

UNIVERSITÀ DEGLI STUDI DI TORINO

DIPARTIMENTO DI BIOTECNOLOGIE

DOTTORATO DI RICERCA IN MEDICINA MOLECOLARE

CICLO: XXXIII

**"Co-expression network analysis to unravel breast cancer
molecular features"**

Aurora Savino

TUTOR: Prof. Valeria Poli

COORDINATORE DEL DOTTORATO: Prof. Francesco Novelli

ANNI ACCADEMICI: 2017-2021

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA: BIO/11

Abstract

Biological systems are complex in nature, their behavior being governed by the interactions of many molecular components. Accurately depicting a complex biological system like cancer requires knowledge about elements' interactions, such as the regulatory layer described as Gene Regulatory Networks (GRN), often approximated with gene co-expression.

In this Thesis, I used co-expression networks to study breast cancer (BC) molecular mechanisms. Building upon a well clinically annotated transcriptome dataset of ~2000 primary breast tumors I defined modules of tightly co-regulated genes, which are robust across datasets and correlated with BC high grade, early onset, and poor prognosis. This approach permitted identifying gene modules distinguishing tumors originating metastases to different anatomical sites or developing resistance to therapies (e.g. anthracyclines). This descriptive analysis reveals its potential in understanding BC molecular features and sets the basis for the development of predictive models for clinical progression, site of metastasis, and resistance to therapy, based on gene expression.

With the same approach, I dissected the differences in network connectivity across different BC subtypes, taking network context-dependence into account. I focused on the features distinguishing the most aggressive basal-like tumors from the networks of the other subtypes, aiming at identifying molecular mechanisms conferring basal-like BC its aggressiveness and resistance to therapy. To improve robustness and statistical power, all the analyses performed in this Thesis were confirmed in independent datasets.

Crucial in this work has been the demonstration that network centrality (kWithin) is a suitable measure to identify relevant genes in this biological system, a pivotal observation for gene prioritization. Indeed, kWithin correlates with clinical features such as disease-free survival, gene essentiality for cancer cell lines' survival and, most importantly, it provides a mean to select potential upstream regulators of a module with high reliability.

Going a step further in the dissection of BC biology, I considered tumor composition at the cell-type level. Indeed, the tumor microenvironment, comprising growth factors, cytokines, and cells of different origin, is an important player in determining tumor growth, disease progression, and drug resistance. Hence, I created a database of 48 transcriptomic datasets of laser capture microdissected (LCM) breast tumors (<https://aurorasavino.shinyapps.io/metalcm/>) and identified genes robustly and consistently changing their expression either in the stroma or in the epithelium, obtaining a more detailed description of cancer than with bulk samples.

Lastly, to study tumor-stroma crosstalk, I defined a novel method, which I call “crossWGCNA”, adapting the WGCNA algorithm with an alternative measure of distance between nodes (genes). Applying it to the collected LCM datasets led to the identification of several secreted factors with high interaction scores, such as NOTCH2NLA and IL6ST, involved in Notch signaling and in the response to IL6, both signaling pathways already implied in cancer.

In conclusion, using gene co-expression networks I identified groups of genes with a high potential to play a crucial role in basal-like BC and possibly amenable to drug targeting, not only for their specific functions but also for their ability to regulate survival-related gene networks. Additionally, I proposed an extension of the WGCNA method to study tumor-microenvironment interactions at the gene expression level. Through the systematic collection of available transcriptomic data, I built robust inferences about gene expression changes in breast cancer and provided useful resources for the research community. These approaches could be successfully extended to other BC subtypes, and potentially to all tumor types for which enough expression data are available, permitting to address many additional biological questions.

Contents

1. Introduction.....	5
1.1. Breast cancer	5
1.2. Microenvironment.....	5
1.3. Biological networks.....	6
1.4. Meta-analyses	13
1.5. Aims of the thesis	14
2. Results	16
2.1. Co-expression networks in breast cancer	16
2.2. Meta-analysis of microdissected breast tumors reveals genes regulated in the stroma but hidden in bulk 56	
2.3. Meta-Analysis of bulk tumors.....	73
2.4. crossWGCNA: a co-expression-based method to infer genes involved in inter-tissue communication.....	80
3. Discussion	90
4. Methods	102
4.1. Plots and Statistical Analyses.....	102
4.2. Networks in BC	102
4.3. Metanalyses.....	108
4.4. crossWGCNA.....	114
5. Acronyms	118
6. References	121
7. Supplementary Figures.....	143
8. Acknowledgments	154

1. Introduction

This Thesis will focus on the exploration of breast cancer molecular features, especially interactions between genes that can be inferred from transcriptional data. Breast tumors will be here first considered as a whole, with the analysis of bulk tumor samples, and then in their composition, analyzing their epithelial and stromal components and investigating novel methodologies for studying their crosstalk.

Assessing the robustness of each result will be central in this work, which also comprises two large-scale meta-analyses of available transcriptomic data that permitted to derive robust observations on gene expression changes associated with cancer progression and limited to the stroma or the epithelium.

At the core of this Thesis are gene co-expression networks, which will serve as a tool applied to most of the analyses, employed to define functionally relevant clusters of genes and to prioritize genes to study as potential therapeutic targets.

1.1. Breast cancer

Breast cancer (BC) is one of the leading causes of mortality in women worldwide, for which no general efficacious treatment is available due to disease heterogeneity. BC is usually classified according to the immunohistochemical assessment of the expression of Progesterone and Estrogen Receptors (ER-positive BC), human epidermal growth factor receptor 2 (HER2-positive) and the proliferation marker Ki67 [1,2]. Tumors negative for all three receptors are defined as triple negative. Alternatively, gene expression profiling of the tumor (prediction analysis of microarray or PAM 50 assay) permits the classification in five molecular subtypes, partially overlapping with histological characterization: luminal A, luminal B, basal-like, HER2-positive and normal-like. These are correlated with prognosis and response to treatments [3]. In particular, basal-like subtype does not respond to targeted treatments such as hormonal blockage or Herceptin and shows poor outcome and high frequency of TP53 (Tumor Protein p53) mutations [4].

1.2. Microenvironment

The tumor microenvironment is an important player in determining tumor growth, disease progression, and drug resistance [33–35]. It is a composite environment comprising growth factors, cytokines, and cells of different origin such as fibroblasts, endothelial cells, and immune cells [36]. These components can support or inhibit tumor growth and are involved in multidirectional

crosstalk among each other and with tumor cells that can influence their behavior in supporting cancer. Indeed, the pro-tumoral functions of Cancer Associated Fibroblasts (CAFs) and immune cells in the tumor microenvironment are being studied as therapeutic targets [37,38].

Understanding the biology of each component of the tumor milieu is necessary to obtain a complete picture of tumors but obtaining compartment-specific gene expression profiles is laborious, and therefore most high-throughput datasets are based on bulk tissues. Nevertheless, relying on samples composed of cell admixtures may hide cell type specific signals and may create confounding effects. For example, tumor composition due to sampling variation substantially impacts genomic data [39] and tumor subtype definition [40]. Moreover, differences in the prognostic role of the same gene when measured in different compartments have been reported [41–43].

To overcome these limitations, a number of approaches have been introduced: (i) laser capture microdissection (LCM) is commonly employed to separate cell compartments that are histologically well defined [44]; (ii) spatial transcriptomics has permitted the obtaining of spatially-resolved profiles of stroma-rich and stroma-poor regions in tumor tissues [45]; (iii) single-cell techniques have permitted the distinguishing of transcriptomic profiles of different cell types within a tumor [46] and the dissection of the CAFs' transcriptional heterogeneity [47]; (iv) computational methods have been designed to deconvolve the contribution of each cell type to the final bulk gene expression profile in silico [48]. Despite all of these methods being valuable, single-cell techniques are affected by loss of information resulting from dropouts and zero-inflation, and due to the cost, they are usually only applied to screen a few tumors in a single study, impeding the correlation of gene expression profiles with clinical features [49]. Similar problems apply to spatial transcriptomics. Deconvolution methods, on the other hand, rely on strong assumptions and depend on the quality of the specific signatures of the cell type, which are applied as input in the model. Finally, LCM, despite not granting single-cell resolution, represents a good compromise to disentangle the specific contribution of the tumor epithelium and microenvironment and collect information on many clinically distinct samples.

1.3. Biological networks

Biological systems are complex in nature, their behavior being governed by the interactions of many molecular components (e.g., coding and non-coding RNAs, proteins), through several regulatory layers [5] (e.g. promoter binding, miRNA-mRNA interaction, post-translational modifications). Cancer is no exception, being the result of multiple perturbations within a single cell

that also affect cell-cell and cell-microenvironment communication. Each perturbation does not act in isolation but is influenced and in turn influences the whole system, with reciprocal relationships occurring between most components [6].

Therefore, accurately depicting a biological system such as a cancer cell requires knowledge about elements' interactions, and in particular about the regulatory layer described as Gene Regulatory Networks (GRN). Gene Regulatory Networks, like all biological networks, represent biological components as nodes and their interactions, either physical or functional, as edges (Figure 1). GRNs are the ideal reconstruction of interactions between genetic elements, comprising the activity of TFs on their targets' expression, post-translational modifications influencing a protein's impact on other elements of the network, epigenetic modifications altering transcription and many additional levels of regulation. While the overall goal of most biological network studies is the inference of GRNs, this is a complex and laborious task that is generally approached by setting some simplifying assumptions and by analyzing one kind of relationship at a time, based on physical or other kinds of interactions, as described below.

One caveat is that, despite the general assumption that physical or genetic interactions indicate shared functions or the belonging to the same molecular pathway(s), this is not necessarily true, since the information used for edges' inference is not a direct measure of functional relationship. Moreover, each network type estimates only part of the overall GRN structure (e.g., transcriptional regulation, protein-protein interactions), losing the information hidden to the specific data type used for its construction, which could be revealed combining the results obtained with additional data sources.

The most frequently studied biological networks based on physical interactions are protein-protein interaction networks (PPIs), where nodes are proteins and links indicate direct binding, but also the binding of a transcription factor (TF) to its targets' promoters can be represented in a network. A particular type of biological network are pathways, curated and deposited in repositories such as KEGG [7], where nodes can be either proteins or small molecules and edges indicate a variety of interactions, among which enzymatic reactions. Genetic interactions such as synthetic lethal interactions can also be studied as networks [8], assuming it to indicate that the two genes belong to the same pathway. Additionally, methods have been developed to build a variety of biological networks based on metabolic, single nucleotide polymorphisms (SNPs) and phenotypic data (reviewed in [9]).

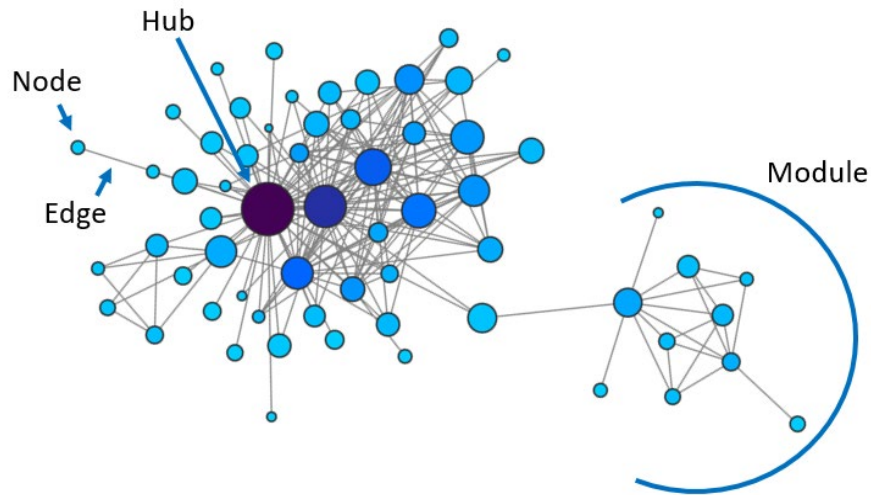


Figure 1. Example of a network, indicating nodes, edges, centrally located genes (hubs) and groups of tightly connected genes (modules)

High-throughput gene expression assays are often used to infer functional relationships between genes from correlations between their expression levels, building the so-called Gene Co-expression Networks (Figure 1). The potential relevance of this method to estimate GRNs is supported by the knowledge that genes with similar transcriptional expression profiles are likely to be regulated through similar mechanisms and to participate in the same functions, and/or to physically interact [10–13]. This information can also be combined with other information such as, for example, transcription factors binding and/or PPIs, to obtain a more complete and accurate representation of molecular elements' interrelationships [14].

Specific open-access databases have been created to store published physical and functional networks. Some of the most popular are summarized in Table 1.

Table 1. Some of the most popular databases of biological networks.

Database	Type of network	Link	Reference
DIP	PPI	https://dip.doe-mbi.ucla.edu/dip/Main.cgi	[15]
MINT	PPI	https://mint.bio.uniroma2.it/	[16]
IntAct	PPI	https://www.ebi.ac.uk/intact/	[17]
BioGRID	PPI	https://thebiogrid.org/	[18]
STRING	Various	https://string-db.org/	[19]
COXPRESdb	Co-expression	https://coexpresdb.jp/	[20]
SEEK	Co-expression	http://seek.princeton.edu/	[21]

1.3.1. Gene co-expression networks

Gene networks have proved to be a valuable tool to understand some general principles governing biological systems, revealing that gene co-expression is organized in a non-random fashion, with highly interconnected genes (hubs) and groups of tightly co-regulated genes (modules) (Figure 1). These principles have been first identified in simple unicellular organisms such as yeast but have then been demonstrated to be preserved across evolution.

Seminal studies on *Saccharomyces cerevisiae* have established that there is a relationship between genes' topological features in the network and their function [22,23]. In fact, they have shown that biological networks have a scale-free topology, with many genes having few interactors and a few genes having many interactors (central nodes or hubs) [22] (Figure 1). Centrality distinguishes the most important nodes for network's integrity: hub genes' removal has the strongest effect in inducing lethality [22,23]. This has important implications for the ability to predict the effect of interfering with specific genes on the resulting biological phenotype.

The same concepts have been extended to other organisms including humans and applied to the study of disease [24–27]. In Mendelian disorders, for example, genes harboring disease mutations with dominant phenotype display considerably higher network connectivity than genes with recessive phenotype [27,28], and cancer driver genes tend to be centrally located in protein-protein interaction networks [26]. These observations give important indications for genes' function prediction and for the prioritization of therapeutic targets [29,30].

Since biological networks are intrinsically modular [31], with genes belonging to the same module usually sharing common functions [32], identifying modules across multiple networks can help discovering patterns, and shed light on the underlying mechanisms shaping specific biological systems. In particular, much interest has been directed to the identification of shared modules in different conditions, also in cancer biology [33]. Therefore, the organization of gene co-expression networks can inform on genes and biological processes that are relevant for cell behavior, facilitating predictions on specific genes knock-downs (KD) and drug treatments' effects.

Several methods have been proposed for modules inference [34], with clustering methods being particularly suited for global networks analyses and interpretation, despite not permitting the definition of finer grained modules with possibly overlapping genes. WGCNA (Weighted Gene Coexpression Network Analysis) is a co-expression modules identification method based on clustering that assumes a scale-free topology and defines gene connectivity based on topological features [35,36].

However, gene and protein networks provide a static snapshot of molecular interactions within a tissue, while biological systems are highly dynamic. Perturbations caused by evolutionary changes, disease, environmental stresses, and oncogene activation result in rewiring the network's topology [37,38]. A simple example of changes in co-expression is Max, a transcription factor, which acts either as an activator or a suppressor on the same target genes depending on its binding partner, either Myc or Mad [39]. Accordingly, Max expression exhibits either positive or negative correlation with its targets depending on the conditions. The study of such changes can give valuable insights on altered pathways and disease mechanisms [40], providing information on those weakly differentially expressed genes that are regulated at a different level (e.g., post-translational) [41]. In fact, several reports confirmed that differentially expressed and differentially co-expressed (DC) genes tend not to overlap, and hence reflect distinct regulatory processes [42]. Also, the differential expression of lowly expressed genes such as TFs might be difficult to retrieve, while it could be evidenced by the consistent change in co-regulation of a group of their targets, identified as a DC module

Finally, the integration of various data types can build upon co-expression networks to infer the sources of variability responsible for the observed differential co-expression. As represented in the simplified model of Figure 2, the loss of co-expression between a transcription factor and its target can be accounted for by intervening regulatory mechanisms such as the transcription factor post-translational modifications, mutations, its association to alternative co-factors or epigenetic control (DNA methylation, non-coding RNAs), not necessarily leading to a detectable differential expression at the mRNA level. Adding this information to the model helps, on one side, to explain the molecular processes involved in the co-expression change and, on the other side, to identify the functional correlates of less experimentally accessible regulatory layers (e.g., protein phosphorylation). Indeed, the many regulatory mechanisms defining the properties of a biological system can't be enclosed in their entirety in co-expression networks based on transcriptomic data only, as this does not account for regulations mediated by alternative splicing, RNA editing, sncRNA/lncRNA. These regulatory mechanisms have proven to be essential in shaping cancer behavior [43,44], and, specifically in breast cancer, splicing isoforms have been shown to define prognostic signatures, regulate proliferation and invasion [45–47]. Co-expression networks allow indirectly exploring such post-transcriptional mechanisms by analyzing the regulation of the involved enzymes [48] but their molecular description of the system is still not exhaustive.

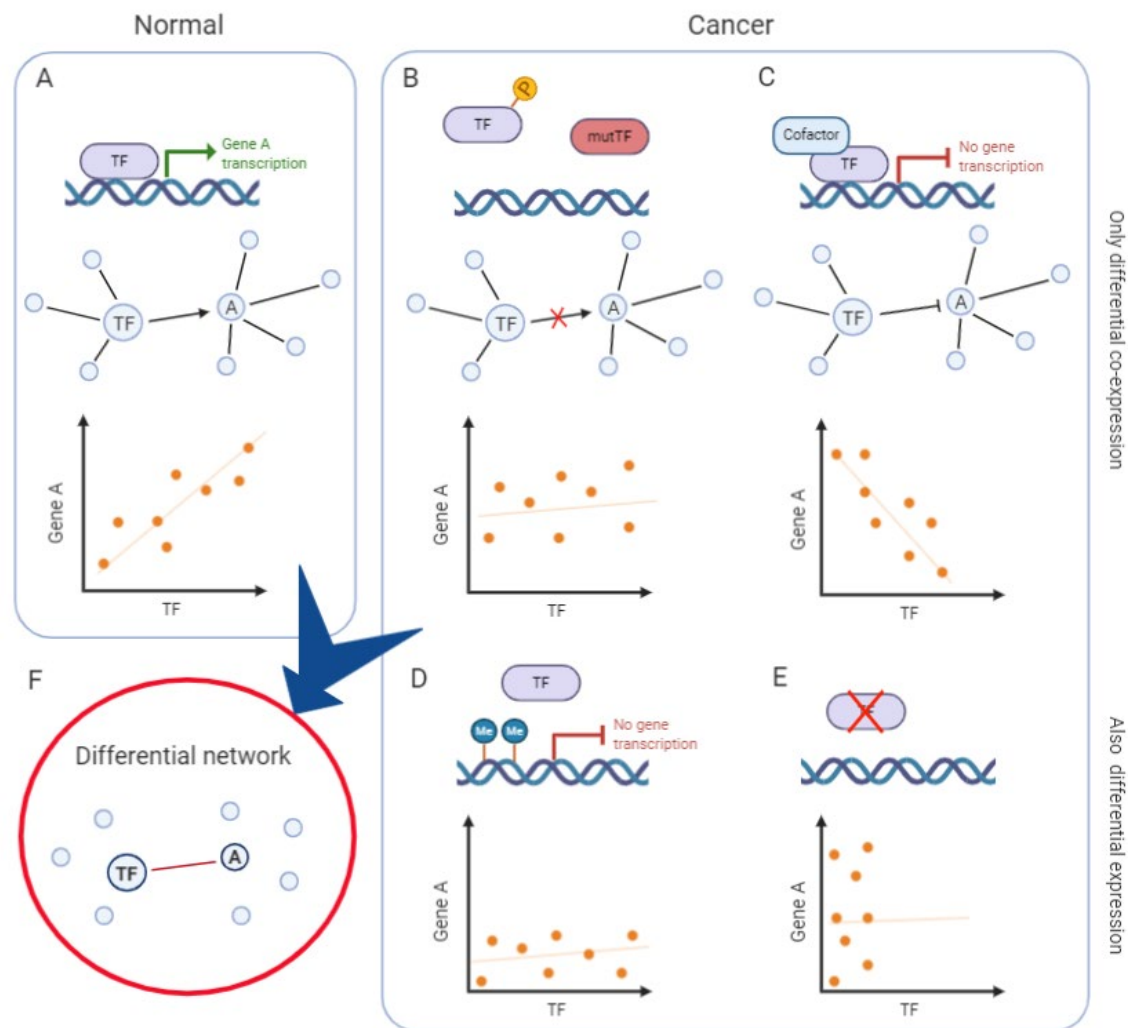


Figure 2. Mechanisms underlying differential co-expression and their impact on transcripts' correlation. Example of a simplified network in normal tissue (A) and in cancer (B, C,D,E), which comparison leads to define a differential network (F). A) Example of a simplified co-expression network in normal tissue, with a transcription factor (TF) activating its target Gene A. TF and A are connected by an edge and their expression is positively correlated. B) Post-translational modifications (e.g., phosphorylation) or mutations of the TF impede its binding to gene A promoter. Hence, TF does not activate gene A expression: they are not transcriptionally correlated and therefore not linked in the co-expression network. C) The presence of a cofactor alters the activity of the TF, that switches from acting as an activator to being a repressor of gene A expression. Hence, TF and gene A are negatively correlated. In the network, they can be represented as having an inhibitory edge or as not being linked. D) Epigenetic mechanisms (e.g., DNA methylation) inhibit TF binding and gene A expression. TF and gene A are not correlated, and they are not linked in the co-expression network. E) In the absence of the TF expression, gene A expression does not depend on its regulatory activity. The TF and gene A are not correlated nor linked in the co-expression network. These regulatory mechanisms and others alter the relationships between the TF and gene A, changing their correlation at the transcriptional level. In some cases, one of the two genes is also differentially expressed (D, E), in other cases only co-expression changes (B,C). Reconstructing the differential network between “normal” and “cancer” conditions here exemplified would lead to defining a network where only the TF and gene A are linked by an edge indicating a decreased correlation in cancer (F).

1.3.2. Networks in breast cancer

Network inference has been applied to the study of breast cancer for the identification of prognostic modules and centrally connected genes as potential therapeutic targets [49–54]. Importantly, often the identified co-expression modules have been found enriched for immune-

system-related functional categories [55], pointing to the relevance of surrounding or infiltrating cells in the biology of the whole system. Nevertheless, most published works are purely descriptive and do not exploit the wealth of information hidden in the constructed networks for the definition and investigation of specific biological hypotheses. Additionally, the problem of differences in network structure across conditions has motivated the definition of differential co-expression analysis methods [56–58], but network structure and connectivity have never been compared across breast cancer subtypes. However, this investigation could drive the generation of new hypotheses about the molecular features conferring, for example, aggressiveness to the basal-like subtype.

1.3.3. Networks to infer cell-cell communication

Biological systems can be often dissected in multiple components: different organs, tissues, or cell types that, at different levels, interact with each other and contribute to defining the system's properties and behavior. These interactions are manifest also at the gene expression level, with the expression of a gene in one cellular compartment influencing gene expression in the other. For nearby compartments, this can be explained by direct cell-cell contacts or by paracrine signals that activate signaling cascades in the target cells. For tissues or organs that are far apart, hormones might be the vehicles.

Despite their well-recognized role, experimentally studying these communications is laborious, and mostly limited to the analysis of one signal-receptor interaction at a time.

Therefore, some methods have been designed to study this problem, and applied in particular to single cell RNA-seq data ([59–63], Reviewed in [64]).

Most studies have focused on ligands expressed by sender cells and receptors expressed by receiver cells, to ease the interpretation, and compute “communication” scores between cells that are based on the assumption that the presence of ligand-receptor pairs expressed on the sender-receiver cells, respectively, implies a functional interaction. To improve the sensitivity of the methods and the quality of the resulting interactions, manually curated repositories of ligand-receptor pairs have been generated (e.g., in [64]).

Besides gene expression, additional information has been included, such as metabolite secretion data [65], protein complexes involving multiple subunits [61,63,66], or downstream signaling gene products and gene regulatory networks [67,68], using expression data of interacting cells to infer the effects of sender-cell ligands on receiver-cell expression.

In these methods, the expression levels of ligand-receptor are combined to produce a score (either binary or continuous) that is used to define an interaction ‘active’ or ‘inactive’ [64].

As an alternative measure of interaction, co-expression between ligand-receptor genes has been employed [69], positing that for some autocrine pathways, ligand and receptor are regulated by coupled mechanisms at the level of transcription. In this early work, applied to bulk data, examples of ligand-receptor pairs with significantly correlated expression in five cancer gene expression datasets were retrieved, observing that chemokines are often involved in these potential autocrine pathways [69]. Nevertheless, it must be noted that the grounding assumption of this method is strong, and prone to lead to false negative results.

The tumor-stroma crosstalk has been analyzed in very few papers using these and similar methods. For example, CCCExplorer was employed to build directed and connected network from ligands-receptor interactions to transcriptional factors (TFs) and their target genes, using DEGs between normal tissue and tumor [65,70].

Gene co-expression networks have been widely employed and proven valuable in inferring genome wide gene interactions but have been exclusively applied to individual tissues or to samples comprising a non-separable admixture of cell types. Nevertheless, the principle of co-expression networks could be extended to the study of inter-cellular communication, to provide a simple way of selecting genes mediating inter-tissue interactions, which could be then assessed experimentally. Indeed, co-expression has been applied by Oh et al. [71], describing the increased presence of self-loops between cancer epithelium and stroma, when compared to normal breast tissue. Nevertheless, self-loops (high correlation between the expression of a gene in the stroma and the same gene in the epithelium) might be indicative of external influences (either genetic changes or hormonal signaling) governing gene expression in similar ways in the two tissues, and not of stroma-epithelium crosstalk.

It is therefore clear that the currently available methods, despite their strengths in being applicable to single cell data and their ease of interpretation, leave room for improvements, particularly in relaxing the underlying assumptions.

1.4. Meta-analyses

High-throughput analyses of gene expression hold great promise for the identification of biomarkers of clinical status, with the potential of predicting outcome, response to therapy or informing researchers about molecular mechanisms underpinning disease onset and progression [72]. Nevertheless, lists of candidate genes obtained through transcriptome-based studies have

proved difficult to reproduce [73–77], raising a note of caution about conclusions that can be driven by single experiments. Sample collection and processing methods, protocols and platforms may impact on the resulting gene signatures, making them non-overlapping between studies [78]. Additional variability may be introduced by patient heterogeneity, not sufficiently represented in small samples.

To solve these issues, the vast amount of information present in gene expression databases such as Gene Expression Omnibus, ArrayExpress and EGA (European Genome Archive) [79–81] can be integrated to improve the quality of gene signatures. The advantage is twofold: on one side, a larger sample size permits increasing statistical power; on the other side, merging data obtained with different experimental settings facilitates removing single-experiments' biases, improving robustness [82,83]. Meta-analyses serve this scope, providing a quantitative approach to combine the results of studies investigating the same biological system [76]. Several methods, based on different statistics, have been proposed [84,85]: i) aggregating gene lists based on p-value [86]; ii) effect size [87], and iii) gene rankings [88].

Cancer biology is an extremely prolific field, often with tens of independent studies analyzing the same biological or clinical question in different cohorts. Hence, meta-analyses here show the most of benefit. In their simplest, yet significant, form, meta-analyses have been applied to assess the reliability of specific genes as diagnostic and prognostic markers [89,90], while whole transcriptomic datasets have been employed to the unbiased evaluation and refinement of prognostic signatures [91–96], to identify patients' subgroups [97–101], markers of metastatic tumors [102] and of resistance to treatments [103].

1.5. Aims of the thesis

In this work, I aimed at investigating breast cancer molecular mechanisms, prioritizing clinically relevant regulatory hubs as potential therapeutic targets.

Reconstructing breast cancer transcriptional networks, I developed a platform to address specific questions for BC in general and for specific BC subtypes: i) Which are the genes possibly driving clinical features? ii) Can this tool help identifying upstream regulators or downstream targets of specific signaling pathways? iii) Are there specific gene interactions explaining phenotypic differences between cancer subtypes?

In addition, I aimed at dissecting tumor-stroma crosstalk, looking for robust compartment-specific gene expression changes and developing a novel computational method to infer inter-tissue gene expression coupling.

The developed networks, databases and methods represent a tool to address the many different biological questions, such as those outlined above, identify key regulatory genes, and drive experimentally testable hypotheses.

2. Results

2.1. Co-expression networks in breast cancer

2.1.1. Network construction

I applied WGCNA (Weighted Gene Co-expression Network Analysis) [35,36] to a wide and extensively clinically annotated primary breast tumor transcriptome dataset (METABRIC)[104], defining 21 modules of co-regulated genes across all patients (20 plus the “Unconnected” module, comprising genes not included in any co-expression cluster, Figure 3).

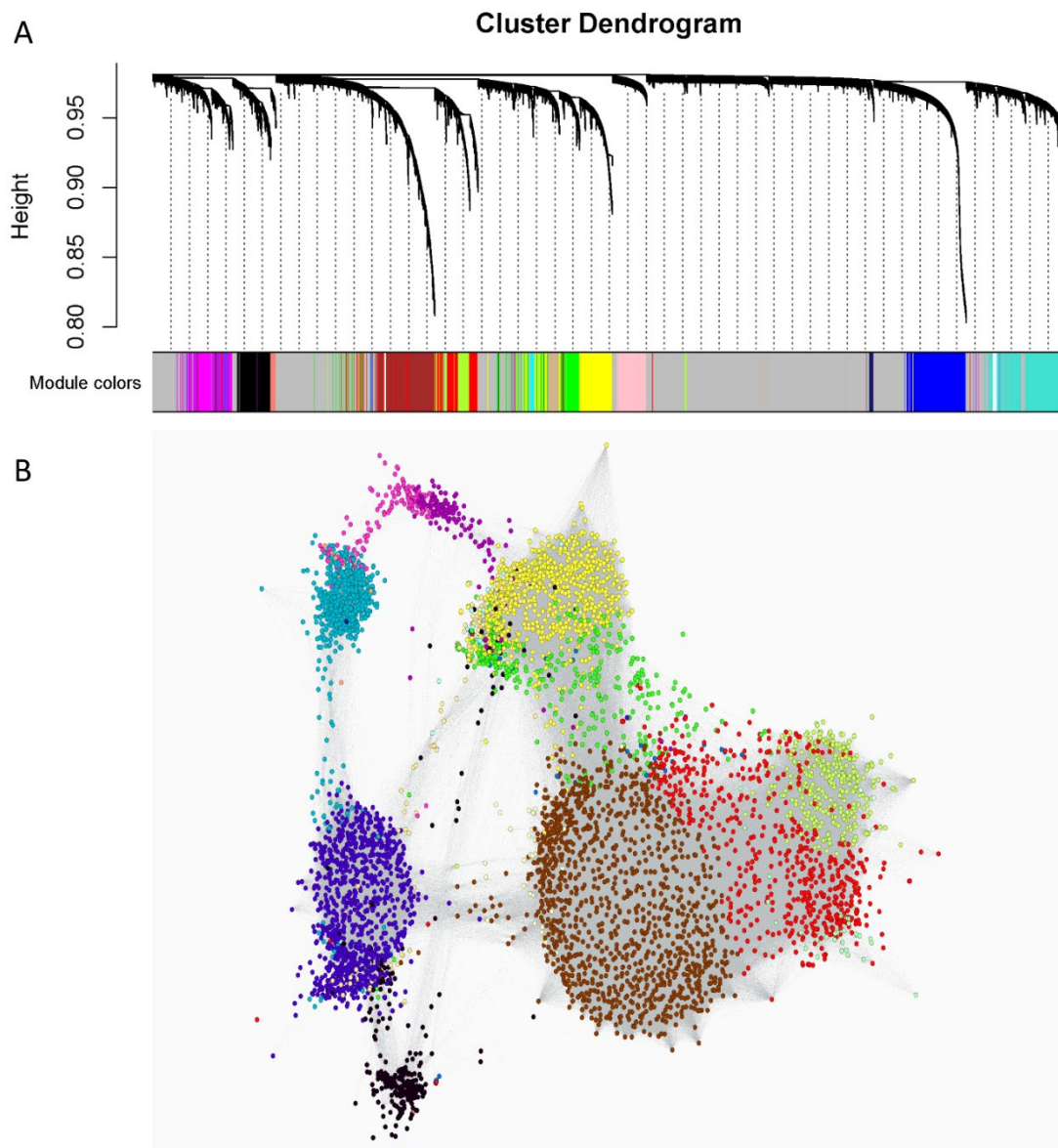


Figure 3. A) Clustering of genes performed with WGCNA, each color corresponds to a module. B) Gene co-expression network represented with Gephi: nodes are genes and edges are represented only for the strongest connections (Topological Overlap ≥ 0.05) in the network. Different colors correspond to different modules.

Modules' names were assigned by analyzing the Hallmarks sets defined by the Molecular Signature Database (MSigDB, <http://software.broadinstitute.org/gsea/msigdb>), choosing the gene set displaying the most significant overlap with the module's genes. NC denotes the modules for which no significant enrichment was found. These groups of genes were confirmed to be robust, being significantly co-regulated both in 7 independent primary breast tumor datasets and in a panel of breast cancer cell lines (<http://cancergenome.nih.gov/>, [105], Figure 4A), and their co-expression being preserved also when considering protein expression data (Figure 4B). In particular, the connectivity of modules' proteins is significantly higher than by chance in the two mass spectrometry datasets. In the RPPA dataset obtained on TCGA (The Cancer Genome Atlas) samples, the connectivity is higher than chance for proteins of the Estrogen_response, Allograft_rejection and E2F_targets modules, while it does not reach significance for all the other modules. It is important to note, nevertheless, that this dataset comprises only 188 proteins, hence limiting the analysis of connectivity for most modules which are not sufficiently represented in this protein selection.

