**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 diseased patient. The aim of this project is to apply an example of human genomic workflow on a patient, characterizing both tumor and control samples. We carried out a computational workflow with the aim to characterize the presence or absence of amplifications, duplications, SNPs and SNVs.  
i  
i  
i  
i  
i  
I 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 marked the duplicates using the *MarkDuplicates* tool from Picard. This tool exploits the CIGARs to infer the presence of duplicates. Specifically, thanks to the argument *REMOVE\_DUPLICATES* set to true, it is possible to create a new .bam file without the duplicates found.

The process of variant calling was possible thanks to the tools *BCFTOOLS* and *GATK*.  
We combined (pipe |) *bcftool mpileup* with *bcftool call* to investigate the presence of variations. We explicated the parameters *-f*, needed for the human reference file in the FASTA format; the parameter *-v* (variants-only), to output variant sites only and finally the parameter *-c* (consensus-caller), to use the original samtools/bcftools calling method. We also operated the variant calling using GATK. We specifically used the tool *UnifiedGenotyper.* The obtained files where then analyzed using *vcftools*. We filtered the variants based on the quality, using a threshold for the minimum quality at 20 (--*minQ*) and a threshold for the minimum mean depth at 30 (*--min-meanDP)*.

We then proceeded the analysis operating a process of variant annotation, a crucial step in linking sequence variants with changes in phenotype. We specifically used SnpEff, a variant effect predictor program that is able to categorize each variant. We operated the annotation of both the *GATK* and the *BCF* .vcf files. Thanks to the tool *SnpSift* it was possible to add a layer of annotation to the variants. We specifically used two different files of annotations, *hapmap\_3.3.b37.vcf,* and *clinvar\_Pathogenic.vcf.* Specifically, the second file collects information on medical conditions with a genetic basis.

Thanks to a process of somatic variant calling we were able to identify SNPs and SNVs. This was possible by implementing in our workflow the tool *Varscan.v2.3.9*. We initially used the program *mpileup2snp* with a p-value threshold of 0.01.to generate a file containing SNPs found in the control sample. We then filtered this file using *vcftools,* applying the same thresholds as seen above.  
We then focused on the somatic point mutations, this was possible by specifying the setting *somatic* of the tool *Varscan.* For this passage it was necessary to input both the control and the tumor sample pileup files. We filtered again using *vcftools* and then we operated a process of annotation. This was possible via the tool *SnpEff*.

We operated a process of ancestry analysis to investigate the mixture of the genome of the patient. This analysis was possible thanks to the tool *EthSEQ*, specifically thanks to the package EthSEQ in R studio, via the command *ethseq.Analysis* and given in input the .vcf files obtained by *Varscan*, we were able to proceed with the analysis.

Applying an algorithm of *Circular Binary Segmentation (CBS)* we were able to obtain information on the somatic copy number abberations. In order to apply CBS we initially extracted information on the coverage susing *samtool mpileup* combined with the *Varscan copynumber* tool. The obtained output can be used to transform the coverage into information on the copy number (amplifications, deletions and homozygous deletions). This was possible using the tool *copyCaller* of *Varscan*. We finally used the Rscript CBS.R, also given to us by Professor Ciani and Professor Francesca Demichelis, to perform the Circular Segmantation. Specifically in this script the library *DNAcopy* is used and a process of segmentation is performed via the command *segment.*

We then estimated the purity and the ploidy of our sample. We decided to perform this analysis since it is important to understand if our sample is an admixture of clones and if the sample has ploidy greater than one. Thanks to the tool *CLONET,* which uses information on the log2R and the Beta value, it was possible to obtain information on the clonality of the tumor sample. Using the tool *TPES* it was possible to estimate the purity from SNVs data. We specifically used an RScript, given to us by Professor Ciani, called CLONET.R which use the library *TPES* and *CLONETv2* to operate an estimation of the purity and ploidy.

We finally decided to use the *SNP Panel Identification Assay (SPIA),* a Rstudio package that determines cell line identities starting from data from SNPs. Specifically, we used a Rscipt that was given to us by Professor Francesca Demichelis and called *esercitazione\_SPIA\_code\_to\_upload.R*.

