Reconstructing Gene Expression from Clinical and Genetic Panel Data for Predictions of Tumor Microenvironment Features & Response to Immune Checkpoint Inhibitor Therapy

ZEPHYR AI

Felicia Kuperwaser, Sunil Kumar, Dillon Tracy, Jeff Sherman, Andrey Chursov, Maayan Baron and Emily Vucic
Zephyr AI: 1800 Tysons Blvd Suite 901, McLean, VA, 22102 | https://www.zephyrai.bio

Background: The development of immune checkpoint inhibitor (ICI) therapy has fundamentally changed the landscape of cancer treatment. While ICIs have exhibited remarkable efficacy across diverse cancer types, the majority of cancer patients do not respond to these therapies [1]. Tools to better identify patients who would benefit from ICI therapy are urgently needed to facilitate personalized care. Models for ICI response that incorporate tumor microenvironment (TME) features in addition to molecular data have demonstrated improved predictive power of patient response to therapy [2, 3]. These features reflect the coordinated activity of multiple cell types and therefore are best captured by mRNA expression. Transcriptional profiles are not, however, readily assayed in clinical settings. Extracting TME features from molecular data already collected in clinical settings provides an opportunity to bridge the gap between predictive models that rely on these features and their translation into clinical practice, as well as enhance the clinical utility of real-world datasets.

Methods: We developed a machine learning (ML) model to reconstruct tumor gene expression profiles using genetic information from clinically available commercial NGS panels and embeddings [4] generated by a language model (Fig 1). This model was trained on publicly available data including ~8000 tumors representing 32 cancer types [5] and validated in additional heterogeneous patient cohorts.

Results: Gene expression reconstruction using this model was highly correlated with true expression (mean correlation per sample = 0.927, [0.9263 - 0.9284, 95% CI, N=1184]). We applied these data to the prediction of a set of TME signatures, previously associated with response to ICI therapy [6, 7, 8] and that describe TME composition and phenotypes (mean correlation per sample = 0.804, [0.8018, 0.8070, 95% CI, N=7555]). We demonstrate how reconstructed TME signatures are predictive of survival and provide interpretable biological insight into differences in patient outcomes across these cohorts.

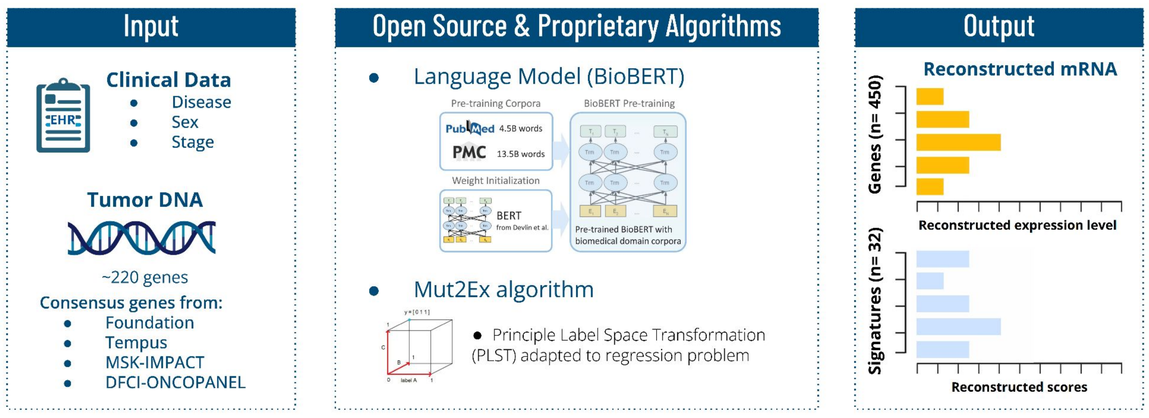

Figure 1 | **Gene expression reconstruction from real world data.** Clinical features and genetic panels are used to reconstruct expression of a selected set of ~450 genes and 32 tumor microenvironment (TME) signatures predictive of response to immune checkpoint inhibitor (ICI) therapy.

Zephyr's ML method reconstructs expression and TME features with high accuracy

We trained a model using DepMap clinical and genomic data to reconstruct gene expression for a set of ~450 genes selected for utility in reconstructing true TME signatures. Our method achieved higher correlation with true expression for most genes compared to other methods (Fig 2A-B). A subsequent model trained on TCGA data, showed consistent reconstruction across all cancer subtypes (Fig 2C). Using this reconstructed expression, we predicted a set of TME features and achieved comparable results to predictions from true expression (Fig 2D). Despite varying feature correlations, our method consistently predicted feature scores similarly from reconstructed or true expression (Fig 2E).

