RUNNING THE PIPELINE (Part I)

Javier Perales-Patón
jperales@cnio.es

Fátima Al-Shahrour
[falshahrour@cnio.es]
Elena Piñeiro-Yáñez
[epineiro@cnio.es]
Pedro Fernandes
[pfern@igc.gulbenkian.pt]

cnio
Centro Nacional de Investigaciones Oncológicas

Translational Bioinformatics Unit
CNIO. Madrid, Spain.

INSTITUTO GULBENKIAN DE CIÊNCIA

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 suffering 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:

- ~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 suffering ovarian cancer.

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

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

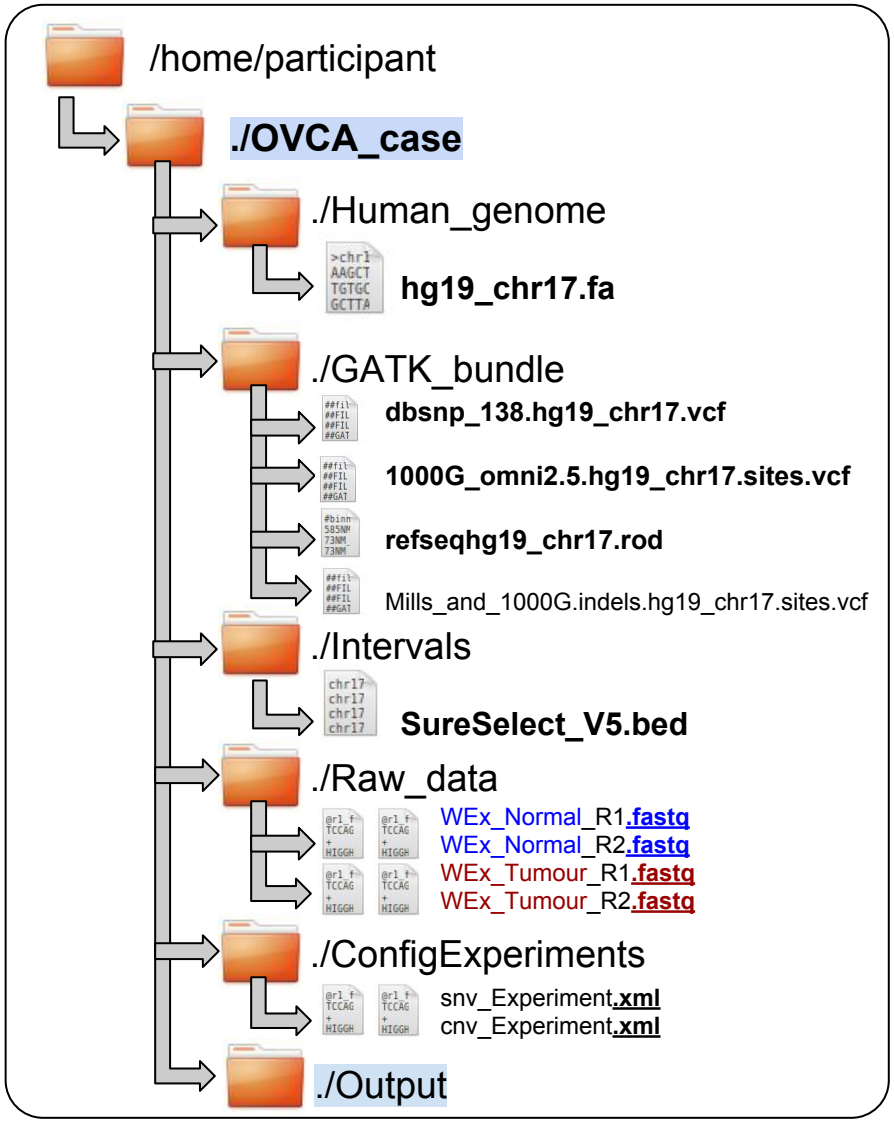
Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

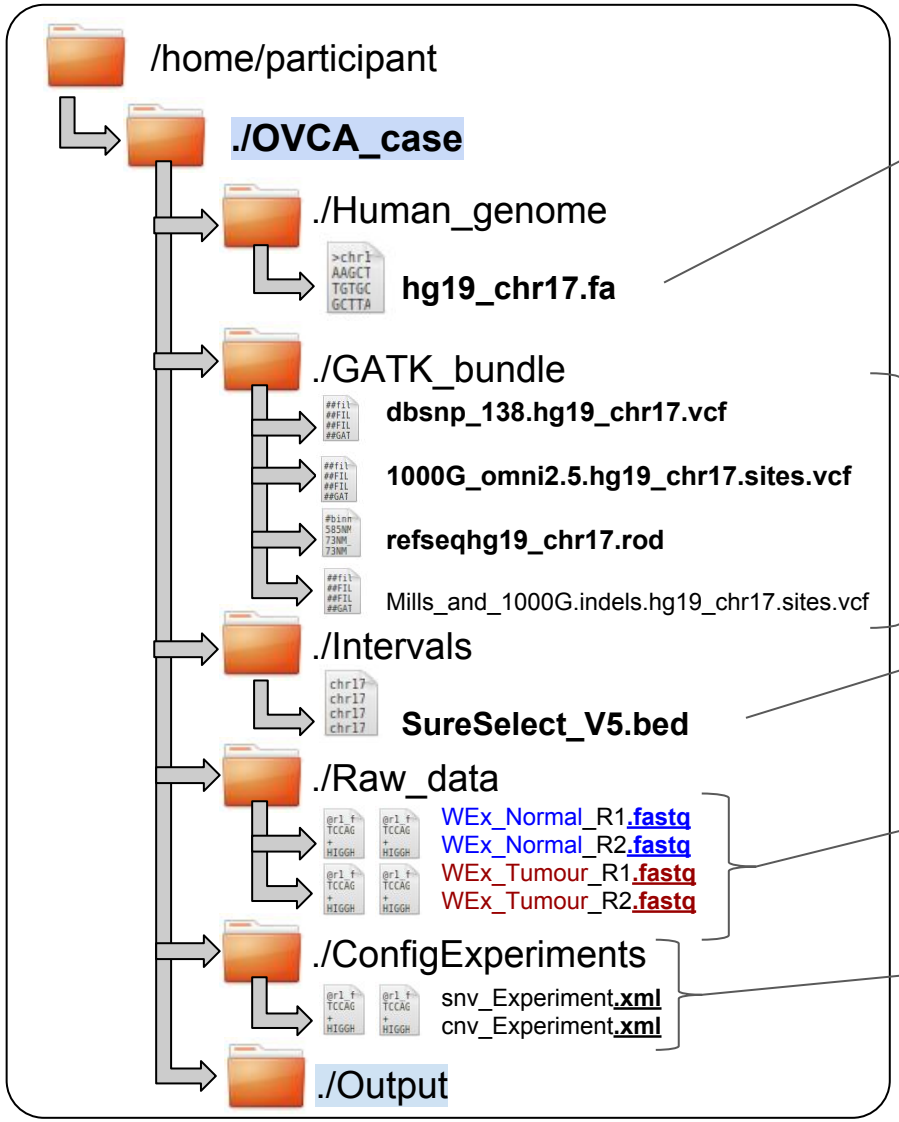
Expected outcome:

- ~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



Human reference genome.
Alignment on a consensus genome reference.
Source: UCSC / 1000G project

`dbsnp_138.hg19_chr17.vcf` dbSNP release before 1000G phase I are set of known SNP to be used in *Base Recalibration* and annotation of VCF files (ID field in output files).

