Heterogeneity in breast cancer:

Integration of cell-patient data to tackle tamoxifen resistance

Mathematical Biology Research Group Talks
31st March 2023 - Bayes Centre

bcam
basque center for applied mathematics

EXCELENCIA
SEVERO
OCHOA

CIC bioGUNE
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Ph.D. Student in Mathematics and Statistic at the University of the Basque Country

Supervisors:

Prof. Elena Akhmastkaya - Group leader in Modeling and Simulation in Life and Material Sciences at Basque Center of Applied Mathematics

Dra. María Vivanco - Group leader in the Cancer Heterogeneity Lab at CICbioGUNE

Currently visiting in the University of Edimburgh:

Dr. Victor Elvira - Reader in Statistics and Data Science at the School of Mathematics

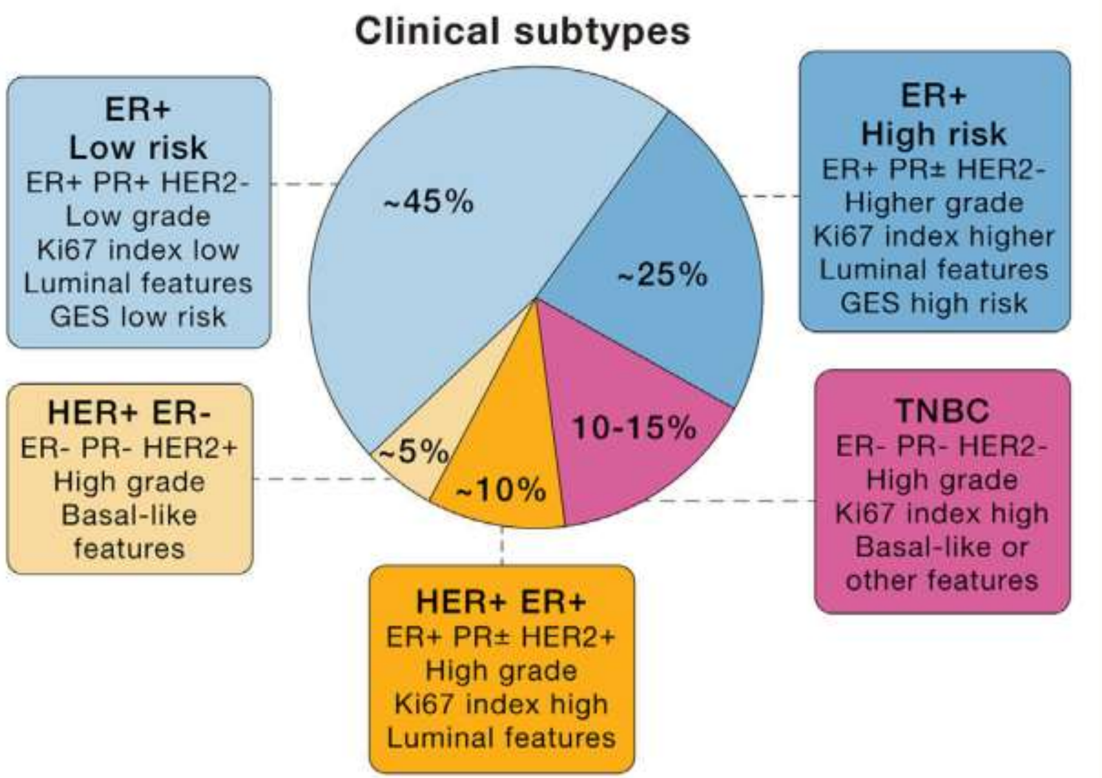
bcam basque center for applied mathematics

CIC bioGUNE CENTRE FOR COOPERATIVE RESEARCH IN BIOSCIENCES

EXCELENCIA SEVERO OCHOA

Research interest	
Mathematics	**Biology**
Bayesian Inference	Transcriptomics/Genomics
Hamiltonian Monte Carlo Techniques	Breast cancer
Efficient symplectic integrators	Prediction of risk

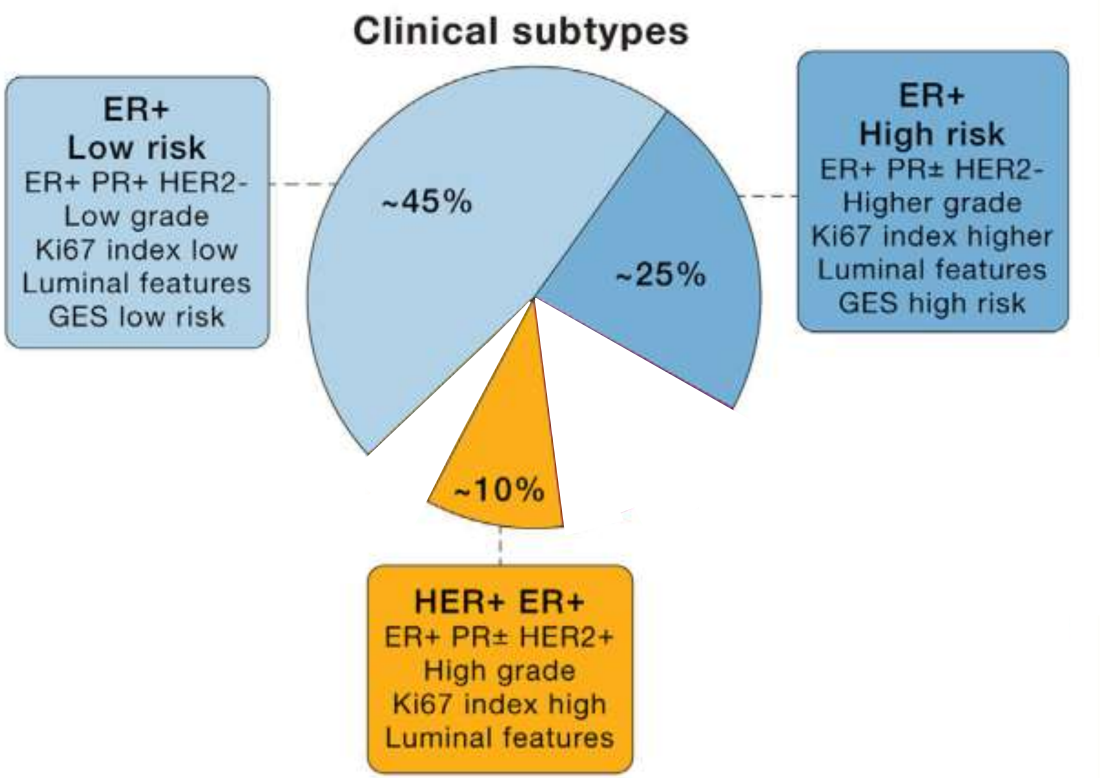
Introduction - Breast cancer is the most prevalent cancer in women



- There are 6 major clinical subtypes, determined by ER, PR and HER2 status
- **Prognosis** and **possible treatments** depend on the subtype

Image:
Nolan, E., Lindeman, G. J., & Visvader, J. E. (2023). Deciphering breast cancer: from biology to the clinic. Cell.

Introduction - The majority of breast cancers (BC) are ER-positive (> 70%)

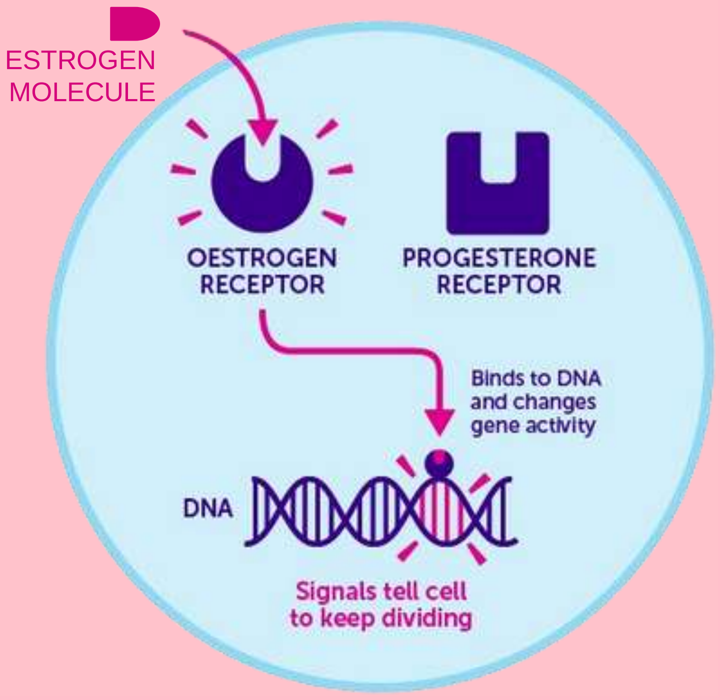


- ➢ There are 6 major clinical subtypes of BC, determined by ER, PR and HER2 status
- ➢ **Prognosis** and **possible treatments** depend on the subtype
- ➢ **70% of them are ER+**, as they express the estrogen receptor
- ➢ These can be treated with **hormone therapy**

Image:
Nolan, E., Lindeman, G. J., & Visvader, J. E. (2023). Deciphering breast cancer: from biology to the clinic. Cell.

Introduction – Estrogen is a hormone that controls the development of breast cells

ESTROGEN FUELS GROWTH AND CELL DIVISION OF BREAST CANCER CELLS

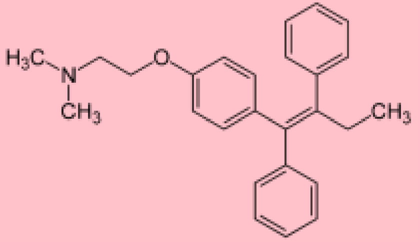

ESTROGEN MOLECULE
OESTROGEN RECEPTOR
PROGESTERONE RECEPTOR
Binds to DNA and changes gene activity
DNA
Signals tell cell to keep dividing

