CHG Project

Baldinelli Sara, De Pietri Letizia, Faggin Gaia, Spadazzi Roan

Rationale

The assessment and identification of somatic aberrations and variations in the genome of an oncologic patient is a fundamental step in characterizing whether such mutations influence tumor events. This project is focused on analyzing and comparing Control versus Tumor DNA sequences from the same individual while trying to reach various conclusions in terms of mutation clinical relevance, with a particular focus on somatic ones. After performing quality checks and recalibrating our BAM files, we worked towards describing and outlining somatic events such as SNVs, Indels, and Copy Number Alterations and annotating them to give us insights into their function and impact. Moreover, we performed an ancestry analysis and additional steps to account for tumor purity and ploidy.

Computational workflow

In the first step of the analysis we utilized samtools to sort and index the two BAM files and to check the mapping and pairing of the reads. In **Table 1** some statistics regarding the reads are shown.

	Total	Mapping to forward strand	Mapping to reverse strand	Paired reads	Mapped reads	Mapped and paired
Control	19720171	9864318	9855853	19708438 (99,94%)	19658191 (99,68%)	19613806 (99,46%)
Tumor	15039503	7521200	7518303	15029250 (99,93%)	15023437 (99,89%)	15019614 (99,87%)

Table 1. Summary of reads statistics for both Control and Tumor samples.

Throughout our analysis, we used the *human_g1k_v37.fasta* as a reference and, when useful, a vast genomic region for the annotation (*Captured_Regions.bed*).

After that, we performed an indel-based realignment through the GenomeAnalysisTK tool: first we identified the regions to realign exploiting the RealignerTargetCreator option; then, we performed the actual realignment using IndelRealigner.

The deduplication step follows, using the MarkDuplicates pipeline of the Picard tool and indexing with samtools the output files. The last quality-related procedure was recalibrating the BAM files with GenomeAnalysisTK, firstly using BaseRecalibrator, then PrintReads with the -BQSR to also generate a .table that was an input of a second recalibration run in order to have a direct comparison before and after recalibration using the AnalyzeCovariates command; plots with various statistics were outputted in .pdf format.

After getting the fully processed BAMs, the following steps were dedicated to generating variant call files. First of all, using UnifiedGenotyper (from GenomeAnalysisTK), we produced Control and Tumor GATK vcf files, which then were filtered with vcftools using options --minQ 20 --max-meanDP 200 and --min-meanDP 5. The annotation of the called filtered variants followed using snpEff, then SnpSift twice consecutively, firstly with hapmap_3.3.b37, then with clinvar_Pathogenic to annotate clinically relevant variants.

Next, the pileup of reads at single genomic positions were generated using the samtools mpileup, and the minimum mapping quality for an alignment to be used was set to 1. VarScan2 was utilized to report germline, somatic, and LOH events (with –min-var-freq of 0.2 and –min-freq-for-hom of 0.8) and somatic variant filtering was performed using the SnpSift tool. Using the snpEff, we conducted somatic variant annotation. Following the same procedure previously described for somatic variants, we proceeded with somatic copy number calling through the VarScan2 tool (copynumber) applied to the output of samtools mpileup. The SCNA captured regions output was given to copyCaller to output CN calls.
Successively, to perform purity and ploidy estimation, we first gave in input to bcftools call the results of bcftools mpileup (with depth annotated) getting new vcf files that were later filtered for heterozygous SNP sites only. The ASEReadCounter option from the GenomeAnalysisTK tool was applied to the two BAM files to obtain allele counts and, lastly, CLONETv2 and TPES R packages were exploited to assess purity and ploidy values..
Finally, we run ancestry analysis using RunEthSEQ.R with the specified SS2.Light.Model.gds model on both Control and Tumor samples.

Results and discussion

Variants analysis

Point mutations: out of the more than 14.000 SNPs, we report that 0.13% of them have high impact and the majority (81.9%) are classified as modifiers. Moreover, 44.0% of all SNPs are missense variants. Among all point mutations, 41.1% of them fall into intronic regions, likely a consequence of the fact that introns span bigger regions and are less conserved, and exonic regions come in second (18.8%). Nucleotide transitions (A-G, C-T) are 2.5 times more common than transversions due to the substitution involving bases of similar chemical structure. The annotation step gave us insights into some of these variants, particularly a high impact one present on chromosome 17 (at position 41246494) substituting a C into an A. Clinically, this mutation has been associated with hereditary breast and ovarian cancer syndrome (or a predisposition to developing it). It is interesting to note that a direct comparison of this point mutation between the two samples shows a significant increase in the amount of reads mapping to the modified nucleotide in the Tumor one (Control: 40% A / 60% C; Tumor: 82% A / 18% C) with similar read counts. Despite this dissimilarity, this mutation has been classified as germline due to the lack of a significant allele frequency difference.

