

Day4_Mapping_Exercise1_CORRECTED

October 6, 2023

1 Let's doing the same mapping with each sample sequenced in illumina

We have 18 samples in the SHORT_READS directory. For each sample, we want to perform all the steps to obtain the last bam used to call SNPs.

The trick here is to use a **for loop in bash**, in which all the unique commands to be carried out for a sample are grouped together. And if we have 3 samples, these commands will be executed 3 times, for example. We cans

We will use a for loop : * first with two samples * secondly with the 18 samples : 1613 1868 4752 5417 685 716 A1 A8 B1 B218 B8 C218 C2 D119 E2 E318 G11 H44

1.0.1 Before starting, remove all the files generated in the results directory

```
[6]: cd /home/jovyan/SV_DATA/RESULTS
ls
```

```
1613.flagstat          1868.flagstat
1613.onlymappedpaired.bam 1868.onlymappedpaired.bam
1613.onlymappedpaired.bam.bai 1868.onlymappedpaired.bam.bai
1613.sorted.bam        1868.onlymappedpaired.flagstat
1616_1868.bcf          1868.sorted.bam
1616_1868.vcf
```

```
[8]: rm *bam* *flagstat
ls
```

```
rm: cannot remove '*bam*': No such file or directory
rm: cannot remove '*flagstat': No such file or directory
1616_1868.bcf 1616_1868.vcf
```

1.0.2 Let's try our first for loop

```
[17]: for sample in 1613 1868
do
    echo "$sample"
done;
```

1613
1868

```
[18]: for sample in 1613 1868
do
echo "---> I can start a series of instructions for the same sample :"
echo -e "\t Here the sample **${sample}** \n"
done;
```

---> I can start a series of instructions for the same sample :
Here the sample **1613**

---> I can start a series of instructions for the same sample :
Here the sample **1868**

Gradually add each analysis step:

- mapping with bwa mem
- then running samtools flagstat
- then filtering with samtools view and samtools index

```
[ ]: for sample in 1613 1868
do
echo -e "##### ${sample} "

echo -e "#### Mapping "
# PUT BWA MEM COMMAND
bwa-mem2 mem -t 8 -R "@RG\tID:${sample}\tSM:${sample}" ../REF/GCA_002220235.
↪1_ASM222023v1_genomic.fna ../SHORT_READS/${sample}_R1.fastq.gz ../
↪SHORT_READS/${sample}_R2.fastq.gz |
    samtools sort -@ 8 -o ./${sample}.sorted.bam

echo -e "### Stat "
# PUT SAMTOOLS FLAGSTAT COMMAND
samtools flagstat ./${sample}.sorted.bam > ./${sample}.flagstat

# PUT SAMTOOLS VIEW TO FILTER BAM
echo -e "### Filtering"
samtools view -bh -@8 -f 0x02 -o ./${sample}.mappedpaired.bam ./${sample}.
↪sorted.bam

# PUT SAMTOOLS INDEX
echo -e "### Indexing bam"
samtools index ./${sample}.mappedpaired.bam

done;
```

List the content of the RESULTS directory - `ls -lht`

```
[48]: ls -lht /home/jovyan/SV_DATA/RESULTS
```

```
total 660M
drwxr-sr-x 2 jovyan users 4.0K Oct  3 20:00 3_bam_filtered
drwxr-sr-x 3 jovyan users 4.0K Oct  3 20:00 2_bam_flagstat
drwxr-sr-x 2 jovyan users 4.0K Oct  3 20:00 1_bam_raw
-rw-r--r-- 1 jovyan users  46K Oct  3 19:39 1868.mappedpaired.bam.bai
-rw-r--r-- 1 jovyan users 137M Oct  3 19:39 1868.mappedpaired.bam
-rw-r--r-- 1 jovyan users  519 Oct  3 19:39 1868.flagstat
-rw-r--r-- 1 jovyan users 158M Oct  3 19:39 1868.sorted.bam
-rw-r--r-- 1 jovyan users  46K Oct  3 19:38 1613.mappedpaired.bam.bai
-rw-r--r-- 1 jovyan users 154M Oct  3 19:38 1613.mappedpaired.bam
-rw-r--r-- 1 jovyan users  521 Oct  3 19:38 1613.flagstat
-rw-r--r-- 1 jovyan users 167M Oct  3 19:38 1613.sorted.bam
-rw-r--r-- 1 jovyan users  36M Oct  3 15:34 1616_1868.vcf
-rw-r--r-- 1 jovyan users  8.8M Oct  3 15:01 1616_1868.bcf
```

