Transcriptional signatures and machine learning provide a framework for better understanding estrogen receptor positive breast cancer

Carlos Ronchi, Cathrin Brisken

Abstract

Estrogen receptor (ER) is the key component in ER positive breast cancer (BC). It is a common target of endocrine therapies among patients whose tumor cells express ER. Among ER+ BC patients, several will recur despite the use of endocrine therapy, therefore the need to understand the biology of the tumors and the risk of the patients before any recurrence happens. In this paper we show that among ER+ BC patients that received only endocrine therapy, molecular estrogen signaling signatures are prognostic on TCGA, METABRIC and SCANB (over 10000 patients in total). Even among tumors with ER IHC percentage greater than 90%, estrogen signatures are prognostic, showing a discrepancy between ER IHC and molecular data. In order to better understand the molecular biology of these tumors, we developed a new pipeline using unsupervised machine learning algorithms to integrate and interpret publicly available datasets in a single sample manner, such that patient samples can be used regardless of the cohort of origin and sequencing platform. To validate the pipeline, we used SCANB (RNAseq), normal mammary gland samples (RNAseq), patient derived xenografts (PDX, RNASeq) and data from the POETIC trial (microarray, ER+ only patients). We show that the molecular landscape, where samples are integrated and embedded together, captures the different PAM50 molecular subtypes and does not depend on clinical variables, such as age, tumor stage and tumor purity. Moreover, all patients from the POETIC trial whose embeddings in the molecular landscape are on the basal region were non-responders. We further show using the POETIC data that looking at the neighborhoods in the molecular landscape of a responder and non-responder we see differences in estrogen signaling, highlighting potential differences in the treatment. PI3K AKT MTOR signaling and G2M checkpoint are two pathways that are drivers of the molecular landscape and we showed that they are also are prognostic for RFS on SCANB and METABRIC, showing possible alternative treatments with drugs currently in the market for early stage BC patients. To finish we derive a new

risk score as an alternative to the commercially available signatures that depends purely on the position of the sample in the molecular landscape and on clinical factors and provides additional information when used in combination with the risk of recurrence (ROR).

Introduction

Breast cancer is the most diagnosed tumor worldwide (1). It is a very heterogeneous disease that can be subdivided in different histological and molecular subtypes. Patients that are diagnosed with breast cancer are classified based on the expression of the estrogen, progesterone receptors and HER2 membrane protein. The estrogen receptor positive (ER+) breast cancer subtype is the most frequent found in clinical practice (more than 70% of the cases) (2). Tumor cells of this subtype contain the estrogen receptor whose main mechanism is the target of selective estrogen receptor modulator (SERM), selective estrogen receptor degrader (SERD) and aromatase inhibitors (AI). To receive endocrine therapy, tumors need to be estrogen receptor positive, which according to the ASCO guidelines are defined as having at least 1% of the tumor cells expressing estrogen recepetor (3). It has been shown that the patients in the low spectrum of ER positivity, from 1% to 10%, do not benefit as much as patients with 10% to 100% of ER positive cells in the tumor (4).

Transcriptomics and genomics have revolutionized cancer research. The new tools opened the option to better understand the molecular underpinnings of cancer biology. In breast cancer, molecular subtyping through the use of next generation sequencing technologies and qRT-PCR have been used to aid clinicians when taking the decision of giving chemotherapy for patients, avoiding potential overtreatment of patients. Researchers have developed gene signatures that are able to assign a risk score. Such score is based on survival data found in the literature and the higher the risk score of a patient, the more likely the recurrence, therefore the use of additional chemotherapy in first-line therapy is advised. Some of the risk scores are already recommended in international guidelines (10) which are already being used in the clinics (5–9). However, these signatures do not provide a possible explanation on why a patient should receive additional therapy. What is known is that they are composed of several submodules pathways, such as estrogen, proliferation and HER2 signaling and that some signatures are correlated to some of its submodules. For example, the recurrence score (RS) is associated with the estrogen module and the Risk of Recurrence (ROR) signature is correlated with its proliferation module (10). An alternative to the signatures only approach is to use RNA-sequencing in the clinics (11). The SCANB consortium showed that it is possible to use mutational and gene expression based biomarkers within one week of tumor surgery (12–14) in the clinics. They also showed that it is possible to use RNA-seq to calculate the Prosigna risk scores and risk stratification (15), showing the potential all in one use of RNA-seq instead of using different assays from different companies.

Even though there are several new analytical techniques to understand the risk of recurrence for ER+ BC patients, there is still a lack of understanding of the patient's tumor molecular

biology. There are challenges to overcome in the pathway analysis for patients individually and how to compare patients molecularly when considering complete transcriptomics data. There is no tool that allows to integrate patients in a single sample way and that does not require large amounts of data. In most cases integration is a one step procedure and cannot be updated. The common tools used in the RNA-seq community for batch effect removal are (16–18) and new samples cannot be integrated in a straightforward manner. The only way is to re-run the procedure together with the new sample. Such procedure is not usually feasible as there are not enough samples to estimate batch effects across groups and therefore correct them.

We propose here a new pipeline for integrating publicly available datasets and developing a framework for personalized medicine and the understanding of molecular pathways at the patient level. By developing a normalization and embedding technique, we show that it is possible to integrate publicly available molecular datasets, such as TCGA, SCANB, METABRIC, microarray data of some patients from the POETIC trial, patient derived xenografts and healthy breast tissue (11,19–22) into a common space, called molecular landscape. By using a special normalization and principal component analysis we can combine the datasets together explore the full molecular landscape of RNA-sequencing and microarray samples together. The new molecular landscape allows us to add new one single sample independently, showing its molecular subtype without the need of any other classification tool that has been validated in another dataset or centering techniques. It is possible to compare patients in a small neighborhood of the molecular landscape, providing a context to better understand the biology of the individual tumor. Moreover, we propose the use of different gene sets from the hallmarks collection of the Molecular Signature Database (MSigDB) (23,24), enabling understanding of individual patient tumor biology.

Alternative treatments are necessary for early stage ER+ BC patients. Currently only chemotherapy or immunotherapy is used as a second line therapy for these patients. Here we show that G2M checkpoint and PI3K AKT MTOR signaling pathways are important drivers of the molecular landscape and are prognostic in METABRIC and SCANB. Drugs are available for these pathways, but only for metastatic BC. The data suggests that early stage BC patients might benefit from these drugs.

To conclude the framework we also developed a risk score signature that depends not on a specific subset of genes but on the position of the molecular landscape and clinical factors. We show that when used in combination with the ROR score it is still statistically significant and provides additional information.

A R package is provided on github (github.com/chronchi/molecular_landscape) with instructions on how to use the newly created molecular landscape and how to integrate new samples and calculate the risk scores.

Methods

Cohorts

The breast cancer cohorts TCGA, SCANB, METABRIC and POETIC (11,19–21) were used to calculate the principal component analysis (PCA) embeddings and perform the validation. The samples from TCGA, SCANB and METABRIC are representative of the different breast cancer molecular subtypes in each respective country of the study. For these three cohorts, only samples from primary breast cancer were used. Each cohort has over 1000, 1000 and 8000 samples respectively, totalling over 10000 samples from primary breast cancer. The POETIC trial data comprises only post-menopausal women that were diagnosed with primary breast cancer and are considered to have an estrogen receptor positive status (IHC) in the initial diagnosis.

TCGA was downloaded from Firebrowse. The version 3 of SCANB data (StringTie FPKM Gene Data unadjusted) was downloaded from Mendeley https://data.mendeley.com/datasets/yzxtxn4nmd (25). METABRIC was downloaded directly from cBioPortal. POETIC data was downloaded from GEO (accession code GSE105777) using the R package GEOquery. Details about the download and preprocessing steps are described in https://chronchi.github.io/transcriptomics.

Selection of highly variable genes

In order to perform the PCA, we sub selected 1000 common and highly variable genes in the TCGA and METABRIC cohorts. For each gene and in each cohort separately, the coefficient of variation (CV) was calculated:

$$\text{CV} = \frac{\text{standard deviation}}{\text{average expression level}}.$$

The 1000 genes with highest average CV were selected.

qPCR-like normalization

Given a list of 44 stable genes across different cancers (26) and 1000 genes selected, all 1044 genes were ranked from lowest to highest expression for each sample separately and the rankings were divided by the average ranking of the 44 stable genes.