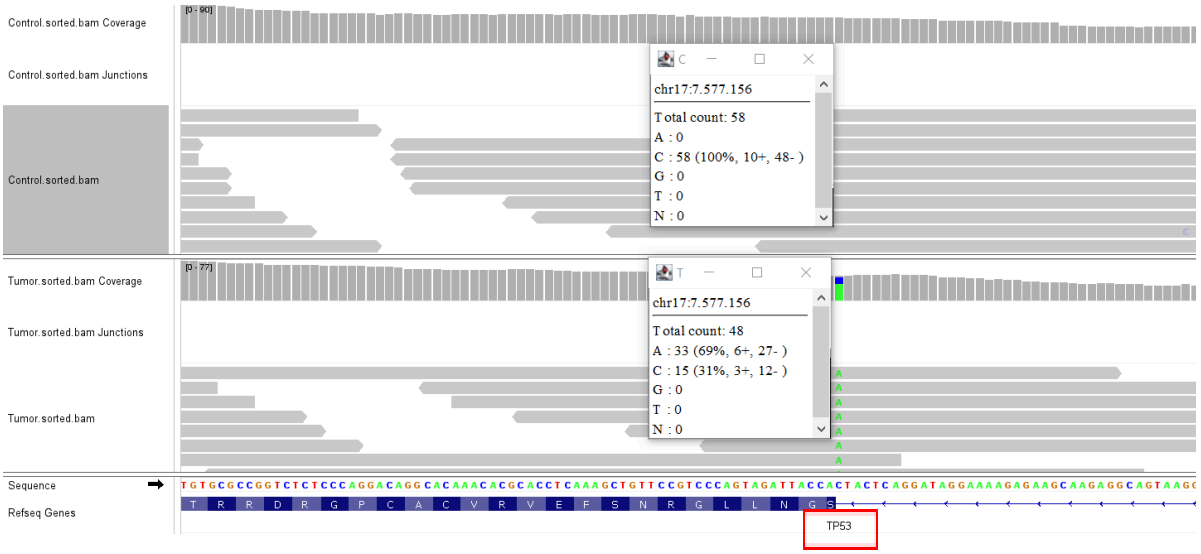

Figure 1. IGV snapshot showing nucleotide count differences at position chr17:7577156 between Tumor (below) and Control (above) of a somatic point mutation. Note that coverage values are comparable in the two conditions. The represented region belongs to the TP53 gene.

Regarding point mutations reported as somatic, we identified two having high impact; one of them is present on chromosome 17 (at position 7577156) between an exon of TP53 (whose number is transcript-dependent) and the adjacent intronic region, with differences in nucleotide frequency (Control: 100% C; Tumor: 69% A / 31% C) with comparable read counts (**Figure 1**). The relevance of this gene and its well known contributions to cancer development (including breast cancer) are important to note when considering this variant. Finally, we report the presence of 8 high impact loss of heterozygosity mutations.

Insertions and deletions: 1452 indels were reported, out of which 662 insertions and 791 deletions. Similarly to point mutations, they are mainly found in intronic regions (49.0%), but much fewer are found in exonic ones (4.3%). Out of all indels, 1.4% have high impact (proportionally more than point mutations due to them having higher chance of being deleterious) and the majority are classified as modifiers (94.6%). Analyzing the ones with high impact, 3 of them are reported as somatic (all in exonic regions when visualized with IGV) and 5 as loss of heterozygosity.

Copy Number Variations: after generating the segmentation plots for both collections of genomic regions for further analysis (using both .bed files), we get plots representing the $\log_2$R ratio between Control and Tumor copy numbers for each specified region of interest. Comparing a sample with itself would center this distribution around 0: any deviation from it can suggest copy number alterations. In our case we see a vast prevalence of the segments of interest displaying potential heterozygous deletions ($\log_2 R = \log_2(½) = -1$), in other words the loss of one of the two copies of an allele containing one or more informative SNPs, and other deletion-like segments around $\log_2 R \sim -0.5$. These could probably be a consequence of subclonal events, where the deletion is not found in all reads, therefore centered around values with a lower copy number ratio. The purity of the sample itself may have played a role in generating noise or affecting those numbers therefore these hypotheses are addressed with the next step of purity and ploidy estimation.

Purity & ploidy estimation

Purity and ploidy estimation is an important step in analyzing sequencing data; indeed, both the quantification of the variant allele fraction (VAF) of SNVs and the assessment of somatic copy number aberrations are influenced by the DNA admixture and by the ploidy: the more normal cells are present, the more diluted these estimates will be, while a tumor polyploidy will have the opposite effect on VAF and copy number aberrations values.

