Abalign Manual

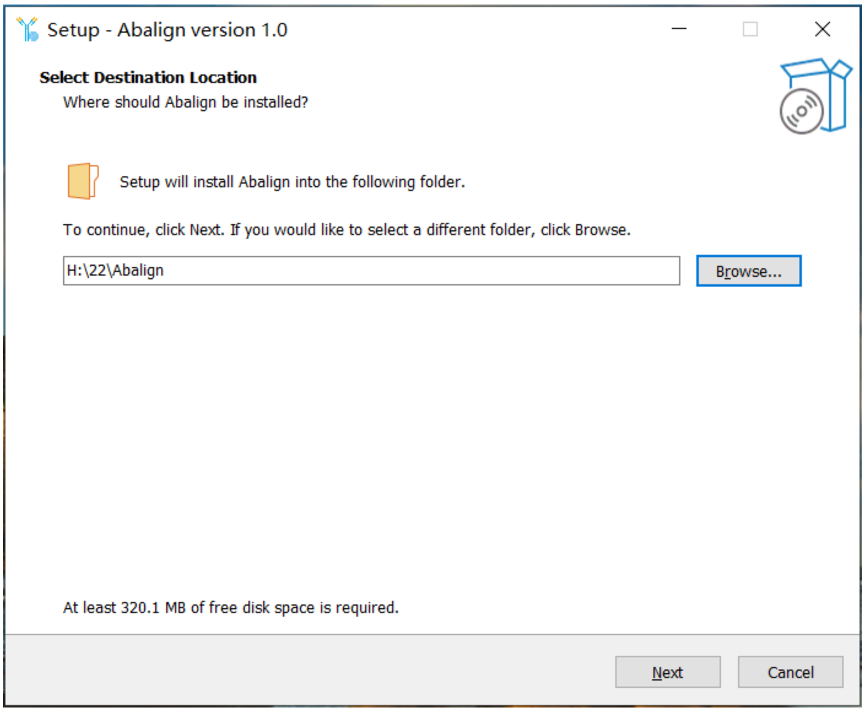
- Notice
- Reference

Introduce

Multiple sequence alignment has long been used as a powerful tool 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 prediction methods. The existing multiple sequence alignment methods are extremely difficult to align the highly variable regions of antibody or B-Cell Receptor (BCR) sequences without the prior knowledge of antibody gene recombination and hypermutation in the process of maturation. This multiple sequence alignment tool, named Abalign, which integrates heuristic knowledge of the standardized antibody sequence numbering schemes, including IMGT, KABAT and Chothia systems. The alignment follows the well-characterized patterns of the conserved or highly variable positions known by immunology studies. Hence the alignment result is consistent with the structural and immunological knowledge. Abalign was implemented in a user-friendly software with interactive and visual interface, which supports the multiple sequence alignment, as well as sequence clustering, antibody numbering, CDR delimiting, constructing phylogenetic tree, sequence similarity heat map, V-gene determination and abundance analysis by just clicking the buttons. Abalign allows the high-throughput analysis for BCR sequencing data, which can be finished in 850 minutes for 1 Gb DNA fasta sequences through a single thread by a notebook equipped with I7 10870H. Abalign is a powerful and efficient tool for biological researchers to analyze massive amounts of BCR or antibody sequences and get new discovery in immunoinformatic study.

Setup (for windows)

Double click **Align Setup.exe** to enter the installation, click **Browse** to select the installation path, and click **Next**.



Select Destination Location

Where should Abalign be installed?

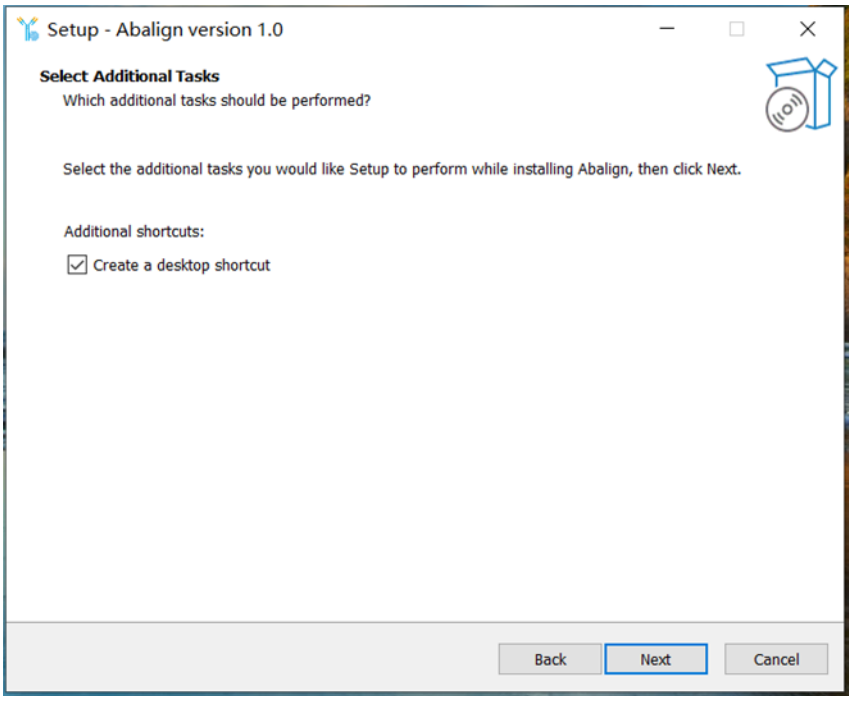
Setup will install Abalign into the following folder.

To continue, click Next. If you would like to select a different folder, click Browse.

H:\22\Abalign

At least 320.1 MB of free disk space is required.

Select whether to **Create a desktop shortcut**, click **Next**, and click **Install** to install.



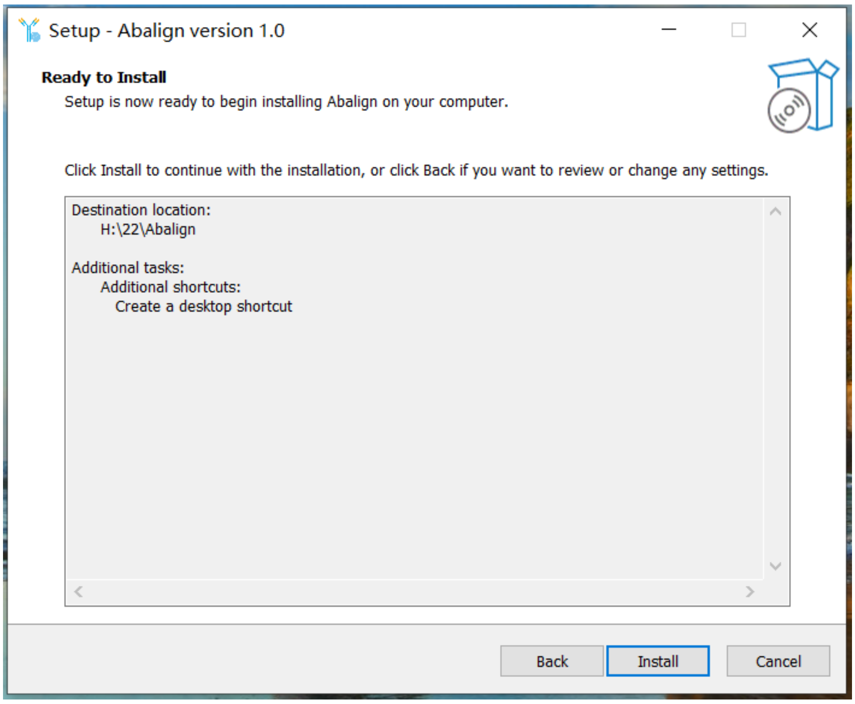
Select Additional Tasks

Which additional tasks should be performed?

Select the additional tasks you would like Setup to perform while installing Abalign, then click Next.

