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] Department of Biology & Department of Computer Science & Electrical Engineering,

Massachusetts Institute of Technology

Novartis Institute of Biomedical Research Summer of Science Internship

[#] Corresponding author, Email: eiru.kim@novartis.com (E.K.), dean.lee@novartis.com (D.L.)

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.

Keywords: tumorigenesis, differential expression analysis, machine learning, scRNA-sequencing

Author's Note: This piece was written for the 2025 Summer of Science Internship at the Novartis Institutes for Biomedical Research.

Background

Preclinical models are important tools in cancer research, acting as a bridge between laboratory discoveries and clinical advancements. Among these models, cancer cell lines offer a widely accessible and reproducible platform for studying tumor biology and screening therapeutic compounds. However, their utility is limited by the environment of cell culture systems which fail to replicate the complexity of an in vivo tumor microenvironment. Factors such as the composition of basal media, absence of systemic signals like cytokines, and the lack of interactions between malignant and surrounding stromal or immune cells all contribute to inherent biases in cell line research. These shortcomings highlight the need for more representative systems in cancer modeling.

To address these limitations, researchers have utilized xenograft models, where human cancer cell lines or tumor samples are transplanted into immunocompromised mice. Xenograft models, including cell line-derived xenografts (CDX) and patient-derived (PDX), provide a more physiologically relevant system to study tumor biology and evaluate drug efficacy (Zanella et al., 2022). By mimicking the cellular and molecular dynamics of human tumors in vivo, these models provide valuable insights into disease mechanisms.

However, an important challenge persists: many cancer cell lines are unable to establish tumors when transplanted into xenograft mice. This phenomenon, which is observed across multiple cancer types, highlights fundamental gaps in our understanding of tumorigenesis in xenografts.

To address this, the goal of this project is to identify molecular and cellular biomarkers that can predict a cancer cell line's ability to establish xenografts. Such insights could guide

experimental modifications to improve the likelihood of engraftment, providing researchers with more reliable tools for studying cancer progression and therapeutic response.

Through a combination of transcriptomic profiling, enrichment analyses, predictive modeling, and comparative analyses with patient tumor data, we aim to investigate the determinants of engraftment in xenograft systems. First, we will identify transcriptomic differences between engrafting and non-engrafting cell lines, leveraging published datasets to find differentially expressed genes (DEGs) and enriched pathways. The Cancer Cell Line Encyclopedia (CCLE) from the DepMap resource provides transcriptomic profiles of hundreds of cancer cell lines and enables integrated exploration of gene expression patterns and functional genomic data. Additionally, we incorporated data from Jin et. al (2020), which provides experimental evidence linking engraftment potential with cancer cell lines.

Next, we aim to predict the engraftment potential of cell lines using computational approaches, such as machine learning methods, to nominate cell lines for experimental validation. Finally, we will extend these findings to human patient tumor scRNA-seq data, exploring whether subpopulations of malignant cells resemble the gene expression profiles of engrafting versus non-engrafting cell lines. This approach hopes to learn more about the molecular drivers of xenograft compatibility, inform future xenograft model development, and bridge the gap between preclinical models and patient-derived tumor biology.

Methods

Differential Expression & Enrichment

To investigate the transcriptomic differences between cancer cell lines with varying engraftment potentials in xenograft models, we utilized the dataset from Jin et al. (2020). The

study featured a dataset of 488 barcoded cancer cell lines, which were subcutaneously injected into immunodeficient mice to evaluate engraftment success. Barcode abundance was measured pre- and post-injection to evaluate engraftment success. This approach enabled a categorization of cell lines into engrafting and non-engrafting groups based on their ability to form tumors in vivo.

To address differential expression and enrichment analyses, raw RNA-seq read counts was obtained from the Cancer Cell Line Encyclopedia (CCLE). Specifically, the dataset used was the CCLE RNA-seq gene counts file (CCLE_RNAseq_genes_counts_20180929.gct). Using the pyDESeq2 v0.5.2 Python package (Muzellec et al., 2023), we performed differential gene expression analysis between the two groups of cell lines. The tool enables analysis of bulk RNA-seq data using methods adapted from the R DESeq2 package. After identifying DEGs, a literature review was conducted to contextualize our findings and verify the potential relevance of the observed gene expression differences in the context of tumorigenesis.

We proceeded with gene set enrichment analysis (GSEA) using the fgsea v3.21 package in R (Korotkevich, 2019) to determine which biological pathways were significantly enriched among the DEGs. The human Hallmark Gene Set was used for this analysis and was obtained from the Human MSigDB Collections (Liberzon et al., 2015). Both data processing and differential expression analyses were repeated twice, initially using the base dataset and subsequently incorporating location as a covariate.

Model Development

We leveraged five statistical and machine learning techniques to predict the engraftment potential of cell lines and classify cell lines based on observed gene expression profiles. First, we

performed feature engineering by filtering the DEGs used for training the models based on their average TPM (transcripts per million) values, retaining only genes with average TPM > 1 to ensure biologically meaningful features. The TPM values were obtained from the Cancer Cell Line Encyclopedia (OmicsExpressionProteinCodingGenesTPMLogp1BatchCorrected.csv).

Next, we selected appropriate models for regression and classification tasks, including Linear Regression, Random Forest Regression, Logistics Regression, Random Forest Classification and Naïve Bayes. The scikit-learn v1.7 library was used to train these models (Pedregosa, 2011) and the curated data was divided into training and validation sets using an 80/20 train-test split. Each model was trained on the training set and validated for its performance on the validation set, with hyperparameter tuning conducted where necessary to optimize predictive power.

To evaluate model performance, we utilized various metrics to ensure accuracy and reliability. Specifically, we compared overall accuracy between models and visualized confusion matrices to assess the proportion of correct classifications versus misclassifications. Additionally, receiver operating characteristic (ROC) curves were generated to evaluate the models' ability to distinguish between classes and to compute the area under the curve (AUC) for further comparative analysis of predictive performance (Extended Figure 2, 3). These metrics provided a comprehensive overview of model performance for both regression and classification tasks.

Breast Cancer scRNA-seq Atlas

We analyzed single-cell RNA sequencing (scRNA-seq) data from human breast cancer tumors derived from three studies (Bassez et al., 2021, Subramanian et al., 2005, Qian et al., 2020).

The genes identified from the logistic regression model were applied to single-cell RNA sequencing (scRNA-seq) data from human breast cancer tumors to identify malignant epithelial subpopulations with varying engraftment potential. We identified genes with positive and negative coefficients from our logistics regression model. Using these gene sets, we calculated signature scores using computational tools from the Scanpy library v1.1 to assess how well the transcriptional profiles of the cells align with specific phenotypes such as high or low engraftment capacity.

