Catalogs

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 clone 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**.

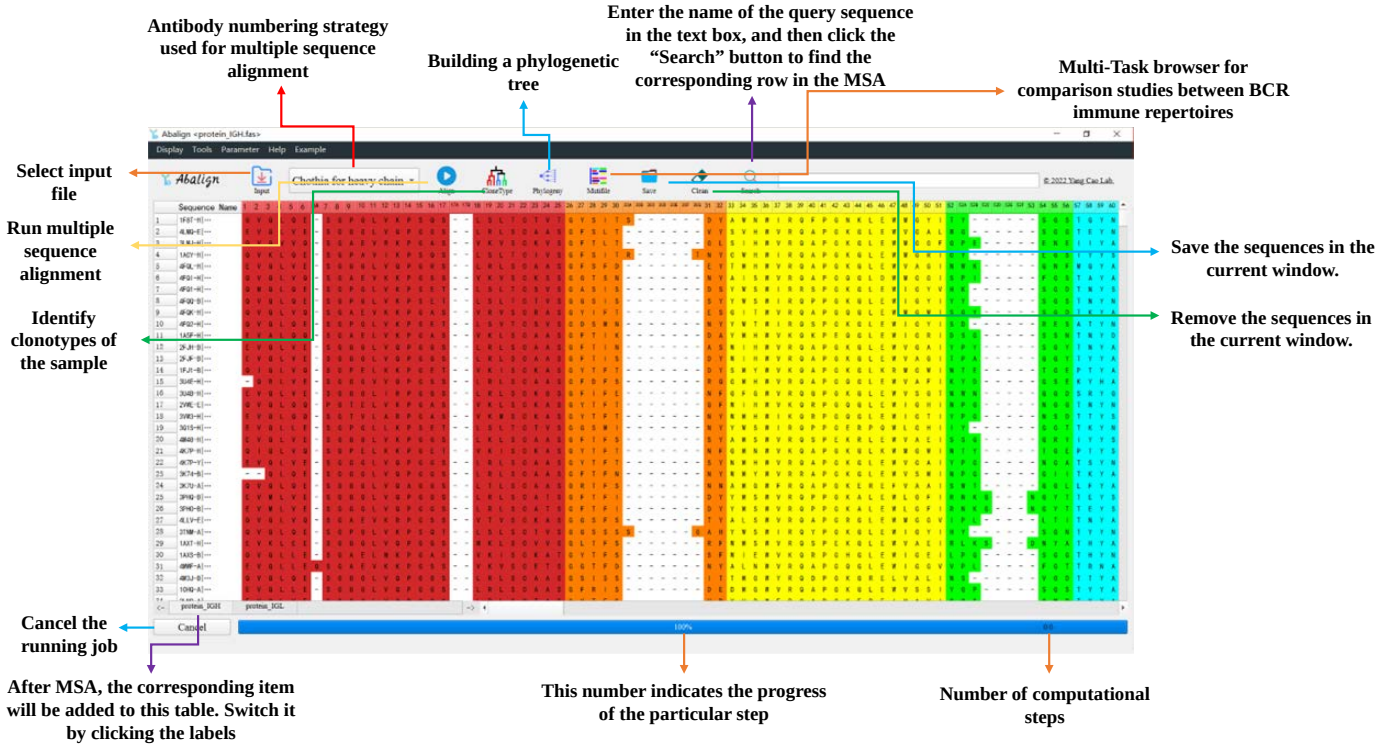
Linux: After extracting the files, go to the extracted folder, find **Abalign_installer.run,** and execute the following command in the terminal to start the installation guide.

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

macOS: After extracting the files, go to the extracted folder, Double-click **Abalign.dmg** to enter the installation. Move the Abalign icon to the Applications folder on the right to complete the installation.

Note: When running Abalign under macOS, if a dialog box 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



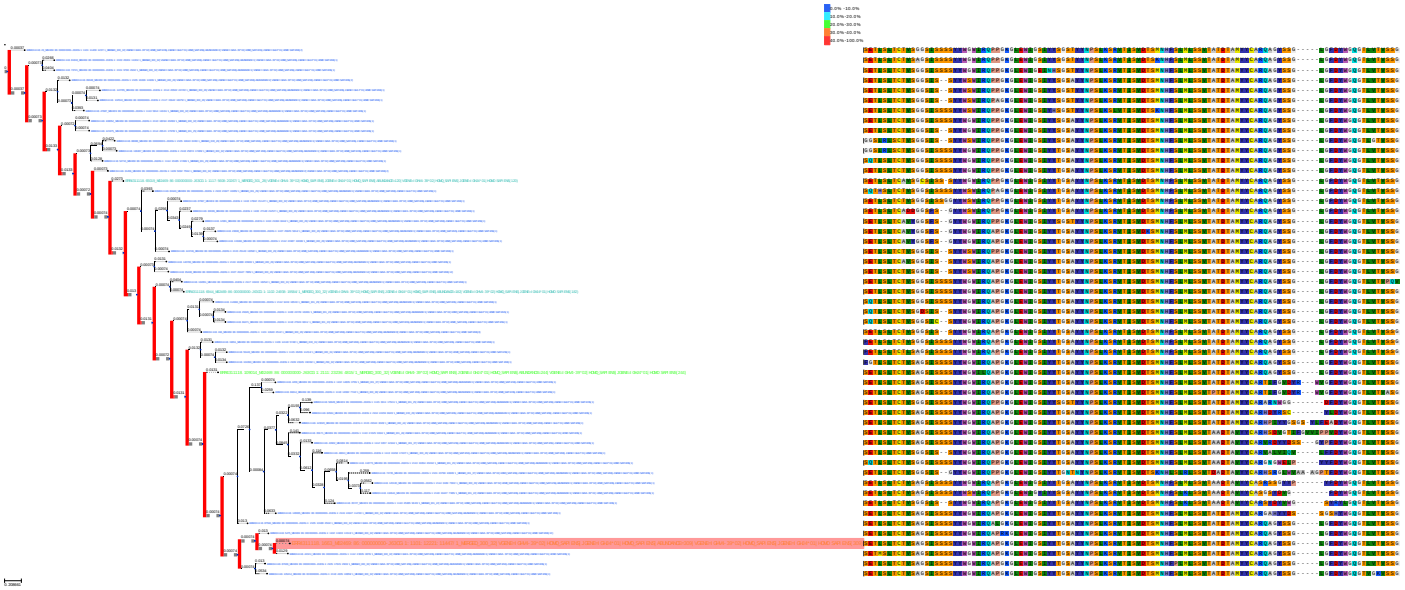
Tool Buttons

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**.

CloneType: 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 CDR.

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, in this dialog you can adjust the parameters for building the phylogenetic tree.



The figure on the left is the phylogenetic tree of the selected sequences, and high lighted 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.

MultiTask: 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: Remove the sequences in the current window.

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.

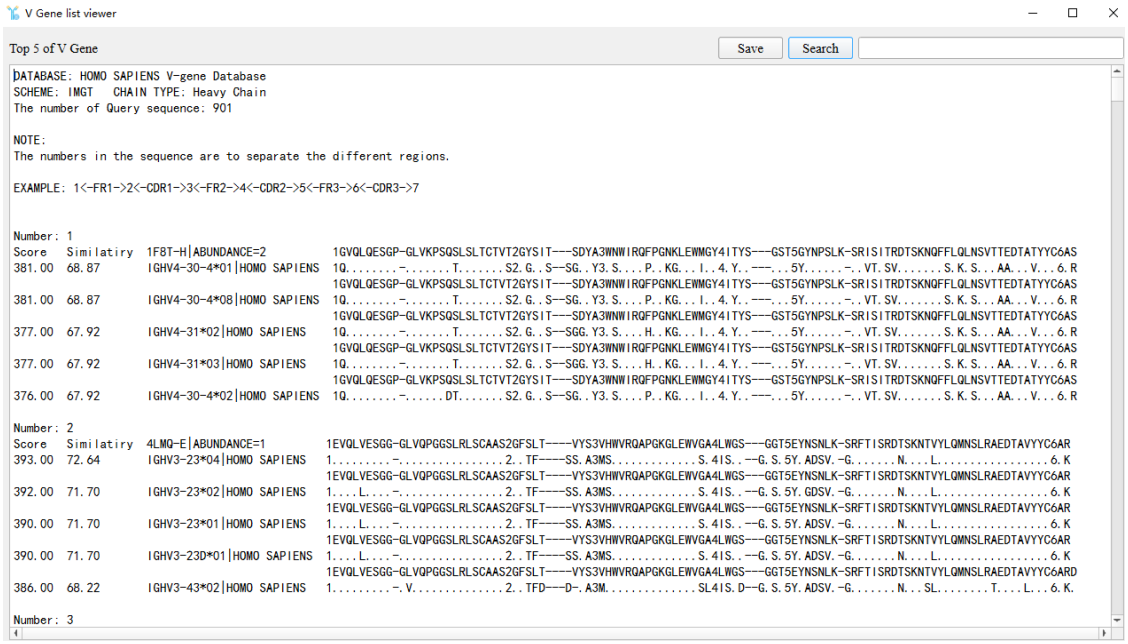
Display by CloneType: Selecting this option will display the specific clonotypes in the sample, which needs to clonetype screening first.

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.

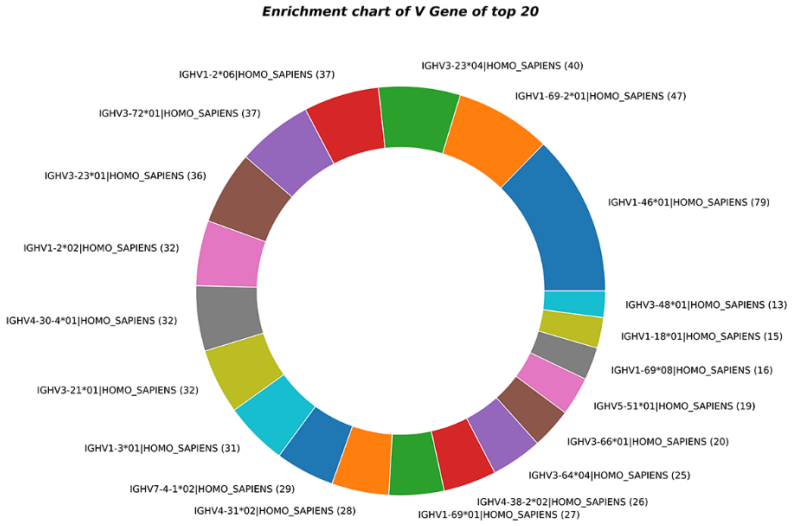
Tools

V Gene: There are two editable options in this menu."Species" provides the V gene species to choose, if you select "Homo sapiens" then only the human V germline gene database will be matched with antibody sequences."Homo sapiens" is used by default, and this option can also be adjusted in “Align parameter”. When identifying antibody heavy/light chains, species with only light/heavy chains will be unable to selected. "Display V Gene list" shows the alignment details of each sequenc with top 5 V germline genes.



