UNIVERSITÀ DI TRENTO

Characterization of Somatic Events in Tumor Samples

Computational Human Genomics
2024/2025

Alex Callegaro, Alberto Catalano, Federico Cavasin, Alessandro Fiume

1. Project rationale

The objective of the study is to identify and characterize the relevant genomic aberrations present in an oncogenic patient, to find possible correlations between the variations and the tumoral process. To do so, paired-end NGS data referred to chromosomes from 15 to 18 collected from both healthy and tumoral samples were compared and analyzed, thus highlighting the somatic variants (single-nucleotide variants (SNVs) and somatic copy number aberrations (SCNAs)) associated with the tumor sample. Through variant annotation, the somatic aberrations were linked with their molecular impact, associating the genomic information with the oncogenic phenotype. Moreover, purity and ploidy analysis were performed on the sample, and lastly, ancestral analysis allowed the estimation of the ethnic origin of the sample.

2. Methods

Sequencing data from tumor and matched control samples were analyzed using a combination of Bash-based tools and custom R scripts. Detailed statistics and coverage were first obtained using *Samtools* [1], and only reads with a mapping quality greater than 60 were retained for downstream analysis. Realignment around indels was performed with *GATK* [2] using the modules *RealignerTargetCreator* and *IndelRealigner*, with the human reference genome *human_g1k_v37.fasta*. This step minimizes misalignment artifacts, particularly in regions surrounding insertions and deletions. Base quality score recalibration (BQSR) was then conducted using *GATK*'s *BaseRecalibrator* and *PrintReads*, incorporating known sites from the *hapmap_3.3.b37.vcf* to correct systematic sequencing errors. The quality of recalibration was assessed using *GATK AnalyzeCovariates* by comparing covariate metrics before and after BQSR. The covariates used were ReadGroup, QualityScore, Context, and Cycle. Duplicate reads, indicative of PCR amplification biases, were identified and marked with *Picard*'s *MarkDuplicates* [3]. Final post-processing quality checks and coverage assessments were again carried out using *Samtools*.

Variant calling was performed on both tumor and normal BAM files using *bcftools mpileup* [1], with parameters set to include variants with a minimum base quality of 20, a minimum mean depth of exactly 5, a maximum mean depth of 200, and exclusion of indels. Identified variants were annotated with *SnpEff*, using the *hg19kg* database, and further enriched through *SnpSift* using annotation sources such as *HapMap* and *ClinVar_Pathogenic.vcf* [4]. Custom filtering criteria were applied to retain all clinically significant variants. Among non-clinically relevant variants, those with high or moderate impact were preserved, while low-impact variants were limited to those annotated in splicing regions. Modifier variants were retained only if located within 5′ and 3′ untranslated regions (UTRs), transcription factor binding sites, or regulatory elements. Somatic variant calling was conducted using *VarScan v2.3.9 somatic* [5], comparing matched tumor and control pileup files to identify somatic single-nucleotide variants (SNVs). Resulting variants were filtered with *vcftools* [6], retaining only those with a minimum mean depth of 30. Annotation and impact-based filtering were carried out following the same procedures applied to germline variants.

Somatic copy number variants (CNVs) were identified using *VarScan v2.3.9 copynumber*, with the output further processed through *VarScan copyCaller* to classify chromosomal segments as gains, losses, or neutral regions. Segmentation plots were generated using a custom R script based on the *DNAcopy* package [7]. In addition, ancestry analysis was performed using *EthSEQ* [8], a SNP-based tool designed to assign population ancestry through principal component analysis (PCA) projection onto a reference panel.

Estimation of tumor purity and ploidy was carried out using *GATK ASEReadCounter*, which quantified allele-specific read counts at heterozygous SNP loci in both tumor and control samples, applying thresholds for mapping quality (≥ 20), base quality (≥ 20), and mean depth (≥ 20). The filtered dataset was used as input for *CLONET v2* [9-11]. Mutational signature analysis was conducted using *COSMIC SigProfilerAssignment* [12], and visualization of specific variants was performed with *IGV* [13].

3. Results

3.1. Data pre-processing

From the alignment process, 15,029,250 reads were identified for the tumor sample, while the control contained 19,708,438 reads. From the mapping quality (MAPQ) distribution of the two samples, about 23% of the reads presented a MAPQ = 0, while 73% of the reads had MAPQ = 60. The remaining 4% of the sequences were comprehended between the values 1 and 59 [**Figure S1**]. Given the high value of mapping quality, only reads with a MAPQ of 60 were retained, providing a solid basis for the downstream analysis. After filtering, 10,922,204 and 14,195,705 reads were maintained, respectively, for the tumor and control samples.

Through realignment, potential hidden indels are made explicit, removing artifactual gaps that may impair the analysis. Specifically, in the tumor sample, 5,972 reads were realigned, while in the control, 7,362 reads required attention. BQSR allows the correction of possible biases in the base quality score assignment performed by the base caller. After the procedure, the average quality score for the bases fell by ~ 0.5 points in both cases, and no biases were observed in the calibrated score distribution [**Figure 1**]. With the deduplication step, all the duplicate reads that may have been created during the experimental procedure were removed. Specifically, 11.08% of the reads for the tumor and 12.84% for the control sample were not considered.

