实验项目 1-1：基因组测序模拟【应用】

一、分析平台：

1. 硬件平台：（硬件配置）CPU 1.70GHz　2.40GHz
2. 系统平台：（操作系统及其版本号）WIN10 专业版、Ubuntu
3. 软件平台：（软件系统及其版本号，若是在线分析平台，还需要提供 URL 地址）R3.4.3 、SUBLIME TEXT

二、实验步骤：

1、基因组测序模拟工具相关文献资料的调研：

通过检索公共搜索引擎或专业数据库（PubMed），查阅不少于 5 种不同基因组测序模拟工具，并对其功能特征进行描述，至少精度其中一个工具的相关算法文献。

2、基因组测序模拟软件下载：

2.1、从 ART 官网 (http://www.niehs.nih.gov/research/resources/software/biostatistics/art/) 下载 ART 系列（ART Illumina/454/SOLiD read simulator）软件。

3、测序数据下载：

3.1、从 Genbank 数据库的 Genome 子库中下载某个物种的基因组序列，建议下载的基因组规模较小的物种基因组序列，如真菌类。

3.2、利用 NCBI SRA Toolkit 从 Genbank 数据库的 SRA 子库中下载不同测序平台的测序结果数据（fastq 格式）【优先选择基因组 DNA 测序结果】。

①illumina

prefetch --option-file SraAccList.txt

fastq-dump --I –split-files SRR6823092

②454（检索 SRA 数据库，roche 454）

fastq-dump SRR6846999

③SOLID

prefetch --option-file SraAccList_solid.txt

fastq-dump SRR619952

4、基因组测序模拟：

4.1、使用 art 系列软件，对下载基因组序列进行全基因组测序模拟（每个系列至少模拟一种测序模式），截图保存程序运行结束后的屏幕回显结果。

①art_454.exe:

[single-end read simulation]

art_454 -s ./mydata_454/NC_010321.fasta ./mydata_454/single_dat 50

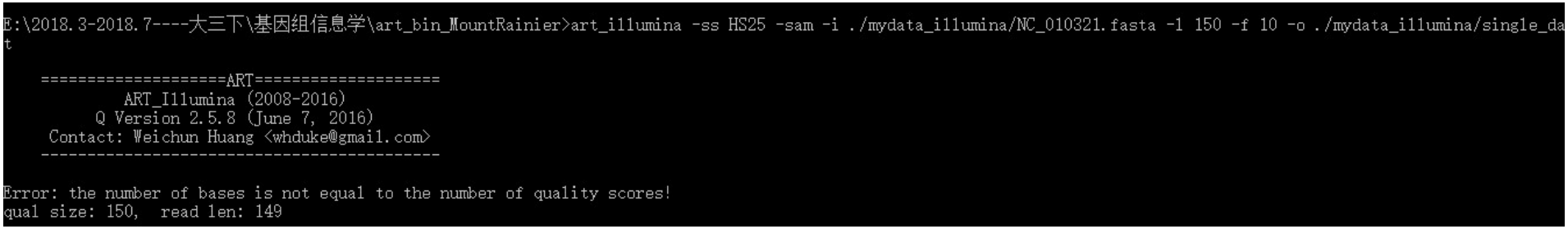
```
E:\2018.3-2018.7----大三下\基因组信息学\art_bin_MountRainier>art_454 -s ./mydata_454/NC_010321.fasta ./mydata_454/single_dat 50
==================================================================
                     ART_454 (Version 2.6.0)
                  Simulation of 454 Pyrosequencing
     Copyright (c) 2008-2015, Weichun Huang. All Rights Reserved.
==================================================================

                          Single-end simulation

Total CPU time used: 385.784

The random seed for the run:   1521006928

Parameters Settings
       number of flow cycles:           100
       fold of read coverage:           50X

454 Profile for Simulation
       the built-in GS-FLX profile

Output Files

  FASTQ Sequence File:
       ./mydata_454/single_dat.fq

  SAM Alignment File:
       ./mydata_454/single_dat.sam

  Read Coverage File:
       ./mydata_454/single_dat.stat
```

图表 1.454 模拟测序操作

②art_illumina.exe:

[single-end read simulation]

art_illumina -ss HS25 -sam -i ./mydata_illumina/NC_010321.fasta -l 150 -f 10 -o ./mydata_illumina/single_dat



图表 2.illumina 模拟测序操作

③art_solid.exe:

[singl-end 25bp reads simulation at 10X coverage]

art_SOLiD -s ./mydata_solid/NC_010321.fasta ./mydata_solid/single_dat 25 10

```
E:\2018.3-2018.7----大三下\基因组信息学\art_bin_MountRainier>art_SOLiD -s ./mydata_solid/NC_010321.fasta ./mydata_solid/single_dat 25 10
==================================================================
                   ART_SOLiD (Version 1.3.3)
        Simulation of Applied Biosystems' SOLiD Sequencing
   Copyright (c) 2008-2015, Weichun Huang. All Rights Reserved.
==================================================================

                   Single-end simulation

Total CPU time used: 139.939

The random seed for the run:   1520986184

Parameters Settings
       fold of read coverage:  10X
       read length:    25

SOLiD Error Profile for Simulation
       the built-in 35bp error profile

Output Files

  FASTQ Sequence File:
       ./mydata_solid/single_dat.fq

  MAP Alignment File:
       ./mydata_solid/single_dat.map

  SAM Alignment File:
       ./mydata_solid/single_dat.sam
```

图表 3.solid 模拟测序操作

4.2、查看模拟运算的输出结果文件：节选结果文档的第一页内容存放实验结果部分【可以截图】。

5、基因组测序模拟的数据模型（profile）创建

5.1、利用 art 系列软件，根据下载的不同测序平台的测序结果数据，创建模拟基因组测序的数据模型（profile）。

①454

perl art_profiler_454 ./454_dat_dir ./454_profile_dir fastq

②illumina

./art_profiler_illumina ./illumina_profile/out ./illumine_dat fastq

5.2、查看模拟运算的输出结果文件：节选结果文档的第一页内容存放实验结果 部分【截图】。

5.3、与该软件自带的数据模型（profile）进行对比分析，有何异同之处。【部分 R 语言脚本见附录】

###R 语言代码：

```
## ①454 平台模拟
title1<-read.table("Book1-454-profile.txt",head=F,sep="\t")
x<-rep(title1[1,],title1[2,])
#每一个打分值 重复出现的”频数”次
title2<-read.table("self-454-profile.txt",head=F,sep="\t")
y<-rep(title2[1,],title2[2,])

png("profile_454.png")
hist(unlist(y),breaks=100,freq=F,col="green",ylim=c(0,0.015),main="SRR6846
998(.fastq)与该软件
自带的数据模型(profile)对比分析",xlab="score")
hist(unlist(x),breaks=100,freq=F,col="red",add=T)
dev.off()

res <- cbind(as.numeric(x),as.numeric(y))
chisq.test(res,correct=T)
```

##② illumina 平台模拟【见附录】

三.实验结果：

1. 基因组测序模拟工具相关文献资料的调研

共查找到 PacBio、Wessim、VarSim、 metasim 和 GemSIM 五种相关工具，下面为其简单介绍

①PacBio

Rhoads, Anthony, and Kin Fai Au. "PacBio Sequencing and Its Applications." *Genomics, Proteomics & Bioinformatics* 13.5 (2015): 278–289. *PMC*. Web. 18 Mar. 2018.
PMC4678779

介绍： PacBio 测序在目标 DNA 分子复制过程中的捕获其序列信息。该模板称为 SMRTbell，是一种封闭的单链环状 DNA，是通过发夹衔接子与目标双链 DNA（dsDNA）分子的两端连接而形成。当一个 SMRTbell 样本被加载到一个称为 SMRT 单元的芯片上时，一个 SMRTbell 扩散到一个为光检测提供最小的可用体积的零模式波导（ZMW）的排序单元中。在每个 ZMW，单个聚合酶固定在底部，其可以结合到 SMRTbell 的任一发夹衔接并启动复制。将产生不同发射光谱的四种荧光标记的核苷酸添加到 SMRT 细胞中。当一个碱基被聚合酶所控制时，产生一个光脉冲来识别碱基。SMRT 单元的所有 ZMW 中的复制过程由光脉冲的"电影"记录，并且对应于每个 ZMW 的脉冲可以被解释为一系列碱基（称为连续长读，CLR）。最新的平台 PacBio RS II 通常生成长度为 0.5-4 h 的序列记录。由于 SMRTbell 形成一个封闭的环，在聚合酶复制靶 dsDNA 的一条链之后，它可以继续掺入衔接子的碱基，然后连接另一条链。如果聚合酶的寿命足够长，两条链都可以在单个 CLR 中进行多次测序（称为"通过"）。在这种情况下，CLR 可以通过识别和剪切适配器序列来拆分为多个读取（称为子读取）。单个 ZMW 中的多个亚读数的共有序列产生具有更高准确性的循环共有序列（CCS）。如果目标 DNA 太长而无法在 CLR 中进行多次测序，则不能生成 CCS 读数，而只会输出一个子读数。由于 PacBio 测序是实时进行的。

②Wessim

Kim, Sangwoo, Kyowon Jeong, and Vineet Bafna. "Wessim: A Whole-Exome Sequencing Simulator Based on *in Silico* Exome Capture." *Bioinformatics* 29.8 (2013): 1076–1077. *PMC*. Web. 18 Mar. 2018.
PMC3624799

介绍： Wessim 使用一个 fasta 文件的样本基因组序列，例如人类参考组装和基因组目标区域。Wessim 首先从指定的目标区域生成随机的 DNA 片段，这是由一个包含目标区域的坐标或一组用于捕获片段的探测序列指定的。每一个片段被长度和 GC-content 进一步过滤，以重现潜在的偏差。最后，NGS 读取是使用主要仿真平台从选定的 DNA 片段中生成的。

③VarSim

Mu, John C. et al. "VarSim: A High-Fidelity Simulation and Validation Framework for High-Throughput Genome Sequencing with Cancer Applications." *Bioinformatics* 31.9 (2015): 1469–1471. *PMC*. Web. 18 Mar. 2018.

PMC4410653

介绍： VarSim 工作共分为两步。①模拟：一个摄动二倍体基因组是通过将变异插入到用户提供的参考基因组中产生的。然后，这个被摄动的基因组中进行模拟之后读取。读取是使用正在考虑的二级分析管道进行处理的；②验证：VarSim 通过存储在 read 名下的元数据来验证对齐。所有可能的真正的读校准位置都存储在元数据中。这使得 VarSim 能够验证对 SVs 的中断点的重叠。而且，每一个对齐都用它生成的区域的类型进行注释，这样就可以只验证重叠的特定类型的变体。如果它靠近任何一个真实的位置，则称为校正。

④metasim

Richter, Daniel C. et al. “MetaSim—A Sequencing Simulator for Genomics and Metagenomics.” Ed. Dawn Field. *PLoS ONE* 3.10 (2008): e3373. *PMC*. Web. 18 Mar. 2018.

PMC2556396

介绍： MetaSim 以一组已知的基因组序列和丰富的平台作为输入。这个平台决定了为模拟选择了哪些基因组序列，以及在数据集中的每个基因组序列的相对丰富度。MetaSim 集成了 NCBI 分类法 27 的“诱导树视图”，该视图可用于交互式地选择分类法和分类法的内部节点，以配置它们的相对丰度。此外，用户还可以使用物种模拟器模拟单个基因组序列的“进化”种群。这一特点旨在模拟现实世界中许多不同的、但是亲缘关系紧密的物种可以共存。

⑤GemSIM

McElroy, Kerensa E, Fabio Luciani, and Torsten Thomas. “GemSIM: General, Error-Model Based Simulator of next-Generation Sequencing Data.” *BMC Genomics* 13 (2012): 74. *PMC*. Web. 18 Mar. 2018.

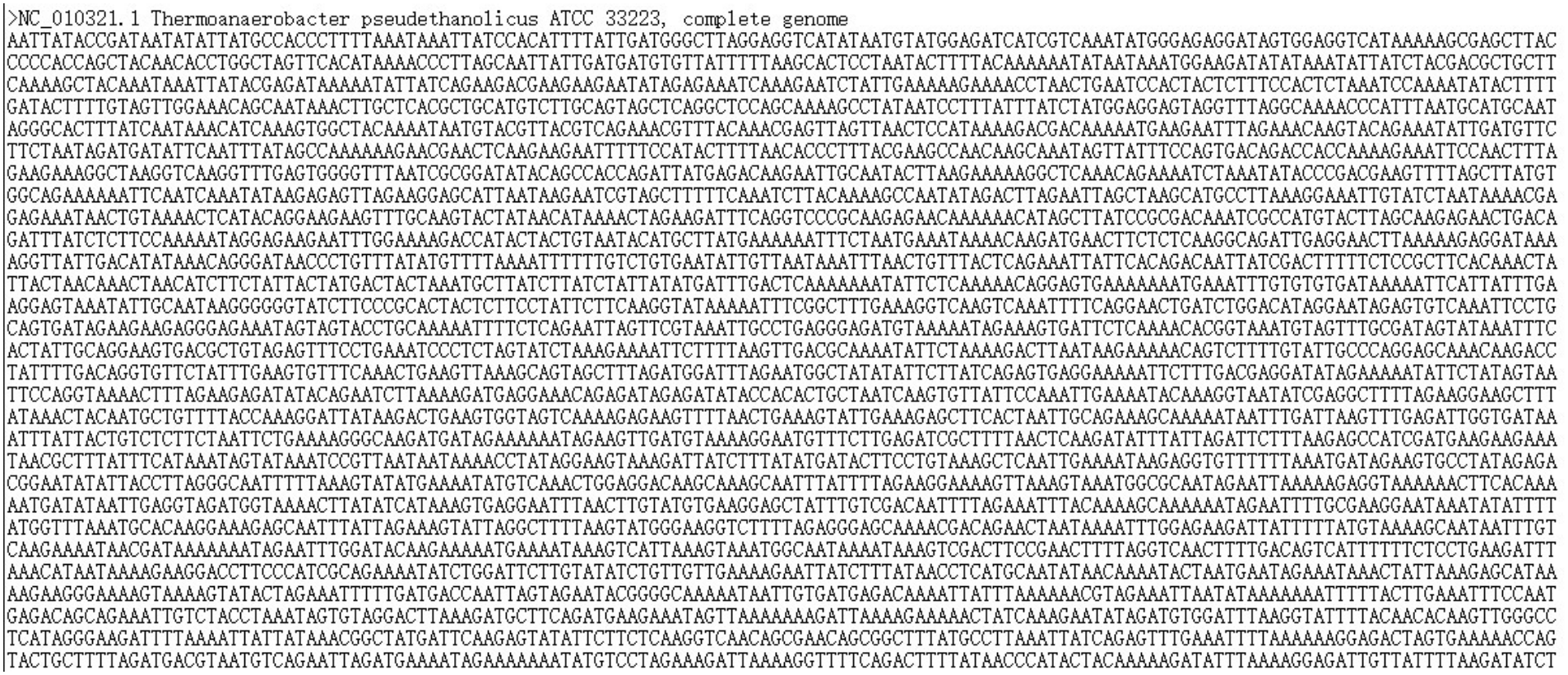
PMC3305602

介绍： 通过从两个不同的 illumina 序列运行和一个 roche/454 运行的错误模型来演示 GemSIM 的值，并比较和对比每一次运行的结果错误配置。总的错误率有很大的差异，在单独的 Illumina 运行中，每对的第一和第二读之间，以及来自 Illumina 和 roche/454 技术的数据集之间。在 roche/454 中，in 德尔的频率明显高于 Illumina，这两种技术都在每次读取结束时出错率增加。