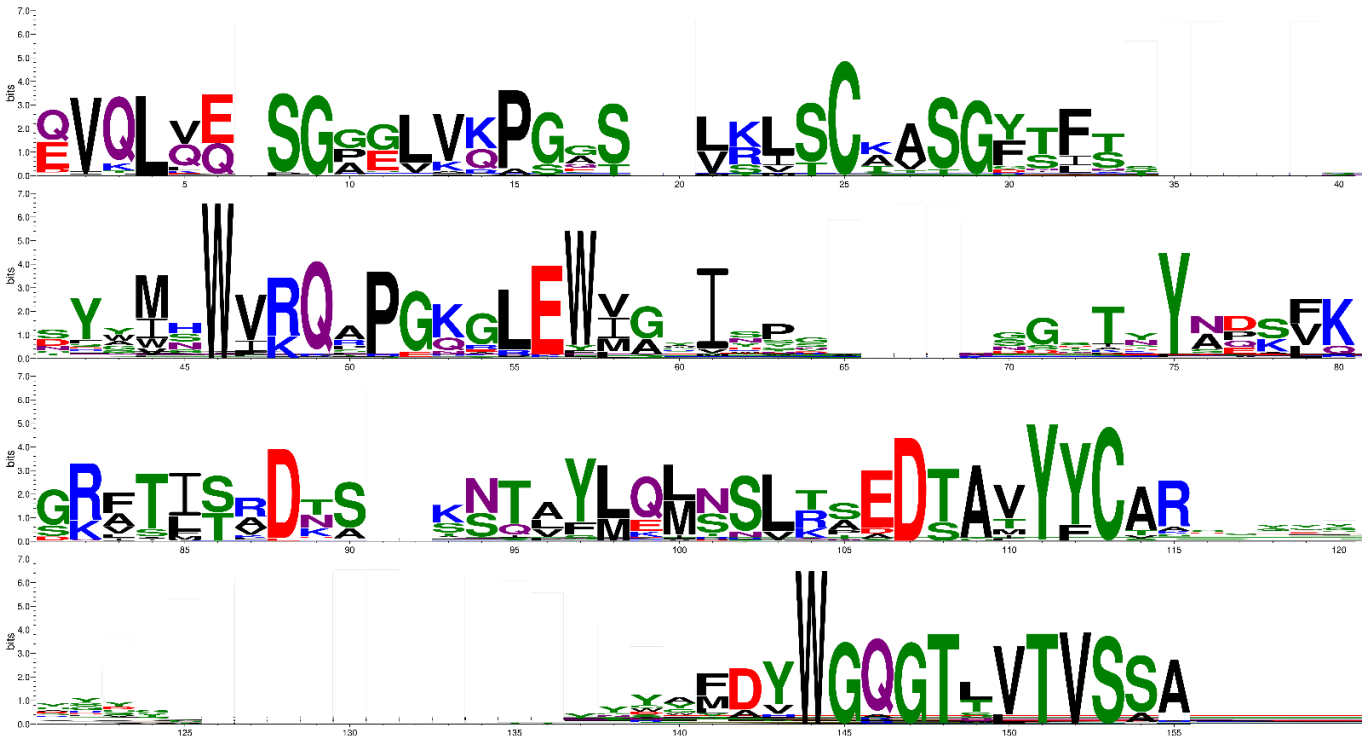
This table records the alignment score, similarity, and pairwise alignment of each sequence with the germline V genes. Abalign selects the V gene with the highest score to match with the sequence (if the score is the same, the similarity is compared). Besides, Abalign also allows users to search for the specific sequence and save the file.

Abundance: There are three options in this menu. "V Gene abundance" shows the top 20 V genes in abundance. "Sequence abundance" shows the top 20 sequences in abundance. "Region abundance" menu shows the abundance of top 20 sequences in a particular region.



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

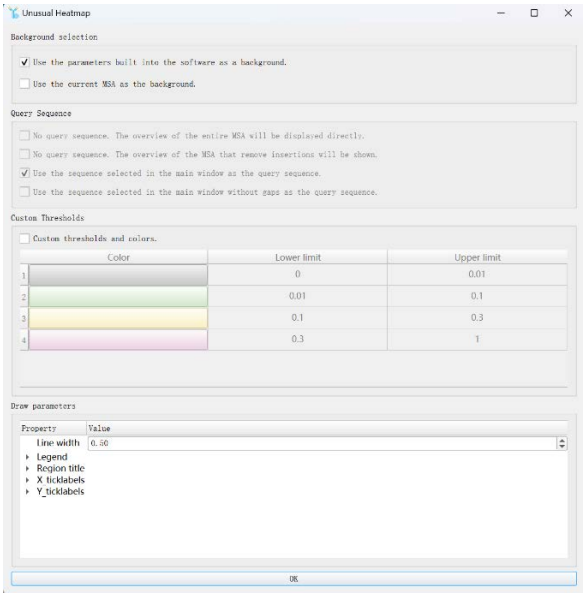
Seqlogo (only for linux): There are three options in this menu. "By Entropy" option takes the currently displayed MSA (if "Display by region" or "Display by gene" is used then the MSA will be changed accordingly) as input to generate a Seqlogo plot with entropy as the y-axis. "By Frequency" is used as input to generate a Seqlogo plot with frequency as the Y-axis for the currently displayed MSA. "Color" option changes the rendering mode of the Seqlogo plot residue colors.



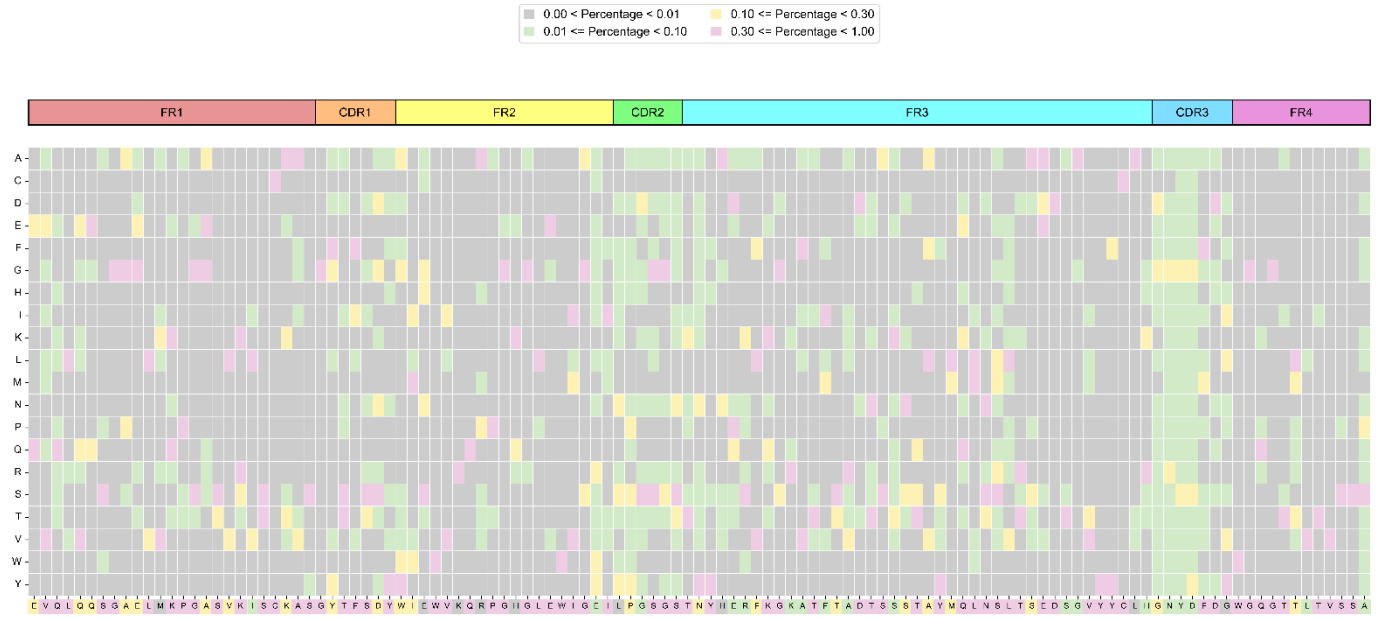
The ordinate indicates the entropy of amino acids, and the abscissa indicates the position in the variable domain. The larger the letter of the amino acid, the greater the entropy value.

Unusual Residue: This option shows the proportion of residues at different positions in the MSA. A residue ratio difference graph constructed from tens of millions of human sequences downloaded from OAS is built into the software and can be compared to this graph by

selecting the query sequence to aid in antibody humanization. In addition, users are also allowed to build a residue ratio difference graph with their own dataset for comparison.

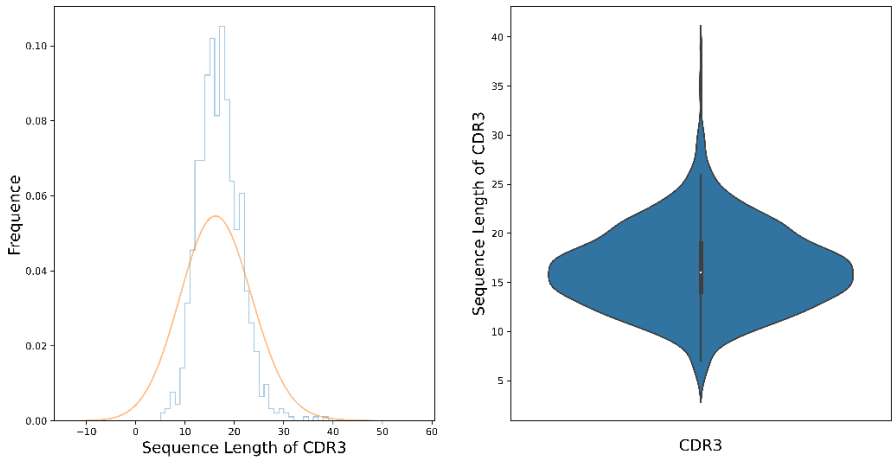


"Background selection" is used to select the background library which compares to the query sequence."Query Sequence" is used to determine whether to enter a query sequence and whether to remove gaps from the query sequence."Custom Thresholds" is used to adjust the thresholds of amino acid distribution and the colors they represent."Draw parameters" are used to modify drawing parameters.



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 being considered “Unusual Residue” (Marked in gray)by default.

