July 18th, 2025

Computational Human Genomics Project

Maccarone Giulia, Rizzolo Maria Luce, Salsano Anna, Sanetti Caterina

Project Rationale

In oncology, large-cohort genomic studies are the gold standard to identify clinically actionable patterns in cancer. However, single-patient analyses remain critical for precision medicine, enabling the characterization of patient-specific aberrations that may drive tumorigenesis. The aim of this project is to characterize the somatic genomic landscape of an oncological patient, through computational analysis of matched tumor and normal DNA samples, defining how these aberrations influence the disease and their potential clinical relevance.

Computational Workflow

Input files used were two sorted and indexed BAM files containing aligned sequencing reads from two distinct DNA samples from the same patient: a tumor and a matched normal control. The reference genome used for all alignment and analysis steps was *human_g1k_v37.fasta*. Our analysis began with a data pre-processing phase to refine the alignment data and correct for systematic errors, ensuring the highest quality input for downstream variant detection. The first step was to address alignment artifacts around small insertions and deletions (indels). We used the *Genome Analysis Toolkit (GATK) RealignerTargetCreator* tool on the input BAMs to first identify genomic intervals susceptible to misalignments. Subsequently, the *IndelRealigner* tool performed a localized de novo realignment of reads within these target intervals for both samples, generating more accurate read placements. All pre-processing and analysis steps were focused on specific genomic areas defined by the *Captured_Regions.bed* file. We implemented *GATK*'s *Base Quality Score Recalibration (BQSR)* to correct for systematic sequencing errors. The *BaseRecalibrator* tool created an error model by comparing our data to known polymorphic sites in *HapMap 3.3*. This model was then used by *PrintReads* to rewrite the quality scores in the BAM files. Finally, we used *AnalyzeCovariates* to generate plots that verified the effectiveness of the correction. As a final pre-processing step, PCR duplicates, which are non-independent reads that can bias allele frequencies, were discarded entirely using *Picard's MarkDuplicates* tool. The resulting clean BAM files were then indexed with *samtools* to enable rapid retrieval of data from any genomic region. To identify somatic copy number alterations (SCNAs), we generated a pileup with *samtools mpileup* and processed it with *VarScan2* to compare read depths between the tumor and normal samples. The output was then segmented using the *Circular Binary Segmentation (CBS)* algorithm. Finally, to interpret these results, we annotated the copy number aberrations by intersecting the *SCNA.copynumber.called.seg* with a curated list of known cancer-related genes (*CancerGenesSel.bed*). In parallel, we independently called germline variants on each sample using both *bcftools* and *GATK*'s *UnifiedGenotyper*. After filtering for high-confidence SNPs *(minQ 20, minDP 30*), we assessed concordance between the callers and then functionally annotated all variants using *SnpEff* and *SnpSift*. The primary goal of identifying somatic mutations was achieved using *VarScan2*'s *somatic* calling mode. After generating separate pileup files from the final processed BAMs using *samtools mpileup*, we ran *VarScan2 somatic* to compare the tumor and normal pileups directly, identifying variants present at a significant frequency in the tumor but absent or at very low frequency in the normal sample. This produced a VCF file of high-confidence somatic SNPs and indels, which was subsequently annotated. To determine the tumor's purity, ploidy and clonal structure, we integrated our SCNA results with allele-specific read counts using *CLONET.R*

script. An initial attempt to generate these counts from control heterozygous SNPs was unsuccessful due to the limited number of such sites in our tumor panel. We therefore used the *HapMap 3.3* VCF as a source of known polymorphic sites for *GATK*'s *ASEReadCounter*. This provided a much denser set of data points, which, after applying strict filters *(minDP 20, minMQ 20, minBQ 20*), was used to accurately estimate the tumor's composition. Within the *CLONET.R* script, we also conducted a *Tumor Purity Estimation* from SNVs *(TPES)* analysis. The analysis concluded with a mutational signature analysis on the somatic SNVs with the *COSMIC SigProfiler* tool to infer the biological processes that drove tumorigenesis. Then, we visualized specific genomic events in the *Integrative Genomics Viewer (IGV)* to manually validate the supporting read evidence. For deeper context, we then interrogated a key mutated gene in the *cBioPortal* database to assess its clinical relevance.

