项目报告：精分患者与健康人肠道微生物组rRNA的初步分析

- 姓名：王俊豪
- 学号：522150910010

1. 摘要

本项目以14个精神分裂症（SCZ）和健康对照（HC）的肠道微生物组16S RNA单向测序数据作为分析对象，进行了一系列的生物信息学分析流程。其中上游分析包括包括数据下载、质量控制、剪切低质量数据、数据去噪等；下游分析包括特征表和分类表生成、物种分类注释、alpha多样性分析、差异表达分析等。项目实现了两组之间较为简单的比对结果，显示了两组肠道菌群的多样性和丰富度，揭示了两组之间存在显著差异表达的feature。项目仍存在一些不足之处，如样本数量局限、测序质量不高等，未来可以进一步通过beta分析、feature序列研究、多组学分析得出更多差异性特征。希望通过这个项目，能够初步揭示 SCZ 与 HC 之间的微生物组差异，为精分的生物标志物及疾病机制研究做下基础。

2. 前言

- **项目背景**

  精神分裂症是“最具致残性”的一种精神障碍，出于其疾病机制的复杂性和异质性，目前，精神分裂症在临床上仍然缺乏一套客观的生物标志物指标，没有有效且可靠的治疗手段。因此，探寻精分患者相对于正常人存在显著差异的生物标记尤为重要，对于精神分裂症的早期识别、干预，改善精神分裂症的预后都有深刻的意义。

  Kyung Hee University于近期提交在NCBI的一项研究ID 1119695 - BioProject - NCBI (nih.gov)，通过`Illumina ISeq 100`对于精分患者和健康对照的肠道微生物组进行了16S rRNA 单向测序，16S rRNA基因是编码细菌和古细菌核糖体RNA的一部分，其序列高度保守但包含足够的变异区段，这使得16S rRNA基因成为微生物群落分析和系统发育研究的常用分子标记。原项目共有14组SRR（7HC + 7SCZ），旨在揭示健康对照组和精神分裂症患者之间肠道微生物组差异。本项目对于这份公共数据集进行了初步的生信分析，显示了两组之间可能存在的差异性。

- **项目思路及工作概要**

  本项目以14组SRR数据集为研究对象，进行的分析步骤主要有：

  - 数据集下载和解压
  - 质量分析
  - 质量控制和修剪
  - 数据去噪
  - 生成特征表、差异表
  - 物种分类注释
  - Alpha多样性分析
  - 差异表达分析

  并由此得到系列分析文件，分析结果主要以`.qzv` `.png`格式呈现。

3. 数据集与方法

3.1 项目涉及的分析工具及数据集

1、**SRA Toolkit**: 用于下载SRA数据。

2、**FastQC**: 用于质量控制，**MulitiQC**生成综合质控报告。

3、**Trimmomatic**: 用于数据剪切。

4、**QIIME2**: 用于大部分数据分析和可视化。

5、**DADA2**（内置于QIIME2中）: 用于去噪和特征表生成。

6、**SILVA数据库**: 用于物种注释。

7、**R**和**DESeq2**: 用于基因表达差异分析。

8、项目分析的数据集及相关信息：Run Selector :: NCBI (nih.gov)

	Run	BioSample	AvgSpotLen	Bases	Bytes	Collection_Date	Experiment	isolate	Library Name	ReleaseDate	Sample Name
1	SRR29282889	SAMN41662312	300	12.57 M	4.32 Mb	2019-09-18	SRX24800062	SCZ	SCZ_3	2024-06-04	Schizophrenia Patient3
2	SRR29282890	SAMN41662311	300	10.30 M	3.72 Mb	2019-08-27	SRX24800061	SCZ	SCZ_2	2024-06-04	Schizophrenia Patient2
3	SRR29282891	SAMN41662310	300	9.60 M	3.36 Mb	2019-07-30	SRX24800060	SCZ	SCZ_1	2024-06-04	Schizophrenia Patient1
4	SRR29282892	SAMN41662309	292	10.21 M	5.91 Mb	2020-01-13	SRX24800059	Control	HC_7	2024-06-04	Healthy Control7
5	SRR29282893	SAMN41662308	293	8.23 M	4.77 Mb	2020-01-02	SRX24800058	Control	HC_6	2024-06-04	Healthy Control6
6	SRR29282894	SAMN41662307	292	10.35 M	5.90 Mb	2019-11-13	SRX24800057	Control	HC_5	2024-06-04	Healthy Control5
7	SRR29282895	SAMN41662306	293	7.89 M	4.41 Mb	2019-10-22	SRX24800056	Control	HC_4	2024-06-04	Healthy Control4
8	SRR29282896	SAMN41662305	292	7.66 M	4.35 Mb	2019-09-18	SRX24800055	Control	HC_3	2024-06-05	Healthy Control3
9	SRR29282897	SAMN41662316	300	12.17 M	4.17 Mb	2020-01-13	SRX24800054	SCZ	SCZ_7	2024-06-04	Schizophrenia Patient7
10	SRR29282898	SAMN41662315	300	13.42 M	4.53 Mb	2020-01-02	SRX24800053	SCZ	SCZ_6	2024-06-04	Schizophrenia Patient6
11	SRR29282899	SAMN41662314	300	11.08 M	3.75 Mb	2019-11-13	SRX24800052	SCZ	SCZ_5	2024-06-04	Schizophrenia Patient5
12	SRR29282900	SAMN41662313	300	10.23 M	3.49 Mb	2019-10-22	SRX24800051	SCZ	SCZ_4	2024-06-04	Schizophrenia Patient4
13	SRR29282901	SAMN41662304	300	17.61 M	5.84 Mb	2019-08-27	SRX24800050	Control	HC_2	2024-06-04	Healthy Control2
14	SRR29282902	SAMN41662303	300	16.97 M	5.69 Mb	2019-07-30	SRX24800049	Control	HC_1	2024-06-04	Healthy Control1

3.2 项目工作流程

3.2.1 计算环境的建立

- 安装 Miniconda 用于环境管理，便于创建QIIME2环境

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
bash Miniconda3-latest-Linux-x86_64.sh -b -p ~/project/miniconda
export PATH="~/project/miniconda/bin:$PATH"
conda init
```

- 创建并激活QIIME2环境

```
wget https://data.qiime2.org/distro/amplicon/qiime2-amplicon-2024.5-py38-linux-conda.yml
conda env create -n project2503 --file qiime2-amplicon-2024.5-py38-linux-conda.yml
conda activate project2503
```

- 安装R

```
sudo apt update -qq
sudo apt install --no-install-recommends software-properties-common dirmngr
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc |
sudo tee -a /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu
$(lsb_release -cs)-cran40/"
sudo apt install --no-install-recommends r-base
```

- 安装DESeq2

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("DESeq2")
```

