# Computational Human Genomics Project

Group members

Richard Anderson (mat. 229576), Diego Barquero Morera (mat. 229577), Annalisa Xamin (mat. 232715) and Vilma Ovaskainen (mat. 233688)

## 1. Introduction

The following report details a workflow for pre-processing of next-gen sequencing data and the different types of follow up analysis that can be used in an academic and clinical context for the purposes of better understanding tumor evolution.

## 2. Methods

We analyzed data coming from a specific breast cancer patient and its respective control that were retrieved from the Cancer Genome Atlas database (TCGA). The data had already had some pre-processing applied to it, for instance duplicate sequences had already been removed and only data pertaining to regions within chromosomes 15, 16, 17 and 18 have been retained.

Firstly, we used **samtools** (v1.7) [1] to sort and index the BAM files. The module *sort* allows to sort alignments by the leftmost coordinates and create a new BAM file. The output obtained was then processed with the module *index*. Successively, some general statistics were obtained with the module *flagstat:* in particular, we were able to get the number of properly paired reads.

To perform the realignment, we used **GATK** (v3.8-1-0) [2]. We used the tool *RealignerTargetCreator* to identify what regions needed to be realigned and then, with the tool *IndelRealigner*, the actual realignment was performed. As a reference genome we used the assembly version GRCh37 from fasta file human_g1k_v37.fasta. We focused this realignment over the genomic interval saved as Captured_Regions.bed.

GATK was also used to perform the recalibration step. In this case, we first used the tool *BaseRecalibrator* to detect systematic errors in base quality scores and recalibrate them, then the module *PrintReads* to write the recalibrated data to file and finally *AnalyzeCovariates* was used to make before/after plots.
The workflow for the recalibration begins with building the recalibration model. To do that, the *BaseRecalibrator* module requires as input the known sites (hapmap_3.3.b37.vcf) and the previously sorted and realigned BAM file. We also specified the genomic interval to focus the recalibration and the reference sequence. The recalibration table we generated as output was then used along with the realigned BAM file as input for the *PrintReads* module to get the recalibrated BAM file. After this first recalibration, we built the after model to evaluate remaining errors. To do that, the module *BaseRecalibrator* generates as input the recalibration table, the realigned BAM file and the known sites. We obtained a second recalibration table containing the remaining errors. Finally, we generated the before-after plots using the *AnalyzeCovariates* module, which used the two recalibration tables as input.

To perform the variant calling we used the module *UnifiedGenotyper* from GATK. This tool requires as an input the recalibrated BAM file. We also specified the reference genome and the genomic interval (saved in Captured_Regions.bed) to focus on. We obtained a .vcf file as output, from which we filtered out specific

variants using **vcftools** (v0.1.15)[3]. We filtered the variants specifying the following parameters: `--minQ 20` to include only sites with quality value above the threshold of 20; `--max-meanDP 200` and `--min-meanDP 5` to include only sites with mean depth values (over all included individuals) greater than or equal to the "--min-meanDP" value and less than or equal to the "--max-meanDP" value; `--remove-indels` to remove any indel sites; `--recode` and `--recode-INFO-all` to respectively generate a new VCF file after applying the filtering options specified by the user and to define an INFO key name to keep in the output file. From the VCF file obtained after the filtering step, we selected only the heterozygous SNPs: we used *grep* to select the variants with PL 0/1 (where PL is the “normalized” Phred-scaled likelihoods of the possible genotypes; in a diploid organism). Then, the result was saved as a new VCF file to proceed afterwards with the variant annotation.

The VCF files were then sequentially annotated using variant callers, using pre-existing databases. It is a crucial step in linking sequence variants with changes in phenotype. In our case, we used **SnpEff** (v4.3t)[4] to annotate and predict the effects of genetic variants on genes and proteins (such as amino acid changes). We also used **SnpSift** (v4.3t)[5] to annotate the SNPs using information from the Hapmap 3.3 and Clinvar Pathogenic databases.

In order to determine the ancestry of the patient we used **EthSEQ** [6]: in particular, we used the recalibrated BAM file of the control sample as input for RunEthSEQ.R script, for which a pre-exisitng model was supplied.

To identify somatic copy number we used **Varscan2** (v2.3.9) [7]: first we used samtools mpileup as the input for the *copynumber* module (which determines relative tumor copy number from tumor-normal pileups). Then, this output was analyzed with the *copyCaller* module (which performs the GC-adjust and processes copy number changes from VarScan *copynumber* output). This output of this analysis was saved and subsequently used to help estimate the tumor purity.

