

基于扩增子数据的系统发育树的构建和展示

Construction and display of phylogenetic tree based on amplicon data

周欣^{1,2}, 马紫英^{1,2}, 祁智慧³, 刘永鑫⁴, 蔡磊^{1,2,*}

¹真菌学国家重点实验室, 中国科学院微生物研究所, 北京; ²生命科学学院, 中国科学院大学, 北京;

³国家粮食和物资储备局科学研究院 北京; ⁴植物基因组学国家重点实验室, 中国科学院遗传与发育生物学研究所

*通讯作者邮箱: cail@im.ac.cn

摘要: 随着高通量测序技术的发展, 基于扩增子和宏基因组测序的微生物组学研究技术已经成为研究土壤、动植物及海洋等环境微生物多样性及功能的主要手段。基于扩增子的微生物组数据集, 往往能获得数千至上万个 OTUs (可操作分类单元), 我们需要从中筛选获得高丰度及核心微生物类群进行系统发育树的构建和展示。系统发育树又名分子进化树, 是生物信息学中描述不同生物或者不同基因之间进化关系的方法。通过系统学分类分析, 可以帮助研究者推测生物的进化历程和亲缘关系。本文主要介绍基于 IQ-TREE、MUSCLE、USEARCH10 等软件的下载安装、使用方法和步骤以及结果分析, 实现从扩增子数据集的提取、数据处理到系统发育树的构建和美化等流程, 方便研究者能更高效准确地实现基于扩增子数据的系统发育树构建以及下游系统发育树的编辑和展示, 为发表高水平研究论文提供技术支持。

关键词: OTUs, 系统发育树, 微生物多样性, 扩增子测序, iTOL

仪器设备

普通个人电脑 (Windows10 系统 64 位版、CPU ≥ 双核、内存 ≥ 4 G、硬盘 ≥ 20 GB)

软件和数据库

1. gitforwidnows 2.23.0 (<http://gitforwindows.org>)
2. R 4.0.3 (<https://www.r-project.org>)
3. Rstudio 1.2.5019 (<https://www.rstudio.com/products/rstudio/download>)
4. USEARCH v10.0.240 (<https://www.drive5.com/usearch/download.html>)
5. MUSCLE (<http://www.drive5.com/muscle/>)

6. IQ-TREE v2.0.3 (<http://www.iqtree.org>)

7. trimAL (<http://trimal.cgenomics.org/downloads>)

软件的安装和使用

一、首先在 C 盘根目录新建名为 bin 的目录，其具体位置为 C:\bin。

二、USEARCH 软件的下载和安装

USEARCH 软件 (<http://www.drive5.com/usearch/download.html>) 是 Robert C. Edgar 开发的一款超快的扩增子数据分析软件，在序列比对、OTU 聚类、多样性分析等多领域广泛应用 (Edgar, 2013)。

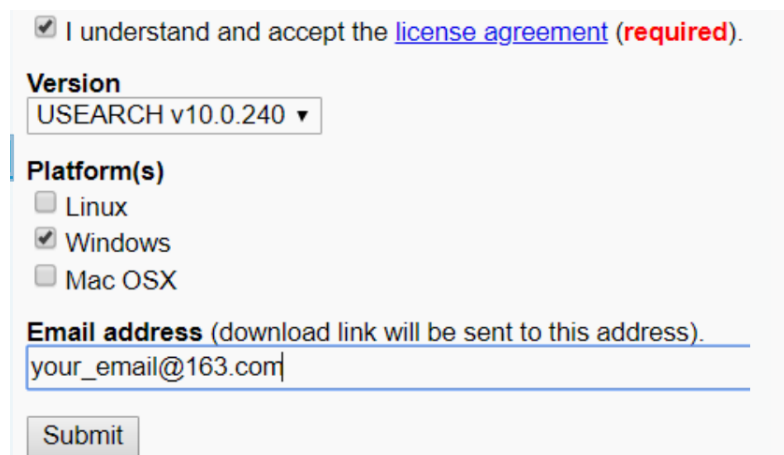


图 1. USEARCH10 软件的下载

注：32 位 usearch 为免费版，但限制使用内存 4GB，64 位为收费版本，没有内存使用限制。选择接受许可协议，版本必须选择 v10.0，选择填写邮箱，提交收到链接，下载后改名为 usearch10 并将软件放在 Windows10 系统中 C:\bin 目录中。

三、R 语言的下载安装和使用

R 语言是目前生物学、经济学等领域最流行的统计分析语言，下载最新版 R 语言 (下载页面: <https://cran.r-project.org>)；点击 Download R for Windows 完成 R-4.0.3.win.exe 安装程序的下载；双击安装程序，建议语言选择英文安装。



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

图 2. R 语言的下载

四、IQ-TREE 的下载安装和使用

IQ-TREE 软件是 2015 年发表的一款快速准确进行最大似然法 (Maximum Likelihood, ML) 构建系统发育树的软件 (Nguyen 等, 2015), 目前已经更新到 v2.2.2 版本。IQ-TREE 软件应用一种快速、有效的随机算法, 在近似的计算时间内具有比 RAxML 软件和 PhyML 软件更高的精确度。此外 IQ-TREE 软件的模型选择速度比 jModelTest 快 10-100 倍, 其自展支持率估算比 RAxML 软件快 10-40 倍并且支持宏基因组等大数据计算 (Minh 等, 2020)。IQ-TREE 软件的下载界面: (<http://www.iqtree.org>), 选择下载最新版 64 位的 IQ-TREE (v2.2.2) 软件 (如图 3), 解压后将其放在 C:\bin 目录中。



