# Abstract

# Introduction

# Methods

## Cohorts

The breast cancer cohorts TCGA, SCANB, METABRIC and POETIC (1–4) were used to calculate the PCA embeddings and scores. The samples are of primary breast cancer samples from different molecular subtypes and age distribution. TCGA was download from Firebrowse, SCANB data was downloaded from GEO (accession code GSE96058) using the package GEOquery (5) in R along the clinical data from the https://oncogenomics.bmc.lu.se/MutationExplorer/ website. METABRIC was downloaded directly from cBioPortal. POETIC data was downloaded from GEO (accession code GSE105777) using the R package GEOquery. More 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 standard deviation of that gene was calculated. Then an average standard deviation was calculated using the formula below.

$$\text{average sd} = \sqrt{\frac{(sd^2_{TCGA} + sd^2_{METABRIC})}{k}}$$

## qPCR-like normalization

Since samples are coming from different platforms, they need to be scaled in a way that they are comparable. For this we developed a new way to scale the data based on the ranking of the samples. Given a list of 44 stable genes across different cancers (6) and the 1000 genes selected previously, all genes were ranked from lowest to highest expression for each sample separately and the rankings were divided by the average ranking of the stable genes. Stable genes are considered the housekeeping genes.

### PCA embedding

Using the normalized data, a total of 1000 random samples coming from TCGA and METABRIC were selected to perform the initial PCA. This is an unsupervised learning method, therefore there is no need to label the samples with respect to some category. The package PCAtools (7) in R was used to perform the PCA and to obtain the loadings for downstream analysis. The embedding for new individual samples is obtained by multiplying the loadings matrix with the normalized data for that sample. Since the normalization procedure is performed sample-wise, this step is independent of the number of samples. If there are missing genes in a sample, 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 (8) was applied along with the $SET_{ER/PR}$ signature (9) and the hallmark collection from the molecular signature database (10,11). 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 witho only intercept was fit using `rstanarm` (12). The package `tidybayes` (13) was used to extract the draws and put them in a tidy format.

### Code availability

The code used to generate all the analysis is available on https://github.com/chronchi/molecular_landscape. Descriptions for a docker image to reproduce the analysis are available on the github repository.

## 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 (10) and $SET_{ER/PR}$ from (9). The individual scores for each patient sample

are shown in Figure 1 (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 (b) shows the forest plots for each cohort individually when calculating the hazard ratio for the $SET_{ER/PR}$ estrogen signaling signature. In all the three cases, the hazard ratio for $SET_{ER/PR}$ 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.

## 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 (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 (b)). A combination of the third and fourth components shows a good distinction among the PAM50 molecular subtypes (Figure 2 (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 (d) 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-.