- 安装必要的分析工具

```
conda install sra-tools
conda install -c bioconda fastqc
pip install multiqc
conda install -c bioconda trimmomatic
```

3.2.2 下载数据

`download_data.sh`

```
bash download_data.sh ~/project SRR29282889 SRR29282890 SRR29282891 SRR29282892
SRR29282893 SRR29282894 SRR29282895 SRR29282896 SRR29282897 SRR29282898
SRR29282899 SRR29282900 SRR29282901 SRR29282902
```

此步骤在运行目录下创建`data`目录，存放下载的SRR数据集以及对应解压后的`.fasaq`文件。

3.2.3 质量控制

`quality_control.sh`

```
bash quality_control.sh ~/project
open ~project/fastqc_results/multiqc_report.html
```

此步骤在运行目录下创建`fastqc_results`目录，使用`fastqc`命令得到质控文件并解压缩，并且使用`multiqc`生成质控总报告的HTML文件。

3.2.4 剪切低质量数据

`trim_data.sh`

```
bash trim_data.sh ~/project
```

此步骤通过`Trimmomatic`工具，根据质控文件的信息，对`data`下的所有`.fasaq`文件进行质量控制和修剪，生成修剪后的文件并存放在`trimmed_reads`目录中，完成数据的初步预处理。

3.2.5 去噪、生成特征表

- 先创建 manifest 文件`manifest.tsv`

```
sample-id       absolute-filepath
SRR29282889_trimmed
/home/bio/project/trimmed_reads/SRR29282889_trimmed.fastq
```

```
SRR29282890_trimmed
/home/bio/project/trimmed_reads/SRR29282890_trimmed.fastq
SRR29282891_trimmed
/home/bio/project/trimmed_reads/SRR29282891_trimmed.fastq
SRR29282892_trimmed
/home/bio/project/trimmed_reads/SRR29282892_trimmed.fastq
SRR29282893_trimmed
/home/bio/project/trimmed_reads/SRR29282893_trimmed.fastq
SRR29282894_trimmed
/home/bio/project/trimmed_reads/SRR29282894_trimmed.fastq
SRR29282895_trimmed
/home/bio/project/trimmed_reads/SRR29282895_trimmed.fastq
SRR29282896_trimmed
/home/bio/project/trimmed_reads/SRR29282896_trimmed.fastq
SRR29282897_trimmed
/home/bio/project/trimmed_reads/SRR29282897_trimmed.fastq
SRR29282898_trimmed
/home/bio/project/trimmed_reads/SRR29282898_trimmed.fastq
SRR29282899_trimmed
/home/bio/project/trimmed_reads/SRR29282899_trimmed.fastq
SRR29282900_trimmed
/home/bio/project/trimmed_reads/SRR29282900_trimmed.fastq
SRR29282901_trimmed
/home/bio/project/trimmed_reads/SRR29282901_trimmed.fastq
SRR29282902_trimmed
/home/bio/project/trimmed_reads/SRR29282902_trimmed.fastq
```

`dada2_processing.sh`

```
bash dada2_processing.sh ~/project
```

此步骤使用 `QIIME 2` 中的 `DADA2` 插件进行去噪处理。创建 `dada2_output` 目录存放分析结果：

代表性序列文件 `rep-seqs.qza`

特征表 `table.qza`

去噪统计文件 `denoising-stats.qza`

同时生成对应的可视化文件 `.qzv`

3.2.6 分类物种注释

`taxonomy_classification.sh`

```
bash taxonomy_classification.sh ~/project
```

此步骤中，首先下载SILVA文件 `silva-138-99-nb-classifier.qza`，使用 `QIIME 2` 中的 `feature-classifier classify-sklearn` 命令进行物种分类，分析的文件是 `rep-seqs.qza`，产生一个分类结果文件 `taxonomy.qza`。

3.2.7 Alpha 多样性分析

`alpha_diversity.sh`

```
bash alpha_diversity.sh ~/project
```

此步骤使用`qiime diversity alpha`计算健康对照组和精神分裂症患者之间的Shannon多样性指数，并生成Alpha多样性组间显著性可视化结果，评估组间微生物多样性是否存在显著差异。

3.2.8 差异表达分析

- 先准备元数据文件`metadata.tsv`

```
treatment
SCZ
SCZ
SCZ
HC
HC
HC
HC
HC
SCZ
SCZ
SCZ
SCZ
HC
HC
```

在这里，metadata文件提供数据组别信息，需要保持行的排列与特征表列的排列一致。

`prepare_deseq2_input.sh`

```
bash prepare_deseq2_input.sh ~/project
```

这个bash脚本处理之前得到的特征表，用于准备DESeq2分析，使用`qiime`和`biom`工具将特征表`table.qza`转换为TSV文件，存储于目录`deseq2_input`下。

`deseq2.R`

```
Rscript deseq2.R
```

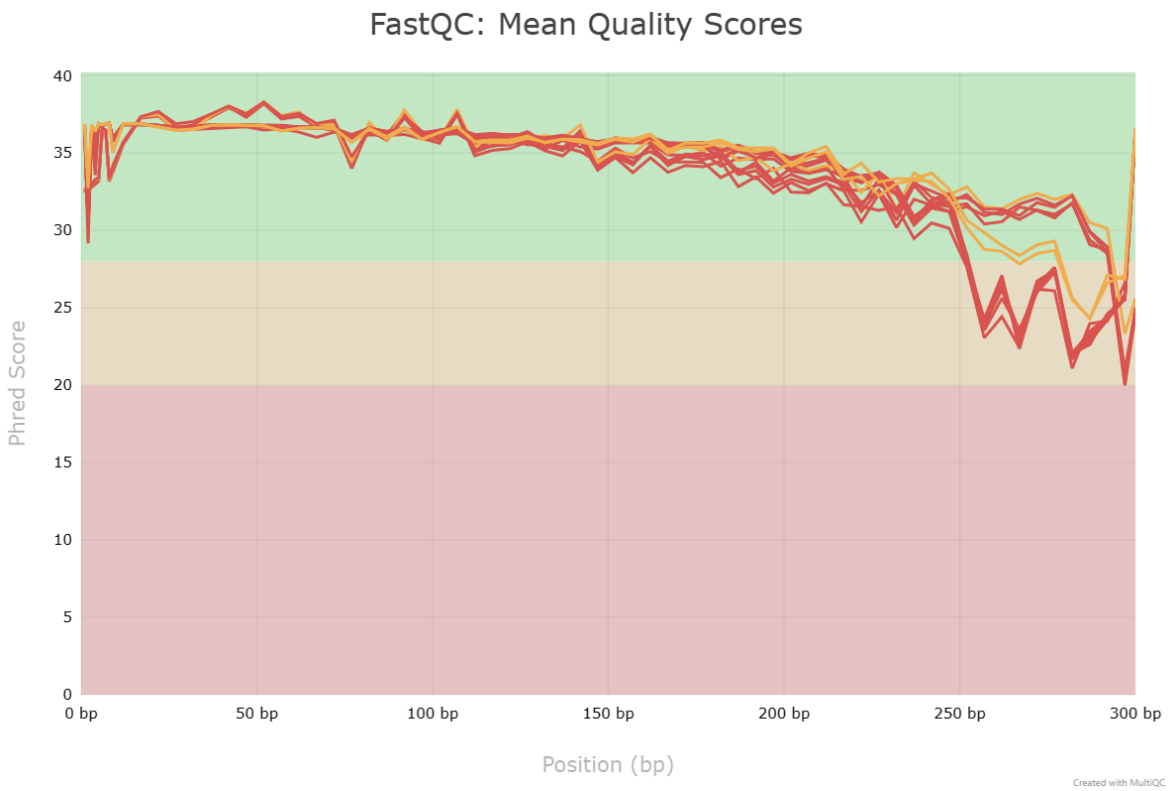
这个R脚本使用`DESeq2`工具，读取TSV格式的特征表为countData，meta文件为colData，执行DESeq2分析，并将输出结果用火山图呈现。

