UNIVERSITÀ
DI TRENTO

Analysis of partial genomic data from tumor tissue and matched normal sample

Elisa Bugani, Stefano Cretti, Maurizio Gilioli

INTRODUCTION

In the context of this project, part of the **Computational Human Genomics course**, we analyzed part of the genomic data (portions of chromosomes from 15 to 18) obtained through NGS of two samples collected from the same patient, a tumor tissue sample and a matched normal sample. Throughout this report we will use the term Sample when an operation is performed on both tumor and control sample; we will instead use Control and Tumor when the operation is specific to either of them. Flowcharts of the analysis' steps and commands are provided as attachments.

METHODS AND RESULTS

1. Determine the number of properly paired reads in both Tumor and Normal BAM files

As a first step in our analysis pipeline, the two unsorted BAM files containing the aligned reads were sorted based on alignment position, and then indexed. After that, we checked the number of properly paired reads using **samtools**[1]. In the Control file we observed the presence of 19′720′171 reads in total, with 19′576′046 of them properly paired (99.33% of the total). Instead, in the Tumor file we observed the presence of 15′039′503 reads in total, with 14′979′936 of them properly paired (99.67% of the total). Overall, a high percentage of the reads were properly paired.

2. Realign and recalibrate both BAM files

The **realignment and recalibration** processes were then sequentially performed on both the samples. Using the **GATK**[2] modules RealignerTargetCreator and IndelRealigner, respectively to identify realignment targets and to perform the actual realignment step, we realigned both BAM files. New BAM files were obtained, in which the new CIGAR scores of the realigned reads substituted the original ones, kept with 'OC' tags. Using the OC tags we determined the number of realigned reads, amounting to 3′158 realigned reads in the Control file and 2′267 in the Tumor one.

2.1. Determine the number of reads realigned in both BAM files

The BaseRecalibrator and PrintReads modules from GATK were then used on the realigned files to obtain the recalibration tables and the recalibrated BAM files (in which the original qualities are retained using the 'OQ' tag), taking into account a set of known variable sites (from Hapmap[3]).

The BaseRecalibrator module was then run again to produce an after model using the realigned BAM files and the initial recalibration tables, generating new recalibration tables. Then, using the AnalyzeCovariates module we produced before/after plots from which we expected the post-recalibration quality scores to fit the empirically-derived quality scores very well, since systematic biases should have been removed (see figures 1 and 2).

2.2. Generate and comment before-after plots (recalibration workflow)

As a result, we observed that 12′900′694 reads were recalibrated in the Control sample and 9′583′597 reads were recalibrated in the Tumor sample. The plots (figures 1 and 2) show that no major change was produced after recalibration; this is probably due to the fact that the original BAM files already had a good distribution of qualities. Nevertheless, by looking at the distribution of the data of the after model, it can be noticed how the recalibration process eliminated the remaining part of systematic bias, since the reported quality scores match properly the empirical quality scores. After having recalibrated the data, we performed deduplication of our samples (**PICARD**[4]), although it was not strictly required.

3. Identify all heterozygous SNPs in the patient using GATK

Using the GATK module UnifiedGenotyper, we obtained the VCF file containing all the **germline variants** (obtained comparing the Control to the reference sequence). The variants were filtered using vcftools, specifying a minimum quality of 20 and a range from 5 to 200 for the mean depth of coverage. We also removed the indels and focused on the SNPs. After filtering, 8′742 sites were kept out of 9′929 starting sites.

3.1. Which identified SNPs are in the list clinvar_Pathogenic.vcf

We then proceeded in annotating the SNPs; firstly using **SnpEff**[5] and the database hg19kg, then using **SnpSift**[5] and a list of pathogenic variants obtained from Clinvar[6]. We found the presence of a single SNP

