**病原微生物检测分析平台 PathFinder**

**技术手册 V1.1**

（2019 年 11 月 15 日）

**一 软件的总体设计描述：**

本技术手册是为宏基因组测序分析平台搭建和使用的手册。主要目的是利用高通量测序方法对微生物进行宏基因组测序，通过搭建生物信息学分析流程，来检测样本中的病原微生物。主要内容包括相应分析软件安装、微生物参考基因组序列整理和分析脚本说明。

**二 总体结构和模块设计**

# 总体结构包括数据分析环境搭建、参考基因组构建和分析流程搭建。

1. **数据分析环境搭建**

整个平台操作环境是基于Linux系统环境下搭建的。本平台建议安装Centos 7版本。具体操作如下：

1. 查看具体操作系统环境

进入Linux系统，具体命令为：

cat /etc/redhat-release

可获得当年平台的具体操作系统型号和版本。

1. 建立系统平台工作目录

在系统平台工作目录下，建立PathFinder文件夹，作为当前平台的工作目录。具体命令为:

pwd //获得当前工作目录

cd $pwd //进入当前工作目录

mkdir PathFinder //建立PathFinder文件夹，作为平台工作目录。

1. 建立系统平台所需软件源代码存放目录

在系统平台工作目录下，建立src文件夹，作为系统平台所需软件源代码存放目录。具体命令为：

cd $pwd //进入当前工作目录

mkdir src //建立src文件夹，作为平台软件源代码存放目录。

1. 建立系统平台所需软件安装目录

在系统平台工作目录下，建立software文件夹，作为系统平台所需软件安装目录。具体命令为：

cd $pwd //进入当前工作目录

mkdir software //建立src文件夹，作为平台所需软件安装目录。

1. 系统平台所需软件安装配置

(5).1 测序序列质控软件Trimmomatic下载安装：

具体在Linux系统环境下操作命令为：

网址：http://www.usadellab.org/cms/?page=trimmomatic

解压：unzip Trimmomatic-\*\*.zip

拷贝：cp -rf Trimmomatic-\*\* ./software

(5).2 BWA 下载配置安装

具体在Linux系统环境下操作命令为：

网址：http://bio-bwa.sourceforge.net/

解压：tar -jxvf bwa-\*a.tar.bz2

进入：cd bwa-\*a

编译：make

拷贝：cp bwa ./software/bwa

(5).3 samtools下载配置安装

具体在Linux系统环境下操作命令为：

下载：http://samtools.sourceforge.net/

解压：tar -jxvf samtools-\*.tar.bz2

进入：cd samtools-\*

编译：

./configure --prefix=$workdir/samtools

make all all-htslib

make install install-htslib

(5).4 bedtools 下载配置安装

具体在Linux系统环境下操作命令为：

下载：http://bedtools.readthedocs.io/en/latest/

解压:tar -jxvf bedtools2-\*.tar.bz2

进入：cd bedtools2

编译：make

拷贝：cp -r bin ./software/bedtools/

(5).5 taxonkit 下载配置安装

具体在Linux系统环境下操作命令为：

下载：

http://bioinf.shenwei.me/taxonkit/

解压：tar -xzvf taxonkit\_linux\_amd64.tar.gz

拷贝：cp taxonkit ./software/taxonkit

(5).6 下载配置安装 taxdump：

下载: ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz

解压:

mkdir dmp

Tar -xzvf taxdump.tar.gz -C dmp

copy "names.dmp" and "nodes.dmp" to data directory: "~/.taxonkit"

1. **参考基因组构建**

2.1人类基因组序列下载

下载地址：https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml

文件：GRCh38\_latest\_genomic.fna.gz 文件

存放路径：

./H\_sapiens

解压缩：

gunzip GRCh38\_latest\_genomic.fna.gz 文件

建立 bwa index 文件：

bwa index GRCh38\_latest\_genomic.fna

1.2 细菌基因组序列下载整理

构建工作目录：

./NCBI\_complete\_GbBac

下载步骤：

1. Get the list of assemblies

wget ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/assembly\_summary.txt

1. Parse the addresses of complete genomes

awk -F '\t' '{if($12=="Complete Genome") print $20}' assembly\_summary.txt > assembly\_summary\_complete\_genomes.txt

