Reconstructing a latent representation of gene expression from genomic alterations to improve clinical utility of real-world clinicogenomics data

ZEPHYR AI

Sunil Kumar, Felicia Kuperwaser, Dillon Tracy, Jeff Sherman, Emily Vucic and Maayan Baron

Zephyr AI: 1800 Tysons Blvd Suite 901, McLean VA 22102 | www.zephyrai.bio

Abstract # 3519

BACKGROUND

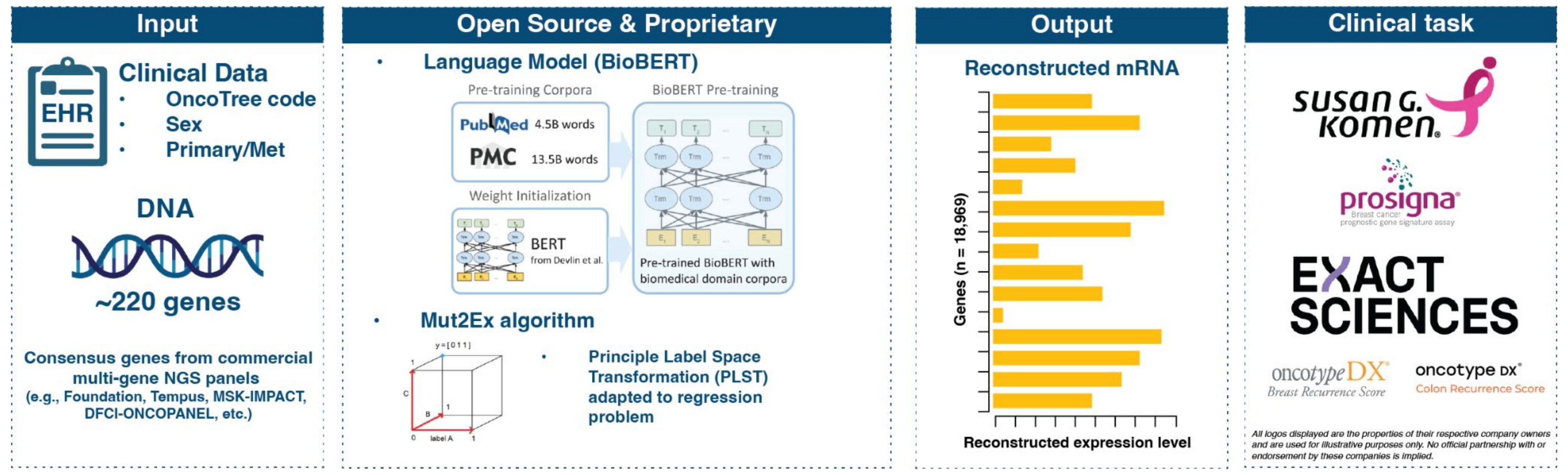
- Patient datasets with clinical and molecular information are ideal for studying tumor biology and developing robust machine learning (ML) models for predicting outcome and treatment response. These data however rarely exist in real-world settings or in sufficient quantities within research contexts.
- Large publicly available datasets like The Cancer Genome Atlas (TCGA), which provide multi-omic profiles for diverse cancer types, have greatly facilitated development of novel therapies and personalized medicines. However, the absence of patient outcome data tied to treatment limits the applicability of these data for understanding and modeling treatment response.
- Real-world clinicogenomics cohorts, such as the AACR Project GENIE, on the other hand are typically very rich in clinical annotations, including treatment regimens and outcomes measures. These data, however, are sparsely annotated for patient tumor molecular profiles, rarely exceeding ~100's of genes profiled.
- We hypothesized that it would be possible to reconstruct latent tumor mRNA representations from limited genomic and clinical data available in real-world clinicogenomic cohorts, and that these reconstructed expression profiles would be useful for a variety of clinically meaningful downstream applications.

METHODS

We developed an ML model (Mut2Ex) to reconstruct tumor gene expression profiles using genetic information available on commercial next generation sequencing panels using a regression-adapted Principle Label Space Transformation (PLST), along with embeddings from minimal clinical information (OncoTree code, sex and stage) generated by a language model (**Fig. 1**). Mut2Ex was trained on ~1200 DepMap cell lines across 26 cancer types to reconstruct whole transcriptome mRNA expression profiles. These profiles were generated for ~10,000 tumors from TCGA and ~180,000 tumors from AACR Project GENIE and applied to a variety of clinical tasks.