First, the CCLE log-transformed TPM dataset was preprocessed to ensure only genes with TPM values > 1 were retained and batch effects were corrected for comparability across samples. We employed the `sc.tl.score_genes` function from Scanpy (Wolf et al., 2018) to calculate signature scores. These scores were computed for each gene set (positive and negative) across all cells using the specified expression layer (`layer="log2_couns_scvi"`) from an AnnData object.

A positive marker score represents the activity level of genes associated with high engraftment capacity in each cell while a negative marker score represents the activity level of genes associated with low engraftment capacity in each cell. The ratio (difference) between the positive and negative scores were computed cell by cell which quantifies the balance between the transcriptional signals supporting high vs. low engraftment.

To assess the statistical relevance of the observed signatures, a random control was generated. A vector of random ‘positive-negative’ score differences was created by repeatedly (20 iterations) selecting random subsets of the gene set and computing their difference scores. This control helps account for biases introduced by gene set size, variability, or technical noise.

Using the computed signature scores, UMAPs were plotted ('sc.pl.umap') to visualize the actual gene set ratios and the average random difference.

This approach captured both inter-tumor (between patients) and intra-tumor (within a single tumor) heterogeneity, providing deeper insights into the cellular states and molecular programs most likely to drive tumorigenesis.

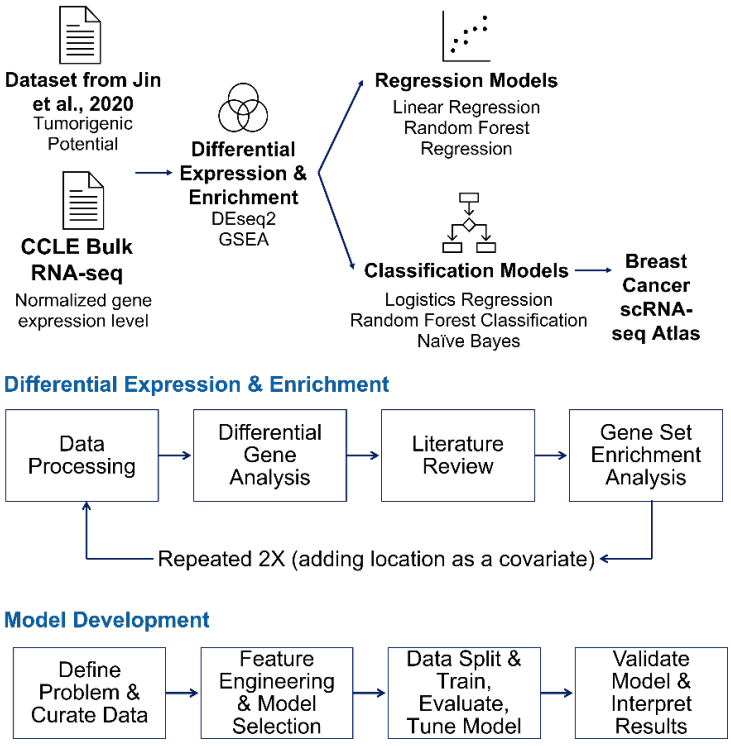


Figure 2. Graphical summary of methods

Results

The distribution of engrafted and non-engrafted cell lines from the Jin et al. (2020) dataset is illustrated in Figure 1, which also depicts the measured engraftment potential of the 76 cell lines that successfully formed tumors.

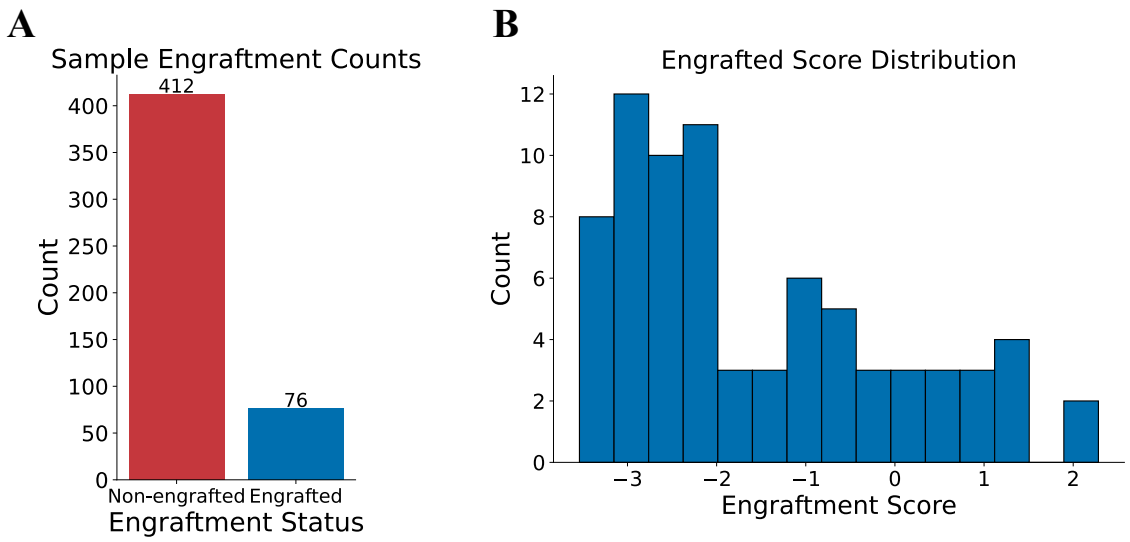


Figure 1: Distribution of Engraftment Scores. **(A)** Counts of non-engrafted and engrafted cell lines from the dataset. **(B)** The distribution of engraftment potential that did engraft, ranging from -3 to +2.

Differential Expression & Enrichment

The analysis of differential gene expression and gene set enrichment has yielded insights into the molecular biomarkers associated with the engraftment potential of cancer cell lines. These findings are presented as a volcano plot seen in Figure 3, which incorporates a location covariate, and in Extended Figure 1, where the covariate is excluded. The location covariate was added to account for potential variability that could influence gene expression patterns and engraftment outcomes.