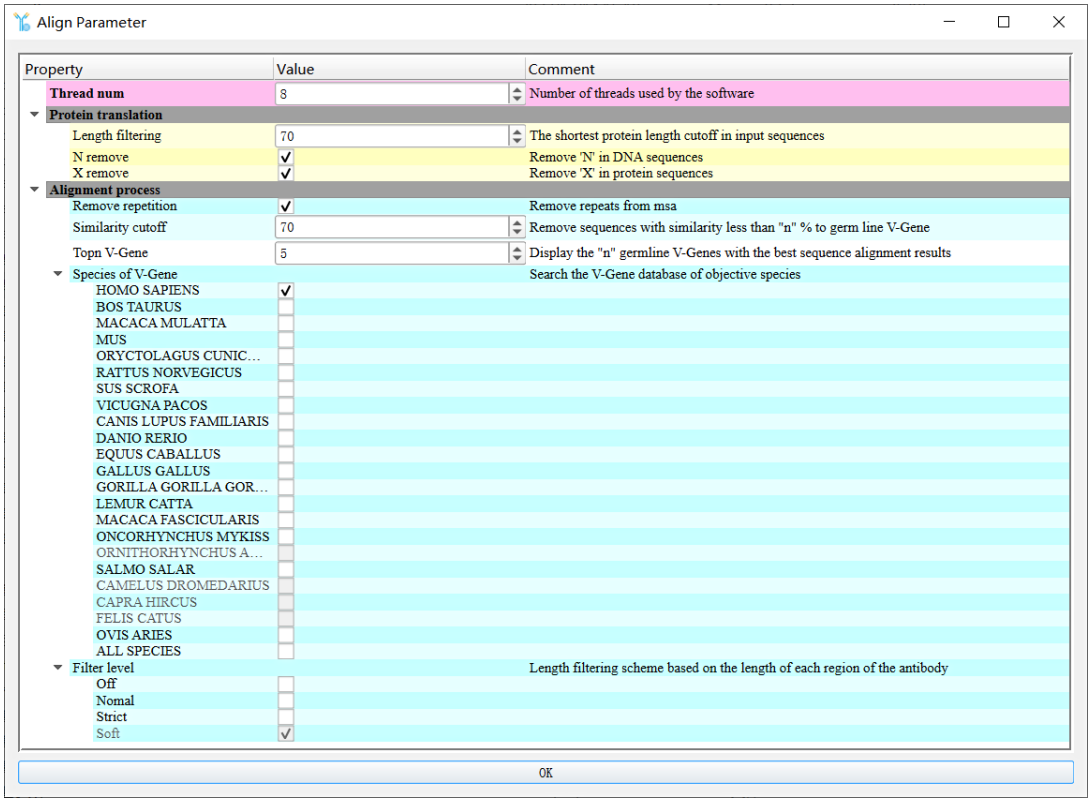
Length Distribution: There are eight options in this menu, for each of the seven regions of the antibody variable region and the full length of the antibody variable region. Click on the different options to see the distribution of the length of the different regions of the entire MSA.



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: Clicking on this option will bring up a parameter options window where you can adjust the parameters related to the alignment.



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

Temporary Path: Clicking on this option will bring up a dialog box where you can change the path of the temporary files exported by Abalign. Please restart the software after changing the temporary path.

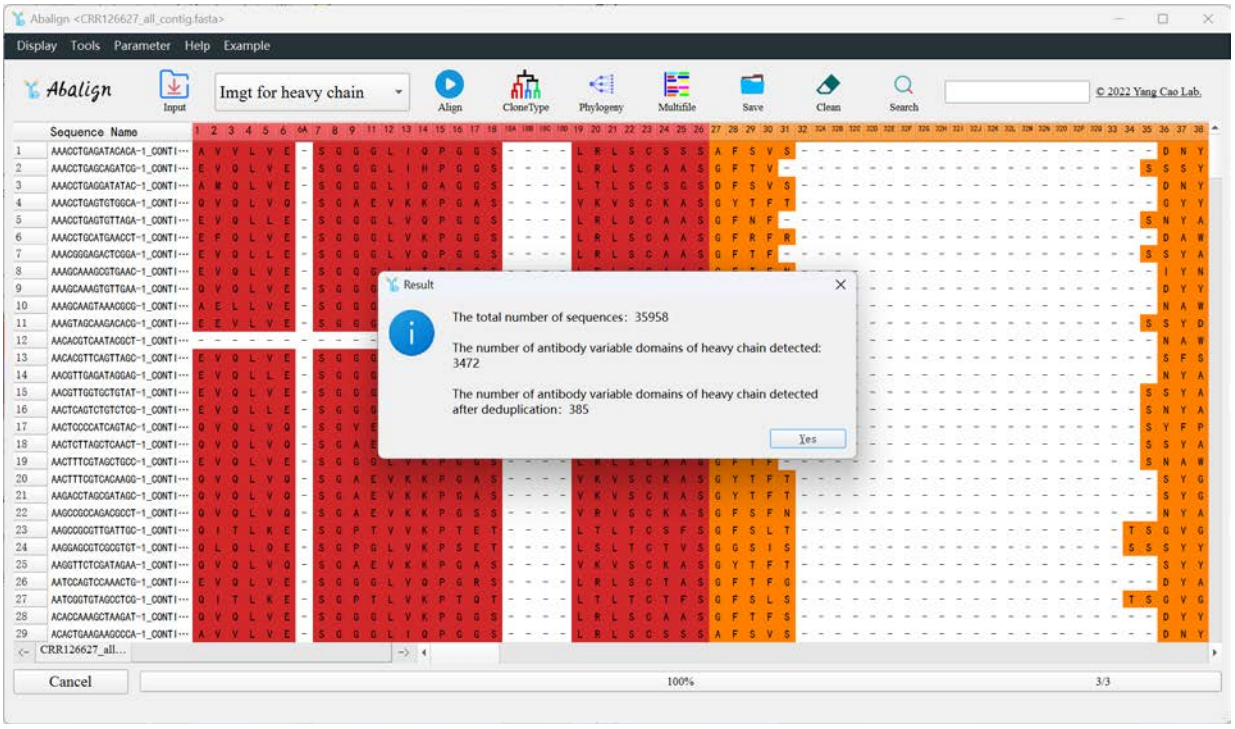
Example: Multiple antibody sequence files (DNA or amino acid) are provided for users to use, aiming to quickly familiarize users with Abalign. "Example" contains two options, which are used for single-file mode and multi-file mode respectively, and users can also find example files under the "example" folder in the installation path.

Usage case

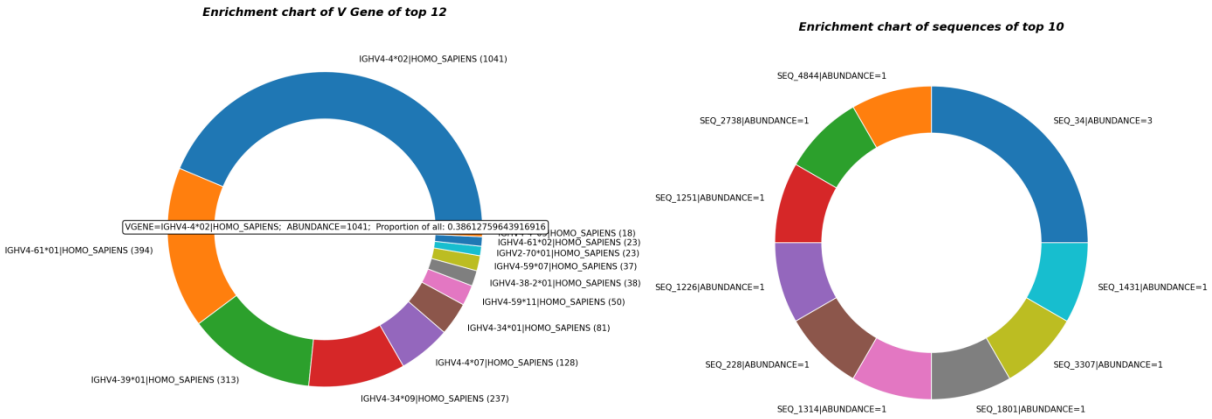
In this use case, we use the example in the software to demonstrate, you can click Example->SingleFile in the menu bar to load the example file, or select the input file through the Input button in the toolbar.

Step 1 Input file: select the fasta file that needs to be processed after clicking input.(For the example, we simply click on Example->SingleFile in the menu bar to load the file)

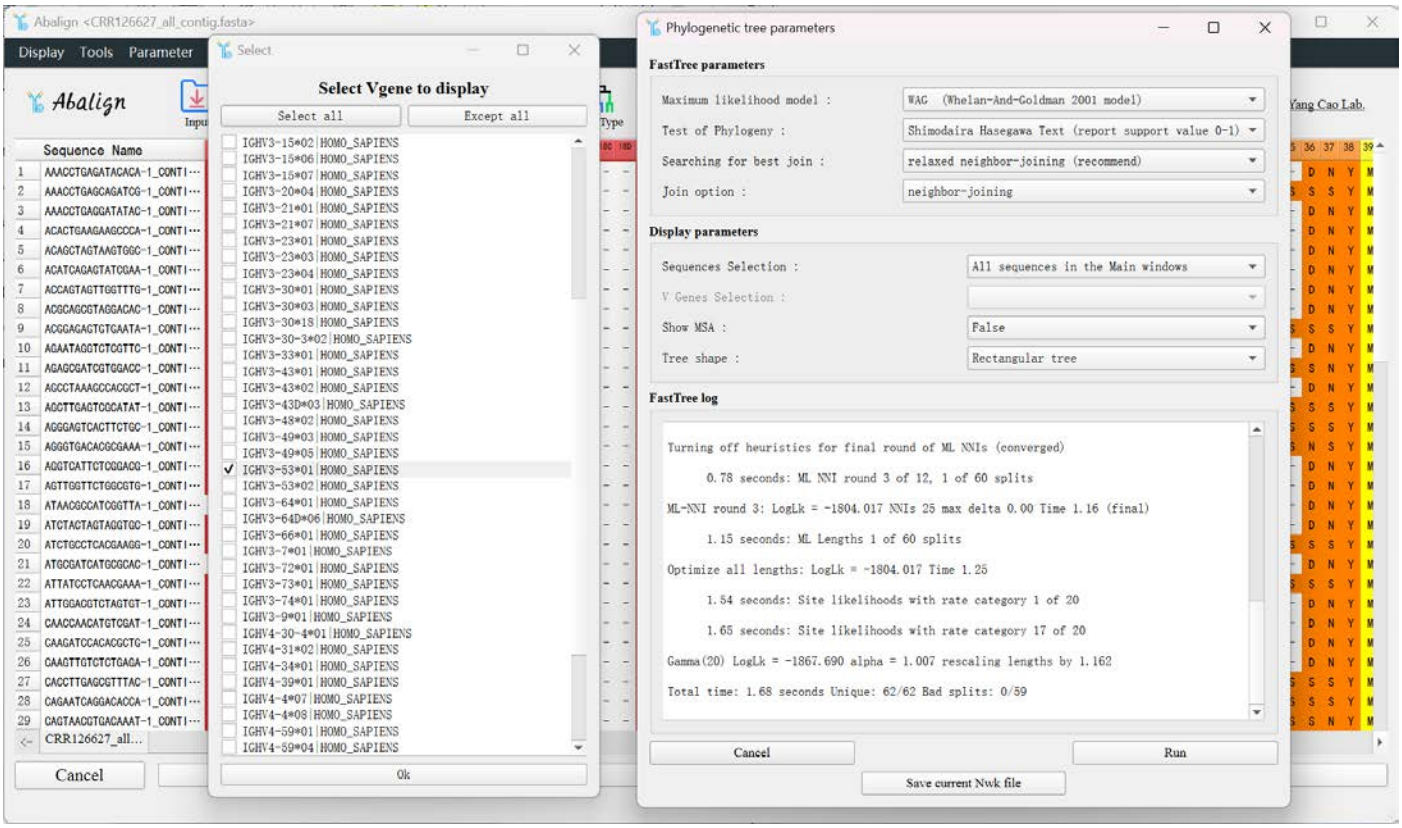
Step 2 Search for antibody variable domain and multiple sequence alignment: After file loading is completed, click “Align” to run the program, the progress bar below the program will show the progress of the program, if the file is too large, the progress bar is not updated in a short time is normal. After the comparison, a dialog box will be popped, and the total number of input sequences will be displayed in the dialog box. The number of sequences in the variable region of the antibody and the number of sequences after deduplication will be detected.