图 3. IQ-TREE 软件的下载

五、MUSCLE 的下载安装和使用

MUSCLE 软件是一款快速多重序列比对软件, MUSCLE 软件具有比 CLUSTALW 等软件更快的比对速度以及精确度, 它能在数分钟内完成数百条序列的比对。迄今, MUSCLE 软件已经被引用了超过 37000 次, 是生物学领域中最广泛使用的软件之一 (Edgar, 2004)。MUSCLE 软件的下载界面 (图 4):

(<http://drive5.com/muscle>), 选择下载最新版的 MUSCLE 软件, 解压后将其放在 C:\bin 目录中。

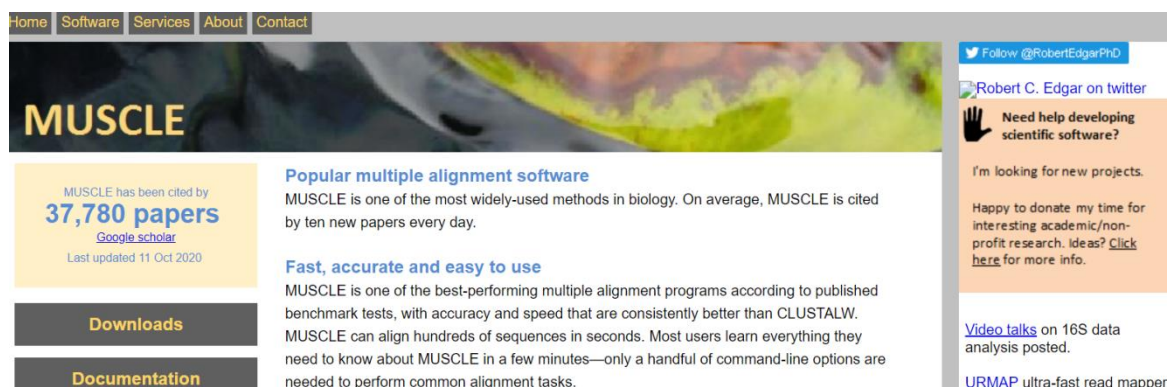


图 4. MUSCLE 软件的下载

六、trimAL 的下载安装和使用

完成精确的多序列比对后, 通常要过滤掉一些低质量以及高变异度的序列区域。trimAL 软件能快速, 精确切除和过滤低质量以及高变异度的序列, 仅保留进化保守的区域用于后续分析。

trimAL 软件的下载界面 (图 5): (<http://trimal.cgenomics.org/downloads>), 选择下载最新版的 trimAL 软件, 解压后将其放在 C:\bin 目录中。



图 5. trimAL 软件的下载

七、Rstudio 的下载安装和使用

Rstudio 的下载页面: (<https://rstudio.com/products/rstudio/download>)。从网页中选择下载最新版的 Rstudio, 如 RStudio Desktop 1.3.1093, 双击安装程序进行默

认安装。Rstudio 安装完成后，按如图所示步骤调出“Terminal”界面，然后在 Terminal 窗口中输入 ls (LS 的小写)，按回车进行测试，如果出现“command not found”错误，请按照下图重新进行操作和设置 (如图 6)。

- 安装好 “git for windows” 软件；
- 按步骤 (8) 配置 Windows 10 环境变量；
- 照右图中的箭头，设置“Terminal”为 “Git Bash” ；
- 点击 “Apply” 和 “OK” 按照下图，打开新的 Terminal；
- 如果还是无法调用可重启 Rstudio。

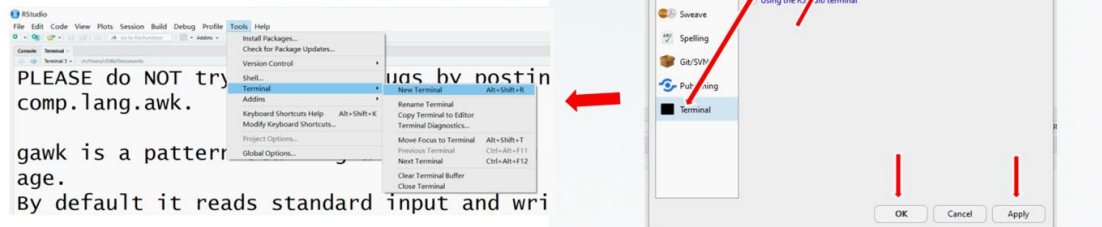


图 6. Rstudio 软件的设置及 Terminal 的调用

八、“git for windows”的下载安装和使用

“git for windows”软件 (v2.28.0) 是一款能在 Windows 系统下运行的命令行工具，能在 Windows 下运行部分 Linux 代码 (下载页面: <https://gitforwindows.org/>)，按照默认参数右键管理员安装 Git-2.28.0-64-bit.exe 即可(如图 7)。