通过对混合细菌单体型的模拟序列数据分析，研究了这些不同剖面对低频 snp-calling 精度的影响。一般而言，snp-calling 使用 VarScan 只对频率大于 3%的 snp 是精确的，独立于使用错误模型来模拟数据。误差剖面间的差异与 VarScan 的“最小质量”参数相互作用强烈，从而导致不同的测序运行的最优设置。

3.测序数据下载：

- 3.1. Genbank 数据库的 Genome 子库中下载的高温厌氧杆菌的基因组序列



图表 4.NC_01032 基因序列

4.2

①art_454:

1).single_dat.stat

0	13	13
1	0	13
2	0	13
3	0	13
4	0	13
5	0	13
6	0	13
7	0	13
8	0	13
9	0	13
10	0	13
11	0	13
12	0	13
13	0	13
14	0	13
15	0	13
16	0	13
17	0	13
18	0	13
19	0	13
20	0	13
21	0	13
22	0	13
23	0	13
24	0	13
25	0	13
26	0	13
27	0	13
28	0	13
29	0	13
30	0	13
31	0	13
32	0	13
33	0	13
34	0	13
35	0	13

图表 5. art_454 单端测序模拟结果（1）

2).single_dat.fq

single_dat.fq - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
@NC_010321.1-157520
TGGAATAGCTTCTTTTTTAAAATAATCACCAAAAGAAAAGTTACCTAGCATTTCAAAAAAGGTACCGTGACGTGCTGTTTTACCTACTCTTTCAATATCAGGAGTTCTTATGCATCTTTGGCAAGTAGTTACTCGTTTGCTAGGGGGAGT+
CC=1GCGGGGGGCJJJGCJJJJJJJ=JJJJ=GJGJJJJGJGJGGJGJGGJGG=GGGJGCGJGGCCGGGCGCGG8GCCJGGGGCGCGGGCGCGCGGGGGCGJCGCGGGGCGCGGGGGCGGG=GGCC=GCGGGGCGGCCGGC=GGGCG=GGC@NC_010321.1-
157519TTGTTGTTTTATGGTTGCCCAGTTTTCTTTGAGATACGGTAGAAGTTCATCTACTTCCATCCCATCGACTCCATGGCTTCCTTTATTTGCAACAACGCGCTTGTATGCTGCTCTCATGTTTCCTCGTTCTACTATCATTTCAAGCATCTT+
CC8GGGGGGGGGGJJJJJCJGGJ8JJJJJJJJJJJJJGGJGJJCGJGJGG1GJJJJJ=GGGGGJJGGCGJGGGGGGCCGGCCGGGGGGGGGGGGGGCGG=CJCCGGGGGCCCGCGGGGGGCGGGGGGGGGGG=GGCG=CCCGCGCGGGGG@NC_010321.1-
157518AAGGCTAAAGTGGAAAATGAGTTTAAAAAAGAGGGTAAAAAGGTAGAATGGGGAGCAAGAAGAGCCATTCTTGGGGATTTATCTCCTGCTGCTACTTTTATAGCTTACAATTACAATACTCCTGTTGATGTAGAAGAATTTAAAAAAGAG+
CCCCGGGGGGGGGGJ=JJJGGCJCGJJGJGJJJJJJJJJJJGGJCJGGGCGCJCGJ(GGJGGJGGJ=C1GGGGJCGGCGGGCG=GGGGGCGCGGGCGGGGCCCGGGCGGGGGG=GGCG=G==G=GCGG=CCGCG=GGG=G=G8GGGGGGG@NC_010321.1-
157517AGCTAAATTTAAAGTAGCTAAAGCAGAAGATTTTTTGAAATATGCTGAAAATAAAATATTAAATGAAAAATGGTCACCAGATGCAGTTGTAGGCTATTGTAAAAAAGACCCAAGCTGGAATAATAAAACTATTGTTTGTACTAAAACACT+
CC=GGCGCGGGGGGCJGJGJJCJJG=CJ=GJJ=JJGJJGJJJJCGJ=GCJGG=GGJJCG1GGG=GCCGGGCGGCGGCCGGG(GGC=GCGGGG=CGGGGGGJCGCCGCGGG=GGGCCCCGGG=GGGGGGCGGGGG=CGGGGGGGG=GCGG=@NC_010321.1-
157516TTTTTTAATTTGCATTTTATCACCTGAAAAATCTATTTGATTTTTATTTTGTTCATTGTACGCCATATTTTTGTCATATTTTTTTCTATCATATCTCTATAACGTCTATATTTTTTCATGTCTATCTTAGGTATTATATATGCTGCAGTT+
CCCGGCGCGGG=GJJJJCGJGGGJGJJJCCJGJJJJJJJJGGGJJJGGCJJGGJJGGGGJGGGCGGJGC1JGCGGG=GCG8CGCGGGGGGGGGGGGGCGCCJCGGGGGCGGGGCCGGGG=GGGGGCGGGCGGGC==GCC=GCCGGGGGGGC@NC_010321.1-
157515CTAAAATTATAGAATCTCCTTCTTCTATCATTGATGCAGCAATTTTACCTATCTTCCTTTTTGCCTCTAGTTCCACATCTTCTTTTTCAAAAAATGCAGGCTCATAATGAGTGCCAACATTCAATATCGCGCCACCGTGAGTCCTTTGTA+
CCCGGGGGGGGGGJJJJJJJJJJGGJCJJJGG=JJGJJJGJCGGG8CCJGG(GGGJJJJGJJGGGGJCCCGJJJGGG=CGCGCGGGGGG=GGGGCGGGG=JGGGGGCC1GGC8GC8=GCGGCGGG=CGGGGGGCG=GGGGGCGGGCCG@NC_010321.1-
157514TAAACTCGGCTACGATGTAACGATCTTTGAAGCTCTTCACAAACCCGGTGGCGTACTGGTATACGGTATTCCTGAATTCAGGCTACCGAAAATTATAGTGGAACAAGAAGTAAAGTTTATACAACAGCTGGGCGTAGAAATTAAAACCAA+
CC=GGGGGGGGGGCJJJJG(JJJJJJJCJCGJJJJJGGGGJJJGJJ8GJGG=1CGGG=8G8GG==JGJGGJJGG8GG8=GCGCCGGGCGGGGCGGGGGGCJCC=GGCGGGGGGGGCGGGGGGGGG1GGGGGCGGGGG=CGCGG=1GGG@NC_010321.1-
157513AGAAATTATAGAATTAGATGAAGAAATAATAAATCAAGCTATAGAAGAATCAAAGGAGAATGCCAATTACCTCAGAAAAATAATCAGAAAGAAAAAACATGCTCTTCTTCCAGAAGTAGAAAAAACTTTATCTGCTCTTTCGAGTGTTTT+
CC1GGGGGGGGGGGGGJJJGJJJJJGCJJJJJJGJ1JJGJGJJJJGGGGCGGG=JJGJJGJJGGGGCGGJCGGGG=8J=CCGGGCGC=G=G1GGGGGGCGJCGGGGCCGGGGCGGGC=G=GGGGG8CCG18CGGGGGGG8CGGGCCGGG@NC_010321.1-
157512TATCACTGAAGCACAAAGCGTCCAAGAAATAAACATAAGAGGCAACAAAGCCTTCTTAATCACAAGTAAGTATTCTCCTGATACGGTTACACTTGCTTGGAATGAAAATGGAGTAATGTATGCCTTGTCTATGTTCGGTACAAGTTCTTC+
CCCGGGCGGGGGGJJJJJ1JGJGJGJCJJJGGJCJJJGGJGJGJGJJJGJGJJJCCCGJGGJGGJ=CGGJGJGCGGCGGCCGGGCGCGGCGGGGG=GGCCJG=GGGGC8GCCGGGGGCCG8=GGGGGGCGCGGG8CCGGGG(GGCGCCCG@NC_010321.1-
157511CGCACAACCTGAGCTCGAATAGTTTTTCCCATTTACACTTCTCCCCTCTAACTAAAAATAAAAATTCCGTATTTATTTATCGGTTTAAATCGGCAAAAATTTAATAGAAATAAGCGGTTCTTCTTACAAAACACTAAAAAGAACCGCTTT+
CCCGCGGGGGGGGJJJJJCJJGJJJGJGJJJGJGJ1JGGJCGJJGJJJC1JGJJJCJJGGJGGGGGGGGGGGCCGG88CGCGGCGGGGCCGCGGGGGGCGJGC=CGC=GCCCCCG8GGCGGGCGGCGGGG8CCGCGGGGCGGC==GGGG=@NC_010321.1-
157510AATGAAGATTTAAAATCAGTAAGTGGTTTATCAGGTGGAGATGCAAAAAAACTATATGAAAGAATAATCATTGGCAATACTCTTTCTTGTGAAACAATGGCAAAAGCTGCTGCTTCTGCTCTTGCAGTGACAGAAGTAAATGCTTCTATG+
=C1GGGGGGG1GGJJGJJJGCGJJJGGG8JCJGJJJJ=GJJG=JJGGGJCJ1JCGJJGJGGGJ(JGCGG8GGGGJGCCGGCGGC8GGGGGGGG1G=CGGGJCGGC=GGGGCGGCGCGGGGGGG8GGCCCGCGGCGGCGGC=CCGCG=GGG@NC_010321.1-
157509TTATTTCATTGAAAGAGAAATTCTATTTCTCTCAATATCAACACTTAAAACTCTAACTTTTACAATGTCTCCCACTGAAACCACATCAAGAGGATGTTTTATGTAATTTTGAGACATCTCAGAAATGTGAACAAGCCCATCAGTGTGAAC+
CCCGGGG1GGGGGJJJJJJJJGJCJJGJGCJJJGGJJCJ==JJJ=JJCJJJJJJGJCJJCJJGGJCGCG=JGGGCG(GGGGCC1CCGGGGGGCGGGGCCCCCGGGGGGCGGG8C=GGGGG=GGCGGGG=CGCGGGCGCCG=1GC=G==CC@NC_010321.1-
157508AAATGCTTTGCATATTCCTTTCACACAGCCTGCGATAAAAATGGATTCATTTTAGCAGCGAAAGTAACAAGCGCAAACTTGCATGATAGCGTAATGTTCGAAGAAGTATTAGATGAAGTCATAAAGAATGTAGGAAAACCCAAGGCAGTA+
CCCG=GCGGGGGG=GJJJJJCJJGJGGCJJGGJJGCCGJJJGGJ=GGGJGJGCJGJCGGGCGGGGCJGGGGG8GGGGGGGGGGCGGGGCG=GGGGGGCCGJ8CCC=1CGGGG=GGGCGGCGGCGGCCCGG=GGGGGGGGGGGGGC1CGGG@NC_010321.1-
157507AAGAAGCTTATTCAATCAAGGAAATCTTTTAATTTTTTGCTCCTGCTGGGATGCCTCAATTTTCTCAAAGCCTTTGCTTCAATTTGCCTTATGCGTTCTCTTGTAACATTAAACTCTTTTCCTACTTCCTCTAAAGTTCTTGCTCGACCG+
CCC1GGGGGGGGGGJGJGCJJGJGJCGJ8JJJJGCJJGJGJGJJJJGJJJGGCGCCGJJCGGGGGJGJ1JCJCCGG=8G1GGGCGCCG1GCGGGGCGCGCG8JGGCGCG=GGGGCGGCCGGCGCCCGCGG=GCCGCCGGGGGGCCCGGGGG@NC_010321.1-
157506TATAAAATTTATGGAAAAGTTAGGGGCCTATGTAATACCCTACAGTAACAATGTCACTTCTCCAAGAGATATGAGCATGTATATGAAAAATCTTTTAGATTATGTAAATGCTCATCCGGATACTGCAGGAGAATTGATGTATTATCTTAA+
CCCGGGGGCGGGGGJGJJGJJJCJGCJJJGJJJ=J8GJJJ=GJGJJG8JJJG8GCC=GGGJGG1CGCGCGGGGGGGGGGGGGGC=CCGGGGCCGGGGGGGJCGCGGGGGGCGGGGGGGCCCCGCGGGGCG=GGCGG8G=GCCC=CCCC=C@NC_010321.1-
157505TCCCTCTGGTAGCTGAATAAATGCTCCAAAATCTGTGATGTTTAACACAGTGCCCTCAACAACATCTCCAACCTCAAATGGCATATAAATAAAGATCCTCCCTCTAAAAATTTTATTAATGAGTTCATTATATATTAAAATACAAAATAT+
C=1GGGGCGGGGGJJ1J8CJJJJGJJJJGJJJJJGGJJGJGGJ=GGJCJJJG11GGC1GGJGGGGGGGGGGJJ1GGCCGCC=GGGGGCGCCGGCGCG=GG=GGGGCCGGGGCCCGGGGG==GG1GC(CGCCGGCCCCCGGGGGCGCCGCG@NC_010321.1-
157504TATTCTACAAGCCTGTAGTCTAAATAAGGTACTCTTGCTTCAAAACCTACTGCCATACTCATTTTATCCATTCTGTCAAGAAGCATGGGCATAAATCTTGTTAGGTTAAGGTTAAACATTTCCCTCATTCTGGTATTTTTGGGGTCTTCT+
CCCGGGGGCGGGGGJJJJJCJJJGJ(JCGGJJJJJGGJJGGGGGJJJGJ8JJGGJ=JGGGJGG(CGJCGCGGGJCGGGGGGG8GGG1CG=GGGGGGGCG8JCCGGGCGCCCG(G=GGGGGGGGGGGGCGCG=GGCCCC8GGGG=GCGGGG@NC_010321.1-
157503ACTTTTTGGAGGGAGAATGAGGGGGAATGATGGAAGGATTAATGTGGATAGCGCCCGTCTCGGGCTTACTTGCACTTTTGTTTGCCTTCATTAACGTCATGAGGGTTGTCAAGAGCCCTGCAGGGAACGAGCGTATGCGCGAAATATCCG+
CCCGGGGGGGGGGGJJGJJJGJJJJJJJJGGGJGCJJJGJGJGJGJJCJJGJJJGJCGJJJJGJGGJCGCGGJ1GCGG1GCGCCGGGGCCGCGC=GCGGGCC8GGCGGGGGCCGGGGGGGGGCGGGCGGGGGCGG=GGCGGGGGGGG=GG@NC_010321.1-
157502CTAAAAAACAAATAAAATTTAGGGGGCTTAAAAAATGAATCTTGAACAAATTGTATTCACTATTATATCTCACGCAGGAAATGCACGAAGCATTTGTTTTGATGCATTAAAACATGCAAAAGAAGGCAACTTTGTTGAGGCAGAAAATTG+
```

图表 6.art_454 单端测序模拟结果（2）

3).single_dat.sam

