R Project

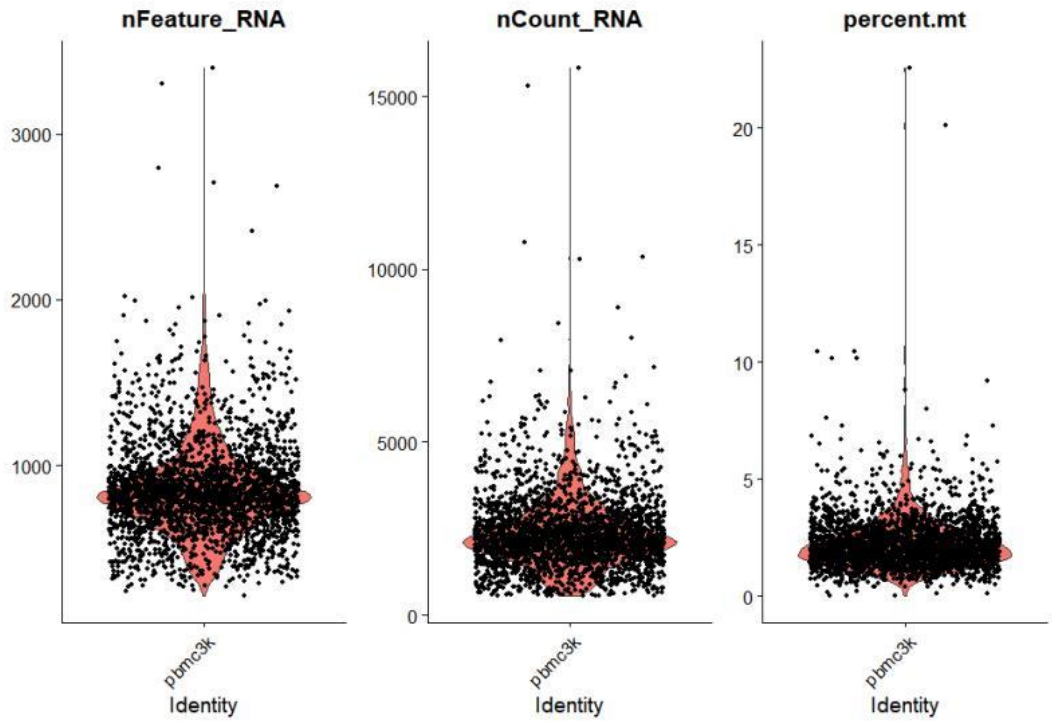
1 数据来源

我的 project 分为两个部分：单细胞聚类分析和宏基因组装箱分析。其中单细胞聚类分析数据集来自于 10xgenomics 平台[1]提供的外周血单核细胞（PBMC）数据集，包含 2700 个单细胞在 Illumina NextSeq 500 平台上测序获得的 gene、barcode 数据。宏基因组装箱分析数据集来自于 NCBI[2]提供的宏基因组 fastq 数据，测序平台是 Illumina HiSeq 4000。