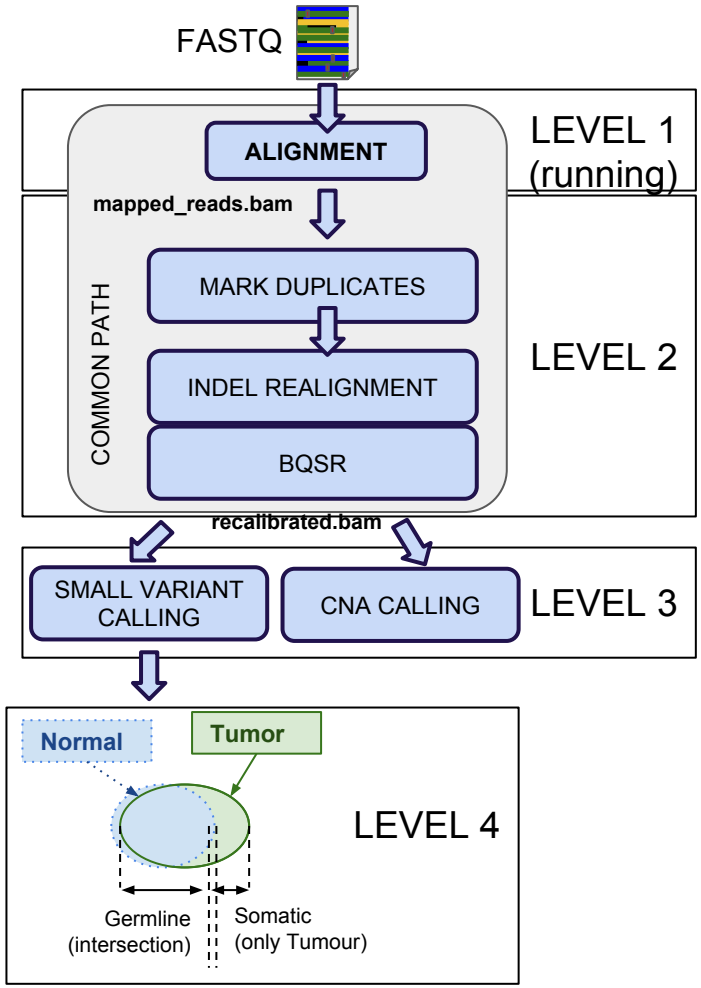
`1000G_omni2.5.hg19_chr17.sites.vcf` (1000G phase I) and **`Mills_and_1000G.indels.hg19_chr17.sites.vcf`** are set of known indels to be used for *local realignment*.

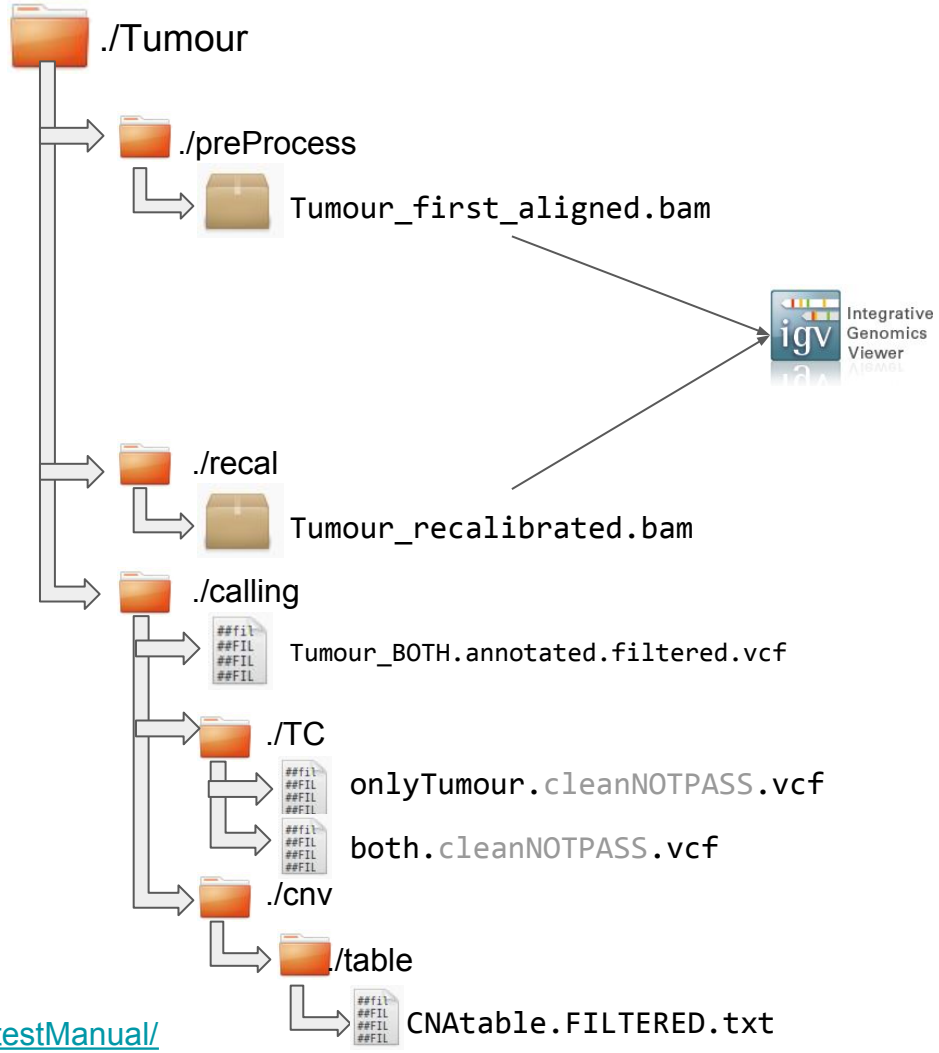
WEx Library design.
Predefined genomic regions of interest.
Source: manufacturer

Raw exome sequencing data.
Patient's sample data, generated by the sequencing machine.
Source: Collaborator / Consumer provider

Experiment files.
Source: Templates from RUbioSeq (see Manual).