RESULTS

- Reconstructed mRNA expression by Mut2Ex was highly correlated with true expression in cell lines (r = 0.9342, [0.9328-0.9357, 95% CI, N=164]). Compared to true expression, reconstructed profiles recapitulate sub-clusters within cancer types, PAM50 subtyping in breast tumors, survival signatures in colorectal tumors and multiple oncogenic signatures in a pan-cancer manner.
- Analysis of reconstructed expression for AACR Project GENIE tumors revealed expected enrichment of known driver genes within expression subtypes and enrichment of oncogenic signatures associated with distinct clinical outcomes in a cancer type specific manner.



Input

Clinical Data
- OncoTree code
- Sex
- Primary/Met

DNA

~220 genes

Consensus genes from commercial multi-gene NGS panels (e.g., Foundation, Tempus, MSK-IMPACT, DFCI-ONCOPANEL, etc.)

Open Source & Proprietary

- Language Model (BioBERT)

Pre-training Corpora: PubMed 4.5B words; PMC 13.5B words

BioBERT Pre-training

Weight Initialization: BERT from Devlin et al.

Pre-trained BioBERT with biomedical domain corpora

- Mut2Ex algorithm
- Principle Label Space Transformation (PLST) adapted to regression problem

Output

Reconstructed mRNA

Clinical task

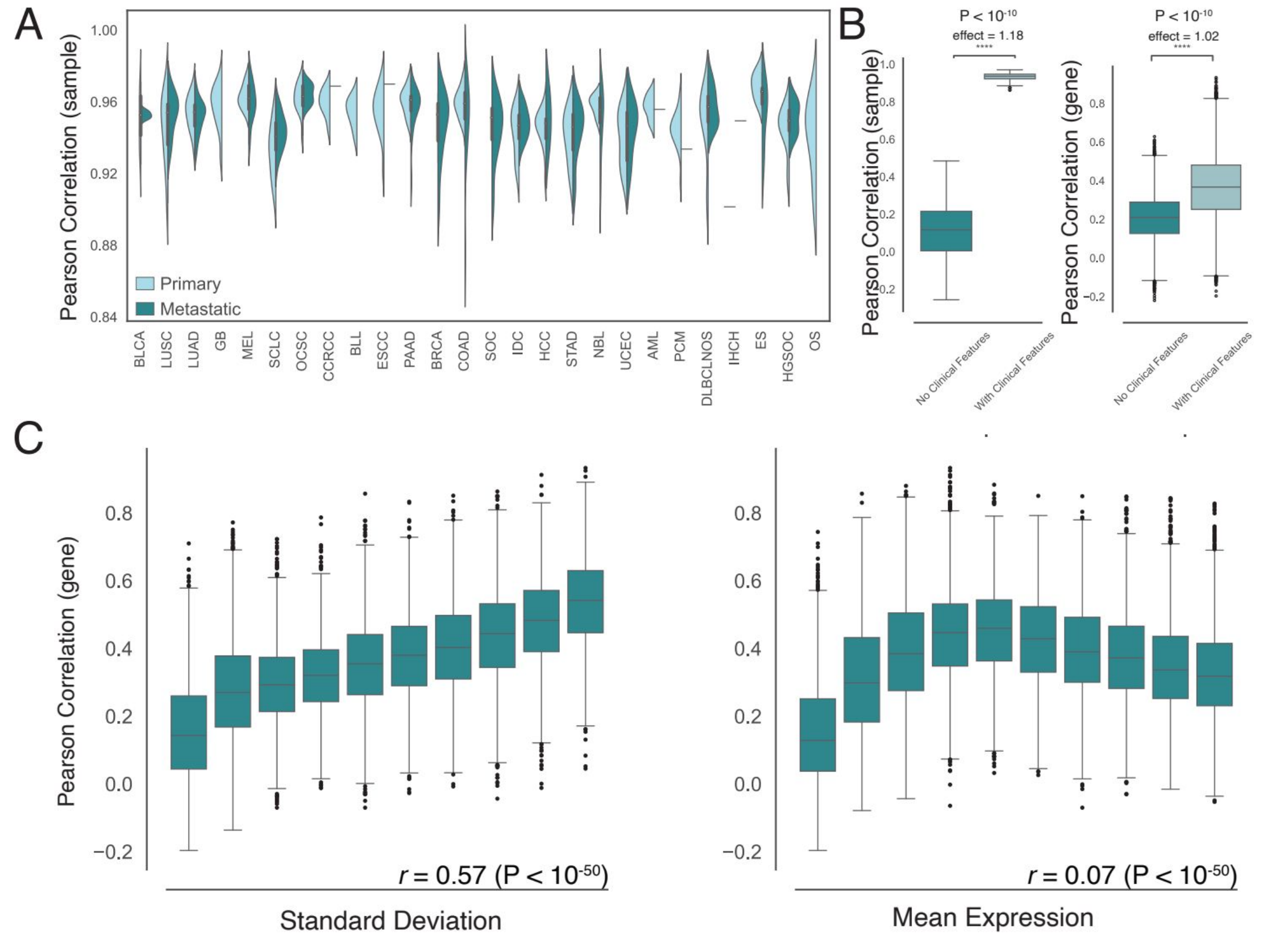
All logos displayed are the properties of their respective company owners and are used for illustrative purposes only. No official partnership with or endorsement by these companies is implied.