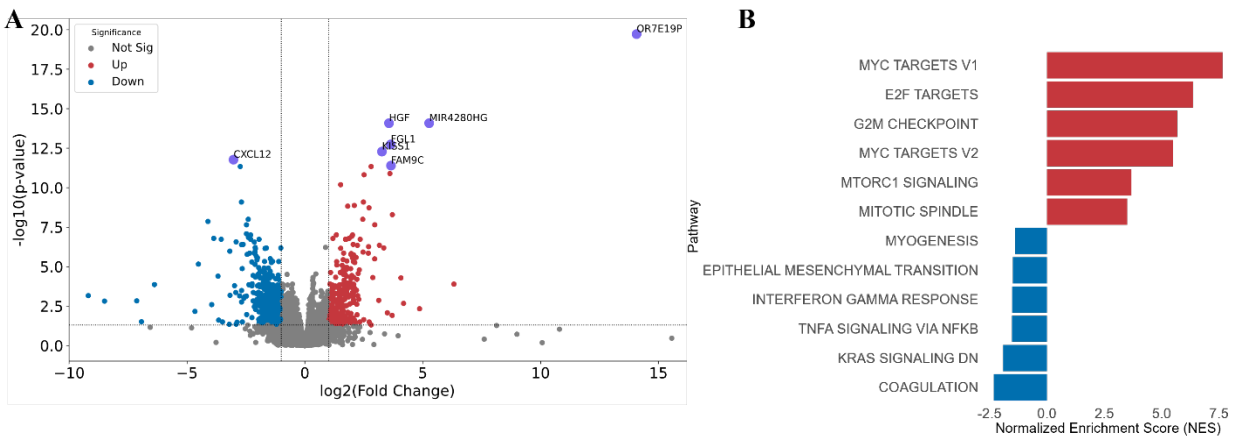


Figure 3. Differential Gene Expression and Pathway Analysis with Location Covariate (A) Volcano plot showing significantly upregulated and downregulated genes. **(B)** Gene set enrichment analysis reveals upregulation of the MYC signaling pathway and downregulation of the KRAS signaling pathway.

Specific genes such as CXCL12, HGF, FGL1, KISS1, OR7E19P, MIR4280HG, and FAM9C were significantly upregulated or downregulated. HGF, KISS1, FAM9C, OR7E19P, and MIR4280HG were identified as significantly upregulated genes in cell lines with higher engraftment potential. HGF, in combination with insulin-like growth factor 2 (IGF2), has been reported to drive tumorigenesis in epithelial ovarian cancer (Chu et al., 2023). Upregulation of KISS1 signaling has been shown to promote tumor growth in estrogen receptor-negative breast cancer (TNBC) and hepatocellular carcinoma (Dragan et al., 2020). Similarly, FAM9C is frequently overexpressed in human hepatocellular carcinoma (Zhou et al., 2013). Conversely, CXCL12 was markedly downregulated in cell lines with higher engraftment potential. Reduced CXCL12 expression is a well-established feature of acute myeloid leukemia (AML) and other hematopoietic neoplasms. (Wang et al., 2021). To date, there is no published literature implicating OR7E19P or MIR4280HG as contributors to tumorigenesis.

Gene Set Enrichment Analysis

The results of gene set enrichment analysis (GSEA), showcasing the normalized enrichment scores (NES) for various pathways, are highlighted in Figure 3B. This analysis underscores the functional pathways activated or suppressed in cell lines with different engraftment potential. Pathways such as MYC targets V1 and V2, E2F targets, G2M checkpoint, and mitotic spindle are highly enriched in cell lines that successfully engraft. These pathways are associated with cell proliferation, survival, and growth which are critical components of tumorigenesis. Conversely, pathways including myogenesis, epithelial-to-mesenchymal transition (EMT), interferon gamma response, TNFα signaling via NF-kB, KRAS signaling, and coagulation are among the most suppressed. These pathways are related to differentiation, inflammatory responses, and cell migration, which might inhibit tumor growth in xenografts.

Model Development

Using the differentially expressed genes identified from the previous analysis, various machine learning approaches were employed to predict the engraftment potential of cancer cell lines. We initially evaluated both classification and regression frameworks to determine the optimal modeling strategy.

The accuracy of various classification and regression models across a range of DEGs is presented in Figure 4. The classification models (Logistic Regression, Random Forest, and Naive Bayes) consistently exhibited superior performance compared to their regression counterparts. Logistic Regression and Random Forest maintained accuracy above 0.85 across most DEG sets, while Naive Bayes demonstrated stable performance exceeding 0.80. In contrast, Linear Regression and Random Forest Regressor showed worse performance, with weaker accuracy and instability seen with fluctuating scores across DEG ranges. These results indicated that

classification approaches were better suited for predicting engraftment potential than regression-based methods.

Building on these findings, we next optimized feature selection and compared classification model performance. After the models were trained, they were tested on all cancer cell lines in the CCLE (n = 1673). The effect of increasing the number of DEGs on model performance is shown in the data (Figure 5), with the number of predicted engrafted cell lines plotted against the number of DEGs used. As feature dimensionality increased, the Random Forest classifier exhibited signs of overfitting when DEG counts exceeded 100, seen by declining performance and increasingly unstable predictions with larger feature sets. Conversely, both Logistic Regression and Naive Bayes models maintained stable predictions across varying numbers of DEGs. The choice of 100 DEGs as our optimal feature set was determined when observing that model accuracy plateaued at approximately this threshold for both Logistic Regression and Naive Bayes classifiers (Supplementary Table 1). Additional features beyond 100 DEGs provided diminishing returns in predictive power while introducing the risk of overfitting. The list of predictions from the Logistics Regression and Naïve Bayes with 100 DEGs models are seen in Supplementary Table 2.

Model performance was validated using an independent dataset of 200 cancer cell lines from Novartis internal records, comprising 174 cell lines that successfully established CDX models and 26 that failed to engraft. The comparison of prediction scores between successful and failed engraftment cases using the Logistic Regression and Naive Bayes models with 100 DEGs is shown in Figure 6. In Figure 6A, the Logistic Regression model yields significantly higher prediction scores for cell lines that successfully induced CDX models (Engraft Yes) compared to those that failed (Engraft No; $p = 0.007$), supporting its reliability in distinguishing engraftment

potential. Analogous results for the Naïve Bayes model are seen in Figure 6B, which similarly differentiates successful engraftments with elevated scores ($p = 0.012$). These findings demonstrate that both models, when trained on 100 DEGs, can reliably predict engraftment potential in an independent validation cohort.

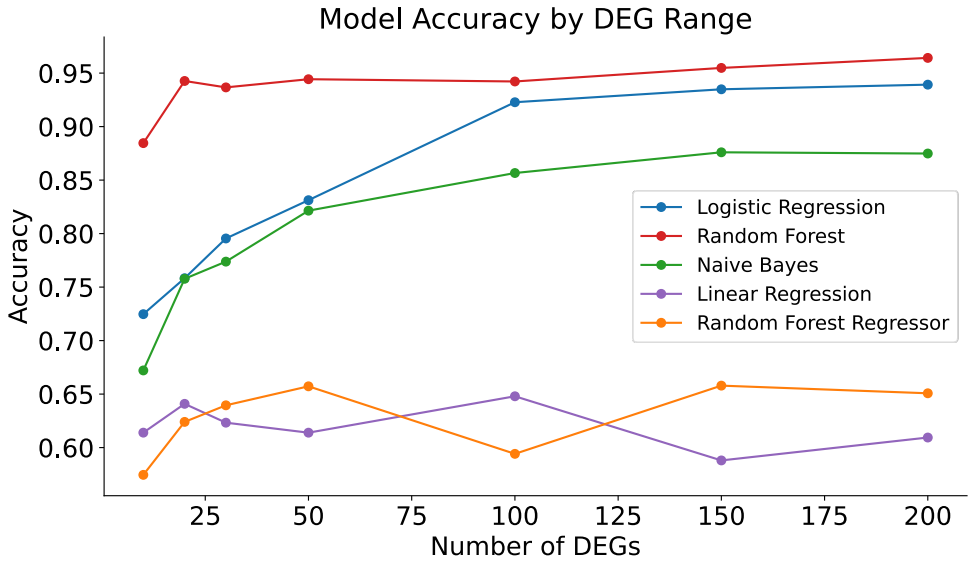


Figure 4. Model Accuracy by DEG Range. A classification target of engraftment yes/no and a regression target of engraftment potential were used with a threshold set at -3.86 (mean value of the training data).