2 单细胞聚类分析

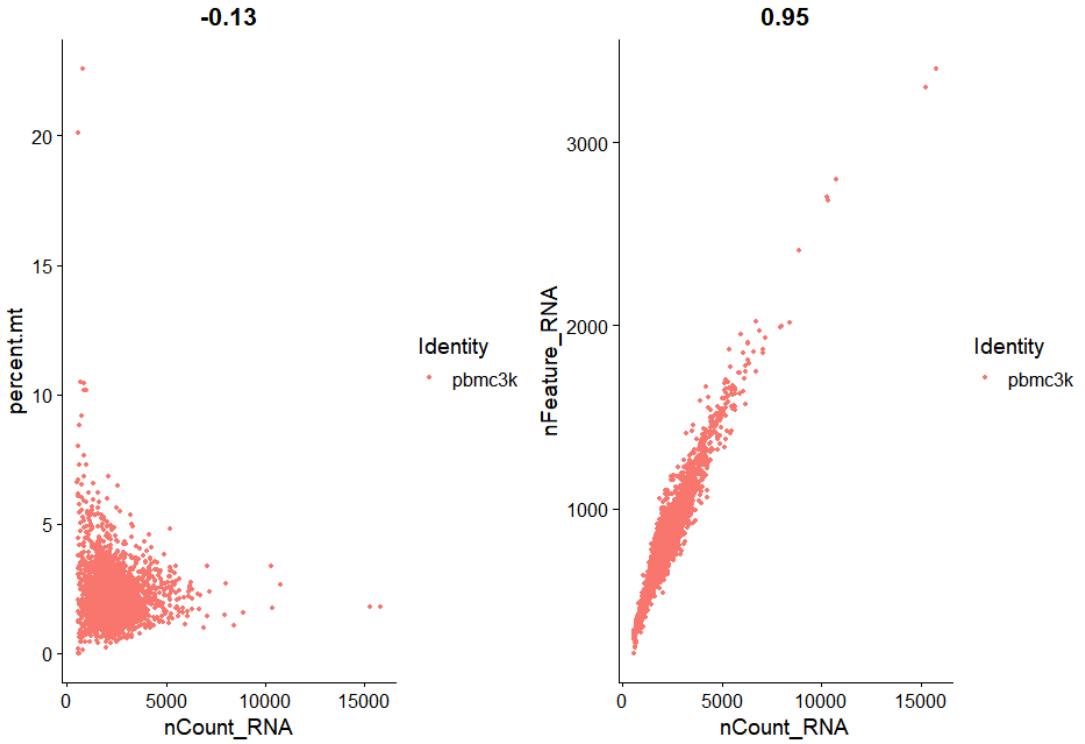
2.1 数据的质控和筛选

以特征 RNA 的拷贝数、RNA 的数量和线粒体的含量为 Y 轴，以单细胞编号为 X 轴做出小提琴图：



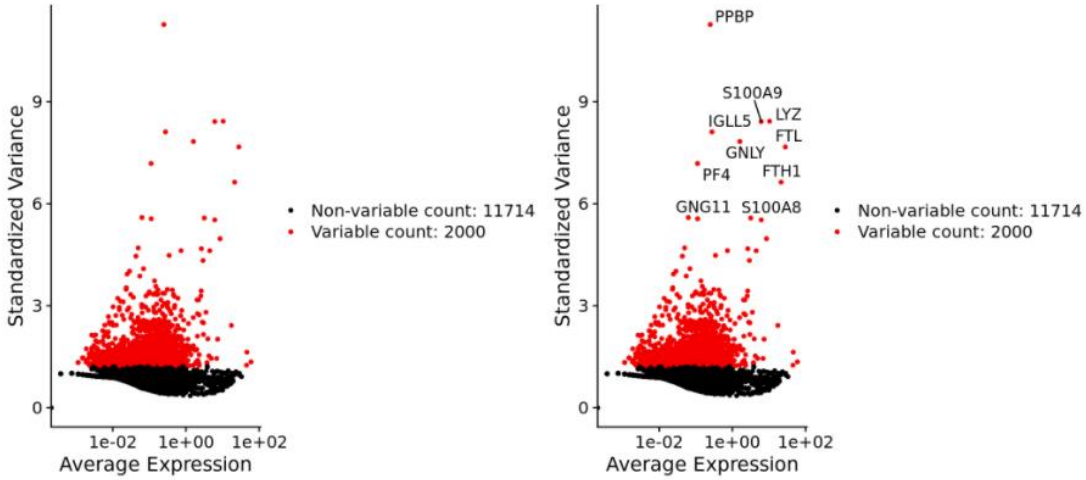
筛选出特征 RNA 拷贝数在 200-2500 之间；线粒体的含量>5%的单细胞；

