Systematic and Robust Prediction of Gene Expression Using Comprehensive Epigenomic Profiling of Plasma cfDNA

Khoi Nguyen*[1], Justin Finkle*[1], Anthony D'Ippolito*[1], Jonathan Beagan*[1], Aparna Gorthi*[1], Travis Clark[1], Amy Donahue[1], Baovy Tran[1], Tyrone Tamakloe[1], Charlene O'Brien[1], Michael Coyne[1], Jenna I. Wurster[1], Kristian Cibulskis[1], Corrie Painter[1], Mike Zhong[1], Humphrey Gardner[1], Jacob Berchuck[2], Jayne Stommel[3], SMMART Clinical Trials Program[3], Gordan Mills[3], J. Carl Barrett[1], Matthew L. Eaton[1]

[1]Precede Biosciences, Boston, MA, [2]Winship Cancer Institute of Emory University, Atlanta, GA, [3]SMMART Clinical Trials Program, Knight Cancer Institute, Oregon Health & Science University, Portland, OR

BACKGROUND

Accurate prediction of RNA transcript levels from circulating tumor DNA (ctDNA) offers a promising approach for dynamic non-invasive cancer biomarker characterization and longitudinal monitoring.
By leveraging epigenomic signals proximal to individual genes, we developed transcript level machine learning prediction models for breast and prostate cancer using simulated plasma and validated them in patient plasma.

METHODS and RESULTS

In breast cancer, our models predicted expression for a panel of 2622 key breast cancer oncogenes and transcriptional drivers in simulated plasma at 10% ctDNA or lower, enabling clinically relevant prediction of cancer gene transcription. At 10% ctDNA, we can evaluate all models in ~40% of stage IV patients, but many models remain performant at lower ctDNA levels, expanding the evaluable patient population.

These models are tuned to separate healthy plasma signal from cancer cells' contribution to the plasma. The panel is enriched for genes important for transcriptional regulatory identity and breast cancer subtypes, as well as targets for antibody drug conjugates (ADCs) and radio immune conjugates (RICs), including HER2, PSMA, Nectin-4, B7-H4, MET, and EGFR.

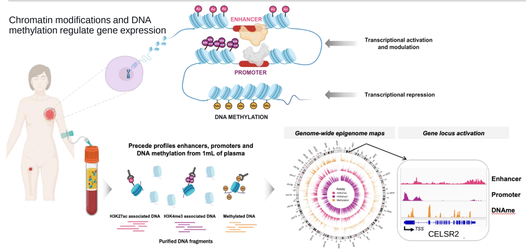
Our models demonstrated direct applicability in patient plasma with orthogonal validation. For example, they accurately predicted prostate cancer PSA and PSMA PET SUVmean, as well as breast cancer HER2, ER, and PR status. Additionally in a cohort of patients with matched tissue RNA-seq there was a strong correlation between RNA-seq and plasma-based predictions.

Further improvements were achieved with bespoke feature engineering and multi-gene aggregation. These strategies improved AUC for predicting HER2, ER, and PR IHC positivity in breast cancer, as well as the performance of our predictor for PSMA PET SUVmean. Similarly, for a multi-gene DLL3 predictor in lung cancer, we reduced the ctDNA threshold from 4% to 1.9% ctDNA by incorporating predicted expression of three highly correlated genes.

RESULTS

Figure 1. Epigenomic Platform

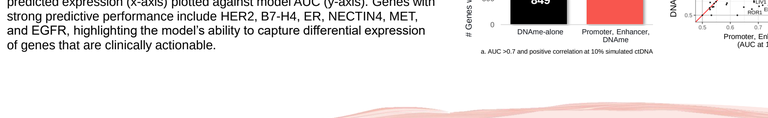
Comprehensive Epigenomic Platform Offers Dynamic Resolution into Target & Pathway Biology from 1ml of Plasma



Cell free DNA derived from tumors exist in circulation as chromatin fragments that faithfully maintain tumor-associated epigenetic modifications on the histones and DNA. Antibodies against active enhancers, active promoters and DNA methylation are used to enrich for associated DNA fragments from 1mL plasma and sequenced to capture the underlying transcriptional state of the tumor cells (Baca et al.).

