华大基因纯测序项目报告

2022/7/19

目 录

1 项目信息

项目编号：Default
样本数量：21

2 数据统计

从测序仪产生的原始读数包含接头、未知或低质量的碱基。原始数据的统计数据如下所示。

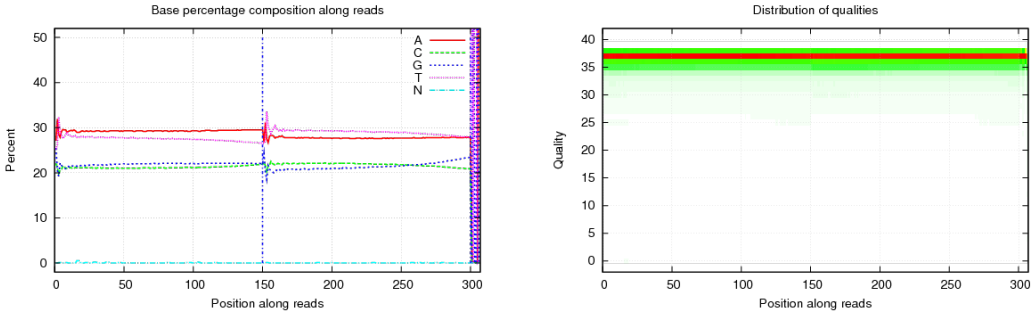
样本	读长	Q20(%)	Q30(%)	GC含量(%)	序列数量	碱基数量
GFP-Pei	150;150;6	96.50;96.68;65.01	90.78;90.56;56.30	47.24;47.14;67.52	46,975	14,374,350
MS22	150;150;6	96.78;98.43;92.43	91.48;95.49;85.60	36.73;37.34;49.72	8,230,122	2,518,417,332
MS23	150;150;6	97.99;97.91;85.46	94.24;94.20;72.91	49.60;50.14;50.37	2,481,656	759,386,736
MS24	150;150;6	98.08;98.14;89.27	94.54;94.82;79.60	48.92;49.51;50.28	2,237,195	684,581,670
MS25	150;150;6	98.06;98.05;87.52	94.28;94.55;77.30	50.18;50.65;48.67	3,029,000	926,874,000
MS26	150;150;6	96.93;98.50;93.59	91.80;95.50;87.39	32.70;33.03;49.96	8,807,374	2,695,056,444
MS27	150;150;6	96.66;98.37;92.38	91.13;95.07;86.27	30.63;31.14;33.91	7,249,856	2,218,455,936
Tn5-EC	150;150;6	97.36;96.53;28.88	92.72;90.86;15.74	44.93;44.73;55.96	52,986	16,213,716
Tn5-EP	150;150;6	92.82;96.19;82.48	81.49;88.27;77.88	40.51;40.49;49.46	196,593	60,157,458
Tn5-KP	150;150;6	95.50;97.59;86.76	88.03;92.45;81.71	42.35;42.37;50.44	220,746	67,548,276
Tn5-L	150;150;6	97.46;97.32;97.17	93.03;92.60;94.67	43.16;44.42;50.04	2,148,302	657,380,412
a	150;150;6	97.98;98.18;97.05	94.20;94.48;93.16	43.13;43.23;33.53	14,540,106	4,449,272,436
b	150;150;6	97.91;98.42;95.63	94.01;95.20;90.93	42.87;42.94;49.85	10,147,031	3,104,991,486
c	150;150;6	98.06;98.07;97.38	94.35;94.15;93.84	45.89;45.97;49.99	10,328,265	3,160,449,090
d	150;150;6	98.06;98.47;90.78	94.59;95.48;82.74	44.56;44.24;83.21	17,743,683	5,429,566,998
dMcphC1	150;150;6	98.04;98.07;96.00	94.27;94.08;90.84	43.77;43.82;66.22	77,186,065	23,618,935,890
dMcphC2	150;150;6	98.06;98.24;96.16	94.31;94.55;90.95	43.50;43.55;50.18	76,998,905	23,561,664,930
dMcphRA1	150;150;6	97.94;98.22;97.66	93.97;94.49;93.96	42.76;42.82;50.05	71,590,959	21,906,833,454
dMcphRA2	150;150;6	98.07;98.24;98.54	94.37;94.62;96.41	42.92;42.97;49.97	44,720,493	13,684,470,858
e	150;150;6	98.28;98.57;91.53	95.24;95.82;83.55	44.01;43.78;82.65	22,826,546	6,984,923,076
f	150;150;6	97.98;98.64;97.21	94.27;95.88;93.97	42.47;42.58;50.02	23,779,271	7,276,456,926

