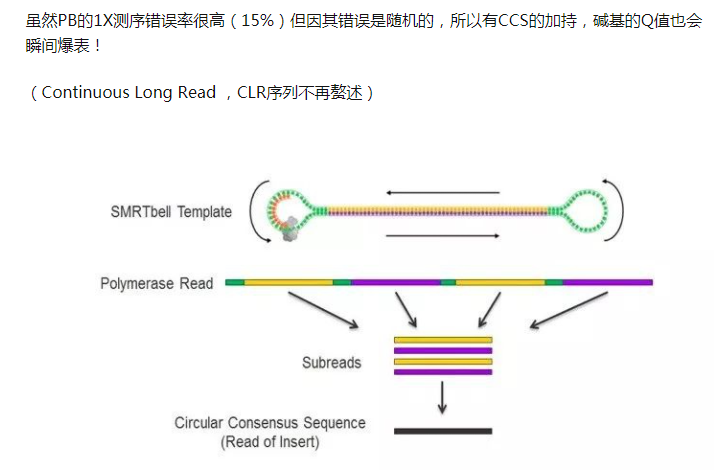


在收集mRNA时，通常用磁珠连接的oligodT吸附polyA达到收集的目的。为了转化成cDNA，需要设计上下游引物，以mRNA为模板扩增为cDNA，这是3’和5’引物的用途。然后在含引物的两端连接单链发卡状的接头，构建文库。

原始数据的每一个subreads的组成：5’引物--转录本--polyA--3’引物。



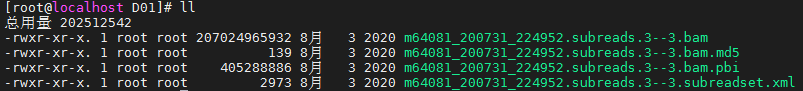
CCS步骤：聚类相近的共有序列，设置NumPasses阈值筛选subreads。

Lima步骤：对所有的reads去除引物和barcode，去除嵌合序列(转录组文库在构建过程中可能会产生嵌合体，即同一个ZMW中两个转录本嵌合到一起)。，把序列的正负链整合一致。我们的测序没有barcode。

Cluster步骤：去除polyA，去除连环结构的序列。层级聚类。

Polish步骤：利用arrow-model，将多个subreads合并成一个read，被合并的序列应该是属于同一个cDNA的subreads。

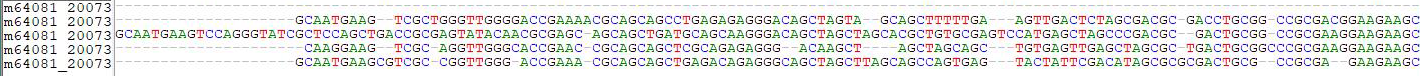
Polish完成后输出高质量和低质量序列，后续需要校正。



上图是原始数据的目录。打开m64081\_200731\_224952.subreads.3--3.bam。



上图是polyA和3’-primer.



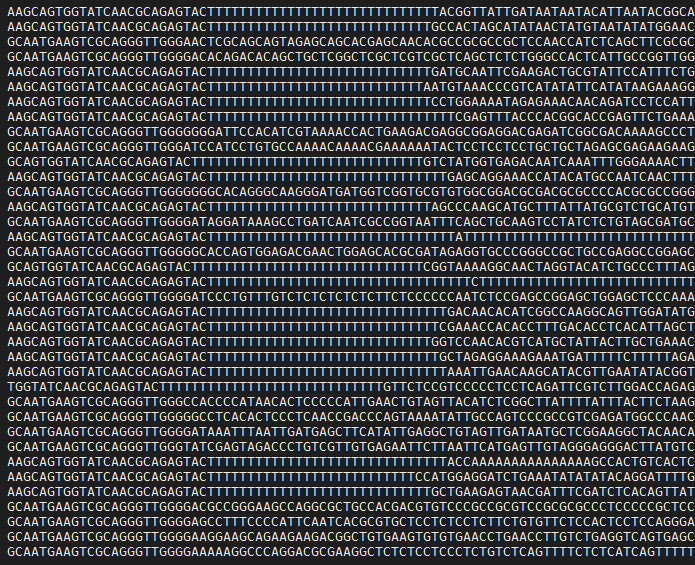
上图是G-cap和5-primer.