Figure 2. Training & Testing of Predictive Gene Locus Expression Models in Simulated Cancer Plasma

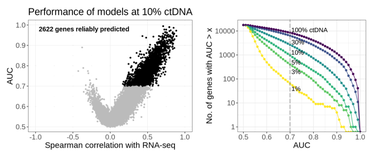
A) Training Models to Predict Tumor Gene Expression



(Left) Schematic representation of the approach used to generate data to train gene locus expression models. Simulated cancer plasma is created by mixing tumor-derived chromatin with healthy plasma.

(Right) For each gene, an expression model is trained to predict tumor RNA transcript levels as a function of gene locus activity in plasma, integrating enhancer signals, promoter accessibility, and DNA methylation (DNAme).

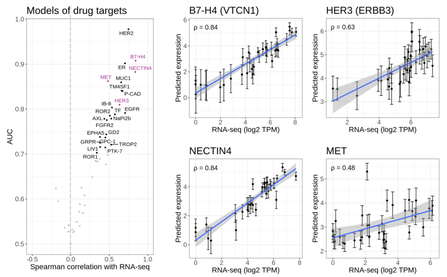
B) Cross-validated Performance of Predictive Gene Expression Models Trained Across a Panel of Breast Cancer Samples



(Left) Performance of models for all genes, evaluated in simulated 10% ctDNA plasma samples. Each point represents a gene model, with Spearman correlation between RNA-seq and predicted expression (x-axis) plotted against model AUC (y-axis). 2,622 genes (black points) met the threshold of AUC >0.7 and positive Spearman correlation, indicating reliable prediction of tumor RNA levels from cfDNA-based epigenomic signals at 10% ctDNA.

(Right) The number of genes meeting different AUC thresholds across varying ctDNA fractions (100%, 30%, 10%, 5%, 3%, 1%). Predictive accuracy declines at lower ctDNA fractions, but many genes remain robustly predicted at clinically relevant levels.

C) Performance of Models for Clinically Relevant Drug Targets in Breast Cancer

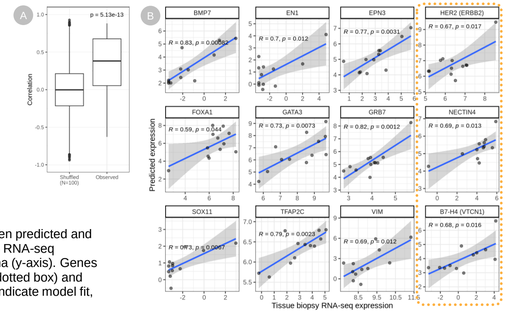


(Right) Correlation between actual and predicted values of expression in simulated 10% ctDNA plasma samples for 4 drug targets: B7-H4 *(VTCN1), NECTIN4*, *MET*, and HER3 *(ERBB3).* The x-axis represents actual expression measured by RNA-seq in a tumor sample, while the y-axis shows the model's predicted RNA expression from cfDNA epigenomic features in simulated 10% ctDNA plasma samples generated from that tumor. Error bars represent the standard deviation of the model predictions across simulated biological replicates.

(Left) Cross-validated performance of predictive models for key drug targets in simulated 10% ctDNA plasma samples, including targets of antibody-drug conjugates (ADCs), hormone therapy, and other oncology drugs. Each point represents a gene, with Spearman correlation between tumor RNA-seq and predicted expression (x-axis) plotted against model AUC (y-axis). Genes with strong predictive performance include HER2, B7-H4, ER, NECTIN4, MET, and EGFR, highlighting the model's ability to capture differential expression of genes that are clinically actionable.