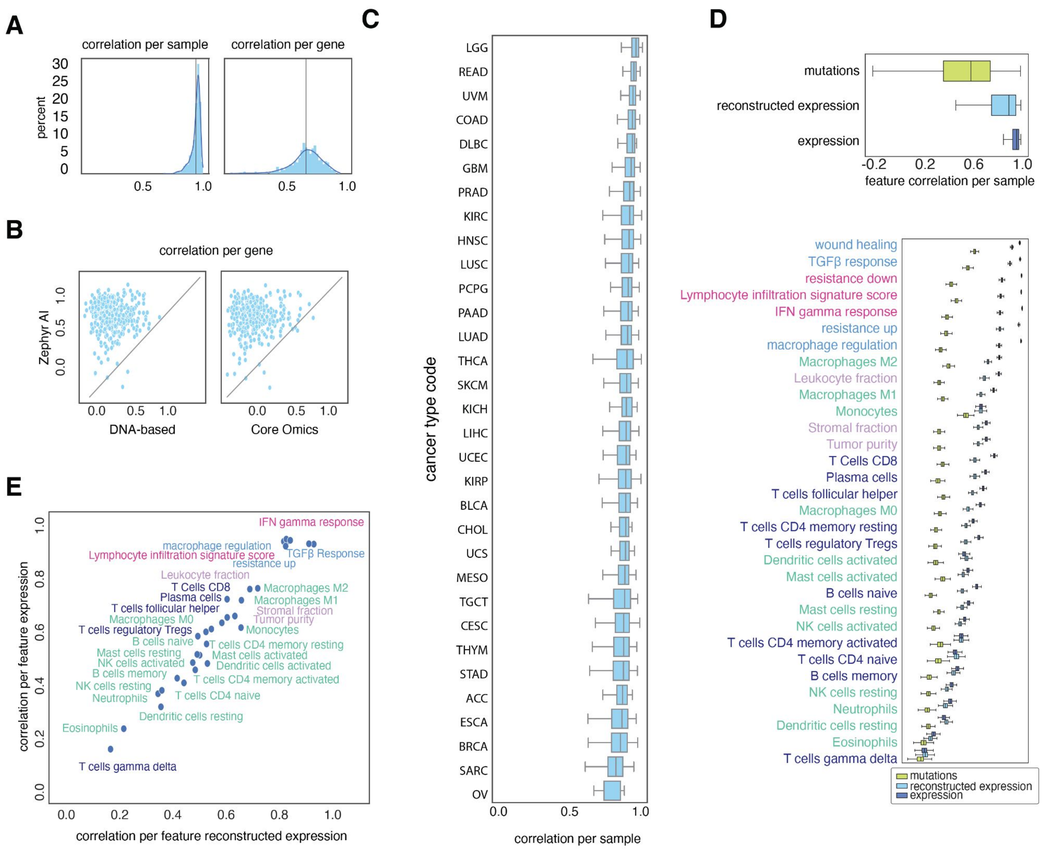

Figure 2 |**Zephyr AI methods reconstruct expression and TME features with high accuracy. A)** Distributions of correlation per sample (left, mean correlation = 0.927, [0.9263 - 0.9284, 95% CI, N=1184]) and per gene (right, mean correlation = 0.672, [0.6660 - 0.6778, 95% CI, N=400]) of reconstructed expression of selected genes in DepMap data. **B)** Zephyr AI method correlation per gene compared to DNA-based (left) and Core Omics (right) reconstruction approaches. **C)** Correlation per sample by cancer subtype of reconstructed TCGA expression. Central band of the boxplot shows the median, boxes represent the IQR, and whiskers represent the 5th and 95th percentiles. **D)** Correlation per sample of reconstructed TME features from mutations (mean correlation = 0.528, [0.5244, 0.5308, 95% CI, N=7555]), reconstructed expression (mean correlation = 0.804, [0.8018, 0.8070, 95% CI, N=7555) and true expression (mean correlation = 0.942, [0.9418, 0.9432, 95% CI, N=7555]) (top). Correlation per feature of reconstructed TME features from mutations, reconstructed expression and true expression (top). Central band of the boxplot shows the median, boxes represent the IQR, and whiskers represent the 5th and 95th percentiles. Features labels are colored by regulatory (light blue) or inflammatory (pink) summary signatures, total tumor fractions (lavender), or innate (green) or adaptive (dark blue) cell type fractions (bottom). **E)** Correlation per feature derived from reconstructed versus true expression. Features are colored as in D (top).