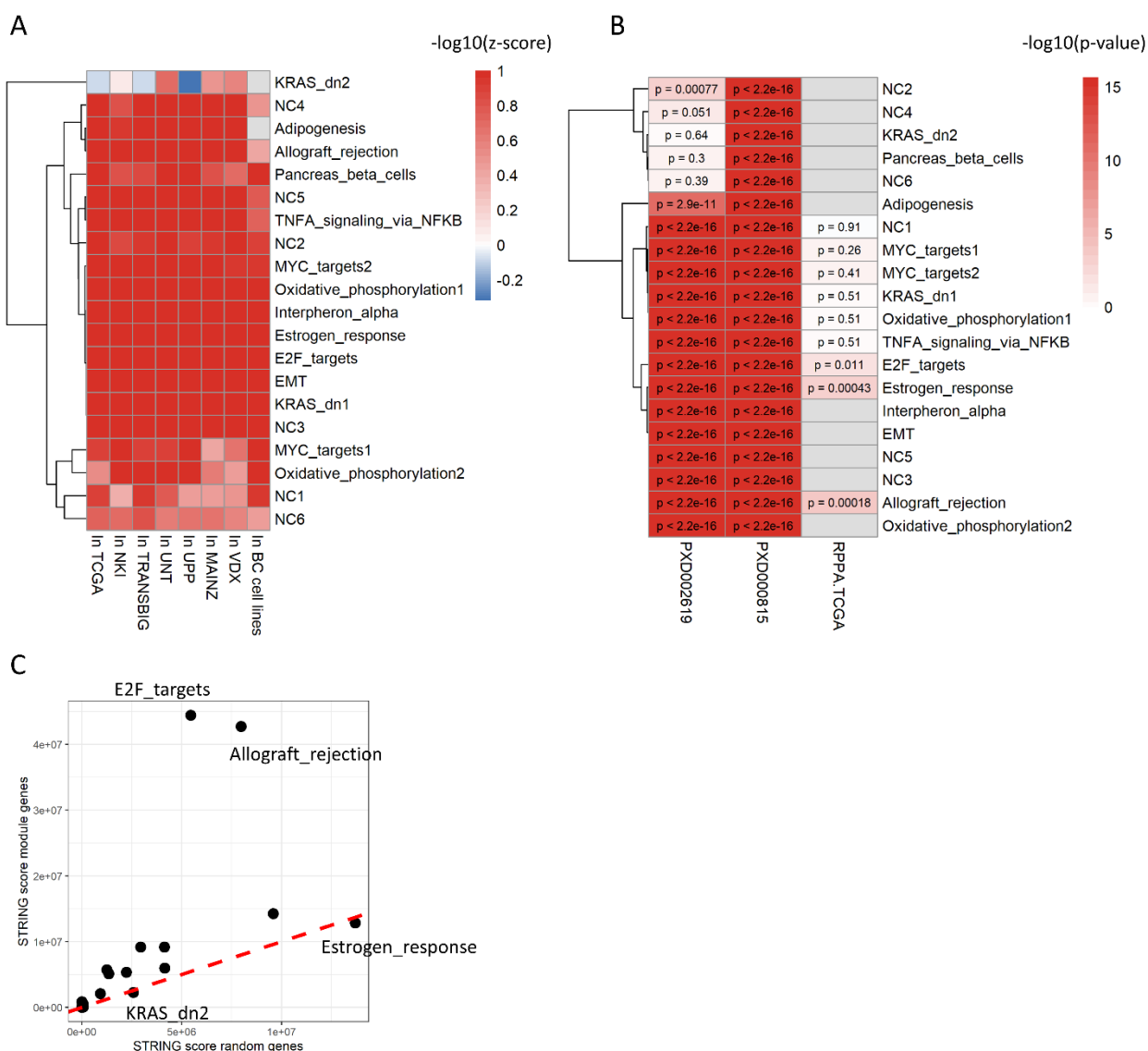


Figure 4. Breast cancer co-expression network are robust. A) Heatmap summarizing the z-score of modules' preservation across 7 independent primary BC transcriptome datasets and one BC cell lines dataset. Datasets are indicated in columns, modules are in rows and the color represents the $-\log_{10}$ of z-scores, with red corresponding to higher values, and hence higher preservation. B) Heatmap showing the connectivity of modules (rows) in proteomic datasets (columns). The color indicates the $-\log_{10}$ of the p-value for a Kolmogorov-Smirnov test comparing the correlations of modules' genes with those of random genes. C) The co-expression modules are more tightly connected in the STRING database than random genes, as obtained by summing the scores of the STRING database for each pair of interacting genes in one module (y axis) and comparing it with the corresponding measure obtained with an equal number of random genes (x axis). In this dot plot each dot represents a module, and some relevant modules' names have been added next to the corresponding dot.

Moreover, the relationships between genes within each module were confirmed by modules' enrichment in interacting proteins, as obtained from the STRING protein-protein interaction database [106] (Figure 4C).

Based on these networks, I searched for groups of genes most likely involved in determining clinical features and being particularly relevant for the aggressive basal-like subtype. Moreover, I

exploited network's topology to select genes likely involved in maintaining network's coherence, which were experimentally tested.

2.1.2. Transcriptional networks are correlated with clinical features

To identify transcriptional networks of clinical relevance, I considered modules' relationship with several clinical features: tumor grade, age at onset, patient's survival, site of relapse, and tumor progression after treatment with anthracyclines, indicative of acquired resistance.

2.1.2.1. *Grade and age at diagnosis*

To determine the relationship between a module and grade or age at diagnosis (proxy for age at onset) I calculated the correlation between modules' expression in each patient and the clinical feature. As a measure of a module's expression, I used the module eigengene, i.e. the ideal gene that best represents the whole module, based on the projection of each sample on the first principal component obtained with module's genes [35]. Many modules showed a significant correlation with both grade and age at diagnosis (Figure 5, Figure 6A, B). For example, the E2F_targets module positively correlates with tumor grade (Figure 5A), while the Estrogen_response module shows opposite correlations (Figure 5B). Indeed, all but three modules showed a significant correlation, either positive or negative, with tumor grade (Figure 6A), and an opposite trend with respect to age at diagnosis (Figure 6B). Moreover, this correlation was observed in six additional primary breast tumor transcriptomic datasets, confirming the robustness of this result (Figure 6A, B). In particular, E2F_targets is the top module considering the combination of significance between positive correlation with grade and negative correlation with age at diagnosis (Figure 6A, B). Thus, tumors with high expression of this module have features of high aggressiveness since they develop earlier and with a higher grade than tumors' average [107]. On the contrary, the Estrogen_response module is negatively correlated with grade and positively correlated with age at diagnosis (Figure 6A, B), indicating that its expression is representative of the least aggressive tumors.

2.1.2.2. *Survival*

To further test the relationships with clinical features I calculated the correlation between module eigengenes and patients' disease specific survival: again, I found modules associated with poor (Figure 5A) or good (Figure 5B) prognosis, with some of these associations consistently preserved across six independent datasets (e.g., NC4 and E2F_targets modules, Figure 6C). The same E2F_targets and Myc_targets2 modules that are correlated with high grade and early onset

are also correlated with worse prognosis, and the opposite is true for the Estrogen_response and NC4 modules (Figures 5,6).

2.1.2.3. Molecular subtype

Finally, I examined whether modules' expression was significantly different among breast cancer molecular subtypes (Figure 7) and found that the modules KRAS_dn1 and E2F_targets are the top overexpressed modules in the highly aggressive basal-like tumors (p-values $<2.2 \times 10^{-16}$), while the Estrogen_response module is overexpressed in both luminal subtypes (p-values $<2.2 \times 10^{-16}$) (Figures 7,8). The E2F_targets module is also overexpressed in the HER2 subtype when compared with all other breast cancer samples (p-value $<2.2 \times 10^{-16}$).

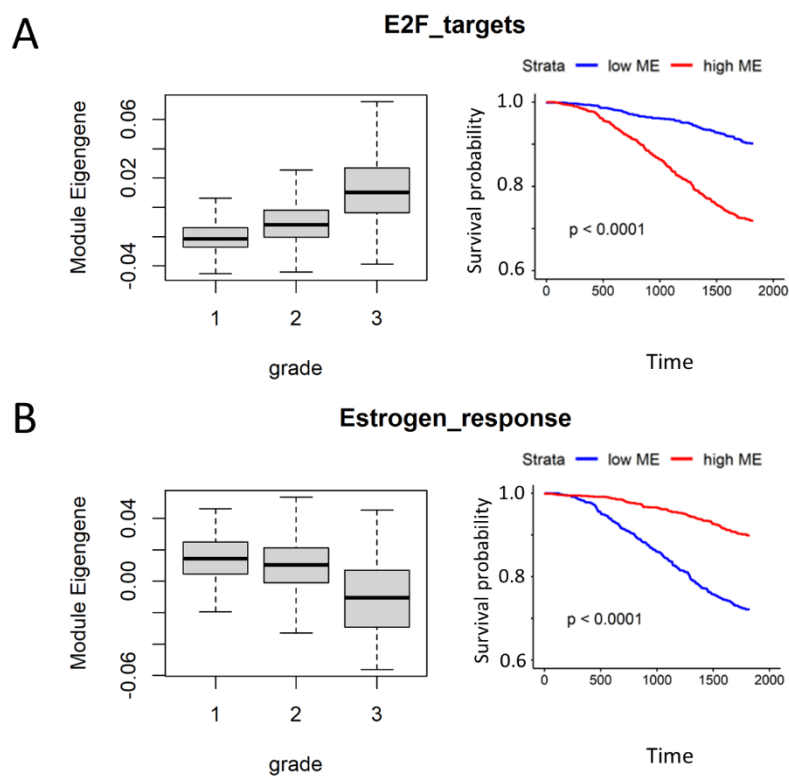


Figure 5. The E2F_targets module correlates with high grade and poor prognosis (A), and the Estrogen_response module correlates with low grade and good prognosis. Patients have been divided in two groups based on the median ME for the corresponding module.

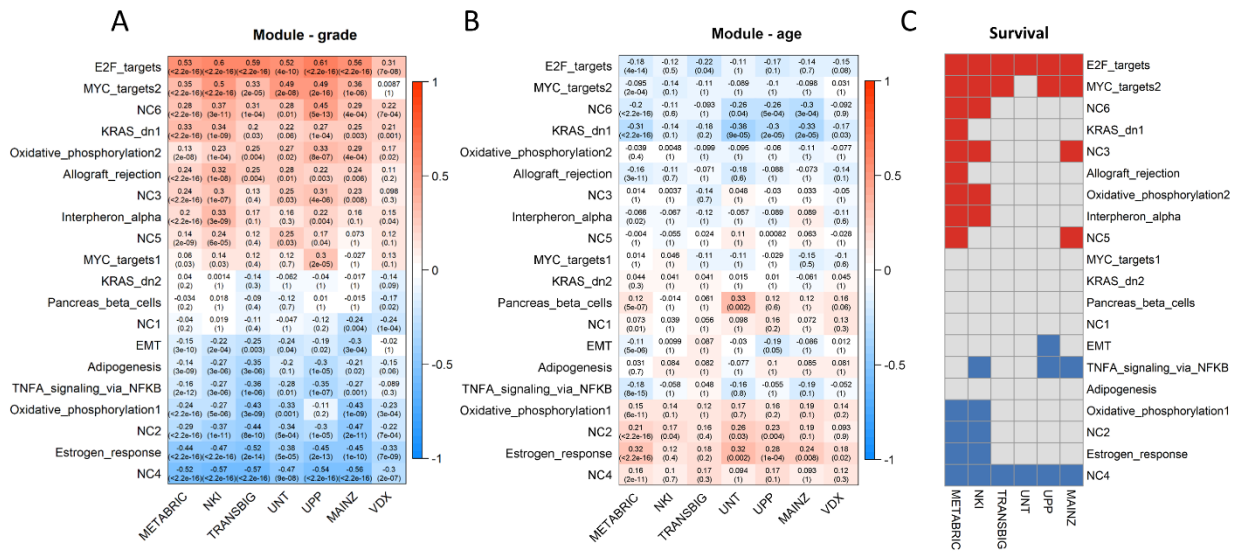


Figure 6. A, B) Correlation of modules' expression, represented by module eigengene, and tumor grade (A) and age at diagnosis (B). Each row corresponds to a different module, columns correspond to datasets. Red indicates positive, and blue negative correlations. For each combination of module-clinical feature, correlation and p-value are shown in the corresponding box. C) Correlation with survival probability in six independent primary BC datasets. Red indicates a significant correlation with poor prognosis, while blue indicates a significant correlation with good prognosis.

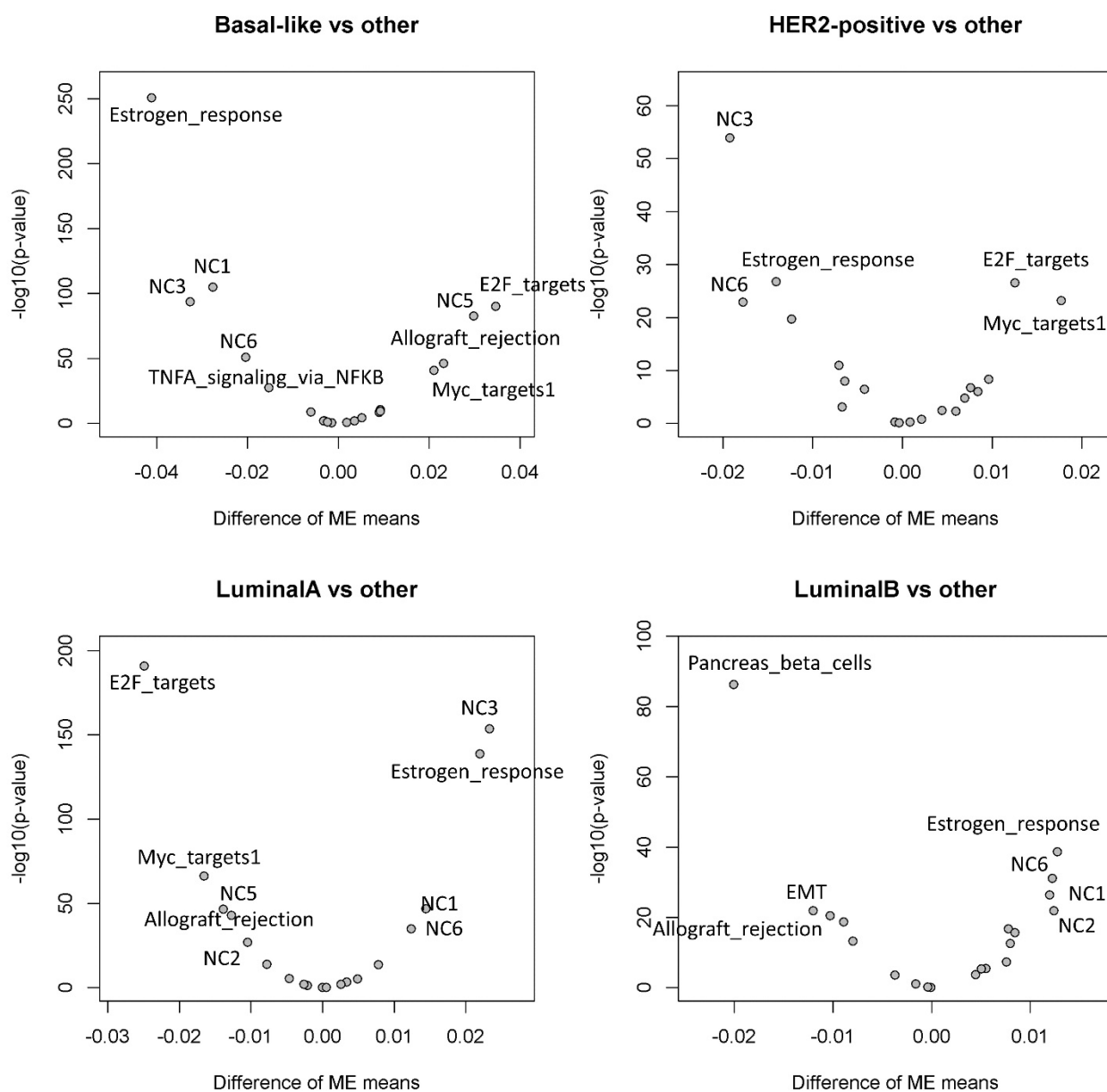


Figure 7. Difference of modules' expression across subtypes. The expression of each module in a subtype (specified in plot title) is compared with the expression of the same module in all other subtypes pooled together. X-axis shows the difference between module eigengene between the two groups, y-axis represents the significance for the comparison.

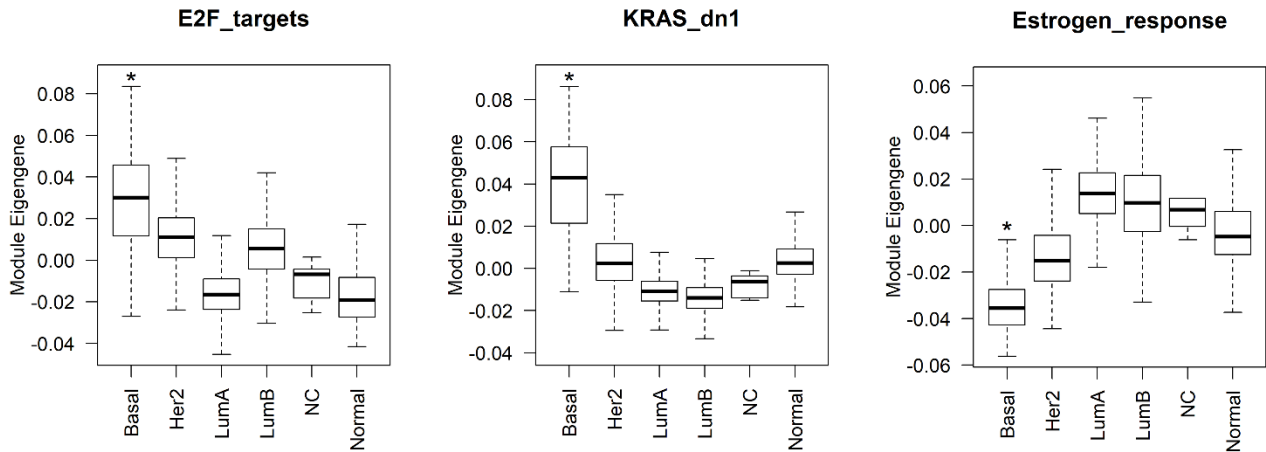


Figure 8. E2F_targets, KRAS_dn1 and Estrogen_response modules' expression across molecular subtypes.

E2_targets, KRAS_dn1 are significantly up-regulated in basal-like tumors, and Estrogen_response is significantly down-regulated in basal-like tumors. * p-value for the comparison of basal-like and the rest of the tumors pooled together $<2.2 \times 10^{-16}$.

2.1.2.4. Correlation with clinical features does not solely depend on the subtype

The observation that modules' expression depends on the tumor molecular subtype raises the possibility that some or all of the correlations with clinical features observed above are indeed entirely driven by the differential expression across subtypes, which, themselves, differ in grade, age at diagnosis (Figure 9) and prognosis [3]. To rule out this possibility, I fitted bivariate models with module eigengenes and tumor subtype as covariates and with the clinical feature (either grade, age at onset or 5 years disease free survival) as dependent variable. Figure 10 shows the p-values and the coefficients for the module eigengene term in such models, confirming that modules' relationship with clinical features does not solely depend on the tumor molecular subtype. On the contrary, modules provide additional information explaining the analyzed features across subtypes.

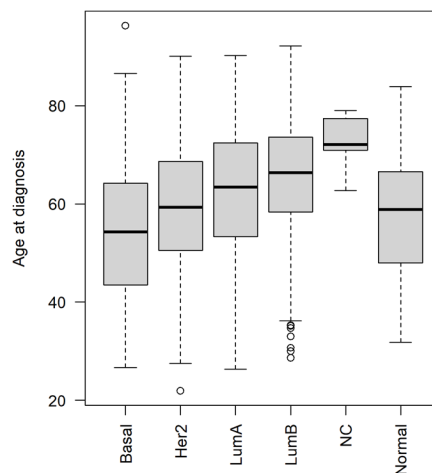


Figure 9. Relationship between age at diagnosis and tumor molecular subtype.

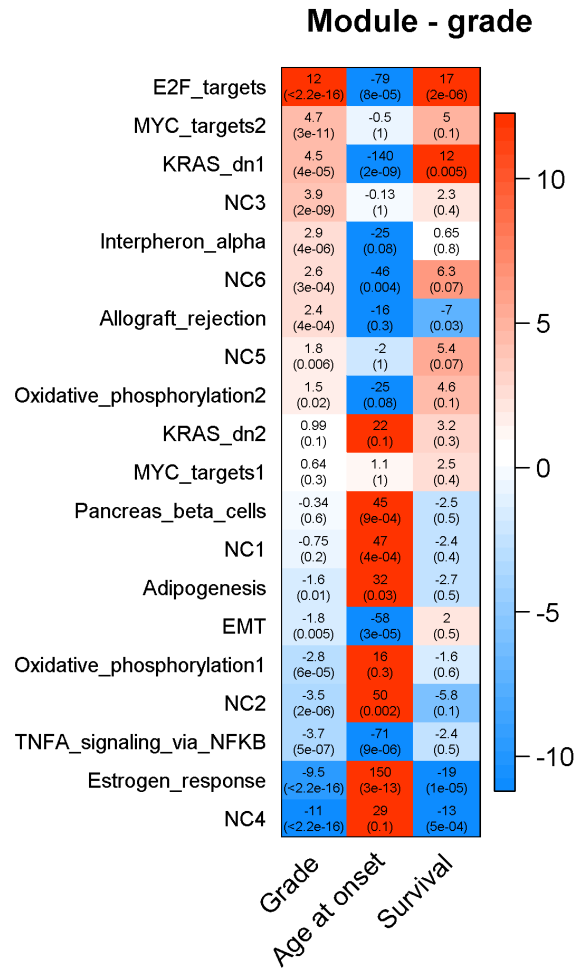


Figure 10. Modules correlate with clinical features independently of the tumor subtype. The heatmap shows coefficients and corresponding p-values for module eigengenes (rows) in models with clinical features (columns) as dependent variables, and tumor subtype as a categorical covariate.

2.1.2.5. Metastasis

To further explore the potential usefulness of defined co-expression networks in distinguishing clinically different tumors, hence helping in understanding BC heterogeneity at the molecular level and potentially setting the ground for predictive models of clinical progression, I tested whether any of the modules' expression in the primary tumor could indicate the subsequent development of metastases to specific anatomical sites. In the METABRIC cohort, annotated metastasis sites comprise lung, bone, brain, liver, lymph nodes and pleura. For validation purposes I collected four additional primary breast tumor datasets annotated for the site of metastasis (GSE12276, GSE2034, GSE2603, GSE5327, [108–111]), comprising tumors developing metastases in lung, bone, or brain. Hence, I limited the analysis to these three sites.

First, I compared modules' expression across tumors later developing a metastasis to the lung, bone or brain using the METABRIC cohort, where the modules have been initially defined. This resulted in modules either positively or negatively associated with metastases in brain and bone, and only one module negatively associated to metastases in the lung (Figure 11A). For example, the E2F_targets module has higher expression in tumors developing brain metastases and lower expression in tumors with metastases to the bone, while the Estrogen_response and NC4 modules show the opposite trend (Figure 11A). I then tested whether the identified trends could be confirmed in the four independent microarrays: most modules' expression associated with bone and lung metastases were consistently observed also in the external datasets (green dots in Figure 11A). For example, the E2F_targets module has lower expression in tumors with bone metastases in the two independent datasets comprising tumors developing this kind of metastasis (Figure 11B), and the Estrogen_response module shows the opposite trend (Figure 11C). Comparing the individual metastasis classes, these three modules display a significant difference in module eigengene between bone and the other two classes, while no difference is detected between brain and lung (Figure 11D). On the other side, none of the associations with metastases to the brain could be replicated. It must be noted, nevertheless, that the number of primary tumors giving origin to brain metastases was smaller than the other two groups, and that the lower statistical power could contribute to the lack of significance.

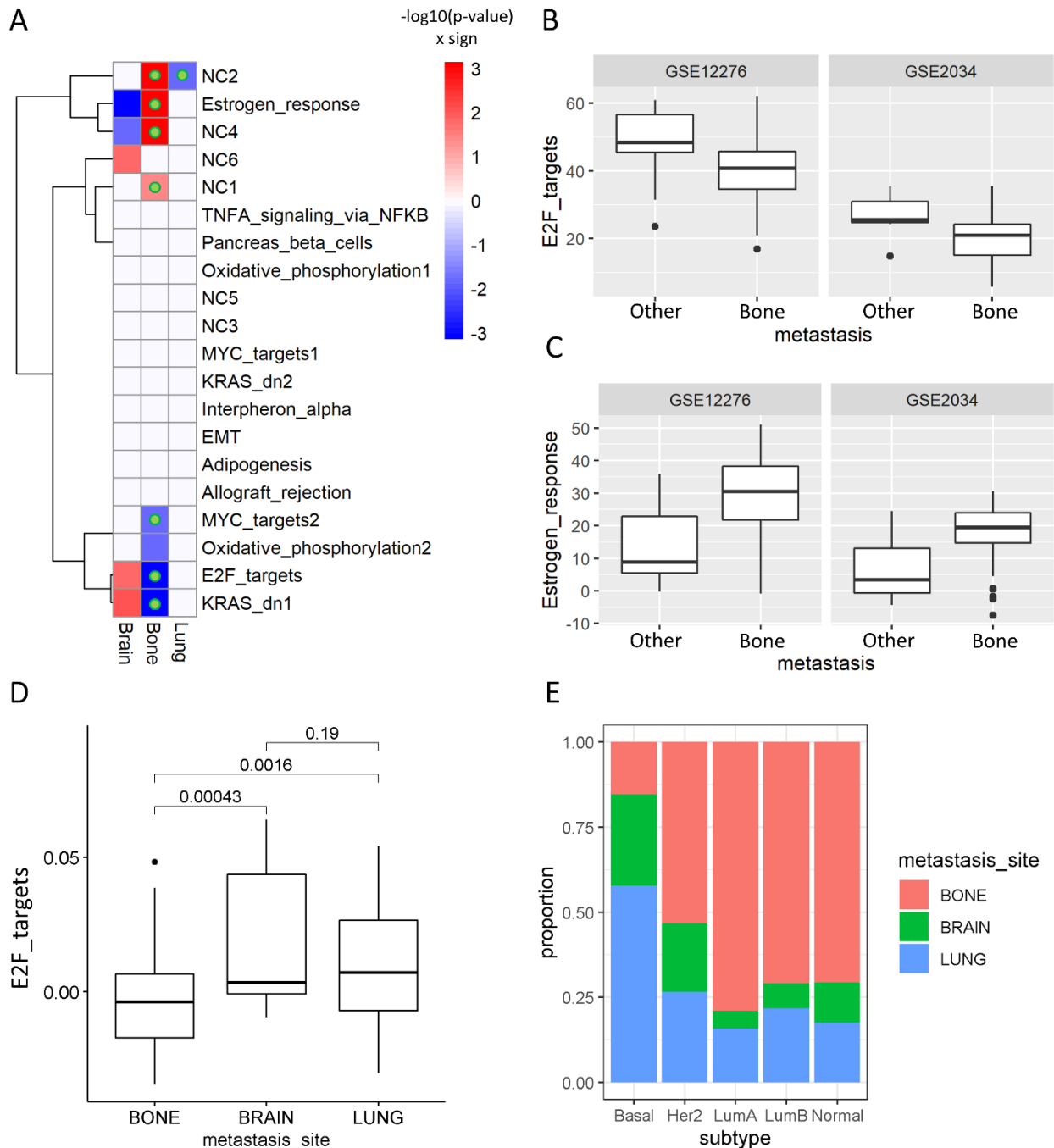


Figure 11. A) Heatmap summarizing the p-values for the differential activity of each module in tumors giving rise to metastases to the brain, bone or lung when compared to the other two sites. Rows are modules, columns are metastasis sites. The color indicates both the p-value and the sign of the differential activity (red: higher module eigengene for tumors metastasizing to the site indicated in the column name, blue: lower module eigengene). Green dots label those associations that are validated in the external microarrays. B, C) Module eigengene of the E2F_targets (B) or Estrogen_response (C) module in tumors metastasizing to the bone or not in the two external datasets annotated for this metastasis site. D) Proportion of metastases to the bone, brain, or lungs by tumor molecular subtype. E) Boxplot showing E2F_targets' expression (quantified as the module eigengene) in tumors metastasizing to the bone, brain, or lungs.

Nevertheless, analyzing the proportion of metastases to different anatomical sites across BC molecular subtypes, the differences are manifest (Figure 11E). This observation raises the possibility that the differences here reported are entirely driven by the subtype. To test this hypothesis, I fitted

linear bivariate models, adding the molecular subtype as a covariate, to explain the presence/absence of metastasis to one of the three analyzed anatomical sites. With this correction, the relationship between E2F_targets, KRAS_dn1 or Estrogen_response and bone metastases is not any more significant, while the relationship with brain metastases holds true ($p=0.007$; 0.013 ; 0.005 , respectively). Hence, the differences here observed are not entirely driven by the subtype.

2.1.2.6. Resistance

Finally, I tested whether modules could contribute to identifying tumors most likely responsive or resistant to specific treatments. As a proof of concept, I focused on the therapy with anthracyclines (a class of drugs intercalating with the DNA and interfering with DNA and RNA synthesis, especially in highly cycling cells), for which a large amount of data is available.

I collected 8 datasets from the Gene Expression Omnibus [79] comparing primary tumors with/without complete clinical response to anthracyclines or comparing cell lines before and after a chronic treatment with these drugs and their consequent acquisition of resistance (Table 2). For each dataset, I compared each module's expression, quantified through the projected module eigengene, between sensitive and resistant samples with a Wilcoxon Signed Rank test, obtaining a p-value for higher expression, and a corresponding p-value for lower expression in resistant samples of each module. Despite the high variability observed across different datasets, some modules, such as NC4, are associated with resistance to anthracyclines in most datasets, and others, such as the E2F_targets module, are associated with sensitivity in most datasets (Figure 12A). To have an overall measure of modules' association with resistance across all datasets, I collapsed the p-values with the Fisher's combined probability test [98], obtaining a single p-value for each module's association with resistance, and a corresponding p-value for each module's association with sensitivity ("up" and "down" annotations in Figure 12A, respectively). As expected, the top significant modules resulted to be NC4 and E2F_targets. Overall, the NC4, Estrogen_response, NC2, Oxidative_phosphorylation1 and TNFA_signaling_via_NFKB modules are higher in tumors resistant to anthracyclines, while the modules E2F_targets, MYC_targets2, Oxidative_phosphorylation2, KRAS_dn1, NC3, NC5 and NC6 are higher in tumors sensitive to anthracyclines.

Table 2. Datasets analyzing gene expression of tumors/cell lines resistant/sensitive to anthracyclines.

Dataset	System	Specifics	Drug	Condition 1	Condition 2
GSE87455	Primary tumor	HER2-negative	Epirubicin	Cycle 2 (chemo only)	Surgery
GSE22226	Primary tumor		Doxorubicin	Pathological complete response	No pathological complete response
GSE16446	Primary tumor	ER-negative	Epirubicin	Pathological complete response	No pathological complete response
GSE25055	Primary tumor	HER2-negative	Doxorubicin	Pathological complete response	No pathological complete response
GSE54326	Cell line	MDA-MB-231, MCF7, SKBR3 and ZR-75-1	Epirubicin	Parental cell line	Resistant cell line
GSE162187	Primary tumor		Epirubicin	Sensitive	Resistant
GSE47013	Cell line	MDA-MB-231	Doxorubicin	Parental cell line	Resistant cell line
GSE155478	Cell line	MCF7	Adriamycin	Parental cell line	Resistant cell line

To assess the robustness of this result, I checked the correlation of the modules NC4 and E2F_targets with breast cancer cell lines' sensitivity to the anthracycline doxorubicin, obtaining cell lines' gene expression profiles from the Cancer Cell Line Encyclopedia and the sensitivity to doxorubicin from the GDSC drug sensitivity screening [112]. In this context, NC4 expression shows a positive correlation with doxorubicin's IC50 (Figure 12B), while E2F_targets is negatively correlated with it, even if just as a trend (Figure 12C). Accordingly, TCGA tumors treated with anthracyclines and showing complete response or progressive disease display the corresponding trend of differential expression, despite the lack in significance, likely due to the small sample size.

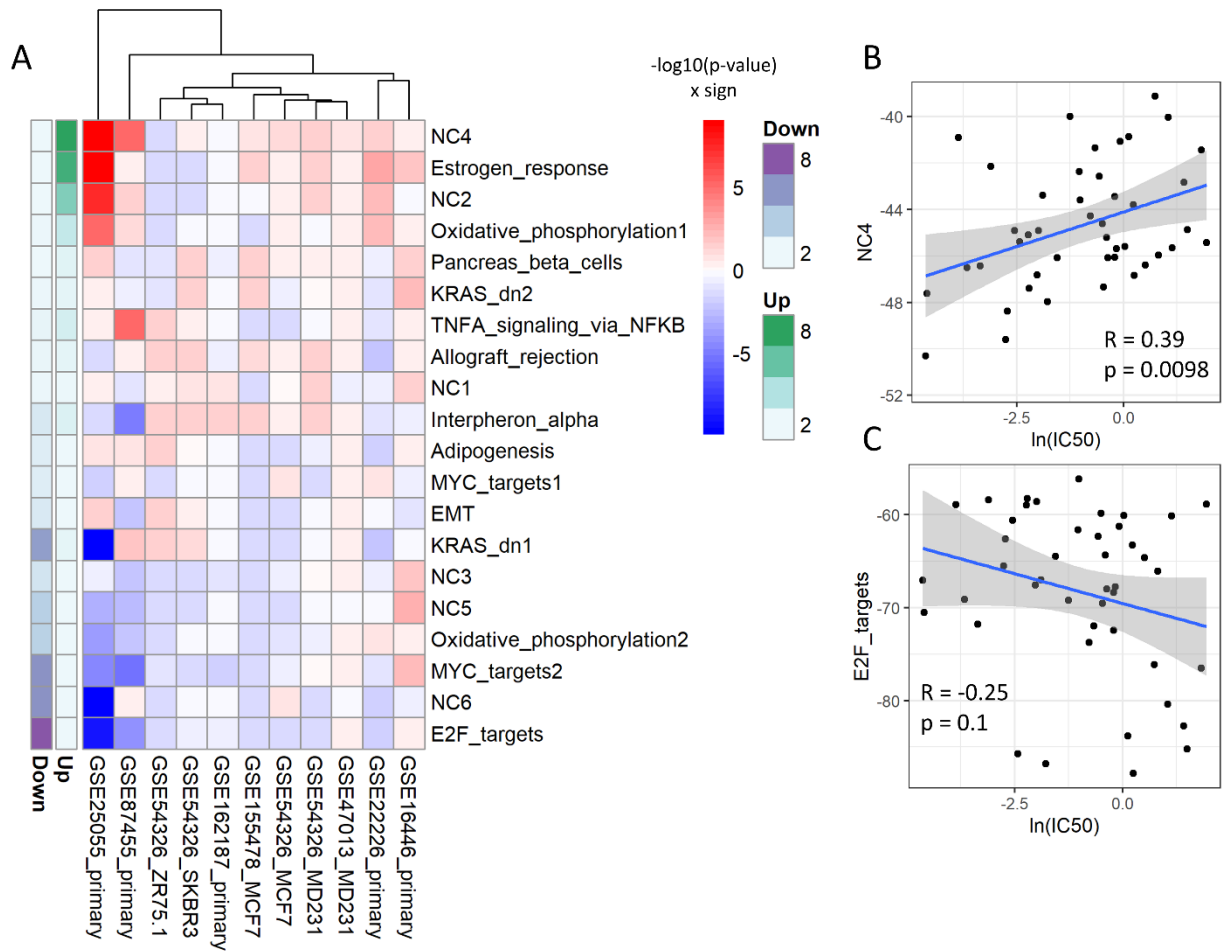


Figure 12. Modules' expression correlates with resistance to anthracyclines. A) Heatmap summarizing the differential expression of each module (rows) between breast tumors sensitive or resistant to anthracyclines, and between breast cancer cell lines before and after the acquisition of resistance to anthracyclines, as obtained with different transcriptome datasets (columns). Color intensity indicates the significance of positive (red) or negative (blue) association with resistance. The color code on the left side, indicates the overall p-value for positive (green) or negative (purple) association with resistance, obtained through the combination of individual datasets' p-values through the Fisher method. B, C) Correlation between the IC50 of doxorubicin across breast cancer cell lines and NC4 (B) or E2F_targets (C) expression.

2.1.3. Transcriptional networks to build predictive models

In the previous sections, modules' expression has been revealed to be related with several clinical features and to be consistently associated with sensitivity/resistance to drugs. Therefore, modules could serve as the basis for building predictive models of tumors' clinical progression. As a proof of concept, I have developed a cross-validated model of disease-free survival based on module eigengenes or modules' hubs, and either considering all patients from the METABRIC cohort or only patients belonging to a specific molecular subtype.

Models using all patients show good concordance indexes, while subtype-specific models are significantly less accurate (Figure 13). Overall, as expected, module eigengenes, combining the expression levels of many genes, are the most suitable predictors of survival, permitting better than

random predictions for the global model and also for the Basal-like and the Luminal B specific models. On the contrary, reducing the number of measured genes to the top 5 hubs results in better-than-random performances only for the model with all patients and for the Luminal B specific model.

In the clinical setting, reducing the number of genes to be measured for predictive purposes would be extremely useful, reducing costs and hence permitting wider applicability of the predictor. Hence, additional efforts could be devoted to introducing feature selection algorithms to prune out the meaningless modules or hubs from the final model.

Nevertheless, it is worth noting that a 5-hub model appropriately predicts Luminal B patients' survival. This result, despite the relatively low concordance index (Figure 13), is particularly interesting given that subtype-specific models are rare and generally less appropriate than models applicable to the whole cohort of BC patients.

Moreover, with the appropriate amount of data, similar predictive models, and not merely descriptive models as shown in the previous paragraphs, could be built to predict the site of metastasis or the response to the treatment.