```
@HD VN:1.4  SO:unsorted
@SQ SN:NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome  LN:2362816
@PG ID:02   PN:ART_454  CL:art_454 -s NC_010321.fasta ./mydata/single_dat 20
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-341772-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1739966-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1706146-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1709656-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2335800-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-670868-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2056870-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1223419-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2213320-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1824338-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2314397-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2101776-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-414546-1    0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1813855-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1582451-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1545808-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1952160-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1140101-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-126462-1    0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-701053-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2098167-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-228331-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-714339-1    0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-519275-1    0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-935064-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-970270-1    0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1602925-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2137105-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-2077953-1   16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-1274007-1   0   NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-318185-1    16  NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
```

图表 7.art_454 单端测序模拟结果（3）---汇总分析

②art_illumina:

1).single_dat.aln

```
##ART_Illumina  read_length 150
@CM art_illumina -ss HS25 -sam -i ./mydata_illumina/NC_010321.fasta -l 150 -f 10 -o ./mydata_illumina/single_dat -rs 1520985769
@SQ NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome 2362816
##Header End
>NC_010321.1    NC_010321.1-157520  1761454 +
TTTGCCAGTCTTTCTTGTAATTTTTCTCTATCATAATCAGAGGTAGTTTCCTCAATTTGAGCCTTAATTTGATTTACTCTCTTCTTAATTTCTGCAGCATCGCCTGCTCCGCCTACAATTGTGGTATTCTCTTTTGTAACTTTTACTTGT
TTTGCCAGTCTTTCTTGTAATTTTTCTCTATCATAATCAGAGGTAGTTTCCTCAATTTGAGCCTTAATTTGATTTACTCTCTTCTTAATTTCTGCAGCATCGCCTGCTCCGCCTACAATTGTGGTATTCTCTTTTGTAACTTTTACTTGT
>NC_010321.1    NC_010321.1-157519  238523  -
GTTTGTATAAGGCCCTTAAAGATAGGTTTGGGGAAATAGCCAGTATAAATTAATTTCGAAAAATATACTTGTAAAAGTAAAAACACAATATTAAAGGGGAATGTTAAAGGGGGTGAGATAAAATCACAAAATTGTTAAATTAAAAATTTA
GTTTGTATAAGGCCCTTAAAGATAGGTTTGGGGAAATAGCCAGTATAAATTAATTTCGAAAAATATACTTGTAAAAGTAAAAACACAATATTAAAGGGGAATGTTAAAGGGGGTGAGATAAAATCACAAAATTGTTAAATTAAAAATTTA
>NC_010321.1    NC_010321.1-157518  818824  -
GTTGCGTGGAAAAAAGCTGCTGACAATATGTACCTTCCTTATTCAAAAGAACTTGATATTATACCACAGGATGACAGTTTTTTGTATAAAGAAAGGATAACAGTGGATGAAATACCTGAGGACCAATTTCCACTTTTATTGCACTGGCAT
GTTGCGTGGAAAAAAGCTGCTGACAATATGTACCTTCCTTATTCAAAAGAACTTGATATTATACCACAGGATGACAGTTTTTTGTATAAAGAAAGGATAACAGTGGATGAAATACCTGAGGACCAATTTCCACTTTTATTGCACTGGCAT
>NC_010321.1    NC_010321.1-157517  1806808 +
GTTGTACCAAAAGAGGAGTACCTCTTATAATTTCAACATAGCTTGCGCCTATAAAGCTTATAGGTTTTATAGAAGACATTTTAATAAGAGCAACAATCAGACCTAATATAACTCCTATAGTCACTGCTAAAAAGGTGAGTTTTAAAGTCA
GTTGTACCAAAAGAGGAGTACCTCTTATAATTTCAACATAGCTTGCGCCTATAAAGCTTATAGGTTTTATAGAAGACATTTTAATAAGAGCAACAATCAGACCTAATATAACTCCTATAGTCACTGCTAAAAAGGTGAGTTTTAAAGTCA
>NC_010321.1    NC_010321.1-157516  2290633 +
AAGTTACTTAAAGACTTTTCAGGTAAAGCTCCGGTCTCATACCTTGCAATAGAAGCTTCGCCTATGCCCAAAAGTTTCGCAAAAGCCCTTTGACTTAAGCCGTAATTTTCTCTAATATTTCTTATTTCCTTTGTTGAAAGTATGCCGTGC
AAGTTACTTAAAGACTTTTCAGGTAAAGCTCCGGTCTCATACCTTGCAATAGAAGCTTCGCCTATGCCCAAAAGTTTCGCAAAAGCCCTTTGACTTAAGCCGTAATTTTCTCTAATATTTCTTATTTCCTTTGTTGAAAGTATGCCGTGC
>NC_010321.1    NC_010321.1-157515  1136231 +
AAAATGAAGGAGAGCAAGTTTTGACCTATAAAGTACGTGTATGGACGGAAGGTAGAATGAACAAAATCGTACGTTACGGTTATGCGGAGATTACATTTAATATTAATGTAAATATTACGAAGATGGGTTCTGATAAGAGTGCAAATTATA
AAAATGAAGGAGAGCAAGTTTTGACCTATAAAGTACGTGTATGGACGGAAGGTAGAATGAACAAAATCGTACGTTACGGTTATGCGGAGATTACATTTAATATTAATGTAAATATTACGAAGATGGGTTCTGATAAGAGTGCAAATTATA
>NC_010321.1    NC_010321.1-157514  2183124 -
CCTGTGCGTTATTATTTGTTATCATTCTTAGCGCTTTTTTTAGTTTGGCTATATCCATAACGTACATGGCAGAATTTATCTCATGGATGCTTTTTTCTTCTTCCGTTGCATCTTTATCTTCCACAATCCTTATAACGTTTCCACTTTTAT
CCTGTGCGTTATTATTTGTTATCATTCTTAGCGCTTTTTTTAGTTTGGCTATATCCATAACGTACATGGCAGAATTTATCTCATGGATGCTTTTTTCTTCTTCCGTTGCATCTTTATCTTCCACAATCCTTATAACGTTTCCACTTTTAT
>NC_010321.1    NC_010321.1-157513  2015840 -
CTACTAAAATAATATACCACATACTTCCCAAATACACAATAAATTTAATCCCTTTGACAAAGTTCCAAAAGTACTTTCTAATTCTTCCTTGTCAACCATAAAACACCTTCCTTTGATTTTCCCCTTAACATTATAACATAAATTTTTAAC
CTACTAAAATAATATACCACATACTTCCCAAATACACAATAAATTTAATCCCTTTGACAAAGTTCCAAAAGTACTTTCTAATTCTTCCTTGTCAACCATAAAGCACCTTCCTTTGATTTTCCCCTTAACATTATAACATAAATTTTTAAC
>NC_010321.1    NC_010321.1-157512  1016032 -
ATTAATTCTATTCAGTCAAATATAAAACTGAAGACAGGGGAAGTCAGTGAAAAGCTTTTAAGAGGTAGACACACAACAAGAAGTGTAGAACTTTTATCTCTTGATTTTGGAGGGTATGTTTTAGATACTCCTGGTTTTACTGCTTTGACA
ATTAATTCTATTCAGTCAAATATAAAACTGAAGACAGGGGAAGTCAGTGAAAAGCTTTTAAGAGGTAGACACACAACAAGAAGTGTAGAACTTTTATCTCTTGATTTTGGAGGGTATGTTTTAGATACTCCTGGTTTTACTGCTTTGACA
>NC_010321.1    NC_010321.1-157511  694299  -
AAATTACCAGCGATTATTCCTATCGTGTTGTATAATGGTGTTAATAGATGGACTGCTTCTTTGAGCTTTAAAGAAACGATAGATTCATATCAATTATTTGGAGAAAATATAATAGATTTCAAGTATATTCTAATTGACGTAAACAGATAT
AAATTACCAGCGATTATTCCTATCGTGTTGTATAATGGTGTTAATAGATGGACTGCTTCTTTGAGCTTTAAAGAAACGATAGATTCATATCAATTATTTGGAGAAAATATAATAGATTTCAAGTATATTCTAATTGACGTAAACAGATAT
```

图表 8.art_illumina 单端模拟结果（1）---序列对比

2).single_dat.fq

single_dat.fq - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-371297-1
TAtATTTTAAGTTACaTCATAATAAGTTTCCTTTTGAATAATTTTnTnCTGTACTATGCTTTAAAAGTAacGGGGAAGATTTTGTTTAAAGAGTATTGGAATGAgAGTGTgTTATTAGAGTTGGGTACTTAtATAGTTATGATTCCnGTTTCATTACTTCTTGCATA
TATTTATTTTAAACATAGCCTTgTGTTTTTTTTGTGTTATCTTTAAaTCCACTTATATTCAaAGnATATGCTTTTAGGATGATTAGAGaTTTGATTAAGGCAAATAAAAGGCTTAACgCCATATATn+,**-@@@@;8-88--,-,-,DD(88,FFF??
33<<D93*11....!*!--*,--,-*-)666BDDDD,-,-....77-,BBBB,FFF444)D-,-CCFF??-,,,-,-D-*99,FF--+,;;<<?,-*B@D,-,*,66,),)HG@@!,CCC-,::--II,AAI',DD,,FFF,;;;;,,,--,-,--88
--,5//////,,,,,,--:@@CC!,GG+IB@,DDCC,),-,)I-+,,*6666IHH,DD,CCI,,D,DDD,D33IIAA):::,..22CCICC88I!FF,-,D,,,@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223,
complete genome-2138044-1
TCCTCTAAAGAACCCATATCAACCAAaATAATGATTCCTTTGGAGGTAcCTATATTCTTTAAATAAccTTGAACTTTTAGCATAATTTCCTGTGTTaACATCTCAATTGGCnTATCAAAAGCTTCAAAAACAAACTGTCCCAAAAGCCTATTGGtGACGCTGGCTAT
GCTACTGGCAGTTaAAtAACCATGGGnTATTATTtTAgAATTTATGAGATTTGTATTGTCACTTTTATTTAAACTTTT++@?I,,HHH-4>799-,,D)66IIDD!-*55,-),,77FFD56-==,I!-,),D72DIII000,::--II-::,3111++,,-
CC@CC66D-,D//,,--,ID-66@@88I,-,,-***+,+>AI22222-FFF,,I-AAA4455->>--<EFF!,--,--@BI(+,-I,,-,II,,-;;)@@-CCEE,IEEE'I,EE*<9!-,!??666,-I,,,88:D-,==,-',-GGGG*?@<HHH,5555
@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-698991-1
GGGACAnTTTcATGAGCAAAAAGAAATAGATTTTGCCgCAGGTTGTTaTATTTTAATnCATAAATCAGTATTTGAcAAAAtAGGATTTCTTGCTGAAGAATATTTTCTTTACTTCGAAGATACAGATTTTTGCGTGAGAGCTAAAAGAGCAGGnATTAAAAcCTTAT
aTGAACCTTCTGcGAAACTTTGGCATAAGGTTAGTAGCtCTACaGGTGGTGAGGAATCATTAATAACATTaTACTgTGGgAATAGAAAtAGGCTTTATTTTAATGATAAATTTAA+HHH,+I,HHH--,--,D555//DHHH,-D-1...,99!-)=<DD+--!--
444211D!,-->(=I-)--(CCCI,!<<==,,::-@CC-AA+,-I;7DII,-....DCCC,(EED-47-,--,,,,=====,,,,-,-,-,,1111,-,,,CC-,66???=)-??,-!+,33FFDD-,,,+<<=-;;;EE--,IIAA22,-(*-,,,,,,)DD-
FFI-,--CC)-,FF66D<9-,21!I,-**,II!66-DDB<<,,DD,DEE-7666GG--,,FFFBBEFF@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-576182-1
GCAGCAGTGAGGTTTGaAAAGTCAAAAGTAGGAAGGCGtGCttTAATTACTTCTTTAGGtAAAGCATTAGAAGCTTTAGAGGGAAAAGaAGGCACAGTTaTAAGCTTATcAAAATAAATTAAAAATTTTTCACAAAAAAACgATTGACAAAGAGAGATAAAAGTGAGA
GAATAAAGTTAAAATTGAA+,),,D-,---2/???,!???-,+4444+,,DD;=DD,-!D-,!-HHFF,-76,800,??!666,,,DD(-CCD+BBB,-,006::::-!,0-I,,,-;;,-44,(II,-!1133,HHH995:::::::85DD,EEEEED)!
*HH*,,---,D,--(-3300(-,,,-IDG,,,,,554444II-::@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-627465-1
TAAGTTCCTGCTTCTTGGGGCAGGGGCTATGGGAGTGgTTAAATACCGCCTGCTGGGAATAGGCACTCTCTGAAAGGGCGGnCTATAAATAGCCCAGCCTTCcAAACTgGTgaAGTTTTGC+,@@D7;;;,*-DD-;;,,,,--CCCC,---
000I---!;;332,,7:-FF-*--HHF,,-,67I----,,DI322666)II,-,D-000,),<<?-,66FF,,HHH-,,-,,!I,;;;;-,@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete
genome-1670614-1
GCCTTCTTTAACATAgTTGGTTAGTGTTGATCTTGATATCCGCAGTAGTTCcAAAAGTTCTTTTGCTTTCATATTTtTATCGCCTCATAGATATTATcGCAACAACCAGCAAAGTGAAGTAAATcTtTTTTGTTAAAAATTTTAGcATGTTTAAACAAATAGGTTGA
TTTTTAGTACTATATTATaTnATTTTATTcCCTACTTATAAAAAAGTACAAAAAAGGgAATGCTTC+,IIGG,366FF-,,+!?AFFII-,-(CC,ID(55,,,,-::+--,-I,DD*,EEEE-=?-AAAA,,333*--,111!D-D,IIID,-,---D-CC-+!
+DII-GD??,,-111-*DII--?>A,-I(////,HH:::::AAAA+-!,,+IIIHHH,544,IIICC--,,,,,,--,-,,,,>=,,--)-6777,FF(DD,-IA@,,110000,D+-////..@B,FF-,,==D@NC_010321.1 Thermoanaerobacter
pseudethanolicus ATCC 33223, complete genome-585981-1
AaAAtACGATAaAAATTTTAAAAAAGAACTtATAAAAATTCTTTTAACAATGGAAGGCaTAAGnGTTgAATATTCTACAGGAaACGGAAATGACGCATTAGAAtCgTTAATTAAAATTAAACCGGACGTTGCTATTTTAGATATAGGATTACCAGGAATATCGGGCA
TTGAGGTGGCGAAAAGATTAGAGAATACATGCCTTTTTTTTGGAAATa+*!.....-,,,,C>>.....,,,,,,,77,-,-,====;BB*5577II-@@,;;IEDD-,*;;-,,9/!AA,,FF,D-,-CC,,D-HH<<;,-,I-,,EE-III*,!
FFFF768888DBDDD==FA-,,EFI-,)GGGG-,,--,,CC,FF*BB,;;00-++-99:,)CCD(AB)A5,D;;==,*FF--+-CCD-)-,,FF<<AAAA;;II888,!@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC
33223, complete genome-1063268-1
AGAAGAAAAtATTATAAGGTAAAtATTACTATTAGAAAAAAAATGAAATAATTTCTATTAAAAAAGCTGCTTCAAAACAAnACAAcTATGGTAgtGaATTAGATATAGGAACTACTACTATTtCTaCTTATTTGTATGACTTAGATaAAGGTAAGGAAGTAGATGTG
TATTgnTGTTTAAAaCCACAGAGAACTTTTGGGGcTGATGTTATAACGAGAATAACTTATAGTATTTCTAATTCTCAAGGtTTATATCAGTTGCATTCgGCTgaAA++,DD,9666),+GG-D44@@+444,D11,,-
III,,-------,'111DFF;;;-,)88//11,,,D,-,::,A@AA-DD,*ICC!,DD=;--,!-!-IG(,-*I(IFF66,,,-,,,I-CC,I-,,66,CBBD,--*--HH--,D)CC88DFF00CC-,*I-,D,,,->>,,,,IIIDDD*33,,-,-
IEEDFFFF1111--*,--CC,-DD--*,22-CCI99-,-,*,444--II44---HHAA!88,,)*,D,>>-D,??I!,,D,,II@NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome-
665933-1
GGGCcAATTTTTTGCCGACTTTTGGATGTTTtGGGAGAAgAACGAcAACAATAGAAAGAGCATTGCAGATATAAAAAAGCACCCCCACTTTTTATGCGTcgAAAGTGGGGGTGTTGTCATTCTAnATATAAGTTaCCTATGATGTTAAAGTATCTAACtGTTTTTTT
TTGCCCAaTACATTTAATTGAAGAATTAACCTAAAAAATAATTTtTGAATTTTTCcTATTAACGTATTGCAAAATTAATAAAATATA+??@,!AA;;;;;6,:8,--4224>>,-,111!000D-DD-AA,I,!==-ED,-,886D-,DD--,,,,,-+-
GGGGF@D,,////.D-46666--,,--*!666,-,,,,,,-II--D---,,+,+--,II*DD-AA,-,(D,DOO...-,,D,,::+-D---------,HHH-!---*355DDII,CC+54II;;FF,DDDDDD-22DDD!D->8EEEEE,!,-BB>>-I--44-
```

