Analysis Protocol

1. Pre-processing BAM files:

- Sorting:

```
samtools sort Control.bam > Control.sorted.bam
```

```
samtools sort Tumor.bam > Tumor.sorted.bam
```

The default sort by position.

- Indexing:

```
samtools index Control.sorted.bam
```

```
samtools index Tumor.sorted.bam
```

This aims to speed up future search

- Counting reads:

```
samtools view -c Control.sorted.bam
```

We have 19720171 reads

```
samtools view -c Tumor.sorted.bam
```

We have 15039503 reads

- Checking quality of reads:

- Reads that have a mapping quality > 30:

```
samtools view -c -q 30 Control.sorted.bam
```

We have 15210703 reads

```
samtools view -c -q 30 Tumor.sorted.bam
```

We have 11678731 reads

- Reads that have a mapping quality > 25

```
samtools view -c -q 25 Control.sorted.bam
```

We have 15703445 reads

```
samtools view -c -q 25 Tumor.sorted.bam
```

We have 12020882 reads

-> We can easily state that the reads are associated with high quality.

- General statistics

```
samtools flagstat Control.sorted.bam
```

99.75% mapped

0.23% singletons

10030 reads with mate mapped to a different chr (with mapQ>=5)

```
samtools flagstat Tumor.sorted.bam
```

99.96% mapped

0.03% singletons

7572 reads with mate mapped to a different chr (with mapQ>=5)

- Detailed statistics:

```
samtools stats Control.sorted.bam > stat.control.txt
```

```
less stat.control.txt
```

Average quality: 30.4

Pairs with other orientation: 3280

Insert size standard deviation: 79.4 with an average insert size of 235.5

```
samtools stats Tumor.sorted.bam > stat.tumor.txt
```

```
less stat.tumor.txt
```

Average quality: 31.4

Pairs with other orientation: 2586

Insert size standard deviation: 75.3 with an average insert size of 216.1

- Explore coverage statistics

Single base sum coverage per region (samtools bedcov is used to have an idea on how many reads of the .bam map on the region kept in the .bed file):

```
samtools bedcov Captured_Regions.bed Control.sorted.bam > BEDCov.Control.CR.txt
```

```
samtools bedcov Captured_Regions.bed Tumor.sorted.bam > BEDCov.Tumor.CR.txt
```

Captured_Regions.bed: contains info on the chromosomes: 15, 16, 17, and 18. Includes the regions that have been selected for this experiment.

```
less BEDCov.Control.CR.txt
```

```
less BEDCov.Tumor.CR.txt
```

Output the number of reads mapped within the selected region of the genome.

2. Realignment

- Create the .intervals (contains which regions can be realigned):

```
java -jar ../tools/GenomeAnalysisTK.jar -T RealignerTargetCreator
```

```
-R ../Annotations/human_g1k_v37.fasta -I Control.sorted.bam -o
realigner.intervals.Control -L Captured_Regions.bed
```

```
java -jar ../tools/GenomeAnalysisTK.jar -T RealignerTargetCreator
-R ../Annotations/human_g1k_v37.fasta -I Tumor.sorted.bam -o
realigner.intervals.Tumor -L Captured_Regions.bed
```

-> -R indicates the path of the reference (fasta of the human genome)

- Perform the realignment:

```
java -jar ../tools/GenomeAnalysisTK.jar -T IndelRealigner -R
../Annotations/human_g1k_v37.fasta -I Control.sorted.bam
-targetIntervals Control.realigner.intervals -o
Control.sorted.realigned.bam -L Captured_Regions.bed
```

```
java -jar ../tools/GenomeAnalysisTK.jar -T IndelRealigner -R
../Annotations/human_g1k_v37.fasta -I Tumor.sorted.bam
-targetIntervals Tumor.realigner.intervals -o
Tumor.sorted.realigned.bam -L Captured_Regions.bed
```

All the positions in which there is a hidden indel are now realigned

- Count how many reads were realigned

Using the OC tag that is used to maintain the original CIGAR

```
samtools view Tumor.sorted.realigned.bam | grep OC | wc -l
```

Number of reads that have been realigned: 2267

```
samtools view Control.sorted.realigned.bam | grep OC | wc -l
```

Number of reads that have been realigned: 3158

3. Quality control -> Recalibration

- BaseRecalibrator: model the error modes and recalibrate qualities.

