Abalign 手册

介绍

长期以来，多序列比对一直是研究蛋白质家族进化、结构和功能特性的有力工具。与普通的蛋白质家族相比，抗体或 BCR 序列存在高度可变的结构域，这使得现有的多序列比对方法不能很好地对抗体进行分析。近年来，随着 COVID-19 的全球大流行，BCR 测序的数据量大幅度增加，这使得对 BCR 进行多序列比对分析的需求日益增长。为了解决这个问题，我们基于 AbRSA[1]开发了一个多序列比对方法 Abalign，它集成了标准化抗体序列编号方案的启发式知识，包括 IMGT、KABAT 和 Chothia 策略。比对遵循免疫学研究中已知的保守位点或高度可变位点的先验知识，比对结果与蛋白结构、免疫学知识高度一致。Abalign 被集成到了一个对用户友好的软件中，它具有交互式和可视化的界面，并且支持多序列比对、序列聚类、抗体编号、CDR 区域划分、构建系统发育树、序列相似性热图、V 基因测定和丰度分析等功能，全程操作只需点击鼠标即可实现。Abalign 允许对 BCR 测序数据进行高通量分析，对于 500MB 的 DNA FASTA 序列，420 分钟内即可完成分析（AMD 2990WX，单线程）。我们在 Linux 上测试了 10G 的数据，在 Windows 上测试了 8G 的数据，并且都得到了正确的结果。Abalign 将会对免疫信息学和制药公司分析大量 BCR 或抗体序列，取得新的研究发现起到很大的帮助。

用法

软件组成

注：请勿修改 lib 文件夹的任何内容

可执行文件(UI)：软件根目录下名为 Abalign_ui 的文件即为可视化窗口程序，双击运行。

lib 文件夹：请勿对软件目录下的 lib 文件夹做任何修改，lib 文件夹内包含一系列配置文件以及实现多序列比对功能的二进制文件。

example 文件夹：存放了 FASTA 格式的 DNA 以及氨基酸的序列文件供用户测试。

基础功能

文件输入：程序支持 FASTA 格式的氨基酸序列文件和核酸序列文件，可通过点击 **Input** 按钮选择输入文件或直接将文件拖入到文本框内以输入文件。

编号策略：通过点击”**IMGT for heavy chain**”下拉菜单选择抗体编号策略（支持 Chothia，Kabat，IMGT 策略）和抗体序列类型（选择重链会检测出输入序列中所有的抗体重链并过滤掉轻链和非可变区序列，轻链同理）。

多序列比对：点击 **Align** 按钮以查找抗体的可变区并用查找到的可变区序列进行多序列比对，比对过程中会查找与每条可变区序列最相似的 V-Gene 及该 V Gene 所属的物种，并将 V Gene 名和物种名展示于对应的序列名后。运行结束后每个位点的残基编号将会展示在上方，并默认将抗体可变区的（FR1, CDR1...CDR3, FR4）共七个区域渲染为不同的颜色，若需更改渲染方式可点击 **Disaplay->Sequence render mode**.

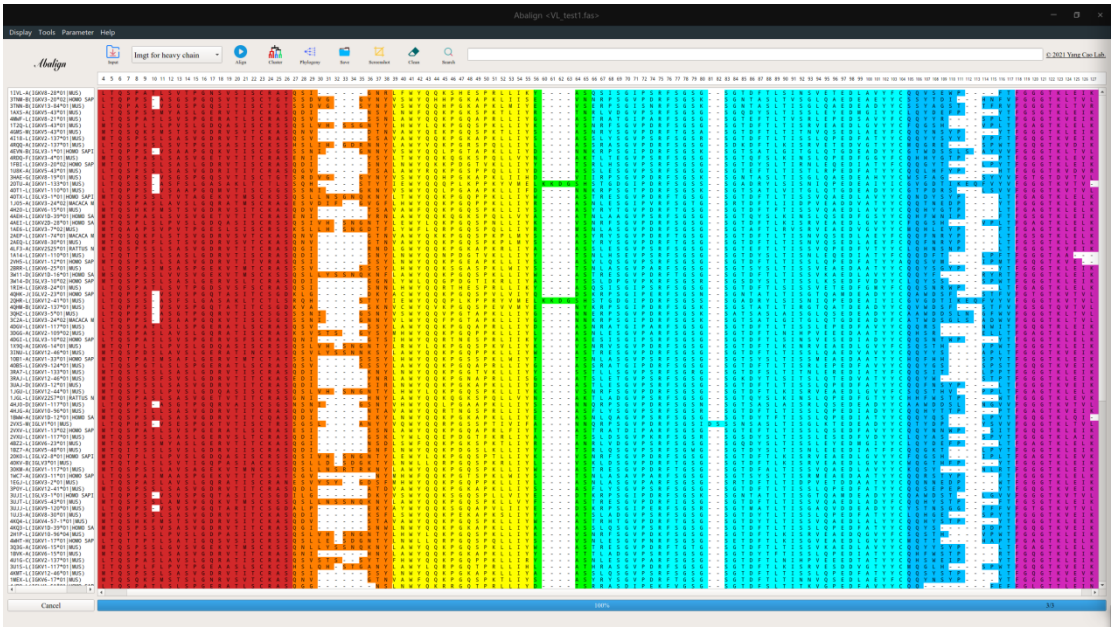


图 1. 多序列比对结果，抗体可变区的不同区域渲染为不同颜色

文件保存：点击 **Save** 按钮后，可将当前界面的内容保存为.fasta 和.temp 文件。.fasta 存储了多序列比对结果，.temp 存储了多序列比对结果并以”*”作为可变区各区域的分隔。

终止比对：点击 **Cancel** 按钮，可终止正在进行的多序列比对。

可选功能

注：所有可选功能均要求先完成多序列比对。

层次聚类：点击 **Cluster** 按钮以运行层次聚类。该功能会将 Align 后的多序列比对数据进行层次聚类，聚类结束后会显示层次聚类树状图，序列将根据树状图进行重排，同类序列会相邻排列，且序列名会渲染为相同颜色。层次聚类的参数可通过菜单栏中的 **Parameter** 选项更改。

