Predicting Tumorigenic Potential of Human Cancer Cell Lines in Mice

Dominique Dang[1,2], Eiru Kim[1], Dean Lee[1]

[1]Oncology Data Science, Novartis Institutes for Biomedical Research, Cambridge, MA, USA; [2] Massachusetts Institute of Technology

Abstract

Cancer cell lines are foundational tools in preclinical research, yet many fail to form tumors in xenograft models, limiting their translational utility. To address this gap, bulk RNA sequencing data was integrated with experimentally validated tumorigenicity outcomes to identify transcriptomic features associated with successful xenograft formation. Differential expression analysis uncovered key genes distinguishing tumorigenic from non-tumorigenic lines. Using these features, we trained a logistic regression classifier that achieved a median test set accuracy of 0.84 (84% CI: [0.78, 0.90]) across 1,000 bootstrapped samples. From our model, this approach introduces a new dimension for prioritizing cell lines based on their likelihood of forming tumors in vivo. These findings offer a practical framework for enhancing model selection and improving the translational relevance of cancer research.

Background

- Cancer cell lines are useful cancer models in drug development, but they are grown in vitro and thus do not capture crosstalk between malignant and non-malignant cells
- Cell line-derived xenograft models can be used to better represent patient tumors in vivo
- However, many cell lines fail to induce tumorigenesis in these xenograft models
- We hypothesize that there are biomarkers shared across these cell lines that explain whether they do or do not form xenografts
- The dataset used includes 488 barcoded cancer cell lines that were injected into mice subcutaneously
- Tumor barcodes were quantified to compare their abundance pre- and post-injection, assessing tumorigenic potential (Jin et al., 2020)

Prevalence of Non-Engrafted Samples Compared to Engrafted Samples

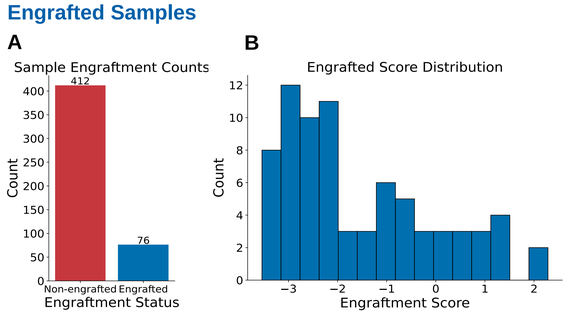
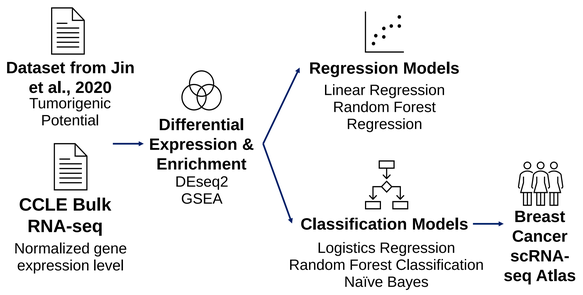


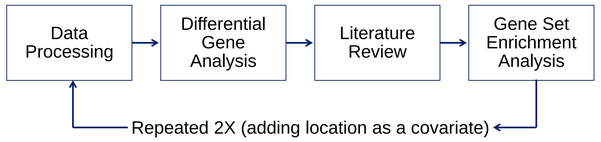
Figure 1: Distribution of Engraftment Scores. (A) Counts of non-engrafted (412) and engrafted (76) cell lines from the dataset. **(B)** The distribution of engraftment potential that did engraft, ranging from -3 to +2.