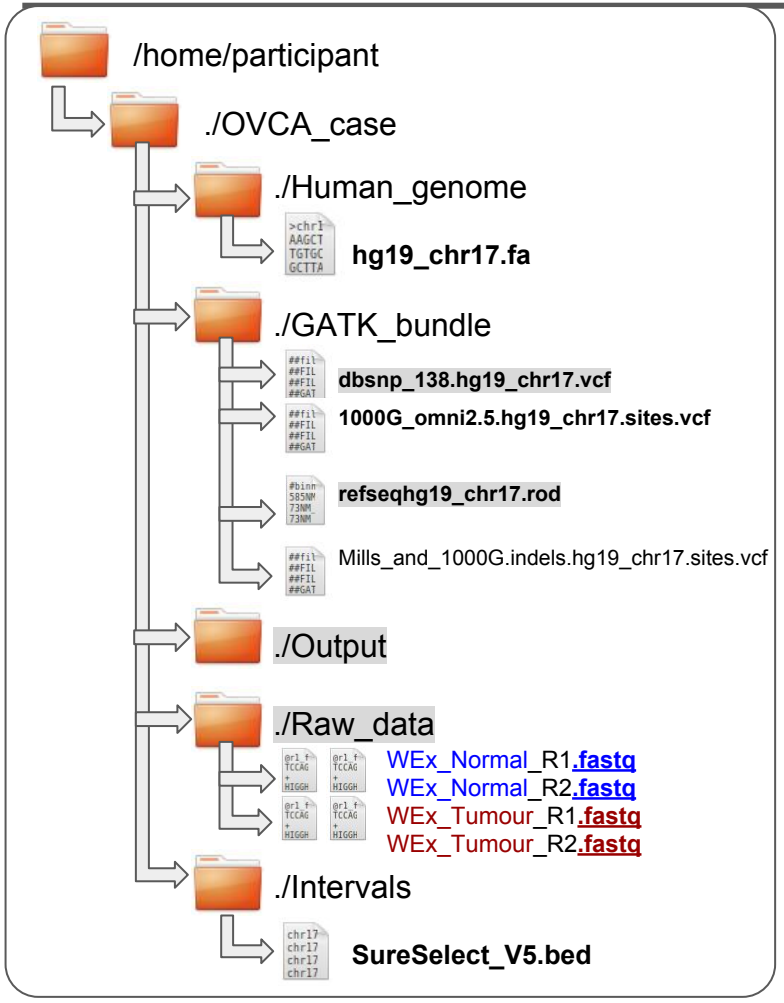
Exome analysis (OVCA) :: output files





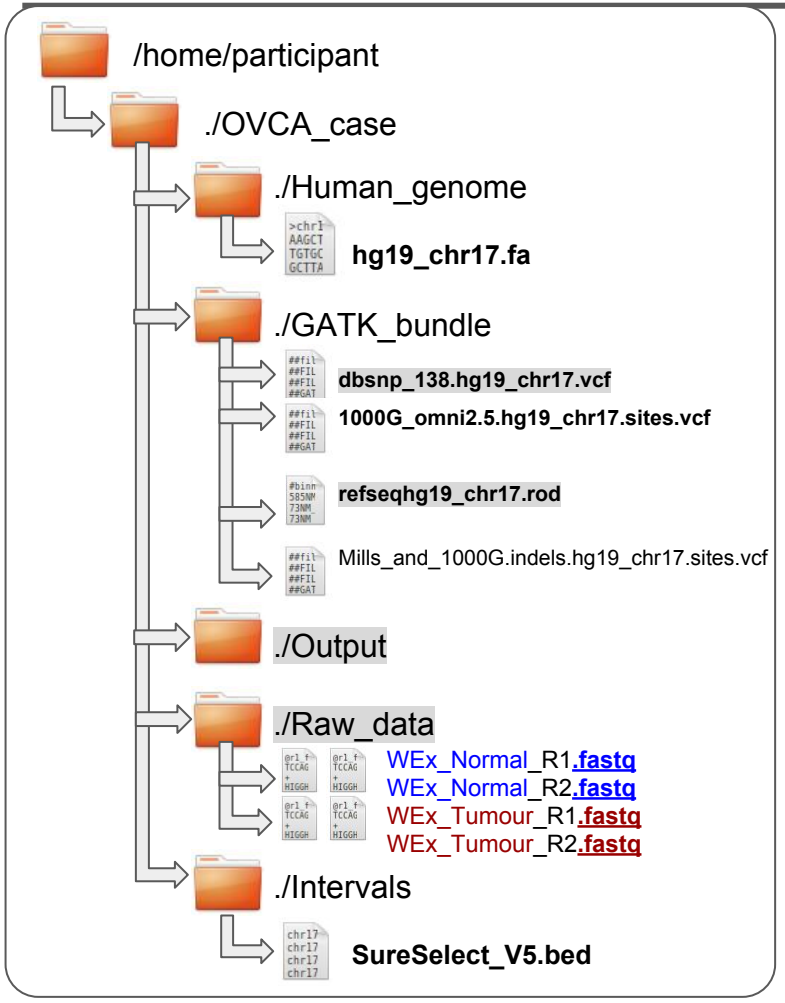
https://sourceforge.net/projects/rubioseq/files/LatestManual/

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>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

