

RUNNING THE PIPELINE (Part I)

Javier Perales-Patón
jperales@cniio.es



Translational Bioinformatics Unit
CNIO. Madrid, Spain.

Fátima Al-Shahrour
[\[falshahrour@cniio.es\]](mailto:falshahrour@cniio.es)
Elena Piñeiro-Yáñez
[\[epineiro@cniio.es\]](mailto:epineiro@cniio.es)
Pedro Fernandes
[\[pfern@igc.gulbenkian.pt\]](mailto: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.
 - Capture and Library construction.

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 SureSelect V5
Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

Expected outcome:

- ~docs 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.

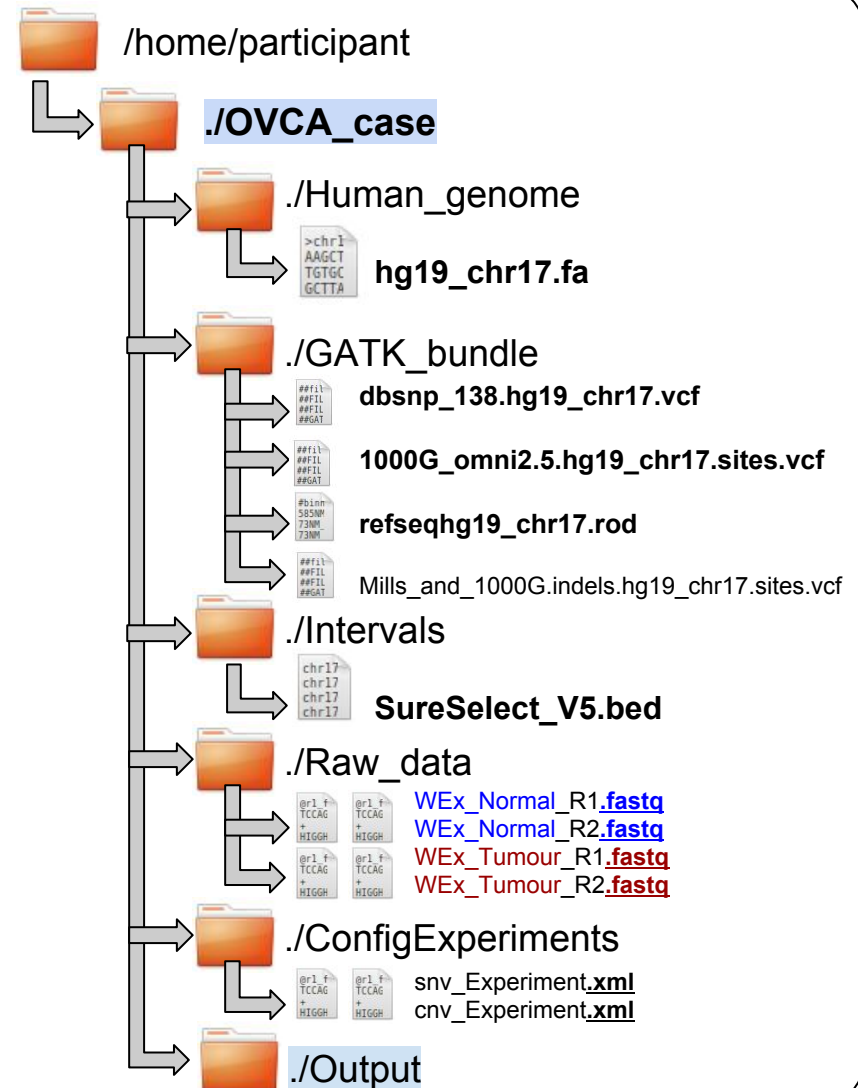
Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