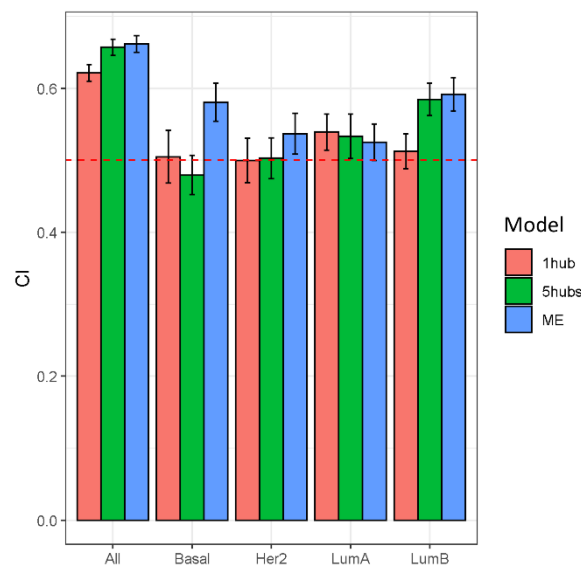


Figure 13. Concordance index (CI) for the models predicting 5-years DSS in the METABRIC cohort, using all module eigengenes (ME, light blue), the module eigengene based on the first 5 hubs (5hubs, green) or the top centrally located gene (1hub, pink). Models have been generated on all patients or separating patients by tumor subtype.

2.1.4. Transcriptional networks represent both tumor epithelium and stroma

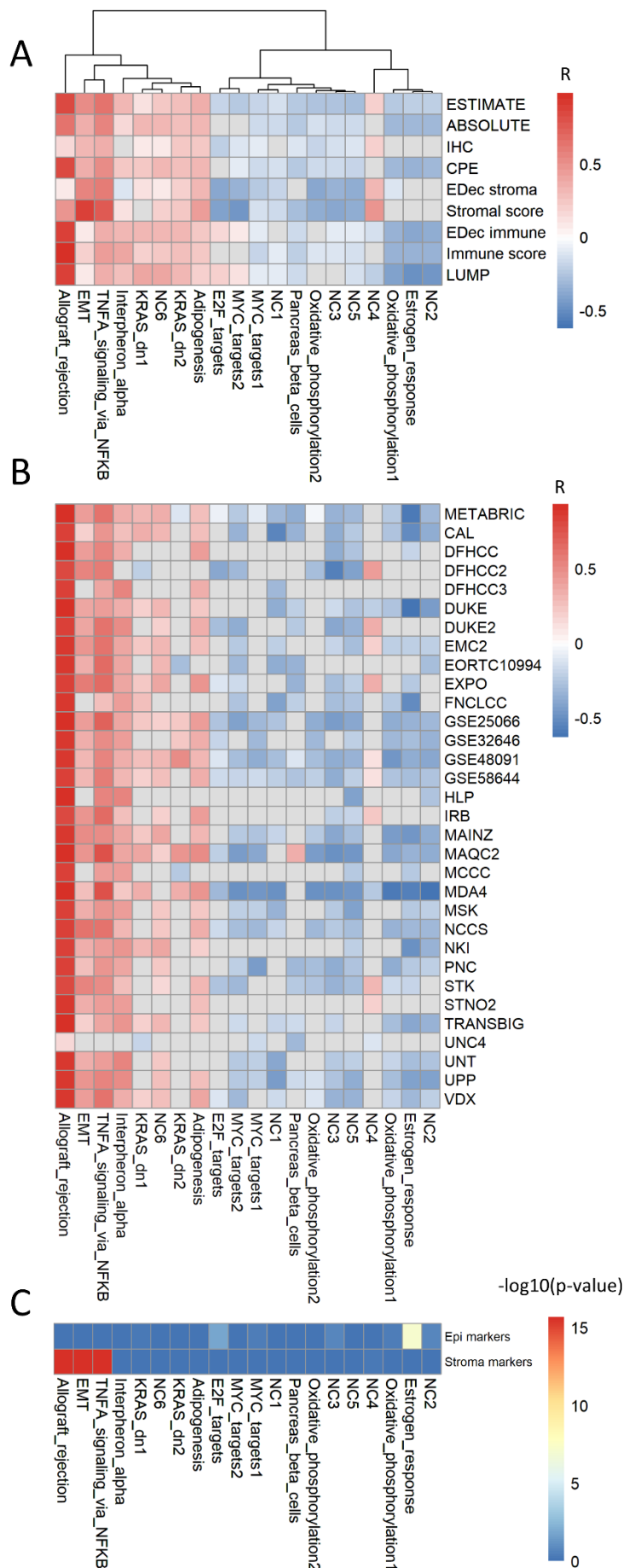
The gene co-expression network here analyzed has been obtained from primary breast tumors' gene expression profiles, likely comprising mostly cancer epithelial cells, but also microenvironmental cells such as cancer associated fibroblasts, endothelial cells, or immune cells. Hence, at least part of the gene expression signal employed to build the network is likely linked to

the proportion of these cell types infiltrating the tumor, and to their gene expression patterns in different individual tumors.

To dissect this relationship, I employed the estimation of cell types' proportions in TCGA samples obtained with various algorithms [113,114] and tested modules' correlations with specific cell types' abundance. As shown in Figure 14A, most modules are, to various degrees, correlated with immune or stromal percentage. In particular, the *Allograft_rejection* module strongly correlates with immune infiltration, while the *EMT* and *TNFA_signaling_via_NFKB* modules correlate with the presence of stromal cells. On the other side, the *E2F_targets* and the *MYC_targets1* modules negatively correlate with stromal percentage.

To assess the robustness of these observations, I used an additional pool of 33 transcriptome datasets of primary breast tumors [115] and estimated the tumor purity as previously shown [113], obtaining a good agreement with previous results on the TCGA cohort (Figure 14B). The correlation of *Allograft_rejection*, *EMT* and *TNFA_signaling_via_NFKB* with the stroma is also confirmed by the high overlap of these modules with stromal-specific markers (Figure 14C) identified through a comprehensive meta-analysis of laser capture microdissected breast tumors (Paragraph 2.2, [116]). Similarly, the *Estrogen_response* and the *E2F_targets* modules were confirmed to be associated with the epithelial and not with the stromal component (Figure 14C). Finally, these observations were further validated through the comparison of projected modules' eigengenes across different cell types profiled through single cell RNA-sequencing of breast tumors (GSE118390, GSE123837, PRJEB35405), showing that some modules comprise genes typically expressed in specific cells of the microenvironment.

Hence, some modules correlate with non-epithelial cells' percentage in primary tumors and likely represent, at least in part, the molecular regulatory networks of these cell types. A finer dissection of the relevance of each module for different cell types can be obtained through the analysis of single cell RNA-seq data, separating epithelial cells, fibroblasts, and immune cells. Instead of considering the overall expression through module eigengenes, which does not necessarily imply co-expression, the connectivity of each module across cell types can be evaluated, indicating whether the corresponding genes are co-regulated or not. It is interesting to note that, computing the module preservation in epithelial cells across 10 different patients from 4 independent datasets, not only the *E2F_targets* module, but also the *Allograft_rejection* module appears to be co-expressed in epithelial cells (Figure 15), despite its correlation with stromal percentage (Figure 14).



This indicates that the same module is relevant for multiple cell types and potentially represents co-expression *across* cell types, due to cell-cell interactions influencing reciprocal gene expression.

This observation motivated the development of a co-expression method to infer cell-cell interactions, which will be described in Paragraph 2.4.

Figure 14. Modules correlate with stromal percentage. A) Pearson's correlation between estimated stromal and immune percentages (in different rows), calculated with 9 different algorithms, and modules' eigengenes in the TCGA breast cancer dataset. Red indicates positive and blue indicates negative correlation. Non-significant relationships are in grey. EDec (Epigenomic Deconvolution of stromal or immune percentage), ABSOLUTE (based on somatic copy-number data), IHC (haematoxylin and eosin staining); LUMP (leukocytes unmethylation for purity); stromal and immune score (based on ssGSEA using gene expression profiles of 141 stromal or immune genes); CPE (combining ABSOLUTE, ESTIMATE, IHC and LUMP scores). B) Pearson's correlation between estimated stromal percentage (ESTIMATE algorithm) and modules' eigengenes across 33 primary breast tumors' datasets. Non-significant relationships are in grey. C) Enrichment of modules' genes in epithelial or stromal markers derived from a meta-analysis of breast cancer LCM datasets. Red indicates high significance.

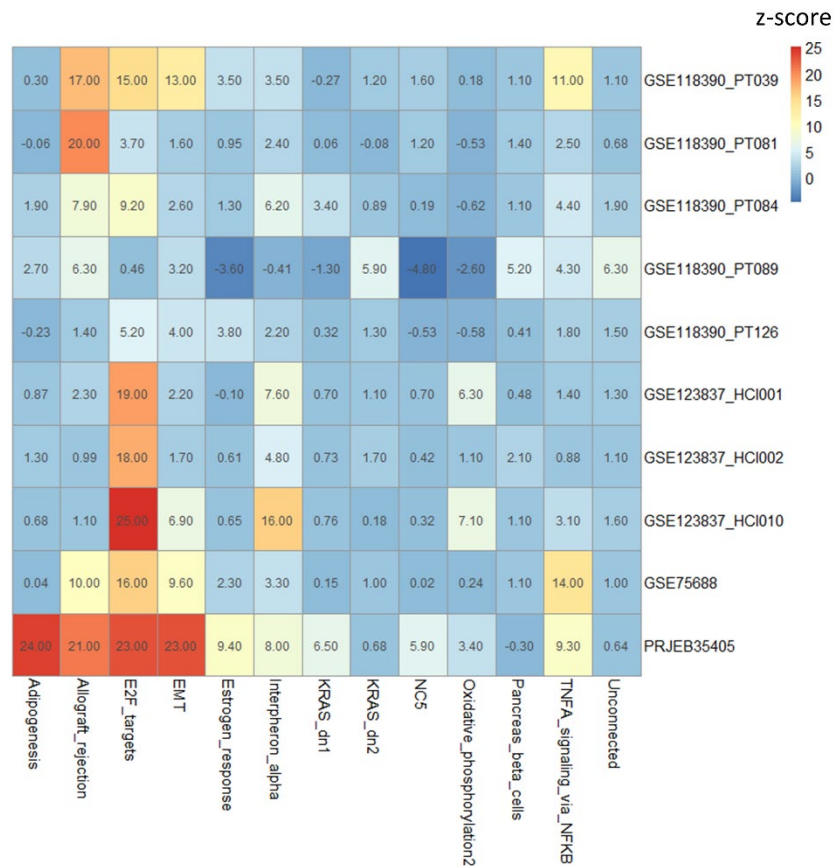


Figure 15. Module preservation in single cell datasets of breast cancer epithelial cells. Rows are different BC patients; columns represent different modules. Red indicates the z-score for the preservation of each module in the epithelial cells of the corresponding patient.

2.1.5. Identification of basal-like specific networks

The complexity of reconstructing functionally relevant regulatory networks is partly because genes may play multiple functions and that their expression, interactions, and significance can be strongly influenced by different conditions, such as, for example, gender or tumor subtype. Moreover, in dissecting the relationships between networks' activity and clinical features, the results can be influenced by the co-linearity of multiple covariates, such as age at diagnosis and subtype, both correlating with modules' expression, but clearly inter-dependent (Figure 9). This problem had already been addressed in Paragraph 2.1.2.4, showing that in multivariate models, considering the tumor's molecular subtype, the relationship between modules' expression and clinical features is preserved (Figure 10).

Here, to distinguish the effects of modules' expression from those of the tumor subtype with even more confidence, I analyzed each subtype separately building co-expression networks from samples belonging to the same BC molecular subtype. This analysis has the additional advantage of

permitting to account for potential differences in network structure between subtypes, hence avoiding misleading inferences, as described in the Simpson Paradox [117] (Figure 16).

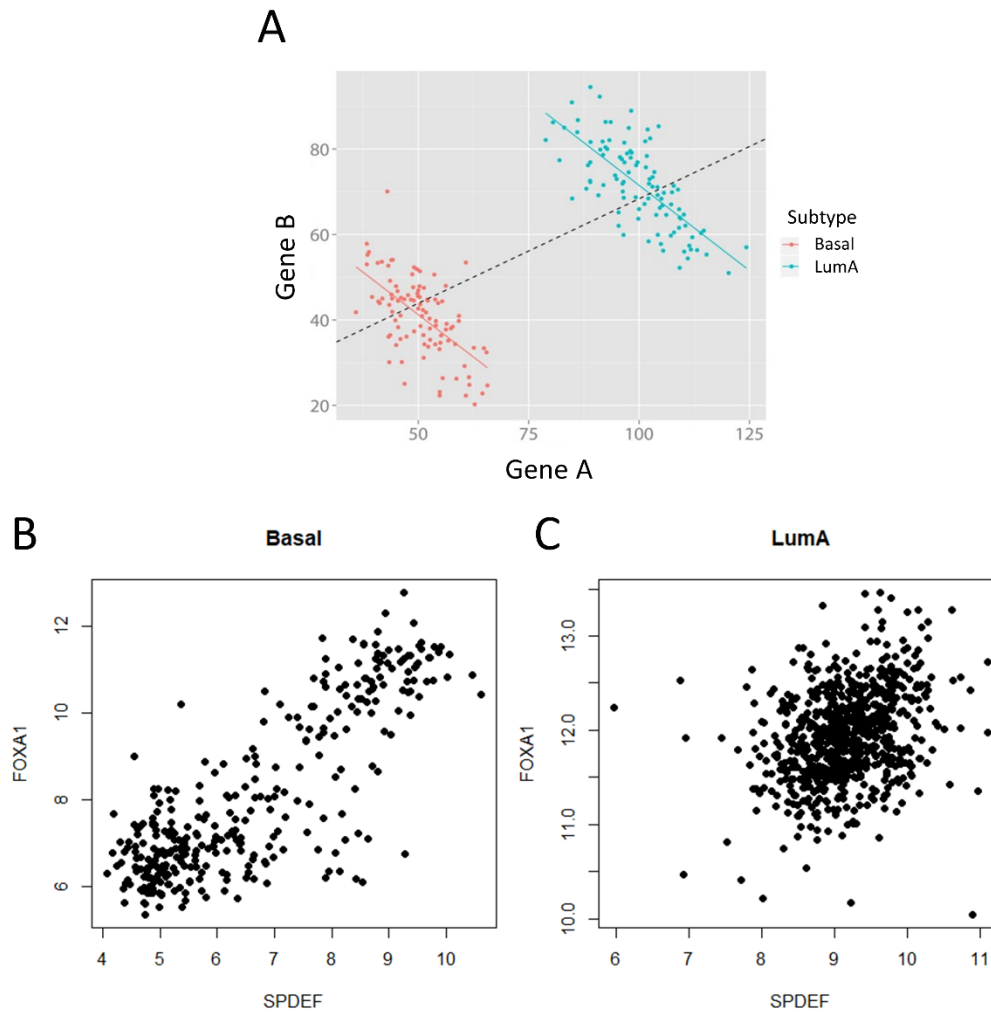


Figure 16. Exemplification of the Simpson Paradox. A) Two ideal genes display overall positive correlation but are negatively correlated within Basal-like or Luminal A subtypes. Correlation between FOXA1 and SPDEF is strikingly different between Basal-like tumors (B) and Luminal A tumors (C) in the METABRIC cohort.

To dissect the specificity of gene networks in established molecular subtypes of breast cancer, WGCNA was applied to each set of samples classified as basal-like, luminal A, luminal B, HER2-positive. This approach has two advantages: i) having a more homogenous set of samples as input, WGCNA can more reliably identify gene networks based on gene interactions and not on differential gene expression between sets of samples (determined by different molecular subtypes, in this case); ii) it permits identifying differences between network structures across subtypes. Hence, 20 gene modules were identified in basal-like tumors, 21 in luminal A, 19 in luminal B, 10 in HER2-positive tumors. I then focused on the basal-like network, searching for differences in network structure that could account for the higher aggressiveness of this specific subtype. Therefore, I selected basal-like specific gene modules based on two criteria: at least 40% of module genes falling

in the “unconnected” module in other subtypes’ networks (Suppl. Fig. 1); module genes’ intramodular connectivity (kWithin, Figure 17) being higher in basal-like than in other subtypes both in METABRIC, where networks have been defined, and in the independent dataset TCGA (Figure 18A, B). In this analysis, networks derived from normal-like tumors were not used due to the low number of samples in TCGA (23) that does not lead to a reliable network inference. With this approach, three gene modules more tightly co-regulated in basal-like tumors than in the other subtypes were identified. These were named, according to the same criteria used for global modules, b (for basal-like) E2F_targets, b_Estrogen_response and b_Estrogen_response_early.

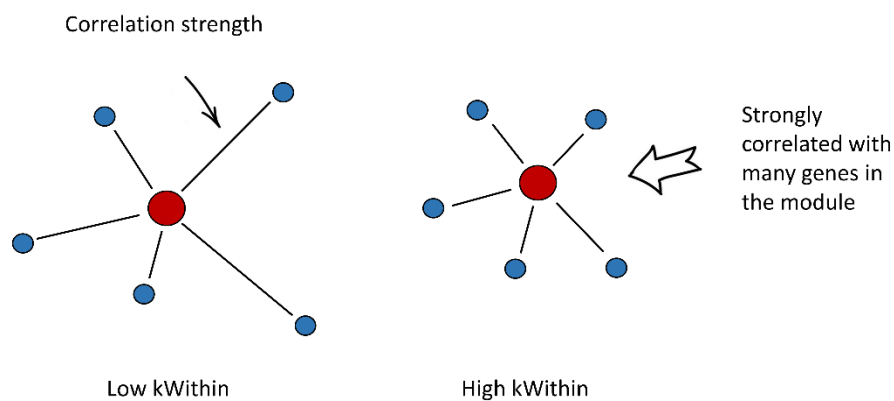


Figure 17. kWithin quantifies the overall connectivity of genes with their neighbors in the same module. The red node on the left has low connection strength (long edges) with its neighbors and therefore low kWithin, while the red node on the right has high connection strength (short edges) and high kWithin.

Notably, expression of the b_E2F_targets module genes is up-regulated in basal-like tumors while that of the b_Estrogen_response module genes is down-regulated, implying that modules co-regulation and expression levels are independent (Figure 18C). The b_Estrogen_response module comprises genes specific of ER-positive tumors, such as ESR1 (Estrogen Receptor 1) itself, FOXA1 (Forkhead Box A1) and GATA3 (GATA Binding Protein 3), which are known to act in the same network in ER-positive breast tumors [118]. Indeed, the b_Estrogen_response module is enriched for the global Estrogen_response module genes (234/389 genes, Fisher test $p\text{-value} < 2.2 \times 10^{-16}$), which are in turn up-regulated in luminal molecular subtypes, in agreement with the down-regulation in basal-like tumors (Figure 18C).

These results can be interpreted as luminal specific genes being consistently down-regulated in basal-like tumors, potentially through shared mechanism, while being spread across different pathways upon activation, thus becoming less consistently regulated.

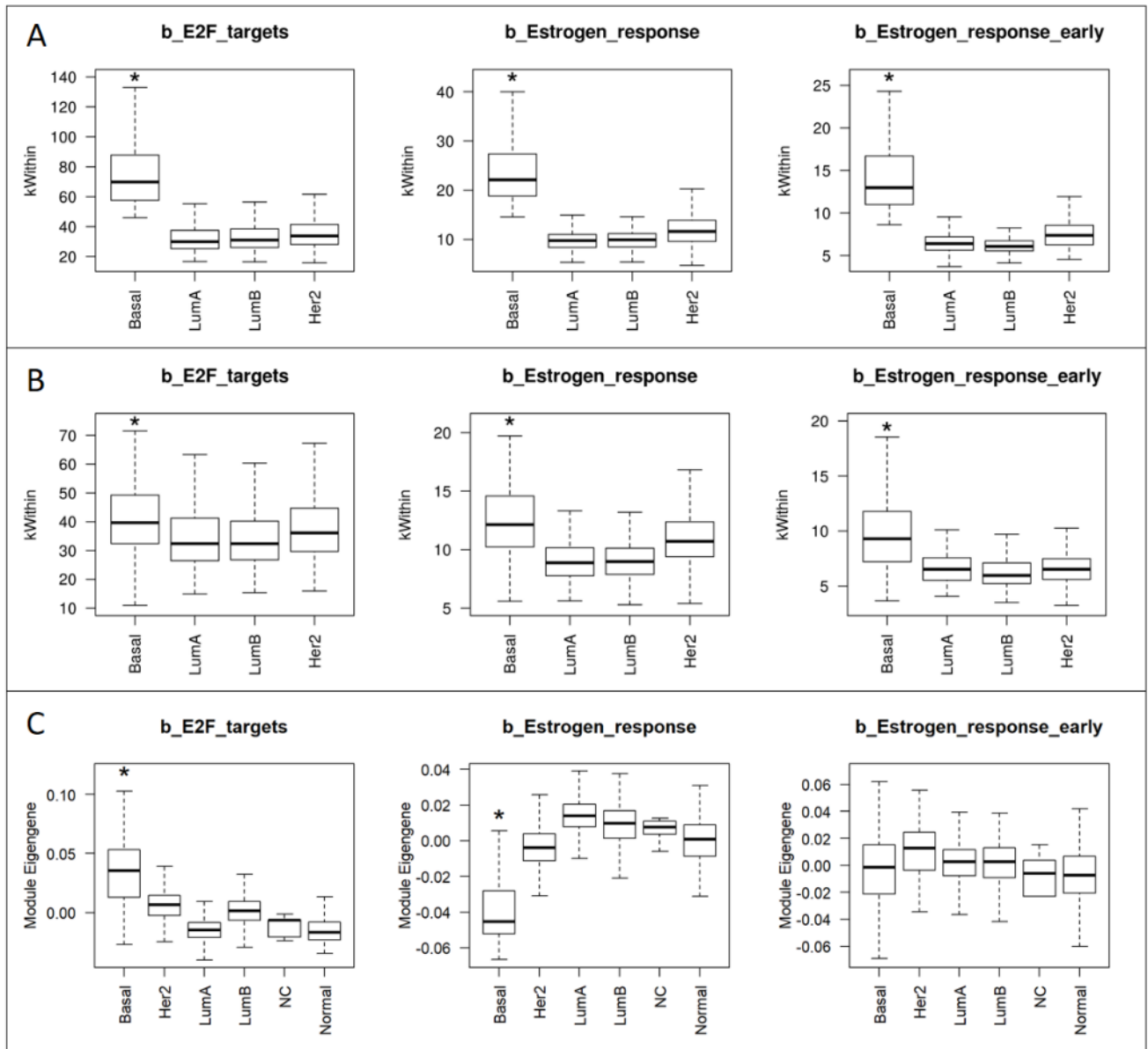


Figure 18. Intramodular connectivity of b_E2F_targets, b_Estrogen_response, b_Estrogen_response_early across BC subtypes as obtained calculating the kWithin in Metabric (A) and in TCGA (B). (C) Expression of the three modules represented by Module Eigengene across BC subtypes. * p-value for the comparison of basal-like tumors and all the other subtypes pooled together $< 2.2 \cdot 10^{-16}$.

Of note, assessment of robustness and correlations with clinical features were performed similarly to global modules (Figure 4, Paragraphs 2.1.2.1-2), confirming that all modules except for b_KRAS_dn are dataset independent (Figure 19), and showing that the b_E2F_targets module is correlated with high grade, early onset and low disease specific survival in basal-like tumors (Figure 20A,B), and that the correlations with grade and age at onset are preserved across independent datasets as a trend, despite a lack of individual-dataset's significance likely due to the reduced number of samples with respect to the analysis of global modules with all patients' data (Figure 20C,D).

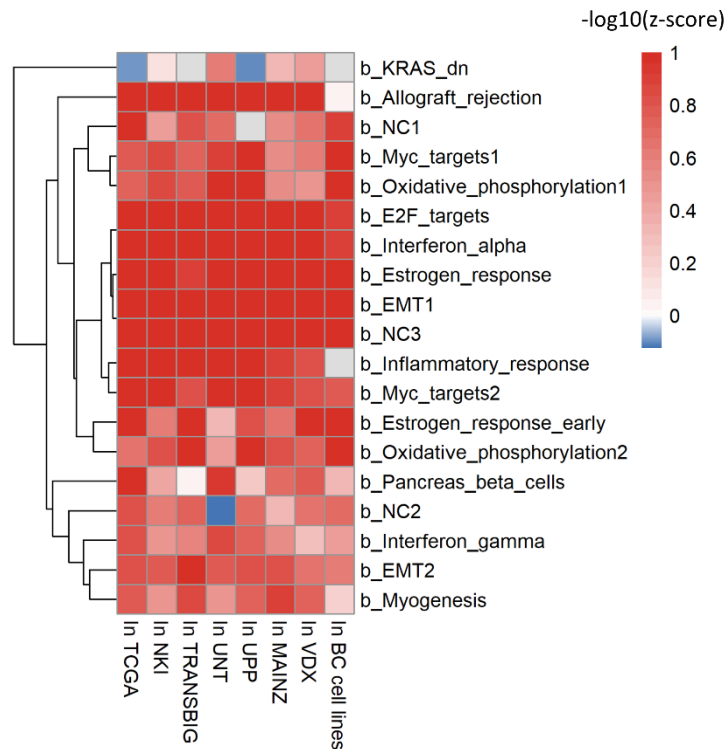


Figure 19. Basal-like breast cancer co-expression network are robust. Heatmap summarizing the z-score of modules' preservation across 7 independent primary BC transcriptome datasets and one BC cell lines dataset. Datasets are indicated in columns, modules are in rows and the color represents the $-\log_{10}$ of z-scores, with red corresponding to higher values, and hence higher preservation.

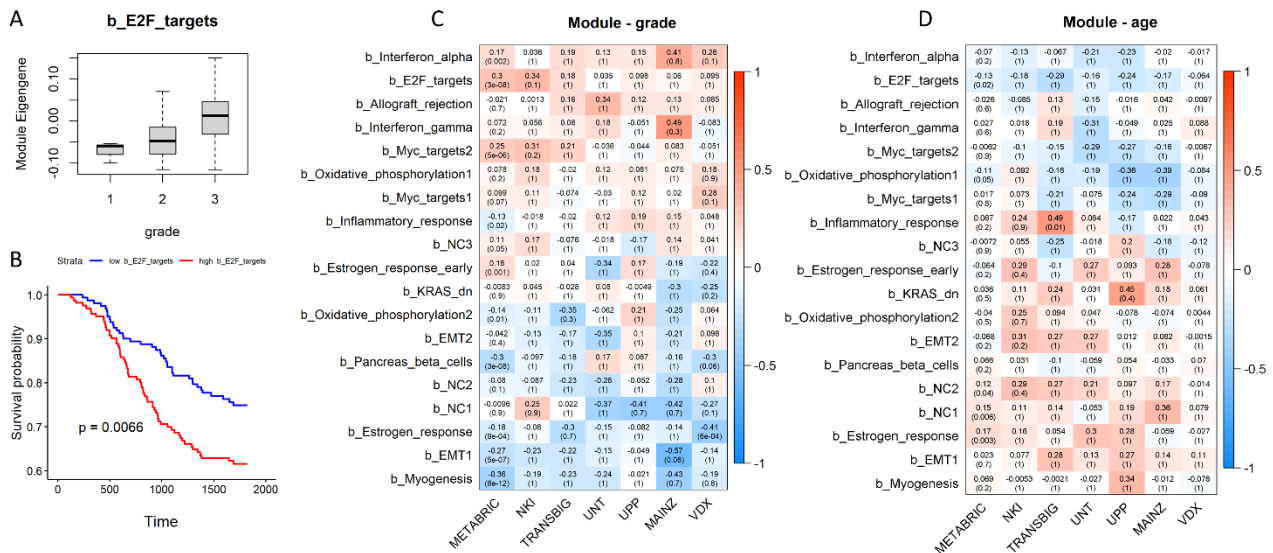


Figure 20. The b_E2F_targets module correlates with high grade (A) and poor prognosis (B) in basal-like tumors. C, D) Correlation of modules' expression, represented by module eigengene, and tumor grade (C) or age at diagnosis (D). Each row corresponds to a different module, columns correspond to datasets. Red indicates positive, and blue negative correlations. For each combination of module-grade, correlation and p-value are shown in the corresponding box.

2.1.6. Regulatory mechanisms identified from network topology

Having thus identified gene networks of particular interest, both across all samples or basal-like specific, an important question is how to exploit network structure to select genes relevant for networks' maintenance, hence identifying potential therapeutic targets and informing on breast cancer molecular mechanisms.

The most obvious candidates in maintaining network's coherence are hubs, already proved to be enriched in cancer driver genes for protein-protein interaction networks [26]. In co-expression networks, based on transcriptional data, this network-structure-based selection can be further refined with the intuition that transcription factors (TFs) are the most likely candidates regulating the co-expression. Hence, the network can be overlaid with *a priori* information, filtering hubs for their belonging to the TF class.

For example, using the intramodular connectivity (kWithin, exemplified in Figure 17) as a measure of centrality, it results that E2F_targets module's hubs are cell cycle-related genes, such as CDCA5 (Cell Division Cycle Associated 5), TPX2 (Targeting Protein for Xklp2) and CENPA (Centromere Protein A), while the most central TF is FOXM1 (Forkhead box M1), scoring as the 14th gene of the module based on the kWithin. Indeed, this TF is well known to be involved in cancer progression and therapy resistance, in particular in BC (reviewed in [119]). The most central genes in the E2F_targets module are overexpressed in basal-like tumors and associated with poor prognosis (Suppl. Fig. 2, 3), confirming that they are good representative of module's features. The first hub of the KRAS_dn1 module, another module associated with aggressiveness (Figure 6) is the TF FOXC1 (Forkhead box C1), a gene well-known to be involved in BC development and metastasis, overexpressed in basal-like breast cancer, and associated with poor prognosis (reviewed in [120]). The second and third hubs are ROPN1 (Rhopilin Associated Tail Protein 1), a cancer-testis antigen involved in sperm motility, and its paralog ROPN1B. Little is known about ROPN1 in breast cancer, but it has been associated with a SOX10 (SRY-box Transcription Factor 1) signature shared with salivary adenoid cystic cancer [121]. Interestingly, SOX10 scores as the 8th hub in the same KRAS_dn1 module and has been reported to be a marker of triple negative breast cancer [122] and involved both in stem cell activity and EMT in mammary epithelial cells [123]. Next is VGLL1 (Vestigial Like Family Member 1), the 6th gene in the kWithin ranking in the KRAS_dn1 module. It is a transcriptional co-activator that binds to several members of the TEAD (TEA Domain Transcription Factor) family of TFs to modulate the Hippo signaling pathway and it is involved in cell cycle regulation. Its association with basal-like breast cancer has only been reported once [124]. As observed for the E2F_targets module,

also in KRAS_dn1 module the most central genes are overexpressed in the basal-like subtype and correlated with poor survival (Suppl. Fig. 2, 3). Notably, the KRAS_dn1 module is split into two subnetworks connected by a single node, represented by the PSAT1 gene (Phosphoserine aminotransferase 1, involved in serine biosynthesis) (Figure 21A). Interestingly, one of the two subnetworks is highly enriched in poor prognosis related genes, and PSAT1 expression itself is significantly enriched in basal-like BC (Figure 21B) and correlates with poor prognosis (p-value: 4.5×10^{-10} , Figure 21C). This gene is therefore a good candidate to shape the KRAS_dn1 module and to coordinate the expression of the module's genes involved in aggressiveness.

centrally connected genes are MLPH (Melanophilin) and P4HTM (Prolyl 4-Hydroxylase, Transmembrane), whose possible role in BC is poorly characterized.

2.1.6.1. *Proof of concept: centrality identifies functionally relevant nodes*

Hubs are the most natural candidates as regulators of modules' activity and hence as the most functionally relevant nodes in each network, and, as commented above, previous data support this idea. Nevertheless, testing this concept is of important to build a gene-selection strategy based on solid and not merely speculative knowledge.

Hence, a few tests were performed to assess the significance of centrality in selecting functionally relevant genes in breast cancer. The centrality measure used in this context is the kWithin, commonly employed in co-expression networks based on WGCNA, which quantifies the overall connectivity of a gene with all the other genes in the same module. Indeed, it is defined as the sum of the adjacencies, which indicate connection strengths (Figure 17).

To work on the most reliably mechanistically driven modules (and not subtype-driven ones) I chose to apply the tests on the basal-like modules whenever appropriate, obtaining three independent *in silico* lines of evidence supporting the choice of hubs as the most likely relevant for network's functioning:

- Centrality correlates with the significance of relationship with patients' prognosis
- Centrality correlates with relevance for fitness in BLBC (Basal like Breast Cancer) cell lines
- TF hubs specifically regulate their own module's expression

Nevertheless, for the first test the global modules have been preferred, to increase the amount of survival data that could be used.

As first evidence for the ability of kWithin to prioritize functionally relevant genes, I have considered how well each gene correlates with survival in patients' data. Each gene has been independently used to stratify patients in two groups of equal size based on its expression, and differences in survival probability have been assessed, giving as output a sign (higher or lower survival probability corresponding to higher gene expression) and a p-value. Using this information, many modules display a significant correlation between gene centrality and its association with either good or poor prognosis. For example, in the E2F_targets module the most central genes are most confidently correlated with poor prognosis (Figure 22A), while in the Estrogen_response module the same is true for good prognosis (Figure 22B). Overall, 50% of the modules display a significant correlation between these two measures (Figure 22C). It is interesting to note that the

only two modules showing a negative correlation between the centrality and the relationship with survival are likely microenvironmentally-related modules (Allograft_rejection, TNFA_signaling_via_NFKB, see Figure 14).

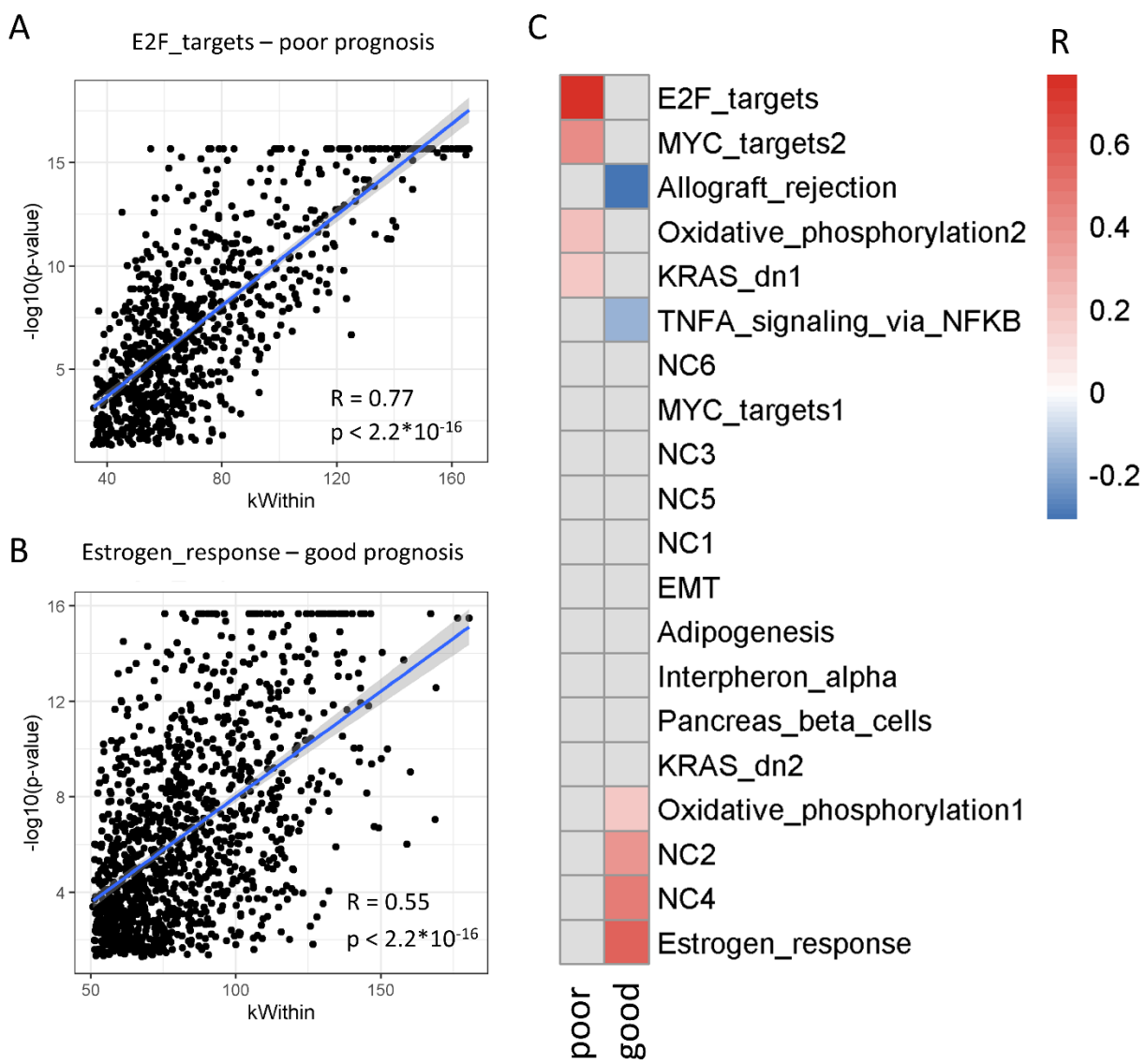


Figure 22. Nodes' centrality within some modules correlates with the confidence of the relationship with the survival. A, B) Correlation between the intramodular connectivity (kWithin) the p-value for the relationship of nodes' expression with survival for the E2F_targets module (A) and the Estrogen_response module (B). C) Summary of the correlation between the intramodular connectivity (kWithin) and the p-value for the relationship of nodes' expression with survival for each module. Only significant correlations are shown.

