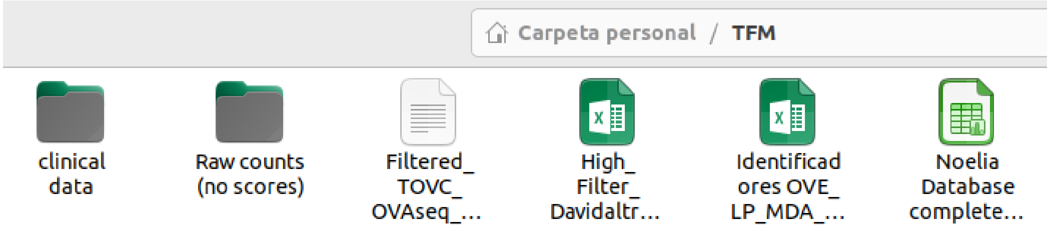
# Scripts and variables coding

## Input data used in the scripts


Carpeta personal / TFM
clinical data
Raw counts (no scores)
Filtered_ TOVC_ OVAseq_...
High_ Filter_ Davidaltr...
Identificad ores OVE_ LP_MDA_...
Noelia Database complete...

## Directories with scripts and description (https://github.com/nsn1992/TFM_Ovarian_Cancer)

### - Preprocessing_clinical_data

1. **LP_clinical.ipynb, OVE_clinical.ipynb, MDA_clinical.ipynb, RVB_clinical.ipynb**: These scripts do the prefiltering of the initial clinical records (clinical data directory) provided by the different centers and hospitals (LP=La Paz Hospital, OVE= Virgen del Rocio hospital, MDA: MD Anderson Center, RVB= Red Valenciana de Biobancos).
2. **LP_preunify.ipynb, OVE_preunify.ipynb, MDA_preunify.ipynb, RVB_preunify.ipynb:** These scripts unify the notation for the different clinical variables across all centers and hospitals.

### - Preprocessing_immunogenic_data

3. **TILs_raw_counts.ipynb:**
   This script generates the 'Genomic_tils_final_database.xlsx' file using input files with TILs raw counts ('Raw counts (no scores)' directory) and a database containing genetic, genomic, and TILs scores ('Noelia Database complete.ods' file) to clean and unify the data.

### - Intregration_data

4. **Combined_cohorts_final.ipynb**:
   This script generates the final working database ('Samples_alltypedata_annotated.xlsx') and other databases of interest. It integrates and standardizes the format of all data: clinical, genomic, genetic, and TILs information. This script uses as input the preprocessed clinical files (obtained after executing the scripts to preprocess clinical data: 1 and 2) and a file with genomic, genetic, and immunogenic information generated by script 3 ('Genomic_tils_final_database.xlsx').

   **The process for generating the final working database has been automated using the **'DB_TFM.sh'** script and includes the execution of the scripts contained in the previous directories in the order indicated (1,2,3 and 4).

## - Plots_statistics

The following scripts use as input the final working database as well as the mutations files (see input data) without filtering (‘Filtered_TOVC_OVAseq_all.txt’) or filtered for driver mutations (‘High_Filter_Davidaltreads40VAF0.1.xlsx’) when this type of analysis was performed.

5. **Definitive Oncoplots.R**:
   Generates oncoplots for EOC and CCOC driver mutations.
6. **Heatmap_MMRd_MMRmutations.R**:
   Creates a heatmap showing driver mutations in MMR genes for MMRd samples.
7. **Descriptive_analysis_genomics_definitive.ipynb**:
   Performs individual descriptive analyses of genomic and immunogenic features according to MMR status and histotype (EOC, CCOC, or the entire cohort).
8. **Integrated analysis 1.ipynb**:
   Analyzes TILs (raw counts) versus genomic variables, including scatter plots, visualizations, and statistical analysis.
9. **Integrated analysis 1b.ipynb**:
   Explores intraepithelial TILs (raw counts) versus clinical variables (All, MMRp, MMRd) using boxplots and statistical tests.
10. **Integrated_analysis2.ipynb**:
    Examines genomic variables versus clinical variables (boxplots and statistics) across EOC, CCOC, and the entire cohort, irrespective of MMR status (MMRp + MMRd samples).
11. **Integrated_analysis3.ipynb**:
    Analyzes genomic variables versus clinical variables (boxplots and statistics) for EOC, CCOC, and the entire cohort, focusing exclusively on MMRp samples.
12. **Barplots.ipynb**:
    Generates bar plots representing the percentages of samples with driver gene mutations by histotype and MMR status.
13. **Drivers_percentages_by_clinical.ipynb**:
    Creates plots showing the percentages of samples with mutations in the top 20 genes (from the entire cohort) across categories of different clinical variables (FIGO stage, differentiation grade, and residual disease) analyzed by histotype and MMR status.
14. **Allele_frequency_MMR_POLE.ipynb**:
    Studies allele frequencies in the unfiltered mutations file to determine the VAF threshold for filtering. A VAF > 0.1 was applied using an external R script (adapted from group code, not included). The filtered file with driver mutations is subsequently analyzed in this script to evaluate allele frequencies in MMR genes and POLE/POLD1 for MMRd samples.
15. **Survival_analysis_genomics_genetics.ipynb**:
    Performs survival analyses based on genetic and genomic variable categories. Includes results presented in Table 9 and some Figures in the Annex.
16. **Survival_analysis.ipynb**:
    To do survival analysis based on clinical and prognostic factors, including age at diagnosis, histotype, FIGO stage, residual disease, and tumor differentiation grade. This script covers

both univariate and multivariate analyses using the Cox proportional-hazards model (Table 2).

17. **Descriptive_analysis_clinical_individual_all.ipynb**:
    Conducts descriptive analyses to generate results for:

- Table 1: Clinical and histopathological features by histotype and for the entire cohort.
- Table 3: Clinical features by histotype and MMR status.
- Table 5: Immunogenic analysis by histotype and MMR status.

## Variable Notation used in scripts

### Genomic variables (Name in database: description)

1. 'MSI_sensor2': MSI score . It refers to the percentage of MS regions with MSI.
2. 'TMB': TMB score or total number of non synonymous coding mutations per megabase (Mb).
3. 'SUM ID2+ID7': Sum of indel signature values ID2 and ID7.Percentage of observed mutations in the tumour sample that can be attributed to ID2+ID7 mutational signatures.
4. '%genome_altered': Percentage of genome altered.
5. 'CNV': Number of copy number aberrations (CNAs)/events. One event is considered when a DNA segment has more or fewer than 2 copies (the normal value).

### Immunogenic variables (Name in database: description)

1. 'TILs_raw_ep': Number of TILs (raw counts) in the intraephitelial selected section of the tissue microarray (TMA).
2. 'TILs_raw_tu': Number of TILs (raw counts) in the intratumoral selected section of TMA.

### Clinical variables (Name in database: Definition and categories)

1. 'FIGOL': FIGO stage of the tumor. 2 categories: Localized (1,2), Advanced (3,4).
2. 'FIGOa': FIGO stage of the tumor. 4 categories: 1,2,3,4.
3. 'RESIDUALa': Residual disease status after surgery. 2 categories: Yes (Complete resection), No (Incomplete resection).
4. 'GRADE': Differentiation grade of the tumor. 3 categories: Well differentiated (Low grade or grade G1), Moderately differentiated (Medium grade or G2), Poorly differentiated (High grade or G3). In some analysis medium and high grade categories were grouped.
5. 'HISTOLOGY': Histotype of ovarian cancer. 2 categories: EOC (0) and CCOC (1).
6. 'MMR_final_status': MMR classification of the sample.2 categories: MMRp and MMRd.