Figure 3. Gene Expression Predictions in Clinical Patient Plasma Validate in Tumor RNA-seq

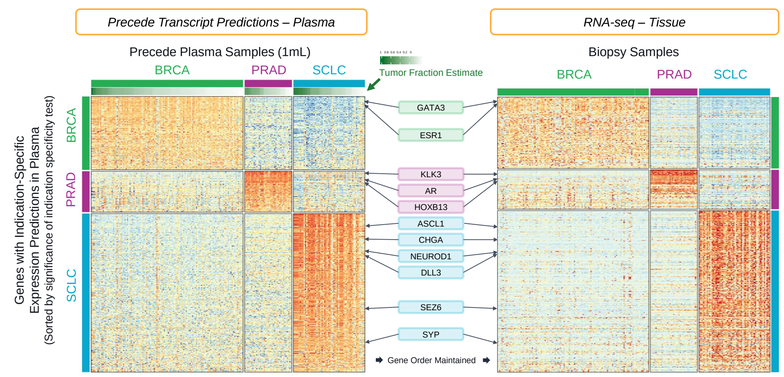
A, B) Correlation Between Model-predicted Gene Expression in Breast Cancer Patient Plasma and RNA-seq from Tumor Biopsies of the Same Patients



A) We used our models to predict gene expression from patient plasma samples (N=12, ctDNA >3%) for which we have matched tumor biopsies. Boxplots show the distribution of Pearson correlation between predicted and measured RNA-seq expression across genes (Observed), compared to the distribution of correlations in a randomized control (Shuffled, N=100). Model predictions are significantly more correlated with RNA expression than expected by chance.

B) Examples of 12 genes with strong correlation between predicted and measured expression. Scatter plots show tumor biopsy RNA-seq expression (x-axis) vs predicted expression from plasma (y-axis). Genes include ADC targets (e.g., HER2, NECTIN4, B7-H4 in dotted box) and breast cancer-relevant markers. Blue regression lines indicate model fit, and shaded regions represent confidence intervals.

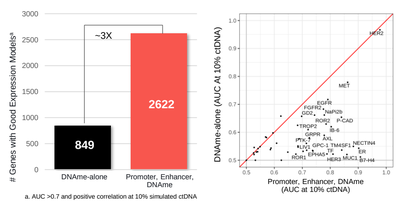
C) Gene Expression Predictions in Patient Plasma Recapitulate Indication-specific Transcriptional Programs at Scale



(Left) Heatmap showing model-predicted gene expression in plasma samples from breast (BRCA), prostate (PRAD), and small cell lung cancer (SCLC) patients. Genes displayed in the heatmap had higher predicted expression in plasma samples from one indication compared to the others, and include well known cancer type markers such as GATA3, KLK3, and ASCL1.

(Right) Heatmap showing publicly available RNA-seq expression data of tumor samples from the same indications, for the same genes shown on the left (TCGA Research Network, George et al.). Despite no direct matching between plasma and these tumor samples, 94% of the genes identified in plasma as indication-specific were significantly elevated in the expected cancer type in tissue biopsies. Further, gene network structures were highly similar between plasma predictions and tumor RNA-seq (Mantel test, p<2.2e-16).

D) Comprehensive Epigenomics Enables More Powerful Predictive Models Compared to Using DNA Methylation Alone



a. AUC >0.7 and positive correlation at 10% simulated ctDNA

(Left) Comparison of number of genes for which we can generate predictive gene expression models (AUC >0.7 and positive correlation at 10% simulated ctDNA) when using our comprehensive epigenomic platform measuring promoter, enhancer, and DNA methylation (DNAme) signal, compared to using DNA methylation only.

(Right) Comparison of model AUC for key drug targets at 10% simulated ctDNA. AUC for models trained using data from our comprehensive platform is plotted on the x-axis, while AUC for models trained using DNA methylation data only is on the y-axis. Most targets are below the diagonal line, indicating that our complete platform increases performance.