格式说明：
1. 样本: 样本名
2. 读长: 测序序列长度
3. Q20 (%): 质量值大于20的碱基在序列中的占比
4. Q30 (%): 质量值大于30的碱基在序列中的占比
4. GC含量(%): C和G碱基在序列中的占比
5. 序列数量: 原始序列数量
6. 碱基数量: 原始碱基数量

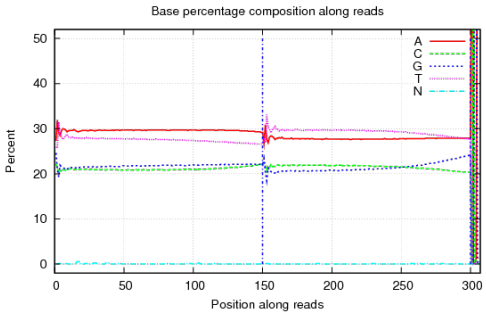
3 数据质控

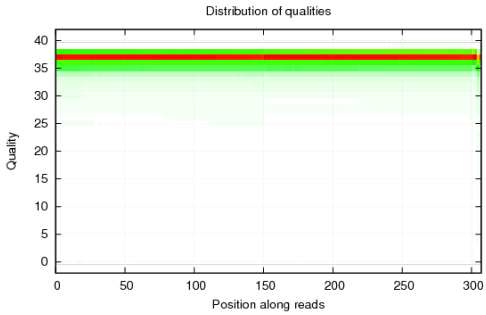
数据过滤中碱基比例和质量值分布如下图所示（如果一个样本有多个lane，则只显示其中一个）。左图是样本碱基百分比分布，右图是样本序列的质量分布。

