# RUNNING THE PIPELINE (Part I)

Javier Perales-Patón jperales@cnio.es



Translational Bioinformatics Unit CNIO. Madrid, Spain.

Fátima Al-Shahrour

[falshahrour@cnio.es]

Elena Piñeiro-Yáñez

[epineiro@cnio.es]

Pedro Fernandes

[pfern@igc.gulbenkian.pt]



## Practical: VARIANT DETECTION DAY1 CASE STUDY

- Variant calling:
  - Detection of somatic SNV and indels
  - Detection of gene copy-number variants
- Quality Control on:
  - Sequencing data.

### Overview of the case study: Exome analysis (OVCA)



## Patient with ovarian cancer.

Whole-exome sequencing data from two samples from the patient:

- Tumour sample.
- Matched normal sample (healthy tissue) from epithelium.

**Library protocol**: Agilent <u>SureSelect V5</u> Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

### **Expected outcome:**

- ~docens germ-line variants.
- A few somatic cancer mutations (SNV, indel or CNA).

NOTE: This data was simulated and reduced in order to perform the computational analysis in 30 minutes.

### Overview of the case study: Exome analysis (OVCA)



## Patient with ovarian cancer.

Whole-exome sequencing data from two samples from the patient:

- Tumour sample.
- Matched normal sample (healthy tissue) from epithelium.

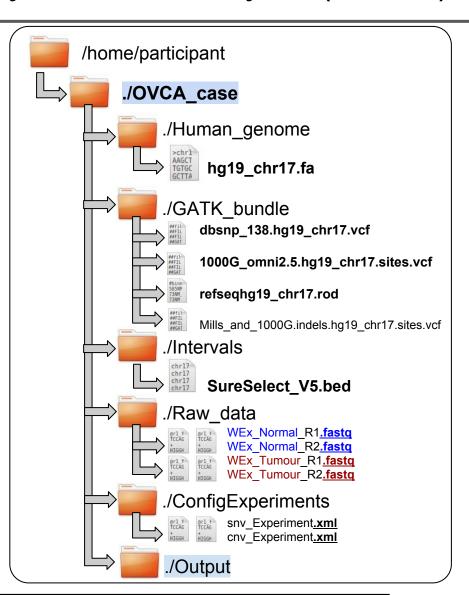
**Library protocol**: Agilent <u>SureSelect V5</u> <u>Human All Exons</u>.

Sequencing platform: HiSeq 2000 (Illumina)

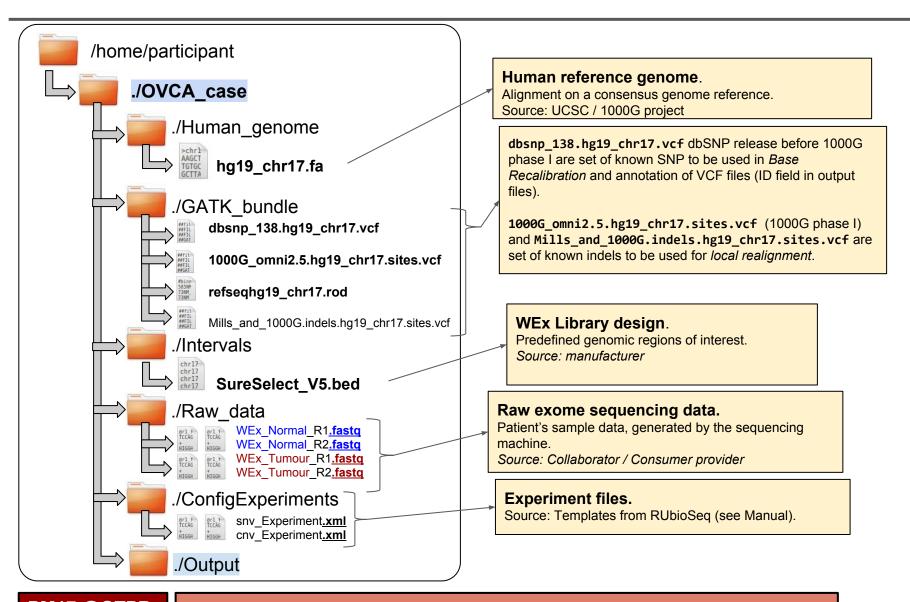
### **Expected outcome:**

- ~docens germ-line variants.
- A few somatic cancer mutations (SNV, indel or CNA).

