Computational Human Genomics Project

Variant Calling and Annotation, Ancestry analysis, Tumor Ploidy and Purity assessment.

Linda Cova and Enrico Frigoli

Github: https://github.com/enricofrigoli/chg_project.git

Computational Human Genomics
Prof. Demichelis, Prof. Romanel
University of Trento
A.Y. 2021-2022

Introduction

The study of somatic and germline variations (SNPs, SNVs, indels, CNVs and SVs) is of key importance in the field of cancer genomics, as these lead to altered gene function in cancer. Although many tools and pipelines have been developed in recent years to tackle the problem, these variations are still challenging to identify. Here, we describe the use of different tools to analyse germline variants, somatic variants, ancestry, tumor ploidy and purity, and to perform additional analysis on provided data.

Methods

To perform the required analysis, two setups were used: the first one (hereafter setup **1**) involves the use of a provided Virtual Machine that already included GATK (v3.8-1-0), Picard (v2.22.3), VarScan (v2.3.9), and snpEff (v4.3t); since some of the included tools are no longer considered best-practice and variant calling works better by combining different tools, a second setup was used (hereafter setup **2**), consisting of GATK (v4.2.6.1) run in a Docker container with Picard (v2.27.1), and SnpEff (5.1d). Both setups share custom R scripts, the provided raw BAM files, few files containing the reference genome, the .bed files containing the captured regions and the .vcf files for variant annotation. The code with additional details can be found on github. The two provided BAM files contain already selected fractions of data from a breast cancer sample and matched control from TCGA (TCGA-A7-A4SE).

Data inspection

Each of the two BAM files can be inspected either using `samtools stats [filename]` or `samtools flagstat [filename]`, for detailed and general statistics respectively.

Preprocessing of BAM files

In order to be used for subsequent analysis, BAM files containing raw mapped reads must undergo several preprocessing steps; however, different variant calling tools require different preprocessing procedure. A common first step is deduplication: this allows to remove technical duplicates among the pool of reads that could bias the discovery of variants. After sorting and indexing with `samtools`, deduplication was performed with `MarkDuplicates` tool of Picard, using as additional parameters `-ASSUME_SORT_ORDER coordinate`, `-M Control_dedup_metrics.txt` and `-REMOVE_DUPLICATES true`.

Common steps after duplicates removal are realignment and recalibration; while recalibration is always needed before performing downstream analysis, different tools may require or not realignment. Here, realignment was performed with `RealignerTargetCreator` and `IndelRealigner` (setup 1) specifying as parameter the reference genome and the .bed files containing the captured regions. The number of realigned reads was retrieved with `grep "OC" | wc -l` piped from the output of `samtools view [filename]`. Recalibration was performed on the realigned files with `BaseRecalibrator` and `PrintReads` (setup 1), and on the raw files with `BaseRecalibrator` and `ApplyBQSR` (setup 2). In both cases the results of the recalibration were plotted using `AnalyzeCovariates`, together with a second recalibration on already recalibrated files that was run as a quality control. The number of recalibrated reads can be retrieved with `samtools view [filename] | grep "OQ" | wc -l`, where "OQ" stands for *Original Quality*, that was set to be emitted during recalibration using the parameter `--emit-original-quals`.

Germline Variant Calling

Germline variants (SNPs and indels) were called differently based on the setup in use. With the setup 1, variants were called with `UnifiedGenotyper` using, as additional input, the reference genome and the .bed file containing the captured regions. SNPs calls were filtered and selected (omitting indels) with `vcftools`, using the following input: `--minQ 20 --max-meanDP 200 --min-meanDP 5` to filter variants with min quality of 20, max depth across samples of 200 and min of 5, `--remove-indels --recode --recode-INFO-all` to remove indels, output the filtered vcf and keep all INFO variables respectively.

Since `UnifiedGenotyper` is now considered a legacy tool, a more updated one was also used. For this purpose, we decided to use `HaplotypeCaller` (setup 2) following the pipeline recommended by GATK for small germline variant discovery (Figure S1), that recommends to call variants without a realignment step; this is the case since the tool also performs a sort of assembly in active regions, removing the need of upstream realignment. Following GATK recommendation, the raw callset was filtered with `CNNScoreVariants`, that annotates the obtained VCF with scores generated from a CNN (Convolutional Neural Network) using a 2D model with pre-trained architecture, indicating the model's prediction of the quality of each variant, based on the genomic context. Indels were then removed to keep only SNPs using `SelectVariants` with `--select-type-to-include SNP` as parameter. The callset was further refined using `FilterVariantTranches`, that filters variants based on the score assigned by `CNNScoreVariants` using as additional resource the `hapmap_3.3.b37.vcf` file.

In both cases the number of heterozygous SNPs was calculated with basic bash commands: `cat [filename.vcf] | grep "0/1" | wc -l`, as "0/1" is the annotation of heterozygous SNPs in VCF files. The vcf file containing only heterozygous SNPs was retrieved using `grep -E "(^ #|0/1)"`. The intersection between the two callsets was computed with `bedtools intersect` and `vcftools --diff-site`, to see what was the concordance between the two different variant caller.

The refined callsets were annotated with `SnpEff`[1] and its submodule `SnpSift`[2], using the `hapmap_3.3.b37.vcf` file and `clinvar_Pathogenic.vcf`. SnpSift `filter` function was also used to select clinically significant variants annotated in Clinvar with the following expression `"(exists CLNSIG)"`.

