华大基因
BGI

# 项目简报

# 目录

# 1 项目信息

本项目的基本信息见下表:

| 项目信息 | 描述 |
| --- | --- |
| 项目编号 | F18FTSSCKF3059_MUSapkM |
| 测序区域 | 16S-V4 |
| Tag数据量 | 50000 |
| 注释数据库 | GreenGene |

表1 项目信息

# 2 实验流程

取质量合格的基因组DNA样品30ng及对应的融合引物配置PCR反应体系，设置PCR反应参数进行PCR扩增，使用Agencourt AMPure XP磁珠对PCR扩增产物进行纯化并溶于Elution Buffer，贴上标签，完成建库。使用Agilent 2100 Bioanalyzer 对文库的片段范围及浓度进行检测。检测合格的文库根据插入片段大小，选择HiSeq平台进行测序。

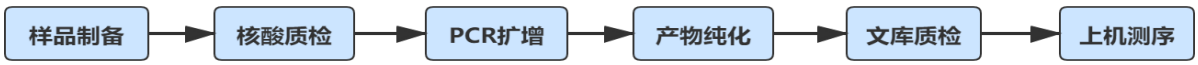


图1 实验流程.

# 3 信息分析流程

下机数据过滤，剩余高质量的Clean data用于后期分析；通过reads之间的overlap关系将reads拼接成Tags；将Tags聚类成OTU并与数据库比对、物种注释；基于OTU和注释结果进行样品物种复杂度分析，组间物种差异分析，以及关联分析与模型预测等。

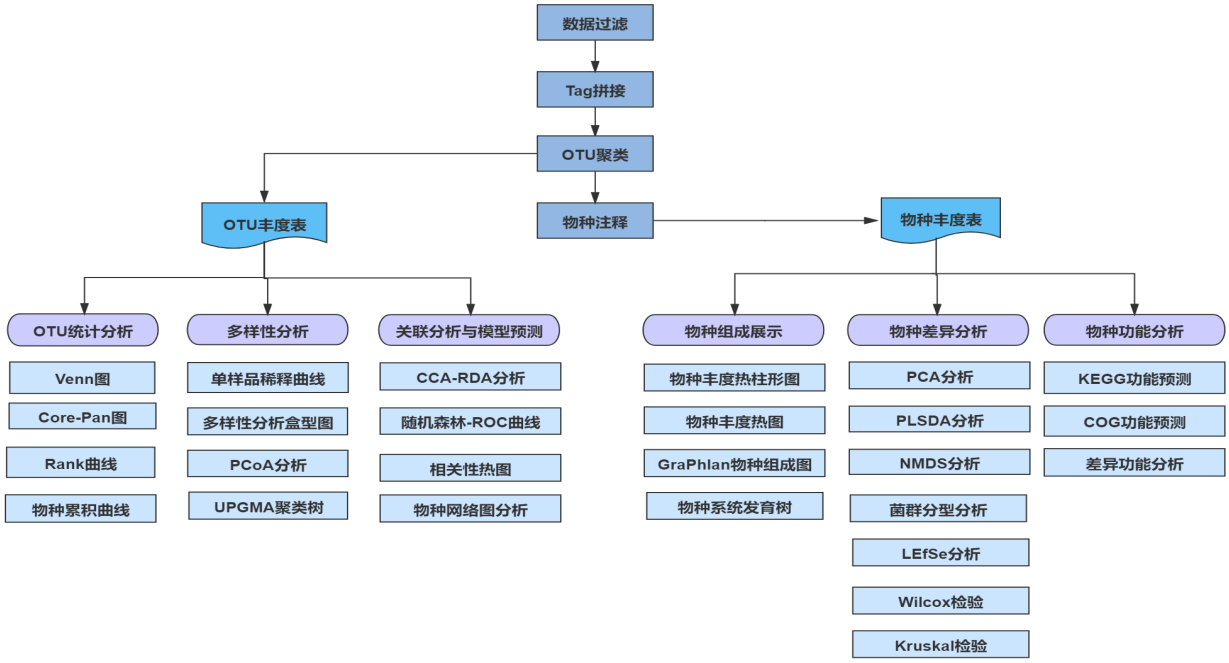


图2 信息分析流程.

## 4 数据过滤

对原始的测序数据进行如下处理，获得Clean Data，具体步骤如下：

1)去除质量值≤20的碱基数达到20%的reads；

2)去除含N的reads；

3)将能匹配到引物的reads，截取掉引物和接头污染，得到最终的Cleandata。

| sample_name | read_length | rawData_size | adapter_percent | NBase_percent | ployBase_percent | lowQuality_percen |
|---|---|---|---|---|---|---|
| S3 | 298:298 | 41.03 | 0.0 | 0.065 | 2.884 | 3.549 |
| Y1 | 300:298 | 43.87 | 0.0 | 0.101 | 6.869 | 4.771 |
| D0 | 296:299 | 41.63 | 0.0 | 0.054 | 3.109 | 5.494 |
| Y6 | 299:297 | 43.69 | 0.0 | 0.096 | 5.706 | 6.488 |
| Y9 | 300:296 | 47.16 | 0.0 | 0.031 | 7.135 | 10.929 |
| S15 | 293:295 | 40.88 | 0.0 | 0.022 | 0.931 | 5.05 |
| Y30 | 299:294 | 42.46 | 0.0 | 0.018 | 0.791 | 6.926 |
| L1 | 299:298 | 44.05 | 0.0 | 0.085 | 5.957 | 6.208 |
| Y3 | 296:298 | 43.13 | 0.0 | 0.068 | 4.502 | 7.165 |
| L15 | 298:295 | 40.04 | 0.0 | 0.01 | 0.775 | 3.807 |
| L12 | 294:295 | 46.17 | 0.0 | 0.01 | 5.88 | 10.824 |
| S30 | 293:294 | 43.48 | 0.0 | 0.017 | 1.599 | 8.116 |
| L30 | 298:294 | 44.4 | 0.0 | 0.016 | 3.482 | 8.651 |
| D15 | 300:295 | 47.13 | 0.0 | 0.03 | 4.956 | 13.852 |
| L75 | 298:293 | 39.66 | 0.0 | 0.01 | 0.515 | 3.51 |
| S1 | 297:293 | 41.28 | 0.0 | 0.014 | 2.253 | 5.044 |
| L9 | 299:296 | 41.89 | 0.0 | 0.025 | 1.494 | 6.573 |
| Y20 | 295:294 | 43.71 | 0.0 | 0.023 | 2.849 | 8.024 |
| L50 | 294:293 | 42.2 | 0.0 | 0.011 | 2.758 | 6.582 |
| Y75 | 299:293 | 40.65 | 0.0 | 0.023 | 1.268 | 4.776 |
| D6 | 300:297 | 40.13 | 0.0 | 0.106 | 1.633 | 2.981 |
| Y12 | 295:295 | 45.37 | 0.0 | 0.01 | 5.473 | 9.962 |
| L3 | 294:297 | 45.59 | 0.0 | 0.047 | 8.462 | 7.522 |
| L0 | 294:298 | 39.44 | 0.0 | 0.07 | 1.35 | 2.35 |
| S9 | 293:296 | 41.93 | 0.0 | 0.011 | 0.929 | 7.296 |
| D50 | 296:294 | 45.77 | 0.0 | 0.009 | 2.671 | 10.909 |
| S12 | 298:296 | 43.11 | 0.0 | 0.007 | 3.547 | 7.194 |
| Y50 | 295:293 | 45.25 | 0.0 | 0.017 | 6.839 | 8.67 |
| S50 | 297:294 | 46.85 | 0.0 | 0.016 | 4.223 | 12.418 |
| D3 | 297:298 | 42.18 | 0.0 | 0.072 | 3.853 | 5.806 |
| D75 | 300:293 | 46.83 | 0.0 | 0.043 | 8.13 | 9.953 |
| S6 | 293:297 | 38.84 | 0.0 | 0.074 | 0.349 | 1.947 |
| S0 | 297:299 | 39.44 | 0.0 | 0.066 | 0.183 | 3.252 |
| Y15 | 299:295 | 40.62 | 0.0 | 0.021 | 1.064 | 4.798 |
| S75 | 293:293 | 41.47 | 0.0 | 0.023 | 2.759 | 4.739 |
| Y0 | 295:298 | 43.21 | 0.0 | 0.063 | 5.241 | 6.89 |
| D1 | 293:298 | 39.81 | 0.0 | 0.078 | 1.763 | 2.48 |
| S20 | 297:295 | 86.51 | 0.0 | 0.016 | 16.502 | 38.358 |
| D20 | 296:295 | 60.64 | 0.0 | 0.02 | 12.683 | 23.521 |

表2 数据过滤统计表

第一列样本名称，第二列reads长度，第三列未过滤的rawadata，第四列N的比例，第五列poly碱基的比例，第六列低质量的比例，第七列高质量数据的cleandata，第八列cleandata/rawdata，第九列rawdata的条数，第十列cleandata的条数，第十一列匹配到primer的reads条数，第十二列匹配到primer的reads与rawdata的比例。

## 5 Tags连接

如果聚类方法用的是USEARCH，要用这里的tag进行OTU聚类，如果聚类方案用的是DADA2，那么此处的tag仅作为数据质控，tag将在DADA2包中处理。

序列拼接使用软件FLASH（Fast Length Adjustment of Short reads，v1.2.11），利用重叠关系将双末端测序得到的成对reads组装成一条序列，得到高变区的Tags。拼接条件如下：