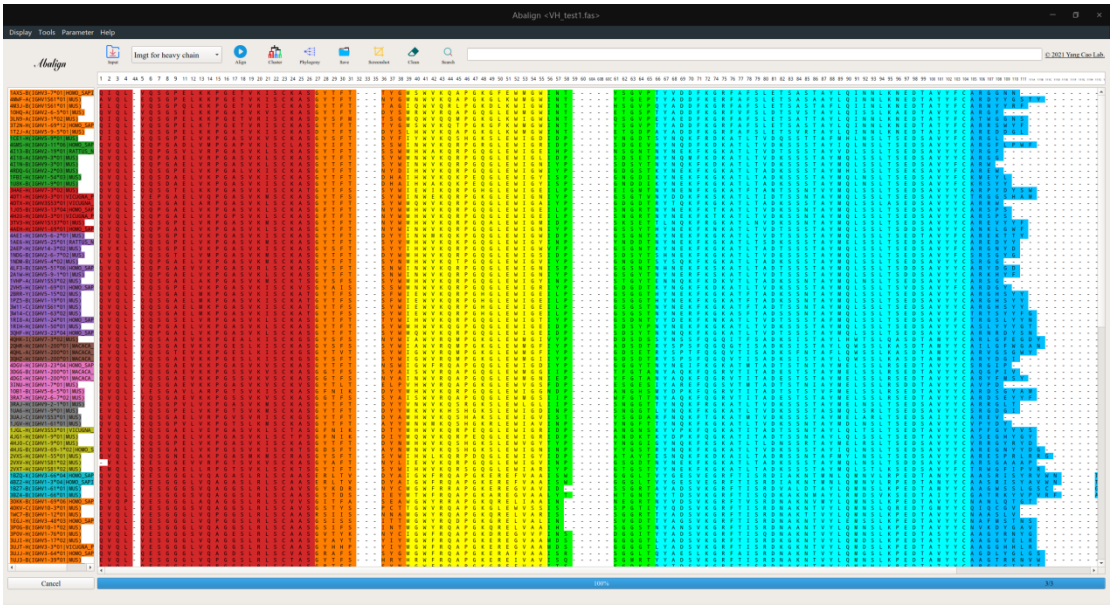


图 2. 抗体可变区层次聚类结果

系统发生树：本功能利用 FastTree 软件（最大似然法）[2]构建 nwk 文件，并用 Ete3[3] 进行可视化。点击 **Phylogeny** 后会弹出系统发生树的参数框，点击 **Run** 开始构建系统发生树，下方的文本框会输入日志信息。运行完成后会弹出系统发生树的可视化结果，如果需要保存当前进化树的.NWK 文件，可点击参数框下方的 **Save** 按钮进行保存。

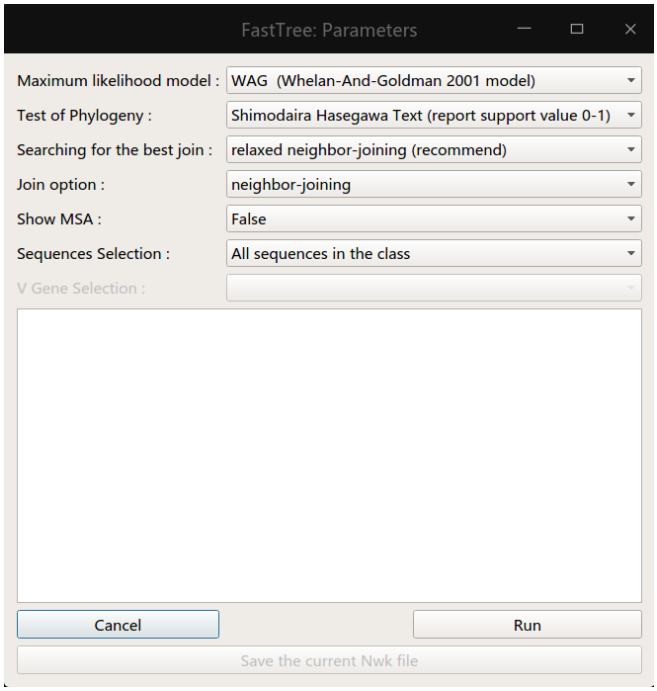


图 3. 系统发生树参数面板

(a) Maximum likelihood model: 最大似然法的氨基酸替换模型，包括 WAG, JJT, LG

(b) Test of Phylogeny: 系统发生树的分支检验方法，包括由 FastTree 定义的 Shiodaira Hasegawa Text 和 Bootstrap(1000x)

(c) Searching for best join: 使用邻接法构建系统发生树粗略拓扑结构的方式，默认使用 relaxed neighbor joining 速度较快，exhaustive search 速度较慢，search the visible set only 速度最快

(d) Join option: 构建粗略拓扑结构所使用的邻接法，默认为 neighbor-joining，除此之外还有 BioNJ.

(e) Show msa: 会将多序列比对结果和进化树一起展示

(f) Sequences Selection: 默认为 All sequence in the class，会用当前显示的所有序列构建系统发生树，Sequences belonging to a specific V Gene,将会用属于 V Gene Selection 中选中的 V Gene 的序列构建系统发生树

(g) V Gene Selection: 当 Sequences Selection 为 Sequences belonging to a specific V Gene 时可用，里面的选项包含当前页面所有序列的 V Gene 种类