Introduction – Hormone therapies target the estrogen receptor to impede growth

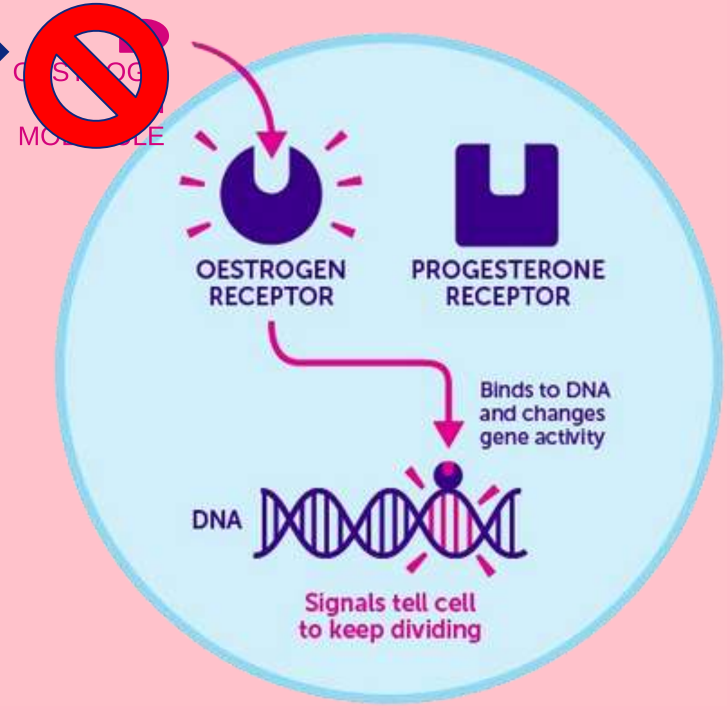
Tamoxifen

Applied as a **5-year treatment** after surgery

↓ Relapse by ~50%
Mortality by ~30%



ESTROGEN FUELS GROWTH AND CELL DIVISION OF BREAST CANCER CELLS

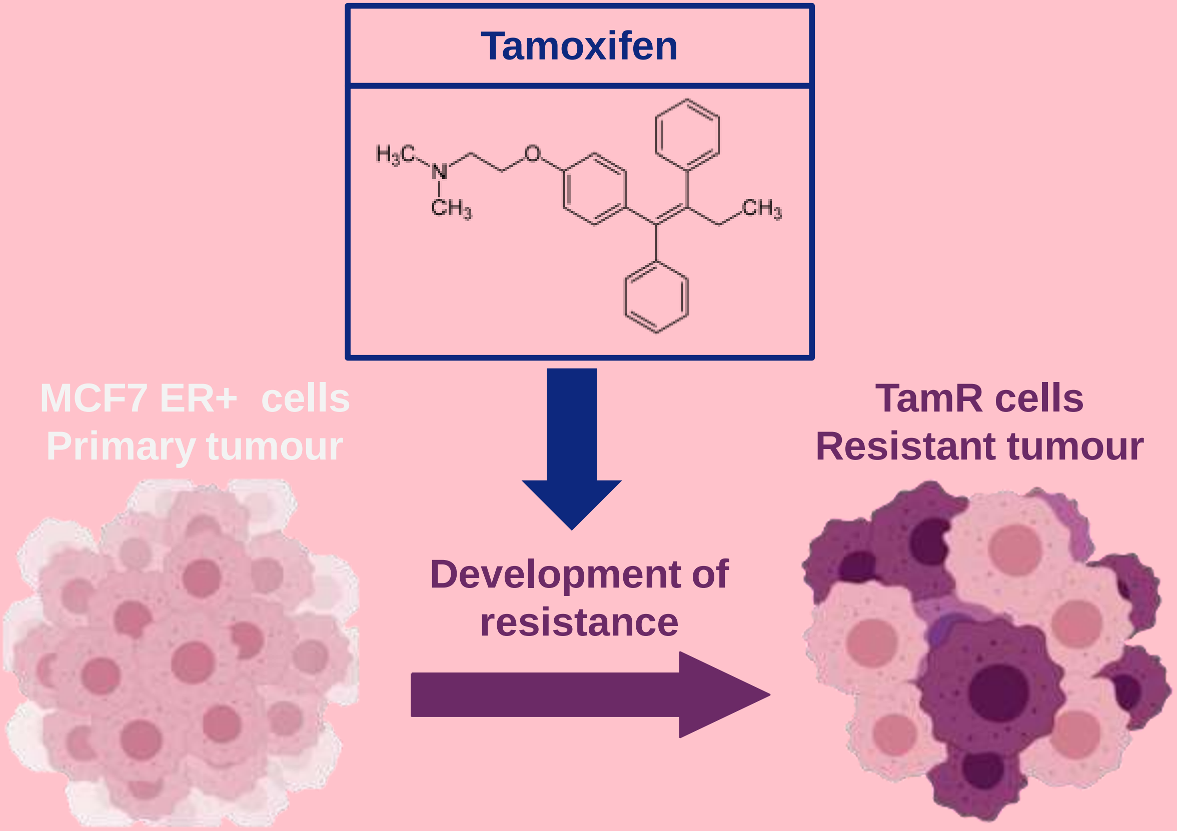


OESTROGEN MOLECULE
OESTROGEN RECEPTOR
PROGESTERONE RECEPTOR
Binds to DNA and changes gene activity
DNA
Signals tell cell to keep dividing

- ➢ As an **antagonist**, tamoxifen binds to the estrogen receptor, keeping the estrogen from binding to it
- ➢ Alternatively, other hormone therapies look to inhibit the synthesis of estrogen in the first place
- ➢ Between 30%-50% of treatment can generate a **resistant response** where it doesn't work and treatment time is crucially wasted

Szostakowska,M. et al. (2019). Resistance to endocrine therapy in breast cancer: molecular mechanisms and future goals. Breast Cancer Research andTreatment, 173, 489-497.

Introduction – Some cells become resistant to the treatment and continue dividing



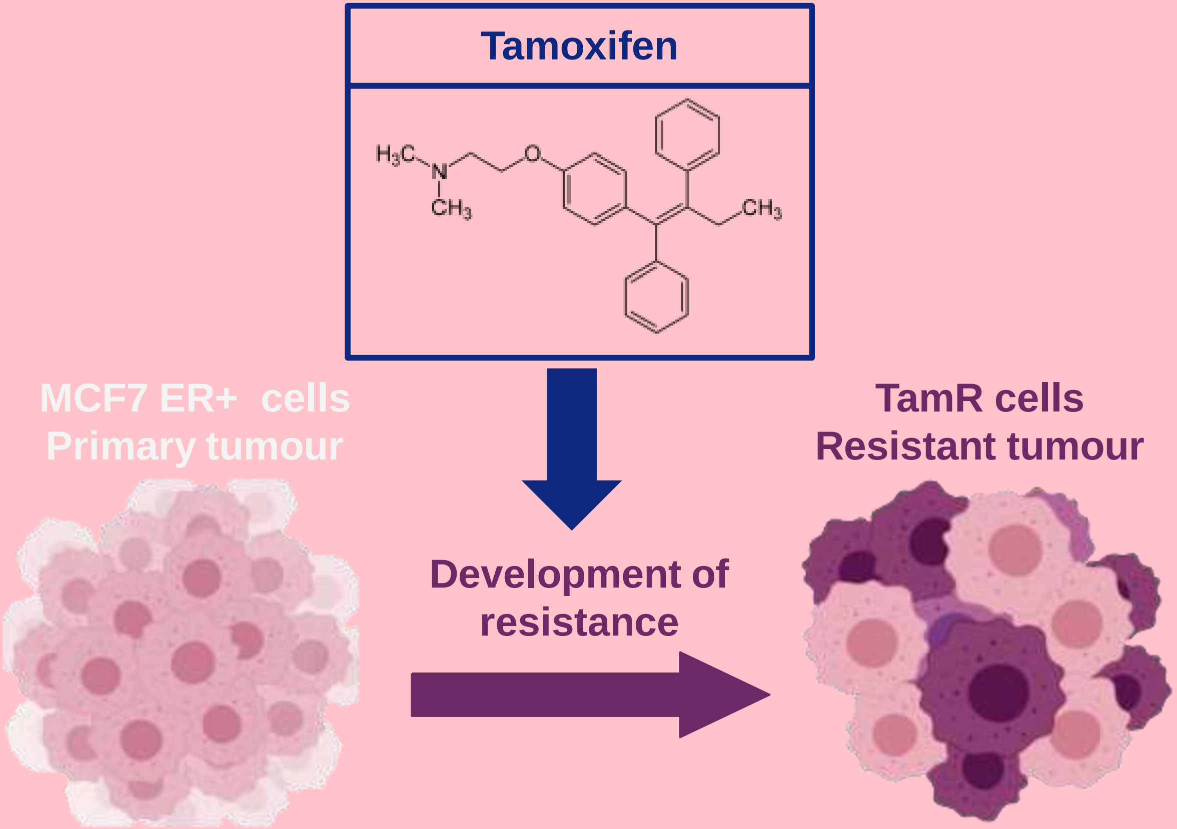
Tamoxifen

MCF7 ER+ cells
Primary tumour

Development of resistance

TamR cells
Resistant tumour

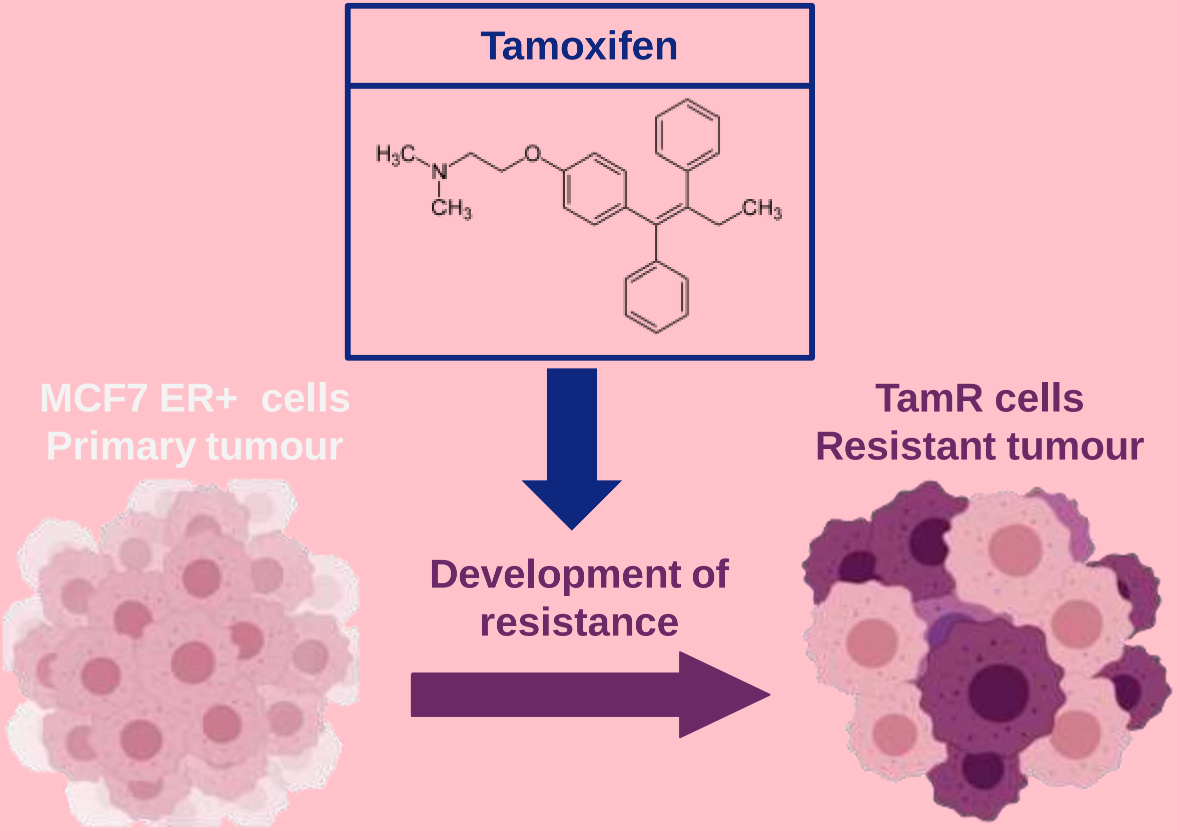
Introduction – Some cells become resistant to the treatment and continue dividing



How do we characterize a sample?