做出散点图反映特征 RNA 的拷贝数、RNA 的数量和线粒体的含量这三种变量之间的相关关系：



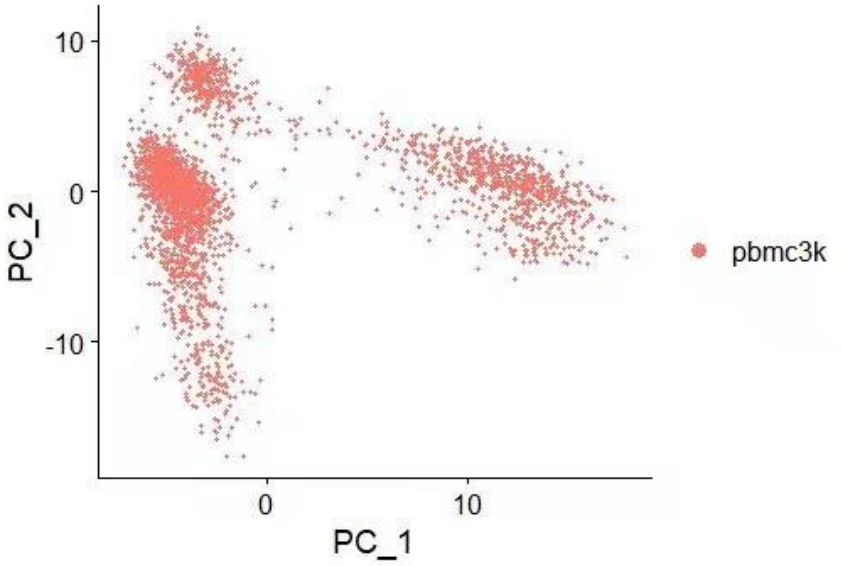
2.2 识别高度可变基因

以基因的平均表达量为 X 轴，基因的可变性为 Y 轴做出散点图，找出高度可变性基因（图中为可变性在前 2000 的基因，红色），并标注出可变性在前十的基因（右图）：