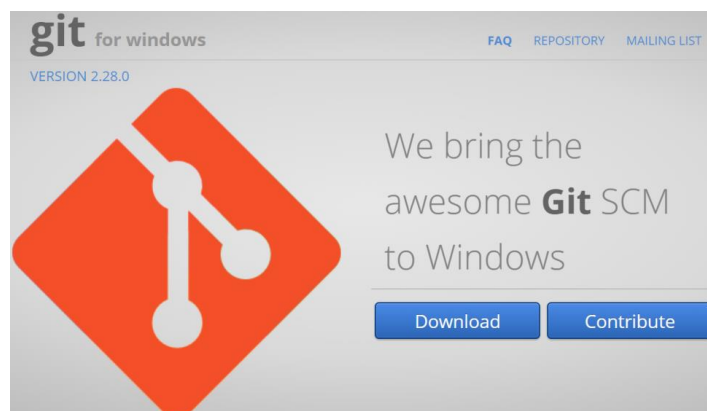


图 7. “git for windows”软件的下载

九、添加程序位置至 Windows 系统中的环境变量

我的电脑-右键属性-按右侧截图操作，测试是否安装成功：在 RStudio 的 Terminal 下输入：usearch10，按回车，如有出现 USEARCH10 的版本信息，则表明安装成功。若不成功，可检查环境变量配置，按下图进行操作 (如图 8)：

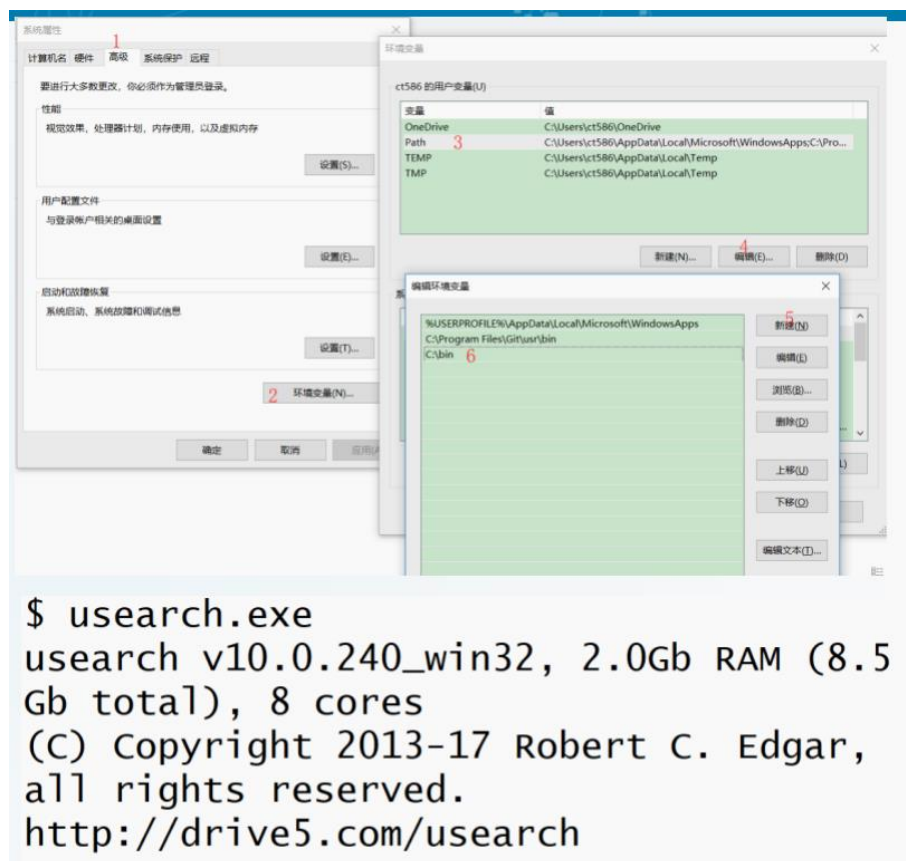


图 8. Windows10 系统的环境变量的设置

实验步骤

一、数据获得和操作流程概况

准备输入数据

本文中所有的测试数据集、所需软件、R 脚本以及生成的结果均已上传到百度网盘，如需测试和使用可点击如下链接进行下载和使用（百度网盘链接：链接：链接：https://pan.baidu.com/s/15MIJzV0_5kNV3VF_o0rKZw 提取码： 5ru9）。具体整个技术流程概况以及各个软件的功能与彼此之间的联系如图所示(如图 9)：



图 9. 系统发育树基本操作流程

二、数据处理相关的命令行操作均在 Rstudio “Terminal” 界面进行

1. 首先在 C 盘新建名为 Tree_data 目录

#切换到 Tree_data 目录中

\$ cd C:\Tree_data

\$ mkdir -p result/tree

\$ cd result/tree

2. 研究者可以根据实际情况 0.001 到 0.01 的阈值筛选高丰度 OTU。

#统计 OTU 表中 OTU 数量，代码如下：

\$ tail -n+2 ../otutab.txt | wc -l

#按相对丰度 0.2 %筛选高丰度 OTU，代码如下：

\$ usearch10 -otutab_trim ../otutab.txt -min_otu_freq 0.002 -output otutab1.txt

#统计筛选 OTUs 表特征数量，代码如下：

\$ tail -n+2 otutab1.txt | wc -l

#提取 ID 用于提取序列，代码如下：

\$ cut -f 1 otutab1.txt | sed '1 s/#OTU ID/OTUID/' > otutab_high.id

3. 在进行完 OTU 筛选后要根据 OTUs 的 ID 提取每个 OTUs 对应的 fasta 格式的代表性序列，手动整理物种注释和分组信息信息表 annotation.txt，如图 10 所示。