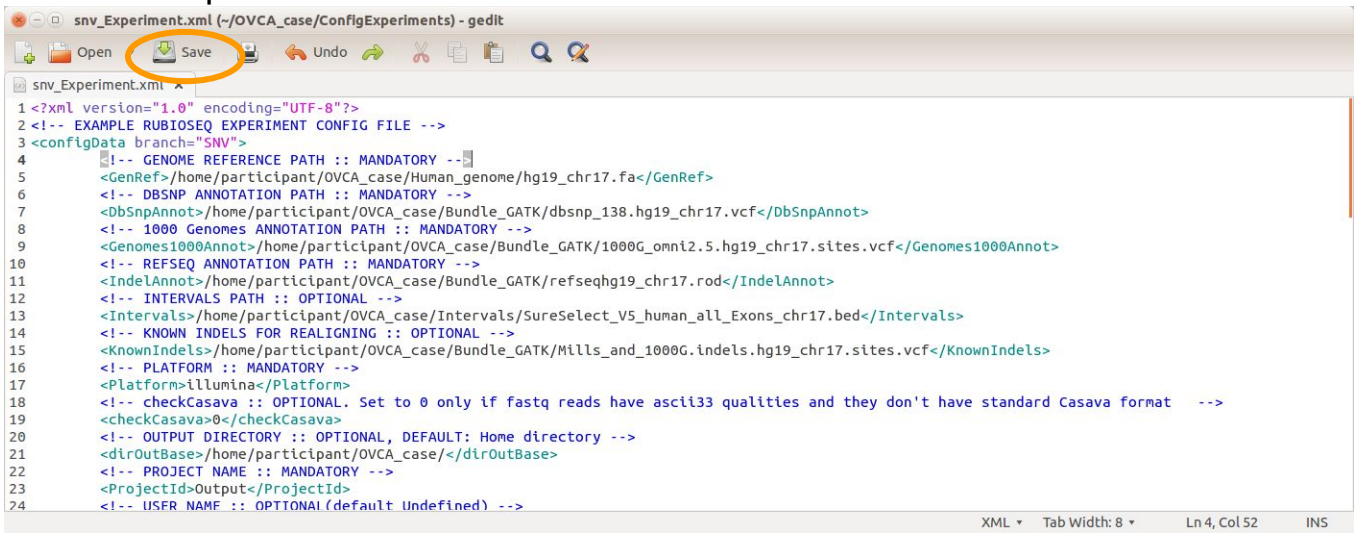



```
        [ . . . ]

        <!-- 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 -->
        <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&lt;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/Sotware/RUbioSeq+/RUbioSeq3.7/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/Sotware/RUbioSeq+/RUbioSeq3.7/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.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
fastqc: 1
minQual: 100

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

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

Level 0

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 1: Tumor

[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

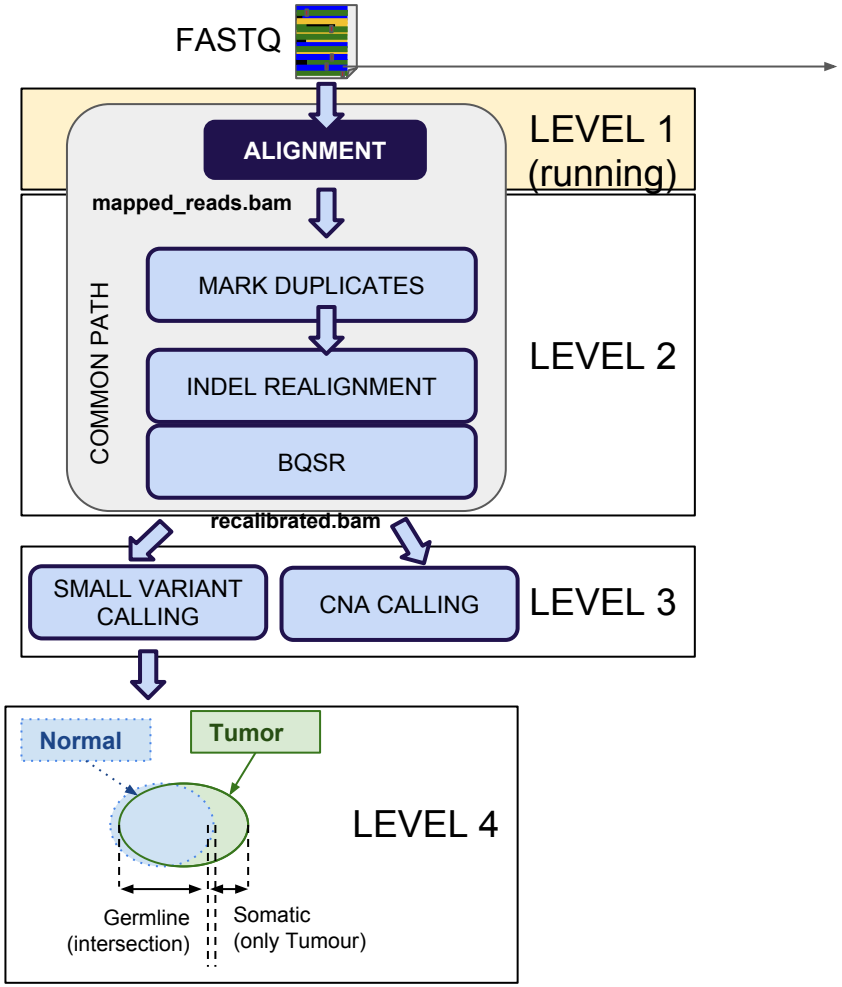
Level 1: Normal

[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

Level 4: Tumor-Matched Normal

Hands-on Quality Control