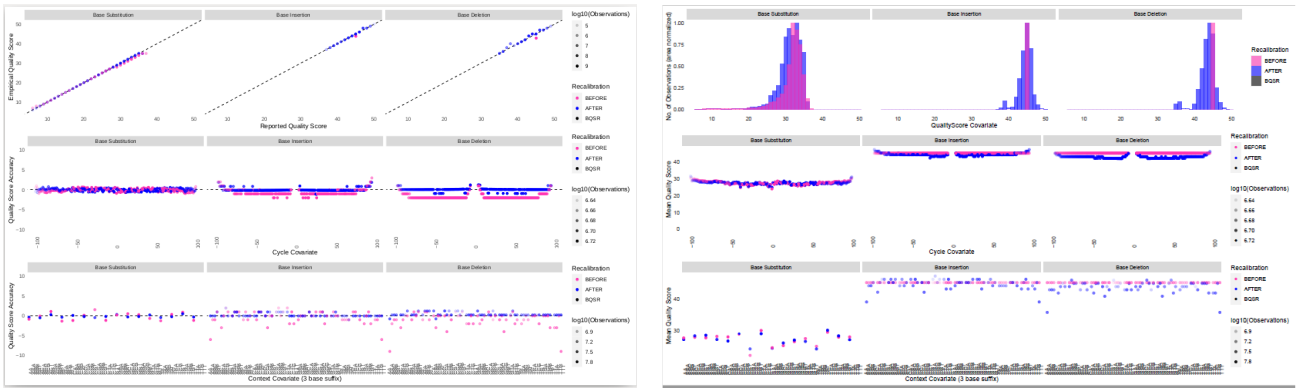


Figure 1. Results obtained after recalibration of the Control sample

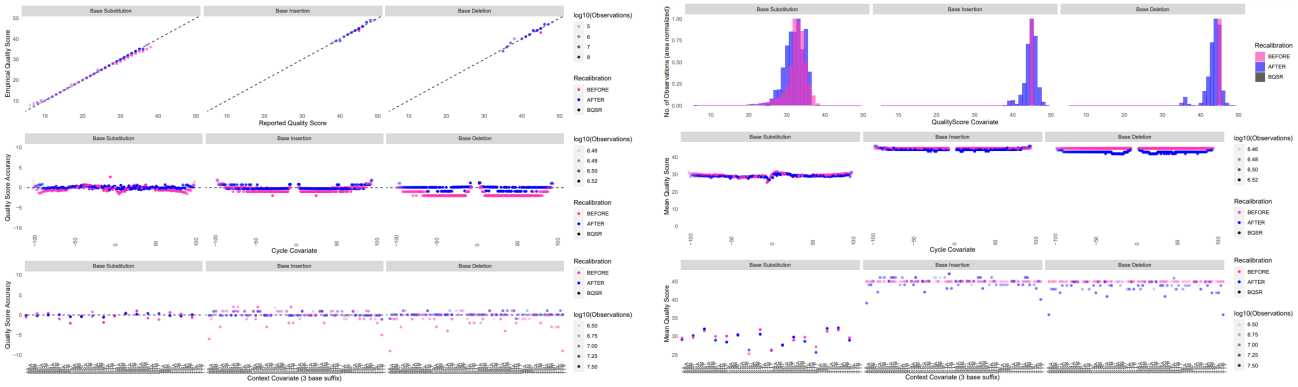


Figure 2. Results obtained after recalibration of the Tumor sample

belonging to the Clinvar pathogenic list, with variation **ID: 54108**, namely an alteration of the **BRCA-1 gene**, interpreted in several studies as pathogenic. In particular, it is linked with breast-ovarian cancers. The SNP position is chr17:43094477; here, a cytosine is substituted by an adenine or a guanine (the latter being our case), resulting in the acquisition of a premature stop codon by BRCA-1 (further information can be obtained at Clinvar link and dbSNP link).

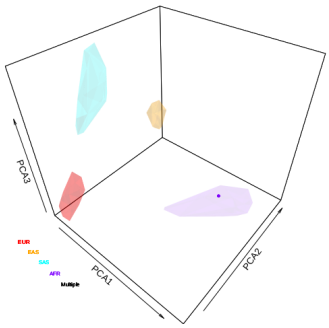


Figure 3. EthSEQ PCA plot

*4. (4.1, 4.2) **Determine the ancestry of the patient***

To determine the ancestry of the patient, we used the function ethseq.Analysis contained in the **EthSEQ**[7] R package. EthSEQ is a PCA-based tool for ancestry analysis which integrates the target model with a reference model, containing data of known populations, namely Africa, Europe, South Asia and East Asia. The output of EthSEQ is a standard PCA 3D plot, produced using the first three components, in which the colored polygons represent clusters of reference samples belonging to specific geographic areas. As we could notice both from the various graphs and from the report output file in TXT format, our target sample is located inside the polygon representing Africa, leading us to conclude that the patient is likely to be African.

4.3. Explore and comment EthSEQ output

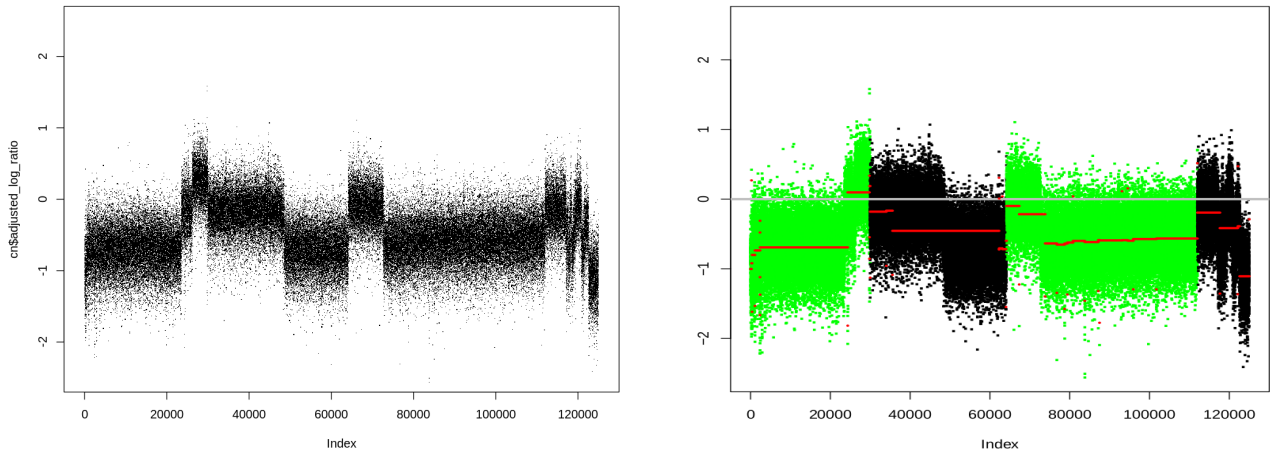
*5. **Identify somatic copy numbers using Varscan workflow***

We used **Varscan**[8,9] to identify somatic copy number variations. Firstly, the pileups for both Tumor and Control samples were computed using samtools; then those files were analyzed with Varscan copynumber function, which generates a file containing a pre-segmentation. The default parameters of the function are an α value of 0.01, a minimum coverage of 10 and a minimum quality of 15. On the obtained output file, the Varscan copyCaller module was used in order to apply some thresholds and adjustments. In particular, it allows to filter copy number calls by minimum coverage and/or region size, it adjusts raw copy

number (log2) values for GC content, classifies each region as amplification (gain), deletion (loss), or neutral based on your preferred log2 thresholds, recenters raw copy number data if neutral segments are not on the $log_2 = 0$ axis[10].

We used the provided **CBS.R** script, which includes functions from the **DNAcopy R library**[11], to plot the raw ratios and the adjusted ones and to perform the segmentation.

5.1. Determine the number of segments compatible with heterozygous deletions and explain how/why you selected that segments



(a) Adjusted raw ratio **(b)** Segmentation

Figure 4. DNAcopy plots

Looking at the graph and the SEG output file, we notice 66 segments with different copy numbers in the two samples. A hemizygous deletion would be represented by a segment with an average log2 ratio belonging to the interval [-0.5, -1.5], meaning an amount of Tumor DNA which is roughly half of the Control sample; we therefore have 37 regions compatible with hemizygous deletion.

To identify somatic point mutations we used Varscan's somatic module which scans the pileup files of the Control and Tumor samples simultaneously. At each position filtered with default parameters of coverage (8× for normal, 6× for tumor), the program analyzes the differences between the Control sample, the Tumor sample and the reference sequence. We obtained 14250 point mutations of which 168 somatic.

6. Identify somatic point mutations using Varscan workflow

We then used **bedtools**[12] to intersect the list of DNA repair genes, the list of heterozygous SNPs of our sample that are annotated as pathogenic in Clinvar and the list of segments with somatic hemizygous deletion. We obtain that 19 DNA repair genes show some form of somatic hemizygous deletion, while only one also presents a pathogenic SNP according to Clinvar; that gene is the aforementioned BRCA-1. Considering that BRCA-1 is a tumor suppressor gene, a major role in the oncogenesis could have been played by the loss of function of this gene, which happened according to the two-hit model (the first due to the premature stop codon SNP, the second due to the complete deletion of the other allele).

7. Determine which DNA repair genes overlap both heterozygous deletions and heterozygous SNPs of the patient that are in Clinvar

Using bedtools we then intersected the list of DNA repair genes with the list of somatic point mutations (obtained by filtering the list of point mutations obtained via Varscan); only two genes come up, namely FANCI and TP53. Intersecting then with the list of segments with somatic hemizygous deletion we did not get any result, meaning that no DNA repair gene has both a somatic point mutation and a somatic hemizygous deletion.

8. Determine which DNA repair genes overlap both heterozygous deletions and somatic point mutations of the patient

Purity and ploidy estimation was performed at first using **CLONET**[13], a tool that exploits the distribution of the allelic fraction of heterozygous SNPs to compute β values, representing the proportion of neutral reads, to perform the estimations.

