数据产出

1 数据产出

测序完成后，对原始数据进行数据过滤，数据过滤包括去污染，去接头及去除低质量数据。过滤后得到的clean data产量统计如表1-1。

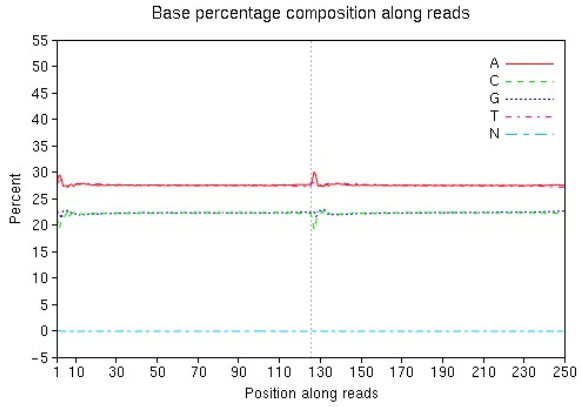
表**1-1** Clean data数据统计

No.	Sample Name	Read length(bp)	Clean Reads	Clean bases	Q20(%)	GC(%)
1	188	125	480427706	60053463250	98.12;95.04	44.72
2	283	125	495470300	61933787500	97.96;93.33	45.26

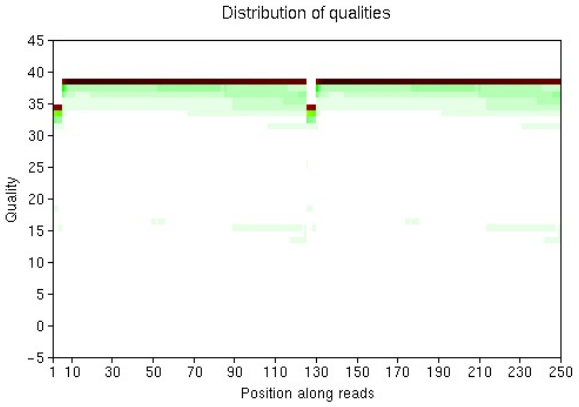
数据质控

2 数据质控

数据过滤后的reads每一位置的碱基分布和质量值分布如图。

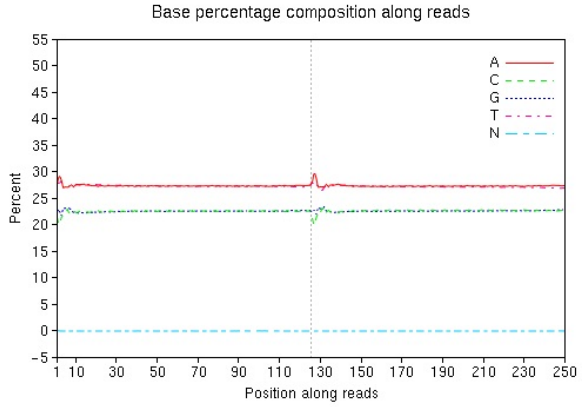


a)过滤后reads每个位置的碱基分布

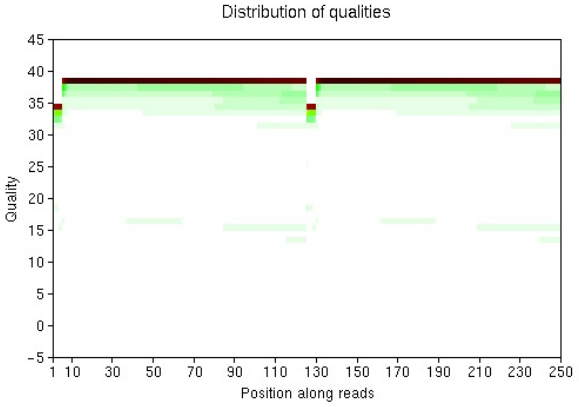


b)过滤后reads每个位置上碱基的质量值分布

样品188质控图



a)过滤后reads每个位置的碱基分布



b)过滤后reads每个位置上碱基的质量值分布

样品283质控图

流程说明

1 原始序列格式说明

测序得到的原始图像数据经base calling转化为序列数据，我们称之为raw data或raw reads，结果以FASTQ文件格式存储，包含reads的序列以及reads的测序质量。在FASTQ格式文件中每个read由四行描述，如下：

33质量体系的数据
@SIM:1:FCX:1:15:6329:1045 1:N:0:ATCCGA
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>##=><9=AAAAAAAAA9#:<#<<<<????#=
或者是64质量体系的数据
@A80GVTABXX:4:1:2587:1979#ACAGTGAT/1
NTTTGATATGTGTGAGGACGTCTGCAGCGTCACCTTTATCGGCCATGGT

+
BTTMKZXUUUddddddddddddddddddddddddddddddaddddddd^WYYU

每个序列共有4行，第1行和第3行是序列名称（有的fq文件为了节省存储空间会省略第三行“+”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，用字符表示并一一对应于第2行每个碱基；比如33质量体系的数据中，字符A对应的ASCII值为65，那么其对应的碱基质量值是65-33=32；在64质量体系的数据中，字符c对应的ASCII值为99，那么其对应的碱基质量值是99-64=35；表1-1为Illumina HiSeq™平台测序错误率与测序质量值简明对应关系。具体地，如果测序错误率用E表示，碱基质量值用sQ表示，则有下列关系：

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

表**1-1** Ilumina HiSeq™ 平台测序错误率与测序质量值简明对应关系

测序错误率	测序质量值	对应**33**质量体系字符	对应**64**质量体系字符
5%	13	.	M
1%	20	5	T
0.1%	30	?	^

2 数据产出统计项说明

表**2-1** 数据产出统计项说明

统计项	说明
Sample Name	样品名
Insert Size (bp)	测序片段长度
Read length (bp)	read长度
Clean Reads	过滤后的reads数
Clean bases	过滤后的碱基数
Q20 (%)	质量为20的碱基比例，即测序错误率为1%的碱基比例
GC (%)	GC数/总碱基数