1. Make dir for data

mkdir GbBac

mkdir GbBac\_cds

mkdir GbBac\_rna

1. Make download.sh to get data

for next in $(cat assembly\_summary\_complete\_genomes.txt); do wget

-P GbBac "$next"/\*genomic.fna.gz; done

1. Remove cds and rna

Mv GbBac/\*cds\_from\_genomic.fna.gz GbBac\_cds

Mv GbBac/\*rna\_from\_genomic.fna.gz GbBac\_rna

6)Extract data

zcat GbBac/\*.gz > Gb\_Bac.fa

1.3 去除质粒序列：

1. Get all sequence ID

grep ">" Gb\_Bac.fa >GbAc\_IDs.txt

1. Remove plasmid ID

grep "plasmid" -v GbAc\_IDs.txt > GbAc\_IDs\_noplasmid.txt

1. Get sequences without plasmid by rmPlasmidSeq.pl

perl rmPlasmidSeq.pl

1. Copy to new 新文件夹 RefGenome

cp Gb\_Bac\_noplasmid.fa ./ RefGenome

1. 建立 bwa index 文件：

bwa index Gb\_Bac\_noplasmid.fa

1.4 细菌系统发育分类注释

下载 genbank ID 对应物种信息：

https://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/nucl\_gb.accession2taxid.gz

存放目录：

./NCBI/taxonomy

提取细菌序列对应物种 ID：

perl GetSeqTaxID.pl

cut seqTaxID.txt > taxids.txt

提取细菌序列对应物种分类树：

taxonkit lineage -t taxids.txt

taxonkit reformat lineage.txt | cut -f 1,3 > lineageRF.txt

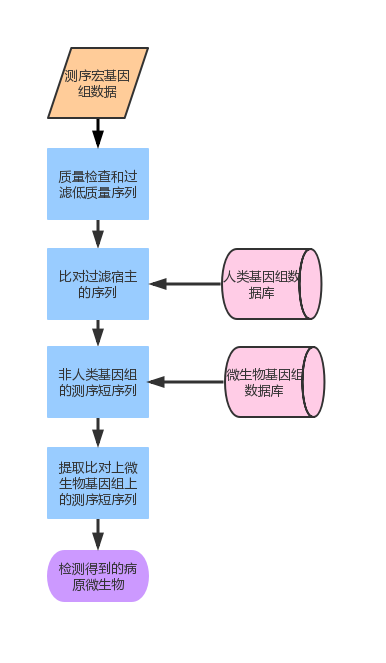
sort -u lineageRF.txt > lineageRF.txt

整理生成序列对应物种分类注释文件：GetSeqTaxon.pl

perl GetSeqTaxon.pl > Gb\_Bac\_Taxon.txt

1. **分析流程搭建**

**软件流程设计图如下：**

****

软件名： PathFinder

主要程序包：PathFinder.sh praseResult.R

主要用法：

Usage: ./DoPathFinder.sh <inputFileName> <FolderNameForSeq>

<inputFileName>: 一个文件里面包含输入文件的名称

例如：vi input.txt

cat input.txt:

filename\_R1.fq filename\_R2.fq filename

<FolderNameForSeq> : 一个文件夹包含输入的fastq文件

mkdir data

ls data:

filename\_R1.fq filename\_R2.fq

输出文件：<InputFile>\_SepeciesCount.csv (top 20 species)

 具体分析步骤：

1. “Trimmomatic” to remove adaptors and low quality reads

去除接头和低质量测序短序列

1. “bwa” to mapping clean reads to human genome

与人类基因组进行比对

1. “samtools” to screen non-human reads

寻找未必对上人类基因组的测序短序列

1. “bedtools” to get fastq file with the filterd non-human reads

通过过滤，得到非人类基因组上的测序短序列文件

1. “bwa” to mapping non-human reads to microbiome genomes

将得到非人类测序短序列与微生物基因组序列进行比对

1. “samtools” to get unique mapping reads

提取比对上微生物基因组的测序短序列

1. praseResult.R to get the final results

得到最终结果

具体使用代码见流程脚本 PathFinder.sh

结果文件：

“SampleID”\_SpeciesCount.csv

格式

"物种名", "匹配该物种 reads 条数", "界", "门", "纲", "目", "科", "属", "种"