- ➢ **RNA-seq** analyzes gene expression by measuring the abundance of RNA transcripts
- ➢ Transcripts serve as **templates** for protein synthesis so they **regulate cell functions**
- ➢ RNA-seq offers a **picture** into the state of a cell and its **activity**
- ➢ An usual RNA-seq provides information on **over 24.000 transcripts/genes**
- ➢ Is this where the heterogeneity appears? **NO** Cell models are replicable and differences can be controlled to a certain degree

Introduction – Some cells become resistant to the treatment and continue dividing



Tamoxifen

MCF7 ER+ cells
Primary tumour

Development of resistance

TamR cells
Resistant tumour

<u>How do we compare two biological states?</u>

- By taking **RNA-seq of two different conditions** we can study how the **abundance** of genes/transcripts in each of them

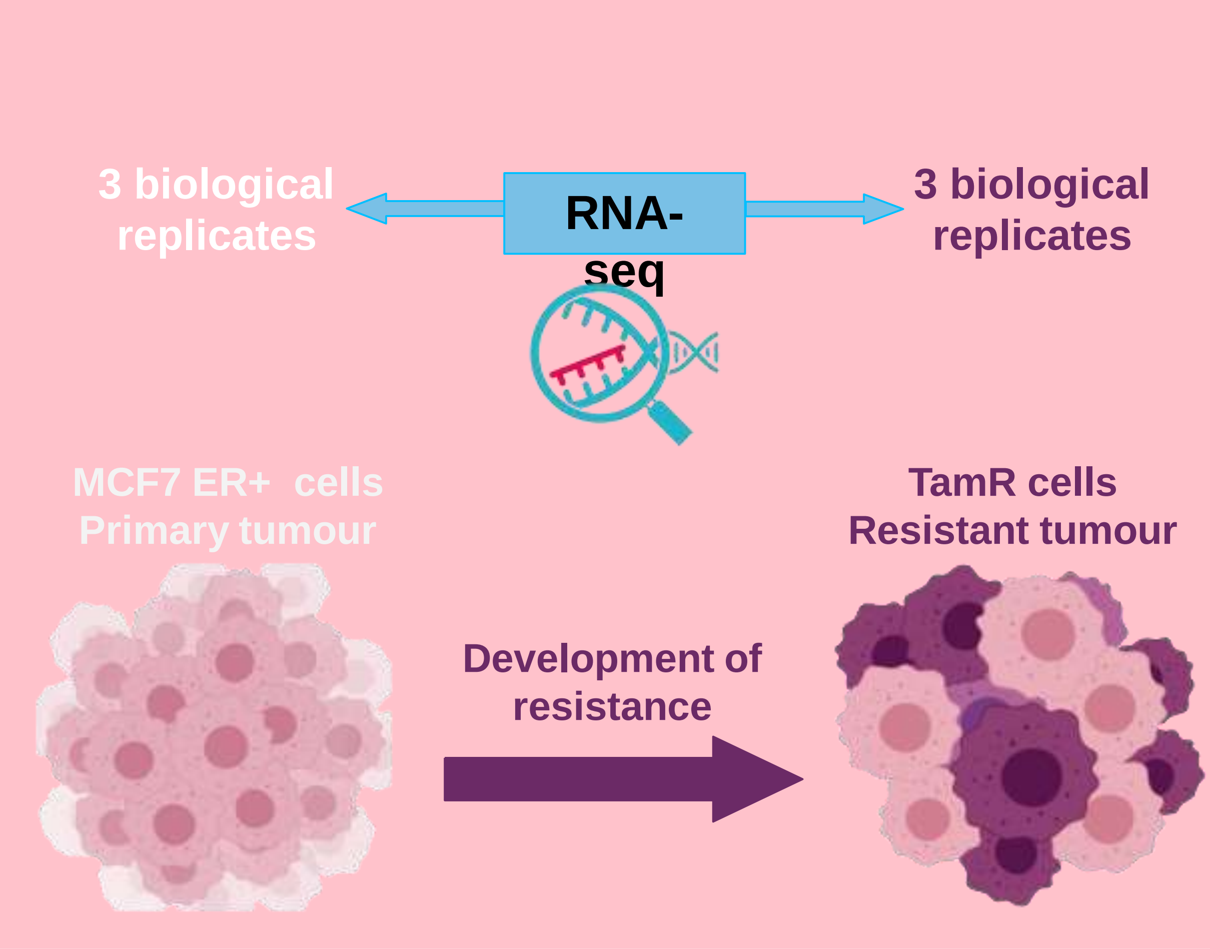
- **Differential Gene Expression** is measured in **Fold Change**, or how much abundant a feaature is in one sample over the other

- For **cells**, replicating an experiment can produce multiple instances or **replicates** that should give **homogenous** outputs

- For **patients**, differences between them are bigger (state of disease, external factors, age) creating a more **heterogeneous** landscape.

- It is important to tackle this heterogeneity to identify **problem specific biomarkers (genes)**

Data – Cell models are good for controlled experiments in homogeneous environments



3 biological replicates

RNA-seq

3 biological replicates

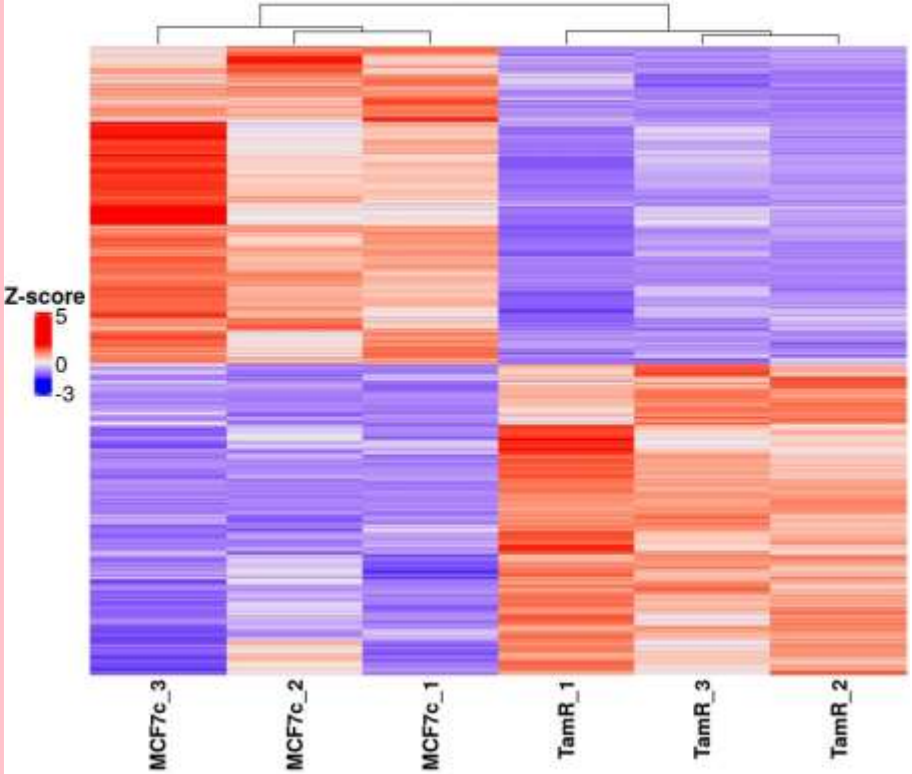
MCF7 ER+ cells
Primary tumour

Development of resistance

TamR cells
Resistant tumour

Heatmap of biological replicates

- ➢ Clear distinctions arising from induced changes



Data – Patients are heterogeneous in their type of disease and conditions

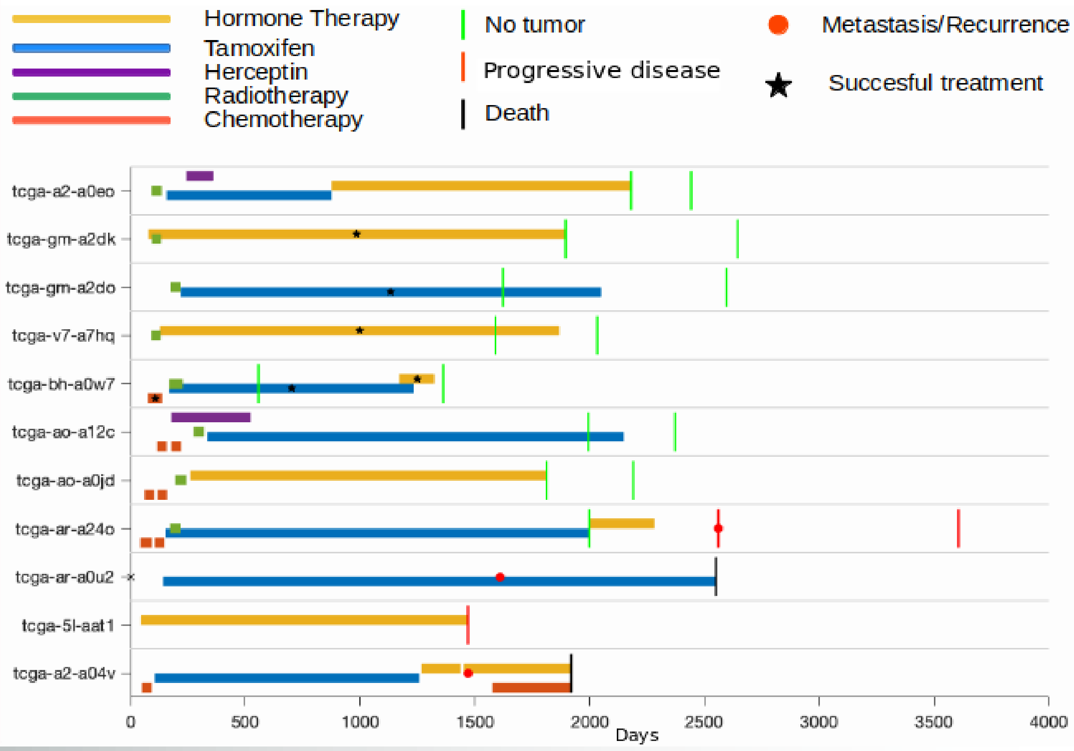
The Cancer Genome Atlas (TCGA)

- ➢ **Public database** with >1000 BC patients from USA
- ➢ RNA-seq + Extensive clinical records
- ➢ This allows a proper **cleaning and classification** of patients where the administered treatment was irregular or inconsistent
- ➢ Resulting cohort of patients with **tamoxifen** or other **hormone therapies** and their **response to treatment**

Tamofixen	**Hormone therapies**
➢ 25 Good Responders ➢ 12 Resistant	➢ 87 Good Responders ➢ 40 Resistant

A patient's journey

➢ Helps classifying patients and reducing heterogeneity by removing patients with non-cancer related issues



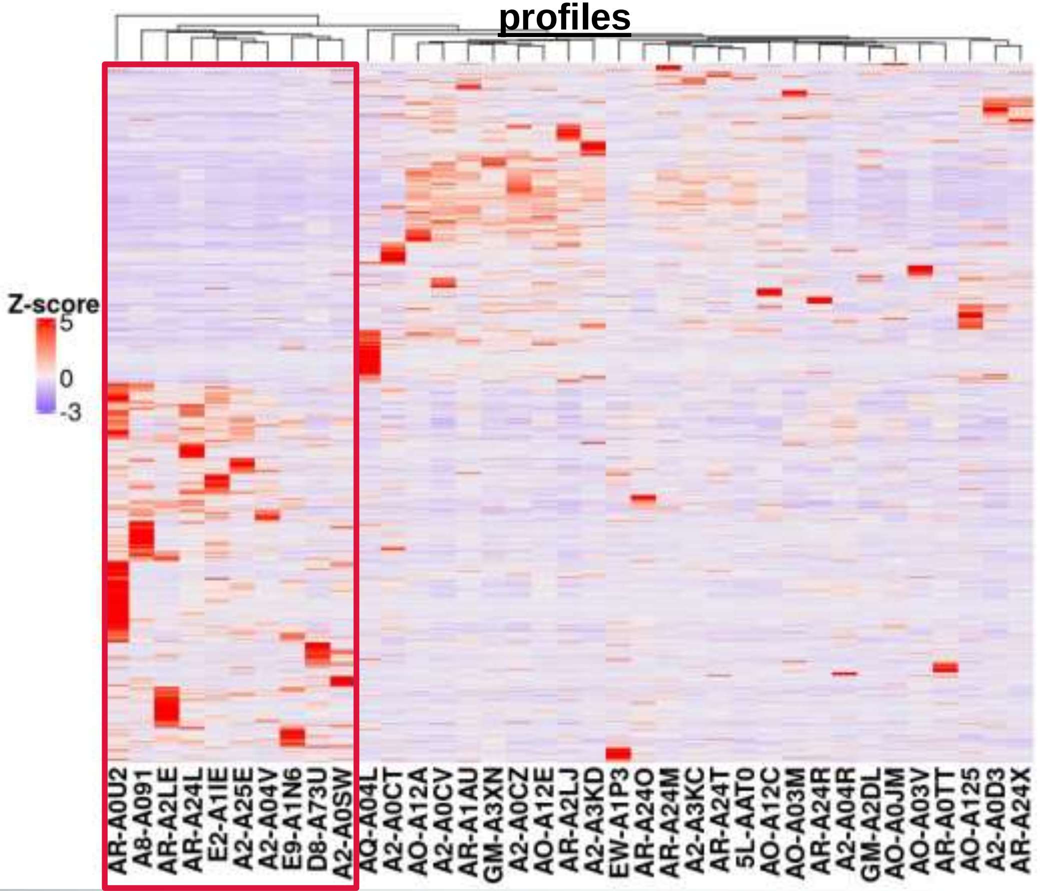
Data – Heterogeneity is clearly present in the gene heatmap