#筛选高丰度菌/指定差异菌对应 OTUs 的代表性序列，代码如下：

```
$ usearch10 -fastx_getseqs ../otus.fa -labels otutab_high.id -fastaout high_otus.fa
$ head -n 10 annotation.txt
```

OTUID	Kingdom	Phylum	Class	Order	Family	Genus	Species	KO	OE	WT	All
ASV_657	Bacteria	Actinobac	Actinobac	Actinomyc	Thermom	Unassigne	Unassigne	3.34	3.78	5.19	4.10
ASV_2	Bacteria	Proteobac	Betaprote	Burkholde	Comamor	Pelomona	Pelomona	5.68	2.65	3.35	3.89
ASV_3	Bacteria	Proteobac	Gammapr	Pseudomyc	Pseudomyc	Rhizobact	Rhizobact	2.21	2.38	3.17	2.59
ASV_12	Bacteria	Bacteroid	Flavobact	Flavobact	Flavobact	Flavobact	Flavobact	2.18	2.84	2.67	2.56
ASV_4	Bacteria	Proteobac	Gammapr	Pseudomyc	Pseudomyc	Rhizobact	Unassigne	2.27	1.66	2.58	2.17
ASV_8	Bacteria	Actinobac	Actinobac	Actinomyc	Streptomy	Streptomy	Unassigne	1.17	2.15	1.80	1.71
ASV_6	Bacteria	Proteobac	Betaprote	Burkholde	Unassigne	Unassigne	Unassigne	2.64	1.04	1.42	1.70
ASV_18	Bacteria	Actinobac	Actinobac	Actinomyc	Pseudomyc	Lentzea	Lentzea_fl	1.38	1.39	1.69	1.49
ASV_9	Bacteria	Actinobac	Actinobac	Actinomyc	Micromor	Actinoplat	Unassigne	1.27	1.38	1.38	1.34

图 10. annotation.txt 文件中包含的内容

三、序列对齐及系统发育树的构建

#构建进化树，实现高丰度菌的进化树的分组信息展示与美化。

#起始文件为 result/tree 目录中 high_otus.fa (序列)、annotation.txt (物种和相对丰度)文件

Muscle 软件进行序列比对和对齐，代码如下：

```
$ cd Tree_data/result/tree
$ muscle -in high_otus.fa -out otus_aligned.fa
```

#trimAL 软件进行低质量以及高变异度的序列的过滤和修剪，代码如下：

```
$ trimal -in otus_aligned.fa -out otus_aligned_trimed.fa -gt 0.95
```

#利用 IQ-TREE 软件进行 ML 系统发育树的构建，代码如下：

```
$ mkdir -p iqtree
$ iqtree -s otus_aligned_trimed.fa -bb 1000 -redo -alrt 1000 -m MFP -nt AUTO -
pre iqtree/training_otus
```

#参数简介：

-m 参数：指定模型选项，MFP 表示 ModelFinder Plus(自动默认)

-redo 参数：之前运行成功后生成了相应的文件，指定 redo 会重新跑一遍覆盖之前的文件；

-pre 参数：将结果输入到 iqtree 文件夹中，且生成文件的前缀为 training_otus；

-alrt 参数：是否启用 SH-aLRT 检验。

四、iTOL 网站进行系统发育树的编辑和展示

在运用 iTOL 在线工具来进行系统发育树的美化之前，首先要使用“table2itol.R”这个 R 包 (<https://github.com/mgoecker/table2itol>) 生成用于系统发育树编辑和美化的注释文件。然后访问并登陆 iTOL 网站 (<http://itol.embl.de/>)，上传 otus.nwk，再拖拽以下命令行生成的不同注释文件（分别位于“plan1”、“plan2”和“plan3”三个文件夹）于 iTOL 主界面的系统发育树图上即完成系统发育树的美化。

plan1 生成外圈颜色、形状分类和丰度文件，代码如下：

```
$ cd Tree_data/result/tree
```

```
$ Rscript ../../script/table2itol.R -a -c double -D plan1 -i OTUID -l Genus -t %s -w 0.5 annotation.txt
```

plan2 生成丰度柱形图注释文件，代码如下：

```
$ Rscript ../../script/table2itol.R -a -d -c none -D plan2 -b Phylum -i OTUID -l Genus -t %s -w 0.5 annotation.txt
```

plan3 生成热图注释文件，代码如下：

```
$ Rscript ../../script/table2itol.R -c keep -D plan3 -i OTUID -t %s otutab.txt
```