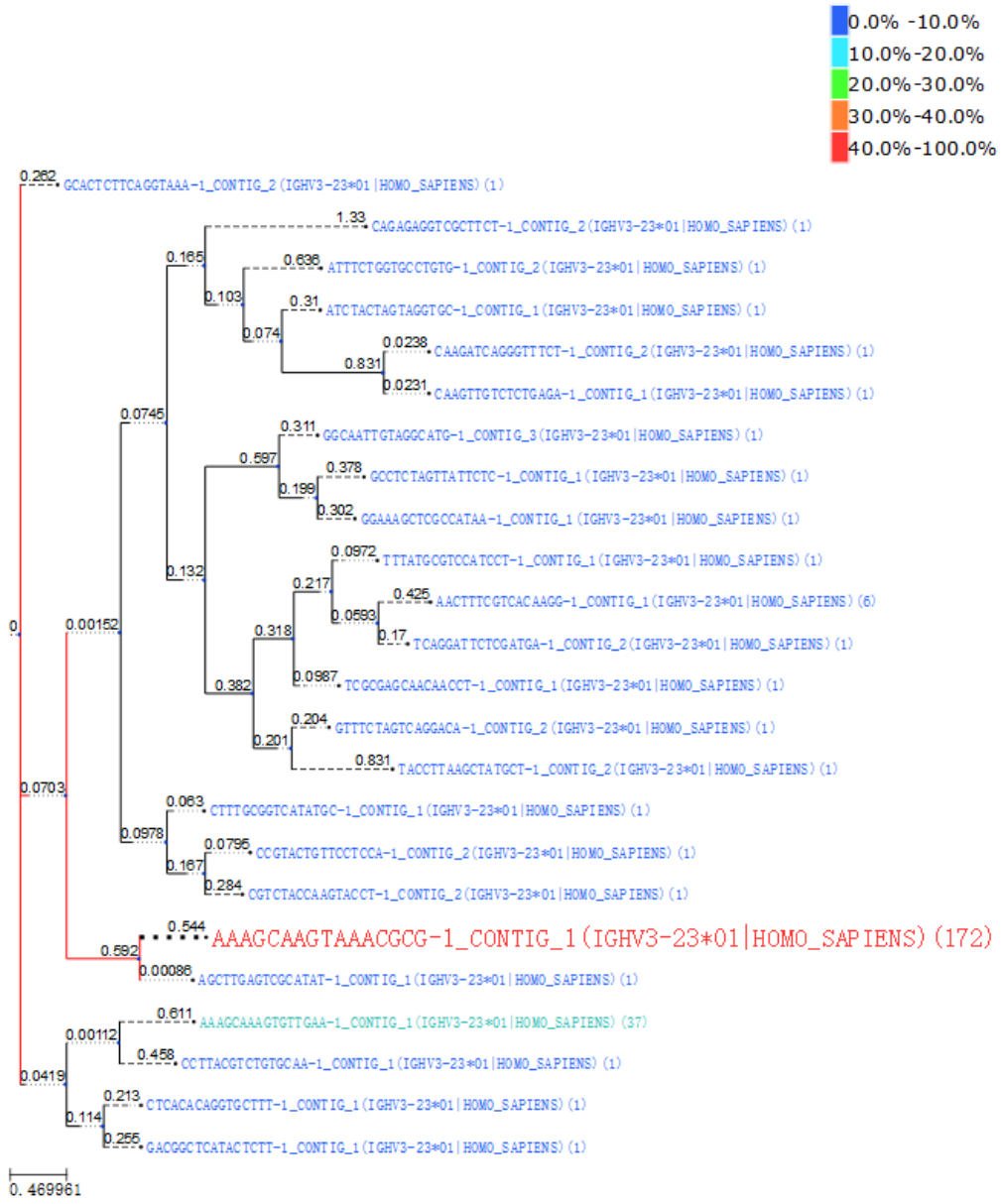


图 4. 按照 V Gene 构建的系统发生树可视化结果

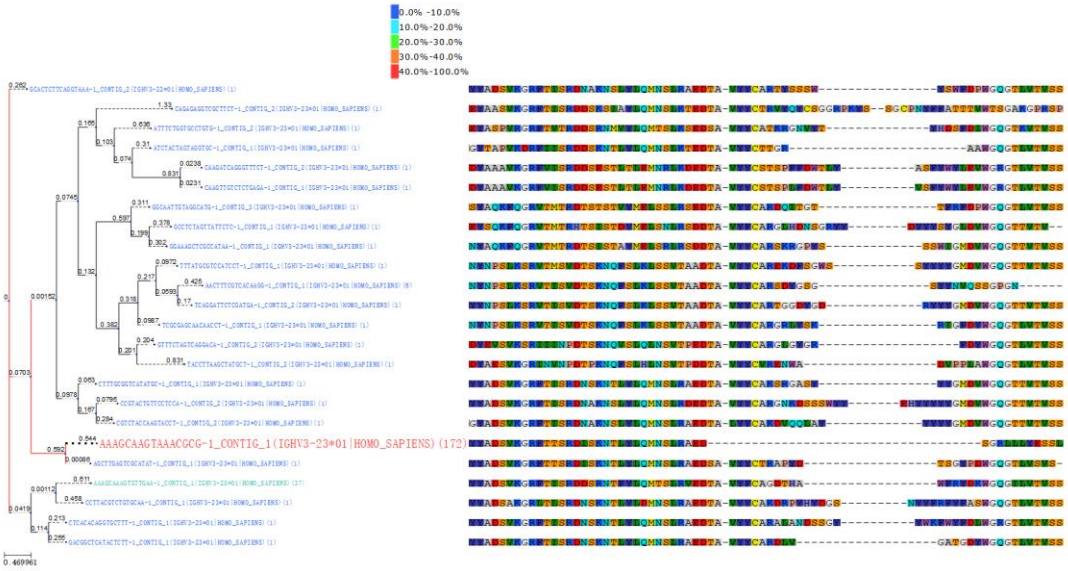


图 5. 带有多序列比对结果的系统发生树

本程序中自动将根节点设置为系统发生树的所有序列中与种系 V Gene 最相似的序列，该种系 V Gene 为丰度最高的序列所属的 V Gene。树中红色分支代表根节点到丰度最高的序列的进化路径。程序以序列的丰度来渲染叶节点的标签，该叶节点的序列数量占所有叶节点序列总数的比例越高，叶节点标签的颜色越接近红色且字体越大。