Its inputs are the realigned BAM file and the hapmap with all known SNPs. We limit the recalibration within the selected regions (.bed file).

```
java -jar ../tools/GenomeAnalysisTK.jar -T BaseRecalibrator -R
../Annotations/human_g1k_v37.fasta -I
../Realignment/Control.sorted.realigned.bam -knownSites
../Annotations/hapmap_3.3.b37.vcf -o recal.table.Control -L
Captured_Regions.bed
```

```
java -jar ../tools/GenomeAnalysisTK.jar -T BaseRecalibrator -R
../Annotations/human_g1k_v37.fasta -I
../Realignment/Tumor.sorted.realigned.bam -knownSites
../Annotations/hapmap_3.3.b37.vcf -o recal.table.Tumor -L
Captured_Regions.bed
```

- PrintReads: write recalibrated data to a BAM file. Original qualities are retained with the OC flag.

```
java -jar ../tools/GenomeAnalysisTK.jar -T PrintReads -R
../Annotations/human_g1k_v37.fasta -I
../Realignment/Control.sorted.realigned.bam -BQSR
recal.table.Control -o Control.sorted.realigned.recalibrated.bam
-L Captured_Regions.bed --emit_original_quals
```

```
java -jar ../tools/GenomeAnalysisTK.jar -T PrintReads -R
../Annotations/human_g1k_v37.fasta -I
../Realignment/Tumor.sorted.realigned.bam -BQSR recal.table.Tumor
-o Tumor.sorted.realigned.recalibrated.bam -L
Captured_Regions.bed --emit_original_quals
```

- The process is repeated to build the after model to evaluate remaining error

```
java -jar ../Tools/GenomeAnalysisTK.jar -T BaseRecalibrator -R
../Annotations/human_g1k_v37.fasta -I
../Realignment/Control.sorted.realigned.bam -knownSites
../Annotations/hapmap_3.3.b37.vcf -BQSR recal.table.Control -o
after_recal.table.Control -L Captured_Regions.bed
```

```
java -jar ../Tools/GenomeAnalysisTK.jar -T BaseRecalibrator -R
../Annotations/human_g1k_v37.fasta -I Tumor.sorted.realigned.bam
-knownSites ../Annotations/hapmap_3.3.b37.vcf -BQSR
recal.table.Tumor -o after_recal.table.Tumor -L
Captured_Regions.bed
```

- AnalyzeCovariates: before and after plots are made based on recalibration tables.

```
java -jar ../Tools/GenomeAnalysisTK.jar -T AnalyzeCovariates -R
../Annotations/human_g1k_v37.fasta -before recal.table.Control
-after after_recal.table.Control -csv recal.Control.csv -plots
recal_Control.pdf
```

```
java -jar ../Tools/GenomeAnalysisTK.jar -T AnalyzeCovariates -R
../Annotations/human_g1k_v37.fasta -before recal.table.Tumor
-after after_recal.table.Tumor -csv recal.Tumor.csv -plots
recal_Tumor.pdf
```

4. Deduplication

Marking the duplicates can be done in two different ways:

- MarkDuplicates from Picard, the golden standard.
- *markdup* from samtools (requires the addition of mate tags to the BAM file through the fixmate command)

The samtools command is faster than the Picard one as it exploits the Cigar of the mate read to correct with a simple iteration. However the Picard command retains more reads because samtools' command removes all the reads that have a mate mapped to a different chromosome, removing in this way structural variants.

- Use Picard MarkDuplicates:

```
java -jar ../../Tools/picard.jar MarkDuplicates
```

```
I=Control.sorted.bam O=Control.sorted.dedup.bam

REMOVE_DUPLICATES=true TMP_DIR=/tmp
METRICS_FILE=Control.picard.log ASSUME_SORTED=true

samtools index Control.sorted.dedup.bam
```

The number total reads: 21355813 -> the percentage of duplicates where 13.8%

```
java -jar ../../Tools/picard.jar MarkDuplicates
I=Tumor.sorted.bam O=Tumor.sorted.dedup.bam
REMOVE_DUPLICATES=true TMP_DIR=/tmp METRICS_FILE=Tumor.picard.log
ASSUME_SORTED=true

samtools index Tumor.sorted.dedup.bam
```

The number total reads: 18084804 -> the percentage of duplicates where 12.2%

-> REMOVE_DUPLICATES as true because we want to remove them

-> METRICS_FILE -> File to write duplication metrics on

5. Somatic Copy Number Calling

- Pileup + calculating coverage

```
samtools mpileup -q 1 -f ../Annotations/human_g1k_v37.fasta
Control.sorted.realigned.recalibrated.debup.bam
Tumor.sorted.realigned.recalibrated.debup.bam | java - jar
../Tools/VarScan.v2.3.9.jar copynumber --output-file SCNA
--mpileup 1
```

- Calculating copy numbers

```
java -jar ../Tools/VarScan.v2.3.9.jar copyCaller SCNA.copynumber
--output-file SCNA.copynumber.called
```

5237 regions are tagged as amplified (log2 > 0.25)

34757 regions are tagged as copy neutral

121880 regions are tagged as deleted (log2 < -0.25)