The plot in **Figure 2a** provides us with a space for copy number analysis: the logR signal is shown on the x axis, while the Beta value is depicted on the y axis. It can be observed that the majority of segments cluster around a logR of -0.5 and a Beta value of 0.5, indicating a hemizygous deletion leading to a proportion of neutral reads of 50%. We can identify another cluster that possesses a copy for each allele (logR=0, Beta=1) and it can be categorized as wild type. Between these two clusters, a group of segments is observed: they constitute subclonal LOH events. We can also detect the presence of a segment reporting 2 copies for allele A and no copies for the other; this is a CN-LOH event in which one allele is lost and the other is duplicated, the logR is indeed still around 0. Note that some noise is present in the plot as not all the outliers can be considered as subclonal events.

The purity estimated by CLONET equals 0.65, while the one computed by TPES is 0.71. Despite the fact that the TPES measurement is based on 4 SNVs only, the obtained value approaches the one obtained using CLONET. These estimates are consistent with the results shown in **Figure 2b**; the plot shows two VAF peaks at 0.334 and 0.244 that correspond, respectively, to a clonal peak which does not coincide to around the 0.5 value due to the lower purity of the sample, and to a possible subclonal peak of lower density. Note that, when applying the smoothing procedure (Density function variation plot) the subclonal peak disappears; indeed, in the AF histogram we do not observe an enhanced discontinuity between the two peaks. This indicates that the subclonal event might have occurred immediately after the clonal one.

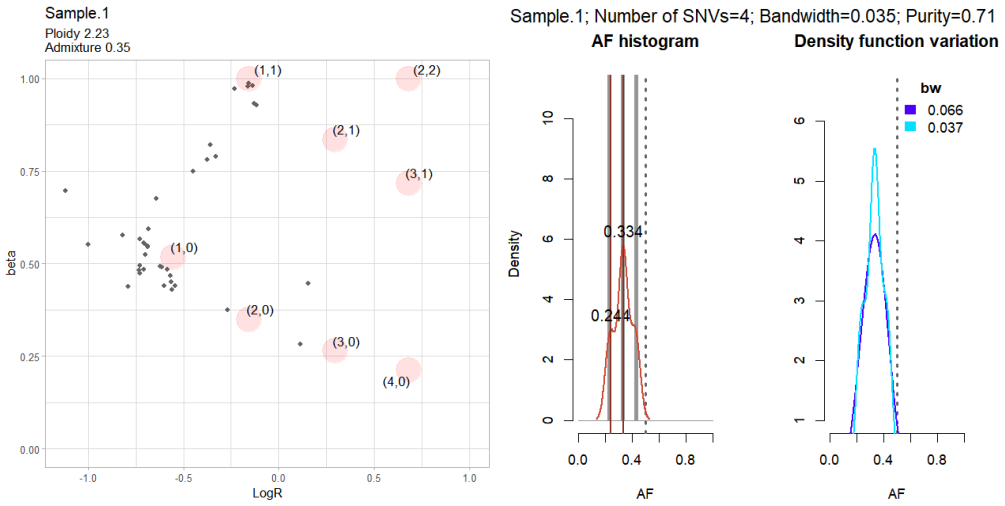

Figure 2. (a) CLONET plot, the sample shows a ploidy of 2.23 and an admixture of 0.35; segments are distributed in a space defined by logR on the x axis and Beta value on the y axis. **(b)** TPES plot, variant allele fractions (x axis) with corresponding densities (y axis) are represented. Left: AF histogram showing both a clonal and a subclonal peak. Right: VAF distribution smoothing by kernel density estimation.

Ancestry analysis

Ancestry analysis was performed in order to infer the ethnicity of the patient: EthSEQ is capable of comparing BAM files against models deriving from whole exome sequencing data outputting a 3D-PCA plot (**Figure 3**). From the image, it can be inferred that both samples are inside the African cluster (confirmed also by visualizing the 2D plots), with a slight difference between them likely caused by the previously described variations between Control and Tumor.

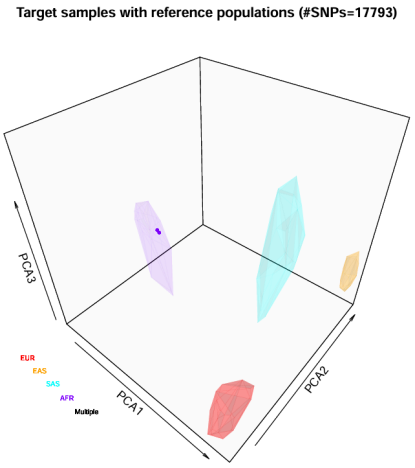

Figure 3. EthSEQ 3D-PCA-plot, both samples are clearly within the African cluster (purple).

Conclusions

In conclusion, this investigation led to the identification of some potential causative variants, including a somatic point mutation found in the TP53 gene, 3 somatic indels and copy number variation events. VAF estimates and copy number aberrations values are compatible with (and can be explained by) the purity estimate of our sample. Moreover, after an ancestry analysis, it could be assessed that the individual belongs to the African cluster.