Computational Human Genomics Project (A.Y. 2023-24)

Project rationale

Genetic abnormalities, both inherited and somatic, are the driving force for tumor development. Through the identification and characterization of events like point mutations and chromosomal structural alterations, it's possible to gain relevant insight into the mechanisms that lead to the disease. In this project, we focused on the analysis of the genome of a single oncologic patient, from whom both a normal and a tumor sample were collected. By implementing a pipeline that exploits several computational tools, our aim is to obtain insightful statistics and to pinpoint relevant genomic aberrations.

Computational workflow

First steps involve the preprocessing of both the control and tumor corresponding BAM files, that contain information about the alignment of the genomic reads over the reference. The reference human genome used in this and all the subsequent analysis is the GRCh37 from the 1000 Genomes project.

Sorting and indexing was performed using the *samtools* utility: through these operations, the genomic files are sorted by position, and then an index file (.bai) is generated to allow for fast recovery of specific locations and subsequent processing.

Afterwards, using the *RealignerTargetCreator* and *IndelRealigner* tools from the *GATK* utility, we identified deletions or insertions that were previously missed and performed a realignment of the reads. Due to the fact that the BAM files only cover a limited region of the genome (corresponding mostly to the chromosomes 15, 16, 17 and 18), we also provided a bed file containing the regions captured during the sequencing process.

Base quality scores were adjusted through a process of recalibration, exploiting the *BaseRecalibrator* and *PrintReads* tools from *GATK* to obtain more accurate PHRED scores. To keep track of known polymorphic sites, which could be interpreted as sequencing errors, we also submitted a file containing the SNPs documented in HapMap 3.

Finally, duplicates were isolated and removed from the BAMs using the *MarkDuplicates* tool from the *Picard* suite, which uses the CIGARs contained in the files to identify duplicate reads.

The presence of copy number alterations was assessed by analyzing the coverage detected for different genomic regions. A combination of *samtools mpileup* and *VarScan2 copynumber* and *copyCaller* was used to obtain information about the coverage from both the tumor and the control, and to compare corresponding reads from both files to infer the copy number. Finally, applying a process of circular binary segmentation (CBS) on this output through the R library *DNAcopy* it was possible to gain insight about the presence of copy number aberrations.

The differences between the reference and the patient's genome were assessed through variant calling, which was performed using both *BCFtools* and *GATK*. For the first method we combined *BCFtools mpileup*, that generates genotype likelihoods (the probability of observing the sequencing data given a particular genotype at each genomic position), and *call,* that makes the variant calling itself. For the second one we used *GATK UnifiedGenotyper*.

These tools produce VCF files, which store genetic variant information like positions and genotypes. Finally, we used *vcftools* to filter the variants based on the quality and operate a comparison of the results.

A process of variant annotation was then performed to obtain information about the phenotypic consequences of the previously identified genetic variants. *SnpEff*, a variant effect predictor

program, was used, and in particular the tool *SnpSift* allowed us to identify the most significant ones. Two different files were used during the annotation process: the file containing the SNPs from HapMap3, and a file from ClinVar containing information about pathogenic variants.
SNPs and SNVs were identified through a process of somatic variant calling, using *VarScan2*. We used *samtools mpileup* to generate the pileup files, which were then filtered by *mpileup2snp* to isolate the SNPs. *vcftools* was used to filter the pileup files from the indels and limit the analysis to regions within a given mean depth (5 < DP < 200). Somatic variants were isolated by comparing the control and tumor samples using *VarScanv2 somatic* setting. Somatic variant annotation was then performed using *SnpEff*.
Ancestry analysis was performed in order to infer the population of origin of the patient. We used the *EthSEQ* R library using the BAM files with all the pre-processing applied.
Finally, purity and ploidy estimation was performed using *CLONETv2*, an R package which exploits both the log2ratio between tumor and control and the beta statistics to obtain information about the purity and clonality of the tumor sample.

Relevant results

Preprocessing

As a proof of concept, the statistics between the original tumor BAM and the processed and corrected one were compared. After the realignment, the number of unmapped reads dropped drastically (~70%) and the number of mismatches went from 5 million to about 2 million. Due to this change, there is an improvement of the error rate (mismatches over bases mapped).
After recalibration, we didn't see a significant drop of the quality, while with deduplication the number of sequences went from ~15 Mln to ~8 Mln.

Somatic copy number calling

The analysis of copy number was restricted to the regions that had coverage at least equal to 20. The average log2ratio of copy number change was ~-0.54, indicating that there is a prevalent loss of genetic material throughout the sequenced regions of the tumor sample.
From the analysis, 5237 regions are tagged as amplified (log2 > 0.25), while 34757 regions seem to be copy neutral. 121880 regions appear to be deleted (log2 < -0.25), spanning about 12 Mln base pairs: in particular, 376 regions (about 33k base pairs) present a homozygous deletion in the tumor. In Figure 1, we report the output of the CBS process, divided by chromosome: as previously stated, the most represented events are hemizygous deletions.