Results

Preprocessing

Initial alignment assessment confirmed high data quality, with over 99% of reads mapped and properly paired for both tumor and control samples. A significant disparity in sequencing depth was observed, with the control library containing approximately 31% more raw sequences than the tumor (19.7M vs. 15.0M reads). Following this, data refinement began with local realignment, which targeted and corrected 6,562 tumor and 8,048 control genomic regions. *BQSR* then successfully mitigated systematic sequencing errors, evidenced by the convergence of reported and empirical qualities (mismatch quality corrected from Q29.1 to Q29.0 in the tumor). Finally, the deduplication results show excellent library quality, with very low and consistent duplication rates of 3.85% for the control and 4.30% for the tumor sample, indicating high data reliability for downstream analysis.

Informative Somatic Events

Somatic event analysis revealed a highly unstable tumor genome, defined by both critical SNVs and extensive SCNAs. Out of a total of 14,251 SNVs detected across the samples, 168 were identified as somatic mutations. Functional annotation of these somatic events with *SnpEff* showed that the majority were located in non-coding regions, with 31.5% classified as intronic and 15.5% as intergenic. *SnpEff* also highlighted a small but critical subset of these mutations, classifying four as having a *HIGH* predicted impact on protein function. Among these, the most significant finding was a high-impact splice acceptor variant in the tumor suppressor gene *TP53* in *chr17*, which exhibited the highest tumor allele frequency of all high-impact mutations at 65.85%. Additionally, three other variants were classified as high-impact, including two in *MYO5B* and one in *GCGR* (**Table 1**)**.** The investigation of SCNAs through segmentation analysis revealed numerous large-scale deletions affecting key cancer-related genes, generally highlighting log2ratio ~-1, suggesting hemizygous deletions and with some probability of being subclonal events. The most significant losses (**Table 2**) were observed on *chr16*, which included a focal deletion containing *CBFA2T3* (*seg.mean* = -0.71) and on *chr15*, where a ~53.7 Mb deletion (*seg.mean* = -0.69) spanned a series of relevant genes including *PML* and *TP53BP1*. *Chr17* displayed multiple distinct deletion events, leading to the loss of critical tumor suppressors such as *BRCA1* (*seg.mean* = -0.59) and genes from key signaling pathways like *ERBB2* (*seg.mean* = -0.62). In contrast to these strong losses, the analysis identified only one region of slight copy number gain on *chr15* (*seg.mean* = +0.097), which included the *IGF1R* gene.

1

Chromosome	Position	Gene Name	Change	Variant Type	Impact	%Tumor Allele Freq
chr17	7577156	TP53	C>A	splice_acceptor_variant&intron_variant	HIGH	65.85
chr17	79768908	GCGR	G>C	structural_interaction_variant	HIGH	59.21
chr18	47363931	MYO5B	G>C	structural_interaction_variant	HIGH	23.26
chr18	47364009	MYO5B	G>A	structural_interaction_variant	HIGH	24.10
chr15	20739739	GOLGA6L6	A>G	missense_variant	MODERATE	23.4
chr15	20740192	GOLGA6L6	G>A	missense_variant	MODERATE	28.63
chr15	82934374	GOLGA6L10	C>G	missense_variant	MODERATE	25.0
chr15	82934619	GOLGA6L10	A>G	missense_variant	MODERATE	20.17
chr15	82934997	GOLGA6L10	T>C	missense_variant	MODERATE	28.57
chr15	85402487	ALPK3	C>G	missense_variant	MODERATE	60.44

2

Chromosome	Start	End	seg.mean	Genes in region
chr16	88104996	89256915	-0.707	CBFA2T3
chr15	28929734	82620613	-0.6926	CAPN3,ADPGK,RPAP1,TP53BP1,TCF12,MAP2K1,PML,RASGRP1, MYO9A
chr17	34495876	38926267	-0.6163	MLLT6,LASP1,MED1,RARA,ERBB2,ACACA
chr17	28088031	34493383	-0.6	SLFN13,PSMD11,MYO1D
chr17	47210017	48750516	-0.597	COL1A1,SPOP
chr17	39216193	45096552	-0.5925	PSME3,BRCA1,EFTUD2,STAT3
chr17	48752796	62888760	-0.5729	CD79B,DDX5,SRSF1,RNF43,BRIP1
chr17	62891966	81052511	-0.5655	AXIN2,PRKAR1A,CANT1,GRB2
chr16	4322588	87970103	-0.4553	TNFRSF17,IL21R,TAOK2,MAPK3,CYLD,HERPUD1,NLRC5,CMTR2, CIITA,ERCC4,FUS,CTCF,CDH1,ZFHX3,CDH11
chr18	28908043	55287899	-0.4164	MBD1,SMAD2,MAPRE2,SETBP1,SMAD4