The identification of somatic point mutations required the execution of samtools mpileup on both Tumor and Control recalibrated BAM files. The pileup files generated were used as input for the *somatic* module of Varscan2. As parameters we specified the output file for SNP calls with `--output-snp`; the output file for indel calls with `--output-indel` and then choose VCF as output format with `--output-vcf 1`.

The **DNAcopy R package** [9] implements the circular binary segmentation (CBS) algorithm to segment DNA copy number data to identify changes in copy number across genomic regions. Briefly, the output from the varscan copy number output was used as input into the package. This was then processed with the `CNA` function, its output was then processed by the `smooth.DNA` function and finally this data was used by the `segment` function. Within this function, the `sdundo` option was selected as the `undo.splits` method. `Minwidth`, `alpha` and `undo.SD` variables were adjusted until there appeared to be fit that followed the major trends seen in each chromosome. Larger values of Minwidth counterintuitively appeared to produce shorter sections with similar log2 ratio values and more stringent undo.SD parameters appeared to be less able to deal with the major breaks seen around indices 50000 and 75000. Alpha values were used to reduce the occurrence of very small regions where the mean log2 ratio appeared to be quite different from the flanking regions despite no obvious differences in distribution. A final alpha value was decided as it appeared to reduce the occurrences of these small regions but still follow some of the trends observed in chromosome 18.

**Bedtools** (v2.26.0)[8] allows the intersection, merging, counting, complementing, and shuffling of genomic intervals from multiple files in widely-used genomic file formats such as BAM, BED, GFF/GTF, VCF. In our project, we used the method `intersect` to determine which DNA repair genes (saved in

DNA_Repair_Genes.bed) overlap both heterozygous deletions and heterozygous SNPs of the patient that are in Clinvar. We used *grep* to filter out the heterozygous deletions from the SCNA.copynumber.called file obtained in the previous step. To perform the intersection, we specified the files to compare with the options `-a` and `-b`. The output was then saved in a new BED file.

Then, we determined which DNA repair genes overlap both heterozygous deletions and somatic point mutations of the patient. As before, firstly we used *grep* to filter the somatic point mutations from the somatic.pm.vcf. Then, we used *bedtools* to intersect the somatic point mutations with the DNA repair genes and then the heterozygous deletions with the somatic point mutations. The output was saved in a new BED file.

To determine the purity of the tumor sample we used both the **CLONET** [10] and **TPES** [11] R packages with the mpileup and somatic copy number files as the inputs for both. For CLONET, the functions, `compute_polidy` and `compute_beta_table`, were first run and used in the `compute_dna_admixture`. All three outputs were then used in the `check_ploidy_and_admixture function` to generate a visualization of the data. Only somatic mutations from the somatic point mutation pilup-file were used for the TPES analysis. The `TPES_purity` function was used to obtain a summary of what data was used to calculate TP. For this we used the following parameters in with the TPES_purity function: `RMB = 0.47, maxAF = 0.6, minCov = 10, minAltReads = 10, minSNVs = 1`. In particular: `RMB` is the Reference Mapping Bias Value (we kept the default value); `maxAF` is the filter on the allelic fraction (AF) distribution of SNVs (this is necessary to be sure to keep only heterozygous SNVs. Clonal and subclonal SNVs, which have an AF greater than maxAF, will be removed); `minCov` is the minimum coverage for a SNV to be retained; `minAltReads` is the minimum coverage for the alternative base of a SNV to be retained; `minSNVs` is the minimum number of SNVs required to make a purity call.

**SPIA R package** [12] was used to determine the similarity of Tumor and Normal samples, taking advantage of the Hapmap annotations that were common in both the Control and Tumor genomes. Thresholds for genotyping were the following: homozygous for reference base if AF < 0.2, homozygous for alternative base if AF > 0.8 and heterozygous if 0.2 < AF < 0.8.

## 3. Results and Discussion

The number of properly paired reads in the normal BAM file was 19 613 806 (99.33% of the reads) and 14 979 936 (99.67 %) for the tumor BAM file, which means nearly all reads in both files were in the correct orientation and passed various quality control thresholds such as mapping to the same chromosome. For the control and tumor BAMs, 3158 and 2267 reads were realigned, respectively. After recalibration 65.9% of the reads in normal realigned BAM-file had their base quality score changed and 63.7% for the tumor realigned BAM-file. The before-after plots of the recalibration process are shown for both normal and tumor files in figures 1 and 2. In both samples, we noticed a higher correlation between reported quality score and empirical quality score after recalibration, indicating that the process increased the accuracy of reported quality. The same is observed for cycle covariate (position of base in the read) and context covariate (base context), as the distribution is spread further from zero before the recalibration.

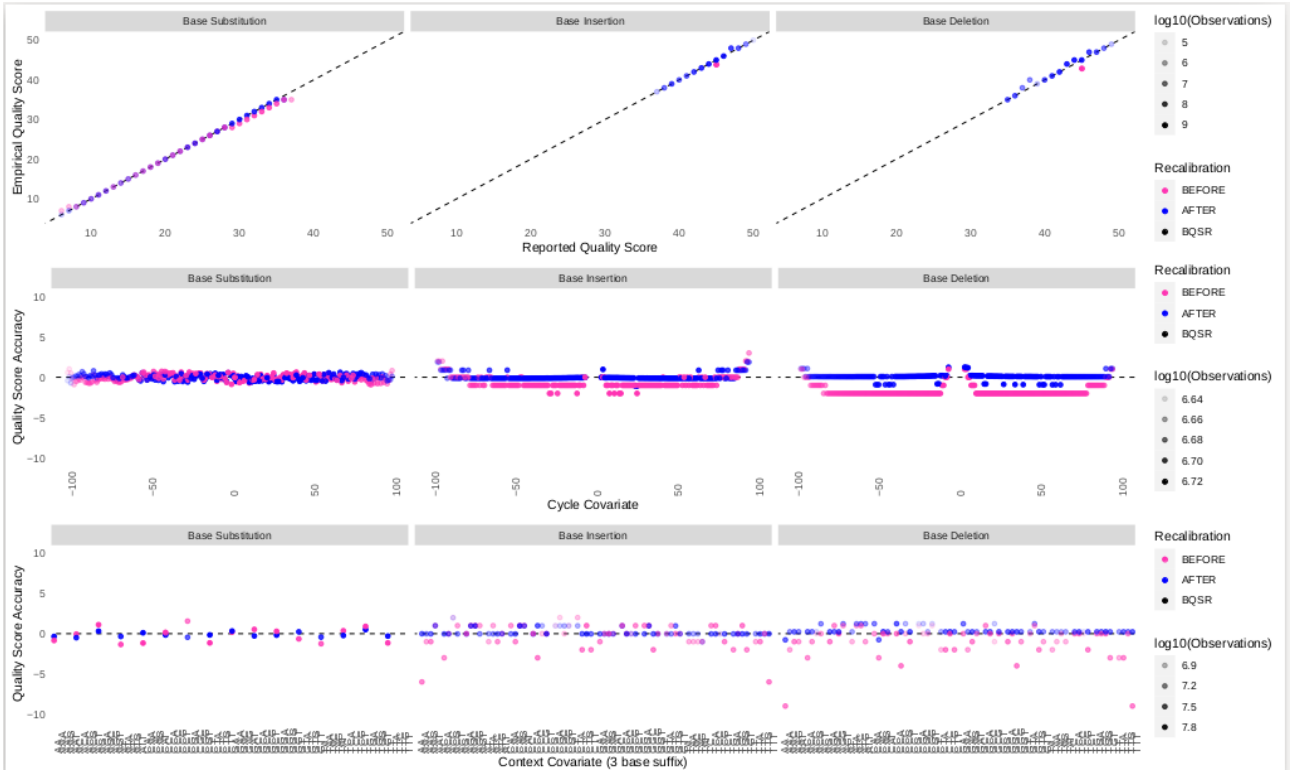


*Figure 1. Recalibration of quality distributions for normal sample.*

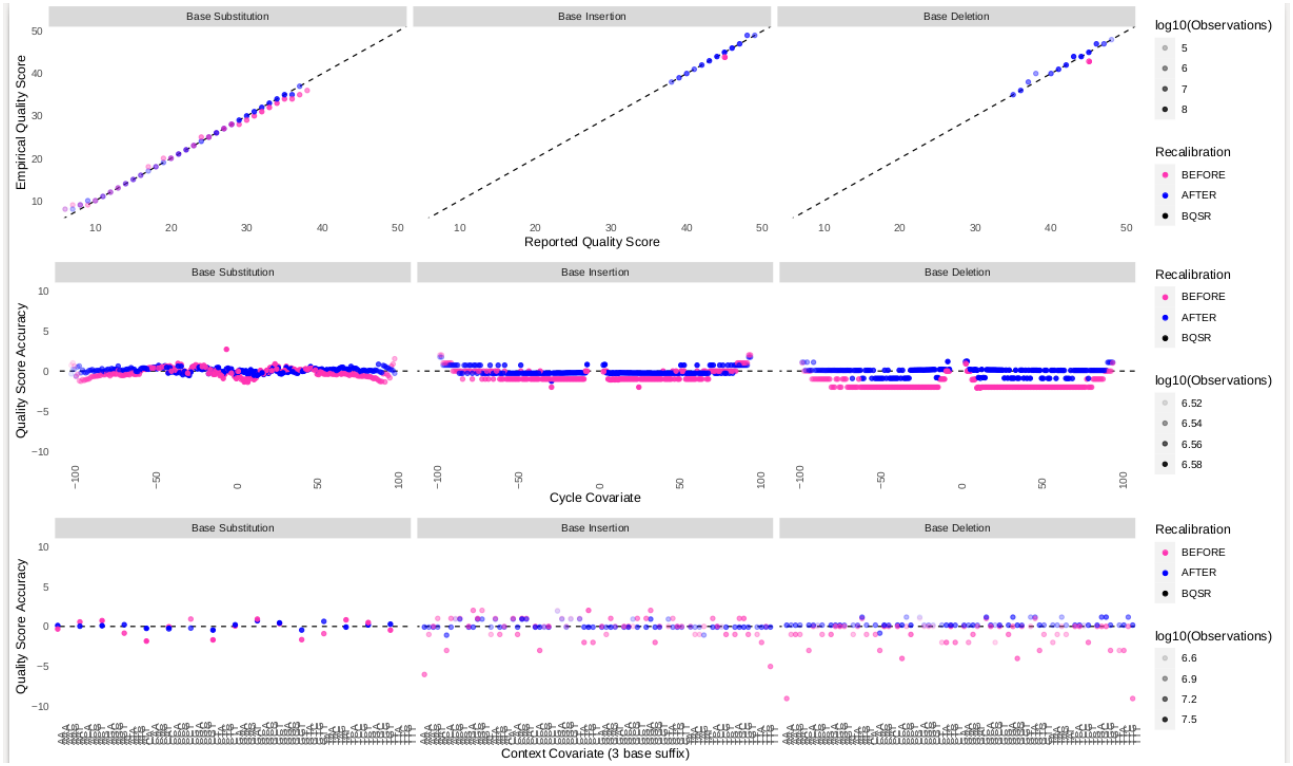


*Figure 2. Recalibration of quality distributions for tumor sample.*

## 3.1 Identification of germline single nucleotide variants

The total number of SNPs in the control BAM file was 9941, from which heterozygous were 7105. For the tumor BAM file the number of SNPs was 9601, from which 6719 were heterozygous. For both control and tumor sample, we identified one SNP with clinical significance classified as "pathogenic". This SNP (17:41246494, C⟶ A) is located in the gene BRCA1, which encodes for the breast cancer 1 susceptibility protein. The SNP is classed as nonsense mutation as it changes the amino acid coding codon to a stop codon leading to a premature termination of translation.

## 3.2. Ancestry of the patient

The patients ethnitical background was determined using the EthSEQ tool comparing it to a supplied model file. Based on the sample clustering the data suggests the patient is of African ancestry (see figure 3).

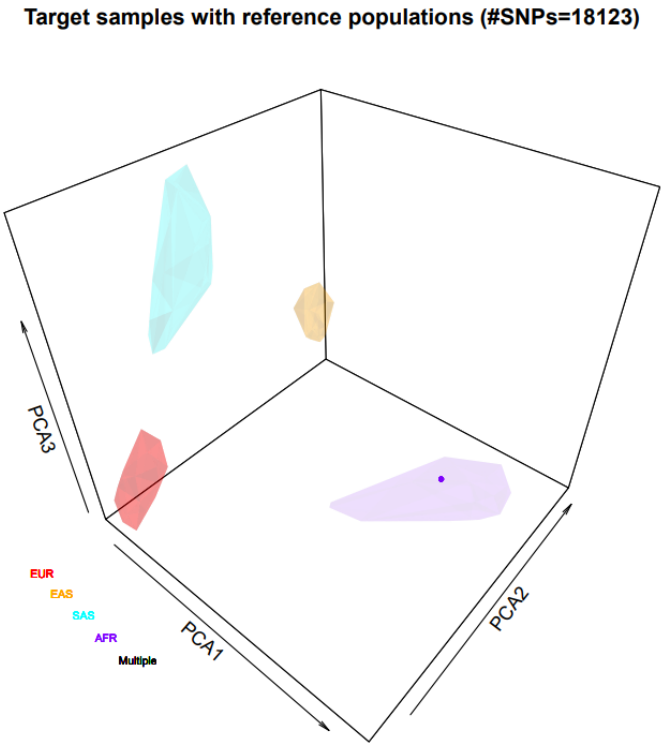


*Figure 3. Principal component analysis (PCA) of patient ancestry.*

## 3.3 Somatic copy number aberrations

Circular binary segmentation (CBS) was used to identify copy number aberrations in the tumor sample. The analyzed regions span exon coding regions from chromosomes 15 to 18. The result of the CBS analysis is presented in figure 4. Large stretches are observed to have undergone loss of heterozygosity (LOH) as the segments (log2 ratio of tumor signal over control signal) in the figure 4 are close to -1, rather than 0. This is because for a 100% pure sample from a diploid tumor, a log2(T/N) score of -1 would correspond to copy number state 1, meaning that an allele had been lost. However, it's rare that a tumor sample has a purity of 100%, therefore, a value less than -1 of log2 ratio is observed for the copy number state 1. Given a tumor purity estimation from 65% – 95% (see section 3.7), we expect LOH events to have log2(T/N) ratio between

-0.5 – -1. Based on this expectation, it can be observed that large numbers of LOH events within each of the chromosomes.

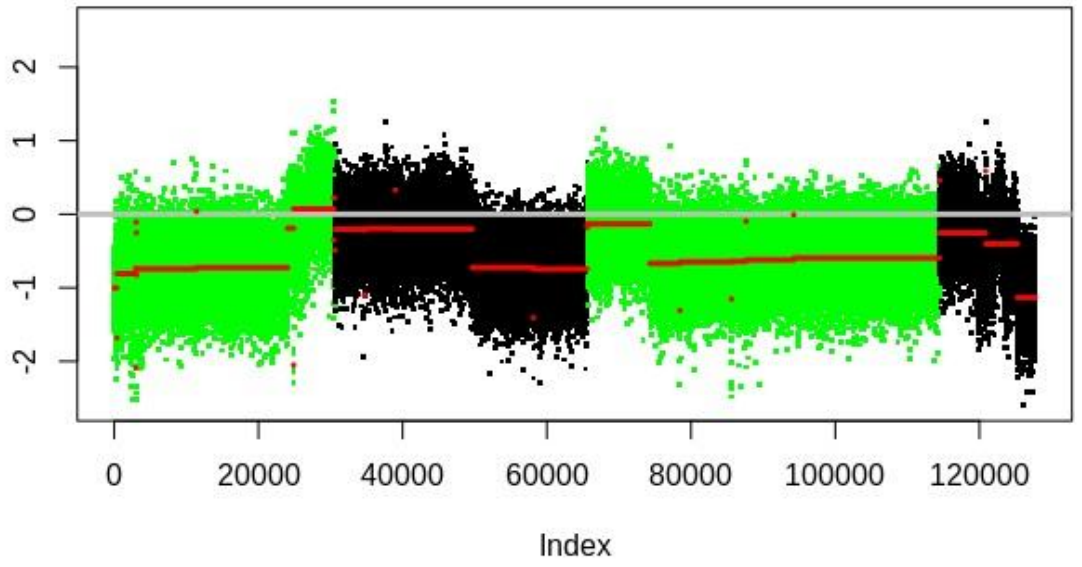


*Figure 4. CBS analysis for breast cancer patients. Genomic positions are presented on the x-axis. Y-axis is the log2 ratio of tumor/normal signal. Red lines represent the mean log2 ratio for that region, mean width was set to 2, alpha threshold of 0.00001 and undosd parameter of 2sd. These settings allowed the results of the analysis to more closely resemble the global patterns that can be seen by visually inspecting the plot. Chromosomes are arranged sequentially with chromosome 15 on the left and different colors have been used to differentiate each chromosome from their neighbor.*

## 3.6 Somatic point mutations

Varscan 2 was used to identify different classes of mutations: which were somatic point mutations (SPM), germline mutation or LOH. The results are summarized in table 1. As a false positive correction method was not implemented, the number of mutation calls are also presented after applying two generic P-value thresholds (with P< 0.001 being the most stringent). For the tumor sample, LOH was the most common call, which is to be expected as the genomic regions analyzed showed high levels of monoallelic deletions, figure 4.

*Table 1. Summary from Varscan 2 point mutation analysis*

| Type | Total[1] | # P< 0.05[2] | # P< 0.001[2] |
|---|---|---|---|
| Germline[a] | 10716 | 9213 | 9117 |
| Somatic[b] | 219 | 177 | 74 |
| LOH[b] | 2882 | 2882 | 1878 |
| Unknown[b] | 16 | 15 | 12 |

[1]Analysis was performed without a P-value cutoff
[2]No. of point mutations that had an un-adjusted P-value less than the value indicated
[a]Variant_p_value parameter used
[b]Somatic_p_value parameter used

## 3.7. Aberrations in the DNA repair genes

Bedtools was used to search for an overlap between known DNA repair genes and identified regions harboring a copy number deletion, including monoallelic deletions (all the regions with a log2(T/N) significantly lower than 0) (figure 5). In total, 1174 monoallelic deletions were observed to overlap with 32 DNA repair genes, namely: BLM, BRCA1, BRIP1, CDH1, DUT, EME1, EME2, ENDOV, ERCC4, FAN1, FANCA, FANCI, LIG3, MPG, NEIL1, NTHL1, PALB2, PER1, POLG, POLI, RAD51, RAD51C, RAD51D, RBBP8, RDM1, RECQL5, RPA1, SLX1A, SLX4, SPOP, TP53 and TP53BP1 genes.

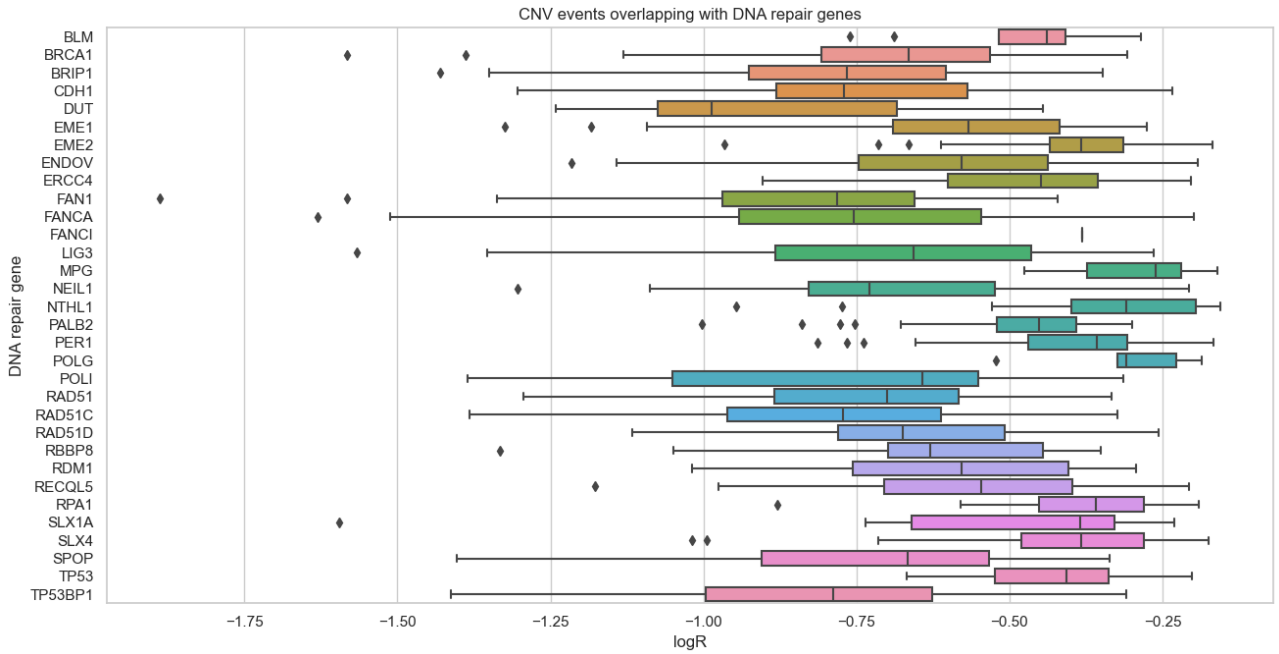


*Figure 5. Distribution of logR values for the monoallelic deletions found to overlap with DNA repair genes.*

Furthermore, the monoallelic deletion regions overlapped with the previously identified pathogenic SNP, BRCA1 (section 3.1), indicating that only the pathogenic variant of the allele is remaining. As the patient has lost the allele harboring the healthy gene, only the truncated protein will be expressed. We then searched for whether any of the observed SPMs occurred in DNA repair genes. Based on the analysis, only two genes, FANCI and TP53, were found to have a SPM located within their region. Other DNA repair genes affected by both SPMs and deletion events were not identified.

## 3.8. Estimation of tumor purity

Tumor purity (TP) was estimated by two different methods. First we used CLONET, which estimated the admixture of the tumor sample to be 0.36, corresponding to TP (1- admixture) of 64%. TPES estimated TP to be 92%. The estimation of TP by TPES relies on the distribution of the variant allele frequency (VAF) of single nucleotide variants (SNVs) in copy number neutral segments of the genome, which are filtered from the segmentation data with thresholds i.e. $-0.1 < \log2(T/N) < 0.1$. To determine TP in our data TPES only used two SNVs in its calculation. According to the authors of TPES, there is a high concordance between TP estimation of CLONET and TPES when the number of applicable SNVs is >9 [11]. Therefore, in the case of our data we expect the TP estimation to be more trustworthy by CLONET than TPES, as TPES did not have the minimum required number of SNVs to perform an accurate estimation of TP.

In figure 6, genomic segments are plotted in beta-logR space by CLONET. Red circles represent the expected areas for segment clustering with an allele copy number specified in the brackets. The majority of the genomic segments in figure 6, cluster around copy number (1,0) with a minor cluster at (1,1). We hypothesize that the bulk of tumor cells contribute to the segment in (1,0) corresponding to LOH, whereas the segments clustering around (1,1) are representative of a diploid genome and are the result of admixture from healthy tissue and tumor cells who still habour a WT genotype in some genomic regions.

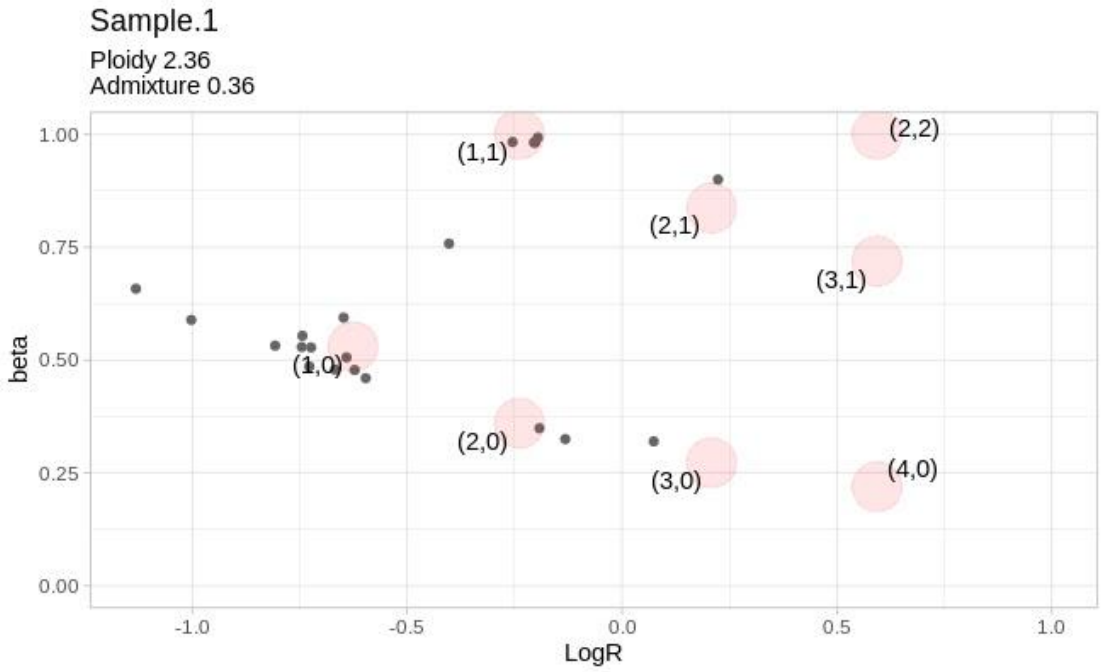


*Figure 6. Beta - LogR space for copy numbers after ploidy and purity adjustment. Beta represents the percentage of neutral reads in a given genomic position.*

## 3.9. Genetic distance of tumor and control sample

Genotyping of samples was performed based on the frequency of observing an alternative base in a read in respect to total number of reads (allelic frequency, AF). After obtaining the allelic fraction for each SNP position in the hapmap_3.3.b37.vcf file, a total of 13727 SNPs were found to be in common for both samples. Result of the SPIA test is shown in figure 7. The tumor and control samples had a genetic distance of 0.068, which correspond to matching cell lines. As the samples were coming from the same patient, a matching genetic origin was expected with some variation due to somatic aberrations in tumor tissue.