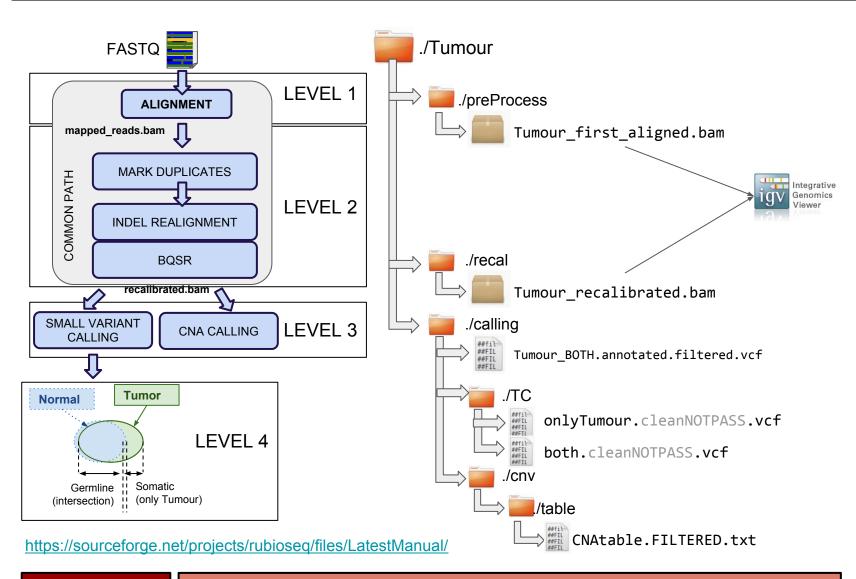
NOTE: This data was simulated and reduced in order to perform the computational analysis in 30 minutes.