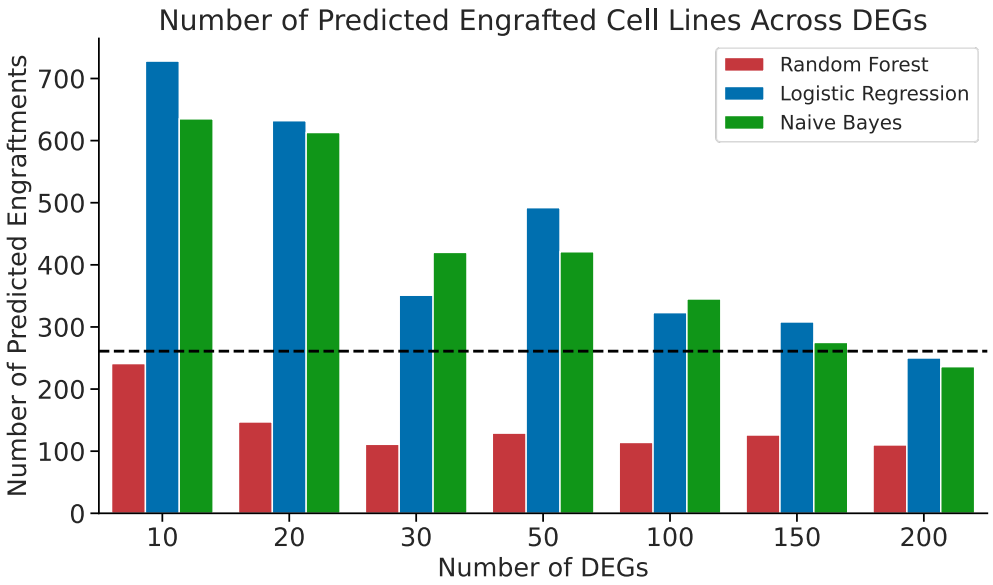


Figure 5. Predicted Engraftment of CCLE Cell Lines (n = 1,673) by DEG Range. The dotted line indicates the expected number of predicted engrafted lines (n = 261) based on the proportion of engrafted samples in the training set (15.6%, 76 out of 488 cell lines).

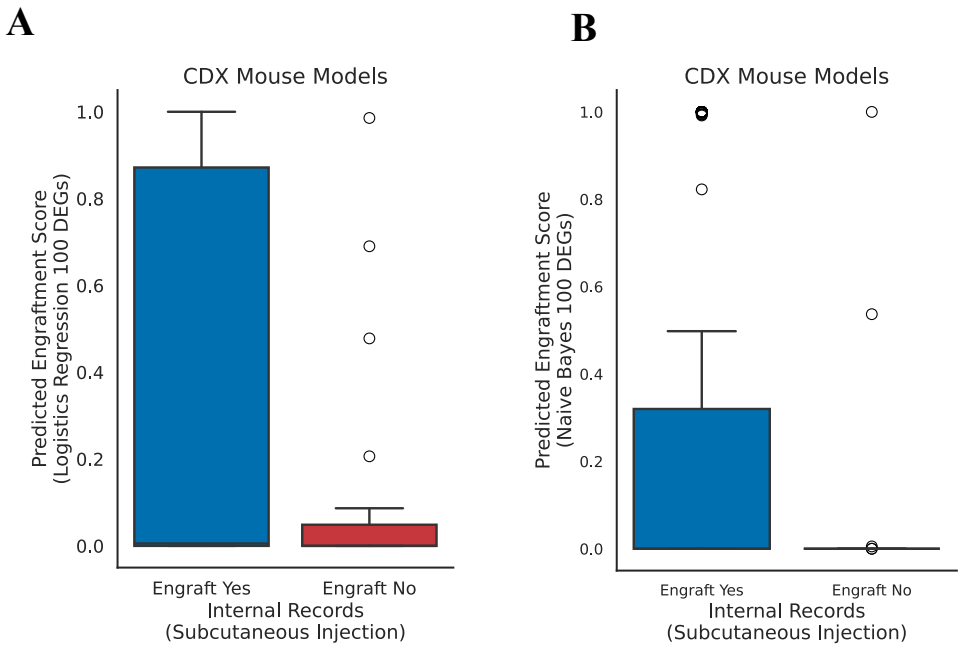


Figure 6. Boxplots of the validation of prediction scores for successful and failed CDX models using Novartis internal records for (A) Logistics Regression (p = 0.007) and (B) Naïve Bayes (p = 0.012) with 100 DEGs.

Breast Cancer scRNA-seq Atlas

In the final part of our project, we leveraged single-cell RNA sequencing (scRNA-seq) data to directly examine patient tumor samples, with a focus on identifying subpopulations of malignant cells that exhibit varying engraftment signature scores. This analysis was particularly aimed at exploring the engraftment potential across different molecular subtypes of breast cancer epithelial cells (Figure 7A).

