A CRISPR Knockout Framework for Cancer Research

A DepMap-Centered Tutorial and Analysis

Anika Thatavarthy Param Somane Andrii Dovhaniuk Elizabeth Murphy

Department of Medicine
University of California, San Diego

Winter 2025

GitHub Repo:
github.com/ChestnutKurisu/MED263_Final_Project_WI25

Table of Contents

1.1 CRISPR: A Revolutionary Gene-Editing Tool

What is CRISPR?

- CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is a revolutionary gene-editing technology.
- Originates from a bacterial immune system; uses Cas proteins (e.g., Cas9) to target and cut DNA.

How It Works (High-Level):

- **Guide RNA (gRNA):** A designed RNA sequence that directs Cas9 to the target DNA sequence.
- **Cas9 "molecular scissors":** Cuts the DNA at the specified genomic locus.
- **Repair Mechanisms:**
  - Non-Homologous End Joining (NHEJ) $\rightarrow$ can introduce indels and knock out the gene.
  - Homology-Directed Repair (HDR) $\rightarrow$ can insert new genetic material precisely.

1.2 Visual Overview of CRISPR Editing

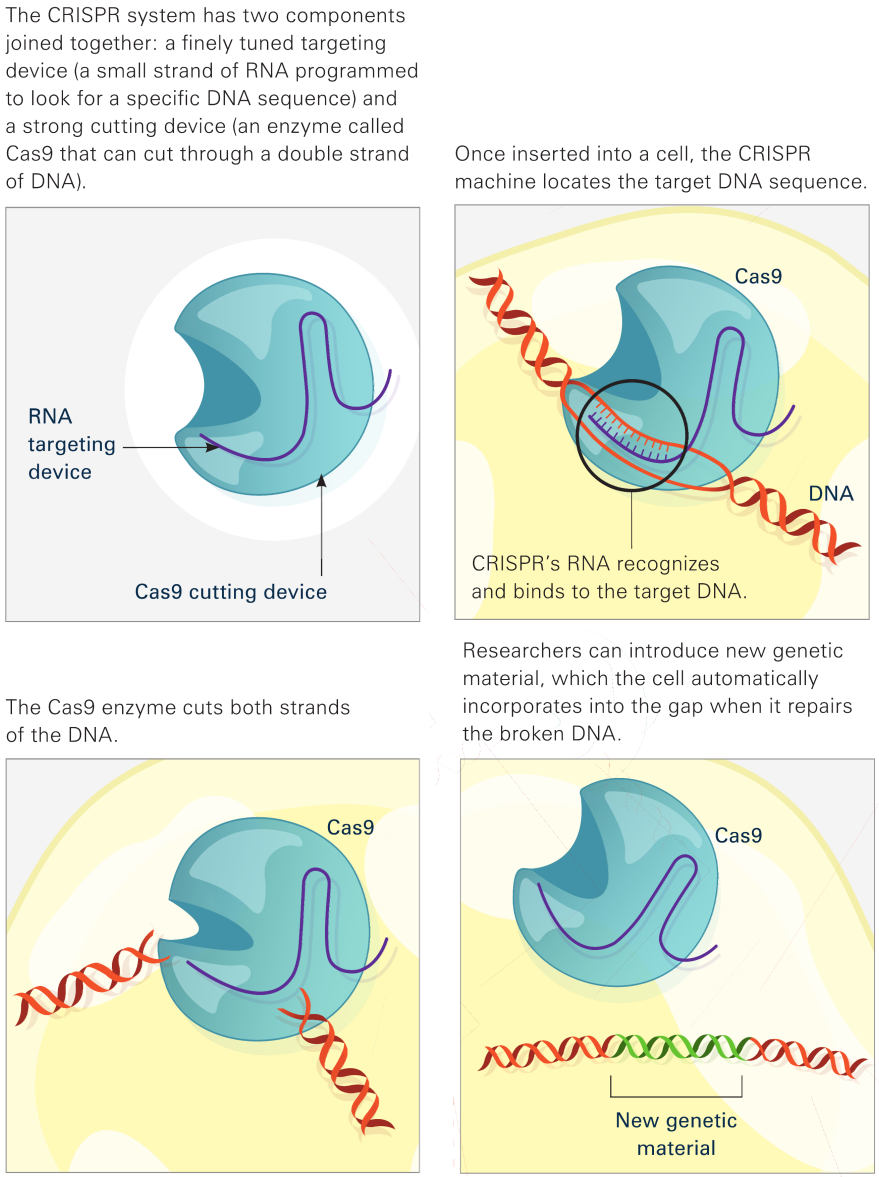
Key Steps:

- **1.** CRISPR RNA (purple) + Cas9 (blue) locate target DNA.
- **2.** Cas9 makes a cut in the double-stranded DNA.
- **3.** Cell repairs the break.

Why It Matters:

- Enables highly targeted gene modifications.
- Potential for treating genetic diseases, creating knockout models, etc.

Image Source: NIGMS, ID 3719.



1.3 The DepMap Project

The Cancer Dependency Map (DepMap):

- A large-scale resource identifying essential genes for tumor survival using genome-wide CRISPR knockout screens.
- Integrates multi-omics: copy number, gene expression, mutations, etc.
- Public portal: depmap.org/portal
- dbGaP: phs003444.v2.p1

Why DepMap Matters:

- Systematic "dependency" data across hundreds of cancer cell lines.
- Fuels precision oncology by linking genomic features to gene essentialities.

Note on ModelID:

- Each DepMap cell line is assigned a unique `ModelID`.
- Ensures consistent referencing across CRISPR effect, dependency, CN, and expression data.

1.4 Key Score Definitions (CERES, Chronos, etc.)

CERES Scores (*Main metric used in DepMap*)

- Corrects for copy-number effects in CRISPR-Cas9 screens.
- Scale: Negative = essential, near 0 = non-essential.
- Large negative scores imply strong depletion (i.e., gene knockout is detrimental).