Somatic Variant Calling

As it was done for germline small variants, also somatic ones were called using two different tools: VarScan2 and Mutect2.

To perform somatic point mutations calling with the setup 1, both Control and Tumor files were given as input to `samtools mpileup`, along with the reference genome, to compute the pileups, which were then used to call variants with `VarScan somatic` using `--output-snp somatic.pm --output-vcf 1` as parameters to select for SNPs and produce a .vcf file. The callset at this point included also germline variants, so only somatic ones were filtered using `grep`. The filtered callset was then annotated with SnpEff specifying `hg19` as internal dataset to be used.

The setup 2 instead involved the use of Mutect2 (from GATK): as parameters were given the BAM files, the bed file containing captured regions, the reference genome, and the filtered callset of germline variants found with `HaplotypeCaller` (setup 2). The callset was filtered using `FilterMutectCalls` and annotated with snpEff using again the internal hg19 dataset (the use of `Funcotator` is recommended but its complexity places it beyond the purposes of this analysis).

Somatic CNVs were called with VarScan (setup 1) running the `copyCaller` command on pileups file; the resulting output was processed in R using the package `DNAcopy` (version 1.68.0) [3]. A circular binary segmentation was performed to point out regions with copy-number aberrations with the following parameters: 2 as minimum number of markers in a segment, 3 as minimum standard deviation between splits, *alpha*=0.05 as the significance threshold to distinguish two segments and 100 as the number of permutations used to compute the P-value.

Variants in DNA Repair Genes

The identification of SCNAs overlapping with DNA repair genes was performed in R: the segments containing DNA repair genes were individuated and their mean log2 ratio was analysed to identify copy-number aberrations. With the same purpose, starting from the bed file containing heterozygous deletions and intersecting it with the file `DNA_Repair_Genes.bed`, another .bed file was retrieved containing DNA repair genes that were subject to somatic heterozygous deletions. Another intersection was performed between the callset of heterozygous pathogenic SNPs (containing only the one in *BRCA1*) and the new bed file, using again `bedtools intersect`.

Since only one variant was pathogenic, another intersection was perfomed between the .bed file previously obtained and the entire callset of heterozygous SNPs found with `HaplotypeCaller` (setup 2). This allowed to retrieve also other germline SNPs that were not annotated as pathogenic but still residing into fragments that later underwent somatic heterozygous deletions (found in the Tumor sample).

The same approach was also used to see whether called somatic small variants (point mutations and indels) were present in the subset of DNA repair genes harboring somatic heterozygous deletions. For this purpose, the callset generated with Mutect2 (setup 2) was used, since Varscan (setup 1) identified far less somatic variants. The output of the intersection between the callset and `DNA_Repair_Genes.bed` was formatted in bash and used as input to a custom R script that allowed to intersect the list of DNA repair genes with the one of heterozygous deletions (as could have been done with `bedtools intersect`) to generate a plot to show the number of somatic variants (point mutations and indels) for each DNA repair gene harboring a somatic heterozygous deletion.

Ancestry Analysis

The ancestry analysis was performed with the R package `EthSEQ`[4]. The control sample preprocessed (setup 1) BAM file was provided as target input. The reference model was provided in gds format with the `SS2.Light.Model.gds` file. The function `ethseq.Analysis` relies on the `ASEQ` R package to genotype the BAM file on positions indicated by the reference model and creates a target model. Only SNPs that satisfy the requirements given as parameters are retained (minimum base quality = 20, minimum read quality = 1, minimum read count =10) and a PCA is performed on the aggregated target and reference models. For each ethnicity, an area is individuated in the 3D space and the sample is classified based on where its principal components fall in this space.

Purity and Ploidy estimation

Tumor purity was inferred with the R packages `CLONETv2` (version 2.2.1)[5] and `TPES` (version 1.0.0)[6], that rely respectively on somatic copy-number changes and somatic point mutations.

Read counts per allele were computed with the ASEReadCounter tool (GATK) on the processed control and tumor BAM files (setup 1) considering the SNP sites individuated in the "germline variant calling" section. These were coupled with the segmentation data in order to retrieve the *beta* table, following the CLONETv2 workflow. The beta values are the percentages of neutral reads per segment and are useful to infer tumor ploidy, which in turn is computed considering the log ratio between tumor and control coverage within each segment LogR, normalized over the ratio between the mean tumor and control coverage. With the ploidy and beta tables it is possible to calculate the admixture: the percentage of non-tumor cells in a tumor sample, equivalent to 1-purity. The standard parameters were maintained for all the CLONETv2 functions.

On the other hand, the TPES workflow requires as input the SNVs identified in the "somatic variant calling" section with VarScan, in addition to the ploidy and segmentation data. This tool selects copy-number neutral segments, and, after applying a threshold on allelic fraction (maxAF=0.55) and reference mapping bias (RMB=0.47), identifies putative clonal SNVs and exploits them to assess tumor purity.

SPIA test

Read counts per allele were obtained with ASEReadCounter (GATK), giving as input the preprocessed (setup 1) BAM files for the tumor and control samples and considering all the SNP sites provided by the hapmap_3.3.b37.vcf file. The SPIA test is an assay aimed at the individuation of different cell lines, usually employed to discriminate between samples from different patients or to confirm that multiple samples have the same origin. The test was performed with the R package SPIAssay (version 1.1.0)[7], which uses the genotype of the provided SNPs (classified as homozygous reference, homozygous alternative or heterozygous) of both samples to compute a distance score and assigns a SPIA score that classifies the two samples as *similar*, *different* or *uncertain*.

