高学月

(全) WES (全外显子测序) 分析流程

(/apps/ utm_sc banner



黄晶_id (/u/060676b0fa06) (+ 关注)

2019.01.24 17:50* 字数 1139 阅读 609 评论 2 喜欢 12

(/u/060676b0fa06)

1.安装软件配置环境

先安装conda, 相当于先给我们的电脑安装一个软件管家

安装conda的方法

先去清华镜像站搜索你想安装版本的下载链接
mkdir src && cd src
wget -c https://mirrors.tuna.tsinghua.edu.cn/anaconda/miniconda/Miniconda2-4.5.12-Lir

回车后系统会自动下载,下载完成后会有一个 Miniconda2-4.5.12-Linux-x86_64.sh bash文件

bash Miniconda2-4.5.12-Linux-x86_64.sh #这样就相当于激活了Miniconda这个软件,软件安装成功。

更改镜像源配置如下:

#直接复制粘贴运行这些代码,镜像就配置好了
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/pkgs/free
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/conda
conda config --add channels https://mirrors.tuna.tsinghua.edu.cn/anaconda/cloud/biocc
conda config --set show_channel_urls yes

然后就可以用conda来安装一系列软件了

注意:安装软件前一定先搜索一下,conda有没有这个软件或者这个软件是不是你以为的名字

+

 \bigcirc

(/apps/

utm sc

banner

```
conda create -n wes python=2 bwa #创建wes环境 conda info --envs #查看当前环境命令,看看是否添加上了wes环境 source activate wes #启动当前环境 **必须保证所有的软件都是安装在 wes 环境下面** conda install sra-tools conda install samtools conda install -y bcftools vcftools snpeff conda install -y multiqc qualimap conda install -y gatk4
```

2.raw fastq

从NCBI上下载SRA数据转成fastq,此时就得到了原始的fastq数据。(下载sra数据特别慢)

```
## 找到SRA数据,下载SRR list
nohup cat SRR_Acc_List.txt|while read id; do (prefetch ${id});done &
## sra转为fastq
ls SRR5660416.sra |fastq-dump -gzip --split-3 -0 ./ SRR5660416.sra #单个测试
cat >sra.sh #写脚本
ls SRR* |while read id; do (fastq-dump -gzip --split-3 -0 ./ ${id}) 1>./${id}.sra.log
nohup bash sra.sh & #挂后台运行
```

```
drwxrwxr-x 4 jhuang jhuang 4.0K Jan 25 07:25 ../
-rw-r--r-- 1 jhuang jhuang 1.8G Jan 25 14:54 SRR5943132.sra
-rw-r--r-- 1 jhuang jhuang 1.8G Jan 25 14:54 SRR5943133.sra
-rw-rw-r-- 1 jhuang jhuang 1.4G Jan 25 15:38 SRR5943131_2.fastq.gz
-rw-rw-r-- 1 jhuang jhuang
                          97 Jan 25 15:46 sra.sh
-rw----- 1 jhuang jhuang
                          0 Jan 25 15:47 nohup.out
-rw-rw-r-- 1 jhuang jhuang
                          81 Jan 25 16:14 SRR5943132.sra.sra.log
-rw-rw-r-- 1 jhuang jhuang 1.4G Jan 25 16:14 SRR5943132_2.fastq.gz
drwxrwxr-x 2 jhuang jhuang 4.0K Jan 25 16:14 ./
-rw-rw-r-- 1 jhuang jhuang
                          81 Jan 25 16:41 SRR5943133.sra.sra.log
-rw-rw-r-- 1 jhuang jhuang 1.3G Jan 25 16:41 SRR5943133 1.fastq.gz
           jhuang jhuang 1.4G Jan 25 16:41 SRR59
```

得到原始的fastq数据

3.质控得到dean fastq

软件:

- fastqc
- cutadapt
- Trim Galore

fastqc质量报告



高学月

(/apps/

utm_sc

banner

```
mkdir 1.fastq_qc && cd 1.fastq_qc
fastqc /home/qmcui/7E5240_L1_A001.L1_1.fastq.gz /home/qmcui/7E5240_L1_A001.L1_2.fastq
```

```
(wes) vip11 20:28:11 ~/project/wes/sra/tmp
[$ ls
7E5240_L1_A001.L1_1_fastqc.html 7E5240_L1_A001.L1_2_fastqc.html
7E5240_L1_A001.L1_1_fastqc.zip 7E5240_L1_A001.L1_2_fastqc.zip
(wes) vip11 20:28:13 ~/project/wes/sra/tmp
```

fastqc得到的文件

去接头

```
mkdir 1.fastq_qc && cd 1.fastq_qc
#单个测试
trim_galore --phred33 -q 25 -e 0.1 --length 36 --stringency 3 --paired -o ~/ ../sra/t
```