On the pre-processed files, we extracted the coverage statistics for the two cases. Considering only the captured regions, in which the sequencing was focused, we obtained an average mean coverage of 51.43 and 67.22, respectively, for the tumor and control sample [**Figure S2**].

3.2. Variant analysis

We identified the single-nucleotide variants present in the two samples. For the tumor sample, 23,176 mutations were found, while in the control, we observed 28,372 single-nucleotide modifications. Among these, the majority was tagged as silent (57.19% and 56.91% respectively), with only a small percentage of nonsense mutations (0.40% and 0.35%). Moreover, through variant annotation, most of the observed variants were identified to have a “modifier” impact (88.6% and 89.97%), corresponding to the lowest impact level, assigned to non-coding regions. On the other side, few high-impact mutations were found (0.08% and 0.06%). Interestingly, the control sample showed a higher number of variations in its sequences, probably due to the generally higher number of reads. Nevertheless, the tumor sample was associated with a higher number of high-impact mutations. In the two samples, a very rare germline variant in the BRCA1 gene (allele frequency = 0.002% in *ExAC*) was identified. This variant, classified as pathogenic in the *ClinVar* database, involves a C > A substitution that introduces a premature stop codon at position 352 (p.Glu352*), leading to loss of gene function. The BRCA1 gene is an E3 ubiquitin-protein ligase that plays a central role in DNA repair, and its mutation is usually associated with breast cancer [14]. Through *VarScan*, 11,387 variants were observed, with a total number of effects of 45,195. Among these, 65 were identified as somatic, and 2,521 were tagged as LOH. The high number of LOH mutations may hide a large number of deletions in the tumor genome. Notably, a possible deletion also involved the BRCA1 gene, determining the mutated gene to be the only one present in the cancer tissue.

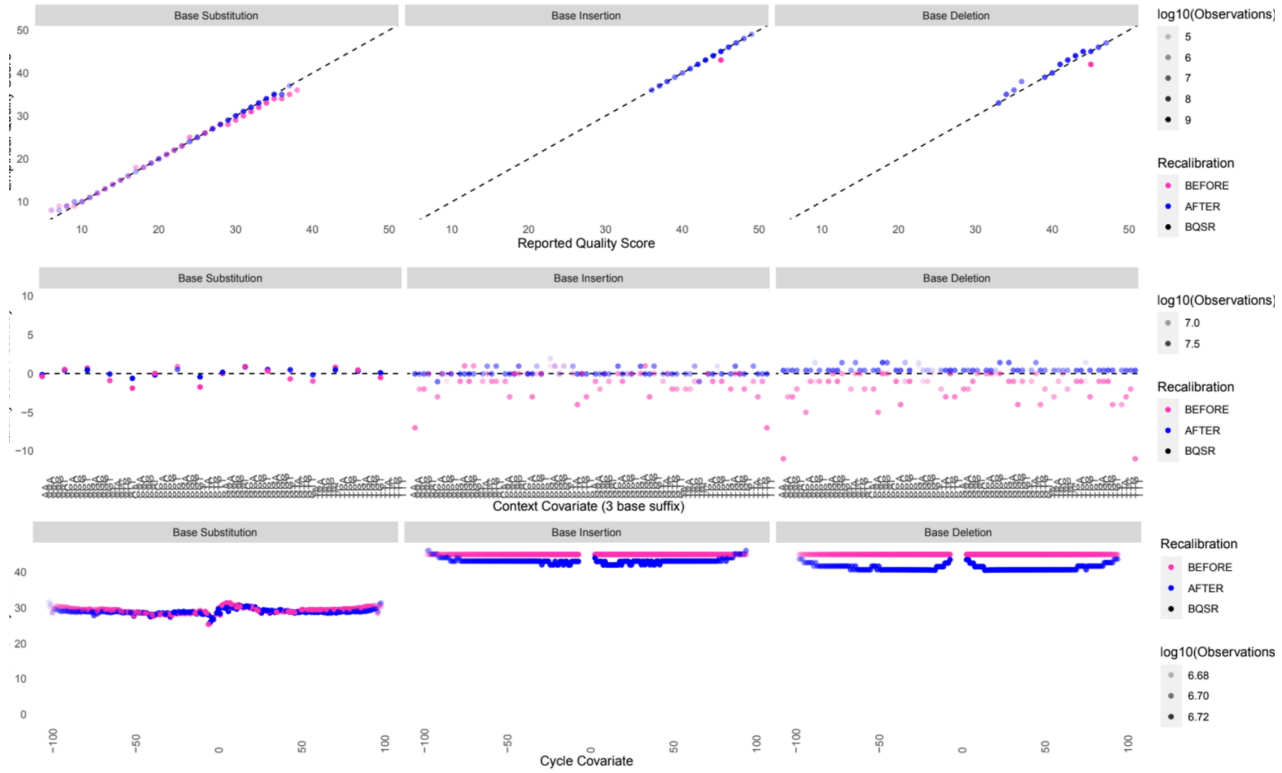

Figure 1. BQSR for the tumor sample. The violet dots represent the recalibrated data. It is possible to observe how the data are adjusted based on the analyzed covariates. Control sample showed similar plots.