Chronos Scores (*Alternative model to CERES*)

- An alternative model that refines gene fitness effects over time.
- Often yields a heavier tail for highly essential genes.
- Generally correlates with CERES but can differ for common essentials.

Dependency Probability (*Likelihood-based ranking of gene essentiality*)

- Interpreted as the likelihood a gene is essential & helps identify context-dependent vulnerabilities.
- Ranges from 0 (not dependent) to 1 (strongly dependent).

Gene Expression (*mRNA abundance & dependency correlation*)

- Typically measured in log2(TPM+1).
- High expression can correlate with "oncogene addiction," but not always a direct proxy for essentiality.

1.5 Schematic of the CERES Model

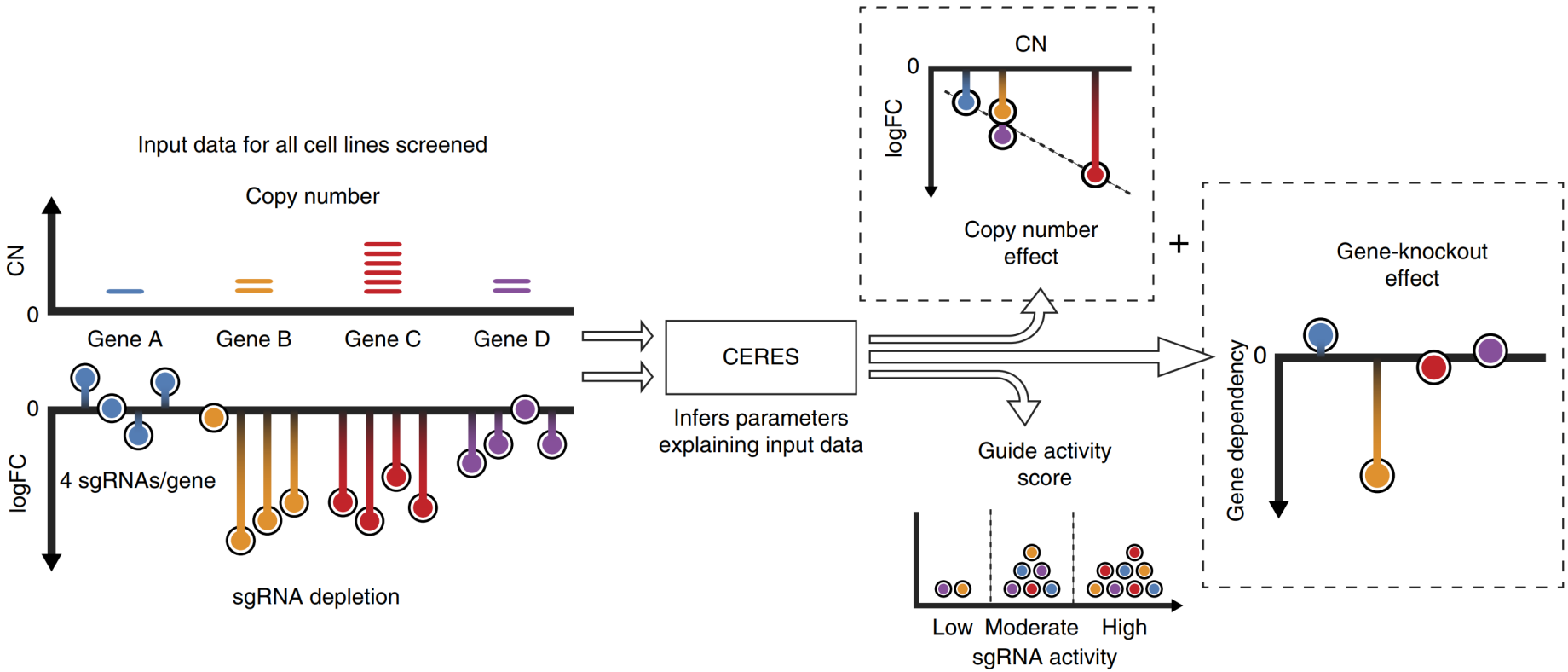


Figure: Schematic of the CERES computational model. As input, CERES takes sgRNA-depletion and copy number data for all cell lines screened. During the inference procedure, CERES models the depletion values as a sum of gene-knockout and copy number effects, multiplied by a guide activity score parameter. It infers the parameters (copy-number effect, guide efficiency, etc.) that best explain the observed CRISPR depletion data (maximum-likelihood of the observed data).

Source: Meyers *et al.*, *Nature Genetics*, 2017.

Follow Along with the Tutorial

Google Colab Notebook:

- We have an interactive Jupyter/Colab notebook that walks through the data loading, merging, EDA, and modeling steps.
- *Link:* **Colab Notebook**
- Download or open directly to run each cell and replicate the results.

2.1 Data Preprocessing Workflow

Goals:

- Subset large DepMap files to genes of interest.
- Standardize column names (e.g., `ModelID`).
- Prepare consistent CSV files for downstream EDA and modeling.

Script: `filter-preprocess-data.py` (GitHub Link)

- **Select Genes:**
  - Curated list of core cancer genes (e.g., `TP53`, `EGFR`, `KRAS`, `BRCA1`, etc.).
- **Chunk Reading:**
  - DepMap CSVs can be hundreds of MBs.
  - Reads in chunks of 5000 rows, concatenates selected columns.
- **Outputs:**
  - `CRISPRGeneEffect.csv`, `CRISPRGeneDependency.csv`
  - `OmicsCNGene.csv`, `OmicsSomaticMutationsMatrixDamaging.csv`
  - `OmicsExpressionProteinCodingGenesTPMLogp1.csv` (plus stranded version)

2.2 Preprocessing Steps

Key Highlights:

1. **Fix Headers:**
   - The script checks for blank or unnamed columns and renames them to `"ModelID"`.
2. **Filter by Genes:**
   - Retains only columns matching the user-provided list (e.g., `EGFR`, `BRCA1`, etc.).
3. **Save Cleaned CSVs:**
   - Each subset is saved in an output directory (e.g., `DepMap-preprocessed/`).