The second test made use of genome-wide screenings of genes' relevance for fitness, where each gene was knocked-down in a panel of breast cancer cell lines, either through the CRISPR-Cas9 technology (Clustered Regularly Interspaced Short Palindromic Repeats) or siRNAs, and then the proliferative ability of each cell line was measured. In these screenings, cells are transfected with a pool of vectors, so that each cell is silenced for a random gene in the pool. Then, cells' proliferation is inferred by measuring the presence of vectors in cells' DNA after letting them to grow. In this

context, cells transfected with a vector inducing the silencing of an essential gene will die or grow very slowly so that the specific vector will be depleted in the final pool.

Here, I collected 3 independent datasets of this kind, generated in independent projects [128–130], and calculated the correlation between a gene’s centrality within its module (kWithin) and its logFC (the logarithm of the fold change for the corresponding silencing vector). In Figure 23, the results of these tests are summarized in a separate heatmap for each dataset: in all of them, the most central genes in the b_E2F_targets and b_Myc_targets2 modules show higher relevance for fitness (lower logFC) than peripheral genes.

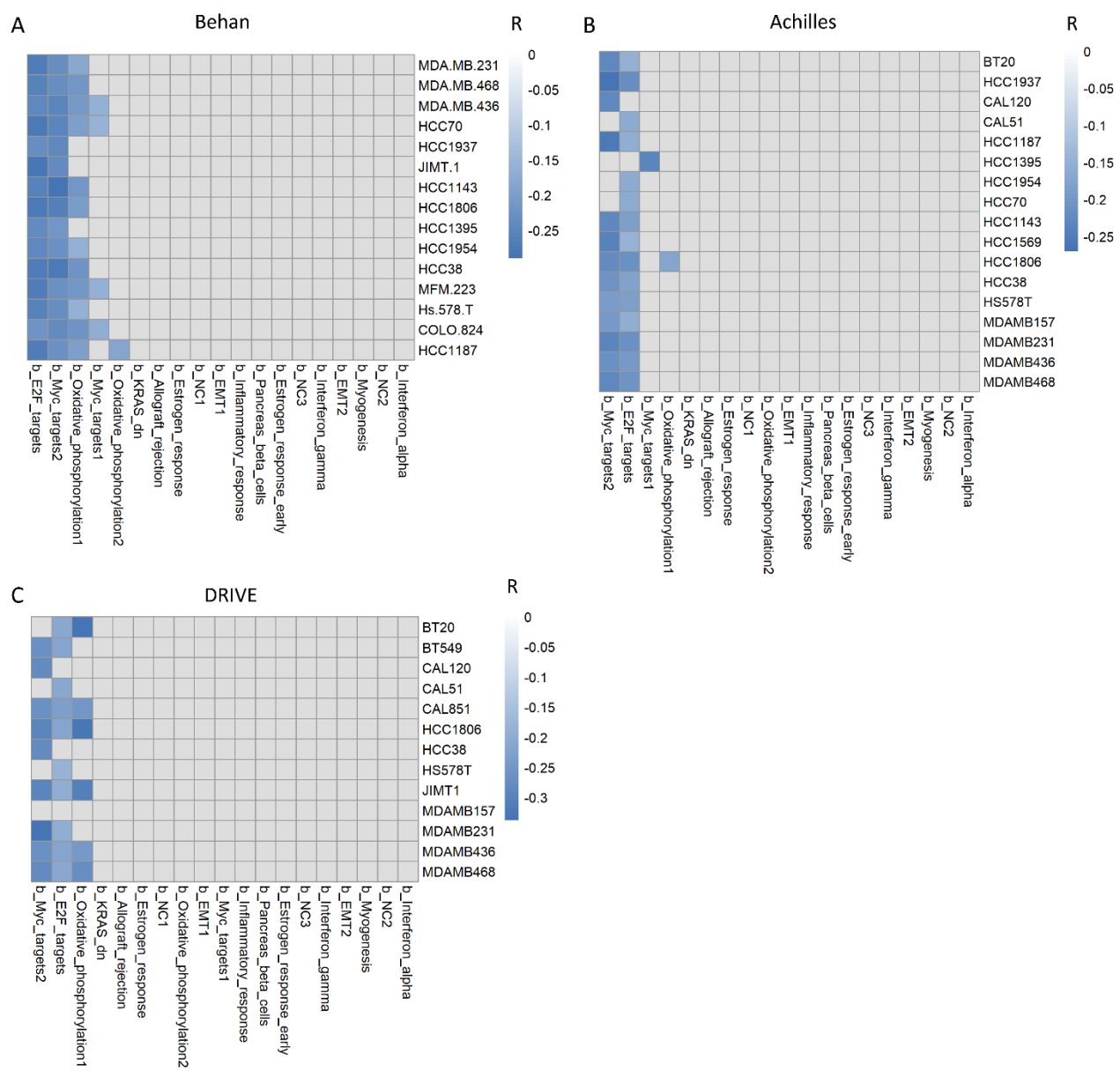


Figure 23. Correlation between gene centrality (kWithin) in each module (columns) and the logFC of the corresponding guide/siRNA in cultured cells of the cell line indicated in the rows. A) CRISPR screening dataset from [128], B) RNAi screening from [129], C) RNAi screening from [131]. Non-significant tests are in grey, blue indicates negative correlation.

Finally, I tested hubs' relevance in network's stability focusing on the b_E2F_targets module, which was identified as the most promising module in explaining basal-like tumors' aggressiveness (Paragraph 2.1.5). Since the undirected networks here generated do not permit to reconstruct upstream and downstream genes in regulatory connections, genes with high kWithin could both be downstream targets of many module genes or upstream regulators. In order to enhance the likelihood of identifying upstream regulators, I specifically searched for TFs. Strikingly, of the top 25 TFs within the b_E2F_targets module, 75% are correlated with BC poor prognosis and 92% are overexpressed in basal-like tumors (Table 3).

Table 3. Top centrally located transcription factors in the b_E2F_targets module. Genes in bold have been chosen for experimental validations. (OE=Overexpressed, BL=Basal-like)

Gene	survival (p-value)	OE in BL (p-value)	Expression datasets
FOXM1	$<2.2*10^{-16}$	$<2.2*10^{-16}$	GSE2222, GSE55204, GSE25741
PTTG1	$<2.2*10^{-16}$	$<2.2*10^{-16}$	GSE48928 (luminal only)
TEAD4	$1.4*10^{-8}$	$<2.2*10^{-16}$	
EZH2	$2.1*10^{-7}$	$<2.2*10^{-16}$	GSE103242, GSE48979, GSE36939
E2F3	$2.8*10^{-11}$	$<2.2*10^{-16}$	
FOXC1	0.098	$<2.2*10^{-16}$	GSE73234
PARP1	0.0085	$6.3*10^{-15}$	GSE34817
TFDP1	0.015	$<2.2*10^{-16}$	
CEBPG	$8.1*10^{-6}$	$<2.2*10^{-16}$	
CHCHD3	$1.3*10^{-6}$	$<2.2*10^{-16}$	
NOLC1	0.00058	$<2.2*10^{-16}$	
HMGA1	$2*10^{-15}$	$<2.2*10^{-16}$	GSE35525
UHRF1	$1.5*10^{-10}$	$<2.2*10^{-16}$	
TAF5	0.0031	$<2.2*10^{-16}$	
SOX9	0.15	$<2.2*10^{-16}$	
BOLA1	0.4	1	
TCF7L1	0.054	$<2.2*10^{-16}$	GSE38893
SSRP1	0.00017	$<2.2*10^{-16}$	GSE92281
ZNF232	0.2	$5.4*10^{-14}$	
ZNF165	0.28	0.005	
ELF5	0.005	$<2.2*10^{-16}$	GSE30405
RCOR2	$1.4*10^{-7}$	$<2.2*10^{-16}$	
NR2C2	$2.5*10^{-6}$	$<2.2*10^{-16}$	
ZBTB5	0.038	$<2.2*10^{-16}$	
DNMT1	0.03	$<2.2*10^{-16}$	

I then hypothesized that TF hubs of the b_E2F_targets module could be central players in up-regulating the expression of the genes belonging to the network. To test this hypothesis, I searched for published experiments reporting gene expression datasets obtained in BC cell lines

upon modulating selected b_E2F_targets TF hubs, in order to assess hubs correlation with global regulation of module's genes. I found 19 datasets collectively studying 9 different TFs. In 16/19 cases, indeed, genes which expression was positively correlated with that of the selected TFs showed a significant overlap ($p\text{-value} < 0.001$) with genes in the b_E2F_targets module (Figure 24), validating the initial hypothesis. The only exception was EZH2 (Enhancer of Zeste Homolog 2), for which contrasting results were obtained in different datasets. Nevertheless, the hypothesis was validated for FOXM1, PTTG1 (Pituitary Tumor-Transforming Gene 1 Protein), FOXC1, PARP1 (Poly(ADP-Ribose) Polymerase 1), HMGA1 (High Mobility Group Protein A1), TCF7L1 (Transcription Factor 7 Like 1), SSRP1 (Structure Specific Recognition Protein 1) and ELF5 (E74-Like Factor 5), analyzed across several BC cell lines and giving consistent results upon down- and up-regulation (Figure 24). For example, ELF5 knock-down in the basal-like BC cell line HCC1937 leads to the specific down-regulation of b_E2F_targets module genes, while its overexpression in luminal BC cell lines (MCF7, T47D) induces the up-regulation of the same b_E2F_targets module genes.

Having thus proved that modules' hubs are functionally relevant and that TF hubs regulate their own module, this concept could be exploited in some more specific contexts, as will be described in the following paragraphs.

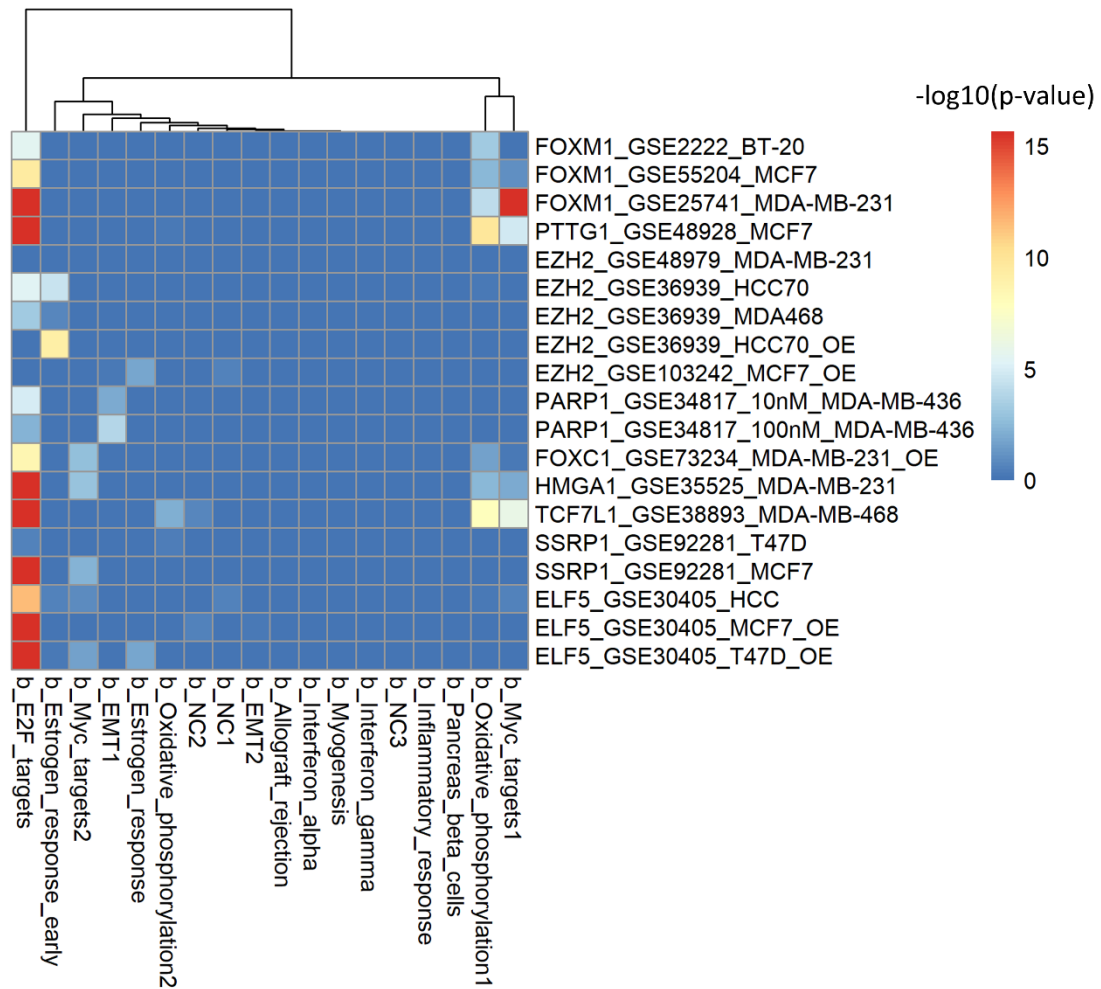


Figure 24. TF hubs in the b_E2F_targets module regulate genes in the same b_E2F_targets module. The heatmap shows the significance of the enrichment (Fisher test) of differentially expressed genes upon hubs modulation (in the same direction of hubs expression) in modules genes. Values are $-\log_{10}$ of p-values, scaled by row. Each row corresponds to a different dataset where the indicated TF has been down- or up-regulated in a BC cell line (OE= overexpression). Columns corresponds to different basal modules.

2.1.6.2. ZNF532 is a potential common target of STAT3 and WNT/PCP pathways

As a first application, I investigated whether the BC networks could be used to define potential targets of specific pathways. The reasoning behind this idea was that if a gene is transcriptionally activated by a pathway, then it will be co-expressed with other genes of the pathway. Therefore, pathway genes and potential targets will be part of the same network. Here, I searched for potential common targets of the STAT3 and WNT/PCP (Planar Cell Polarity) pathways, which display several overlapping functions and can interact in a feed-forward loop [132–134].

The TF STAT3 (Signal Transducer and Activator of Transcription 3) is considered as an oncogene and correlates with poor prognosis. It is implicated in many functions fundamental for

cancer progression such as proliferation, survival, invasiveness, and metastasis [135,136]. Similarly, Wnt signaling has been implicated in malignant transformation [137]. Depending on the type of ligands involved, Wnt signaling has been classified in canonical and non-canonical, with the non-canonical PCP pathway being involved in defining cell shape and movement and inducing cancer progression [138–140], particularly in basal-like BC.

The non-canonical Wnt5a (Wingless 5a) and Wnt5b ligands are overexpressed in basal-like breast cancer cell lines [138]. Moreover, we recently showed that the two pathways converge on the activation of the common downstream target RhoU, promoting cell motility in basal-like human breast tumor cells [141]. Hence, here I aimed at identifying potential additional common targets of the two pathways that could mediate their functions in determining cancer aggressiveness.

As a first step, I collected a set of gene signatures of the two pathways [142–151] and I determined which, if any, of the BC modules was enriched in at least one signature of both pathways. Not surprisingly, this analysis pointed to the EMT network, enriched for STAT3 and non-canonical WNT signatures with p-values of 0.001 and 0.02, respectively [142,151]. Via the analysis of a BC ChIP-seq dataset [152], I determined that the EMT module is enriched for genes displaying STAT3 binding at their promoter in the basal-like BC cell lines HCC1143 and MDA-MB-157 (p-values: <0.004 and $<3 \times 10^{-12}$, respectively). Moreover, I analyzed an RNA-seq dataset of the basal-like MDA-MB-231 cell line upon STAT3 knock-down, and another one of the luminal cell line MCF7 upon overexpression of the WNT/PCP ROR2 co-receptor (Receptor Tyrosine Kinase Like Orphan Receptor 2) [147,152]. I selected the genes in the intersection between those down-regulated upon STAT3 knock-down and up-regulated following ROR2 overexpression, hence positively correlated with both STAT3 and WNT/PCP, and then calculated the overlap of gene networks for these STAT3/ROR2 regulated genes, finding a significant enrichment for the EMT module (25 genes, p-value= 2.15×10^{-5}), followed by the KRAS_dn1 module (p-value=0.001). Therefore, it can be concluded that the EMT network includes a significant number of genes regulated by both pathways, confirming previous analyses.

Three TFs (ZNF532, RUNX2 - Runt-related transcription factor 1- and MSX1) are included in these 25 genes. Interestingly, although the role of ZNF532 (Zinc Finger 532) in cancer is not known, this factor is overexpressed in basal-like tumors (Figure 25A), significantly correlating with early tumor onset (Pearson's correlation=-0.17, p-value= 2×10^{-14}) and with poor prognosis (p-value: 0.004) (Figure 25B). This factor represents therefore an interesting candidate as a downstream mediator of the

STAT3 and WNT/PCP pathways, potentially involved in regulating all or a subset of the other 24 common target genes identified above.

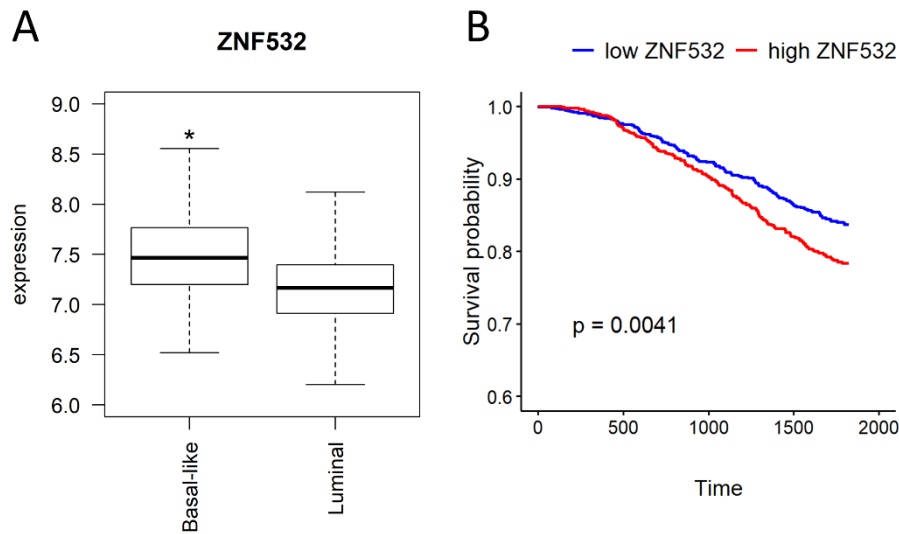


Figure 25. A) ZNF532 expression in basal-like vs luminal (A and B) tumors (* p-value<2.2*10⁻¹⁶). B) Survival curves of patients divided based on ZNF532 expression showing that higher ZNF532 expression significantly correlates with poor prognosis.

2.1.6.3. Identify regulators of CAF-related signature

In the tumor microenvironment, Cancer Associated Fibroblasts (CAFs) become activated by cancer cells and increase their secretory activity to produce soluble factors that contribute to tumor cells proliferation, invasion, and dissemination to distant organs. The pro-tumorigenic transcription factor STAT3 and its canonical inducer, the pro-inflammatory cytokine IL-6, act conjunctly in a positive feedback loop that maintains high levels of IL-6 secretion and STAT3 activation in both tumor and stromal cells. Recently, we have demonstrated that STAT3 is essential for the pro-tumorigenic functions of murine breast cancer CAFs both *in vitro* and *in vivo*, and identified a STAT3 signature significantly enriched for genes encoding for secreted proteins (*Manuscript under revision in Oncogene*).

To identify potential regulators of the STAT3 CAF signature, up-stream, or down-stream of STAT3, I employed the previously defined co-expression networks, looking for modules enriched in the signature itself and then selecting corresponding TF. Two modules resulted significantly enriched in STAT3 signature genes: the EMT and the TNFA_signaling_via_NFKB. Not surprisingly, and in line with the cellular system in which the signature was derived, these were amongst the modules most clearly correlated with stromal or immune percentage in primary breast tumors (Figure 14). I then filtered the TFs belonging to these modules according to their expression in CAFs (expression higher than the median of the genes) and ranked them based on their kWithin.

To validate the potential link between STAT3 activity and the selected TFs, I used an external single cell RNA-seq dataset of breast cancer associated fibroblasts [153]. With this, I computed the correlation between SOCS3 (Suppressor of Cytokine Signaling 3) expression, used as a proxy for the activity of the post-translationally regulated STAT3, and the selected TFs expression, and compared it with random genes. The shift of correlations toward positive values (Figure 26A) is the first evidence supporting a link between these TFs and STAT3 in cancer associated fibroblasts. Then, I tested the correlation between selected TFs and the estimated stromal percentage in the TCGA cohort, resulting in significant positive correlations for all genes except for THRA (Thyroid hormone receptor alpha) (Figure 26B, C).

This suggests that the selected TFs, and especially the top ranking PRRX1 (Paired Related Homeobox 1) and KLF12 (Krupper Like Factor 12), are interesting candidate regulators of the STAT3 CAF signature.

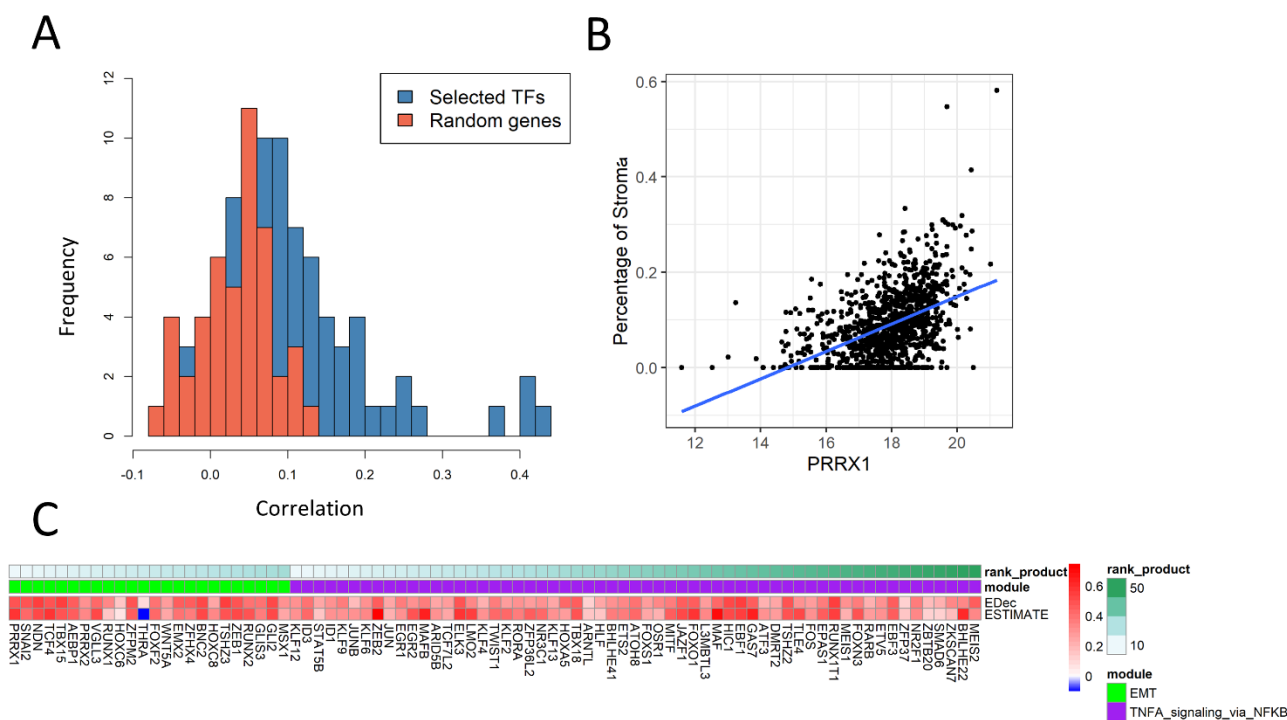


Figure 26. A) Correlation between TFs or random genes and SOCS3 expression across BC CAFs. B) As an example, PRRX1 expression in primary breast tumors from the TCGA cohort is shown, correlating with the estimated percentage of stroma (ESTIMATE algorithm). C) A positive correlation can be observed for most selected TFs using two independent algorithms for stromal percentage estimate. Purple and green bars indicate the module to which TFs belong, while different shades of green represent the rank product of expression in CAFs and centrality in the module.

2.1.7. Additional regulatory layers: microRNA

Dissecting regulatory mechanisms of identified clinically relevant co-expression modules could be helpful in gaining insights about the molecular mechanisms behind their functioning, but also in identifying additional actionable targets.

Therefore, I looked for microRNAs regulating modules' expression. I limited the analysis to assuming the microRNAs' inducing the degradation of the target gene's mRNA, and hence expecting a negative correlation between miRNA-targets' expressions.

For this, I used matched miRNA-mRNA expression datasets, where miRNAs and mRNAs were profiled in the same samples, identifying five independent suitable datasets, comprising the METABRIC dataset itself (METABRIC, GSE22220, GSE28884, GSE59595, GSE81002, [154–158]). For each of those datasets I performed an unbiased search of negatively correlated microRNAs and module eigengenes. The miRNA-module pairs being consistently negatively (and significantly) correlated in all six datasets were additionally filtered using lists of predicted microRNAs' targets, deposited and summarized in the multiMiR database [159]. On top of this selection, I required a significant overlap between module's genes and predicted targets.

Based on these very stringent criteria, I identified four microRNAs, putative regulators of the EMT and the NC4 modules: miR-29c-3p regulating the EMT module, and miR-20a-5p, miR-93-5p, and miR-106b-5p regulating the NC4 module. In Figure 27, the correlation between their expression and module eigengenes, strikingly consistent for all of the five datasets, is shown.

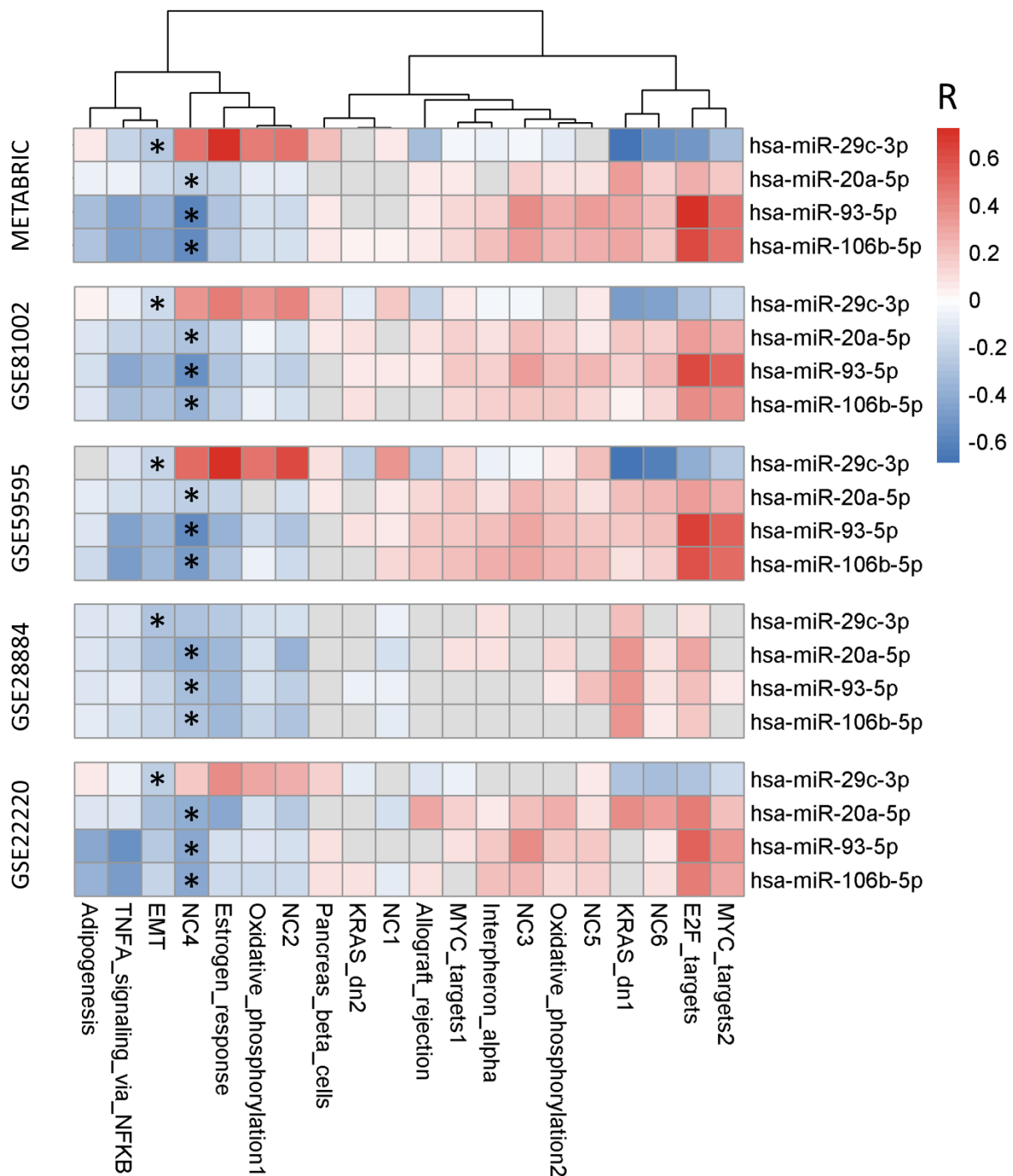


Figure 27. Correlation between modules' and microRNAs' expression across the five primary BC datasets. Modules are indicated in the columns, while the selected microRNAs are in the rows. Red indicates positive, while blue negative correlation. Not significant correlations are in grey ($p > 0.05$). Asterisks marks the module-microRNA pairs selected in the analysis based on consistent negative correlation across the five datasets and on the enrichment of module's genes in predicted microRNA's targets.

2.1.8. Identifying means of disrupting aggressiveness modules

Having identified a module that is more tightly connected in the basal-like subtype than in any other subtype, and that is correlated with higher tumor grade and poor prognosis in basal-like

tumors (the b_E2F_targets), the next step was to find ways of disrupting its connectivity for therapeutic purposes. In this direction, I exploited network's topology to find novel potential therapeutic targets and made use of drug-target databases to identify drugs with many targets in the module, hence likely able to disrupt it.

2.1.8.1. Targeting hubs

In Paragraph 2.1.6.1 I have proven that TF hubs regulate their own module, making use of publicly available data on a few hubs. In the light of these results, I expect that additional TFs in a central position within the b_E2F_targets module could orchestrate the expression of the same basal-like specific gene module, therefore likely playing a role in defining the phenotypic features of this subtype.

In particular, I selected the most central TFs not fully studied in BLBC (no available transcriptome profiling of basal-like BC cell lines after their up- or down-regulation), all of which correlate with poor prognosis: PTTG1, TEAD4, E2F3 (E2F Transcription Factor 3), TFDP1 (Transcription Factor DP-1), CEBPG (CCAAT Enhancer Binding Protein Gamma) (Figure 28).

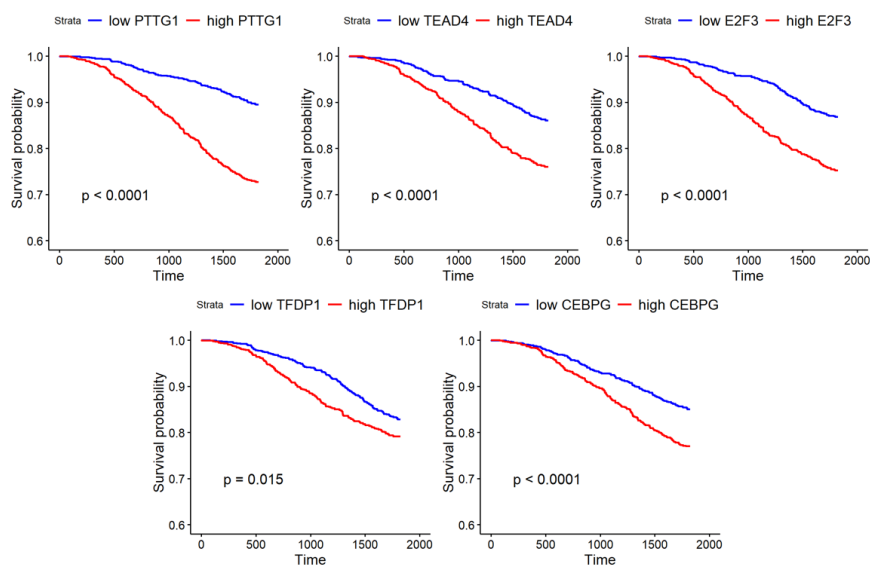


Figure 28. Kaplan-Meier survival curves showing that patients with higher expression of PTTG1, TEAD4, E2F3, TFDP1 or CEBPG have worse prognosis

To experimentally validate the relevance of these genes in maintaining basal-like tumors' phenotype, functional assays were performed in the laboratory of Prof. Valeria Poli on basal-like breast cancer cell lines silenced for each of them, separately. As model systems, the MDA-MB-231 and MDA-MB-468 were employed. Five different shRNAs per gene and cell line were transfected in the cells and screened for their ability to effectively silence the target gene, to select at least two

independent efficient shRNAs per gene. Proliferation and migration assays were used as outcome measures of functional relevance. Results are summarized in Table 4.

Table 4. For each gene (in rows), the results of two independent silencings (indicated with different acronyms, e.g. P1 and P2 indicate two different shRNAs employed to silence PTTG1 in two independent batches of cells) are indicated for each functional assay and cell line (columns). Green indicates that the silencing induces a significant impairment of proliferation or migration, while red indicates a significant increase in proliferation or migration. White indicates inconclusive results (not significant), while some experiments are ongoing.

Functional analysis				
	Proliferation		Migration	
	MDA-MB-231	MDA-MB-468	MDA-MB-231	MDA-MB-468
PTTG1	P1	P1	P1	P1
	P4	P4	P4	P4
TEAD4	T1	T1	Ongoing	Ongoing
	T4	T4		
E2F3	E1	E1		E1
	E5	E5	E5	E3
TFDP1	TF4	TF1	TF4	Ongoing
	TF5	TF4	TF5	
CEBP	C2	Ongoing	C2	Ongoing
	C5		C5	

Moreover, to assess the relationship between hubs' expression and module genes' expression, genome wide transcriptomic profiling of silenced cell lines was performed and analyzed similarly to previous *in silico* tests (Paragraph 2.1.6.1, Figure 24). Again, genes being down-regulated upon TF hub's silencing were tested for their overlap with modules' genes, expecting the strongest enrichment to be observed for the b_E2F_targets module, to which tested hubs belong. Surprisingly, the observed result is strikingly different than expected: the most significant enrichment is observed for the b_EMT1 module for 4/5 hubs, and only the silencing of E2F3 induces a specific down-regulation of the b_E2F_targets module (Figure 29A). Indeed, up-regulated genes upon TFDP1 silencing are enriched for b_E2F_targets (Figure 29B). Interestingly, the b_E2F_targets and the b_EMT1 modules are strongly negatively correlated across METABRIC samples, making this result even more counter-intuitive (Figure 29C-E).