Results

Data inspection and pre-processing The number of properly paired reads was retrieved from the statistics generated for the two BAM files (Control.bam and Tumor.bam), and were 19576046 (99.33%) and 14979936 (99.67%) respectively. After realignment, the number of realigned reads (setup 1) was 2445 for the Control file and 1937 for the Tumor file.

The results of the recalibration steps are summarized in the plots (Figures S2-S5); the process resulted in higher quality score accuracy and slightly lower mean quality score. As expected, recalibration worked better on realigned BAM files, since artifactual SNPs are replaced with real indels (recalibration for the setup 2 not shown). Although successful, recalibration plots usually show better results with increasing size of input data: since this analysis was performed on small data, a higher noise in recalibration plots is expected.

Germline variant calling Since two germline variants caller were used, it was possible to compare the resulting callsets: the filtered one obtained with UnifiedGenotyper contained 6480 heterozygous SNPs, while the filtered one obtained with HaplotypeCaller contained 6937 heterozygous SNPs; 5933 of them were common between the two files.

When looking for pathogenic variants (which is by definition a variant that is well characterized and strongly associated with a high increase in the risk of developing a pathology), both callers identified only the variant in BRCA1 gene, associated with hereditary breast and ovarian cancer syndrome (Figure 1). This is indeed the type of cancer from which the tumor sample was retrieved.

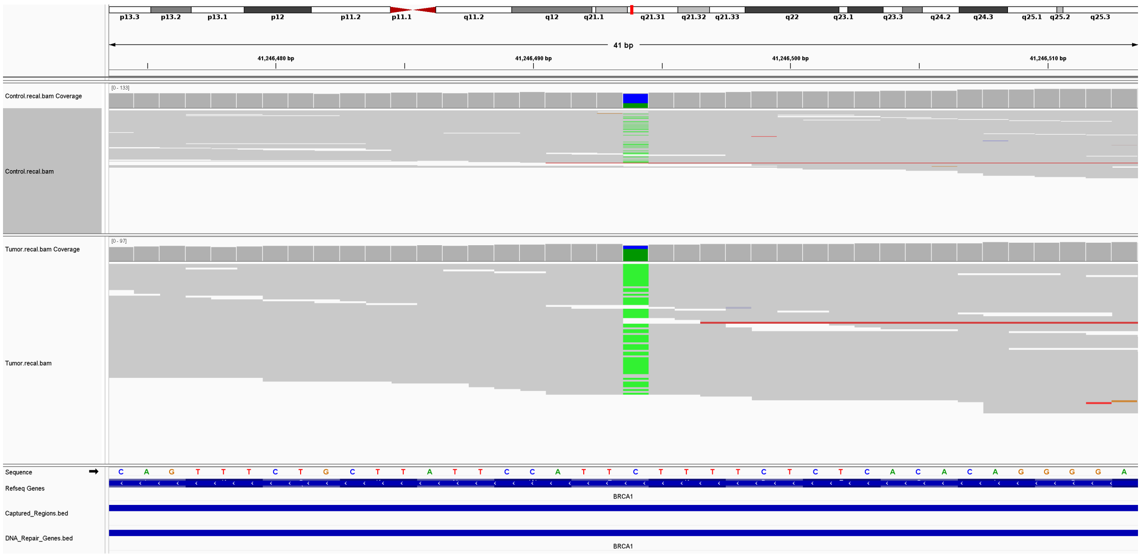


Figure 1: Snapshot from IGV on the SNP in BRCA1 associated with hereditary breast and ovarian cancer at position chr17:41246494; see Results section for the biological implication of the change in the allelic fraction of the alternative base.

Ancestry analysis The ancestry analysis determined that the patient is of African ethnicity. The first three principal components (respectively: 0.053, 0.003 and -0.003) place the control sample perfectly into the 3D space associated with the African population (Figure 2).

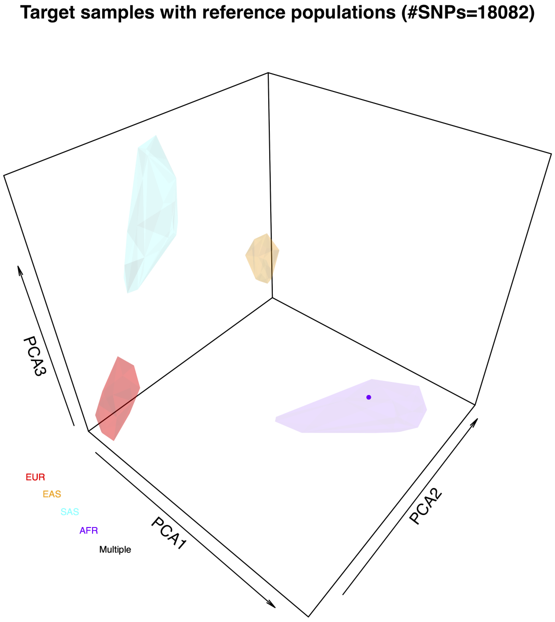


Figure 2: EthSEQ output: the sample (purple dot) falls into the AFR area

Somatic variant calling The results of the somatic copy-number calling are depicted in Figure 3. Most of the segments present a log2 ratio value lower than zero, indicating that some deletions occurred. Since log2 ratio is measured as the base-2 logarithm of the tumor-related signal over the control signal, the expected value for heterozygous deletions is -1. However, due to confounding factors such as tumor purity and ploidy, the measured log2 ratios are likely to be shifted and it is safer to consider as heterozygous deletions the segments with values between -1.5 and -0.5. In particular, 101 out of 175 individuated segments present mean log2 ratio compatible with this kind of aberration.