The Cancer Genome Atlas (TCGA)

- **Public database** with >1000 BC patients from USA
- RNA-seq + Extensive clinical records
- This allows a proper **cleaning and classification** of patients where the administered treatment was irregular or inconsistent
- Resulting cohort of patients with **tamoxifen** or other **hormone therapies** and their **response to treatment**

Tamofixen	**Hormone therapies**
➢ 25 Good Responders ➢ 12 Resistant	➢ 87 Good Responders ➢ 40 Resistant

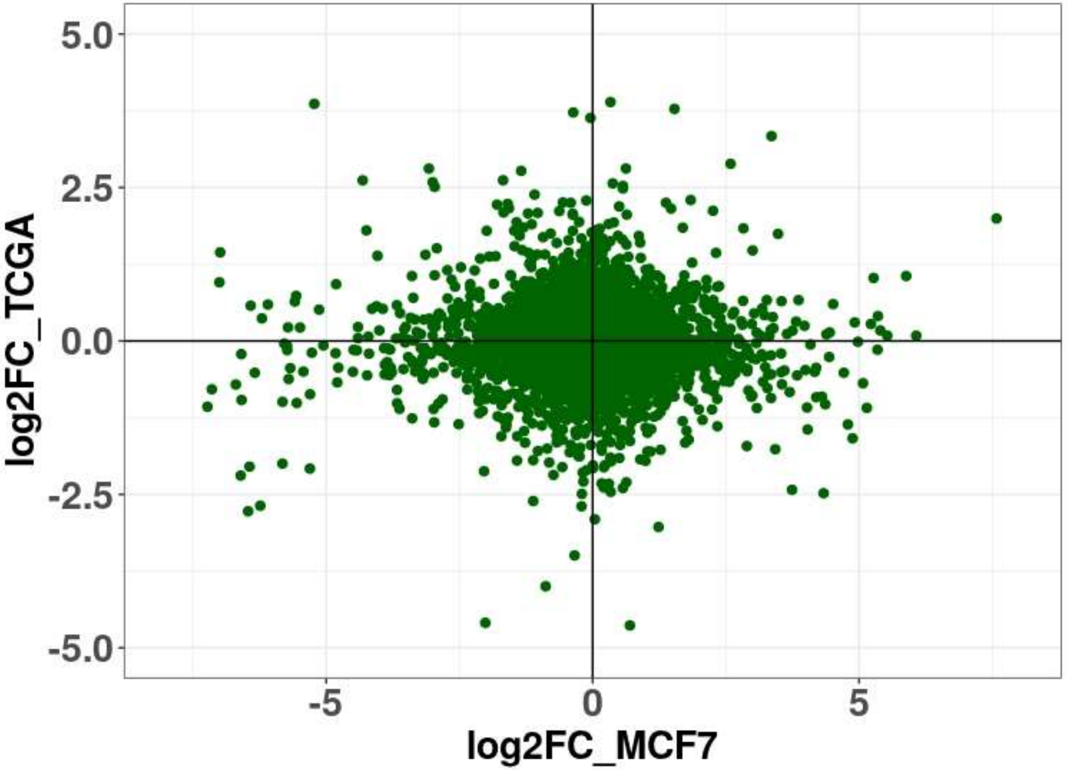
Tamoxifen treated patients profiles



Analysis – Comparing cell and patients profiles

We can look at the **distribution of differentially expressed genes** across patients and cells.

By **filtering out** non-relevant genes we can try to identify which ones behave similarly in these two **comparable resistance scenarios**



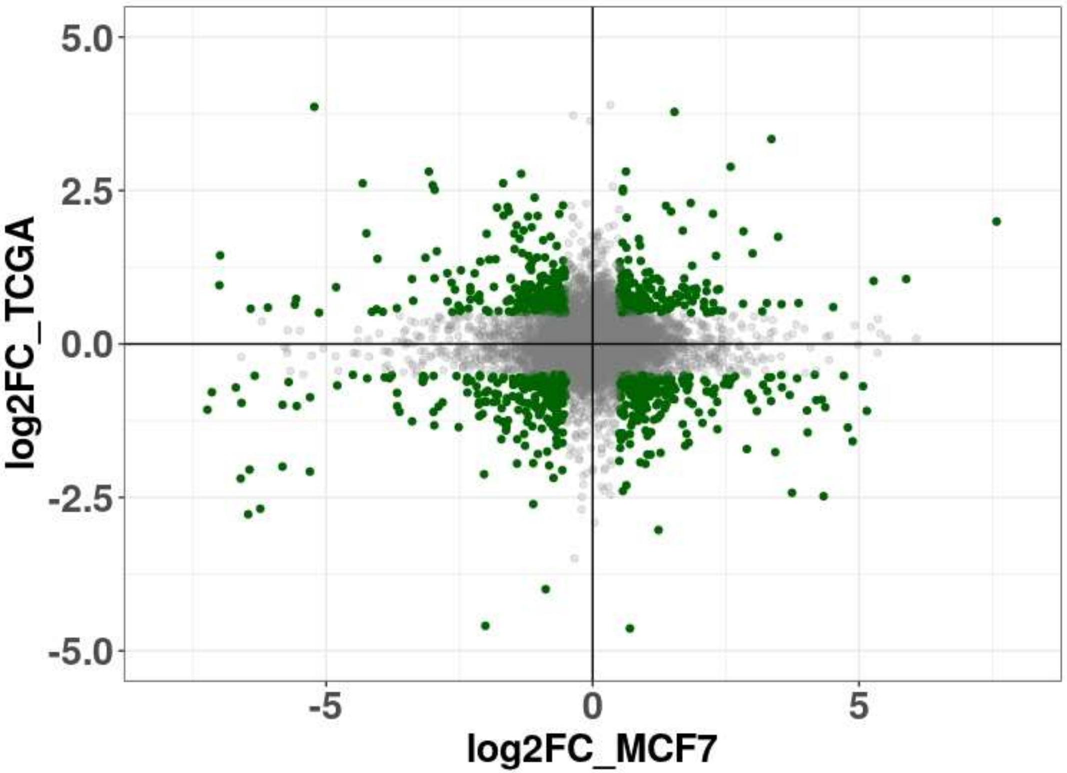
Analysis – Comparing cell and patients profiles

We can look at the **distribution of differentially expressed genes** across patients and cells.

By **filtering out** non-relevant genes we can try to identify which ones behave similarly in these two **comparable resistance scenarios**

Filters:

- ➢ Genes with | $\log_2$ Fold Change | > 0.5



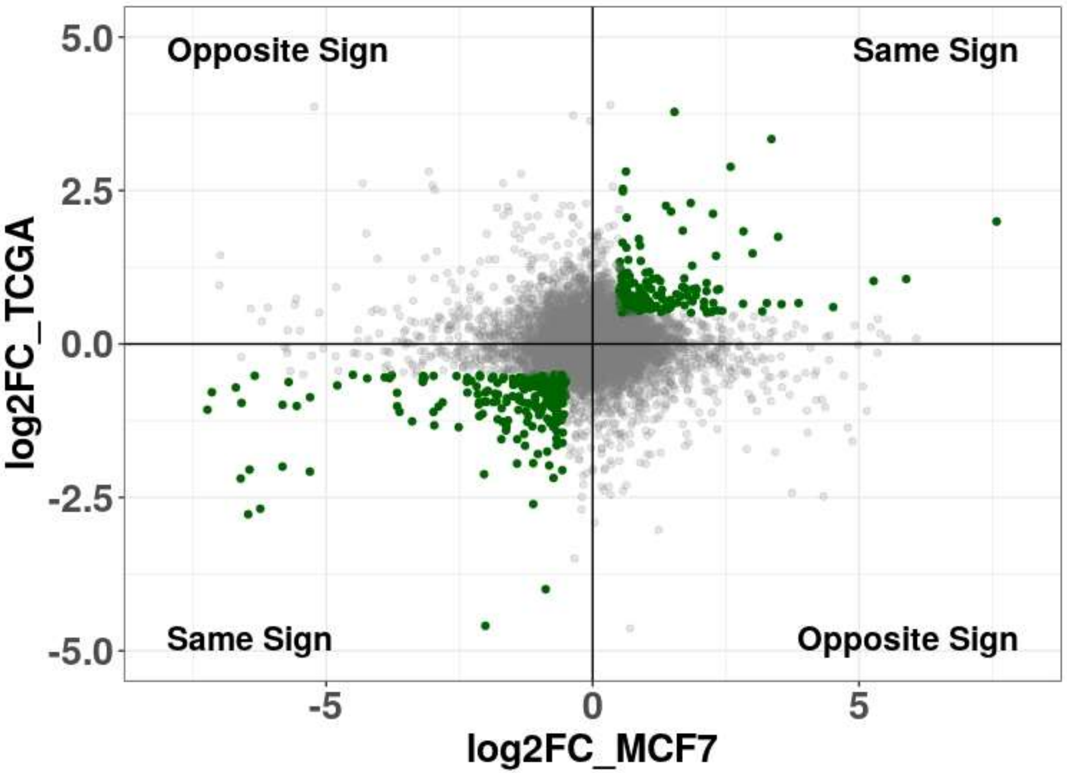
Analysis – Comparing cell and patients profiles

We can look at the **distribution of differentially expressed genes** across patients and cells.

By **filtering out** non-relevant genes we can try to identify which ones behave similarly in these two **comparable resistance scenarios**

Filters:

- ➢ Genes with | $\log_2$ Fold Change | > 0.5
- ➢ Genes expressed in the same direction



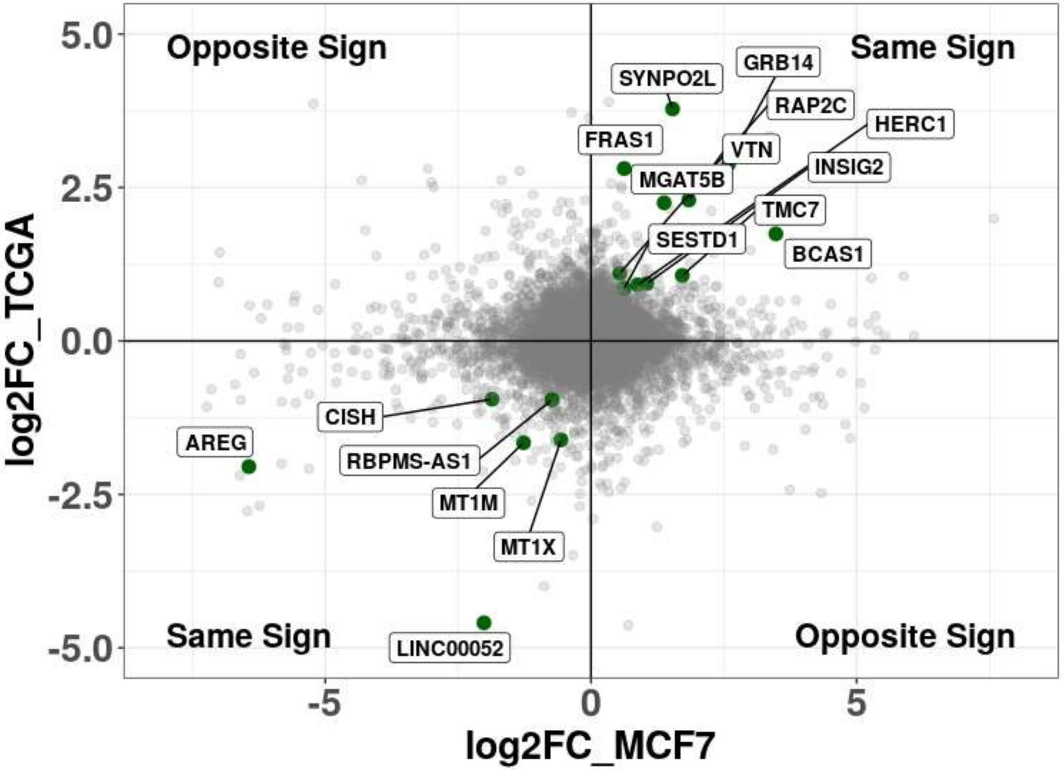
Analysis – Comparing cell and patients profiles

We can look at the **distribution of differentially expressed genes** across patients and cells.

