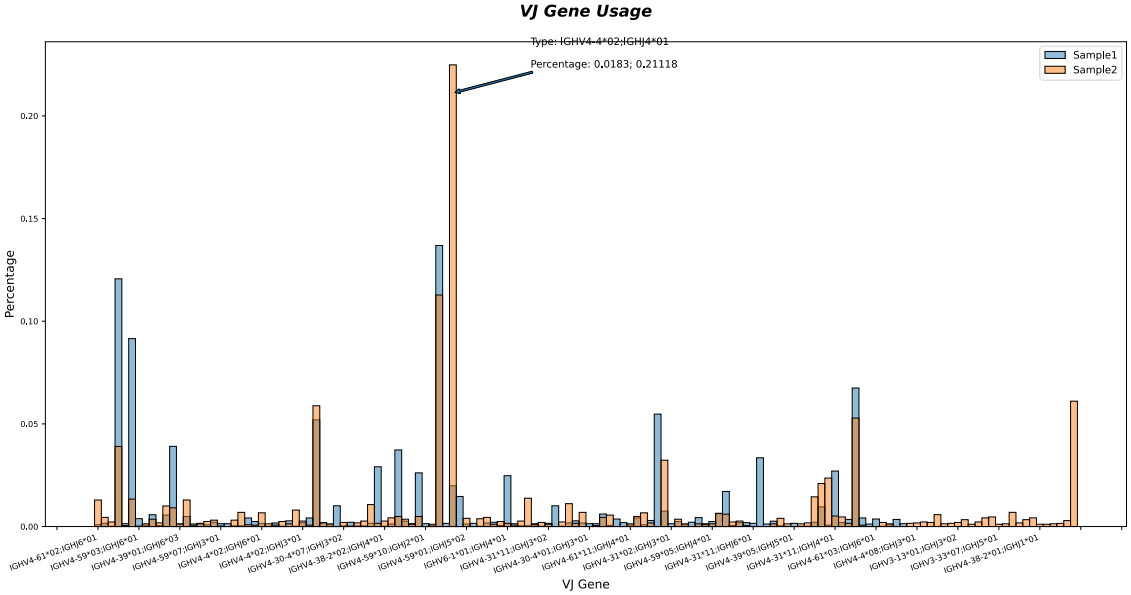


Figure 27. An example of VJ gene abundance histogram. The x-axis represents the combination of VJ gene, and the y-axis represents the proportion of the specific gene. The bars are colored differently to represent different samples. Details are displayed when hovering.

Figure 28. An example of VJ gene abundance scatter plot. The x-axis represents the type of V gene, and the y-axis represents the type of J gene. Each point represents a specific VJ gene combination, with the size of the point indicating its proportion. The points are colored differently to represent different samples.

Clonotype Analysis: There are four options in this menu. **"Clonotype Distribution"** displays the distribution of clonotype in the selected samples and draw the distribution bar plot (Fig. 29). **"Clonotype Changes"** displays the clonotype changes in the selected samples and draw the difference bar plot (Fig. 30). **"Public Clonotypes"** displays clonotypes that are shared between different samples and draw the public clonotypes bar plot (Fig. 31). **"Public Clonotypes between groups"** is used to display the expanded clonotypes that are shared between different groups. After clicking this button, **"Group Selector"** will pop up (Fig. 32), which is used to group different samples and adjust the parameters for screening expanded clonotypes. Public expanded clonotype result will be presented as an upset plot (Fig. 33).