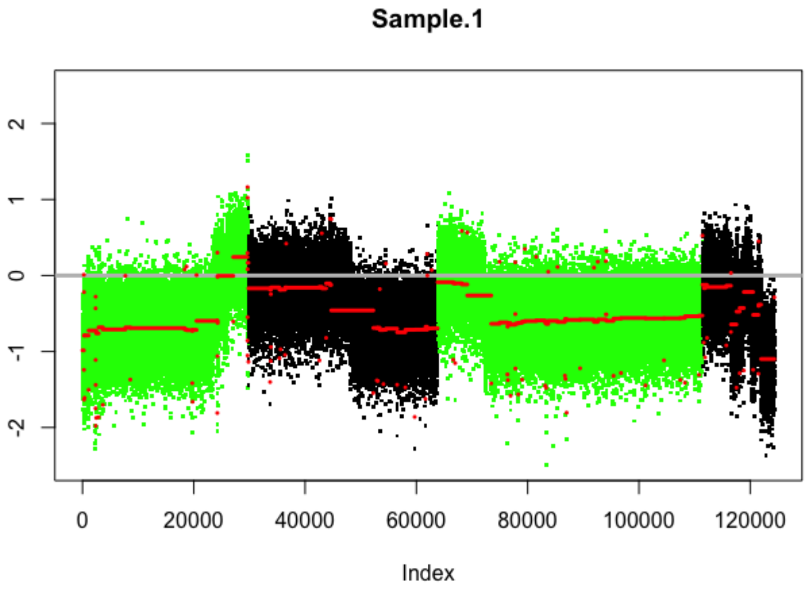


Figure 3: Circular binary segmentation results: red lines and dots indicate the mean log2 ratio of each segment, the colors divide the four analysed chromosomes.

As for germline small variants, somatic point mutations were also called using two different tools. VarScan2 (setup 1) found 167 somatic point mutations, while Mutect2 (setup 2) found 641 point mutations; among them, 30 were found in common. An example of somatic point mutations found by Mutect2 in *TP53* is reported in Figure S6. These low numbers are normal in this context since the BAM files from which variants were called represents only a small portion of the entire data belonging the original sample from TCGA. Also, significant differences in the number and types of variants identified are expected when comparing callset produced by different caller: piping them together is indeed a strategy to reduce the number of false negatives. Since Mutect2 claims to be more sensitive than other methods (especially for low allelic fraction and low read support events) while maintaining high specificity, we used this callset for the subsequent steps.

Variants in DNA repair genes If a DNA repair gene is characterized by the presence of small heterozygous variants associated with an increased risk of pathology, or even pathogenic ones, the accumulation of somatic aberrations in the copy number status (deletions, in particular) in those segments poses a significant threat with respect to the susceptibility to diseases. This seems to be what happened in this sample. In fact, the only pathogenic variant identified in the callset of germline heterozygous SNPs is the one in *BRCA1*, as already mentioned in the previous paragraph. Moreover, the intersection between segments with heterozygous deletions and DNA repair genes identified *BRCA1* as one of the genes affected by the copy-number aberrations (Figure 4). As can be seen in Figure 1, the support for the alternative base (that encodes for the pathogenic SNP) is higher in the tumor sample harboring the somatic heterozygous deletion in *BRCA1*: the combination of these aberrations is very likely related to the disease.

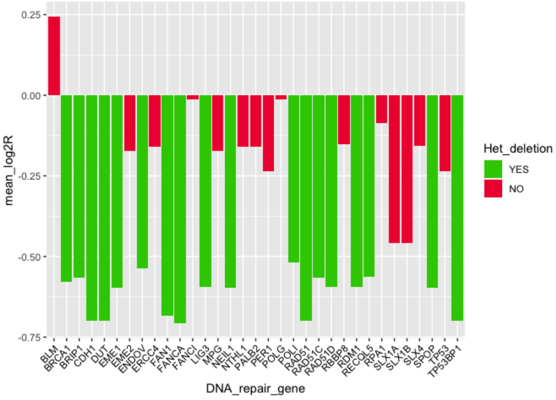
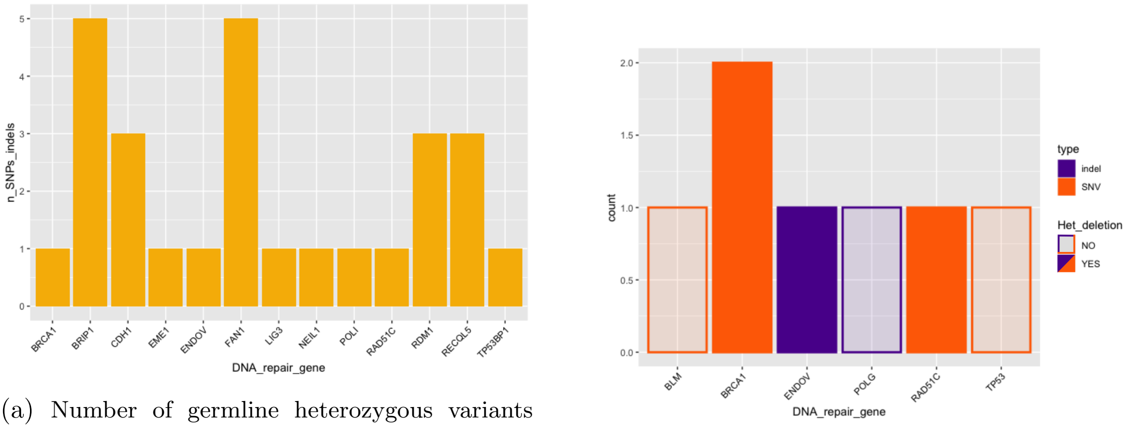