界面截图：点击 **Screenshot** 按钮可将比对结果截图保存，图片格式默认为 png 格式。注：若序列过多可能无法截图成功。

字符搜索：在 **Search** 按钮后的文本框中输入需要查询的字符串，点击 **Search** 按钮后会在序列名和序列中搜索该字符串出现的位置。

菜单栏

展示

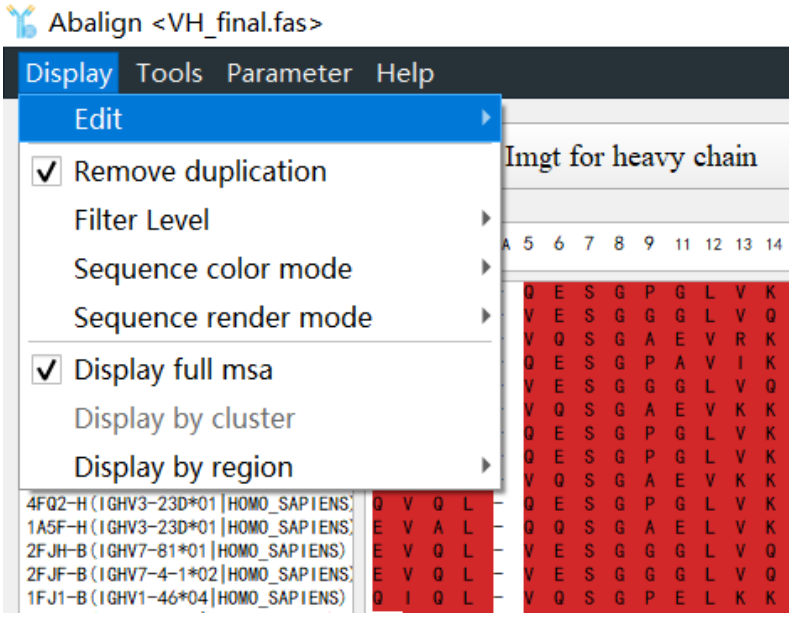


图 6. Display 选项

去重：默认勾选，勾选后在 Align 过程中会去掉查找到的抗体可变区序列中的重复序列。

序列过滤：默认为 **Normal**，勾选 **Normal** 或 **Strict** 后在 Align 过程中会对抗体可变区各个区域的长度进行限制，勾选 **Off** 则关闭限制，若不在限制范围内则序列被过滤掉。Strict 的限制比 Normal 更加严格。

切换序列渲染模式：点击 **Display**，将鼠标移动到 **Sequence color mode** 上，可选择 **Light mode**（颜色更加鲜艳）和 **Soft mode**（颜色更加柔和）两种颜色模式。点击 **Display**，将鼠标移动到 **Sequence render mode** 上，可选择 **Color by region**（抗体序列不同区域渲染为不同颜色）和 **Color by amino**（不同的氨基酸渲染为不同颜色）。

选择展示区域：点击 **Display**，选择 **Display full msa**（默认）可展示整个 msa 比对结果；将鼠标移动到 **Display by region** 上，可选择展示抗体的不同区域；点击 **Display by cluster**（需完成 Cluster 操作）可通过弹出选项框选择需要显示的类。

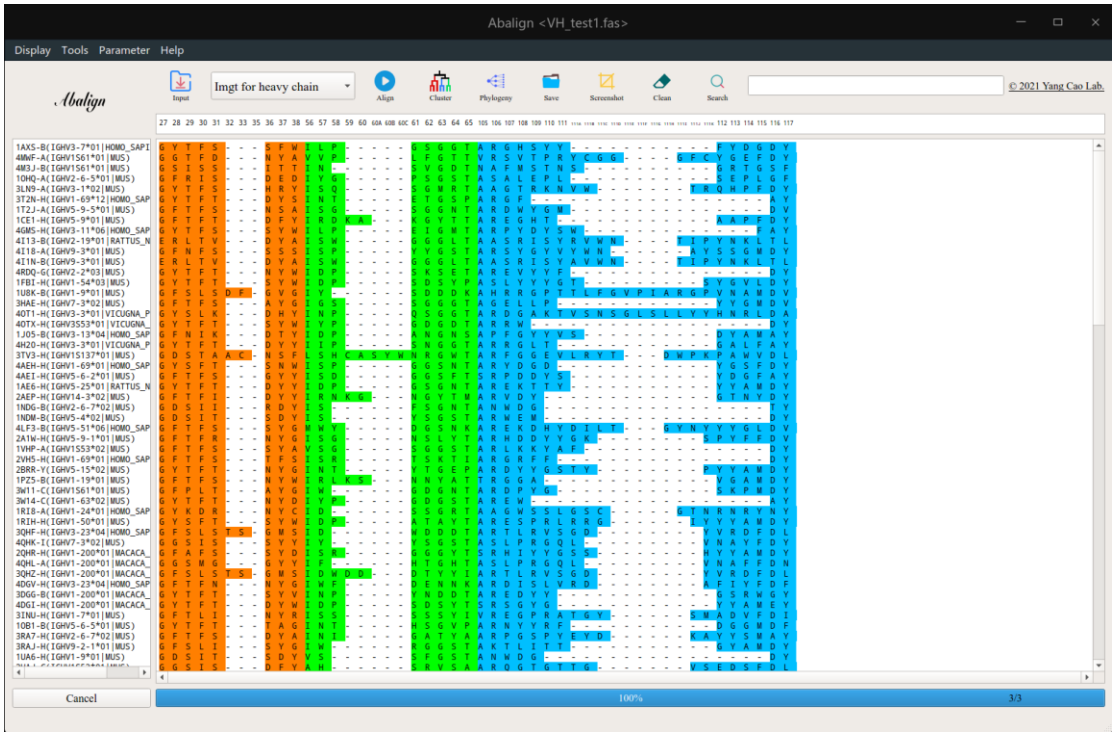


图 7. 仅显示抗体可变区的 CDR1,CDR2,CDR3 区域

工具

查找 V Gene：默认选择无法更改，可查找每条可变区序列最相似的 V Gene 及该 V Gene 所属的物种，并将 V Gene 名和物种名展示于对应的序列名后。

热图：点击 **Heatmap** 后可根据当前序列的相似性生成热图，并且可将热图进行聚类，若序列数量较大则需等待一段时间。所有的 Heatmap 均以当前多序列比对文本框的内容为基础，若使用 Display 更改了当前显示的序列，则 Heatmap 会发生相应的改变。

