Computational Human Genomics Project

Teamwork: Andrea Tonina, Lorenzo Santarelli, Gloria Lugoboni

1. Project Rationale

The aim of this project is to apply an example of human genomic workflow on a patient, characterizing both tumor and control samples. Thanks to the applied workflow we were able to characterize germline and somatic variants, determine the ancestry of the patient, and study the tumor ploidy and purity.

2. Computational Workflow

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

A realignment process was performed to identify all the positions in which hidden deletions are present. This is possible via the GATK tool, which bases its functionality on the default CIGAR line. First, *RealignedTargetCreator* is applied to the BAM file to identify which regions must be realigned. Then, *IndelRealigner* performs the actual realignment at the target intervals. *human_g1k_v37.fasta* is used as Homo Sapiens genome reference in the analysis. The analysis was also limited to a set of target genome regions, specifically, to chromosomes 15, 16, 17, and 18 (information contained in the file *Captured_Regions.bed*).

A process of recalibration was operated to assign accurate quality scores to each sequence, adjusting the PHRED quality scores. This process requires four steps that include the use of the *BaseRecalibrator* tool, which models the errors and generates a *BaseRecalibrator Table* that contains information on the corrections 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 the analysis.

The duplicates were individuated using the *MarkDuplicates* tool from Picard. This tool exploits the CIGARs to infer the presence of duplicates. Thanks to the argument *REMOVE_DUPLICATES* set to true, it is possible to create a new .bam file without the found duplicates.

The process of variant calling was possible thanks to *BCFTOOLS* and *GATK*.
bcftool mpileup was combined (pipe |) with *bcftool call* to investigate the presence of variations. We also operated the variant calling using GATK. The tool *UnifiedGenotyper* was used. The obtained files were then analyzed using *vcftools*. The variants were filtered 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)*.

The process of variant annotation, a crucial step in linking sequence variants with changes in phenotype, was effectuated using *SnpEff*. Thanks to *SnpSift*, it was possible to categorize each variant. We operated the annotation of both the *GATK* and the *BCF* .vcf files. We specifically used two different files of annotations, *hapmap_3.3.b37.vcf,* and *clinvar_Pathogenic.vcf.* The second file collects information on medical conditions with a genetic basis.

Via the process of somatic variant calling, using the tool *Varscan.v2.3.9* it is possible to identify SNPs and SNVs. A p-value threshold of 0.01 was used in *mpileup2snp* 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. Specifying the setting *somatic* of the tool *Varscan* it is possible to focus on somatic events. For this passage, it is necessary to input both the control and the tumor sample pileup files. The output was filtered and annotated using the tool *SnpEff*.

To investigate the mixture of the genome of the patient an ancestry analysis was performed. This analysis was performed using the package *EthSEQ* in R studio. The command *ethseq.Analysis,* which requires as input the .vcf files obtained by *Varscan*, was used for the analysis.

Applying an algorithm of *Circular Binary Segmentation (CBS)* it was possible to obtain information on the somatic copy number aberrations. First, information on the coverage is extracted using *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 is possible using the tool *copyCaller* of *Varscan*. We finally used the R library *DNAcopy* to perform a process of segmentation via the command *segment.*

The purity and the ploidy of the sample were estimated using the tools *CLONET* and *TPES.* CLONET uses information on the log2R and the Beta value to obtain information on the clonality of the tumor sample. *TPES* estimates the purity from SNVs data. Thanks to the R library *TPES* and *CLONETv2* it is possible to estimate the purity and ploidy.

Thanks to the R package *SPIA* (*SNP Panel Identification Assay)* it was possible to determine cell line identities starting from data from SNPs.

3. Relevant Results and Interpretation

Pre-processing

Both the files were associated with high average quality (>30) and a percentage of mapping major than 99.75%. During the step of pre-processing, the .bam files needed for future analysis were obtained. A total of 2267 reads in the Tumor sample and 3158 reads in the Control sample were realigned. The quality of the reads was corrected and the duplicates were eliminated.

Variant Calling and Variant Annotation

The analysis was restricted to chromosomes 15, 16, 17, and 18. Regarding the Control sample, a total of 9036 variants were found, 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 0.2% of the found SNPs were connected to a high impact. The missense/silent ratio was equal to 0.81 and the percentage of nonsense mutation was 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 was equal to 2.63 and the most registered transitions were transitions G->A and C->T.
Thanks to annotation analysis it was possible to identify a BRCA1 mutation already known (**Fig 1**). A point mutation found in position 41246494, characterized by the C->A transversion, associated with the familiar breast-ovarian cancer, leading to the hereditary cancer-predisposing syndrome. It is a nonsense mutation causing the generation of a non-coding transcript variant.

