# Abstract

# Introduction

# Methods

# Results

## Estrogen receptor is a clinical continuous variable

We used three independent breast cancer molecular datasets (1–3) to calculate estrogen signaling scores. The estrogen signatures HALLMARK\_ESTROGEN\_RESPONSE\_EARLY and HALLMARK\_ESTROGEN\_RESPONSE\_LATE were extracted from the molecular signature database (4) and from (5). The individual scores for each patient sample are shown in [Figure 1](#fig-01) (a) for each cohort stratified by estrogen receptor status. It shows the scores capture the differences between the two breast cancer subtypes as expected. Moreover, there is a wide range of values in the estrogen receptor positive (ER+) subgroup.

Cox regression was used to determine the hazard ratio of the estrogen signaling in overall survival (OS) for TCGA, SCANB and METABRIC and recurrence free survival (RFS) for METABRIC. Each survival analysis was done independently and adjusted for available clinical variables. Tumor size and number of lymph nodes were used for TCGA and SCANB cohorts. The Nottingham prognostic index (NPI) was used for METABRIC. Age was used in all cohorts as a clinical variable for adjustment. Only ER+ BC patients were used and when possible only those that received endocrine therapy. [Figure 1](#fig-01) (b) shows the forest plots for each cohort individually when calculating the hazard ratio for the estrogen signaling signature. In all the three cases, the hazard ratio for was below 1, with values ranging from 0.23 to 0.61. There is moderate variability for each hazard ratio. This shows the continuous aspect of estrogen receptor status.

|  |
| --- |
| 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

Since each patient has a different ER signaling score, we assumed that patients should be treated individually, not just binned in two big subgroups as ER+ and ER-. Therefore, it is important to consider each patient individually. We developed a single sample batch effect removal method (See methods section for the step by step) to integrate microarray and bulk RNA-seq and create a molecular landscape. The advantage of the method is that given a new sample, it can easily be integrated with all the other previous samples without any retraining.

The biplot in [Figure 2](#fig-02) (a) with the third and fourth components from TCGA and METABRIC samples shows that the samples are well integrated. All samples, including those using for training and validation, are plotted. The third components corresponds to the separation between ER+ and ER- BC patients in both cohorts ([Figure 2](#fig-02) (b)). A combination of the third and fourth components shows a good distinction among the PAM50 molecular subtypes ([Figure 2](#fig-02) (c)). The fourth component is mostly dividing the luminal A and luminal B subtypes, whereas the normal-like subtype is spread across the third and fourth component. This also highlights the fact that one cannot interpret the PCA locations globally, rather when comparing samples one should consider only its neighborhood. As pointed out before, ER status should be considered continuous and not dichotomous, [Figure 2](#fig-02) (d) shows a gradient of the ER signaling score . The higher values are on the far right of the third component, going to negative values as one goes from right to left, i.e., moving from a more ER+ status to the ER-.

|  |
| --- |
| 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

METABRIC and TCGA were used to train and validate the projections. SCANB was used as an external validation cohort. SCANB is well mixed with both METABRIC and TCGA samples ([Figure 3](#fig-03) (a)). ER+ and ER- BC patients are well separated ([Figure 3](#fig-03) (b)) and the procedure can also distinguish the molecular subtypes ([Figure 3](#fig-03) (c)). As an RNA-seq cohort, it is expected that SCANB samples will be closer to TCGA than to METABRIC when removing batch effects, due to platform biases and initial scale of the genes. Biplot of PC1 and PC2 ([Figure 3](#fig-03) (d)) shows that SCANB is closer to TCGA than to METABRIC. It is also in between the two cohorts.

In order to check the robustness of the procedure, we redid all the pipeline 10 times with 10 random sets of patient samples from TCGA and METABRIC, simulating a cross validation process. The PCA embedding is invariant to rotation, translation and reflection ([Figure 3](#fig-03) (e)). Another problem that arises with publicly available datasets, is the fact that there are missing genes. We try to understand the effect of missing genes in the embedding based on their loading values. Ideally if a low amount of genes with high loadings are missing, this should not affect very much the embedding. On the other hand, the more genes missing with high loadings, the more it will impact the embedding. We removed 200 genes in total with a varying proportion of top loading genes (ranging from 0 to 100% in a 5% step). The number of top loading genes missing from the dataset is key for the embedding ([Figure 3](#fig-03) (f)). The higher the proportion the less precise the embedding is.

|  |
| --- |
| 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. Blue dots correspond to the original embedding of a sample and red dots correspond to the new embedding given the new training set. (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

1. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The sweden cancerome analysis network - breast (SCAN-b) initiative: A large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Medicine [Internet]. Springer Science; Business Media LLC; 2015;7:20. Available from: <https://doi.org/10.1186/s13073-015-0131-9>

2. Comprehensive molecular portraits of human breast tumours. Nature [Internet]. Springer Science; Business Media LLC; 2012;490:61–70. Available from: <https://doi.org/10.1038/nature11412>

3. Curtis C, Sohrab P. Shah and, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2, 000 breast tumours reveals novel subgroups. Nature [Internet]. Springer Science; Business Media LLC; 2012;486:346–52. Available from: <https://doi.org/10.1038/nature10983>

4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences [Internet]. Proceedings of the National Academy of Sciences; 2005;102:15545–50. Available from: <https://doi.org/10.1073/pnas.0506580102>

5. Sinn BV, Fu C, Lau R, Litton J, Tsai T-H, Murthy R, et al. SETER/PR: A robust 18-gene predictor for sensitivity to endocrine therapy for metastatic breast cancer. npj Breast Cancer [Internet]. Springer Science; Business Media LLC; 2019;5. Available from: <https://doi.org/10.1038/s41523-019-0111-0>