Computational Human Genomics Project

Characterization of Somatic Events in the Genome of an Oncologic Patient

Andrea Policano
Alberto Lupatin

August 28, 2024

1 Project Rationale

The aim of this project is to assess the difference between the DNA sequences of a control and a tumor sample, both coming from the same patient. Differences will be evaluated in terms of the genetic characteristics or traits that allow us to better distinguish the two samples.

The project will be executed using several computational tools in a virtual Linux environment. This setup allows us to exploit an older version of Java in order to use specific tools.

The project will be divided into six steps, with each one focusing on a particular genomic analysis. Starting with the pre-processing of the given BAM files, the variants will be identified through the following steps: somatic copy number (CN) calling, variant calling, and variant annotation. Ancestry analysis will follow to identify the degree of admixture. Finally, the purity and ploidy will be computed, and the log2R-β space will be plotted.

2 Methods

The provided data at the start of the project were two *".bam"* files resulting from the sequencing of a tumor and a control DNA sequence from the same individual.

All the following analyses were performed in a Linux Ubuntu 18.04 virtual environment using VirtualBox 7.0.16. Java Runtime Environment was used through the *java-1.8.0-openjdk* package. If not specified otherwise, all the scripts were executed in *bash*.

All the scripts (both in *bash* and *R*), the terminal outputs, and the plots of each step can be found on our GitHub page.

In order to improve the accuracy of the results of the following steps and to avoid sequencing biases, some preliminary tools were applied to the given *bam* files.

1. Firstly, **sorting** and **indexing** were performed using the *samtools v. 1.20* package on both control and tumor files by using, respectively, the *sort* and the *index* parameters.

2. Secondly, **realignment** was performed in two steps, aiming to remove position artefacts. Firstly, the genomic regions harbouring position artifacts were found with *RealignerTargetCreator* tool in *GATK* - (Genome Analysis ToolKit v. 3.8-1-0-gf15c1c3ef). Then, output files were exploited by the *IndelRealigner* tool (in *GATK*) to perform the actual realignment.

3. Thirdly, the **base quality score recalibration** was performed to avoid biases due to systematic sequencing errors. This step was done by using 3 tools in *GATK*:
    - *BaseRecalibrator*: generates a recalibration table based on the known polymorphic sites (on the *hapmap_3.3.b37.vcf* provided file)
    - *PrintReads*: writes reads from the sorted *BAM* file that pass the criteria contained in the newly generated recalibration table into a new *BAM* file.
    - *BaseRecalibrator*: redoes the first step implementing not only the known polymorphic sites but also the recalibration table.
    - *AnalyzeCovariates*: evaluates the base quality score recalibration by comparing the initial and the final *bam* files and generating plots from them.
4. Lastly, the marking of the **duplicates** was performed in order to choose the best read among them. To locate and tag the duplicates, the *MarkDuplicates* tool in the *Picard v. 2.22.3* toolbox was used on the already sorted, realigned, and recalibrated *BAM* files. Finally, the output file was indexed through the *samtools index* tool.

In order to identify the genomic variants harbored by the control and the tumor sample, variant calling was performedby exploiting both *BCFTools* and *UnifiedGenotyper* by *GATK*. Additionally, to store the results in a *VCF* file (Variant Call Format), the *vcftools v. 0.1.16* package was used.

The output files from *BCFTools* and *UnifiedGenotyper*, containing the genomic variants, were given as input to *SnpEff v. 5.2c* and *SnpSift v. 5.2c*. *SnpEff* was used to annotate and predict the effects of genetic variants on genes, while *SnpSift* annotates the genomic variants based on the haplotype map (*HapMap 3.3_37*) and *ClinVar* database data.

Somatic CN changes between the control and tumor samples were inferred by performing a pairwise comparison of read depth between the samples at each position exploiting *VarScan v.2.3.9* from the *samtools mpileup v. 1.20* output. Then, the circular binary segmentation algorithm were used through *DNAcopy v. 1.76.0* R library and plots of the results were generated. The plots analyzed the distribution of the log ratio through the implementation of the smooth and segment algorithm.

SNVs in the control sample were found and annotated using *samtools*, *VarScan*, and *vcftools*. Then, to find the somatic single nucleotide variations, *VarScan* was used with the *somatic* parameter. SNVs were then annotated with *SnpEff* and the most relevant genes (i.e. those that are high impact, depth ¿ 20, existing ID) were extracted with *SnpSift*. From the list of extracted genes, enrichment was performed by using *EnrichNet* and *DAVID* online tools.

To identify the population from which the data is derived and its degree of admixture, ancestry analysis was performed using: *EthSEQ v. 3.0.2* package in R, which also returns the ethnicity annotations.

Purity and ploidy values have significant importance in the clinical interpretation of the results. Firstly, variant calling was performed using the *samtools call* function. Then, all the heterozygous genotypes were filtered. Subsequently, *ASEReadCounter* from *GATK* applied (in both control and tumor data) the following filters:

- *minDepth 20*: minimum number of bases
- *minMappingQuality 20*: minimum read mapping quality
- *minBaseQuality 20*: minimum base quality

Lastly, purity and ploidy were calculated and the logR-β graph was depicted in R using the following packages: *data.table v. 1.15.4, CLONETv2 v. 2.2.1, TPES v. 1.0.0.*

3 Results

Results from the variant calling step showed that GATK's tool (*UnifiedGenotyper*) was able to find more variants compared to Bcftools. In particular, the latter found almost half of the variants present in the tumor sample (see table 1 for more info).

	Common	**Only in main**	**Only in second**	**Non-matching**
BCFTools	25677	10066	5705	0
UnifiedGenotyper	28743	14250	10966	0

Table 1: Comparison of VCF files using BCFTools and UnifiedGenotyper

From the somatic CN calling step, we can infer that some regions of the genome under went a hemizygous deletion and that, overall, there is a loss of CN (Figure 1).

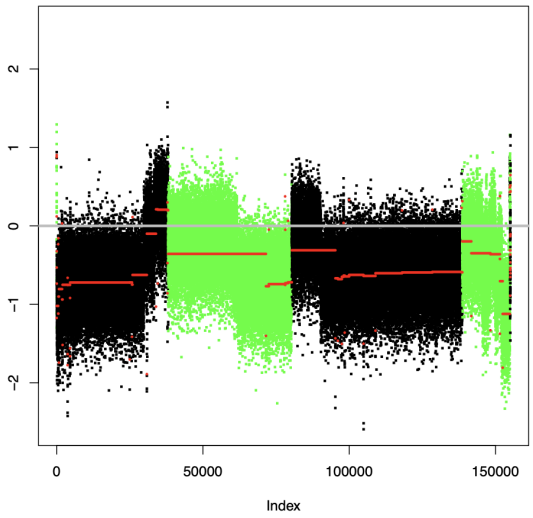

Figure 1:

Plot showing the results from the implementation of CBS algorithm, depicting the genomic region (X-axis) with the log ratio (Y-axis). Also, the computed mean CN level were represented through the red line.

In the Variant Annotation step (Figure 2), similar results between control and tumor groups were observed, with all of them showing a high percentage of silent (56%) and missense (43%) mutations. The transitions/transversions ratios were similar across GATK and BCF, even though the latter was able to find less SNPs in both samples.

Type (alphabetical order)	Count	Percent
DOWNSTREAM	17,021	13.803%
EXON	11,778	9.551%
INTERGENIC	8,868	7.192%
INTRON	64,651	52.429%
SPLICE_SITE_ACCEPTOR	12	0.01%
SPLICE_SITE_DONOR	11	0.009%
SPLICE_SITE_REGION	1,262	1.023%
UPSTREAM	12,280	9.959%
UTR_3_PRIME	5,670	4.598%
UTR_5_PRIME	1,758	1.426%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	20,045	14.066%
EXON	11,853	8.317%
INTERGENIC	12,442	8.731%
INTRON	74,009	51.933%
SPLICE_SITE_ACCEPTOR	18	0.013%
SPLICE_SITE_DONOR	18	0.013%
SPLICE_SITE_REGION	1,251	0.878%
UPSTREAM	15,037	10.552%
UTR_3_PRIME	6,008	4.216%
UTR_5_PRIME	1,827	1.282%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	19,075	14.514%
EXON	11,876	9.037%
INTERGENIC	12,121	9.223%
INTRON	65,090	49.527%
SPLICE_SITE_ACCEPTOR	23	0.018%
SPLICE_SITE_DONOR	23	0.018%
SPLICE_SITE_REGION	1,232	0.937%
UPSTREAM	14,825	11.28%
UTR_3_PRIME	5,425	4.128%
UTR_5_PRIME	1,732	1.318%

Type (alphabetical order)	Count	Percent
DOWNSTREAM	15,556	14.355%
EXON	11,429	10.547%
INTERGENIC	8,254	7.617%
INTRON	53,928	49.766%
SPLICE_SITE_ACCEPTOR	22	0.02%
SPLICE_SITE_DONOR	18	0.017%
SPLICE_SITE_REGION	1,172	1.082%
UPSTREAM	11,382	10.504%
UTR_3_PRIME	4,986	4.601%
UTR_5_PRIME	1,616	1.491%

Figure 2: Comparison between the tumor and control groups variants. We can see the regions in which the variants were annotated, with the "Intron_variant" having the highest percentage, indicating that the majority of variants were in fact present in intronic regions. The order of the tables is respectively: Top-left → Control_BCF; Top-right → Control_GATK; Bottom-left → Tumor_GATK; Bottom-right → Tumor_BCF

By integrating the data with *SnpEff* and *SnpSift* tools, a list of clinically significant genes were extracted. While the enrichment analysis did not provide any significant results, an inference of the tumor type can be made since *BRCA1* gene, among all the genes, have a direct correlation with breast and ovarian cancer in women [1].

Furthermore, from the ancestry analysis, we observed that the data were associated with European ethnicity (Figure 3).

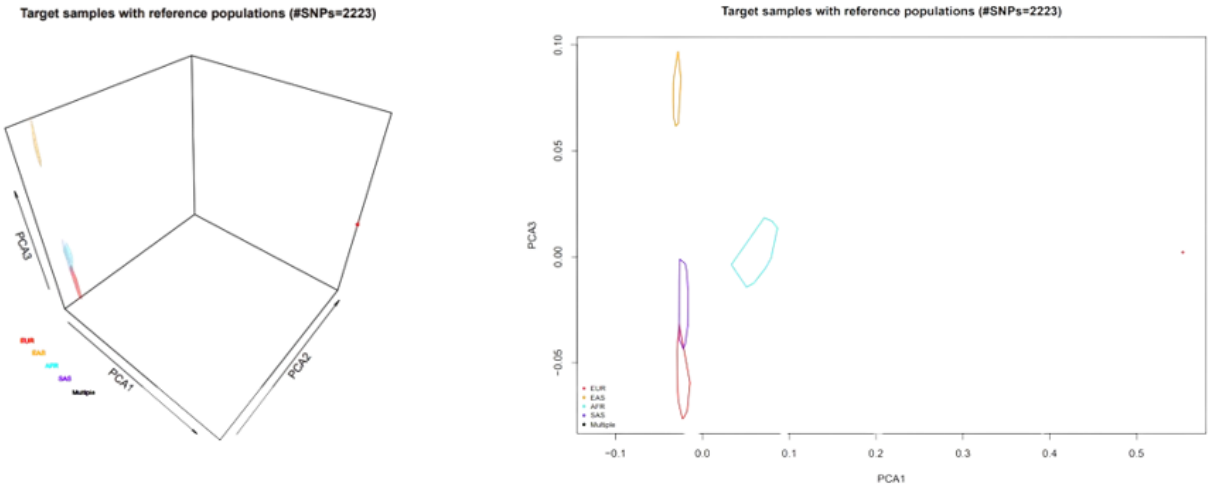

Figure 3: Results from the ancestry analysis obtained through the EthSEQ software, showing a higher level of correlation with European ethnicity.

According to the plot 4 obtained from purity and ploidy estimation, the tumor sample has an increased ploidy (2.44), indicating a slight overall gain in genetic material. Additionally, the low admixture (0.37%) suggests that there is a significant amount of normal cells in the sample, thereby possibly altering the results. Lastly, the clusters in the plot itself show that there are some segments with no significant CN changes (logR = 0), clusters with CN gains (logR >0), and segments with CN losses (logR <0).

4 Troubleshooting

When variant calling was performed, an issue was encountered when comparing the *vcf* files with control and tumor data. This error was due to the different content of the files in terms of genomic positions. In order to solve this problem, R and Python scripts (uploaded in the GitHub) were developed to filter out the sites that differed in the two files.

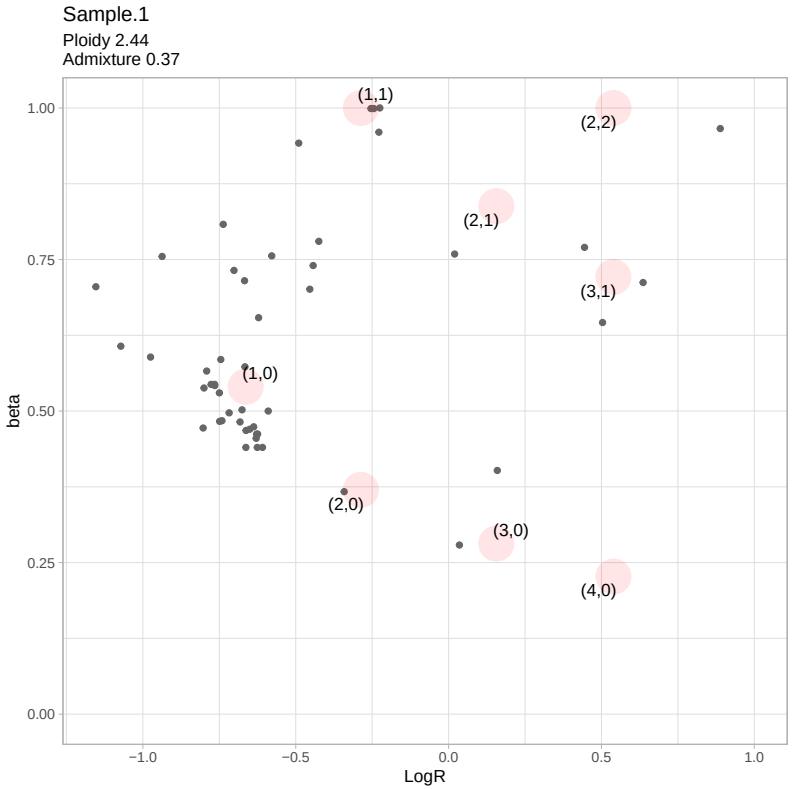

Figure 4: plot showing the relationship between Log2R (x-axis) and β (y-axis) values. The x-axis represents the Log2 ratio of tumor versus control read counts. The y-axis represents the β-AF, ranging from 0.0 to 1.0. Ploidy is indicated at 2.44 and admixture at 0.37. Data points are distributed across the plot, with notable clusters at various LogR and β values.

References

[1] D Ford, D F Easton, and J Peto. Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *Am. J. Hum. Genet.*, 57(6):1457–1462, December 1995.