图表 9.art_illumina 单端模拟结果（2）---fastq

3).single_dat.sam

```
@HD VN:1.4  SO:unsorted
@SQ SN:NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome  LN:2362816
@PG ID:01   PN:ART_Illumina CL:art_illumina -ss HS25 -sam -i ./mydata_illumina/NC_010321.fasta -l 150 -f 10 -o ./mydata_illumina/single_dat -rs 1520985769
NC_010321.1-157520  0   NC_010321.1 1761455 99  150=     *   0   0   TTTGCCAGTCTTTCTTGTAATTTTTCTCTATCATAATCAGAGGTAGTTTCCTCAATTTGAGCCTTAATTTGATTTACTCTCTTCTTAATTT
NC_010321.1-157519  16  NC_010321.1 2124144 99  150=     *   0   0   TAAATTTTTAATTTAACAATTTTGTGATTTTATCTCACCCCCTTTAACATTCCCCTTTAATATTGTGTTTTTACTTTTACAAGTATATTTT
NC_010321.1-157518  16  NC_010321.1 1543843 99  150=     *   0   0   ATGCCAGTGCAATAAAAGTGGAAATTGGTCCTCAGGTATTTCATCCACTGTTATCCTTTCTTTATACAAAAAACTGTCATCCTGTGGTATA
NC_010321.1-157517  0   NC_010321.1 1806809 99  150=     *   0   0   GTTGTACCAAAAGAGGAGTACCTCTTATAATTTCAACATAGCTTGCGCCTATAAAGCTTATAGGTTTTATAGAAGACATTTTAATAAGAGC
NC_010321.1-157516  0   NC_010321.1 2290634 99  150=     *   0   0   AAGTTACTTAAAGACTTTTCAGGTAAAGCTCCGGTCTCATACCTTGCAATAGAAGCTTCGCCTATGCCCAAAAGTTTCGCAAAAGCCCTTT
NC_010321.1-157515  0   NC_010321.1 1136232 99  150=     *   0   0   AAAATGAAGGAGAGCAAGTTTTGACCTATAAAGTACGTGTATGGACGGAAGGTAGAATGAACAAAATCGTACGTTACGGTTATGCGGAGAT
NC_010321.1-157514  16  NC_010321.1 179543  99  150=     *   0   0   ATAAAAGTGGAAACGTTATAAGGATTGTGGAAGATAAAGATGCAACGGAAGAAGAAAAAAGCATCCATGAGATAAATTCTGCCATGTACGT
NC_010321.1-157513  16  NC_010321.1 346827  99  47=1X102=    *   0   0   GTTAAAAATTTATGTTATAATGTTAAGGGGAAAATCAAAGGAAGGTGCTTTATGGTTGACAAGGAAGAATTAGAAAGTACTTTTGGA
NC_010321.1-157512  16  NC_010321.1 1346635 99  150=     *   0   0   TGTCAAAGCAGTAAAACCAGGAGTATCTAAAACATACCCTCCAAAATCAAGAGATAAAAGTTCTACACTTCTTGTTGTGTGTCTACCTCTT
NC_010321.1-157511  16  NC_010321.1 1668368 99  150=     *   0   0   ATATCTGTTTACGTCAATTAGAATATACTTGAAATCTATTATATTTTCTCCAAATAATTGATATGAATCTATCGTTTCTTTAAAGCTCAAA
NC_010321.1-157510  16  NC_010321.1 2013463 99  111=1X38=   *   0   0   TCTCTTTTACTGTCCCTTCAAAAAGATGAGGCTGCTGAAAAACCATTCCAACTCTCCTTCTTAAGTCTATCACATCTATTGTTTTAA
NC_010321.1-157509  0   NC_010321.1 2246289 99  150=     *   0   0   TTTCTAAAATATCCTTTGCCGTTTTGGGCATAAATCAAGAAAGAAAGGTATCTCTTTAAAAGATCGAGGTATTTTTCATCTTTGTAAACTT
NC_010321.1-157508  0   NC_010321.1 2108424 99  150=     *   0   0   TTTTTGGGTCTTTTTTATCGCTTTTTGTGGGAGAATTGTCATCTAATTCTTTGCTCCTCTTTACATGGAATGGATTAACCAATACTACCTT
NC_010321.1-157507  0   NC_010321.1 1783302 99  150=     *   0   0   CAAAATTTCCTTCCTCCATAAAAGATTTAGGAAACTCAATAGCTTCCATATTCCCCAGCTTTGCAGCCCTTTCAATAGCAGGTCCTCCAGC
NC_010321.1-157506  16  NC_010321.1 1872353 99  150=     *   0   0   GCACAACCTGAGCTCGAATAGTTTTTCCCATTTACACTTCTCCCCTCTAACTAAAAATAAAAATTCCGTATTTATTTATCGGTTTAAATCC
NC_010321.1-157505  0   NC_010321.1 892733  99  150=     *   0   0   TAAAAGAGAATATGGATAGACCGGAAAAAGTAAATAAGCTAATAAACTTTGGTCTTACCGCTGCTTATTATTATGTGAGCTTTTTCAAAGA
NC_010321.1-157504  0   NC_010321.1 2111380 99  150=     *   0   0   TATCTTTCCATTGTCAGTGTAGAGCATTTTGGGTATTCCCCTTCTTAGCACTGCTTCTTTAAATGAGTCTCTTAAAGCTAAAAAATTCTGC
NC_010321.1-157503  0   NC_010321.1 1302434 99  150=     *   0   0   AAGTCATCGTCAGTAAATTCTTTTACCCCTCTTGATAAGGCTAAAGTTCCTAACCTGTCTCCACCGCCATTGACAGGTACTATAGTCAAAA
NC_010321.1-157502  16  NC_010321.1 924460  99  54=1X95=    *   0   0   GTATATATAATGGCAAAAGACATGTTAGATACTGTAAAGAAAGAAGCTAACCTTACAGATTATGAAATAGTTGCTGTATTTAAAGGA
NC_010321.1-157501  16  NC_010321.1 894969  99  150=     *   0   0   GTGCGGGCCTTCAAAAAAAGGATATTACATCAATAGTAGCAACAGGGTATGGAAGAGTAAGCATTCCCTTTGCAGATAGAATTATAACTGA
NC_010321.1-157500  0   NC_010321.1 579509  99  112=1X37=   *   0   0   GGAAGAAGTGGCAGCTTTAATATTAAAGGAAGCTAAACAGAGTTTGGAGAGGGAAATTGGCGAAGAAATAGTAAATGCGGTTATAAC
NC_010321.1-157499  16  NC_010321.1 1775156 99  35=1X114=   *   0   0   AATTAAGCTTATACCATATTGTATTTTTGCCATTTGAATATCCGTCAAATTTTGAGTTTTAAACAGTCTTTGTGTCAACTTCTCAGC
NC_010321.1-157498  16  NC_010321.1 1856706 99  150=     *   0   0   ATGCAAATCTAAAGTTTCGTAAATTGGTATGTTTTCGTATTGTGCTACTACATATCTTAGACGATCATTATCATACACTAACCTTCCAATT
NC_010321.1-157497  16  NC_010321.1 1718986 99  150=     *   0   0   AAACCTTTCAAAGGAATCTTAGGGATGTCACCTTTTTCTACCAATTTGTTAAGTGCATCTGCCAATGCTATAAATGAGCCTGTAATAGAAC
NC_010321.1-157496  16  NC_010321.1 248980  99  150=     *   0   0   CGCCAAAATAATTTACATTAGCGATGGTGCCACCAGGTTTAACAATCTTAACTGCTGTAGCCATAATGTCAGCATTTCCTCCAGCGATGAT
NC_010321.1-157495  16  NC_010321.1 621547  99  150=     *   0   0   CAAAAGAAGAACTTGACCCTAAAAAATACGGAGTAATAGTAAGAAGCGGCTATCGTTCAGGATTGCTTTTACCTGACTTAGAAGGAGTAGA
NC_010321.1-157494  0   NC_010321.1 1617749 99  56=1X93=    *   0   0   GGAAATATCGATTTAATACTATTGTAAATTTTTCTAATTTCTATATTCACTTTTTCATTATACTTTTCATAATATAATGGAAACAAA
NC_010321.1-157493  0   NC_010321.1 1267030 99  150=     *   0   0   TGGATTATTTCCTCCAATTTTTTTATTGCCTATTTTGACTTCGCGAGTAATTTTTCTCATTTTTATCACCTTTTAAAAAATGCGTAGAATA
NC_010321.1-157492  16  NC_010321.1 491037  99  36=1X4=1X108=   *   0   0   TCTATTGGGGCTCATAAACACGGCAGAGTAGTAGATTTAAAGGATGCTTTAAATGGGGAAGGGCCTTTTTCTCCAGAAAGAGC
NC_010321.1-157491  16  NC_010321.1 1794263 99  150=     *   0   0   GTAAAAGGCTTTTTAGATTTTCTTTAAAAACTATTATTGCCTTTTCTTCTGCTTTTTCCGTAAGAGTATTCCTTATTTCCCTCTCTATTGA
NC_010321.1-157490  16  NC_010321.1 2272536 99  150=     *   0   0   CTCATCATATCGTAAGAAATTTCCGGCTTTATCAACCTCAAAATAGTCACTTCTAAAAGGGTTAACGGAGATATAGCATATCTAACTTCTT
```