样本 a

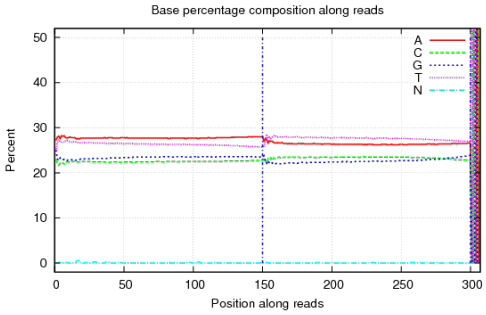


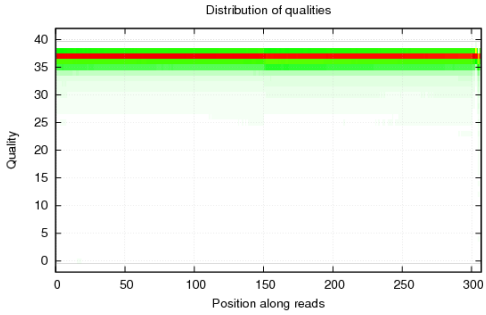
样本 b



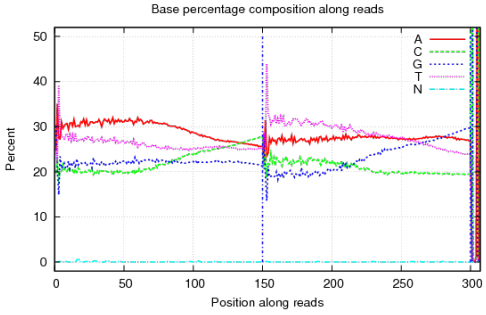


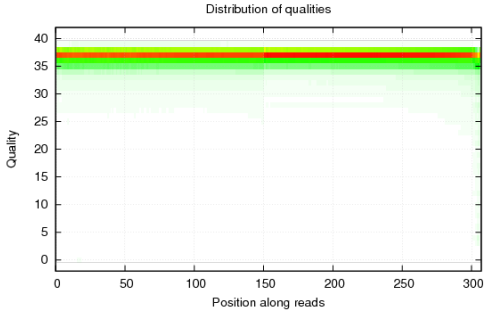
样本 c



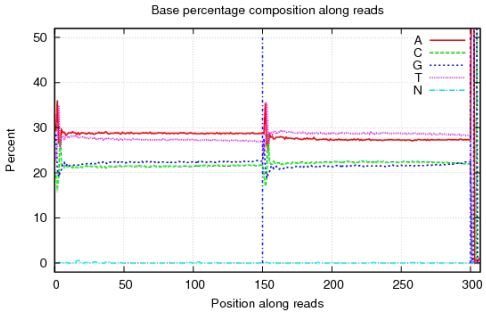


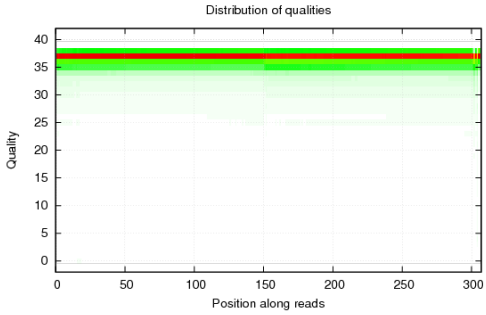
样本 d



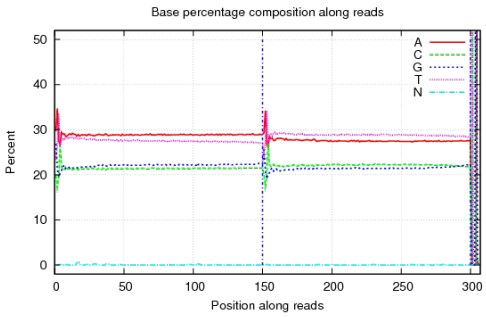


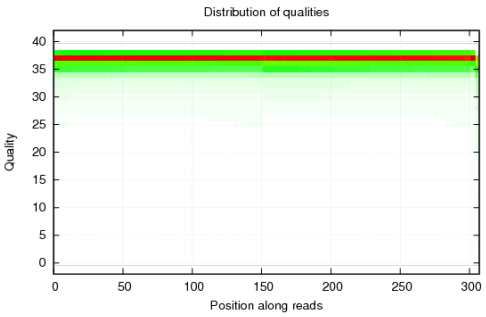
样本 dMcphC1



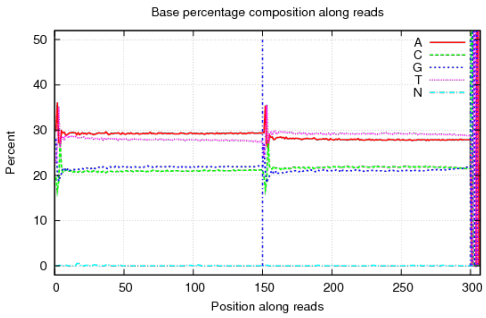


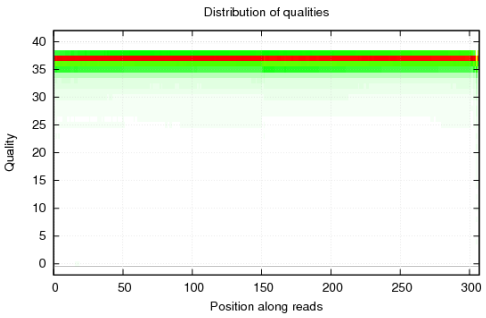