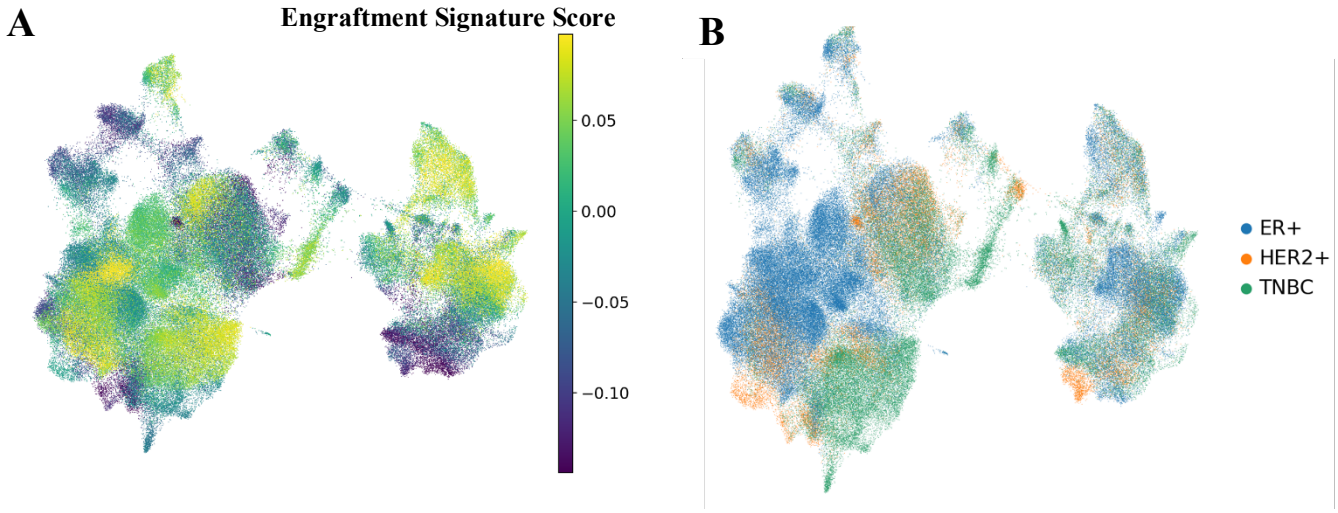


Figure 7. Gene Signature Score in Epithelial Cells and Fibroblasts. (A) Gene signature score derived from filtered DEGs suggest clusters with higher engraftment potential. **(B)** Cells colored by breast cancer molecular subtype.

Discrete epithelial clusters were observed with elevated scores, suggesting that tumorigenic programs are not uniformly expressed but confined to transcriptionally distinct cellular states. A subset of fibroblasts also exhibits higher scores, which could reflect stromal programs that facilitate engraftment. The same manifold colored by breast cancer subtypes is shown in Figure 7B, however the engraftment potential does not clearly correlate with any specific subtype.

These findings suggest that engraftment potential in breast cancer xenograft models is characterized by specific cellular clusters rather than distinct molecular subtypes. Our scRNA-seq analysis provided critical insights into the cellular heterogeneity within patient tumors and identified specific subpopulations of epithelial cells with high engraftment potential. These findings pave the way for several future directions such as validating high-potential epithelial clusters in xenograft models and applying these methodologies to scRNA-seq data from other cancer types to explore subpopulation variations in engraftment potential. Additional scRNA-seq findings can be integrated with patient clinical outcomes to enhance the relevance of engraftment signatures in therapeutic and prognostic applications.

Discussion

This study provides novel insights into the molecular determinants of xenograft compatibility through integrated transcriptomic analysis and machine learning approaches. Our findings advance understanding of tumor engraftment in three key areas: identification of predictive biomarkers, development of robust classification models, and characterization of cellular heterogeneity in engraftment scores in patient tumors.

Our differential expression analysis revealed distinct transcriptomic signatures that distinguish engrafting from non-engrafting cancer cell lines. The upregulation of genes such as HGF, KISS1, and FAM9C in successfully engrafting lines aligns with established roles of these factors in promoting tumorigenesis and cell survival. Conversely, the downregulation of CXCL12 in engrafting lines is consistent with its known association with poor prognosis in hematologic malignancies. The enrichment of MYC signaling pathways and cell cycle progression programs in engrafting lines further supports the concept that proliferative capacity is a critical determinant of xenograft success, while the suppression of differentiation and

inflammatory response pathways may facilitate tumor establishment in the immunocompromised environment.

The superior performance of classification over regression approaches emonstrates that engraftment potential is best conceptualized as a binary outcome rather than a continuous variable. Our systematic optimization identified 100 DEGs as the optimal feature set, balancing predictive accuracy with biological interpretability while avoiding overfitting. The statistical significance of model predictions in distinguishing successful from failed engraftments ($p < 0.05$ for both Logistic Regression and Naive Bayes) from internal Novartis data provides confidence in the clinical utility of this approach. These models offer researchers a data-driven framework for prioritizing cell lines most likely to succeed in xenograft studies, potentially reducing experimental costs and improving research efficiency.

The application of our engraftment signatures to single-cell breast cancer data revealed notable heterogeneity in predicted engraftment potential, an observation that underscores the complexity of tumor biology. We found that specific epithelial clusters exhibited enriched high engraftment scores, which was independent of molecular subtype classification. This aligns with clinical observations of differential xenograft take rates and suggests that our transcriptomic signatures effectively capture biologically relevant differences in tumorigenic potential that translate from cell line models to patient tumor biology.

Several limitations warrant consideration. Our reliance on a single primary dataset (Jin et al.) for model training may introduce dataset-specific biases and limit generalizability across different experimental conditions. The significant class imbalance (15.6% engraftment rate) in our training data poses inherent challenges for minority class prediction, though our external validation suggests adequate model performance despite this limitation. Our transcriptomic

approach, while comprehensive, captures only gene expression patterns and does not account for post-translational modifications, epigenetic factors, or metabolic states that may critically influence engraftment success.

The immediate translational value of this work lies in providing the research community with validated tools for cell line selection in xenograft studies. The methodological framework developed here can be readily extended to other cancer types. Such expansion would enhance the utility of predictive modeling across diverse tumor types and provide insights into tissue-specific mechanisms of tumorigenesis.

Conclusion

This study successfully identified molecular biomarkers that determine cancer cell line engraftment potential in xenograft models through integrated transcriptomic analysis and machine learning. Our differential expression analysis revealed distinct gene signatures associated with successful engraftment, including upregulation of proliferative pathways and downregulation of differentiation programs. The Logistic Regression classifier, optimized with 100 differentially expressed genes, achieved robust predictive performance with external validation confirming its reliability across 200 independent cell lines.

Extending these findings to patient tumor biology, our single-cell RNA sequencing analysis demonstrated that engraftment signatures capture clinically relevant heterogeneity across epithelial cells. This translation from cell line models to patient samples validates the biological relevance of our approach and suggests broader applicability to understanding tumor biology.

The practical impact of this work lies in providing the cancer research community with validated, data-driven tools for optimizing xenograft model selection. By enabling researchers to prioritize cell lines with higher likelihood of successful engraftment, our approach promises to improve experimental efficiency, reduce costs, and enhance the translational relevance of preclinical studies. The framework established here creates a foundation for extending predictive modeling to other cancer types and patient-derived xenograft systems, ultimately advancing our understanding of tumorigenesis and improving cancer therapeutic development.