#参数简介

-a：找不到输入列将终止运行（默认不执行）；

-c：将整数列转换为 factor 或具有小数点的数字；

-t：偏离提示标签时转换 ID 列；

-w：颜色带，区域宽度等；

-D：输出目录；

-i：OTUs 列名；

-l: OTUs 显示名称如种/属/科名。

注：当需要标注的颜色过多时，R 脚本会采用形状+颜色的方式对类别进行区分。

结果与分析

1. IQ-TREE 运行完成后会在 iqtree 文件下生成多个文件，主要包括程序运行日志 training_otus.log、ML 树文件 (含有 UFBoot 或 BP/SH-aLRT 评估分支置信度) 和系统发育树文件 training_otus.contree，本文测试数据中生成的系统发育树文件名称为 training_otus.contree (如图 11)。

training_otus.bionj	2019/7/15 12:33	BIONJ 文件	3 KB
training_otus.ckp.gz	2019/7/15 12:34	好压 GZ 压缩文件	77 KB
training_otus.contree	2019/7/15 12:34	CONTREE 文件	4 KB
training_otus.iqtree	2019/7/15 12:34	IQTREE 文件	34 KB
training_otus.log	2019/7/15 12:34	文本文档	14 KB
training_otus.mldist	2019/7/15 12:33	MLDIST 文件	99 KB
training_otus.splits.nex	2019/7/15 12:34	NEX 文件	24 KB
training_otus.treefile	2019/7/15 12:34	TREEFILE 文件	5 KB

图 11. IQ-TREE 软件建树生成的结果文件

2. 首先进入 iTOL 在线网站，点击右上角注册 (仅限新用户) 和登陆，(如图 12):

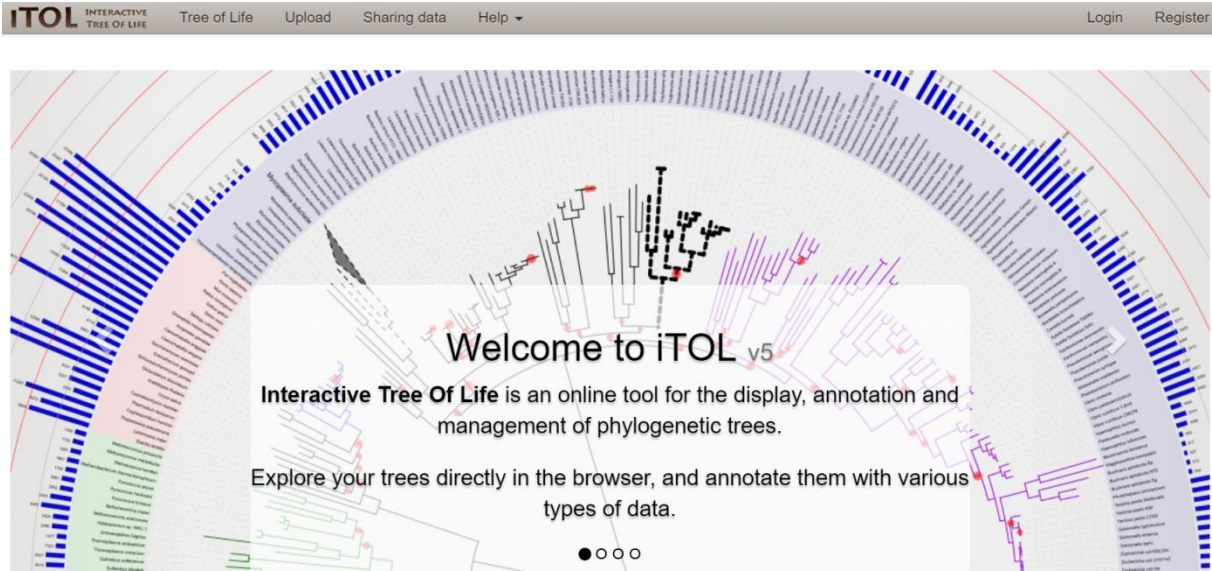


图 12. iTOL 在线网站主界面

3. 完成账号登录后，点击 My Tree 按钮，进入个人主界面，然后点击“Upload tree files”上传文件夹中的 training_otus.contree 文件，获得如下树形图 (图 13)。导入

树文件之后，可以在 iTOL 在线网站的右上角选择下图红框中的“Basic”和“Advanced”进行系统发育树的编辑（比如树形的变换、自展支持率的显示、字体大小和颜色的调整、分支的位置变换等）。iTOL 在线网站具有非常强大的系统发育树的编辑和美化功能，研究者可根据自己的需求进行各种个性化的调整，关于编辑的具体使用方法可以参考 iTOL 的官方帮助文档 (<https://itol.embl.de/help.cgi>) 和 iTOL 官方视频资料 (https://itol.embl.de/video_tutorial.cgi)。