376 regions are tagged as homozygous deletion

- CNA, segmentation and visualization

```
Rscript CBS.R
```

6. Variant Calling

Available tools to perform this task are:

- bcftools
- GATK

- bcftools

```
bcftools mpileup -Ou -a DP -f
../../Annotations/human_g1k_v37.fasta Control.sorted.bam |
bcftools call -Ov -c -v > Control.BCF.vcf
```

```
bcftools mpileup -Ou -a DP -f
../../Annotations/human_g1k_v37.fasta Tumor.sorted.bam | bcftools
call -Ov - c -v > Tumor.BCF.vcf
```

- GATK

```
java -jar ../../Tools/GenomeAnalysisTK.jar -T UnifiedGenotyper -R
../../Annotations/human_g1k_v37.fasta -I Control.sorted.bam -o
Control.GATK.vcf -L chr20.bed
```

```
java -jar ../../Tools/GenomeAnalysisTK.jar -T UnifiedGenotyper -R
../../Annotations/human_g1k_v37.fasta -I Tumor.sorted.bam -o
Tumor.GATK.vcf -L chr20.bed
```

The output of the analysis is a VCF file.

- Analysis of VCF files

- Filtering on the quality >= 20 (BCF)

```
vcftools --minQ 20 --min-meanDP 30 --remove-indels --vcf
Control.BCF.vcf --out Control.BCF --recode --recode INFO-all
```

After filtering, 11085 out of 18497

```
vcftools --minQ 20 --min-meanDP 30 --remove-indels --vcf
Tumor.BCF.vcf --out Tumor.BCF --recode --recode INFO-all
```

After filtering, 8453 out of 17179.

Includes only sites with mean depth values greater than or equal to the "--min-meanDP" value and less than or equal to the "--max-meanDP" value. These options require that the "DP" FORMAT tag is included for each site.

- Filtering on the quality >= 20 (GATK)

```
vcftools --minQ 20 --min-meanDP 30 --remove-indels --vcf
Control.GATK.vcf --out Control.GATK --recode -- recode-INFO-all
```

After filtering, 9034 out of 9929

```
vcftools --minQ 20 --min-meanDP 30 --remove-indels --vcf
Tumor.GATK.vcf --out Tumor.GATK --recode -- recode-INFO-all
```

After filtering, 7884 out of 9598

- Operate the comparison between files

```
vcftools --vcf Control.BCF.recode.vcf --diff
Control.GATK.recode.vcf --diff-site
```

```
vcftools --vcf Tumor.BCF.recode.vcf --diff Tumor.GATK.recode.vcf
--diff-site
```

```
vcftools --vcf Control.BCF.recode.vcf --diff Tumor.BCF.recode.vcf
--diff-site
```

```
vcftools --vcf Control.GATK.recode.vcf --diff
Tumor.GATK.recode.vcf --diff-site
```

7. Variant Annotation

- SNPEff

-BCF

```
java -Xmx4g -jar ../Tools/snpEff/snpEff.jar -v hg19kg
../05_VariantCalling/Data/Control.BCF.recode.vcf - s
Control.BCF.recode.ann.html > Control.BCF.recode.ann.vcf
```

-GATK

```
java -Xmx4g -jar ../Tools/snpEff/snpEff.jar -v hg19kg
../05_VariantCalling/Data/Control.GATK.recode.vcf -s
Control.GATK.recode.ann.html > Control.GATK.recode.ann.vcf
```

About 50% of calls in the intronic regions.

About 30% of calls in the exonic regions.

Other calls are in the regulatory regions.

- SnpSift

- Annotation with Hapmap

```
java -Xmx4g -jar ../Tools/snpEff/SnpSift.jar Annotate
../Annotations/hapmap_3.3.b37.vcf Control.BCF.recode.ann.vcf >
Control.BCF.recode.ann2.vcf
```

```
java -Xmx4g -jar ../Tools/snpEff/SnpSift.jar Annotate
../Annotations/hapmap_3.3.b37.vcf Control.GATK.recode.ann.vcf >
Control.GATK.recode.ann2.vcf
```

- Annotation with Clinvar

```
java -Xmx4g -jar ../Tools/snpEff/SnpSift.jar Annotate
../Annotations/clinvar_Pathogenic.vcf Control.BCF.recode.ann2.vcf
> Control.BCF.recode.ann3.vcf
```

```
java -Xmx4g -jar ../Tools/snpEff/SnpSift.jar Annotate
../Annotations/clinvar_Pathogenic.vcf
Control.GATK.recode.ann2.vcf > Control.GATK.recode.ann3.vcf
```