PCA embedding

Using the normalized data as described in the qPCR-like normalization, a total of 1000 random samples (fixed seed in R) coming from TCGA and METABRIC were selected to perform the initial PCA, using PCAtools (27) in R. PCA was performed without centering and scaling since the data is already centered and scaled for all genes and samples. The embedding for new individual samples is obtained by multiplying the loadings matrix with the sample normalized expression. For samples where not all the 1044 genes are available, the normalization is performed and the missing genes are padded with 0.

Scoring strategies

For the 4 big cohorts, TCGA, SCANB, METABRIC and POETIC, GSVA (28) was applied along with the $SET_{ER/PR}$ signature (29) and the hallmark collection from the molecular signature database (23,30). Default parameters were used in the `gsva` function from the GSVA package.

Average neighborhood scores

In order to calculate the posterior distribution of the average scores in each neighborhood, a linear regression with only intercept (`score ~ 1`) was fitted using `rstanarm` (31) and the function `stan_glm` for each pathway individually. When applying the function `stan_glm`, we used four chains and a prior normal distribution with location 0 and scale equals to 1. The package `tidybayes` (32) was used to extract the draws and put them in a tidy format.

Survival analysis

Survival analysis was done using cox regression with the `survival` package from R. The variables used for adjustment were age, tumor stage and number of positive lymph nodes for TCGA and SCANB. For METABRIC the variables used were Age and the Nottingham Prognostic Index (NPI). Overall survival was performed for all three cohorts and recurrence free survival for METABRIC and SCANB. The causal model used here is that the confounders affect both the outcome and the pathway of interest. It could be that there are unmeasured confounders that are affecting the pathways and outcomes at the same time, leading to a biased result. For METABRIC the variables age and Nottingham prognostic index were used for adjustment. For SCANB and TCGA the variables age, node stage and tumor stage were used for adjustment.

Risk score

We used cox regression on data of estrogen receptor positive breast cancer positive patients that received only endocrine therapy. A subset of samples from METABRIC was used as a training cohort. The rest of the samples was used as a test to evaluate the prognostic value of the signature and then SCANB was used as a validation cohort. The cox regression was calculate by having the following covariates: PC3, PC4, tumor size, node status (negative if no node, positive otherwise) and age. The risk score was then calculate as the sum of the coefficients multiplied by the respective values for each patient.

$$\begin{aligned} -0.09 \times \text{PC3} + 0.08 \times \text{PC4} + 0.02 \times \text{Tumor size} \\ +0.58 \times \text{Node status} + 0.003 \times \text{Age} \end{aligned}$$

In order to compare the risk score and its clinical utility we performed a likelihood ratio test with the full model (risk score, binary category coming from ROR, tumor size, age and node status) and the base model (without the risk score).

A risk score using only the principal components was calculated and the formula is provided below.

$$-0.09 \times \text{PC3} + 0.09 \times \text{PC4}$$

Code availability

The code used to generate all the analysis is available on https://github.com/chronchi/molecular_landscape. To fully reproduce the images in this paper check the instructions on how to run the docker image and RStudio session at the github repository. An online version with a website containing all the analysis and code can be found on https://chronchi.github.io/molecular_landscape. To check the code in the previous link click in source code on each chapter separately.

Results

Estrogen receptor signaling is a clinical continuous variable

We hypothesized that estrogen receptor (ER) signaling measured by scores from bulk RNA-seq samples is prognostic among patients that received only endocrine therapy. One way to define estrogen signaling is to calculate scores derived from gene expression levels of bulk RNA-sequencing. This provides a continuous measure for estrogen signaling that can be

used to predict the impact of endocrine therapy. We used three independent breast cancer RNA based datasets, TCGA, SCANB and METABRIC, (11,19,20) to calculate estrogen signaling scores. Each of these cohorts comprises more than 1000 untreated patient tumors . The estrogen signatures HALLMARK_ESTROGEN_RESPONSE_EARLY and HALLMARK_ESTROGEN_RESPONSE_LATE were extracted from the molecular signature database (23) (MSigDB) and $SET_{ER/PR}$ from the paper (29). The signatures from MSigDB each contain 200 genes and the latter signature has 18 genes that are associated with estrogen and progesterone receptor signaling. The individual scores for each patient sample are shown in Figure 1 (a) for each cohort stratified by estrogen receptor status. There is a difference in the scores between the estrogen receptor (ER) positive and negative subgroups for each cohort separately. The ER- has mostly negative ER signaling scores. On the other hand, the ER+ group has scores that range from negative to positive, covering the whole spectrum of possible scores. These results are reproducibly seen in all the three cohorts Figure 1 (a).

Among the ER+ BC patients we have several possible ER immuno histochemistry (IHC) percentage values. It could be that high ER IHC percentage samples are also high ER signaling score. We show that there is a small correlation between the molecular ER signaling score and ER percentage (Figure 1 (b)) and that among patients with high ER IHC percentage there is a continuum of ER signaling scores. Among those patients with high ER percentage (over 90%), the distribution of ER signaling scores is all over the spectrum, indicating a mismatch between IHC and molecular measures (Figure 1 (c)).

Next we wanted to evaluate the predictive power of the molecular ER signaling scores, given that there is a mismatch between ER IHC percentage and molecular ER signaling score based on the SCANB dataset. Cox regression was used to determine the hazard ratio (HR) of the ER signaling in overall survival (OS) for TCGA, SCANB and METABRIC and recurrence free survival (RFS) for METABRIC and SCANB. 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. To better understand the effect of estrogen receptor signaling in the clinics, samples from ER+ BC patients that received only endocrine therapy (for SCANB and METABRIC), or samples with the luminal A and B PAM50 molecular subtype (for TCGA). Such sampling may avoid a bias on the coefficients due to chemotherapy, as patients who receive chemotherapy tend to have a worse outcome compared to those that do not. Figure 1 (d) shows the forest plots for each cohort individually when calculating the hazard ratio for one of the estrogen signaling signature. In all the three cases, the hazard ratio for estrogen early was below 1, with values ranging from 0.23 to 0.61 (TCGA: 0.23, 0.05-0.99 CI; SCANB: 0.22, 0.10-0.45 CI; METABRIC: 0.56, 0.34-0.89 CI). In all three cases the upper part of the confidence intervals (CI) are below 1, with TCGA having the widest CI, possibly because we could not subset based on endocrine therapy, only by molecular subtype, and a low number of events. This shows the continuous aspect of estrogen receptor signaling. For both hallmark estrogen response late and SET ER/PR, hazard ratios were below 1 in all cases (Table S1), with a HR below 1 meaning that the higher the score the better the chances of survival for a patient. In this case in which data

from patients that received only ET was used, we could argue that patients that have higher ER signaling score have a better response to endocrine therapy.

Given that the ER percentage could have an impact in the previous results, we only selected patients that have high ER IHC percentage (above 90%) and performed survival analysis once again on the SCANB cohort (Figure 1 (e)). The HR for ER percentage is 1.01 (CI 0.88-1.03), as expected. But even among those patients the two estrogen signatures have a HR below 1 (OS: 0.26, 0.10-0.67; RFS: 0.48, 0.03-1.02). This further shows that ER IHC does not provide a complete description of endocrine therapy sensitivity and scores can replace it at those cases.

Single sample integration preserves relevant breast cancer properties

Given that not all patients with high ER IHC percentage will respond the same to endocrine therapy, we aimed to develop a pipeline and framework where one can better understand the molecular pathways of the breast cancer for a single patient given transcriptomics data. We developed a single sample batch effect removal method (see methods section) to integrate microarray and bulk RNA-seq and create a molecular landscape, which is an embedding of the molecular data into a common space for all samples. The advantage of the method is that given a new sample, it can easily be integrated with all the other previous samples without retraining.

The first step is to normalize the samples to the same scale. The average ranking of house-keeping genes distribution is similar among the two RNA-seq datasets, SCAN-B and TCGA and their mean values are higher than the mean of the distribution on METABRIC (Figure S1 (a)). The normalization preserves the distinctions of gene expression levels between ER+ and ER- breast cancer samples, ESR1 and TFF1 being such examples (Figure S1 (b)).

The biplot in Figure 2 (a) with the third and fourth components from TCGA and METABRIC samples shows that the samples are overlapping across the two cohorts. All samples, including those used for training and validation, are plotted.