图表 10.art_illumina 单端模拟结果（3）---汇总分析

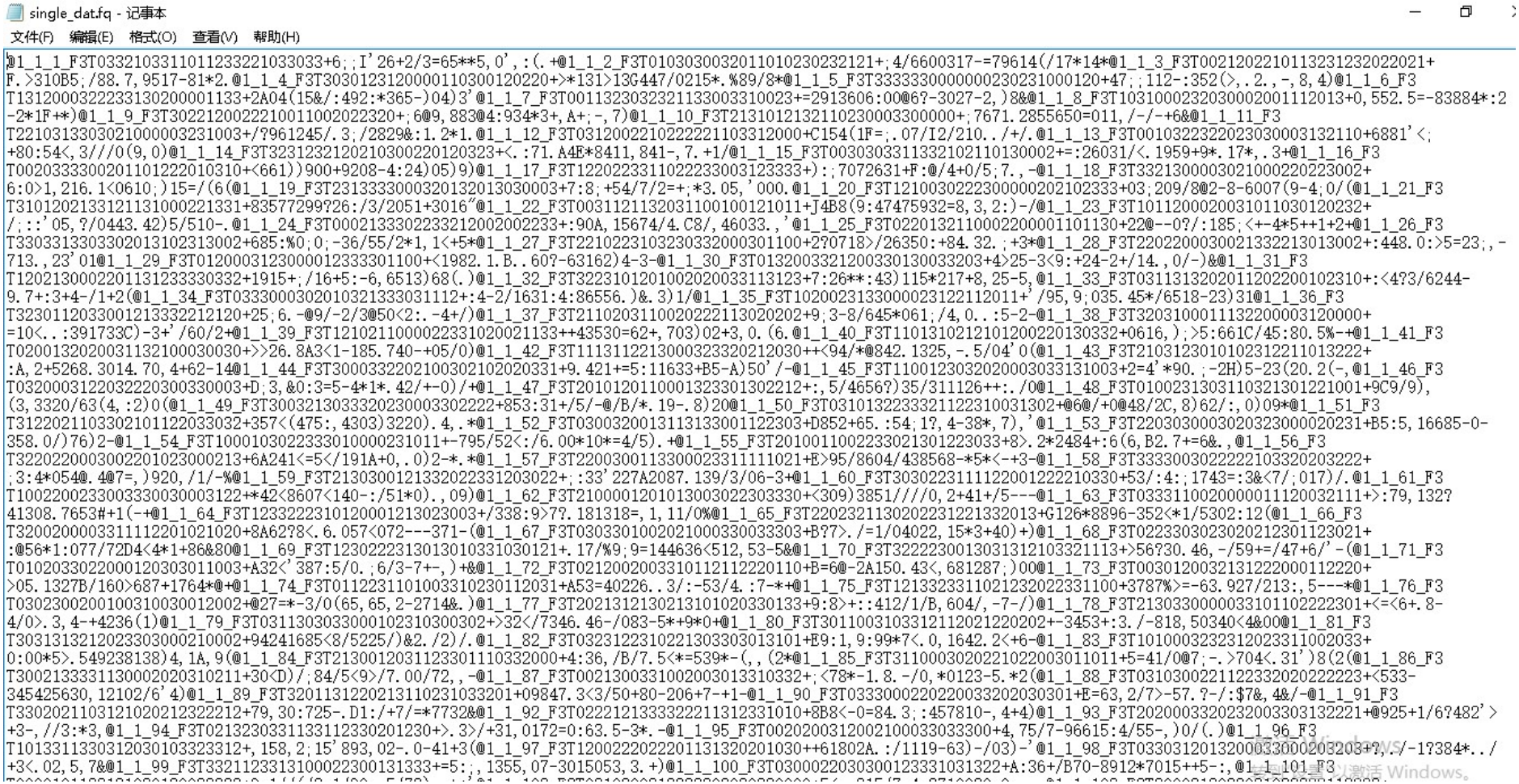
③art_solid:

1).single_dat.map

```
##ART_SOLiD read_length 25
@CM art_SOLiD -s ./mydata_solid/NC_010321.fasta ./mydata_solid/single_dat 25 10
@SQ NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome 2362816
##Header End
NC_010321.1 1_1_1_F3    1859254 -   2   13  02  17  32
NC_010321.1 1_1_2_F3    366745  -   5   4   20  8   13  17  13  21  32  24  31
NC_010321.1 1_1_3_F3    1883699 +   0
NC_010321.1 1_1_4_F3    632683  -   0
NC_010321.1 1_1_5_F3    1437418 +   2   9   30  18  31
NC_010321.1 1_1_6_F3    189862  +   3   7   03  14  03  24  13
NC_010321.1 1_1_7_F3    435970  +   1   3   31
NC_010321.1 1_1_8_F3    1846491 -   1   8   03
NC_010321.1 1_1_9_F3    961715  +   0
NC_010321.1 1_1_10_F3   1840506 +   0
NC_010321.1 1_1_11_F3   2003472 +   0
NC_010321.1 1_1_12_F3   352323  +   0
NC_010321.1 1_1_13_F3   806753  +   3   4   03  7   03  20  03
NC_010321.1 1_1_14_F3   551632  -   1   22  23
NC_010321.1 1_1_15_F3   495604  -   0
NC_010321.1 1_1_16_F3   1679487 -   3   4   03  5   13  23  01
NC_010321.1 1_1_17_F3   38794   -   0
NC_010321.1 1_1_18_F3   1620718 -   0
NC_010321.1 1_1_19_F3   970945  -   2   4   03  20  03
NC_010321.1 1_1_20_F3   2143868 +   1   4   30
NC_010321.1 1_1_21_F3   269254  -   0
NC_010321.1 1_1_22_F3   1852332 +   0
NC_010321.1 1_1_23_F3   1183955 +   1   14  01
NC_010321.1 1_1_24_F3   86602   +   0
NC_010321.1 1_1_25_F3   1648408 -   1   24  10
NC_010321.1 1_1_26_F3   521130  +   4   4   13  7   03  18  32  24  02
NC_010321.1 1_1_27_F3   1712441 +   1   17  10
NC_010321.1 1_1_28_F3   1449892 -   0
NC_010321.1 1_1_29_F3   377850  -   0
NC_010321.1 1_1_30_F3   1683020 +   0
NC_010321.1 1_1_31_F3   287642  +   0
NC_010321.1 1_1_32_F3   1008586 +   0
```

图表 11.art_solid 单端模拟结果（1）---map

2).single_dat.fq



图表 12.art_solid 单端模拟结果（2）---fastq

3).single_dat.sam

```
@HD VN:1.4  SO:unsorted
@SQ SN:NC_010321.1 Thermoanaerobacter pseudethanolicus ATCC 33223, complete genome  LN:2362816
@PG ID:03   PN:ART_SOLiD    CL:art_SOLiD -s ./mydata_solid/NC_010321.fasta ./mydata_solid/single_dat 25 10
1_1_1_F3    16  NC_010321.1 503538  99  12X13=  *   0   0   ataataactcgctgttgtattgata   +.(:,'0,5**56=3/2+62'I;;6
1_1_2_F3    16  NC_010321.1 1996047 99  8X9=4X4=    *   0   0   cagtcgaatccaacaagcccggcca   *41*71/(41697=-7130066/4;
1_1_3_F3    0   NC_010321.1 1883700 99  25= *   0   0   ttcaggagttgtagcagcttcttca   F.>310B5;/88.7,9517-81*2.
1_1_4_F3    16  NC_010321.1 1730109 99  25= *   0   0   ttcttcaaattgtttttcatcaatt   *8/98%.*5120/744G31>131*>
1_1_5_F3    0   NC_010321.1 1437419 99  9=16X   *   0   0   atatatttttttcggatggggtcc   47;;112-:352(>,.2.,-,8,4)
1_1_6_F3    0   NC_010321.1 189863  99  7=7X10=1X   *   0   0   gcaggggctctatgccttttttgtat   2A04(15&/:492:*365-)04)3'
1_1_7_F3    0   NC_010321.1 435971  99  3=22X   *   0   0   ttgtagccgatcacgcccgcaaagc   =2913606:00@6?-3027-2,)8&
1_1_8_F3    16  NC_010321.1 516301  99  17X8=   *   0   0   atggacacccttttaagcttttgcc   )*+F1*2-2:*48838-=5.255,0
1_1_9_F3    0   NC_010321.1 961716  99  25= *   0   0   aagactttctcaaacaaaggagctt   ;6@9,883@4:934*3+,A+;-,7)
1_1_10_F3   0   NC_010321.1 1840507 99  25= *   0   0   catggtcatcacctaaaataaaaaa   ;7671.2855650=011,/-/-+6&
1_1_11_F3   0   NC_010321.1 2003473 99  25= *   0   0   ctggcataattcaaaaaatcgttta   /?961245/.3;/2829&:1.2*1.
1_1_12_F3   0   NC_010321.1 352324  99  25= *   0   0   tactttctggagagacaatactttt   C154(1F=;.07/I2/210../+/.
1_1_13_F3   0   NC_010321.1 806754  99  4=3X13=5X   *   0   0   ttggctcgaggattaaaatgctgtt   6881'<;+80:54<,3///0(9,0)
1_1_14_F3   16  NC_010321.1 1811160 99  3X22=   *   0   0   gctaagttctttaacttcagctgct   /1+.7,-148,1148*E4A.17:.<
1_1_15_F3   16  NC_010321.1 1867188 99  25= *   0   0   cttttaccacttgatacataattaa   +3.,*71.*9+9591.</13062:=
1_1_16_F3   16  NC_010321.1 683305  99  2=19X4= *   0   0   aacggttctcaacaagggcgcggaa   )9)50)42:4-8029+009))166<
1_1_17_F3   16  NC_010321.1 2323998 99  25= *   0   0   tatatcatttatctccacgctcctc   -,.7;5/0+4/@:F+1362707;:)
1_1_18_F3   16  NC_010321.1 742074  99  25= *   0   0   ctttagaagaaaacttaaaaatgat   (6(/=51);0160<1.612,1>0:6
1_1_19_F3   16  NC_010321.1 1391847 99  5=16X4= *   0   0   taaaattacctacctaaaatatacg   .000',50.3*;+=2/7/45+;8:7
1_1_20_F3   0   NC_010321.1 2143869 99  4=21X   *   0   0   gacccggagatttttcctggatat   03;209/8@2-8-6007(9-4;0/(
1_1_21_F3   16  NC_010321.1 2093538 99  25= *   0   0   tgcgtcttttgcactgcgtcctggt   "6103+1502/3/:62?99277538
1_1_22_F3   0   NC_010321.1 1852333 99  25= *   0   0   ttacagtgcttacaaacccagttgt   J4B8(9:47475932=8,3,2:)-/
1_1_23_F3   0   NC_010321.1 1183956 99  14=11X  *   0   0   ggtgaaaagggcaacaattgaagct   /;::'05,?/0443.42)5/510-.
1_1_24_F3   0   NC_010321.1 86603   99  25= *   0   0   tttcatattctatcagggaaagata   :90A,15674/4.C8/,46033.,'
1_1_25_F3   16  NC_010321.1 714384  99  1X24=   *   0   0   ttacaacaaaaagaaaacagcaaga   +2+1++5*4-+<;581:/?0--@22
1_1_26_F3   0   NC_010321.1 521131  99  4=14X6=1X   *   0   0   attatgcggcggaacgttcgtaaag   685:%0;0;-36/55/2*1,1<+5*
1_1_27_F3   0   NC_010321.1 1712442 99  17=8X   *   0   0   ctggagcaatcggcgaaaattgttt   2?0718>/26350:+84.32.;+3*
1_1_28_F3   16  NC_010321.1 912900  99  25= *   0   0   gaaatggcagatactttaaaagaag   10'32,.317-,;32=5>:0.844:
1_1_29_F3   16  NC_010321.1 1984942 99  25= *   0   0   tttgttatatcaaaaatcattttca   -3-4)26136-?06..B.1.2891<
1_1_30_F3   0   NC_010321.1 1683021 99  25= *   0   0   tgctttatcagggcggtaaataggc   4>25-3<9:+24-2+/14.,0/-)&
```

图表 13.art_solid 单端模拟结果（3）---汇总分析

5.2 根据下载的不同测序平台下载测序结果数据，创建模拟 基因组测序的数据模型，模拟运算的输出结果文件

5.2.1------454 结果生成三个文件：

①length_dist

```
##454 read profile for ART 454 simulator
##the empirical distribution of 454 read length
##FORMAT
##len_1        len_2  len_3  ...
##freq_1       freq_2 freq_3 ...
```

48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68
69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131
132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152
153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173
174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194
195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215
216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236
237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257
258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278
279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299
300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320
321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341
342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362
363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383
384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404

图表 14.454 模拟结果（1）---length_dist

② qual_1st_profile

```
##454 read profile for ART 454 simulator
##homopolymer-length dependent quality score distribution of the 1st-base of homopolymer runs
##FORMAT
##1enth_of_homopolymer  quality_score_1 quality_score_2 ...
##1enth_of_homopolymer  freq_of_score_1 freq_of_score_2 ...
```

1	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40						
1	3103349	54461	405708	417756	212734	42891346		14985573		14487403		3956569	14270331		8102397	10948054		19810567		
14773226		6419945	8131935	13651044		10518681		11265280		12822120		5724856	13568646		10432566		6508579	14196583		5972030
14884357		12367228		3749362	16956996		4753661	62236723		6521913	14508261		79534645							
2	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40						
2	94658	19384	186625	138436	76408	11715423		4396766	3805159	903658	4389976	3344042	4074539	1520494	4246001	2078218	3123536	1453546	3659651	2050161
2412722	2237751	3136581	2242630	2375921	1791861	1483901	3358544	2804429	1000813	4103192	1043587	10809314		1658449	3014877	16861111				
3	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40						

图表 15.454 模拟结果（2）---qual_1st_profile

③ qual_mc_profile

```
##454 read profile for ART 454 simulator
##the empirical transition distribution of the homopolymer-length dependent 1st order Markov model of quality scores
##FORMAT
##homopolymer_length    previous_base_position  previous_base_quality_score     next_base_quality_score_1       next_base_quality_score_2       ...
##homopolymer_length    previous_base_position  previous_base_quality_score     freq_next_base_qual_score_1     freq_next_base_qual_score_2     ...
2       1       0       0
2       1       0       94658
2       1       7       7       8       9       10      13
2       1       7       18431   562     60      319     12
2       1       8       7       8       9       10      11      12      13      14      15      16      17
2       1       8       338     170182  11754   871     675     2710    33      50      7       4       1
2       1       9       7       8       9       10      11      12      13      14      15      16      17      18      19
2       1       9       28      8396    109494  4231    3906    10830   120     1200    201     9       13      4       4
2       1       10      7       8       9       10      11      12      13      14      15      16      17      18      19      20      21
2       1       10      121     553     3129    62328   4222    1632    1276    2274    215     359     219     65      2       4       9
2       1       11      8       9       10      11      12      13      14      15      16      17      18      19      20      21      22      23      24      25
26      28      30
2       1       11      415     1538    2447    10911760        120247  302344  18928   250369  50204   8883    23199   17163   4951    1907    743     204     57
5       53      5       1
2       1       12      8       9       10      11      12      13      14      15      16      17      18      19      20      21      22      23      24      25
26      27      28
```

图表 16.454 模拟结果（3）---“qual_mc_profile”