1)最小匹配长度15 bp；

2)重叠区域允许错配率为0.1。

| SampleName | TotalPairsReadNumber | ConnectTagNumber | ConnectRatio(%) | AverageLengthAndSD |
|---|---|---|---|---|
| D0 | 63010 | 62989 | 99.97 | 255/19 |
| D1 | 63336 | 63220 | 99.82 | 241/34 |
| D15 | 63010 | 63009 | 100.00 | 283/18 |
| D20 | 63431 | 63430 | 100.00 | 271/28 |
| D3 | 62923 | 62922 | 100.00 | 247/35 |
| D50 | 63537 | 63537 | 100.00 | 236/27 |
| D6 | 62799 | 62799 | 100.00 | 255/31 |
| D75 | 63196 | 63194 | 100.00 | 240/32 |
| L0 | 63328 | 63303 | 99.96 | 254/17 |
| L1 | 62797 | 62796 | 100.00 | 256/33 |
| L12 | 63380 | 63005 | 99.41 | 282/24 |
| L15 | 63202 | 63200 | 100.00 | 283/23 |
| L20 | 62857 | 62856 | 100.00 | 253/35 |
| L3 | 62681 | 62298 | 99.39 | 243/45 |
| L30 | 63313 | 63233 | 99.87 | 271/32 |
| L50 | 63641 | 63640 | 100.00 | 257/36 |
| L75 | 63371 | 63323 | 99.92 | 229/25 |
| L9 | 62963 | 62963 | 100.00 | 282/21 |
| S0 | 62877 | 62874 | 100.00 | 251/28 |
| S1 | 63513 | 63266 | 99.61 | 236/29 |
| S12 | 63061 | 63060 | 100.00 | 286/16 |
| S15 | 63539 | 63297 | 99.62 | 258/36 |
| S20 | 63212 | 63211 | 100.00 | 255/35 |
| S3 | 62899 | 62899 | 100.00 | 258/27 |
| S30 | 63694 | 63685 | 99.99 | 232/33 |
| S50 | 63367 | 63359 | 99.99 | 235/29 |
| S6 | 63350 | 63324 | 99.96 | 269/35 |
| S75 | 63536 | 63524 | 99.98 | 230/36 |
| S9 | 63571 | 60262 | 94.79 | 266/31 |
| Y0 | 63228 | 63158 | 99.89 | 259/13 |
| Y1 | 62701 | 62701 | 100.00 | 250/30 |
| Y12 | 63510 | 59834 | 94.21 | 260/28 |
| Y15 | 63124 | 63123 | 100.00 | 281/22 |
| Y20 | 63607 | 63602 | 99.99 | 251/36 |
| Y3 | 63105 | 63103 | 100.00 | 263/27 |
| Y30 | 63186 | 63147 | 99.94 | 233/29 |
| Y50 | 63707 | 63707 | 100.00 | 269/34 |
| Y6 | 62911 | 62911 | 100.00 | 273/30 |
| Y75 | 63315 | 63187 | 99.80 | 248/35 |

表3 tag拼接统计表

第一列样本名称，第二列高质量的clean reads条数，第三列拼接为tag的reads条数，第四列tag拼接率。

## 6 OTU聚类结果统计

OTU聚类有两种方法，Usearch和DADA2。Usearch:按照97%序列相似性聚类生成OTU；DADA2：通过去噪的序列以100%的相似度聚类来生成ASV序列，这里统称为OTU。

Usearch:

利用软件USEARCH（v7 .0.1090）将拼接好的 Tags聚类为OTU。其主要过程如下：

1) 利用UPARSE在97 %相似度下进行聚类，得到OTU的代表序列；

2) 利用UCHIME（v4.2.40）将PCR扩增产生的嵌合体从OTU代表序列中去除；

(16S和ITS采取和已有的嵌合体数据库进行比对的方法去除嵌合体。18S采取De novo的方法去除嵌合体

16S嵌合体数据库：gold database（v20110519）

ITS嵌合体数据库：UNITE（v201407 03），分为ITS全长，ITS1和ITS2，按测序区域进行选择）

3) 使用usearch_global方法将所有Tags比对回OTU代表序列，得到每个样品的OTU的丰度统计表。

DADA2：利用软件QIIME2中的DADA2（Divisive Amplicon Denoising Algorithm）方法去噪，获得Amplicon Sequence Variants (ASVs)，ASV为100%相似的序列。进而得到特征表（Feature，对ASV/ASV等的统称）。其主要过程如下：1) 利用qiime tools import导入过滤后的双端序列；