Additional shortcuts:

☑ Create a desktop shortcut



Ready to Install

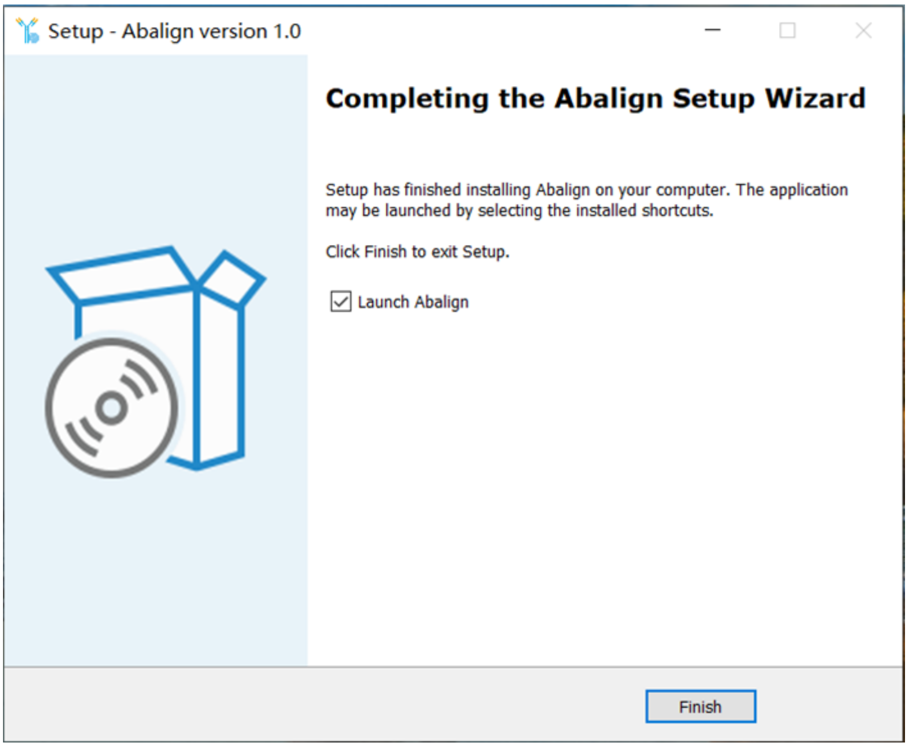
Setup is now ready to begin installing Abalign on your computer.

Click Install to continue with the installation, or click Back if you want to review or change any settings.

Destination location:
H:\22\Abalign

Additional tasks:
Additional shortcuts:
Create a desktop shortcut

Please wait for the installation to complete, select whether to **Launch Abalign**, and click **Finish** to exit the installation program.



Completing the Abalign Setup Wizard

Setup has finished installing Abalign on your computer. The application may be launched by selecting the installed shortcuts.

Click Finish to exit Setup.

☑ Launch Abalign

If you need to uninstall the software, click **unins000.exe** under the installation folder or uninstall or change the program through windows to complete the uninstall.

Usage

Software constitution

Note: Do not modify any content of lib folder

Executable file (UI): The file named Abalign_ui in the software root directory is the visual window program, double-click to run.

lib folder: Do not make any changes to the lib folder under the software directory, which contains a series of configuration files and binary files that implement multiple sequence alignment.

example folder: The sequence files of DNA and amino acids in FASTA format are stored for user testing.

Basic functions

File input: The program supports the amino acid sequence file or nucleic acid sequence file in FASTA format. The input file can be selected by clicking the **Input** button or dragged directly into the text box to input the file.

Numbering strategy: Select the antibody numbering strategy (support Chothia, Kabat, IMGT strategy) and the type of antibody sequence by clicking the "**IMGT for heavy chain**" drop-down menu (select the heavy chain will detect all the antibody heavy chains in the input sequence and filter out the light chain and the non-variable region sequence, the light chain is the same)

Multiple sequence alignment: Click the **Align** button to find the variable domain of the antibody and use sequences to execute multiple sequence alignment. During the alignment, the V Gene and the species that are most similar to each variable region sequence are found, and the V Gene name and species name are displayed after the corresponding sequence name. After running, the residue number of each site will be displayed above, and the seven regions of the antibody variable domain (FR1, CDR1... CDR3, FR4) are rendered into different colors by default. If you need to change the rendering method, please click **Display- > Sequence render mode**.

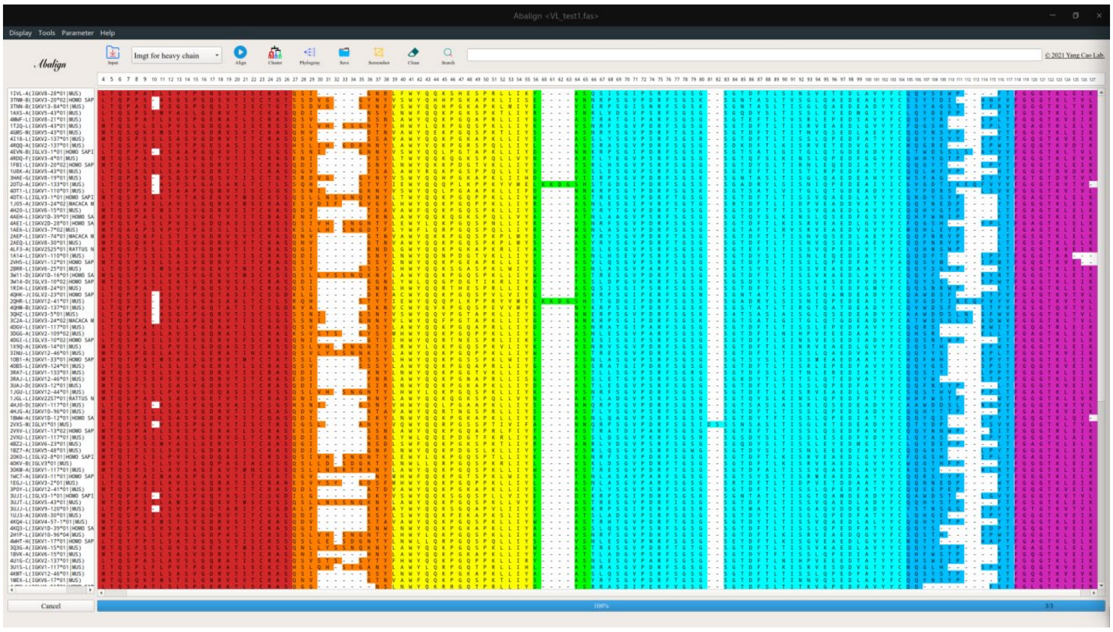


Figure 1. Multiple sequence alignment results.

File Save: Click the **Save** button to save the contents of the current interface as .fasta file and .temp file. The former stored the multiple sequence alignment results and the latter stored the multiple sequence alignment results with '*' as a separation of the various regions in the variable region.

Terminate process: Click the **Cancel** button to terminate the ongoing multiple sequence alignment.

Optional functions

Note: All optional functions require multi-sequence alignment first.

Hierarchical clustering: Click the **Cluster** button to run hierarchical clustering. This function will hierarchically cluster the multiple sequence alignment data after Align. After the clustering, the hierarchical clustering tree will be displayed. The sequence will be rearranged according to the tree, and the same sequence will be arranged adjacently, and the sequence name will be rendered as the same color. The parameters of hierarchical clustering can be changed by the **Parameter** option in the menu bar.

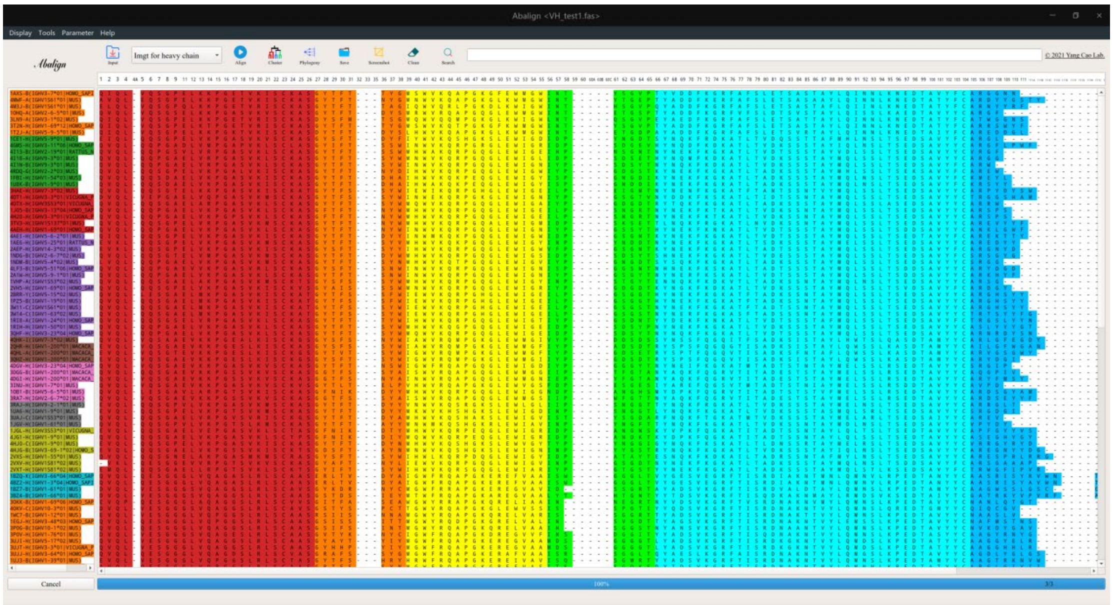
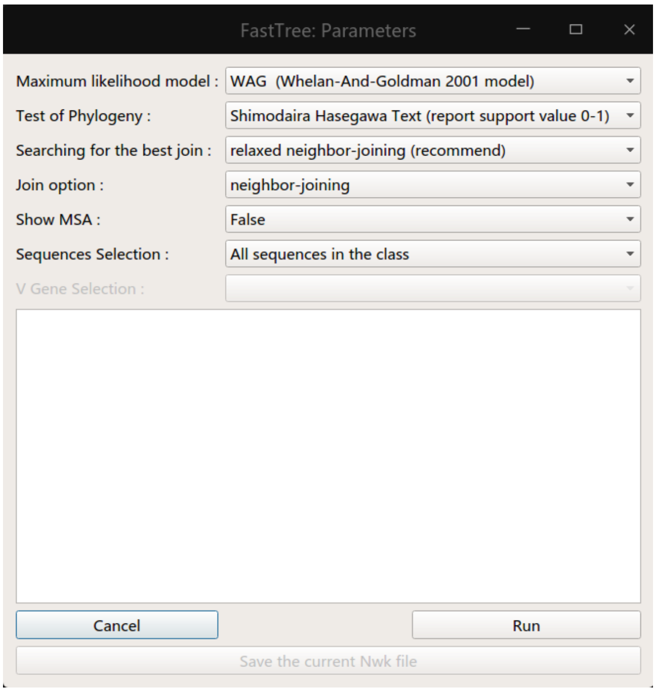