5.2.2------illumina 双端测序生成两个文件 R1 和 R2：

```
.   0   3   19  26  28  29  31  32  33  34  35
.   0   7312    134958  198804  241975  272527  327305  392469  753332  2050292 4524133
.   1   3   19  26  28  29  31  32  33  34  35
.   1   743 107373  157942  205729  239554  288392  363188  676957  1820004 4524133
.   2   3   19  26  28  29  31  32  33  34  35
.   2   817 101872  147142  188517  223589  272829  355437  659164  1766018 4524133
.   3   3   16  26  28  29  30  31  32  33  34  35
.   3   893 101169  143580  177711  213953  218989  267217  363421  639973  1737218 4524133
.   4   3   16  26  28  29  31  32  33  34  35
.   4   992 104009  146027  177898  214454  265185  353766  642129  1738096 4524133
.   5   3   15  16  17  25  27  28  29  30  31  32  33  34  35  36  37  38  39
.   5   1196    16719   23835   88200   88206   133320  138675  167835  214463  246944  300764  303594  362451  480892  565004  779099  1308666 4524133
.   6   3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   6   1338    15132   21541   83372   83391   83392   118755  124390  142086  184239  196221  238777  241679  296749  393301  473707  612807  928713  4524133
.   7   3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   7   1447    13473   19599   79583   79657   79665   109815  115345  129171  169467  178172  216922  219744  268800  352964  424349  545284  778656  4524133
.   8   3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   8   1540    13377   19423   78979   79104   79126   107461  113039  126447  166741  174375  211804  215471  260692  342835  410722  530976  732445  4524133
.   9   3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   9   1632    11635   17466   77876   78056   78121   105160  110286  122588  162586  169968  207365  209981  254903  332450  396933  512855  709125  4524133
.   10  3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   10  1728    12013   17916   77874   78169   78304   105098  110313  122458  162841  169494  205978  208713  252844  330511  394365  509592  704072  4524133
.   11  3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   11  1817    12197   18126   78652   79060   79329   105370  110561  122891  164307  169378  205914  208611  253110  329764  393039  508395  701500  4524133
.   12  3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
.   12  1894    13666   20368   86223   86751   87239   114091  119856  133097  174964  180689  217931  220673  264510  342974  406672  522297  717170  4524133
.   13  3   15  16  17  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
```

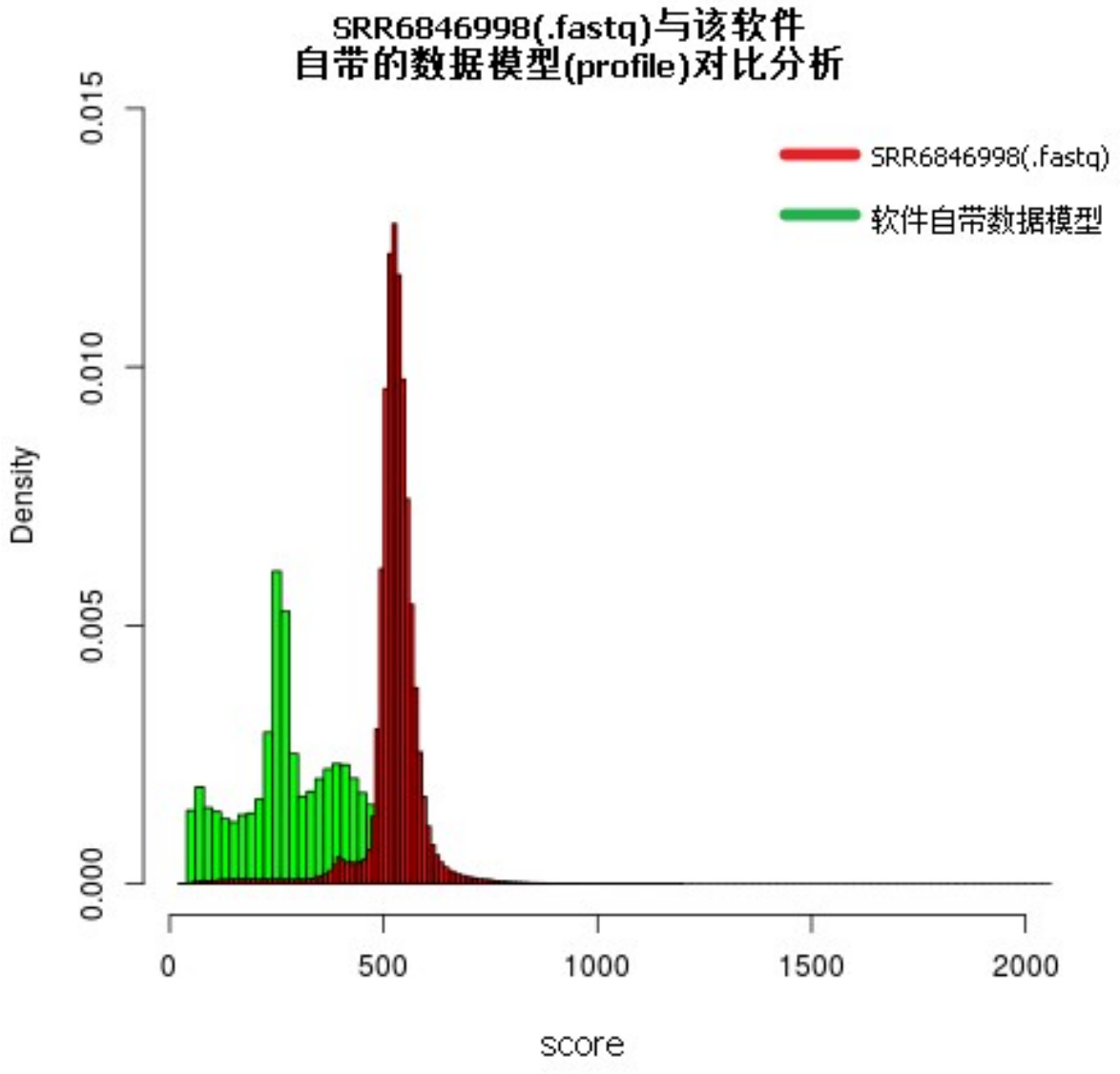
图表 17.illumina 模拟生成文件 outputR1

.	0	3	19	26	28	29	31	32	33	34	35										
.	0	8919		229730		309142		407755		450252		506400	612590	1083807	3016106	4524133					
.	1	3	19	26	28	29	31	32	33	34	35										
.	1	8944		204378		267698		350783		399660		448308	563756	955946	2695520	4524133					
.	2	3	19	26	28	29	31	32	33	34	35										
.	2	9083		215647		282717		371198		418696		468375	587384	956171	2589843	4524133					
.	3	3	16	26	28	29	31	32	33	34	35										
.	3	9266		207241		266795		342358		390195		438366	560734	918721	2548499	4524133					
.	4	3	16	26	28	29	31	32	33	34	35										
.	4	9399		206510		265309		332323		378862		426968	549741	892925	2533153	4524133					
.	5	3	15	16	17	25	27	28	29	30	31	32	33	34	35	36	37	38	39		
.	5	9555		42425		59373		207183		207188		264006	272088	299529	351780	373040	425828	429021	487909	579193	653706	781742	1001892	4524133	
.	6	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	6	10236		41079		57287		208109		208148		208149	265256	271934	299333	351165	372923	425732	428725	484368	576226	648342	777070	1002998	4524133
.	7	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	7	10948		38599		54732		202213		202387		202391	258021	264052	291393	341859	363202	416085	418952	471819	563526	633799	763659	990586	4524133
.	8	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	8	11629		41236		56712		205818		206087		206141	260986	267628	293834	345713	365904	418903	423273	476708	570409	643430	775607	993498	4524133
.	9	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	9	12299		39615		55572		210777		211252		211369	266680	272851	299204	351295	370941	424644	427671	481090	572979	644883	775486	995943	4524133
.	10	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	10	13035		40781		56603		208660		209298		209611	265102	271655	298177	350341	370160	424108	427123	480731	572580	643883	774220	995255	4524133
.	11	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	11	13671		41819		57768		211796		212701		213336	267880	274678	301525	355042	374015	427768	430738	485010	576758	648244	779856	1001540	4524133
.	12	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	12	14351		45088		61290		211537		212730		213769	267302	275283	303558	358034	374712	426587	430307	483315	577606	649872	782291	996423	4524133
.	13	3	15	16	17	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	
.	13	14944		52061		68462		216993		218546		220311	277393	286081	326346	382125	396167	447606	450827	505488	608179	678762	843197	1059545	4524133

图表 18.illumina 模拟生成文件 outputR2

5.3.模拟结果与该软件自带的数据模型（profile）进行对比分析

①art_profile_454



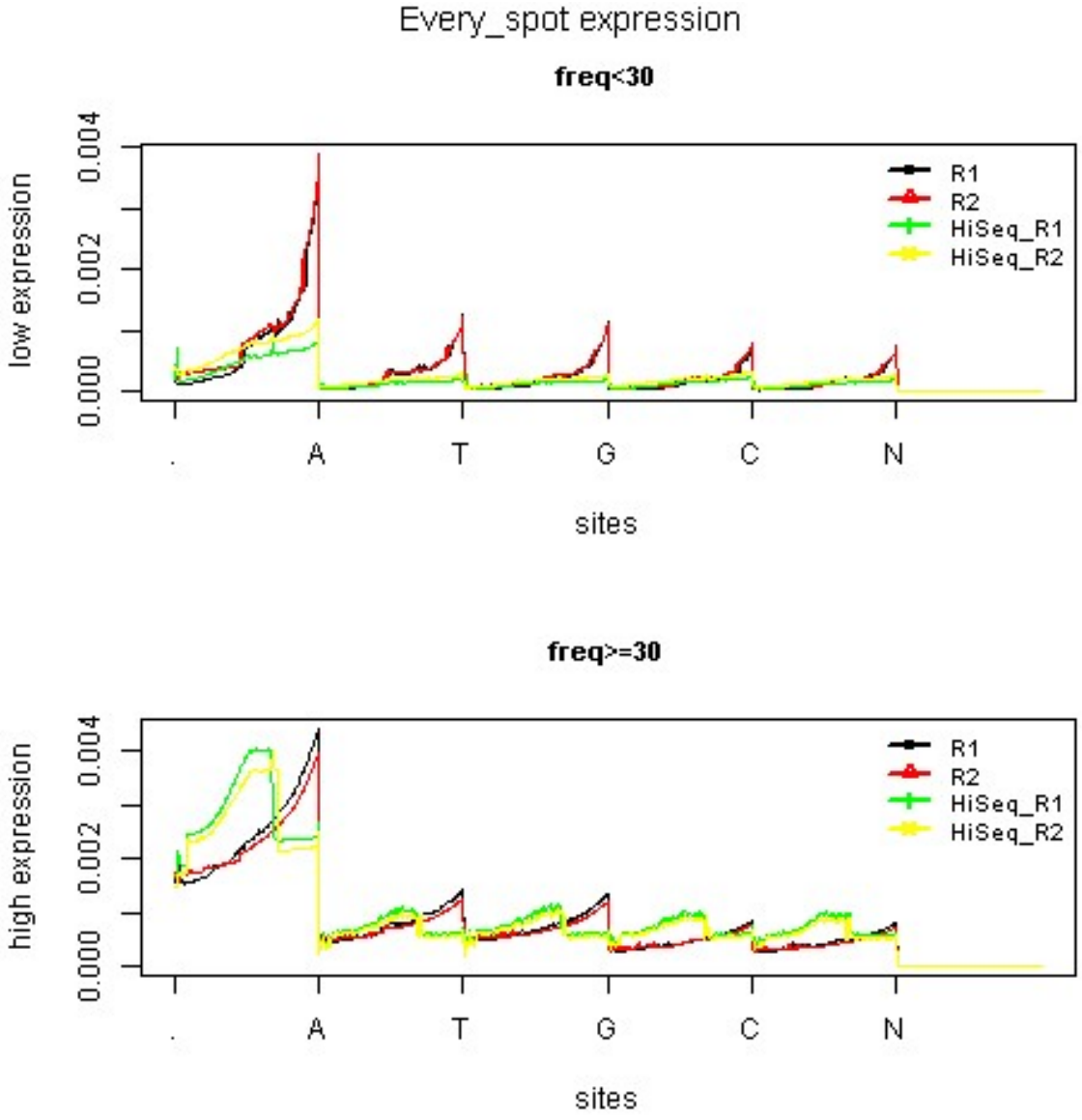
图表 19.454 平台的对比验证结果

两组数据进行卡方检验结果：

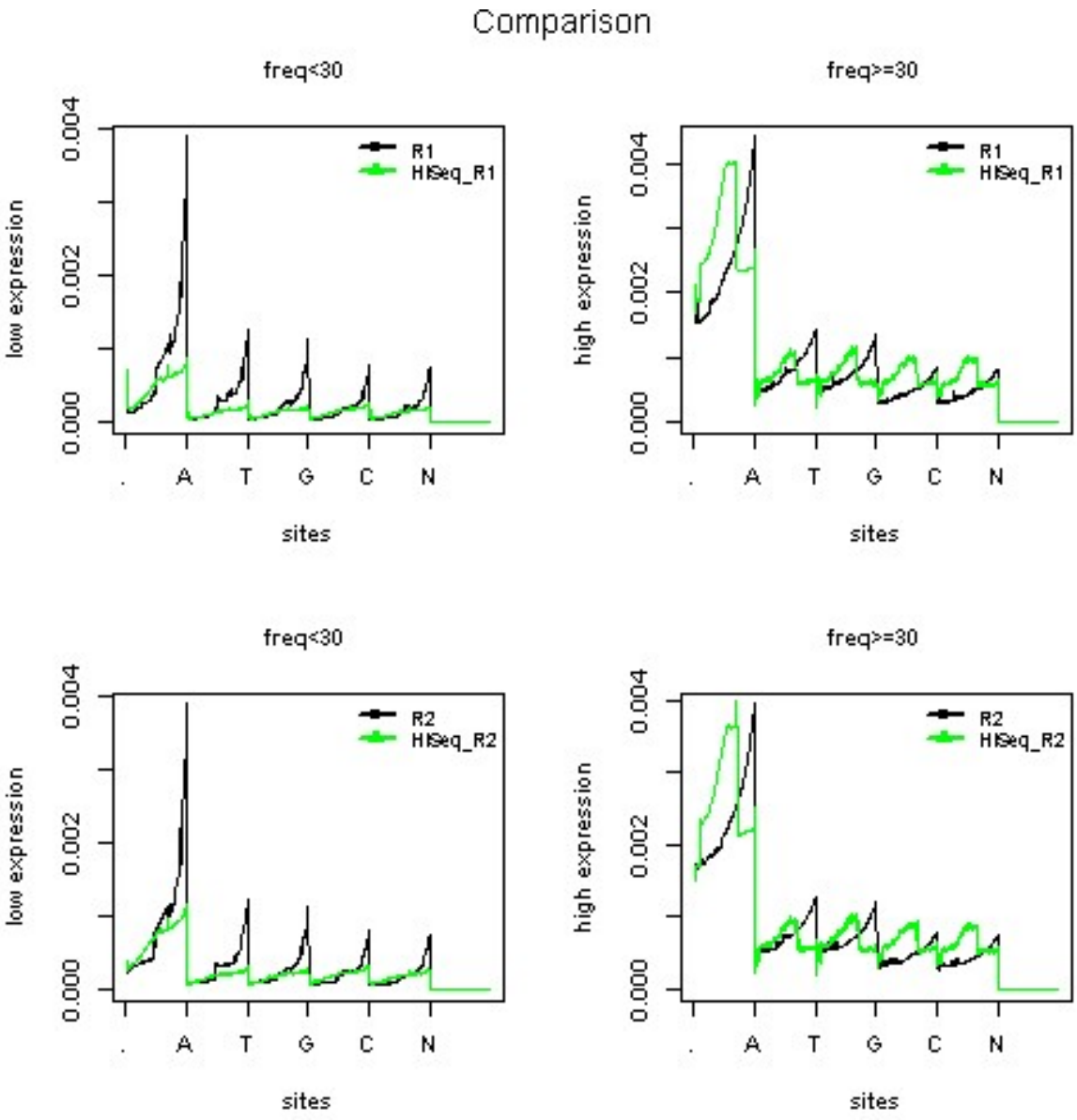
X-squared = 3586600000, df = 90327000, p-value < 2.2e-16