By **filtering out** non-relevant genes we can try to identify which ones behave similarly in these two **comparable resistance scenarios**

Filters:

- ➢ Genes with | $\log_2$ Fold Change | > 0.5
- ➢ Genes expressed in the same direction
- ➢ Differential expression significance test FDR>0.1

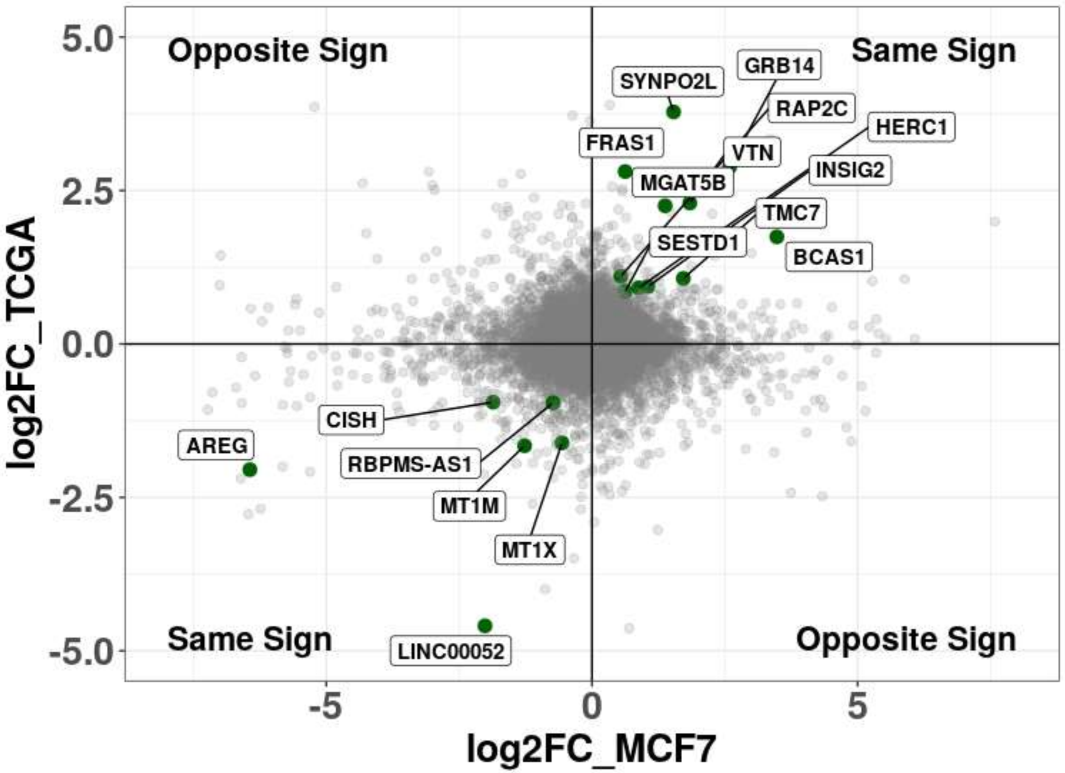


Analysis – Gene signatures

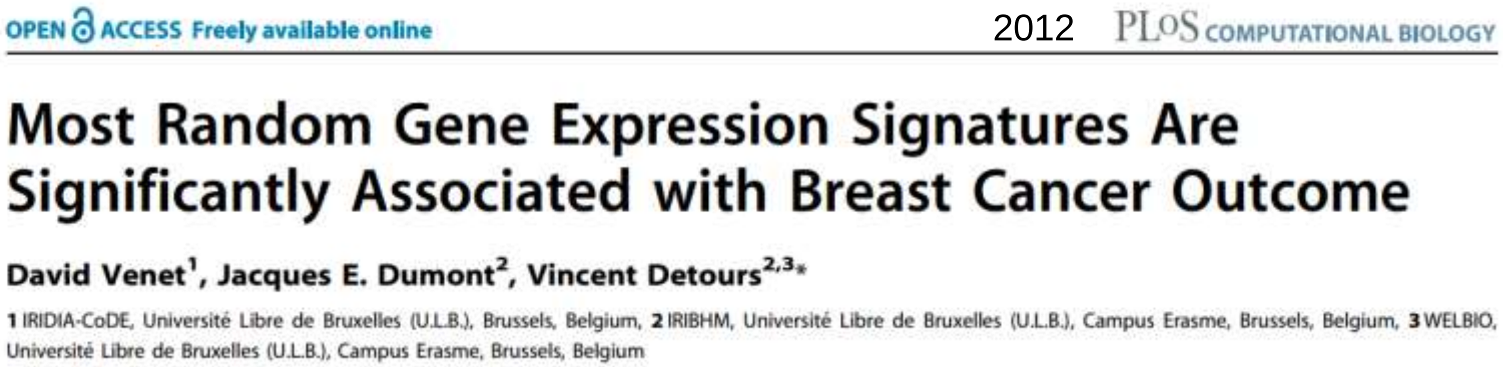
A **collection of genes** that can be used to **represent** or identify some **biological process or clinical condition** is called a **gene signature**

We were able to use the homogeneous cell data to select 17 genes related to tamoxifen resistance in the heterogenous patient dataset

17 Gene Signature! DONE!
Bring down the curtain!



Analysis – The prevalence of random gene signatures



OPEN ACCESS Freely available online

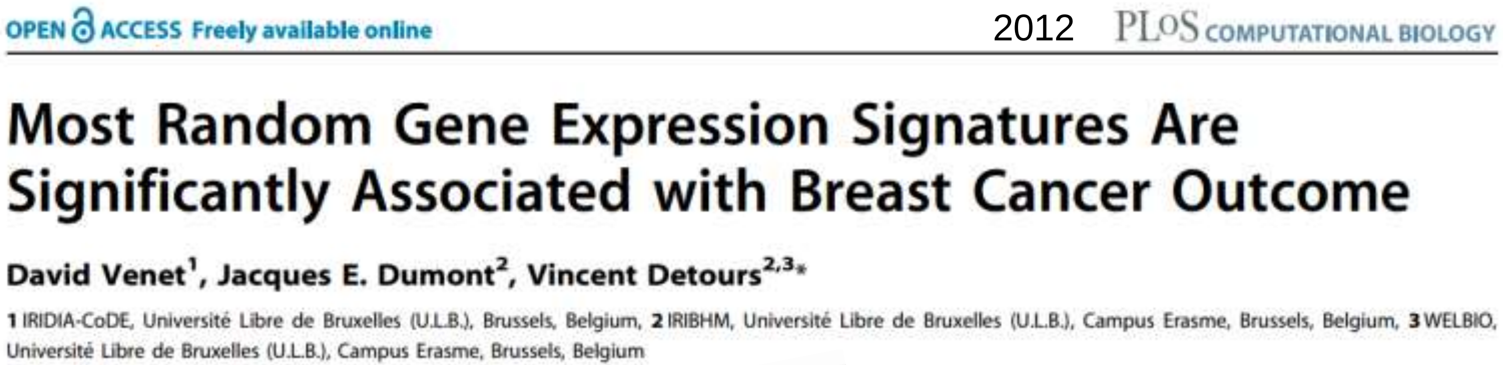
2012 PLoS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

[1] IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, [2] IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, [3] WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

Analysis – The prevalence of random gene signatures



2012

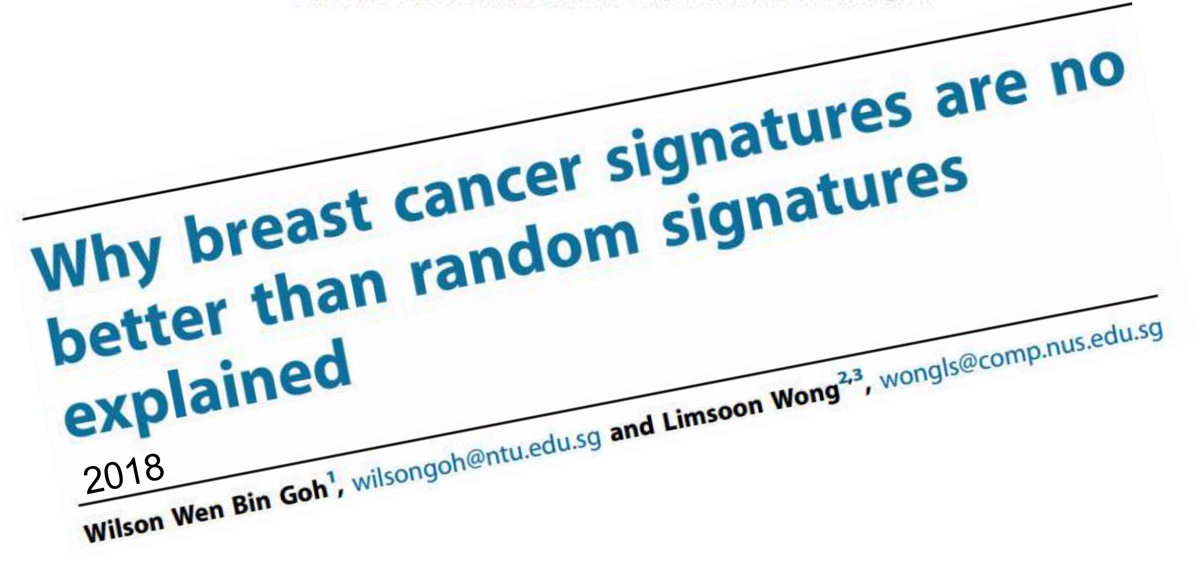
OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

1 IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, 2 IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, 3 WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium

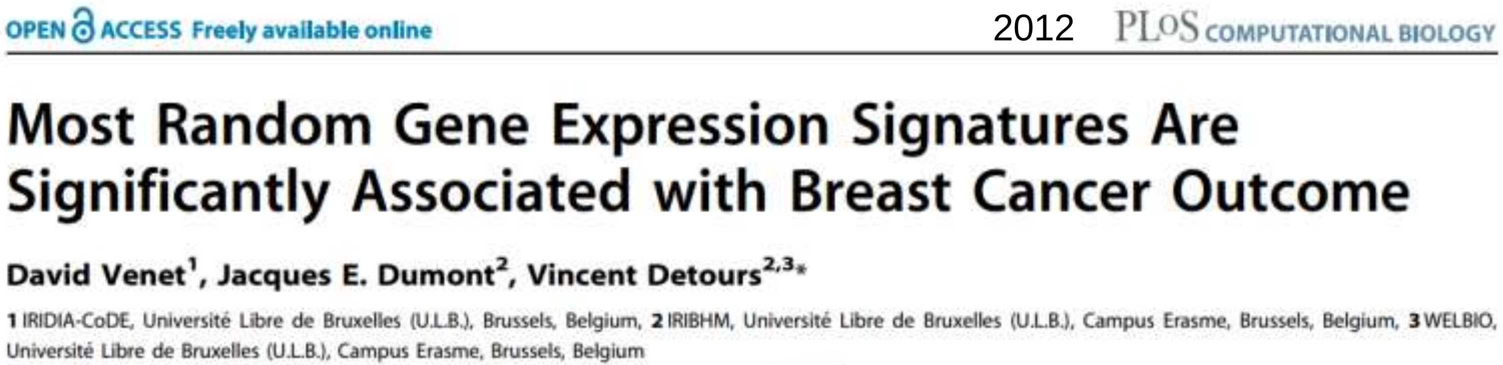


Why breast cancer signatures are no better than random signatures explained

Wilson Wen Bin Goh[1], wilsongoh@ntu.edu.sg and Limsoon Wong[2,3], wongls@comp.nus.edu.sg