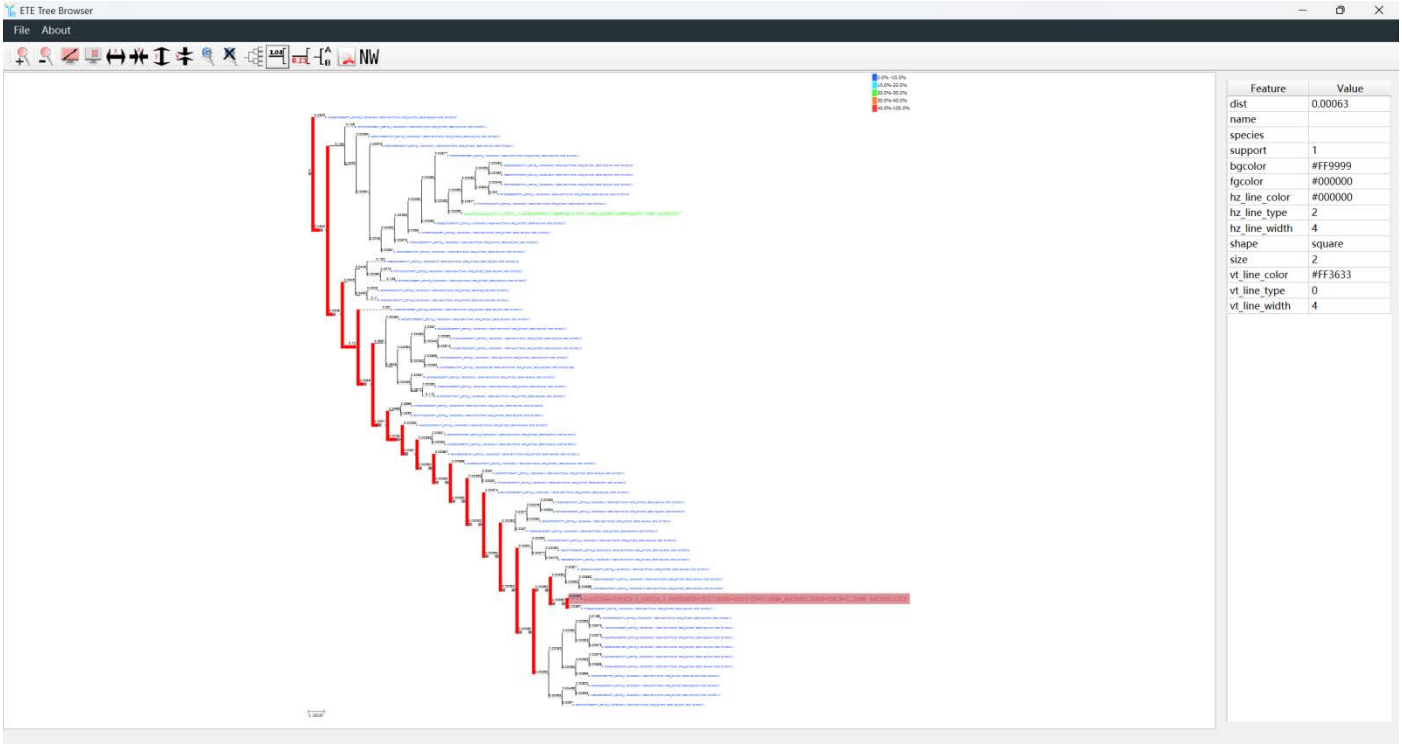
Step 3 looks at multiple sequence alignments: click the “Tools” button in the top menu bar and move the mouse to “Abundance”. Click “V Gene abundance” or “Sequence abundance” to obtain the abundance map of V Gene and sequence. Somatic hypermutation occur during antibody maturation and the final mature antibody is highly expressed in the body. Therefore, it is meaningful to study antibodies with high abundance.



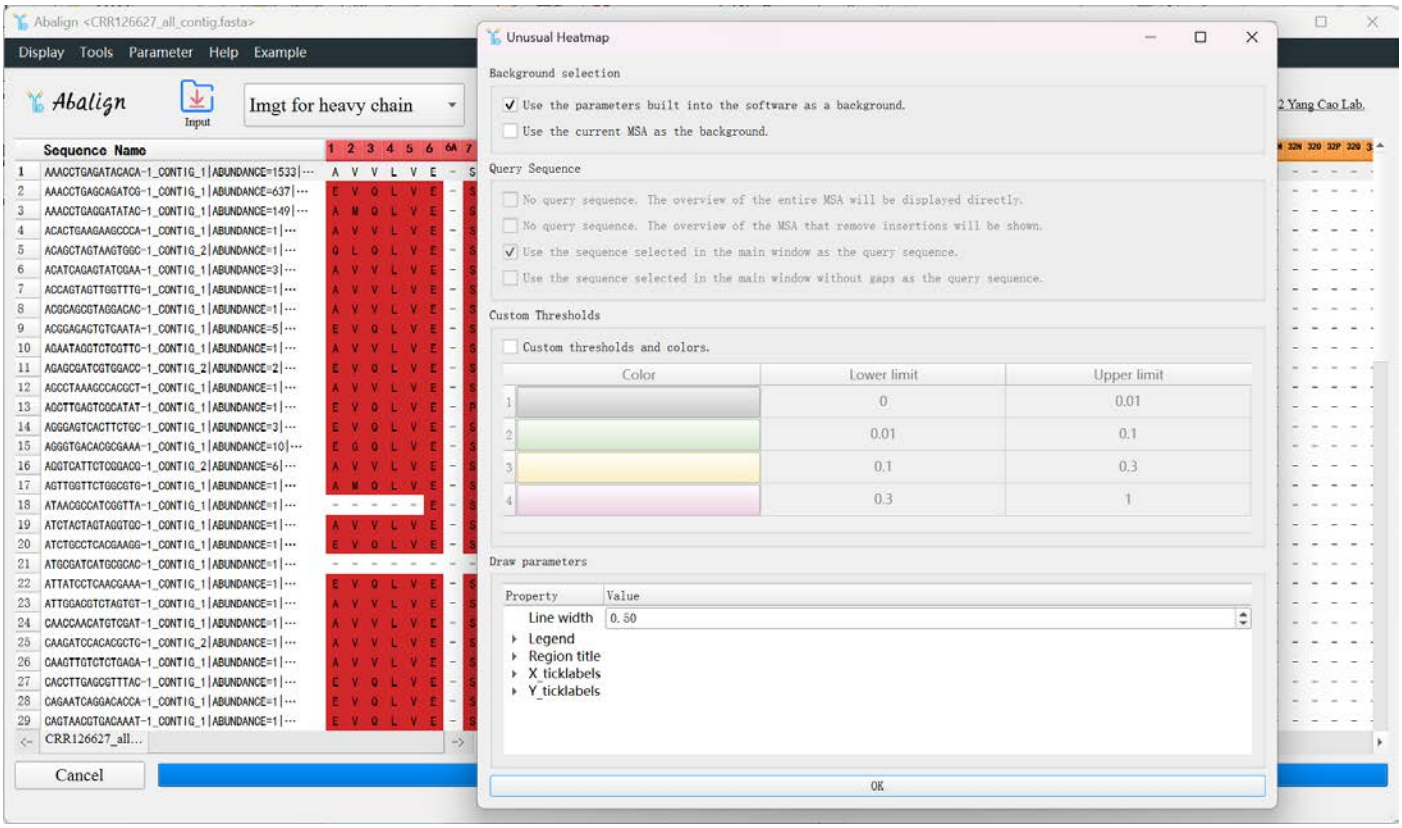
Step 4 Build phylogenetic tree: according to the sequence abundance information, build the tree with the V Gene of the high abundance sequence. In this case, the most abundant sequence belongs to V Gene “IGVH3-53 * 01”. We can click Display->Display by Genes ->Display by Vgene in the menu bar to select the V Gene “IGVH3-53*01”, and then click the “Run” button in the Phylogenetic tree parameters window to build the phylogenetic tree of sequences belonging to this V Gene. If you are satisfied with the tree building results, click the “Save the current Nwk file” button in the tree building menu to save the .nwk file of the current system tree.



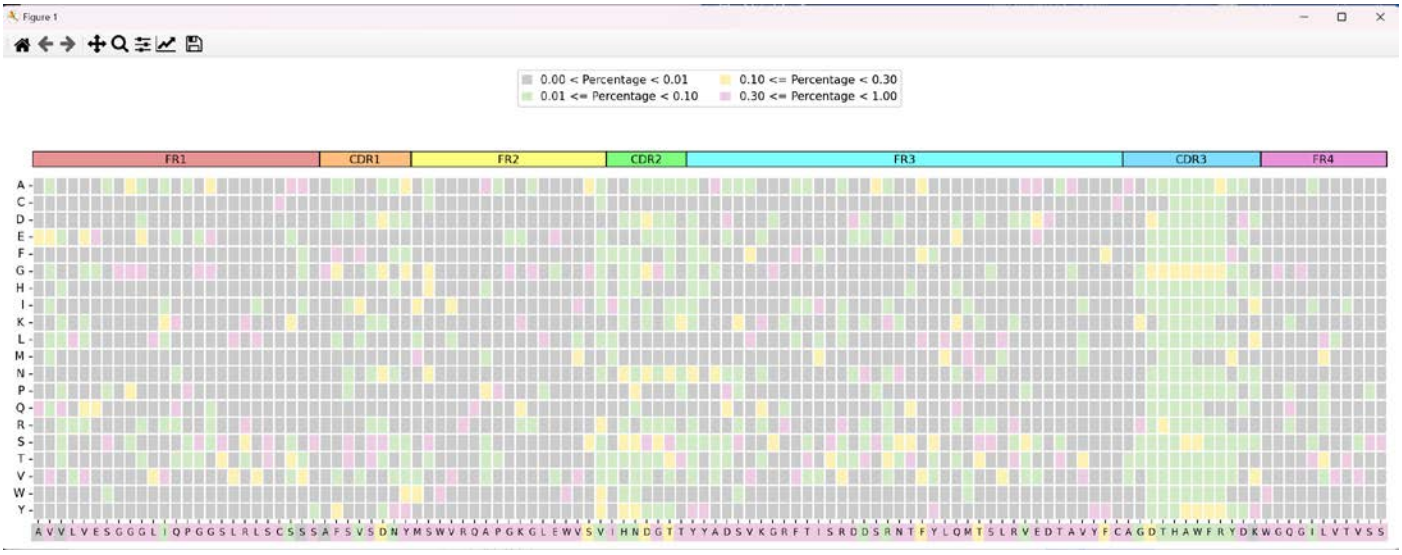
Step 5 View the phylogenetic tree: Once the evolution tree is built, a visual window pops up that adjusts the view and displays other information about the tree through the button above the visual window. After selecting the node of the tree, you can also modify the attributes of the node, such as the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