Zephyr AI model predicts TME features associated with clinical features

We next examined patterns across cancer subtypes in TCGA data and a curated real-world dataset with ICI-related patient outcomes. Visualization using tSNE based on reconstructed expression and TCGA features captured relevant clinical information, supporting the reliability of reconstructed data for clinical analysis (Fig 3A, B). We successfully recovered known relationships, such as the link between tumor mutational burden and T cell subtypes in lung squamous cell carcinoma (LUSC, Fig 3C) [9], and the correlation between wound healing and resistance signatures in melanoma (MEL, Fig 3D) [10]. Reconstructed features also explained survival differences. In skin cutaneous melanoma (SKCM) samples, high expression of immunoregulatory signatures correlated with worse outcomes, as previously reported (Fig 3E) [6]. PCA of bladder urothelial carcinoma (BLCA) samples revealed B cell differentiation as the primary source of variance (PC1), likely reflecting the cell of origin in this cancer type (Fig 3F, left and middle), whereas PC2 captured differences in resistance (Fig 3F, right). An inverse relationship was observed between presence of B cells, a lower 'resistance up' signature and survival (Fig 3F bottom and 3G), consistent with prior findings [11].

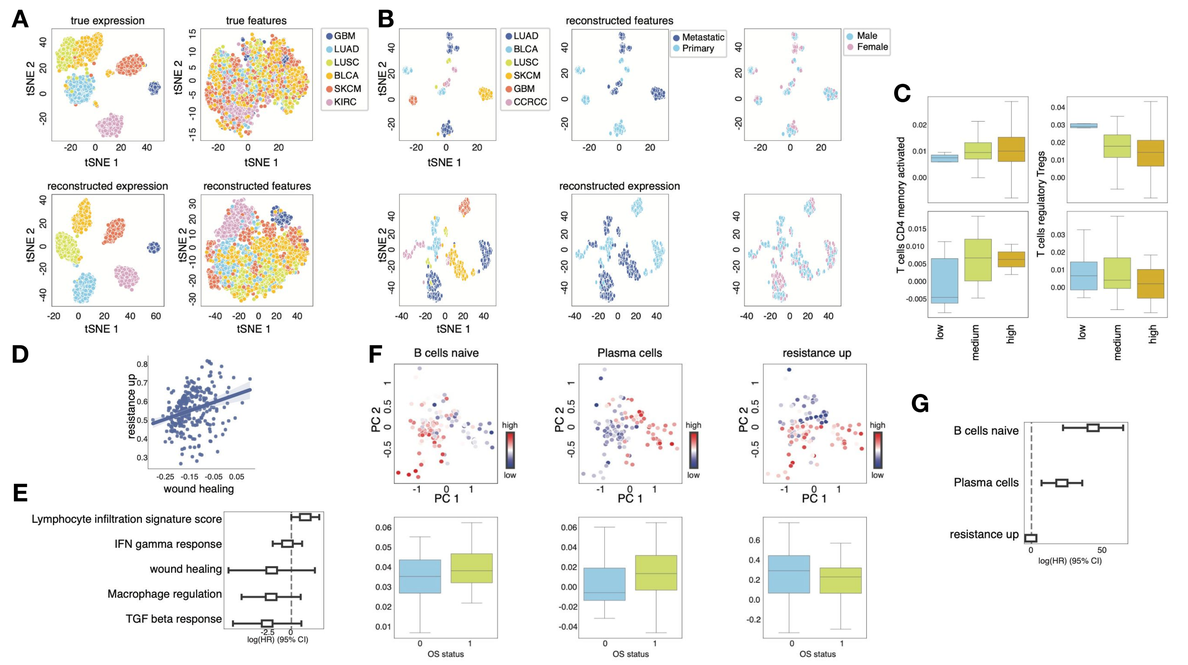

Figure 3 | **Reconstructed TME features explain clinical features and survival differences in real world data. A)** tSNE on TCGA true expression (top left, 446 genes) and true TME features (top right), reconstructed expression (bottom left, 446 genes) and reconstructed TME features (bottom right) colored by cancer subtype. **B)** tSNE on samples curated from our real world data cohort on reconstructed expression (top, 446 genes) and reconstructed TME features (bottom) colored by cancer subtype (left), primary versus metastatic status (middle) and sex (right). **C)** Boxplots of predicted fractions of T cells CD4 memory activated (left) and T cells regulatory Tregs (right) in LUSC samples from TCGA (top) and our curated RWD cohort (bottom) stratified by tumor mutational burden (x axis). Central line of the boxplot shows the median, boxes represent the IQR, and whiskers represent the 5th and 95th percentiles. **D)** Scatterplot of reconstructed wound healing versus resistance up resconstructed