## 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

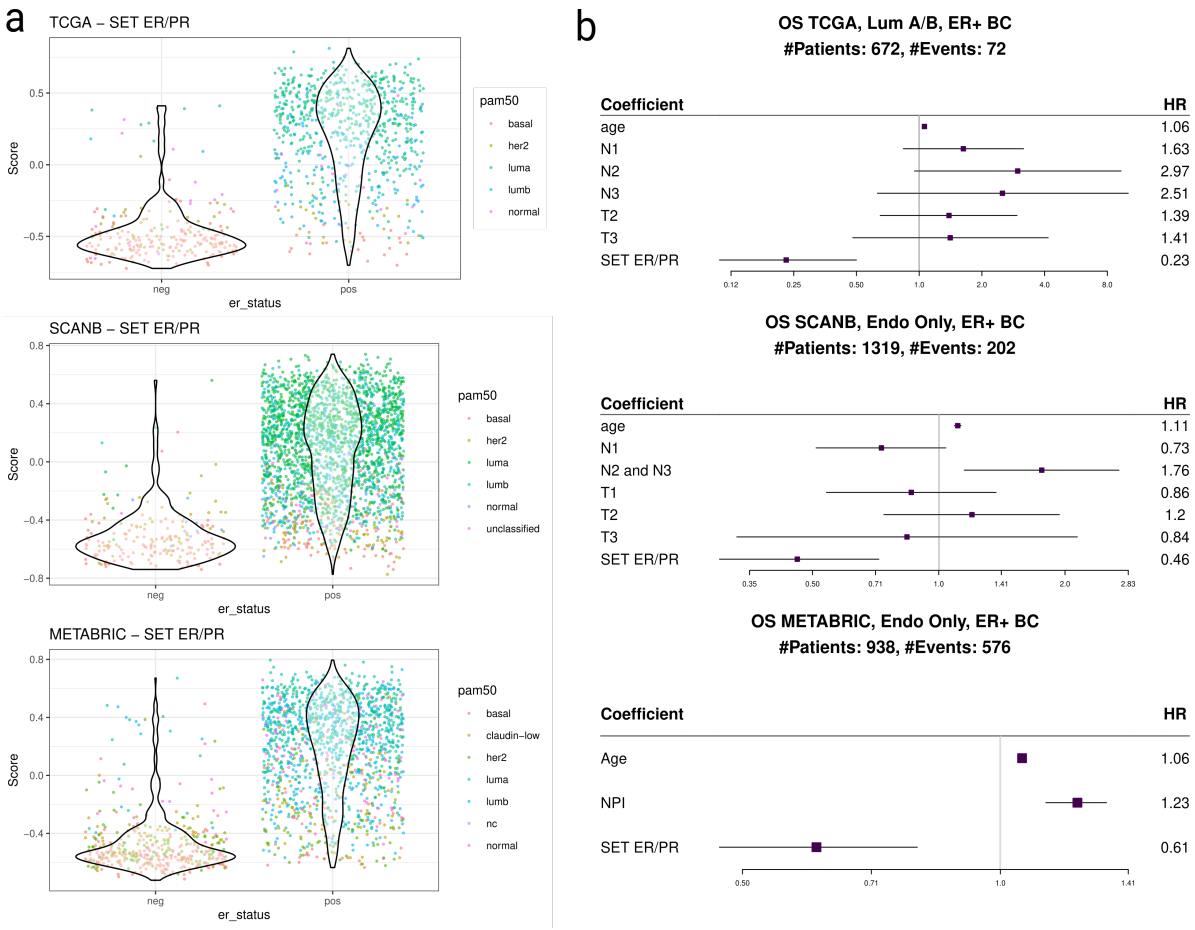


Figure 1: Scores and survival analysis results from TCGA, SCANB and METABRIC cohorts. (a) GSVA scores for the SET ER/PR signature for each cohort. Each point corresponds to a patient sample and they are divided by estrogen receptor status. (b) Forest plot of the survival analysis for each cohort separately.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