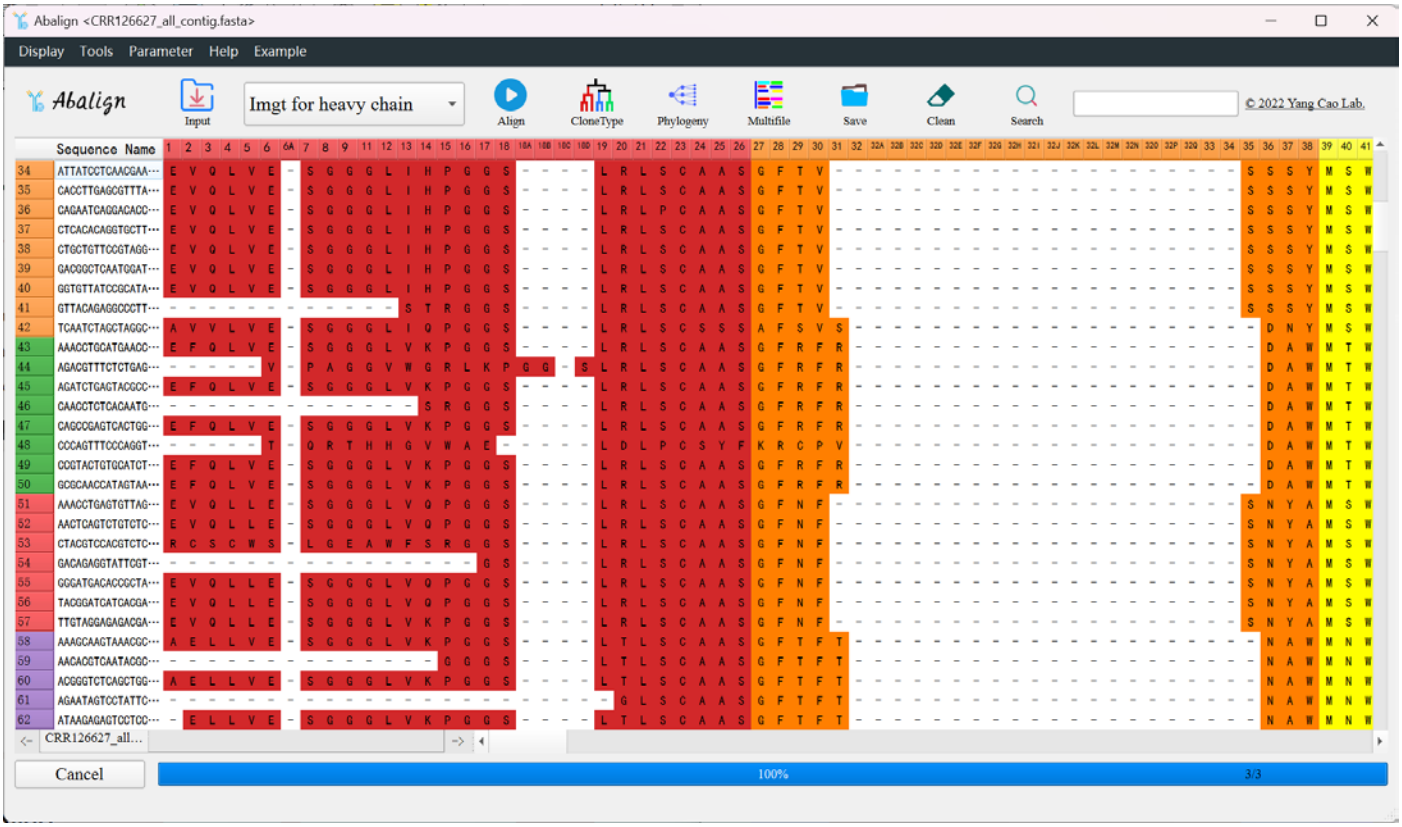
Step 7 Antibody Humanization Assist: Through the above analysis, we can clearly find that certain sequences have higher sequence abundance and a certain evolutionary trend. If we want to know the frequency of residues corresponding to these sequences in human antibodies, we can click Tools->Unusual Residue to get insights. In the results, residues of different frequencies will be displayed in different colors, and users can customize the threshold and color.



By default, low frequency residues will be shown in gray, if such residues are present in the FRs of the query sequence, consider changing them to the corresponding high frequency residues for antibody humanization

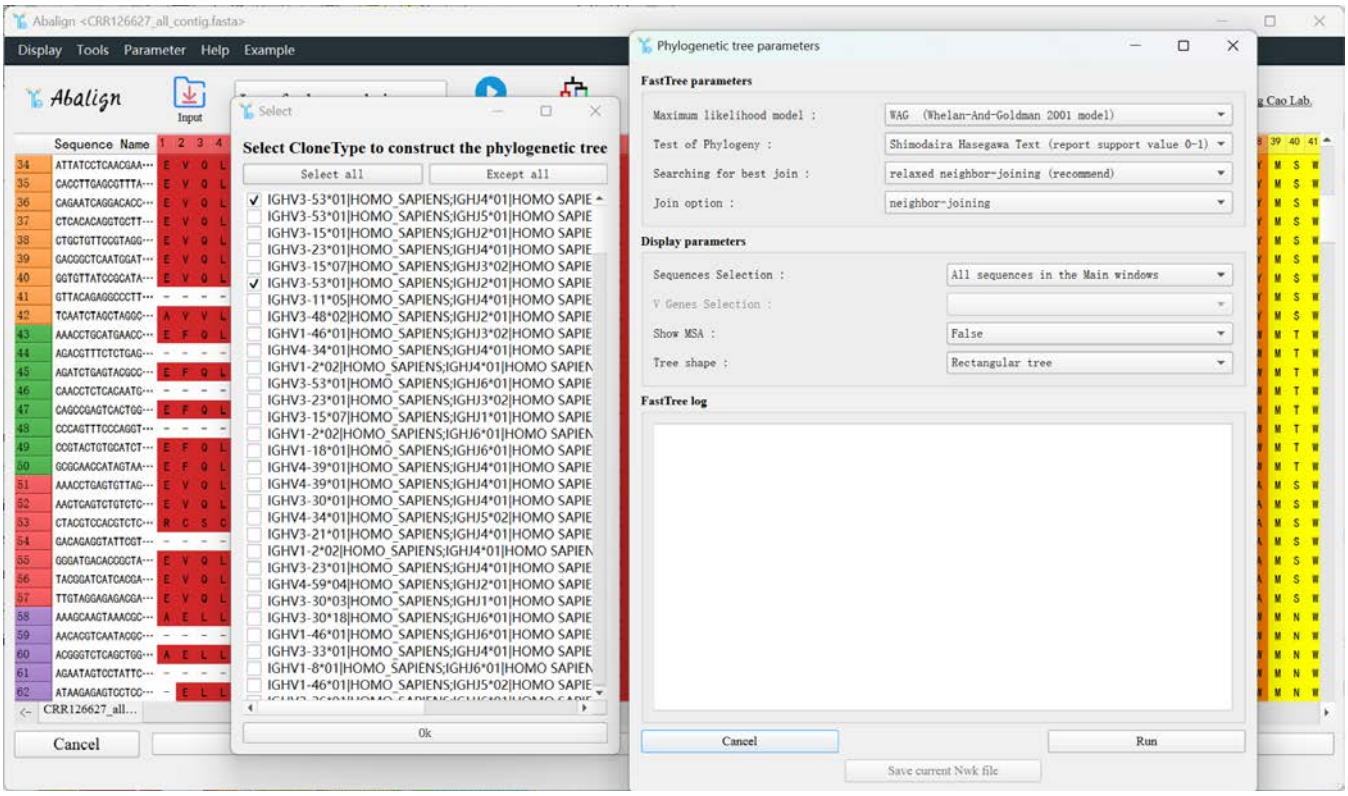


Step 8 Clonotype Matching: Once you click the "Clonetype" button in the toolbar, you can sort the sequences by clonetype, and sequences with the same clonetype will be lined up together and the sequence names will be rendered with the same color.



Step 9 Build of phylogenetic tree by clonetype: Once the Clonetype button has been clicked, we can display the sequences belonging to a specific clonetype individually in the

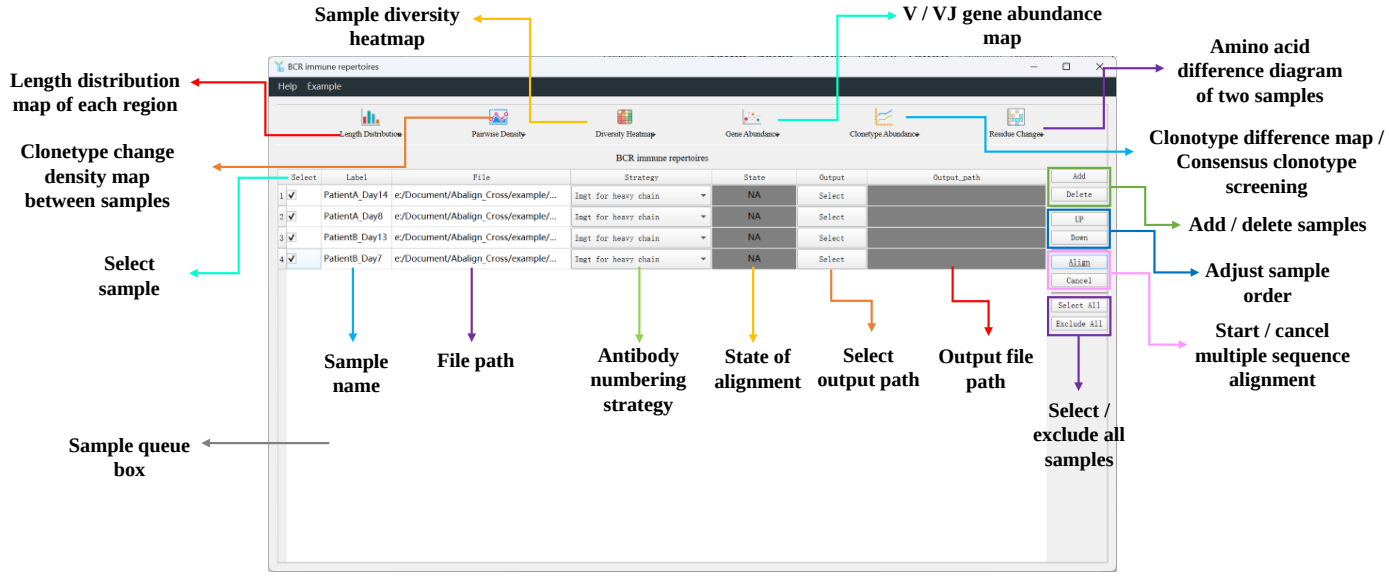
main window. Then, a phylogenetic tree can be constructed with these sequences individually.



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 Clonetype, in addition to saving all sequences, we can personalize and save the sequences we need.

Multi-File Navigator

Navigator layout



Usage

Add file: Click "Add" to select the target file, then the corresponding file will be displayed in the main window, and the filename can be modified by changing the value of "Label".