Between the somatic variants, only one was identified with a high impact, affecting the TP53 gene [15]. The coding gene plays an important role in the regulation of the cell cycle, DNA repair, and apoptosis, and its malfunction is often associated with tumorigenesis. The gene was also of interest in a loss of heterozygosity process, which caused only the transcript containing a germline upstream mutation to be present.
Inside [**Table 1**] are reported somatic mutations accumulated in some of the genes of interest in DNA repair processes or tumorigenesis.

It is worth reporting that, among the germline mutations, many were concerning important genes involved in the genome and telomere stability maintenance, hence presumably determining a favorable environment for cancer development.

The mutational signature analysis performed on the somatic variations showed a majority of mutations from T to C, especially inside the ATA triplet [**Figure S3**], and the mutational pattern was mainly associated with the clock-like SBS5 signature, representing a steady accumulation of mutations over time.

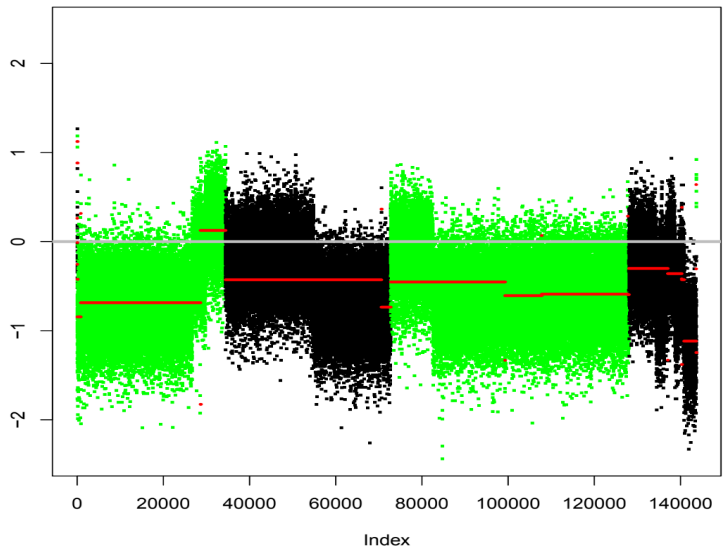

Figure 2. Genomic segments on the x-axis, adjusted log2R on the y-axis. The chromosomes are differentiated by color. It is possible to observe the general low value of log2R hovering around -1, indicating a homozygous deletion. The distance of the mean value from the expected -1 can be explained by the admixture of the sample or subclonal deletions.

Gene	Mutation	Effect Tag	Biological role
RPA1	G > A, LOH	MODIFIER	Binds and stabilizes single-stranded DNA intermediates that form during DNA replication.
RAD51D	C > A, LOH	LOW	Involved in the homologous recombination repair pathway.
WRAP53	G > T, LOH	MODIFIER	RNA chaperone that plays a key role in telomere maintenance.
BPTF	A > G, S	MODERATE	Facilitate access to DNA during replication, transcription, and repair.
GPS1	T > G, S	MODIFIER	Involved in various cellular and developmental processes, and also in the regulation of p53.

Table 1. Relevant genes showing somatic mutations or loss of heterozygosity events are reported.

3.3. Somatic copy number aberrations analysis

Using *VarScan*, it was possible to observe how the tumor sample contained diffused hemizygous deletions in the chromosomes 15, 16, 17, and 18 [**Figure 2**]. The average log2R was -0.50, indicating the general reduction of signal for the tumor sample.

3.4. Ploidy and Purity Allele-Specific Analysis

The *CLONETv2* analysis of the provided sample yielded an estimated tumor ploidy of 2.46 and a DNA admixture of 0.42. The corresponding logR-beta plot [**Figure 3**] reveals a complex genomic landscape with both clonal and subclonal copy-number events. Since this interpretation is derived from a sparse dataset of only twelve genomic segments, its statistical power is inherently limited. Therefore, the following characterization of the model's output should be considered preliminary and not a definitive conclusion on the biological state of the sample.

The most cohesive cluster in the dataset consists of three segments. Its position in the logR-beta space is intermediate to that of the diploid wild-type (1,1) and the clonal hemizygous deletion (1,0) predictions. This intermediate positioning is consistent with a hemizygous deletion being present in approximately 50% of the tumor cells for these specific segments, defining a clear subclonal event.

The model identifies several plausible, albeit isolated, clonal genomic events. These include a segment representing the diploid baseline (1,1), a segment exhibiting a complex gain-LOH alteration (3,0), and another indicating a homozygous deletion (0,0). The alignment of these observed data with the model's theoretical predictions for pure integer copy-number states is noted; however, these isolated observations lack the statistical power of clustered data and are therefore of limited analytical value.