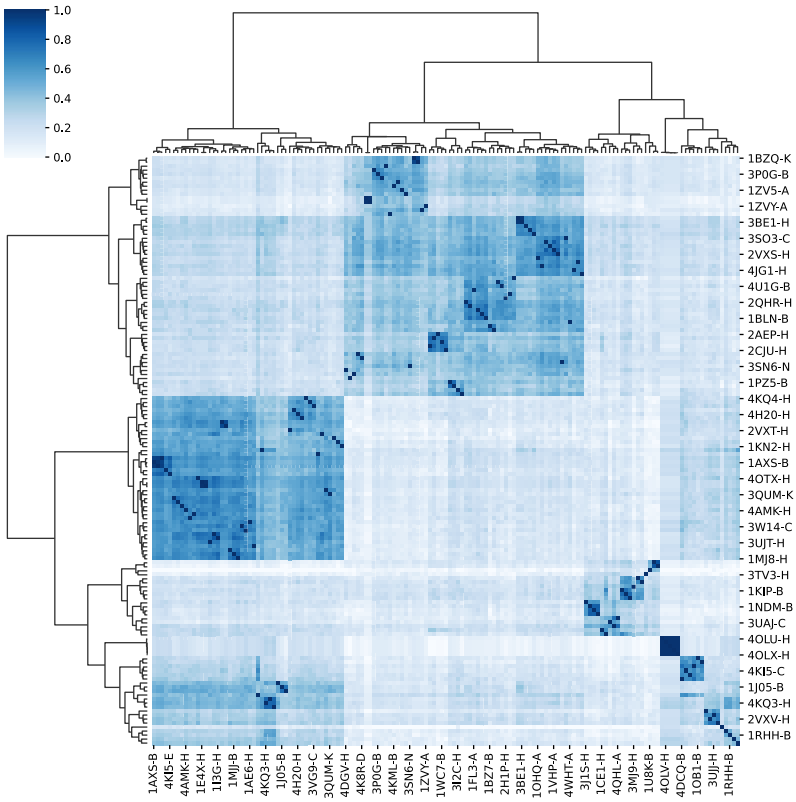


图 8. 经过层次聚类的热图

序列标识图：点击 **Tools**，将鼠标移动到 **Seqlogo** 上，点击 **By Entropy** 可获得以熵为纵坐标的 Seqlogo，点击 **By Frequence** 可获得以频率为纵坐标的 Seqlogo。将鼠标移动到 **Color** 选项上，可更改 Seqlogo 的渲染模式。所有的 Seqlogo 均以当前多序列比对文本框的内容为基础，若使用 Display 更改了当前显示的序列，则 Seqlogo 会发生相应的改变。为保证兼容性，所有 Seqlogo 均为 PDF 格式。

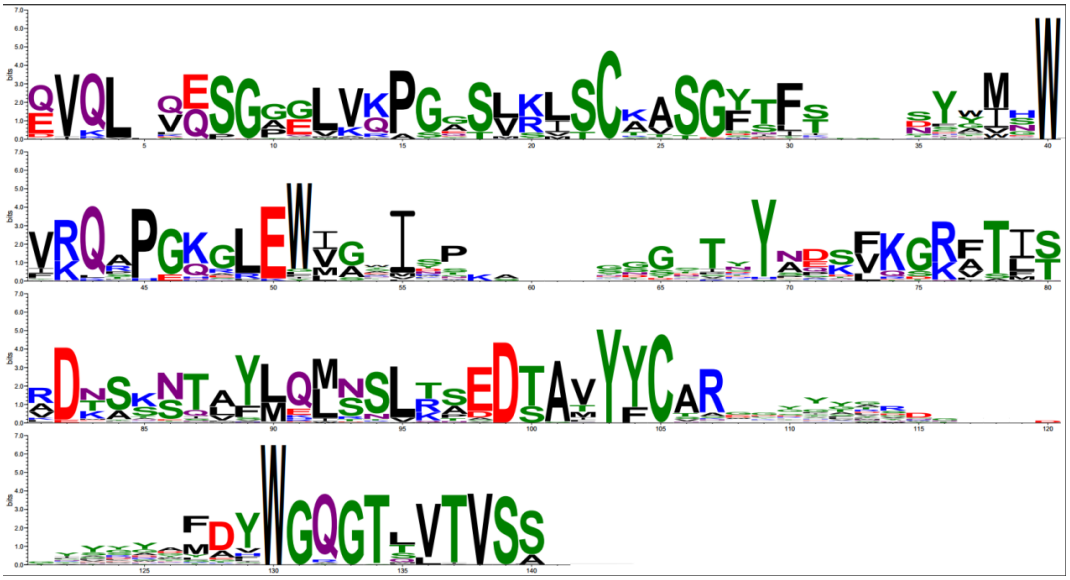


图 9. 整个抗体可变区的 Seqlogo，以熵值为 y 轴

丰度图：点击 **Tools**，将鼠标移动到 **Abundance** 上，可见 **V Gene abundance** 和 **Sequence abundance** 选项，点击后可生成 V Gene 丰度图和序列丰度图

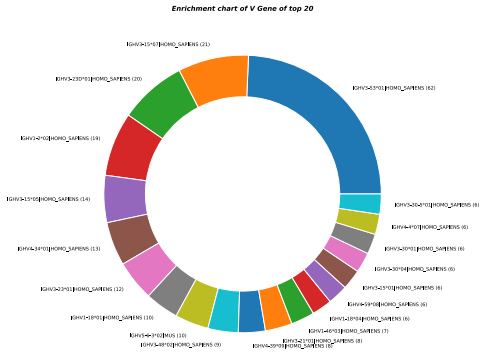


图 10. V Gene 丰度图

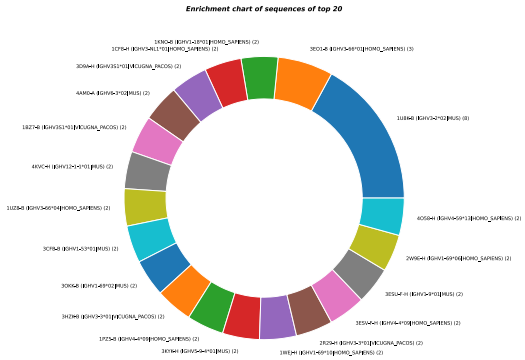


图 11. 序列丰度图

参数

点击菜单栏中的 **Parameter** 后，会弹出一个对话框，可在对话框中修改聚类罚分、聚类权重（如将 CDR3 区域权重增加，则聚类过程中会更加注意 CDR3 区域的相似度），和 Hierarchical Clustering Dendrogram 的各项参数。