我们来看下这些数字的含义:

- •去除接头
- 去除低质量碱基(-q 25)
- •最大允许错误率(默认-e 0.1)
- 去除<36的reads(--length 36)
- 切除双端的overlap>3的碱基
- 去除reads以对为单位(--paired)

```
#多个循环 trim_galore
ls /home/qmcui/*1.fastq.gz>./1;cat 1
ls /home/qmcui/*2.fastq.gz>./2;cat 2
paste 1 2 >config
cat >trim.sh
cat config|while read id;do arr=(${id}); fq1=${arr[0]}; fq2=${arr[1]}; echo $fq1 $fq2$
nohup bash mvadaptor.sh &
```

下面跑流程的时候我们用少量sample序列做分析,所以我们每一步会给它先赋值

4.比对mapping



比对目的:

- 将打断测序的reads比回参考基因组
- 得到比对结果sam文本,用于后续分析

软件:

bwa(Burrows-Wheeler Aligner对准器)bwa 是一款将序列比对到参考基因组上的软件

比对策略:

- index先建立索引
- mem算法(maximal exact matches)

bwa软件的作用是将序列比对到参考基因组上,在比对之前,首先需要对参考基因组建立索引。

#1. 建立project路径
mkdir 0.config 2.mapping
#2.建立索引
ln -s /public/biosoft/GATK/resources/bundle/hg38/Homo_sapiens_assembly38.fasta /home/
bwa index /home/jhuang/project/wes/0.config/hg38.fa
人类参考基因组 hg38.fa

索引建立好之后, 会生成5个文件, 后缀分别为 .bwt / .pac / .ann / .amb / .sa

索引建好之后我们开始用baw比对,比对结果是要生成排好序的bam,有两个策略:

- bwa生成sam转bam,然后再排序。
- bwa比对时直接利用管道符(')直接生成排序bam (推荐)

这里这两种方法都介绍,但跑的时候建议用第二种

bwa首先生成sam文件:

先赋值
sample="7E5241"
INDEX="/home/qmcui/database/reference/index/bwa/hg38" (比对基因组所放的位置)
fq1="/home/qmcui/7E5239.L1_1_val_1.fq.gz"
fq2="/home/qmcui/7E5239.L1_2_val_2.fq.gz"
(这里的的fq1、fq2是经过过滤得到的fastq文件)

**单个样品比对
bwa mem -t 5 -R "@RG\tID:\$sample\tSM:\$sample\tLB:WGS\tPL:Illumina" \$INDEX \$fq1 \$fq2 :
##`1>7E5241.sam 2>7E5241.log`条命令是说正确的运行结果放入命名为7E5241.sam的文档,错误的放

再将上步生成的sam文件转成bam

bam文件是二进制文件,占内存小,一般存bam删sam

(/apps/ utm_sc banner

章 高学月

+

 \bigcirc

≪

(/apps/ utm_sc

banner

```
## sam转bam
samtool view -bS -h 7E5241.sam > 7E5241.bam
或者
samtool view -bS -h 7E5241.sam -o 7E5241.bam
```

bam查看方法

```
samtool view -h smaple.bam |less -S
或者
samtool view smaple.bam |less -S
```

bam文件排序 (sort)

sort目的:

- •用于后续软件分析
- 提取bam序列

软件: samtools(sort参数)

```
## 建立路径
mkdir 3.sort_bam && cd 3.sort_bam
## 排序bam
samtools sort -@ 5 ../2.mapping/7E5241.bam -o 7E5241.sort.bam
```

下面介绍一步到位生成sort.bam

```
## 先赋值
sample="7E5241"
INDEX="/home/qmcui/database/reference/index/bwa/hg38" (比对基因组所放的位置)
fq1="/home/qmcui/7E5239.L1_1_val_1.fq.gz"
fq2="/home/qmcui/7E5239.L1_2_val_2.fq.gz"
##
bwa mem -t 5 -R
"@RG\tID:$sample\tSM:$sample\tLB:WGS\tPL:Illumina" $INDEX
$fq1 $fq2 | samtools sort -@ 5 -o $sample.sort.bam -
### bam统计
*单个测试
samtools flagstat /home/qmcui/7E5241.chr1_2.sort.bam
*多个
ls *.bam | while read id ;do (samtools flagstat $id > $(basename$id ".bam").stat);dor
```