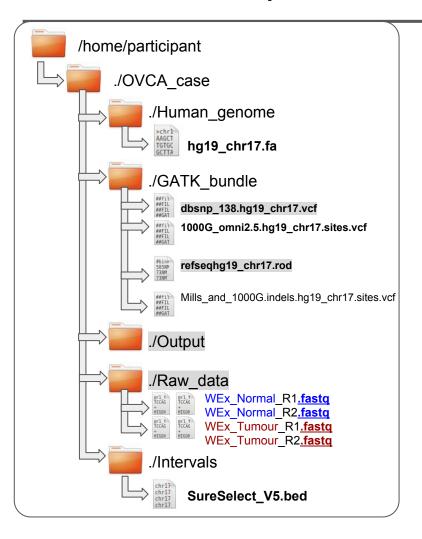
## Exome analysis (OVCA) :: Input files



## Exome analysis (OVCA) :: output files

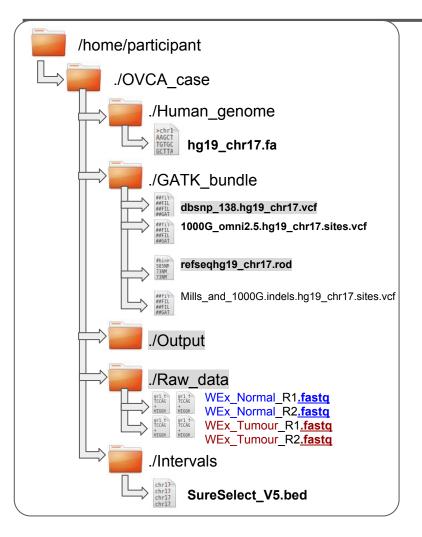


## Create the experiment file (SNV) :: ~ 30 minutes



```
<?xml version="1.0" encoding="UTF-8"?>
<!-- EXAMPLE RUBIOSEQ EXPERIMENT CONFIG FILE -->
<configData branch="SNV">
         <!-- GENOME REFERENCE PATH :: MANDATORY -->
         <GenRef> path to hg19.fa </GenRef>
         <!-- DBSNP ANNOTATION PATH :: MANDATORY -->
         <DbSnpAnnot> path to dbsnp version.vcf </DbSnpAnnot>
         <!-- 1000 Genomes ANNOTATION PATH :: MANDATORY -->
         <Genomes1000Annot> __path_to_1000G_omni2.5.hg19.sites.vcf___</Genomes1000Annot>
         <!-- REFSEQ ANNOTATION PATH :: MANDATORY -->
         <IndelAnnot> path to refseqhg19 chr17.rod </IndelAnnot>
         <!-- INTERVALS PATH :: OPTIONAL -->
         <Intervals> path to WExLibrary.bed </Intervals>
         <!-- KNOWN INDELS FOR REALIGNING :: OPTIONAL -->
         <KnownIndels> path to Mills and 1000G.indels.hg19.sites.vcf /KnownIndels>
         <!-- PLATFORM :: MANDATORY -->
         <Platform>illumina</Platform>
         <!-- checkCasava :: OPTIONAL. -->
         <checkCasava>0</checkCasava>
         <!-- OUTPUT DIRECTORY :: DEFAULT: Home directory -->
         <dirOutBase>/home/participant/OVCA case/</dirOutBase>
         <!-- PROJECT NAME :: MANDATORY -->
         <ProjectId>Output</ProjectId>
         <!-- USER NAME :: OPTIONAL(default Undefined) -->
         <UserName>participant</UserName>
         <!-- RAW DATA PATH :: MANDATORY -->
         <InDirPreProcess>/home/participant/OVCA case/Raw data/</InDirPreProcess>
         <Sample>
                  <!-- SAMPLE NAME :: MANDATORY -->
                  <SampleName>Tumor</SampleName>
                  <SampleFiles>WEx Tumour</SampleFiles>
                  <!-- SUFFIX :: MANDATORY -->
                  <SampleSuffix>.fastq</SampleSuffix>
                  <!-- READ TYPE - 1: single-end 2:paired-end :: MANDATORY -->
                  <SampleType>2</SampleType>
         </Sample>
         <Sample>
                  <!-- SAMPLE NAME :: MANDATORY -->
                  <SampleName>Normal</SampleName>
                  <SampleFiles>WEx Normal</SampleFiles>
                  <!-- SUFFIX :: MANDATORY -->
                  <SampleSuffix>.fastq</SampleSuffix>
                  <!-- READ TYPE - 1: single-end 2:paired-end :: MANDATORY -->
                  <SampleType>2</SampleType>
         </Sample>
```

## Create the experiment file (SNV) :: ~ 30 minutes



```
[...]
         <!-- CALL TYPE :: OPTIONAL (default BOTH) -->
         <CallingType>BOTH</CallingType>
         <!-- GATKoutputMode - variants: EMIT VARIANTS ONLY, others: EMIT ALL SITES,
EMIT ALL CONFIDENT SITES ::default (EMIT VARIANTS ONLY) -->
         <GATKoutputMode>EMIT VARIANTS ONLY</GATKoutputMode>
         <!-- Clean dbSNP output entries :: 1, clean dbSNPs (default 0) :: OPTIONAL -->
         <rsFilter>0</rsFilter>
         <!-- RUbioSeq Mode Values: 0: standalone multisample, 1: joint multisample
execution (default 0) :: OPTIONAL-->
         <RUbioSeq Mode>0</RUbioSeq Mode>
         <!-- Run fastqc analysis :: 1, run analysis (default 0) :: OPTIONAL -->
         <fastgc>0</fastgc>
         <!-- Run TEQC analysis :: 1, convert file (default 0) :: OPTIONAL -->
         <!--<bedTEQCFlag>0</bedTEQCFlag>-->
         <!-- VEP analysis :: 1,execute analysis (default 0) :: OPTIONAL -->
         <VEPFlag>0</VEPFlag>
         <!-- Tumor/Control flag :: OnlyTumor and Germline analyses:: 1 (default 0) ::
OPTTONAL -->
         <TCFlag> 1 </TCFlag>
         <!-- Markduplicates flag (WARNING:: ONLY FOR ADVANCED USERS) Enable
markduplicates step :: 1, Disable markduplicates step :: 0 (default 1) :: OPTIONAL -->
         <MDFlag>1</MDFlag>
         <!-- Min phred-scaled confidence threshold for calling (default 30.0) -->
         <standCallConf>30.0</standCallConf>
         <!-- Min phred-scaled confidence threshold for emitting (default 30.0)-->
         <standEmitConf>30.0</standEmitConf>
         <!-- Queue project :: OPTIONAL (default none) -->
         <queueSGEProject>none</queueSGEProject>
         <!-- Whole exome and target sequencing analyses filtering -->
                  <!-- VCF Depth filter :: OPTIONAL -->
                  <DPmin>15
                  <!-- VCF min quality filter :: OPTIONAL -->
                  <minQual>100</minQual>
                  <!-- Optional. ONLY FOR ADVANCED USERS. Hard filter custom name -->
                  <!--<HfilterNameSNP>ODFilter</HfilterNameSNP>-->
                  <!-- Optional. ONLY FOR ADVANCED USERS. Hard filter custom rule -->
                  <!--<HfilterRuleSNP>QD&lt;2.0</HfilterRuleSNP>-->
         </HardFilters>
</configData>
```