2018

Analysis – The prevalence of random gene signatures



2012

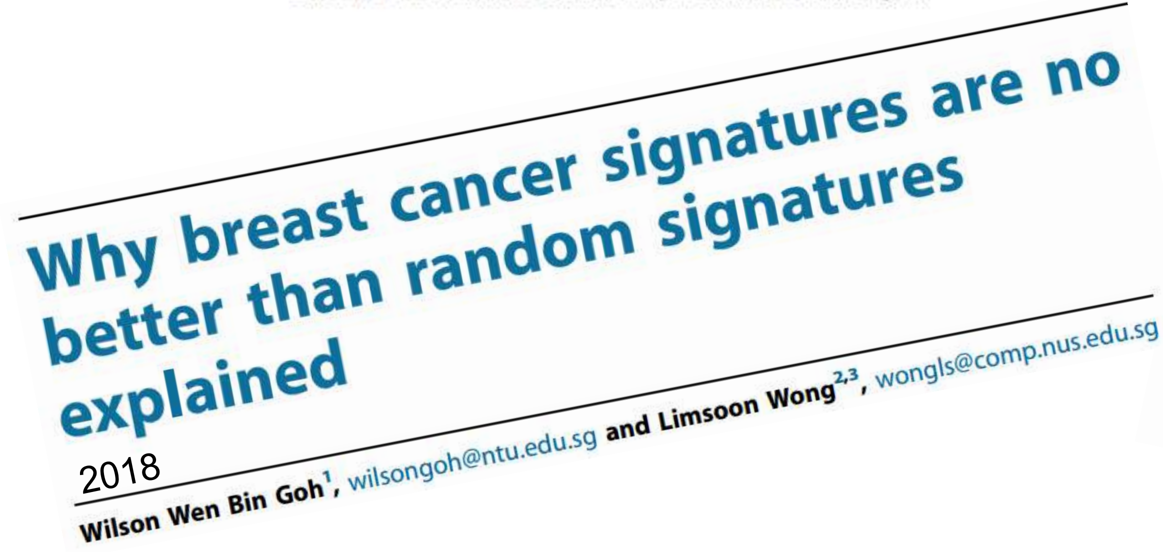
OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome

David Venet[1], Jacques E. Dumont[2], Vincent Detours[2,3]*

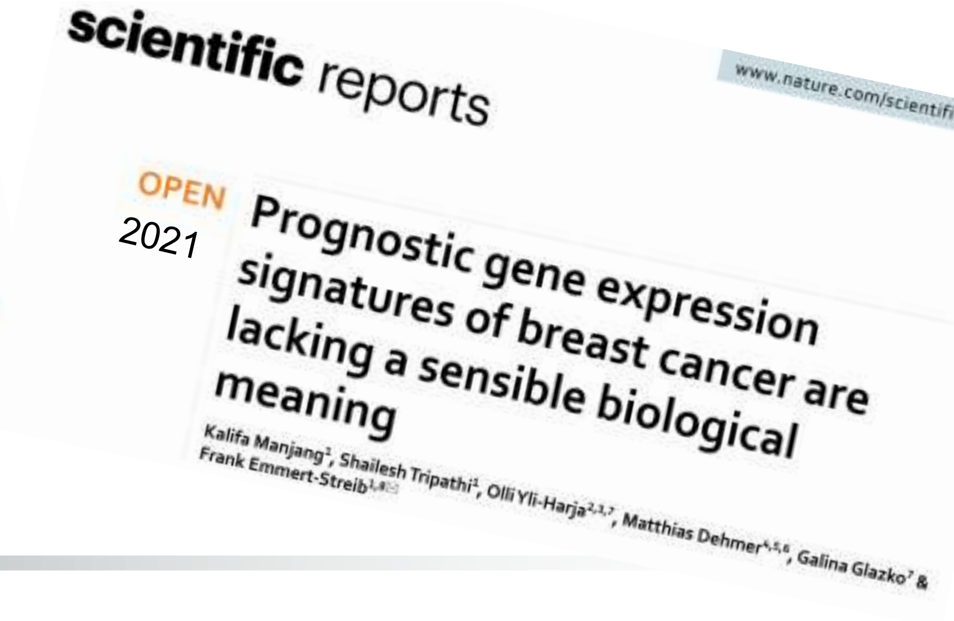
[1] IRIDIA-CoDE, Université Libre de Bruxelles (U.L.B.), Brussels, Belgium, [2] IRIBHM, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium, [3] WELBIO, Université Libre de Bruxelles (U.L.B.), Campus Erasme, Brussels, Belgium



Why breast cancer signatures are no better than random signatures explained

Wilson Wen Bin Goh[1], wilsongoh@ntu.edu.sg and Limsoon Wong[2,3], wongls@comp.nus.edu.sg

2018



scientific reports

www.nature.com/scientifi

OPEN

2021

Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning

Kalifa Manjang[1], Shailesh Tripathi[1], Olli Yli-Harja[2,3,7], Matthias Dehmer[4,5,6], Galina Glazko[7] & Frank Emmert-Streib[1,8]

Analysis

1. How do we address the concerns raised by these papers?

2. Can we actually predict the risk of resistance to treatment?

3. How can we validate our purely computational result?

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

We follow their advices!

1. Often significance comes from the correlation of genes with proliferation.

2. The bigger the signature, the closer to a random one it is (<25 genes is OK)

3. Add biological insight and more testing subjects

4. Check if actually a random signature can replicate our result.

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

We follow their advices!

✓ 1. Often significance comes from the correlation of genes with proliferation.

2. The bigger the signature, the closer to a random one it is (<25 genes is OK)

3. Add biological insight and more testing subjects

4. Check if actually a random signature can replicate our result.

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

We follow their advices!

✓ 1. Often significance comes from the correlation of genes with proliferation.

✓ 2. The bigger the signature, the closer to a random one it is (<25 genes is OK)

3. Add biological insight and more testing subjects

4. Check if actually a random signature can replicate our result.

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

We follow their advices!

✓ 1. Often significance comes from the correlation of genes with proliferation.

✓ 2. The bigger the signature, the closer to a random one it is (<25 genes is OK)

✓ 3. Add biological insight and more testing subjects

4. Check if actually a random signature can replicate our result.

Analysis – Dealing with the issue random gene signatures

1. How do we address the concerns raised by these papers?

We follow their advices!

✓ 1. Often significance comes from the correlation of genes with proliferation.

✓ 2. The bigger the signature, the closer to a random one it is (<25 genes is OK)

✓ 3. Add biological insight and more testing subjects

✓ 4. Check if actually a random signature can replicate our result.

Results

1. How do address the concerns raised by these papers?

2. Can we actually predict the risk of resistance to treatment?

3. How can we validate our purely computational result?

Results – Creating a model for the classification of resistant patients

2. Can we actually predict the risk of resistance to treatment?

Results – Creating a model for the classification of resistant patients

2. Can we actually predict the risk of resistance to treatment?

➢ We should find out if the selected genes (or a subset of them) can identify **resistant patients**

➢ Using a **Bayesian Logistic Regression** model we can estimate the probability of a good or resistant response from patient *i:*

$$y_i \sim Bernoulli(p_i) \;\rightarrow\; logit(p_i) = \theta_0 + \theta_1 g_{i,1} + \cdots + \theta_{D-1} g_{i,D-1}$$

➢ Resulting in a likelihood:

$$L(y_i|\theta, g_i) = \prod_{i=1}^{N} \left[\left(\frac{e^{\theta g_i}}{1 + e^{\theta g_i}} \right)^{y_i} \left(1 - \frac{e^{\theta g_i}}{1 + e^{\theta g_i}} \right)^{1-y_i} \right]$$

Results – Creating a model for the classification of resistant patients

2. Can we actually predict the risk of resistance to treatment?

- ➢ As priors we use Normal distributions, given by the **differential expression values** of each gene in the **cell data** μ:

$$pr(\theta) \sim \mathcal{N}(\mu, \sigma^2)$$

- ➢ So using **Bayes' theorem**, we can obtain the posterior distribution of the parameters θ given the already set likelihood and prior:

$$\overbrace{p(\theta|y_i, g_i)}^{Posterior} \propto \underbrace{p(y_i|\theta, g_i)}_{Likelihood} \overbrace{p(\theta)}^{Prior}$$

- ➢ We characterize each signature using a **Gene Signature Score** that defines a signature of N genes:

$$GSS_i = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{g_{i,n} - \mu_n}{\sigma_n} \right)$$

Results – Creating a model for the classification of resistant patients

2. Can we actually predict the risk of resistance to treatment?

- We used improved **Hamiltonian Monte Carlo (HMC)** techniques to obtain the coefficients for each GSS combination

- HMC employs Hamilton's equation of motion to stay in **Hamiltonian trajectories** in space so that we can **efficiently sample** from the resulting posterior distribution.

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = \frac{\partial H(\theta, p)}{\partial p} = M^{-1}p \; ; \; \frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial H(\theta, p)}{\partial \theta} = -\nabla_\theta U(\theta)$$

- These method allow **efficient explorations** of **complex, high-dimensional spaces** as the trajectories aid the search and subsequent sampling. This makes them ideal candidates for working with -omics data in general.