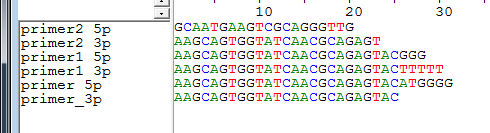
可见在原始数据中m64081\_200731\_224952.subreads.3--3.bam引物序列都存在测序的错误的，因此判断这些序列没有经过CCS处理。

使用CCS对原始数据m64081\_200731\_224952.subreads.3--3.bam进行处理：

ccs m64081\_200731\_224952.subreads.3--3.bam LL.ccs.bam --min-passes 1 --min-length 50 -j 25

获得了输出文件LL.ccs.bam。打开head查看：

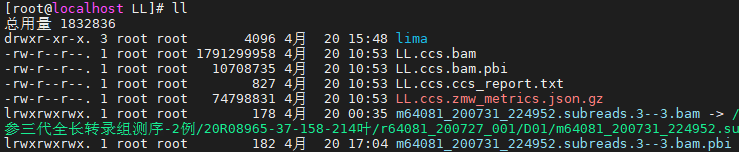




结合公司给的primer判断CCS处理是必要的。

经过CCS处理后引物序列正确且完整，这样才符合下一步Lima去除引物的条件。

CCS步骤完成输出如下文件：



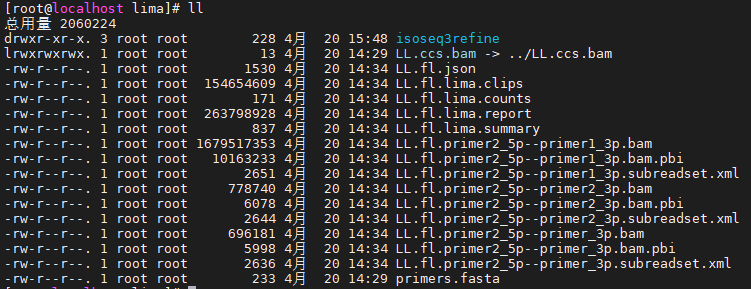
创建lima/

下一步将使用lima去除引物，并统一序列的正负链。

lima LL.ccs.bam primers.fasta LL.fl.bam --isoseq -j 25

输出了三对引物分别的bam。但另外两个文件很小不到1Mb。

lima步骤完成输出如下文件：



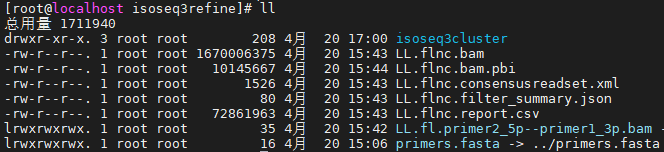
创建isoseq3refine/

samtools merge LL.fl.5-3.bam LL.fl.primer2\_5p--primer1\_3p.bam LL.fl.primer2\_5p--primer2\_3p.bam LL.fl.primer2\_5p--primer\_3p.bam -@ 25

创建isoseq3refine/

isoseq3 refine --require-polya LL.fl.5-3.bam primers.fasta LL.flnc.bam -j 25

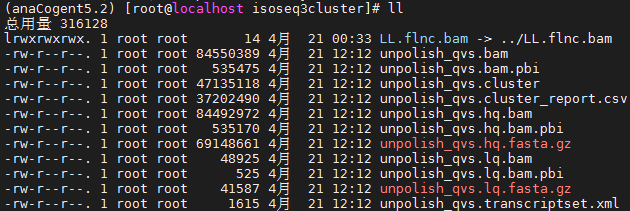
lima步骤完成输出如下文件：



创建isoseq3cluster/

isoseq3 cluster LL.flnc.bam unpolish.bam --verbose -j 25 --use-qvs

isoseq3cluster步骤完成输出如下文件：

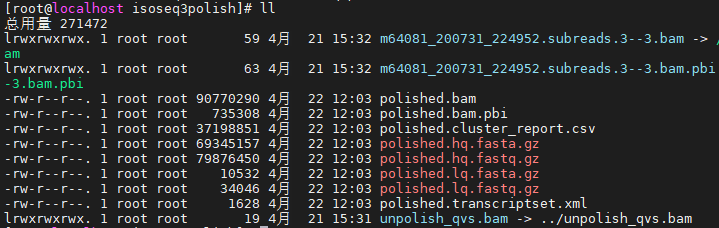


创建isoseq3polish/

isoseq3 polish unpolish.bam m64081\_200731\_224952.subreads.3--3.bam polished.bam -j 25

（在LL/里连接pbi，在isoseq3polish/里链接pbi，pbi是原始文件的pbi，这样才能运行）

isoseq3polish步骤完成输出如下文件：



创建LoRDEC/

lordec-correct -T 25 -2 R20061004\_R1.fq,R20061004\_R2.fq -k 21 -s 3 -i polished.lq.fasta -o polished.lq.corrected.fasta &> mylog.log

注意，输出的polished.lq.corrected.fasta的一条转录本序列是多行显示的，转换成一行。

#fasta转一行

awk '/^>/&&NR>1{print "";}{ printf "%s",/^>/ ? $0"\n":$0 }' test.fasta > test1.fasta

最后合并：

cat polished.hq.fasta polished.lq.corrected.fasta > LL.final.fasta

LL.final.fasta就是处理完之后最终的结果，碱基质量较高，正确率较高。