5.Mark PCR重复

此步的目的:

- 标记/删除PCR重复的reads
- · 为后续call变异位点增加可信度, 去掉假阳性

软件: GATK4(MarkDuplicates)

ಹ

+

```
mkdir 4.gatk_markdup && cd 4.gatk_markdup

sample="7E5241"
gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" MarkDuplicates -I /home/qmcui/7E524
```

(/apps/ utm_sc banner

6.找变异

目的:

• 得到vcf前矫正的bam步骤

参数:

- sample赋值
- -I 输入文件
- -O 输出文件
- 1> 重定向标准输出
- 2> 重定向错误输出到1的文

件内(目的:记录程序运行日志)

软件:

- GATK 子命令FixMateInformation
- samtools 子命令index 建索引

6.1标记flagstat

```
### 建立路径
mkdir 6.gatk_bam && cd 6.gatk_bam
### 找变异
sample="7E5241"
gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" FixMateInformation -I ${sample}_mar
samtools index ${sample}_marked_fixed.bam
```

6.2找变异

```
sample="7E5241"
ref="/home/qmcui/database/reference/index/hisat/hg38/hg38.fa"
indel="/home/qmcui/tmp/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz"
snp="/home/qmcui/dbsnp_146.hg38.vcf.gz"

gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" BaseRecalibrator -R $ref -I ${samp}.
```

6.3矫正bam

```
sample="7E5241"
ref="/home/qmcui/database/reference/index/hisat/hg38/hg38.fa"
indel="/home/qmcui/tmp/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz"
gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" ApplyBQSR -R $ref -I ${sample}_mark
```



高学月

+

 \bigcirc

ಹ

(/apps/ utm_sc

banner

6.4calling variation得到vcf

```
mkdir gatk_single_vcf && cd gatk_single_vcf
ref="/home/qmcui/database/reference/index/hisat/hg38/hg38.fa"
"snp="/home/qmcui/dbsnp_146.hg38.vcf.gz
gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" HaplotypeCaller -R $ref -I ${sample}
```

得到_raw.vcf文件找变异这步就算成功

6.5 得到gvcf文件

3个样本循环 (一个样本1.5h的耗时) 所以把它挂后台运行吧 (nohup) 。

参数:

- ERC 指定类型
- •-R 参考基因组
- -I 输入的bam(bqsr bam)
- dbsnp数据库
- -O 输出文件

软件: GATK

```
## sample是不同的样本,可以把样本写入一个txt文档中
cat bq_bam.txt
7E5239
7E5240
7E5241
^C
cat >bq_bam.sh
ref="/home/qmcui/database/reference/index/hisat/hg38/hg38.fa"
snp="/home/qmcui/dbsnp_146.hg38.vcf.gz"
cat bq_bam.txt| while read sample
do gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" HaplotypeCaller -ERC GVCF -R $redone
nohup bash bq_bam.sh &
samtools index ${sample}.bqsr.bam # 核查bam是否建索引
```

6.6合并上步生成的gvcf, 按染色体来找变异

高学月

^



+



&

参数:

- •-L 区域(可多个)
- •-R 参考基因组
- •-V 一个样本的gvcf
- •-V 再一个, 可追加
- --genomicsdb先生成db文件

软件:

- GATK
- GenomicsDBImport合并gvcf
- GenotypeGVCFs工具对gvcf文件进行 joint genotyping

#先将染色体按条数分开
seq 22|while read id;do echo chr\${id};done >bed.txt
cat >>bed.txt
chrX
chrY
chrM
#合并样本的gvcf
ref="/home/qmcui/database/reference/index/hisat/hg38/hg38.fa"
snp="/home/qmcui/dbsnp_146.hg38.vcf.gz"
cat bed.txt|while read bed;do gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" Genor gatk --java-options "-Xmx20G -Djava.io.tmpdir=./" GenotypeGVCFs \ -R \$ref -V gendb:/,
#合并gvcf
gatk MergeVcfs -I final_chr2.vcf -I final_chr1.vcf -O raw.combine.vcf.gz

7.过滤

由raw vcf经过滤得到HQ vcf

过滤流程

7.1 vcf snp过滤

(/apps/ utm_sc banner

高学月



目的:

• 过滤snp的vcf

参数:

- -select-type 筛选类型
- -V 输入文件
- --filter-expression过滤参数
- -I 输入文件
- •-0 输出
- exclude-filtered true: 过滤

非pass

软件:

SelectVariants 还可以设置---min-indel-size 2、--max-indel-size 5 -R \$ref

mkdir 7.vcf_filter && cd 7.vcf_filter
mv ../raw.combine.vcf.gz ./

##意思就是把上步生成的合并的gvcf文件`raw.combine.vcf.gz `放入到当前目录下。

##提取snp为单独文件,

gatk SelectVariants -select-type SNP -V raw.combine.vcf.gz -O raw.snp.vcf.gz #设置过滤参数进行过滤

-filter-expression "QUAL < 30.0 || QD < 2.0 || MQ < 40.0 || FS > 60.0 || SOR > 3.0

gatk VariantFiltration -V raw.snp.vcf.gz --filter-expression "QUAL < 30.0 || QD < 2.6

(/apps/ utm_somer

青 高学月

标记过滤参数下要被过滤及合格的序列

#过滤

#`--exclude-filtered true`过滤掉非pass的

gatk SelectVariants --exclude-filtered true -V filter.snp.vcf.gz -O filtered.snp.vcf

zless -S filter.indel.vcf.gz|grep -v '^#'|cut -f 7 |sort|uniq -c zless -S filtered.snp.vcf.gz|grep -v '^#'|cut -f 7 |sort|uniq -c #查看过滤前和过滤后文件的内容,第七列标记了pass或是要被过滤的信息

%

+

(/apps/

过滤前后序列数量比较

7.2 vcf indel过滤

步骤与snp过滤相同,过滤参数不同

```
gatk SelectVariants -select-type INDEL -V raw.combine.vcf.gz -O raw.indel.vcf.gz gatk VariantFiltration -V raw.indel.vcf.gz --filter-expression "QUAL < 30.0 | QD < 2 --filter-name "Filter" -O filter.indel.vcf.gz gatk SelectVariants --exclude-filtered true -V filter.indel.vcf.gz \ -O filtered.indel.vcf.gz #过滤后的snp与indel合并(可以不合并)gatk MergeVcfs -I filtered.snp.vcf.gz -I filtered.indel.vcf.gz -O combine.filtered.vc
```

8.ANNOVAR注释

目的:

• 注释vcf

参数:

- convert2annovar perl脚本生成注释需要文件
- annotate_variation 注释

软件:

• annovar 的perl脚本

8.1输入文件格式转换

```
# convert2annovar.pl 可将多种格式转为ANNOVAR使用.avinput格式;
# -format vcf4 表明输入文件`filtered.indel.vcf.gz`是vcf格式的内容
# -allsample 将输入filtered.indel.vcf.gz中3个样本按每个样本输出到单一的.aviput文件
dir=/home/qmcui/software/ANNOVAR/annovar/
$dir/convert2annovar.pl -format vcf4 -allsample filtered.indel.vcf.gz -outfile anno
```

8.2 ANNOVAR注释功能



高学历



(/apps/ utm_sc banner

高学月

小礼物走一走,来简书关注我

赞赏支持



▮被以下专题收入,发现更多相似内容

+ 收入我的专题

(/p/61f2e7a765bd?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommenc 老板亲自传授的文献阅读方法 (/p/61f2e7a765bd?utm_campaign=maleski...

前言 论一个气场合适的导师的重要性。想起一个微信朋友前截图 A: 导师牛不牛,师兄师姐牛不牛,跟你牛

不牛没有关系 B: 你错啦,导师牛不牛,师兄师姐牛不牛,很大程度上决定我牛不牛。 演示招式 其实有时...

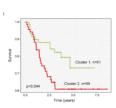
꼐 春卷00 (/u/6323855c2fa2?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

(/apps/ utm sc banner

高学历

(/p/c5fb82f91e08?



utm campaign=maleskine&utm content=note&utm medium=seo notes&utm source=recommenc 对方向扔了一个赤果果的生信分析SCI思路 (/p/c5fb82f91e08?utm_campai...

作者: 白介素2 很久没有写稿子了, 今天给大家分享一篇纯生信的SCI,思路简单粗暴。 文章2017年发表在 Oncotarget上(虽然已经牺牲了),但是我们可以取其精华呀,只要还有值得我们学习的地方,我们就挖掘出...