References

Arafeh, R., Shibue, T., Dempster, J.M. *et al.* The present and future of the Cancer Dependency Map. *Nat Rev Cancer* 25, 59-73 (2025). https://doi.org/10.1038/s41568-024-00763-x

Baron, M., Tagore, M., Wall, P., Zheng, F., Barkley, D., Yanai, I., Yang, J., Kiuru, M., White, R. M., & Ideker, T. (2025). Desmosome mutations impact the tumor microenvironment to promote melanoma proliferation. *Nature Genetics*, *57*(5), 1179–1188. https://doi.org/10.1038/s41588-025-02163-9

Bassez, A., Vos, H., Van Dyck, L., Floris, G., Arijs, I., Desmedt, C., Boeckx, B., Vanden Bempt, M., Nevelsteen, I., Lambein, K., Punie, K., Neven, P., Garg, A. D., Wildiers, H., Qian, J., Smeets, A., & Lambrechts, D. (2021). A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer. *Nature Medicine*, *27*(5), 820–832. https://doi.org/10.1038/s41591-021-01323-8

Chu, T., Aye Aye Khine, Wu, N. Y., Chen, P., Chu, S., Lee, M., & Huang, H. (2023). Insulin-like growth factor (IGF) and hepatocyte growth factor (HGF) in follicular fluid cooperatively promote the oncogenesis of high-grade serous carcinoma from fallopian tube epithelial cells: Dissection of the molecular effects. *Molecular Carcinogenesis*, *62*(9), 1417–1427. https://doi.org/10.1002/mc.23586

Dragan, M., Nguyen, M.-U., Guzman, S., Goertzen, C., Brackstone, M., Dhillo, W. S., Bech, P. R., Clarke, S., Abbara, A., Tuck, A. B., Hess, D. A., Pine, S. R., Zong, W.-X., Wondisford, F. E., Su, X., Babwah, A. V., & Bhattacharya, M. (2020). G protein-coupled kisspeptin receptor induces metabolic reprograming and tumorigenesis in estrogen receptor-

negative breast cancer. *Cell Death and Disease*, *11*(2). https://doi.org/10.1038/s41419-020-2305-7

DepMap, Broad (2025). DepMap Public 25Q2. Dataset. depmap.org

Ilie, M., Nunes, M., Blot, L., Hofman, V., Elodie Long-Mira, Butori, C., Selva, E., Merino-Trigo, A., Vénissac, N., Jérôme Mouroux, Vrignaud, P., & Hofman, P. (2014). Setting up a wide panel of patient-derived tumor xenografts of non–small cell lung cancer by improving the preanalytical steps. *Cancer Medicine*, *4*(2), 201–211. https://doi.org/10.1002/cam4.357

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

Korotkevich G, Sukhov V, Sergushichev A (2019). "Fast gene set enrichment analysis." *bioRxiv*. doi:10.1101/060012, http://biorxiv.org/content/early/2016/06/20/060012.

Li, H., Zhu, Y., Tang, X., Li, J., Li, Y., Zhong, Z., Ding, G., & Li, Y. (2015). Integrated Analysis of Transcriptome in Cancer Patient-Derived Xenografts. *PLOS ONE*, *10*(5), e0124780. https://doi.org/10.1371/journal.pone.0124780

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, Jill P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Pal, B., Chen, Y., Vaillant, F., Capaldo, B. D., Joyce, R., Song, X., Bryant, V. L., Penington, J. S., Di Stefano, L., Tubau Ribera, N., Wilcox, S., Mann, G. B., kConFab,

Papenfuss, A. T., Lindeman, G. J., Smyth, G. K., & Visvader, J. E. (2021). A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO Journal*, *40*(11), e107333. https://doi.org/10.15252/embj.2020107333

Qian, J., Olbrecht, S., Boeckx, B., Vos, H., Laoui, D., Etlioglu, E., Wauters, E., Pomella, V., Verbandt, S., Busschaert, P., Bassez, A., Franken, A., Bempt, M. V., Xiong, J., Weynand, B., van Herck, Y., Antoranz, A., Bosisio, F. M., Thienpont, B., & Floris, G. (2020). A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research*, *30*(9), 745–762. https://doi.org/10.1038/s41422-020-0355-0

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene Set Enrichment analysis: a knowledge-based Approach for Interpreting genome-wide Expression Profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

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

Virginie Dangles-Marie, Pocard, M., Richon, S., Louis-Bastien Weiswald, Franck Assayag, Saulnier, P., Jean-Gabriel Judde, Jean-Louis Janneau, Auger, N., Validire, P., Dutrillaux, B., Françoise Praz, Bellet, D., & Marie-France Poupon. (2007). Establishment of

Human Colon Cancer Cell Lines from Fresh Tumors versus Xenografts: Comparison of Success Rate and Cell Line Features. *Cancer Research*, *67*(1), 398–407. https://doi.org/10.1158/0008-5472.can-06-0594

Wang, S., Xu, Z., Jin, Y., Ma, J., Xia, P., Wen, X., Mao, Z., Lin, J., & Qian, J. (2021). Clinical and prognostic relevance of *CXCL12* expression in acute myeloid leukemia. *PeerJ*, *9*(9), e11820–e11820. https://doi.org/10.7717/peerj.11820

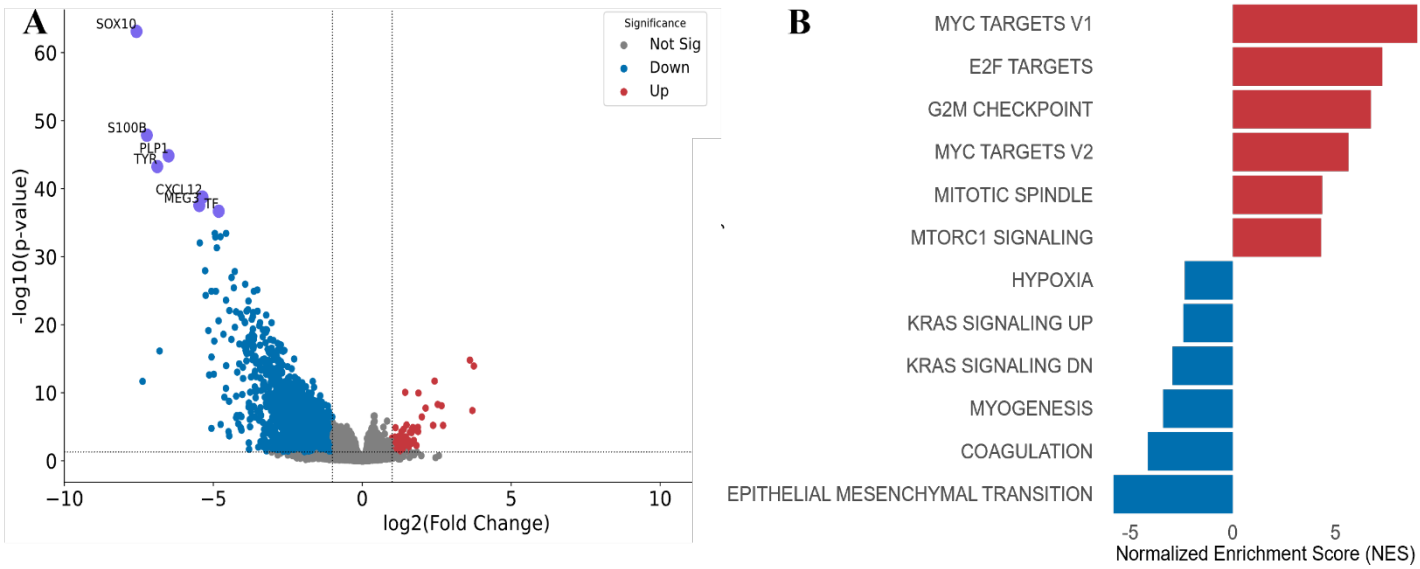
Wolf, F., Angerer, P. & Theis, F. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0 Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Xing Yi Woo, Giordano, J., Srivastava, A., Zhao, Z.-M., Lloyd, M. W., Roebi de Bruijn, Suh, Y.-S., Patidar, R., Chen, L., Scherer, S. D., Bailey, M. H., Yang, C.-H., Cortes-Sanchez, E., Xi, Y., Wang, J., Jayamanna Wickramasinghe, Kossenkov, A. V., Rebecca, V. W., Sun, H., & R. Jay Mashl. (2021). Conservation of copy number profiles during engraftment and passaging of patient-derived cancer xenografts. *Nature Genetics*, *53*(1), 86–99. https://doi.org/10.1038/s41588-020-00750-6