Figure 4: Mean log2Ratio of segments containing DNA repair genes

By intersecting the entire callset of germline SNPs (found with the setup 2, instead of considering only the clinically significant one) with the newly generated .bed file containing the DNA repair genes with heterozygous deletions (see methods), a higher number of variants were individuated (figure 5a). However, it is likely that these variants do not affect greatly the patient's conditions, since none of them were annotated as pathogenic by Clinvar. They can still be associated to disease susceptibility, although this cannot be concluded without further annotation data.

The same workflow was applied to the somatic variant callset (from Mutect, setup 2), this time without considering the clinical relevance and dividing the callset into indels and point mutations; the number of variants with respect to each DNA repair gene (harboring an heterozygous deletion) are reported in Figure 5b.



(a) Number of germline heterozygous variants found across DNA repair genes that underwent a somatic heterozygous deletion. Since only the SNP in BRCA1 was found to be pathogenic, all the callset was used to find the number of germline small variants within each DNA repair gene.

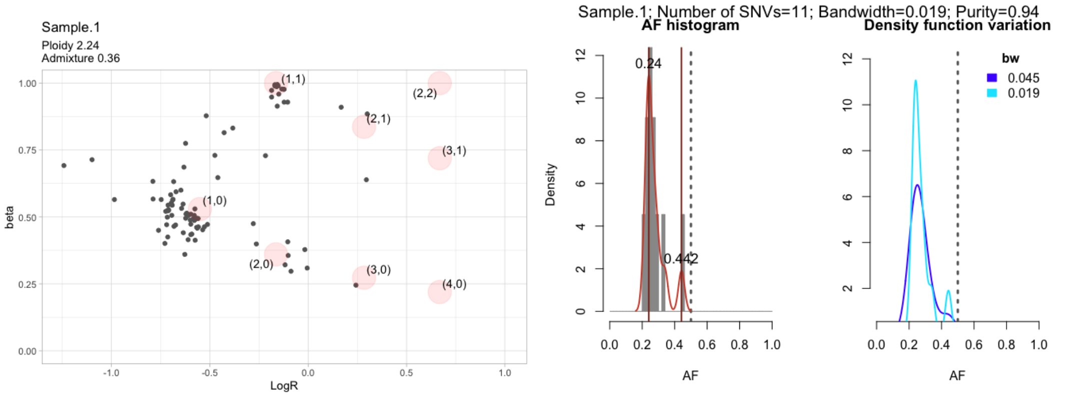
(b) Number of somatic point mutations and indels found in DNA repair genes that harbor a somatic heterozygous deletion

Figure 5: DNA repair genes overlaps with variants.

Tumor purity and ploidy Tumor ploidy and admixture computed with `CLONETv2` resulted of respectively 2.24 and 0.36. Consequently, the estimated tumor purity is 0.64. The log2 ratio values for each considered variant, obtained in the previous section, were plotted against the newly-calculated beta values to better discriminate between different copy-number alterations.

As confirmation of the previous results, a large number of the considered segments fall into a segment with heterozygous deletion, indicated in the figure 6a by the label *(1,0)* representing the copy-number of the major and minor alleles. Allele-specific copy-numbers are represented clearly in the supplementary figure S7.

In contrast, the analysis carried out with `TPES` produced different results, with an estimation of tumor purity equal to 0.94. However, as stated by the `TPES` results summary (6b), only 11 SNVs were considered for this step, a number far too small to have a reliable purity estimation. Moreover, it was observed that this result is quite variable depending on the threshold set on the reference mapping bias (by default set at RMB=0.47).



(a) CLONET results: beta values plotted over the normalized log2 ratio values for each segment

(b) TPES results: the first picture shows allelic fraction distribution of putative clonal and subclonal SNVs within copy-number neutral segments

Figure 6: Tumor purity and ploidy estimation

SPIA analysis SPIA analysis was performed considering the SNPs of the hapmap file which were in common between the tumor and control and gave as result a distance of 0.18809 between the samples, classified as *uncertain*. This score was expected to be low, since both samples come from the same patient. The uncertainty is probably given by the mutations found in the tumor cells, that bring some difference between the two.

Conclusions

Our study was aimed at identifying and characterizing germline and somatic variants, determining the ancestry of the patient, and study the tumor ploidy and purity of a provided breast cancer sample retrieved from TCGA. First, we were able to properly pre-process the given data, following the requirements of different callers; then, both legacy tools and best-practice ones were used to call germline SNPs, somatic CNVs and somatic point mutations. Second, by intersecting the obtained information, we were able to properly assess the status of a pathogenic variant (in *BRCA1*) associated with the tumor that the patient suffers from, linking the presence of this germline heterozygous variant with a subsequent somatic heterozygous deletion. Finally, the tumor sample was characterized in its purity, and the ploidy of the tumor was computed.

Supplementary Figures

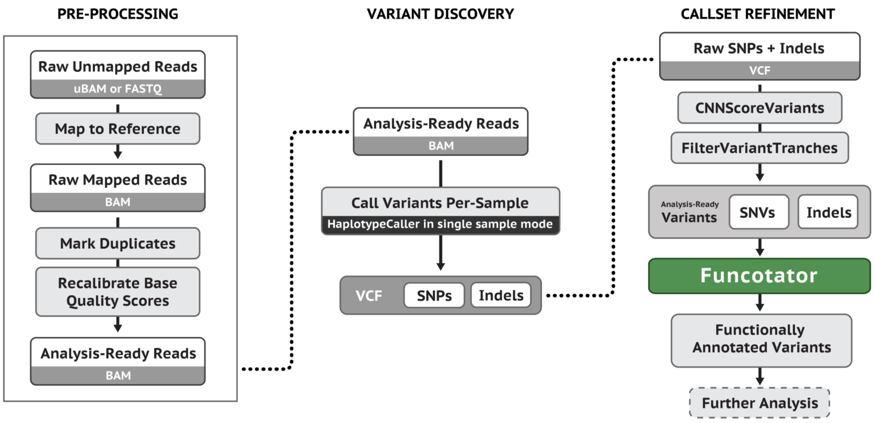


Figure S1: Best-practice workflow for small germline variant discovery taken from GATK website (link)

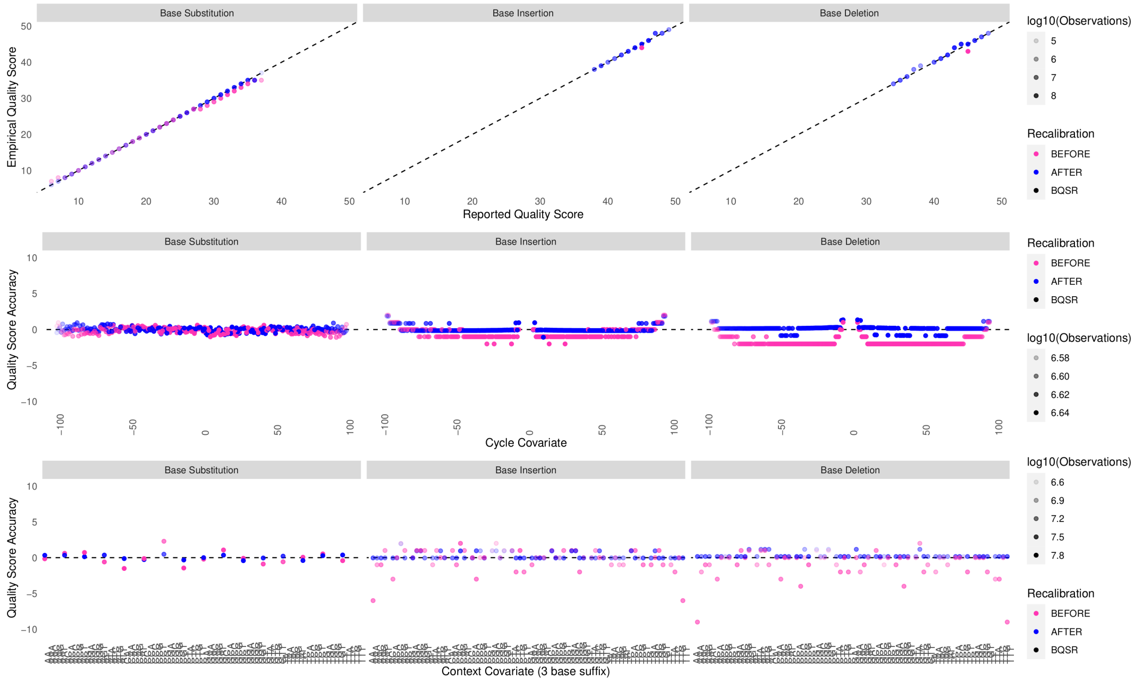


Figure S2: Recalibration control sample (1/2). The ”noisy” distributions were expected here and in the following before/after recalibration plots, since the amount of data processed was small.