Nevertheless, an alternative model could fit all the presented results: the perturbation of the network could more likely induce gene expression changes in genes closely connected to the perturbed hub, without any specific implication on the sign of relationship (Figure 30). Indeed, the

differentially expressed genes upon the TF hubs silencing are more strongly correlated (in absolute value) with the hub than random genes (Kolmogorov-Smirnov test $p < 0.05$).

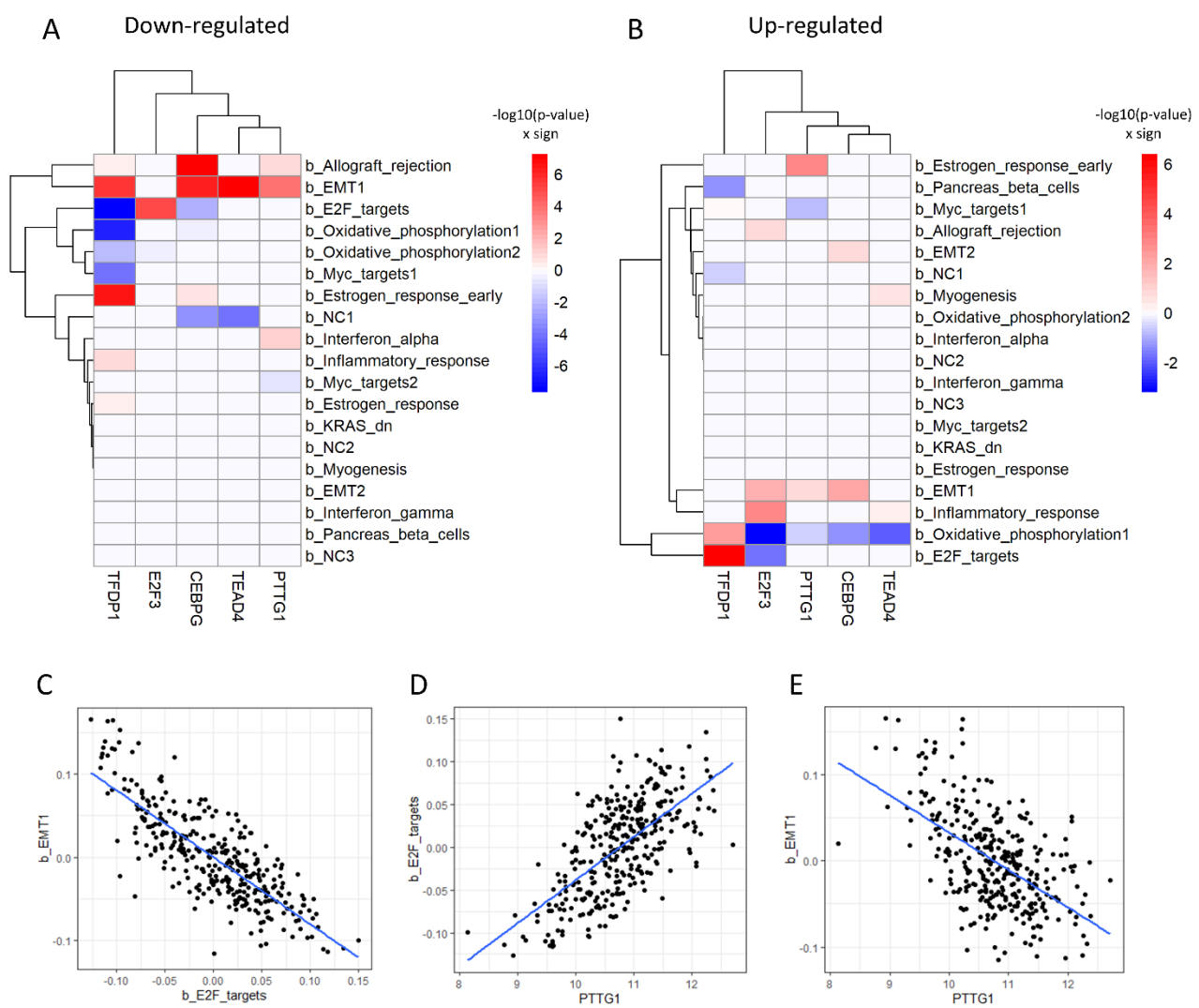


Figure 29. Fisher test genes down-regulated (A) or up-regulated (B) upon silencing for modules’ genes, indicating enrichment (red) or depletion (blue). Modules are shown in rows, silenced hubs in columns. C) *b_EMT1* and *b_E2F_targets* are negatively correlated, and also single *b_E2F_target*’s hubs, such as PTTG1, show the same trend, being positively correlated with *b_E2_target*’s module eigengene (D) and negatively correlated with *b_EMT1* (E).

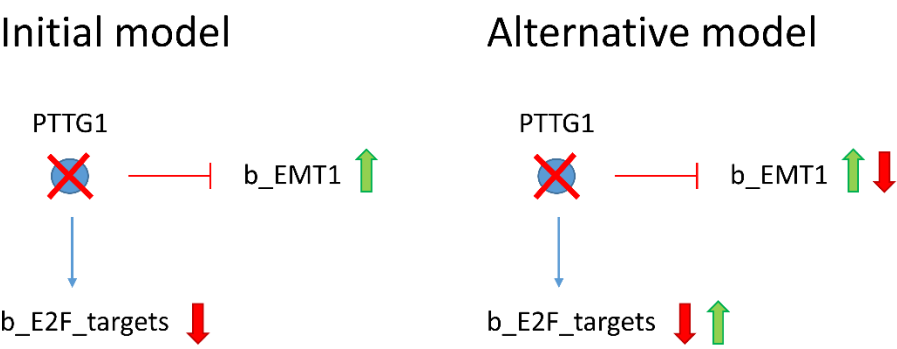


Figure 30. Two models for the effects of hubs’ silencing. In the initial model, the silencing of the TF hub PTTG1, part of the *b_E2F_targets* module, induces a down-regulation of the module it is part of, while it induces the up-regulation of the *b_EMT1* module, negatively correlated with *b_E2F_targets*. In the alternative model, PTTG1 silencing could

either induce up- or down-regulation of both the b_E2F_targets and the b_EMT1 modules, with which it is tightly connected (either positively or negatively).

2.1.8.2. Drugs for repurposing

To identify drugs potentially able to disrupt the b_E2F_targets module, I parsed the ChEMBL database [160], comprising information about small molecules and their targets obtained from more than 80 000 papers, binding measurements, and functional assays for a total of 1.4 million of assays. I programmatically retrieved all drug-target relationships for genes in the module, filtered for drugs with targets specifically enriched in genes of the b_E2F_targets module (Fisher-test $p < 0.05$) and being at least in phase II of clinical trials. Thus, I obtained a list of 35 drugs, promising candidates for the disruption of the b_E2F_targets module and hence as potentially effective treatments for breast cancer (Table 5). Among these, the compound Ellagic acid has the highest number of distinct targets in the module (18), two of which are amongst the first 50 hubs and 5 amongst the first 100 hubs. Ellagic acid is an organic heterotetracyclic compound found in many fruits and vegetables, including raspberries, strawberries, cranberries, and pomegranates. It has a role as an antioxidant, but no clinical trials are investigating its effectiveness in treating or preventing breast cancer, possibly in combination with routinely used drugs.

Table 5. List of drugs targeting the b_E2F_targets module. Max phase indicates the highest phase of clinical trials in which the drug was employed; distinct targets is the number of targets in the module; kSum is the sum of the kWithin of all the targets in the module; the last three columns indicate the number of targets amongst the top 100, 50 or 10 hubs of the module.

ChEMBL ID	Name	Max phase	Distinct targets	p-value	kSum	first100 hubs	first50 hubs	first10 hubs
CHEMBL6246	ELLAGIC ACID	2	18	0.002572	1608.755	5	2	0
CHEMBL483158	ALISERTIB	3	16	0.039702	1785.441	9	5	1
CHEMBL363648	TAK-715	2	15	0.046292	1697.184	9	5	1
CHEMBL362558	LY-2090314	2	11	0.033283	1194.229	6	3	0
CHEMBL507361	PD-0325901	2	10	0.022321	1189.713	7	4	1
CHEMBL1173055	RUCAPARIB	4	9	0.009065	1043.806	5	3	0
CHEMBL564829	MILCICLIB	2	9	0.001634	968.6889	4	2	0
CHEMBL451887	CARFILZOMIB	4	9	0.004171	674.0449	0	0	0
CHEMBL354349	ACRIFLAVINE	4	7	0.032316	647.6279	2	1	0
CHEMBL2103840	DINACICLIB	3	6	0.00052	600.5183	1	0	0
CHEMBL2146123	METHACYCLINE HYDROCHLORIDE	4	6	0.014039	493.6053	1	1	0
CHEMBL33864	Lipoic acid, alpha	3	6	0.014039	414.6265	0	0	0
CHEMBL402548	DANUSERTIB	2	6	0.035696	688.4098	4	2	0
CHEMBL41286	DIACEREIN	4	6	0.035696	517.286	1	0	0
CHEMBL1200467	HYDROXYZINE PAMOATE	4	5	0.021223	435.8745	1	0	0

CHEMBL1230165	SILMITASERTIB	2	5	0.024892	545.6828	1	1	0
CHEMBL3301610	ABEMACICLIB	4	5	0.002074	475.8526	0	0	0
CHEMBL575448	BMS-754807	2	5	0.006442	547.7998	2	2	0
CHEMBL550495	METHYLENE BLUE	4	4	0.03823	349.2214	0	0	0
CHEMBL109480	TANESPIMYCIN	3	4	0.03823	344.888	0	0	0
CHEMBL1185568	DITHIAZANINE	4	4	0.021947	349.2214	0	0	0
CHEMBL1200758	AMPICILLIN SODIUM	4	4	0.03221	312.4555	0	0	0
CHEMBL191083	METHYLENE BLUE	4	4	0.03823	349.2214	0	0	0
CHEMBL231884	DIOSMIN	4	4	0.03823	312.2744	0	0	0
CHEMBL3544942	RONICICLIB	2	4	0.000248	408.3765	0	0	0
CHEMBL232201	BENZIODARONE	4	3	0.023743	266.4964	1	0	0
CHEMBL25202	TAMIBAROTENE	2	3	0.048864	277.645	1	1	0
CHEMBL3545110	RIBOCICLIB	4	3	0.001067	274.081	0	0	0
CHEMBL1231160	PEVONEDISTAT	3	2	0.012285	138.4352	0	0	0
CHEMBL2103879	GANETESPIB	3	2	0.023503	159.8977	0	0	0
CHEMBL425	OLSALAZINE	4	2	0.012285	146.4918	0	0	0
CHEMBL464	ETRETINATE	4	2	0.037481	152.3602	0	0	0
CHEMBL706	CETYL ALCOHOL	4	2	0.037481	146.4918	0	0	0
CHEMBL216913	MITOGUAZONE	3	2	0.023503	160.5776	0	0	0
CHEMBL997	IBANDRONIC ACID	4	2	0.023503	184.8701	0	0	0

2.2. Meta-analysis of microdissected breast tumors reveals genes regulated in the stroma but hidden in bulk

As discussed in the introduction, gene expression profiling of tumors is an essential approach for the selection of biomarkers and the investigation of the molecular mechanisms of cancer, but transcriptomic results are often difficult to reproduce due to technical biases, sample heterogeneity, or small sample sizes. Combining many datasets can help to reduce artefacts and improve statistical power. Therefore, I aimed at creating a comprehensive resource of transcriptomic datasets investigating breast cancers, focusing on microdissected tumors, which enable distinguishing the contribution of the tumor microenvironment from that of cancer cells. I defined robust lists of differentially expressed genes and described their relationships with clinical features in each cellular compartment, identifying clinically relevant markers that can only be retrieved by measuring their expression in the sole tumor microenvironment [116].

2.2.1. Database construction

I collected 48 transcriptomic datasets of breast tumors or breast hyperplasias deposited in the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>), selecting experiments where different cellular compartments were separated prior to RNA extraction. Most of the datasets (43) derive from laser capture microdissected primary tumors, while a minority measure gene expression of cancer associated fibroblasts (CAFs) grown in vitro, derived from either primary tumors or from mouse models. Overall, I collected 2144 samples, 2048 of which derive from primary tumors. The complete list of datasets is available in Table 6.

Table 6. Datasets in the LCM database

Dataset	Reference
GSE10780	[161]
GSE10797	[162]
GSE111662	[163]
GSE118432	[164]
GSE12622	/
GSE13293	[165]
GSE1378	[166]
GSE141828	[167]
GSE14548	[168]
GSE16873	[169]
GSE20437	[170]
GSE21947	[171]
GSE22513	[172]
GSE26910	[173]
GSE28796	[174]
GSE28821	[174]
GSE29270	[175]
GSE31192	[176]
GSE33692	[177]
GSE35019	[178]
GSE37614	[179]
GSE38959	[180]
GSE41228	[181]
GSE45581	[182]
GSE4823	[183]
GSE5764	[184]
GSE57968	[185]
GSE5847	[186]
GSE59772	[187]
GSE68744	[188]

GSE72644	[189]
GSE7377	[190]
GSE81838	[191]
GSE83591	[192]
GSE88715	[193]
GSE8977	[194]
GSE90505	[195]
GSE90521	[196]
GSE9574	[197]
GSE7413	/
GSE31138	[198]
GSE15363	[199]
GSE43379	[200]
GSE120129	[201]
GSE129189	[202]
GSE20086	[203]
GSE32518	[204]
GSE45256	[205]

To facilitate the comparability of different experiments, I mapped the probes used for each specific experiment to gene symbols, and I did an extensive and careful harmonization of biological and clinical annotations, as detailed in the Methods (Paragraph 4.3.3). Specifically, I gathered information about cellular compartment, disease status, receptor status (estrogen receptor, progesteron receptor, and HER2 amplification), PAM50 subtype, tumor histology, size, grade, TNM stage as well as the overall pathological stage, node positivity, recurrence, response to treatment, and patient's age at diagnosis, ethnicity, and menopause status (Tables 7 and 8 and Table S1 - on zenodo). Moreover, wherever possible and appropriate, I predicted the PAM50 breast cancer (BC) subtype from gene expression, obtaining a good concordance between subtype and receptor expression (Figure 31) and inferred clinical variables from other available clinical annotations (Paragraph 4.3.3 and Appendix A). Importantly, for 11 of the 48 datasets, I found clinical annotations that were made available by the authors of the original publication but that did not accompany the corresponding dataset in GEO. Their inclusion in the database substantially improved clinical annotations, which initially were relatively scarce. For example, the number of samples annotated for age increased from 530 to 756 and for size from 122 to 378. The complete database comprising gene expression data and metadata is available in Tables S3 and S4 on zenodo (<https://zenodo.org/record/5013252#.YU3j3bgzZPY>), while Table S1 comprises the complete summary of available clinical features.

Table 7. Summary of sample features: number of samples derived from primary tumors or in vitro systems and biological/clinical annotations of primary tumor samples. Only the main levels of each factor are shown. For a complete list, refer to Table S2.

Category	# of samples	# of datasets
System		
Primary tumor	2048	43
In vitro	96	5
Compartment		
Epithelium	1230	32
Stroma	664	21
Vessels	64	4
Adipose	16	1
Disease status		
Invasive BC	990	31
Tumor (other)	296	11
Normal counterpart	370	17
Normal	326	16
ER status		
Positive	502	20
Negative	419	23
PR status		
Positive	306	16
Negative	435	21
Her2 status		
Positive	302	18
Negative	661	24
LN positivity		
Positive	309	15
Negative	228	17
PAM50 subtype		
Basal-like	120	16
Her2+	71	13
LumA	194	16
LumB	140	16
Normal-like	35	10

Table 8. Summary of clinical features for primary tumor samples

Age	
Median (range)	55 (27–94)
# annotated samples	756
# of datasets	18
Size (in mm)	
Median (range)	24 (4–161)
# annotated samples	378
# of datasets	9
Grade	# of samples
I	50

II	189
III	312
# of datasets	12

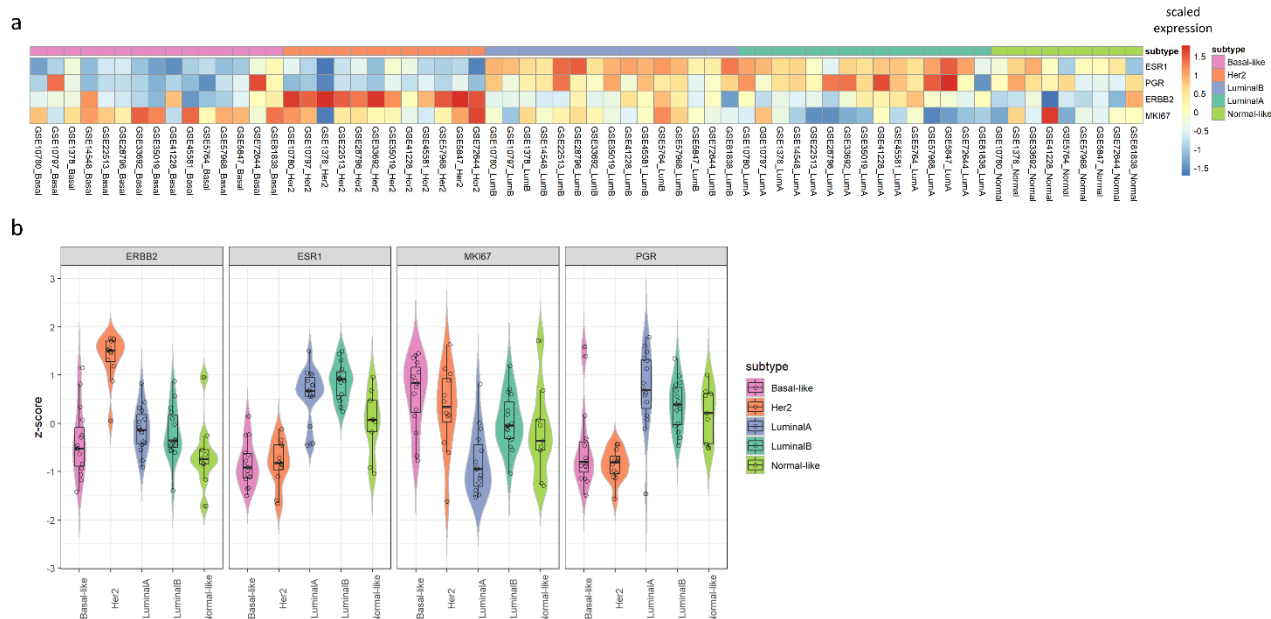


Figure 31. Markers' expression for breast cancer PAM50 subtypes annotated in the original dataset or inferred from gene expression. A) Heatmap showing the scaled expression of ESR1, PGR, ERBB2 and MKI67 averaged across each subtype's samples for each dataset separately. B) Violin plot emphasizing the differences in markers' expression between different subtypes.

2.2.2. Primary and in vitro datasets are only slightly correlated

First, I quantified gene expression changes between normal and tumor stroma in primary invasive ductal carcinomas (IDC), ductal carcinoma in situ (DCIS), or CAFs grown in vitro. As a reference condition, I used samples from cancer patients labelled as histologically normal ("normal counterpart" or simply "counterpart"), or, where available, normal breast tissue from reduction mammoplasty ("normal"). To avoid batch effects, I analyzed each dataset separately and then compared gene expression fold changes to assess similarities and differences between datasets, performing a total of 18 comparisons. The correlations between pairs of comparisons are globally, albeit slightly, positive (Figure 32A, mean correlation of 0.13), indicating intrinsic differences between the datasets. Moreover, different classes of datasets display different degrees of similarity (Figure 32B). Indeed, for example, primary tumors and in vitro samples appeared to be only slightly positively correlated (Figure 32B, C, $p = 0.05$). Similarly, IDC and DCIS samples behave differently (Figure 32D), motivating to keep them separated for meta-analytic purposes in order to limit the biological heterogeneity.

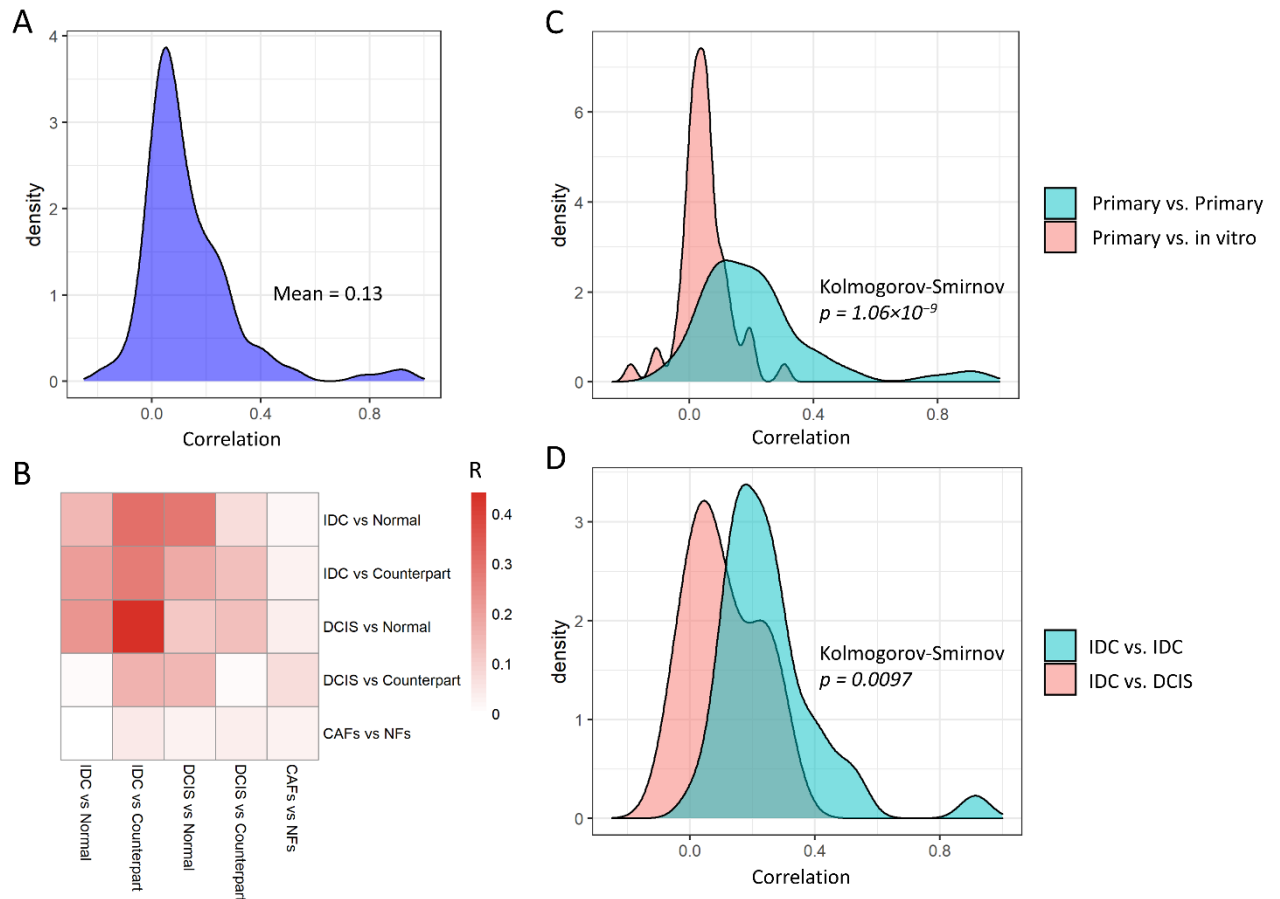


Figure 32. Similarity between stroma datasets of detected genes changing between normal and cancer samples. (A) Distribution of Pearson's correlation between log₂ fold changes (logFC) of genes between cancerous/non-cancerous conditions in each dataset, showing an average correlation of 0.13. All datasets comprising both cancerous and non-cancerous stromal samples were used, and pairwise correlations between datasets are shown. (B) Correlation of logFC between groups of datasets comparing invasive ductal carcinoma (IDC) and normal breast tissue from healthy donors (normal) or histologically normal tissue adjacent to tumor (counterpart), ductal carcinoma in situ (DCIS) vs. normal or counterpart, cancer associated fibroblasts (CAFs) and normal fibroblasts (NFs) grown in vitro. Red indicates a high average positive correlation. (C) Distribution of Pearson's correlations between logFC obtained from datasets sampling primary tumors or comparing logFC obtained from primary tumors or from in vitro experiments. (D) Distribution of Pearson's correlations between logFC obtained from datasets sampling IDCs or comparing logFC obtained from IDCs and from DCIS.

2.2.3. Non-redundant information Is obtained by separating different tissues

I then took advantage of the tissue specificity of the collected datasets to compare tumor epithelium and stroma gene expression behaviors. I thus calculated differentially expressed genes (DEGs) for each dataset comprising normal and tumor samples (or normal counterpart and tumor), analyzing epithelium and stroma separately. I then collapsed the differential gene expression statistics to obtain a global measure of the reliability of the gene expression changes across all of the available datasets. As microarray platform and pre-processing can impact the measure of fold changes in differential gene expression, I chose to apply the widely employed Fisher's method, summing the log-transformed p-values obtained from independent studies [98]. To limit the

heterogeneity of input data, I only employed invasive BC samples for this analysis. In total, I could perform nine comparisons for epithelium and eight for stroma. Strikingly, when comparing the average fold changes for each gene in tumor stroma or epithelium, it results that most genes behave similarly across compartments (Figure 33A), suggesting coordinated gene expression reorganization between the tumor and the surrounding cells. Alternatively, it is possible that, despite the use of LCM, the two compartments have not been perfectly separated, resulting in shared DEGs.

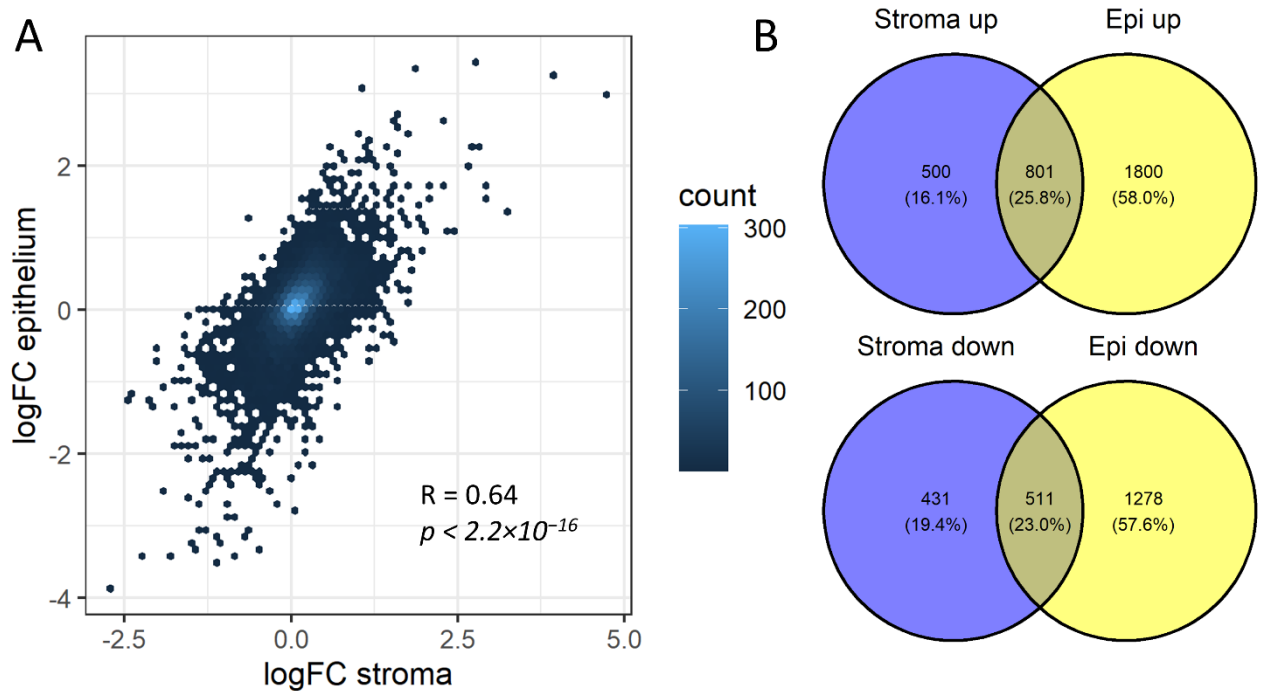


Figure 33. DEGs comparison between stroma and epithelium. (A) Cross-datasets average logFC for genes measured in normal/cancerous stroma (X axis) or in normal/cancerous epithelium (Y axis). Pearson's correlation and p -value are indicated; the color indicates the number of overlapping dots. (B) Venn diagrams comparing significant DEGs (p -adjusted < 0.05) detected in the stroma and in the epithelium.

In the meta-analytic setting, I defined lists of robust DEGs. Specifically, I found 4390 DEGs in the epithelium (2601 up- and 1789 down-regulated in tumors) and 2243 DEGs in the stroma (1301 up- and 942 down-regulated) (Figure 33B). The full list of fold changes and p -values obtained for each dataset and condition and GO categories enriched for each class are available listed on Zenodo (<https://zenodo.org/record/5013252#.YU3j3bgzZPY>, Table S6 and S7). Of note, by combining the information present in several datasets, it was possible to identify 229 DEGs that would not be identified in any individual dataset if analyzed separately.

Although, as mentioned above, changes in gene expression in the tumor stroma and epithelium are globally correlated (Figure 33A). For 17% of the DEGs, there was statistically significant evidence of differential expression in only the stroma/epithelium. Moreover, 104 genes showed significant differential expression in both the compartments but with opposite signs. I posit

that genes that are either regulated in one compartment only or with opposite regulation between compartments, though potentially relevant in tumor progression, may be hidden in bulk datasets due to their regulation in one compartment being confounded by a different regulation in the other.

To test this hypothesis, I selected five classes of genes: (1) genes up-regulated in both the tumor stroma and epithelium (UpBoth); (2) genes with evidence of up-regulation in only one of the two compartments (up-regulated in the epithelium—UpEpi—or in the stroma—UpStr); (3) genes with opposite signs of differential expression in the two compartments (up-regulated in the stroma and down-regulated in the epithelium—StrEpi—or down-regulated in the stroma and up-regulated in the epithelium—EpiStr); (4) genes with evidence of down-regulation in only one of the two compartments (down-regulated in the epithelium—DnEpi—or in the stroma—DnStr); and (5) genes down-regulated in both the tumor stroma and the epithelium (DnBoth). I tested the robustness of these classes with a 2-fold cross validation, obtaining relatively good sensitivity and specificity, especially for the UpBoth and DnBoth classes (Figure 34). It has to be noted that the other three classes might particularly suffer from any difference in the precision and the efficiency of epithelial and stromal cells separation across the various datasets, thus resulting in lower sensitivity. Nevertheless, the specificity is still high also for these classes, indicating that retrieved differentially expressed genes are likely not false positives.

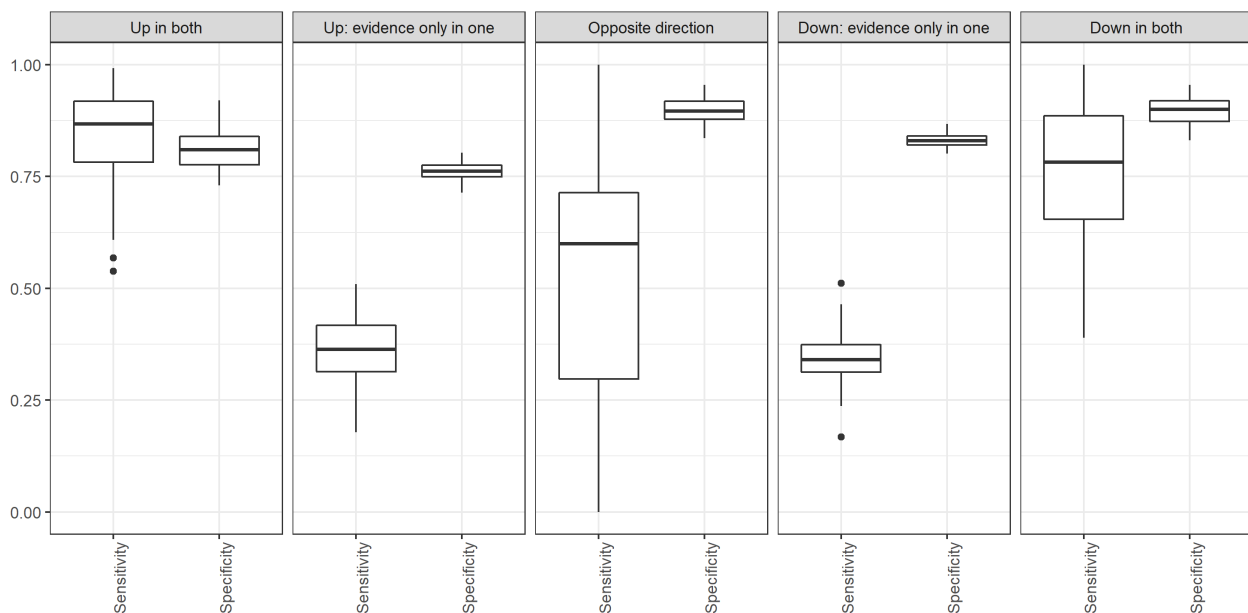


Figure 34. Robustness assessment for the 5 DEG classes. Boxplot showing the sensitivity and specificity obtained through a 2-fold cross validation repeated 100 times for the classes of genes “Up in both” (up-regulated in both tumor stroma and epithelium), “Up: evidence only in one” (up-regulated only in one of the two compartments, but not regulated in the other), “Opposite direction” (opposite sign of regulation in the two compartments), “Down: evidence only in one” (down-regulated only in one of the two compartments, but not regulated in the other), “Down in both” (down-regulated in both tumor stroma and epithelium).

For each gene, I measured the average expression fold changes between normal breast and breast tumor in bulk samples obtained from the TCGA [110]. The classes with the highest fold changes are those comprising the genes that are consistently differentially expressed in both compartments, while the remaining classes show average fold changes closer to zero (Figure 35A), indicating that their genes are not detected as differentially expressed when analyzing bulk tumors. Similar results were obtained with the METABRIC dataset, which due to its extensive clinical annotations, allowed us to show graded relationships with overall patient survival and tumor grade and size in the above classes (Figure 35B-D). Thus, for two independent bulk tumor datasets, I could show that genes regulated in both tissues are more strongly differentially expressed or correlated with clinical features, suggesting the possibility that their expression is more reliably measured in bulk than genes with evidence of regulation in only one tissue or with opposite regulation in the stroma and the epithelium. Indeed, all of the classes of genes here defined show a good robustness (Figure 34) and are therefore likely regulated but missed when analyzing bulk data. A more detailed picture of the DEGs classes indicates that genes with opposite regulation in the stroma/epithelium show correlation with the clinical features in line with their regulation in the epithelium: The EpiStr class increases its expression in bulk tumors when compared to normal breast (Suppl. Fig. 4A). It is enriched with genes correlated with poor prognosis (Suppl. Fig. 4B) and correlates with higher tumor grade (Suppl. Fig. 4C) and size (Suppl. Fig. 4A). The class of genes down-regulated in the epithelium and up-regulated in the stroma shows the opposite trend for its expression in bulk tumors and its correlation with size, though it is not related with patient survival or tumor grade (Suppl. Fig. 4). These observations fit with more epithelium content than stromal content in bulk samples, which hides the signal originating from the tumor stroma. Indeed, the top DEGs in the category of genes that are higher in the tumor stroma and lower in the tumor epithelium are relatively consistently up-regulated in tumor stroma across the available datasets (Figure 36A). However, they all appear significantly down-regulated in tumor samples from the TCGA (Figure 36B), confirming that their regulation in the stroma is not detected in bulk.