scores in melanoma samples from curated RWD cohort (slope = 0.3; p value < 10^{-6}). **E)** CoxPH analysis of reconstructed signatures in SKCM samples from curated RWD cohort. **F)** PCA on BLCA samples from curated RWD cohort on reconstructed TME features colored by B cells naive (left), Plasma cell (middle) and resistance up (right) scores (top). Boxplots of scores for each features stratified by OS status (bottom). Central line of the boxplot shows the median, boxes represent the IQR, and whiskers represent the 5th and 95th percentiles (middle). **G)** CoxPH analysis of features in bladder cohort from F.

Reconstructed TME features provide comprehensive tumor characterization for a large publicly available clinicogenomics cohort from NGS gene panel data alone

AACR Project GENIE is a large-scale RWD oncology clinicogenomics data-sharing initiative spearheaded by the American Association of Cancer Research (AACR) [11]. With data spanning ~180,000 tumor samples (~160,000 patients), it is an invaluable resource for the scientific community, fostering exploration into cancer biology, biomarkers, and therapeutic targets. We reconstructed TME features for 22 cancer types from this cohort.

Preliminary analysis revealed enrichment of TME feature scores based on TMB and sex across various cancer types (Fig 4A, B). Within cancer types, several interesting patterns emerged. In LUAD for example, TME features were strikingly distinct between primary and metastatic tumors (Fig 4C, left). PCA on metastatic lung adenocarcinoma (LUAD) tumors alone revealed two distinct clusters, one comprising only males (Fig 4C, middle), which also showed high expression of reconstructed PDL1 (Fig 4C, right). Correlation between feature scores and PC1 scores revealed further feature enrichments (Fig 4D). For example, patients with high PC1 scores, primarily males with elevated PDL1 expression, showed enrichment in lymphocyte infiltration, suggesting potential immunotherapy responsiveness (Fig 4D, left). In contrast, patients with low PC1 scores exhibited reduced inflammatory markers and cell subtypes, alongside resistance signatures (Fig 4D, middle) and innate immune cell infiltration (Fig 4D, right). Combining innate immune cell-directed therapy with T cell targeted therapy may enhance the inflammatory environment and improve T cell therapy responses in these patients.

