**Computational Human Genomics Project**

Teamwork: Andrea Tonina, Lorenzo Santarelli, Gloria Lugoboni

1. **Project Rationale**

(max 10 lines)

Computational Genomic analysis on samples obtained from a patient. The aim of this project is to apply an example of human genomic workflow on patients characterized by tumor and control samples…. Bla bla spiegare meglio!!

1. **Computational Workflow**

(including details to make it reproducible)

Starting from the two provided BAM files (tumor and control DNA sequences from the same individual) we first analysed the general and detailed statistics using *samtools flagstat* and *samtool stats.*

We then operated a process of realignment, to identify all the positions in which hidden deletions are present. Specifically, this was possible via the GATK tool, that base its functionality on the default CIGAR line. First, *RealignedTargetCreator* is applied to the BAM file to identify which regions need to be realigned. Then, *IndelRealigner* performs the actual realignment at the target intervals. To perform this analysis we used as a reference the *human\_g1k\_v37.fasta,* an Homo Sapiens genome reference file that was built starting from the GRCh37 reference. We also provided a set of target regions in which to limit this protocol, specifically, we limited the analysis to a set of regions in chromosomes 15, 16, 17, and 18 (information contained in the file *Captured\_Regions.bed*).

A process of base quality score recalibration was operated to assign accurate confidence scores to each sequence, specifically adjusting the PHRED quality scores. Specifically, this process requires four steps that include the use of the *BaseRecalibrator* tool, that model the errors and generates a *BaseRecalibrator Table* that tells us how much correction is needed. The recalibrated data is then written to a BAM file thanks to the *PrintReads* tool. The process is then repeated to build the after model to evaluate the remaining errors and finally, it is possible to obtain the before and after plots via the tool *AnalyzeCovariates*. The *hapmap\_3.3.b37.vcf* database, containing information on known polymorphic sites, was selected as the one to exclude regions around known polymorphisms from analysis. We also explicated the argument *–emit-original-qual* to emit original base qualities and obtain a faster process.

we employed VarScan v2.3.9 to identify SNPs solely within the control sample's mpileup file, with a significance threshold set at a pvalue of 0.01. The output of this step was a VCF file specifically containing the detected SNPs. For a comprehensive analysis, we utilized VarScan v2.3.9 in somatic mode, taking both the tumor and control mpileup files as inputs. This allowed us to identify somatic variations (SNVs) present in the tumor sample. The output was another VCF file specifically capturing the SNVs. Additionally, we applied VarScan v2.3.9 in somatic mode to detect insertions and deletions (indels). This process resulted in a separate VCF file containing the identified indels. We employed `vcftools` to filter the VCF file containing SNVs, setting a minimum meanDP of 30. Subsequently, we annotated the somatic events using snpEff, generating a new VCF file for SNVs. Additionally, a detailed report in the form of an HTML file was generated. From the newly annotated VCF, we performed two additional annotations using SnpSift. The first annotation involved the hapmap\_3.3.b37.vcf file. Then, based on the output of this operation, we performed another annotation using clinvar\_Pathogenic.vcf. This process resulted in the creation of the somatic.pm.ann3.clinvar.vcf file. We further applied two filtering steps to this file. The first filter aimed to identify SNVs with a high impact, while the second filter focused on identifying clinically relevant variants.

identify all somatic events (e.g., SNV, CN), and define their clinical relevance. It is advisable to run quality checks, sample identity checks, visual inspections, and tumor population characterization. You are not required to use tools other than those utilized during the classes.

1. **Relevant Results and Interpretation**

NB graphs need to have a goooddd description

Both the files were associated to high average quality (>30) and a percentage of mapping major than 99.75%.

1. **Pitfalls and Criticism**

(max 10 lines).