2) 利用qiime dada2 denoise-paired命令将导入后的双端序列基于DADA2的方法构建特征表；

3) 利用qiime tools export将特征表转换成可以直接查看的格式；

| sample_name | tag_num | otu_num |
|---|---|---|
| Y30 | 63058 | 61 |
| L0 | 63233 | 97 |
| L12 | 62993 | 63 |
| Y6 | 62878 | 61 |
| Y9 | 62828 | 62 |
| L9 | 62957 | 40 |
| Y12 | 59529 | 57 |
| D50 | 63278 | 48 |
| L1 | 62670 | 90 |
| S30 | 63658 | 93 |
| S50 | 63012 | 64 |
| D75 | 63174 | 67 |
| L30 | 63209 | 59 |
| S12 | 63057 | 52 |
| S75 | 63517 | 66 |
| Y75 | 62998 | 85 |
| D20 | 63419 | 64 |
| L20 | 62728 | 63 |
| L3 | 62231 | 108 |
| D15 | 62962 | 61 |
| D3 | 62869 | 87 |
| S1 | 63145 | 115 |
| Y0 | 63081 | 88 |
| D1 | 63076 | 92 |
| D0 | 62856 | 101 |
| L50 | 63557 | 68 |
| S9 | 56464 | 57 |
| S15 | 63124 | 69 |
| Y20 | 63334 | 103 |
| S3 | 62785 | 98 |
| Y3 | 63015 | 82 |
| Y15 | 63044 | 70 |
| Y50 | 63278 | 62 |
| L75 | 63109 | 73 |
| D6 | 62443 | 62 |
| L15 | 63158 | 73 |
| S0 | 62526 | 89 |
| Y1 | 62552 | 95 |
| | | |

| S20 | 63195 | 77 |
|---|---|---|

表4 OTU结果统计表

第一列样本名称，第二列tag条数，第三列OTU个数。

# 6 OTU注释

得到OTU代表序列后，通过RDP classifer（v2.2）软件将OTU代表序列与数据库有比对进行物种注释，置信度阈值设置为0.6。

比对数据库：

16S（包括细菌与古菌）：Greengene（默认）： V201305[8]；RDP： Release9 201203[6]

18S 真菌：Silva（默认）： V119[7]

ITS 真菌：UNITE（默认）：Version6 20140910[9]

对注释结果进行如下过滤：

1.去除没有注释结果的OTU；

2.去除注释结果不属于分析项目中的物种。例如，样品为细菌16S，如果OTU注释上古菌则去除。

剩余的OTU方可用于后期分析。

# 7 物种组成分析

物种柱状图可以直观的展示各样本物种组成及比例，反映样本间物种的变化情况。通过与数据库进行比对，对OTU进行物种分类，并分别在门、纲、目、科、属、种水平对各样品作物种丰度柱状图（门水平画所有物种的柱状图，从纲水平开始，将物种丰度在所有样品均低于0.5%和没有分类的物种全部合并成Others）。为了得到每个OTU 对应的物种分类信息，采用RDP classifier 贝叶斯算法对OTU 代表序列进行分类学分析，并在界门纲目科属种水平统计各样本的群落组成。

本报告只展示所有样品属水平的结果。

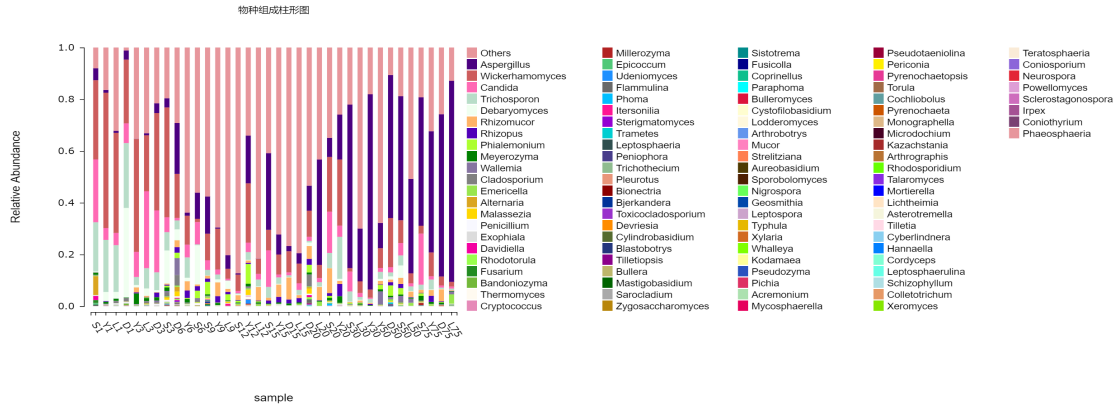


图3 属水平物种丰度柱状图。

物种丰度柱状图。横坐标是样本名称，纵坐标是注释到的物种相对丰度。该分类水平未注释到的合并为Unclassified，丰度在所有样品均低于0.5%和没有分类的物种全部合并成Others。