sgRNA文库质检测序分析结题报告

1 生信分析流程

以PCR的方式进行illumina平台的测序文库构建，文库检测合格后，不同文库按照有效浓度及目标下机数据量的需求混样后进行测序。测序完成获得测序深度足够的原始数据后，即开始进行生物信息分析，该过程主要包括以下三个阶段。

（1）sgRNA 序列提取以及数据质控：根据建库方法提取测序数据中包含 sgRNA 序列的数据，保留20nt的靶向识别相关核心序列，得到 sgRNA核心序列的fastq文件,并进行数据量统计与质量评估，以确保满足后续分析的要求；

（2）sgRNA 比对与数目统计：质控后的数据用于比对，并进行比对率、sgRNA 测序深度和sgRNA及基因覆盖度评估，达到要求方可进行下一步分析；

（3）sgRNA的均一性：使用skew ratio以及sgRNA read count的AUC曲线评估文库sgRNA的均一性，最终评判文库的质量。

2 基本分析

2.1 测序数据评估

测序仪CCD扫描的四种荧光通道下的光学信号通过内置软件RTA实时Basecalling得到bcl 数据，测序完成后使用Illumina 官方软件Bcl2fastq(v2.17) 根据样本Index序列进行数据拆分并转换成 Fastq格式。测序仪HCS软件的Primary Analysis 过程会根据reads前25个cycle 的荧光信号的纯度确定是否 Pass Filter（前25个cycle的 Chastity 值小于 0.6 的碱基不超过 2 个），PF clusters经Bcl2fastq转换之后测序结果以FASTQ 文件格式存储，称为PF data 或 Raw data。对于双端测序 (paired-end) 数据，序列数据由两个 FASTQ文件构成，分别保存每一端测序的读段（read）。

FASTQ格式文件中每条read由四行组成：

@GWZHISEQ01:289:C3Y96ACXX:6:1101:1704:2425 1:N:0:GGCTAC
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTG
+
AAFFFFFKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKKFFFFFF

每个序列共有4行，第1行和第3行是序列名称（第3行一般省略“＋”后面的序列名称），由测序仪产生；第2行是序列；第4行是序列的测序质量，每个字符对应第2行每个碱基，每个字符对应的ASCII值减去33，即为该碱基的测序质量值，比如@对应的ASCII值为64，那么其对应的碱基质量值是31。根据Illumina最新版本Real Time Analysis软件，碱基质量值范围为0到41。

2.1.1 sgRNA 数据提取与统计

根据建库方法定制in-house脚本针对双端测序数据提取20bp的sgRNA序列，得到sgRNA Clean Data，用于进行后续分析。根据sgRNA的在测序数据中的分布情况，统计原始测序序列的组成，如表1所示。

Table 1 原始数据与clean data的组成

Sample_name	Raw_data	Clean_data	Percent
200445_D10216_1	8,489,407	8,209,828	96.71%

2.1.2 测序质量与错误率分布统计

碱基质量与测序错误率有关，受样品、测序试剂、测序仪本身等多因素影响。对于Illumina测序平台，测序质量值和测序错误率分布呈现2个特点:（1）测序质量值随着测序长度的增加而降低，错误率相应升高;（2）每个read的前几个碱基测序质量一般较低，错误率相应较高，这是由于边合成边测序过程初始阶段，测序仪荧光感光元件对焦速度较慢，获取的荧光图像质量较低，导致碱基识别错误率高。此现象为Illumina 高通量测序平台的共有特征。一般而言，测序数据的Q30比例需大于80%，可用于后续分析。测序错误率分布检查用于检测在测序长度范围内，有无异常的碱基位置存在高错误率，比如中间位置的碱基测序错误率显著高于其他位置。一般情况下，每个碱基位置的测序错误率都应该低于0.5%。 如果测序错误率用e表示，测序的碱基质量值用Qphred表示，Illumina 平台的测序错误率与碱基质量值的关系有如下公式：

$$Qphred = -10\log_{10}(e)$$

由公式可知碱基质量值越高，测序错误率越低，对应关系如下表：

Table 2 碱基质量值对应表

Q.score	Base.calling.correct.rate	Erro_rate
Q40	99.99%	1/10000
Q30	99.9%	1/1000
Q20	99%	1/100