### Check that everything is ready for running the pipeline

Save the experiment file

```
snv_Experiment.xml (~/OVCA_case/ConfigExperiments) - gedit
                              ♦ Undo ♦ 🔏 📋 🖺 Q 📿
snv_Experiment.xmt
 1 <?xml version="1.0" encoding="UTF-8"?>
 2 <!-- EXAMPLE RUBIOSEQ EXPERIMENT CONFIG FILE -->
3 <configData branch="SNV">
          <!-- GENOME REFERENCE PATH :: MANDATORY -->
          <GenRef>/home/participant/OVCA_case/Human_genome/hg19_chr17.fa</GenRef>
          <!-- DBSNP ANNOTATION PATH :: MANDATORY -->
          <DbSnpAnnot>/home/participant/OVCA_case/Bundle_GATK/dbsnp_138.hg19_chr17.vcf</DbSnpAnnot>
8
          <!-- 1000 Genomes ANNOTATION PATH :: MANDATORY -->
          <Genomes1000Annot>/home/participant/OVCA_case/Bundle_GATK/1000G_omni2.5.hg19_chr17.sites.vcf</Genomes1000Annot>
10
          <!-- REFSEQ ANNOTATION PATH :: MANDATORY -->
11
          <IndelAnnot>/home/participant/OVCA_case/Bundle_GATK/refseqhg19_chr17.rod</IndelAnnot>
          <!-- INTERVALS PATH :: OPTIONAL -->
          <Intervals>/home/participant/OVCA_case/Intervals/SureSelect_V5_human_all_Exons_chr17.bed</Intervals>
13
14
          <!-- KNOWN INDELS FOR REALIGNING :: OPTIONAL -->
          <KnownIndels>/home/participant/OVCA_case/Bundle_GATK/Mills_and_1000G.indels.hg19_chr17.sites.vcf</knownIndels>
15
16
          <!-- PLATFORM :: MANDATORY -->
          <Platform>illumina</Platform>
18
          <!-- checkCasava :: OPTIONAL. Set to 0 only if fastq reads have ascii33 qualities and they don't have standard Casava format
19
          <checkCasava>0</checkCasava>
20
          <!-- OUTPUT DIRECTORY :: OPTIONAL, DEFAULT: Home directory -->
          <dirOutBase>/home/participant/OVCA_case/</dirOutBase>
21
          <!-- PROJECT NAME :: MANDATORY -->
23
          <ProjectId>Output</ProjectId>
          <!-- USER NAME :: OPTIONAL(default Undefined) -->
```

Getting help to run the pipeline

### Launch the SNV and Indel calling



open a terminal, and execute the cmd:

\$ perl /home/participant/Software/RUbioSeq3.8.1/RUbioSeq.pl --analysis variantCalling --config /home/participant/OVCA case/ConfigExperiments/snv experiment.xml