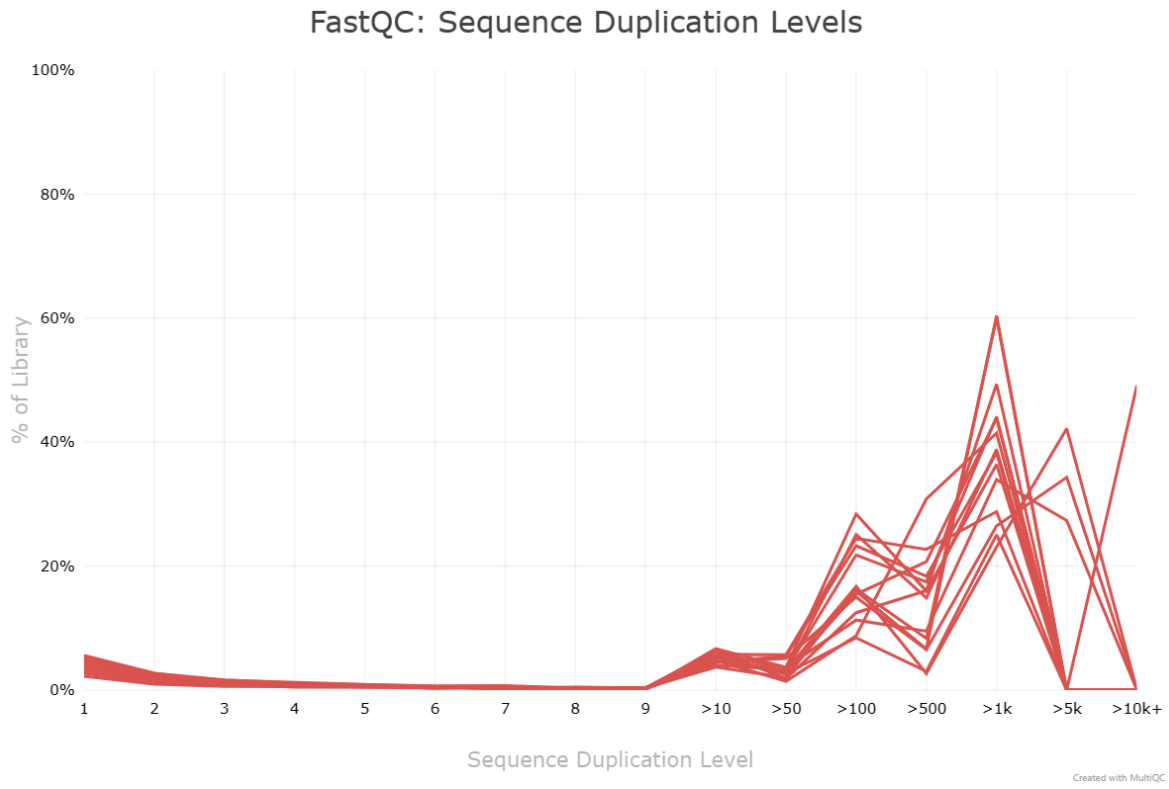
4. 分析结果

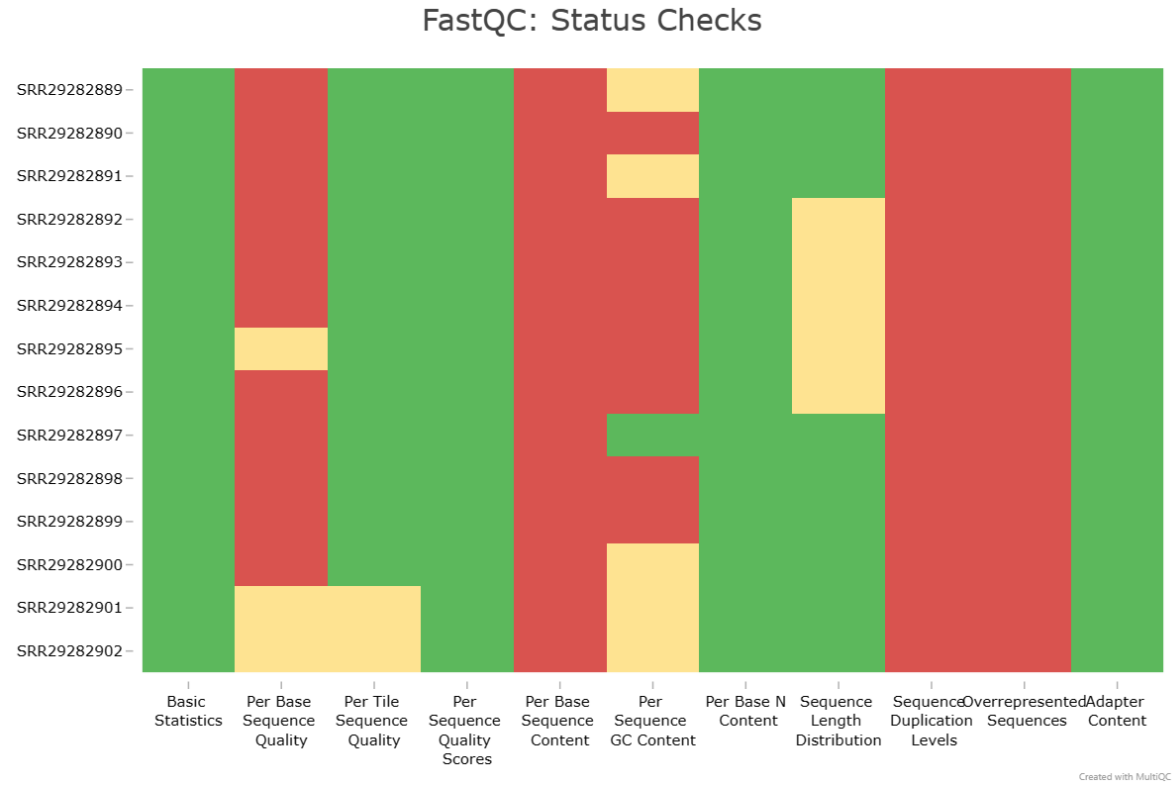
4.1 质控报告

```
open ~project/fastqc_results/multiqc_report.html
```

