1. 不同RNA类型的.bed文件制作
   1. 根据RNA类型，将RNA分类
   2. 针对每种类型RNA，查找gtf文件，得到.bed文件

* .bed文件组成：染色体名称 起始位点 终止位点
  1. 参考（可咨询常贵）：

/hwfssz5/ST\_HEALTH/P20Z10200N0041/leichanggui/cfRNA/ref/bio\_RNA\_bed

1. 不同RNA类型的bam提取
   1. 进入bam所在文件夹，利用samtools建立索引

cd /zfssz3/ST\_MCHRI/BIGDATA/P20Z10200N0041/cfRNA/hill/mushitage/fast\_version\_v2/hisat

###建索引

samtools index SGI75460075Aligned.sortedByCoord.out.bam

本次需要测试的bam文件夹地址为：

|  |  |  |
| --- | --- | --- |
| 样本类型 | Bam文件夹 | 原始数据路径 |
| Hill | /zfssz3/ST\_MCHRI/BIGDATA/P20Z10200N0041/cfRNA/hill/mushitage/fast\_version\_v2/hisat | /zfssz3/ST\_MCHRI/BIGDATA/P20Z10200N0041/cfRNA/hill/mushitage/fast\_version\_v2/fastp |
| Zhongzhong | /zfssz3/ST\_MCHRI/BIGDATA/P20Z10200N0041/cfRNA/cfRNADATA/zhongzhong0207/hisat | /jdfssz1/ST\_HEALTH/P20Z10200N0041/Lishaogang/cfRNASample/zhongzhong0207/01\_quantitative\_analyses/result |
| Jingshen | /zfssz3/ST\_MCHRI/BIGDATA/P20Z10200N0041/cfRNA/cfRNADATA/jingshen0212/hisat | /jdfssz1/ST\_HEALTH/P20Z10200N0041/Lishaogang/cfRNASample/jingshen0212/01\_quantitative\_analyses/result |

* 1. 根据.bed文件，提取相应类型RNA的bam

samtools view -hb 输入bam所在路径/SGI75460075Aligned.sortedByCoord.out.bam -L 制作的.bed >输出结果路径/SGI75460075\_HBB.bam

* 1. 统计bam文件里的所有reads的长度分布
     1. 参考方案
        + Bam转.fq:

samtools view 输入bam所在路径/SGI75460075\_HBB.bam | awk '{print $1}' | sort | uniq > 输出结果路径/SGI75460075\_star\_name.list

sed -i 's/$/\/1/g' \*list 给每行后面添加一个/1

* + - * 回原始数据（路径见表格）找fq

/share/app/seqtk/1.3/seqtk subseq /zfssz8/CNGB\_DATA/BGISEQ01/DIPSEQ/DIPSEQT1/P20Z10200N0041\_Temp/SGI75460075/221125\_SEQ123\_DP8450007087BR\_L01\_DSH-cfRNA-lib7003/DP8450007087BR\_L01\_43\_1.fq.gz /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/hill/test\_stringtie/test\_tpm/SGI75460075\_star\_name.list > /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/hill/test\_stringtie/test\_tpm/SGI75460075\_star\_target\_1.fq

/share/app/seqtk/1.3/seqtk subseq /zfssz8/CNGB\_DATA/BGISEQ01/DIPSEQ/DIPSEQT1/P20Z10200N0041\_Temp/SGI75460075/221125\_SEQ123\_DP8450007087BR\_L01\_DSH-cfRNA-lib7003/DP8450007087BR\_L01\_43\_2.fq.gz /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/hill/test\_stringtie/test\_tpm/SGI75460075\_star\_name.list > /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/hill/test\_stringtie/test\_tpm/SGI75460075\_star\_target\_2.fq

for NAMELIST in `ls -1 /jdfssz1/ST\_HEALTH/P20Z10200N0041/zhangyifan1/20230221\_RNAlength/fangzhen20230222/clean\_50/bam/`

do

type=`echo $NAMELIST | awk -F '.' '{print $1}'`

for id in `cat /jdfssz1/ST\_HEALTH/P20Z10200N0041/zhangyifan1/20230221\_RNAlength/fangzhen20230222/clean\_50/clean\_50\_${type}\_name.list`

do

grep -A 3 $id "/hwfssz5/ST\_HEALTH/P20Z10200N0041/zhangyifan1/07\_similation/01\_jianyue/2ban\_data/data/final/50/50\_merge\_0220.r1.fq" >> /jdfssz1/ST\_HEALTH/P20Z10200N0041/zhangyifan1/20230221\_RNAlength/fangzhen20230222/clean\_50/fq/50\_clean\_${type}.r1.fq

done

done

* 1. 画直方图

统计fastq文件中reads长度

for rnatype in lncRNA miRNA miscRNA Mt\_tRNA Mt\_rRNA other protein\_coding pseudo rRNA scRNA snoRNA snRNA tRNA

do

awk 'NR>1 && (NR-2)%4==0{print length}' /jdfssz1/ST\_HEALTH/P20Z10200N0041/zhangyifan1/20230221\_RNAlength/hill/SGI75460004/fq/SGI75460004\_$rnatype.fq >> /jdfssz1/ST\_HEALTH/P20Z10200N0041/zhangyifan1/20230221\_RNAlength/hill/SGI75460004/fq/length/SGI75460004\_${rnatype}\_length.txt

done

仿真数据：

|  |  |
| --- | --- |
| /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/zhuanli/no\_rRNA/result/clean\_50/step3\_hisat2/clean\_50\_accepted\_hits.sorted.bam |  |
| /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/zhuanli/result/clean\_50/step3\_hisat2/clean\_50\_accepted\_hits.sorted.bam |  |
| /jdfssz1/ST\_HEALTH/P20Z10200N0041/Wangyingying/cfRNA/zhuanli/no\_rRNA/result/complex\_50\_40T/step3\_hisat2/complex\_50\_40T\_accepted\_hits.sorted.bam |  |