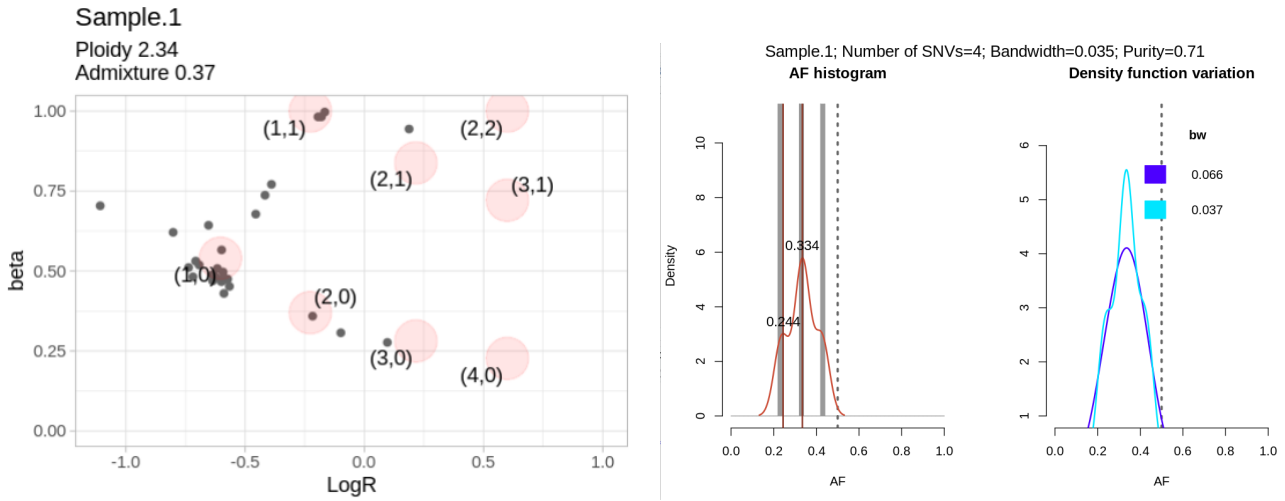
By using the ASEReadCounter module of GATK, a table containing the allele counts of only heterozygous SNPs, germline related, is obtained in CSV format. A minimum depth of 20, minimum mapping quality and base quality of 20 were used. For each heterozygous position, the reference and alternative allele are reported,

9. Determine purity of the Tumor sample using both CLONET and TPES tools

their counts, the quality values and other information. The same procedure was repeated for the Tumor sample, taking into account the heterozygous sites of the Normal sample.

The obtained table was then passed to CLONET.R script containing functions related to the CLONETv2 R package. Using them, we found our sample to have a ploidy of 2.34, a purity of about 0.63 (obtained admixture value = 0.37).

Through the check_ploidy_and_admixture function, we obtained the following plot (figure 5a), where each point represents one of the segments in the table obtained by using the compute_beta_table function. The large red circles represent expected (beta, LogR) values corresponding to the estimated ploidy and DNA admixture[13]. Some points are present inside the (1,1) cluster, representing copy number neutral segments; instead, several dots are clustered in the (1,0) region, indicating the presence of several segments which underwent heterozygous deletions. The plot supports also the presence of at least one event of copy number neutral Loss Of Heterozygosity (LOH). Several points are reported to be in non-circled regions,indicating subclonal events. Only a few gains were observed, one of which is coupled with LOH.



(a) Purity assesment with CLONET

(b) Purity estimation with TPES

Figure 5. Two different values of purity were obtained: 0.63 using CLONET, 0.71 using TPES

Instead of using somatic copy number alterations to estimate our purity, we now wanted to make use of **TPES**[14,15], which calculates tumor purity analyzing the allelic fraction distributions of the SNVs. Using the TPES R package we estimated a value of purity of 0.71, included in the range [0.71, 0.81].

Looking at the graph 5b on the left, the distribution of the allelic fractions is shifted; two peaks are observed, respectively with an AF value of 0.244 and 0.334. The rightmost peak, closests to the center of the graph, is used to estimate the level of admixture, since it represents the event with the highest clonality. The leftmost peak, closer to zero, represents instead a subclonal event.

The purities computed using CLONETv2 and TPES were a little different, with CLONET estimating a lower range of values. CLONET estimated a purity of 0.63, while TPES estimated a purity of 0.71. Overall the two values of estimated purity are quite similar even if not the same; this type of result is not uncommon. Moreover, it should be considered that the number of SNVs used by TPES is quite low (4 SNVs) when compared to the minimal number of SNVs for which the tool provides good correlation with other tools, and therefore reliable purity estimates.[14].