总的质控报告可在MultiQC 报告 (wsl.localhost)查看，包括GC含量、质量评分、高代表性序列、重复水平等等。







从分析结果可看出，异常部分几乎都在序列方面，数据的碱基序列质量似乎并没有很完美，特别是在较长序列中。这可能是由于数据集的测序结果尚未得到充分检验，意味着去噪需要处理的工作量较大。

4.2 去噪数据

`denoising-stats.qzv`可通过view.qiime2.org查看

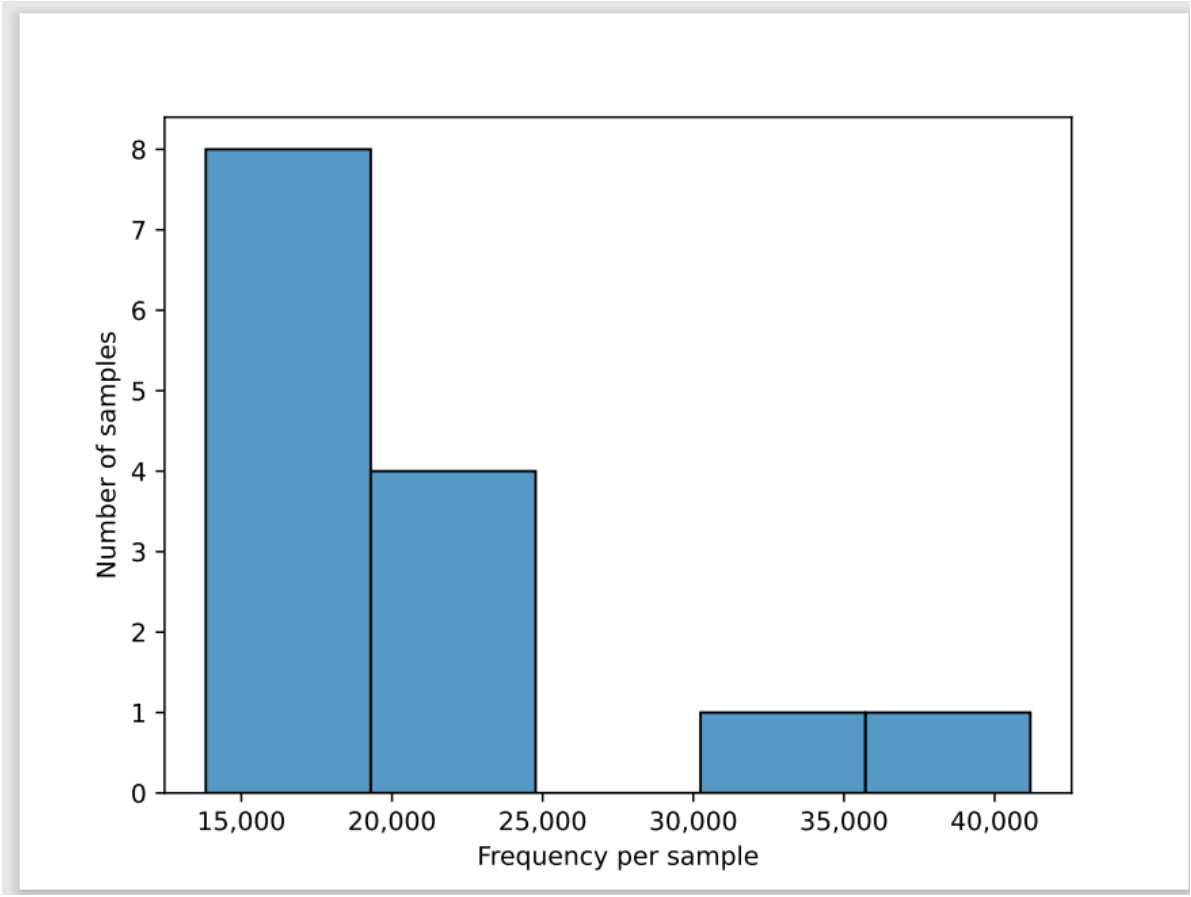
```
qiime tools view denoising-stats.qzv
```

