CHG project

Filippo A. Mirolo, Vittoria Ossanna, Alessandro Pilli

Rationale

The present study takes in account two BAM files (tab-delimited text file that contains sequence alignment data) and tries to evaluate the differences at the genome level. The two files are representation of sequencing alignment of the chromosomes 15, 16, 17 and 18, of a patient defined as control and one in a tumorous state.
The comparison display of the processed files after data curation, described in the "Computational workflow", led to the evaluation of some chromosomal aberrations (SNVs and CNVs) in the region regarding also some DNA repair genes. This probably has driven the cells into the cancerous state.

Computational workflow

In order to perform the analysis described above, the initial files went through a process of realignment, recalibration and deduplication.
For each of the two initial BAM files - Tumor and Control - we initially performed sorting and indexing through `samtools`. Next, we used `GenomeAnalysisTK` tools for realigning operations with the `RealignerTargetCreator` and `IndelRealigner` options separately. This process has been done by using "`human_g1k_v37.fasta`" as reference and "`Captured_Regions.bed`" to limit the intervals of the chromosomes selected in the study. Next step for data recalibration is using the suite `GenomeAnalysisTK`, in particular using `PrintRreads` and `BaseRecalibrator` processes. The same genome of reference has been used. The third step we performed is deduplication, which has been conducted using `picard` suite. At this point we have the fully processed files (realigned, recalibrated and deduplicated) ready for the next analysis. From now on, we will refer to these files as "processed BAM files".
After the clearance of the data we performed the variant calling of the somatic point mutations using the `VARSCAN` tool. To annotate those spots, we created first the `pileup files` of both the control and tumor BAM files processed as described in the step above. We continued the analysis filtering the vcf file, produced by `VARSCAN`, with `vcftools` and the following parameters: `–maxDP 200` and `–minDp 8`.
Copy number was also taken into account, to study it we used again the `VARSCAN` tool and `human_g1k_v37.fasta` as reference. The results are shown with a segmentation plot performed with an R script. Major modifications have been seen and validated through `IGV`.
To give an appropriate interpretation of some results we had to estimate the purity of the tumor sample, to exploit it we used `GenomeAnalysisTK`, `ASEReadsCounter` and an R script that runs `CLONET` and `TPES`, two alternative tools that assess purity and ploidy based either on copy number status of genes or SNVs.
Lastly we decided also to perform an ancestry analysis for completeness through the tool `EthSEQ` (v3.0.2). The parameters used for this analysis were: `SS2.Light.Model.gds` as the reference model and the two file resulting from the `VARSCAN` analysis (`Control.VARSCAN.vcf` and `Tumor.VARSCAN.vcf`).

For any further details on parameters of the tool used and for the R scripts cited before, please refer to this Drive folder in which we provide detailed information.

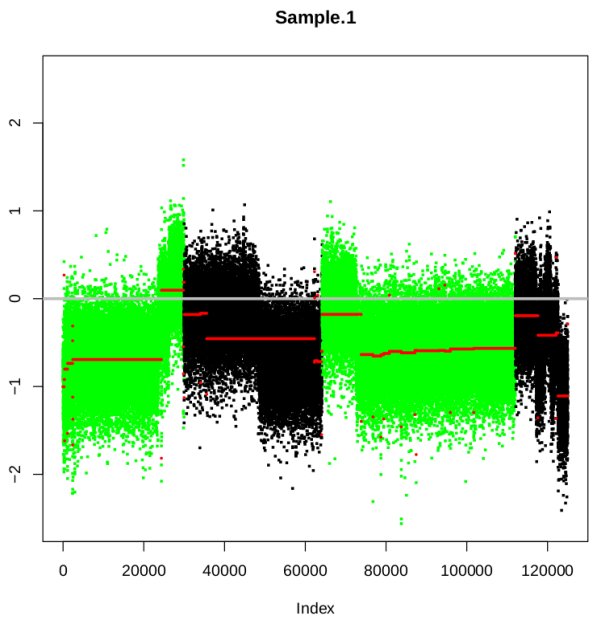
Results

SNV analysis

From the analysis of the SNVs, we report that for the variants found in the tumor against the control we have: 0.12% high impact mutations, 10.27% low impact mutations, 6.78% moderate impact mutations, 82.82% modifier mutations.
From this tool we get that the majority of mutations are modifier mutations, but this annotation actually corresponds to an unknown effect on the protein. Next relevant information from this analysis are that most of the SNVs detected correspond to silent (54.667%) alteration, followed by missense variations (44.981%) and just a few nonsense alterations (0.352%). Lastly, the mutations detected from this analysis underline that these events are mainly associated with intronic variants (40%), next most common regions are exon (18%) and downstream regions (17%).
To sum up, from the analysis concerning SNVs, we can conclude that most of the variations detected have an unknown effect of the protein encoded in the genome. The majority of the alterations are related to intronic regions, so not directly related to the functional structure of a protein but possibly to regulatory regions making an impact on the expression of the genes.

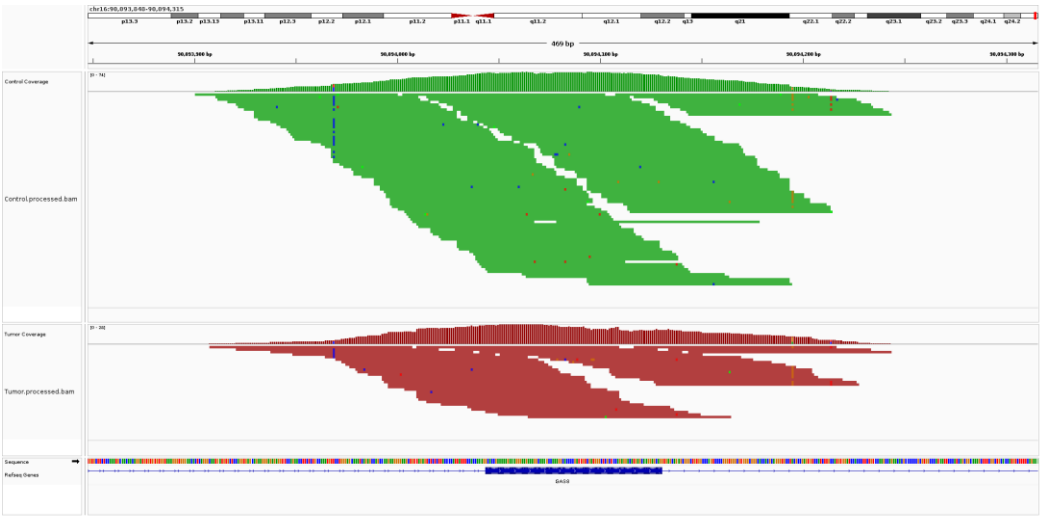
CNV analysis

In order to assess copy number variation events, we used the library "`DNAcopy`" from R. We report the plot of the segmentation of the genomic regions analyzed in this project. From a theoretical point of view, most of the chromosomal region analyzed should have a log2R (on the y axis) equal or centered in zero, otherwise it is evidence of copy number alteration of the region.



From this segmentation we can see that most of the fragments detected by this analysis correspond to deletions. The segments detected around a log2R = -1 correspond to a heterozygous deletion, hence the tumor sample completely lost one of the two alleles containing the informative SNP. In addition, we can see that some of the segments detected lie around a log2R around -0.5. This could be due to a subclonal event, therefore the deletion is not present in all the reads coming from the sequencing, but just a part of them. Lastly, the segments that slightly differ from zero could be interpreted as wild type segments, since - as we will discuss later - we are not dealing with a 100% pure tumor sample.
From the list of regions detected as altered in copy number, we looked for segments that comprehend genes involved in DNA repair. As expected, a lot of them present in the list are actually in an altered copy number state. Among them we report that genes such as BRCA1 and TP53 present a value that corresponds to a deletion. Also, we looked for graphical

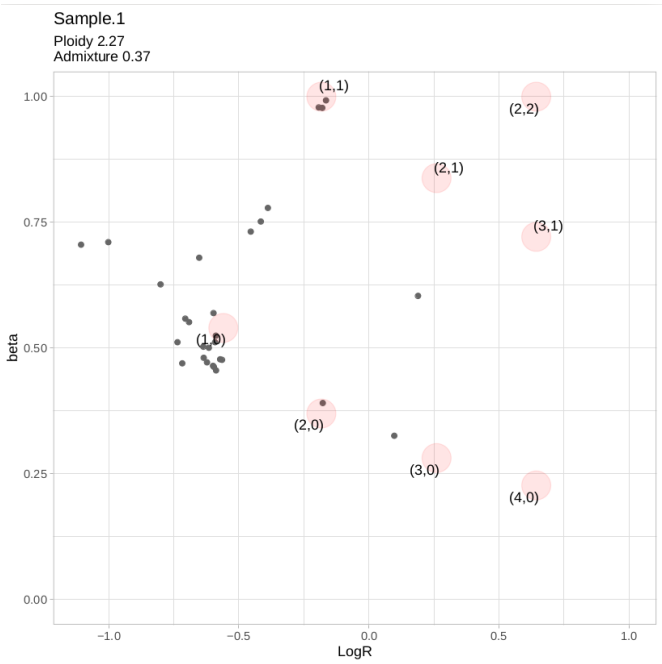
confirmation on IGV for the most altered regions. We report a screen of the gene GAS8 (Growth Arrest-Specific Protein 8) which shows an evident deletion in the tumoral sample.