样本read的不同测序位置碱基平均质量值和平均测序错误率统计如图下图所示。

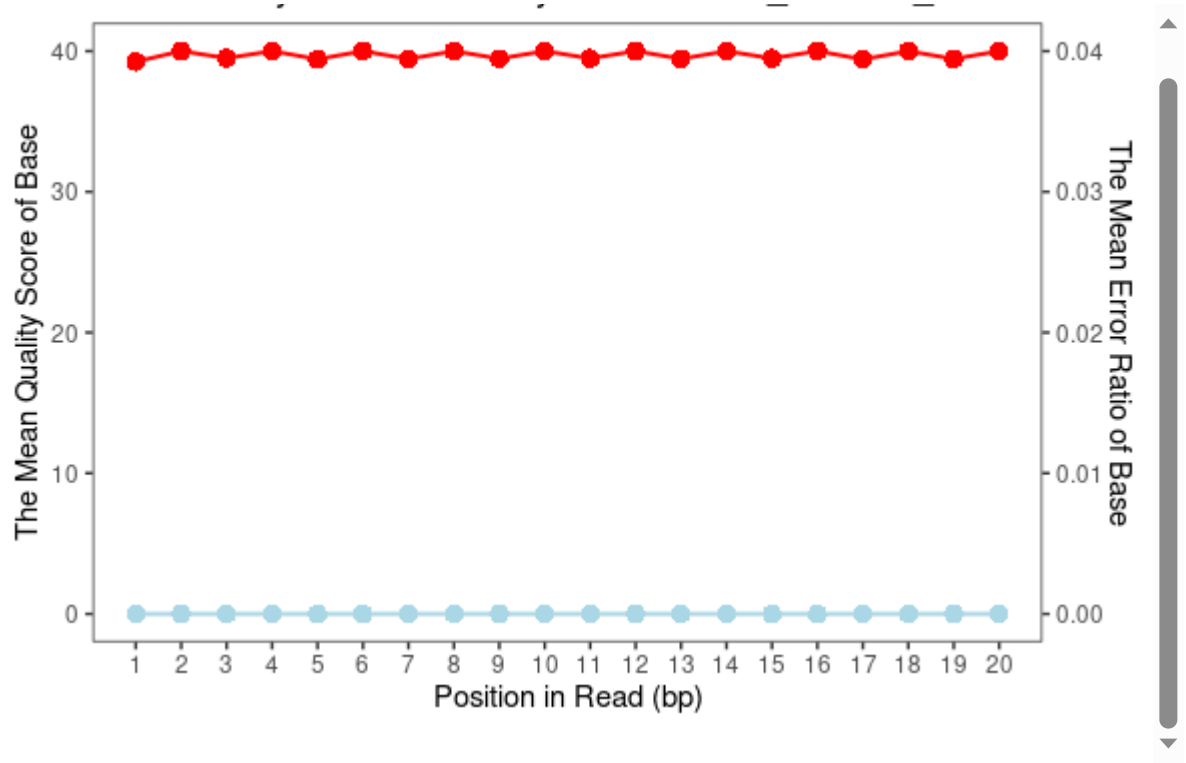


Figure 1 碱基质量值与错误率

注：碱基质量值与错误率：红色点线表示碱基平均测序质量，浅蓝色点线表示碱基平均错误率。

横坐标为read中的碱基位置，纵坐标为对应碱基位置的平均质量值；1-20bp为sgRNA核心序列。

2.1.3 碱基含量分布统计

碱基含量分布统计用于检测有无 AT、GC分离现象。鉴于测序序列的随机性和碱基互补配对原则，理论上文库的每个测序循环中GC及AT含量应分别相

等，且在整个测序过程基本稳定不变，呈水平线。扩增偏好、测序偏好、测序深度不足等原因可能会造成测序数据碱基含量分布异常。

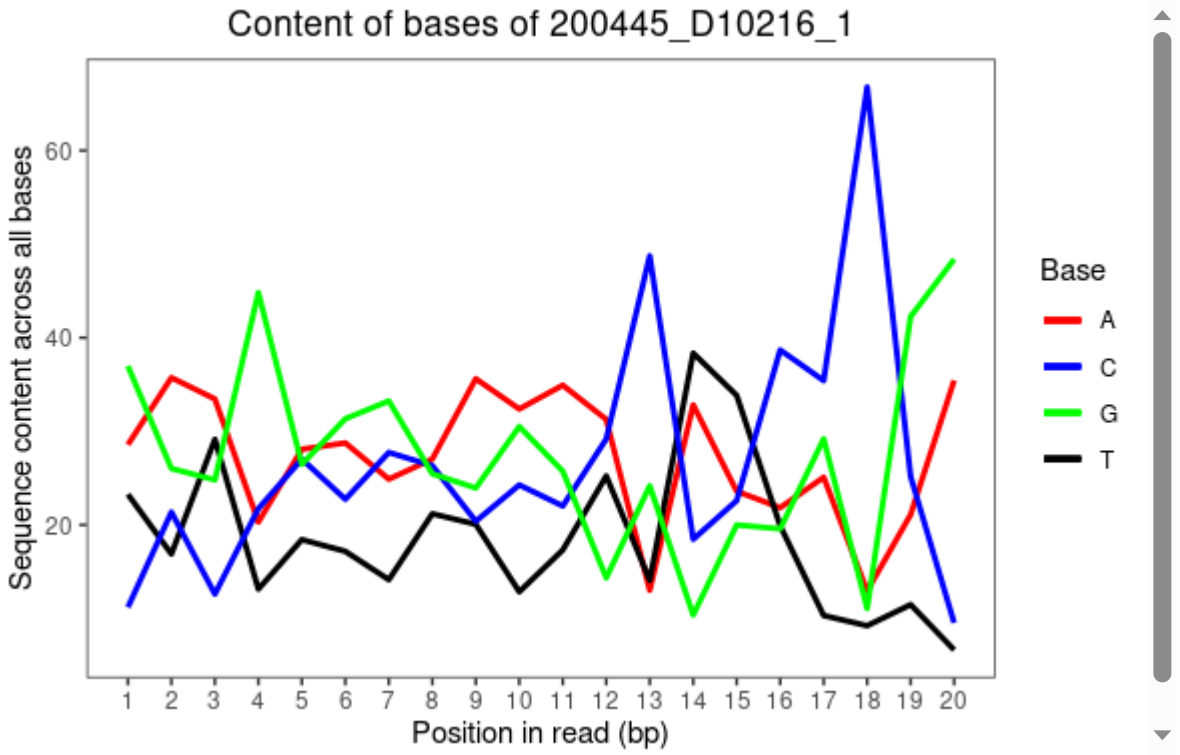


Figure 2 碱基含量分布图

注：碱基含量分布图：

横坐标为reads碱基位置，纵坐标为对应类型碱基在总碱基中的百分比；1-20bp为gRNA 核心序列。

2.2 比对与Readcount分析

采用In-house 脚本将提取的靶向序列比对到 sgRNA library参考序列，并对Library中的所有序列的比对结果进行统计，包括完全匹配的 reads 数目，并统计测序深度、覆盖度、比对率等方面指标，该结果可反映数据的可靠性和准确性。

2.2.1 sgRNA 密度分布

核密度估计（kernel density estimation）是在概率论中用来估计未知的密度函数，属于非参数检验方法之一，通过核密度估计图可以比较直观的看出样本数据本身的分布特征。一般情况下，较于实验组样本，Baseline 和质粒样本 Readcount 更高，其对应的核密度估计图更向右集中。