Methods

Graphical Summary



Differential Expression & Enrichment



Model Development

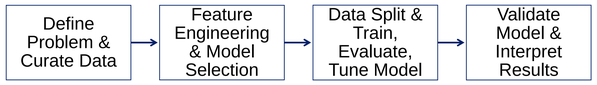


Figure 2. Summary of Methods

Results

Differentially expressed genes (DEGs) were identified in engrafted cell lines after controlling for location

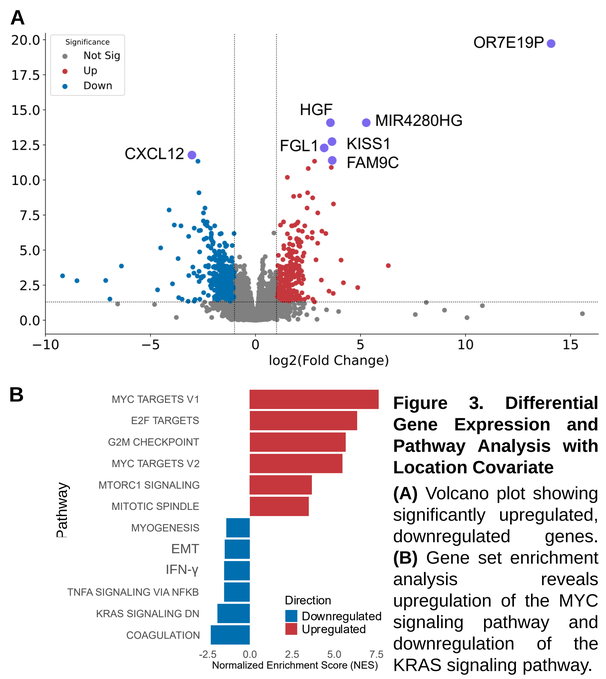


Figure 3. Differential Gene Expression and Pathway Analysis with Location Covariate

(A) Volcano plot showing significantly upregulated, downregulated genes. **(B)** Gene set enrichment analysis reveals upregulation of the MYC signaling pathway and downregulation of the KRAS signaling pathway.

Classification models perform better than regression models

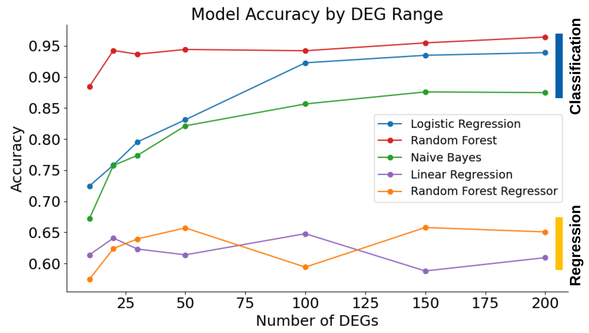


Figure 4. Model Accuracy by DEG Range. A classification target of engraftment yes/no and a regression target of engraftment potential were used with a threshold set at -3.86 (mean value of the training data). Linear Regression and Random Forest Regressor showed lower accuracy levels than the classification models.

Increasing the number of DEGs lead to random forest model overfitting

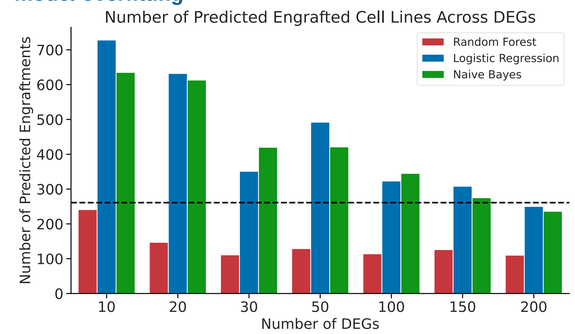


Figure 5. Predicted Engraftment of CCLE Cell Lines (n = 1,673) by DEG Range. The dotted line indicates the expected number of predicted engrafted lines (n = 261) based on the proportion of engrafted samples in the training set (15.6%, 76 out of 488 cell lines). Random Forest performance declines with increasing DEGs, suggesting overfitting to the training set.

Prediction scores for cell lines that successfully generated CDX models tend to be higher compared to those that failed

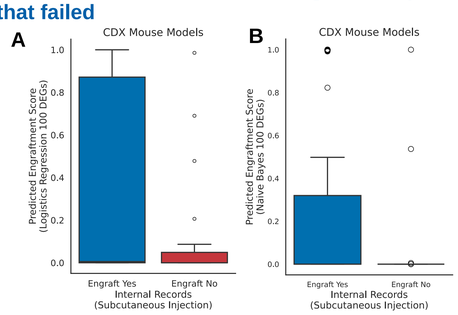


Figure 6. Boxplots of the validation of prediction scores for successful and failed CDX models using Novartis internal records for **(A)** Logistics Regression (p = 0.007) and **(B)** Naïve Bayes (p = 0.012) with 100 DEGs.