BONUS... only if you have the time to write more “complex” code... We want to group all the files generated by the same software for each sample in a single : * the `1_bam_raw` directory that contains all bam files * the `2_bam_flagstat` directory that contains all files generated by samtools flagstat * the `3_bam_filtered` directory that contains all filtered bam files

```
[ ]: ##### ADDING FILE TREE
res_dir="/home/jovyan/SV_DATA/RESULTS"
bam_raw_dir="${res_dir}/1_bam_raw"
bam_stat_dir="${res_dir}/2_bam_flagstat"
bam_filtered_dir="${res_dir}/3_bam_filtered"

data_dir="/home/jovyan/SV_DATA"
ref="${data_dir}/REF/GCA_002220235.1_ASM222023v1_genomic.fna"

cd $res_dir
mkdir -p $bam_raw_dir
mkdir -p $bam_stat_dir
mkdir -p $bam_filtered_dir

for sample in 1613 1868
do

    fq1="${data_dir}/SHORT_READS/${sample}_R1.fastq.gz"
    fq2="${data_dir}/SHORT_READS/${sample}_R2.fastq.gz"

    echo -e "#####\n Mapping ${sample} "
    # PUT BWA MEM CDE ADAPTED
```

```

bwa-mem2 mem -t 8 -R "@RG\tID:${sample}\tSM:${sample}" ${ref} ${fq1} ${fq2}
↪| samtools sort -@ 8 -o ${bam_raw_dir}/${sample}.sorted.bam

echo -e "#####\nFlagstat ${sample} "
# PUT FLAGSTAT
samtools flagstat ${bam_raw_dir}/${sample}.sorted.bam > ${bam_stat_dir}/
↪${sample}.flagstat

# PUT FILTER
echo -e "#####\n Filtering ${sample} "
samtools view -bh -@8 -f 0x02 -o ${bam_filtered_dir}/${sample}.mappedpaired.
↪bam ${bam_raw_dir}/${sample}.sorted.bam

# PUT INDEX
samtools index ${bam_filtered_dir}/${sample}.mappedpaired.bam

done;

```

List recursively the content of the RESULTS directory - ls -R

[43]: `ls -lRh /home/jovyan/SV_DATA/RESULTS`

```

/home/jovyan/SV_DATA/RESULTS:
total 660M
-rw-r--r-- 1 jovyan users 521 Oct 3 19:38 1613.flagstat
-rw-r--r-- 1 jovyan users 154M Oct 3 19:38 1613.mappedpaired.bam
-rw-r--r-- 1 jovyan users 46K Oct 3 19:38 1613.mappedpaired.bam.bai
-rw-r--r-- 1 jovyan users 167M Oct 3 19:38 1613.sorted.bam
-rw-r--r-- 1 jovyan users 8.8M Oct 3 15:01 1616_1868.bcf
-rw-r--r-- 1 jovyan users 36M Oct 3 15:34 1616_1868.vcf
-rw-r--r-- 1 jovyan users 519 Oct 3 19:39 1868.flagstat
-rw-r--r-- 1 jovyan users 137M Oct 3 19:39 1868.mappedpaired.bam
-rw-r--r-- 1 jovyan users 46K Oct 3 19:39 1868.mappedpaired.bam.bai
-rw-r--r-- 1 jovyan users 158M Oct 3 19:39 1868.sorted.bam
drwxr-sr-x 2 jovyan users 4.0K Oct 3 19:40 1_bam_raw
drwxr-sr-x 3 jovyan users 4.0K Oct 3 19:42 2_bam_flagstat
drwxr-sr-x 2 jovyan users 4.0K Oct 3 19:41 3_bam_filtered

/home/jovyan/SV_DATA/RESULTS/1_bam_raw:
total 324M
-rw-r--r-- 1 jovyan users 167M Oct 3 19:40 1613.sorted.bam
-rw-r--r-- 1 jovyan users 158M Oct 3 19:41 1868.sorted.bam

/home/jovyan/SV_DATA/RESULTS/2_bam_flagstat:
total 8.0K
-rw-r--r-- 1 jovyan users 521 Oct 3 19:40 1613.flagstat
-rw-r--r-- 1 jovyan users 0 Oct 3 19:35 1615.flagstat