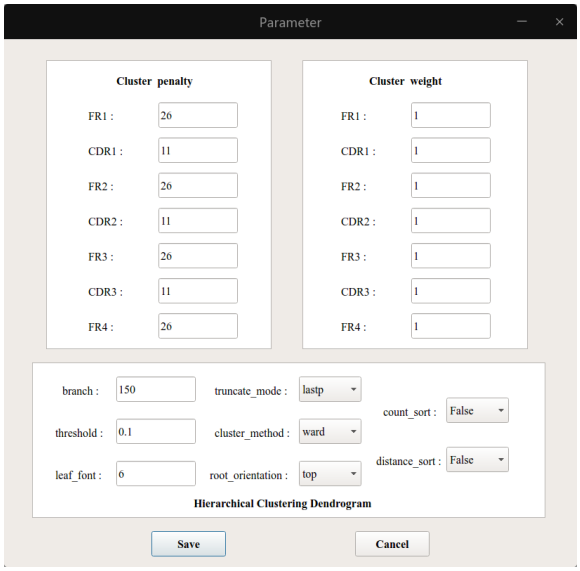


图 12. Hierarchical Clustering Dendrogram 参数

(a) branch: 代表层次聚类树显示的分支数量，该参数和 truncate_mode 参数一起生效

(b) threshold: 代表聚类划分的阈值，若该值为 0.1 则将 Y 轴总长（Y 轴总长代表序列间的最大差异）的十分之一作为阈值划分不同的类，聚类的序列数量越多，该值应该设的越小

(c) leaf_font: 表示分支标签字体的大小

(d) cluster_method: 表示聚类时计算两类之间距离所使用的方法；

(e) root_orientation: 表示树根所在方向

(f)count_sort: 表示是否将聚类结果按照每一类的数量进行排序；

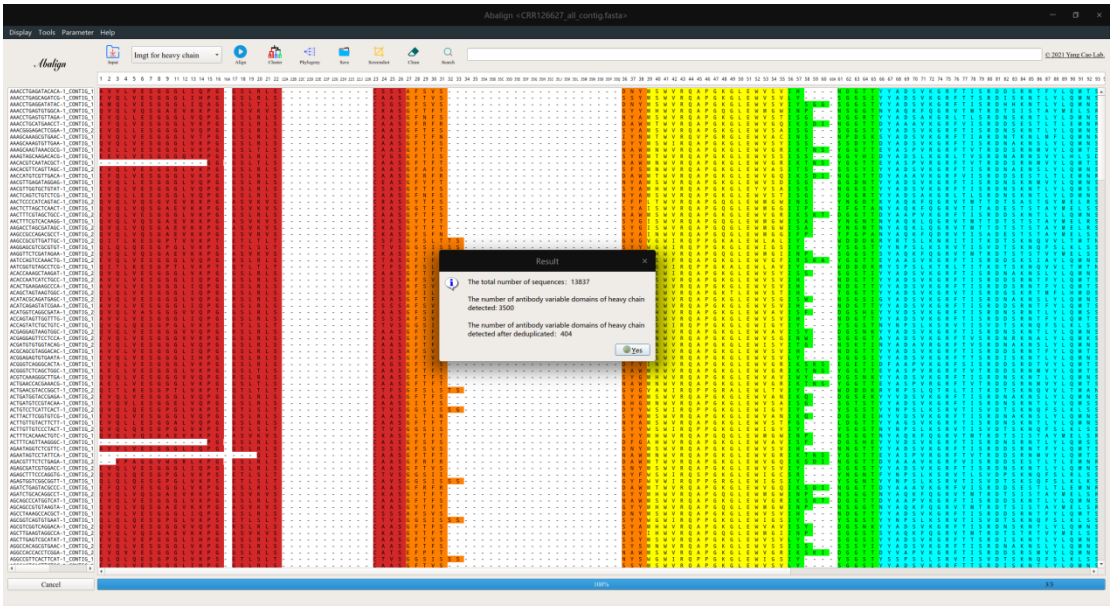
(g) distance_sort 表示是否将聚类结果按照每一类的距离进行排序。

实例

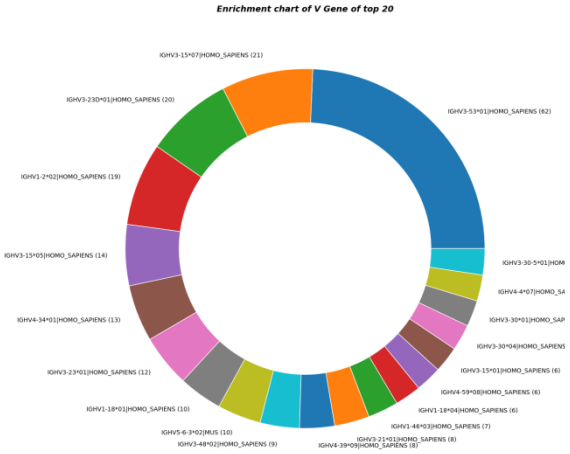
在本例中，我们使用新型冠状病毒患者单细胞测序所得的样本进行演示。

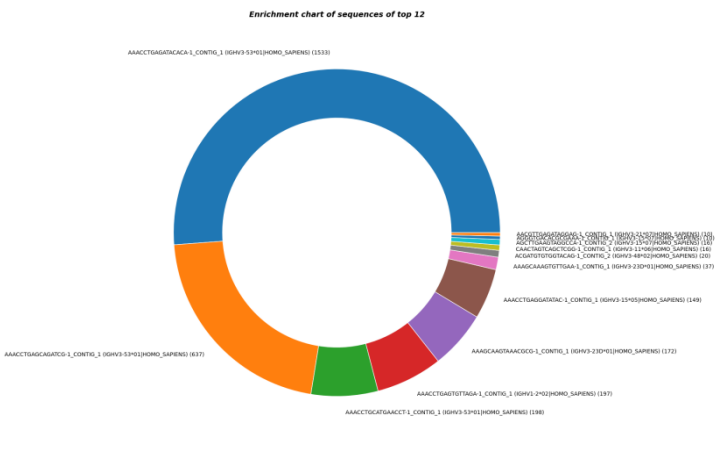
Step 1 输入文件: 点击 Input 后选择需要进行处理的 fasta 文件。