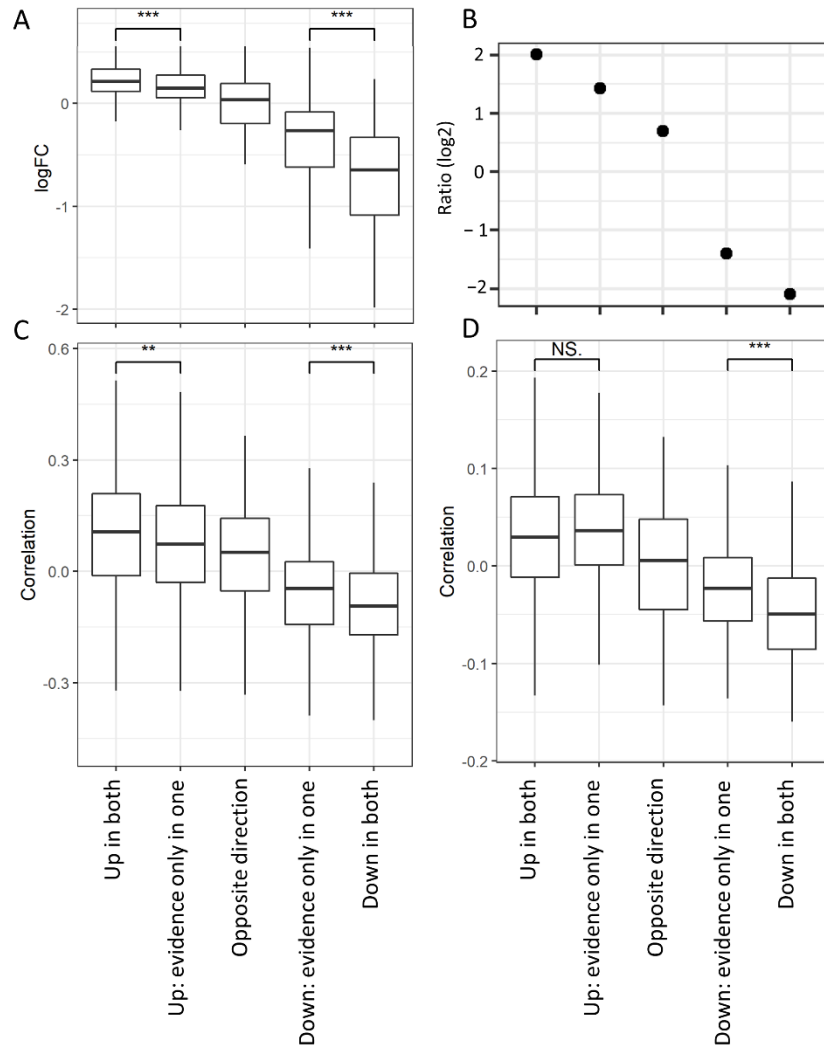


Figure 35. The 5 classes of DEGs in bulk samples. Genes regulated in both tumor epithelium and stroma, significantly regulated only in one of the two compartments and with opposite regulation in the two compartments show different degrees of: (A) differential expression between tumor and normal samples in the bulk samples of the TCGA cohort; (B) ratio between the number of genes significantly correlated with poor prognosis and the number of genes significantly correlated with good prognosis in the METABRIC cohort. High values indicate that many genes are correlated with poor prognosis, while inversely, negative values indicate that many genes are correlated with good prognosis; Spearman's correlation with tumor grade (C) and size (D). Significance was assessed with the Wilcoxon test (** <0.01, *** <0.001, NS = Not Significant).

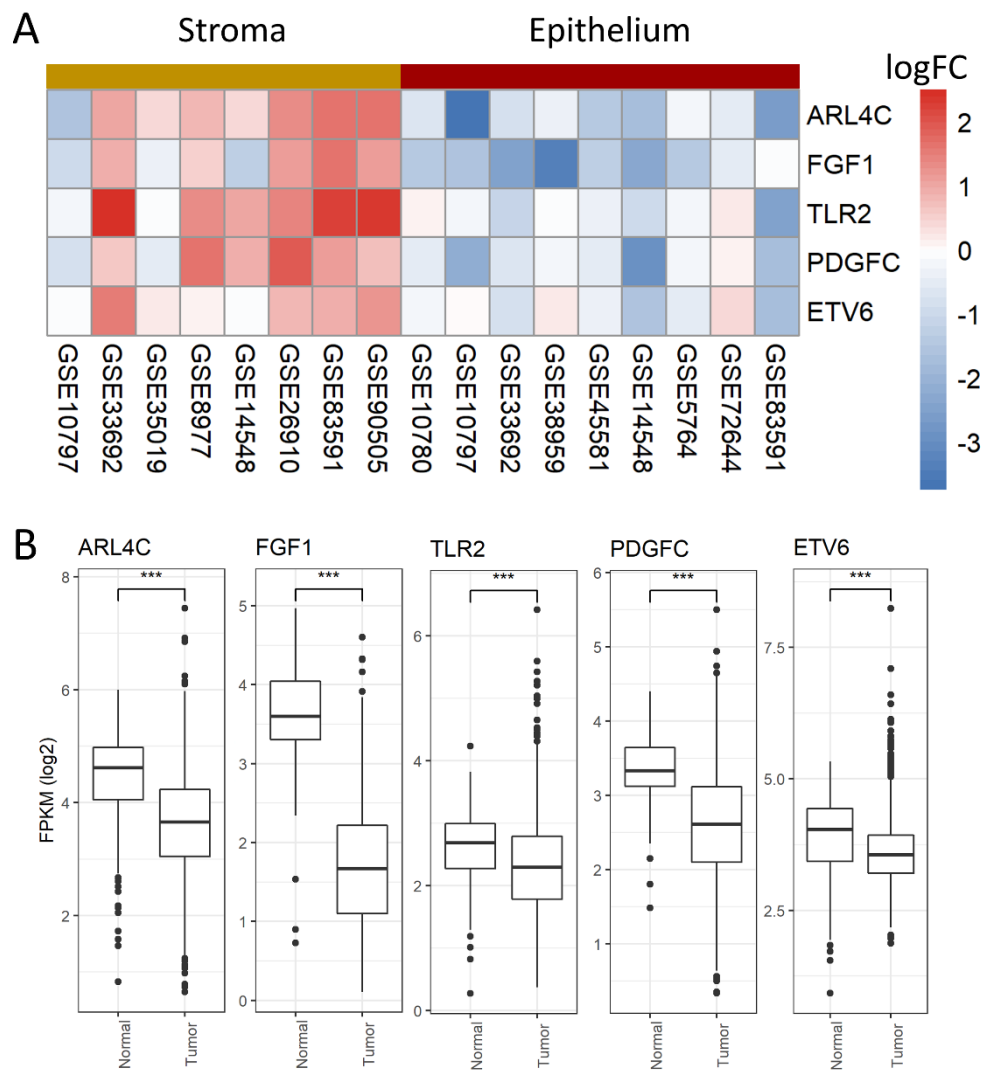


Figure 36. The opposite regulation of genes in tumor stroma or epithelium is hidden in bulk samples. (A) Top five significantly differentially expressed genes with higher expression in tumor stroma and lower expression in tumor epithelium when compared to respective compartments in normal breast tissue. Rows correspond to the five selected genes and columns to GEO IDs of datasets where the comparisons were possible and that were merged in the meta-analysis. On the left, datasets comparing normal and tumor stroma are shown, while on the right, there are datasets comparing normal and tumor epithelium. The color in the heatmap indicates logFC value for the corresponding gene and dataset. (B) Gene expression changes detected in bulk samples from the TCGA dataset for the five selected genes tested with the Wilcoxon test (***) $p < 0.001$.

To test the potential relevance of genes that are differentially expressed in one compartment only as compartment-specific clinical markers, I computed their correlation with the tumor grade or age at onset when measured in the stroma or in the epithelium of invasive BC, identifying a differential correlation in line with the compartment in which each gene class is regulated. Indeed, genes that are up-regulated in the stroma are more strongly correlated with a higher grade and earlier onset when measured in the stroma than in the epithelium, while genes that are down-regulated in the stroma show the opposite trend (Figure 37), supporting the hypothesis of their potential compartment-specific clinical relevance, which cannot be assessed without separating

them. Similar trends are observed with node positivity and size, even if some tests do not reach statistical significance (Figure 37). Nevertheless, I did not observe this consistent relationship for genes that were regulated in the epithelium (Suppl. Fig. 5A). Accordingly, the classes of genes with opposite regulation in the stroma and the epithelium display an opposite relationship with clinical features when measured in either the stroma or in the epithelium (Suppl. Fig. 5B), confirming that these classes of DEGs might also be relevant for tumor progression but differentially regulated in the two compartments.

I further showed that with our compartment-specific gene expression database and the use of meta-analysis, it is possible to identify relationships between genes and clinical features that could not be identified otherwise. As an example, I took the 50 most down-regulated genes in tumor stroma vs. normal stroma (DnStr class). Twelve of these genes were also significantly negatively correlated with tumor grade when measured in the stroma, and most of them (75%) displayed a stronger relationship with grade in the LCM data than in bulk. I identified UPB1 (beta-ureidopropionase 1) as negatively correlated with grade in the tumor stroma but not in bulk (METABRIC dataset, Figure 38). Of note, despite the relatively small number of samples annotated for clinical features and the high within- and between-dataset variability, I was able to improve statistical power by combining multiple datasets (Figure 38, left panel). In one case (the HSD11B2 gene, hydroxysteroid 11-beta dehydrogenase 2), the correlation in the stroma and in bulk shows opposite trends (Suppl. Fig. 6). This could be due to the confounding effect of multiple cell types present in mixed samples. Conversely, I identified NECAP2 (NECAP endocytosis associated 2) as up-regulated in tumor stroma and correlated with higher tumor grade when measured directly in the stroma (p -value = 0.01, average p = 0.30) but only slightly when measured in bulk (p -value = 0.03, p = 0.05).

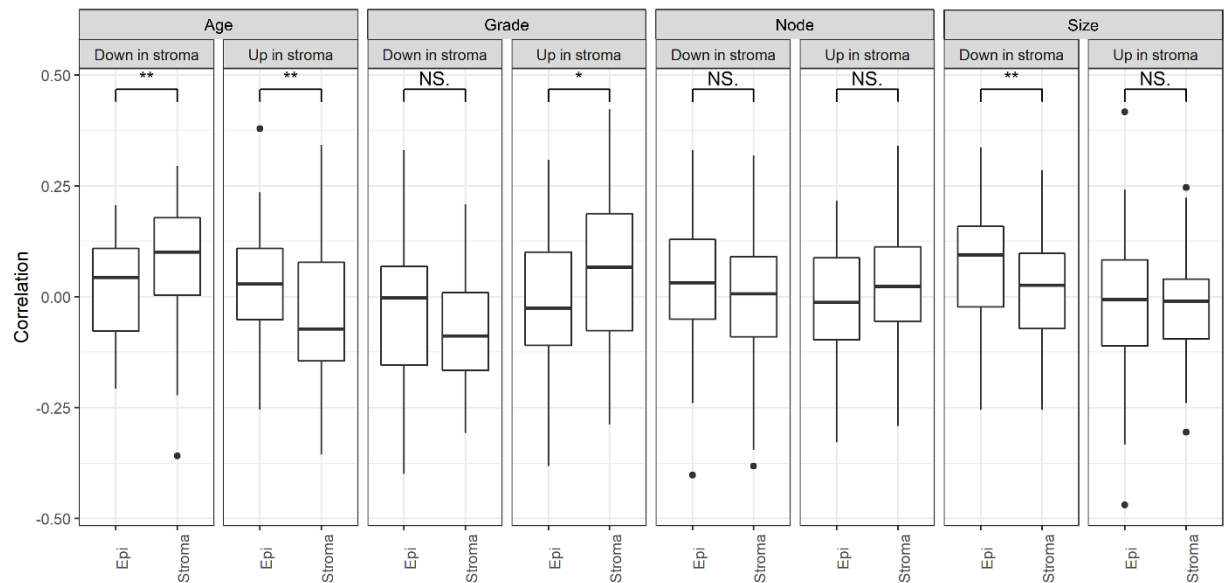


Figure 37. Compartment-specific relationship with clinical features for genes regulated in tumour stroma. Spearman's correlation between gene expression and age at onset, tumour grade, lymph node status, and size for the classes of DEGs regulated only in tumour stroma, calculated for each dataset separately and then averaged. The correlation was computed when their expression was measured directly in the stroma or in the epithelium and was compared with the Wilcoxon test (* <0.05, ** <0.01, NS = Not Significant).

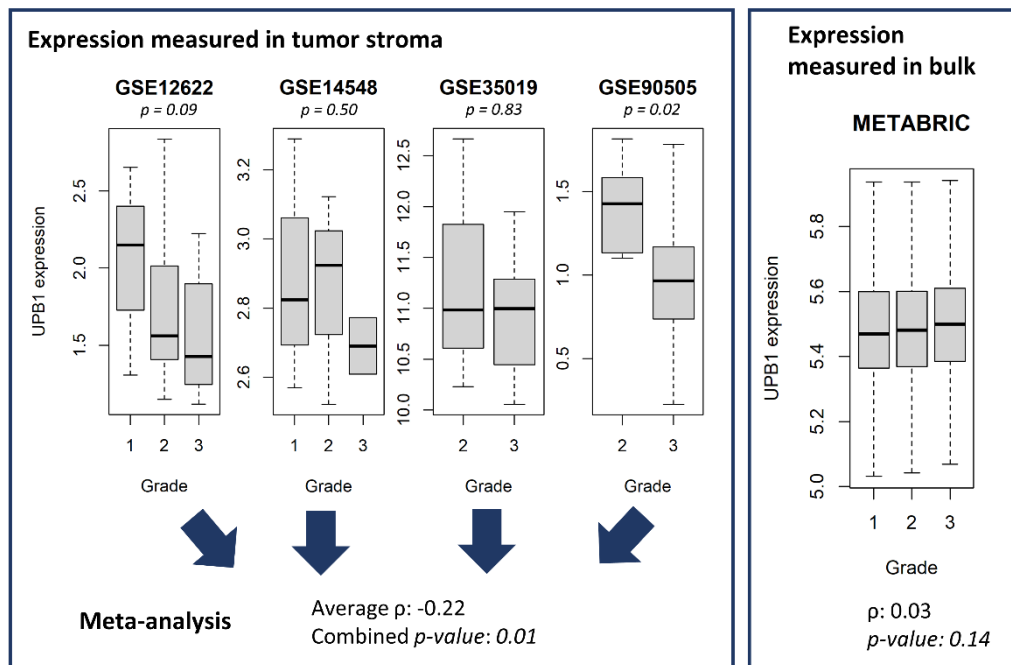


Figure 38. UPB1 is correlated with lower grade in the stroma, but not in bulk. Left panel: UPB1 expression in the 4 datasets with annotation for grade in the tumour stroma. The *p*-value for the correlation between UPB1 expression and tumour grade is indicated above each boxplot. Collapsing *p*-values with a meta-analysis, the statistical power increases, and the *p*-value reaches significance. Overall, the correlation between UPB1 expression and tumour grade in stroma is significantly negative. Right panel: no significant relationship between UPB1 expression and tumour grade in bulk in the METABRIC cohort.

Considering the overlap with gene ontology categories, genes in the UpBoth class show enrichment for mitochondrion-related categories, the extracellular matrix (ECM), and antigen processing, while UpEpi genes are enriched for cell cycle and DNA repair. No categories are enriched for UpStr genes. Counter-intuitively, DnBoth genes are enriched for angiogenesis-related categories. DnEpi genes belong to the “cornification” category, while no biological processes are over-represented in DnStr genes. Interestingly, genes with opposite regulation and over-expressed in tumor stroma are enriched for cytokine secretion and Toll-like receptor 2 signaling.

I repeated the meta-analysis to identify robust DEGs in tumor blood vessels (datasets used: GSE15363, GSE31138, GSE7413, and GSE43379), obtaining 13 up-regulated genes and 1 down-regulated gene.

Despite the difficulty of accurately detecting the signal deriving from specific cellular compartment in bulk, as discussed above, I also showed in bulk samples that the epithelial and vascular signatures are independent predictors of a patient’s disease-free survival (DFS) (Figure 39). This, again, points to the relevance of cell type specific signatures in describing tumor biology. Indeed, higher expression of up-regulated genes in the epithelium or in the blood vessels independently correlate with poor prognosis in the METABRIC BC cohort (Figure 39A), while down-regulated genes in the blood vessels define the only significant signature of good prognosis (Figure 39B). I could not detect any relationship between the stromal signatures and patient DFS. This result can be attributed to the presence of multiple cell types in bulk samples, confounding compartment-specific signals. Nevertheless, it is also possible that the stromal signatures here defined, despite being correlated with tumor grade and age at onset, are not correlated with patient survival.

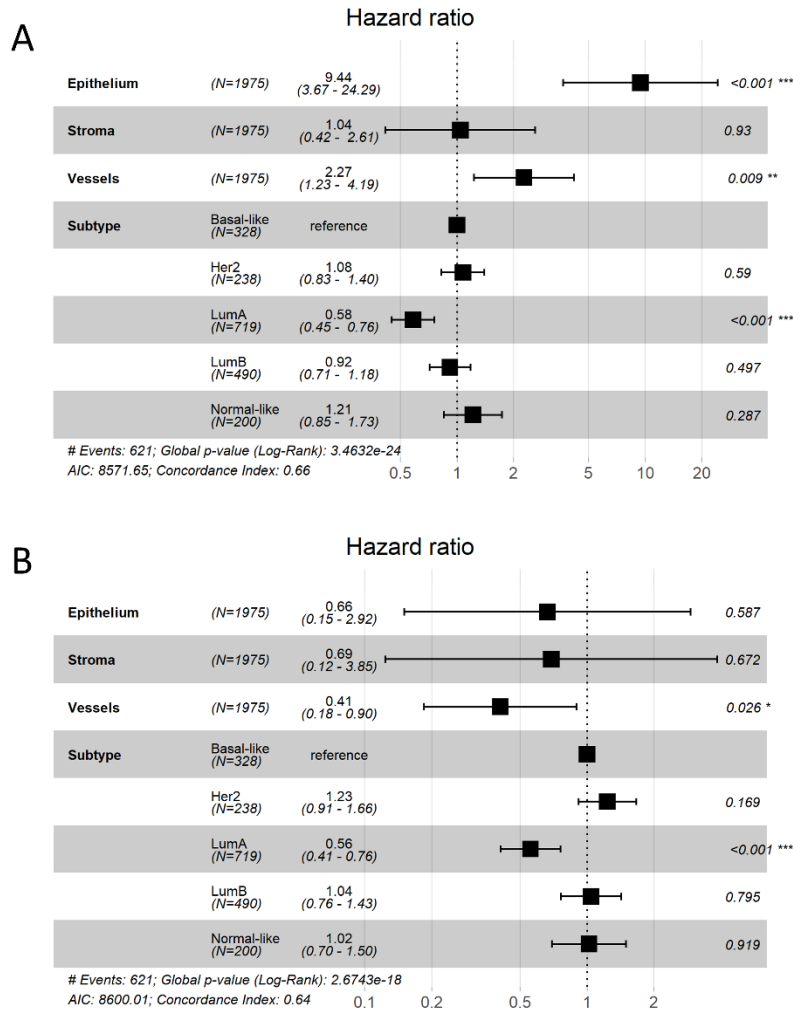


Figure 39. Survival models with epithelial, stromal, and vessel DEGs. Forest plots of multivariate Cox models of disease-free survival with PAM50 subtypes and expression levels of genes in gene signatures (A) up- or (B) down-regulated in tumor epithelium, stroma, or vessels when compared to corresponding normal tissues (* <0.05, ** <0.01, *** <0.001).

2.2.4. Stromal and epithelial markers to impute cell proportions from bulk samples

Computational methods to estimate cell type proportions in bulk transcriptomes often require gene expression signatures of the cell types of interest. Amongst the first proposed and most applied methods is ESTIMATE [111], which based on a single sample GSEA (ssGSEA) of stromal and immune signatures to infer their proportions from the transcriptomes of cell admixtures. I computed ssGSEA on primary BC transcriptomes of the METABRIC cohort based on the stromal and epithelial markers obtained through a meta-analysis comparing tumor epithelium and stroma from our gene expression database (datasets detailed in the methods). I observed that the epithelial signature is positively correlated with the clinician-defined measure of cellularity (Spearman's rho= 0.12, p-value = 2.8×10^{-7}), while the stromal signature is negatively correlated with it (Spearman's rho= -0.25, p-value < 2.2×10^{-16}) (Figure 40), indicating that these marker lists are representative

of the corresponding cell compartment and are appropriate to infer cell proportions. Moreover, our stromal signature shows a stronger correlation with cellularity than the estimated tumor purity obtained with the original ESTIMATE signatures (for which Spearman's $\rho = -0.15$, $p\text{-value} = 1.4 \times 10^{-11}$).

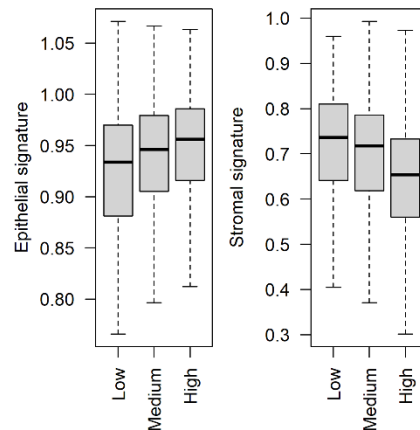


Figure 40. Correlation between epithelial or stromal markers and cellularity in the METABRIC cohort. Epithelial and stromal signatures were obtained through a meta-analysis comparing the expression patterns of tumor stroma and epithelium. The signatures were then used to build stromal and epithelial scores representing cell proportions for each tumor sample in the METABRIC cohort and correlated with clinician annotation of cellularity.

2.2.5. Potential use of the database and web platform

Due to the number of datasets and annotation categories, our meta-analysis offers many combinations of variables that can be selected for comparison to address specific questions. In order to make this resource available to the scientific community and to facilitate the choice of the datasets to analyze, I indicated the annotations available for each dataset (Figure 41 and Suppl. Fig. 7) and some of the comparisons that can be made using a sample from a specific dataset (Suppl. Fig. 8-9). For example, there are 10 different datasets with estrogen receptor status annotations in stroma samples (Figure 41), three datasets that can be used to compare the stroma of IDC and DCIS (Suppl. Fig. 8B), and three datasets permitting the comparison of Basal and LumA stroma in invasive BC (Suppl. Fig. 9A).

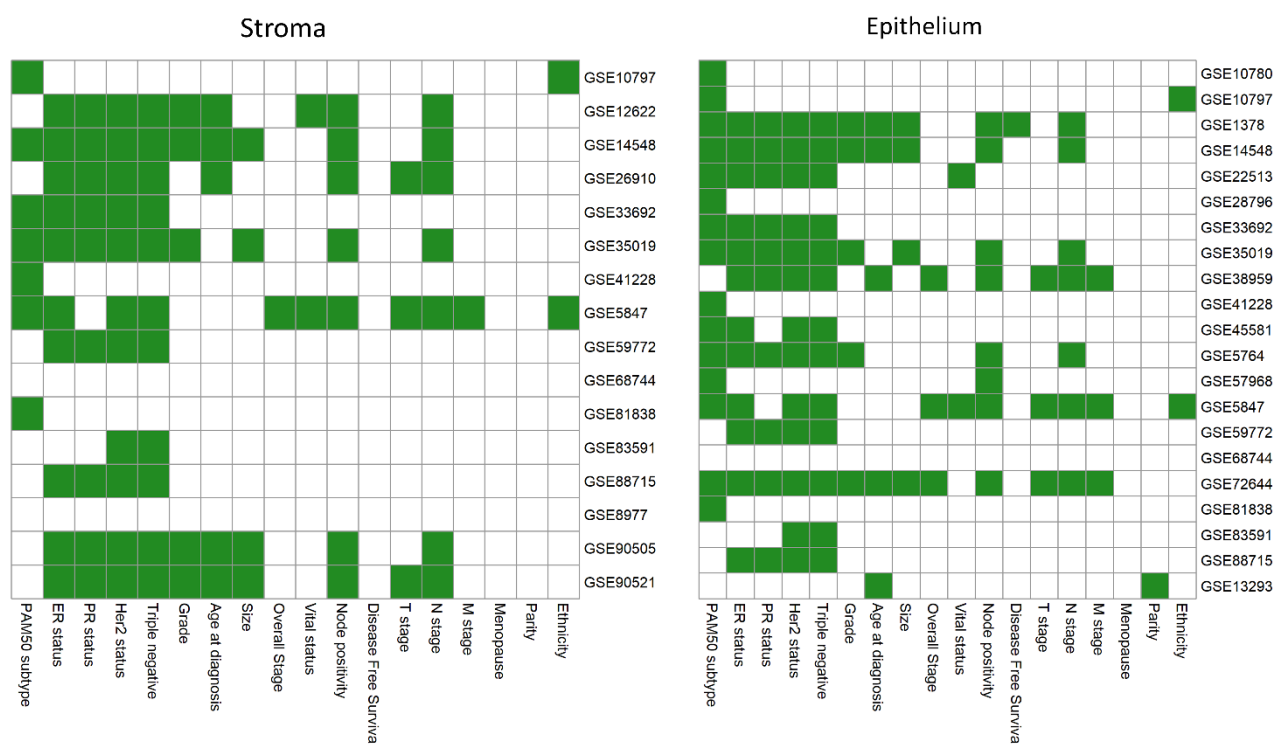


Figure 41. Summary of annotations available for each dataset, separating stromal and epithelial samples. In these matrices, only invasive BC samples have been considered. Green indicates the presence of the corresponding clinical information (column) in the dataset (row).

To ease the interrogation of the whole collection of datasets, I created a web resource that provides a user-friendly interface, which permits to generate lists of DEGs between two conditions and the testing of enrichment for user-provided gene lists (<https://aurorasavino.shinyapps.io/metalcm/>). As a use-case example, we tested the enrichment of up-regulated genes in the tumor stroma for five signatures of the pro-oncogenic transcription factor signal transducer and activator of transcription (STAT) 3 activation. I set the parameters for comparing tumor stroma and normal stroma of any BC subtype, with a p-value threshold of 0.05. I then loaded the gene lists corresponding to the five signatures one at a time and obtained the Fisher test enrichment p-value seen in the “Enrichment” tab. I observed that the overlap is significant for 3/5 of the signatures when tested in the stroma, while none show enrichment for up-regulated genes in the tumor epithelium. This observation is suggestive of the different roles of STAT3 in the two compartments, linking tumorigenesis and its up-regulation in the stroma. A lack of significance in the epithelium due to higher variability in the available datasets cannot be excluded, but it is unlikely given that the number of DEGs detected in tumor epithelium was higher than in tumor stroma (Figure 33B).

2.3. Meta-Analysis of bulk tumors

Even though LCM data provide a greater level of detail, an incredible wealth of information is publicly available regarding gene expression profiles of bulk primary tumors. A simple search on GEO of human breast cancer datasets (expression profiling by array and expression profiling by high throughput sequencing) outputs 3516 series. Indeed, initial efforts for combining these data in a unique resource with harmonized metadata have led to the development of resources such as MetaGxBreast [115], which has been extensively used in the previous sections of this Thesis. Nevertheless, MetaGxBreast comprises only 39 of the available datasets, and gathers only a few of the annotated clinical features. Therefore, I aimed at complementing this database with transcriptome datasets deposited in GEO but not yet included, creating a resource with a richer set of manually curated clinical annotations.

I here present a preliminary collection of 55 BC transcriptome datasets (listed in Table 9), for a total of 18 362 samples. From GEO, only datasets with at least 100 samples and at least 5000 analyzed genes were considered. Data pre-processing was performed as for the LCM database. Moreover, samples with a Pearson's correlation > 0.99 were de-duplicated, as previously done for the MetaGxBreast resource [115]. This permitted, for example, to delete redundant datasets (e.g., GSE21653, GSE23720 both comprised in GSE31448) and samples present in multiple datasets (e.g. 40 samples in both GSE4611 and GSE26971).

With respect to the LCM database, additional features were retrieved and harmonized, such as the presence of BRCA1 or BRCA2 mutations, ki67 status, presence of p53 mutations, the Body Mass Index of the patient and the kind of treatment received (annotated as presence/absence of endocrine treatment, radiotherapy, or adjuvant/neoadjuvant treatment) (Tables 10,11, Figure 42).

Table 9. List of the collected BC transcriptomic datasets.

GSE102484	[206]
GSE103091	[207]
GSE103746	[208]
GSE105777	[209]
GSE10885	/
GSE10886	[210]
GSE115577	[211]
GSE123845	[212]
GSE130788	[213]
GSE131769	[214]
GSE136644	[215]
GSE137356	[216]

GSE142102	[217]
GSE150576	[218]
GSE158309	[219]
GSE16201	[220]
GSE17705	[221]
GSE20624	[222]
GSE20685	[223]
GSE22220	[154]
GSE22820	[224]
GSE23768	[225]
GSE26639	[226]
GSE26971	[227]
GSE28884	[155]
GSE31448	[228]
GSE32603	[229]
GSE36771	[230]
GSE36772	[230]
GSE41119	[231]
GSE45725	[232]
GSE4611	[233]
GSE48091	[234]
GSE49175	[235]
GSE49481	[236]
GSE54002	[237]
GSE54219	[238]
GSE58215	[239]
GSE59595	[156]
GSE65216	[240]
GSE69986	/
GSE70905	[241]
GSE70947	[241]
GSE76250	[242]
GSE76275	[243]
GSE78958	[244]
GSE81538	[245]
GSE8193	[246]
GSE86166	[247]
GSE86948	[248]
GSE87049	[249]
GSE87411	[250]
GSE87455	[251]
GSE93601	[252]
GSE96058	[245]

Table 10. Summary of the number of samples and datasets annotated for each clinical feature.

Category	# of samples	# of datasets
Disease status		
Tumor	17299	55
Normal counterpart	935	6
Normal	128	7
ER status		
Positive	2513	29
Negative	8145	27
PR status		
Positive	2525	18
Negative	4423	21
Her2 status		
Positive	1414	24
Negative	6878	25
LN positivity		
Positive	3679	20
Negative	3740	17
PAM50 subtype		
Basal-like	2386	43
Her2+	1884	44
LumA	5265	44
LumB	3500	44
Normal-like	978	42
Menopause status		
Pre-menopause	241	5
Peri-menopause	18	2
Post-menopause	537	5
P53 mutation		
No	125	2
Yes	69	2
Ki67 status		
Positive	1474	6
Negative	1293	8
TNBC		
Yes	539	3
No	8779	31
Ethnicity		
African American	322	
European American	304	
Caucasian	254	
Mexico Hispanic	159	
Neoadjuvant therapy		
Yes	344	3
No	112	2
Adjuvant therapy		
Yes	530	3

No	253	3
Radiotherapy treated		
Yes	118	2
No	54	2
Endocrine treated		
Yes	2841	4
No	715	2
BRCA1 mutated		
Yes	97	4
No	397	4
BRCA2 mutated		
Yes	28	3
No	355	3

Table 11. Clinical features' distribution.

Age	
Median (range)	57.2 (22-96)
# annotated samples	9677
# of datasets	24
Size (in mm)	
Median (range)	20 (0-140)
# annotated samples	5428
# of datasets	12
Grade	# of samples
I	1082
II	3404
III	3133
# of datasets	20
BMI (Body Mass Index)	
Median (range)	25 (16-46.8)
# annotated samples	953
# of datasets	4

This breast cancer transcriptome database could be profitably applied to address many biological questions, permitting the assessment of results' robustness. Here, I showed a few use cases:

- Validate the differential connectivity of Basal-like modules across BC subtypes
- Identify genes robustly differentially expressed between high- and low-grade tumors.

2.3.1. Differential connectivity of Basal-like modules

As shown in Paragraph 2.1.5, three basal-like modules are more tightly connected in the basal-like subtype than in the others: the `b_E2F_targets`, `b_Estrogen_response` and `b_Estrogen_response_early`. This result has been obtained by comparing the `kWithin` of modules' genes across different subtypes in the METABRIC and in the TCGA dataset. Here, I tested the robustness of this result making use of the additional 55 transcriptome datasets of primary breast tumors collected as part of this database. To have a reliable measure of the adjacency and hence of the `kWithin` of each gene, I restricted the analysis to the datasets with at least 10 samples per tumor subtype, excluding the least abundant Normal-like subtype, not used in the initial modules' selection (Paragraph 2.1.5). Figure 43 displays the p-values of Wilcoxon ranked sum tests for each dataset-module combination. The `b_E2F_targets`, `b_Estrogen_response` and `b_Estrogen_response_early` modules are amongst the most consistently over-connected in the basal-like subtype, together with the `b_Allograft_rejection` and the `b_EMT1`. The remarkable variability between datasets points to the importance of performing analyses on multiple cohorts. Indeed, the same `b_E2F_targets` module that has been extensively studied in this Thesis does not appear to be more tightly connected in the basal-like subtype in 6/40 datasets and dissecting the reason for this difference could inform on relevant variables distinguishing cohorts and influencing the biological conclusions that can be driven from this kind of experiments. On the other side, the `b_Allograft_rejection` and `b_EMT1` modules, which were initially excluded because of the non-consistent difference in connectivity between basal-like and other subtypes in the METABRIC and TCGA cohorts, result here amongst the most strongly over-connected in the basal-like subtype, with `Allograft_rejection` showing this difference in 35/40 of the collected datasets.

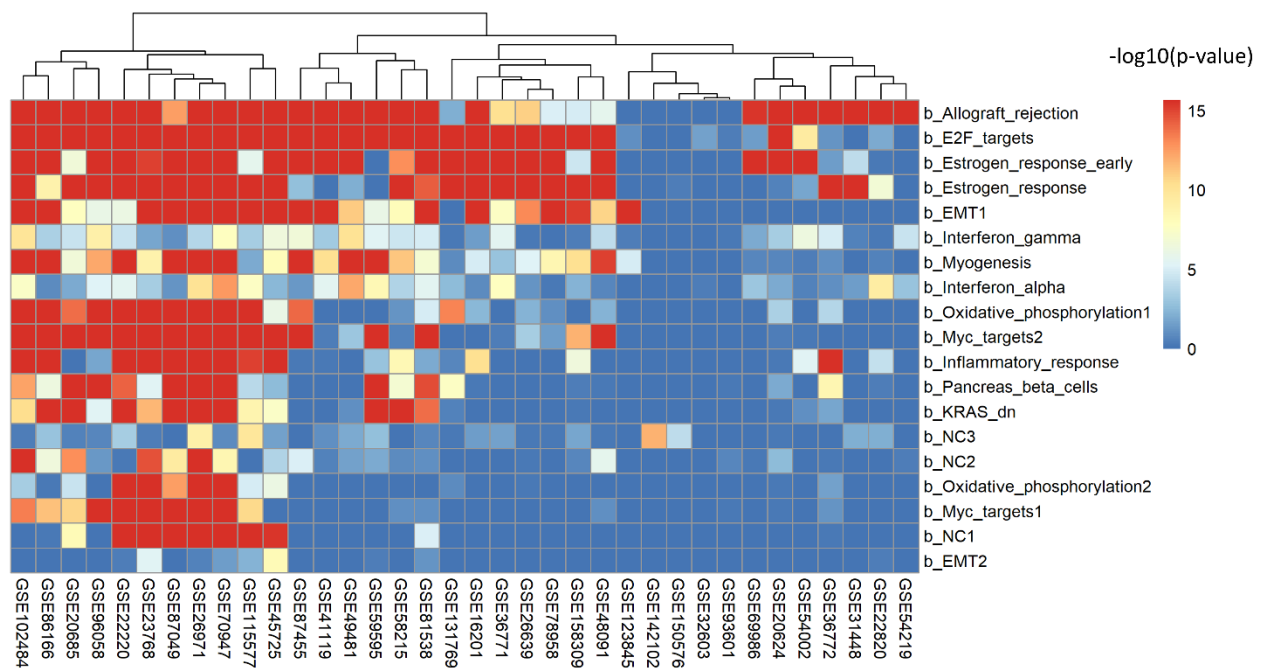


Figure 43. Heatmap showing the p-values for Wilcoxon ranked sum tests comparing genes' kWithin in the basal-like subtype vs other subtypes in each gathered dataset (columns) and for each basal-like module (rows). Red indicates higher significance. Modules are ordered based on the number of significant tests.

