Genomic characterization of a cancer sample

Nicola Perotti (240191), Chiara Rosati (239094), Thomas Sirchi (239007)
Supervisors: prof. Francesca Demichelis, dr. Yari Ciani

1.0 Project Rationale

We started from two BAM files containing alignment information for a control and a tumor sample from the same individual (The Cancer Genome Atlas reference: TCGA-BRCA/TCGA-A7-A4SE) instead of considering an entire cohort. Indeed, we can still obtain significant results even from a single patient. The following analysis we aim to detect somatic events to characterize the cancer genome and to define how these aberrations influence the onset of the disease. At first, we performed pre-processing steps, for instance realignment around indels, recalibration of base qualities and duplicates removal. Then, we identified Single Nucleotide Variants (SNVs) and Somatic Copy Number Variations (SCNVs). Variant calling (SNP detection) was also performed on control to better distinguish between real somatic variants and SNPs wrongly identified as SNVs. As a final step, we estimated purity and ploidy for the tumor sample.

2.0 Computational Workflow

2.1 Pre-processing steps

All following analyses have been conducted on the regions from chromosome 15 to 18. We sorted and indexed the two BAM files which then have been realigned around indels using GATK3 to optimize further analysis. Regions to be realigned have been selected through RealignerTargetCreator with *humanG1Kv37* as reference. The actual realignment was performed using IndelRealigner with previous parameters and the output file from the preceding command.
Next, we recalibrated base qualities using GATK and setting BSQR as a method for Phred quality score adjusting. We used BaseRecalibrator twice, at first to obtain a recalibration table, then we computed recalibration scores. We used hapmap_3.3.b37.vcf as known polymorphic sites to be excluded. Then we applied recalibrated scores to the aligned reads through PrintReads and we used AnalyzeCovariates to check how qualities changed after recalibration.
The final step for pre-processing consisted in the removal of duplicates, we used MarkDuplicates from picard setting REMOVE_DUPLICATES and ASSUME_SORTED as true.
The tumor and control samples ethnic identities were characterized using EthSEQ, an R package which annotates individuals' ancestry.

2.3 SNPs and somatic variant calling

In our variant detection workflow, we started by generating two mpileup files using samtools. These files represented the tumor and control samples. Subsequently, we employed VarScan v2.3.9 to identify SNPs solely within the control sample's mpileup file, with a significance threshold set at a p-value 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.

2.4 Copy Number Variations

We detected somatic copy number aberrations (SCNAs) using VarScan.v2.3.9 set on copynumber and giving both tumor and control mpileup files as inputs. We obtained a file with information about copynumber events. Then VarScan.v2.3.9 was set on copyCaller to adjust previous results and classify somatic events into gain, loss or neutral. The resulting filename is Somatic.copynumber.called. Starting from Somatic.copynumber.called, we used functions from the R package DNAcopy to obtain segmentation plots.

2.6 Purity-Ploidy Estimation

We gathered all heterozygous SNPs from the control VCF file into a new file named Control.het.vcf. Then we used ASEReadCounter from GATK to compute read counts for alleles in heterozygous SNPs (both in tumor and control) setting a minimum threshold for mapping quality, base quality and depth of 20. We obtained two CSV files, one for tumor and the other one for control containing information about chromosome, position, alternative and reference alleles and coverage.
Eventually, we used *CLONET* and *TPES* in an R script. Using the two *CSV* files we computed Allelic Fraction (AF). From the *SCNA* analysis we obtained a file with information about segmentation which has been used, together with the two CSV files, to compute beta values.
Then, beta values have been used to compute ploidy and admixture. We also computed the corrected log2(R) to show the Log(R) beta plot to assess clonality and ploidy. TPES has been used to assess tumor purity from SNVs.

