Highlights on DepMap Gene Dependency Data Ingestion

Kirill Tsukanov
Senior Full Stack Developer & Data Engineer
ktsukanov@ebi.ac.uk

Open Targets Perturbation Catalogue WP1/2/3 Technical Meeting

2025-02-04

DepMap Data Model

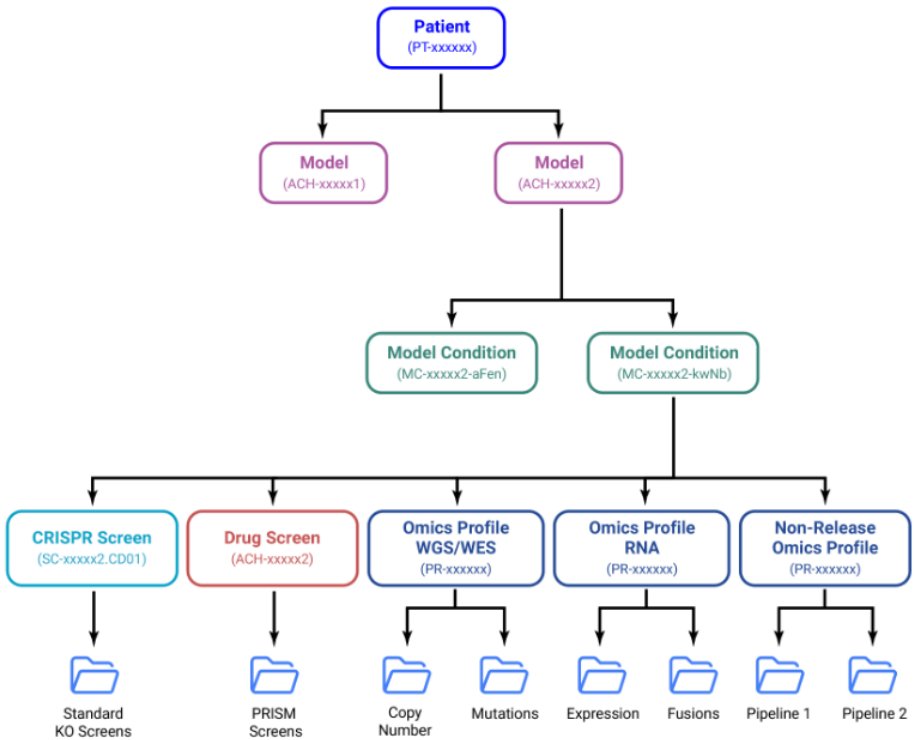
DepMap Data Structure

At the top of the hierarchy is Patient.

Models are a collection of cells derived from a single biopsy of the Patient. Each Patient can have one or more derived Models.

CRISPR and sequencing data are generated from a Model Condition. Each CRISPR Screen receives a Screen ID. Each sequencing datatype (e.g. wgs, rna, wes, etc.) receives an Omics Profile ID. Non-release Omics datasets (OLINK, ATAC-Seq) also receive an Omics Profile ID, but are not considered part of the bi-annual DepMap Release Dataset.

Although data is generated from Model Condition, DepMap Release data are indexed at two principal levels: Models and Screens/Profiles.



DepMap Data Overview

- 1,178 cancer screens $\times$ 17,916 genes
- Data consists of CRISPR knock-out screens to identify which gene disruptions slow the growth of specific cancer cell lines.
- Should we use effect size or probability?
  - Effect size provides granular information on the impact of gene knockout on cancer cell suppression.
  - However, DepMap recommends using **probabilities** as they incorporate **screen quality**, unlike effect sizes.