```

```
-rw-r--r-- 1 jovyan users 519 Oct  3 19:41 1868.flagstat
```

```
/home/jovyan/SV_DATA/RESULTS/3_bam_filtered:
```

```
total 291M
```

```
-rw-r--r-- 1 jovyan users 154M Oct  3 19:40 1613.mappedpaired.bam
-rw-r--r-- 1 jovyan users  46K Oct  3 19:40 1613.mappedpaired.bam.bai
-rw-r--r-- 1 jovyan users 137M Oct  3 19:41 1868.mappedpaired.bam
-rw-r--r-- 1 jovyan users  46K Oct  3 19:41 1868.mappedpaired.bam.bai
```

1.0.3 Let's try our previous loop on all the samples

```
[ ]: bcftools mpileup --threads 8 -f ../REF/GCA_002220235.1_ASM222023v1_genomic.fna
↳ *paired.bam | bcftools call --threads 8 -cv -Oz -o 1613_1868.bcf | ls -lRh /
↳ home/jovyan/SV_DATA/RESULTS
```

```
[ ]: res_dir="/home/jovyan/SV_DATA/RESULTS"
bam_raw_dir="${res_dir}/1_bam_raw"
bam_stat_dir="${res_dir}/2_bam_flagstat"
bam_filtered_dir="${res_dir}/3_bam_filtered"

data_dir="/home/jovyan/SV_DATA"
ref="${data_dir}/REF/GCA_002220235.1_ASM222023v1_genomic.fna"

cd $res_dir
mkdir -p $bam_raw_dir
mkdir -p $bam_stat_dir
mkdir -p $bam_filtered_dir

for sample in 1613 1868 4752 5417 685 716 A1 A8 B1 B218 B8 C218 C2 D119 E2 E318
↳ G11 H44
do

    fq1="${data_dir}/SHORT_READS/${sample}_R1.fastq.gz"
    fq2="${data_dir}/SHORT_READS/${sample}_R2.fastq.gz"

    echo -e "#####\n Mapping ${sample} "
    # PUT BWA MEM CDE ADAPTED
    bwa-mem2 mem -R "@RG\tID:${sample}\tSM:${sample}" -t 8 ${ref} ${fq1} ${fq2}
↳ | samtools sort -@ 4 -o ${bam_raw_dir}/${sample}.sorted.bam

    echo -e "#####\nFlagstat ${sample} "
    # PUT FLAGSTAT
    samtools flagstat ${bam_raw_dir}/${sample}.sorted.bam > ${bam_stat_dir}/
↳ ${sample}.flagstat

    # PUT FILTER
    echo -e "#####\n Filtering ${sample} "
```

```

    samtools view -bh -@4 -f 0x02 -o ${bam_filtered_dir}/${sample}.mappedpaired.
↪bam ${bam_raw_dir}/${sample}.sorted.bam

# PUT INDEX
samtools index ${bam_filtered_dir}/${sample}.mappedpaired.bam

done;