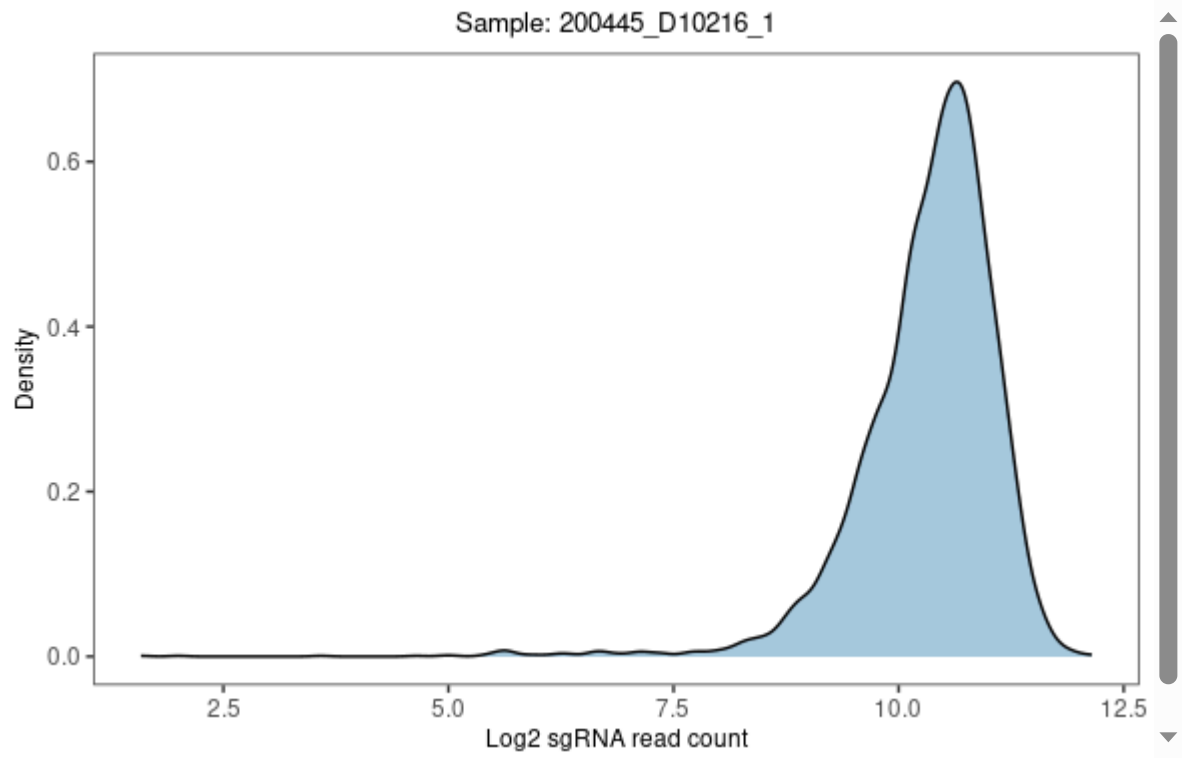


Figure 3 sgRNA Readcount核密度图

注：sgRNA Readcount核密度图：

横坐标为基因组上的单个sgRNA的Readcount的Log2值，纵坐标为分布密度，即各sgRNA Readcount的分布概率。

2.2.2 基因Readcount箱形图

使用箱线图展示所有样本中检测到的靶向基因所对应Readcount的分布情况，箱线图的离散程度的变化也反应了相应处理条件下样本的基因富集或丢失的效果。一般情况下，实验组的Readcount分布更为离散。