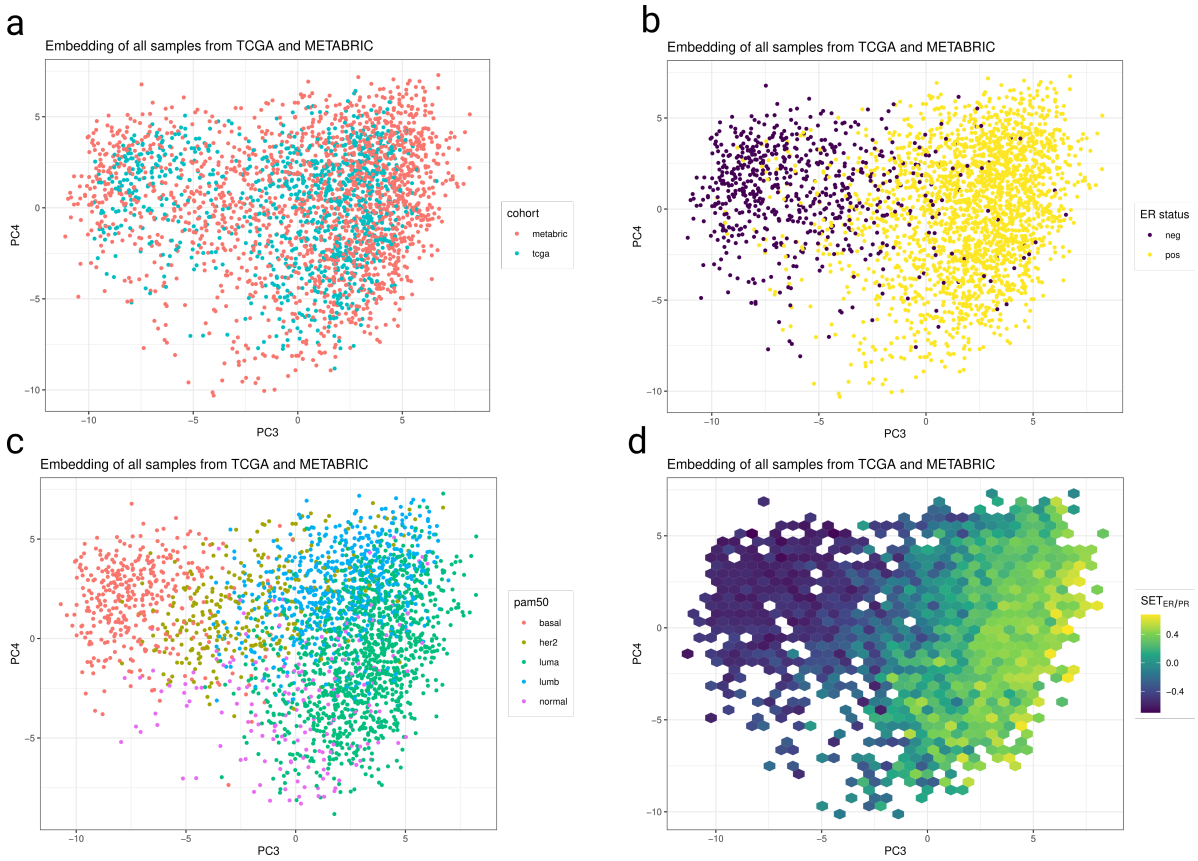


Figure 2: (a) Biplot using the third and fourth components on TCGA and METABRIC samples. Colored by cohort. (b) Same as **a**, colored by ER status. (c) Same as **a**, colored by PAM50 molecular subtype. (d) Hex grid calculated on the biplot of the fourth and third component. Each hex is colored based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

samples (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)). 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 (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 (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 (f)). The higher the proportion the less precise the embedding is.

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

Since the molecular landscape relies in a single sample embedding, we can add samples from any cohort with relative good data. The POETIC trial (14) was a trial that evaluated the use of perioperative aromatase inhibitors in ER+, postmenopausal BC patients. Its primary endpoint was time to recurrence. They sent for microarray hybridization matched samples from baseline (before treatment) and at surgery (after an average of 14 days of treatment) (4). There are also untreated patients, used to control for sample processing artefacts. Moreover, 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.

The molecular landscape can shed light on the differences between responders and non responders. We embedded the POETIC trial samples using the procedure (Figure 4 (a) left). The samples are spread across the whole molecular landscape, showing that patients indeed have different molecular biological properties. Moreover, given the available information, the patients that are ER+ and in the left part of the landscape (ER- patients), are all non responders (Figure 4 (a) right). This highlights the importance to look more carefully to ER+ patients. We selected two patients, a responder and non responder that are close in the embedding (Figure 4 (b)) to highlight their molecular differences and see what is their context. Figure 4 (c) 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 (d)) and also