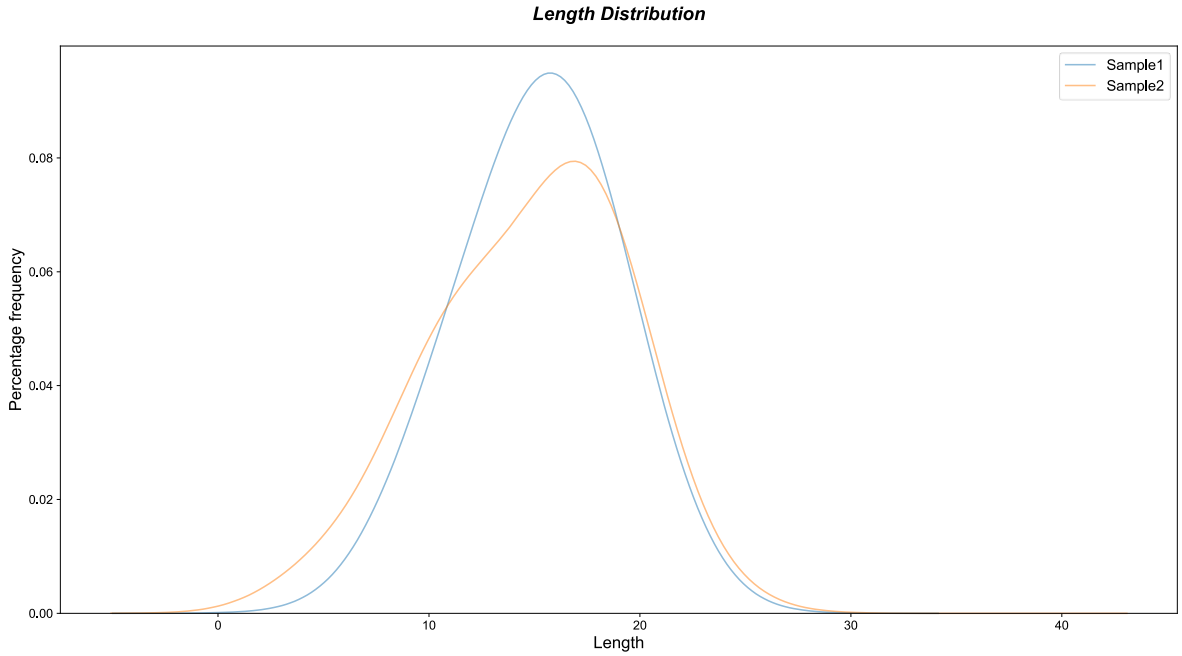
Delete file: Check the file in "Select", then click "Delete".

Multiple sequence alignment: Check the file in "Select", 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 "Select" in "Output" to determine the output path.

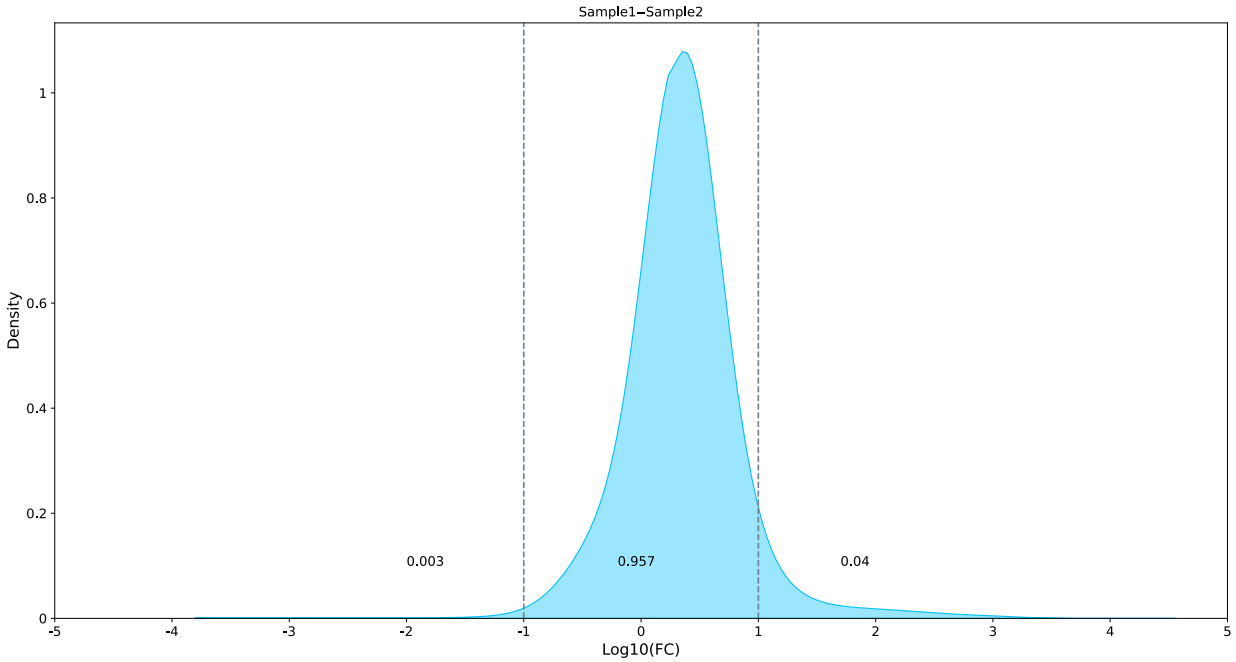
Tool buttons

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



The abscissa represents the length of the sequence, the ordinate represents the proportion of the certain length sequences, and the lines of different colors represent different samples. The plot data can be saved by clicking “Save Sources”.

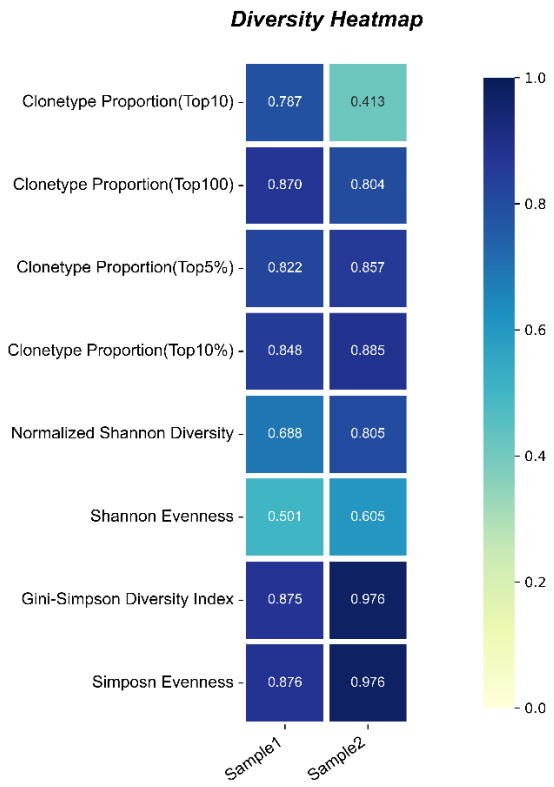
Pairwise Density: Count the changes of clonetype between two selected samples and draw the density map.



The abscissa represents the base 10 logarithm of the clonotype fold change. After removing the low abundance clonetypes, clonotypes are divided into different groups according to log10(FC). The number represents the proportion of the clonetype contained in each group. The plot data can be saved by clicking “Save Sources”.

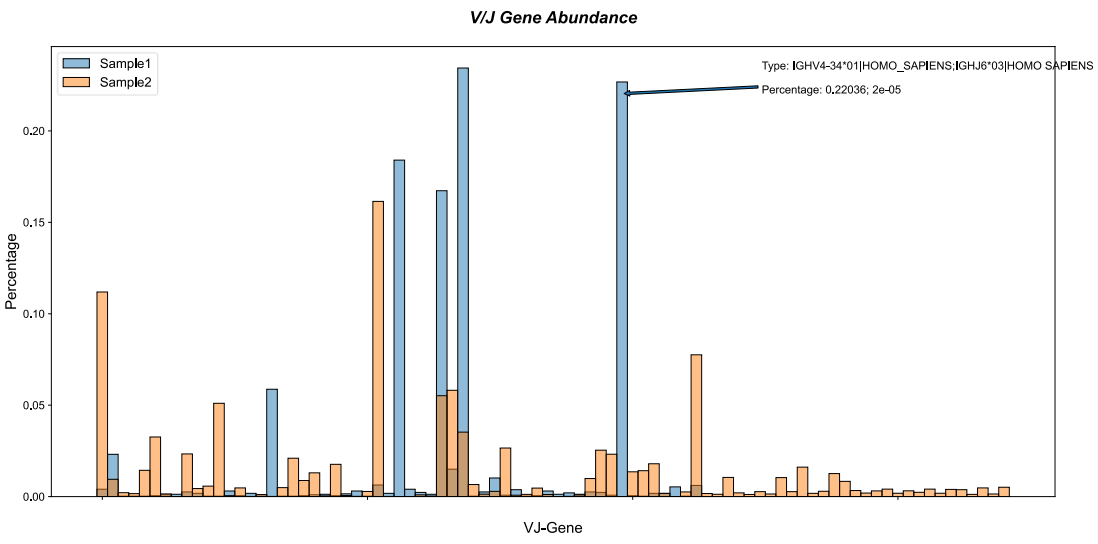
Diversity Heatmap: Calculate the diversity index of clonotypes among selected samples and draw the heatmap.

Diversity Heatmap: Calculate the diversity index of each antibody region for selected samples and draw the heatmap. The region includes FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4, full length, CDR1-FR4(FR1 region will be incomplete when the sequencing data quality is poor).

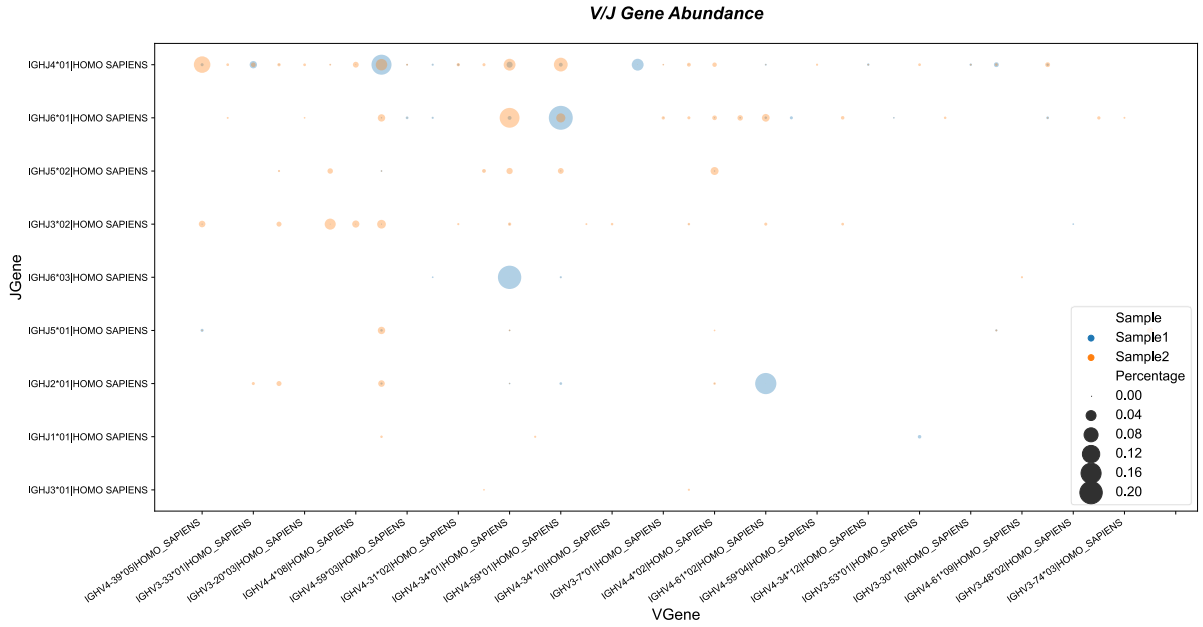


The abscissa represents different samples, the ordinate represents the diversity index, with the more biased dark blue indicating a larger value and the more biased light green indicating a smaller value. The plot data can be saved by clicking “Save Sources”.

