**Readme - CAR-T analysis**

The biosurf folder is a bit chaotic. All the CAR-T related files are in **“biosurf/cartcontent”**.

The Rproject associated with biosurf (and the one to open when working with CART as well is **BioSurf.Rproj**)

**“data” folder:**

* cartcontent/data/FASTA\_sequences\_uniprot/canonical\_and\_isoforms 🡪 contains the FASTA files for all the targets that have been evaluated in the CART analyses. Each FASTA file was downloaded from UniProt and contains both the canonical and the variants for the target. As of 2024, we decided to only focus on 4 targets: CD7, CD19, EGFR and CD19.
* cartcontent/data/isoforms 🡪 contains one excel and one csv file for each target. The .xlsx files have been downloaded from ensembl genome browser (and converted into .csv). They contain ensembl transcript IDs for each target + additional information (e.g. if the transcript is protein coding). [This is an example for CD7](https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000173762;r=17:82314868-82317608).
* cartcontent/data/expr\_sub.Rdata 🡪 it contains two tables.

1. ensg\_expr: genes on the rows. On the columns there are the TCGA cancers, and the corresponding normal tissue from GTEX.This table contains the 33 genes that are used as targets for cancers with a correspondence in TCGA.
2. enst\_expr: transcripts (of the different gene isoforms) on the rows. On the columns there are the TCGA cancers and the corresponding normal tissues from GTEX.

* cartcontent/data/pheno.rds 🡪 table mapping each TCGA or GTEX sample to the corresponding cancer/tissue (category)
* cartcontent/data/gte.Rdata 🡪 list with all the transcripts for each gene. It also contains genomic coordinates.
* cartontent/target\_tcga.txt 🡪 list of targets that we originally planned to analyse. Needed to calculate the essentiality score (script 003).

Files that I don’t think are needed anymore, but that still are in the folder:

* cartcontent/data/data-2020-11-26.csv 🡪 I received this file in November 2020. It contains info on the targets that we evaluated the first time we run the CAR-T analyses.

**“scr” folder:**

The scripts are numbered based on the order they should be run in. The scripts named “00N\_...” are the first one that should be run and are preliminary to the analyses.

* 001\_biomart\_script.R 🡪 script to convert ensembl gene IDs and transcript IDs into Hugo gene symbol.
* 002\_msa\_script.R 🡪 script performing the multiple sequencing alignment. It is needed to make the protein topologies plot, where for each isoform we idenitify the ecto-, endo- and transmembrane- domains.
* 003\_target\_essentiality.R 🡪 script to calculate the target essentiality.
* 00\_functions\_2024.R 🡪 scripts with the functions needed to perform the target analysis in script 02\_Targets\_evaluation\_figures\_paper\_2024.Rmd. It is the newest version.
* 00\_functions\_old.R 🡪 the old version. I am keeping it because it contains the functions needed to make the plots that are currently in the website. My idea was to delete it once we have completed the target evaluation for the targets of interest and we have updated the website/paper consequently.
* 01\_data\_wrangling.R 🡪 script for data wrangling.
* 02\_Targets\_evaluation\_figures\_paper\_2024.Rmd 🡪 Rmarkdown with the analyses of the 4 selected targets (EGFR, CD7, MSLN, CD19).
* 02\_Targets\_evaluation\_figures\_paper\_2024.htlm 🡪 html resulting from the corresponding script.
* All the scripts named TARGET.Rmd (e.g. ALPP.Rmd) 🡪 old scripts. Here I apply the workflow to each target, and generate the figures that are currently on the biosurf website.

**“results” folder:**

* 01\_wrangled\_data 🡪 result of the 01\_data\_wrangling.R script
* aligned\_sequences 🡪 result of the msa script.
* BepiPred 🡪 analysis results from BepiPred (epitope prediction)
* DeepLoc 🡪 analysis results from DeeplLoc2. It predicts subcellular location of the target. For the old analyses, only the canonical isoform was investigated, but for the new ones the variants were included. 🡪 used to make heatmap of cellular location in script 02.
* Deeptmhmm 🡪 analysis results from DeepTMHMM. It predicts the targets/variants topology. 🡪 used to make protein topology plot in script 02.
* html\_docs 🡪 html documents generated by the target-specific Rmds.
* NetOGlyc 🡪 analysis results from NetOGlyc.
* NetPhosPan 🡪 analysis results from NetPhosPan.
* SignalP 🡪 analysis results from SignalP
* Topcons 🡪 analysis results from Topcons. Used to predict protein topology before switching to DeepTMHMM.
* CD19 delta exon 2 analysis 🡪 3 folders. One with images of canonical CD19 and delta exon2 variant. The other one is the output from the alphafold colab that was used to predict the delta exon2 variant structure. Note that AlphaFold2 does not give high confidence predictions for CD19 delta exon2, so we run the analysis again with [AlphaFold3](https://golgi.sandbox.google.com/). The results of AF3 are in the “Alphafold3\_results” folder, which also contains a readme file. Lars and I have been in touch with Magnus H. Høie (PhD student) to get better insights in both the delta exon2 variant and the single aa variant analyses. Further info/updates on the delta exon2 analysis are in the CAR-T report.
* CD19 single variant aa analysis 🡪 contains results for the single aa analysis. Also in this case we relied on Magnus. However, we cannot use the alphafold 2 (or alphafold 3) predictions, because it does not predict single aa variations well. Basically what is in this folder needs to be discarded and replaced with better analyses of the impact of the single aa variation on CD19. The rationale behind these analyses is that in the paper “Clinical Impact of Single Nucleotide Polymorphism in CD-19 on Treatment Outcome in FMC63-CAR-T Cell Therapy”, the authors claim that the L174V variation, present in ~50% of the population, causes therapy resistance in patients treated with CD19-CAR-Ts. When we analyse the predicted impact of the single aa variantion however, we cannot verify this hypothesis. Our hypothesis is that their patient cohort was too small to actually conclude what they do. The paper can be further examined to identify other weaknesses/observations if needed.
* ranges 🡪 contains rds filed used to make the targte topology plots in the old analyses. They are not updated. Not to be used in new analyses. The code creating these files is in 00\_functions\_old.R.
* NOT ON GIT: 003\_crispr\_target\_essentiality.rds. File resulting from script 003. Too big to be on git. It can be generated again running the scripts.

**Other files in the cartcontent folder:**

* cartcontent/Clinical trials and cancer types per CART target.xlsx 🡪 Excel file including the clinical trials where EGFR, CD7, MSLN, and CD19 have been evaluated.
* cartcontent/Report on choosen CAR-T targets.docx 🡪 report on the CAR-T targets.
* Readme – CART analysis 🡪 this file.