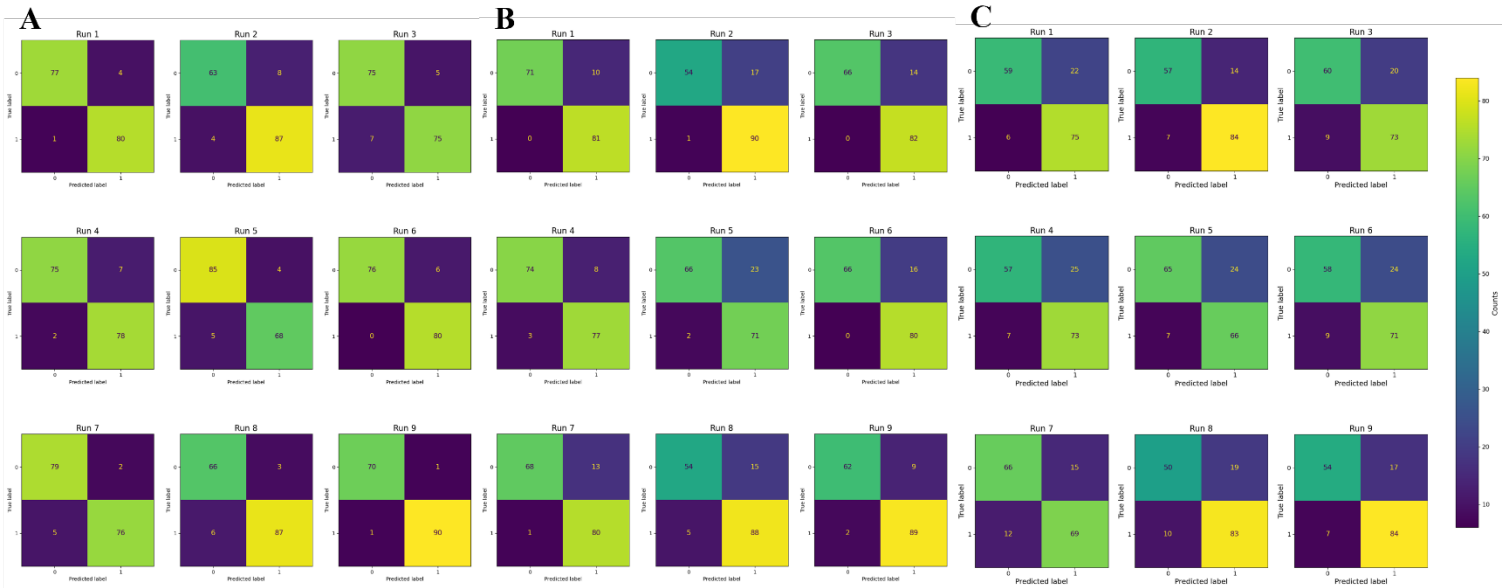
Zanella, E. R., Grassi, E., & Trusolino, L. (2022). Towards precision oncology with patient-derived xenografts. *Nature Reviews Clinical Oncology*, *19*(11), 719–732. https://doi.org/10.1038/s41571-022-00682-6

Zhou, J.-D., Shen, F., Ji, J.-S., Zheng, K., Huang, M., & Wu, J.-C. (2013). FAM9C plays an anti-apoptotic role through activation of the PI3K/Akt pathway in human hepatocellular carcinoma. *Oncology Reports*, *30*(3), 1275–1284. https://doi.org/10.3892/or.2013.2592