2.3 以高度可变基因为特征，对该数据集进行 PCA 分析。

可视化细胞与特征间的 PCA（下图以 PC_1/PC_2 为例）：

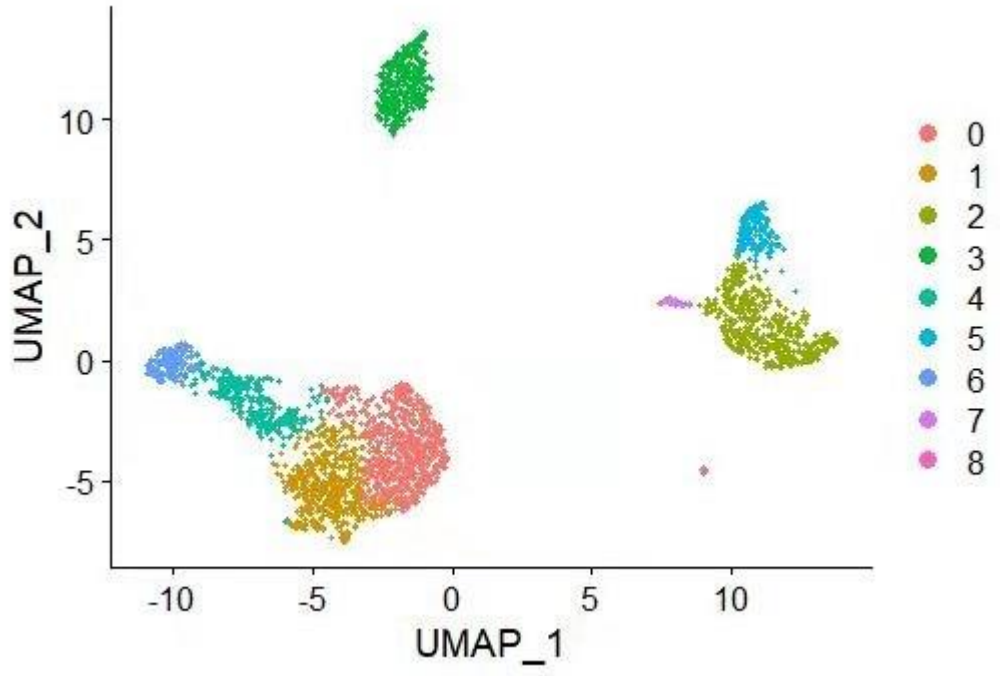


2.4 以高度可变基因为特征，对细胞进行聚类

使用 FindClusters()函数聚类细胞，结果如下：

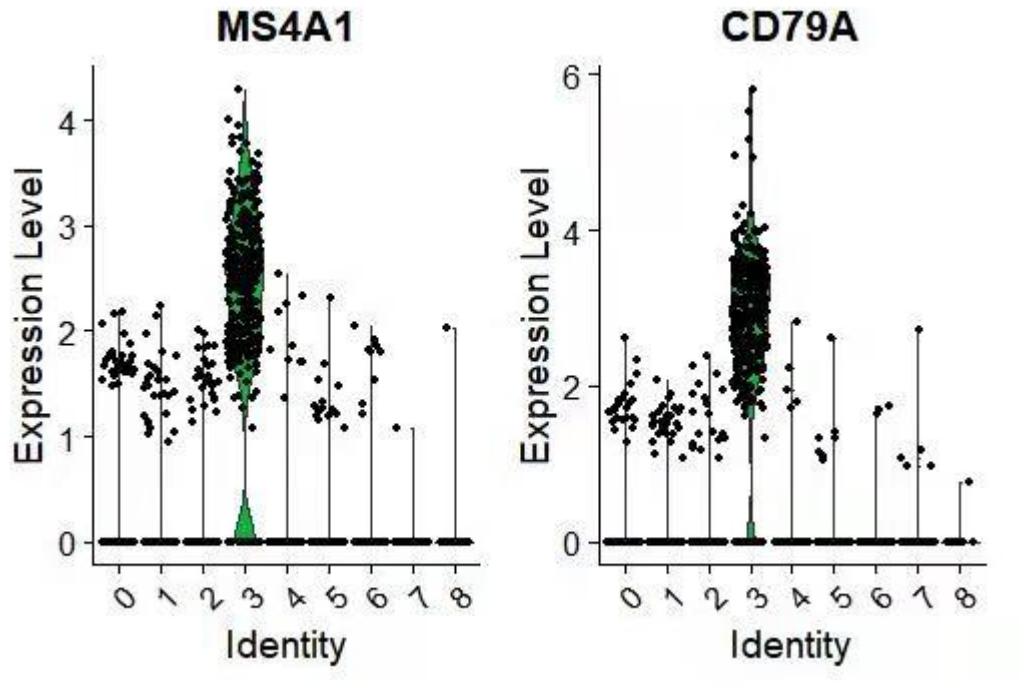
```
> pbmc <- FindClusters(pbmc, resolution = 0.5)
Modularity Optimizer version 1.3.0 by Ludo Waltman and Nees Jan van Eck

Number of nodes: 2638
Number of edges: 96033

Running Louvain algorithm...
0%   10   20   30   40   50   60   70   80   90   100%
[----|----|----|----|----|----|----|----|----|----|
**************************************************|
Maximum modularity in 10 random starts: 0.8720
Number of communities: 9
Elapsed time: 0 seconds
```