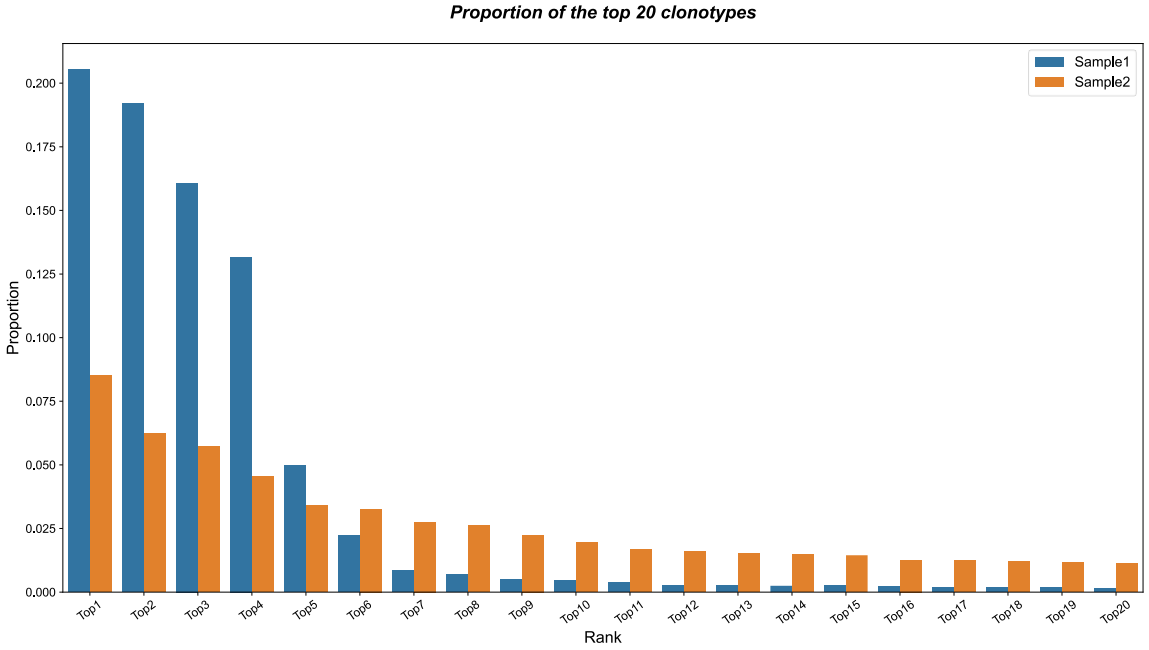


Figure 29. An example of clonotype distribution. The x-axis represents the top 20 clonotypes, and the y-axis represents the proportion of clonotypes. The bars with different colors represent different samples.

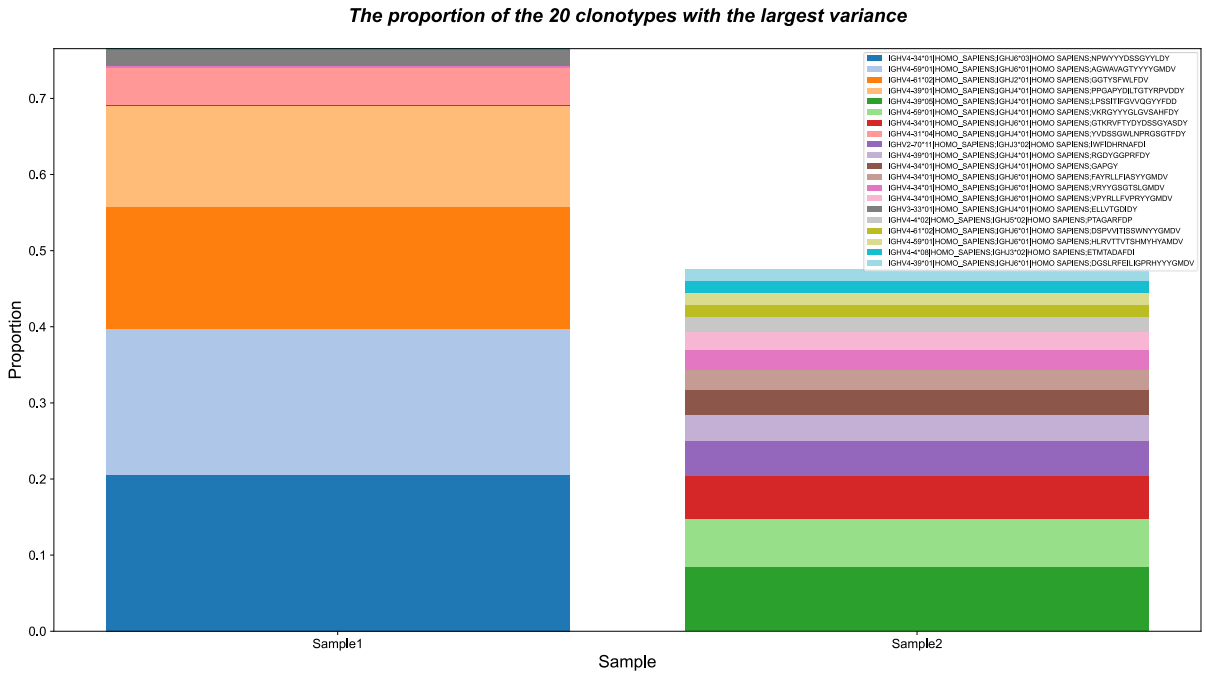


Figure 30. An example of clonotype changes. The x-axis represents different samples, and the y-axis represents the proportion of the top 20 clonotypes that vary across selected samples. The plotted data can be

saved by clicking on **"Save Sources"** button. After saving, Abalign generates a file named **"clonotype_changes.csv"**, which records the abundance and changes of all clonotypes of selected samples.

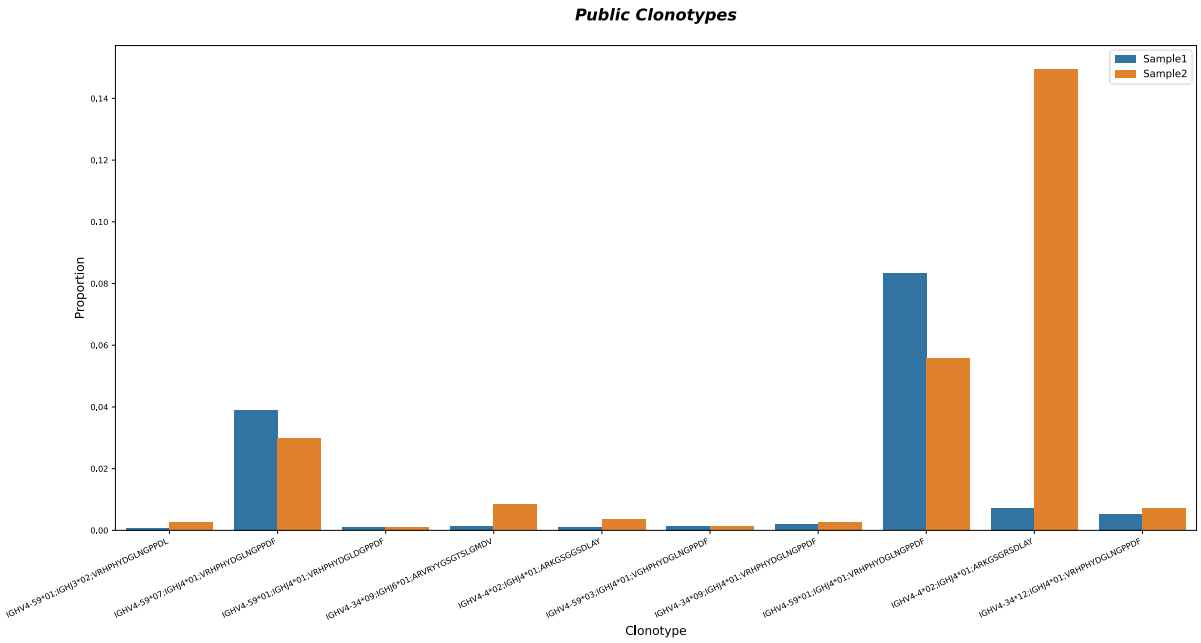


Figure 31. An example of public clonotypes of selected samples. The x-axis represents the specific clonotype, and the y-axis represents their proportion. Bars with different colors represent different samples. The plotted data can be saved by clicking on **"Save Sources"** button. After saving, Abalign generates a file named **"Public_clonotype.csv"**, which records the top N clonotypes shared among all samples.

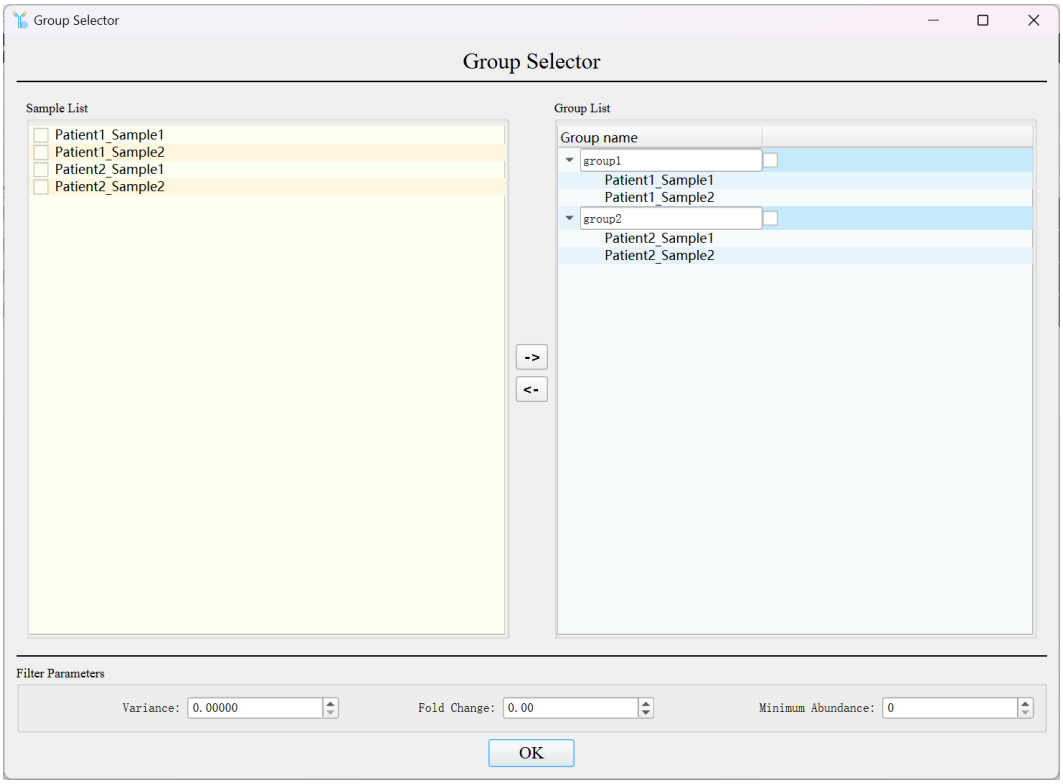


Figure 32. Group Selector. **"Sample List"** displays all the samples that are to be analyzed and have been selected in the BCR repertoires management window. **"Group List"** shows the groups created by users. Users can create a group by selecting samples from the **"Sample List"** and clicking on the move button located in the middle of the interface. Similarly, users can remove a group by selecting it from the **"Group List"** and clicking on the move button. **"Filter Parameters"** are used to control the conditions for expanded clonotypes.

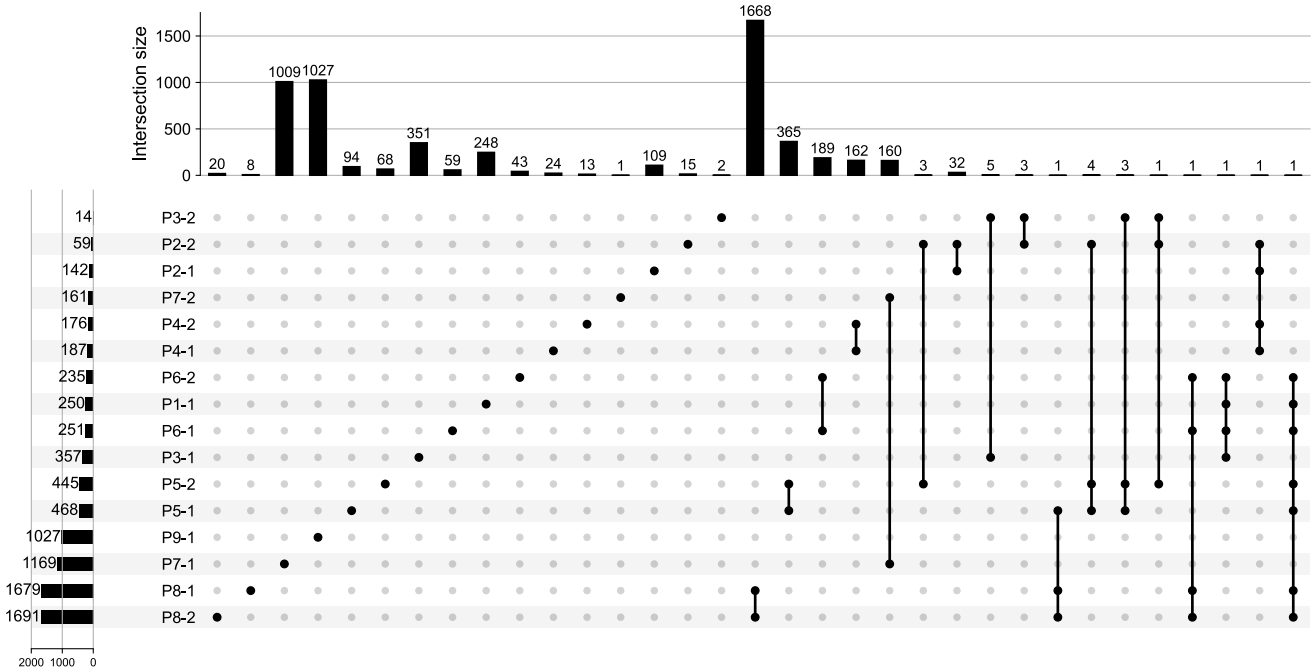


Figure 33. An example of public clonotypes between groups. The left bar represents the number of clonotypes expanded in each group (e.g., one patient represents one group), while the upper bar represents the number of public clonotypes between different groups. The set information corresponding to the public clonotypes is shown as a line, with the dot on the line representing the group within the set. The plotted data can be saved by clicking on **"Save Sources"** button. After saving, Abalign generates several files, including multiple files named **"group.csv"** that record changes of clonotypes in each group, and one file named **"between_groups.csv"** that records details of the public clonotypes.

Residue Preference: Count the preferences of residue between the two samples. After clicking on this button, a dialog will pop up (Fig. 34), which is used to adjust the parameters for residue preference (Fig. 35).

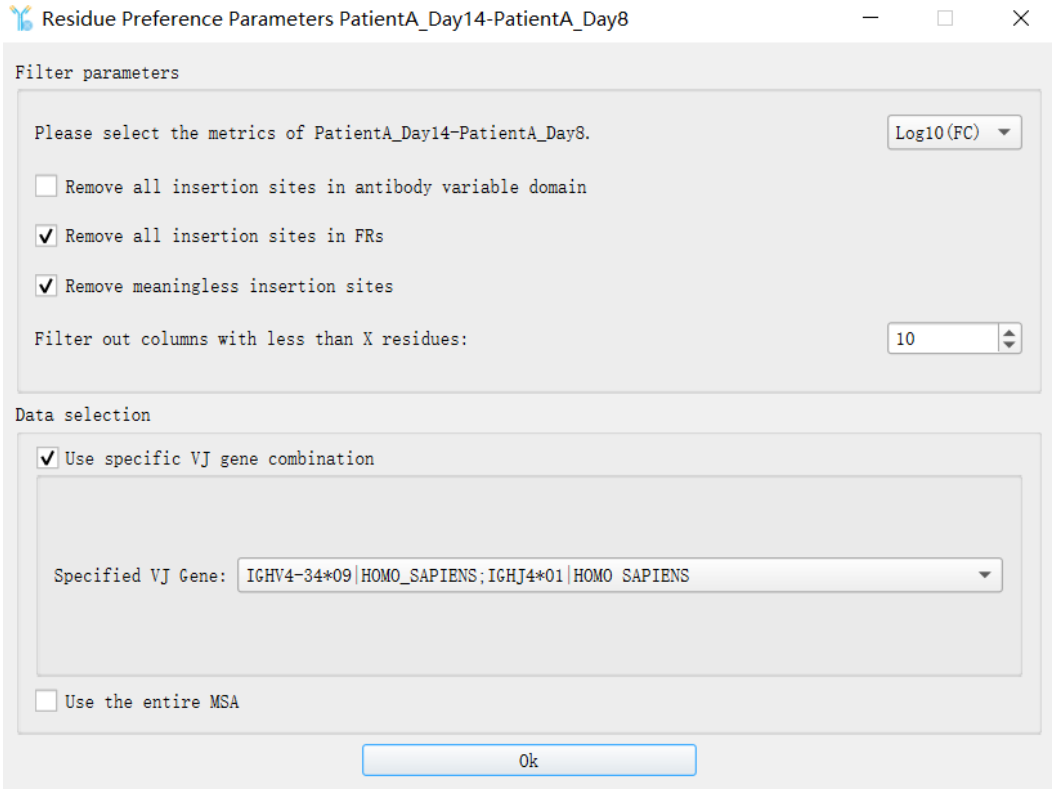


Figure 34. Residue Preference Parameters Dialog. **"Filter Parameters"** control which residue difference positions are displayed. Users can choose whether to retain insertion positions or nonsense positions (the positions with residue occurrences lower than X). **"Data Selection"** controls the data for analysis, allowing users to select the entire MSA or a specific VJ gene combination.

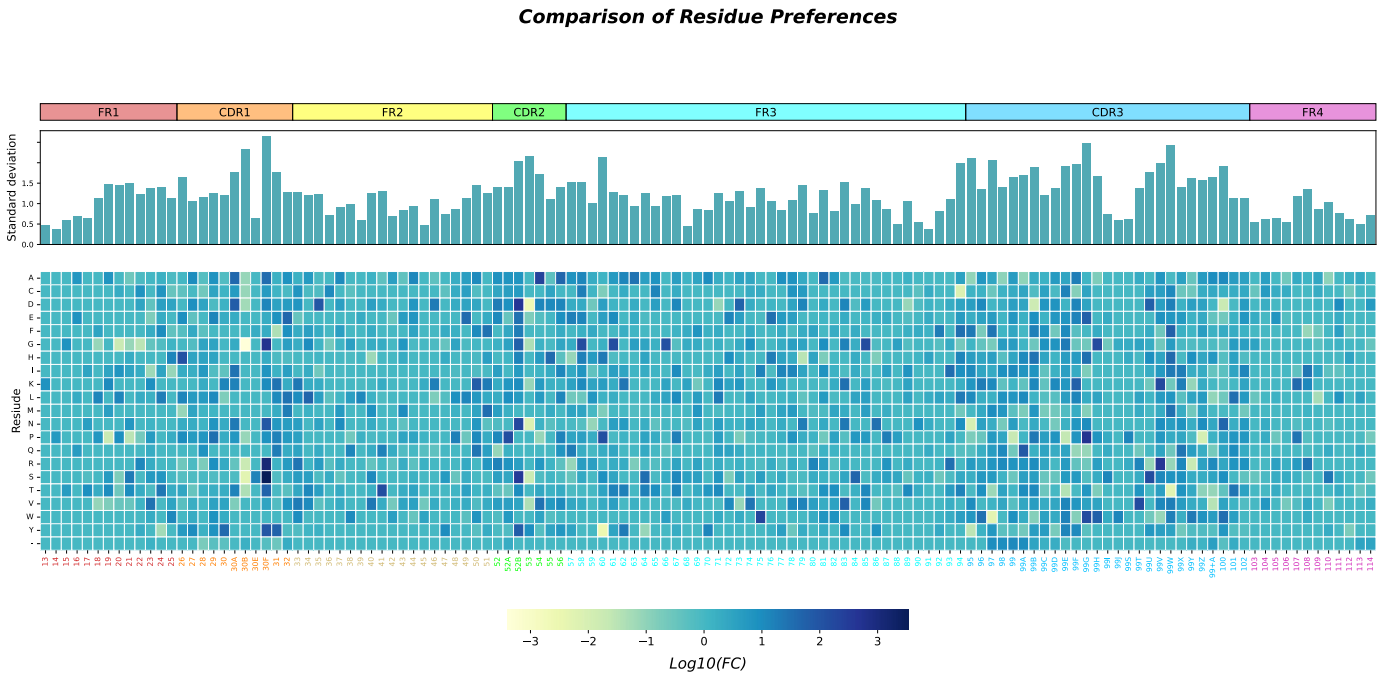


Figure 35. An example of residue preference map. The ribbon at the top uses different colors to distinguish FRs and CDRs. The x-axis at the bottom represents each position in the variable domain. The upper y-axis represents the standard deviation, while the lower y-axis represents 20 residues and gaps. The histogram above represents the standard deviation of residue differences at each position, with larger values indicating stronger residue selection effects. 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 plotted data can be saved by clicking on **"Save Sources"** button. After saving, Abalign generates a **csv** file that records the residue preferences of each position in the variable domain.

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] Sievers F, Higgins D G. Clustal Omega, accurate alignment of very large numbers of sequences[M]//Multiple sequence alignment methods. Humana Press, Totowa, NJ, 2014: 105-116.

[3] Katoh K, Standley D M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability[J]. Molecular biology and evolution, 2013, 30(4): 772-780.

[4] Edgar R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput[J]. Nucleic acids research, 2004, 32(5): 1792-1797.

[5] Hershberg U, Luning Prak E T. The analysis of clonal expansions in normal and autoimmune B cell repertoires[J]. Philosophical Transactions of the Royal Society B:

Biological Sciences, 2015, 370(1676): 20140239.

[6] 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.

[7] 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.

[8] 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.