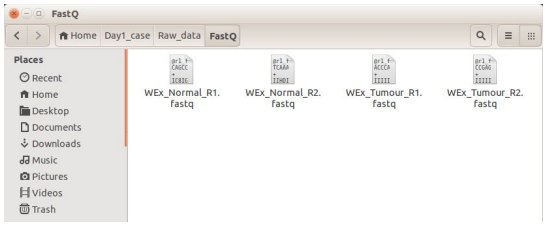
We will perform the Quality Control assessment in the raw data in the meantime the alignment (Level 1 from the pipeline) is running in background.

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)

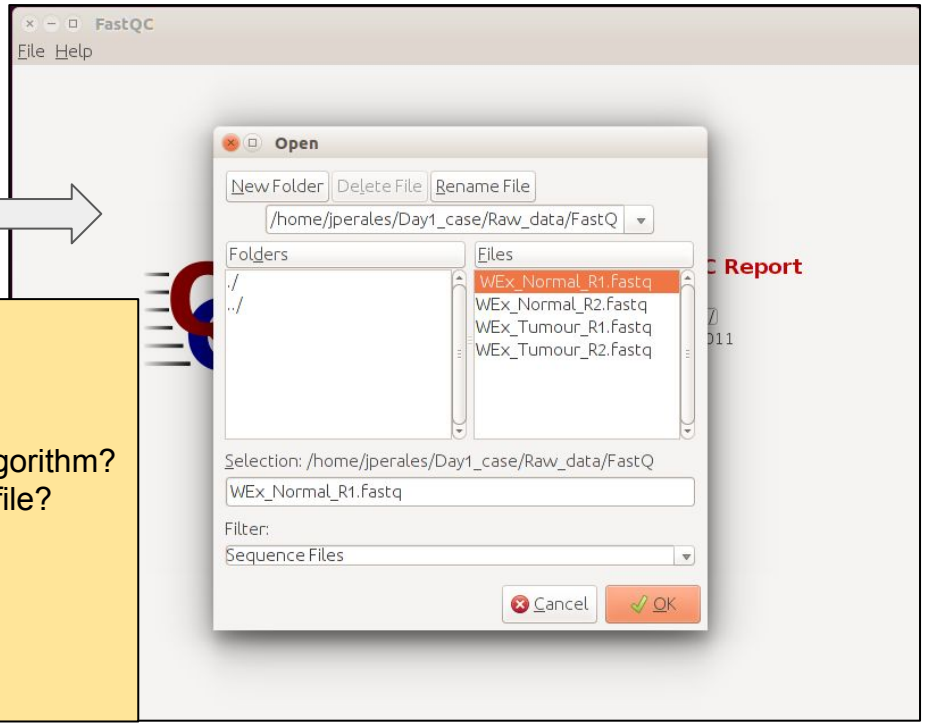


We must open the QC software: FastQC

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

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

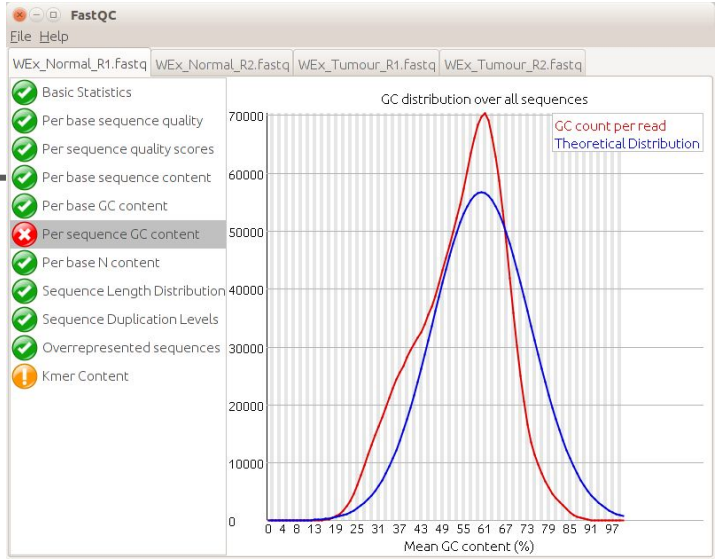
A: Sanger / Illumina 1.8+ (1.9)

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

A: Quite well.

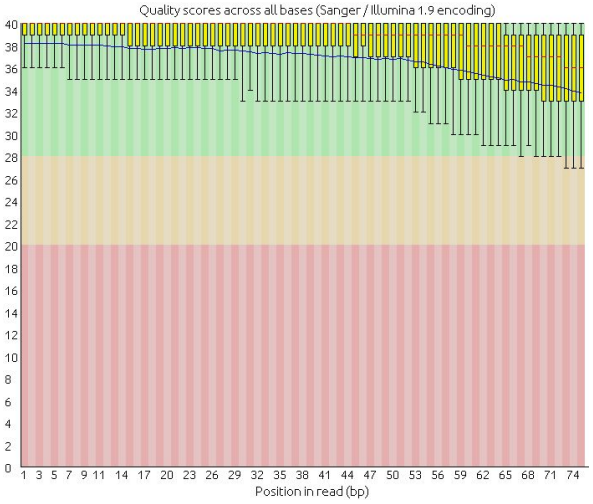
Q: Is there any plot with an error or a warning? Why?

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



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

Coffee Break time

16:00 - 16:30

RUNNING THE PIPELINE (Part II)

Javier Perales-Patón
jperales@cnio.es

Fátima Al-Shahrour