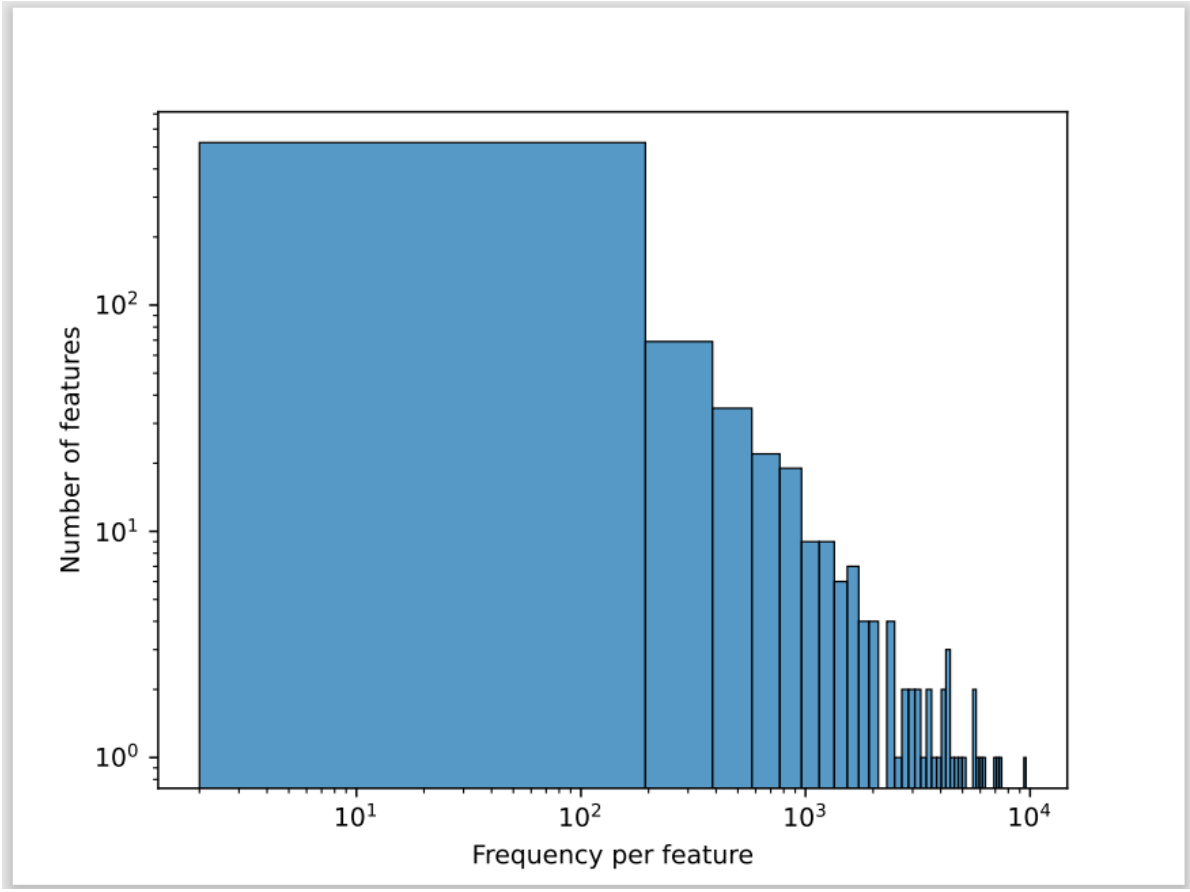
sample-id #q2:types	input numeric	filtered numeric	percentage of input passed filter numeric	denoised numeric	non-chimeric numeric	percentage of input non-chimeric numeric
SRR29282889_trimmed	41581	25584	61.53	24931	23473	56.45
SRR29282890_trimmed	34013	16221	47.69	15989	15850	46.6
SRR29282891_trimmed	31778	18281	57.53	17865	17834	56.12
SRR29282892_trimmed	34065	18980	55.72	18441	17012	49.94
SRR29282893_trimmed	27462	15741	57.32	15343	13821	50.33
SRR29282894_trimmed	34609	21009	60.7	20414	18369	53.08
SRR29282895_trimmed	26314	17106	65.01	16855	16240	61.72
SRR29282896_trimmed	25729	16169	62.84	15814	15058	58.53
SRR29282897_trimmed	40255	24586	61.08	24119	21977	54.59
SRR29282898_trimmed	44426	27063	60.92	26817	22886	51.51
SRR29282899_trimmed	36762	22278	60.6	21849	19819	53.91
SRR29282900_trimmed	33874	20234	59.73	19986	16781	49.54
SRR29282901_trimmed	58131	43067	74.09	42517	41186	70.85
SRR29282902_trimmed	56013	39282	70.13	38651	33209	59.29

4.3 特征表

`table.qzv`

14个样本的Feature总数为741，频率分布如下图。





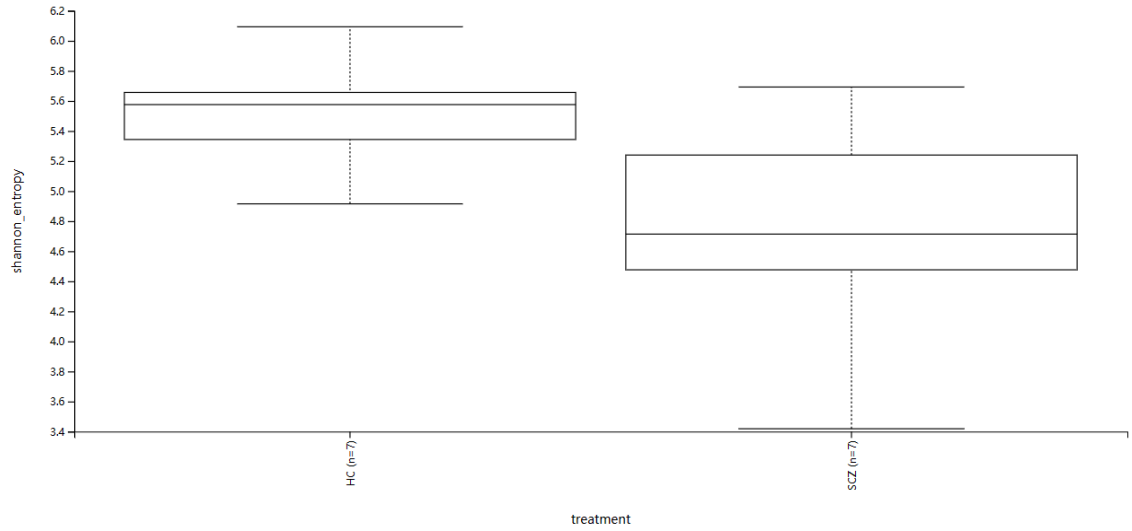
4.4 物种分类结果

`taxonomy.qzv` (部分)

Feature ID #q2:types	Taxon categorical	Confidence categorical
6bdba2e00ab254293e94924589f1af05	d__Archaea;p__Euryarchaeota;c__Methanobacteria;o__Methanobacteriales;f__Methanobacteriaceae;g__Methanobrevibacter;s__	0.9999843713110732
955b6220d747d0dca797409b5e9c8353	d__Archaea;p__Thermoplasmatota;c__Thermoplasmata;o__Methanomassiliicoccales;f__Methanomethylophilaceae;g__Candidatus_Methanogranum;s__	0.8249414669747043
d43fe36e7ea953a921cb3b9a3f59a5c0	d__Bacteria	0.9968081507905789
ad5142d8bf99ad6756131d3db134bb51	d__Bacteria	0.9999675854807285
6ca11c046114a833087d812757c1b2df	d__Bacteria	0.999765848858868
0d6eedd8738c80c0763887707b2a2a35	d__Bacteria	0.9998582690693227
a69a044d6d462d56ca5387c43065c9bf	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__	0.9999871890781193
82a7b8ad1d04ec4ef39267bcfff149fb	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__	0.9999764783481268
7a64ba09ca843b1c82583d9b55e664c1	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces;s__	0.9999712614270425
9e758ed6ac4a6312ae5756d72b279412	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Bifidobacterium;s__	0.9999042939316943
624096e2f53535c37d7e55d60fbfd3c8	d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Bifidobacterium;s__	0.9998543553600582

