* 1. 16S rDNA扩增子数据分析实践

本16S rRNA基因测序数据分析流程包括：首先对样本测序数据进行质控与预处理；然后得到样本的OTU聚类与物种分类信息；在查看样品的细菌种类后，统计每个样品的差异菌种。

此流程所有命令都需在WSL-Ubuntu的bash终端下运行（以下命令前面的“$”代表bash提示符，不需要输入）。运行Ubuntu之前，先在电脑C盘创建目录win16s，并将样本测序数据文件(raw\_reads.zip)放在此目录下。

注意：所有输入文件名，根据自己保存的文件名调整。

* + 1. 相关软件与文件

(1)安装VSEARCH

$git clone https://github.com/torognes/vsearch.git

$cd vsearch

$./autogen.sh

$./configure

$make

$sudo make install #默认安装在/usr/local/bin/目录

(2) 安装RDP classifier

1)下载RDP Classifier(只处用版本v2.2)，下载网址：https://sourceforge.net/projects/rdp-classifier/files/rdp-classifier/

2)解压到windows系统的C盘win16s目录：

$unzip rdp\_classifier\_2.2.zip

$ls /mnt/c/win16s/rdp\_classifier\_2.2/

3)下载RDP Gold数据库(rdp\_gold.fa)

$wget -c <http://drive5.com/uchime/rdp_gold.fa>

* + 1. 分析流程运行命令

(1)准备测序数据(raw reads)

本实验需要把所有原始数据放在一个文件夹，一个样品的双端测序有两个reads文件，默认以"\_R1"与"\_R2"分别代表正向与反向测序reads （如果文件名不对，需要重新命名）。

$cd /mnt/c/win16s/ #切换到工作目录

$unzip raw\_reads.zip

$ls raw\_reads

显示文件列表，如SAA\_R1.fastq.gz, SAA\_R2.fastq.gz等六个样本的两端测序数据的压缩文件。测序公司提供的数据一般已去除barcode和primer序列。

(2)合并双端测序序列

如果样品数量较小，可以对每个样品分别进行合并，一次合并一个样品的两端测序数据：

$vsearch --fastq\_mergepairs raw\_reads/SAA\_R1.fastq.gz --reverse raw\_reads/SAA\_R2.fastq.gz --fastqout SAA.merged.fastq

$vsearch --fastq\_mergepairs raw\_reads/SAB\_R1.fastq.gz --reverse raw\_reads/SAB\_R2.fastq.gz --fastqout SAB.merged.fastq

# Merging reads 100%

28741 Pairs

19633 Merged (68.3%)

9108 Not merged (31.7%)

如果样品数量较多，可将以上处理命令写到一个shell脚本，一次运行合并多个样品数据。

(3)序列质量控制

一般可先用FastQC检查测序reads的质量信息，注意测序reads长度。本练习所用测序序列为16S rRNA v3-v4区片段大约为400bp左右。这里将合并后reads长度< 380 bp的reads都过滤掉。

$vsearch --fastq\_filter SAA.merged.fastq --fastq\_maxee 1.0 -fastq\_minlen 380 --fastq\_maxns 0 --fastaout SAA.filtered.fasta

$vsearch --fastq\_filter SAB.merged.fastq --fastq\_maxee 1.0 -fastq\_minlen 380 --fastq\_maxns 0 --fastaout SAB.filtered.fasta

#33624 sequences kept (of which 0 truncated), 7616 sequences discarded.