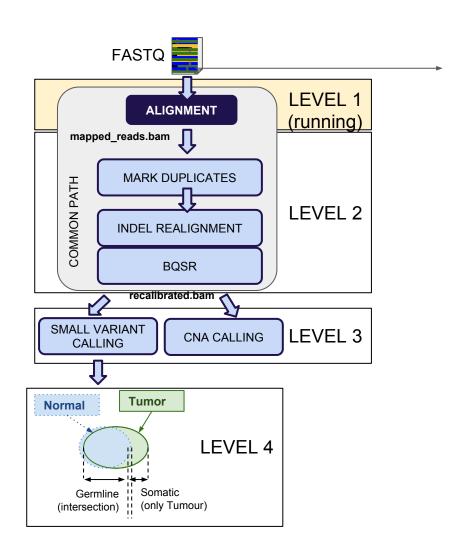
#### VARIANT CALLING ANALYSIS bwaPath: /local/participant/Soft/NGS/bwa-0.7.10/ javaRam: -Xmx16G samtoolsPath: /local/participant/Soft/NGS/samtools-0.1.19/ BFASTPath: /opt/NGS/bfast+bwa/0.7.0b/bin/ gatkpath: /local/participant/Soft/NGS/GenomeAnalysisTK-3.1-1/ The analysis is using picardPath: /local/participant/Soft/NGS/picard-tools-1.107/picard-tools-1.107 these parameters (you CallingType: BOTH RUbioSeq Mode: 0 just input them) IndelAnnot: /home/participant/OVCA case/Bundle GATK/refseqhg19 chr17.rod MDFlag: 1 checkCasava: 0 Genomes1000Annot: /home/participant/OVCA case/Bundle GATK/1000G omni2.5.hg19 chr17.sites.vcf InDirPreProcess: /home/participant/OVCA case/Raw data/ Intervals: /home/participant/OVCA case/Intervals/SureSelect V5 human all Exons chr17.bed minQual: 100 EMIT ALL SITES for TC analysis activated Directory /home/jperales/OVCA case//Output/ exists Executed command perl /home/iperales/Soft/RUbioSeg+/RUbioSeg3.7/variantCalling/,./common/indexReference.pl /home/iperales/Soft/samtools-0.1.19/ Level 0 /home/jperales/OVCA case/Human genome/hg19 chr17.fa -Xmx4G > /home/jperales/OVCA case//Output//log S0.txt 2>&1 Executed command perl /home/jperales/Soft/RUbioSeq+/RUbioSeq3.7/variantCalling/../common/sampleAlign.pl /home/jperales/OVCA case//Output/Tumor /home/jperales/OVCA case/Raw data/ /home/jperales/Soft/bwa-0.7.10/ /home/jperales/Soft/samtools-0.1.19/ /home/jperales/Soft/picard-tools-1.107/ /home/jperales/Soft/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1d5609cc/bin/ /home/jperales/OVCA case/Human genome/hg19 chr17.fa WEx Tumour .fastq Tumor jperales Level 1: Tumor illumina Tumor Output 2 4 -Xmx4G 1 /home/jperales/Soft/FastQC/ 0 0 > /home/jperales/OVCA case//Output/Tumor/log S1 WEx Tumour.txt 2>&1 [Level 2, Level 3 on Tumor sample] Executed command perl /home/jperales/Soft/RUbioSeq+/RUbioSeq3.7/variantCalling/../common/sampleAlign.pl /home/jperales/OVCA case//Output/Normal /home/jperales/OVCA case/Raw data/ /home/jperales/Soft/bwa-0.7.10/ /home/jperales/Soft/samtools-0.1.19/ /home/jperales/Soft/picard-tools-1.107/ Level 1: Normal /home/jperales/Soft/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1d5609cc/bin/ /home/jperales/OVCA case/Human genome/hg19 chr17.fa WEx Normal jperales/ovcA case/Human g illumina Normal Output 2 4 -Xmx4G 1 /home/jperales/Soft/FastQC/ 0 0 > /home/jperales/OVCA case//Output/Normal/log S1 WEx Normal.txt 2>&1 [Level 2. Level 3 on Normal sample] Executed command perl /home/jperales/Soft/RUbioSeq+/RUbioSeq3.7/variantCalling/postProcess.pl /home/jperales/OVCA case//Output/Tumor/calling/TC I evel 4 /home/jperales/OVCA case//Output/Tumor/calling/TC/onlyControl.vcf /home/jperales/OVCA case/Human genome/hg19 chr17.fa 0 1 > /home/iperales/OVCA case//Output/Tumor/log\_S4\_OC.txt 2>&1 Tumor-Matched

Normal

/home/jperales/OVCA case//Output/Tumor/log S4 B.txt 2>&1

/home/jperales/OVCA case//Output/Tumor/calling/TC/bothTC.vcf/home/jperales/OVCA case/Human genome/hg19 chr17.fa 0 1 >

Executed command perl /home/jperales/Soft/RUbioSeq+/RUbioSeq3.7/variantCalling/postProcess.pl /home/jperales/OVCA case//Output/Tumor/calling/TC

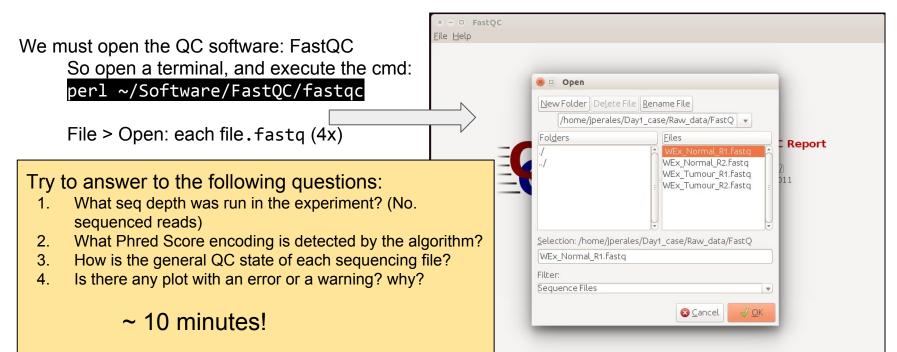


We will perform the Quality Control assessment in the raw data in the meantime the alignment (Level 1 from the pipeline) is running in background.