The largest group of segments, comprising seven data points, is located in the proximity of the predicted (1,0) copy-number configuration. This group is characterized by significant dispersion rather than forming a cohesive cluster, and it represents the primary source of ambiguity in the analysis. The observed spread presents two competing interpretations. The first possibility is that the spread represents a true biological spectrum of multiple, independent homozygous deletions. The clonal fractions for these events are not uniformly high; rather, they vary considerably, with some segments positioned much closer to a subclonal status while others approach a fully clonal deletion. The second possibility is that this dispersion indicates a fundamental poor fit of the estimated ploidy and admixture for the majority of genomic segments. The model may be failing to describe the central tendency of this group accurately. Resolving whether this dispersion reflects true biological variance or a model limitation would necessitate an analysis of a more comprehensive dataset.

The prevalence of segments corresponding to hemizygous deletions is consistent with findings from conventional total copy number analysis methods, which also suggested widespread genomic losses. Nonetheless, a counterintuitive finding emerged: the estimated global ploidy value was relatively high, at 2.46. This reflects the specific methodology employed by *CLONETv2*.

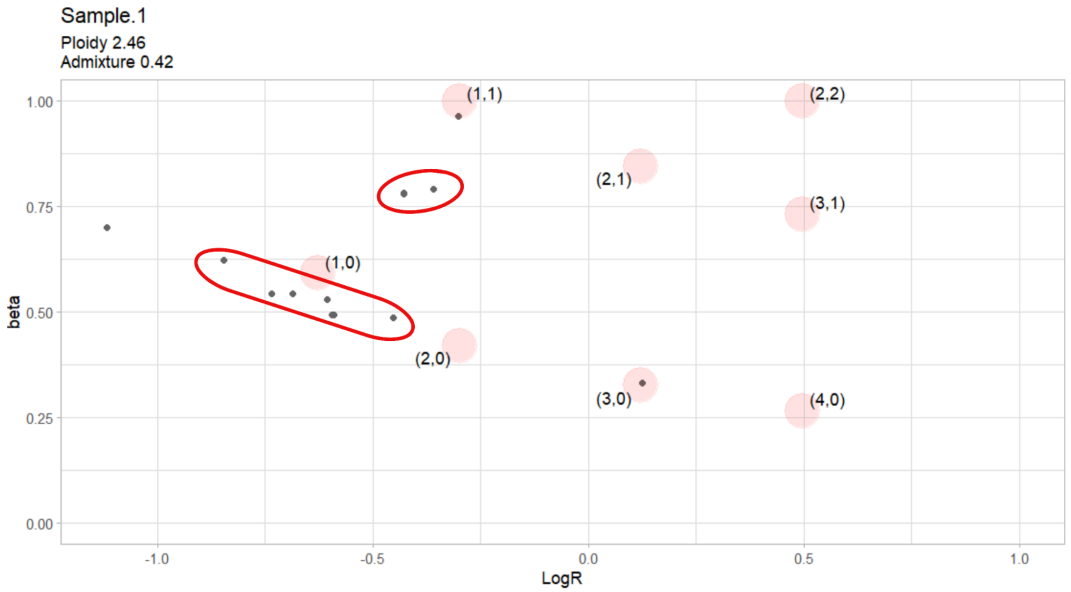

Figure 3. Beta-vs.-logR plot. Each gray dot represents a genomic segment. Light-red circles indicate the expected coordinates for pure clonal copy-number states, based on the estimated ploidy (2.46) and DNA admixture (0.42).

Rather than a simple genome-wide average of copy numbers, the algorithm calibrates the baseline ploidy by referencing only the allelically balanced genomic regions (where beta ≈ 1). The resulting ploidy value is therefore intended as a precise baseline for subsequent allele-specific calculations, not as a descriptive average of the entire tumor genome's copy-number state (which is overall in a low ploidy state).

3.5. Ancestry Analysis

An ancestry analysis was conducted using the PCA-based tool *EthSEQ*. Within the PCs space, the individual's genetic data did not cluster distinctly with any of the reference ethnicity panels [**Figure S4**], suggesting a potentially admixed genetic background. A subsequent distance-based evaluation estimated contributions as EUR (19.89%), EAS (19.64%), SAS (17.97%), and AFR (42.5%), indicating the closest component by this metric was African.

However, this analysis is considered highly unreliable due to significant technical constraints. The primary limitation is the severely restricted genomic data, which is insufficient for robustly capturing the overall genetic background needed for accurate ancestry determination. Furthermore, PCA is not the most suitable method for precisely quantifying complex admixture, and the "percentages" derived from distances when an individual falls outside reference clusters are pseudo-admixture estimates rather than true statistical proportions. Consequently, the results obtained should not be treated as definitive. To reliably assess ancestry and accurately estimate admixture components, especially in potentially admixed individuals, it is essential to use comprehensive genome-wide genetic data alongside appropriate model-based inference tools.