E) ctDNA Content in Patients with Metastatic Breast Cancer

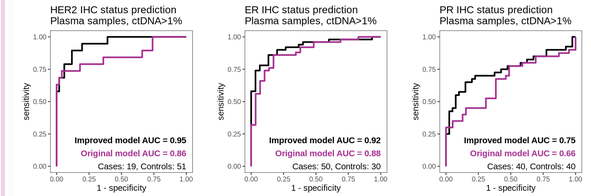
Min. % ctDNA	% of patients
0.01%	95%
0.51%	80%
1.02%	75%
1.44%	70%
2.76%	60%
3%	59%
5%	52%
5.86%	50%
10%	40%
20%	30%
30%	20%

Percent of patients with stage IV metastatic breast cancer (N=725) with at least a specified fraction of plasma ctDNA, as measured by Precede's proprietary epigenomic-based tumor fraction estimation method.

This is the estimated size of the evaluable patient population for a test that is predictive at a given minimum ctDNA fraction.

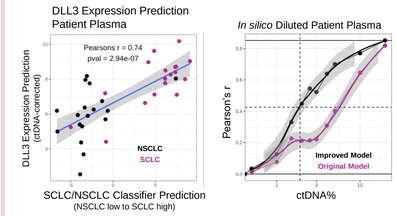
Figure 4. Development of Clinically Relevant Tests with Gene Expression Models

A) Predicting HER2, ER, and PR IHC Status in Breast Cancer from Plasma >1% ctDNA



Original locus expression models (purple) can predict IHC status. Refined models (black) outperform original models, improving AUC through feature engineering (HER2) and incorporation of additional loci (ER, PR).

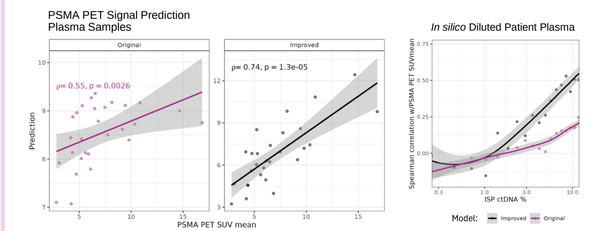
B) Predicting DLL3 Expression in Lung Cancer from Plasma



(Left) DLL3 expression predictions in plasma correlate with lung cancer histology (non-small cell vs small cell lung cancer, NSCLC vs SCLC). Predicted expression is higher in samples with reported SCLC histology (purple dots). Plasma samples were also classified using an SCLC vs NSCLC prediction model. Discordant cases where the classifier predicts SCLC for a reported NSCLC sample also show elevated DLL3, while SCLC cases predicted as NSCLC show low DLL3 expression.

(Right) High ctDNA patient plasma samples were diluted (*in silico*) to range of low ctDNA% to test concordance of predictions for same sample at varying ctDNA levels. DLL3 model refined by incorporating additional loci (black) maintains stronger correlation with source sample compared to the original model (purple), improving model precision.

C) Predicting PSMA-PET SUV Mean in Prostate Cancer from Plasma



(Left & Center) PSMA model predictions correlate with imaging-based quantification of PSMA (PET SUV mean). Refined PSMA model (black) improves correlation with PET SUV mean (R=0.74 vs 0.6 for original model in purple).

(Right) Refined model retains predictive power at lower ctDNA%, showing stronger correlation in *in silico* dilutions.

CONCLUSIONS

Comprehensive epigenomic profiling enables accurate prediction of tumor-specific RNA transcription in cfDNA, even at low ctDNA levels. This approach enhances ADC target profiling and enables tumor transcriptional profiling in patients from blood. Expanding beyond proof-of-concept models promises improved predictive accuracy at lower tumor fractions and broader applications in diagnostics and therapeutic targeting.

Contact J. Carl Barrett, PhD
Precede Biosciences
Email: carl.barrett@precede.bio

References
Baca et al., *Nature Medicine*, 2023 (PMID: 37865722)
TCGA Research Network: https://www.cancer.gov/tcga
George et al., *Nature*, 2015 (PMID: 26168399)