We will carry out a QC on the Case study raw data. Remember the data:

- Whole-exome sequencing (Illumina platform)
- paired-end sequencing (2 samples, 2 files each)





Manual: <a href="http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/Webpage:">http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/Webpage:</a> <a href="http://www.bioinformatics.babraham.ac.uk/projects/fastqc/">http://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a>

Try to answer to the following questions: Q: What seq depth was run in the experiment? (No. sequenced reads) A:
Q: What Phred Score encoding is detected by the algorithm? A:
Q: How is the general QC state of each sequencing file? A:
Q: Is there any plot with an error or a warning? Why? A:

### Try to answer to the following questions:

Q: What seq depth was run in the experiment? (No. sequenced reads)

A: ~ 1.4 Million reads. read length=75 pb.

Q: What Phred Score encoding is detected by the algorithm?

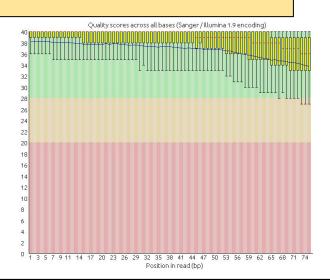
A: Sanger / Illumina 1.8+ (1.9)

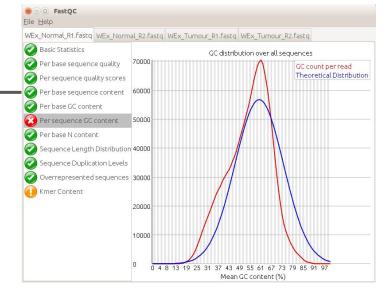
Q: How is the general QC state of each sequencing file?

A: Quite well.

Q: Is there any plot with an error or a warning? Why?

A: Per sequence GC content.





Basic	sequence stats
Measure	Value
Filename	WEx_Normal_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1400000
Filtered Sequences	0
Sequence length	75
%GC	53

Relative enrichment over read length



1 4 7 11 16 21 26 31 36 41 46 51 56 61 66 71 Position in read (bp)

Sequence	Count	Obs/Exp	Obs/Exp	Max Obs
AAAAA	296110	4.711	5.297	14
TITIT	241160	3.593	4.088	2

Coffee Break time

16:00 - 16:30

# RUNNING THE PIPELINE (Part II)

Javier Perales-Patón jperales@cnio.es



Translational Bioinformatics Unit CNIO. Madrid, Spain.

Fátima Al-Shahrour

[falshahrour@cnio.es]

Elena Piñeiro-Yáñez

[epineiro@cnio.es]

Pedro Fernandes

[pfern@igc.gulbenkian.pt]



## Practical: VARIANT DETECTION DAY1 CASE STUDY (Part II)

### Overview of the case study: Exome analysis (OVCA)



## Patient with ovarian cancer.

Whole-exome sequencing data from two samples from the patient:

- Tumour sample.
- Matched normal sample (healthy tissue) from epithelium.

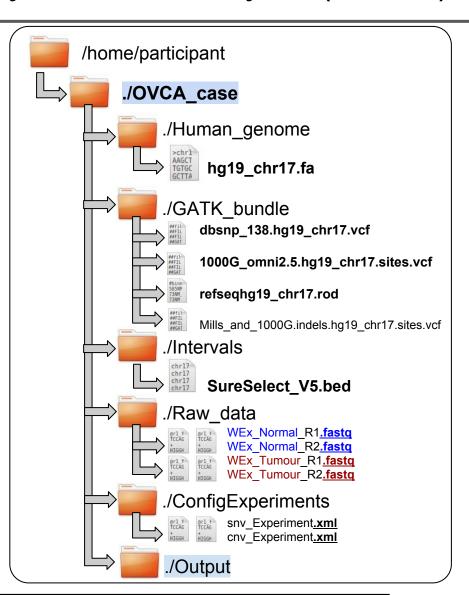
**Library protocol**: Agilent <u>SureSelect V5</u> <u>Human All Exons</u>.

Sequencing platform: HiSeq 2000 (Illumina)