Missing Values

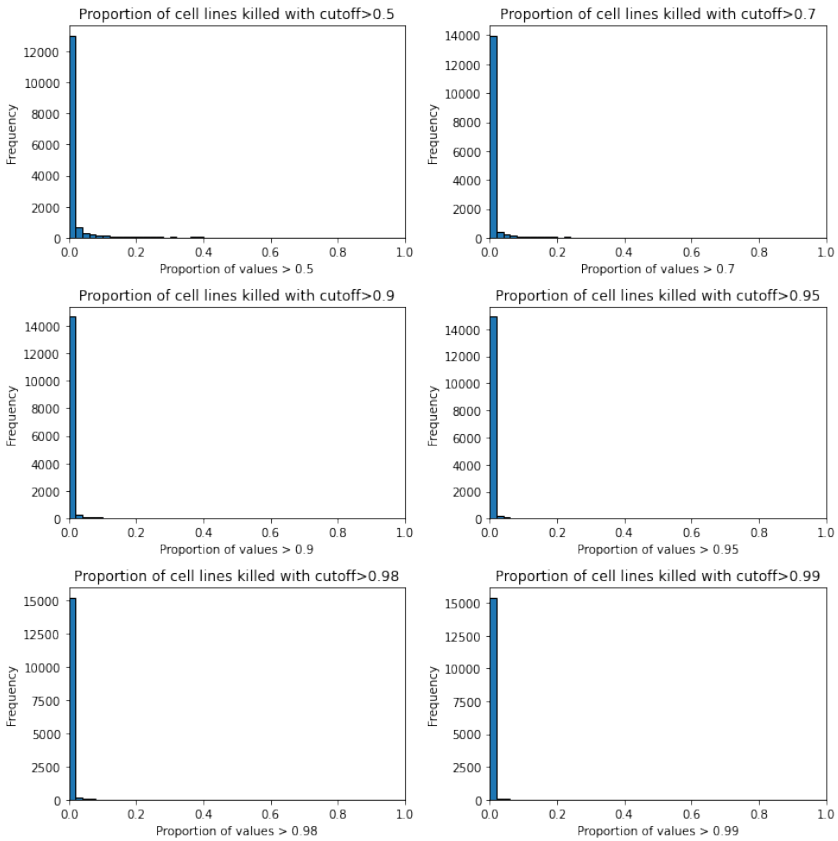
- 829 out of 17,916 genes (4.6%) have substantial missing data (absent in 5-30% of screens, possibly due to batch effects).
- Removing these genes results in a complete dataset of 1,178 screens $\times$ 17,087 genes, with no missing values.

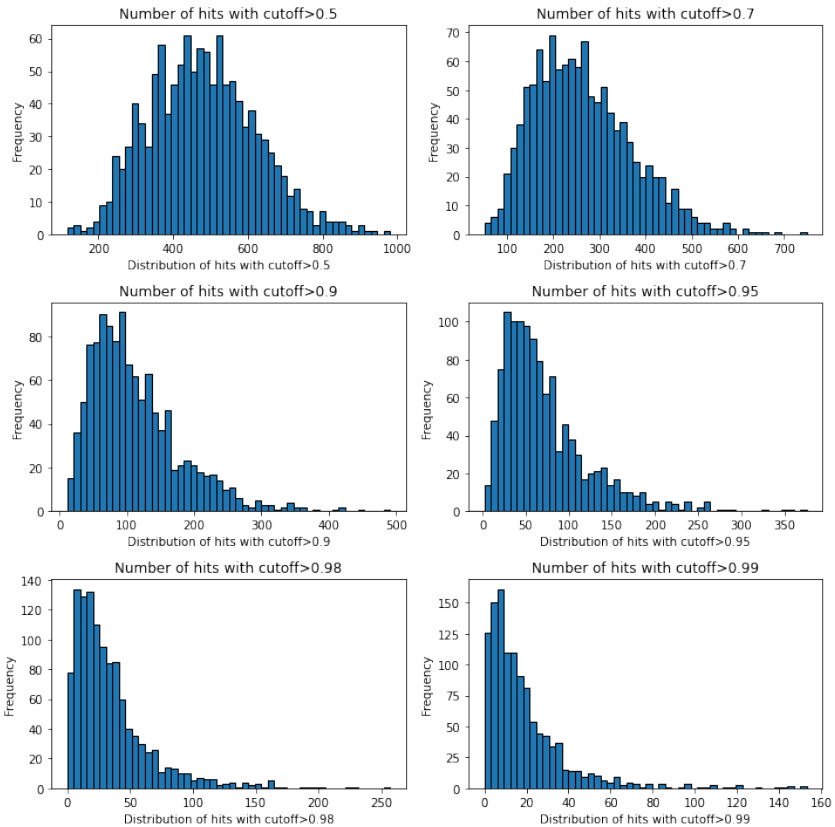
Common Essential Genes

- Some genes exhibit broad dependency patterns—knocking them out suppresses nearly all cancer lines (and likely healthy cells too!).
- Various filtering approaches:
  - Clustering
  - Literature on common essential genes
  - Two datasets from DepMap
- Currently using `CRISPRInferredCommonEssentials.csv`, though this may change.
- This reduces the dataset to 15,633 genes.