👰 白介素2 (/u/3019130068f1?

utm campaign=maleskine&utm content=user&utm medium=seo notes&utm source=recommen

(/p/2055db183907?



utm campaign=maleskine&utm content=note&utm medium=seo notes&utm source=recommenc RNA-seq 数据分析最佳实战(综述) (/p/2055db183907?utm_campaign=...

知识的学习没有一蹴而就,没有捷近,扎实的学习是唯一的捷近。 一篇RNA-seq分析流程的综述,全面而详 细!深度好文,可用来反复阅读。初学者用于把握RNA-seq真个流程及各个流程选择上的差异。已经开始...



dandanwu90 (/u/4467fc299229?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

(/p/ce51b519767f?

utm campaign=maleskine&utm content=note&utm medium=seo notes&utm source=recommenc 零代码也能发3分的SCI? 请收下这个套路 (/p/ce51b519767f?utm_campai...

作者: 白介素2 经常有小伙伴跟我抱怨, 没时间没精力去学代码, 太复杂了。当 然也有些小伙伴确实打起精神开始学,从网上搜了一大堆R语言资料,和生物...



🤼 白介素2 (/u/3019130068f1?

utm campaign=maleskine&utm content=user&utm medium=seo notes&utm source=recommenc

(/p/e209219eccb0?



nic profiling identifies expressed genes associat nmed cell death of nucell 10 biloba L.

+

 \bigcirc

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommend 转录组+? 分析+? 实验=2区文章 (/p/e209219eccb0?utm_campaign=male...

随着高通量测序技术的发展,越来越多的研究者或多或少的做了一些测序项目,其中尤以转录组测序类最

多。 今天小编就给大家带来一篇2月28日发表在BMC Plant Biology(影响因子3.93,中科院分区二区)上...



utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

(/apps/ utm sc banner

(/p/d2b226b14799?



utm campaign=maleskine&utm content=note&utm medium=seo notes&utm source=recommenc 《釜山行》观后感 (/p/d2b226b14799?utm_campaign=maleskine&utm_c...

七月二十日在韩国上映的电影《釜山行》,首映就突破了记录。那时身边的很多人都在向我推荐,我都没空 去看,趁着这次国庆放假,我找来了资源打算慢慢观影。晚上一个人在电脑房里观看,本以为是部很吓人...

★ 汝辈本浴火 (/u/fa915dcf83cb?

utm campaign=maleskine&utm content=user&utm medium=seo notes&utm source=recommen

我还是想成为这样的老师 (/p/cc9eb9d9fc63?utm_campaign=maleskine&...

期中考试已结束,我所带的班级考试结果不尽人意。 在教学上,我始终暗示自己,不争名夺利,做好自己的 本职工作。然而,我又不得不反省自己的教学方式,我现在的这种方式真的适合这些学生吗?我到底该如...

宓美人 (/u/5c80166a0bee?

utm campaign=maleskine&utm content=user&utm medium=seo notes&utm source=recommenc

(/p/690e6b563d39?



utm campaign=maleskine&utm content=note&utm medium=seo notes&utm source=recommenc 我为什么要打12345? (/p/690e6b563d39?utm_campaign=maleskine&ut...

柯桥区老小区改造工程如火如荼地进行着,也许过些时候真的会比原来好?但这只是也许,因为眼前,许许 多多的普通百姓遭殃受罪,年纪大的老人们,被改造引发的诸多事情而气得血压升高。我早几天为我住的4...

4 胡亚红 (/u/d1559aa1b7ff?)

utm campaign=maleskine&utm content=user&utm medium=seo notes&utm source=recommenc

(/p/433dd1d568a6?





utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommenc 在这个"成三"的浪潮中,他连成双的勇气都没有 (/p/433dd1d568a6?utm_c...

1、"风哥,下个月我结婚,记得过来喝杯喜酒。""风仔,我失恋又热恋了,你什么时候不用再喂狗粮?""小

风,在我们村,你这个年纪都在结婚生娃了,要不改天阿姨给你介绍个姑娘处处?"面对过往的同学,好友...



utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

(/apps/ utm_sc

banner

车体部焊装二区"大干六十天,奔腾向前冲"活动所记 (/p/34cbea3749a3?ut...

奔腾T77车型从上线试生产到下线仪式,再到今天的上市销售,经历了一个又一个艰难的阶段跳跃。作为奔 腾事业本部的生命之车,部里的全体员工日以继夜,为了奔腾T77的顺利量产而奋战。为了更好的激励员...

张树超_hzcj (/u/8b4dc8dd8367?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc