本流程共包含5个独立的脚本。最终目的是生成各个样品的核基因序列。

其大致的原理为：首先将某个样品 (假设为样品a) 的原始测序数据 (reads) 回帖 (map) 至参考基因组上，之后使用gatk软件检测样品a在参考序列各个位点上应该是何种碱基类型，同时将检测的结果输出成一条新的fasta序列，该序列与参考基因组拥有完全相同的长度。所有待检测的样品都会经过以上步骤，并生成条与参考基因组长度完全相同的序列，最后，将这些序列合并成一个文件，既是建树矩阵。

**基本运行环境配置：**

本流程所有脚本均需在Linux操作系统环境下运行。脚本使用Python3语言编写，常规Linux发行版均自带该语言解释器。这里推荐安装Python的模块管理软件Anaconda (https://www.anaconda.com)。如果为Anaconda添加Bioconda源，则可以非常方便的安装很多生物信息学软件。配置方法如下：

Bash终端中输入：

conda config --add channels bioconda

conda config --add channels conda-forge

**依赖软件安装：**

(1) **bwa：**使用该软件将原始测序数据回帖至reference上面。

**安装：**conda install bwa -c bioconda

(3) **samtools：**处理bwa软件生成的bam文件，删掉可以map到基因组两个不同位置的read (这种区域在基因组内不是单拷贝的，所以要删除掉)

**安装：**conda install samtools -c bioconda

(2) **gatk：**根据回帖之后的bam文件，推算精确的各个位点碱基信息

**安装：**conda install gatk4 -c bioconda

(2) **biopython：**读取fasta文件

**安装：**conda install biopython -c bioconda

**需要准备的输入文件：**

1. 参考基因组 (reference) 文件：需要fasta格式。建议其中包含的序列 (contig) 条目不要太多，一些太短的contig可以直接删除掉，只留下组装比较完整的，染色体级的contig。但如果参考基因组拼装的结果较差，里面碎片化的序列较多，则可以直接将这些碎片的序列连接至几条较长的序列。

2. 高通量二代测序原始数据文件：需要双端测序 (Pair-end) 的fastq格式 (或者压缩后的fastq格式) 文件。

**具体步骤：**

进入脚本所在的文件夹后，按顺序运行以下命令：

(1) 构建参考基因组 (reference) 的索引文件。

将包含参考基因组的fasta文件拷贝入脚本所在文件夹中 (示例文件名为“ref.fasta”)，在终端中使用如下命令运行脚本：

python 1\_index\_reference.py -r ref.fasta

该命令为构建参考基因组文件“ref.fasta”的索引文件

(2) 批量将测序文件回帖 (map) 至参考基因组上。

将所有的样品原始序列数据拷贝入本文件夹中 (示例文件名为sp1\_1.fastq、sp1\_2.fastq、sp2\_1.fastq、sp2\_2.fastq等)，在终端中使用如下命令运行脚本，脚本会批量将所有样品分别回帖 (map) 至参考序列上。回帖完成之后，脚本会自动调用samtools软件，将基因组中多拷贝的部分删除。

python 2\_bwa\_map.py -r ref.fasta -t 6 -p 1.fastq -m 2.fastq

参数“-r”为参考序列文件名，与上一步骤相同；参数“-t”为允许程序使用的线程，根据计算机实际情况进行设置；参数“-p”为正向测序数据独特的文件后缀，本例中为“1.fastq”，其含义为，所有正向测序数据都是以这串符号结尾的，脚本将批量处理所有符合这种设定的数据文件；参数“-m”为反向测序数据独特的文件后缀。

(3) 批量运行gatk软件。

脚本将使用gatk软件，批量处理上一步骤产生的所有bam格式文件，确定各个位点上明确的碱基类型。使用如下命令运行脚本：

python 3\_gatk.py -r ref.fasta -t 4

参数“-r”为参考序列文件名，与上一步骤相同；参数“-t”为允许程序使用的线程，根据计算机实际情况进行设置，但这一步骤消耗的运行内存较高，如果不是在大型的服务器中运行，建议线程数设置不要超过4。

(4) 将gatk软件生成的vcf文件转化为fasta格式的单条序列。

脚本将批量处理上一步骤中产生的所有vcf文件，将其转化为单条的fasta序列。使用如下命令运行脚本：

python 4\_get\_fasta.py -r ref.fasta

参数“-r”为参考序列文件名，与上一步骤相同；脚本最终生成的文件名为“contigs名称+样品名称”。

至此，所有的参考基因组中的序列所对应的样品的序列都已经生成好。只需要使用cat等命令分别合并相应染色体的序列即可生成矩阵。