Fig. 2. Hierarchical clustering results of antibody variable domain

Phylogenetic tree: This function uses FastTree software (maximum likelihood method)[1] to build .nwk file and visualize it with Ete3. After clicking **Phylogeny**, the parameter box of the phylogenetic tree will be popped up. After clicking **Run**, the phylogenetic tree will be constructed. The text box below will enter the log information. Visualization results of the phylogenetic tree appear after running. If you need to save the .nwk file of the current phylogenetic tree, click the **Save** button below the parameter box to save.


FastTree: Parameters

Maximum likelihood model : WAG (Whelan-And-Goldman 2001 model)

Test of Phylogeny : Shimodaira Hasegawa Text (report support value 0-1)

Searching for the best join : relaxed neighbor-joining (recommend)

Join option : neighbor-joining

Show MSA : False

Sequences Selection : All sequences in the class

V Gene Selection :

Fig. 3. Parameter panel of phylogenetic tree

(a) Maximum likelihood model : Maximum likelihood model for amino acid substitutions, including WAG, JJT, LG

(b) Test of Phylogeny : branch checking methods for phylogenetic trees, including Shiodaira Hasegawa Text(defined by FastTree) and Bootstrap(1000x)

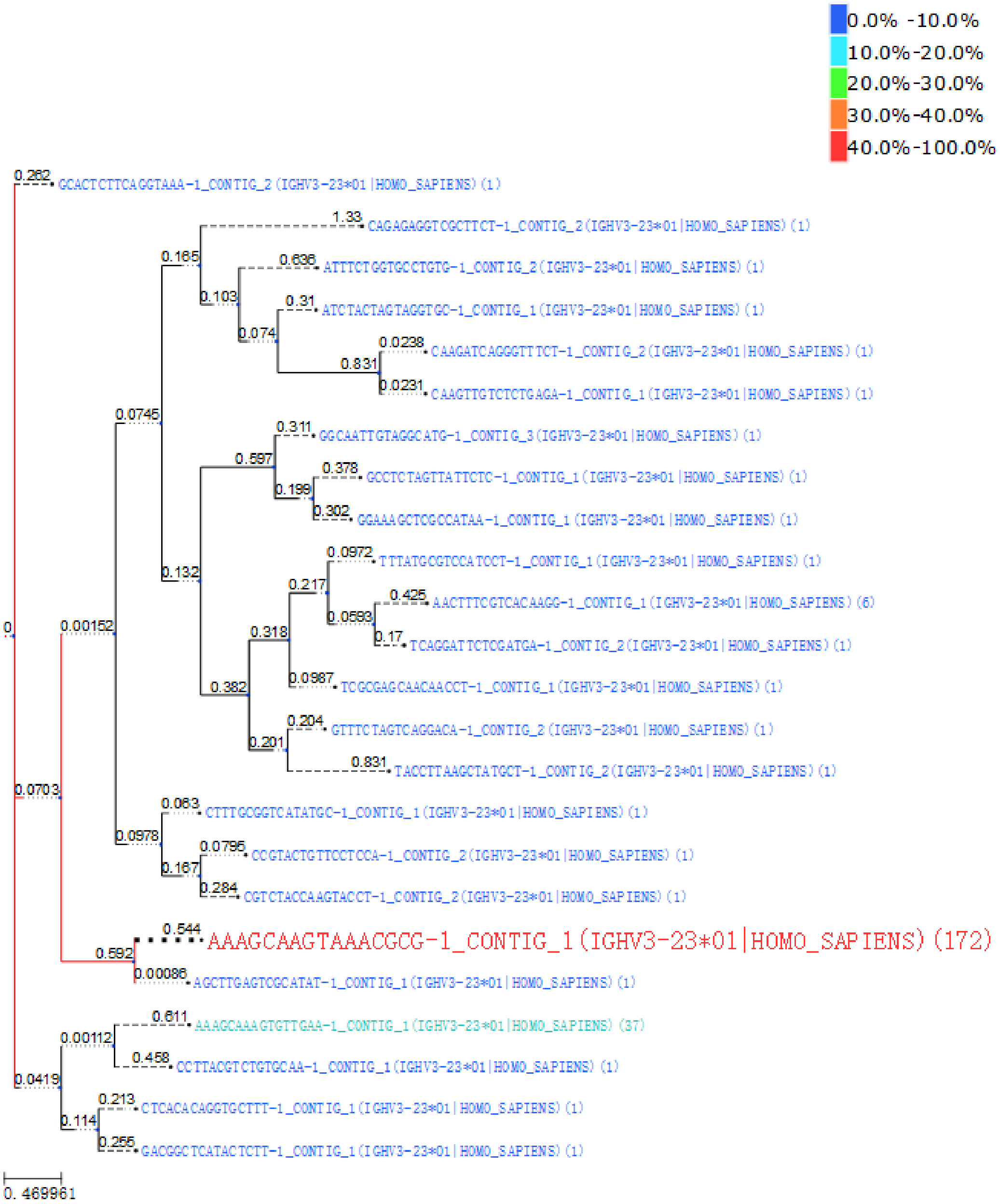
(c) Searching for best join : Using the neighbor-join method to construct a rough tree topology of the system, the default relaxed neighbor joining is faster, the exhaustive search is slower, and the search of the visible set is only the fastest

(d) Join option : The neighbor-join method used to construct a rough topology, defaulting to neighbor - join, in addition to BioNJ.

(e) Show msa : the results of multi-sequence alignment will be displayed with the evolution tree.

(f) Sequences Selection: By default, All sequences in the class, the system tree is constructed with all sequences currently displayed, Sequences belong to a specific V Gene, and the system tree is constructed with sequences belonging to V Gene selected in V Gene Selection.

(g) V Gene Selection: The option contain the V Gene types of all sequences on the current page. Availabled When option of “Sequences Selection” switch to “Sequences belong to a specific V Gene”,



0.0% -10.0%
10.0%-20.0%
20.0%-30.0%
30.0%-40.0%
40.0%-100.0%
0.262
GCACTCTTCAGGTAAA-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
1.33
CAGAGAGGTCGCTTCT-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.165
0.636
ATTTCTGGTGCCTGTG-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.103
0.31
ATCTACTAGTAGGTGC-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.074
0.0238
CAAGATCAGGGTTTCT-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.831
0.0231
CAAGTTGTCTCTGAGA-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.0745
0.311
GGCAATTGTAGGCATG-1_CONTIG_3(IGHV3-23*01|HOMO_SAPIENS)(1)
0.597
0.378
GCCTCTAGTTATTCTC-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.199
0.302
GGAAAGCTCGCCATAA-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.0972
TTTATGCGTCCATCCT-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.132
0.217
0.425
AACTTTCGTCACAAGG-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(6)
0.0593
0
0.00152
0.318
0.17
TCAGGATTCTCGATGA-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.382
0.0987
TCGCGAGCAACAACCT-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.204
GTTTCTAGTCAGGACA-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.201
0.831
TACCTTAAGCTATGCT-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.0703
0.063
CTTTGCGGTCATATGC-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.0978
0.0795
CCGTACTGTTCCTCCA-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.167
0.284
CGTCTACCAAGTACCT-1_CONTIG_2(IGHV3-23*01|HOMO_SAPIENS)(1)
0.544
AAAGCAAGTAAACGCG-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(172)
0.592
0.00086
AGCTTGAGTCGCATAT-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.611
AAAGCAAAGTGTTGAA-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(37)
0.00112
0.458
CCTTACGTCTGTGCAA-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.0419
0.213
CTCACACAGGTGCTTT-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0.114
0.255
GACGGCTCATACTCTT-1_CONTIG_1(IGHV3-23*01|HOMO_SAPIENS)(1)
0. 469961