2.3.2. Differentially expressed genes between high- and low-grade tumors

Similarly to the LCM meta-analysis (Paragraph 2.2), the whole bulk database can be used to identify robust lists of differentially expressed genes between specific conditions across all available datasets. As an example, I compared gene expression between high grade (grade 3) and low grade (grade 1) tumors, performing this comparison in a total of 16 independent datasets in a meta-analytic setting, following the same procedure previously described for the LCM database. The differentially expressed genes are enriched for GO categories such as “immune effector process” and “DNA metabolic process”, for genes up-regulated in high-grade tumors, and “response to growth factor” for genes down-regulated in high-grade tumors, processes known to be involved in tumor aggressiveness.

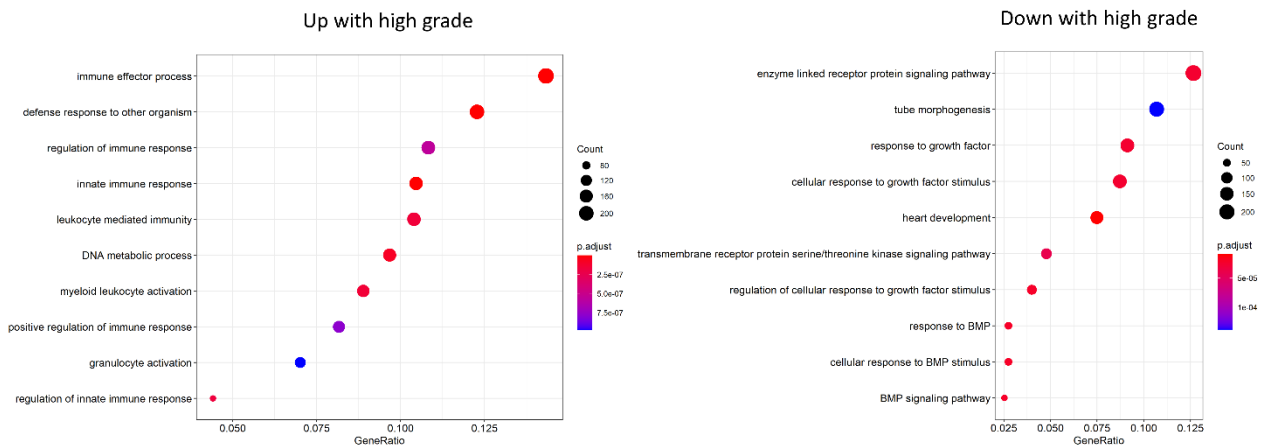


Figure 44. Top 10 Gene Ontology categories enriched in up- and down-regulated genes in high-grade tumors.

2.4. crossWGCNA: a co-expression-based method to infer genes involved in inter-tissue communication

Lastly, to study tumor-stroma crosstalk, I defined a novel method, which I call “crossWGCNA” (cross-tissue WGCNA), based on gene co-expression networks. I adapted the WGCNA algorithm (Weighted Gene Co-expression Network Analysis) modifying the metric used to define nodes’ distance. As for WGCNA, nodes in the network are genes, while edges indicate co-expression of pairs of genes across samples. The main difference with respect to standard co-expression network lies on the fact that correlated genes are here measured in two different tissues. Indeed, the fundamental idea behind this method is that a secreted factor produced by one tissue, for example, can be released and influence the expression of a responder gene in another tissue, and hence the secreted factor in tissue 1 and the responder gene in tissue 2 will be co-expressed (Figure 45A).

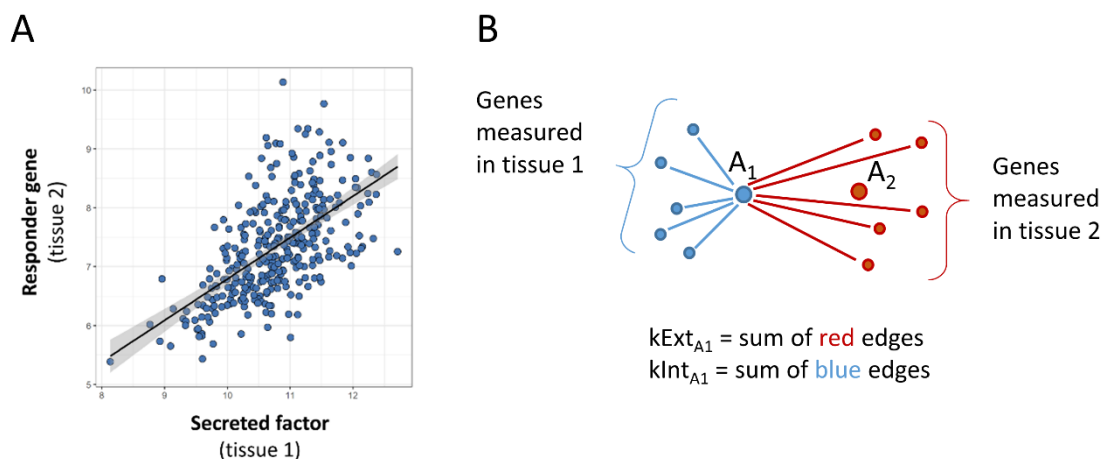


Figure 45. A) The principle underlying crossWGCNA is that a gene expressed by a tissue (tissue 1) can influence the expression of a responder gene in another tissue (tissue 2), and hence their expression levels, measured in tissue 1 and 2, will be correlated. **B)** Nodes in blue represent genes measured in tissue 1, while nodes in red are measured in tissue 2. The same gene, A , is measured in both tissues with distinct expression levels (A_1 and A_2). Edges represent correlations: blue edges are correlations between A_1 and the other genes measured in tissue 1, while red edges are

correlations between A_1 and the other genes measured in tissue 2. The edge between A_1 and A_2 (self-loop) is imposed to zero. The k_{Ext} of the gene A_1 is defined as the sum of red edges, and its k_{Int} is defined as the sum of blue edges.

The method requires as input the transcriptomic profiles of matched samples from the same experimental subjects, deriving from two or more tissues or organs. It then builds a gene adjacency matrix based on genes' inter-tissue pairwise correlations, which it uses as input for WGCNA, finally identifying modules of tightly connected genes. To avoid artefacts due to coordinated gene expression induced by external factors acting on both tissues, self-loops are imposed to zero. In addition, two metrics, which I call k_{ext} and k_{int} , are defined based on the weighted inter-tissue or intra-tissue degree, respectively, and indicate the influence of each gene on within- or between-tissue communication. Modules and genes with high $k_{\text{ext}} / k_{\text{int}}$ represent good candidates as central mediators of interactions.

2.4.1. Application to brain connectivity

To have a first validation of the method's suitability in describing the cross-tissues coordination of gene expression, I applied it to a context where an independent measure of connectivity was available. Therefore, I tested the correspondence between the measure of connectivity produced by the crossWGCNA and the fMRI (functional magnetic resonance imaging) connectivity to quantify brain inter-region communication.

In particular, I analyzed the GTEx transcriptome dataset, comprising gene expression profiles of hundreds of subjects across 11 brain regions (e.g., 255 subjects for brain cortex) and then confirmed the result with an independent transcriptome dataset (GSE25219, [253]), analyzing the same brain regions. Biologically, the strongest communication signal between brain areas is expected to be mediated by chemical signals exchanged at synapses, and hence I restricted this first analysis to synapse-related genes as annotated in the SynGO database (<https://www.syngoportal.org/>). For each pair of brain regions, I applied crossWGCNA to calculate inter-region adjacencies and then obtained an overall interaction score as the sum of these adjacencies. As a second measure of connectivity, the functional interaction between the same pairs of regions was obtained through the analysis of fMRI data (ABIDE project, <http://preprocessed-connectomes-project.org/abide/>, analysis performed by Tommaso Lanciano). The brain areas analyzed in all three datasets (transcriptome and fMRI) were seven: Frontal Cortex (BA9), Cerebellum, Caudate (basal ganglia), Putamen (basal ganglia), Hippocampus, Anterior cingulate cortex (BA24), Amygdala. These two methodologically very different measures of brain regions' interactions show a strong positive correlation (Figure 46A), which is preserved even taking physical distance between brain areas into

account (coefficients in a bivariate linear model with $p < 0.05$). Moreover, it results that the crossWGCNA score is better than inter-region transcriptional similarity in explaining functional coupling (Figure 46B).

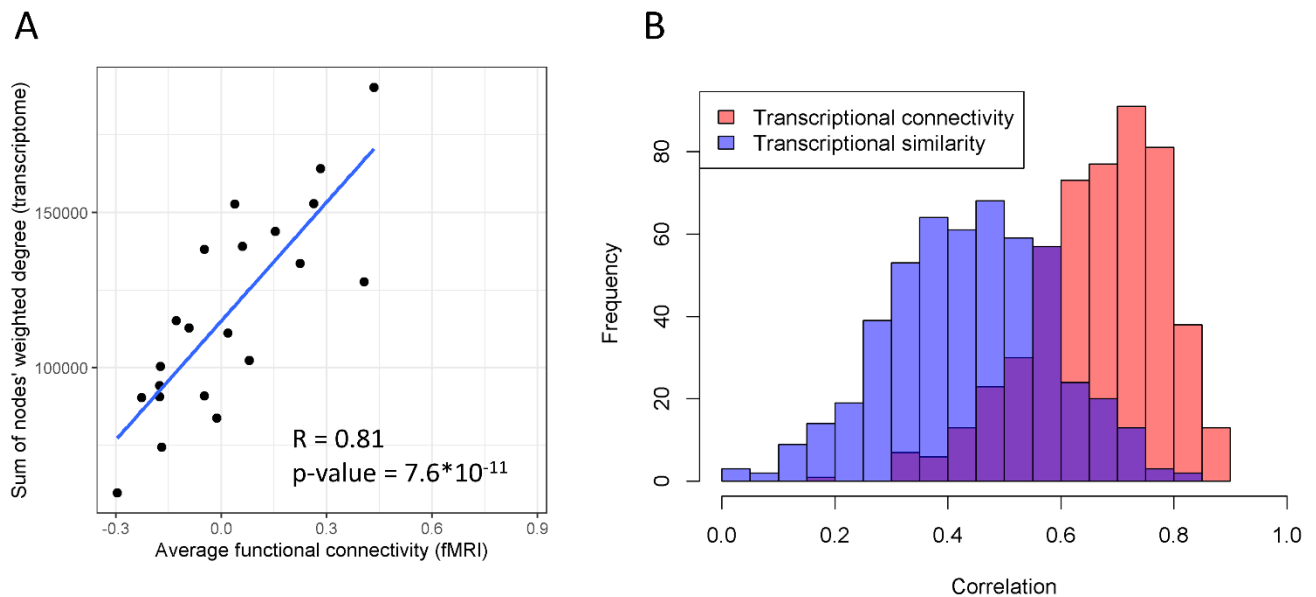


Figure 46. crossWGCNA and fMRI brain connectivity represent the same phenomenon. A) Correlation between functional connectivity for each pair of analyzed brain areas in one healthy subject (x axis) and crossWGCNA measure of connectivity (y axis), obtained as the sum of nodes' weighted degrees. B) Correlation distribution for the whole cohort of healthy subject with fMRI data. In pink is shown the correlation between fMRI connectivity and transcriptional connectivity as for the crossWGCNA method, in blue the correlation between fMRI connectivity and transcriptional similarity between brain areas.

2.4.2. Application to breast tumor epithelium and stroma

The correlation between two biological signals of such different kind, gene co-expression and functional connectivity, but representing the same phenomenon, inter-region communication, provided confidence on the reliability of the method and motivated its application to additional contexts. In particular, I studied tumor-microenvironment interactions in breast cancer, the main biological system analyzed in this Thesis, using data from laser capture microdissected tumors separating stromal and epithelial compartments.

I analyzed six independent LCM datasets with matched stroma-epithelium transcriptome profiles from the same tumor, part of the LCM transcriptome database previously described (Paragraph 2.2, [116]). I applied crossWGCNA on each dataset, obtaining lists and clusters of highly cross-compartment interacting genes. Specifically, for each dataset the algorithm gives as output two cross-compartment interaction scores, one for the stromal genes in their co-expression with epithelial genes (k_{ratio}) and one for the epithelial genes in their co-expression with stromal genes

(k_{Ratio}), plus clusters of genes being measured either in the stroma or in the epithelium, roughly equally represented in each cluster.

As additional validations for the crossWGCNA method, I provided multiple evidence coming from the application to this biological context:

- Functional Enrichment of highly cross-compartment interacting genes
- Comparison with genes regulated upon stroma-epithelium co-cultures
- Comparison with scores obtained with an external single cell RNA-seq dataset

2.4.2.1. Functional Enrichment of highly cross-compartment interacting genes

Genes belonging to the Gene Ontology categories of “G-protein coupled receptors signaling pathway” and “receptor regulator activity” are consistently enriched amongst the genes with the highest k_{Ratio} and k_{Sratio} scores, as shown with a GSEA on genes ranked according to these “inter-tissue interaction” scores. In Figure 47, the top 10 most significantly enriched categories for an exemplificative dataset are shown. The overlap with receptor-signaling categories is in line with their potential relevance in inter-compartment interactions, thus supporting the validity of crossWGCNA to identify such players.

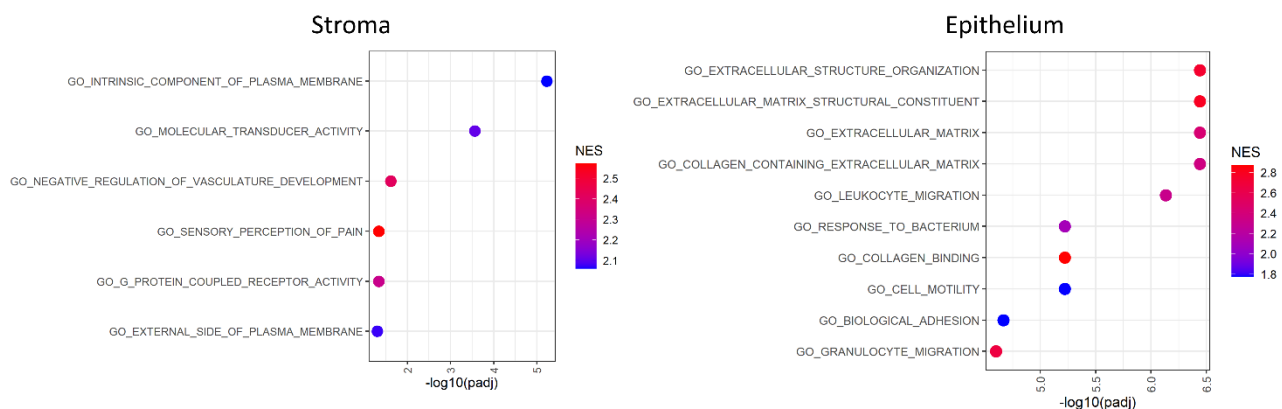


Figure 47. Top 10 significant Gene Ontology categories in a GSEA ranking genes based on k_{Sratio} (left) and k_{Ratio} (right) for the GSE10797 dataset

2.4.2.2. Comparison with genes regulated upon stroma-epithelium co-cultures

It has been previously shown that cancer cells and fibroblasts exchange signals and reciprocally influence their gene expression profiles, with cancer associated fibroblasts often being instructed by cancer cells in supporting their growth. Hence, the ways through which this communication happens have been investigated via co-culture experiments, where the two cell types are put in direct contact or where the culture medium influenced by the growth of one cell type, likely containing molecules released by those cells, is used as growth medium for the second cell type, therefore indirectly influenced by the first one. In some instances, the gene expression

profiles before/after the co-culture have been assessed. The genes changing expression in one cell type upon co-culture are responding to the presence of the other cell type, and I hence used these external datasets to validate the highly communicating genes obtained through crossWGCNA.

I collected 10 transcriptomic datasets of breast cancer cells co-cultured with cancer associated fibroblasts (Table 12), where the transcriptome of either cancer cells or CAFs (or both) had been profiled after culturing in the presence or in the absence of the second cell type. Often, in individual datasets several combinations of cell types have been analyzed. For each of these comparisons, I obtained a list of differentially expressed genes, for a total of 39 DEGs lists for genes modulated in cancer cells, and 31 DEGs lists for genes modulated in CAFs.

I then tested whether the DEGs upon co-culture are amongst the top ranking for k_{Ratio} or k_{ERatio} , performing a GSEA for each DEG list. It must be noted that highly interacting genes do not necessarily change expression levels upon co-culture, and that this test is an over-simplification of a complex mechanism. Nevertheless, of the 210 GSEA performed with DEGs in the epithelium, 89 (42%) are significant ($p\text{-value} < 0.05$), while 44/120 (37%) of tests with stromal DEGs are significant. Randomly shuffling genes' k_{Ratio} and k_{ERatio} , the number of significant tests strongly decreases, confirming that this result is not due to chance (Figure 48).

Finally, I compared the DEG lists with modules' genes in each dataset, performing a Fisher test. Figure 49 shows the proportion of significant tests ($p < 0.05$) for each module in one exemplificative dataset. Of note, modules enriched for DEGs upon co-culture in the epithelium are also enriched for DEGs upon co-culture in the stroma (Figure 49, modules 1, 3, 4, 5, 8). This same pattern is confirmed in 5/6 datasets. Interestingly, genes measured in the epithelium that are part of modules 3 and 4 are enriched for cell-cycle related processes, and they do not significantly overlap with genes measured in the stroma belonging to the same modules. Therefore, the genes expressed in the stroma of these two modules are interesting candidates potentially regulating tumor cells' proliferation.

Table 12. Co-culture datasets employed for validating crossWGCNA.

Dataset	Epithelial cells	Fibroblasts	Type of contact	# DEG lists epi	# DEG lists stroma
GSE106503	MCF10A, MDA-MB-231	CAFs	Extracellular vesicles	0	2
GSE119253	MCF10DCIS	CAFs	Direct	1	0
GSE27018	High- and Low-grade tumors	CAFs (CCdl67TF)	Indirect	2	0

GSE27120	BT474 , LY2, MCF7, MDA- MB-231, MDA-MB- 361, SKBR3, T47D	CAFs (CRL2564)	Direct	0	7
GSE41678	MD231/Cal51	4 different fibroblast strains (HFF1, HFF2, Wi38, CCD1112Sk)	Direct	8	8
GSE4708	MD231	4 different CAFs	Indirect	4	4
GSE60998	MD231	MRC5	Direct	1	1
GSE70884	MCF7, BT474, BT20, MDA- MB-231, MDA-MB- 361, ZR75, SKBR, T47D, MDA-MB-468	CAFs	Conditioned medium	9	0
GSE80333	MCF10DCIS, SUM149PT, MCF7, T47D, and MDA- MB-453	CAFs	Direct	5	0
GSE95461	AU565, SKBR3	4 different CAFs	Conditioned medium	8	0

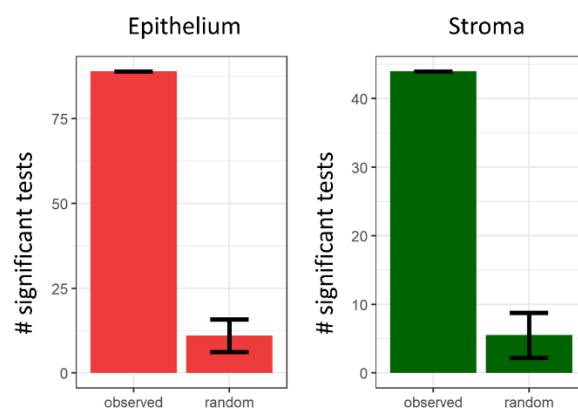


Figure 48. Number of significant GSEA testing whether the DEGs upon co-culture are amongst the top ranking for k_{Ratio} (left) or k_{Sratio} (right) when compared to sets of random genes.

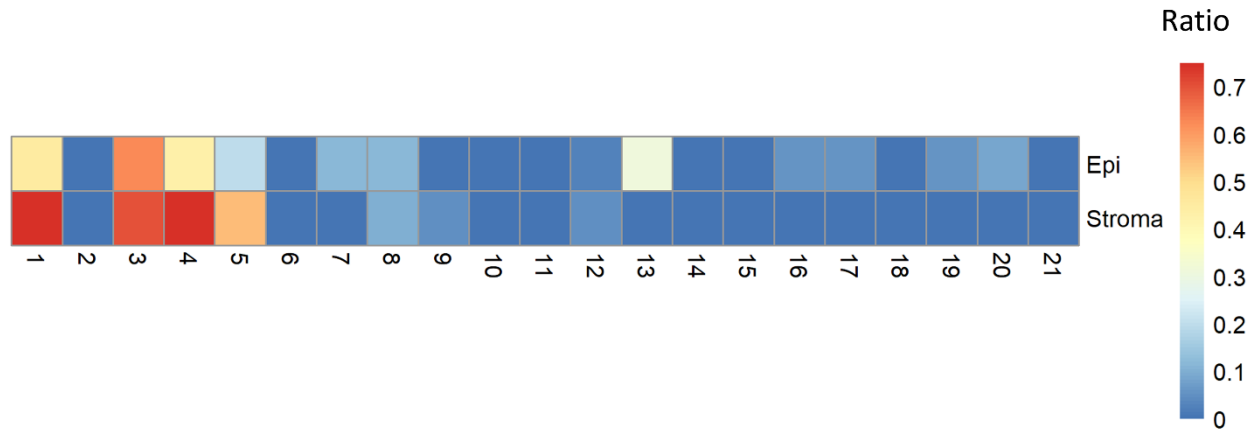


Figure 49. Proportion of co-culture DEG lists enriched in each module, separating genes measured in the epithelium (first row) or in the stroma (second row). Each column is a module. Red indicates a high proportion of significant tests.

2.4.2.3. Comparison with scores obtained with an external single cell RNA-seq dataset

An external validation was provided by the single cell dataset GSE161529 [254], analyzing 33 primary tumors at the single cell level, and hence distinguishing the expression profiles of different cell types. Her2-positive, ER-positive, and triple negative tumors were treated separately to define clusters of cells and assign the “epithelial” and “fibroblast” labels. Epithelial and fibroblast clusters were defined based on the expression of EPCAM (Epithelial Cell Adhesion Molecule), KRT19 (Keratin 19) and KRT18, on one side, and of PDGFRA (Platelet Derived Growth Factor receptor Alpha) and FAP (Fibroblast Activation Protein Alpha) on the other side. Then, all epithelial cells from the same tumor were collapsed by the average in a single gene expression profile. The same was done with fibroblasts, ending up with 33 epithelial and 33 matched fibroblast transcriptomes, one pair for each patient. CrossWGCNA was applied, and genes were ranked based on their k_{Ratio} or k_{Eratio} . These ratios could then be compared with analogous scores obtained through the analysis of LCM data. In particular, the rank product of genes across the six datasets, based either on k_{Ratio} or on k_{Eratio} with the highest ratios being assigned lower (top) ranks, provided an overall score for relevance in stromal or epithelial communication. As shown in Figure 50, the LCM rank product and single cell k_{ratio} display a significant negative correlation, meaning that the scores obtained with crossWGCNA in the LCM and single cell datasets are in overall agreement. Indeed, the top and bottom 100 genes in the LCM rank product significantly differ in their single cell k_{ratio} (Figure 50B, D). Despite the correlation being relatively low, it is to note that in the single cell RNA-seq dataset I have investigated the interaction between the epithelium and the fibroblasts, which are just one of the many cell populations comprised in the stroma. Moreover, single cell data, with the high frequency

of dropouts and zero-inflation, are likely to produce spurious correlations that might negatively affect the result. Nevertheless, the two analyses display a good agreement.

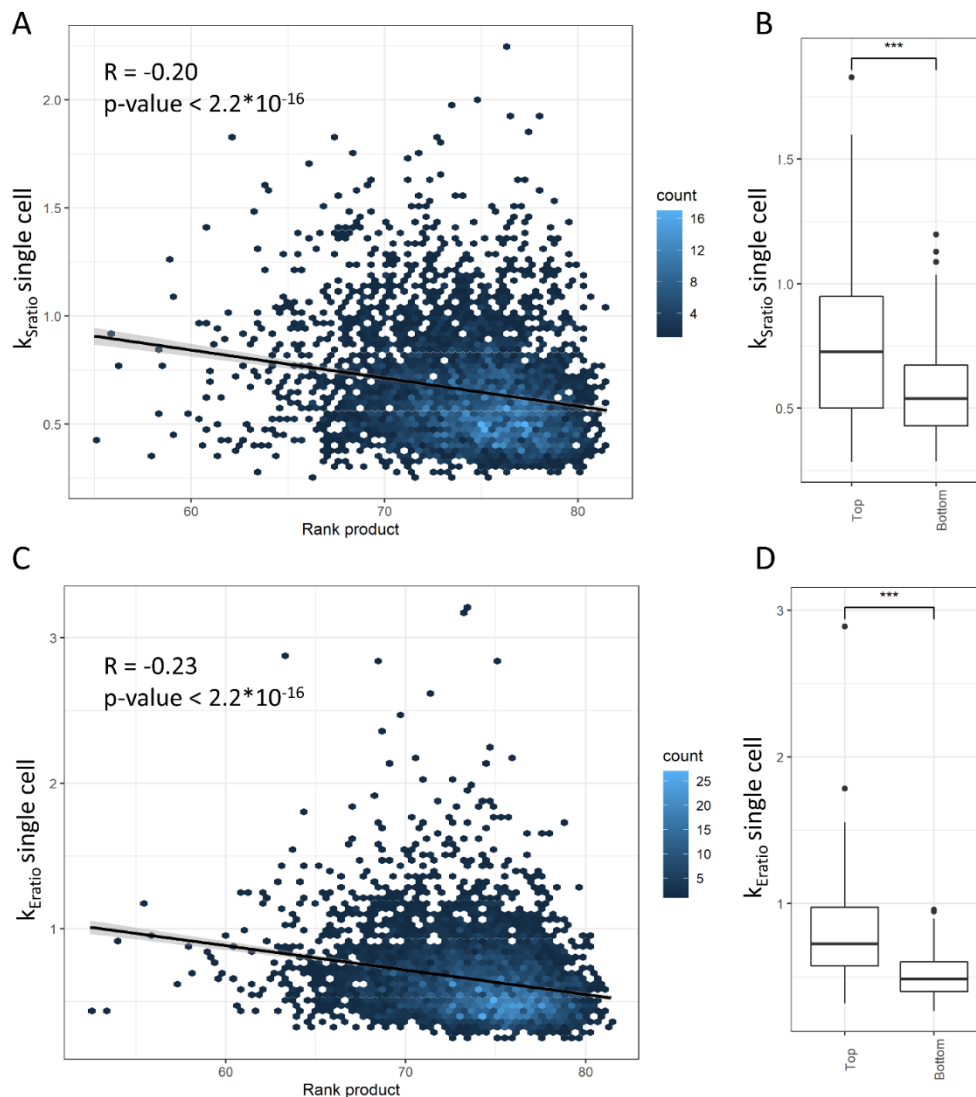


Figure 50. A) Agreement between the LCM and single cell communication scores obtained with crossWGCNA in for stroma/fibroblasts (A, B) and epithelium (C,D). A, C) Relationship between the rank product of genes across the six datasets (x axis), based either on k_{Sratio} or on k_{Eratio} with the highest ratios being assigned lower (top) ranks, and k_{ratio} in the single cell dataset (y axis). B, D) top and bottom 100 genes in the LCM rank product and corresponding single cell k_{ratio}

Looking at the most highly interacting genes in the stroma (as defined through the rank product of k_{Sratio} across the LCM datasets) NOTCH2NLA (Notch 2 N-terminal Like A) is at the top. This gene encodes for a secreted factor involved in Notch signaling, a pathway relevant for cancer progression [255]. Moreover, amongst the top secreted factors we find IL6ST (Interleukin 6 Cytokine Family Signal Transducer), which modulates the response to IL6, an important player in cancer [256].

Intersecting the top 50 interacting genes in the stroma or in the epithelium (LCM based) and the corresponding top 50 genes defined through single cell RNA-seq, the secreted factors CP and

DEFB1 (Defensin Beta 1) appear to be relevant for the interactions in the stroma and in the epithelium, respectively.

Ceruloplasmin (CP) is a multicopper oxidase involved in the acute phase reaction to stress, protecting against free radical-mediated cell injury. It is mainly synthesized in the liver, but it is also elevated in the serum of cancer patients and in malignant tumor cells [257–259]. It is highly expressed in immune cells, particularly in natural killers [260]. Human beta-defensins (DEFB) are produced mainly by epithelia, where they exert an antimicrobial activity [261]. In addition, they mediate the cross-talk with the adaptive immune response by acting as chemokines [262,263]. DEFB1 has been proposed as a tumor suppressor since it promotes cancer cells apoptosis and it is markedly down-regulated in many cancers, despite contradictory results showing its up- or down-regulation in different contexts (Reviewed in Ghosh *et al.* [264]). It has been proposed that it may function as a chemoattractant to immature dendritic cells and T cells, inducing caspase-mediated apoptosis [265].

In addition to identifying genes with a high global inter-compartment communication score, crossWGCNA permits to analyze genes highly correlated with a specific pathway. In this setting, previous knowledge about the relationship between genes, being activators or inhibitors of the pathway, can be included to obtain a more precise and interpretable metric. As an example, I looked for the genes expressed in the stroma most strongly correlated with 10 known oncogenic pathways [266]: cell cycle, Hippo, Myc, Notch, Nrf2, Pi3k, Tgf β , RTK/Ras, p53 and Wnt. For each of these, I obtained the top 5 stromal genes potentially positively influencing their activity in the epithelium or, viceversa, most strongly positively influenced by these pathways being activated in the epithelium (Table 13).

Table 13

Cell cycle	Hippo	Myc	Notch	Nrf2	Pi3k	Tgf β	RTK/Ras	p53	Wnt
MIA	EDN1	NMB	CHI3L2	NDFIP1	INHBB	AGR2	ALCAM	GPC4	MIA
NMU	NMB	TULP3	PSMD3	CLU	MAPT	PON3	PSMD3	CLEC11A	TULP3
KLK7	KLK7	MMP3	SFRP1	PSMD1	MCAM	SCG5	MAPT	MFGE8	S100A7
NMB	PTX3	GFOD1	TULP3	SCUBE2	PSMD3	CXCL1	LAMA3	PMCH	CCL11
CHI3L1	MEGF6	NOTCH4	MFGE8	PYCARD	COL4A5	PTN	SCUBE2	NMB	SFRP1

The top gene for both cell cycle and Wnt signaling pathway is MIA (Melanoma Inhibitory Activity), which has been implicated in cancer invasiveness [267,268]. EDN1 (Endothelin 1) is frequently secreted by many solid tumors, including breast cancer [269,270]. In these tumor cells, EDN1 has been shown to promote cell proliferation, suppress apoptosis, promote metastasis and

angiogenesis [271,272]. Its involvement in the tumor-stroma crosstalk has been investigated, and it is considered one of the soluble factors mediating the bidirectional communication between tumor cells and the microenvironment. In particular, it has been shown that EDN1 released from blood and lymphatic endothelial cells modulates tumor cell behavior, enhancing their motility and plasticity [273]. Similarly, CAF-released EDN1 affects the invasive behavior of tumor cells via a paracrine signaling in oral cancer [274]. Moreover, in line with crossWGCNA inference, EDN1 has been implicated in the activation of Hippo pathway, through the induction of YAP/TAZ target genes [275–278].

Neuromedin B (NMB) acts as an autocrine growth factor and a pro-angiogenic factor. The inhibition of its receptor, NMB receptor (NMB-R), suppresses migration and invasion of the basal like BC cell line MDA-MB-231 [279]. The demonstration of this autocrine loop does not exclude that the microenvironment could contribute through the secretion of NMB.

CHI3L2 (Chitinase 3 Like 2) is produced by tumor-associated macrophages in breast cancer, it attracts the migration of CD14⁺ monocytes and stimulates angiogenesis [280]. Therefore, its correlation with oncogenic pathways in tumor epithelia might be indirect.

INHBB (Inhibin Subunit Beta B) expression has been linked with invasiveness in oral cancer [281], and it has prognostic value in multiple solid tumors, but correlates with better prognosis in breast cancer [282]. AGR2 (Anterior Gradient 2) is a pro-oncogenic secreted from the tumor epithelium [283], exerting paracrine effects on fibroblasts, and inducing endocytosis and activation of beta-catenin [284–286], having therefore being already implicated in tumor-stroma crosstalk.

Finally, I performed a similar analysis to identify genes influenced by or influencing the STAT3-induced gene expression signature that we have recently described in murine breast CAFs (Avalle et al., under revision). The top gene in the epithelium is SLPI (Secretory Leukocyte Peptidase Inhibitor) a secreted inhibitor, which protects epithelial tissues from serine proteases. It is expressed at mucosal surfaces, mainly by epithelial cells [287], and it promotes invasiveness in lung [288], gastric [289] and pancreatic tumor [290]. Moreover, it is expressed in highly metastatic BC cells [291]. Being itself a secreted factor, it is possible that it does not act as an effector of the STAT3 signature in CAFs, but instead as signal from the tumor epithelium inducing the signature in the tumor fibroblasts. Only additional experimental data would help distinguishing these two scenarios.

3. Discussion

Co-expression networks provide a description of gene expression profiles that can be useful to identify sets of functionally related genes. Among the principal aims of network analysis is moreover that of identifying potential regulators of the networks or tightly connected genes (modules) and, consequently, important nodes whose alteration has the potential to disrupt the whole network. Indeed, the number of connections of a gene within a gene expression network was shown to be predictive of its phenotypic relevance in *S. cerevisiae* [22].

In this Thesis, I applied the co-expression network tool to the study of breast cancer molecular mechanisms. Building upon a well clinically annotated transcriptome dataset of ~2000 primary breast tumors (METABRIC,[104]) I applied the method WGCNA (Weighted Gene Co-expression Network Analysis, [36]) and defined modules of tightly co-regulated genes which are preserved in independent transcriptomic and proteomic datasets of primary BC, and which are also confirmed with the STRING protein-protein interaction database [19].

This analysis led to the definition of clinically relevant gene modules, such as the E2F_targets and KRAS_dn1 modules, which correlate with BC high grade, early onset, and poor prognosis. Centrally located within these modules are well-known genes involved in BC progression and metastasis such as FOXM1 and FOXC1 [119,120], but also novel uncharacterized genes like ROPN1, involved in sperm motility and whose role in BC has not been studied despite its being associated with the TNBC marker SOX10 [121–123], one of the most central nodes in the same module.

Striking was the observation that the KRAS_dn1 module is divided in two subnetworks, with the PSAT1 gene linking the subnetwork containing the most central hubs (FOXC1, ROPN1) to the subnetwork enriched in poor prognosis-related genes. PSAT1 is the second enzyme in the serine biosynthesis pathway, important for the generation of many cell building blocks [292–294]. Several enzymes involved in serine biosynthesis, such as SHMT2 (Serine Hydroxymethyltransferase 2) and PHGDH (Phosphoglycerate Dehydrogenase), have been correlated with negative prognosis in BC [292], and PSAT1 itself has been shown to induce proliferation in basal-like BC cell lines [295,296]. Interestingly, the PSAT1 gene appears to be regulated by FOXC1, a central hub of the same KRAS_dn1 module and associated with BC development (GSE73234, Han et al., 2017). Our data suggest that the correlation of PSAT1 with survival in BC may at least partly be due to its central role in shaping the KRAS_dn1 module.

Additionally, this approach permitted identifying gene modules distinguishing tumors metastasizing to different anatomical sites, or more likely to develop resistance to therapies (e.g., anthracyclines).

Breast cancer metastases most frequently localize to bone and lung (Reviewed in [297]), but this frequency varies across molecular subtypes. Indeed, lung relapses were found to be most abundant among the basal-like subtype, whereas bone relapses were found to be primarily associated with the luminal subtypes [298,299].

In this Thesis I have shown that the basal-like over-expressed modules E2F_targets and KRAS_dn1 are associated with less metastases to the bone, while the luminal over-expressed Estrogen_response module is associated with more metastases to the bone, in line with known differences between molecular subtypes. Nevertheless, in the same analysis, which also includes brain metastases, it appears that basal-like tumors differ in the metastatic capacity to the bone, while propensity in metastasizing to lung and brain are overall similar. Therefore, previous and current results can be more correctly interpreted as a difference in the frequency of bone metastases across BC subtypes. Despite a relationship between the metastasis site and the tumor subtype, the different propensity in giving metastases to the brain is not entirely driven by the subtype, and the expression of the E2F_targets, KRAS_dn1 and Estrogen_response modules provide significant additional information in defining the likelihood of metastasis formation to bone, lung, or brain.