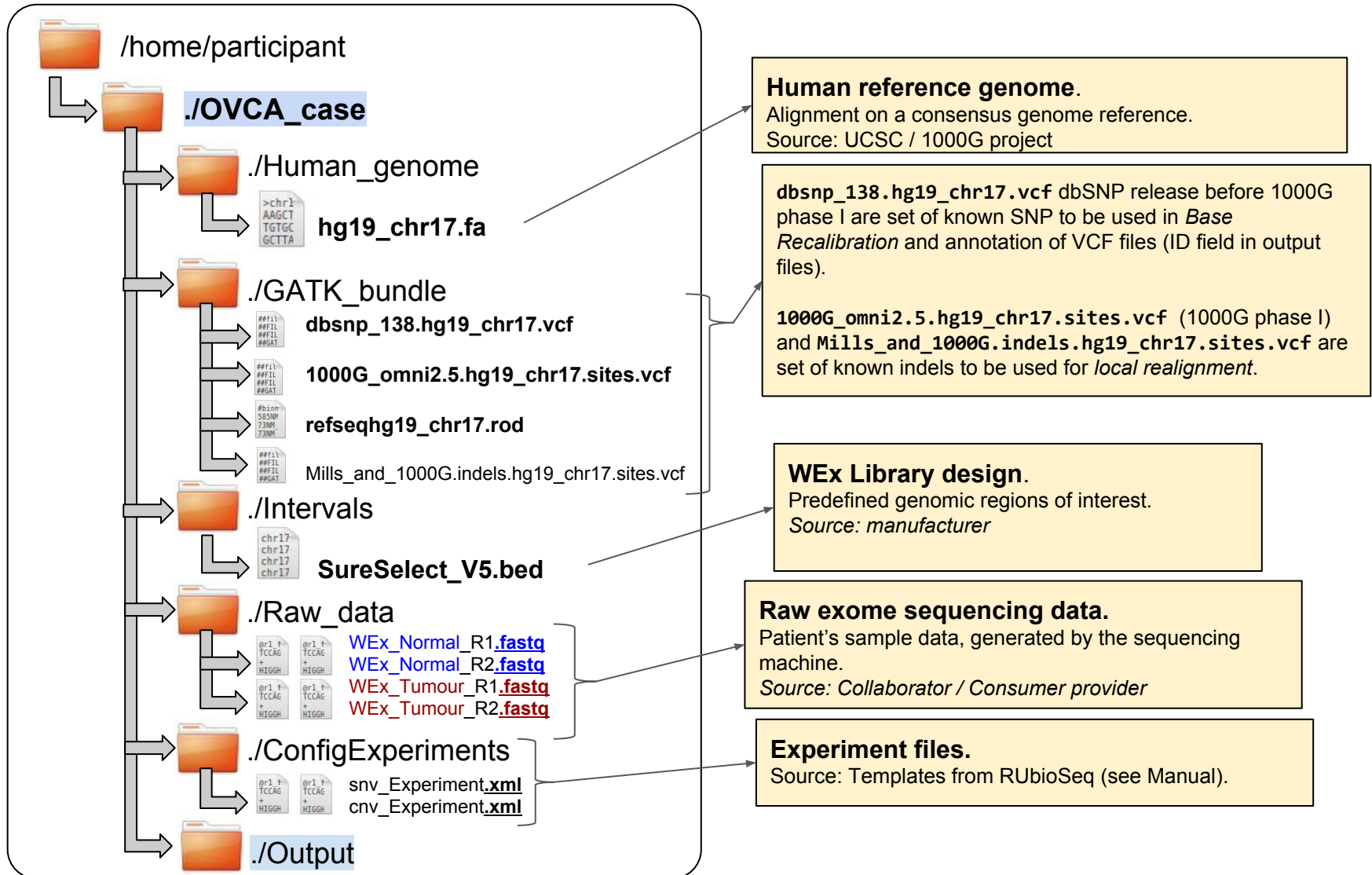
Expected outcome:

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

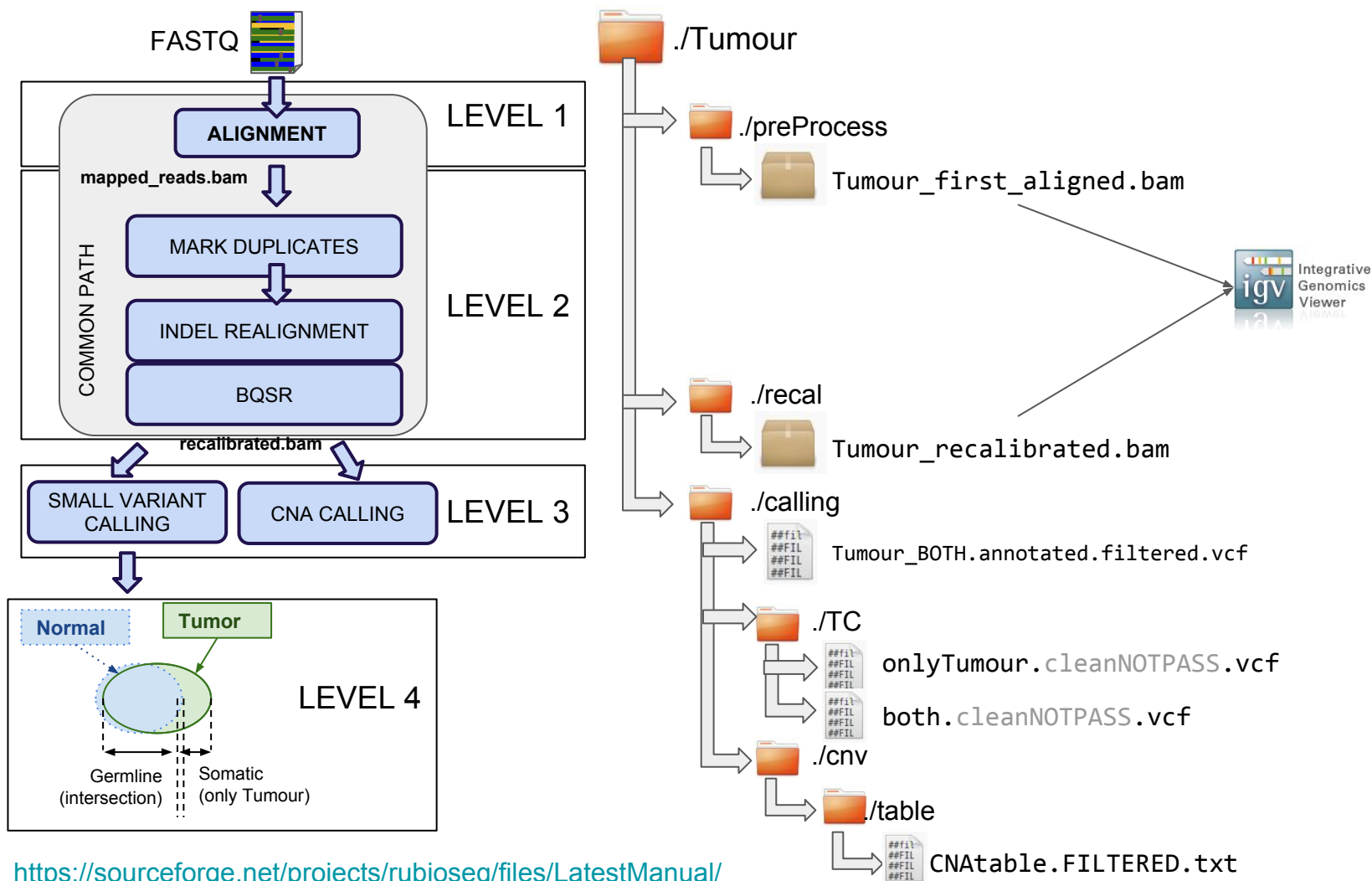
NOTE: This data was simulated and reduced in order to perform the computational analysis in 30 minutes.