EHR
y=[0 1 1]
label A
B
C
Genes (n = 18,969)
Reconstructed expression level
susan G. komen.
prosigna
Breast cancer prognostic gene signature assay
EXACT SCIENCES
oncotypeDX Breast Recurrence Score
oncotype DX Colon Recurrence Score

Figure 1 | Reconstruction of patient tumor gene expression from minimal clinicogenomics data. Patient clinical features including cancer type diagnosis (OncoTree code), sex (Male or Female) and whether a tumor biopsy was sampled from a primary or metastatic site are used to derive sentence embedding vectors for input into a pre-trained biomedical language representation model designed for biomedical text mining tasks, called BioBERT. Corresponding one-hot encoded patient tumor hotspot mutations and high level copy number alterations (amplifications or homozygous deletions) for a set of n≈220 genes commonly profiled on multi-gene commercial next generation sequencing (NGS) panels were input into an adapted Principle Label Space Transformation (PLST) model, to reconstruct an mRNA transcriptome (n=18,969 genes) for a tumor sample. Reconstructed expression profiles can be applied to downstream analyses or mRNA-based clinical tasks, augmenting the utility of RWD cohorts.

Zephyr AI Machine Learning (ML) method reconstructs transcriptomes with high accuracy across multiple tumor types

Our expression reconstruction model, trained on the DepMap dataset, used 720 cell lines for training and 164 for testing. We compared the model's reconstructed expression profiles (18,969 genes) to actual expression in 26 cancer subtypes (**Fig. 2A**).



Including clinical features significantly improved accuracy at sample and gene levels (**Fig. 2B**, $P<10^{-10}$, effect sizes of 1.18 and 1.02, respectively). There was a strong positive correlation between reconstructed and true expression, especially for highly variable genes (**Fig. 2C**, left panel, r=0.66, $P<10^{-50}$), suggesting variability enhances model learning and prediction.

Figure 2 | Evaluation of Reconstructed Gene Expression Accuracy in DepMap Cell Lines by Cancer Type and Impact of Clinical Features. A) Distributions of correlation per sample of reconstructed expression of N=18,969 genes in Depmap cell lines per cancer type, split by primary and metastatic cell lines. [shown on both train and test data due to small sample size] **B)** Boxplots of correlation per sample between real and reconstructed expression (left, mean r = 0.1056 [0.0946 - 0.1167, 95% CI, N=164] without clinical features and mean r = 0.9342, [0.9328-0.9357, 95% CI, N=164] with clinical features) and per gene (right, mean r = 0.2073 [0.2064 - 02081, 95% CI] without clinical features and mean r = 0.3651 [0.3639 - 0.3663, 95% CI] with clinical features) of reconstructed expression of N= 18,969 genes. **C)** Boxplots of correlation per gene binned by standard variation (left) and mean of the true gene expression (right).

Zephyr AI's Reconstruction Model is Robust Across Diverse Commercial NGS Panels

The Zephyr AI 220 gene list was derived from the intersection of genes profiled across various commercial panels (**Fig. 3A, upper**). The performance of the reconstruction model was unaffected by choice of commercial NGS provider or assay (**Fig. 3A, lower**). Notably, while gene hotspot mutation detection varied significantly among commercial NGS providers in the AACR Project GENIE cohort (**Fig. 3B**), they are consistent for the 220 genes used for model input (**Fig. 3C**). A t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of reconstructed expression profiles shows the model output is robust across genomic inputs from various commercial NGS providers and assays, with no distinct clustering by assay type (**Fig. 3D**), while capturing salient clinical and biological features including cancer type (**Fig. 3E**) and expression patterns of key cancer genes (**Fig. 3F**).

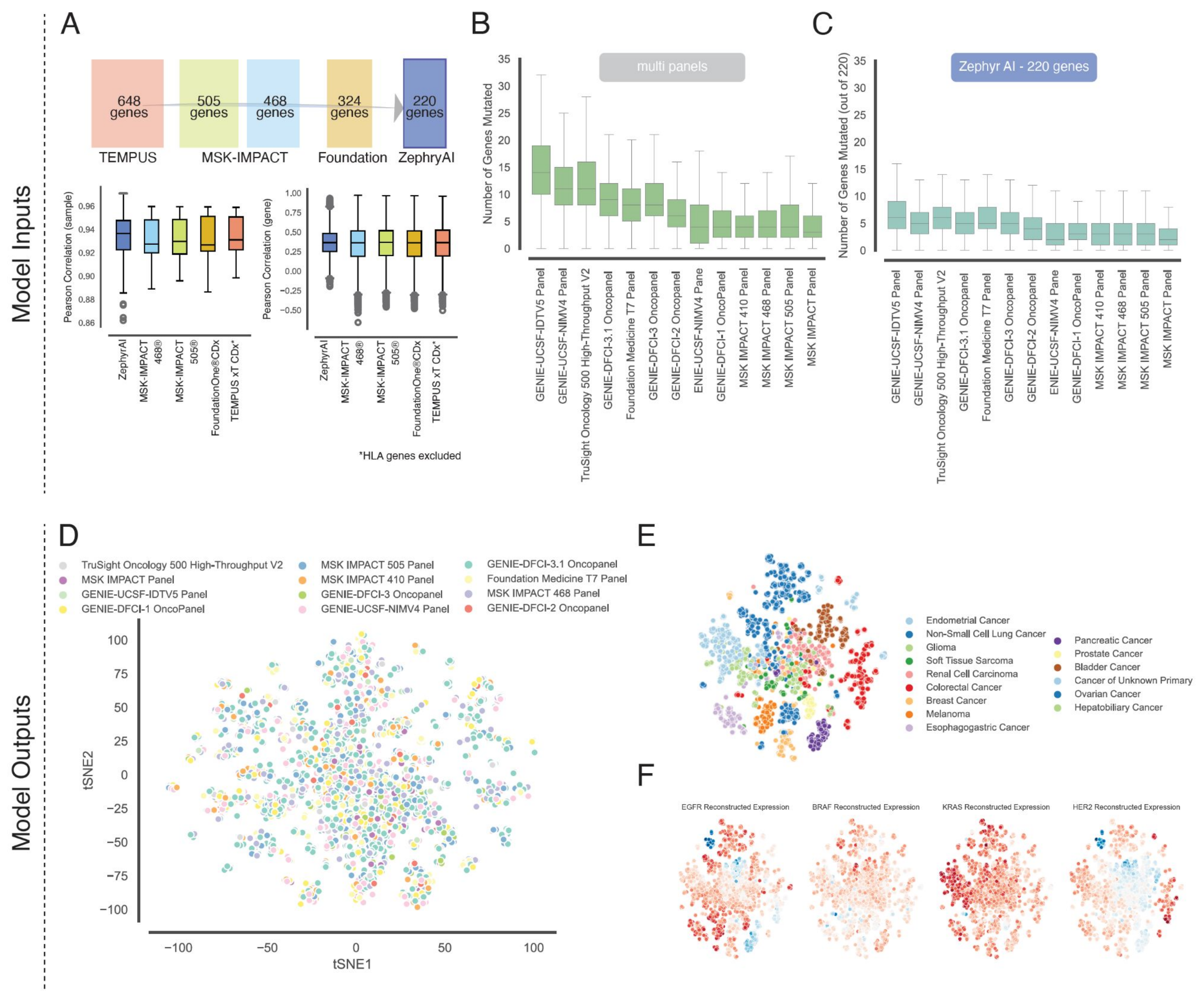


Figure 3 | Consistent Genomic Profiling and Robust Expression Reconstruction Across Commercial Panels. A) Zephyr AI 220 gene list was compiled by intersecting commercial panels (upper panel). Boxplots of correlation per sample between real and reconstructed expression per sample (lower left panel) and per gene (lower right panel) of N=18,969 genes for each of the gene panels tested (N=5). **B)** Boxplots of number of mutations per sample for each cancer type. **C)** Same as (B) but limited to the 220 genes and hotspot mutations that are the input for our Zephyr AI reconstructions model. **D)** tSNE on GENIE reconstructed expression colored by genomic panel. **E)** and by cancer subtype **F)** and expression of key cancer genes

Deriving PAM50 subtyping and other clinical features from reconstructed breast cancer expression profiles is comparable to real expression

To assess the clinical utility of reconstructed expression, we applied our method to 564 breast cancer samples (**Fig. 4A**), using only those features specified in **Fig 1**. We compared the predictive efficacy of reconstructed expression to true expression and mutations for four clinical classifications: stage, HER2 status, ER status, and PAM50 status[2]. Reconstructed expression performed comparably to true expression and outperformed DNA alterations alone in all tasks (**Fig. 4B**)