Engraftment scores vary by molecular subtype in breast cancer epithelial cells

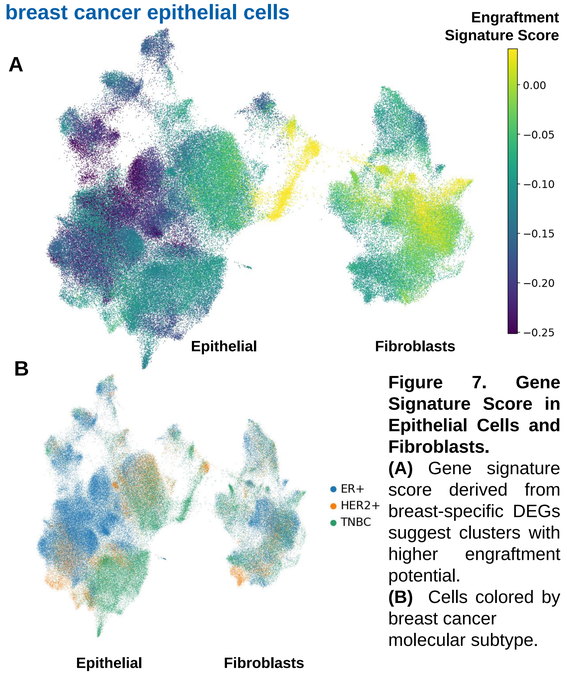


Figure 7. Gene Signature Score in Epithelial Cells and Fibroblasts.

(A) Gene signature score derived from breast-specific DEGs suggest clusters with higher engraftment potential.

(B) Cells colored by breast cancer molecular subtype.

Key Findings

- Engraftment success is associated with distinct gene patterns and influenced by tissue context
- Classification models showed promise and the Logistic Regression model with 100 DEGs was validated with internal records
- Epithelial cells from breast cancer patients have different engraftment scores across breast cancer molecular subtypes and patients
- Next steps include publishing engraftment-predicted cell lines and expand the model to other cancer types and applications, including patient-derived xenograft models

References

- Jin, X., Demere, Z., Nair, K., Ali, A., Ferraro, G. B., Natoli, T., Deik, A., Petronio, L., Tang, A. A., Zhu, C., Wang, L., Rosenberg, D., Mangena, V., Roth, J., Chung, K., Jain, R. K., Clish, C. B., Vander Heiden, M. G., & Golub, T. R. (2020). A metastasis map of human cancer cell lines. Nature, 588(7837), 331–336. https://doi.org/10.1038/s41586-020-2969-2
- Sun, H., Cao, S., Mashl, R. J., Mo, C.-K., Zaccaria, S., Wendl, M. C., Davies, S. R., Bailey, M. H., Primeau, T. M., Hoog, J., Mudd, J. L., Dean, D. A., Patidar, R., Chen, L., Wyczalkowski, M. A., Jayasinghe, R. G., Rodrigues, F. M., Terekhanova, N. V., Li, Y., & Lim, K.-H. (2021). Comprehensive characterization of 536 patient-derived xenograft models prioritizes candidates for targeted treatment. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-25177-3
- Zanella, E. R., Grassi, E., & Trusolino, L. (2022). Towards precision oncology with patient-derived xenografts. Nature Reviews Clinical Oncology, 19(11), 719–732. https://doi.org/10.1038/s41571-022-00682-6

Acknowledgments

I'd like to thank members of Oncology Data Science, especially both of my mentors – Eiru Kim and Dean Lee – for their guidance and for going above and beyond to support my learning and growth here at Novartis.

NOVARTIS | Reimagining Medicine

A SUMMER OF SCIENCE @ NOVARTIS