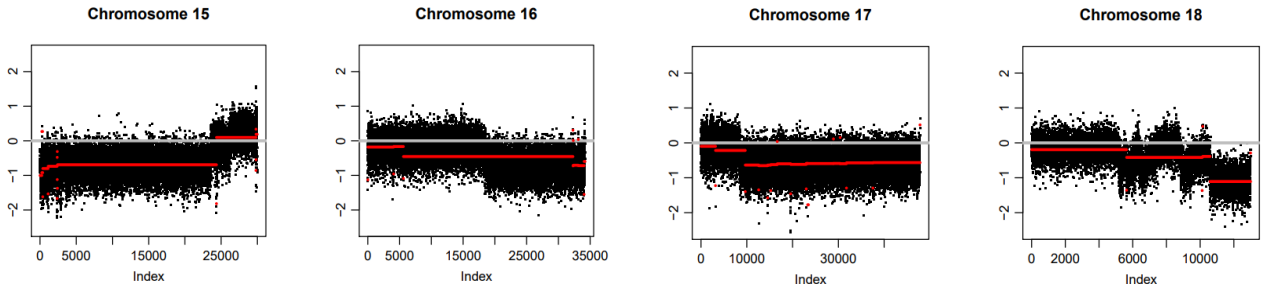

Figure 1. Output of the CBS process.
x-axis: index or position along the genome; y-axis: measurement of copy number ratio.

Variant calling and annotation

Looking at the two files created with BCFtools (both for control and tumor), we found that, out of a total of ~13000 variations compared to the reference genome, about 600 are only present in the tumor sample .

In the case of GATK, we found half of the variations (~8000) compared to the reference, but nevertheless the number of variants found in the tumor sample remained around the same.

Thanks to the annotation steps performed with SnpEff and then the filtering with SnpSift, we look for the variants with High Impact, that seems to be 15 in the control sample and surprisingly 14 in the tumor sample. The variant that is not in common is rs11071990, a nonsense SNV in the 15th chromosome. One of the variants in the 17th chromosome was rs11658717, thought to be modestly associated with breast cancer risk. [1]

In the next filter, we looked for clinical significance based on the annotation ClinVar, and observed only 1 variation as a result. This one also has a mutation between the control and the tumor sample, and it is located in BRCA1 (see in Figure 2). BRCA1 is known to be related to breast and ovarian cancer.

Somatic variant calling and annotation

With the tool VarScan mpileup2, searching for germline variants, we compare the variation annotated in the previous steps. With respect to the BCF annotated file, in the VarScan analysis, were found about a thousand variations that can be recognized as germline, both in control and in tumor sample.

Next, using VarScan somatic, we succeed to find ~3000 LOH and 450 somatic variants. Putting these variants against our reference genome, still 155 point mutations were referred to as somatic, and 130 out of them were located in protein coding genes.

Ancestry analysis

By performing ancestry analysis with EthSEQ on the two BAM genome files (using the SS2.Light.Model.gds) we found that the ethnicity of the patient is African, as shown by the purple dot in Figure 2.

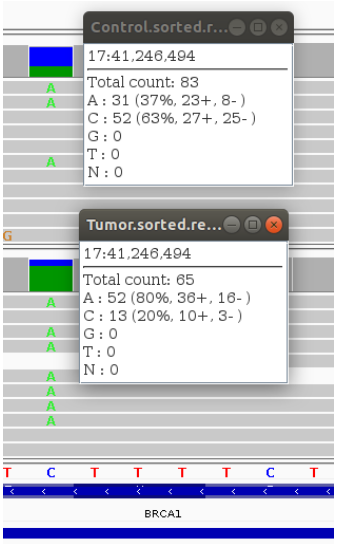

Figure 2.
Position in the genome of the CLNSIG variant

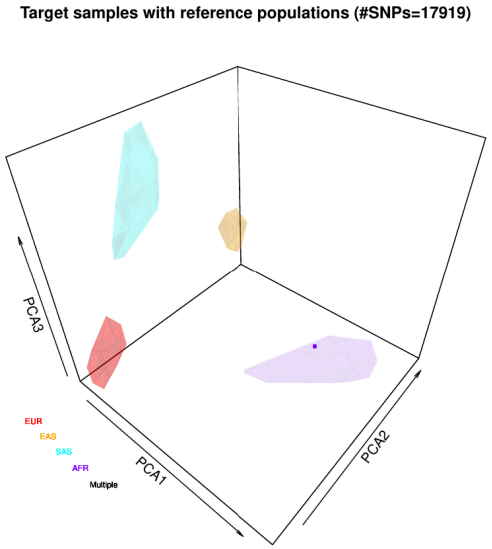

Figure 3.
Output of EthSEQ algorithm

Purity and ploidy estimation

From the output of CLONET, we can infer that the ploidy is estimated to be around 2.27, which is concordant with the extended deletions detected in previous analysis. The reported admixture is 0.37, which indicates that the tumor sample actually contained a significant fraction of normal cells which partially tainted the analysis.