4. Conclusion

In this study, germline and somatic variations were identified and analyzed to determine putative dependencies between the data observed and tumor formation. The quality control and pre-processing steps showed a high quality for the two samples, with sufficient coverage (51.43x in tumor and 67.22x in control) to support a robust analysis of variant calling.

The most relevant features extracted from the data revealed a high-impact germline mutation in the gene BRCA1, causing a premature ending of the transcript and the loss of function of the gene. The condition was further aggravated in the tumor sample by the loss of the WT copy of the gene. Moreover, a somatic mutation in the TP53 gene was observed, causing an increase in the instability of the genome. Considering broader variations, SCNA analysis revealed widespread hemizygous deletions on chromosomes 15, 16, 17, and 18. Chromosomal deletions are a common mechanism for inactivating tumor suppressor genes and can significantly impact disease progression. The ploidy and purity values were estimated to be 2.46 and 0.58 respectively, and we were also able to assess the presence of a probable subclone within the tumor sample.

References

1. Li H. **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509

2. McKenna A, Hanna M, Banks E, et al. **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. Genome Research. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110

3. **Picard Tools - By Broad Institute**. *Github.io*. Published 2025. Accessed July 4, 2025. https://broadinstitute.github.io/picard

4. Cingolani P. **Variant Annotation and Functional Prediction: SnpEff**. *Methods in Molecular Biology*. Published online February 24, 2012:289-314. doi:10.1007/978-1-0716-2293-3_19

5. Koboldt DC, Chen K, Wylie T, et al. **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. *Bioinformatics*. 2009;25(17):2283-2285. doi:10.1093/bioinformatics/btp373

6. Petr Danecek, Auton A, Goncalo Abecasis, et al. **The variant call format and VCFtools**. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

7. Venkatraman ES, Olshen AB. **A faster circular binary segmentation algorithm for the analysis of array CGH data**. *Bioinformatics*. 2007;23(6):657-663. doi:10.1093/bioinformatics/btl646

8. Romanel A, Zhang T, Elemento O, Demichelis F. **EthSEQ: ethnicity annotation from whole exome sequencing data**. *Bioinformatics*. 2017 Aug 1;33(15):2402-2404. doi:10.1093/bioinformatics/btx165

9. Prandi D, Baca SC, Romanel A, et al. **Unraveling the clonal hierarchy of somatic genomic aberrations**. *Genome Biology*. 2014;15(8). doi:10.1186/s13059-014-0439-6

10. Carreira S, Alessandro Romanel, Goodall J, et al. **Tumor clone dynamics in lethal prostate cancer**. *Science Translational Medicine*. 2014;6(254). doi:10.1126/scitranslmed.3009448

11. Romanel A, Tandefelt DG, Conteduca V, et al. **Plasma *AR* and abiraterone-resistant prostate cancer**. *Science Translational Medicine*. 2015;7(312). doi:10.1126/scitranslmed.aac9511

12. Tate JG, Bamford S, Jubb HC, et al. **COSMIC: the Catalogue Of Somatic Mutations In Cancer**. *Nucleic Acids Research.* 2018;47(D1):D941-D947. doi:10.1093/nar/gky1015

13. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. **Integrative genomics viewer**. *Nature Biotechnology*. 2011;29(1):24-26. doi:10.1038/nbt.1754

14. Abbott DW, Thompson ME, Robinson-Benion C, Tomlinson G, Jensen RA, Holt JT. **BRCA1 expression restores radiation resistance in BRCA1-defective cancer cells through enhancement of transcription-coupled DNA repair**. *J Biol Chem*. 1999 Jun 25;274(26):18808-12. doi:10.1074/jbc.274.26.18808

15. Whibley C, Pharoah PD, Hollstein M. **p53 polymorphisms: cancer implications**. *Nat Rev Cancer*. 2009 Feb;9(2):95-107. doi:10.1038/nrc2584