Table: 1. *Top 10 SNVs prioritized by their predicted functional impact and sorted by chromosome to show the most critical point mutations.* ***2.*** *Top 10 non-intergenic SCNA, ranked by absolute seg.mean to highlight the most significant gene losses.*

Mutational Signature Analysis

To better elucidate the mutational processes that have shaped the genomic landscape of the tumor, we performed a mutational signature analysis on the SNVs identified, using the annotated VCF file as input. The deconstruction of the tumor's mutational spectrum revealed an excellent fit (*Cosine Similarity* = 0.966) to a combination of three established *COSMIC* signatures (**Figure 1**). The mutational profile is predominantly driven by *SBS5* (67.73% contribution) and *SBS1* (16.04% contribution). Both are ubiquitous, "clock-like" signatures attributed to endogenous mutational processes that accumulate with age, such as spontaneous deamination of methylated cytosines. Critically, the analysis identified a significant contribution also from *SBS54* (16.23%). This signature is a well-established hallmark of defective DNA repair due to Homologous Recombination Deficiency (HRD) and it is strongly associated with pathogenic mutations in the *BRCA1* and *BRCA2* genes.

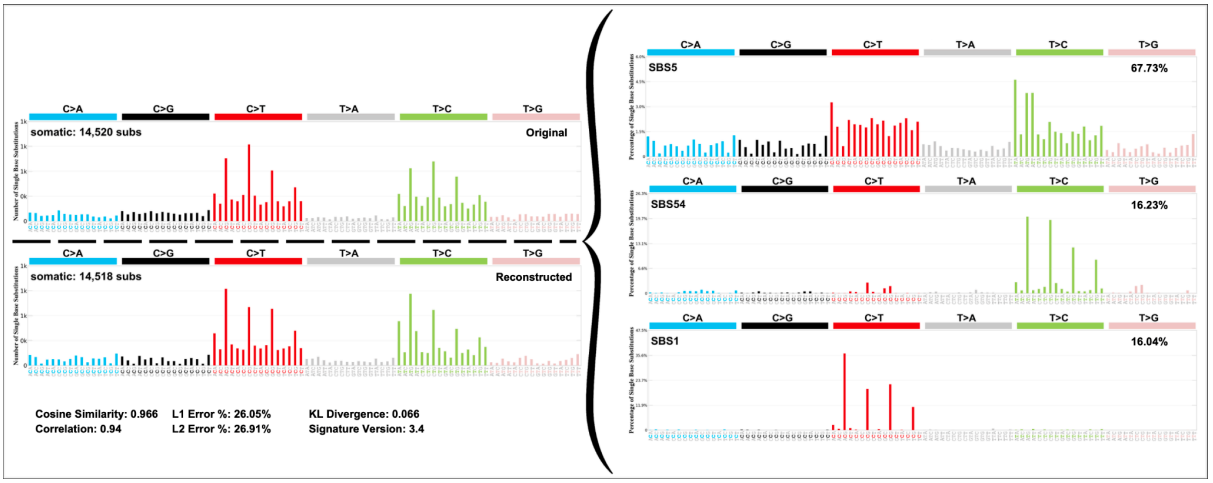

Figure 1: *Deconstruction of the observed mutational spectrum of the tumor (top left) into a combination of three COSMIC signatures (right panel). The reconstructed profile (bottom left) shows a high cosine similarity (0.966) to the original, indicating an excellent model fit. The profile is composed of age-related signatures SBS1 and SBS5, along with SBS54, a signature associated with HRD and BRCA1/2 mutations.*

Purity and Ploidy Estimation

Accurate estimation of tumor purity and ploidy, the proportion of cancer cells in sample and number of chromosomes copies in cancer cells, is fundamental for correct interpretation of somatic mutation data (e.g. the VAF of mutations, that without purity correction would appear a much lower frequency). This step is executed with *CLONET* and *TPES*. *CLONET.R* script

performs copy number analysis (**Figure 2A**), giving a plot with as x-axis *LogR* (relative coverage of the tumor against the control) and as y-axis *beta* (informative on allelic imbalance). Here many clusters can be observed. The wild type, around 0.0 LogR and 1.0 beta, represents the cluster of cells having two copies, one for each allele. The great majority of genomic segments cluster around LogR = -0.5 and beta = 0.5, that represent the cluster of cells with loss of one copy (hemizygous deletion); between these two clusters we can appreciate the LOH subclonal cluster; finally, we can identify some complex events such as CN-LOH where one allele is lost and the other is duplicated (2,0). While *CLONET* estimates a tumor purity of 0.64 and ploidy of 2.27, *TPES* yields a slightly higher purity estimate of 0.71, with both values indicating a predominantly tumor-rich sample containing significant normal cell contamination. The *TPES*-derived AF histogram (**Figure 2B**) shows the distribution of VAFs for the 4 SNVs; two peaks are observed: at 0.33, probably the clonal peak, and 0.24 that may represent the subclonal mutations. The density function variation graph shows a smoothed version of the histogram, highlighting the dominant VAF peaks; it can be noticed that the subclonal peak disappears suggesting noise or a minor subpopulation. We would like to point out that these results are limited by the low number of SNVs considered.