Vgene Abundance: Count the V/VJ gene abundance for selected samples and draw the histogram / scatter plot.

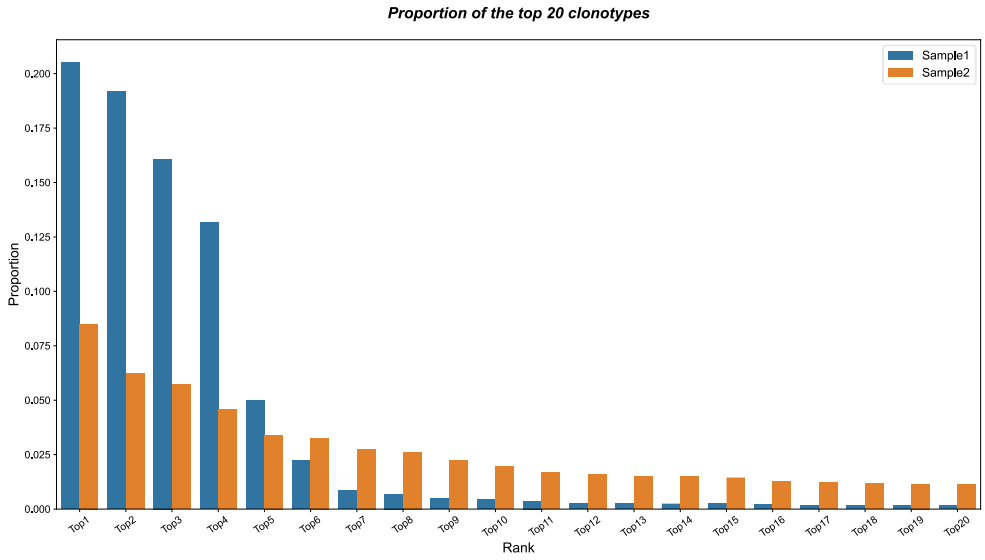


The abscissa represents the type of V/VJ gene, the ordinate represents the proportion of the specific gene, and the different colored bars represent different samples. Details are displayed when hovering.

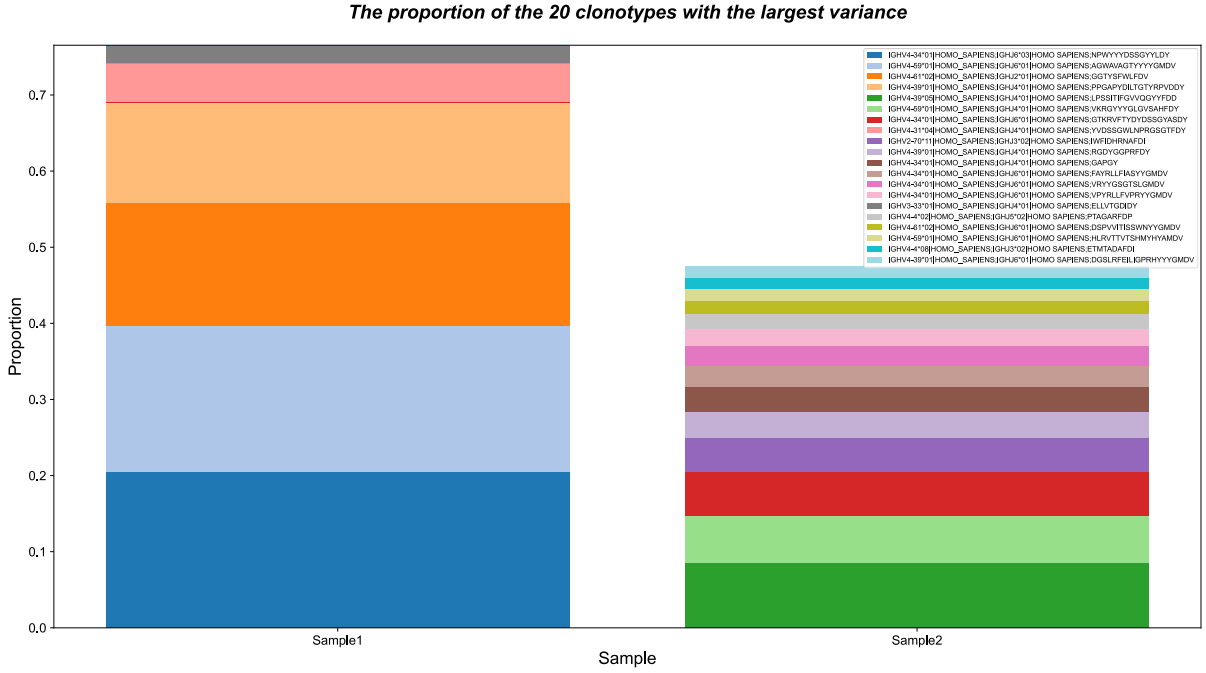


The abscissa represents the type of V gene, the ordinate represents the type of J gene, points of different colors represent different samples, the size of points represents the proportion of the specific VJ gene combination. Details are displayed when hovering.

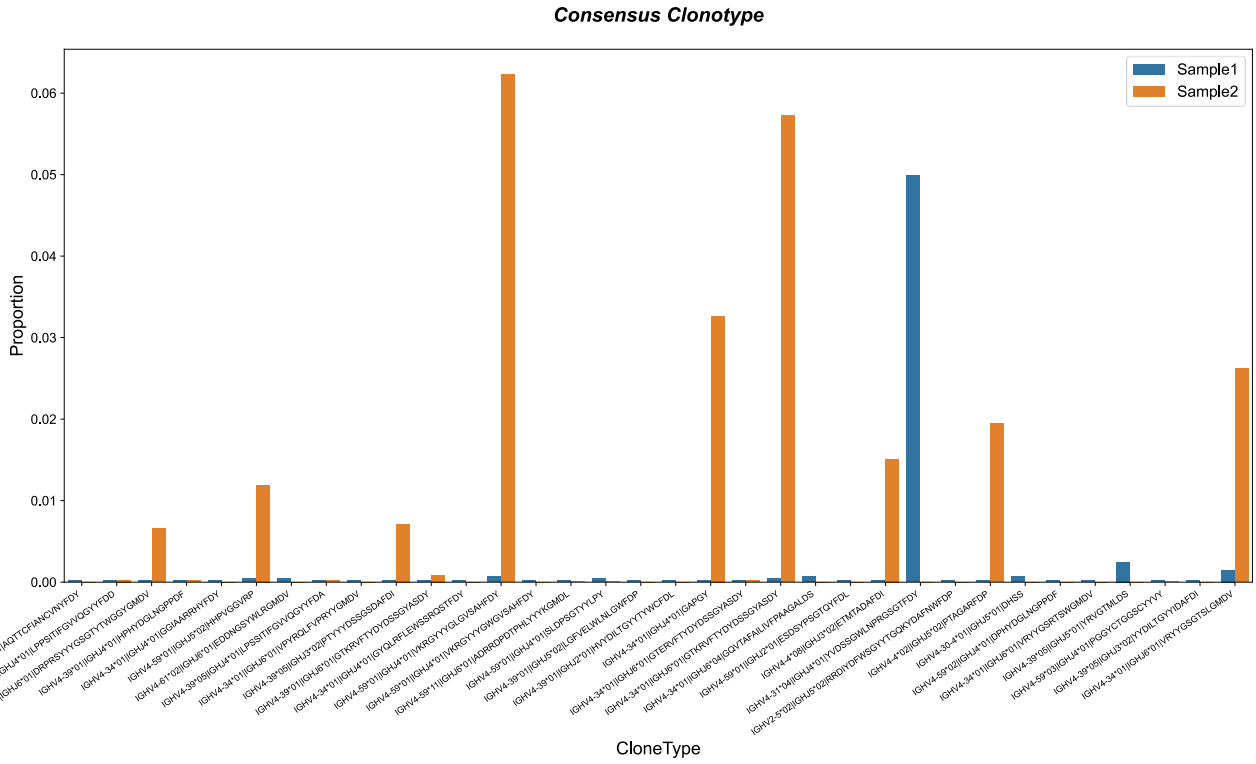
Clonetype Abundance: Count the clonetype abundances for selected samples and draw the distribution bar plot / difference bar plot, and count the common clonotypes among different samples or patients.



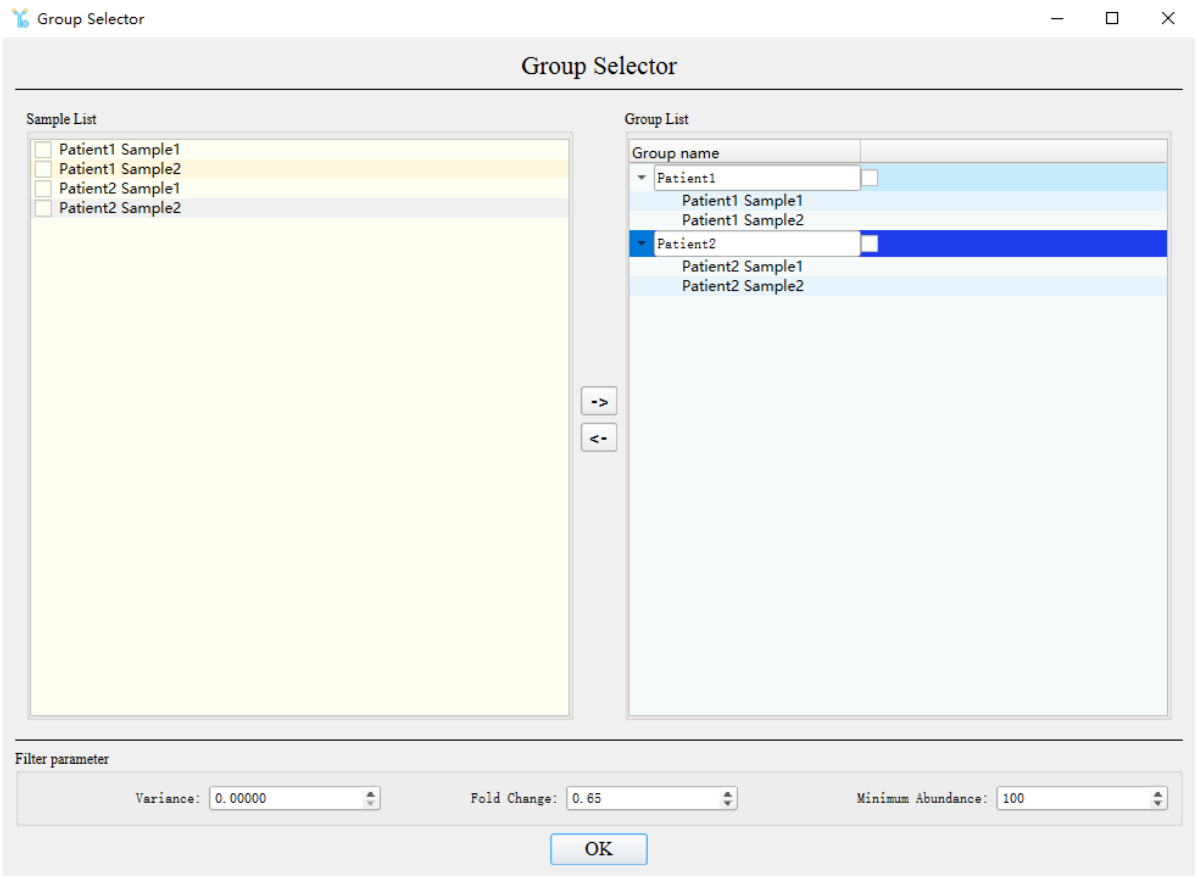
The abscissa represents the top 20 clonotypes, the ordinate represents the proportion of clonotypes, and bars in different colors represent different samples. See output file (clonetype.csv / clonetype_seqs.csv) for clonotype detail.