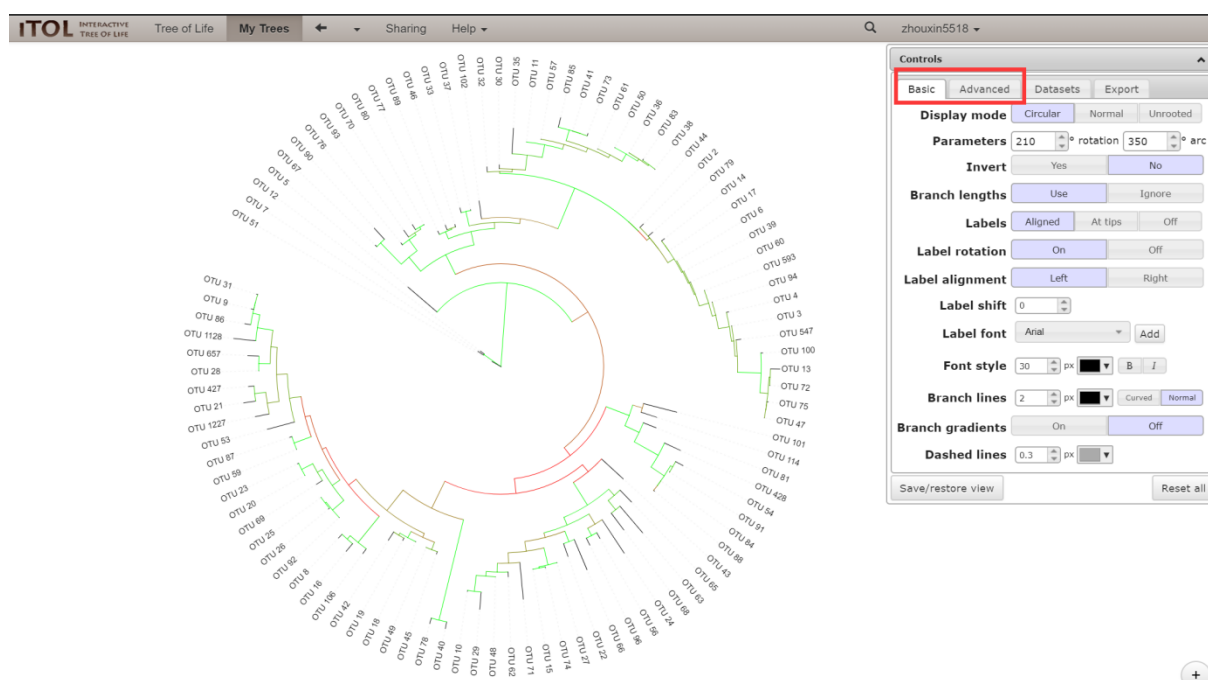


图 13. iTOL 在线网站的系统发育树展示和编辑

4. 系统发育树的编辑：按住鼠标左键，将“plan1”文件夹中的“iTOL_labels-Genus.txt”文件拖到 iTOL 网页的当前主界面上，iTOL 在线网站会自动将所有 OTUs 替换成其物种注释对应的属名，如图所示（图 14）。

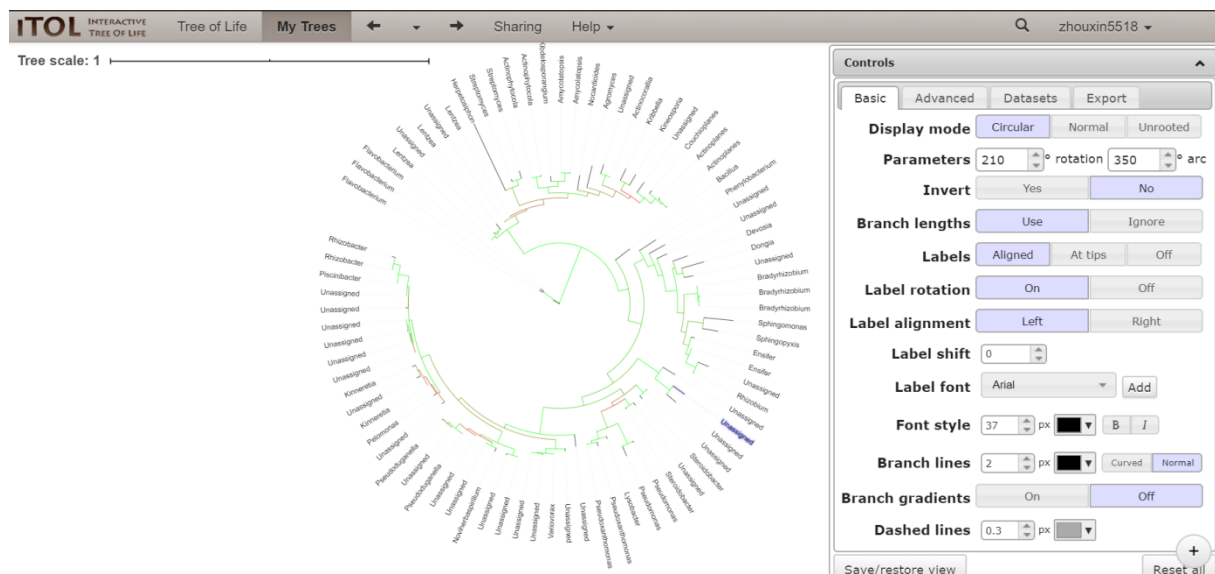


图 14.系统发育树中分支节点名称替换

5. 系统发育树的按微生物门分类水平进行着色编辑：按住鼠标左键，将“plan2”文件夹中的“iTOL_treecolors-Phylum.txt”文件拖到 iTOL 网页的当前主界面上，iTOL 在线网站会自动按微生物门对系统发育树进行着色”，如图所示（图 15）。同时注意，此处着色可分别进行标签、分支和全树着色。