Exome analysis (OVCA) :: Input files

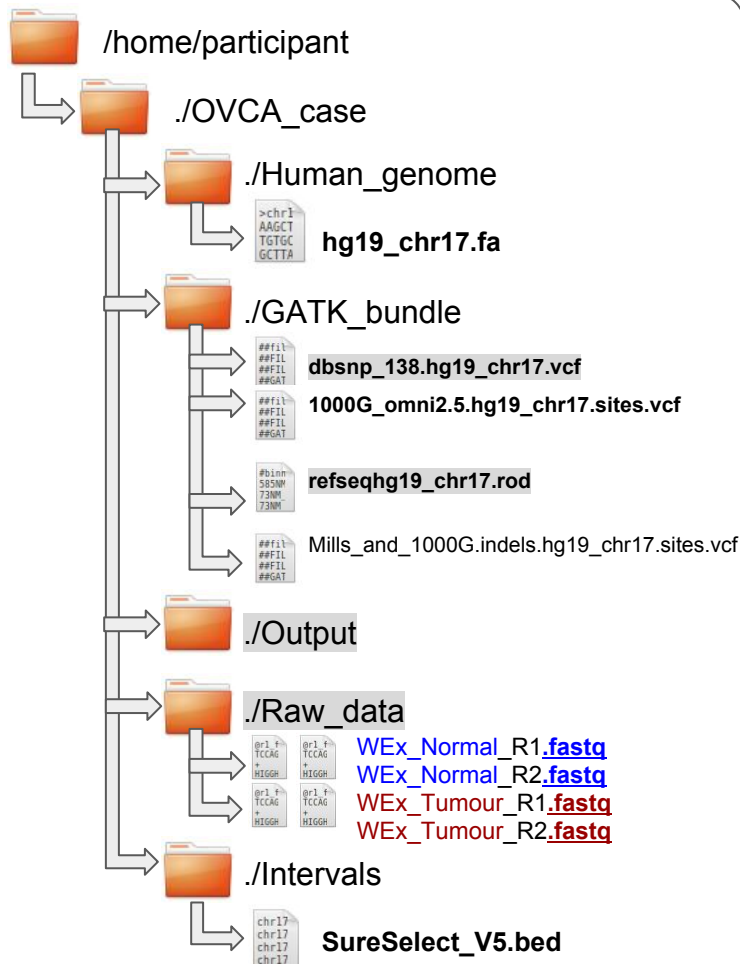


Exome analysis (OVCA) :: output files



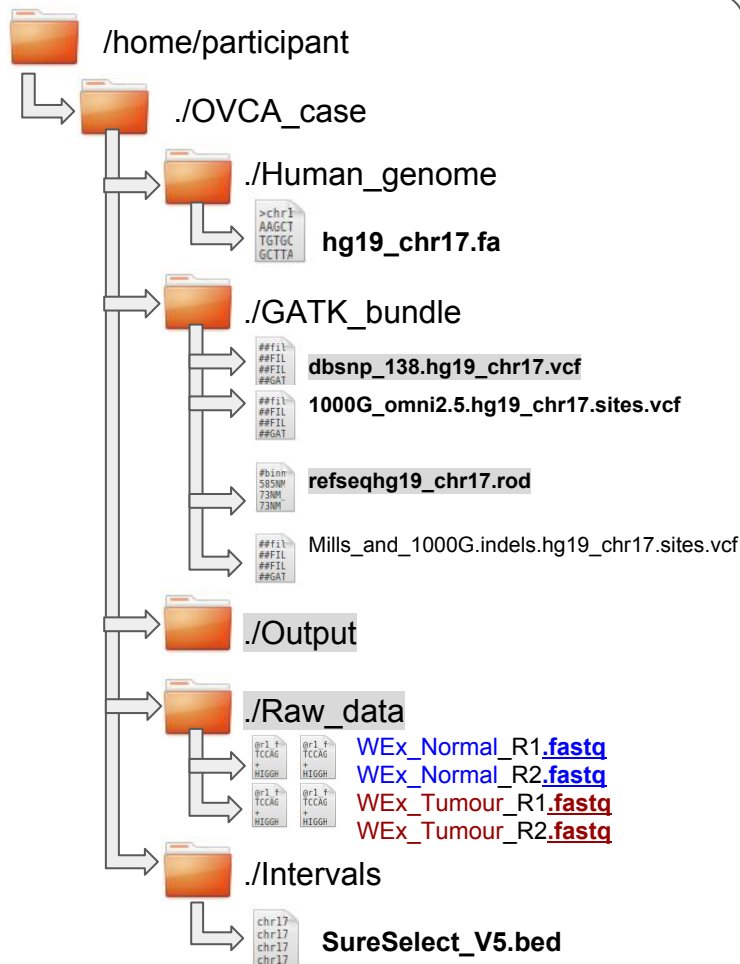
<https://sourceforge.net/projects/rubioseq/files/LatestManual/>