Supplementary Information

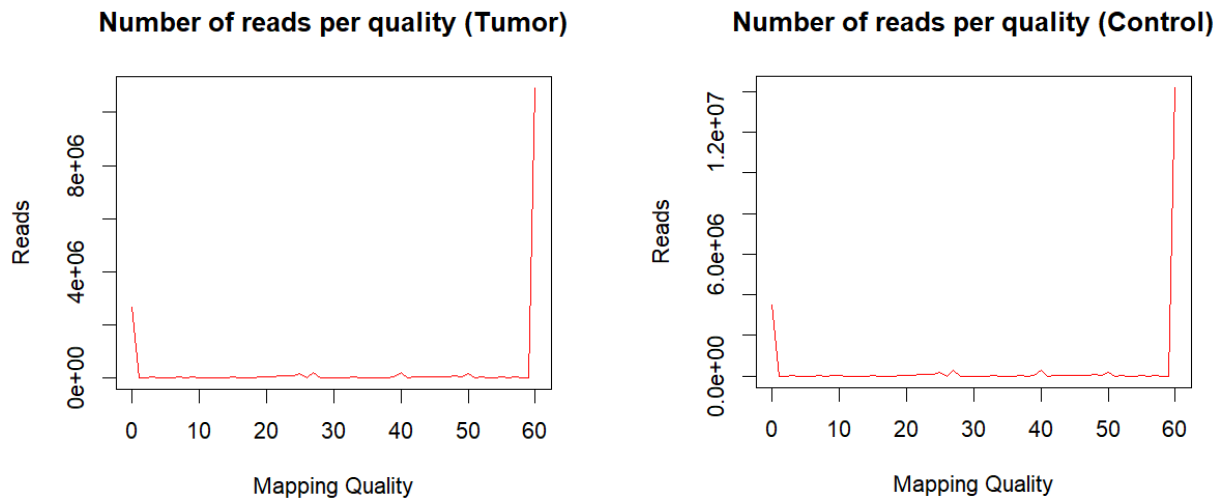

Figure S1. Mapping quality read distribution. The majority of the reads fall at MAPQ = 60, with a distinct peak at 0 as well.

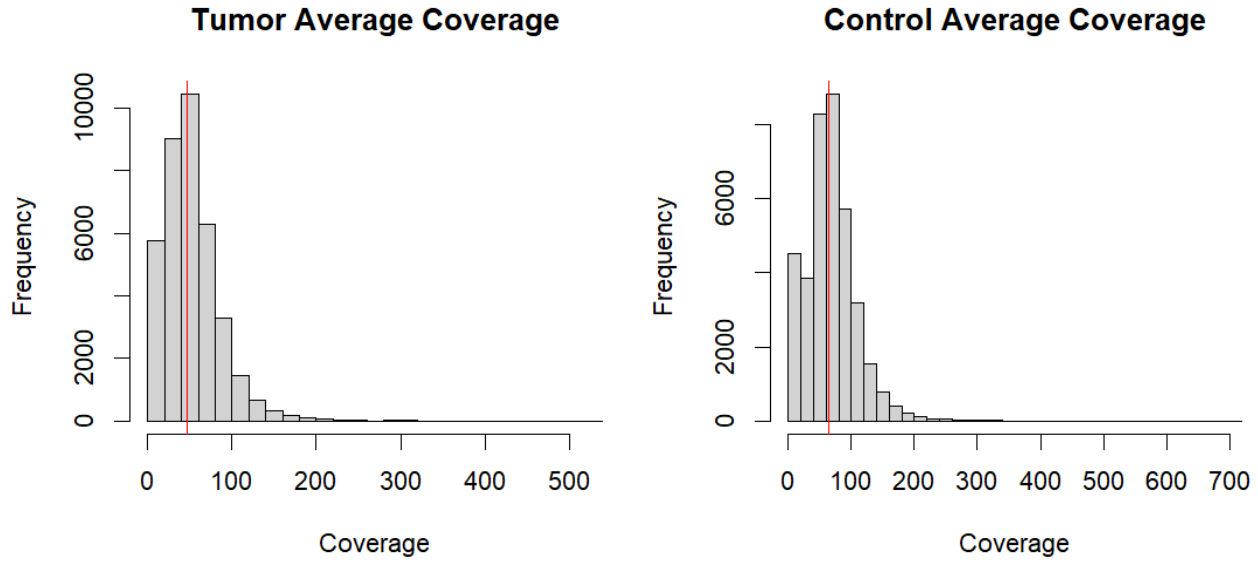

Figure S2. Distribution of the average coverage value for the captured regions. The tumor sample showed a mean average coverage of 51.43, while for the control sample, 67.22 was obtained.

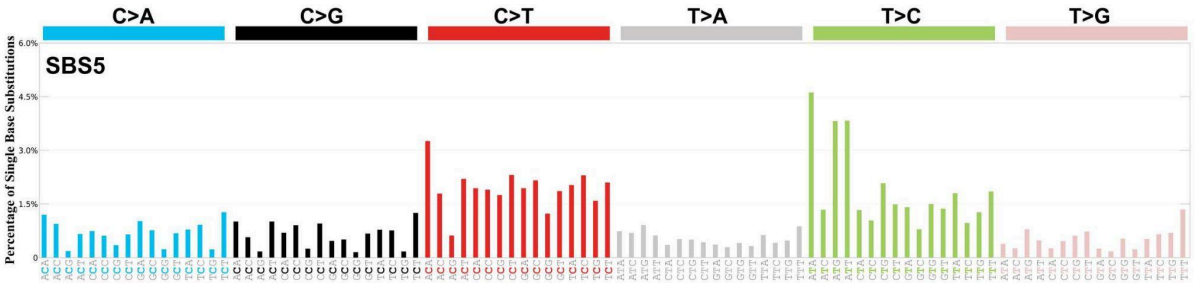

Figure S3. Mutational signature computed from the somatic variations. The mutation T > C is the most common, especially present in the ATA, ATG, and ATT triplets. The signature is associated with the SBS5 type, linked with time-dependent mutation accumulation.