Ploidy and purity estimation

One important step in the study of tumor evaluation is the evaluation of the sample quality. The presence of an excessive quantity of non cancerous cells may drive the study to unwanted results.

Through the tools `CLONET` and `TPES` we obtained the following results.

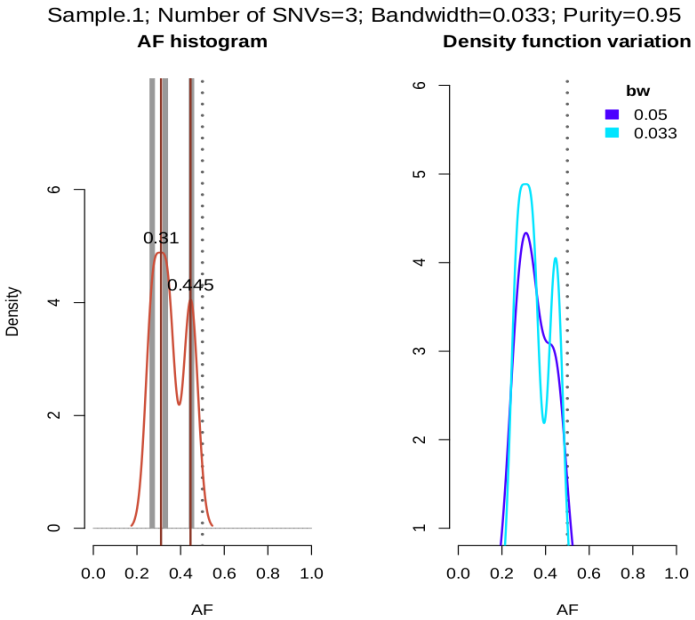


The first one is a distribution of the segments that were visualized in the copy number analysis in a space defined by the evaluated beta and a Log2R (reported to the right). It reports that the majority are part of the group (1,0) corresponding to a loss of heterozygosity (LOH), a part is condensed in (1,1) defining them as Wild-Type (WT). The presence of a segment confined in the space of (2,0) corresponds to a copy number neutral loss of heterozygosity (CNN-LOH) resulting in the loss of an allele and duplication of the remaining one. The presence of points spread too far from the clusters are outliers that we are not able to evaluate, while the three dots cluster in between LOH and WT could be interpreted as a subclonal event.

The purity estimation of `CLONET` is only 63% defining it as impure and so probably some outliers can be explained through the presence of noise.

For the `TPES` results we can observe the curves that represent the number of times a specific allelic fraction is present for each SNV (plot to the right). From the graph we can observe the

presence of two peaks; one near the 0.5 AF value, the other in the proximity of 0.3 with a higher peak. Moreover these results obtained from TPES are in contrast with the ones coming from CLONET, in fact the plot shows a bigger density in the subclonal peak (0.3 AF).
This discrepancy can be justified by a possible evolution of the tumor cells, where the old group is substituted by the new one thanks to the incorporation of new mutations that allow these cells to have a higher fitness in respect to the others. And so the three dots that we previously defined as the subclonal group might be instead representative of the ancestral clone.



It has to be mentioned that TPES is using only 3 SNVs and so these results might be biased and approximated. By changing the values in the analysis function it seemed not possible to increase the number of SNV included in the analysis. Also, the purity of this sample based on the TPES analysis is way higher (95%) but needs to be taken with a pinch of salt for the same reason just discussed.

Ancestry analysis

An important aspect in this kind of analysis is the similarity in terms of ethnicity of the samples. To evaluate them the two samples have been used as part of the analysis completed by the tool EthSEQ under the format of vcf files.
The result reported below identify the two samples as africans

```
sample.id   pop   type  contribution
Control     AFR   CLOSEST EUR(19.73%)|EAS(19.43%)|SAS(17.98%)|AFR(42.85%)
Tumor       AFR   CLOSEST EUR(21.79%)|EAS(20.65%)|SAS(18.96%)|AFR(38.59%)
```

Since both samples coming from tumor and control present a similarity in the measures of their ancestors, we can take this analysis as an indirect confirmation that both the samples have actually been taken from the same individual (the difference can be imputed to the great deletions observed before).

Conclusion

From these analyses we can conclude that the tumoral variation detected are both SNVs and CNVs. From the former aspect we get that many of the mutations have an unknown effect on the functioning of the protein, also the mutations are mostly associated with intronic regions. From the latter, we see that we obtained mostly deletions that comprise several DNA repair genes and cell growth proteins: this is a relevant result from a biological point of view because it could explain the tumoral state of the cells.
From purity and ploidy estimation we could detect the presence of a clonal and a subclonal cluster: also, this result also would reflect the structure of the segmentation of the CNVs analysis.