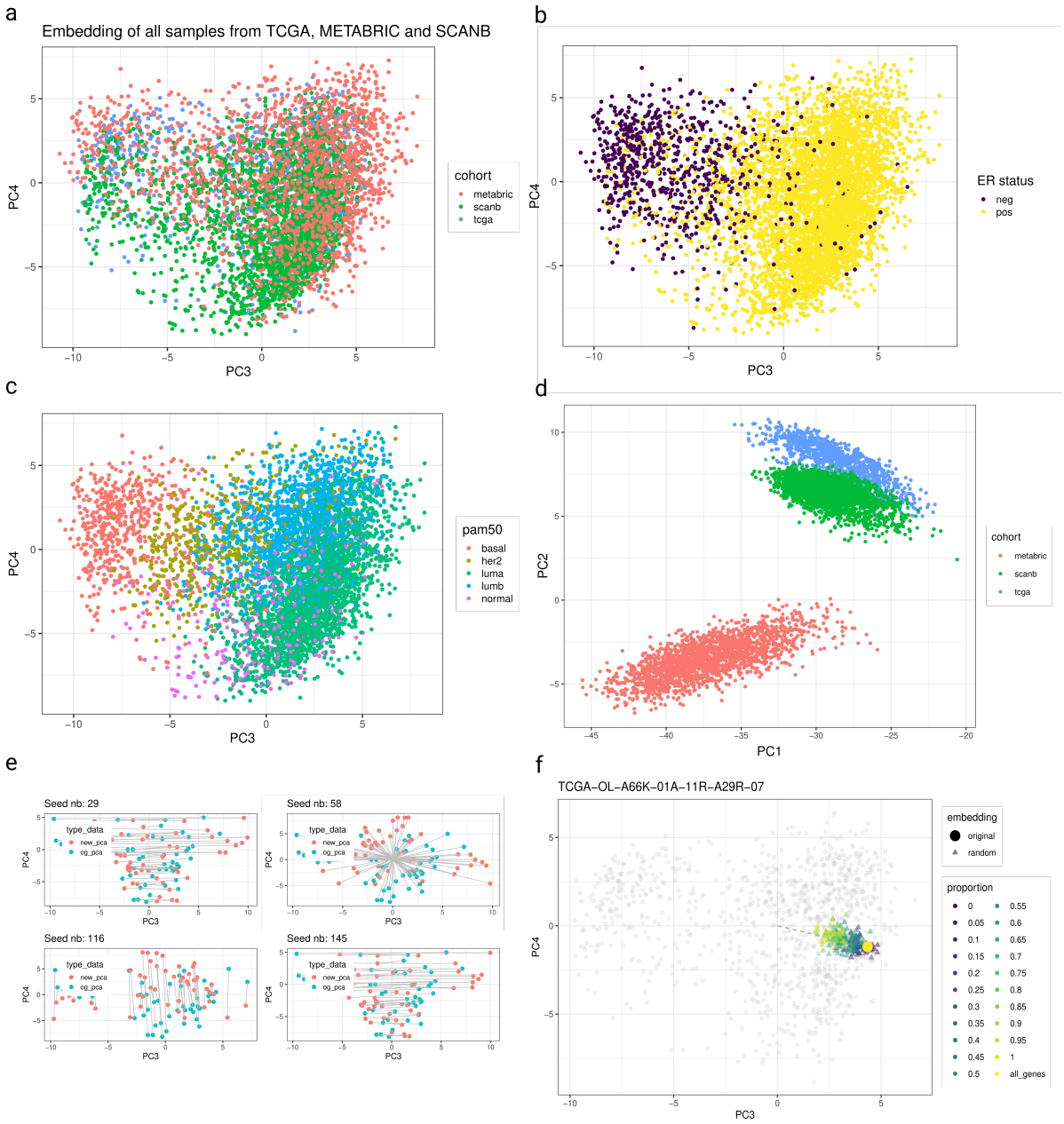


Figure 3: Validation of the molecular landscape with an external cohort (a) Biplot using the third and fourth components and now including all samples from the three cohorts: TCGA, METABRIC and SCANB. (b) Same as **a**, colored by ER status. (c) Same as **a**, colored by PAM50 molecular subtype. (d)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

a higher androgen signaling score (Androgen response). Other pathways and their average posterior distributions are shown for both patients.

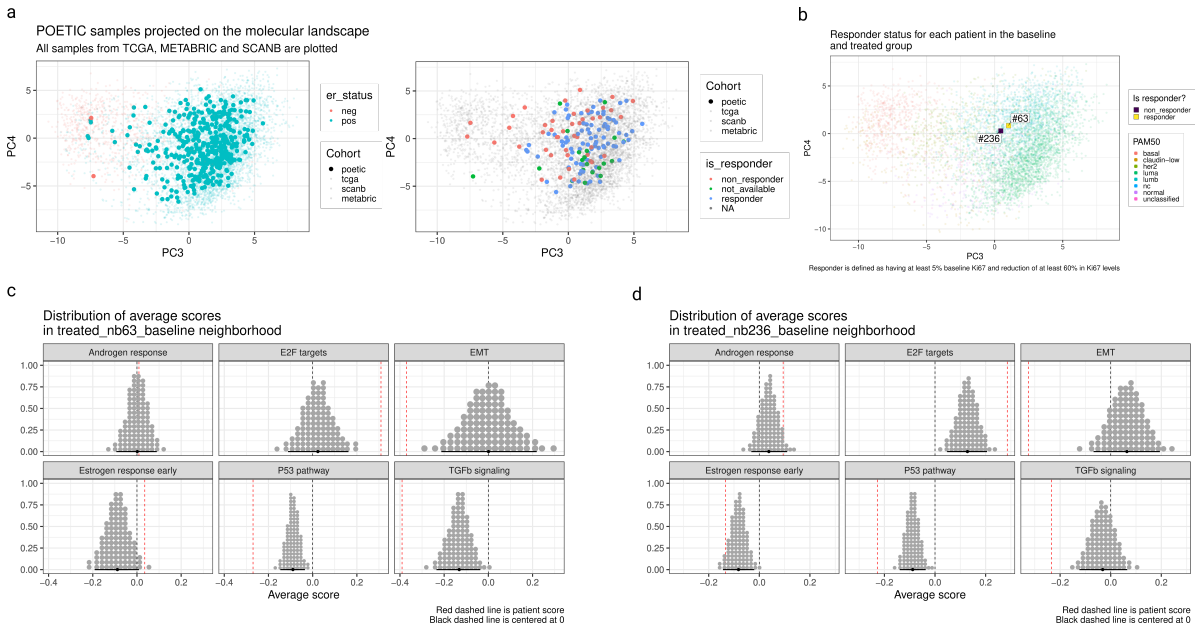


Figure 4: Embedding of the POETIC cohort into the molecular landscape and pathway analysis for patient samples. (a) Biplots of the POETIC samples (baseline and surgery) into the molecular landscape. Left plot is colored by ER status and right plot is colored by molecular subtype PAM50 when available. (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. 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

2. 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

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. 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

5. Davis S, Meltzer P. GEOquery: A bridge between the gene expression omnibus (GEO) and BioConductor. Bioinformatics. 2007;14:1846–7.

6. 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

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

8. 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

9. 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

10. 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

11. 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

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

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

14. 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