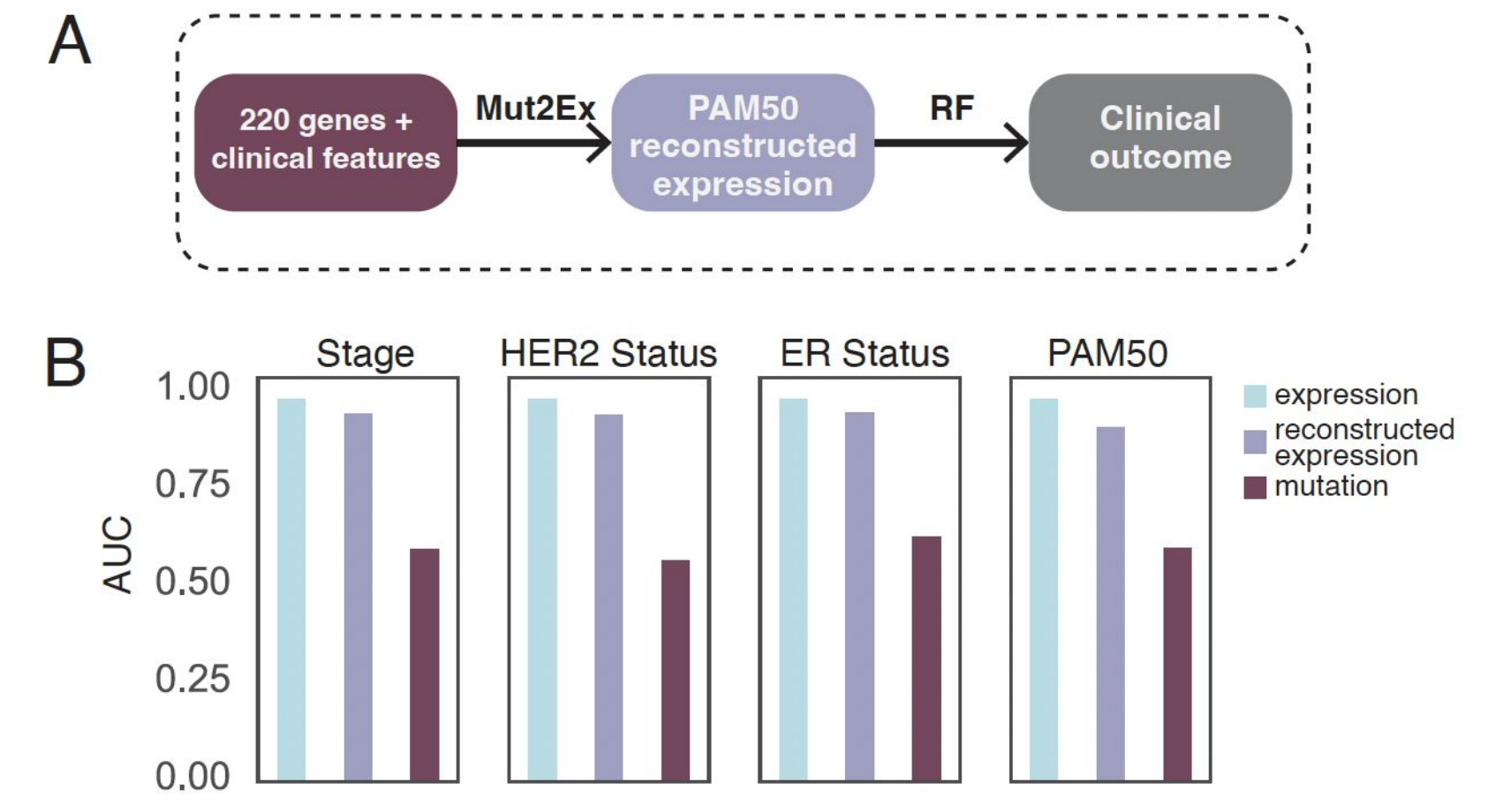


Figure 4 | Evaluating Clinical Utility of Zephyr AI's Reconstructed Expression Model in Breast Cancer. A) Workflow for assessing clinical utility of expression reconstruction model. **B)** Bar plots showing AUCs of classifiers predicting Stage, HER2 Status, ER Status, and PAM50 subtypes (left to right), trained on real expression (teal), reconstructed expression (lavender) or mutations alone (maroon).

Deriving OncotypeDx Signatures from Reconstructed Colorectal Cancer Expression Profiles

Reconstructed expression profiles were generated for 272 colon adenocarcinoma tumors. OncotypeDx signatures[3] were derived using genes from either real RNA sequencing or reconstructed profiles. Overall survival (OS) was compared between patients with high and low risk scores (RS) from real expression (**Fig. 5A**) and reconstructed expression (**Fig. 5B**). Real expression-based signatures showed a 12-month survival increase for low RS patients, while reconstructed expression-based signatures showed a 27-month increase. While a high correlation (r = 0.65, p-value < 10^{-30}) was observed between risk scores from real and reconstructed expression (**Fig. 5C**), discrepancies may contribute to improved survival outcomes in some patients. Indeed, gene set enrichment analysis revealed that high RS from reconstructed expression is associated with STK33 and BMI pathways, whereas high RS from real expression is linked to fatty acid metabolism and AKT/MTOR signaling (**Fig. 5D**).