3.0 Results and Interpretation

The overall coverage for the control sample is 15306.95604 whereas for the tumor sample the coverage is 15172.57443. During data pre-processing 1.771.859 duplicates in control and 1.165.572 in tumor were marked and removed from the original file. According to the analysis conducted by EthSEQ, it has been determined the samples provided are primarily of African origin (CONTROL-EUR(15.54%)|EAS(14.72%)|AMR(16.4%)|SAS(14.83%)|AFR(38.51%),TUMOR-EUR(15.84%)|EAS(14.64%)|AMR(16.34%)|SAS(15.21%)|AFR(37.97%))

3.1 Somatic variant calling

Variants were found in chromosomes 15, 16, 17 and 18 for a total of 14,524 SNVs. The majority of them has been considered as “modifier” (low confidence) and only 0.132% of all variants has been considered of high impact. Almost all variants have been called as missense (44.468%) and silent (55.155%) whereas only a very small percentage of variants is nonsense (0.377%).
Variants are especially concentrated in intron regions (41.599%) and downstream regions (16.283%).

Transitions have higher frequency compared to transversions (Ts/Tv ratio of 2.4263). In particular the main transition is from G to A even though transitions from T to C are only slightly lower than the first one. In 3174 mutations we see a loss of heterozygosity (LOH) of which 169 are indels and the rest are SNPs. In addition we found that 60 mutations were classified as unknown (47 indels and 13 SNPs).

Events were annotated through SnpSift, 46 variants were classified as high impact, while one variant was found to be clinically significant by CLINVAR: a nonsense mutation in chromosome 17 affecting the gene BRCA1, which is involved in the Hereditary breast and ovarian cancer syndrome (Fig. 1).

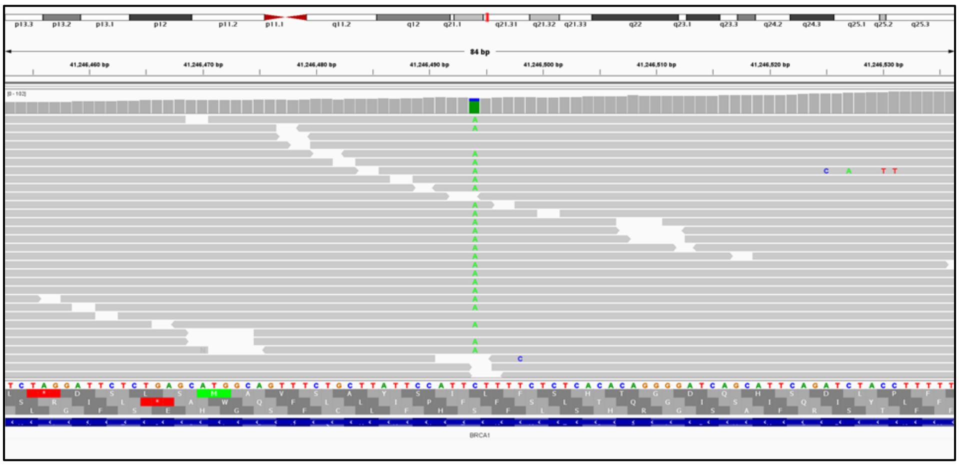

Fig. 1: IGV visualization of BRCA1 gene. A transversion from C to A can be seen (position chr17:41,246,453-41,246,536). Coverage for chr17:41,246,494 : Total count of 65, 52 for A (80%), 13 for C (20%).

3.2 Somatic Copy Number Calling

From VarScan we obtained an average log2R of -0.49, so we can infer that mainly deletion events happened. Particularly 4499 events were classified as gain, 30175 as neutral and 90323 as loss. These results are supported by the segmentation plot in which we mainly observe segments with negative log2R (Fig. 2).

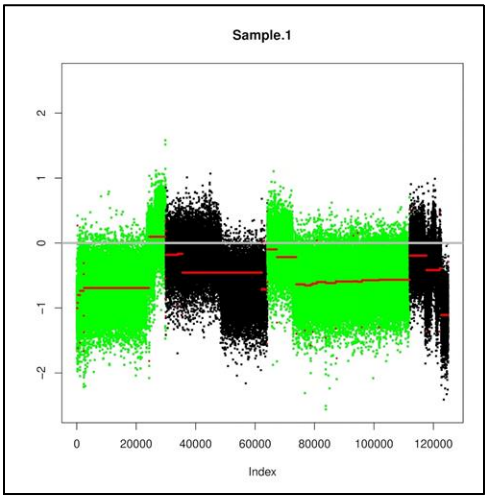

Fig. 2: Segmentation plot showing genomic segments and their Log2Ratio. Mainly losses can be seen, this is specifically the case of hemizygous deletions. The corresponding Log2Ratio for hemizygous deletions should be -1, but we suppose that the contribution of normal cells in the sample slightly shifted the region toward zero (for purity estimation see “Purity and Ploidy Estimation” paragraph).