[falshahrour@cnio.es]
Elena Piñeiro-Yáñez
[epineiro@cnio.es]
Pedro Fernandes
[pfern@igc.gulbenkian.pt]

cnio
Centro Nacional de Investigaciones Oncológicas

Translational Bioinformatics Unit
CNIO. Madrid, Spain.

INSTITUTO GULBENKIAN DE CIÊNCIA

Practical: VARIANT DETECTION
DAY1 CASE STUDY
(Part II)

Overview of the case study: Exome analysis (OVCA)

Patient suffering ovarian cancer.

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

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

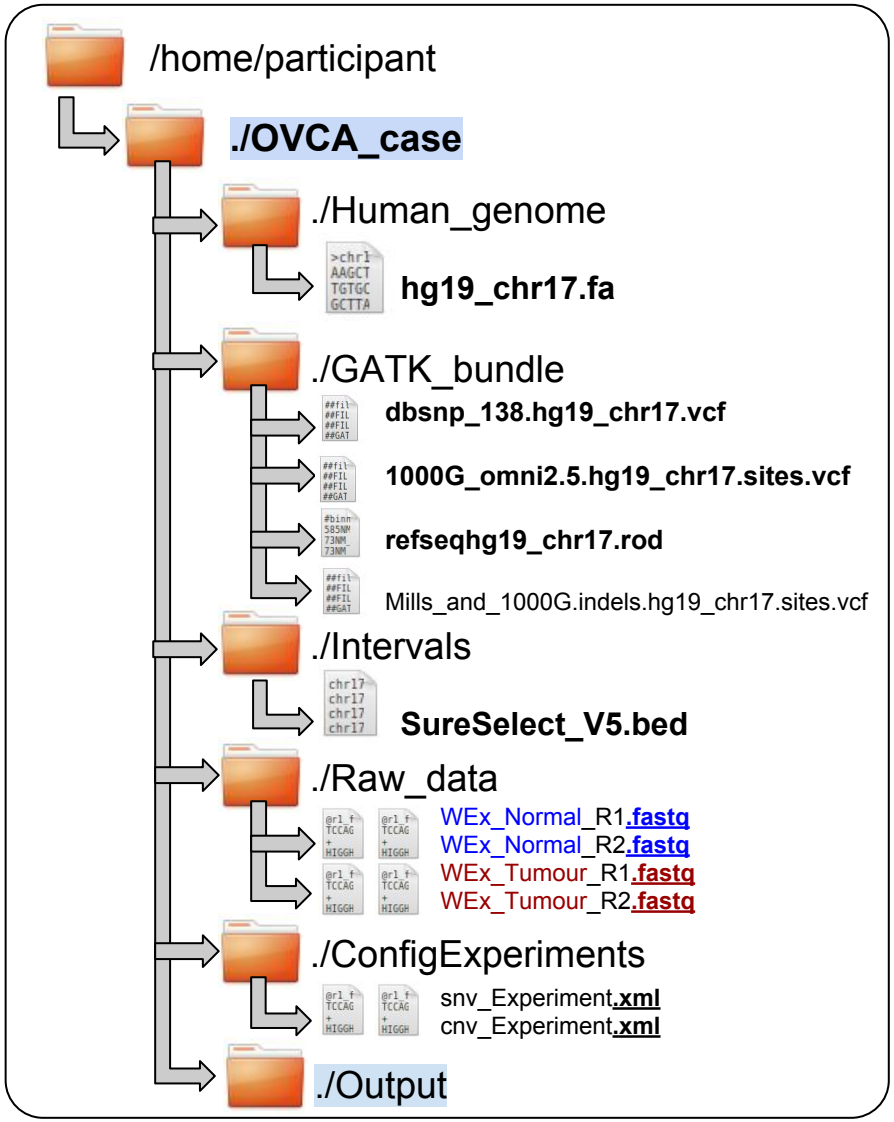
Library protocol: Agilent SureSelect V5 Human All Exons.

Sequencing platform: HiSeq 2000 (Illumina)

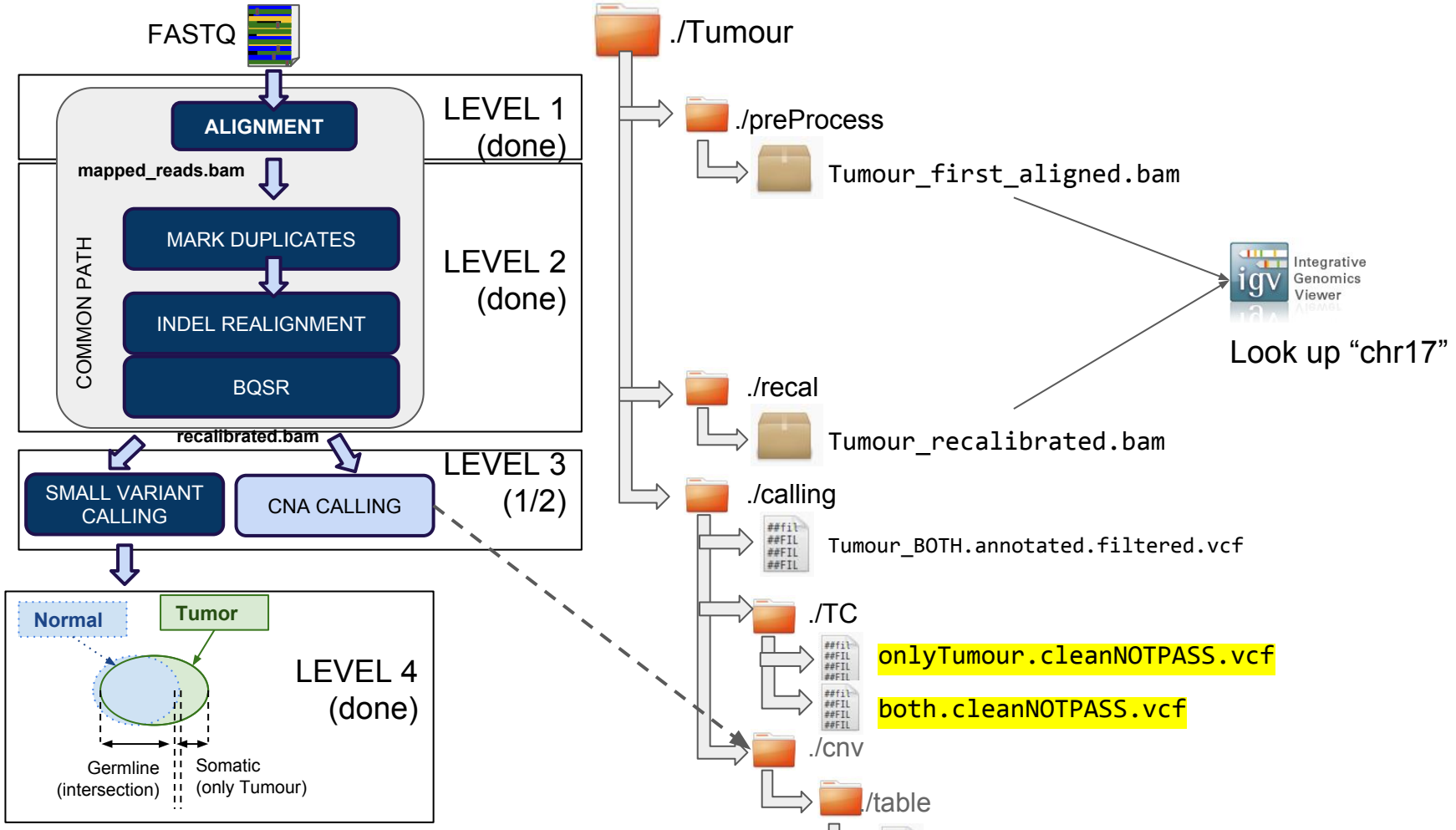
Expected outcome:

- ~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

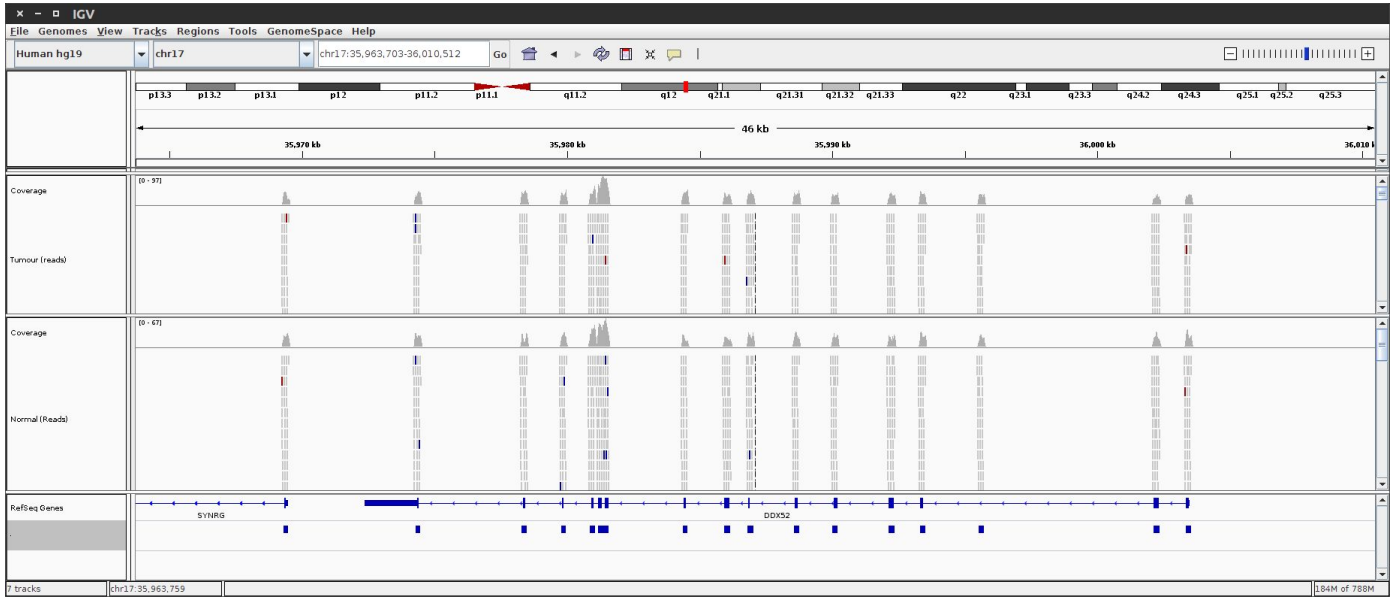


https://sourceforge.net/projects/rubioseq/files/LatestManual/

IGV : Genome Browser

open a terminal, and execute the cmd:

```
$ java -jar /home/participant/Software/IGV_2.3.66/igv.jar
```

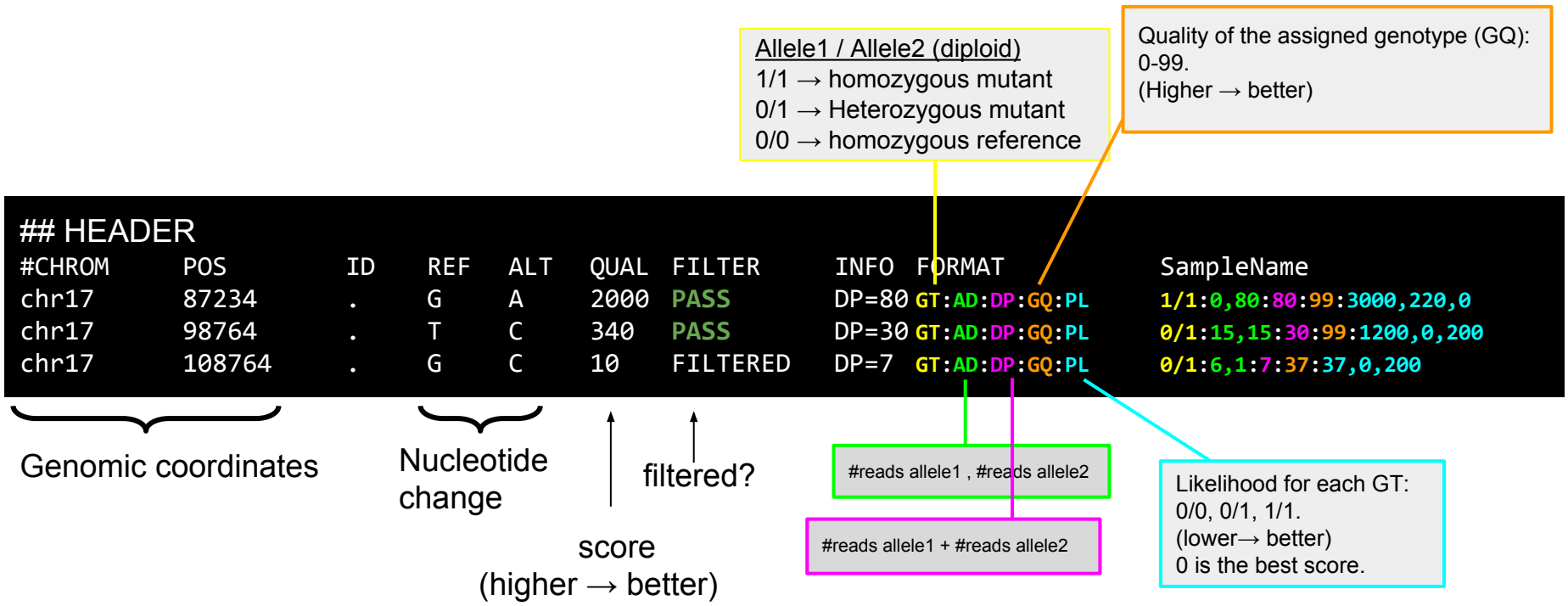


Open the BAM files from each sample:

`recalibrated.bam`

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

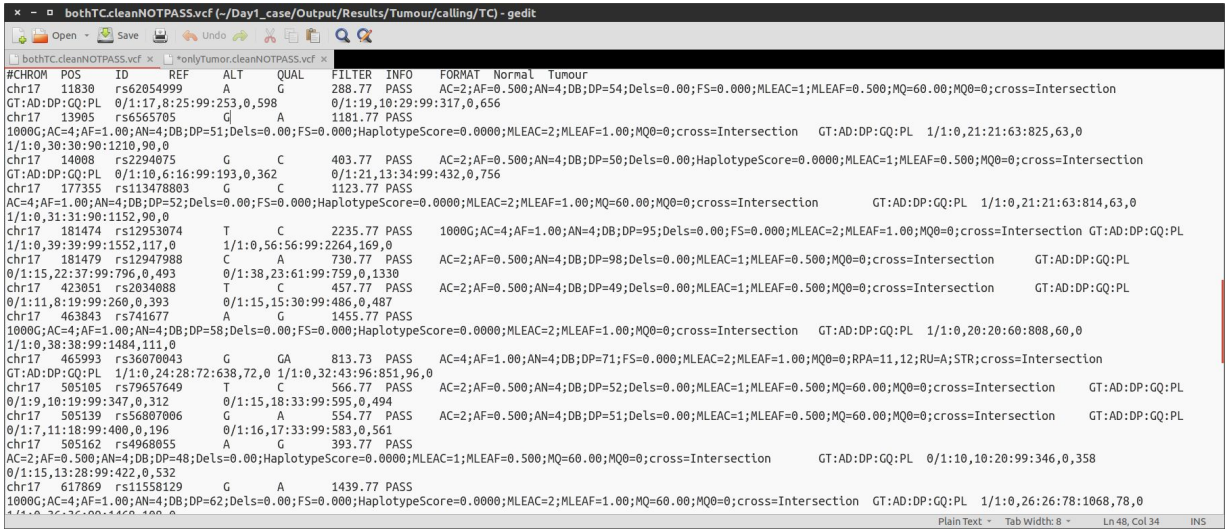
Variant Call Format (vcf) from GATK callers



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

*Case study :: **Point mutation results***

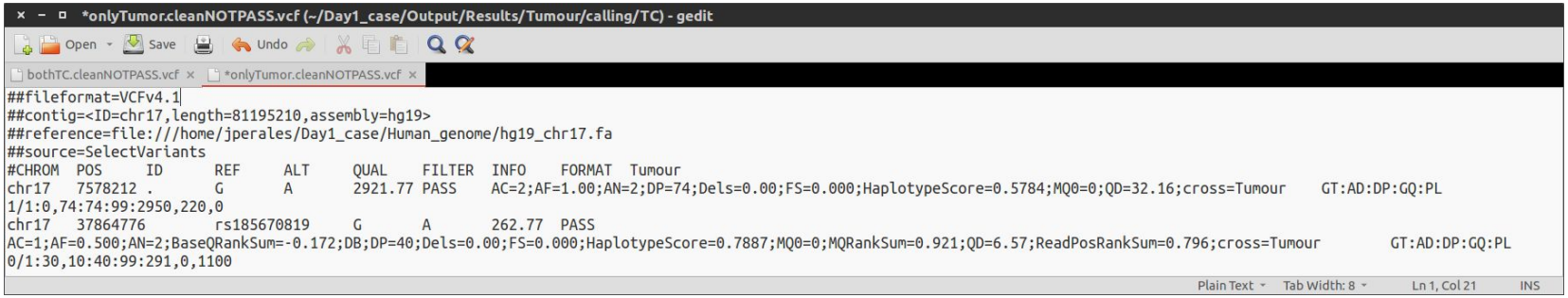
- Germline variants



There were detected 36 germline variants in total:

- 34 Single Nucleotide Variants.
- 2 Indels.

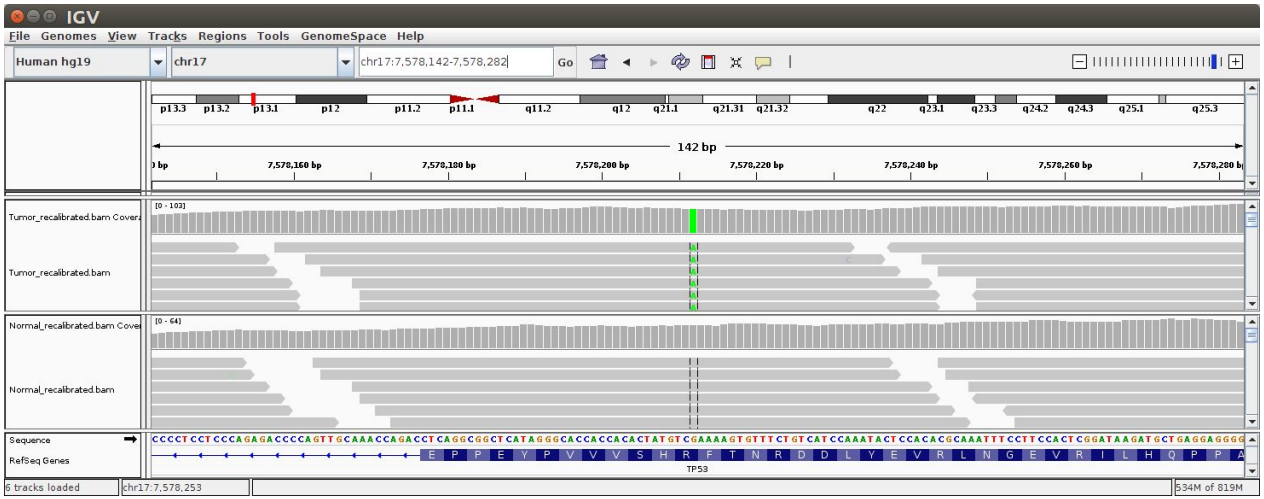
- Somatic variants



There were detected 2 somatic SNVs.

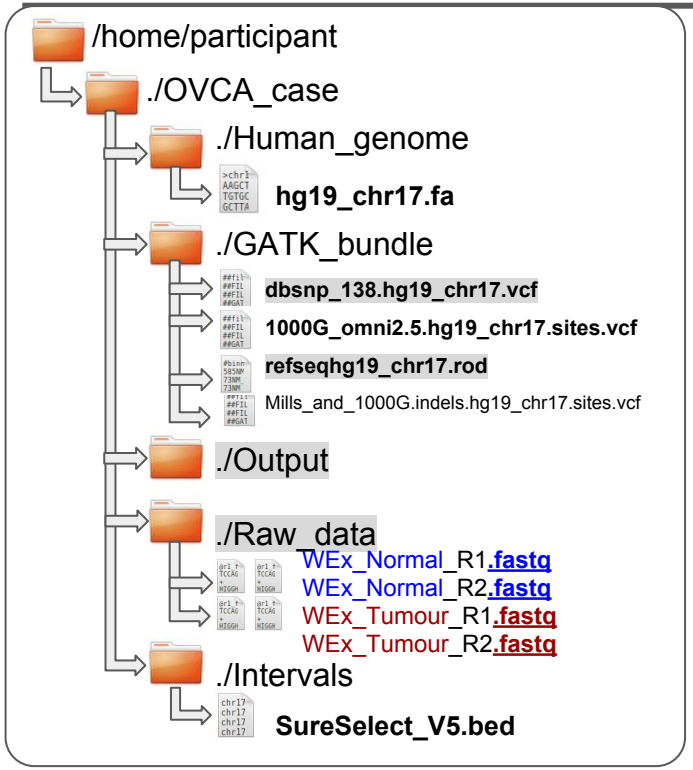
*Case study :: **Alignment of somatic mutations***

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