Somatic Variant Calling

Regarding the Tumor sample, 3164 mutations were identified as LOH, 289 as somatic mutations, and 60 as unknown mutations. These mutations were filtered and only 10261 sites were retained. Most of these mutations were connected to an unknown impact (modifier: 82.174%), while only 0.132% of these were connected to a high impact. Most 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, 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. Performing the Cirular Segmentation it was possible to define segments of positions characterized by a specific copy number. As shown in **Figure 2**, the most represented events are deletions (loss).

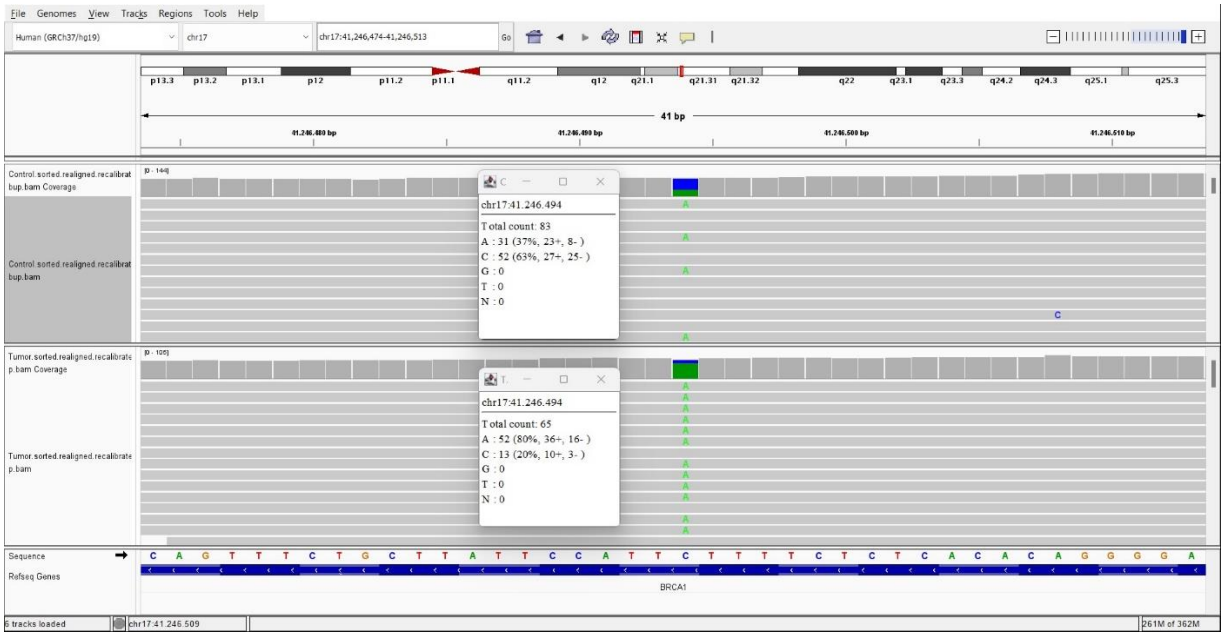

Figure 1: Figure obtained from IGV and showing position 41246494 on chromosome 17, a position found inside the gene BRCA1 that appears mutated both in the control sample and the tumor sample. Specifically, the control sample (above) is characterized by a mutation in 37% of the reads (mutation C->A) while the tumor sample (under) presents the same mutation in 80% of the cases.

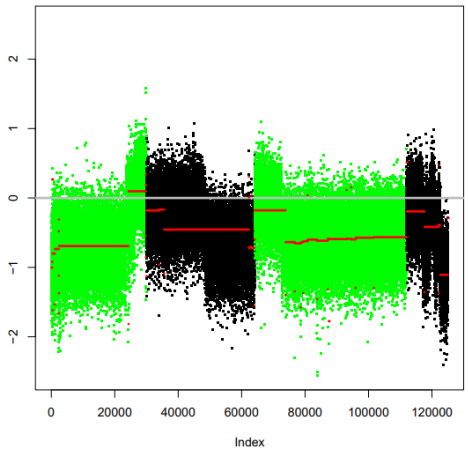

Figure 2: Plot showing the Circular Binary Segmentation results. The x-axis represents the genomic positions while the y-axis shows the Log2Ratio. In red (lines and dots) we have the Log2 ratio value for each segment. It is possible to observe that the majority of the segments have a Log2 ratio value lower than zero, indicating that deletions occurred. A homozygous deletion is associated with Log2 R values at circa -2 while a hemizygous deletion is centered in -1. The effect of contamination of the tumor sample with normal cells causes biases in the computation of the Log2 R. We need to take into consideration factors such as tumor purity and tumor ploidy, therefore we can think that a shift of the distribution of the Log2R toward the zero take place. Specifically, the Log2 of deletions ranges between -0.5 and -1.5. Taking into consideration this information, we can say that our tumor sample is characterized by the presence of deletions and, in a smaller percentage, gains. These results match what we were able to define in the previous step of the analysis.