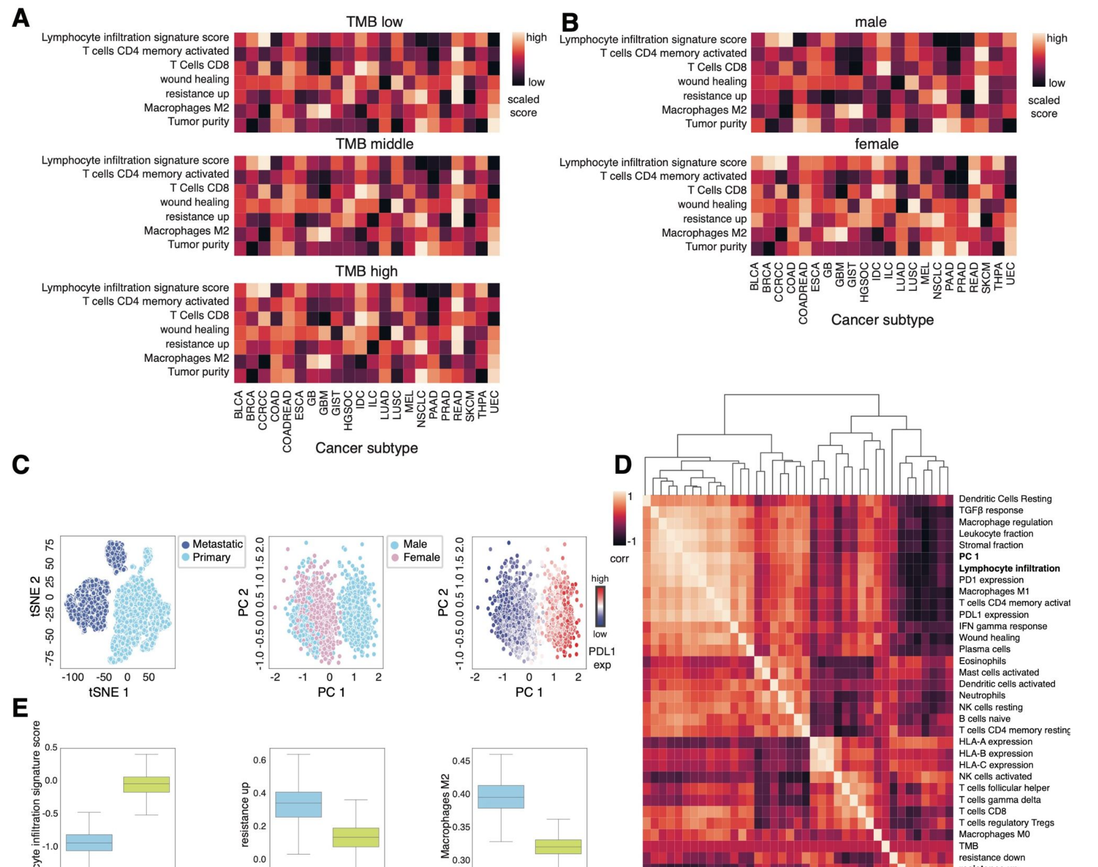

Figure 4 |Reconstructed TME features provide comprehensive tumor characterization in GENIE data. A,B) Heatmap of scaled reconstructed TME features across the most abundant cancer subtypes present in the AACR Project GENIE dataset stratified by TMB **(A)** and sex **(B)**. **C)** tSNE visualization of reconstructed TME features in GENIE LUAD samples colored by primary or metastatic status (top left). PCA visualization of GENIE LUAD metastatic samples colored by sex (top middle) and by expression of PDL1 (top right). **D)** Heatmap of correlations between reconstructed features in GENIE LUAD metastatic samples. **E)** Boxplots of selected feature scores between samples that have high or low PC 1 scores. **BLCA**: Bladder cancer; **BRCA**: Breast cancer; **CCRCC**: Clear cell renal cell carcinoma; **COADREAD**: Colorectal adenocarcinoma; **ESCA**: Esophageal carcinoma; **GB**: Gall bladder cancer; **GBM**: Glioblastoma multiforme; **GIST**: Gastrointestinal stromal tumor; **HGSOC**: High-grade serous ovarian cancer; **IDC**: Invasive ductal carcinoma; **ILC**: Invasive lobular carcinoma; **LUAD**: Lung adenocarcinoma; **LUSC**: Lung squamous cell carcinoma; **MEL**: Melanoma; **NSCLC**: Non-small cell lung cancer; **PAAD**: Pancreatic Ductal Adenocarcinoma; **PRAD**: Prostate adenocarcinoma; **READ**: Rectum adenocarcinoma; **SKCM**: Skin cutaneous melanoma; **THPA**: Papillary thyroid cancer; **UEC**: Uterine endometrioid carcinoma.

Reconstructing TME features from NGS panel data enhances clinical utility of RWD

While RWD may be limited in molecular characterizations, they typically present comprehensive clinical characterizations, including longitudinally collected NGS data and patient outcomes. 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. The expansion of data capabilities to encompass TME features opens exciting new avenues for discovery across numerous applications.

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 for her valuable assistance in designing this poster.

References: **[1]** Shiravand Y et al. Immune Checkpoint Inhibitors in Cancer Therapy. Curr Oncol. 2022 Apr 24;29(5):3044–60. **[2]** Riera-Domingo C et al. Immunity, Hypoxia, and Metabolism-the Ménage à Trois of Cancer: Implications for Immunotherapy. Physiol Rev. 2020 Jan 1;100(1):1–102. **[3]** Yang S et al. Identification of a prognostic immune signature for cervical cancer to predict survival and response to immune checkpoint inhibitors. Oncoimmunology. 2019 Oct 3;8(12):e1659094. **[4]** Lee J et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234–40. **[5]** The Cancer Genome Atlas Program (TCGA) [https://www.cancer.gov/tcga]. **[6]** Jerby-Arnon L et al. A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. Cell. 2018 Nov 1;175(4):984–97.e24. **[7]** Thorsson V, et al. Cancer Genome Atlas Research Network, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG, Shmulevich I. The Immune Landscape of Cancer. Immunity. 2018 Apr 17;48(4):812–30.e14. **[8]** Taylor AM et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018 April 9;33(4): 676-689.e3. **[9]** Charoentong P et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell Reports. 2017 Jan 3;18(1):248-262. **[10]** Hugo W et al. Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. Cell. 2016 Mar 24;165(1):35-44. **[11]** The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine Through An International Consortium, Cancer Discov. 2017 Aug;7(8):818-831 and include the version of the dataset used.

Contact:
maayan@zephyrai.bio

Abstract #1296