```
<?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>
```

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

```
$ perl /home/participant/Sotware/RUbioSeq+/RUbioSeq3.7/RUbioSeq.pl --analysis cnvCalling
    --config /home/participant/OVCA_case/ConfigExperiments/cnv_Experiment.xml --level 3
```

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

./cnv

./table

CNAtable.FILTERED.txt

	Chr	OriStCoordinate	OriEndCoordinate	Mean of LogRatio	Adjusted Mean of LogRatio	SD of LogRatio	Median of LogRatio	Number bases	P-Value	Adjusted P-Value	gain.loss	tumour.rd	normal.rd
NR_073509	chr17	20627912	20628001	2.96	3.28	0.224	2.958	89	3.70E-008	3.97E-005	gain	45.679	5.927
	chr17	30423938	30424028	2.698	3.09	0.337	2.692	90	7.81E-007	5.03E-004	gain	41.633	6.589
BRCA1	chr17	41243353	41245898	-7.014	-6.96	0.376	-7.098	2545	4.07E-021	1.74E-017	loss	0.529	67.81
	chr17	41245903	41246664	-7.084	-7.04	0.159	-7.098	761	1.87E-021	1.20E-017	loss	0.5	68.261
	chr17	41246666	41246844	-6.553	-6.17	0.339	-6.585	178	2.74E-023	3.53E-019	loss	0.5	48.186