In order to check the robustness of the procedure, the analysis was repeated 10 times with 10 random subsets of patient samples from TCGA and METABRIC, simulating a cross validation process. Figure 2 (b) shows the embedding of some of the validation datasets. The blue dots correspond to the original embedding of the samples, the red dots to the new embedding and the lines are connecting the same sample. Whenever a new embedding is performed, all samples are either shifted in the same direction or they are reflected along a common axis, showing the invariance properties of the embedding. (Figure 2 (b)).

Missing genes are a problem inherent to publicly available datasets. We try to understand the effect of missing genes in the embedding based on their loading values, i.e., the importance of each gene individually to the embedding. Ideally if a low number of genes with high loadings are missing, it should not affect the embedding. On the other hand, the more genes missing with high loadings, the more it will impact the embedding, as PCA takes a weighted linear

combination of the gene expression values based on the loadings. We removed 200 genes in total with a varying proportion of top loading genes (ranging from 0 to 100% with a 5% step). The number of top loading genes missing from the dataset is key for the embedding (Figure 2 (c)). The higher the proportion the less precise the embedding is, with the embedding converging towards the origin, i.e., the (0,0) coordinates (Figure 2 (c)).

The third component corresponds to the separation between ER+ and ER- BC patients in both cohorts (Figure 2 (d)). A combination of the third and fourth components shows a good distinction among the PAM50 molecular subtypes (Figure 2 (e)), showing that the embedding captures important molecular information. The fourth component mostly divides 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 the neighborhood.

Figure 2 (f) shows a gradient of the ER signaling score $SET_{ER/PR}$. 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-. Other clinical factors, such as tumor stage, node stage, age, NPI and tumor purity show no influence in the embedding (Figure S2).

Embedding generalizes to a validation cohort

So far only samples from TCGA and METABRIC were used in the training and projection. An external validation cohort is needed. SCANB was used as an external validation cohort. Similar to the previous results, SCANB is also overlapping with METABRIC and TCGA (Figure 3 (a)). ER+ and ER- BC patients are well separated (Figure 3 (b)) and the procedure can also distinguish the molecular subtypes (Figure 3 (c)) on top of the other samples coming from TCGA and METABRIC. We show that the embedding works for the SMC cohort (33) (Figure S3). As an RNA-seq cohort, it is expected that SCANB samples will be closer to TCGA than to METABRIC when removing batch effects, mostly due to platform. Biplot of PC1 and PC2 (Figure 3 (d)) shows that SCANB is closer to TCGA than to METABRIC.

The molecular landscape was also validated using patient derived xenografts (PDX) (22). These PDXs are breast cancer cells derived from patient tumor samples obtained from clinics and injected into the mammary gland of mice (34). By using the MIND model, we can engraft cells coming from ER+ BC tumors. Moreover, to have a succesfull engraftment rate, the cells need to grow so they can establish across the ducts. Therefore, the cells when extracted for RNA-sequencing are usually in a proliferative state. Figure 3 (e) shows the embedding of six different PDX samples with multiple biological replicates (22). These samples are in the scattered around the luminal B region, showing that the landscape captures the biology of the experiment and it shows the intertumor heterogeneity among ER+ BC patients in a research environment. Moreover, we used 66 samples coming from women that performed reduction mammoplasty in Switzerland. Figure 3 (f) shows that these samples are in the region of the normal-like PAM50 molecular subtype, as expected.

Molecular landscape is a tool to understand and explore patient heterogeneity

Since the molecular landscape relies in a single sample to obtain the embedding, we can add samples from any cohort. The POETIC trial (35) was a trial that evaluated the use of perioperative aromatase inhibitors in ER+, postmenopausal BC patients. Its primary endpoint was time to recurrence. Matched samples from baseline (before treatment) and at surgery (after an average 14 days of treatment) (21) were sent for microarray hybridization. There are also untreated patients, used to control for sample processing artefacts. The patients have matched Ki67 percentage levels, which can be considered an indication of how well a patient responded to the endocrine therapy. Patients with more than 5% of baseline Ki67 and a reduction of 60% upon endocrine therapy are considered responders, otherwise they are called non responders.

In order to gain insights on the differences between responders and non responders, we embedded the POETIC trial samples using the procedure and studied molecular landscape (Figure 4 (a) left). The samples are spread across the whole molecular landscape, consistent with patients having different molecular biological properties. Furthermore, the POETIC samples are embedded closer to the METABRIC samples (Fig S4). Given the available information, the BC patients that are ER+ and in the left part of the landscape (ER- patients), are all non responders (Figure 4 (a) right). Moreover, some of the non responder patients are actually in the basal region (Figure 4 (c)). We selected two samples, from patients that were responder and non responder and are close in the molecular landscape (Figure 4 (c)) to highlight their molecular differences and see what is in their context. Figure 4 (d) shows the average posterior distribution of the neighborhood for the responder patient. The responder patient has a ER signaling score higher than the average. On the other hand, the non responder has a smaller ER signaling score than the average (Figure 4 (e)) and also a higher androgen signaling score (Androgen response). Other pathways, such as EMT, E2F targets, P53 and TGFβ signaling along with their average posterior distributions are shown for both patients.

Generating hypothesis: subgroups and alternative treatments

One key aspect of drug discovery is finding the right groups for targeting with the proper drug. With the molecular landscape we can try to identify different subgroups, based on molecular pathway scores. We identify several pathways driving the distinction between samples in the embedding, such as G2M checkpoint, epithelial to mesenchymal transition, DNA repair and PI3K AKT MTOR signaling (Figure 5 (a)). The scores go from one direction to other across the two different principal components, indicating that they complement each other when capturing information. Not only G2M checkpoint is differentiating luminal A and B patients, but also EMT.

Understanding a pathway and its clinical relevance is key for drug development and subtyping. The overall and recurrence free survival calculated from patients that have ER+ BC, received only endocrine therapy (METABRIC: OS/RFS and SCANB: OS/RFS) and have PC3>0 show that G2M checkpoint has a hazard ratio higher than 1 (Figure 5 (b)) and a tight confidence

interval, which reflects the fact the higher the proliferation the worse the outcome, and reflects the subtyping differences between luminal A and B. The PI3K AKT MTOR signaling has a hazard ratio higher than 1 with a wider confidence interval, indicating a possible subgroup among all selected patients that could benefit from additional adjuvant therapy targeting this specific pathway. Current clinical trials (36) are evaluating the use of PI3K-AKT-MTOR inhibitors in advanced metastatic BC patients, here we show weak evidence on the molecular level that such treatment could benefit early stage primary BC patients.

Development of a new risk score and validation

Several risk scores are available, such as the ones from Prosigna, EndoPredict and oncotypeDX. They are all based in a small set of genes that were obtained after several statistical procedures, usually using forward variable selection on cox regression. Here we develop our own risk signature that does not do any kind of subselection, it uses all the information available on the molecular landscape.

To first see if the molecular landscape contained any clinical relevant information we used the third and fourth components separately in a cox regression, adjusted by the clinical variables mentioned in the methods section. In both overall survival and recurrence free survival analysis, the PC3 had a hazard ratio below 1 for the METABRIC, SCANB and SCANB high ER IHC percentage (more than 90%) BC patients that received only endocrine therapy. On the other hand, PC4 had hazard ratios higher than 1 in all cohorts for both OS and RFS (Figure 6 (a)). To further validate the clinical aspect of the molecular landscape, we calculated the euclidean distance between the surgery and baseline matched samples of the POETIC trial data. The average distance of the embedding position is higher in responders than non-responders (Figure 6 (b)).

Based on Figure 6 (a), one can hypothesize that there is a direction in the molecular landscape where patients would be more at risk. We developed a new risk score based on clinical variables (tumor size, age and node status) and the position in the molecular landscape (PC3 and PC4 scores). Table 1 shows the results of the recurrence free survival analysis performed using these variables as coefficients and ER+ BC patients treated with ET only. The risk score developed follows a gradient from the bottom right of the molecular landscape to the upper left of the embedding (Figure 6 (c)) among all the patients treated with endocrine therapy only and that have ER+ BC. To further validate the validate the clinical utility of the score, we compared the new developed risk score with the risk of recurrence score available from the SCANB patients calculated by the nearest centroid technique (15). We see that there is a positive correlation among the scores, moreover there is a population of luminal A patients that present a high risk score that is not captured by the ROR, as by definition luminal A patients are of low/intermediate risk. We further showed that by using the binary categorization of the ROR into high and low/intermediate risk groups (15) there is additional benefit in using the new risk score in a survival analysis (p-value = 0.02). Figure 6 (e) shows that there is a distinction in the distributions of the low/intermediate and high risk patients, but even among

the low/intermediate there are those with risk scores similar to those of high risk, which are by majority luminal B. As another way of validating the risk score, we calculated the risk scores using only the principal components on the POETIC trial data. Responders have a decrease in average in the risk score. Non-responders have almost the same risk score (Figure 6 (f)).

