Computational Human Genomics Project, 2024-2025
Teamwork: Form groups of 3-4 students each.
Deadline: you need to submit your report 1 week before the exam date (yari.ciani@unitn.it; f.demichelis@unitn.it).

Introduction
Most genomic studies in the context of oncology require the characterization of somatic events assessed through the sequencing of bulk DNA tumor samples from patients' cohorts (i.e., multiple patients must be studied to obtain generalizable results and answer specific biological/clinical questions). However, the proper analysis of single patients can provide informative data.

Aim of the project
To detect and interpret the potential clinical relevance of somatic genomic aberrations harboured in the genome of an oncologic patient.

Execution
Starting from the two provided BAM files (i.e., tumor and control DNA seq from the same individual), identify all somatic events (e.g., SNV, CN), and define their clinical relevance. Running quality checks, manual inspections, and tumor population and sub-population characterization is advisable. Please, report the number of somatic variants, their type, and clinical relevance. Show in tabular format the most relevant events, indicating the involved genes, statistical significance, and the variables you consider most relevant. Provide visual representation of the obtained biological/technical results (min 2 figures, max 4 figures). Don't feel limited by the visualizations (and code snippets) used during lessons. Integrate your results with external cBioportal (https://www.cbioportal.org/) and COSMIC (https://cancer.sanger.ac.uk/cosmic) databases.

Each group must submit the following components as part of the assignment:

-**Informative Table of somatic events:** Create a clear, well-organized table summarizing the identified somatic events. Include relevant details such as gene name, variant type, predicted impact, and any annotation information.

-Mutational signature analysis: provide a summary of results and a brief interpretation explaining the biological or clinical relevance of the observed signatures

-Purity and ploidy estimation: present your findings along with graphical visualizations

-Choose one relevant somatic/genomic event and visualize it using IGV: Make sure the visualization: Clearly supports the identified event; is appropriately zoomed and colored to highlight the features of interest.

-Biological relevance contextualization: Write a short text explaining the biological significance of your findings. Use COSMIC and cBioportal repositories as support.

Format of the required project report:
Please comply with the following:

- 5 pages max, font size Arial 11, single line spacing.
- Include figures and related legends (min 2, max 4);
- include the following information: names of the students, project rationale (max 10 lines), summarized computational workflow (you don't need to report each command, but please report meaningful options, such as filtering thresholds), relevant results with related interpretations, and if necessary, pitfalls and criticisms (max 10 lines);
- figure axes, labels, and legends must be correct and complete. Size of figures should be appropriate (visible by human eye in printed format). Figure legends should be clear, informative, and self-contained (the reader should be able to understand the figure just by reading the figure legend).

Notes and Suggestions:

i. You are not required to use tools other than those utilized during the classes.

ii. Provide tables and/or visualizations describing the number and features (for instance AF and clinical relevance for mutations, log2Ratio for CNA) of identified somatic events.

iii. Scripts provided during the lessons are a good starting point. Feel free to optimize the code for the visualization and presentation of the results.

iv. If you need to determine the genotype, you can use ASEReadCounter, then compute allelic fraction (AF) and assign a genotype to each site considering the AF value (use the following thresholds: $AA<0.2$, $0.2 \leq AB \leq 0.8$, $BB>0.8$).

v. Please note that the provided BAM files are limited in genomic size for lighter processing. Still, some intermediary files may take up some GB of space. If you use the provided virtual machine, you may run out of space on disk: remember that some tools allow for setting a max depth parameter to avoid wasting too much space and time during the analysis.

If you have any question, write to yari.ciani@unitn.it

Have Fun!