Figure 4. Visualization results of system tree constructed according to V Gene

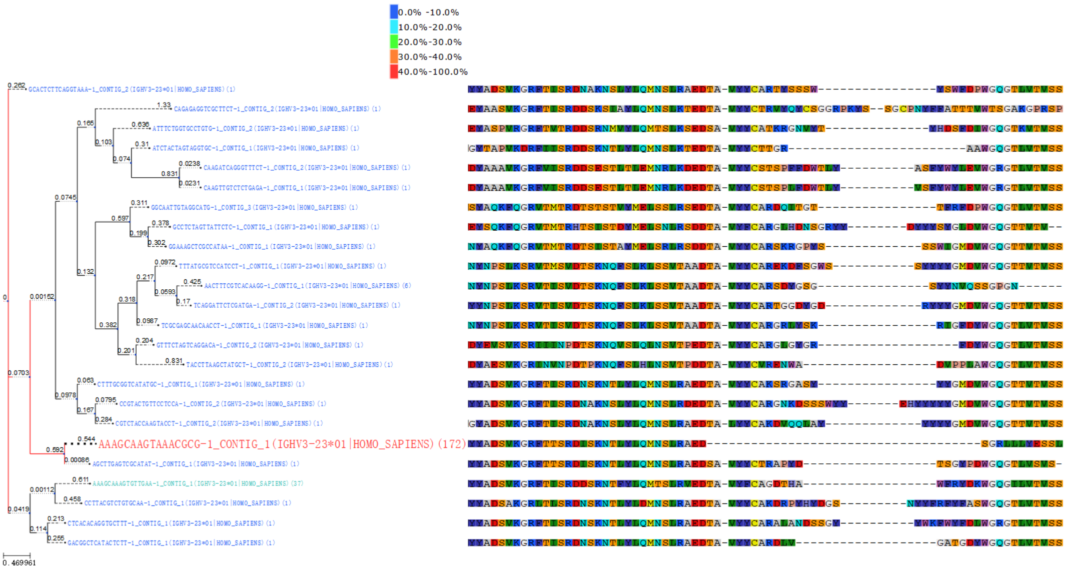


Fig 5. Phylogenetic tree with multiple sequence alignment results

In this program, the root node is automatically set to the sequence most similar to the germline V Gene in all sequences of the phylogenetic tree, and the germline V Gene is the V Gene of the most abundant sequence. The red branch in the tree represents the evolutionary path from the root node to the sequence with the highest abundance. The program renders the label of leaf nodes by the abundance of sequences. The higher the proportion of the number of sequences of leaf nodes in the total number of sequences of all leaf nodes is, the closer the color of leaf node label is to red and the larger the font is.

Interface screenshot: Click the **Screenshot** button to save the result screenshot, which is in .png format by default. Note: Too many sequences may not be successful.

Character Search: Enter the string you want to query in the text box after the **Search** button. Click the **Search** button to search for the location of the string in the sequence name and sequence.

Menu bar

Display

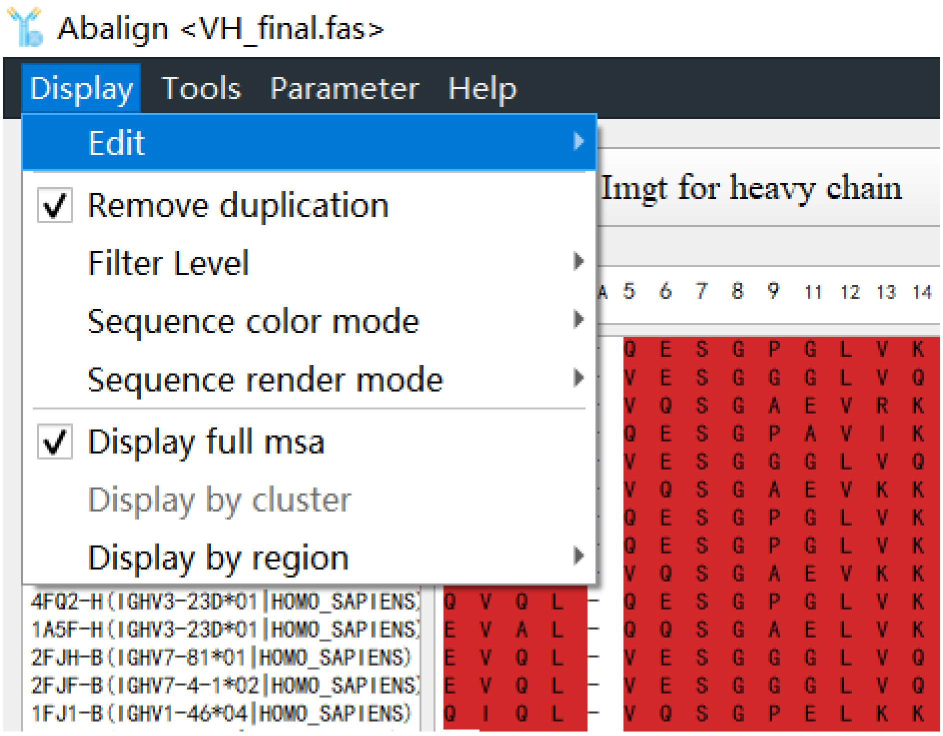


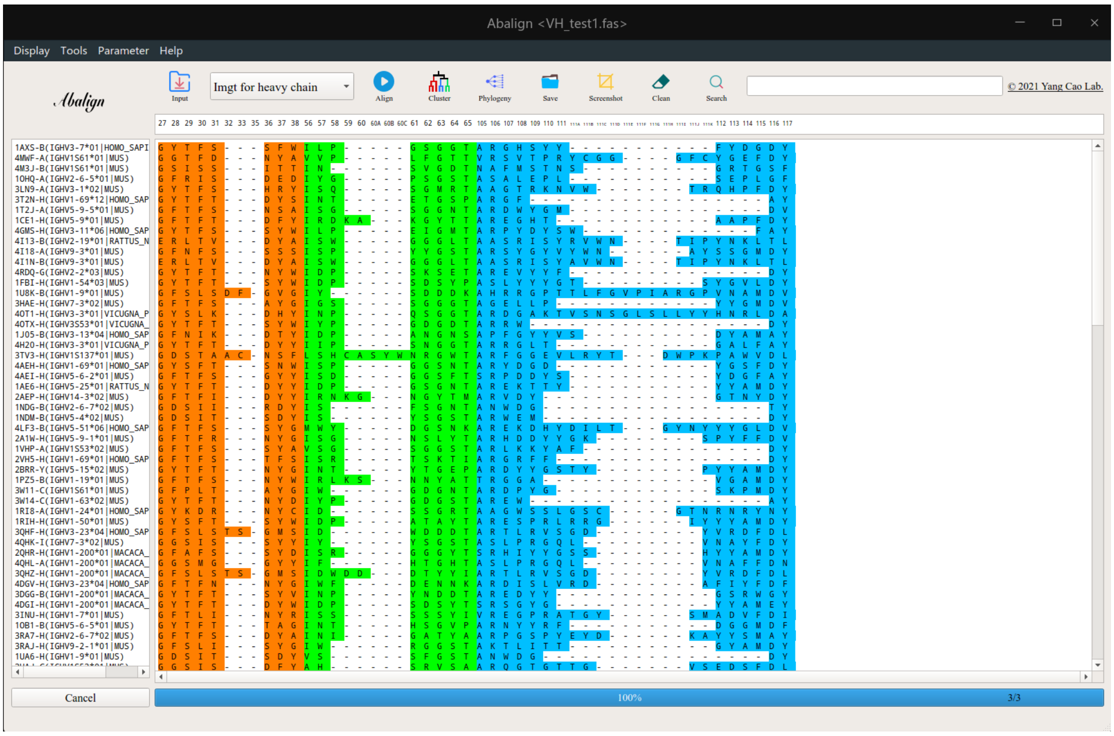
Fig 6. Display options

Remove duplication: By default, repeats in the variable region sequence of the antibody are removed during Align.

Sequence filtering: "**Normal**" is defaulted. After selecting "**Normal**" or "**Strict**", the length of each region in the variable region of the antibody will be limited in the Align process. After selecting "**Off**", the restriction will be closed. If it is not within the limit, the sequence will be filtered out. "Strict" is more restrictive than "Normal".

Switch sequence rendering mode: Click "**Display**" and move the mouse over "**Sequence color mode**", "**Light mode**" (more vivid color) and "**soft mode**" (more soft color) can be selected. Click "**Display**", move the mouse to "**Sequence render mode**", and select "**Color by region**" (different regions of the antibody sequence rendered as different colors) and "**Color by amino**" (different amino acids rendered as different color).