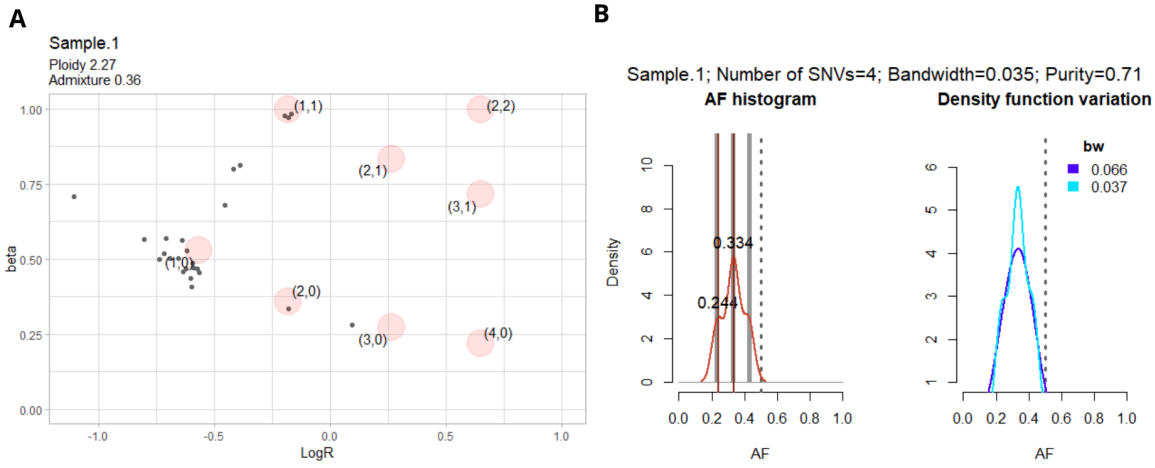

Figure 2: A. *CLONET plot, with LogR as x-axis and beta as y-axis, the gray dots are genomic segments.* ***B.*** *TPES plots, the first individuates the subclonal peak at 0.244 and the clonal peak at 0.334. The second plot validation of results by VAF distribution smoothing by kernel density estimation.*

Relevant Events

Among the variants predicted to have a *HIGH* functional impact, further annotations with *SnpSift* and *ClinVar,* highlighted a clinically significant germline nonsense mutation in *BRCA1* (chr17:41246494 C>A), associated with hereditary breast and ovarian cancer. Although the mutant allele frequency was higher in the tumor (80.20%) than in the control sample (37.63%), suggesting a LOH event in the tumor, its substantial presence in the normal sample confirms its germline origin. Given this inherited background, we focused our somatic analysis on identifying acquired driver mutations. From the high-impact variants, a mutation in the *TP53* gene emerged as the most significant somatic event. As detailed in **Table 1**, the C>A substitution at position chr17:7577156 variant is the top-ranked finding. We can confidently classify this mutation as somatic because it is completely absent in the matched normal sample (100% reference 'C' allele) but present at a high VAF of 69% in the tumor sample, as already predicted by *VarScan2*. This high VAF also indicates that the mutation is clonal, meaning it arose early in tumorigenesis and is present in nearly all cancer cells. To visually confirm these quantitative findings, we inspected the raw sequencing data using IGV (**Figure 3**).