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>Tumour</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>
</configData>
```

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

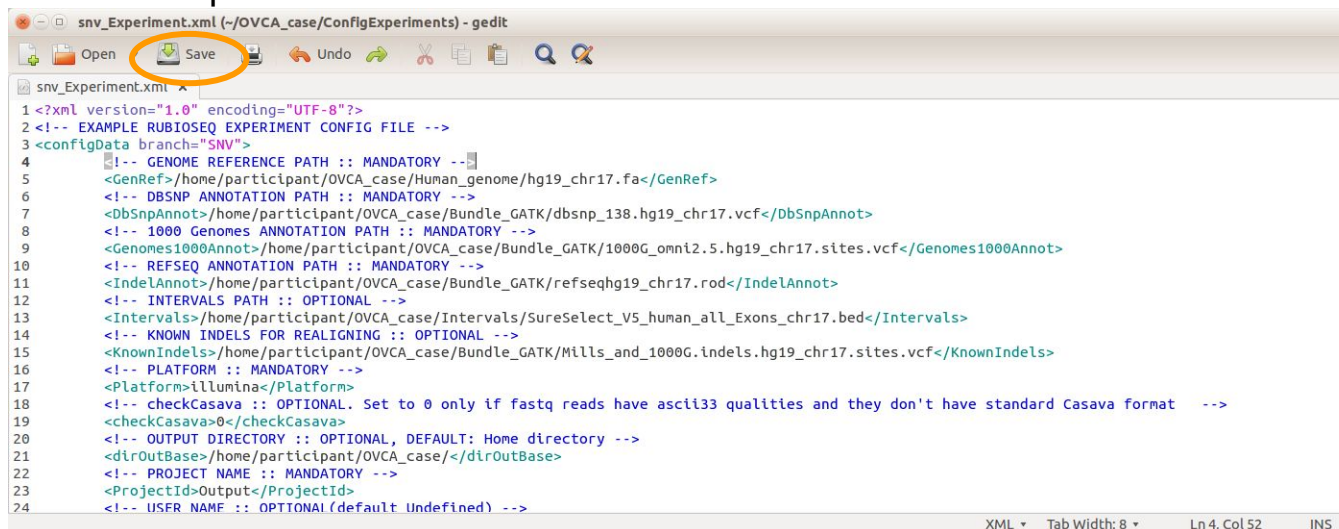


```
[ . . . ]

<!-- 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 -->
<fastqc>0</fastqc>
<!-- 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) ::
OPTIONAL -->
<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 -->
<HardFilters>
  <!-- VCF Depth filter :: OPTIONAL -->
  <DPmin>15</DPmin>
  <!-- VCF min quality filter :: OPTIONAL -->
  <minQual>100</minQual>
  <!-- Optional. ONLY FOR ADVANCED USERS. Hard filter custom name -->
  <!--<HfilterNameSNP>QDfilter</HfilterNameSNP>-->
  <!-- Optional. ONLY FOR ADVANCED USERS. Hard filter custom rule -->
  <!--<HfilterRuleSNP>QD<2.0</HfilterRuleSNP>-->
</HardFilters>
</configData>
```


Check that everything is ready for running the pipeline

- Save the experiment file



- Getting help to run the pipeline

```
$ perl /home/participant/Software/RUBioSeq3.8.1/RUBioSeq.pl -h
RUBioSeq.pl --analysis analysisType --config config_file [--level level_number]
```

Getting help:

[--help]

Analysis Types:

variantCalling : Variant Calling Workflow.(default)

cnvCalling: CNV Calling Workflow.

ChIPseq: ChIPseq workflow.

methylationCalling : Methylation Calling Workflow.

Example:

```
./RUBioSeq.pl --analysis variantCalling --config /dir/config.xml --level 3
```

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/  
picardPath: /local/participant/Soft/NGS/picard-tools-1.107/picard-tools-1.107  
[ ... ]  
TCFlag: 1  
CallingType: BOTH  
RUBioSeq_Mode: 0  
IndelAnnot: /home/participant/OVCA_case/Bundle_GATK/refseqhg19_chr17.rsd  
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  
fastqc: 1  
minQual: 100
```