Select the display area: click "**Display**", select Display full msa (default) to display the entire msa results; move the mouse to the "**Display by region**", you can choose to display different regions of the antibody; click "**Display by cluster**" (after complete the cluster operation) to select the class that needs to be displayed by pop-up the tab.



Imgt for heavy chain

```
                            27 28 29 30 31 32 33 35 36 37 38 56 57 58 59 60 60A 60B 60C 61 62 63 64 65 105 106 107 108 109 110 111 111A 111B 111C 111D 111E 111F 111G 111H 111I 111J 111K 112 113 114 115 116 117
1AXS-B(IGHV3-7*01|HOMO_SAPI G Y T F S - - - S F W I L P - - - - - G S G G T A R G H S Y Y - - - - - - - - - - - F Y D G D Y
4MWF-A(IGHV1S61*01|MUS)     G G T F D - - - N Y A V V P - - - - - L F G T T V R S V T P R Y C G G - - - - G F C Y G E F D Y
4M3J-B(IGHV1S61*01|MUS)     G S I S S - - - I T T I N - - - - - - S V G D T N A F M S T N S - - - - - - - - - - G R T G S F
1OHQ-A(IGHV2-6-5*01|MUS)    G F R I S - - - D E D I Y G - - - - - P S G S T A S A L E P L - - - - - - - - - - - S E P L G F
3LN9-A(IGHV3-1*02|MUS)      G Y T F S - - - H R Y I S Q - - - - - S G M R T A A G T R K N V W - - - - - - - T R Q H P F D Y
3T2N-H(IGHV1-69*12|HOMO_SAP G Y T F T - - - D Y S I N T - - - - - E T G S P A R G F - - - - - - - - - - - - - - - - - - A Y
1T2J-A(IGHV5-9-5*01|MUS)    G F T F S - - - N S A I S G - - - - - S G G N T A R D W Y G M - - - - - - - - - - - - - - - D V
1CE1-H(IGHV5-9*01|MUS)      G F T F T - - - D F Y I R D K A - - - K G Y T T A R E G H T - - - - - - - - - - - - A A P F D Y
4GMS-H(IGHV3-11*06|HOMO_SAP G Y T F S - - - S Y W I L P - - - - - E I G M T A R P Y D Y S W - - - - - - - - - - - - - F A Y
4I13-B(IGHV2-19*01|RATTUS_N E R L T V - - - D Y A I S W - - - - - G G G L T A A S R I S Y R V W N - - - - T I P Y N K L T L
4I18-A(IGHV9-3*01|MUS)      G F N F S - - - S S S I S P - - - - - Y Y G S T A R S Y G Y V Y W N - - - - - - A Y S S G M D Y
4I1N-B(IGHV9-3*01|MUS)      E R L T V - - - D Y A I S W - - - - - G G G L T A A S R I S Y A V W N - - - - T I P Y N K L T L
4RDQ-G(IGHV2-2*03|MUS)      G Y T F T - - - N Y W I D P - - - - - S K S E T A R E V Y Y F - - - - - - - - - - - - - - - D Y
1FBI-H(IGHV1-54*03|MUS)     G Y T F T - - - S Y W I D P - - - - - S D S Y P A S L Y Y Y G T - - - - - - - - - S Y G V L D Y
1U8K-B(IGHV1-9*01|MUS)      G F S L S D F - G V G I Y - - - - - - S D D D K A H R R G P T T L F G V P I A R G P V N A M D V
3HAE-H(IGHV7-3*02|MUS)      G F T F S - - - A Y G I G S - - - - - S G G G T A G E L L P - - - - - - - - - - - - Y Y G M D V
4OT1-H(IGHV3-3*01|VICUGNA_P G Y S L K - - - D H Y I N P - - - - - Q S G G T A R D G A K T V S N S G L S L L Y Y H N R L D A
4OTX-H(IGHV3S53*01|VICUGNA_ G Y T F T - - - S Y W I Y P - - - - - G D G D T A R R W - - - - - - - - - - - - - - - - - - D Y
1J05-B(IGHV3-13*04|HOMO_SAP G F N I K - - - D T Y I D P - - - - - A N G N S A P F G Y Y V S - - - - - - - - - - D Y A M A Y
4H20-H(IGHV3-3*01|VICUGNA_P G Y T F T - - - D Y Y I I P - - - - - S N G G T A R R G L T - - - - - - - - - - - - G A L F A Y
3TV3-H(IGHV1S137*01|MUS)    G D S T A A C - N S F L S H C A S Y W N R G W T A R F G G E V L R Y T - - - D W P K P A W V D L
4AEH-H(IGHV1-69*01|HOMO_SAP G Y S F T - - - S N W I S P - - - - - G G S N T A R Y D G D - - - - - - - - - - - - Y G S F D Y
4AEI-H(IGHV5-6-2*01|MUS)    G F T F S - - - G Y Y I S D - - - - - G G S F T S R P D D Y S - - - - - - - - - - - Y D G F A Y
1AE6-H(IGHV5-25*01|RATTUS_N G Y T F T - - - D Y Y I D P - - - - - G S G N T A R E K T T Y - - - - - - - - - - - Y Y A M D Y
2AEP-H(IGHV14-3*02|MUS)     G F T F I - - - D Y Y I R N K G - - - N G Y T M A R V D Y - - - - - - - - - - - - - G T N Y D Y
1NDG-B(IGHV2-6-7*02|MUS)    G D S I I - - - R D Y I S - - - - - - F S G N T A N W D G - - - - - - - - - - - - - - - - - T Y
1NDM-B(IGHV5-4*02|MUS)      G D S I T - - - S D Y I S - - - - - - Y S G S T A R W E M - - - - - - - - - - - - - - - - - D Y
4LF3-B(IGHV5-51*06|HOMO_SAP G F T F S - - - S Y G M W Y - - - - - D G S N K A R E K D H Y D I L T - - - G Y N Y Y Y G L D V
2A1W-H(IGHV5-9-1*01|MUS)    G F T F R - - - N Y G I S G - - - - - N S L Y T A R H D D Y Y G K - - - - - - - - S P Y F F D V
1VHP-A(IGHV1S53*02|MUS)     G F T F S - - - S Y A V S G - - - - - S G G S T A R L K K Y A F - - - - - - - - - - - - - - D Y
2VH5-H(IGHV1-69*01|HOMO_SAP G F T F S - - - T F S I S R - - - - - T S K T I A R G R F F - - - - - - - - - - - - - - - - D Y
2BRR-Y(IGHV5-15*02|MUS)     G Y T F T - - - N Y G I N T - - - - - Y T G E P A R D Y Y G S T Y - - - - - - - - P Y Y A M D Y
1PZ5-B(IGHV1-19*01|MUS)     G F T F S - - - N Y W I R L K S - - - N N Y A T T R G G A - - - - - - - - - - - - - V G A M D Y
3W11-C(IGHV1S61*01|MUS)     G F P L T - - - A Y G I W - - - - - - G D G N T A R D P Y G - - - - - - - - - - - - S K P M D Y
3W14-C(IGHV1-63*02|MUS)     G Y T F T - - - N Y D I Y P - - - - - G D G S T A R E W - - - - - - - - - - - - - - - - - - A Y
1RI8-A(IGHV1-24*01|HOMO_SAP G Y K D R - - - N Y C I D - - - - - - S S G R T A A G W S S L G S C - - - - - G T N R N R Y N Y
1RIH-H(IGHV1-50*01|MUS)     G Y S F T - - - S Y W I D P - - - - - A T A Y T A R E S P R L R R G - - - - - - I Y Y Y A M D Y
3QHF-H(IGHV3-23*04|HOMO_SAP G F S L S T S - G M S I D - - - - - - W D D D T A R T L R V S G D - - - - - - - - Y V R D F D L
4QHK-I(IGHV7-3*02|MUS)      G G S I S - - - S Y Y I Y - - - - - - Y S G S T A S L P R G Q L - - - - - - - - - V N A Y F D Y
2QHR-H(IGHV1-200*01|MACACA_ G F A F S - - - S Y D I S R - - - - - G G G Y T S R H I Y Y G S S - - - - - - - - H Y Y A M D Y
4QHL-A(IGHV1-200*01|MACACA_ G G S M G - - - G Y Y I F - - - - - - H T G H T A S L P R G Q L - - - - - - - - - V N A F F D N
3QHZ-H(IGHV1-200*01|MACACA_ G F S L S T S - G M S I D W D D - - - D T Y Y I A R T L R V S G D - - - - - - - - Y V R D F D L
4DGV-H(IGHV3-23*04|HOMO_SAP G F T F N - - - N Y G I W F - - - - - D E N N K A R D I S L V R D - - - - - - - - A F I Y F D F
3DGG-B(IGHV1-200*01|MACACA_ G Y T F T - - - S Y V I N P - - - - - Y N D D T A R E D Y Y - - - - - - - - - - - - G S R W G Y
4DGI-H(IGHV1-200*01|MACACA_ G Y T F T - - - D Y W I D P - - - - - S D S Y T S R S G Y G - - - - - - - - - - - - Y Y A M E Y
3INU-H(IGHV1-7*01|MUS)      G F T L I - - - N Y R I S S - - - - - S S S Y I V R E G P R A T G Y - - - - - - S M A D V F D I
1OB1-B(IGHV5-6-5*01|MUS)    G Y T F T - - - T A G I N T - - - - - H S G V P A R N Y Y R F - - - - - - - - - - - D G G M D F
3RA7-H(IGHV2-6-7*02|MUS)    G F T F S - - - D Y A I N I - - - - - G A T Y A A R P G S P Y E Y D - - - - - - K A Y Y S M A Y
3RAJ-H(IGHV9-2-1*01|MUS)    G F S L I - - - S Y G I W - - - - - - R G G S T A K T L I T T - - - - - - - - - - - G Y A M D Y
1UA6-H(IGHV1-9*01|MUS)      G D S I T - - - S D Y V S - - - - - - S F G S T A N W D G - - - - - - - - - - - - - - - - - D Y
                            G G S I S - - - D F Y A H - - - - - - S R V S A A R Q G T G T T G - - - - - - - V S E D S F D L
```