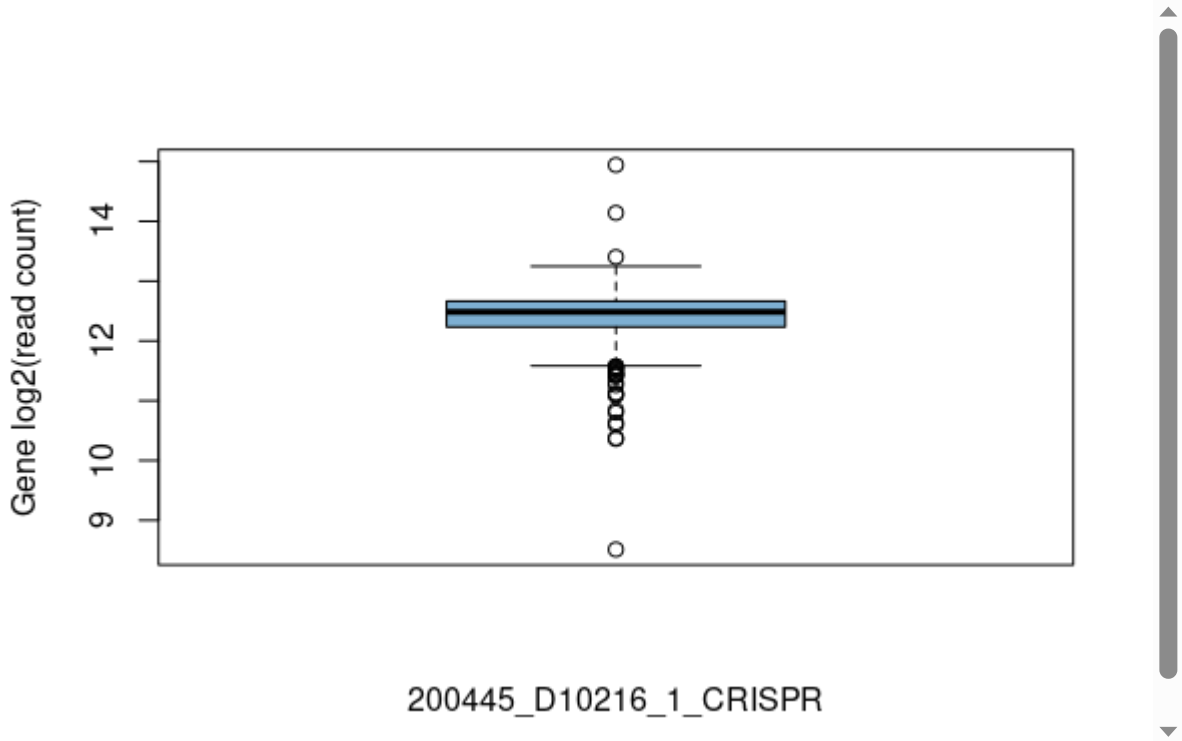


Figure 4 样本基因分布箱形图

注：基因分布箱形图：

横坐标为样本名，纵坐标为gene的Log2 Readcount值。

2.2.3 比对数据统计

将所有样本的比对数据进行汇总如表3所示。

Table 3 样本的比对数据汇总

	MappedReads	Mapping.ratio	TotalsgRNAs	ZerosgRNAs
3	7,152,725	87.12%	4,993	8

注：比对信息统计表：

Read：sgRNA reads总数；

MappedReads：完全比对上sgRNA Library的reads数目；

Mapping ratio：完全比对上sgRNA Library的reads比例

TotalsgRNAs：sgRNA library中sgRNA总数；

ZerosgRNAs：sgRNA文库中丢失的sgRNA数目；

Coverage_rate；文库覆盖度，即测到的 sgRNA占总sgRNA 的比率，对于文库质控，一般要求覆盖度 >90%；

Totalgenes：sgRNA library对应基因总数；

Zerogenes：未能检测到的基因数目。

2.3 测序深度及均一性评估

根据sgRNA序列的比对结果，统计每条 sgRNA 所检测到reads数目，从而评价文库的均一性。文库均一性评估在功能筛选上是非常重要的，可以有效避免sgRNA的缺失导致的筛选阳性靶点的遗漏以及假阳性结果。