卡方检验原假设二者相互独立，检验结果 $p < 0.05$ 拒绝原假设，说明模拟运算结果和软件自带数据模型是相关的。

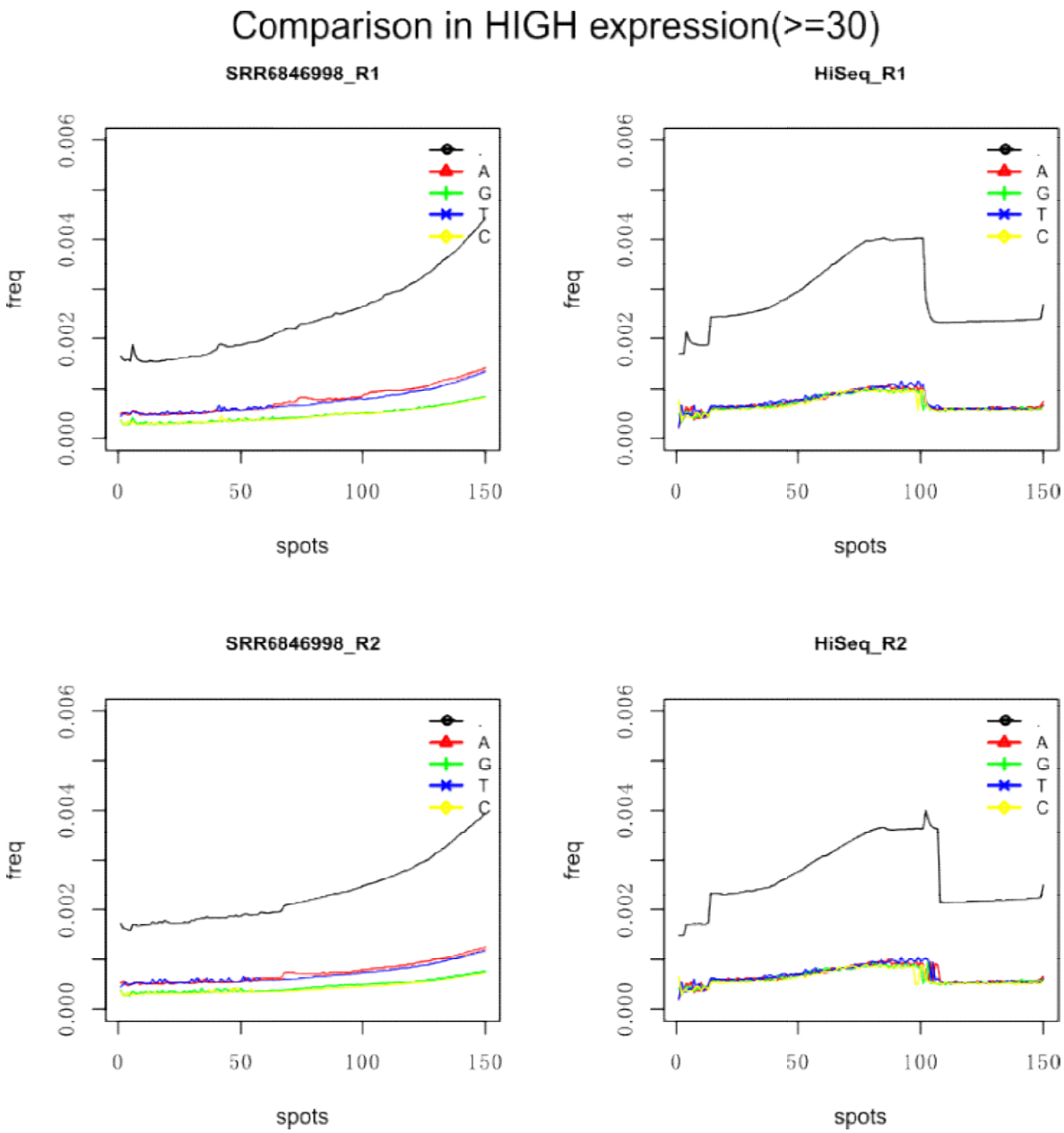
②art_profile_illumina



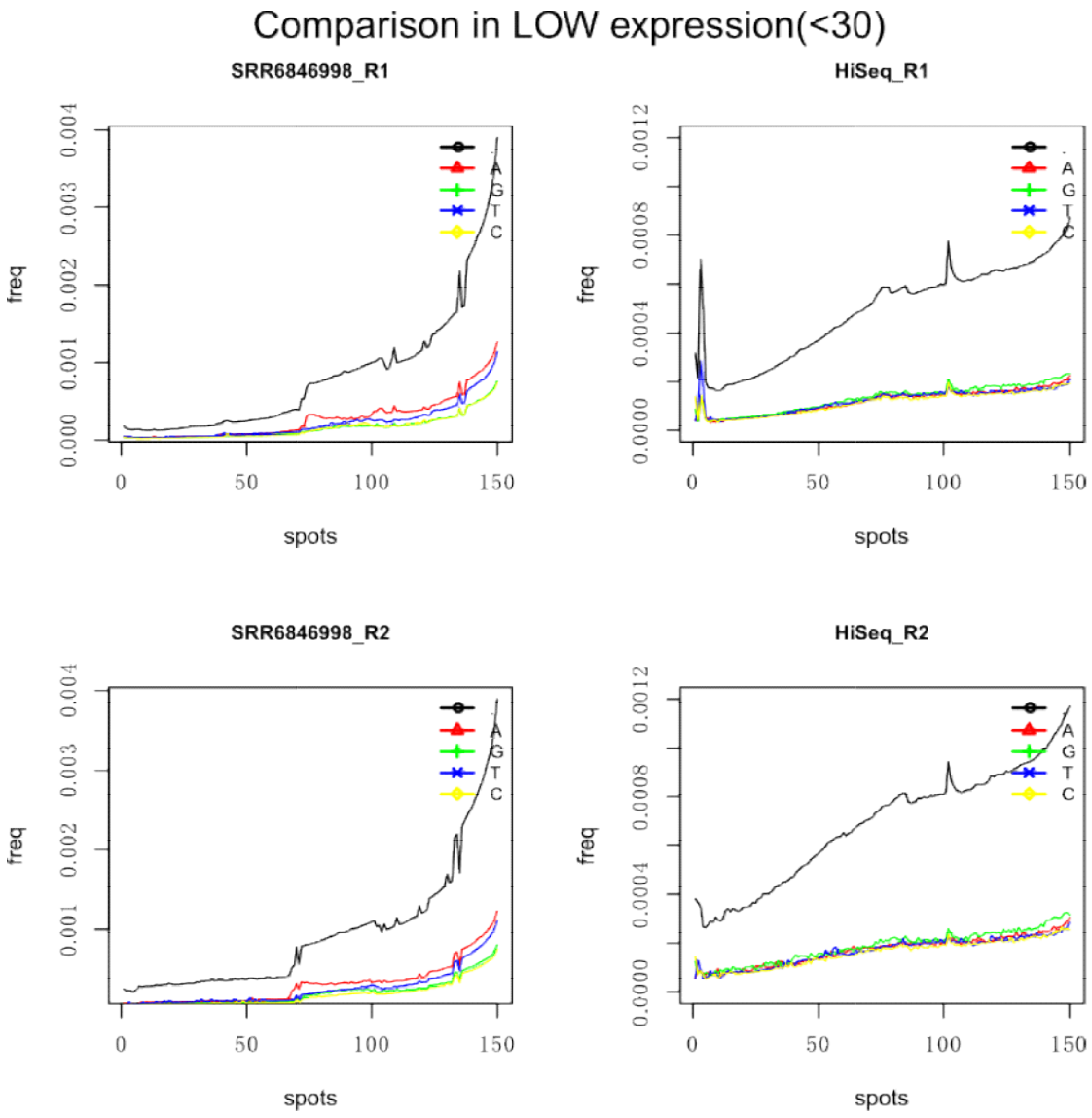
图表 20.illumina 模拟数据和自带数据对比（1）



图表 21.illumina 模拟数据和自带数据对比（2）---低表达和高表达两端结果对比



图表 22.illumina 模拟数据和自带数据对比（3）---高表达中可视化



图表 23. illumina 模拟数据和自带数据对比（4）---低表达中可视化

检验及结果：

```
shapiro.test(R1[,3]-HiSeq_R1[,3])          #p-value < 2.2e-16
shapiro.test(R2[,3]-HiSeq_R2[,3])          #p-value < 2.2e-16

shapiro.test(R1[,4]-HiSeq_R1[,4])          #p-value < 2.2e-16
shapiro.test(R2[,4]-HiSeq_R2[,4])          #p-value < 2.2e-16
```

==》 检验四组数据都不服从正态分布，所以采用秩和检验：

```
wilcox.test(R1[,3],HiSeq_R1[,3])
  #W = 434840, p-value = 0.006789
  wilcox.test(R1[which(R1$base!="A"),3],HiSeq_R1[which(HiSeq_R1$base!="
A"),3])
  #W = 296730, p-value = 0.06473
  wilcox.test(R1[which(R1$base!="T"),3],HiSeq_R1[which(HiSeq_R1$base!="
T"),3]
  #W = 299610, p-value = 0.02854
  wilcox.test(R1[which(R1$base!="G"),3],HiSeq_R1[which(HiSeq_R1$base!=
"G"),3])
  #W = 308910, p-value = 0.0009673
  wilcox.test(R1[which(R1$base!="G"),3],HiSeq_R1[which(HiSeq_R1$base!=
"G"),3])
  #W = 305700, p-value = 0.003542


wilcox.test(R1[,4],HiSeq_R1[,4])
  #W = 339170, p-value = 2.171e-09

wilcox.test(R2[,3],HiSeq_R2[,3])
  #W = 406450, p-value = 0.8954

wilcox.test(R2[,4],HiSeq_R2[,4])
  #W = 351640, p-value = 1.232e-06
```

四、讨论分析：

1.对于 3.2 在 illumina 平台上出现的 error 提示,我认为可能是软件自己存在某些问题

2.对于 454 软件的自带数据和模拟数据检验，结果拒绝卡方检验的原假设说明模拟运算结果和软件自带数据模型是相关的。验证成功

3.对于 illumina 软件自带数据和模拟数据不符的结果我认为原因有可能是 illumina 软件自带的数据和模拟数据预处理有问题
4.下载的用于 art 平台模拟的数据要足够大才能保证检验的正确性
5.在 SRA 数据库下载数据时要注意是单端测序还是双端测序，需使用不同的参数