- Filtering most significant variants

- Search for high impact variants with good depth coverage

```
cat Sample.BCF.recode.ann3.vcf | java -Xmx4g -jar
../Tools/snpEff/SnpSift.jar filter "(ANN[ANY].IMPACT = 'HIGH') &
(DP > 20) & (exists ID)"
```

```
cat Sample.GATK.recode.ann3.vcf | java -Xmx4g -jar
../Tools/snpEff/SnpSift.jar filter "(ANN[ANY].IMPACT = 'HIGH') &
(DP > 20) & (exists ID)"
```

- Search for clinical significance (pathogenic variants)

```
cat Sample.BCF.recode.ann3.vcf | java -Xmx4g -jar
../Tools/snpEff/SnpSift.jar filter "(exists CLNSIG)"
```

```
cat Sample.GATK.recode.ann3.vcf | java -Xmx4g -jar
../Tools/snpEff/SnpSift.jar filter "(exists CLNSIG)"
```

-> Pathogenic variant in BRCA1, both in control and tumor samples, associated with hereditary breast and ovarian cancer syndrome

8. Somatic Variant Calling

Using Varscan

- SNPs

- Pileup

```
samtools mpileup -B -f ../Annota<ons/human_g1k_v37.fasta
Control.sorted.realigned.recalibrated.debup.bam >
Control.sorted.pileup
```

15316013 bases in pileup file

- Generation on vcf file

```
java -jar ../Tools/VarScan.v2.3.9.jar mpileup2snp Control.pileup
--p-value 0.01 --output-vcf 1 > Control.VARSCAN.vcf
```

15796 variant positions reported (15796 SNP, 0 indel)

- Filtering of vcf

```
vcftools --max-meanDP 200 --min-meanDP 5 --remove-indels --vcf
Control.VARSCAN.vcf --out Control.VARSCAN --recode
--recode-INFO-all
```

After filtering, 14936 out of a possible 15796 sites

- Comparing against previous vcf file

```
vcftools --diff Control.BCF.recode.vcf --vcf
Control.VARSCAN.recode.vcf --diff-site
```

Found 10393 sites common to both files.

- SNVs

- Pileup for tumor bam (control done previously)

```
samtools mpileup -q 1 -f ../Annotations/human_g1k_v37.fasta
Tumor.sorted.realigned.recalibrated.debup.bam >
```

```
Tumor.sorted.pileup
```

- Look for somatic variants using both Tumor and Control sample

```
java -jar ../Tools/VarScan.v2.3.9.jar somatic
Control.sorted.pileup Tumor.sorted.pileup --output-snp somatic.pm
--output-indel somatic.indel --output-vcf 1
```

14627535 positions in tumor

14622667 positions shared in normal

12480834 Reference, 12430 Germline, 3164 LOH, 289 Somatic

- Annotation

- SnpEff

```
java -Xmx4g -jar ../../Tools/snpEff/snpEff.jar -v hg19kg
somatic.pm.vcf -s somatic.pm.vcf.html > somatic.pm.ann.vcf
```

- SnpSift filtering

```
cat somatic.pm.ann.vcf | java -Xmx4g -jar
../../Tools/snpEff/SnpSift.jar filter …
```

9. Ancestry Analysis

3 methods available:

- SMARTPCA
- EthSEQ
- fastSTRUCTURE

- EthSEQ

```
Rscript Run.R
```

10. Purity Ploidy estimation

- Pileup

```
bcftools mpileup -Ou -a DP -f ../Annotations/human_g1k_v37.fasta
Control.sorted.bam | bcftools call -Ov -c -v > Control.BCF.vcf
```

```
bcftools mpileup -Ou -a DP -f ../Annotations/human_g1k_v37.fasta
Tumor.sorted.bam | bcftools call -Ov -c -v > Tumor.BCF.vcf
```

- Filtering

```
grep -E "(^#|0/1)" Control.BCF.vcf > Control.het.vcf
```

```
grep -E "(^#|0/1)" Tumor.BCF.vcf > Tumor.het.vcf
```

- Creation of csv files

```
java -jar ../Tools/GenomeAnalysisTK.jar -T ASEReadCounter -R
../Annota<ons/human_g1k_v37.fasta -o recal.Control.csv -I
Control.sorted.realigned.recalibrated.debup.bam -sites
```

```
Control.het.GATK.vcf -U ALLOW_N_CIGAR_READS -minDepth 20
--minMappingQuality 20 --minBaseQuality 20
```

```
java -jar ../Tools/GenomeAnalysisTK.jar -T ASEReadCounter -R
../Annota<ons/human_g1k_v37.fasta -o Tumor.csv -I
Tumor.sorted.realigned.recalibrated.debup.bam -sites
Control.het.vcf -U ALLOW_N_CIGAR_READS -minDepth 20
--minMappingQuality 20 --minBaseQuality 20
```

- Run Clonet and Tpes to assess the allelic fraction

```
Rscript Clonet.R
```