2. Can we actually predict the risk of resistance to treatment?

- A key part of their success is an **efficient integration** of the discretize equation of motion. The integration can be seen as series of **drifts** and **kicks** (moves in ***θ*** and **p**):

$$\theta : \varphi^{\theta}_{\Delta t} = (\theta + \Delta t M^{-1} p, p) \; ; \; p : \varphi^{p}_{\Delta t} = (\theta, p - \Delta t \nabla_{\theta} U(\theta))$$

- We make use of **in-house developed palindromic splitting integration schemes** composed by this sequences of drift and kicks can be used to improve the efficiency of HMC methodologies:

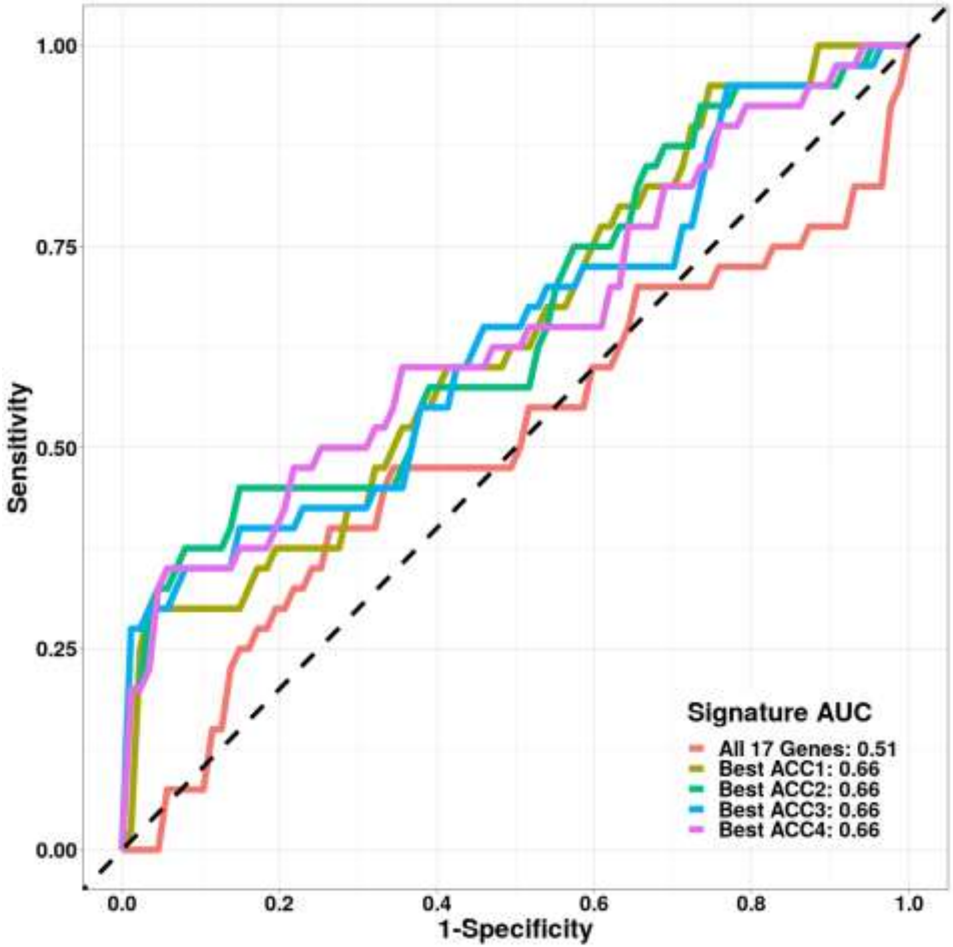
$$\psi_{\Delta t} = \varphi^{\theta}_{b\Delta t} \circ \varphi^{p}_{\Delta t/2} \circ \varphi^{\theta}_{(1-2b)\Delta t} \circ \varphi^{p}_{\Delta t/2} \circ \varphi^{\theta}_{b\Delta t}$$

Results – Improving our initial gene signature

- We used a **Simulated Annealing algorithm** to test 50000 combinations of gene signatures from lengths 1 to 17

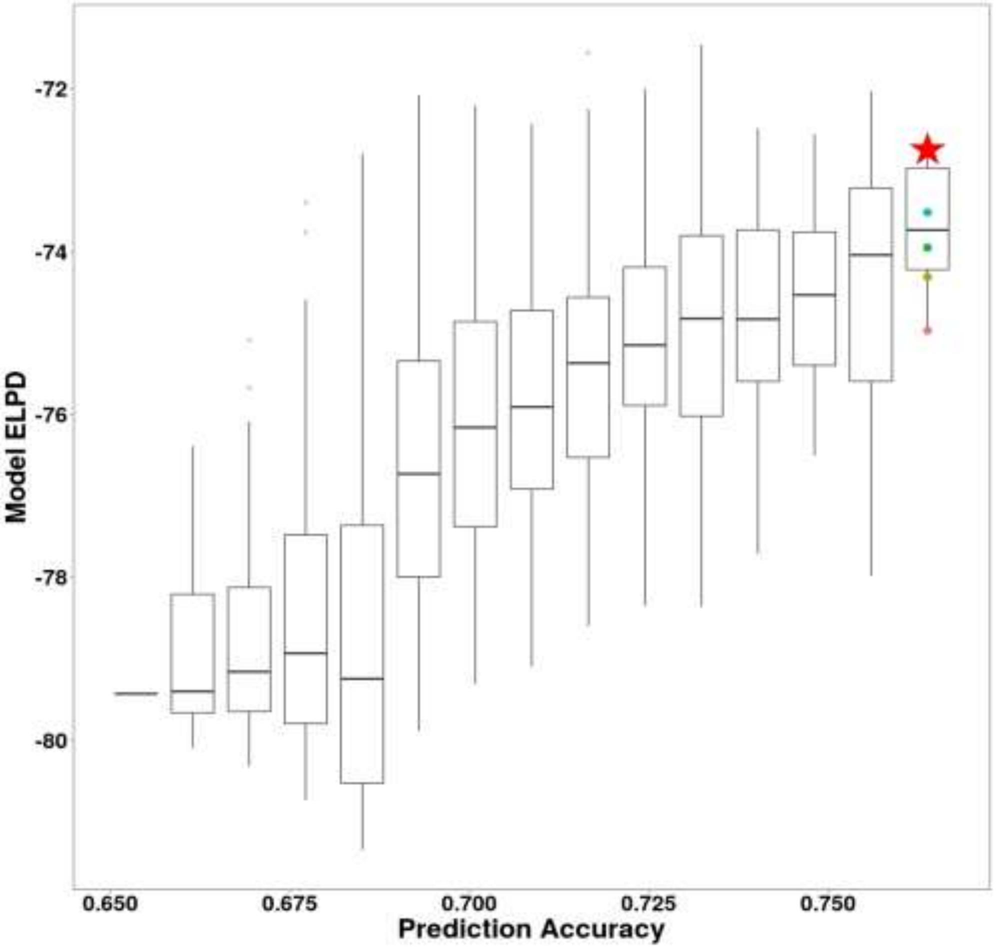
- For **each signature**, we run the model on the hormone therapy cohort (127 patients) and used a **Leave-One-Out algorithm** to assess **accuracy** in prediction

- Among the best gene signatures for classification several provided similarly good accuracy results



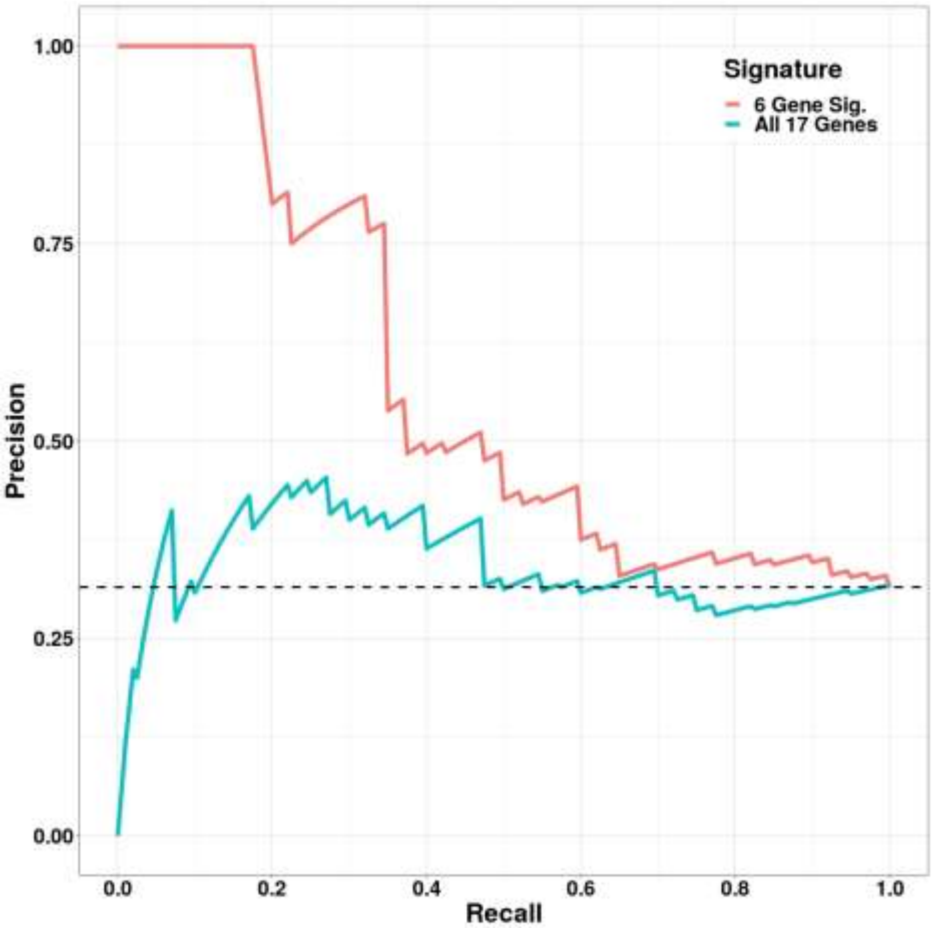
Results – Selecting the best gene signature

- We used a **Simulated Annealing algorithm** to test 50000 combinations of gene signatures from lengths 1 to 17
- For **each signature**, we run the model on the hormone therapy cohort (127 patients) and used a **Leave-One-Out algorithm** to assess **accuracy** in prediction
- Among the best gene signatures for classification several provided similarly good accuracy results
- To resolve these **ties**, we used the Expected Log-pointwise Predictive Density (**ELPD**)
- This **Bayesian specific** metric is used for assessing the goodness of fit and for model comparison



Results – Refinement of the 17 genes into a 6 gene signature

- The dataset is heavily **unbalanced**, with more patients responding well than becoming resistant
- Classifiers need to account for this. A random classifier will **overestimate the amount of resistant patients**
- Medically it is more relevant to **accurately predict a resistant patient** than a good responder (as by default, the assumption is good response)
- Our optimal signature was composed by **6 genes** that accurately classified **81% of their resistant predictions**



Validation

1. How do address the concerns raised by these papers?

2. Can we actually predict the risk of resistance to treatment?

3. How can we validate our purely computational result?

Validation – Two computational and one biological method

3. How can we validate our purely computational result?

Survival Analysis

- ➢ We will use two independent and new patient cohorts
- ➢ Patients with high abundance of the genes in our signature are considered **High** risk
- ➢ Shows the **probability of living without a relapse** over a period of time (10 years) of patients with **High**/**Low** risks

Cox Proportional Hazard Regression

- ➢ Allows the **comparison of multiple covariates** (signatures)
- ➢ Bigger hazard values imply better predictive capabilities for risk

$$h(t) = h_0 + \prod_{n=1}^{N} exp(b_n X_n)$$

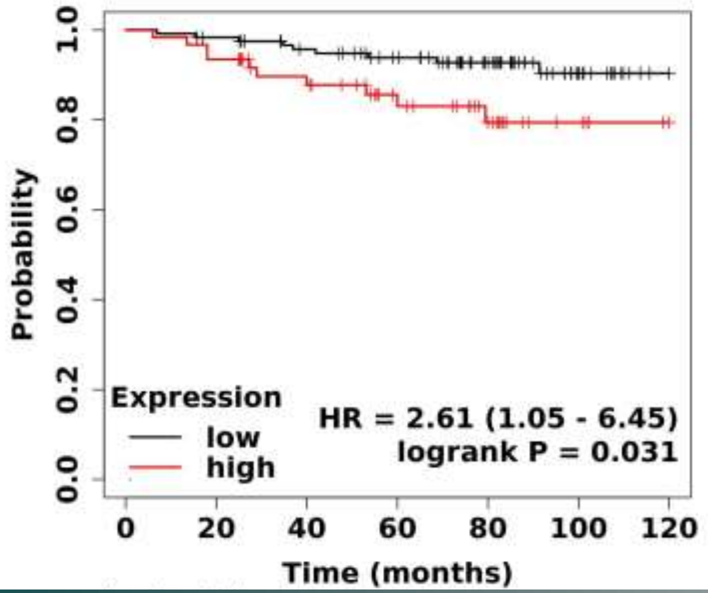
Cell experiments (qPCR)

- ➢ RNA-seq data showed us a picture of the cell in the moment it was sequenced
- ➢ qPCR experiments allows us to measure the abundance of the genes in the signature directly in the cell