Discussion

Personalized medicine is a key topic in medicine and BC (22). The goal of better understanding the molecular underpinnings of the diseases leads to a better allocation of treatments and resources in the patient care. This has been shown to be necessary by using PDX mouse models, where different PDXs respond differently to hormone treatments (22). We have shown here a possible framework to deal with personalized medicine in breast cancer in general with a focus on ER+ BC patients.

Estrogen receptor status is defined as a clinical variable that usually have two or three categories (3). Breast cancers are classified either in ER+ or ER- based on their protein expression levels and IHC. ER+ are those tumors that express more than 1% and for those expressing less than 1% they are considered ER-. This definition depends on the country and guidelines that are used. For example, the ASCO guidelines (3) recommends to use a threshold of 1%, whereas the Swedish national guidelines uses a threshold of 10% for ER status (37). For those tumors that are ER+, they can be subdivided into low ER positive ($1\% < \text{IHC}\% < 10\%$) and simply ER positive. The low ER positive tumors do not usually benefit as much on endocrine therapy. Here we show using cox regression and multiple big study cohorts (11,19,20) with both RNA-seq and microarray data, that ER status is more of a continuous rather than a categorical state. This can aid when evaluating treatment options and avoiding overtreatment. Several gene signatures already are being used, such as OncotypeDX, Mammaprint and Prosigna to assign chemotherapy for those patients with higher risk of recurrence (5–7). This score might be associated with the commercial signatures, as it has been shown that OncotypeDX's estrogen module is highly correlated with the general OncotypeDX signature (10).

Integrating molecular data stemming from different sources is a challenge. On one hand batch effect tools are usually able to remove the batch effects across the different sources of variability (16,17), on the other hand they are not single sample based, meaning each time a new sample comes the algorithm needs to be run again. It is also based on the fact one has enough data in the different datasets, otherwise it skews the possible integration towards one of the datasets. Here we show by using TCGA, METABRIC and SCANB that it is possible to integrate the samples from these cohorts in a single sample manner. The samples show good mixing when using test samples not seen during the training stage. The embeddings preserve key molecular features of breast cancer. PC3 is clearly driven by estrogen receptor signaling where from right to left there is a gradient in the ER scores, such as Estrogen early and $SET_{ER/PR}$. On the

other hand, PC4 is what makes a difference between the molecular subtypes luminal A and B, which in practice differ by proliferation status in terms of Ki67 levels (38).

Moreover, the embeddings in the validation set (SCANB) preserve the key features of breast cancer. Samples are projected by their different PAM50 molecular subtypes and there is a gradient of estrogen signaling pathway from ER+ BC towards ER- BC patients in all the three cohorts combined. The first two components are the batch effect removal components and SCANB is projected closer to TCGA, since both datasets are RNA-seq. On the other hand, part of the POETIC trial was sequenced using microarray dataset (21,35), and we show that it is projected closer to METABRIC as expected according to the first two components. This highlights that the method can capture information from different technologies and remove such batch effects. Moreover all the samples were sent for sequencing in different contexts at different times and populations, showing the power of the embedding method in removing batch effects and capturing truly the biology of breast cancer.

Sometimes when dealing with publicly available datasets, not all of the genes are available due to ethical protocols (11), pre-processing or technical reasons. Therefore we showed how robust the projection is to missing genes with high loadings in the projection. If less than 20% of the genes we can recover almost surely the position of the embedding if all the genes are available. The more genes that are missing, the closer the projection will be to the origin, i.e., the (0,0) coordinate in the embedding of the third and fourth components of the PCA.

To show the clinical validity of the projection, we used a subset of patients from the POETIC trial with microarray and clinical information (21). The samples are projected as expected and surprisingly the samples that are considered to be non responders upon 2 weeks of aromatase inhibition are projected among the ER- BC patients. When looking further upon two different patients that have similar embedding but different response to endocrine therapy, we see that the responder has a higher value of estrogen signaling than the average. On the other hand, the non responder patient has a smaller estrogen signaling score than the average, which suggests a possible explanation for the difference in response. Moreover, estrogen and androgen receptor signaling have been show to be tightly linked (39) and these two patients have different androgen signaling scores. The non responder has a higher score than the average compared to the responder, whose score is just the same as the average.

Complementing the personalized approach, the molecular landscape can be used as a starting point to understand breast cancer molecular subgroups in a more intuititive way using pathways. Using survival analysis we show that some of the pathways that are important for the embedding can be used for subgrouping and hypothesis generation.

A weaknesses of the proposed pipeline is that we rely on GSVA scores, which can be used and compared across different cohorts since they have a representation of all molecular subtypes. A way to circumvent this is by having a small library of RNAseq or microarray samples that are representative of the patient population. This way when scoring a new patient, the scores can be compared across different cohorts. There is still a cost barrier for using RNAseq dataset in the clinical setting, but efforts are being made across the industry to reduce the

sequencing costs and make it more widespread. An example is Alithea genomics, a company aiming to provide large-scale RNA-sequencing by using Bulk RNA barcoding and sequencing (BRB-seq).

In conclusion we provide some evidence that ER status should be considered as a continuous marker rather than a dichotomous one. We also extend this notion to a framework for personalized medicine, where each patient is embedded in some context that can be used for the interpretation of its molecular underpinnings. In this paper we only discussed some pathways, but when analysing patient data, several pathways can be considered and should be taken into account when deciding tumor treatment.

References

1. World Health Organization: Regional Office for Europe. World cancer report. IARC; 2020.
2. Anderson WF, Chatterjee N, Ershler WB, Brawley OW. Estrogen receptor breast cancer phenotypes in the surveillance, epidemiology, and end results database. Breast Cancer Research and Treatment [Internet]. Springer Science; Business Media LLC; 2002;76:27–36. Available from: https://doi.org/10.1023/a:1020299707510
3. Allison KH, Hammond MEH, Dowsett M, McKernin SE, Carey LA, Fitzgibbons PL, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. Journal of Clinical Oncology [Internet]. American Society of Clinical Oncology (ASCO); 2020;38:1346–66. Available from: https://doi.org/10.1200/jco.19.02309
4. Dieci MV, Griguolo G, Bottosso M, Tsvetkova V, Giorgi CA, Vernaci G, et al. Impact of estrogen receptor levels on outcome in non-metastatic triple negative breast cancer patients treated with neoadjuvant/adjuvant chemotherapy. npj Breast Cancer [Internet]. Springer Science; Business Media LLC; 2021;7. Available from: https://doi.org/10.1038/s41523-021-00308-7
5. Cardoso F, Veer LJ van't, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-gene signature as an aid to treatment decisions in early-stage breast cancer. New England Journal of Medicine [Internet]. Massachusetts Medical Society; 2016;375:717–29. Available from: https://doi.org/10.1056/nejmoa1602253
6. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology [Internet]. American Society of Clinical Oncology (ASCO); 2009;27:1160–7. Available from: https://doi.org/10.1200/jco.2008.18.1370
7. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. New England Journal of Medicine [Internet]. Massachusetts Medical Society; 2018;379:111–21. Available from: https://doi.org/10.1056/nejmoa1804710
8. Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer CF, et al. A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors. Clinical Cancer Research [Internet]. American Association for Cancer Research (AACR); 2011;17:6012–20. Available from: https://doi.org/10.1158/1078-0432.ccr-11-0926
9. Jankowitz RC, Cooper K, Erlander MG, Ma X-J, Kesty NC, Li H, et al. Prognostic utility of the breast cancer index and comparison to adjuvant! Online in a clinical case series of early breast cancer. Breast Cancer Research [Internet]. Springer Science; Business Media LLC; 2011;13. Available from: https://doi.org/10.1186/bcr3038

10. Buus R, Sestak I, Kronenwett R, Ferree S, Schnabel CA, Baehner FL, et al. Molecular drivers of oncotype DX, prosigna, EndoPredict, and the breast cancer index: A TransATAC study. Journal of Clinical Oncology [Internet]. American Society of Clinical Oncology (ASCO); 2021;39:126–35. Available from: https://doi.org/10.1200/jco.20.00853

