CCLE = cancer cell line encyclopedia

TCGA = The Cancer Genome Atlas

Link to list of cancer type abbreviations: <https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations>

Background info on gene names --- HGNC <https://www.genenames.org/>

-----

DEMETER2 gene essentiality data:

Data downloaded from

<https://depmap.org/portal/download/>

Select:

RNAi Screens → DEMETER2 Data v5

**README.txt**

Source: https://ndownloader.figshare.com/files/13515380

Has detailed info on the files available in this “directory”

**D2\_Achilles\_gene\_dep\_scores.csv**

DEMETER2 gene essentiality scores

Source: <https://ndownloader.figshare.com/files/11489669>

**RNAseq\_lRPKM\_data.csv**

Gene expression data. Log\_10(RPKM+.001) of protein coding genes. Derived from CCLE data

Source: <https://ndownloader.figshare.com/files/13110677>

→ may want to reduce this down to only the top x most variable genes or get rid of the least variable genes, or filter to include only genes expressed in x% of cell lines

((**D2\_Achilles\_gene\_dep\_scores.csv, RNAseq\_lRPKM\_data.csv**)

These two files have similar (not identical) cell lines and genes. The first is the essentiality scores, the second is the gene expression value

**Sample\_info.csv**

Source: <https://ndownloader.figshare.com/files/11489717>

This is the metadata info for each of the cell lines. Look here for hold-out info. Contains labels of tissue types (disease) and subtypes. Could hold out an entire tissue (disease type), or could hold out one subtype, e.g. all small cell lung vs non-cell lung, or hold out triple-negative breast cancer (TNBC) but leave other types of breast cancer (HER2 amplified, ER-positive)

**BETTER GENE EXPRESSION DATA TO USE:**

**CCLEv1\_hugo\_log2tpm\_58581genes\_2019-04-15.tsv**

Source: <https://xena.treehouse.gi.ucsc.edu/download/CCLEv1_hugo_log2tpm_58581genes_2019-04-15.tsv>

This gene expression data is log of TPM instead of RPKM, which lets you better compare across samples. This file contains more cell lines than we want (933) and the cell lines have different names than we are used to (names are in the format G#####, let’s call them the Treehouse IDs). You will need to convert the names from the Treehouse IDs back to the CCLE/Achilles IDs we are used to using the mapping file below, and then extract the ones we have Achilles data for. This reduces it from 501 to about 414 if my quick computations are correct. I don’t know why there are fewer here, something we could look into later.

**treehouseID\_to\_CCLEID.tsv**

Source: <https://xena.treehouse.gi.ucsc.edu/download/CellLineIndex_CCLE_RNAseq_forComp_051519.tsv> (and then I did a ‘cut’ to extract the two relevant columns)

Mapping of the Treehouse names for the cell lines (“G#####” format) to the CCLE/Achilles names we are used to and see in all the other files we have looked at. Note that this mapping removes the quotation marks that we see in the Achilles files -- this messed me up when I was doing a diff so make sure to standardize

<https://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/>