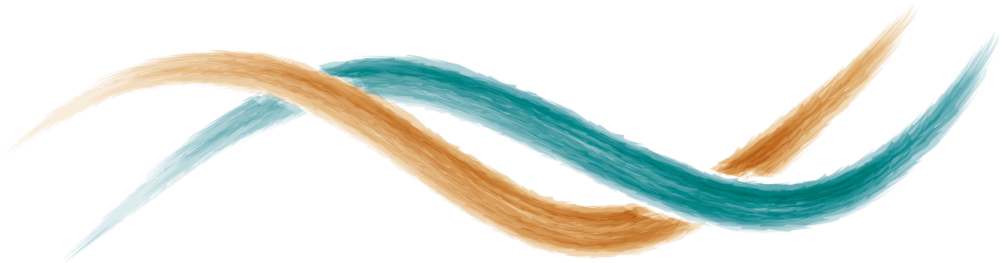
数据质控
结题报告



2023/05/08

目　　录

1 数据质控

1.1 数据统计和分布

PacBio测序的下机数据的保存格式为BAM，下机的Subreads使用PacBio官方软件SMRT Link中的ccs模块进行CCS分析得到HiFi reads，基于HiFi reads进行后续的分析。

数据量信息如下表:

表1 HiFi Reads统计表

Sample	EIL48
HiFi Reads	1,972,775
HiFi Yield (bp)	36,028,539,967
HiFi Read Length (mean, bp)	18,262
HiFi Read Quality (median)	Q31
Max Length	50,574
N50	19,124
GC Content	45.75%

（1）Sample：样本名称;
（2）HiFi Reads：HiFi Reads数目;
（3）HiFi Yield (bp)：HiFi Reads碱基总数;
（4）HiFi Read Length (mean, bp)：HiFi Reads平均长度;
（5）HiFi Read Quality (median)：HiFi Reads质量值中位数;
（6）Max Length：最长的HiFi Reads长度;
（7）N50：HiFi Reads的N50;
（8）GC content: HiFi Reads的GC含量。

绘制各个样本的HiFi Reads准确度分布统计图如下:

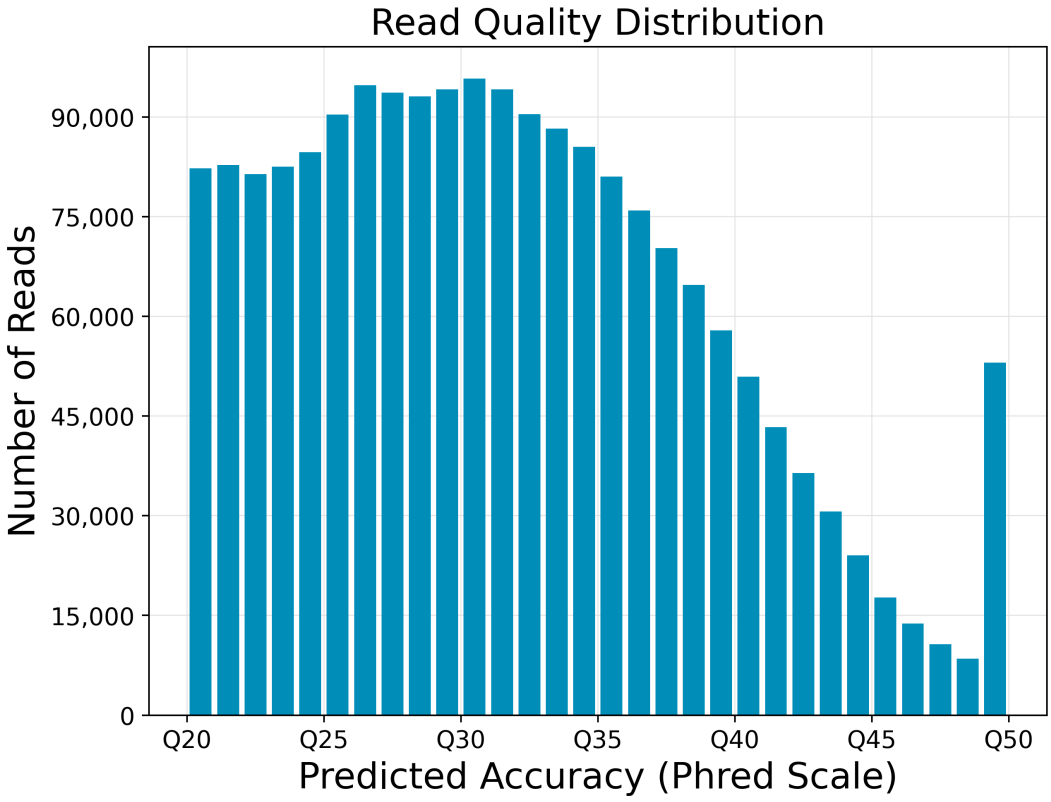


图1 HiFi Reads准确度分布图

上图横坐标为质量值，纵坐标为reads数量，可以通过上图看到全部HiFi Reads整体的质量分布。

绘制各个样本的HiFi Reads长度分布统计图如下:

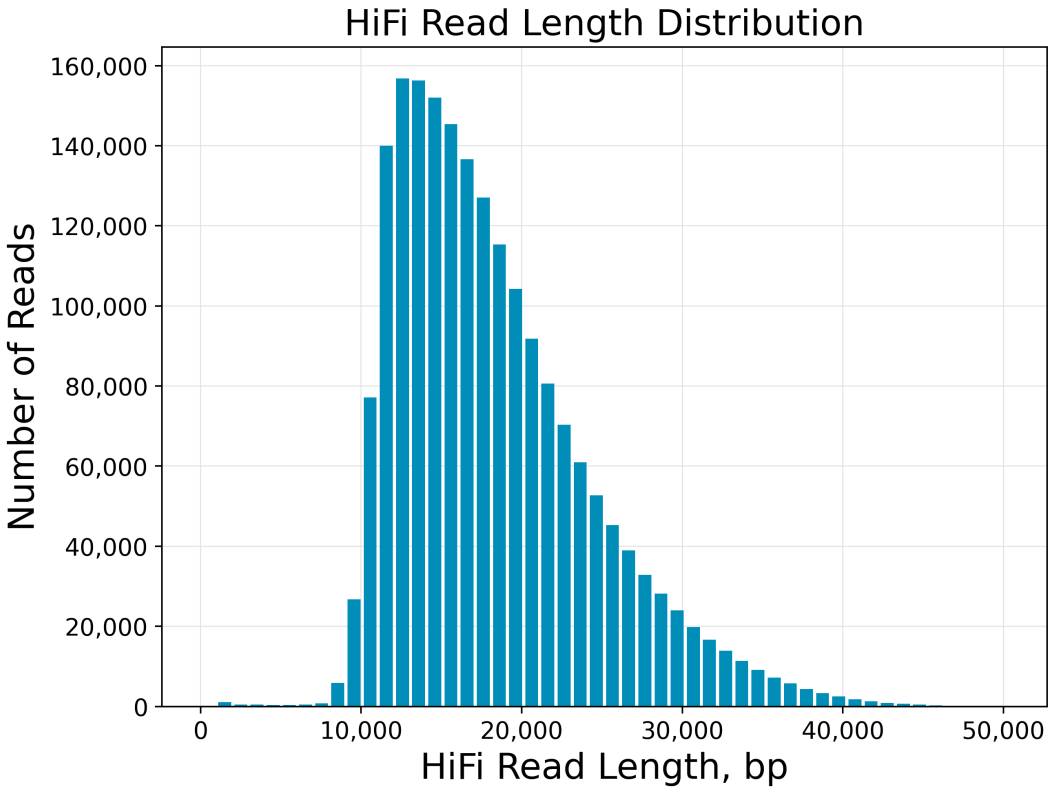


图2 HiFi Reads长度分布图

横轴为HiFi reads长度，纵轴为HiFi reads数量，可以通过此图看到全部HiFi Reads整体的长度分布。