Motivation:

- Reduces file size, speeds up analysis.
- Ensures consistent `ModelID` merges across multiple omics layers.

3.1 Data Loading Overview

Files merged on `ModelID`:

- `CRISPRGeneEffect.csv` (Chronos-corrected effect scores)
- `CRISPRGeneDependency.csv` (Dependency probabilities)
- `OmicsCNGene.csv`, `OmicsSomaticMutationsMatrixDamaging.csv`
- `OmicsExpressionProteinCodingGenesTPMLogp1.csv`
- `Model.csv` for lineage & metadata

Cleaning Steps:

1. Renamed columns to `GENE_eff`, `GENE_dep`, `GENE_CN`, `GENE_mut`, `GENE_expr`.
2. Filtered missing or inconsistent rows.
3. Merged into a final `merged_df`.

3.2.1 ARID1A: Selected Gene of Interest

ARID1A Overview:

- **Gene Name:** AT-Rich Interaction Domain 1A
- **Function:** Core component of the SWI/SNF chromatin-remodeling complex.
- **Clinical Relevance:**
  - Frequently mutated in ovarian clear-cell carcinoma, endometrial carcinoma.
  - Functions primarily as a tumor-suppressor gene.
- **OMIM Link:** OMIM ARID1A (603024)

Why ARID1A?

- Mutations can lead to partial or complete loss-of-function in tumor suppression.
- DepMap reveals variability in ARID1A essentiality across different cell lines.
- ARID1A is a **tumor suppressor**, so loss-of-function mutations can promote tumorigenesis. In CRISPR screens, ARID1A-knockout often shows weaker effects in lines already harboring a damaging mutation.

3.2.2 KRAS: Selected Gene of Interest

KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog)

- One of the most commonly mutated **proto-oncogenes** in human cancer.
- Encodes a GTPase regulating cell proliferation, differentiation, and survival.
- **Activating mutations** often lock KRAS in a GTP-bound "active" state.
- Prevalent in **pancreatic**, **colorectal**, and **lung** cancers, driving oncogenic signaling (e.g., MAPK-ERK).
- **OMIM Link:** OMIM KRAS (190070)

Clinical & Therapeutic Relevance:

- Major driver of tumor initiation and progression in many cancer types.
- Under active investigation for targeted therapies (e.g., KRAS(G12C) inhibitors).
- Mutations can also be associated with certain developmental syndromes (e.g. Noonan).
- KRAS is an **oncogene**; its mutations enhance cell growth and proliferation. CRISPR knockout yields strongly negative scores in cell lines highly dependent on KRAS signaling.

3.3 Exploratory Data Analysis (EDA)

Initial Checks:

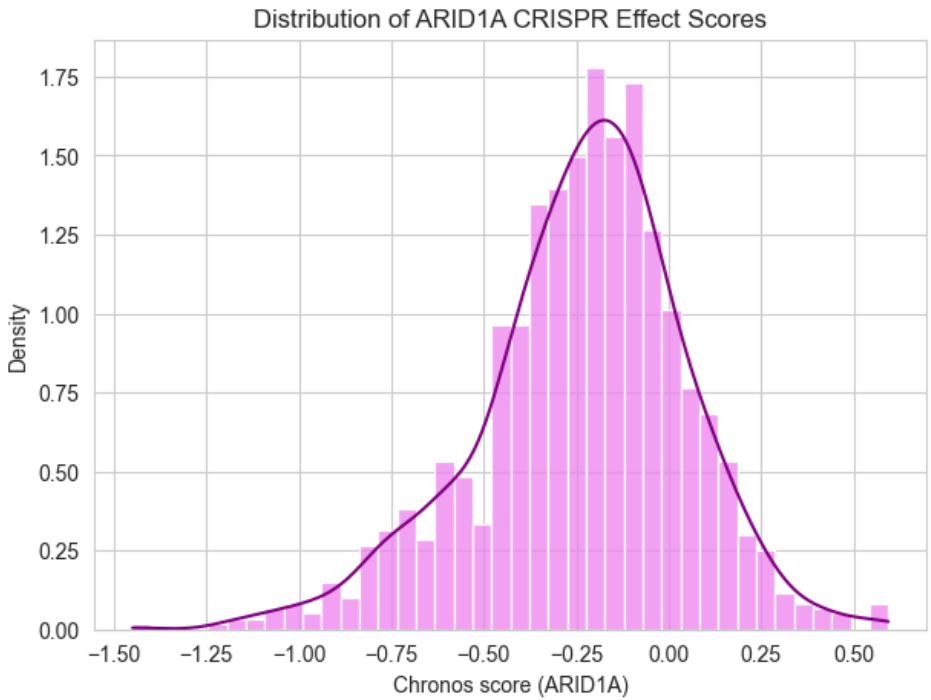
- Summary statistics for effect & dependency scores.
- Gene expression distributions (log2(TPM+1)).
- Mutation frequencies: `ARID1A`: ∼10% mutated lines (total 118/1178); e.g., `ARID1A_mut` distribution: `0 = 1060, 1 = 68, 2 = 50`.

Pitfalls:

- Some lineages with very few samples → less statistical power.
- Expression outliers from highly amplified genes.
- Non-normal score distributions → Mann–Whitney U or other nonparametric tests.

3.4 ARID1A Effect Score Distribution

- Negative Chronos effect implies higher essentiality.
- ARID1A has mean score ≈ -0.236, a small fraction of lines < -1.

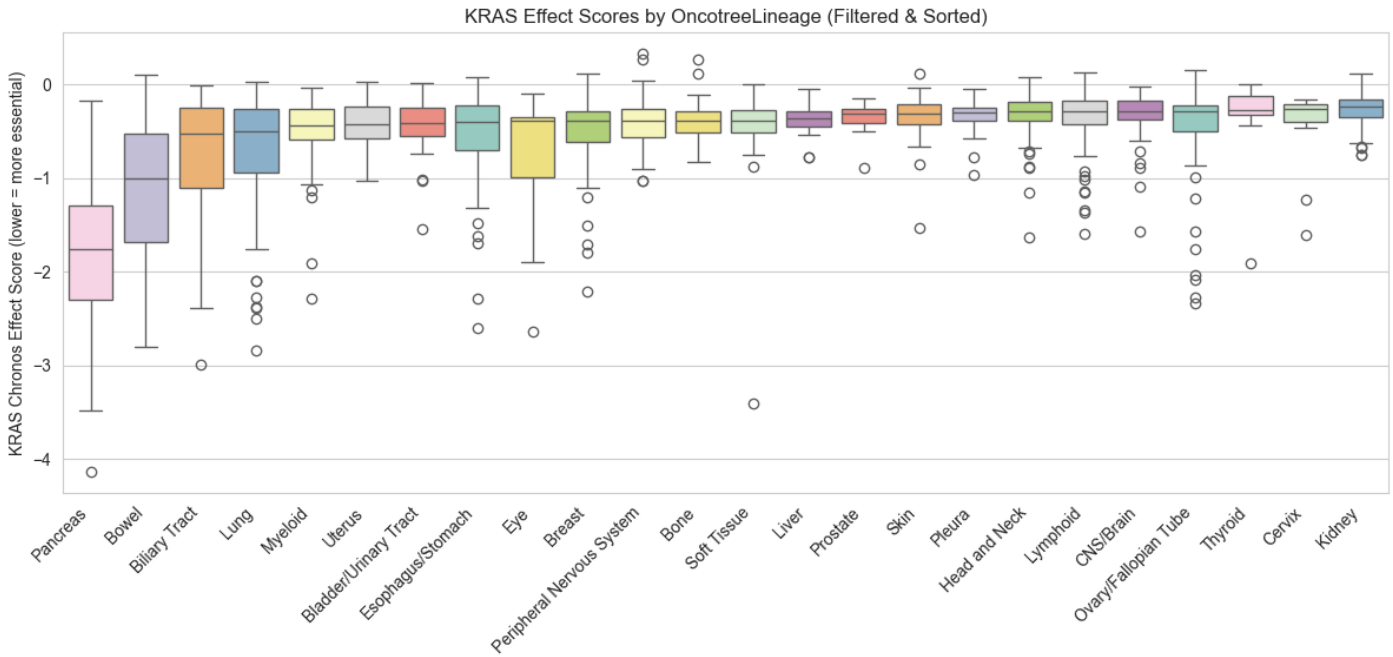


Interpretation:

- Heterogeneous essentiality distribution: many lines near mild essentiality or neutral.
- Strongly negative tail indicates a subset highly dependent on ARID1A.

3.5 KRAS: Lineage-Specific Effect Scores

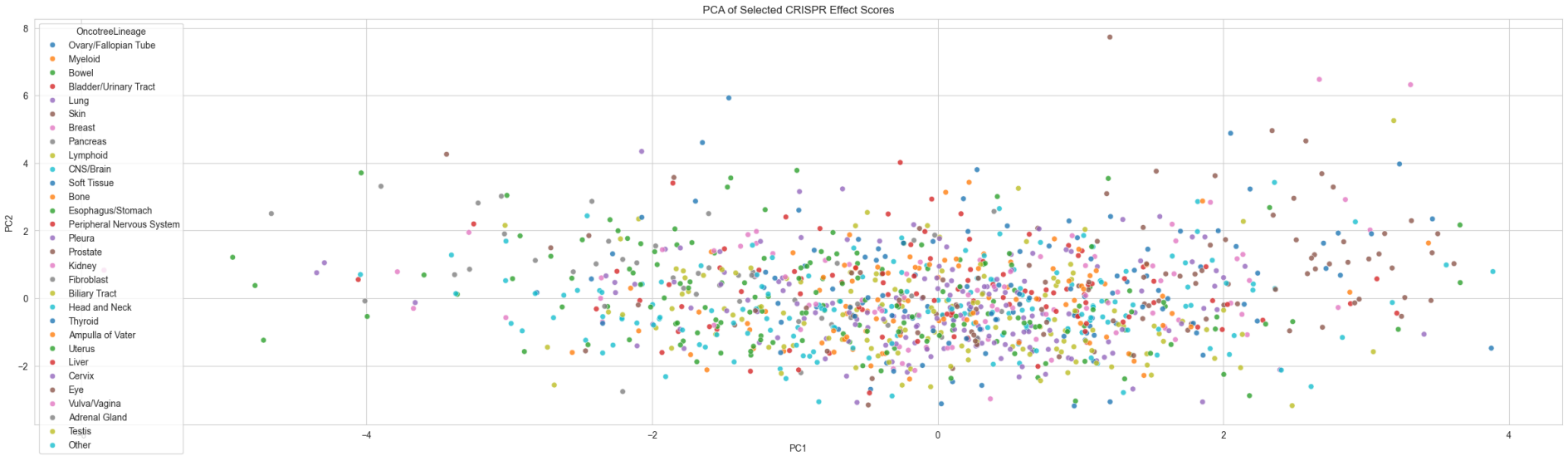
- **Pancreatic** cell lines show the most negative median KRAS effect (≈ -1.76), consistent with strong KRAS dependency.
- **Bowel (colorectal)** lines also rank highly in negative KRAS scores, aligning with frequent KRAS mutations in colorectal cancer.
- **Less negative medians** (e.g., Kidney) suggest lower KRAS dependence, possibly due to alternate oncogenic drivers.
- Overall, these data underscore *lineage-specific KRAS essentiality*, matching the known biology that KRAS is a critical driver in pancreatic and colorectal tumors but may be less central in other tissues.



4.1 PCA Projection

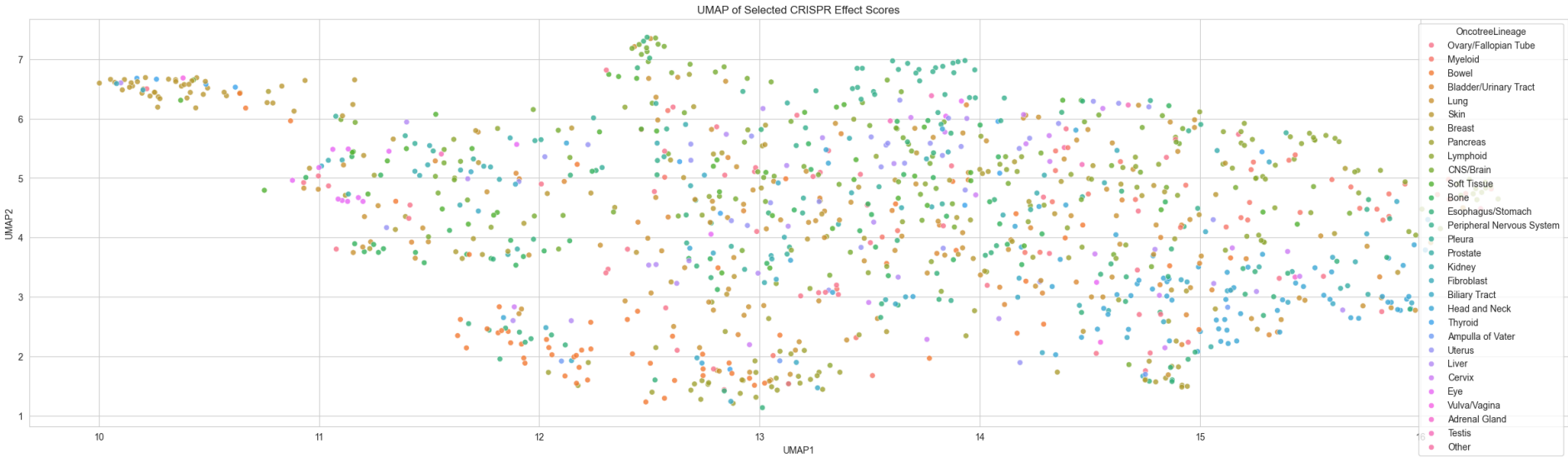
Goal: Summarize thousands of features into a few principal components.

- PC1 + PC2 often explain $< 20\%$ of total variance.
- Tissue-specific clustering not strongly observed.



4.2 UMAP Results

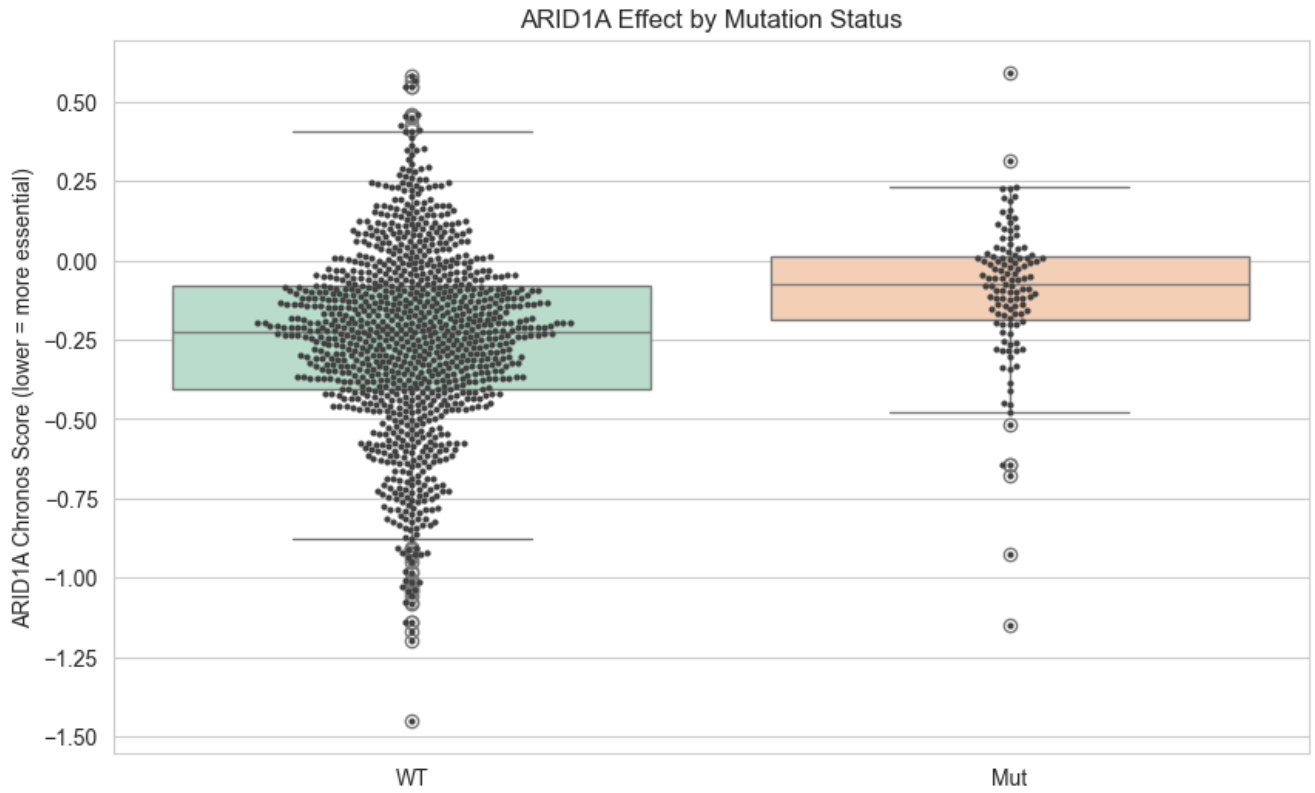
- **UMAP** better preserves local distances; partial lineage clustering but still overlapping.
- Suggests that CRISPR effect scores alone do not strictly separate lineages in 2D.