样本 dMcphC2



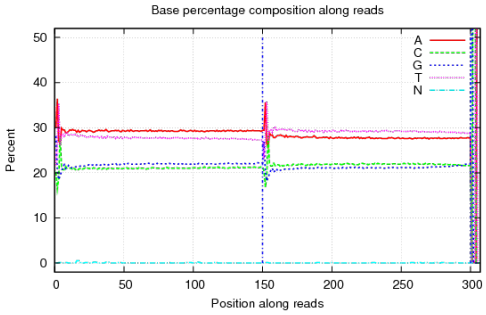


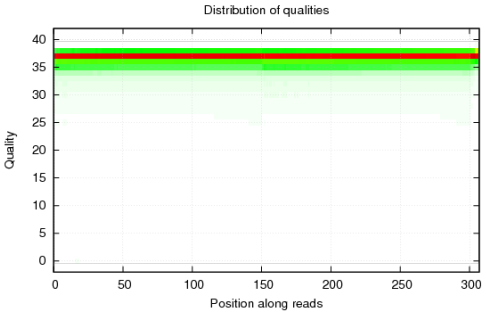
样本 dMcphRA1



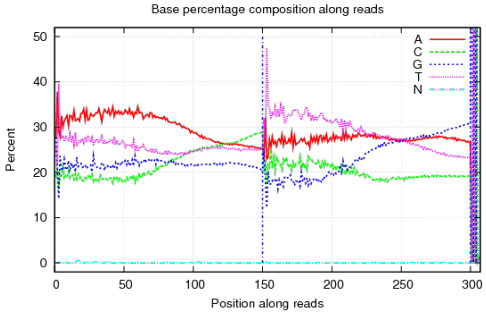


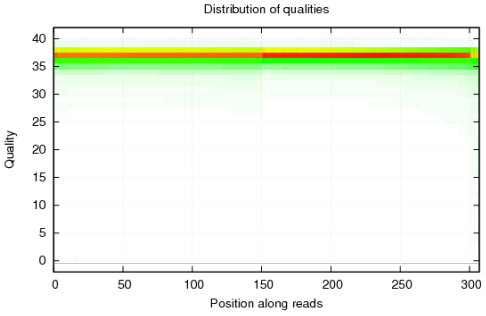
样本 dMcphRA2



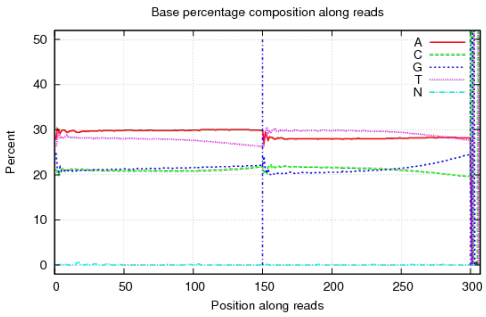


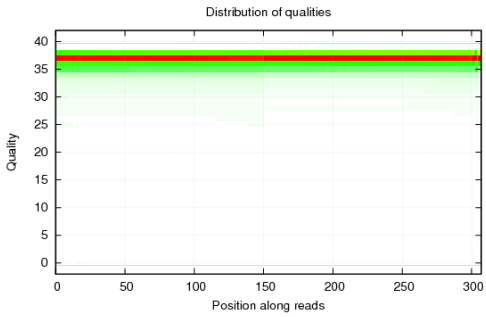
样本 e



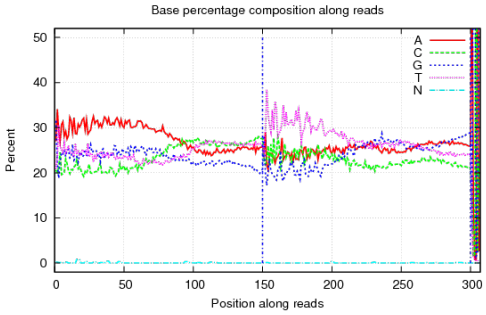


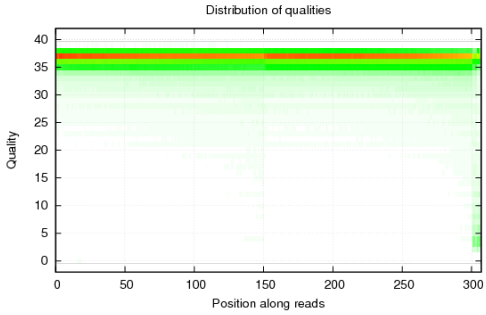
样本 f



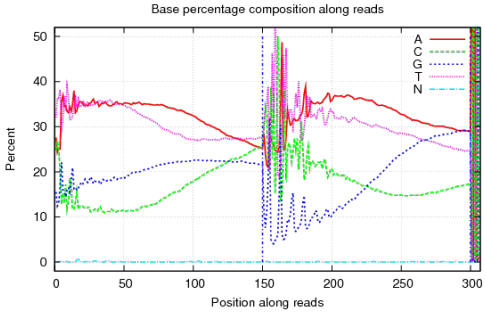


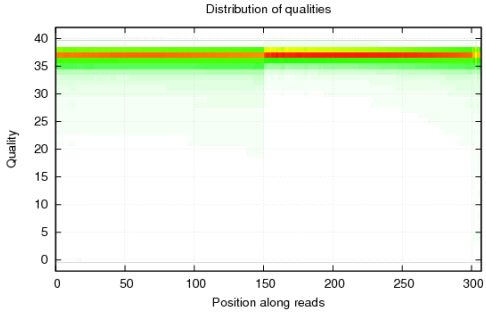