Step 2 寻找抗体可变区及多序列比对: 文件加载完成后，点击 Align 进行比较，程序下方进度条会显示程序进度，若文件过大，进度条短时间不更新是正常现象。比对完成后会弹出一个对话框，对话框中会显示输入序列总条数，检测到抗体可变区的序列条数和去重后的序列条数。

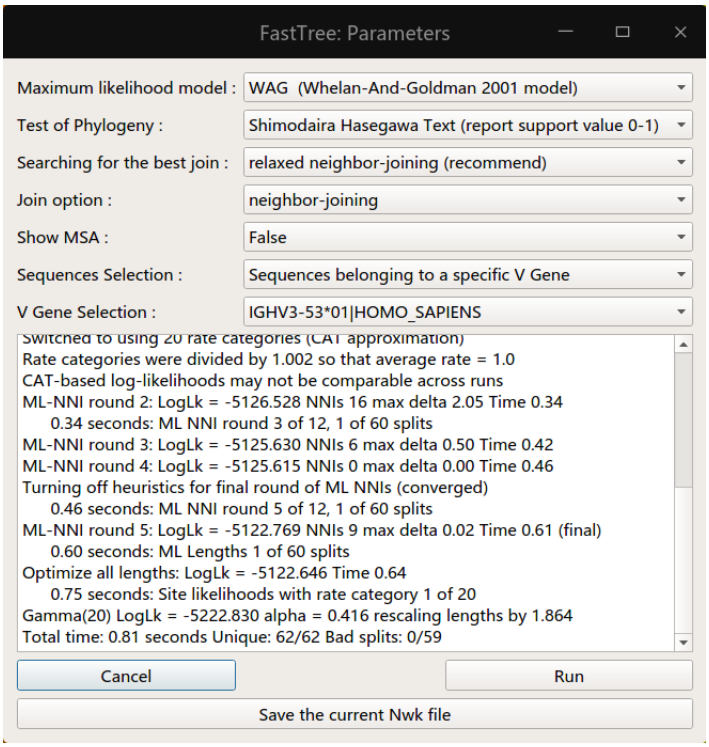


Step 3 查看多序列比对概况：点击顶部菜单栏中的 Tools 按钮，将鼠标移动到 Abundance 上，点击 V Gene abundance 或 Sequence abundance 可获得 V Gene 和序列的丰度图，抗体成熟过程中会产生超突变，最终成熟的抗体会在体内大量表达，因此研究丰度高的抗体是比较有意义的。