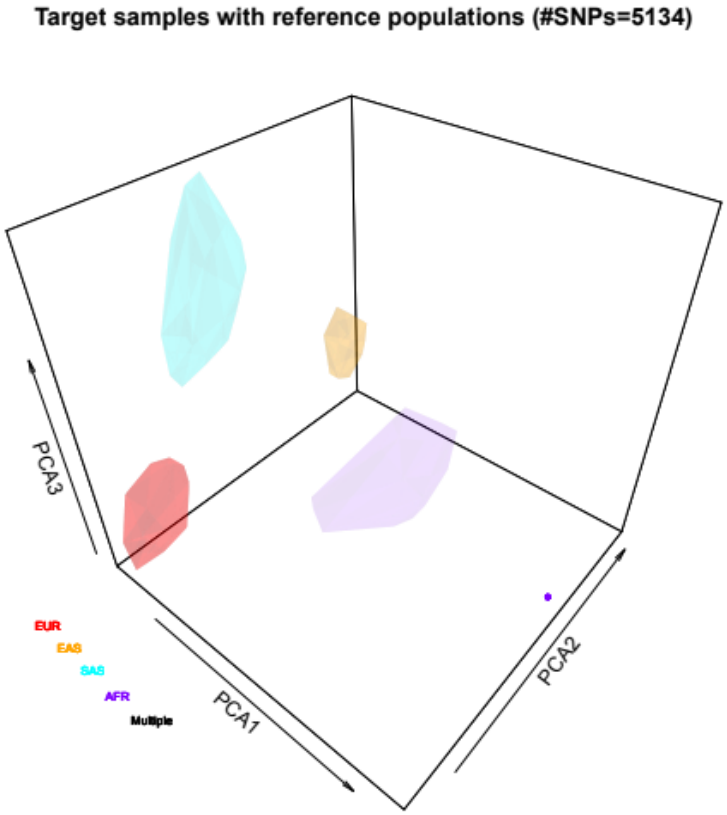

Figure S4. Ancestry Analysis and Population Contribution for Sample TCGA-A7-A4SE-Control. This plot, along with the accompanying table, visualizes the estimated ancestral composition of the `TCGA-A7-A4SE-Control` sample. The sample's "closest population" is identified as African (AFR). However, the detailed contribution analysis reveals a mixed ancestry, with a predominant African component (42.5%), followed by European (19.89%), East Asian (19.64%), and South Asian (17.97%) contributions. This highlights the complex genomic background of the individual.

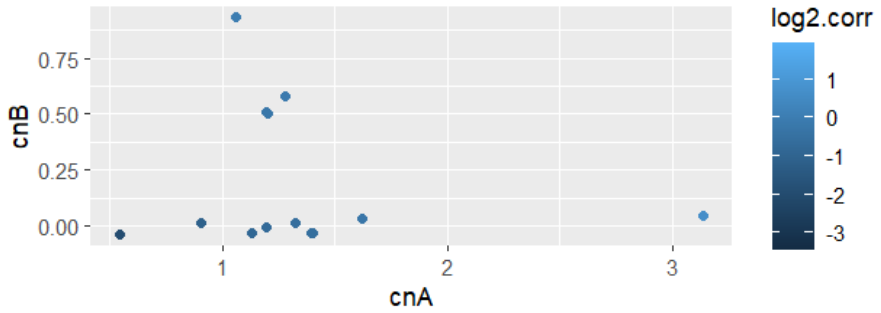

Figure S5. cnA vs. cnB Plot. Blue dots representing genomic segments; cnA: allele-specific number of copies of the major allele; cnB: number of copies of the minor allele; (both cnA and cnB are estimates for the aberrant cell population).