Figure 7. Display only the CDR1, CDR2, CDR3 regions of the antibody variable domain

Tools

Find V Gene: The default selection cannot be changed to find the most similar V Gene for each sequence of antibody variable domain and the species to which the V Gene belongs, and display the V Gene name and the species name behind the corresponding sequence name.

Heatmap: Click on "**Heatmap**" to generate a heatmap based on the similarity of the current sequence, and you can cluster the heatmaps and wait for a while if the number of sequences is large. All Heatmaps are based on the content of the current multi-sequence alignment text box. If "Display" is used to change the current displayed sequence, Heatmap will change accordingly.

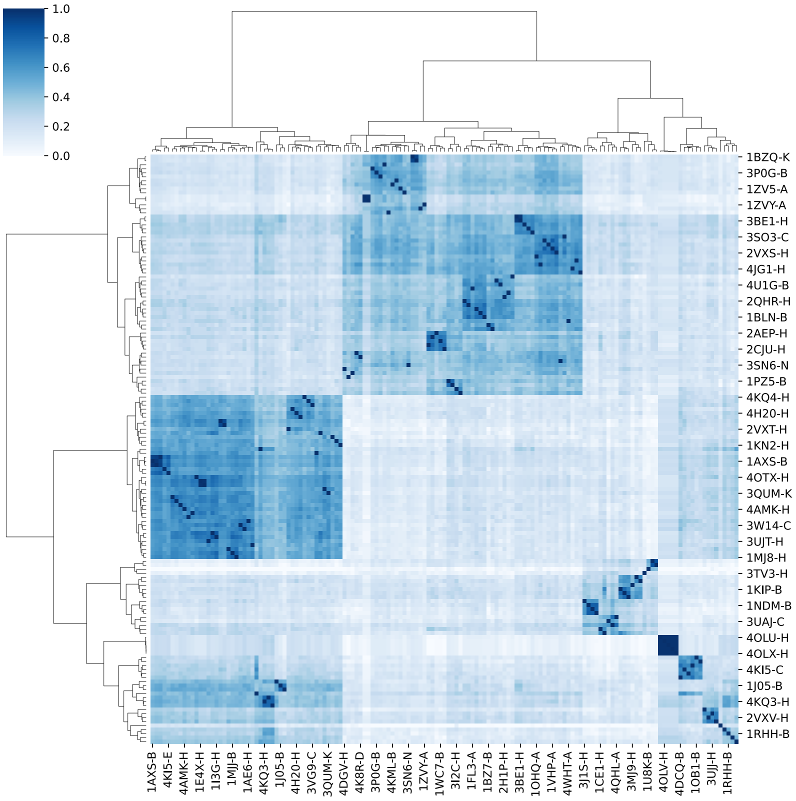
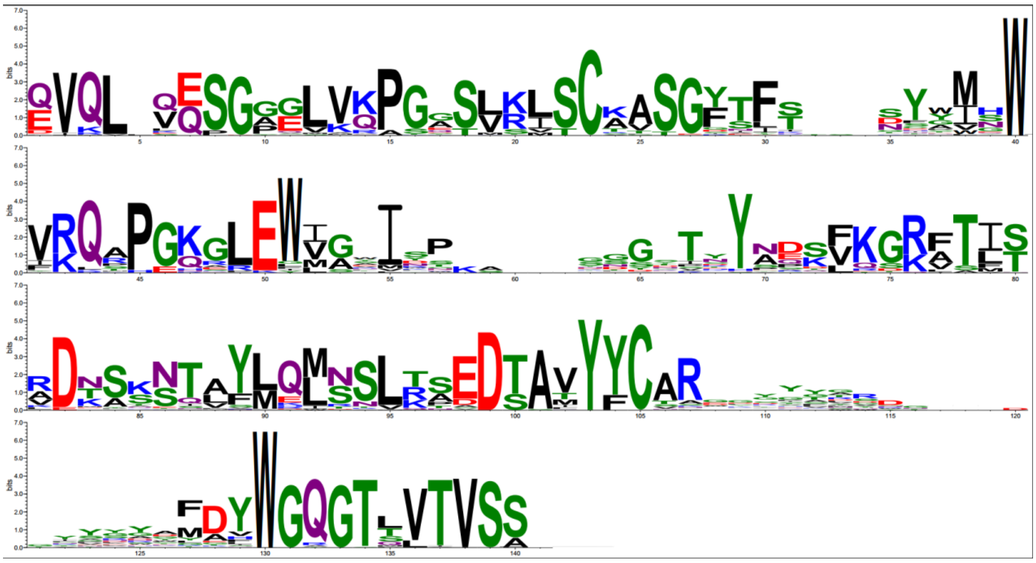


Figure 8. Heatmap after hierarchical clustering

Seqlogo: Click **Tools**, move the mouse to "**Seqlogo**", click "**By Entropy**" to get entropy ordinate Seqlogo, click "**By Frequence**" to get frequency ordinate Seqlogo. Move the mouse to the "**Color**" option to change the rendering mode of Seqlogo. All Seqlogo is based on the content of the current multiple sequence alignment text box. If Display is used to change the current displayed sequence, Seqlogo will change accordingly. To ensure compatibility, all Seqlogo is in



PDF format.

Figure 9. Seqlogo for the entire antibody variable region, with entropy as the y-axis

Abundance map: Click “**Tools**”, move the mouse to “**Abundance**”, see “**V Gene abundance**” and “**Sequence abundance**” options, and generate V Gene abundance and Sequence abundance maps after clicking.

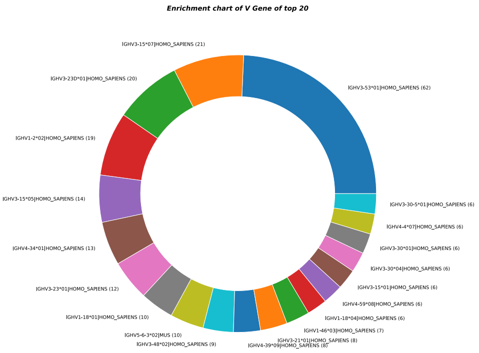


Figure 10. V Gene abundance map

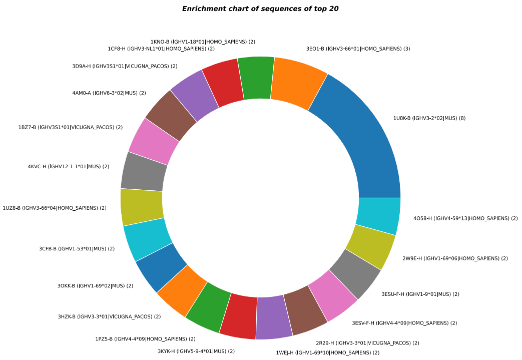
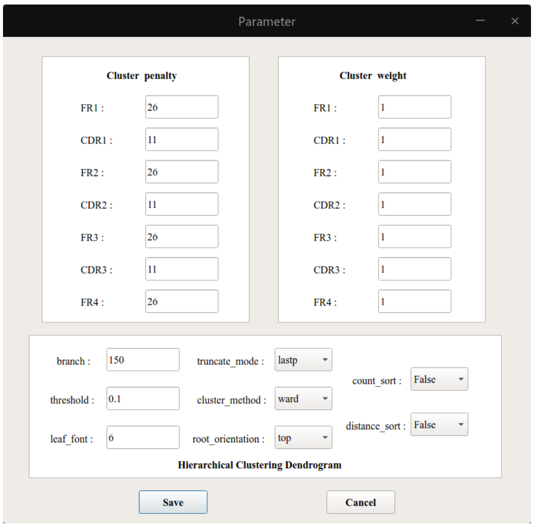


Figure 11. Sequence abundance map

Parameter

After clicking the “**Parameter**” in the menu bar, a dialog box will pop up, which can modify the clustering penalty, clustering weight (if the weight of CDR3 region is increased, the similarity of CDR3 region will be paid more attention to in the clustering process) and the parameters of Hierarchical Clustering Dendrogram.