9.1. Comment results

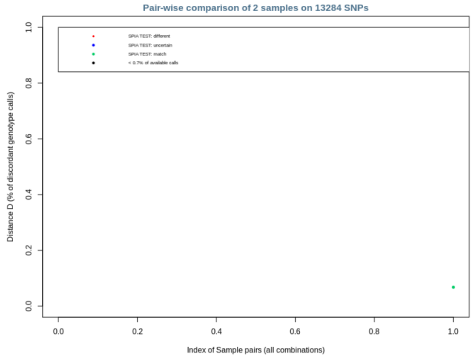


Figure 6. SPIA result

10. Determine similarity of Tumor and Normal samples using SPIA R package

As a last step in our analysis, using GATK ASEReadCounter, we obtained the allelic counts for each site included in the hapmap_3.3.b37.vcf file for both the Tumor sample and the

Normal one. Then, for each position, the AF was computed; the
latter was used to estimate the genotype. After the filtering of all the sites shared by the two samples, SPIA.R was run. The SPIA package, which stands for SNP Panel Identification Assay (**SPIA**[16,17]) is a package that enables an accurate determination of cell line identity from the genotype of single nucleotide polymorphisms, allowing to discern when two cell lines are close enough to be called similar and when they are not.

The graph presented in figure 6 shows the output obtained using the SPIAPlot function (green dot).

The two samples were estimated to be matching when confronting them through the variable sites reported in Hapmap. According to the output, we obtained that the two samples have a distance very close to zero, but not exactly zero, specifically amounting to 0.067856, compatible with the fact that they are obtained from the same individual but they share some differences due to the presence of cancer-related variants.

CONCLUSION

The results of our analysis are rather limited since we analyzed only a small portion of the whole genomic data, hence a larger scale analysis of the same dataset may produce different results. Nevertheless, the steps presented in this project represent an example of a simple pipeline which includes some crucial and common steps for the analysis of genomic data.

References

1. Samtools. http://www.htslib.org/.
2. GATK. https://gatk.broadinstitute.org/hc/en-us.
3. HapMap 3 - Wellcome Sanger Institute. https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html.
4. Picard Tools - By Broad Institute. https://broadinstitute.github.io/picard/.
5. SnpEff and SnpSift. https://pcingola.github.io/SnpEff/.
6. ClinVar. https://www.ncbi.nlm.nih.gov/clinvar/.
7. Romanel, A. & Dalfovo, D. EthSEQ: Ethnicity Annotation from Whole Exome Sequencing Data (2021).
8. VarScan - Somatic Mutation Calling. http://varscan.sourceforge.net/somatic-calling.html.
9. VarScan - Variant Detection in Massively Parallel Sequencing Data - Calling Copy Number Variants from Exome Data. http://varscan.sourceforge.net/copy-number-calling.html.
10. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576, DOI: 10.1101/gr.129684.111 (2012).
11. Seshan, V. E. & Olshen, A. DNAcopy: DNA copy number data analysis. Bioconductor version: Release (3.15), DOI: 10.18129/B9.bioc.DNAcopy (2022).
12. Bedtools: A powerful toolset for genome arithmetic — bedtools 2.30.0 documentation. https://bedtools.readthedocs.io/en/latest/.
13. Prandi, D. & Demichelis, F. Ploidy and purity adjusted DNA allele specific analysis using CLONETv2. *Curr. protocols bioinformatics* **67**, e81, DOI: 10.1002/cpbi.81 (2019).
14. Locallo, A., Prandi, D., Fedrizzi, T. & Demichelis, F. TPES: Tumor purity estimation from SNVs. *Bioinforma. (Oxford, England)* **35**, 4433–4435, DOI: 10.1093/bioinformatics/btz406 (2019).
15. Locallo, A., Prandi, D. & Demichelis, F. TPES: Tumor Purity Estimation using SNVs (2019).
16. Demichelis, F. *et al.* SNP panel identification assay (SPIA): A genetic-based assay for the identification of cell lines. *Nucleic Acids Res.* **36**, 2446–2456, DOI: 10.1093/nar/gkn089 (2008).
17. Adi Laurentiu Tarca, Purvesh Kathri & Sorin Draghici. SPIA: Signaling Pathway Impact Analysis (SPIA) using combined evidence of pathway over-representation and unusual signaling perturbations. Bioconductor version: Release (3.15), DOI: 10.18129/B9.bioc.SPIA (2022).