# Abstract

# Introduction

# Methods

# Results

## Estrogen receptor is a clinical continuous variable

|  |
| --- |
| Figure 1: Scores and survival analysis results from TCGA, SCANB and METABRIC cohorts. (a) GSVA scores for the SET ER/PR signature for each cohort. Each point corresponds to a patient sample and they are divided by estrogen receptor status. (b) Forest plot of the survival analysis for each cohort separately. NPI: Nottingham prognostic index. Ti: i-th stage of tumor. Ni: i lymph nodes with breast cancer cells. |

## Single sample integration preserves relevant breast cancer properties

|  |
| --- |
| Figure 2: (a) Biplot using the third and fourth components on TCGA and METABRIC samples. Colored by cohort. (b) Same as **a**, colored by ER status. (c) Same as **a**, colored by PAM50 molecular subtype. (d) Hex grid calculated on the biplot of the fourth and third component. Each hex is colored based on its average value of the SET ER/PR signature. |

## Embedding is robust to missing genes and generalized to a validation cohort

|  |
| --- |
| Figure 3: Validation of the molecular landscape with an external cohort (a) Biplot using the third and fourth components and now including all samples from the three cohorts: TCGA, METABRIC and SCANB. (b) Same as **a**, colored by ER status. (c) Same as **a**, colored by PAM50 molecular subtype. (d) Biplot using the first and second component of TCGA, METABRIC and SCANB. (e) Embedding of random samples given different training sets for PCA. (f) Biplot of all possible embeddings of sample given a certain proportion of top loadings missing in the dataset. |

## Molecular landscape is a tool to understand and reveal patient heterogeneity

|  |
| --- |
| Figure 4: Embedding of the POETIC cohort into the molecular landscape and pathway analysis for patient samples. (a) Biplots of the POETIC samples (baseline and surgery) into the molecular landscape. Left plot is colored by ER status and right plot is colored by molecular subtype PAM50 when available. (b) Biplot highlighting two patients with similar embedding and different response status. (c) Posterior distributions of the average scores in the neighborhood of the responder patient. Red line corresponds to the patient score. (d) Posterior distributions of the average scores in the neighborhood of the **non**-responder patient. Red line corresponds to the patient score. |

# Discussion

# References