In the graph we can see that most of the dots are clustered around logR = -0.5 and beta = 0.5, representing the presence of hemizygous deletion and some homozygous deletion as reported from the previous results of the copy number analysis. Some dots have logR = 1, but some with beta = 1 (probably due to the admixture) and the others seem to be CN-LOH. A couple of dots land in the region of logR > 0, which could represent the few amplifications detected earlier.

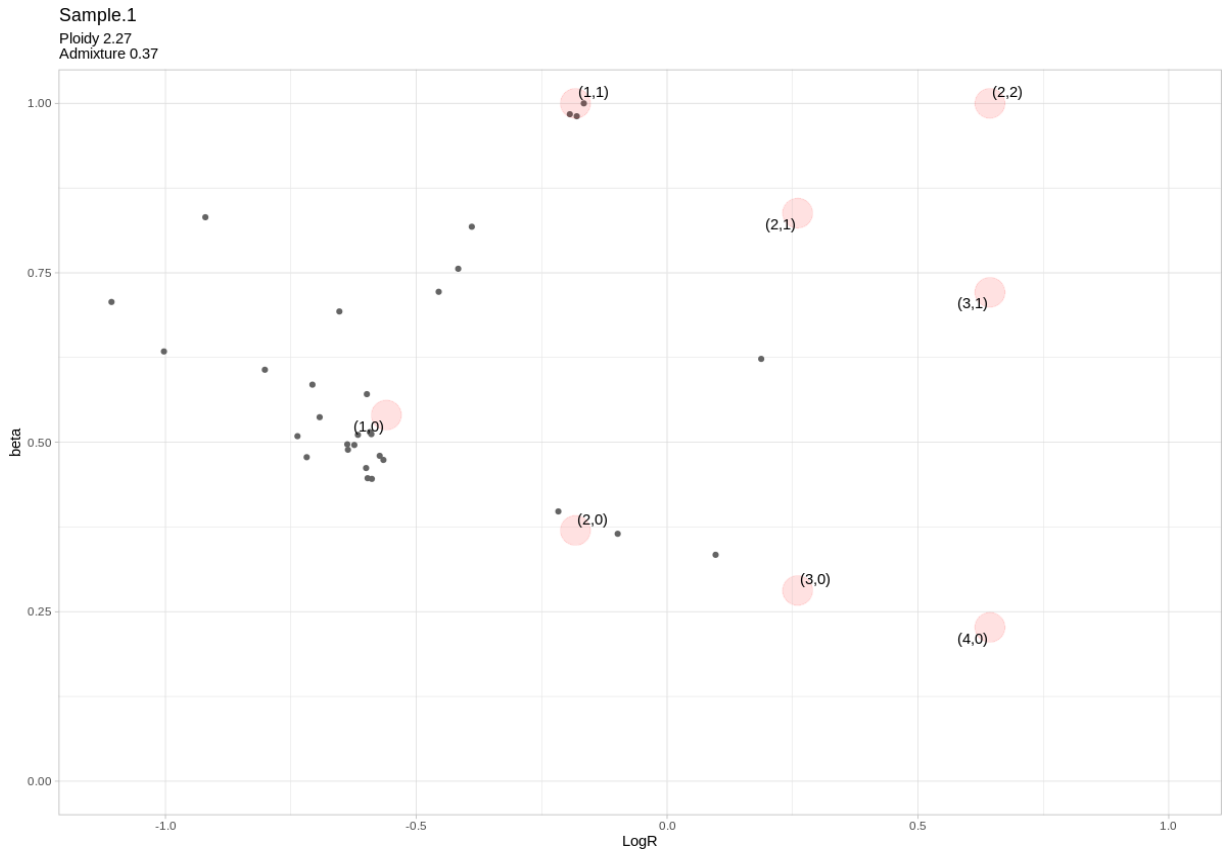

Figure 4. LogR-Beta plot obtained from CLONET. Each dot is a genomic segment. The orange dots represent the possible ploidy states of the sample: (1,1) represent the copy neutral (normal situation), (1,0) hemizygous deletion, (2, 0) copy neutral loss of heterozygosity (CN-LOH).

Conclusions and future directions

Through the implementation of a solid computational pipeline, we managed to pinpoint a few key structural characteristics of the tumor genome, while also characterizing the normal sample to act as a background.

Further refinements on this study could be done by using a more updated and representative reference genome, thus increasing the precision of the ancestry analysis. Another improvement could be the extension of the study toward BAMs that covers more genomic regions instead of only four chromosomes which are limited in size.

References

[1] Caswell, Jennifer L., et al. "Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors." *Human molecular genetics* 24.25 (2015): 7421-7431.