Cluster penalty		Cluster weight	
FR1 :	26	FR1 :	1
CDR1 :	11	CDR1 :	1
FR2 :	26	FR2 :	1
CDR2 :	11	CDR2 :	1
FR3 :	26	FR3 :	1
CDR3 :	11	CDR3 :	1
FR4 :	26	FR4 :	1

branch : 150
threshold : 0.1
leaf_font : 6
truncate_mode : lastp
cluster_method : ward
root_orientation : top
count_sort : False
distance_sort : False

Hierarchical Clustering Dendrogram

Figure 12. Hierarchical Clustering Dendrogram parameters

(a) Branch: It represents the number of branches displayed in hierarchical clustering tree. This parameter takes effect with truncate_mode parameter.

(b) Threshold: represents the threshold of clustering partition. If this value is 0.1, a tenth of the total length of the Y-axis (the maximum difference between sequences) is used as a threshold to partition different classes. The more the number of clusters, the smaller the value should be set.

(c) leaf _ font: Represents the size of the branch label font

(d) cluster_method: represents the method used to calculate the distance between two classes during clustering.

(e) root_orientation: Represents the direction of the root.

(f) count_sort: indicates whether the clustering results are sorted according to the number of each category.

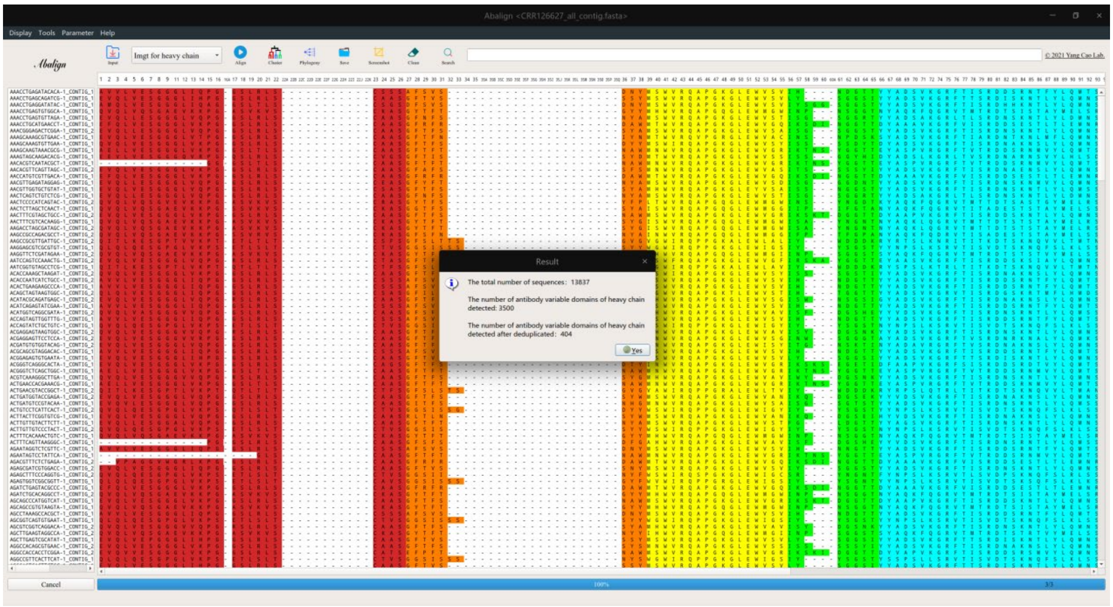
(g) Distance_sort: indicates whether the clustering results are sorted according to the distance of each category.

Example

In this case, we used the samples obtained from single cell sequencing of patients with new coronavirus for demonstration.

Step 1 Input file: select the fasta file that needs to be processed after clicking Input.

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.

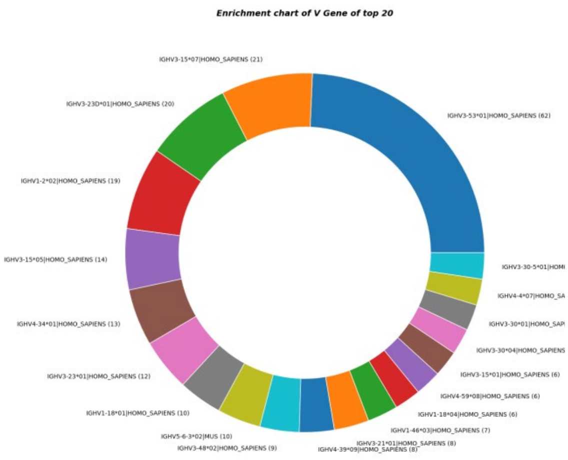


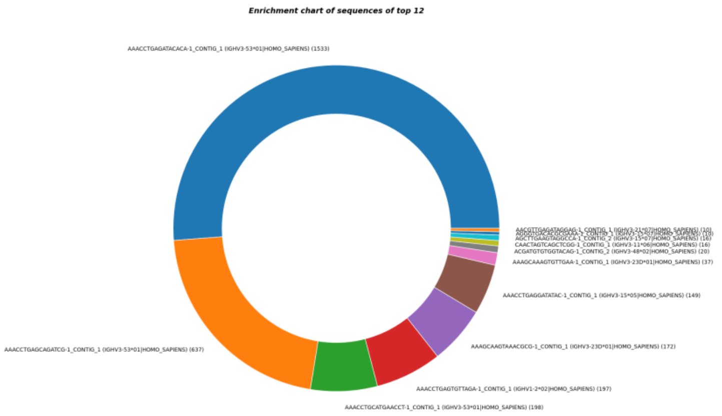
The total number of sequences： 13837

The number of antibody variable domains of heavy chain detected: 3500

The number of antibody variable domains of heavy chain detected after deduplicated： 404

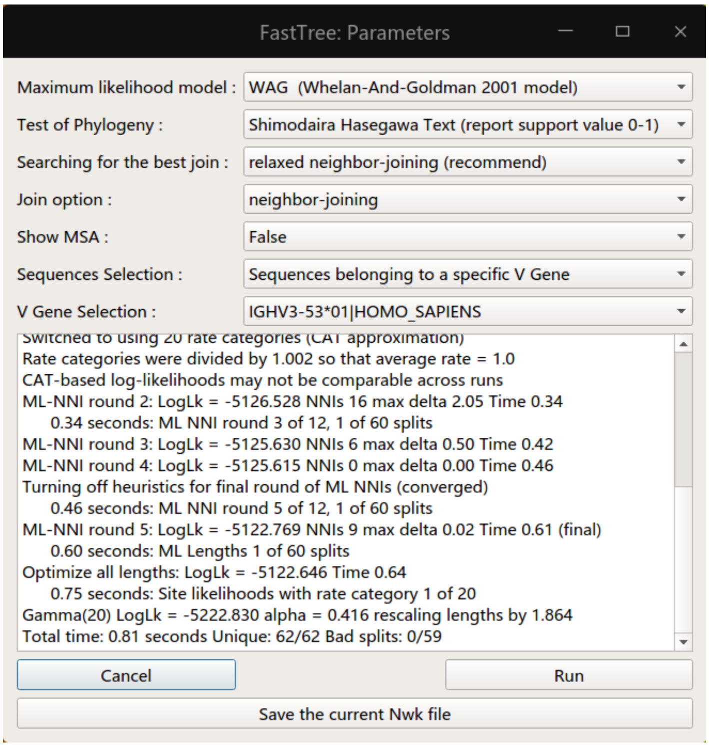
Step 3 looks at multiple sequence alignments: click the “Tools” button in the top menu bar and





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.