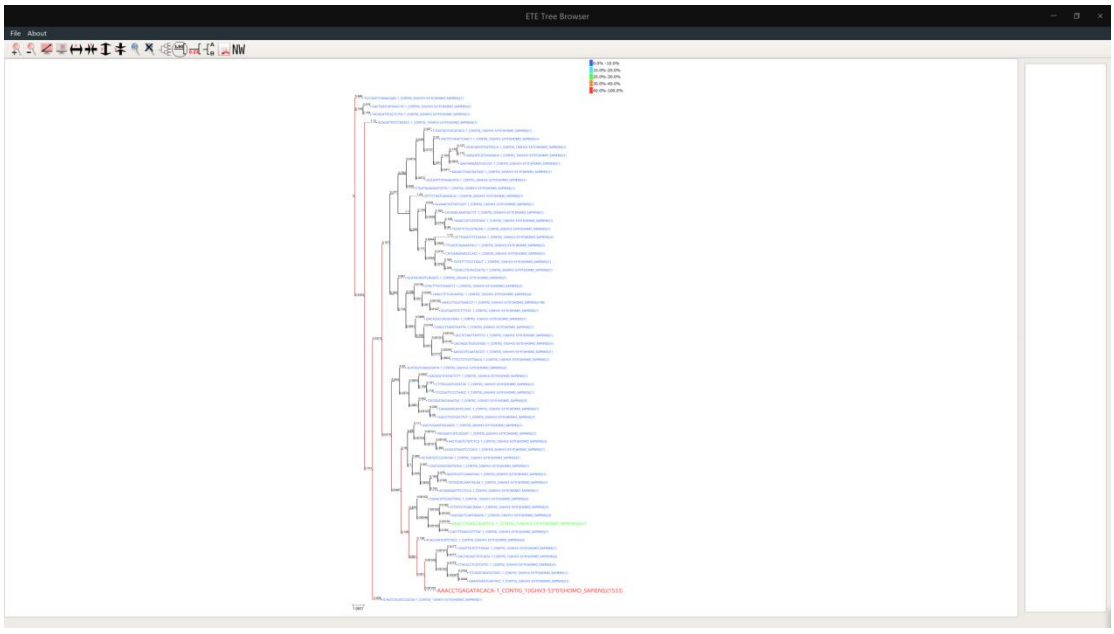




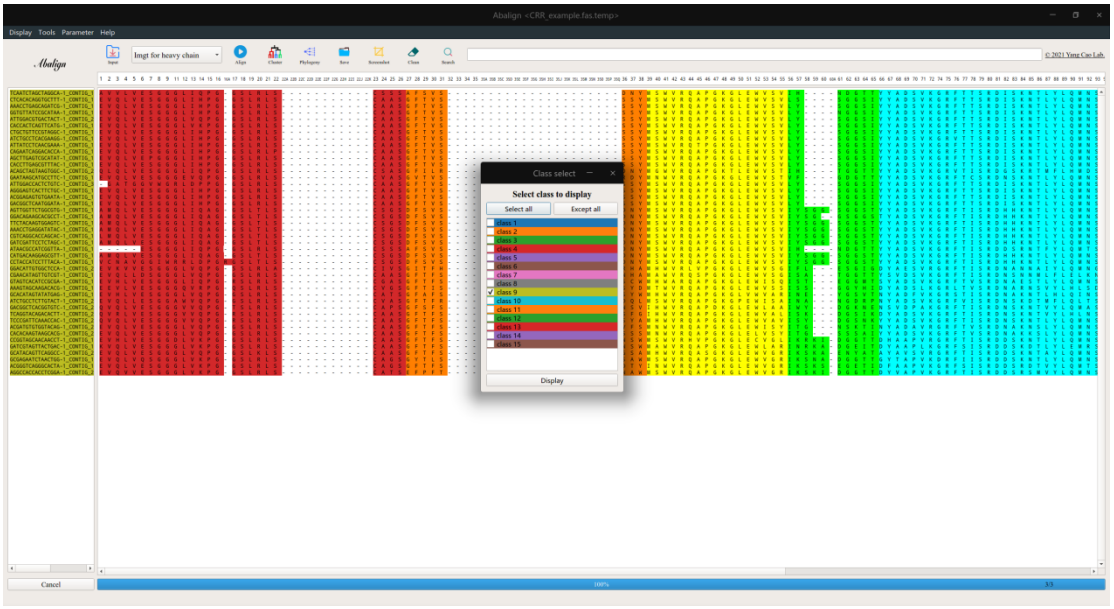
Step 4 构建系统发生树：根据序列丰度信息，以高丰度序列所属的 V Gene 进行建树。在本例中，丰度最高的序列所属的 V Gene 为”IGVH3-53*01”。点击 Phylogeny 打开建树的参数列表，将 Sequences Selection 选项切换为”Sequences belonging to a specific V Gene”，在 V Gene Selection 选项中选择”IGVH3-53*01”，点击 Run 开始构建进化树。如果你对建树结果满意，点击建树菜单中的”Save the current Nwk file”按钮，以保存当前系统发生树的 Nwk 文件。



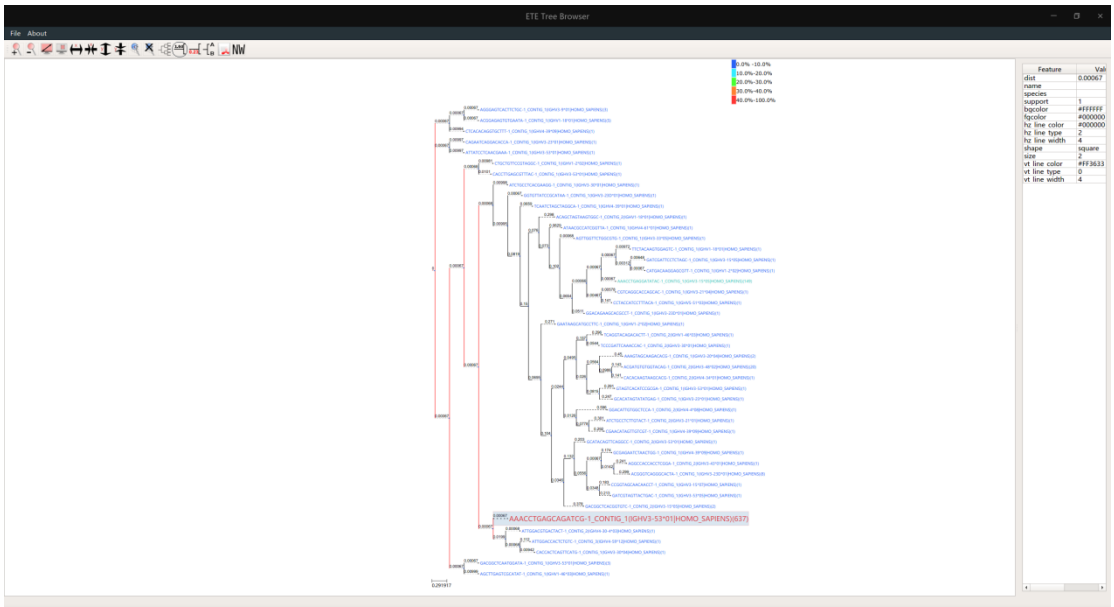
Step 5 查看系统发生树：进化树构建完成后，会弹出一个可视化窗口，可通过可视化窗口上方的按钮调整视角和显示树的其他信息，选择树的节点后还可在窗口的右方子窗口修改节点的属性，例如节点的颜色。



Step 6 层次聚类：点击 Cluster 后，可进行层次聚类，若文件过大会消耗较多的时间请耐心等待。聚类结束后会显示层次聚类的树状图，多序列比对会按照树状图的结果进行重排，序列名也会按照树状图分支的颜色渲染为相同颜色。聚类完成后可点击菜单栏中的 Display-->Display by cluster，选择需要显示的类。



Step 7 用不同的类构建系统发生树：在 Display by cluster 中选择需要用于建树的一类或多类后，再次点击 Phylogeny，将”Sequences Selection”调整为”All sequences in the class”，点击 Run 构建进化树。



注意事项

1. 若程序运行过程中出现未响应现象，请稍作等待，程序仍在运行。
2. 多序列比对是针对输入文件，若进行多序列比对后，再次进行多序列比对，还是对输入文件进行多序列比对，而非比对后序列。

3. 点击 Cancel 按钮只能终止比对的过程，最终加载比对后的序列时，无法通过 Cancel 取消。

4. 层次聚类在大数据量时十分消耗资源与时间，若想对大量序列进行层次聚类，请确保计算机有足够的内存。

5. 若软件画面显示不正常，请调整屏幕的缩放比例。

本软件由四川大学生命科学学院曹洋实验室开发，主要开发人员为宗凡杰，龙晨宇，胡万鑫，曹洋和肖智雄，如果您有任何意见或建议，请联系 cy_scu@yeah.net。

参考文献

[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.