EMIT ALL SITES for TC analysis activated

Directory /home/jperales/OVCA_case/Output/ exists

Executed command perl /home/jperales/Soft/RUBioSeq+/RUBioSeq3.7/variantCalling/./common/indexReference.pl /home/jperales/Soft/samtools-0.1.19/
/home/jperales/Soft/picard-tools-1.107/ /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
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/
/home/jperales/Soft/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1d5609cc/bin/ /home/jperales/OVCA_case/Human_genome/hg19_chr17.fa WEX_Normal .fastq Normal jperales
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
/home/jperales/OVCA_case/Output/Tumor/calling/TC/onlyControl.vcf /home/jperales/OVCA_case/Human_genome/hg19_chr17.fa 0 1 >
/home/jperales/OVCA_case/Output/Tumor/log_S4_OC.txt 2>&1
Executed command perl /home/jperales/Soft/RUBioSeq+/RUBioSeq3.7/variantCalling/postProcess.pl /home/jperales/OVCA_case/Output/Tumor/calling/TC
/home/jperales/OVCA_case/Output/Tumor/calling/TC/bothTC.vcf /home/jperales/OVCA_case/Human_genome/hg19_chr17.fa 0 1 >
/home/jperales/OVCA_case/Output/Tumor/log_S4_B.txt 2>&1

The analysis is using
these parameters (you
just input them)

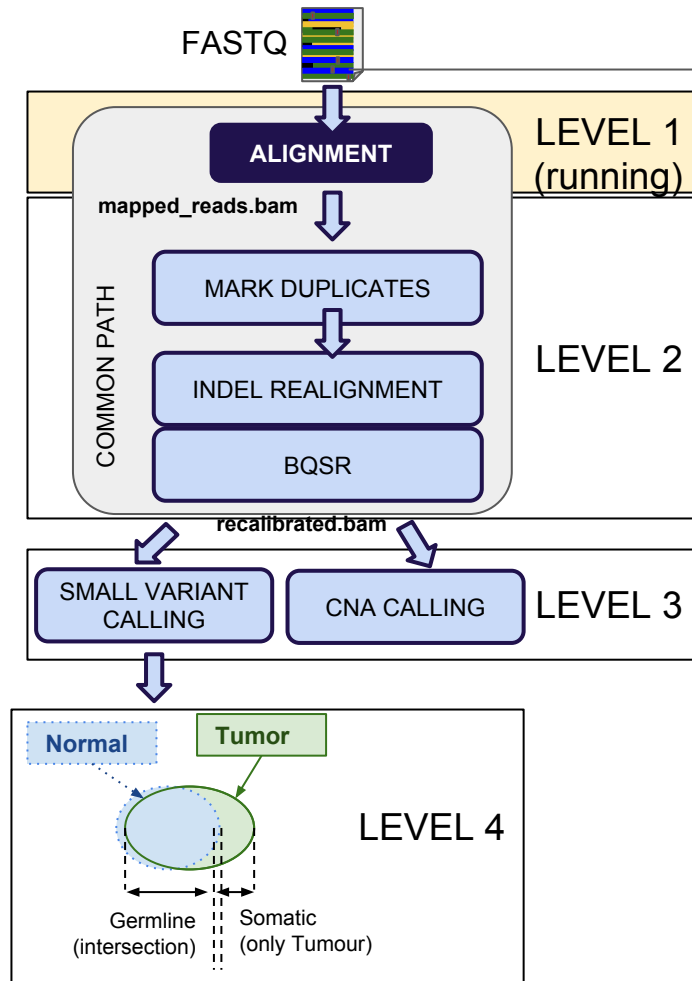
Level 0

Level 1: Tumor

Level 1: Normal

Level 4:
Tumor-Matched
Normal

Hands-on Quality Control



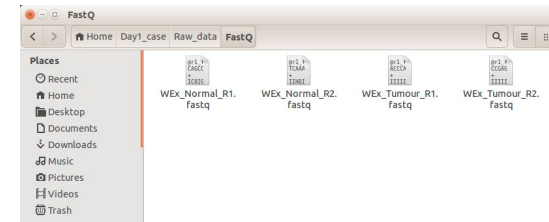
We will perform the Quality Control assessment in the raw data in the meantime the alignment (Level 1 from the pipeline) is running in background.

Hands-on Quality Control

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)



We must open the QC software: FastQC

So open a terminal, and execute the cmd:

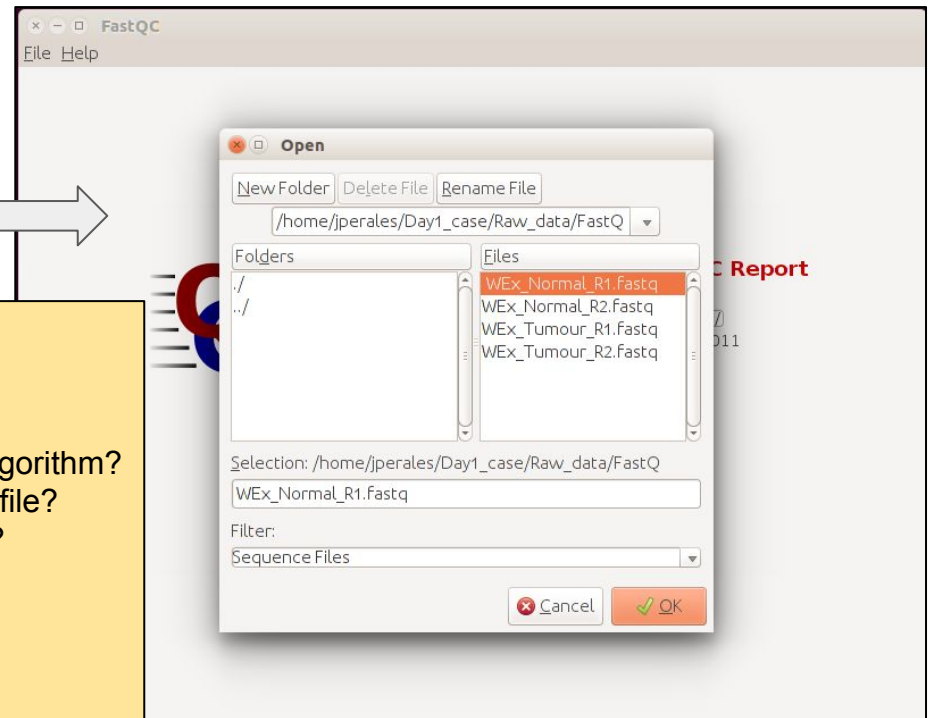
```
perl ~/Software/FastQC/fastqc
```

File > Open: each file .fastq (4x) →

Try to answer to the following questions:

1. What seq depth was run in the experiment? (No. sequenced reads)
2. What Phred Score encoding is detected by the algorithm?
3. How is the general QC state of each sequencing file?
4. Is there any plot with an error or a warning? why?

~ 10 minutes!



Manual: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/>

Webpage: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Hands-on Quality Control

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:

Coffee Break time

16:00 - 16:30

RUNNING THE PIPELINE (Part II)

Javier Perales-Patón
jperales@cniio.es



Translational Bioinformatics Unit
CNIO. Madrid, Spain.

Fátima Al-Shahrour
[\[falshahrour@cniio.es\]](mailto:falshahrour@cniio.es)
Elena Piñeiro-Yáñez
[\[epineiro@cniio.es\]](mailto:epineiro@cniio.es)
Pedro Fernandes
[\[pfern@igc.gulbenkian.pt\]](mailto: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.

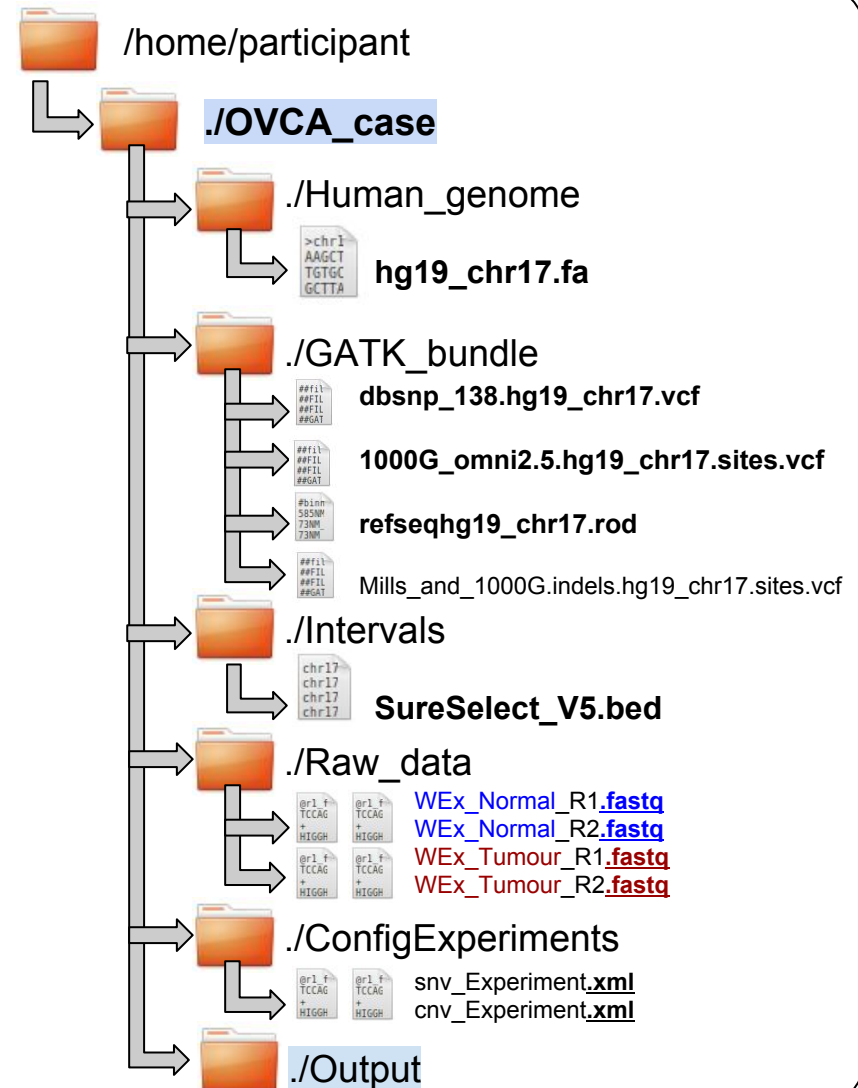
Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