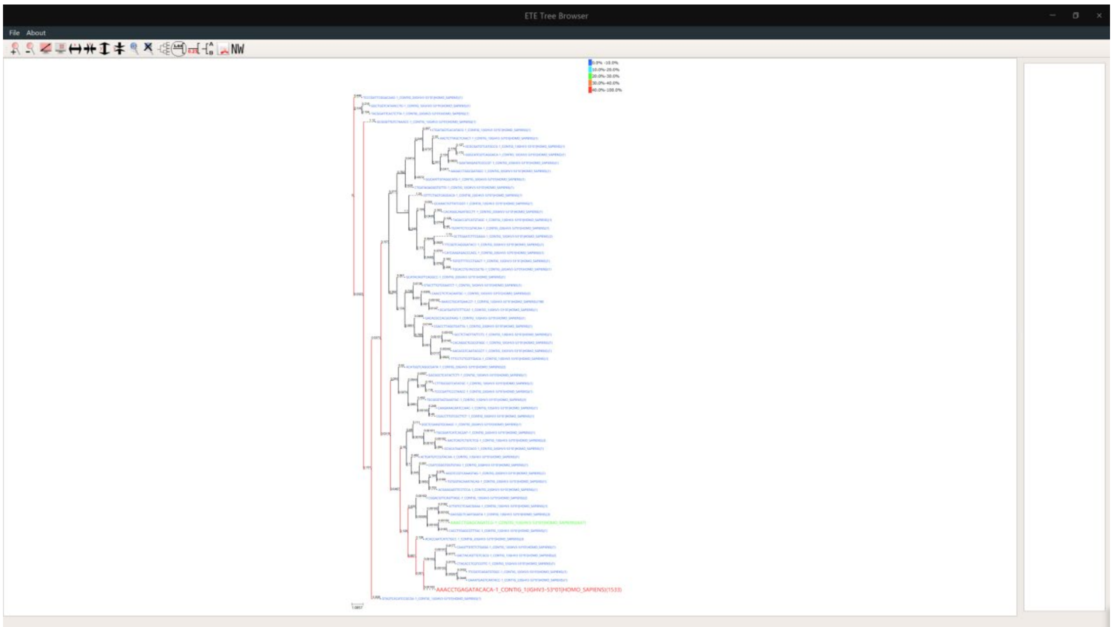
Step 4 Build the phylogenetic tree: according to the sequence abundance information, build the tree with the V Gene of the high abundance sequence. In this case, the most abundant sequence belongs to V Gene “IGVH3-53 * 01”. Click Phylogeny to open the parameter list of tree building, switch the Sequences Selection option to “Sequences belonging to a specific V Gene”, select “IGVH3-53 * 01” in the V Gene Selection option, and click Run to start building the evolutionary tree. 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.



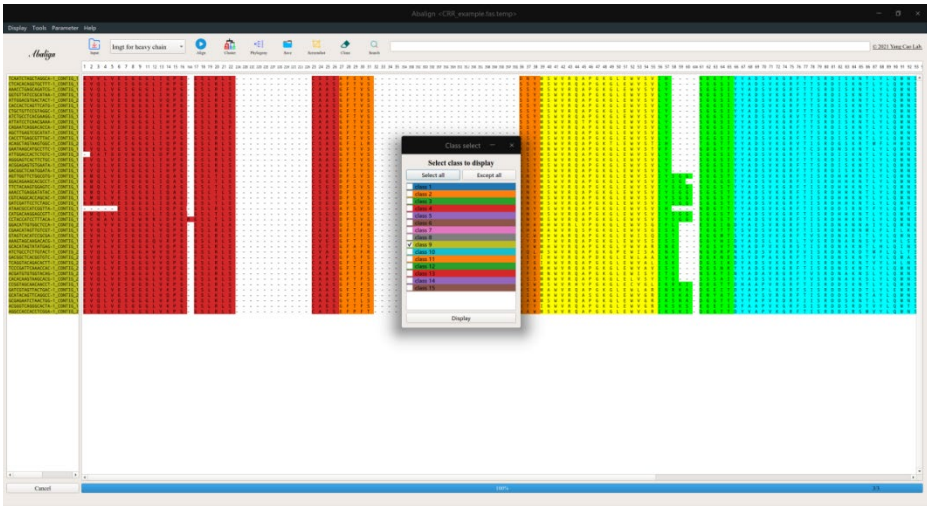
Maximum likelihood model : WAG (Whelan-And-Goldman 2001 model)
Test of Phylogeny : Shimodaira Hasegawa Text (report support value 0-1)
Searching for the best join : relaxed neighbor-joining (recommend)
Join option : neighbor-joining
Show MSA : False
Sequences Selection : Sequences belonging to a specific V Gene
V Gene Selection : IGHV3-53*01|HOMO_SAPIENS

```
Switched to using 20 rate categories (CAT approximation)
Rate categories were divided by 1.002 so that average rate = 1.0
CAT-based log-likelihoods may not be comparable across runs
ML-NNI round 2: LogLk = -5126.528 NNIs 16 max delta 2.05 Time 0.34
      0.34 seconds: ML NNI round 3 of 12, 1 of 60 splits
ML-NNI round 3: LogLk = -5125.630 NNIs 6 max delta 0.50 Time 0.42
ML-NNI round 4: LogLk = -5125.615 NNIs 0 max delta 0.00 Time 0.46
Turning off heuristics for final round of ML NNIs (converged)
      0.46 seconds: ML NNI round 5 of 12, 1 of 60 splits
ML-NNI round 5: LogLk = -5122.769 NNIs 9 max delta 0.02 Time 0.61 (final)
      0.60 seconds: ML Lengths 1 of 60 splits
Optimize all lengths: LogLk = -5122.646 Time 0.64
      0.75 seconds: Site likelihoods with rate category 1 of 20
Gamma(20) LogLk = -5222.830 alpha = 0.416 rescaling lengths by 1.864
Total time: 0.81 seconds Unique: 62/62 Bad splits: 0/59
```

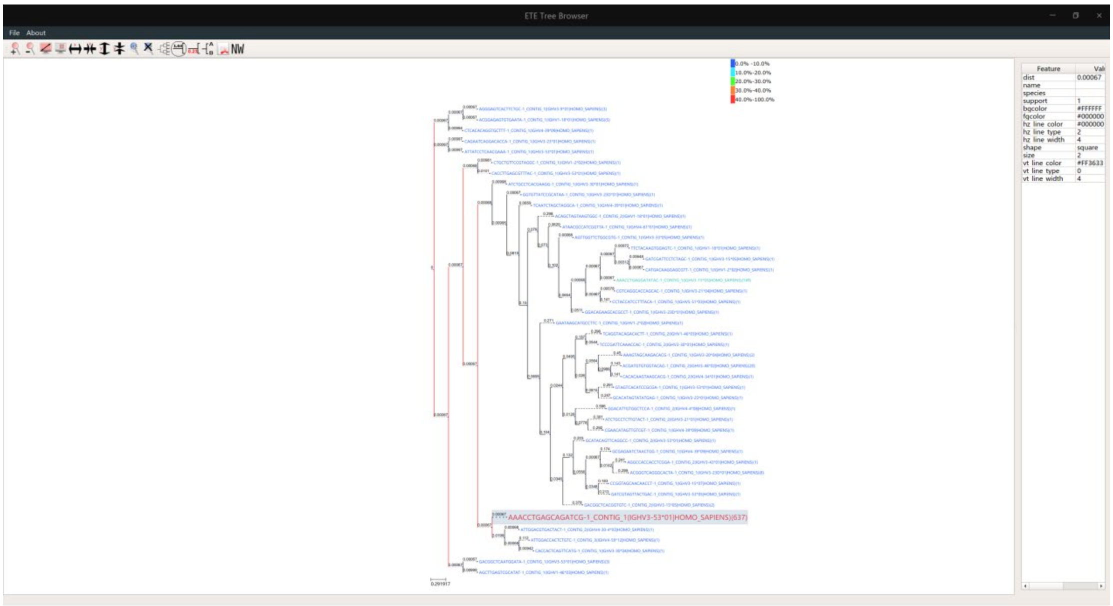
Step 5 View the phylogenetic tree: Once the evolution tree is built, a visual window pops up that adjusts the view and displays other information about the tree through the button above the visual window. After selecting the node of the tree, you can also modify the attributes of the node, such as the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



Step 6 Hierarchical clustering: After clicking Cluster, hierarchical clustering can be carried out. If the file is too large, it will consume more time, please wait patiently. After the clustering, the hierarchical clustering tree graph will be displayed. Multiple sequence alignment will be rearranged according to the results of the tree graph, and the sequence name will be rendered as the same color according to the color of the branch of the tree graph. After the clustering is completed, click Display- > Display by cluster in the menu bar to select the class that needs to be displayed.



Step 7 Building phylogenetic trees with different classes: After selecting one or more classes that need to be used to build trees in Display by cluster, click Phylogeny again, adjust “Sequences Selection” to “All sequences in the class”, and click Run to build phylogenetic trees.



Feature	Val
dist	0.00067
name	
species	
support	1
bgcolor	#FFFFFF
fgcolor	#000000
hz line color	#000000
hz line type	2
hz line width	4
shape	square
size	2
vt line color	#FF3633
vt line type	0
vt line width	4

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. Hierarchical clustering consumes resources and time. To hierarchically cluster a large number of sequences, ensure that the computer has sufficient memory.

5. If the software screen is not displayed properly, adjust the scale of the screen.

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

Reference

[1] 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.