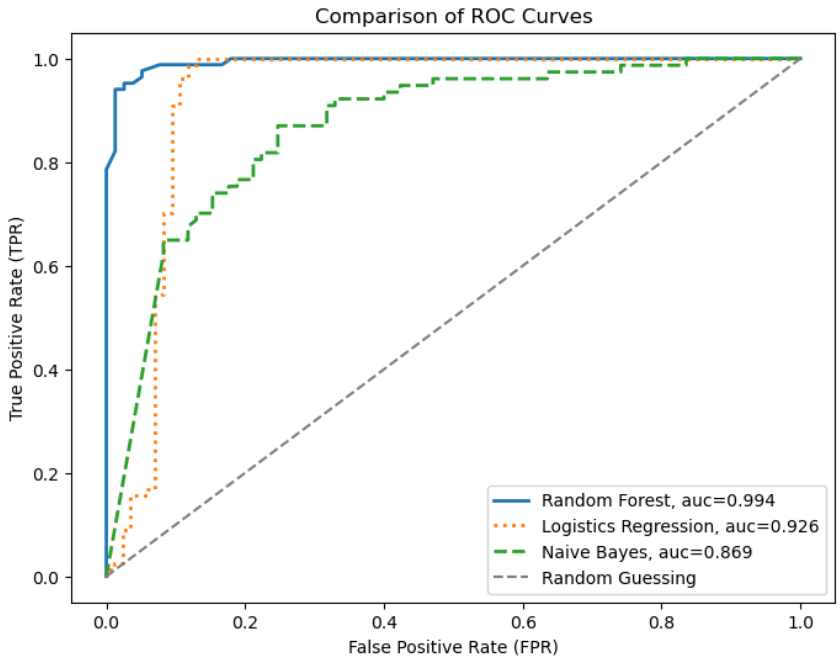
Extended Figures



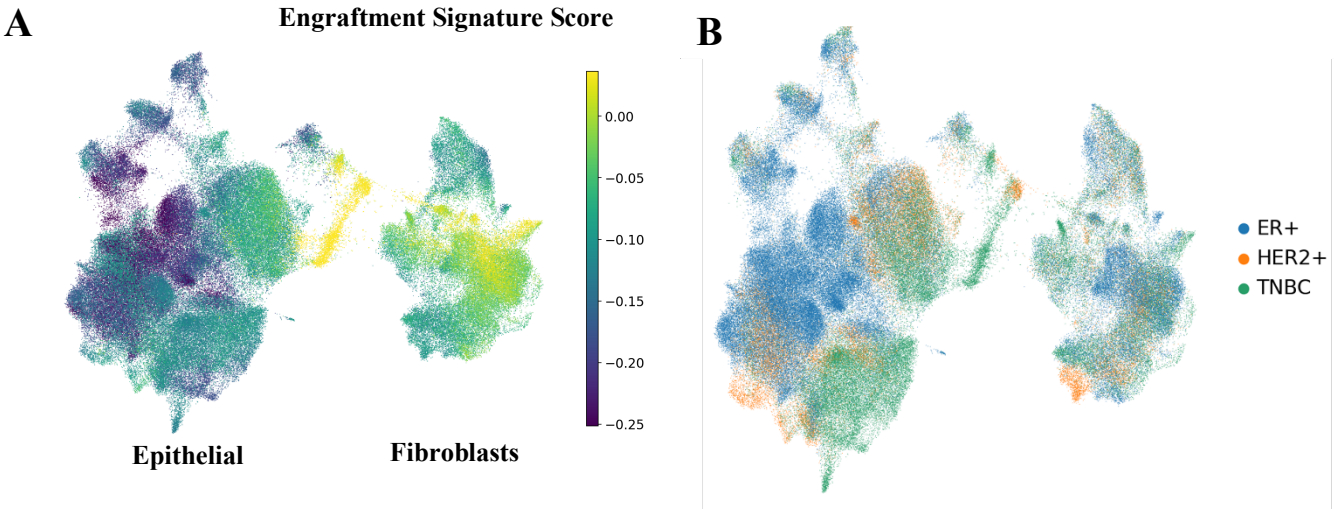
Extended Figure 1. Differential Gene Expression and Pathway Analysis (without location as a covariate) (A) Volcano plot showing significantly upregulated, downregulated genes. **(B)** Gene set enrichment analysis reveals upregulation of the MYC signaling pathway and downregulation of the EMT pathway.



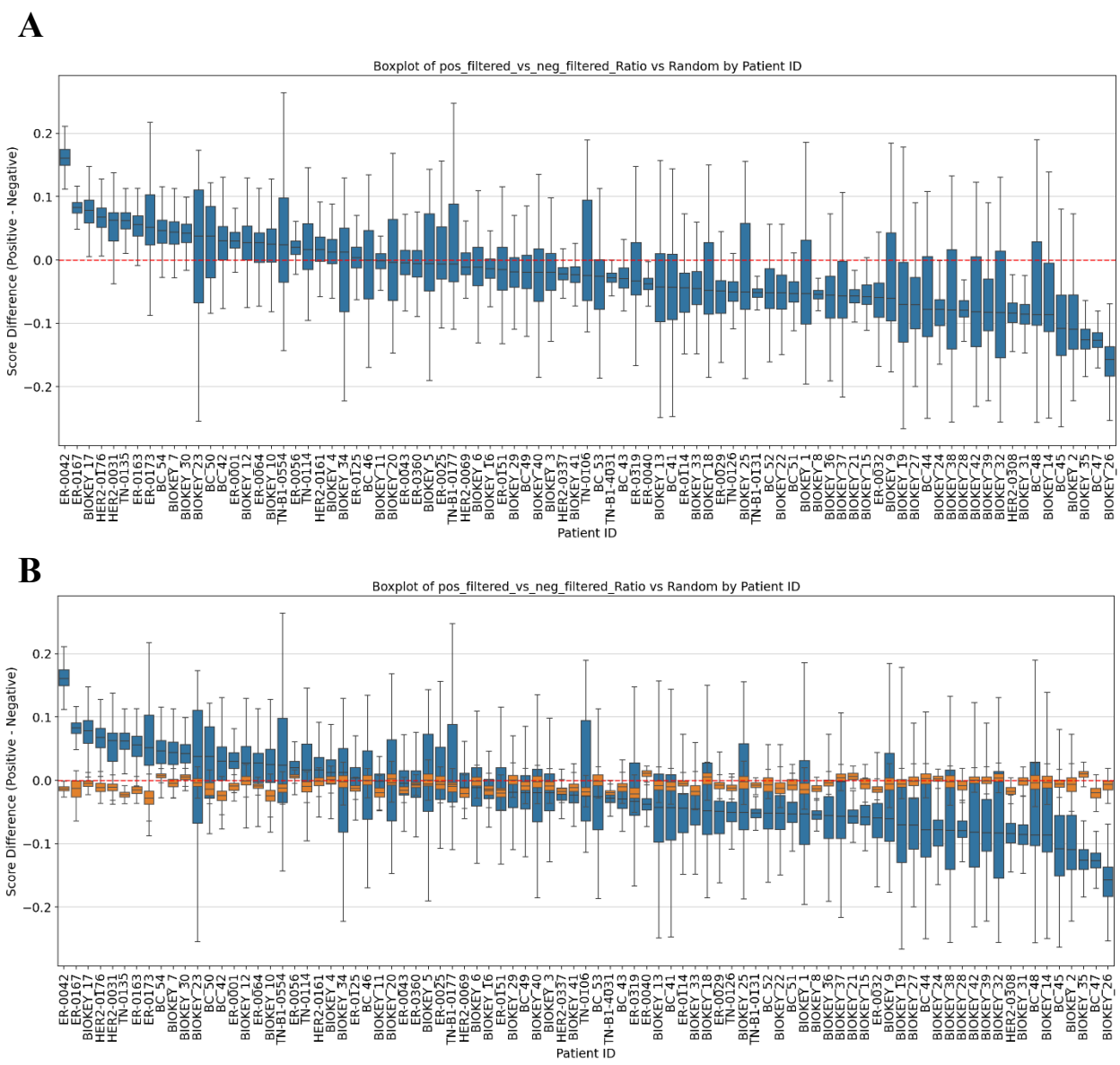
Extended Figure 2. Confusion Matrices of **(A)** Random Forest, **(B)** Logistics Regression, and **(C)** Naïve Bayes models



Extended Figure 3. ROC Curves with AUC values of **(A)** Random Forest, **(B)** Logistics Regression, and **(C)** Naïve Bayes models



Extended Figure 4. 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.



Extended Figure 5. Boxplot of Engraftment Signature Score by Patient ID **(A)** derived from filtered DEGs and **(B)** with random control.