11. 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

12. Dihge L, Vallon-Christersson J, Hegardt C, Saal LH, Häkkinen J, Larsson C, et al. Prediction of lymph node metastasis in breast cancer by gene expression and clinicopathological models: Development and validation within a population-based cohort. Clinical Cancer Research [Internet]. American Association for Cancer Research (AACR); 2019;25:6368–81. Available from: https://doi.org/10.1158/1078-0432.ccr-19-0075

13. Brueffer C, Gladchuk S, Winter C, Vallon-Christersson J, Hegardt C, Häkkinen J, et al. The mutational landscape of the <scp>SCAN</scp> -b real-world primary breast cancer transcriptome. EMBO Molecular Medicine [Internet]. EMBO; 2020;12. Available from: https://doi.org/10.15252/emmm.202012118

14. Dahlgren M, George AM, Brueffer C, Gladchuk S, Chen Y, Vallon-Christersson J, et al. Preexisting somatic mutations of estrogen receptor alpha (ESR1) in early-stage primary breast cancer. JNCI Cancer Spectrum [Internet]. Oxford University Press (OUP); 2021;5. Available from: https://doi.org/10.1093/jncics/pkab028

15. Staaf J, Häkkinen J, Hegardt C, Saal LH, Kimbung S, Hedenfalk I, et al. RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer. npj Breast Cancer [Internet]. Springer Science; Business Media LLC; 2022;8. Available from: https://doi.org/10.1038/s41523-022-00465-3

16. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology [Internet]. Springer Science; Business Media LLC; 2014;32:896–902. Available from: https://doi.org/10.1038/nbt.2931

17. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: Batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics [Internet]. Oxford University Press (OUP); 2020;2. Available from: https://doi.org/10.1093/nargab/lqaa078

18. Fei T, Zhang T, Shi W, Yu T. Mitigating the adverse impact of batch effects in sample pattern detection. Birol I, editor. Bioinformatics [Internet]. Oxford University Press (OUP); 2018;34:2634–41. Available from: https://doi.org/10.1093/bioinformatics/bty117

19. 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

20. 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

21. Gao Q, Elena López-Knowles and, Cheang MCU, Morden J, Ribas R, Sidhu K, et al. Impact of aromatase inhibitor treatment on global gene expression and its association with antiproliferative response in ER+ breast cancer in postmenopausal patients. Breast Cancer Research [Internet]. Springer Science; Business Media LLC; 2019;22. Available from: https://doi.org/10.1186/s13058-019-1223-z

22. Scabia V, Ayyanan A, Martino FD, Agnoletto A, Battista L, Laszlo C, et al. Estrogen receptor positive breast cancers have patient specific hormone sensitivities and rely on progesterone receptor. Nature Communications [Internet]. Springer Science; Business Media LLC; 2022;13. Available from: https://doi.org/10.1038/s41467-022-30898-0

23. 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

24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics [Internet]. Oxford University Press (OUP); 2011;27:1739–40. Available from: https://doi.org/10.1093/bioinformatics/btr260

25. Johan Vallon-Christersson. RNA sequencing-based single sample predictors of molecular subtype and risk of recurrence for clinical assessment of early-stage breast cancer [Internet]. Mendeley; 2023. Available from: https://data.mendeley.com/datasets/yzxtxn4nmd/3

26. Bhuva DD, Cursons J, Davis MJ. Stable gene expression for normalisation and single-sample scoring. Nucleic Acids Research [Internet]. Oxford University Press (OUP); 2020;48:e113–3. Available from: https://doi.org/10.1093/nar/gkaa802

27. Blighe K, Lun A. PCAtools: PCAtools: Everything principal components analysis [Internet]. 2022. Available from: https://github.com/kevinblighe/PCAtools

28. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics [Internet]. Springer Science; Business Media LLC; 2013;14. Available from: https://doi.org/10.1186/1471-2105-14-7

29. 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

30. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics [Internet]. Springer Science; Business Media LLC; 2003;34:267–73. Available from: https://doi.org/10.1038/ng1180

31. Goodrich B, Gabry J, Ali I, Brilleman S. Rstanarm: Bayesian applied regression modeling via Stan. [Internet]. 2022. Available from: https://mc-stan.org/rstanarm/

32. Kay M. tidybayes: Tidy data and geoms for Bayesian models [Internet]. 2022. Available from: http://mjskay.github.io/tidybayes/

33. Kan Z, Ding Y, Kim J, Jung HH, Chung W, Lal S, et al. Multi-omics profiling of younger asian breast cancers reveals distinctive molecular signatures. Nature Communications [Internet]. Springer Science; Business Media LLC; 2018;9. Available from: https://doi.org/10.1038/s41467-018-04129-4

34. Sflomos G, Dormoy V, Metsalu T, Jeitziner R, Battista L, Scabia V, et al. A preclinical model for ERα-positive breast cancer points to the epithelial microenvironment as determinant of luminal phenotype and hormone response. Cancer Cell [Internet]. Elsevier BV; 2016;29:407–22. Available from: https://doi.org/10.1016/j.ccell.2016.02.002

35. Dowsett M, Smith I, Robertson J, Robison L, Pinhel I, Johnson L, et al. Endocrine therapy, new biologicals, and new study designs for presurgical studies in breast cancer. JNCI Monographs [Internet]. Oxford University Press (OUP); 2011;2011:120–3. Available from: https://doi.org/10.1093/jncimonographs/lgr034

36. AstraZeneca S. A Phase III Double-blind Randomised Study Assessing the Efficacy and Safety of Capivasertib + Fulvestrant Versus Placebo + Fulvestrant as Treatment for Locally Advanced (Inoperable) or Metastatic Hormone Receptor Positive, Human Epidermal Growth Factor Receptor 2 Negative (HR+/HER2-) Breast Cancer Following Recurrence or Progression On or After Treatment With an Aromatase Inhibitor. https://clinicaltrials.gov/ct2/show/NCT04305496; 2020.

37. Nationellt vårdprogram Bröstcancer . http://www.swebcg.se/wp-content/uploads/2016/09/nationellt-vardprogram-brostcancer.pdf; 2019.

38. Arima N, Nishimura R, Osako T, Okumura Y, Nakano M, Fujisue M, et al. Ki-67 index value and progesterone receptor status can predict prognosis and suitable treatment in node-negative breast cancer patients with estrogen receptor-positive and HER2-negative tumors. Oncology Letters [Internet]. Spandidos Publications; 2018; Available from: https://doi.org/10.3892/ol.2018.9633

39. Hickey TE, Selth LA, Chia KM, Laven-Law G, Milioli HH, Roden D, et al. The androgen receptor is a tumor suppressor in estrogen receptor–positive breast cancer. Nature Medicine [Internet]. Springer Science; Business Media LLC; 2021;27:310–20. Available from: https://doi.org/10.1038/s41591-020-01168-7

Table 1: Coefficients of the recurrence free survival analysis performed using cox regression on a subset of METABRIC samples coming from ER+ BC patients treated with only ET. CI: confidence interval.