2.5 使用 UMAP 函数进行非线性降维，以对聚类结果进行可视化：

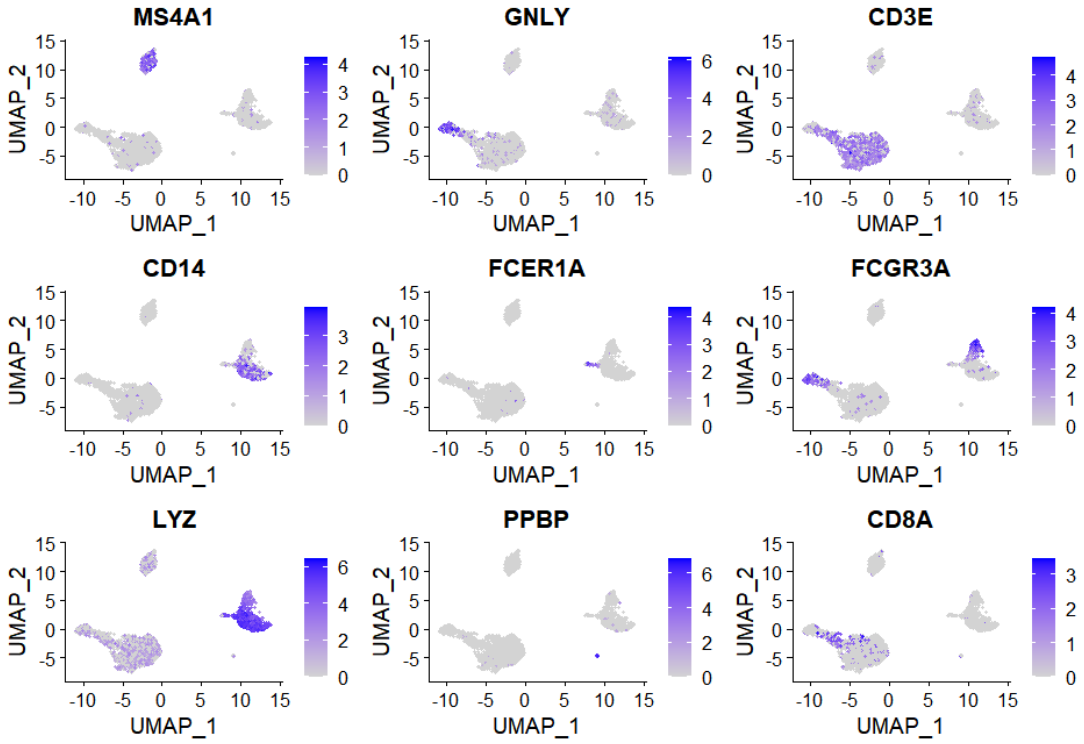


2.6 分析 marker 基因在不同的细胞簇之间的表达水平：

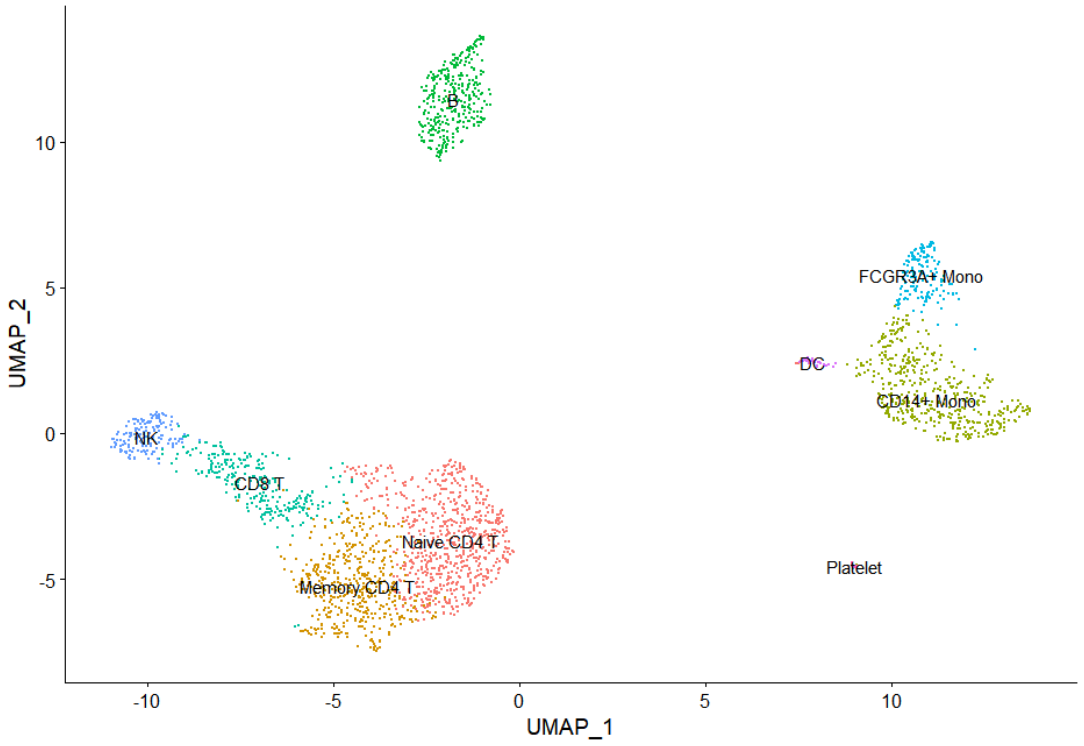
以"MS4A1", "CD79A"基因为例，做出小提琴图展示这两种基因在不同细胞之间的表达水平：