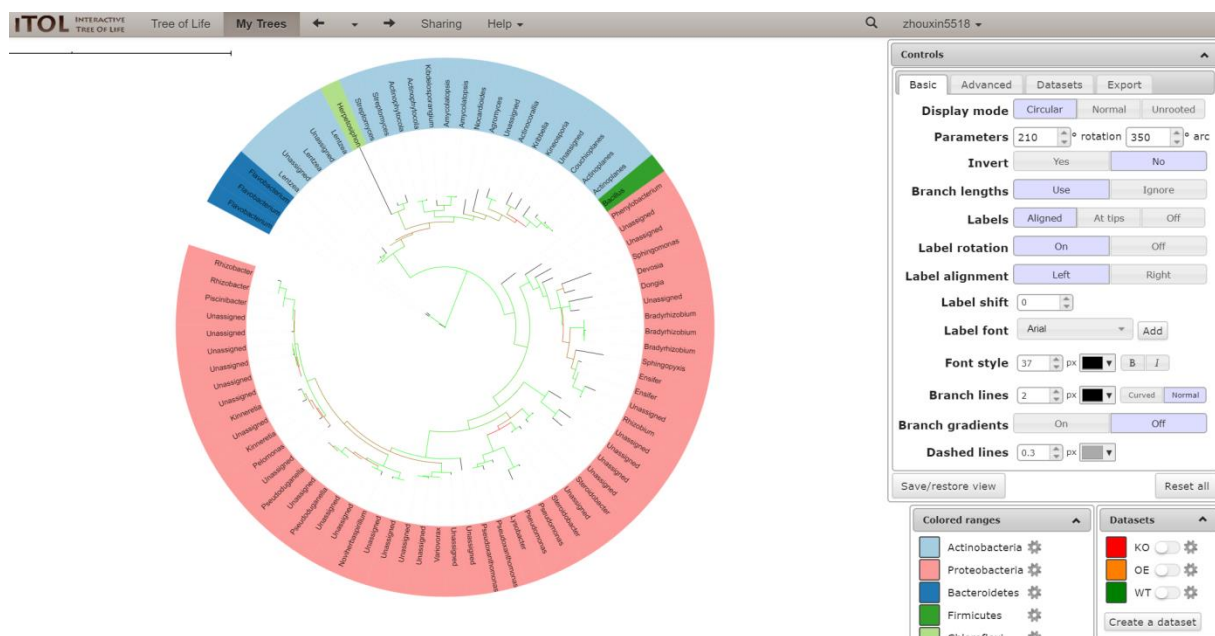


图 15.系统发育树图按门水平进行着色

6. 添加其它系统发育树分组注释：使用三个不同分组的高丰度 OTUs 做进化树，可

以把三个分组的高丰度 OTUs 的相对丰度，用柱状图形式进行展示。按住鼠标左键，将“plan2”文件夹中的“iTOL_simplebar-A.txt”、“iTOL_simplebar-B.txt”和“iTOL_simplebar-C.txt”文件拖到 iTOL 网页的当前主界面上，iTOL 在线网站会自动将“KO”、“OE”和“WT”三个分组的高丰度 OTUs 用柱形图形式进行展现（如图 16）。

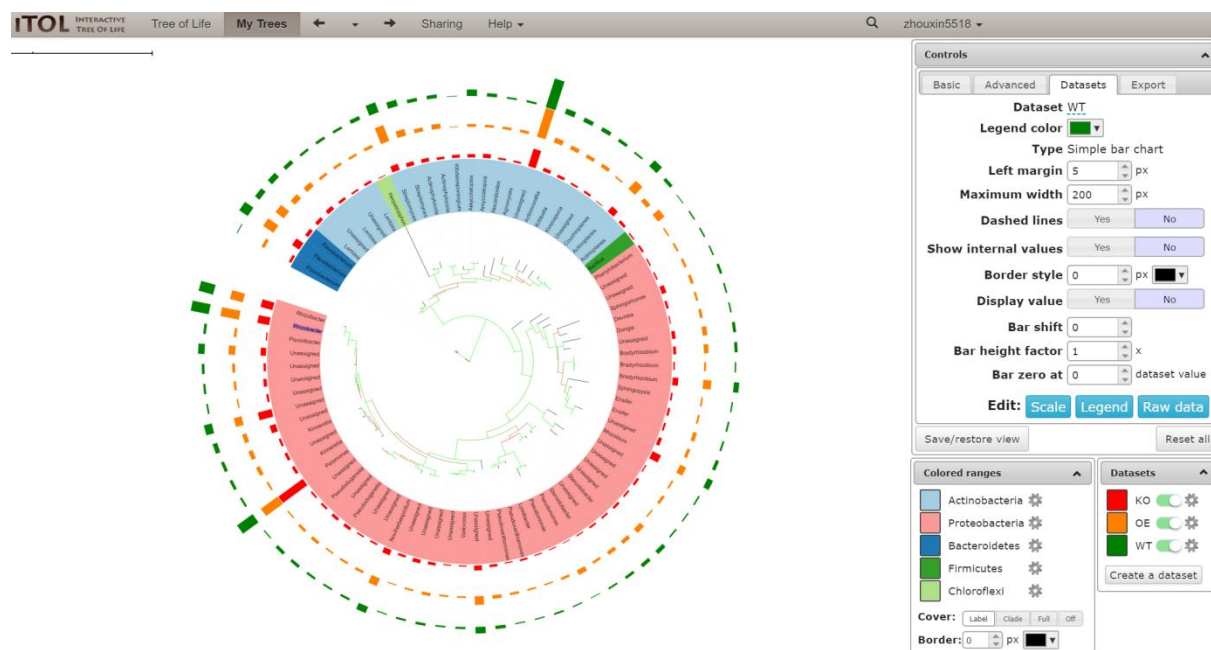


图 16.添加高丰度 OTUs 不同分组柱形图的系统发育树图

7. 在“plan1”、“plan2”和“plan3”文件夹中有很多其它系统发育树美化和编辑的文件。例如，“plan3”文件夹中的热图的添加等，研究者可以根据自己的需求进行灵活添加，最终达到自己的系统发育树展示和研究目的。最后，研究者可以点击 iTOL 在线网站右上角的“Export”按钮导出编辑完成的系统发育树图。本文最终生成的系统发育树图，如下图所示（图 17）。