3.3 Purity Ploidy Estimation

Purity and ploidy estimation are fundamental measurements when characterizing a tumor sample. Purity correction shifts the AF distribution towards zero whereas ploidy correction increases the distance between peaks of the distribution.

CLONET gave as output a log2R-beta plot and estimations for admixture and ploidy (Fig. 3A). Admixture is 0.36 and ploidy is 2.27, thus slightly above what we consider normal. We have previously seen that our sample has mainly losses even though ploidy is higher than 2. We need to consider that our sample is composed not only of cancer cells, but also of normal cells (admixture=0.36) which might contribute a bit to increase the value. Moreover, we see that two points are on the right side of the plot (log2R>0) where gains are found. The majority of points cluster as hemizygous deletions and the surrounding cloud of points can be considered as made of subclonal events. Three points are clustered as (1,1), thus no change in copy number occurred. Only one point can be found at (2,0) and is considered a copy-neutral loss of heterozygosity.

Tumor purity has been estimated through TPES as well which is based on SNVs instead. As previously assessed by CLONET, TPES found a similar tumor purity of 0.71. Looking at the plot on the left, we see that the distribution has its main peak in 0.335 so we have a purity shift of 0.165 (Fig. 3B). This value is proportional to tumor content. Moreover, a second peak can be identified at 0.205 suggesting a subclonal population.

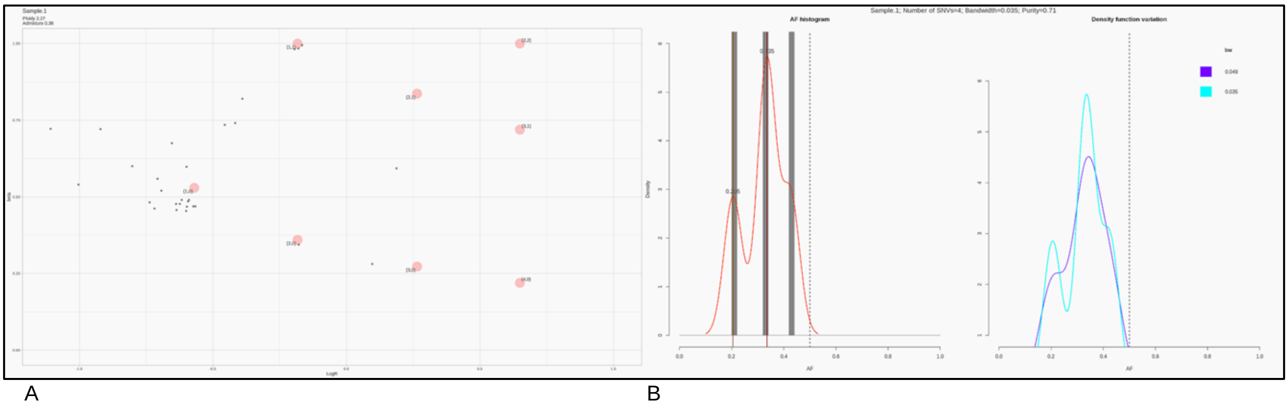

Fig. 3: (A) Clonet plot. The majority of segments cluster around (1,0) suggesting hemizygous deletions. A small cluster can be seen in (1,1) and a single segment in (0,2) which suggests a copy-neutral LOH. Ploidy of 2.27 and admixture of 0.36. (B) TPES plot. The Allelic Fraction (AF) distribution can be seen. The main peak is in 0.335 and its distance from AF=0.5 is proportional to the tumor content. Another peak can be identified in 0.205 corresponding to a subclonal event.

4.0 Pitfalls and criticisms

Future analysis could focus on the entire genome. A targeted analysis is desirable to reduce computational cost and that might be performed identifying an informative region within which it is known crucial information can be retrieved.