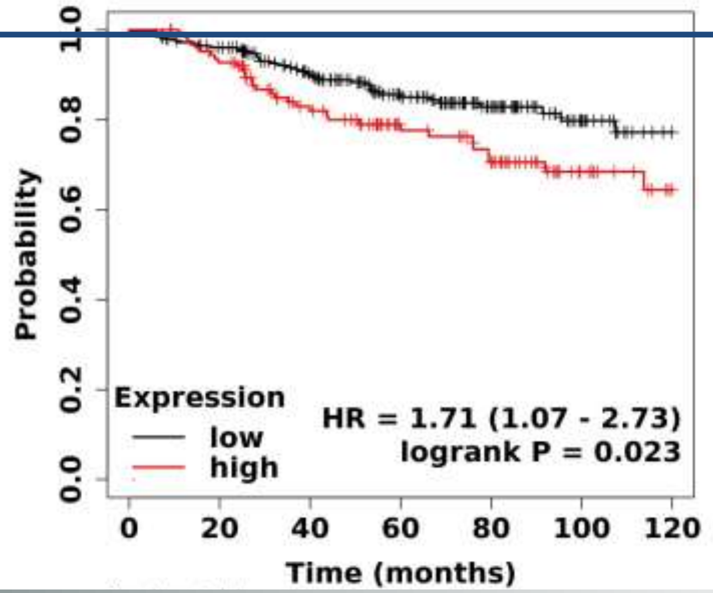
Validation – Survival analysis in the smaller cohort with tamoxifen-specific data (KMplotter)

Survival Analysis

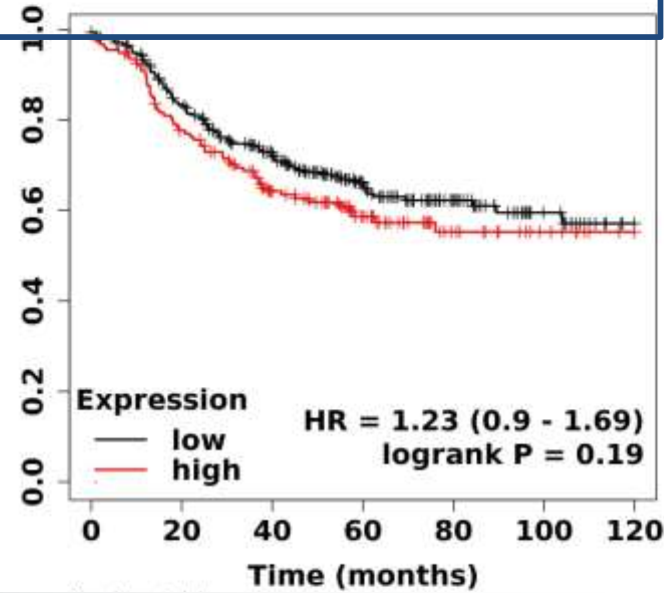
181 tamoxifen treated patients
61 at high risk: >14% less 10y-RFS



385 hormone therapy treated patients
127 at high risk: >16% less 10y-RFS



470 non-susceptible to treatment
Same 10y-Relapse Free Survival



bcam basque center for applied mathematics

EXCELENCIA SEVERO OCHOA

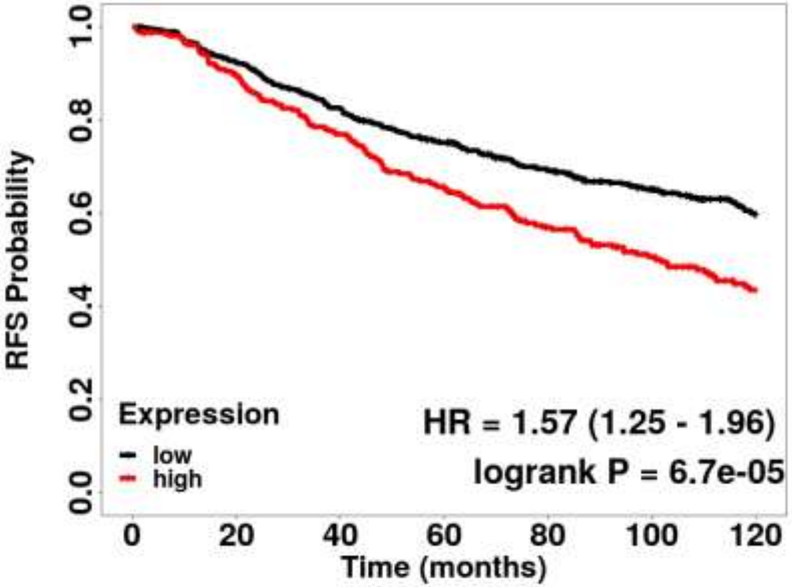
CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Data extracted from:
Győrffy,B. (2021) Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. Computational and Structural Biotechnology Journal, 19, 4101-4109

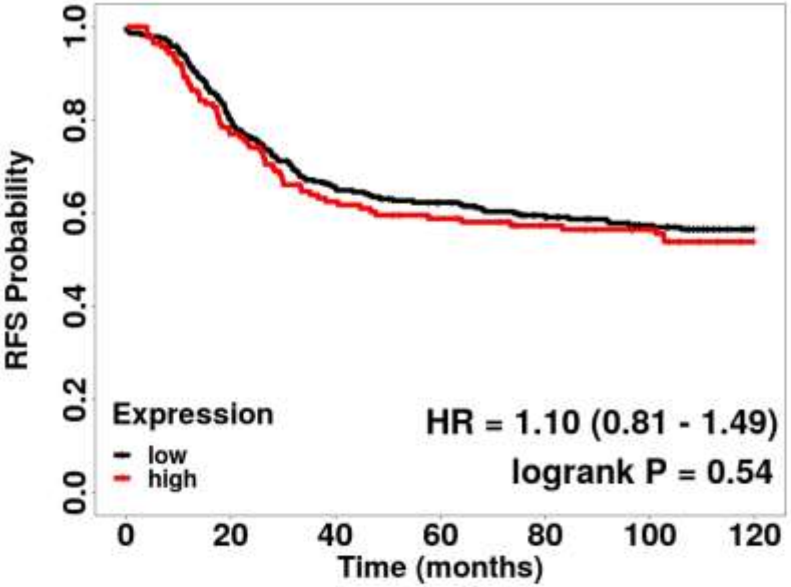
Validation – Survival analysis in the bigger cohort for all hormone therapies (METABRIC)

Survival Analysis

769 hormone therapy treated patients
256 at high risk: >16% less 10y-RFS



429 non-susceptible to treatment
Same 10y-Relapse Free Survival



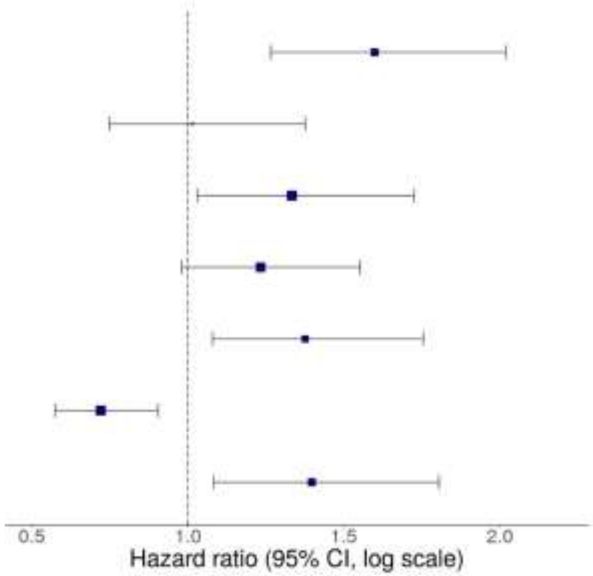
Data extracted from:
Gao,J. et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling, 6, pl1-pl1.

Validation – Our 6 Gene Signature outperformed many established signatures

Cox Proportional Hazard Regression

$$h(t) = h_0 + \prod_{n=1}^{N} exp(b_n X_n)$$

Signature	b_n Hazard Coef (95% CI)	P-value
6 Gene Signature	1.60 (1.27-2.02)	0.000533
5 Candidate Pathways	1.01 (0.75-1.38)	0.951496
SET ER/PR	1.33 (1.03-1.73)	0.457182
HOXB13 / IL17BR ratio	1.23 (0.98-1.55)	0.032555
Men et al 10 Gene Signature	1.38 (1.08-1.75)	0.028390
CRISPR mutant ESR1	0.72 (0.57-0.91)	0.007548
Oncotype DX	1.40 (1.08-1.80)	0.003053



basque center for applied mathematics

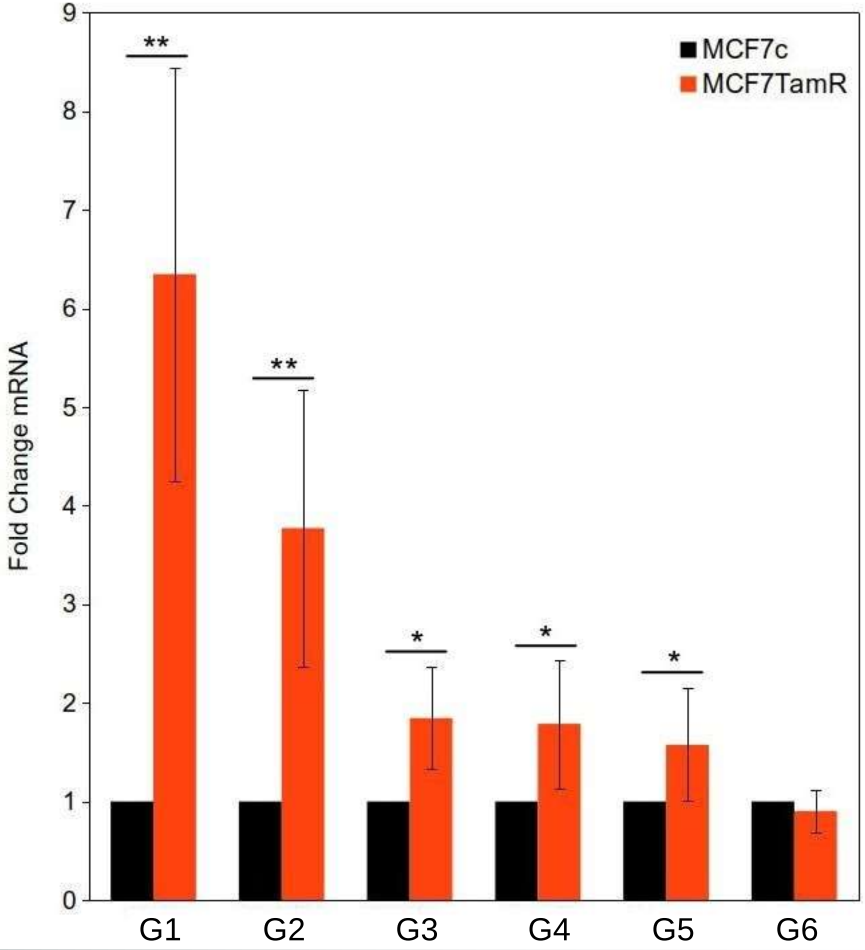
EXCELENCIA SEVERO OCHOA

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Validation – Initial biological confirmation of the computational results

Cell experiments (qPCR)

- ➢ One of the issues with gene signatures was the **lack of biological insight** (only computational results)
- ➢ To address it, we performed **qPCR analysis** of the genes in the signature in **control (black)** and **resistant (red)** cells, which have developed resistance over 48h
- ➢ We see a **significant increase** for **5 out of 6** genes in the signature.
- ➢ More experiments on the effects of **silencing** these genes will be performed to further understand the biological implications of the discovery



Thank you!

Grazas!

Eskerrik asko!

mparga@bcamath.org