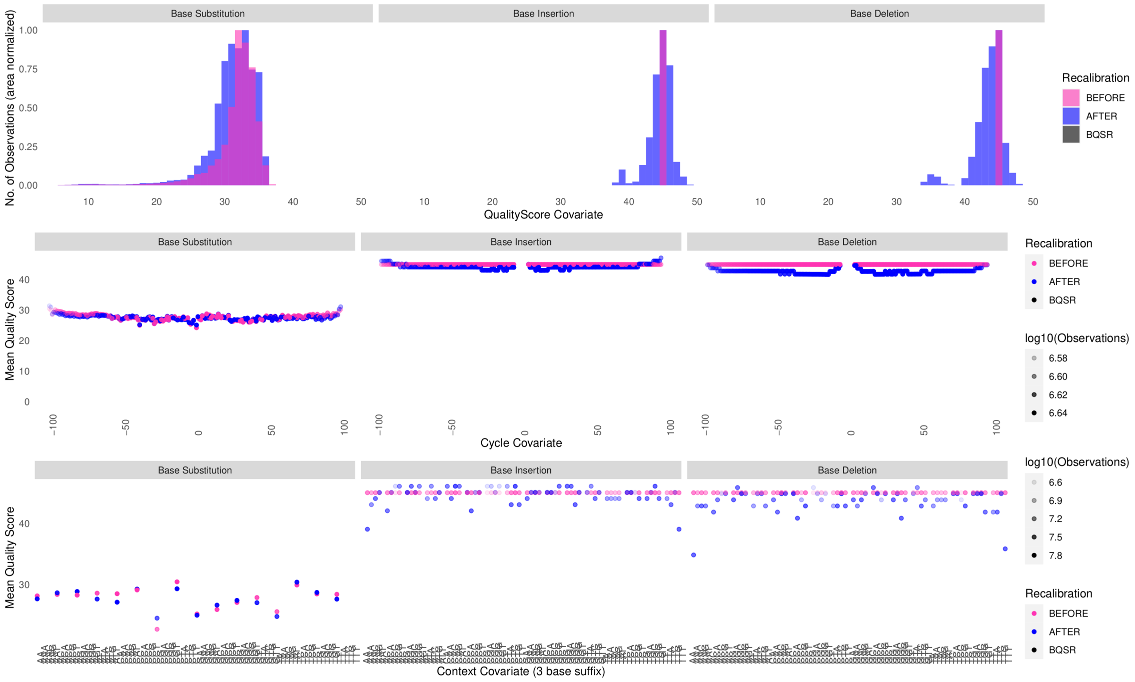


Figure S3: Recalibration control sample (2/2)

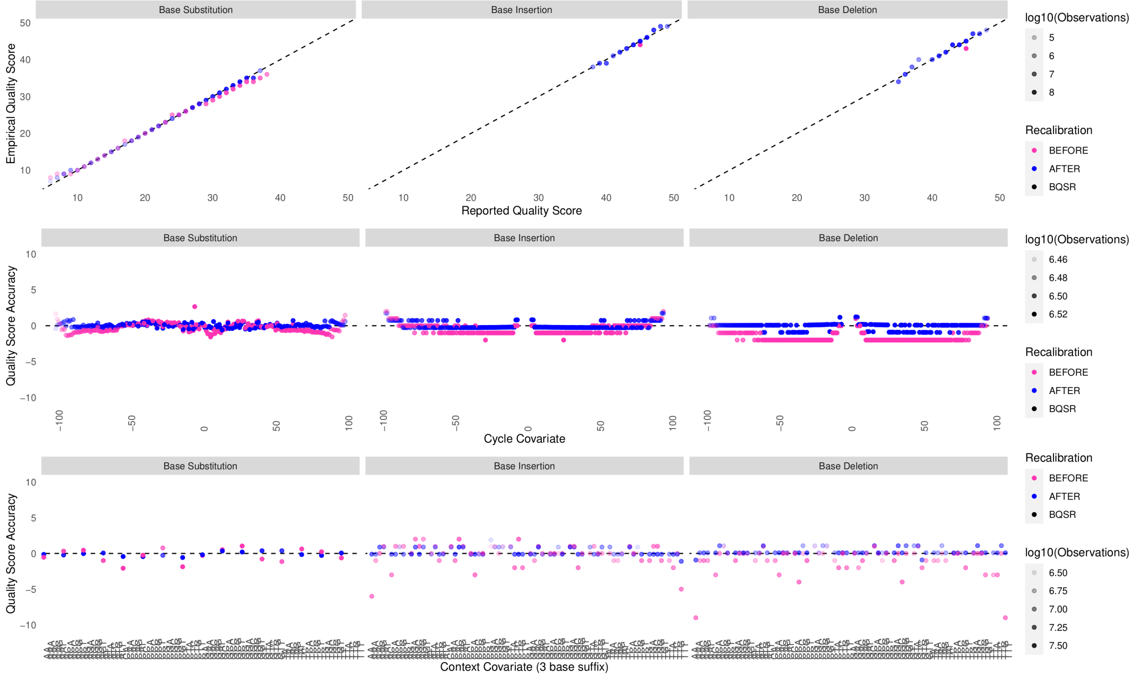


Figure S4: Recalibration tumor sample (1/2)