2.3.1 sgRNA 测序深度柱状图

统计各样本每条 sgRNA数目，并进行可视化展示。该图可以直观的展示sgRNA的测序深度，一般呈泊松分布，如图下图所示。

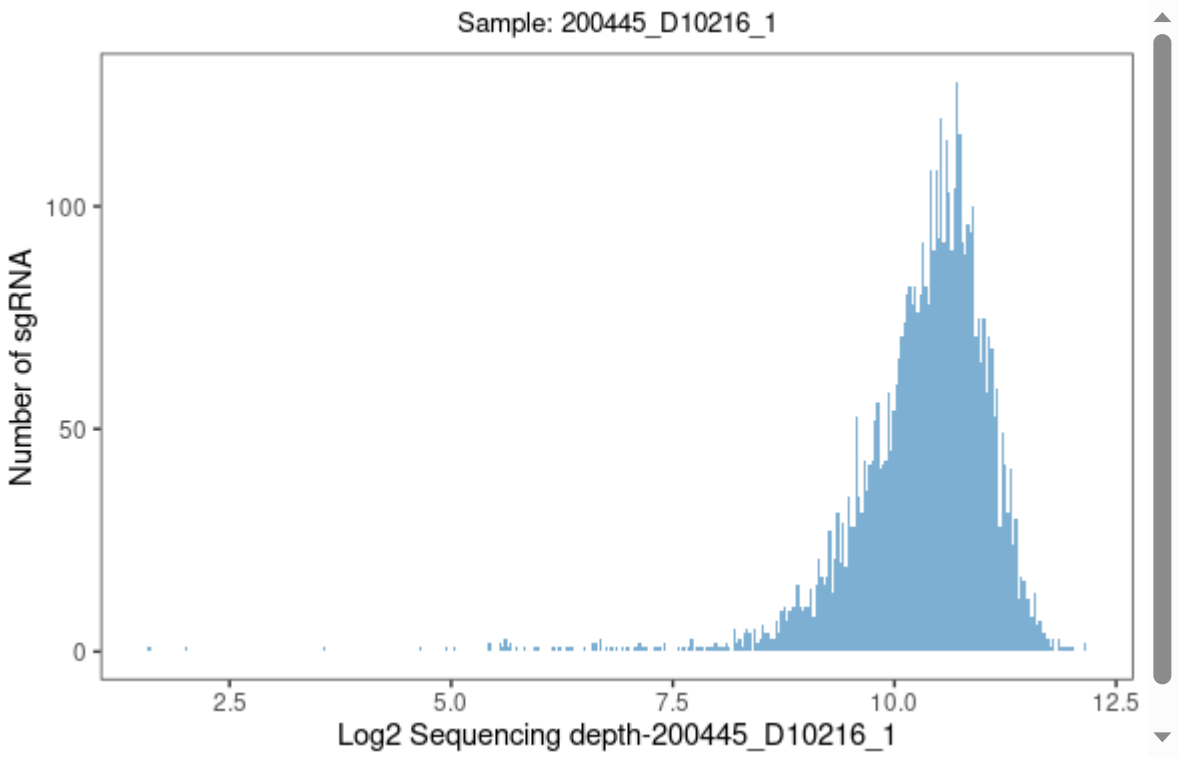


Figure 5 sgRNA 测序深度柱状图

注：sgRNA 测序深度柱状图：

横坐标为测序深度的Log2值，纵坐标为对应的 sgRNA 数目。

2.3.2 均一性斜率图

根据各样本 sgRNA 数目，统计其分布斜率，并进行可视化展示，如图8所示。分布斜率(Skew Ratio)为累积分布达到90%时对应的 sgRNA 数目与累积分布达到10%时对应的sgRNA数目的比值，该值可以直接反应文库的均一性，一般认为，Skew Ratio < 10，则文库均一性合格。