4.5 Alpha多样性分析

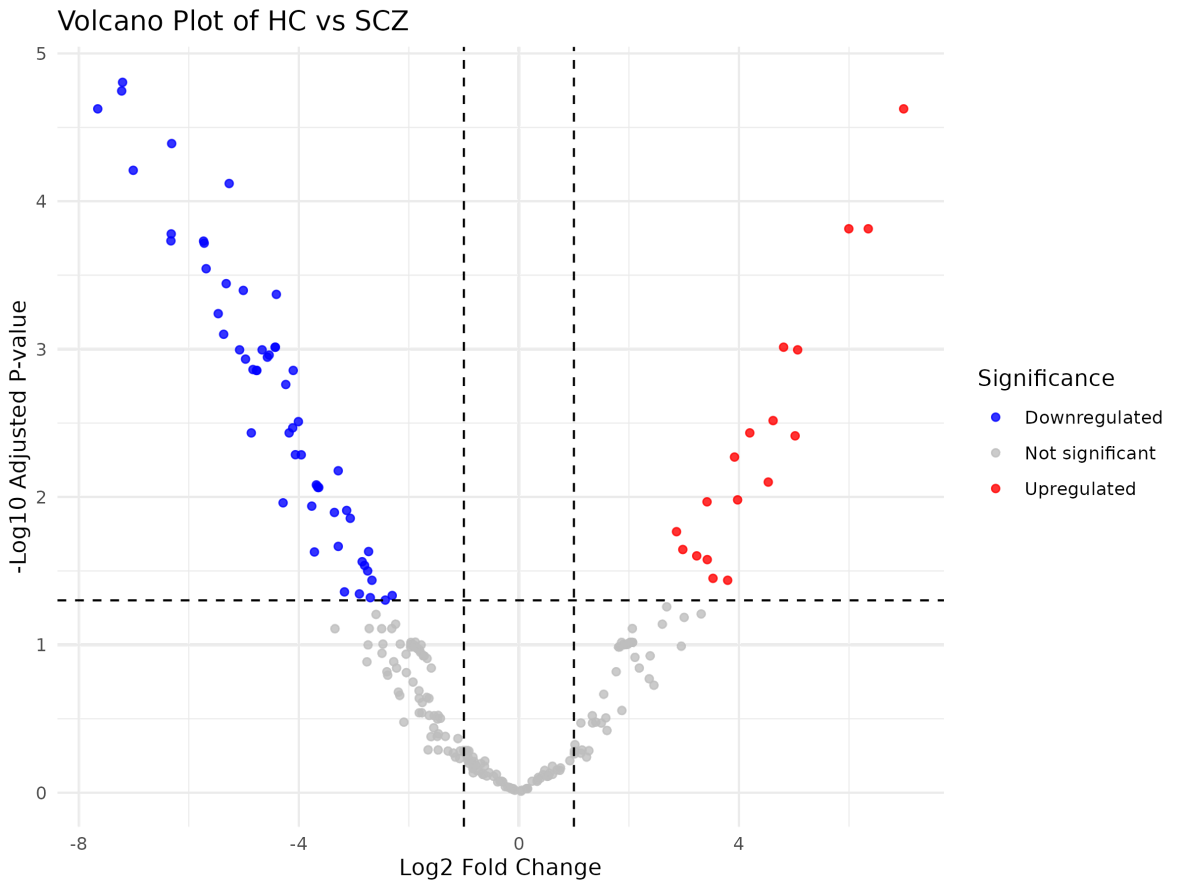
`shannon_group_significance.qzv`



组间p值为0.063919，接近0.05，alpha多样性分析表明HC组肠道微生物具有较高的多样性。

4.6 差异表达火山图

`volcano_plot.png`



火山图中，`log2 fold change`为正数表示HC组相比于SCZ组出现高表达，为负表示低表达。一些出现极端差异表达（`|log2 fold change|` > 6）的Feature为：

#OTU ID	baseMean	log2FoldChange	pvalue
1f2adb648e051fe82d76971b34b370dc	97.49454914	-7.219748605	1.42E-07
7940bf3ab1f7acac346efa1b84ab5d38	80.24945395	6.992931326	3.10E-07
c84942954987db84e43cbf059c8f00bb	52.5355815	-6.319248482	6.60E-06
fcd261fb1145b0daa4b3befdac4d1b8b	96.41245225	-7.203514212	6.23E-08
357db39768977771a015a7d5c7b3f1e4	52.24304066	-6.311110719	8.08E-07
802bfb41f7bc0aa2f2ab41bb1d0c64a1	52.72446369	-6.324679931	8.08E-06
3ba1dc17d6ce32d01c39d4b7f6f0d403	51.61701414	6.349868571	5.48E-06
c1b52f84220653d6061e4004721e5177	84.37965449	-7.009741877	1.47E-06
6154cfd80a89be37c28619ceee523b72	131.5717479	-7.654576732	3.76E-07

5. 总结与讨论

5.1 项目总结

本项目基于16S rRNA高通量测序数据，对健康对照组和精分患者的肠道菌群进行了比较分析。通过数据集预处理、物种分类注释、Alpha多样性分析、差异表达分析，显示了两组肠道菌群的多样性和丰富度，揭示了两组之间存在显著差异表达的feature。肠道微生物是人体代谢的重要参与者，其物种组成与活动能够提供人体的重要信息，这些分析与发现从肠道微生物组的角度揭示精分患者的差异，或许能

为进一步研究疾病发病机制提供了重要的生物标志物和潜在的治疗靶点。

关键点

- **数据预处理**：确保数据的质量和完整性至关重要。使用了`fastqc`检测数据的质量，通过`Qiime2`进行数据过滤，确保后续分析的可靠性。
- **Alpha多样性分析**：这一步评估了样本中物种的多样性和丰富度，HC组肠道微生物具有较高的多样性
- **差异表达分析**：通过比较两组样本中基因的表达水平，寻找显著差异表达的feature。这些差异或许与疾病的发病机制有关，并通过火山图直观展示差异表达。

5.2 讨论

存在的不足

- **样本数量有限**：本项目分析了7组精分患者和7组健康对照组的16S RNA数据，样本数量对于疾病的差异研究来说仍然较少，可能导致结论的代表性不强。
- **测序质量不高**：`fastqc`检测出测序的质量不算很好，可能影响后续分析的可靠性，导致去噪处理量较大，在去噪中也可能丢失一些关键信息。
- **质控检测**：未进行读数的深入验证，确定哪些区域被充分覆盖。在质控和去噪中可能存在遗漏或重复的区域。

后续分析

- **Beta多样性分析**：对数据集进行Beta多样性分析，检测具体是何种物种在两组之间出现显著差异。
- **feature分析**：对于具有显著差异的Feature进行更深入的分析，研究这些序列片段的特征。
- **多组学**：纳入精分患者和健康人的自身RNA-seq数据，进行肠道微生物组与临床症状或其他生化分子之间的相关性分析，多组学探究精神分裂症的生物标志物。

6. 参考文献

- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... & Knight, R. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology, 37(8), 852-857.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15(12), 550.
- Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-D596.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature methods, 13(7), 581-583.