```

List recursively the content of the RESULTS directory - ls -R

```
[ ]: # PUT COMMAND
```

```
[51]: ls -R
```

```

.:
1613.flagstat          1616_1868.vcf          1_bam_raw
1613.mappedpaired.bam  1868.flagstat          2_bam_flagstat
1613.mappedpaired.bam.bai 1868.mappedpaired.bam  3_bam_filtered
1613.sorted.bam        1868.mappedpaired.bam.bai
1616_1868.bcf          1868.sorted.bam

./1_bam_raw:
1613.sorted.bam  716.sorted.bam  B8.sorted.bam  E318.sorted.bam
1868.sorted.bam  A1.sorted.bam  C218.sorted.bam  G11.sorted.bam
4752.sorted.bam  A8.sorted.bam  C2.sorted.bam  H44.sorted.bam
5417.sorted.bam  B1.sorted.bam  D119.sorted.bam
685.sorted.bam  B218.sorted.bam  E2.sorted.bam

./2_bam_flagstat:
1613.flagstat  5417.flagstat  A8.flagstat  C218.flagstat  E318.flagstat
1615.flagstat  685.flagstat  B1.flagstat  C2.flagstat  G11.flagstat
1868.flagstat  716.flagstat  B218.flagstat  D119.flagstat  H44.flagstat
4752.flagstat  A1.flagstat  B8.flagstat  E2.flagstat

./3_bam_filtered:
1613.mappedpaired.bam  A1.mappedpaired.bam  C2.mappedpaired.bam
1613.mappedpaired.bam.bai  A1.mappedpaired.bam.bai  C2.mappedpaired.bam.bai
1868.mappedpaired.bam  A8.mappedpaired.bam  D119.mappedpaired.bam
1868.mappedpaired.bam.bai  A8.mappedpaired.bam.bai  D119.mappedpaired.bam.bai
4752.mappedpaired.bam  B1.mappedpaired.bam  E2.mappedpaired.bam
4752.mappedpaired.bam.bai  B1.mappedpaired.bam.bai  E2.mappedpaired.bam.bai
5417.mappedpaired.bam  B218.mappedpaired.bam  E318.mappedpaired.bam
5417.mappedpaired.bam.bai  B218.mappedpaired.bam.bai  E318.mappedpaired.bam.bai
685.mappedpaired.bam  B8.mappedpaired.bam  G11.mappedpaired.bam
685.mappedpaired.bam.bai  B8.mappedpaired.bam.bai  G11.mappedpaired.bam.bai
716.mappedpaired.bam  C218.mappedpaired.bam  H44.mappedpaired.bam
716.mappedpaired.bam.bai  C218.mappedpaired.bam.bai  H44.mappedpaired.bam.bai

```

```
[50]: pwd
```

```
/home/jovyan/SV_DATA/RESULTS
```

1.1 SNPs calling using all bams just generated

```
[ ]: # PUT COMMAND
```

```
[52]: bcftools mpileup --threads 8 -f ../REF/GCA_002220235.1_ASM222023v1_genomic.fna_
      ↪3_bam_filtered/*.bam | bcftools call --threads 8 -cv -Oz -o all_samples.bcf
```

Note: none of --samples-file, --ploidy or --ploidy-file given, assuming all sites are diploid

[mpileup] 18 samples in 18 input files

[mpileup] maximum number of reads per input file set to -d 250

```
[54]: ls -lrt 2_bam_flagstat
```

```
total 76
-rw-r--r-- 1 jovyan users  0 Oct  3 19:35 1615.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:44 1613.flagstat
-rw-r--r-- 1 jovyan users 519 Oct  3 19:45 1868.flagstat
-rw-r--r-- 1 jovyan users 517 Oct  3 19:45 4752.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:46 5417.flagstat
-rw-r--r-- 1 jovyan users 520 Oct  3 19:47 685.flagstat
-rw-r--r-- 1 jovyan users 515 Oct  3 19:47 716.flagstat
-rw-r--r-- 1 jovyan users 527 Oct  3 19:49 A1.flagstat
-rw-r--r-- 1 jovyan users 520 Oct  3 19:50 A8.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:50 B1.flagstat
-rw-r--r-- 1 jovyan users 443 Oct  3 19:50 B218.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:51 B8.flagstat
-rw-r--r-- 1 jovyan users 524 Oct  3 19:53 C218.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:54 C2.flagstat
-rw-r--r-- 1 jovyan users 524 Oct  3 19:55 D119.flagstat
-rw-r--r-- 1 jovyan users 521 Oct  3 19:56 E2.flagstat
-rw-r--r-- 1 jovyan users 525 Oct  3 19:58 E318.flagstat
-rw-r--r-- 1 jovyan users 519 Oct  3 19:58 G11.flagstat
-rw-r--r-- 1 jovyan users 525 Oct  3 20:00 H44.flagstat
-rw-r--r-- 1 jovyan users 490 Oct  3 20:18 all_stat.csv
```