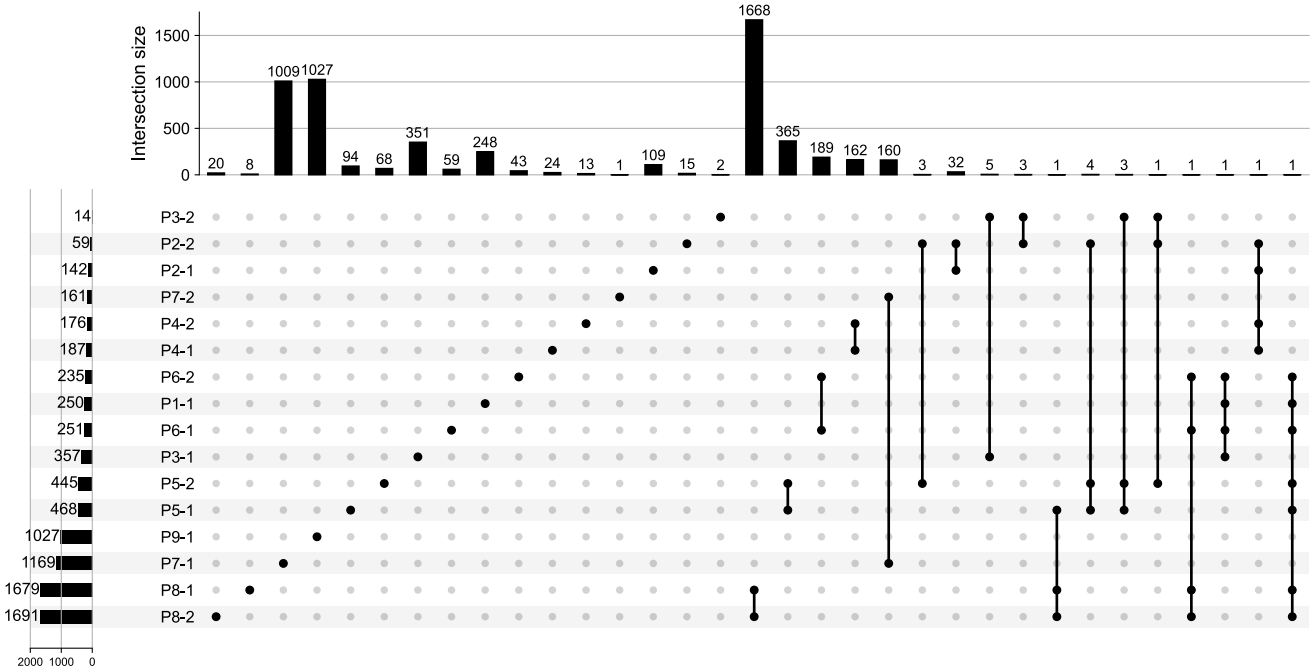
The abscissa represents different samples, the ordinate represents the proportion of the top 20 clonotypes that vary across selected samples. See output file (clonetype_changes.csv) for all clonotypes differences. The plot data can be saved by clicking “Save Sources”.



The abscissa represents the specific clonotype, the ordinate represents the proportion of clonotypes, and bars in different colors represent different samples. See output file (consensus_clonetype.csv) for consensus clonotype detail. The plot data can be saved by clicking “Save Sources”.

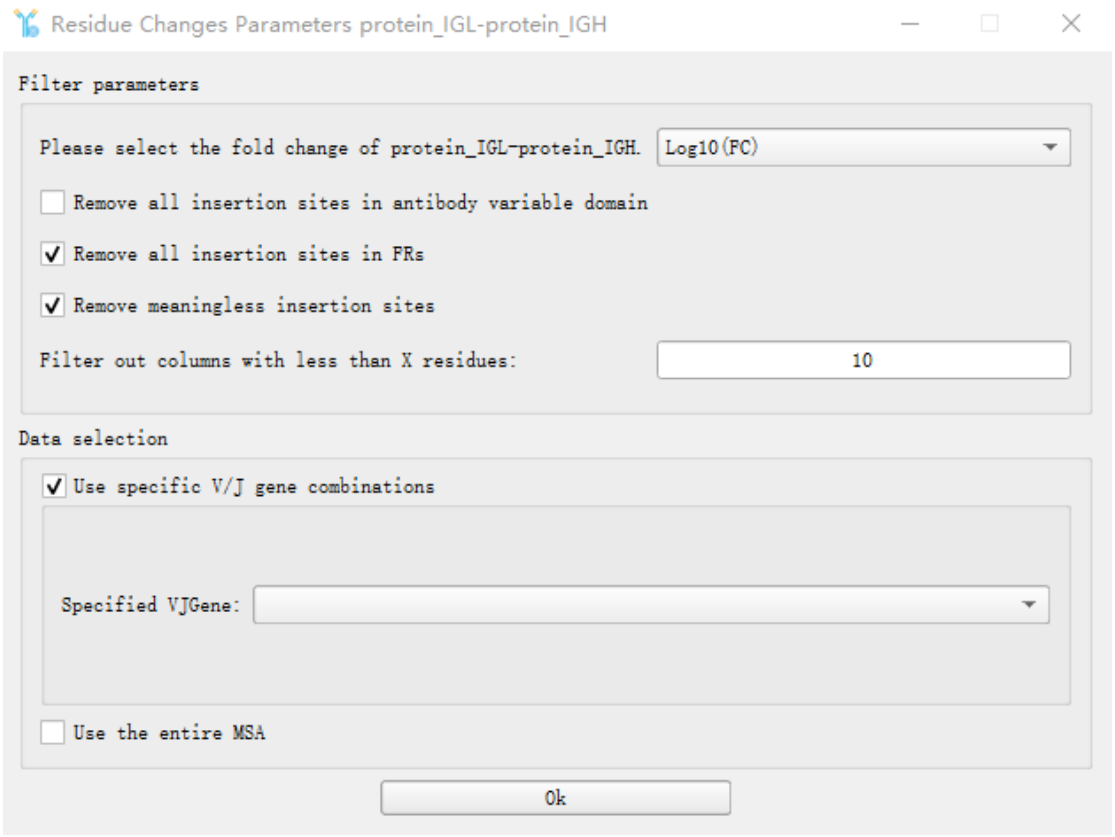


"Sample List" shows all samples to be analyzed, selected through the multi-file navigator. "Group List" is used to display the groups constructed by users, which can be generated by selecting the samples on the left and clicking the move button in the middle. Conversely, ticking the right group and clicking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

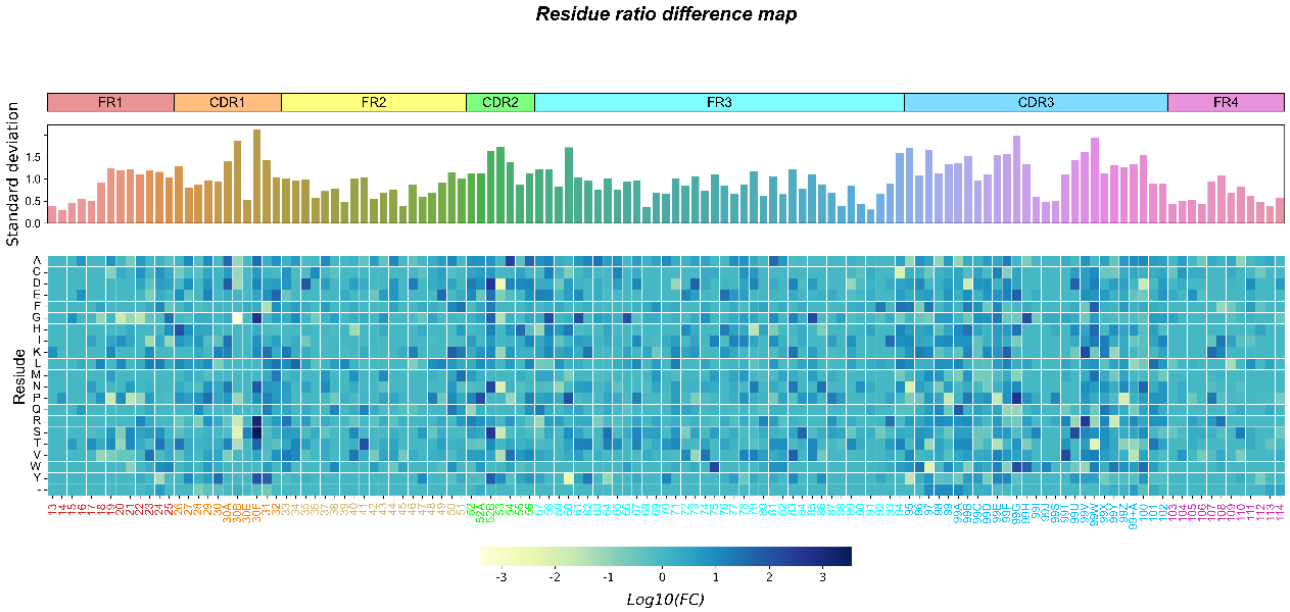


The left bar represents the number of clonotypes expanded in each patient, and the upper bar represents the number of shared clonotypes among different patients. The patient information corresponding to the shared clonotypes is shown as the line below, and the dot on the line represent the patient. It should be noted that if the clonotype appears among multiple patients (eg: Patient A+ Patient B+ Patient C), the clonotype will not be taken into consideration when counting the shared clonotypes among part of the patients (eg: PatientA+ PatientB \ PatientB+ PatientC \ PatientA+ PatientC). See output file (consensus_clonetype_between_groups.csv) for consensus clonotype detail. The plot data can be saved by clicking “Save Sources”.

Residue Changes: Count the changes of amino acids between the two samples and draw a heatmap.



"Filter parameters" are used to control the amino acid difference positions to be displayed. Users can choose the fold change, whether to retain insertion positions or nonsense positions (the position with residue occurrences lower than X). "Data selection" is used to control the sequences for analysis, allowing the user to select the entire sample or a specific VJ gene combination (the common VJ gene combination required within two samples). If you get an error message, make sure you select the same sample numbering strategy and chain type.



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. The ordinate of the lower figure represents 20 amino acids and gaps. The histogram above represents the standard deviation of the difference in the proportion of residues at each position. If the standard deviation at a certain position is larger, it means that the diversity of the position residues is more abundant. 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 allows the user to save plot data via "Save Sources".

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] 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.

[3] 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.