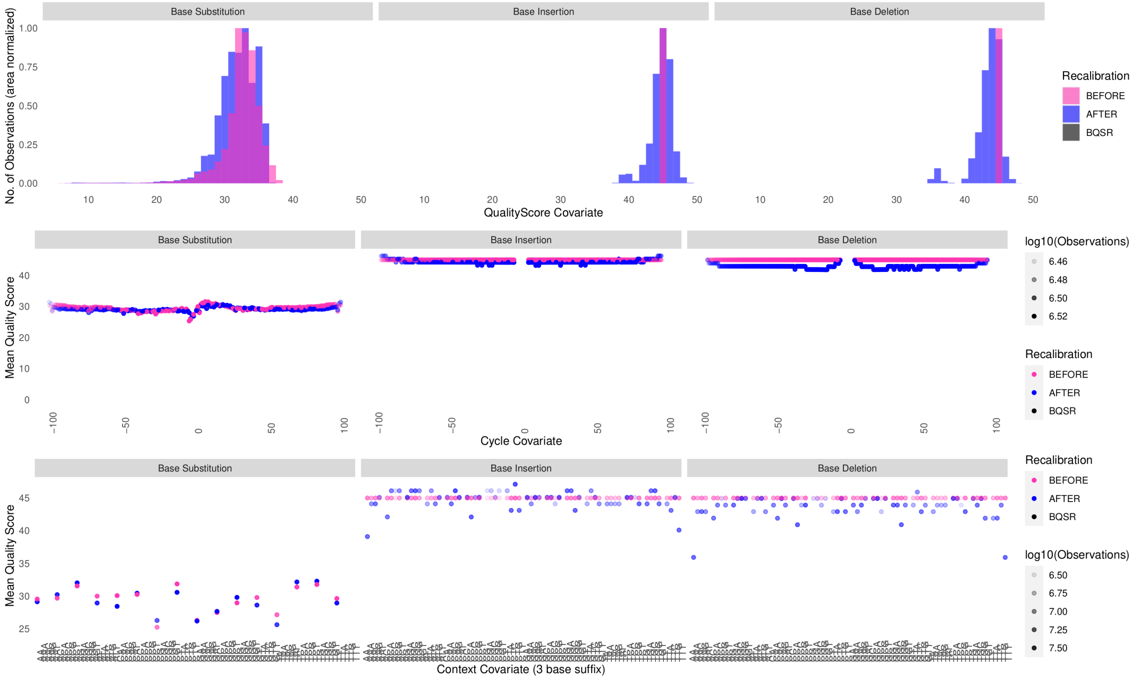


Figure S5: Recalibration tumor sample (2/2)

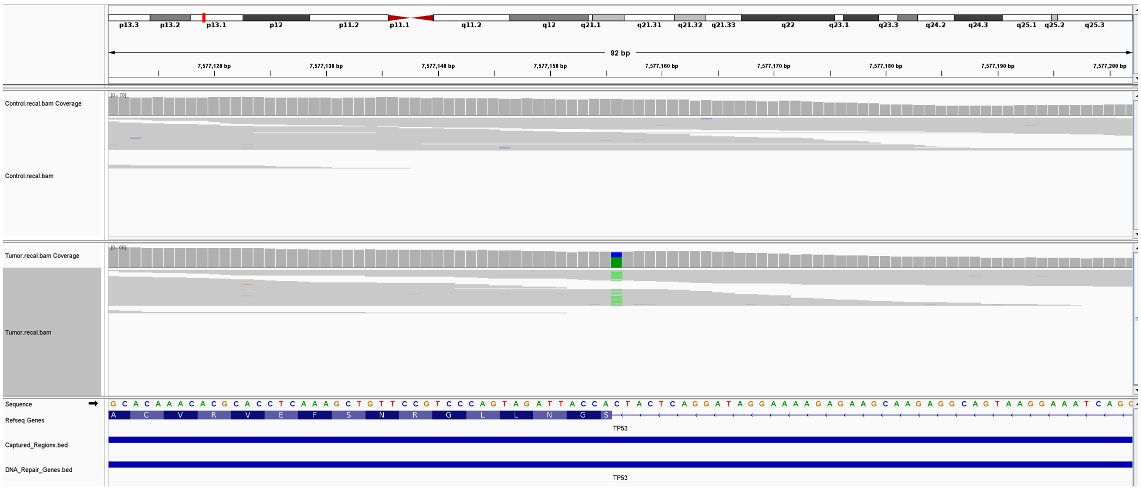


Figure S6: Somatic point mutation identified with Mutect2 in *TP53* at position chr17:7577156

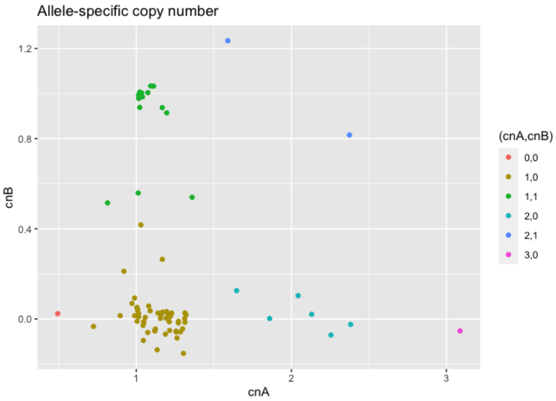


Figure S7: Allele-specific copy-number results obtained with CLONETv2

References

[1] P. Cingolani et al. “A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3”. In: *Fly* (2012).

[2] P. Cingolani et al. “Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift”. In: *Frontiers in Genetics* (2012).

[3] Venkatraman E. Seshan and Adam Olshen. “DNAcopy: DNA copy number data analysis”. In: (2021). R package version 1.68.0.

[4] Alessandro Romanel et al. “EthSEQ: ethnicity annotation from whole exome sequencing data”. In: *Bioinformatics* (2017).

[5] Davide Prandi. “CLONETv2: Clonality Estimates in Tumor”. In: (2021). R package version 2.2.1. URL: `https://CRAN.R-project.org/package=CLONETv2`.

[6] Alessio Locallo, Davide Prandi, and Francesca Demichelis. “TPES: Tumor Purity Estimation using SNVs”. In: (2019). R package version 1.0.0. URL: `https://CRAN.R-project.org/package=TPES`.

[7] Francesca Demichelis and Davide Prandi. “SPIAssay: A Genetic-Based Assay for the Identification of Cell Lines”. In: (2016). R package version 1.1.0. URL: `https://CRAN.R-project.org/package=SPIAssay`.

[8] Ryan Poplin et al. “Scaling accurate genetic variant discovery to tens of thousands of samples”. In: *bioRxiv* (2018).