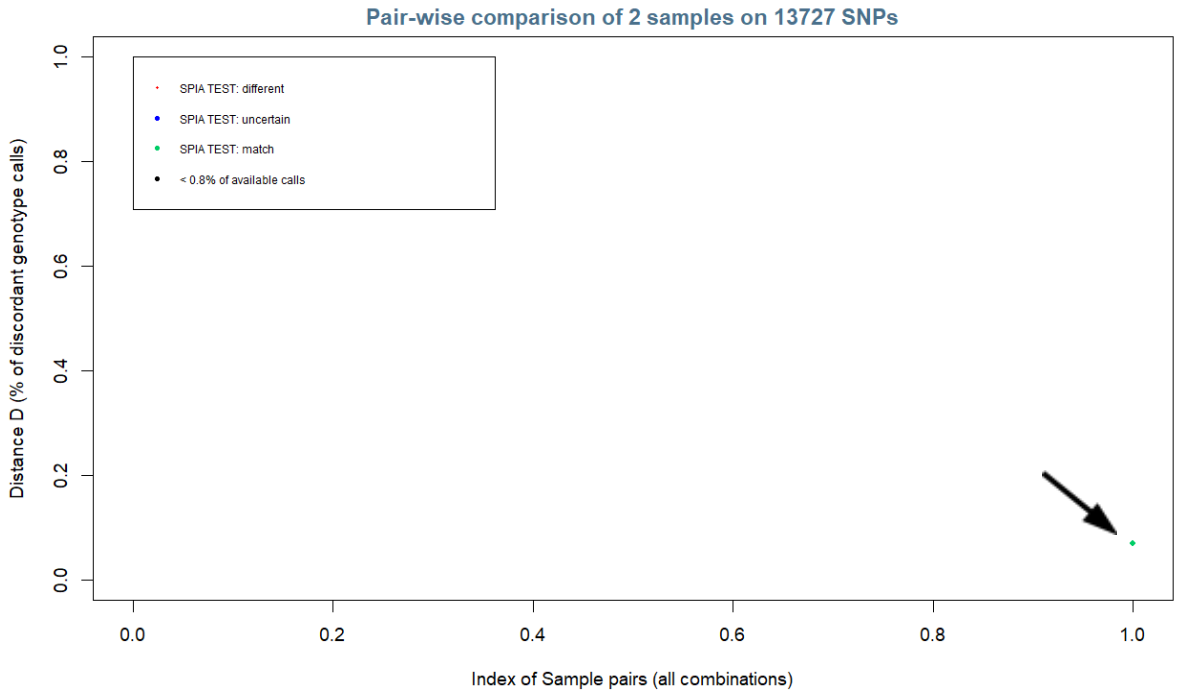


*Figure 7. Comparison of genetic distance between the tumor and control sample.*

## 4. Conclusions

The analysis suggests that the person whose data was used in this study was heterozygous for a nonsense SNP mutation in the BRAC1 gene, alterations in this gene is known to be associated with an increased risk of developing breast cancer. The tumor, at the time that the sample was taken and based on the regions in chromosome 15-18 that were analysed, appear to have developed several thousand LOH mutations which appears to include the loss of other, presumably, functional allele of BRAC1 and 32 other genes associated with DNA repair and regulation of the cell cycle. A further two genes appeared to be affected by SMPs but the functional consequences of these mutations was not assessed.

# References


[1] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008. PMID: 33590861; PMCID: PMC7931819. https://doi.org/10.1093/gigascience/giab008

[2] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19. PMID: 20644199; PMCID: PMC2928508.

[3] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics, Volume 27, Issue 15, 1 August 2011, Pages 2156–2158, https://doi.org/10.1093/bioinformatics/btr330

[4] "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.", Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. Fly (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672 https://doi.org/10.4161/fly.19695

[5] "Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift", Cingolani, P., et. al., Frontiers in Genetics, 3, 2012. https://doi.org/10.3389/fgene.2012.00035

[6] Romanel A, Zhang T, Elemento O, Demichelis F. EthSEQ: ethnicity annotation from whole exome sequencing data. Bioinformatics. 2017 Aug 1;33(15):2402-2404. PMID: 28369222; PMCID: PMC5818140. https://doi.org/10.1093/bioinformatics/btx165

[7] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar;22(3):568-76. Epub 2012 Feb 2. PMID: 22300766; PMCID: PMC3290792. https://doi.org/10.1101/gr.129684.111

[8] Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842. https://doi.org/10.1093/bioinformatics/btq033

[9] Seshan VE, Olshen A (2021). DNAcopy: DNA copy number data analysis. R package version 1.68.0.

[10] Prandi, D., Baca, S.C., Romanel, A. et al. Unraveling the clonal hierarchy of somatic genomic aberrations. Genome Biol 15, 439 (2014). https://doi.org/10.1186/s13059-014-0439-6

[11] Locallo, A., Prandi, D., Fedrizzi, T., & Demichelis, F. (2019). TPES: Tumor purity estimation from SNVs. *Bioinformatics*, *35*(21). https://doi.org/10.1093/bioinformatics/btz406

[12] Francesca Demichelis, Heidi Greulich, Jill A. Macoska, Rameen Beroukhim, William R. Sellers, Levi Garraway, Mark A. Rubin, SNP panel identification assay (SPIA): a genetic-based assay for the identification of cell lines, Nucleic Acids Research, Volume 36, Issue 7, 1 April 2008, Pages 2446–2456, https://doi.org/10.1093/nar/gkn089