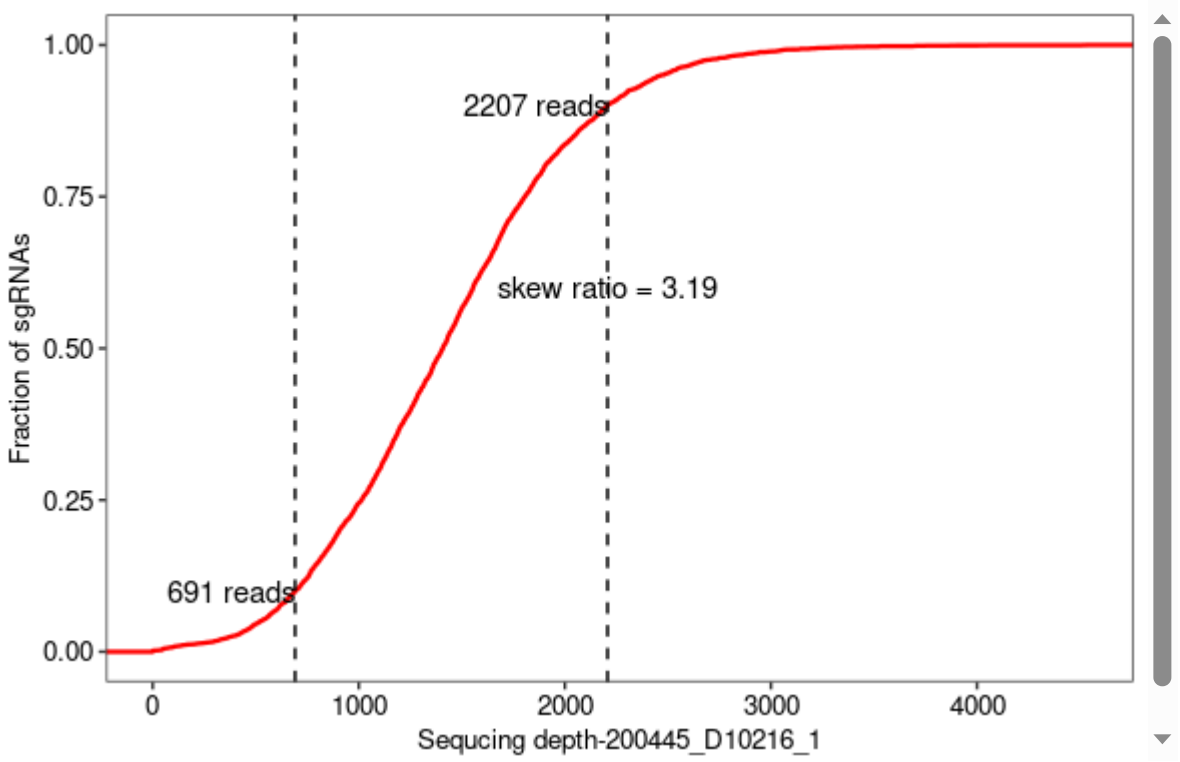


Figure 6 均一性斜率图

注：均一性斜率图：

横坐标为测序深度，纵坐标为 sgRNA 累积频率，左下值为累积频率达到10%时对应的 sgRNA 数目，左中值为达到90%时对应的 sgRNA 数目，Skew Ratio 为二者比值。

参考文献

1. Joung J, Konermann S, Gootenberg JS, Abudayyeh OO, Platt RJ, Brigham MD, Sanjana NE, Zhang F. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nat Protoc. 2017 Apr;12(4):828-863. doi: 10.1038/nprot.2017.016. Epub 2017 Mar 23. Erratum in: Nat Protoc. 2019 Jul;14(7):2259. doi: 10.1038/s41596-018-0063-0. PMID: 28333914; PMCID: PMC5526071.
2. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelson T, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014 Jan 3;343(6166):84-87. doi: 10.1126/science.1247005IF: 44.7 Q1 . Epub 2013 Dec 12. PMID: 24336571; PMCID: PMC4089965.
3. Sanson KR, Hanna RE, Hegde M, Donovan KF, Strand C, Sullender ME, Vaimberg EW, Goodale A, Root DE, Piccioni F, Doench JG. Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun. 2018 Dec 21;9(1):5416. doi:

10.1038/s41467-018-07901-8. PMID: 30575746; PMCID: PMC6303322.