Ancestry Analysis

Thanks to the ancestry analysis we were able to retrieve information on the ethnicity of the sample. The patient is of African ethnicity. The complete stratification that was obtained is the following: EUR(19.84%) | EAS(19.56%) | SAS(18.05%) | AFR(42.55%).

Purity and Ploidy Estimation

Thanks to CLONET and TPES we were able to estimate the ploidy and the purity of the sample. CLONET (**Fig. 3**) is used to infer the ploidy, estimated to be equal to 2.34, and the value of

admixture, estimated to be 0.38. From this data, it is possible to understand that the sample contains a portion of normal cells. The global content of DNA (ploidy) is above what we consider normal. TPES (**Fig. 4**) is used to estimate the purity of the sample based on the shift between the observed major peak (observed VAF) at 0.335 and 0.5 (expected VAF for a pure tumor sample). It is also possible to detect the presence of subclonal events (observed peak at 0.205).

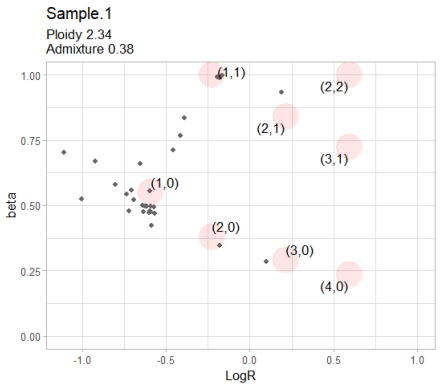

Fig 3. LogR-Beta Plot obtained from CLONET. Each gray dot represents a genomic segment. We can observe that the majority of the points are around (1,0), indicating an event of hemizygous deletion, surrounded by subclonal events. It is possible to observe the presence of amplification events (2,1), as we observed in previous analysis, these represent only a small percentage. We can also define complex events, specifically, clones in which a deletion and a gain occurred, like (2,0) and (3,0), also known as CN-LOH and Gain-LOH. We can also define a small cluster of points (three points) in (1,1) in which no changes in the copy number occurred, these could represent the non-tumoral cells that bring our admixture value to 0.38.

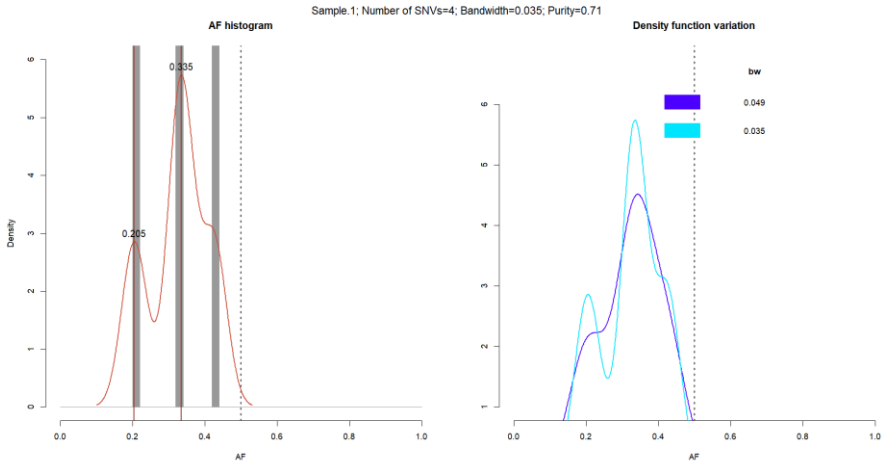

Fig 4. Plots representing the AF distribution (histogram) of clonal and subclonal SNVs (4.a) and the density variation of the AF distribution based on a smoothing correction (4.b) obtained from TPES. In Figure 4.a it is possible to observe the shift of the AF from 0.5 to 0.335 which indicates the presence of normal cells inside the sample.TPES was able to estimate the purity which is equal to 0.71. It is possible to observe a peak around 0.205 which indicates sub-clonal events.

SPIA

From SPIA analysis, based on 13284 SNPs, it was possible to estimate the genetic distance between the tumor and the control samples. From the pairwise comparison, we obtained a distance of circa 0.1. The classification results in a match of the two samples, indicating that the analyzed SNPs are similar in both the control and the tumor samples. Indeed, the two samples were obtained from the same patient.

4. Pitfalls and Criticisms

Further analysis are needed to have a more complete overview of the sample and the tumor. It could be interesting to analyze the familiar condition or obtain familiar clinical records since the point mutation in BRCA1 is connected to a familiar syndrome. It could also be optimal to extend the analysis to the whole genome since we focused on a subset of chromosomes (15-18).