Coefficient	Hazard Ratio	Std error	p-value	CI low	CI high
PC3	0.91	0.0273	7.8e-04	0.87	0.96
PC4	1.08	0.0226	3.6e-04	1.04	1.13
tumor_size	1.02	0.0043	3.7e-07	1.01	1.03
node_statuspos	1.79	0.1476	7.9e-05	1.34	2.39
age	1.00	0.0070	6.5e-01	0.99	1.02

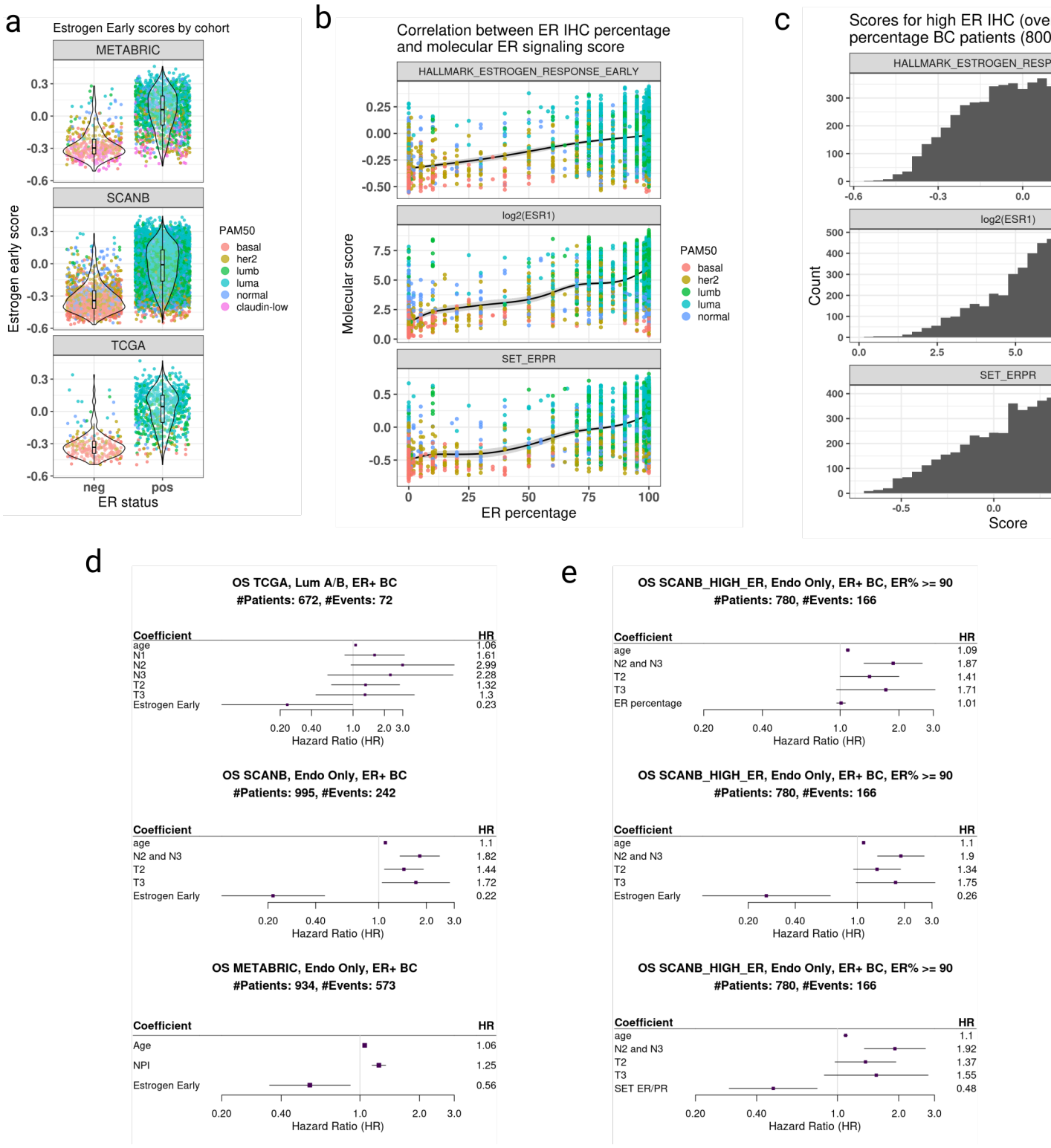

Figure 1: Scores and survival analysis results from TCGA, SCANB and METABRIC cohorts. (a) GSVA scores for the SET ER/PR signature for each cohort. Each point corresponds to a patient sample and they are divided by estrogen receptor status. (b) Correlation of ER percentage with molecular signatures on the SCANB cohort. (c) Distribution of the molecular scores among patients with high ER percentage (more than 90%) from SCANB. (d) Forest plot of the survival analysis for each cohort separately. (e) Forest plot of the survival analysis for high ER percentage patients that were treated only with endocrine therapy from SCANB. NPI: Nottingham prognostic index. Ti: i-th stage of tumor. Ni: i lymph nodes with breast cancer cells.

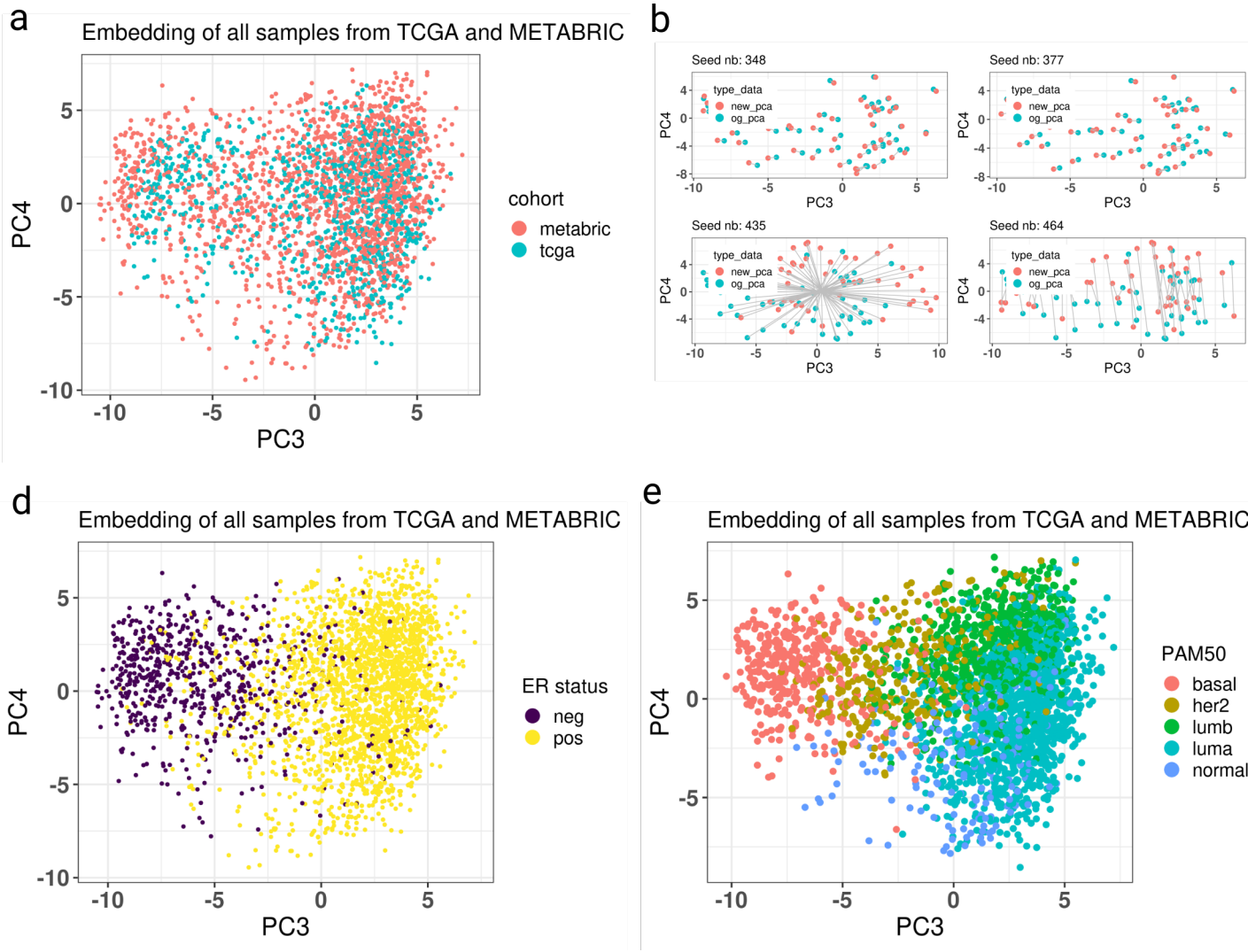

Figure 2: (a) Biplot using the third and fourth components on TCGA and METABRIC samples. Colored by cohort. (b) Embedding of random samples given different training sets for PCA. Blue dots correspond to the original embedding of a sample and red dots correspond to the new embedding given the new training set. (c) Biplot of all possible embeddings of sample given a certain proportion of top loadings missing in the dataset. (d) Same as **a**, colored by ER status. (e) Same as **a**, colored by PAM50 molecular subtype. (f) Hex grid calculated on the biplot of the fourth and third component. Each hex is colored based on its average value of the estrogen response early.