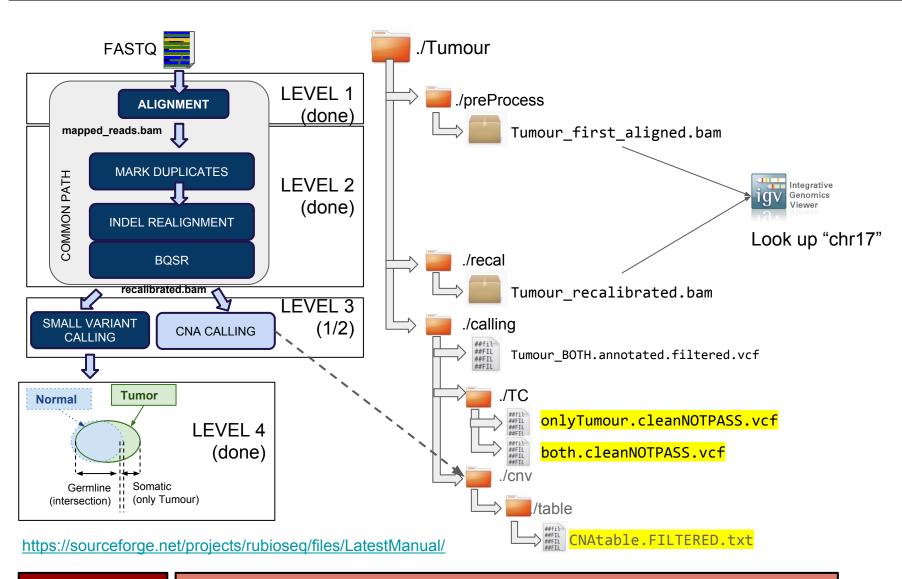
### **Expected outcome:**

- ~docens germ-line variants.
- A few somatic cancer mutations (SNV, indel or CNA).

NOTE: This data was simulated and reduced in order to perform the computational analysis in 30 minutes.



### Exome analysis (OVCA) :: output files

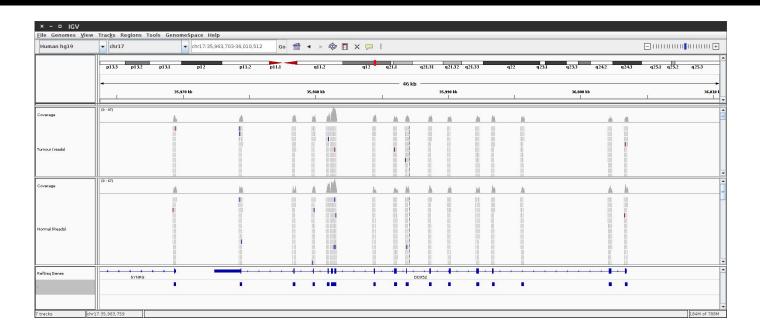




### Integrative Genomics IGV : Genome Browser

open a terminal, and execute the cmd:

\$ igv



Open the BAM files from each sample:

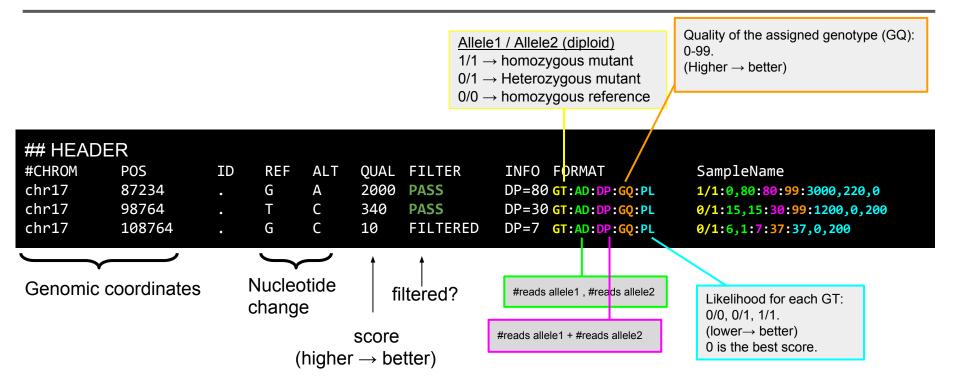


recalibrated.bam

Manual:

http://software.broadinstitute.org/software/igv/UserGuide

### Variant Call Format (vcf) from GATK callers



#### More info.:

https://www.broadinstitute.org/gatk/guide/article?id=1268

### Case study :: Point mutation results

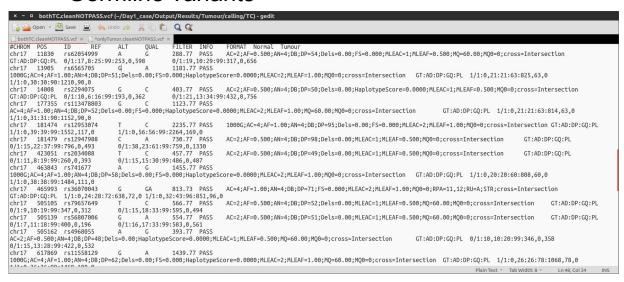
### Germline variants

There were detected germline variants in total:
- Single Nucleotide Variants.
- Indels.