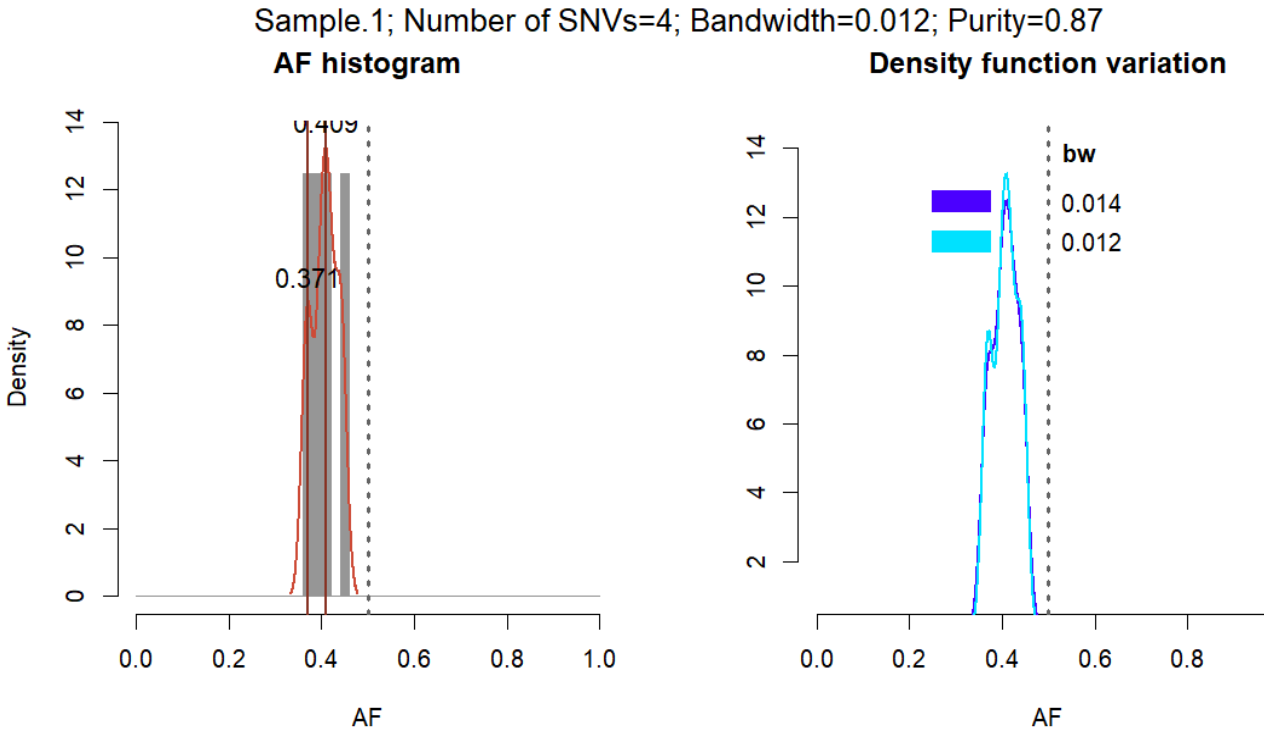

Figure S6. Clonet analysis results for Sample 1. The left panel shows the Allele Factor (AF) histogram with the number of SNVs (4) and the estimated purity (0.87). The right panel illustrates the variation of the density function for different bandwidth (bw) values.

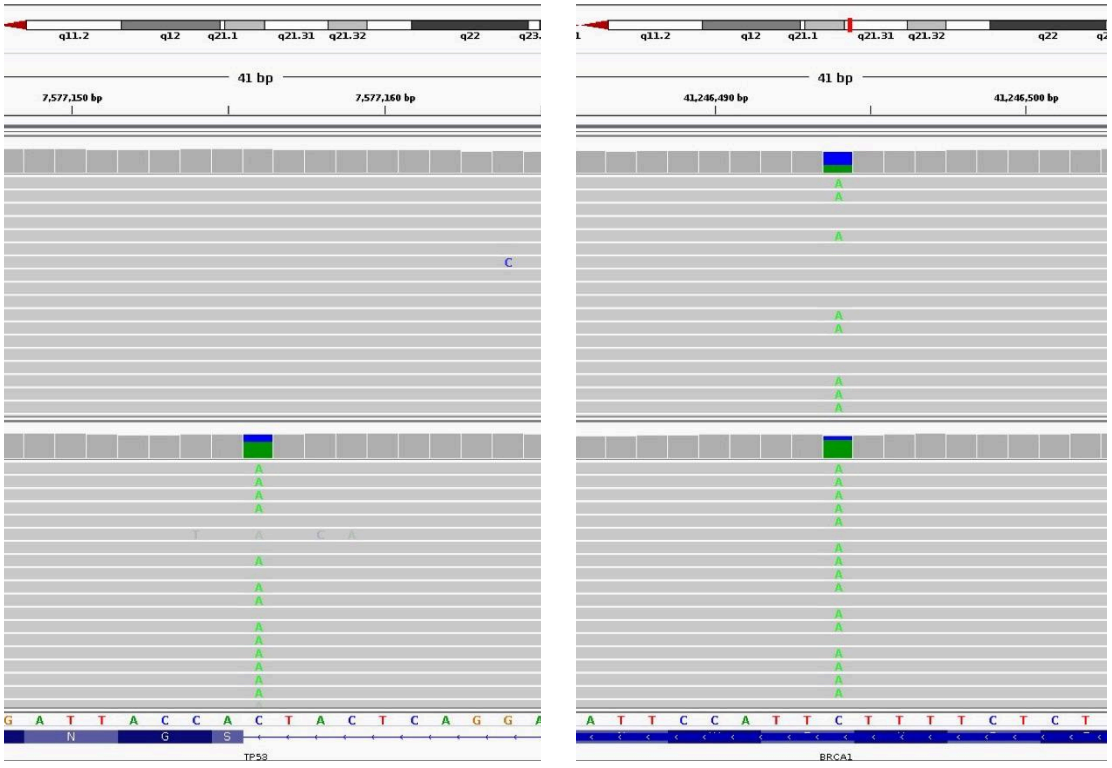

Figure S7: IGV representation of two somatic events. Control sample reads on the top, tumor sample reads on the bottom. On the right, it is possible to observe the SNV that modifies a C into a T at position 7'577'156 of chr 17 in the TP53 gene. On the left, the loss of heterozygosity in the BRCA1 gene is shown. Indeed, almost only reads containing the SNV are present in the tumor sample.