样本 GFP-Pei



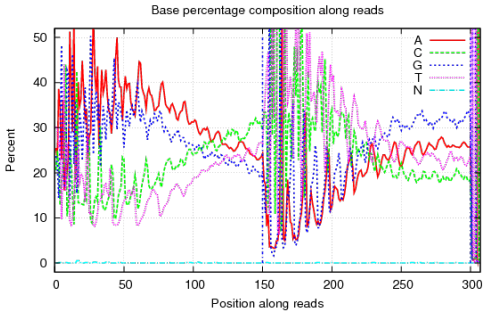


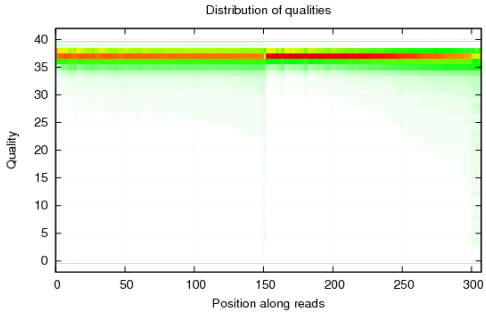
样本 MS22



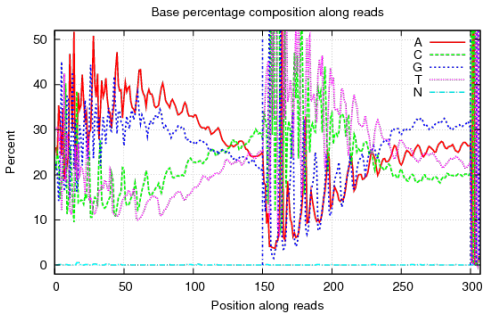


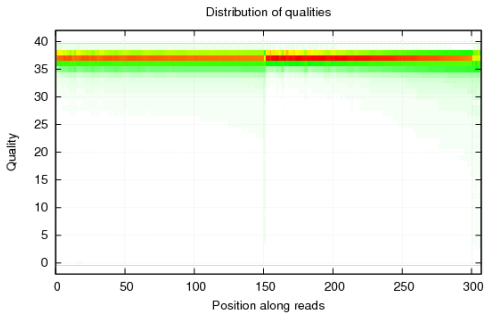
样本 MS23



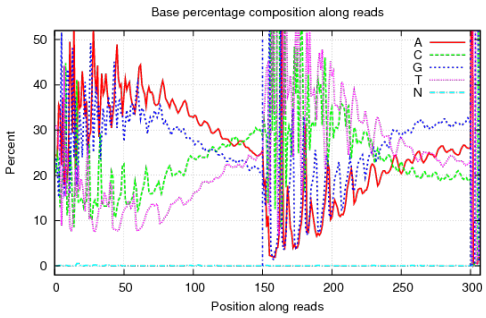


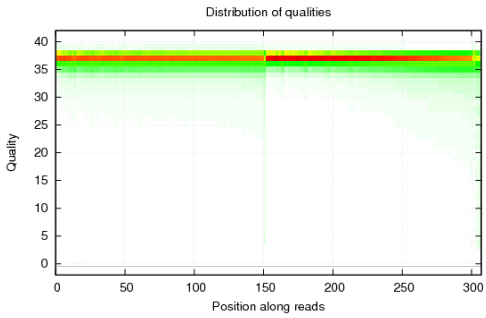
样本 MS24



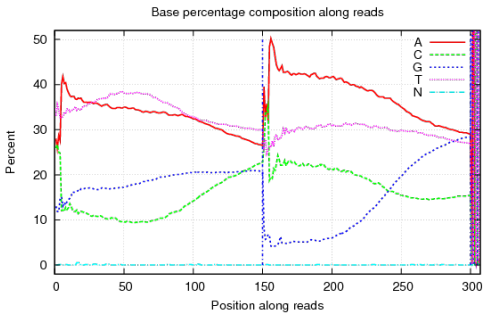


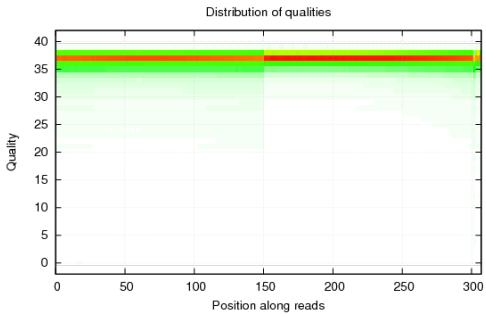
样本 MS25



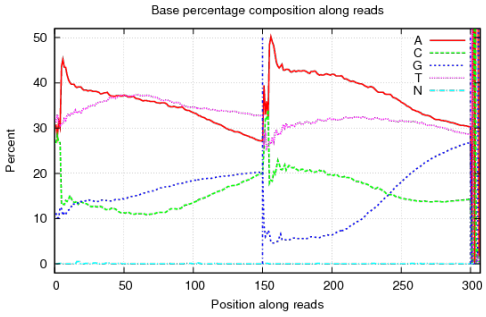


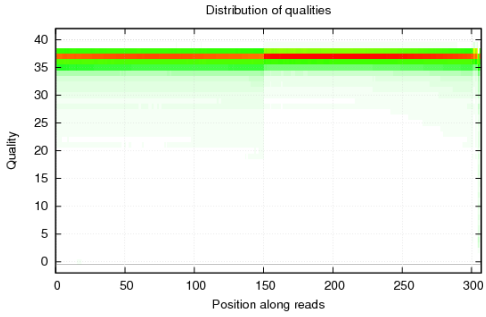