5.1 Mann–Whitney U: ARID1A Mutant vs. WT

Hypothesis: ARID1A knockout has different effects in mutant vs. wild-type lines.

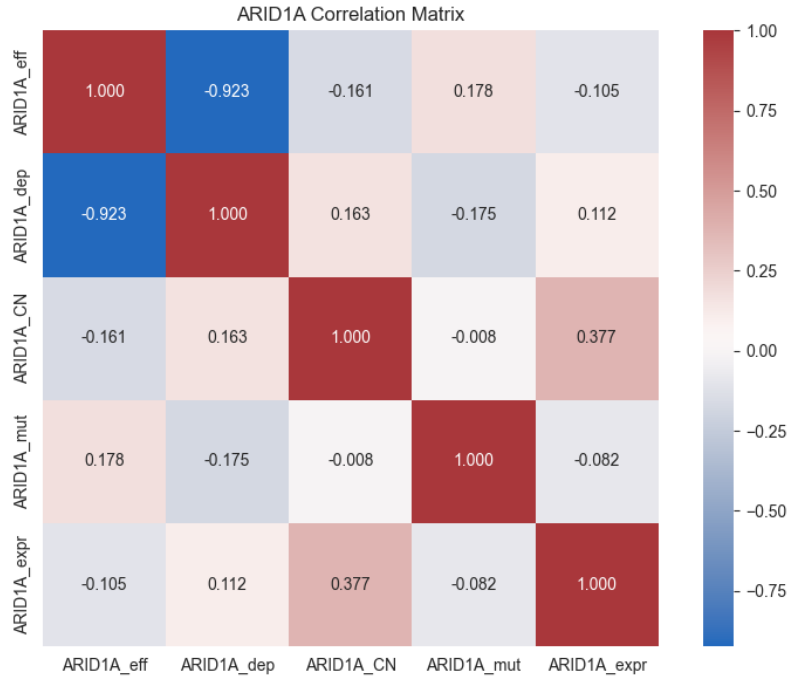
- p-value $\approx 9.17 \times 10^{-11}$ $\rightarrow$ highly significant difference.
- Mutant lines $\rightarrow$ less negative (knockout has weaker impact if gene is already partly inactivated).
- WT lines $\rightarrow$ more negative effect (fully functional ARID1A crucial).



5.2 ARID1A Effect vs. Dependency & Correlation Matrix

Correlation Matrix (ARID1A_eff, _dep, _expr, _mut, etc.)

- $\mathrm{Corr}(\text{ARID1A\_eff}, \text{ARID1A\_dep}) \approx -0.92$.
- $\mathrm{Corr}(\text{ARID1A\_eff}, \text{ARID1A\_expr})$ weaker, near -0.10.



Interpretation:

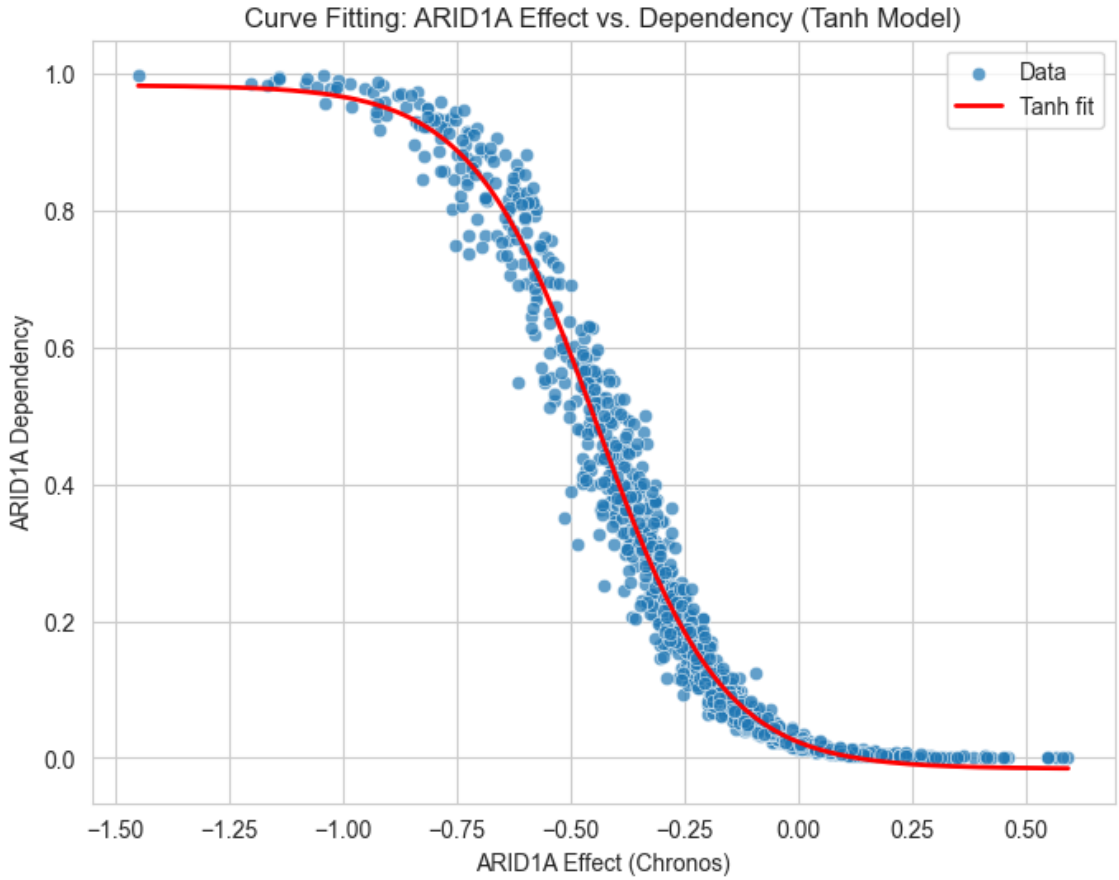
- `ARID1A_dep` is inversely correlated with `ARID1A_eff` (more negative = higher dependency).
- Expression alone does not strongly predict essentiality for ARID1A.