(4)序列去重复(Dereplication)

$vsearch --derep\_fulllength SAA.filtered.fasta --sizeout --minuniquesize 2 --output SAA.derep.fasta

$vsearch --derep\_fulllength SAB.filtered.fasta --sizeout --minuniquesize 2 --output SAB.derep.fasta

#1371 uniques written, 7560 clusters discarded (84.6%)

如将miniuniqusize值设成大于2，可去除低丰度序列，增加计算速度

(5)嵌合体检测(remove chimera)

嵌合体的产生主要是由于PCR过程中模版的不完全延伸。这一步是去除嵌合体，一般需要参考数据库序列文件：

$vsearch --uchime\_ref SAA.derep.fasta --db rdp\_gold.fa --sizein --sizeout --nonchimeras SAA.nochimeras.fasta

$vsearch --uchime\_ref SAB.derep.fasta --db rdp\_gold.fa --sizein --sizeout --nonchimeras SAB.nochimeras.fasta

如果没有参考序列，VSEARCH也可以通过de novo方法去除嵌合体：

$vsearch --uchime\_denovo SAA.derep.fasta --sizein --sizeout --nonchimeras SAA.nochimeras.fasta

(6)OTU聚类(OTU clustering)

OTU (operational taxonomic units)是人为给某一个分类单元(科、属、种等)设置的标志。通常按照97%的相似性阈值将序列划分为不同的OTU，每一个OTU通常被认为是一个微生物物种。选取最长的序列作为OTU的代表。

$vsearch --cluster\_fast SAA.nochimeras.fasta --id 0.97 --centroids SAA.otus.fasta --relabel OTU\_ --sizein --sizeout

$vsearch --cluster\_fast SAB.nochimeras.fasta --id 0.97 --centroids SAB.otus.fasta --relabel OTU\_ --sizein --sizeout

以序列最小相似度97%聚类，得到Clusters: 121

(7)生成OTU表格(Map reads back to the OTU data)

$vsearch --usearch\_global SAA.nochimeras.fasta --db SAA.otus.fasta --id 0.97 --otutabout SAA.otutab.txt

$vsearch --usearch\_global SAB\_nochimeras.fasta --db SAB.otus.fasta --id 0.97 --otutabout SAB.otutab.txt

(8)物种分类注释

采用RDP classifier 将OTU注释为不同物种分类。

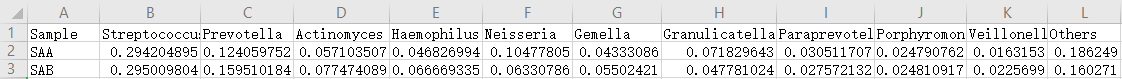
$java -Xmx1g -jar rdp\_classifier\_2.2/rdp\_classifier-2.2.jar -q SAA.otus.fasta -o SAA\_otus\_tax.txt -f fixrank

$java -Xmx1g -jar rdp\_classifier\_2.2/rdp\_classifier-2.2.jar -q SAB.otus.fasta -o SAB\_otus\_tax.txt -f fixrank

注意：这里采用Ribosomal Database Project classifier及其默认数据库做OTU物种分类注释。

(9)OTU物种分类数据整理

在WPS或EXCEL打开上一步得到的OTU分类注释数据\*\_otus\_tax.txt，整理成文件(otus\_taxa.csv)，其中行为样品名，列为菌的属名。只取丰度最高的前10个菌属，其它菌属都合并为“Others”。



(10)菌群丰度可视化

运行R可视化命令前，先将样本文件otus\_tax.csv放到目录（C:/bigbook）。

#Plot relative abundance for all samples

#Stacked Bar Plots in R

library(ggplot2)

library(reshape2)

library(scales)

#set your working directory by either setwd()

setwd("c:/bigbook/")

#upload your data to R

data <- read.table("otus\_taxa.txt", header = T, sep = "\t")

head(data)

#data <- data[,c(0:12)] # select 11 genus

#convert data frame from a "wide" format to a "long" format

datam = melt(data, id = c("Sample"))

datam$value <- datam$value \* 100

#make the plot!

n <- length(levels(datam$variable))

#check n = 11?

cols <- hue\_pal(h = c(0,360) + 15, c = 100, l = 65, h.start = 0, direction = 1)(n)[c(11,7,1,3,5,9,2,6,4,8,10)]

ggplot(datam, aes(x = Sample, y = value)) +

geom\_bar(aes(fill = variable), stat = "identity") +

labs(x = "Sample", y = "Relative abundance (%)") +

scale\_fill\_manual(name = "Genus", values = cols)

查看各个分类水平的菌株分布信息。