1. **Relevant Results and Interpretation**

**Pre-processing**

Both the files were associated to high average quality (>30) and a percentage of mapping major than 99.75%. During the step of pre-processing we were able to prepare the .bam file needed for the future analysis. Specifically we were able to realign a total of 2267 reads in the Tumor sample and 3158 reads in the Control one. We also corrected the quality of the reads and eliminated the found duplicates. Specifically, we identified and removed 13.8% of duplicates in the Control and 12.2% in the Tumor.

**Variant Calling and Variant Annotation**

We restricted the analysis on variants to the chromosomes 15, 16, 17 and 18. We found a total of 7,886 variants in the Tumor sample, majorly concentrated in the chromosome 17 (3027 variants). All these variants were classified and SNPs, the majority (73.2%) linked to a not-known/predictable effect (modifier effect). Only the 0.2% of the variants were connected to an with high putative impact/deleteriousness. The functional class of these SNPs is almost equally divided between missense and silent mutations (ratio missense/silent: 0.8148). Only the 0.43% of the SNPs were connected to a nonsense mutation. The majority of these variants is found in the exon and intron regions (27.9% and 25%), also in the downstream region (19.49%). This support the fact that the tool was not able to predict entirely the impact of these variants (since they are mainly in regions from which proteins are not generated, rather they are in reions that could control the quantity of protein produced). The ratio between transition and transversion is equal to 2.6, the majority of observed base changes are in fact transitions G->A and C->T. DA GUARDARE DETAILS BY GENE

Regarding the Control sample, we found a total of 9036 variants, mainly concentrated in chromosome 17 (3477 variants). These variants were all classified as SNPs, primarily connected to a not-known impact (modifier: 71.6%) Only a 0.2% of the found SNPs were connected to a high impact. We also found that a missense/silent ratio equal to 0.81 and a percentage of nonsense mutation equal to 0.43%. The variants were found in exon regions (29.4%), intronic regions (27.4%), and regions downstream (16.8%). The ratio between transitions and transversion is equal to 2.63 and the most registered transitions are transitions G->A and C->T. DA GUARDARE DETAILS BY GENE

GUARDIAMO SU IGV SOVè BRCA? CHIEDE DI SAPER MUOVERSI SU IGV

**Somatic Variant Calling**

We initially operated a process of definition of the SNPs for the Control sample. Specifically, after the process of filtering, 14936 SNPs were retained. Of these, 10393 sites were common to the one identified in the previous analysis using bcftool.  
Regarding the Tumor sample, we found 3164 mutations identified as LOH, 289 as somatic mutations and 60 as unknown mutations. These were then filtered and only 10261 sites were retained. The majority of these mutations were connected to an unknown impact (modifier: 82.174%), while only 0.132% of these were connected to a high impact. The majority of these mutations were silent or missense (55.2% and 44.5%) and were found in intronic regions (40.7%). Transitions were mainly registered (ratio transition/transversion: 2.43), primarily transitions G->A and C->T.

**Somatic Copy Number Calling**

From the analysis of the Somatic Copy Number we found that 4499 regions were called amplification, characterized by a log2 > 0.25). 30175 regions were called neutral, and 90323 regions were called deletion, characterized by a log2 <-0.25. Only 103 regions were identified as homozygous deletion. Performing the Cirular Segmentation it was possible to define segments of positions characterized by a specific copy number. We can see that the most represented event is the deletions (loss).

**Ancestry Analysis**

**Purity and Ploidy Estimation**

Thanks to CLONET and TPES we were able to estimate information on the ploidy and the purity of our sample. Specifically, from CLONET we were able to infer the ploidy (clones and subclones present in our sample), estimated to be equal to 2.34 and the value of admixture, estimated to 0.38. We also obtained a LogR-Beta plot. Da spiegare bene il plot, per spiegarlo guarda le lezioni di teoria.

TPES was used to infer information on the purity of our sample. We spacifically obtined two plots, one refiguring the

DA COMMENTARE I GRAFICIII

**SPIA**

NB graphs need to have a goooddd description

Grafici che potremmo mettere:

CLONET E TPES di sicuro

* Magari il Circular Segmentation
* IGV
* Al massimo SPIA se funziona

1. **Pitfalls and Criticism**

(max 10 lines).

Un singolo paziente

Qualità super alta

Analisi piu semplice perché i dati iniziali sono ottimi