5.3 ARID1A Effect vs. Dependency

Chronos effect score vs. dependency probability

- Often strongly (negatively) correlated: more negative effect → higher probability of essentiality.
- **Non-linear fit**: a tanh or logistic curve can model the saturation at extremes.

Plot: $\mathrm{ARID1A}_{\mathrm{dep}} \approx 0.4996 \tanh(-3.6348\, \mathrm{ARID1A}_{\mathrm{eff}} - 1.5996) + 0.4838.$



6.1 Single-Gene Random Forest: KRAS Dependency

Goal: Predict `KRAS_dep` $\in [0, 1]$ using expression, copy number, mutation data, and metadata (Oncotree lineage, etc.).

Model: `RandomForestRegressor` with 100 trees, trained on an 80/20 split of 1178 samples.

Results:

- $R^2 \approx 0.2637$ on the held-out test set.
- MSE ≈ 0.078.
- **Top Features**:
  - `KRAS_CN` (copy number)
  - `OncotreeLineage_Pancreas`
  - `OncotreeLineage_Bowel`
  - `MET_expr`, `EGFR_CN`, `PTEN_CN`, etc.

Interpretation:

- Amplifications/deletions of `KRAS` drive dependency strength.
- **Lineage** (pancreatic, colorectal) is crucial: known KRAS-addicted tumors.
- Additional CN changes (e.g. `EGFR`, `PTEN`) modulate KRAS reliance.

6.2 Multi-Gene Regression with Random Forest

Approach: MultiOutput Random Forest to predict multiple gene dependencies simultaneously from the same multi-omics + metadata feature set.

- Retained genes with $\sigma(\text{dep}) \geq 0.25$.
- Trained on all lines with non-missing data ($n \approx 1178$).
- Evaluate each gene's R^2 and MSE individually.

Findings:

- `CTNNB1_dep`: highest $R^2 \approx 0.58$.
- `CDK6_dep`: $R^2 \approx 0.46$.
- `KRAS_dep`: $R^2 \approx 0.26$.
- Tumor suppressors like `BRCA1`, `ARID1A` show lower predictability ($R^2 \approx 0.04$–0.06).

Interpretation:

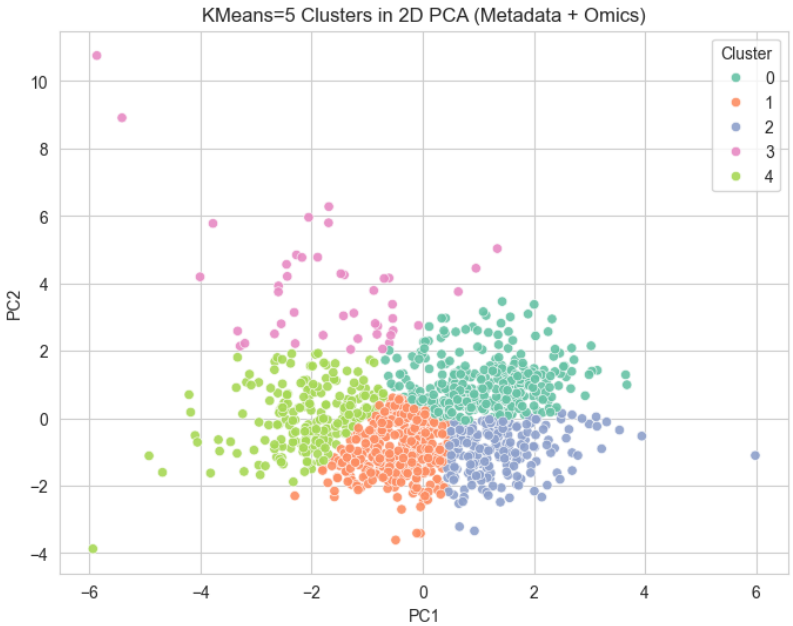
- Genes central to known oncogenic pathways (e.g., β-catenin/Wnt) are more systematically predicted by lineage and multi-omics signals.
- Tumor-suppressor essentialities can be more context-dependent or confounded by partial/inactivating mutations.

6.3 Unsupervised Clustering of Cell Lines

Method:

- Used PCA on expression and CN features (scaled), then KMeans ($k = 5$).
- Clusters do not perfectly match single lineages; some groups are mixed.

- **Cluster 0**: Lymphoid, Skin, Myeloid
- **Cluster 1**: Lung, CNS/Brain, Uterus
- **Cluster 2**: CNS/Brain, Lung, Bone
- **Cluster 3**: Bowel-dominant cluster
- **Cluster 4**: Head/Neck, Stomach, Pancreas



Takeaway: Some lineages cluster together, but there's overlap. Multi-omics signals can partition certain subtypes (e.g., Bowel, Head/Neck).