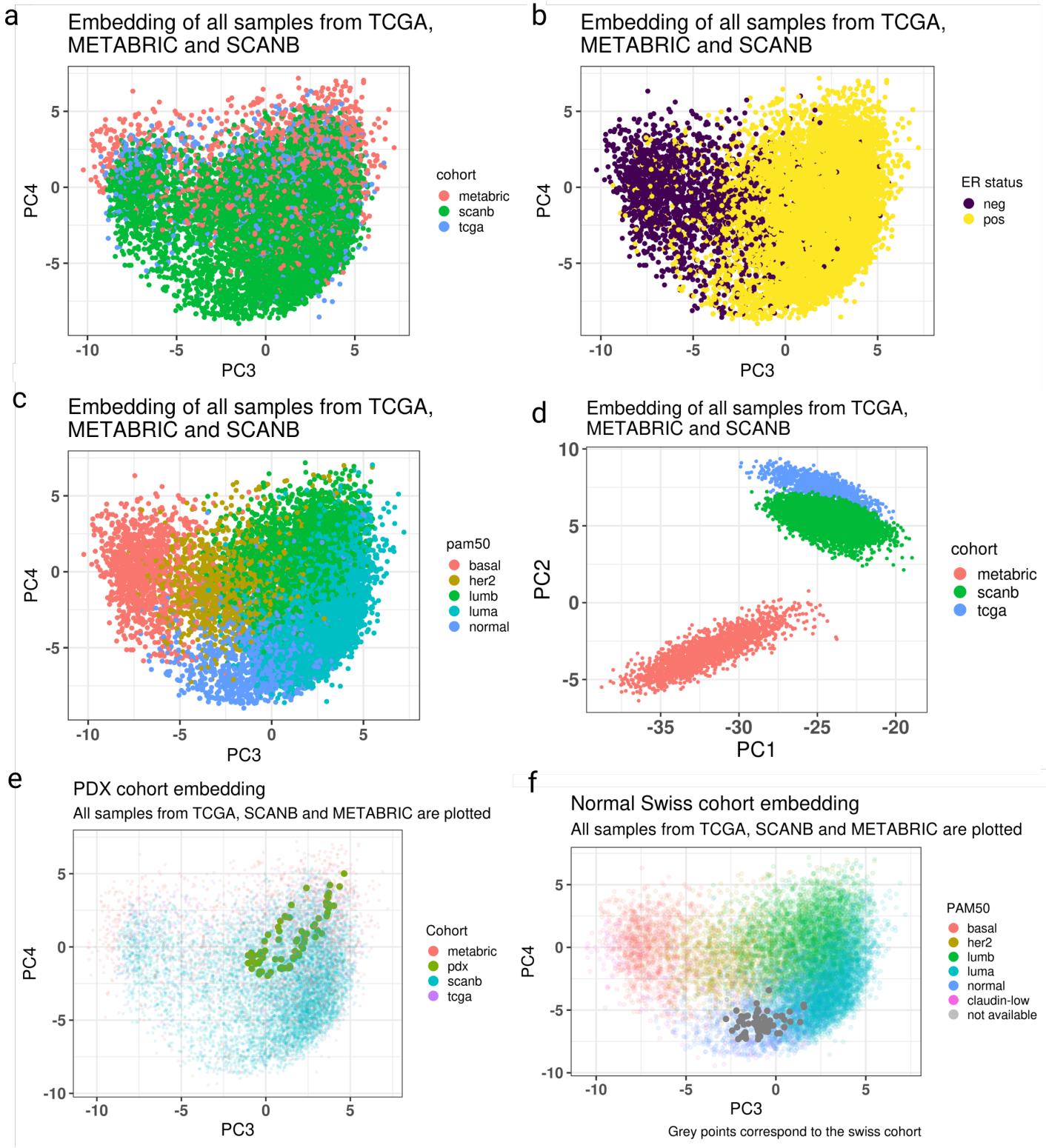

Figure 3: Validation of the molecular landscape with an external cohort (a) Biplot using the third and fourth components and now including all samples from the three cohorts: TCGA, METABRIC and SCANB. (b) Same as **a**, colored by ER status. (c) Same as **a**, colored by PAM50 molecular subtype. (d) Biplot using the first and second component of TCGA, METABRIC and SCANB. (e) Biplot of the PDXs on top of all three big cohorts. (f) Biplot of the normal samples on top of the three big cohorts colored by PAM50 molecular subtype.

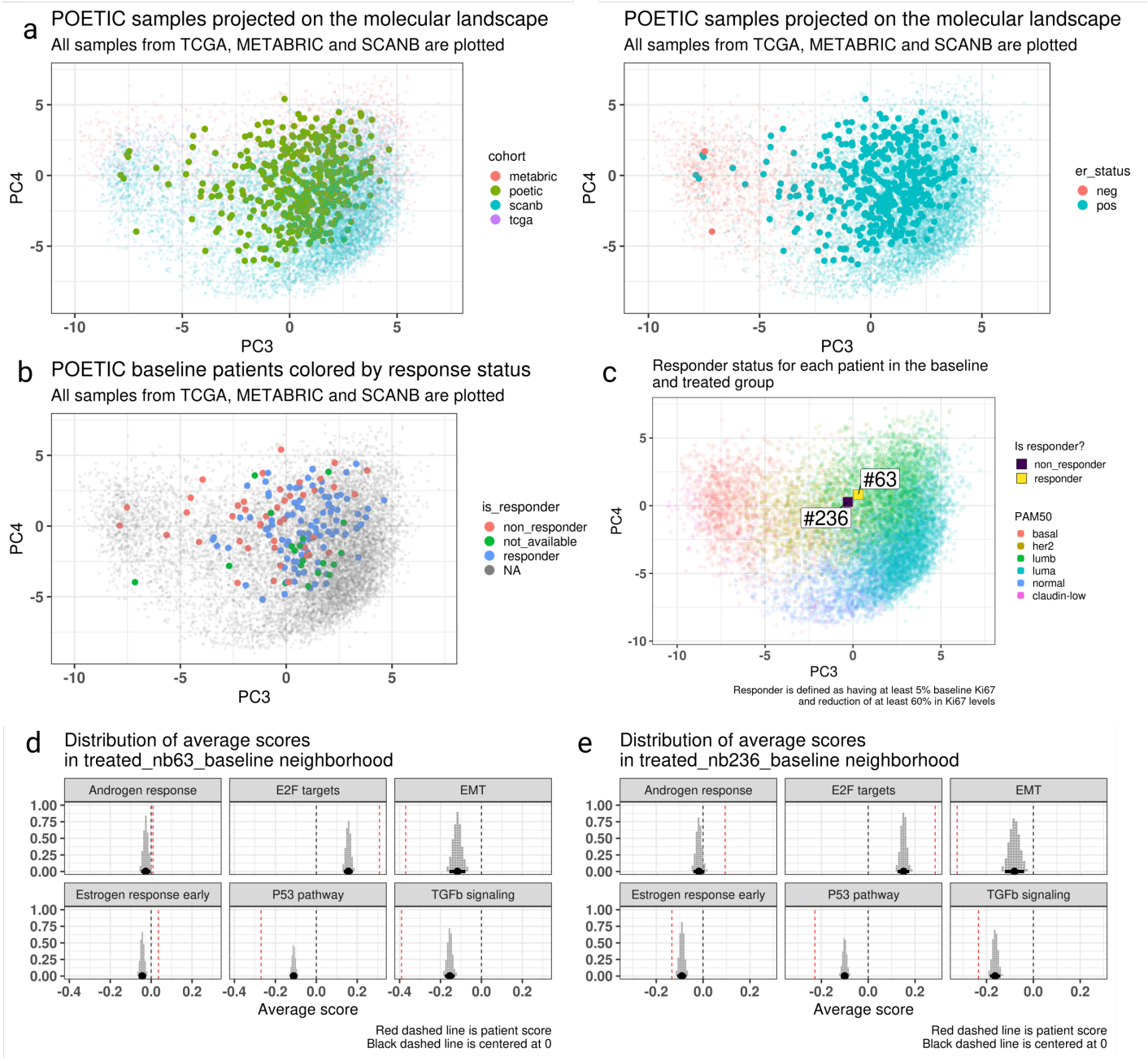

Figure 4: Embedding of the POETIC cohort into the molecular landscape and pathway analysis for patient samples. (a) Biplots of the POETIC samples (baseline and surgery) into the molecular landscape. Left plot is colored by cohort and right plot is colored by ER status. (b) Biplot of the POETIC sample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