将 marker 基因的表达展示在聚类图中：



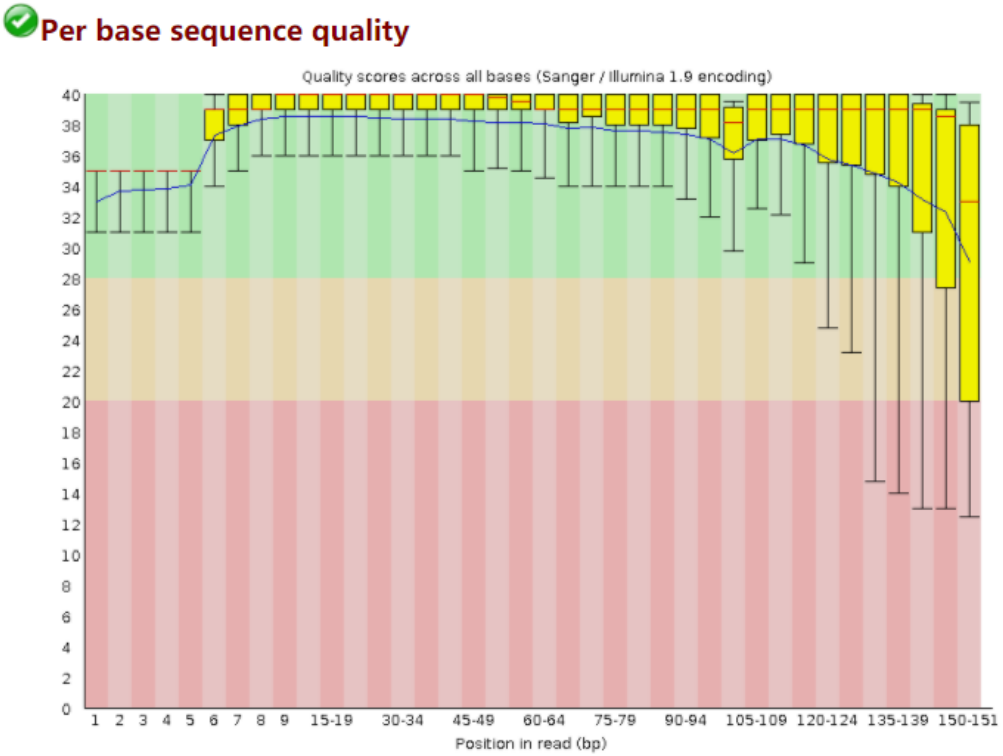
2.7 根据 marker 基因的表达量，识别出 9 种单细胞的类型并标注在聚类结果图上：