【附录：5.3 illumina 平台对比，R 语言预处理数据代码：】

```
#Instrument: Illumina HiSeq 2500
#Layout: PAIRED
setwd("E:/2018.3-2018.7----大三下/基因组信息学/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")
#使用 read.table 时有多行数据不能成功读取
#strand_1 <- read.xlsx2("outputR1.xlsx",sheetIndex=1,head=F,colIndex=29)
##R1
strand_1 <- read.table("outputR1.txt",sep="\t",header=F,fill=T,col.names=1:29)

sum=0
n<-nrow(strand_1)
strand_1 <- strand_1[1:n,]
strand_1[is.na(strand_1)] <- 0
new_1 <- matrix(nrow=nrow(strand_1)/2,ncol=4)
for(i in 1:nrow(strand_1)/2){

    new_1[i,1] <- as.character(strand_1[2*i-1,1])
    new_1[i,2] <- as.numeric(as.character(strand_1[2*i-1,2]))

}
#该组数据第 1510 行对应最后 150 的数据
k <- n/2      #数据量
s1 <- c()     #序列值<30 的对应的数值和的 集合
s2 <- c()     #序列值>=30 的对应的数值和的 集合
for(i in 1:k){      #行
    r=i*2
    sum1<-0
    sum2<-0

    for(j in 3:ncol(strand_1)){
        x <-  as.numeric(as.character(strand_1[r,j]))      #eg.7312
        if(as.numeric(as.character(strand_1[r-1,j])) < 30)
            sum1 <- sum1+x
        if(as.numeric(as.character(strand_1[r-1,j]))>=30)
            sum2 <- sum2+x
    }
```

```
    s1 <- c(s1,sum1)
    s2 <- c(s2,sum2)

}
sum <- sum(s1)+sum(s2)
new_1[,3] <- s1/sum
new_1[,4] <- s2/sum
write.table(new_1,"processed_illumina_R1.txt",sep="\t",row.names=F,col.names=c("base","site
","<30",">30"))



##R2
setwd("E:/2018.3-2018.7----  大  三  下  /  基  因  组  信  息  学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")
strand_2 <- read.table("outputR2.txt",sep="\t",header=F,fill=T,col.names=1:29)

n<-nrow(strand_1)

#把原文件的 1510 行后的清空
strand_2 <- strand_2[1:n,]
strand_2[is.na(strand_2)] <- 0
new_1 <- matrix(nrow=nrow(strand_2)/2,ncol=4)
for(i in 1:nrow(strand_2)/2){

    new_1[i,1] <- as.character(strand_2[2*i-1,1])
    new_1[i,2] <- as.numeric(as.character(strand_2[2*i-1,2]))

}
#该组数据第 1510 行对应最后 150 的数据      碱基列为“N”的直接删除
k <- n/2    #数据量
s1 <- c()    #序列值<30 的对应的数值和的 集合
s2 <- c()    #序列值>=30 的对应的数值和的 集合
for(i in 1:k){    #行
    r=i*2
    sum1<-0
    sum2<-0
    for(j in 3:ncol(strand_2)){
        x <-  as.numeric(as.character(strand_2[r,j]))    #eg.7312
        if(as.numeric(as.character(strand_2[r-1,j])) < 30)
            sum1 <- sum1+x
        if(as.numeric(as.character(strand_2[r-1,j]))>=30)
            sum2 <- sum2+x
    }
```

```
    s1 <- c(s1,sum1)
    s2 <- c(s2,sum2)
}
sum <- sum(s1)+sum(s2)
new_1[,3] <- s1/sum
new_1[,4] <- s2/sum
write.table(new_1,"processed_illumina_R2.txt",sep="\t",row.names=F,col.names=c("base","site
","<30",">30"))
#sum：5647438329
# 5647438329


##数据平台 HiSeq_R1
setwd("E:/2018.3-2018.7----  大  三  下  /  基  因  组  信  息  学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")

strand_2 <- read.table("HiSeq2500L150R1.txt",sep="\t",header=F,fill=T,col.names=1:29)
n <- nrow(strand_2)

#把原文件的 1500 行后的清空
strand_2 <- strand_2[1:n,]
strand_2[is.na(strand_2)] <- 0
new_1 <- matrix(nrow=nrow(strand_2)/2,ncol=4)
for(i in 1:nrow(strand_2)/2){

    new_1[i,1] <- as.character(strand_2[2*i-1,1])
    new_1[i,2] <- as.numeric(as.character(strand_2[2*i-1,2]))

}
#该组数据第 1500 行对应最后 150 的数据      碱基列为“N”的直接删除
k <- n/2     #数据量
s1 <- c()    #序列值<30 的对应的数值和的 集合
s2 <- c()    #序列值>=30 的对应的数值和的 集合
for(i in 1:k){     #行
    r=i*2
    sum1<-0
    sum2<-0
    for(j in 3:ncol(strand_2)){
        x <-  as.numeric(as.character(strand_2[r,j]))     #eg.7312
        if(as.numeric(as.character(strand_2[r-1,j])) < 30)
            sum1 <- sum1+x
        if(as.numeric(as.character(strand_2[r-1,j]))>=30)
            sum2 <- sum2+x
    }
```

```
    s1 <- c(s1,sum1)
    s2 <- c(s2,sum2)
}
sum <- sum(s1)+sum(s2)
new_1[,3] <- s1/sum
new_1[,4] <- s2/sum
write.table(new_1,"processed_HiSeq_R1.txt",sep="\t",row.names=F,col.names=c("base","site","
<30",">30"))


##数据平台  HiSeq_R2
setwd("E:/2018.3-2018.7----  大  三  下  /  基  因  组  信  息  学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")

strand_2 <- read.table("HiSeq2500L150R2.txt",sep="\t",header=F,fill=T,col.names=1:29)
#把原文件的 1500 行后的清空
n<-nrow(strand_2)
strand_2 <- strand_2[1:n,]
strand_2[is.na(strand_2)] <- 0
new_1 <- matrix(nrow=nrow(strand_2)/2,ncol=4)
for(i in 1:nrow(strand_2)/2){

    new_1[i,1] <- as.character(strand_2[2*i-1,1])
    new_1[i,2] <- as.numeric(as.character(strand_2[2*i-1,2]))

}
#该组数据第 1500 行对应最后 150 的数据      碱基列为“N”的直接删除
k <- n/2     #数据量
s1 <- c()    #序列值<30 的对应的数值和的 集合
s2 <- c()    #序列值>=30 的对应的数值和的 集合
for(i in 1:k){     #行
    r=i*2
    sum1<-0
    sum2<-0
    for(j in 3:ncol(strand_2)){
        x <-  as.numeric(as.character(strand_2[r,j]))     #eg.7312
        if(as.numeric(as.character(strand_2[r-1,j])) < 30)
            sum1 <- sum1+x
        if(as.numeric(as.character(strand_2[r-1,j]))>=30)
            sum2 <- sum2+x
    }
    s1 <- c(s1,sum1)
    s2 <- c(s2,sum2)
}
```

```
sum <- sum(s1)+sum(s2)
new_1[,3] <- s1/sum
new_1[,4] <- s2/sum
write.table(new_1,"processed_HiSeq_R2.txt",sep="\t",row.names=F,col.names=c("base","site","
<30",">30"))
```

8237009154
8237009154

```
setwd("E:/2018.3-2018.7----大三下/基因组信息学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")
R1 <- read.table("processed_illumina_R1.txt",head=T,sep="\t")
R1<-R1[which(R1$site!=150),]
#R1 <- R1[which(R1$site!=150),]
R2 <- read.table("processed_illumina_R2.txt",head=T,sep="\t")
R2<-R2[which(R2$site!=150),]
#R2 <- R2[which(R2$site!=150),]
#数据平台
setwd("E:/2018.3-2018.7----大三下/基因组信息学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")

HiSeq_R1 <- read.table("processed_HiSeq_R1.txt",head=T,sep="\t")
HiSeq_R2 <- read.table("processed_HiSeq_R2.txt",head=T,sep="\t")

pdf("illumina_freq_LOW30_comp.pdf")
par(mfrow=c(2,2))
plot(R1[which(R1$base=="."),3],main="SRR6846998_R1",type="l",ylab="freq",xlab="spots",cex.
main=0.9)
lines(R1[which(R1$base=="A"),3],col="red")
lines(R1[which(R1$base=="G"),3],col="green")
lines(R1[which(R1$base=="T"),3],col="blue")
lines(R1[which(R1$base=="C"),3],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

plot(HiSeq_R1[which(HiSeq_R1$base=="."),3],main="HiSeq_R1",ylim=c(0,0.0012),type="l",ylab
="freq",xlab="spots",cex.main=0.9)
lines(HiSeq_R1[which(HiSeq_R1$base=="A"),3],col="red")
lines(HiSeq_R1[which(HiSeq_R1$base=="G"),3],col="green")
lines(HiSeq_R1[which(HiSeq_R1$base=="T"),3],col="blue")
lines(HiSeq_R1[which(HiSeq_R1$base=="C"),3],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
```

```
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

plot(R2[which(R2$base=="."),3],main="SRR6846998_R2",type="l",ylab="freq",xlab="spots",cex.
main=0.9)
lines(R2[which(R2$base=="A"),3],col="red")
lines(R2[which(R2$base=="G"),3],col="green")
lines(R2[which(R2$base=="T"),3],col="blue")
lines(R2[which(R2$base=="C"),3],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

plot(HiSeq_R2[which(HiSeq_R2$base=="."),3],main="HiSeq_R2",type="l",ylim=c(0,0.0012),ylab
="freq",xlab="spots",cex.main=0.9)
lines(HiSeq_R2[which(HiSeq_R2$base=="A"),3],col="red")
lines(HiSeq_R2[which(HiSeq_R2$base=="G"),3],col="green")
lines(HiSeq_R2[which(HiSeq_R2$base=="T"),3],col="blue")
lines(HiSeq_R2[which(HiSeq_R2$base=="C"),3],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")
mtext("Comparison in LOW expression(<30)", side = 3, outer = TRUE,line=-1,cex=1.4)
dev.off()


pdf("illumina_freq_HIGH30_comp.pdf")
par(mfrow=c(2,2))

plot(R1[which(R1$base=="."),4],main="SRR6846998_R1",type="l",ylim=c(0,0.006),ylab="freq",xl
ab="spots",cex.main=0.9)
lines(R1[which(R1$base=="A"),4],col="red")
lines(R1[which(R1$base=="G"),4],col="green")
lines(R1[which(R1$base=="T"),4],col="blue")
lines(R1[which(R1$base=="C"),4],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

plot(HiSeq_R1[which(HiSeq_R1$base=="."),4],main="HiSeq_R1",type="l",ylim=c(0,0.006),ylab="
freq",xlab="spots",cex.main=0.9,)
lines(HiSeq_R1[which(HiSeq_R1$base=="A"),4],col="red")
lines(HiSeq_R1[which(HiSeq_R1$base=="G"),4],col="green")
lines(HiSeq_R1[which(HiSeq_R1$base=="T"),4],col="blue")
lines(HiSeq_R1[which(HiSeq_R1$base=="C"),4],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")
```

```
plot(R2[which(R2$base=="."),4],main="SRR6846998_R2",type="l",ylim=c(0,0.006),ylab="freq",xl
ab="spots",cex.main=0.9)
lines(R2[which(R2$base=="A"),4],col="red")
lines(R2[which(R2$base=="G"),4],col="green")
lines(R2[which(R2$base=="T"),4],col="blue")
lines(R2[which(R2$base=="C"),4],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

plot(HiSeq_R2[which(HiSeq_R2$base=="."),4],main="HiSeq_R2",type="l",ylim=c(0,0.006),ylab="
freq",xlab="spots",cex.main=0.9)
lines(HiSeq_R2[which(HiSeq_R2$base=="A"),4],col="red")
lines(HiSeq_R2[which(HiSeq_R2$base=="G"),4],col="green")
lines(HiSeq_R2[which(HiSeq_R2$base=="T"),4],col="blue")
lines(HiSeq_R2[which(HiSeq_R2$base=="C"),4],col="yellow")
legend("topright",legend=c(".","A","G","T","C"),col=c("black","red","green","blue","yellow"),pc
h=c(1,2,3,4,5),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")

mtext("Comparison in HIGH expression(>=30)", side = 3, outer = TRUE,line=-1,cex=1.4)

dev.off()


png("every_site_expression_illumina.png")
#按位点
par(mfrow=c(2,1))
plot(R1[,3],main="freq<30",xlab="sites",xaxt="n",ylab="low expression",type="l",cex.main=0.9)
lines(R2[,3],col="red")
lines(HiSeq_R1[,3],col="green")
lines(HiSeq_R2[,3],col="yellow")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))
legend("topright",legend=c("R1","R2","HiSeq_R1","HiSeq_R2"),col=c("black","red","green","yell
ow"),pch=c(1,2,3,4),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")


plot(R1[,4],main="freq>=30",xlab="sites",xaxt="n",ylab="high
expression",type="l",cex.main=0.9)
lines(R2[,4],col="red")
lines(HiSeq_R1[,4],col="green")
lines(HiSeq_R2[,4],col="yellow")
legend("topright",legend=c("R1","R2","HiSeq_R1","HiSeq_R2"),col=c("black","red","green","yell
ow"),pch=c(1,2,3,4),lwd=c(2,2),inset=0,cex=0.8,bty="n",box.col="transparent")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))
```

```
mtext("Every_spot expression", side = 3, outer = TRUE,line=-1,cex=1.2)
dev.off()

png("compa.png")
par(mfrow=c(2,2))
plot(R1[,3],main="freq<30",xlab="sites",xaxt="n",ylab="low expression",type="l",cex.main=0.9)
lines(HiSeq_R1[,3],col="green")
legend("topright",legend=c("R1","HiSeq_R1"),col=c("black","green"),pch=c(1,2),lwd=c(2,2),inset
=0,cex=0.8,bty="n",box.col="transparent")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))

plot(R1[,4],main="freq>=30",xlab="sites",xaxt="n",ylab="high
expression",type="l",cex.main=0.9)
lines(HiSeq_R1[,4],col="green")
legend("topright",legend=c("R1","HiSeq_R1"),col=c("black","green"),pch=c(1,2),lwd=c(2,2),inset
=0,cex=0.8,bty="n",box.col="transparent")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))

plot(R2[,3],main="freq<30",xlab="sites",xaxt="n",ylab="low expression",type="l",cex.main=0.9)
lines(HiSeq_R2[,3],col="green")
legend("topright",legend=c("R2","HiSeq_R2"),col=c("black","green"),pch=c(1,2),lwd=c(2,2),inset
=0,cex=0.8,bty="n",box.col="transparent")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))

plot(R2[,4],main="freq>=30",xlab="sites",xaxt="n",ylab="high
expression",type="l",cex.main=0.9)
lines(HiSeq_R2[,4],col="green")
legend("topright",legend=c("R2","HiSeq_R2"),col=c("black","green"),pch=c(1,2),lwd=c(2,2),inset
=0,cex=0.8,bty="n",box.col="transparent")
axis(1,c(0,150,300,450,600,750),labels=c(".","A","T","G","C","N"))

mtext("Comparison", side = 3, outer = TRUE,line=-1,cex=1.2)
dev.off()



setwd("E:/2018.3-2018.7----大三下/基因组信息学
/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/")
R1 <- read.table("processed_illumina_R1.txt",head=T,sep="\t")
R1<-R1[which(R1$site!=150),]
#R1 <- R1[which(R1$site!=150),]
R2 <- read.table("processed_illumina_R2.txt",head=T,sep="\t")
R2<-R2[which(R2$site!=150),]
#R2 <- R2[which(R2$site!=150),]
```

#数据平台
setwd("E:/2018.3-2018.7----大三下/基因组信息学/art_bin_MountRainier/ART_profiler_illumina/illumina_profile/illumina_profiles_self/")
HiSeq_R1 <- read.table("processed_HiSeq_R1.txt",head=T,sep="\t")
HiSeq_R2 <- read.table("processed_HiSeq_R2.txt",head=T,sep="\t")

shapiro.test(R1[,3]-HiSeq_R1[,3]) #p-value < 2.2e-16
shapiro.test(R2[,3]-HiSeq_R2[,3]) #p-value < 2.2e-16

shapiro.test(R1[,4]-HiSeq_R1[,4]) #p-value < 2.2e-16
shapiro.test(R2[,4]-HiSeq_R2[,4]) #p-value < 2.2e-16

wilcox.test(R1[,3],HiSeq_R1[,3])
#W = 434840, p-value = 0.006789
wilcox.test(R1[which(R1$base!="A"),3],HiSeq_R1[which(HiSeq_R1$base!="A"),3])
#W = 296730, p-value = 0.06473
wilcox.test(R1[which(R1$base!="T"),3],HiSeq_R1[which(HiSeq_R1$base!="T"),3]
#W = 299610, p-value = 0.02854
wilcox.test(R1[which(R1$base!="G"),3],HiSeq_R1[which(HiSeq_R1$base!="G"),3])
#W = 308910, p-value = 0.0009673
wilcox.test(R1[which(R1$base!="G"),3],HiSeq_R1[which(HiSeq_R1$base!="G"),3])
#W = 305700, p-value = 0.003542

wilcox.test(R1[,4],HiSeq_R1[,4])
#W = 339170, p-value = 2.171e-09

wilcox.test(R2[,3],HiSeq_R2[,3])
#W = 406450, p-value = 0.8954

wilcox.test(R2[,4],HiSeq_R2[,4])
#W = 351640, p-value = 1.232e-06