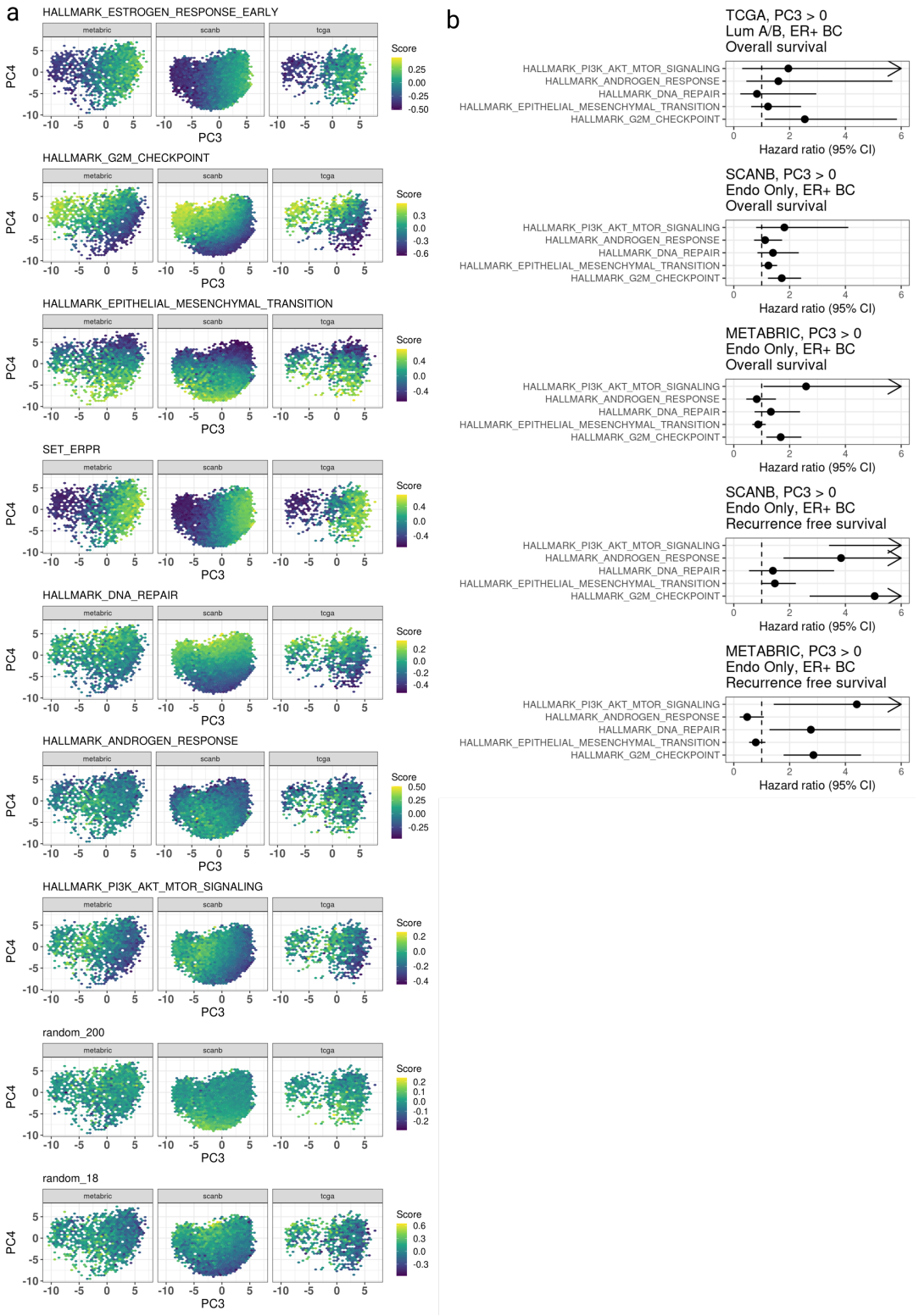

Figure 5: Average scores in hex regions using all three cohorts together and survival analysis in each cohort individually. (a) Biplots of all samples from TCGA, SCANB and METABRIC that were grouped in small hex regions. Colors are depicted as the average score value in each hex region. Selected pathways are shown. (b) Survival analysis results obtained for each pathway individually. The pathways were adjusted as described in the methods section. CI: confidence interval.

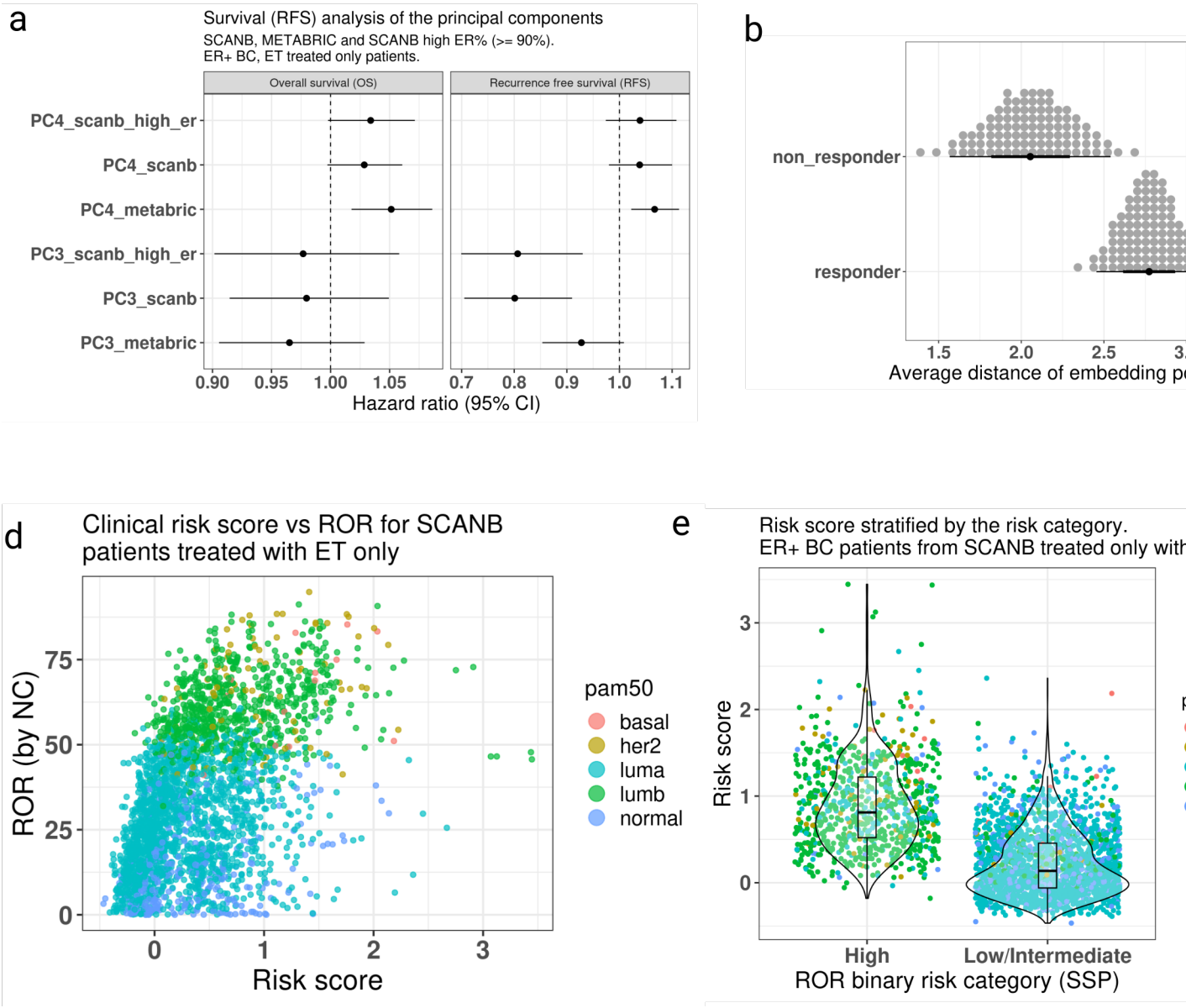

Figure 6: Average scores in hex regions using all three cohorts together and survival analysis in each cohort individually. (a) Biplots of all samples from TCGA, SCANB and METABRIC that were grouped in small hex regions. Colors are depicted as the average score value in each hex region. Selected pathways are shown. (b) Survival analysis results obtained for each pathway individually. The pathways were adjusted as described in the methods section. CI: confidence interval.