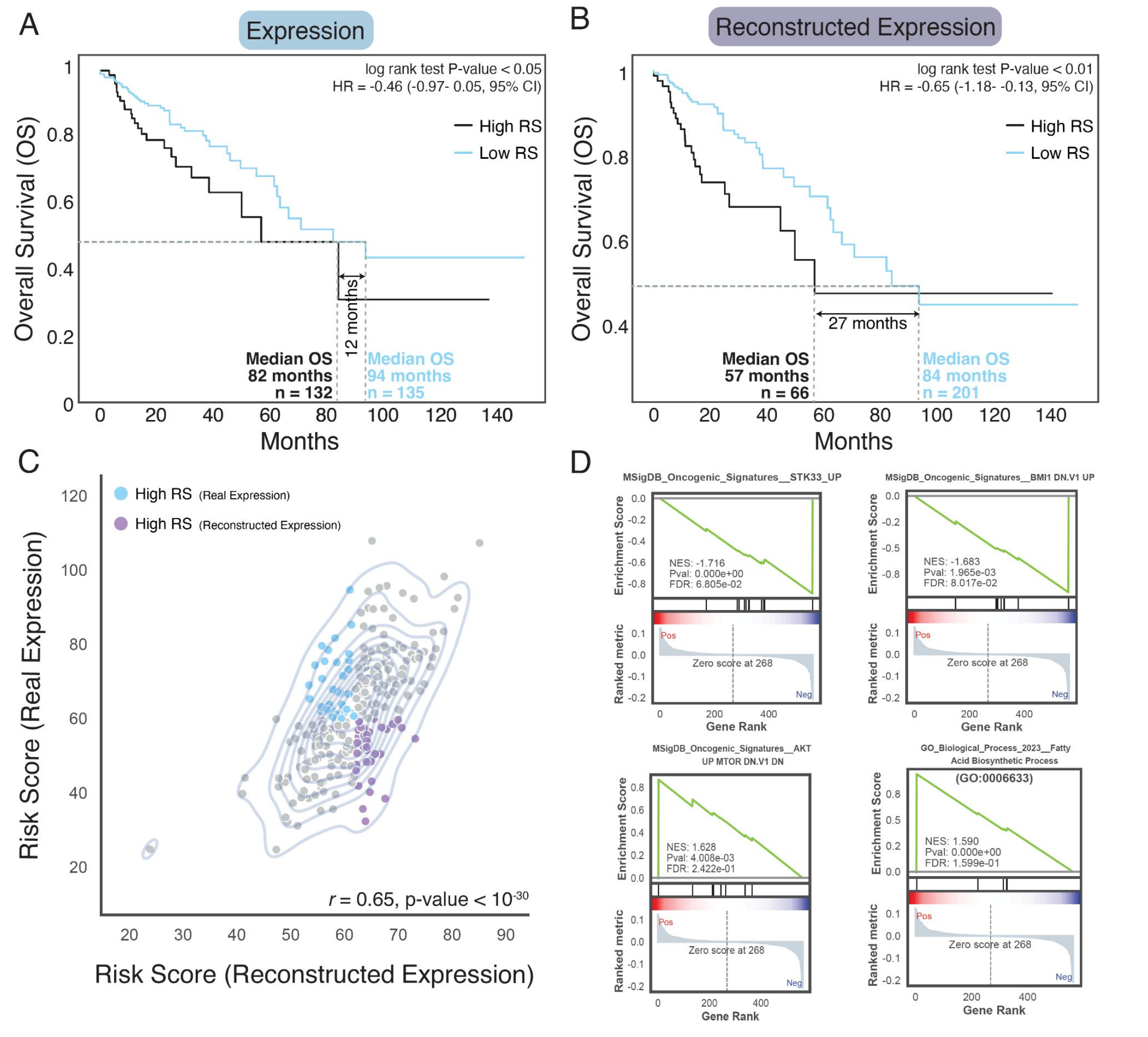


Figure 5 | Derivation of OncotypeDx signatures from reconstructed colorectal cancer expression profiles. A) Kaplan-Meier plots depicting the OS outcomes of patients stratified into low and high-risk groups based on risk scores derived from real expression. **B)** Same as (A) but patients stratified into low and high-risk groups based on risk scores derived from reconstructed real expression. **C)** Scatter plot illustrating correlation between risk scores computed from reconstructed expression and real expression data (r=0.65, p-value < 10^{-30}). Sample groups were categorized into two distinct groups based on risk scores (RS): samples with high RS solely derived from real expression data were denoted in light blue, those with high RS solely from reconstructed expression were indicated in purple, and the remaining samples were assigned to a default group represented by gray. **D)** Gene-set enrichment analysis (GSEA) results comparing the groups defined in (C).

CONCLUSION

Our flexible analytic framework for reconstructing gene expression profiles from clinicogenomics data substantially augments the clinical utility and value of data acquired in real-world settings.

REFERENCES: 1. Consortium, T. A. P. G. et al. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discov. 7, 818–831 (2017). **2.** Bastien, R. R. et al. PAM50 Breast Cancer Subtyping by RT-qPCR and Concordance with Standard Clinical Molecular Markers. BMC Méd. Genom. 5, 44–44 (2012). **3.** Knezevic, D. et al. Analytical validation of the Oncotype DX prostate cancer assay – a clinical RT-PCR assay optimized for prostate needle biopsies. BMC Genom. 14, 690–690 (2013).

ACKNOWLEDGEMENTS

The authors express their gratitude to the Zephyr AI science, engineering, data and business development teams for invaluable technical support and discussion. We also acknowledge the contributions of the authors and organizations cited, with special thanks to AACR Project GENIE, TCGA and the Cancer Dependency Map for essential data resources. We extend our appreciation to Candy Zhu and Jasmine Chu for their valuable assistance in designing this poster.