7. 附录

7.1 项目工作环境详细

```
# 操作系统及环境
Ubuntu 22.04.3
conda 24.3.0
q2cli version 2024.5.0
# 编程环境
Python 3.9.19
R version 4.3.3 (2024-02-29)
# 计算工具
"fastq-dump" version 2.11.3
FastQC v0.11.9
multiqc, version 1.22.2
```

7.2 项目内的脚本详细

`download_data.sh`

```
# 下载数据集
#!/bin/bash

Project=$1

mkdir -p $Project/data
cd $Project/data

for SRR in "$@"; do
    if [ "$SRR" != "$1" ]; then
        prefetch $SRR
        fasterq-dump $SRR
    fi
done
```

`quality_control.sh`

```
# 质量控制
#!/bin/bash

Project=$1
mkdir -p $Project/fastqc_results

for FILE in $Project/data/*.fastq; do
    fastqc $FILE -o $Project/fastqc_results
done

for FILE in $Project/fastqc_results/*.zip; do
        unzip $FILE -d $Project/fastqc_results
done

multiqc $Project/fastqc_results -o $Project/fastqc_results
```

trim_data.sh

```
# 剪切低质量数据
#!/bin/bash

Project=$1
mkdir -p $Project/trimmed_reads

for FILE in $Project/data/*.fastq; do
     /usr/bin/TrimmomaticSE -phred33 \
        $FILE \
        $Project/trimmed_reads/$(basename $FILE .fastq)_trimmed.fastq \
        ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 \
        LEADING:5 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36 \
        HEADCROP:15
done
# 去除序列开头质量值小于5,末尾质量值小于20的碱基;在4个碱基的滑动窗口中，如果平均质量值小于20，则截断序列;保留长度至少为36的序列;去除序列开头的前15个碱基。
```

dada2_processing.sh

```
#!/bin/bash

Project=$1

mkdir -p $Project/dada2_output

# 导入序列数据
qiime tools import \
  --type 'SampleData[SequencesWithQuality]' \
  --input-path $Project/manifest.tsv \
  --output-path $Project/dada2_output/demux.qza \
  --input-format SingleEndFastqManifestPhred33V2

if [ $? -ne 0 ]; then
  echo "Error: Failed to import sequences."
  exit 1
fi

# DADA2去噪
qiime dada2 denoise-single \
  --i-demultiplexed-seqs $Project/dada2_output/demux.qza \
  --p-trim-left 0 \
  --p-trunc-len 240 \
  --o-representative-sequences $Project/dada2_output/rep-seqs.qza \
  --o-table $Project/dada2_output/table.qza \
  --o-denoising-stats $Project/dada2_output/denoising-stats.qza

if [ $? -ne 0 ]; then
  echo "Error: DADA2 denoise-single failed."
  exit 1
fi
```

```
# 特征表汇总
qiime feature-table summarize \
  --i-table $Project/dada2_output/table.qza \
  --o-visualization $Project/dada2_output/table.qzv

if [ $? -ne 0 ]; then
  echo "Error: Feature table summarization failed."
  exit 1
fi

# 序列表格
qiime feature-table tabulate-seqs \
  --i-data $Project/dada2_output/rep-seqs.qza \
  --o-visualization $Project/dada2_output/rep-seqs.qzv

if [ $? -ne 0 ]; then
  echo "Error: Feature table tabulation failed."
  exit 1
fi

# 去噪统计数据表格
qiime metadata tabulate \
  --m-input-file $Project/dada2_output/denoising-stats.qza \
  --o-visualization $Project/dada2_output/denoising-stats.qzv

if [ $? -ne 0 ]; then
  echo "Error: Denoising stats tabulation failed."
  exit 1
fi

echo "dada2 processing completed successfully!"
```

`taxonomy_classification.sh`

```
#!/bin/bash

Project=$1

mkdir -p $Project/taxonomy


# 下载SILVA数据库（如未下载）
if [ ! -f "silva-138-99-nb-classifier.qza" ]; then
  wget -O silva-138-99-nb-classifier.qza \
  https://data.qiime2.org/2024.5/common/silva-138-99-nb-classifier.qza
fi

# 分类物种注释
qiime feature-classifier classify-sklearn \
  --i-classifier silva-138-99-nb-classifier.qza \
  --i-reads $Project/dada2_output/rep-seqs.qza \
  --o-classification $Project/taxonomy/taxonomy.qza

if [ $? -ne 0 ]; then
```

```
  echo "Error: Taxonomy classification failed."
  exit 1
fi

# 分类物种可视化
qiime metadata tabulate \
  --m-input-file $Project/taxonomy/taxonomy.qza \
  --o-visualization $Project/taxonomy/taxonomy.qzv

if [ $? -ne 0 ]; then
  echo "Error: Taxonomy visualization failed."
  exit 1
fi

echo "Taxonomy classification completed successfully!"
```

`alpha_diversity.sh`

```
#!/bin/bash

Project=$1

mkdir -p $Project/alpha_diversity

# 计算Alpha多样性
qiime diversity alpha \
  --i-table $Project/dada2_output/table.qza \
  --p-metric shannon \
  --o-alpha-diversity $Project/alpha_diversity/shannon.qza

# 可视化Alpha多样性
qiime diversity alpha-group-significance \
  --i-alpha-diversity $Project/alpha_diversity/shannon.qza \
  --m-metadata-file $Project/metadata.tsv \
  --o-visualization $Project/alpha_diversity/shannon_group_significance.qzv

echo "Alpha diversity analysis completed!"
```

`prepare_deseq2_input.sh`

```
#!/bin/bash

Project=$1

mkdir -p $Project/deseq2_input

# 导出特征表到biom格式
qiime tools export \
  --input-path $Project/dada2_output/table.qza \
  --output-path $Project/deseq2_input

# 转换BIOM为TSV
biom convert \
```

```
  -i $Project/deseq2_input/feature-table.biom \
  -o $Project/deseq2_input/feature-table.tsv \
  --to-tsv

# 复制元数据文件
cp $Project/metadata.tsv $Project/deseq2_input/metadata.tsv

echo "Feature table and have been prepared!"
```