7. Conclusions & Future Directions

Key Takeaways

- **ARID1A** shows context-dependent essentiality:
  - Mutant lines have weaker knockout effects.
  - Mann–Whitney p-value $\sim 9.17 \times 10^{-11}$.
- **KRAS dependency** strongly linked to lineage (pancreatic, bowel) and copy-number variations (`KRAS_CN`). Single-gene RF yields $R^2 \approx 0.26$.
- **Multi-gene RF** reveals `CTNNB1`, `CDK6` are more predictable ($R^2 > 0.45$), whereas `BRCA1`, `ARID1A` remain difficult to predict ($R^2 < 0.10$).

Next Steps

- Further refine multi-omics (proteomics, epigenetics, drug-response data) for more accurate essentiality predictions.
- Investigate advanced ML (deep neural networks, ensemble methods) for gene dependencies.

8.1 CRISPR in Cancer: Clinical Trials & Immunotherapy

CRISPR-Based Cell Therapies:

- *Ex vivo* approach: Editing T cells (e.g., knocking out **PD-1**, TCR genes) and re-infusing them to enhance tumor targeting.
- Early-phase **CRISPR-CAR T-cell** trials in blood cancers show feasibility & safety.
- Allogeneic (off-the-shelf) CAR T-cells leverage multiplex gene edits to avoid graft-vs-host disease.

In Vivo Gene Editing:

- Researchers exploring direct delivery (via gels, viral/lipid nanoparticles) to *knock out* oncogenes *in situ*, although still early-stage.
- Example: CRISPR gel targeting HPV E6/E7 in cervical lesions.

Proof of Concept:

- First CRISPR therapy (*exa-cel*, for sickle cell) approved in 2023, heralding clinical acceptance of CRISPR-based treatments.
- Similar ex vivo editing strategies are now applied in oncology pipelines.

8.2 Base Editing, Prime Editing, & Off-Target Mitigation

Base Editing:

- Creates precise point mutations (e.g. C-to-T) without a double-strand break.
- Enables complex "multi-gene" edits in T-cells, reducing large genomic rearrangements.
- Already entering clinical trials for T-ALL (quadruple-edited CAR T-cells).

Prime Editing:

- Uses a reverse transcriptase with Cas9 nickase to "search-and-replace" DNA sequences.
- Can theoretically fix a wide range of mutations *in situ*.
- Currently lower efficiency than base editing, but active improvements are ongoing.

Off-Target Effects & Safety:

- High-fidelity Cas9 variants and advanced guide design reduce unintended cuts.
- FDA/EMA approvals require robust off-target screening (e.g., GUIDE-seq) and long-term monitoring.
- *Ex vivo* protocols further minimize immunogenic risk by removing Cas9 protein before infusion.

8.3 Ethical, Regulatory, & Commercial Outlook

Ethical & Regulatory Considerations:

- Worldwide prohibition on **germline** editing (heritable changes); clinical focus remains on **somatic** therapies.
- Regulators (FDA, EMA) demand thorough safety data: off-target analyses, immune response checks, post-trial surveillance.
- Affordability & equitable access remain major concerns as gene-edited therapies tend toward high initial costs.

Industry & Future Directions:

- Biotechs (CRISPR Therapeutics, Editas, Intellia, Beam) lead the push toward commercialization: ex vivo CAR T-cells, in vivo editing, base editing, prime editing.
- Ongoing wave of Phase I/II trials for solid & hematologic cancers. Positive safety + efficacy could drive Phase III expansions.
- Integrating **multi-omics data** (like DepMap) will refine targeting strategies, enabling personalized CRISPR interventions.

References I

DepMap, Broad Institute. (2025). *DepMap: The Cancer Dependency Map Project.* https://depmap.org/portal/

Krill-Burger et al. (2023). *Multi-omics approaches to refine CRISPR-based essentiality.*

NCI Staff. (2017). *CRISPR Gene-Editing Tool May Help Improve Cancer Immunotherapy.* cancer.gov Blog (2017)

NCI Staff. (2020). *How CRISPR Is Changing Cancer Research and Treatment.* NCI Blog (2020)

Ng, M. (2019). *Unleash the Power of CRISPR with Electroporation.* btxonline.com

Meyers, R. M. et al. (2017). *Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens. Nature Genetics* 49, 1779–1784.

Dempster, J. M. et al. (2021). *Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol.* 22, 343.

National Center for Biotechnology Information (NCBI). (n.d.) *DepMap Paper: phs003444.v2.p1.* Study Link (dbGaP)

Wu, T. et al. (2022). *Off-target effects in CRISPR/Cas9 gene editing.* PMC10034092

References II

Liu, Z. et al. (2024). *Recent Advancements in Reducing the Off-Target Effect of CRISPR-Cas9 Genome Editing.* PMC10802171

Anzalone, A. et al. (2022). *Cytosine base editing enables quadruple-edited allogeneic CART cells for T-ALL.* PMC9373016

Trabuchet, C. et al. (2024). *Editorial: Precision oncology in the era of CRISPR-Cas9 technology.* Front. Genet.

Smith, J. et al. (2024). *Enhancing precision in cancer treatment: the role of gene therapy and immune modulation in oncology.* Front. Med.

Park, M. et al. (2023). *Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review.* PMC10199778

Gonzalez, R. et al. (2023). *The Potential Revolution of Cancer Treatment with CRISPR Technology.* PMC10046289

Kim, S. et al. (2022). *Prime Editing: An Emerging Tool in Cancer Treatment.* PMC9574179

Garcia, N. et al. (2023). *Editorial: First Regulatory Approvals for CRISPR-Cas9 Therapeutic Gene Editing for Sickle Cell Disease and Transfusion-Dependent β-Thalassemia.* PMC10913280

Thank You!

Questions or Discussion?