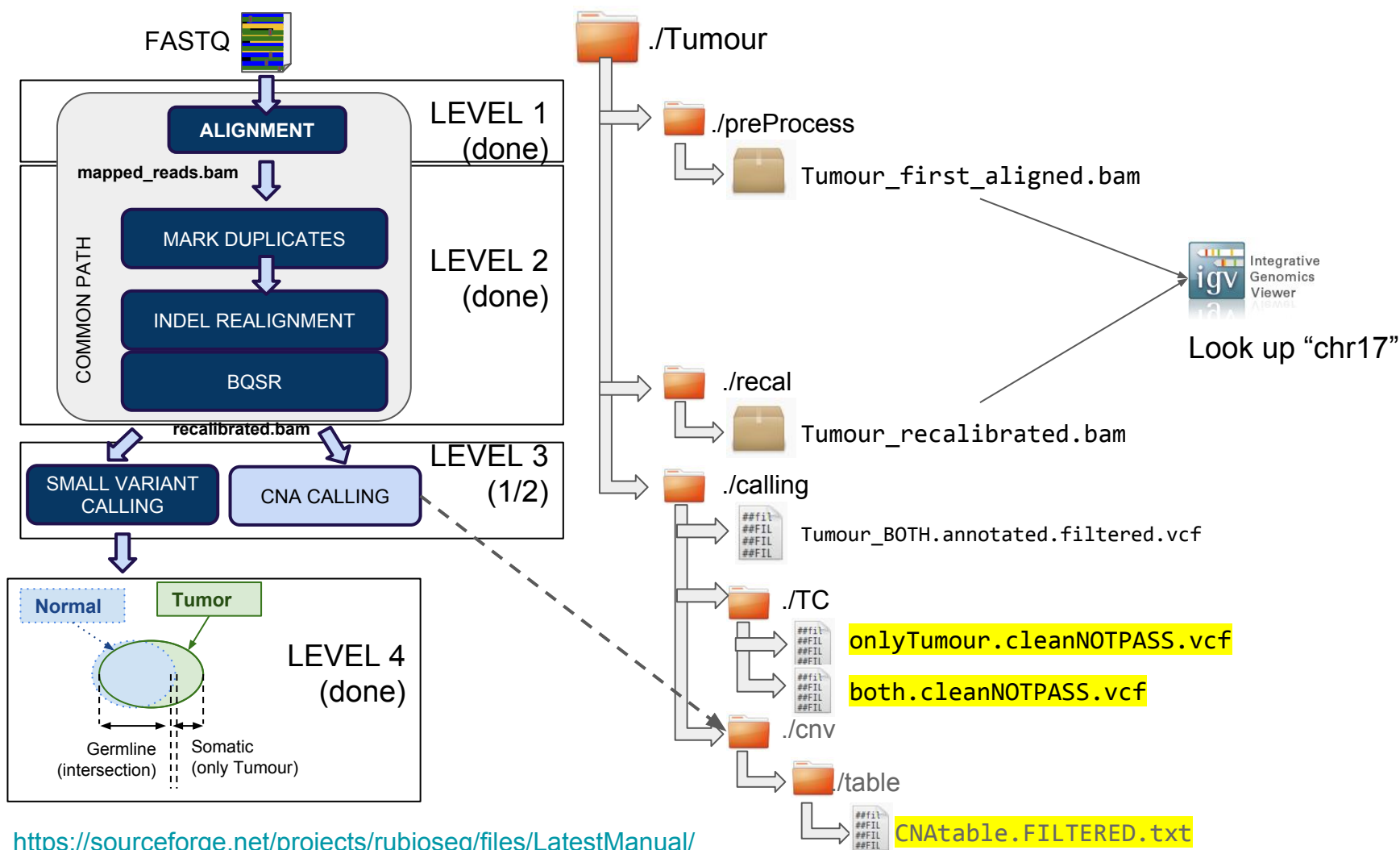
Expected outcome:

- ~docs 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

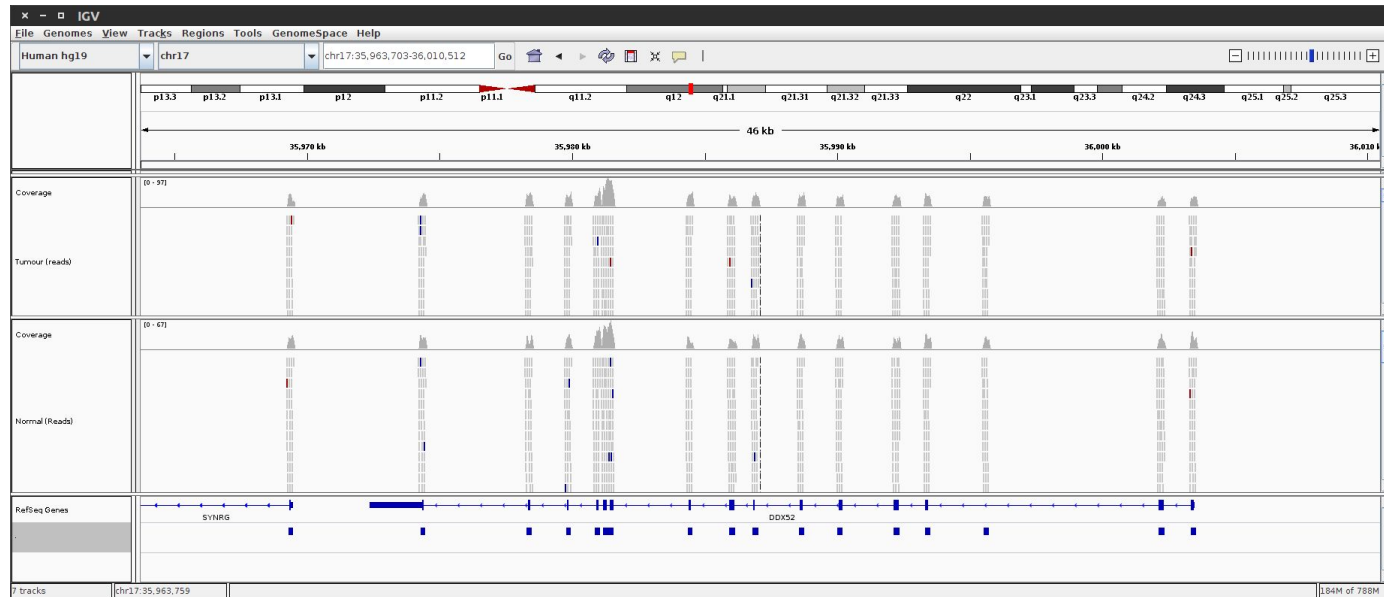




IGV : Genome Browser

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

| ## HEADER | | | | | | | | | | | |
|-----------|--------|----|-----|-----|------|----------|-------|----------------|----------------------------|--|--|
| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | SampleName | | |
| chr17 | 87234 | . | G | A | 2000 | PASS | DP=80 | GT:AD:DP:GQ:PL | 1/1:0,80:80:99:3000,220,0 | | |
| chr17 | 98764 | . | T | C | 340 | PASS | DP=30 | GT:AD:DP:GQ:PL | 0/1:15,15:30:99:1200,0,200 | | |
| chr17 | 108764 | . | G | C | 10 | FILTERED | DP=7 | GT:AD:DP:GQ:PL | 0/1:6,1:7:37:37,0,200 | | |

Allele1 / Allele2 (diploid)
 1/1 → homozygous mutant
 0/1 → Heterozygous mutant
 0/0 → homozygous reference

Quality of the assigned genotype (CQ):
 0-99.
 (Higher → better)

Genomic coordinates

Nucleotide change

score
(higher → better)

filtered?

#reads allele1 , #reads allele2

#reads allele1 + #reads allele2

Likelihood for each GT:
 0/0, 0/1, 1/1.
 (lower → better)
 0 is the best score.

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

There were detected somatic SNVs.

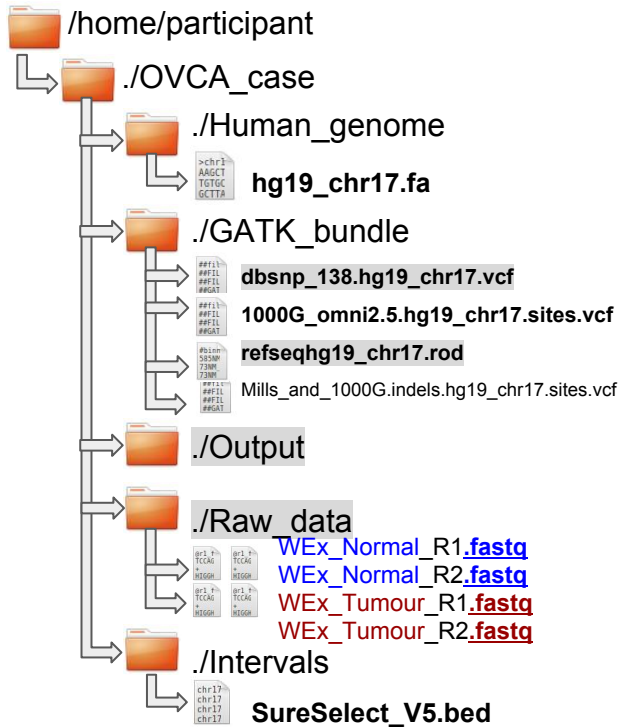
Genes affected and type of mutations

(see alignment using **IGV on chr17**):

- .
- .



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



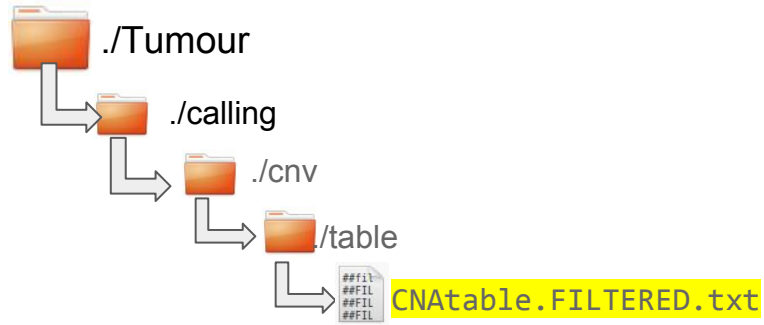
```
<?xml version="1.0" encoding="UTF-8"?>
<!-- EXAMPLE RUBIOSEQ EXPERIMENT CONFIG FILE -->
<configData branch="CNV">
  <!-- GENOME REFERENCE PATH :: MANDATORY -->
  <GenRef>path_to_hg19_fa</GenRef>
  <!-- DBSNP ANNOTATION PATH :: MANDATORY -->
  <DbSnpAnnot>path_to_dbsnp_version.vcf</DbSnpAnnot>
  <!-- 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>
    <!-- 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>
</configData>
```

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



```
$ perl /home/participant/Software/RUBioSeq3.8.1/RUBioSeq.pl --analysis cnvCalling
--config /home/participant/OVCA_case/configExperiments/cnv_experiment.xml --level 3
```

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



There were detected somatic CNVs (see table).

Genes affected and type of mutations
(see alignment using **IGV on chr17**):

- .
- .