### Somatic variants

### Case study :: Point mutation results

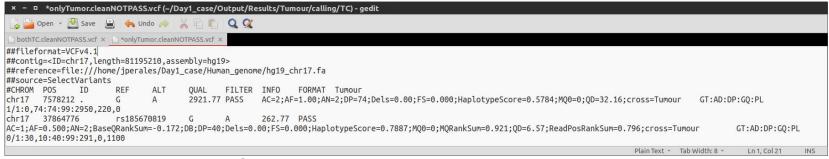
### Germline variants



There were detected 36 germline variants in total:

- 33 Single Nucleotide Variants.
- 3 Indels.

### Somatic variants



There were detected 2 somatic SNVs.

### Case study :: Alignment of somatic mutations

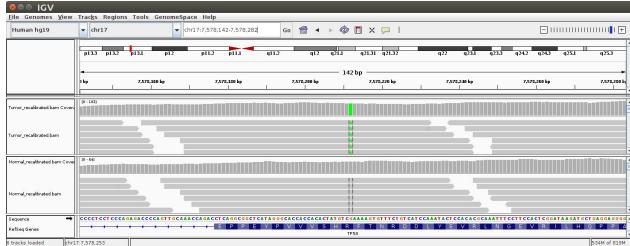
Integrative Genomics Viewer

chr17:7578212-7578212:G>A

- TP53
- Homozygous (AD=0,70)
- Good quality

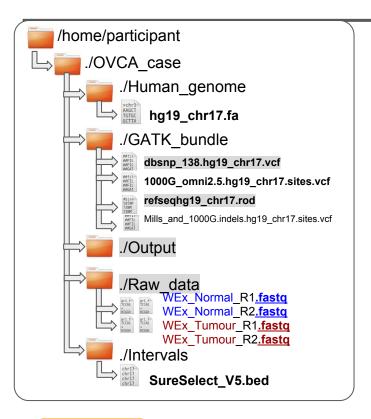
chr17:37864776-37864776:G>A

- ERBB2
- Heterozygous (AD=30,10)
- Moderate quality





## Create the experiment file (CNV) :: ~ 15 minutes



WARNING!! Use '--level 3' in the command because the 'Common Path' has been already done during the snvCalling.

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- EXAMPLE RUBIOSEQ EXPERIMENT CONFIG FILE -->
<configData branch="CNV">
          <!-- GENOME REFERENCE PATH :: MANDATORY -->
          <GenRef> path to hg19.fa </GenRef>
          <!-- DBSNP ANNOTATION PATH :: MANDATORY -->
          <DbSnpAnnot>__path_to_dbsnp_version.vcf___</DbSnpAnnot>
          <!-- REFSEQ ANNOTATION PATH :: MANDATORY -->
          <IndelAnnot> path_to_refseqhg19_chr17.rod__</IndelAnnot>
          <!-- INTERVALS PATH :: OPTIONAL -->
          <Intervals> path to WExLibrary.bed </Intervals>
          <!-- KNOWN INDELS FOR REALIGNING :: OPTIONAL -->
          <KnownIndels> path to Mills and 1000G.indels.hg19.sites.vcf </KnownIndels>
          <!-- PLATFORM :: MANDATORY -->
          <Platform>illumina</Platform>
          <!-- checkCasava :: OPTIONAL. -->
          <checkCasava>0</checkCasava>
          <!-- OUTPUT DIRECTORY :: DEFAULT: Home directory -->
          <dirOutBase>/home/participant/OVCA case/</dirOutBase>
          <!-- PROJECT NAME :: MANDATORY -->
          <ProjectId>Output</ProjectId>
          <!-- USER NAME :: OPTIONAL(default Undefined) -->
          <UserName>participant</UserName>
          <!-- RAW DATA PATH :: MANDATORY -->
          <InDirPreProcess>/home/participant/OVCA case/Raw data/</InDirPreProcess>
                    <!-- SAMPLE NAME :: MANDATORY -->
                    <SampleName>Tumor</SampleName>
                    <SampleFiles>WEx_Tumour</SampleFiles>
                    <!-- SUFFIX :: MANDATORY -->
                    <SampleSuffix>.fastq</SampleSuffix>
                    <!-- READ TYPE - 1: single-end 2:paired-end :: MANDATORY -->
                    <SampleType>2</SampleType>
          </Sample>
          <Sample>
                    <!-- SAMPLE NAME :: MANDATORY -->
                    <SampleName>Normal</SampleName>
                    <SampleFiles>WEx_Normal</SampleFiles>
                    <!-- SUFFIX :: MANDATORY -->
                    <SampleSuffix>.fastq</SampleSuffix>
                    <!-- READ TYPE - 1: single-end 2:paired-end :: MANDATORY -->
                    <SampleType>2</SampleType>
          </Sample>
```



### Case study :: Copy-number variant results