`deseq2.R`

```
library(DESeq2)
library(ggplot2)
library(dplyr)

project_path <- "~/project"

# 读取特征表
counts <- read.table(file.path(project_path, "deseq2_input", "feature-
table.tsv"),
                     header = TRUE, row.names = 1, sep = "\t")

# 转换特征表数据为整型,DESeq2分析只能读取整型
counts <- as.matrix(counts)
counts <- apply(counts, 2, function(x) as.integer(as.numeric(x)))
counts <- counts + 1
# 将特征表列名转换为字符型
colnames(counts) <- as.character(colnames(counts))

# 读取元数据
meta <- read.table(file.path(project_path, "metadata.tsv"),
                   header = TRUE)



# 运行DESeq2分析
dds <- DESeqDataSetFromMatrix(countData = counts, colData = meta, design = ~
treatment)

dds <- DESeq(dds)
res <- results(dds)

dir.create(file.path(project_path, "deseq2_output"), showWarnings = FALSE)

# 保存分析文件
write.csv(as.data.frame(res), file=file.path(project_path, "deseq2_output",
"deseq2_results.csv"))

project_path <- "~/project"
result_file <- file.path(project_path, "deseq2_output", "deseq2_results.csv")

deseq2_results <- read.csv(result_file)

cat("DESeq2 results head:\n")
```

```
print(head(deseq2_results))

# 计算 -log10 p-value，并添加显著性标签
deseq2_results <- deseq2_results %>%
  mutate(log10_padj = -log10(padj),
         significance = case_when(
           padj < 0.05 & log2FoldChange > 1 ~ "Upregulated",
           padj < 0.05 & log2FoldChange < -1 ~ "Downregulated",
           TRUE ~ "Not significant"
         ))

# 绘制火山图
volcano_plot <- ggplot(deseq2_results, aes(x = log2FoldChange, y = log10_padj,
color = significance)) +
  geom_point(alpha = 0.8, size = 1.5) +
  scale_color_manual(values = c("Upregulated" = "red", "Downregulated" = "blue",
"Not significant" = "gray")) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", color = "black") +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", color = "black") +
  theme_minimal() +
  labs(title = "Volcano Plot of HC vs SCZ",
       x = "Log2 Fold Change",
       y = "-Log10 Adjusted P-value",
       color = "Significance")

# 保存火山图为文件
ggsave(file.path(project_path, "deseq2_output", "volcano_plot.png"), plot =
volcano_plot, width = 8, height = 6)

print("DESeq2 analysis completed! ")
```

`workflow.sh`

```
#!/bin/bash

# 设置工作目录
WORK_DIR=~/project
cd $WORK_DIR

# 建立环境
echo "Setting up the environment..."

# 安装Miniconda
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
bash Miniconda3-latest-Linux-x86_64.sh -b -p $WORK_DIR/miniconda
export PATH="$WORK_DIR/miniconda/bin:$PATH"
conda init
source ~/.bashrc

# 安装qiime2并激活环境
wget https://data.qiime2.org/distro/amplicon/qiime2-amplicon-2024.5-py38-linux-
conda.yml
conda env create -n project2503 --file qiime2-amplicon-2024.5-py38-linux-
conda.yml
```

```
conda activate project2503

# 安装R
sudo apt update -qq
sudo apt install --no-install-recommends software-properties-common dirmngr
wget -qO- https://cloud.r-project.org/bin/linux/ubuntu/marutter_pubkey.asc |
sudo tee -a /etc/apt/trusted.gpg.d/cran_ubuntu_key.asc
sudo add-apt-repository "deb https://cloud.r-project.org/bin/linux/ubuntu
$(lsb_release -cs)-cran40/"
sudo apt install --no-install-recommends r-base

# 安装DESeq2
Rscript -e "if (!requireNamespace('BiocManager', quietly = TRUE))
install.packages('BiocManager'); BiocManager::install('DESeq2')"

# 安装必要的分析工具
conda install sra-tools -y
conda install -c bioconda fastqc -y
pip install multiqc
conda install -c bioconda trimmomatic -y

# 下载数据
echo "Downloading data..."
bash download_data.sh $WORK_DIR SRR29282889 SRR29282890 SRR29282891 SRR29282892
SRR29282893 SRR29282894 SRR29282895 SRR29282896 SRR29282897 SRR29282898
SRR29282899 SRR29282900 SRR29282901 SRR29282902

# 质量控制
echo "Performing quality control..."
bash quality_control.sh $WORK_DIR

# 剪切低质量数据
echo "Trimming low-quality data..."
bash trim_data.sh $WORK_DIR

# 去噪、生成特征表
echo "Generating feature table..."
# 创建 manifest 文件
cat <<EOF > $WORK_DIR/manifest.tsv
sample-id       absolute-filepath
SRR29282889_trimmed     $WORK_DIR/trimmed_reads/SRR29282889_trimmed.fastq
SRR29282890_trimmed     $WORK_DIR/trimmed_reads/SRR29282890_trimmed.fastq
SRR29282891_trimmed     $WORK_DIR/trimmed_reads/SRR29282891_trimmed.fastq
SRR29282892_trimmed     $WORK_DIR/trimmed_reads/SRR29282892_trimmed.fastq
SRR29282893_trimmed     $WORK_DIR/trimmed_reads/SRR29282893_trimmed.fastq
SRR29282894_trimmed     $WORK_DIR/trimmed_reads/SRR29282894_trimmed.fastq
SRR29282895_trimmed     $WORK_DIR/trimmed_reads/SRR29282895_trimmed.fastq
SRR29282896_trimmed     $WORK_DIR/trimmed_reads/SRR29282896_trimmed.fastq
SRR29282897_trimmed     $WORK_DIR/trimmed_reads/SRR29282897_trimmed.fastq
SRR29282898_trimmed     $WORK_DIR/trimmed_reads/SRR29282898_trimmed.fastq
SRR29282899_trimmed     $WORK_DIR/trimmed_reads/SRR29282899_trimmed.fastq
SRR29282900_trimmed     $WORK_DIR/trimmed_reads/SRR29282900_trimmed.fastq
SRR29282901_trimmed     $WORK_DIR/trimmed_reads/SRR29282901_trimmed.fastq
SRR29282902_trimmed     $WORK_DIR/trimmed_reads/SRR29282902_trimmed.fastq
EOF

bash dada2_processing.sh $WORK_DIR
```

```
# 分类物种注释
echo "Classifying taxonomy..."
bash taxonomy_classification.sh $WORK_DIR

# Alpha 多样性分析
echo "Performing alpha diversity analysis..."
bash alpha_diversity.sh $WORK_DIR

# 差异表达分析
echo "Preparing DESeq2 input..."
# 创建元数据文件
cat <<EOF > $WORK_DIR/metadata.tsv
treatment
SCZ
SCZ
SCZ
HC
HC
HC
HC
HC
SCZ
SCZ
SCZ
SCZ
HC
HC
EOF

bash prepare_deseq2_input.sh $WORK_DIR
Rscript deseq2.R

echo "All steps completed successfully!"
```