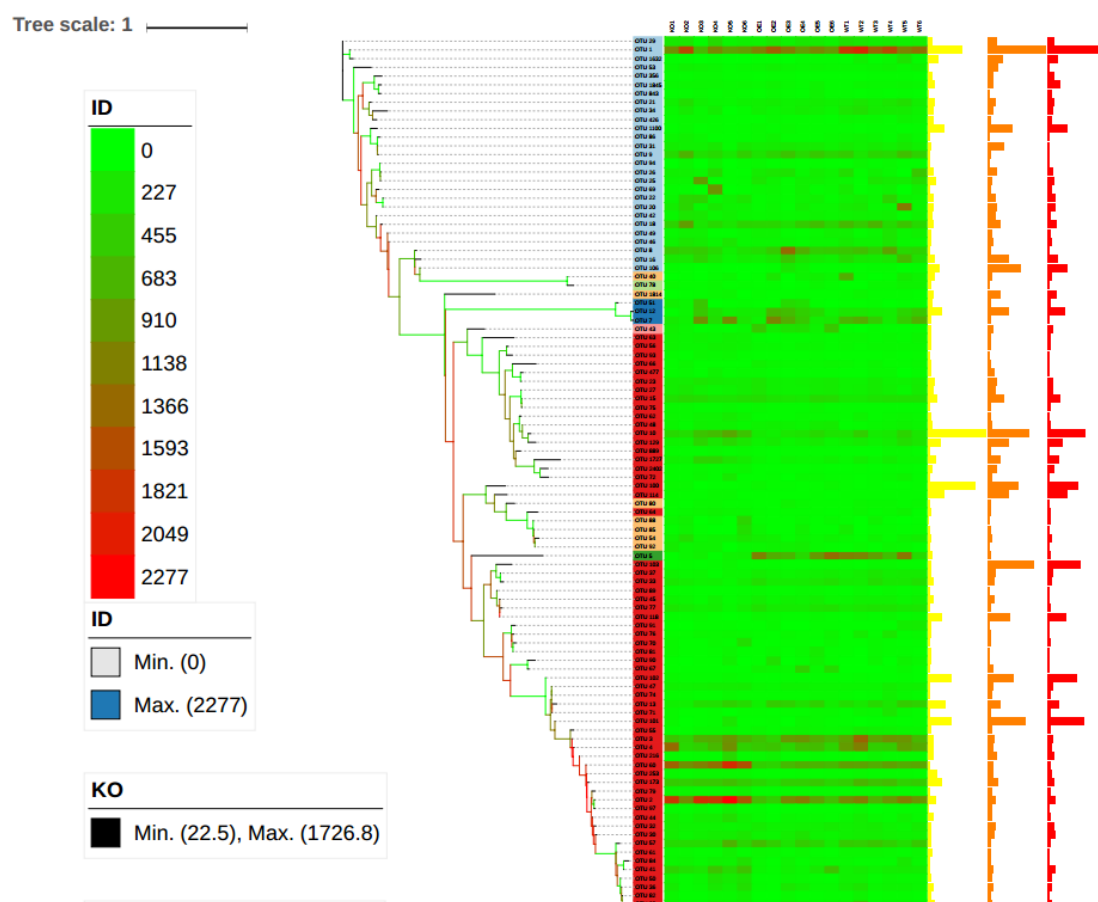


图 17.最终生成的系统发育树图

小结

本文简要介绍了微生物扩增子数据中的高丰度 OTUs 数据的筛选，代表性序列及对应物种注释的获取，以及系统发育树的构建方法。展示了一套完整操作流程，以帮助研究者学习和使用生成接近发表质量要求的系统发育树图的构建方法。研究者在建树过程中可以根据研究领域参考同行文献对系统发育树进行一些细节参数的调整，并使用 AI (Adobe Illustrator)对特殊图形、字符和树进行进一步修改和美化。

致谢

感谢“git for windows”软件开发者提供的 Git 软件(<https://github.com/git-for-windows>)及 mgoeker 在 GitHub 网站上开发和公开分享的 R 语言包“table2itol”(<https://github.com/mgoeker/table2itol>)。本文分析方法已应用于待发表文章“Distribution and variations of mycotoxin producing fungal community in major rice production areas of China”。

参考文献

1. Edgar, R. C. (2004). [MUSCLE: multiple sequence alignment with high accuracy and high throughput](#). 32(5):1792–1797.
2. Edgar, R. C. (2013). [UPARSE: highly accurate OTU sequences from microbial amplicon reads](#). *Nature Methods*. 10(10): 996–998.
3. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. and Lanfear, R. (2020). [IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era](#). *Molecular Biology and Evolution*. 37(5): 1530–1534.
4. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. and Minh, B. Q. (2015). [IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies](#). *Molecular biology and evolution*. 32(1): 268–274.