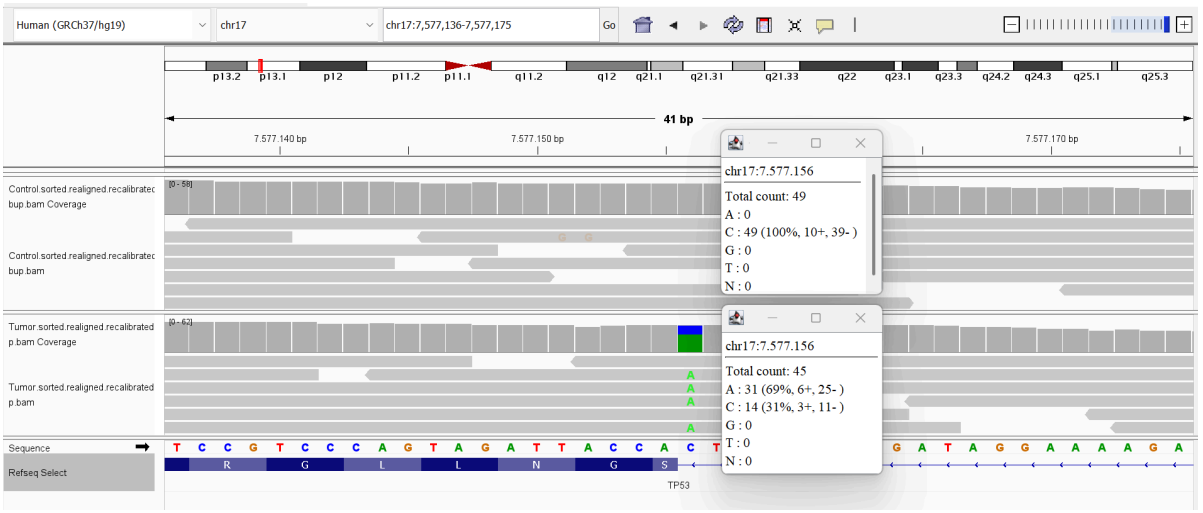

Figure 3: *TP53 locus chr17:7577156 visual representation in IGV, showing Control (above) and Tumor (below) samples nucleotide counts. It can also be appreciated that the coverage of the two samples in the specific position is comparable.*

Biological Contextualization

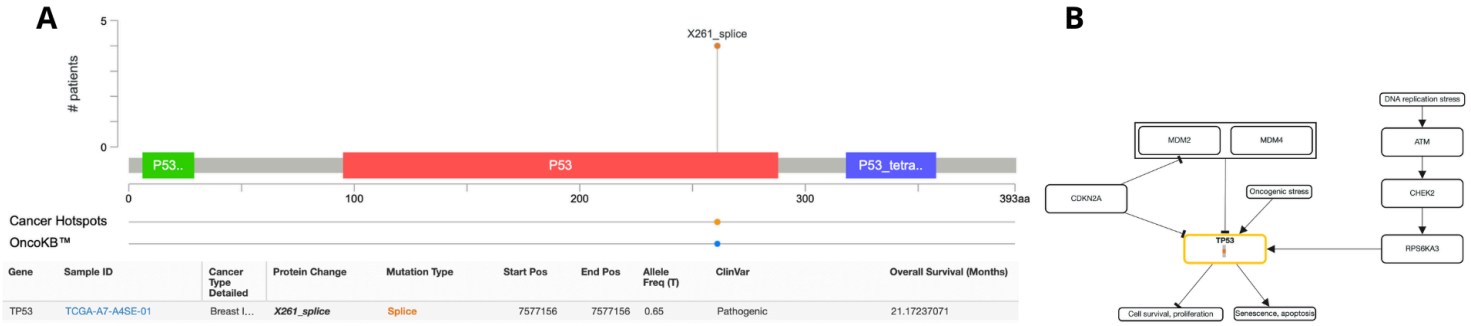

Figure 4: A. *Lollipop plot showing the TP53 X261_splice mutation identified via cBioPortal.* ***B.*** *TP53 signaling pathway, illustrating its central role in DNA damage response and controlling cell survival.*

By querying the *cBioPortal* resource for the *TP53* gene, we identified and validated our somatic event of interest. **Figure 4A** displays *X261_splice* mutation located within the P53 DNA-binding domain. This alteration was found in a patient with Breast Invasive Ductal Carcinoma, whose overall survival was ~21 months since the initial diagnosis, and is validated both as a recurrent hotspot (statistically significant in a population-scale cohort of tumor samples of various cancer types) and in *OncoKB*, where it is classified as a putative driver mutation. This is consistent with the function of *TP53*: the most frequently mutated gene in cancer, with somatic mutations present in 35.1% of samples in the *TCGA PanCancer Atlas*, is a critical tumor suppressor whose central role in key signaling pathways is illustrated in **Figure 4B**. Although specific functional data for the *X261_splice* variant is unavailable, it is considered a loss of function and likely oncogenic mutation: truncating mutations in *TP53* are generally inactivating, associated with poor prognosis and known to promote cancer cell proliferation and survival. To date, there are no specific FDA-approved treatments for patients with breast cancer harboring this particular *TP53* alteration.