样本 MS26



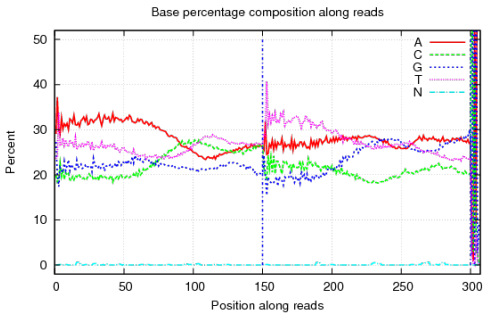


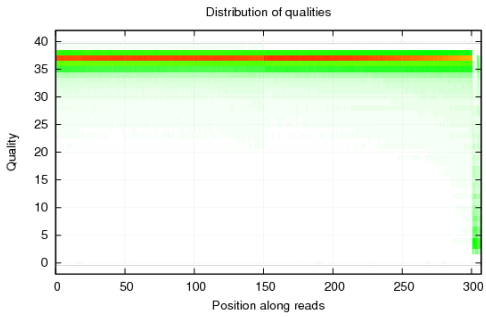
样本 MS27



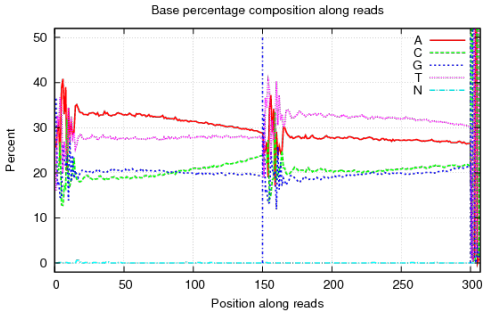


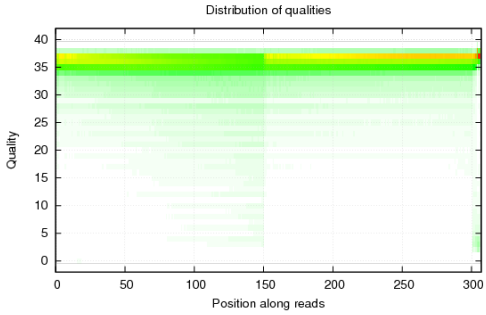
样本 Tn5-EC



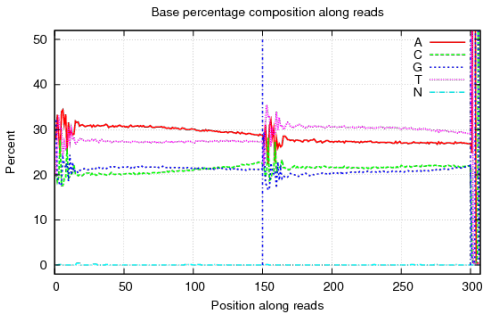


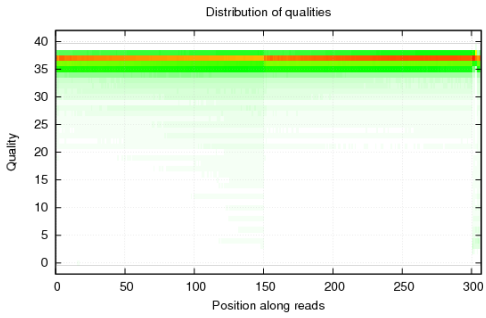
样本 Tn5-EP



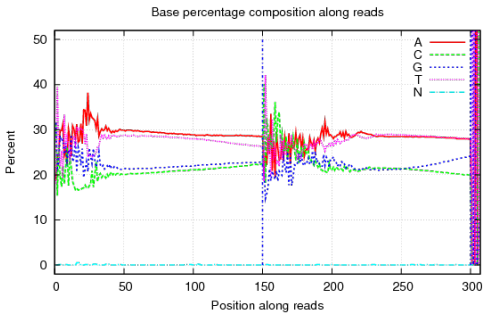


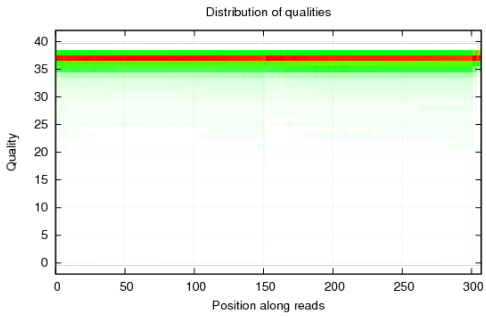
样本 Tn5-KP





样本 Tn5-L





4 帮助文档

原始图像数据转换为序列数据，定义为原始数据并保存为 FASTQ 文件。FASTQ 文件中的每个条目由4行组成：

1. 包含有关测序运行和集群信息的序列标识符。此行的确切内容因使用的BCL到FASTQ转换软件而异；

2. 测序序列(包含碱基：A,C,T,GandN)；
3. 分隔符；
4. 测序质量值，通常是Phred+33编码,使用ASCII字符表示质量分数。

以下是FASTQ文件中单条序列的示例：

```
@V300029029L1C001R0010000210/1
GCGACCCCAGGTCAGTCGGGACTACCCGCTGAAGTCGGAGGCCAAGCGGT
+
FFFCFFFFFFFFFDFEFFFFEFEF0FFFFEFFFFFFFEFFFFFECGFFFF
```

DNBseq测序仪测序错误率与测序质量值的关系如下式所示。具体来说，如果测序错误率记为“E”，DNBseq测序碱基质量值记为“sQ”，关系如下：

$$sQ = -10\log_{10} E$$

测序错误率	测序碱基质量值	Phred +33 编码字符
5%	13	.
1%	20	5
0.1%	30	?