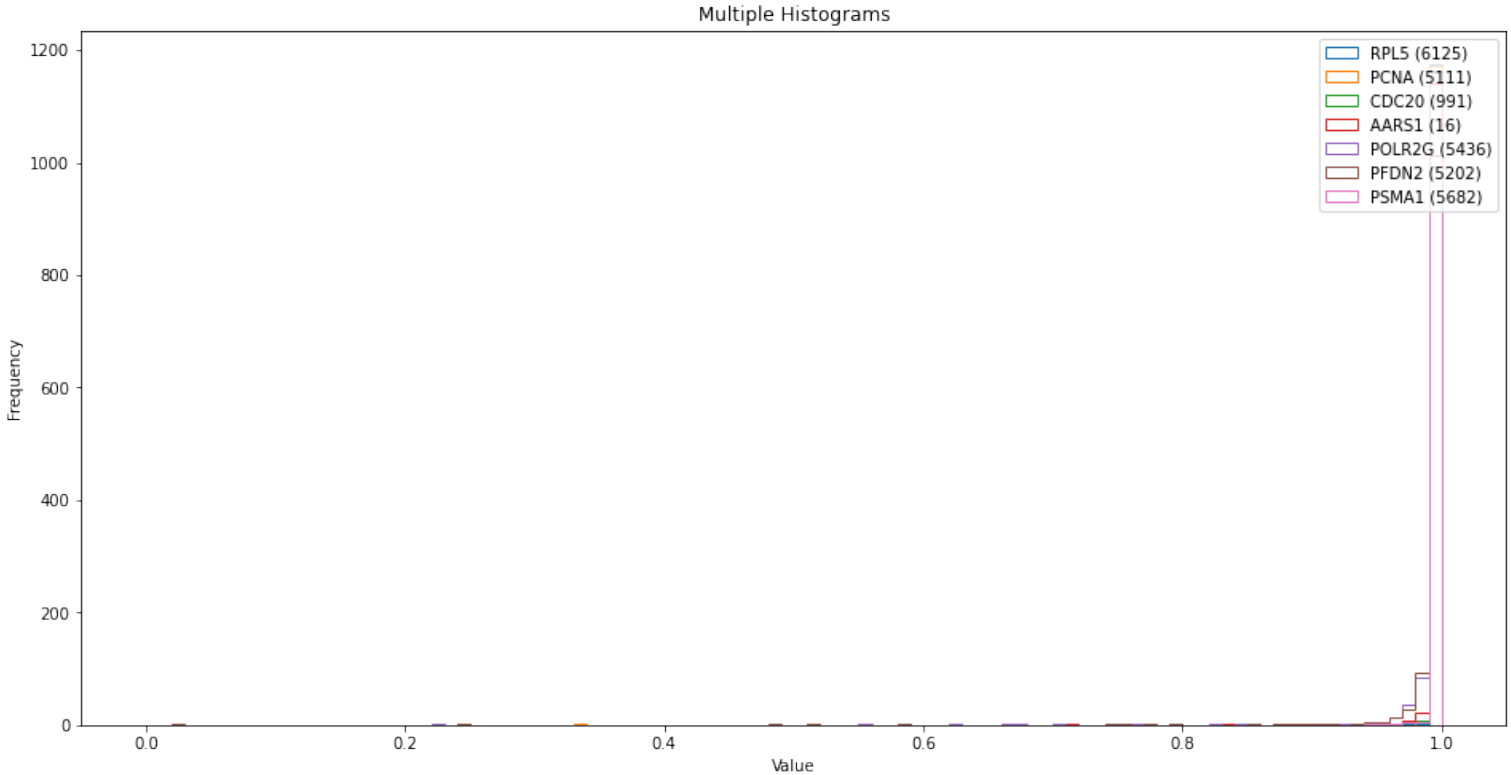
Threshold Selection

- Probabilities cannot be normalized or converted into Z-scores—hence, we need a reasonable threshold.
- Optimizing for cancer cell line suppression and number of gene hits, **95%** appears suitable.
- This threshold and the essential gene list will be user-configurable.





Validation for a Known Subset of Highly Essential Genes



Current Status and Next Steps

- Data processed and ingested into Elastic.
- Back-end under development.
- Front-end concept:
  - Users search for a cancer cell line of interest.
  - Highly dependent genes for the selected cell line are displayed.
  - Genes with associated MaveDB functional information are highlighted, enabling users to connect phenotype to gene and variant analysis.