3 宏基因组装箱分析

3.1 数据的质控和筛选

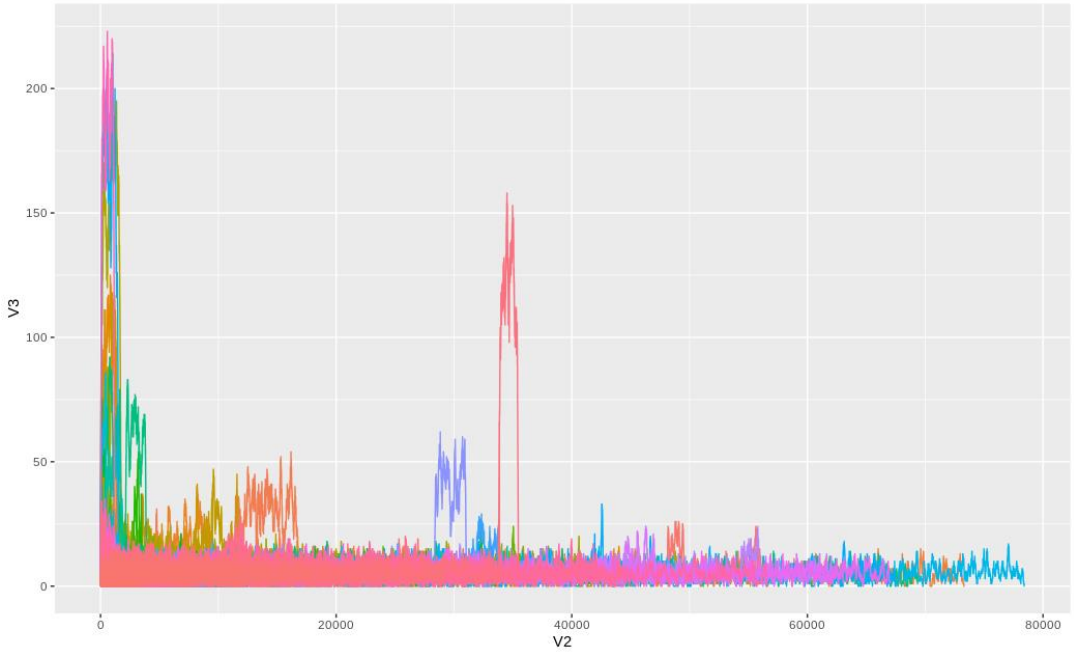
使用 fastqc 软件对原始 fastq 格式的数据进行质控：



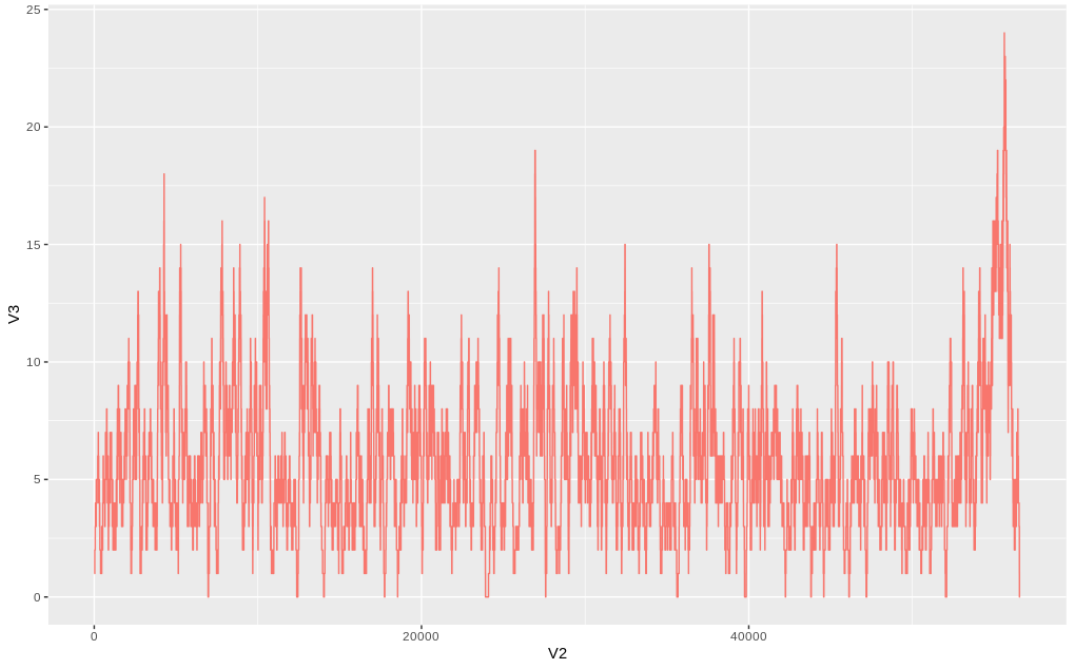
筛选出有 50%碱基测序质量为合格（质量系数>20）的数据。

3.2 计算测序的覆盖率

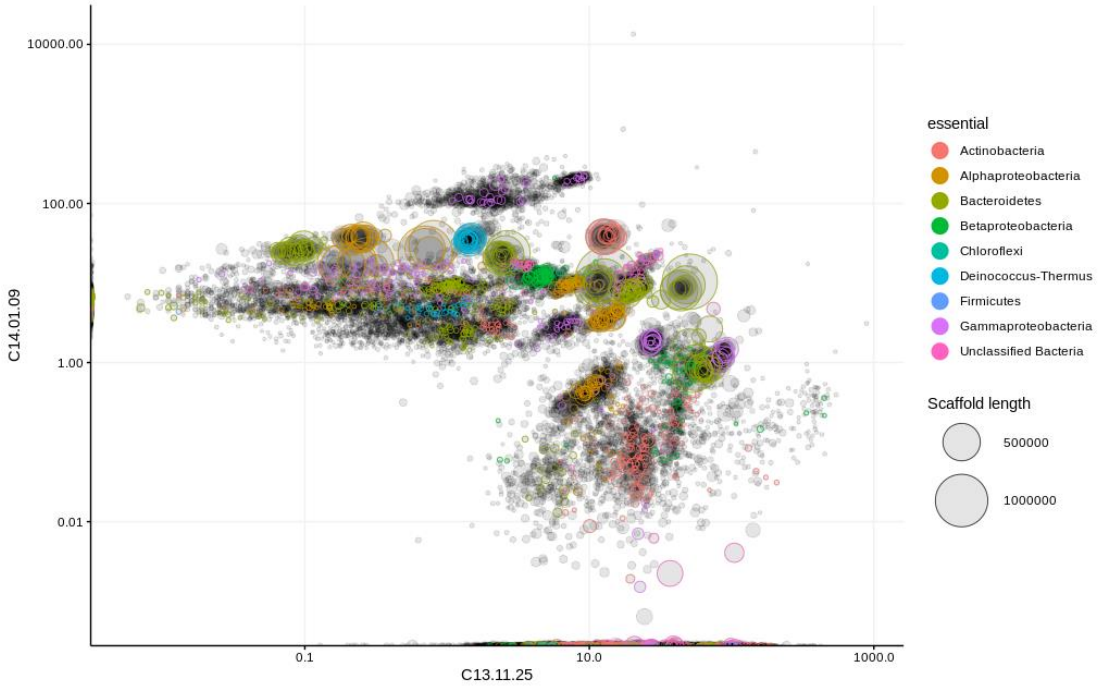
每个 scaffold 的覆盖率：



单个 scaffold 的覆盖率：

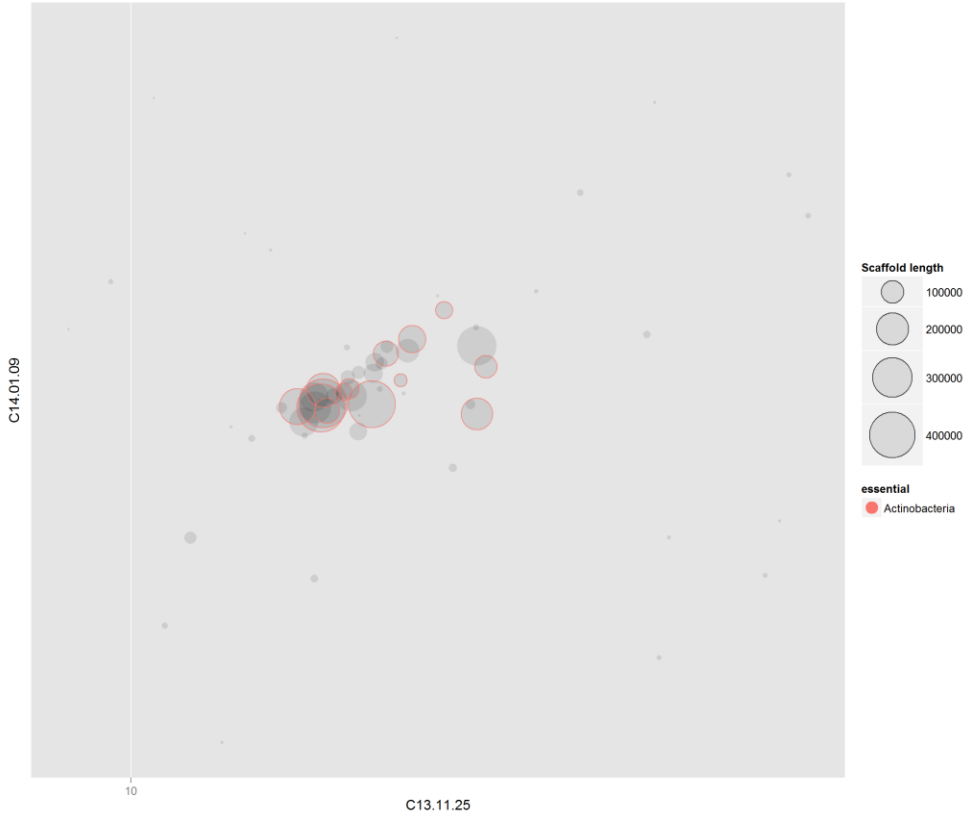


3.3 根据覆盖率、物种注释、16srRNA 等信息，利用 mmgenome 包进行宏基因组装箱分析：



可得出此宏基因组数据中不同细菌的种类，以及拼接出的 scaffold 大小等信息。

3.4 抽取出 Actinobacteria 的数据，进行进一步装箱分析:



4 总结

这两个项目都将单细胞或宏基因组的测序数据加以处理分析，根据测序数据得到单细胞或细菌的类型。其中宏基因分箱分析的整个过程在 linux 系统中完成，项目中用到了数据的

质控和筛选，变量的选择（项目 1 中选择高度可变基因，对细胞进行聚类）。

Reference:

[1] https://www.10xgenomics.com/
[2] https://www.ncbi.nlm.nih.gov/sra
[3] https://www.jianshu.com/p/67d2decf5517
[4] http://madsalbertsen.github.io/mmgenome/
[5] http://blog.sina.com.cn/s/blog_6c0267490102wf25.html
[6] https://blog.csdn.net/u012110870/article/details/82500741
[7] http://www.ehbio.com/Bioinfo_R_course/