To test whether modules could inform on drug effectiveness, I analyzed the ability of co-expression modules to explain the responsiveness of tumors to a well-studied and frequently employed class of therapeutic drugs: the anthracyclines. These compounds intercalate with the DNA and interfere with DNA and RNA synthesis, especially in highly cycling cells, and they have been shown to inhibit the topoisomerase II [300]. The E2F_targets and NC4 modules are the most significantly and consistently correlated with response to anthracyclines across 8 datasets primary BC and cell lines datasets. Specifically, higher expression of the E2F_targets module correlates with better response to the treatment, while the opposite is true for the module NC4. Importantly, this result has been confirmed with data from drug sensitivity screens. Indeed, the gene coding for the topoisomerase II (TOP2) is among the E2F_targets' genes (ranking 94th based on the kWithin). Hence, higher levels of the E2F_targets module correspond to higher levels of anthracyclines' target, which is available to be stabilized on the DNA, thus inducing double-strand breaks.

NC4, despite being recurrently identified as one of the modules most clearly correlated with clinical features, does not display any significant enrichment for MSigDB categories nor Gene Ontology categories. It comprises 89 genes, thus being one of the smallest modules of this BC network. CYBRD1 (Cytochrome B Reductase 1) is the top central gene within this module. It encodes for the Duodenal cytochrome B (Dcytb) also known as cytochrome b reductase 1, which catalyzes the reduction of Fe^{3+} to Fe^{2+} required for dietary iron absorption in the duodenum of mammals [301]. In breast cancer, DCYTB has been shown to correlate with longer survival, but its mechanism of action seems not to depend on the reductase activity, since in BC it does not affect iron intracellular concentration, while it inhibits focal adhesion activation [302]. IGIP (IgA-Inducing Protein Homolog), the second most central gene, is known to enhance IgA secretion from B-cells. The third hub CX3CL1, chemokine (C-X3-C motif) ligand 1, correlates with the number of tumor infiltrating lymphocytes (TILs), including cytotoxic T cells and NK cells. A higher number of TILs is considered an indicator of better clinical prognosis [303]. Moreover, it has been shown that dendritic cells over-expressing CX3CL1 injected intratumorally decrease the tumor growth due to an increase of intratumoral infiltration of CD4^{+} and CD8^{+} T cells and improved overall survival in mouse models [304].

Overall, the function of the top hubs in the NC4 module points to its relevance in regulating the crosstalk between different cells of the microenvironment. In line with this idea is the observation of its positive correlation with stromal content and negative correlation with immune content.

This initial descriptive analysis reveals its potential in understanding breast cancer molecular features and sets the basis for the development of predictive models for clinical progression, site of metastasis, and resistance to therapy, based on gene expression. As a proof of concept, modules' expression proved to be an appropriate predictor of patients' survival both across all patients and in subtype-specific contexts. Many expression-based models have been defined to predict breast cancer patients' survival [305,306], but less attention has been dedicated to subtype-specific models [307,308]. The context-specific co-expression networks here defined could serve as a tool to select the genes most likely relevant for a BC molecular subtype, and hence suitable for the use in predictive models. The models here provided show the feasibility of this approach, despite leaving wide room for improvement, given the relatively low resulting concordance index.

As mentioned above, I additionally dissected the differences in network connectivity across different breast cancer subtypes, comparing the networks obtained with samples of each subtype separately. This analysis is of particular importance when considering that, especially in multicellular

organisms, gene networks are context dependent. Thus, pooling together samples belonging to different groups could result in misleading inferences or hide biologically meaningful signals. I focused on the features distinguishing basal-like networks from networks derived from other subtypes, aiming at identifying molecular mechanisms conferring basal-like BC its aggressiveness and resistance to therapy. Specifically, I defined three modules with stronger intramodular connectivity in the basal-like subtype, with one, called “b_E2F_targets”, being also overexpressed in basal-like tumors and correlated with higher grade and worse prognosis.

Dissecting regulatory mechanisms of clinically relevant co-expression modules could be helpful in gaining insights about the molecular mechanisms behind their functioning, but also in identifying additional actionable targets. For this reason, I overlaid modules' definition with microRNA expression of matched mRNA-miRNA profiling of five independent datasets, and microRNAs' predicted targets, selecting miR-29c-3p, miR-20a-5p, miR-93-5p, and miR-106b-5p as potential regulators of good prognosis-related modules. Particularly interesting is miR-106b-5p, which has been shown to promote tumorigenesis and metastasis[309–311] and to be coordinated with mir-93 [312]. Hence, the analysis here provided could provide insights about the convergent mechanisms through which these two microRNAs induce BC aggressive features, identifying the genes of the NC4 modules as part of their targets.

Crucial in this work has been the demonstration that network centrality (kWithin) is a suitable measure to identify relevant genes in this biological system, a pivotal observation for gene prioritization. Indeed, I showed that kWithin correlates with clinical features such as disease-free survival, gene essentiality for cancer cell lines' survival and, most importantly, that it provides a mean to select potential upstream regulators of a module with high reliability. Therefore, I combined network measures with the results of independent experiments analyzing gene expression changes induced by altering the expression of selected transcription factor (TF) hubs. Strikingly, kWithin-based hubs identification and analysis of published gene expression data following their experimental modulation in BC cell lines showed that the most central TFs act indeed as upstream regulators of the module's genes. This observation represents an empirical validation of the reliability of the kWithin as a selection criterion for hubs that have a high potential of being central regulators of the network structure, and therefore possessing phenotypic relevance. Importantly, among the most central TFs in the b_E2F_targets module are also poor prognosis-related genes for which no experimental studies in BC are available (*preprint deposited on BioRxiv, March 2019* [313]). I selected PTTG1, TEAD4, E2F3, TFDP1, and CEBPG. While PTTG1, TEAD4, E2F3,

TFDP1 have been studied to different degrees in BC, C/EBPG was mainly characterized in myeloid tumors [314–318]. These may represent interesting candidates for phenotypic studies, whose disruption may lead to the dysregulation of the whole basal-like specific module b_E2F_targets, and thus to the inhibition of aggressiveness features of basal-like BC. Experimental validations performed in the laboratory suggest that an alternative model should be considered. Indeed, it results that hubs' silencing induces a gene expression perturbation specifically in the genes that are closer to the hub in the co-expression network, but without any specific association with the sign of their relationship (either positive or negative).

Moreover, through this network-based selection of relevant genes, it was possible to prioritize drugs for repositioning, integrating modules with available resources, such as ChEMBL, annotating drug-target relationships. The compound with the strongest evidence for a regulatory activity on the b_E2F_target module is the ellagic acid, an organic heterotetracyclic compound found in red fruits, with antioxidant activity but currently not in use for BC. It has been employed in clinical trials investigating its effectiveness in treating or preventing the metabolic syndrome, HPV infection, childhood obesity, cardiovascular disease, myelodysplastic syndrome, follicular lymphoma, lung carcinoma, and colorectal cancer (<https://clinicaltrials.gov/>). The one clinical trial on BC patients is investigating its bioavailability in breast tissue.

I applied the same procedure of centrality-based ranking to define potential targets of specific pathways and searched for potential common targets of the STAT3 and WNT/PCP pathways. I found evidence of the involvement of the EMT module in STAT3 and WNT/PCP signaling, since this module is significantly enriched in i) both STAT3 and non-canonical WNT signatures; ii) genes whose promoters are bound by STAT3 in several BC cell lines, and iii) genes down-regulated upon STAT3 knock-down in the basal-like cell line MDA-MB-231, and up-regulated upon ROR2 overexpression in the luminal MCF7 cell line. Within the EMT module, the ZNF532 gene, encoding for a TF correlated with poor prognosis in BC, appears to be a target of both pathways. It can therefore be hypothesized that ZNF532 is a novel common target of the two pathways, and likely a central regulator of the EMT module. This hypothesis can be readily assessed experimentally and may lead to the identification of a novel target with therapeutic potential, particularly in basal-like BC.

Breast cancer heterogeneity is hardly captured in the small sample sizes of most microarray studies, but the strong research effort dedicated to this biological system has led to a rich collection of independent datasets that can be combined to improve robustness and statistical power. Therefore, all the analyses performed in this Thesis were confirmed in independent datasets to

assess their robustness. Moreover, to enable robust findings, I collected available transcriptomic data from the Gene Expression Omnibus database gathering 55 datasets of primary BC for a total of 18362 patients, generating a rich and manually curated resource with carefully harmonized clinical data. The analysis of this resource confirmed that the increased connectivity of the b_E2F_targets module in the basal-like subtype is a robust result, shared across most of the available datasets, and that additional previously undetected modules could be more tightly connected in the basal-like subtype. Importantly, the comparison of multiple datasets contributed in defining which variables (either biological or technical, such as the microarray platform) could influence the observations, reducing misleading interpretations.

Breast cancer is a heterogeneous disease with several cellular components playing specific roles in its development and clinical course. In particular, the microenvironment has been shown to either counteract or promote tumor progression depending on the specific conditions, and important players such as immune cells or cancer associated fibroblasts are the objects of intense study. Importantly, the analysis of bulk tumor samples comprising cell admixtures complicates disentangling the specific behaviors of each cell component. Laser capture microdissection can help in separating the contribution of different cell compartments and can still present some relevant advantages with respect to single cell techniques. Indeed, cell type separation does not rely on set of markers, but it is directly based on histological features. Moreover, the relatively contained cost permits sampling ranges of tumors with different characteristics and the assessment of the relationships between gene expression and clinical features.

As an additional complication, breast cancer heterogeneity is hardly captured in the small sample sizes of most microarray studies, but the strong research effort dedicated to this biological system has led to a rich collection of independent datasets that can be combined to improve robustness and statistical power.

The relevance of the microenvironment in shaping BC gene expression is evident also in this work: co-expression modules correlate with stromal percentage (either fibroblast or immune cells), again showing how gene regulatory processes are linked with the microenvironment. Indeed, gene co-expression and differential co-expression often result in gene modules enriched for immune-related processes or extracellular matrix shapers [319]. Hence, it would be extremely interesting to investigate what cell type-specific co-expression networks could tell about cell type-specific regulatory processes in tumors. Such studies have not yet been performed due to the limited amount of data. Nevertheless, some laser capture microdissection datasets separating stroma and

epithelium comprise enough samples to build such networks, despite not permitting to dissect the differences between subtypes, which has been a central part of this Thesis. On the other side, single cell datasets would permit to compare networks across cell types and tumor subtypes even with a few subjects, but co-expression in the sparse single cell RNA-seq data needs to rely on more specific metrics than correlation [320].

To study BC microenvironment, revealing gene expression changes specific of tumor stroma or epithelium, I gathered 48 transcriptomic datasets of microdissected breast tumors where the stroma and the epithelium were separated prior to RNA extraction to study the distinct behavior of different cellular compartments. I carefully collected and harmonized corresponding biological and clinical annotations to facilitate data integration.

With this tool in hand, I identified genes robustly and consistently changing their expression in breast tumors when compared to normal tissue, either in the stroma or in the epithelium. Analyzing these lists separately, I detected increased expression of cell cycle related genes in the tumor epithelium and of immune-related categories in the tumor stroma. Additionally, I detected the over-expression of the non-canonical Wnt/PCP pathway, which is involved in breast cancer progression [138], synergizing with the STAT3 pathway, contributing to its aggressiveness [141] in both the stroma and the epithelium. Moreover, I observed a decrease in lipid catabolism in the tumor stroma, consistent with a potential metabolic coupling between cancer cells and the microenvironment, with stromal components reducing their consumption to release lipids and feeding cancer cell growth [321].

Comparing differentially expressed genes in the two compartments, genes that are up-regulated in both are enriched in mitochondrial-related and extracellular matrix gene ontology categories. Indeed, the extracellular matrix can act as a reservoir of growth factors, and its remodeling has been associated with metastatic spread [322]. The role of high oxidative phosphorylation in tumors is being increasingly recognized [323,324]. Counter-intuitively, down-regulated genes are enriched for angiogenesis. A similar trend had already been observed [325] and could be explained by the higher resistance of tumor cells to apoptosis under hypoxic conditions, especially in advanced tumors that also show lower microvessel densities than normal tissues [326].

Moreover, genes that are significantly regulated in only one compartment can exhibit a corresponding compartment-specific relationship with clinical features. For example, genes that are only up-regulated in tumor stroma correlate with a higher grade and earlier age at onset when measured directly in the stroma, while their relationship with clinical features is weaker when their

expression is measured in the epithelium, highlighting the specific role of the stromal tissue. At the top of the list were UPB1 and HSD11B2. The latter is the enzyme that converts cortisol in cortisone, which is down-regulated in tumor stroma and correlates with a lower grade. Its decrease across tumor progression might be responsible for high cortisol levels, which have been associated with higher severity and mortality [327]. UPB1 encodes for the last enzyme in the pyrimidine degradation pathway, and its down-regulation might lead to dihydropyrimidine accumulation, linked with EMT [328]. Similarly, NECAP2 is among the top up-regulated genes in tumor stroma only, correlating with a higher grade. NECAP2 is involved in endocytic recycling [329], suggesting a potential role in regulating surface protein localization and cell-cell communication. Of note, the compartment-specific regulation of a number of genes exhibiting opposite regulation in tumor epithelium or stroma would have been missed based on bulk data: genes that are up-regulated in tumor stroma and down-regulated in tumor epithelium are enriched for cytokine secretion and Toll-like receptor 2 signaling, suggesting that the microenvironment produces different sets of cytokines than the tumor itself, possibly reflecting different chemoattraction and hence different immune cell proportions depending on tumor proximity. Indeed, immune cell distribution has been found to be of clinical relevance in cancer [330–332].

To the best of my knowledge, this is the first meta-analysis performed specifically on LCM transcriptomic data. I show how different platforms can be successfully integrated to reveal robust differential expression patterns and to increase statistical power, identifying differentially expressed genes that would not be identified otherwise. It is to note, however, that some of our results need to be cautiously interpreted given that the definition of non-differential genes cannot be given in a statistically rigorous way. Nevertheless, this caveat only applies to a limited part of our work, and the lists of DEGs in tumor epithelium, stroma, and vessels that I provide are indeed robust.

From the analysis of this database, it can be concluded that although the behavior of the epithelium and the stroma at the gene expression level is globally similar, separating compartments permits the identification of gene regulation patterns that could not be detected in bulk. Moreover, integrating many different datasets allowed us to improve statistical power and, despite the small sample size of each dataset, to identify the genes that are correlated with clinical features in a compartment-specific manner.

I showed additional use-cases of the database, such as the selection of epithelial and stromal markers that improves the correlation with cellularity compared to published signatures. Indeed, this refined signature could prove to be profitable in estimating the tumor/stroma composition of

bulk tumors through deconvolution methods. Moreover, the analysis of STAT3 signaling pathway regulation revealed a particularly relevant role in tumor stroma compared to the epithelial counterpart, despite STAT3 being considered as an oncogene in many tumor types, including breast cancer, and was consistent with data showing an opposite role in the two compartments in colorectal cancer. Specific questions can be addressed by performing the wide variety of comparisons permitted by the conditions represented in our database. A particularly interesting application will be the construction of compartment-specific gene regulatory networks. Indeed, cancer gene networks built from bulk transcriptome data are affected by the presence of multiple cell types and often include microenvironment-related gene sub-networks, confounding the identification of cancer cell gene interactions [55]. Therefore, the use of the LCM data collected here will be a valuable resource to build more specific and robust gene co-expression networks.

An important practical corollary of this work is the availability of the database as a resource for other researchers to explore via a simple web platform that includes differential gene expression and enrichment analyses (<https://aurorasavino.shinyapps.io/metalcml/>).

Lastly, to study tumor-stroma crosstalk, I defined a novel method, which I call “crossWGCNA”, adapting the WGCNA algorithm with an alternative measure of distance between nodes (genes) which requires matched samples from the same subjects and takes the tissue of origin into account.

The assumptions on which crossWGCNA is based differ with respect to previous methods, permitting to investigate interactions unbiasedly (without limiting them to ligand-receptor pairs). Moreover, crossWGCNA can include information about downstream signaling pathways influencing a cell’s response to an interaction, and the sign of the relationship between various nodes (activations or inhibitions). Finally, it provides as output not only lists of genes scoring high for their potential relevance in the interaction, but also clusters of highly connected genes, which are more robust and interpretable than single genes.

The major downside of this method is implicit in any method employing correlation: the correlation is an un-directional, non-causal measure of relationship, and hence correlations driven by other external factors than the actual interaction between tissues can’t be excluded. To limit this issue, I added a self-loop removal step in the pipeline, eliminating the correlations most likely driven by genetic or environmental factors shared by all tissues in a subject.

After assessing the suitability of this method to select genes with high potential for inter-tissue crosstalk, I applied it to the collected LCM datasets and identified several secreted factors

with high interaction scores, such as NOTCH2NLA and IL6ST, involved in Notch signaling and in the response to IL6, both signaling pathways already implied in cancer [255,256].

Moreover, in addition to identifying genes with a high global inter-compartment communication score, crossWGCNA permits to analyze genes highly correlated with a specific pathway, including the previous knowledge about the relationship between genes (activators or inhibitors of the pathway). Genes in the stroma being correlated with the activation of oncogenic pathways in the epithelium comprise MIA (Melanoma Inhibitory Activity), implicated in cancer invasiveness [267,268]; EDN1, frequently secreted by many solid tumors, including breast cancer [269,270] and promoting cell proliferation metastasis and angiogenesis [271,272]; and the pro-angiogenic factor Neuromedin B (NMB). For some of these genes, the involvement in the tumor-stroma crosstalk has been investigated. For example, EDN1 released from lymphatic endothelial cells or CAFs enhances tumor cell motility and plasticity [273,274].

In conclusion, using gene co-expression networks I identified groups of genes with a high potential to play a crucial role in basal-like BC and possibly amenable to drug targeting, not only for their specific functions but also for their ability to regulate survival-related gene networks. Additionally, I proposed an extension of the WGCNA method to study tumor-microenvironment interactions at the gene expression level. Through the systematic collection of available transcriptomic data, I built robust inferences about gene expression changes in breast cancer and provided useful resources for the research community. These approaches could be successfully extended to other BC subtypes, and potentially to all tumor types for which enough expression data are available, contributing to addressing many additional biological questions.

Experimental validations of the presented results will allow testing for the reliability of these inferences. For example, having shown the functional relevance of modules' hubs, the same principle could be exploited to select central genes within the NC4 and E2F_targets modules that could serve as pivots for restraining drug resistance. Silencing the hubs of these modules in cancer cells could influence anthracyclines' efficacy, and hence inform on promising candidates for developing novel drug combinations. Similar considerations apply to the crossWGCNA method, which needs experimental validations to assess its validity in dissecting tumor-stroma crosstalk. Silencing CP in cancer associated fibroblasts and then measuring whether their influence on epithelial cells' transcriptomes is affected will provide a first indication on CP relevance for tumor-stroma communication, and therefore give confidence to the method.

Some of the reported analyses could provide useful resources to the scientific community if delivered in easily accessible formats. Therefore, an important outlook will be the development of web resources for accessing breast cancer co-expression networks and the complete list of harmonized primary breast tumor transcriptomic datasets. Similarly, the crossWGCNA method will be implemented in an R package deposited in CRAN or Bioconductor.

Finally, this work opens new perspectives on regulatory principles governing co-expression networks. How gene interactions react upon perturbation is a still unsolved but relevant question, since its solution would help predicting therapeutic responses and developing drug combinations to prevent resistance. Nevertheless, the simplest model of hub silencing leading to the whole module silencing proved here not to be verified, suggesting more complex relationships that will be the ground for future investigations.

4. Methods

4.1. Plots and Statistical Analyses

All statistical analyses were performed with R 4.0.4 [333].

Packages used for plotting are R base graphics, ggplot2 [334], ggsignif (<https://CRAN.R-project.org/package=ggsignif> [335]), ggvenn (<https://CRAN.R-project.org/package=ggvenn> [336]), survminer (<https://CRAN.R-project.org/package=survminer> [337]), ggpubr [338], and pheatmap (<https://CRAN.R-project.org/package=pheatmap> [339]).

4.2. Networks in BC

All analyses were performed using R 3.5.1 and packages obtained from Bioconductor and CRAN [325]. Plots were generated with R base graphics, ggplot2 [334], WGCNA [36] and pheatmap.

4.2.1. Data download and pre-processing

GEO datasets were downloaded using the GEOquery package [341]. Normalized (FPKM) breast cancer data from the TCGA project and corresponding clinical annotations were obtained through TCGA biolinks [342], and METABRIC gene expression data and metadata were obtained from www.synapse.org (syn2133318, syn2133322, syn2133500). Probe names were converted in Gene Symbols and, for gene symbols corresponding to multiple probes, the most expressed probe across all samples was considered. ENSEMBL IDs were converted in Gene Symbols using biomaRt [343].

4.2.2. Network construction

Gene co-expression networks were constructed with WGCNA function blockwiseModules [36], using as parameters: power= 6, corType="pearson", networkType = "signed", minModuleSize = 30, reassignThreshold = 0, mergeCutHeight = 0.25.

4.2.3. Network Robustness assessment

Preservation was obtained with the function modulePreservation in WGCNA using, independently, breast primary tumors data (from the packages breastCancerMAINZ, breastCancerTRANSBIG, breastCancerUPP, breastCancerUNT, breastCancerNKI, breastCancerVDX [344–348]) and BC cell lines data (<http://cancergenome.nih.gov/>, [105]). To assess preservation of Basal-like modules, only Basal-like samples were kept, using the function PAM50pred from the

genefu package [349] to assign the molecular subtype. For plotting, a maximum cutoff of 10 for $\log_{10}(\text{z-scores})$ was employed.

Preservation in proteomic data was obtained calculating the correlations between proteins encoded by the gene of a module and comparing them with random proteins (Kolmogorov-Smirnov test, one sided, with the `ks.test` function [350]). Datasets used: PXD002619, PXD000815, TCGA-BRCA-L4. The STRING database was mined through the package STRINGdb [19].

4.2.4. Module eigengenes

Module eigengenes calculated on the whole dataset [36] were used as a measure modules' expression in each sample. For additional datasets, the corresponding module eigengenes were calculated as the samples' projection on the first principal component in a PCA with the genes of the corresponding module and all METABRIC samples.

4.2.5. Functional enrichment

Functional enrichment was performed using clusterProfiler [99], msigdb [106] and fgsea [351] and considering only categories with an adjusted p-value less than 0.05.

4.2.6. Correlation with clinical features

Pearson's correlation between module eigengene and either grade or age at diagnosis was calculated and p-values were obtained with the function `corPvalueFisher` (Fisher's asymptotic p-value). P-values were adjusted using the Holm method in the function `p.adjust`. Differences in gene or modules' expression across subtypes were assessed via a two-sided t-test on z-scores (for individual genes) or on module eigengenes (for modules). Expression in one subtype was compared with the expression in all other subtypes pooled together.

The correlation between the expression level of a gene or a module was calculated dividing the patients in two groups based on the gene or module's expression median. The function `survdif` of the survival package [352], implementing a log-rank statistic, was used and Kaplan-Meier curves were plotted with the `ggsurvplot` function in the `survminer` package.

4.2.7. Metastases

Annotation for site of relapse for the METABRIC dataset was obtained from EGAF00000245831. Additional datasets employed: GSE12276, GSE2034, GSE2603, GSE5327 [108–111]. The analysis has been limited to samples developing metastases to the lung, bone, or brain.

A Wilcoxon ranked sum test has been employed to compare the module eigengenes of tumors developing or not metastases to one specific site. P-values were then corrected for multiple testing with the `p.adjust` function (`fdr`).

Significant relationships were finally assessed in the additional datasets, calculating corresponding module eigengenes as the samples' projection on the first principal component in a PCA with the genes of the corresponding module and all METABRIC samples. Linear models with the module eigengenes and the dataset as independent variables, and the development of metastases to a specific anatomical site as dependent variable, were fitted and coefficients' p-values and signs taken into for conclusions.

4.2.8. Resistance

Datasets comparing primary tumors with/without complete clinical response to anthracyclines or comparing cell lines before and after a chronic treatment with these drugs and their consequent acquisition of resistance: GSE87455 [353], GSE22226 [354], GSE16446 [355], GSE25055 [356], GSE54326 [357], GSE162187 [358], GSE47013 [359], GSE155478 [360]. For each dataset, I compared each module's expression, quantified through the projected module eigengene, between sensitive and resistant samples with a Wilcoxon Signed Rank test, obtaining a p-value for higher expression, and a corresponding p-value for lower expression in resistant samples of each module. To have an overall measure of modules' association with resistance across all datasets, I collapsed the p-values with the Fisher's combined probability test [98], obtaining a single p-value for each module's association with resistance, and a corresponding p-value for each module's association with sensitivity.

Cell lines' gene expression profiles were obtained from the Cancer Cell Line Encyclopedia [361] and the sensitivity to doxorubicin from the GDSC1 drug sensitivity screening [112] (downloaded from www.cancerxgene.org, version of the 25th of February 2020).

TCGA samples annotated for patient's treatment with anthracyclines ("doxorubicine", "DOXORUBICIN", "doxorubicin", "Doxorubicin", "Adriamycin", "ADRIAMYCIN", "adriamycin", "Adriamycin", "adriamicin") were retrieved, and samples annotated as "complete response" or "clinical progressive disease" were compared with a Wilcoxon ranked sum test.

4.2.9. Stromal percentage

TCGA tumor purity data have been obtained from:
<http://genboree.org/theCommons/documents/569> (EDec),

<https://bioinformatics.mdanderson.org/estimate/disease.html> (Stromal scores), and the TCGAbiolinks R package (ESTIMATE, ABSOLUTE, IHC). TCGA samples corresponding to have been selected for the analyses and matched with the corresponding stromal scores.

In the additional datasets, obtained from MetaGxBreast and pre-processed as described in Paragraph 4.2.1, tumor purity was calculated with single sample GSEA (ssGSEA) (package GSVA [362]) using gene expression profiles of 141 stromal and 141 immune genes, as previously described [113].

Signatures of stroma or epithelium were derived as DEG lists from the metanalysis of LCM BC datasets, as described in Paragraph 2.2.4, and a Fisher exact test was used to assess their enrichment for modules' genes.

4.2.10. Predictive models

Cox models were employed to predict 5 years disease specific survival in a 5-fold nested cross validation (glmnet, family="cox" [363]). Specifically, the samples were randomly divided in 5 groups of equal size, and one kept as test set. The remaining 4 sets, used as training set, were additionally divided in 5 groups of equal size (4 for the nested training set, one as a validation set). The model was trained on the nested training set and 10000 different values for the parameter lambda (equally distributed between 0.0001 and 0.1). The lambda giving the highest concordance index on the merge of all validation sets was used to fit the model on the whole training set. The model was finally evaluated on the test set with the concordance.index function from the survcomp package [364].

Either module eigengenes, the projection of samples on the top 5 hubs' coefficients of module eigengenes or the expression of the top 1 hub were employed as predictive variables.

4.2.11. CRISPR screenings

Data from three genome-wide screenings for gene essentiality in cancer were obtained from the studies Behan *et al.* [128], Tscherniak *et al.* [129], McDonald *et al.* [131]. Data from Behan *et al.* were downloaded from the supplementary files of the original publication, while the other two datasets were downloaded from <https://depmap.org/>, DEMETER D2.

For each combination of dataset, basal-like cell line and module, the correlation between genes' centrality and the logFC of corresponding CRISPR guides or siRNAs detected in the final pool of cells was calculated. The significance was obtained with the function cor.test, and only correlations with $p < 0.05$ were shown.

4.2.12. MicroRNA

Data: METABRIC (EGAF00000245831) and GSE22220, GSE28884, GSE59595, GSE81002 [154–158], with matched miRNA and mRNA profiling. MiRNA names were converted to miRbase version 22 with miRNAVersionConvert function. Only microRNA in common to all datasets were retained.

Pearson correlation between miRNA expression and modules' eigengenes and corresponding p-value were obtained, and only microRNAs significantly ($p < 0.05$) negatively correlated with the same module in all 5 datasets were kept for further analyses. For each microRNA, the predicted targets were obtained from several databases (DIANA, elmmo, microcosm, miranda, mirdb, pictar, pita) using the package multiMiR [159] and filtered for their presence in at least 2 of such databases.

The enrichment between the list of targets and module's genes (Fisher test) was performed only for the miRNA-module pairs selected in the previous step. P-values were finally corrected for multiple testing to obtain FDR.

4.2.13. STAT3 and WNT signatures

A set of STAT3 and WNT signatures was collected [142–144,146,148–151,365,366] and individual enrichments of modules for each signature were calculated with a Fisher's exact test using as background the list of genes not falling in the "unconnected" compartment. P-values were adjusted with the p.adjust function and the Holm method considering all the tests performed. Differentially expressed genes upon STAT3 down-regulation in MDA-MB-231 were defined from [145] [152] using DESeq2 [367] and a p-value cutoff of 0.05. Differentially expressed genes upon ROR2 overexpression in MCF7 were obtained from [140] [147] with $p\text{-value} < 0.05$. The enrichment of modules for genes in the intersection of these lists of differentially expressed genes was calculated with a Fisher's exact test.

4.2.14. ChIP-seq data analysis

ChIP-seq data [152] were mapped to the *H. sapiens* hg19 genome build with bowtie 1.2.1 (parameters: -p 3 -best -strata -m1 -S [368]). Peaks were defined with MACS 1.4.2 [369] and the closest genes to peaks were annotated with the function closestBed (BEDTools suite, [370][354]). Peaks with score ≥ 200 were retained, and those in the range between -1000bp and 0bp from the transcription start site were considered indicative of binding at the promoter. The enrichment for genes in each module was calculated with Fisher's exact test.

4.2.15. Subtype-specific modules

Subtype-specific networks were constructed on basal-like, luminal A, luminal B and HER2-positive tumor samples separately using WGCNA with the same parameters used for the global network. Comparisons of intramodular connectivity (k_{Within}) were performed as follows: k_{Within} were calculated on all combinations of subtype-specific networks and assignments of genes to modules to determine the connectivity of genes of a module, as defined in a subtype, in the other subtypes; distributions of k_{Within} of genes of a module were compared across subtypes using a t-test.

4.2.16. Analysis of TF datasets

Pre-processed gene expression data were downloaded from Gene Expression Omnibus (GSE2222, GSE55204, GSE25741, GSE48928, GSE103242, GSE48979, GSE36939, GSE73234, GSE34817, GSE35525, GSE38893, GSE92281, GSE30405). Lists of differentially expressed genes were obtained with a t-test (function `t.test`) and a p-value cutoff of 0.05. Enrichment of differentially expressed genes for modules' genes were calculated with a Fisher's exact test using as background the list of genes not falling in the "unconnected" compartment in basal-like tumors and which expression could be retrieved based on the specific array used in the dataset.

4.2.17. ChEMBL

ChEMBL (version 24.1) was interrogated using the following SQLite query, to select compounds targeting at least one of the genes in the module (`genes_db`), and at least in phase 2 of clinical trials:

```
"select molecule_dictionary.chembl_id, molecule_dictionary.pref_name, count(distinct
component_synonym), withdrawn_flag, max_phase, group_concat(distinct component_synonym)
as target_list from component_synonyms, target_components, assays, activities,
molecule_dictionary where UPPER(component_synonym) in (select * from genes_db) and
max_phase>1 and component_synonyms.component_id=target_components.component_id and
target_components.tid=assays.tid and assays.assay_id=activities.assay_id and
activities.molregno=molecule_dictionary.molregno group by molecule_dictionary.chembl_id order
by count(distinct component_synonym)"
```

Resulting genes were then filtered using the exact Fisher test for assessing the enrichment of their targets in genes of the module.

4.2.18. In house RNA-seq data analysis

After quality controls (FastQC, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> [371]), data were aligned with hisat2 [372] to the *Homo sapiens* genome assembly (grch38), trimming the first 5 bases at the 5' end, bam files sorted and indexed with samtools, and reads mapping to genes quantified with htseq-count [373] (with default parameters and `-s reverse -r pos`).

Differentially expressed genes between silenced cell lines and respective scramble controls were calculated with DESeq2 [367]. DEGs in common between the two replicates (different shRNAs) for each gene were kept for enrichment analysis. A one-sided Fisher test was employed to calculate the enrichment of DEGs in each module's genes.

4.2.19. Network representations

The most strongly connected nodes (Topological Overlap ≥ 0.02) were retained for individual modules' visualization in Cytoscape 3.7.0 [374]. The whole network was filtered for connections with Topological Overlap ≥ 0.05 and represented with Gephi 0.9.2 [375].

4.3. Metanalyses

Where not otherwise specified, the methods describe the procedure adopted for both LCM and bulk tumor meta-analyses.

4.3.1. Search of Datasets

Transcriptomic datasets of breast tumors analyzed in their stromal compartment were searched on Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) on 20 December 2020, using the search terms "Breast cancer" AND "lcm" or "Breast cancer" AND "stroma" or "Breast cancer" AND "microdissect*" and selecting the study type as "Expression profiling by array" and "Expression profiling by high-throughput sequencing". They were then individually screened to discard datasets not comprising untreated tumor samples. The whole list of datasets comprised in the final database is available in Table 6. Original works describing each dataset can be found in [50–94].

Transcriptomic datasets of bulk breast tumors were searched on Gene Expression Omnibus (GEO, <https://www.ncbi.nlm.nih.gov/geo/>) using the search term "Breast cancer" and selecting the study type as "Expression profiling by array" and "Expression profiling by high-throughput sequencing". They were ranked based on the number of available samples, and then individually

screened to discard datasets not comprising breast tumor samples or datasets already part of the MetaGxBreast database [115]. A provisional list of analyzed datasets is in Table 9.

4.3.2. Data Download and Pre-Processing

GEO datasets were downloaded using the GEOquery package [95]. Normalized (FPKM) breast cancer data from the TCGA data were obtained through TCGA biolinks [96], and METABRIC transcriptome data were obtained from synapse.org (syn2160410). Clinical and biological annotations were obtained using the same methods.

Mapping of the probes to the gene symbols was obtained from the respective platforms' information in GEO for each dataset, and in case of multiple probes mapping to the same gene symbol, the probe with the highest mean expression across the dataset's samples was chosen.

Whenever data were not already log transformed, we applied log2 transformation, adding an offset of 1 when the data minimum value was 0.

Replicates were merged calculating the average of their expression signals before log transformation. In GSE4823, expression values of dye-swap replicates (not log transformed) were inverted before averaging. In GSE8977, to permit variance stabilization, negative values were removed prior to log transformation.

4.3.3. Database Metadata

After defining the biological and clinical annotations to be gathered, Niccolò De Marzo and me independently collected them from GEO. Discrepancies were then individually checked and resolved. Moreover, for datasets GSE14548, GSE16873, GSE20437, GSE21947, GSE22513, GSE26910, GSE33692, GSE35019, GSE38959, GSE5764, and GSE72644, additional clinical annotations were obtained from the tables of the original manuscripts.

To facilitate the comparison of the different datasets, annotations were harmonized as much as possible.

When multiple samples from the same subject were available and were matched to the subject of origin, all clinical annotations were assigned to all samples, including the histologically normal ones.

Details regarding single annotations are provided in Appendix A.

4.3.4. Meta-Analysis Pipeline

The meta-analysis involves the following steps: (1) define the conditions to be compared (e.g., tumor stroma vs. normal stroma); (2) select all datasets in the database comprising samples belonging to both conditions; (3) for each separate dataset, obtain the differentially expressed genes (p-values and logFC); (4) keep only the genes with a probe in all microarray platforms of the selected datasets; (5) collapse the p-values by taking the differential expression sign into account, thus obtaining a p-value for each gene's up-regulation and a p-value for each gene's down-regulation; (6) adjust the resulting p-values for multiple testing. Importantly, as stated in step 3, each dataset was analyzed separately to avoid batch effects.

The procedure for each step is detailed below (Paragraphs 4.3.5, 4.3.7).

4.3.5. Differential Expression

Differential expression was calculated using limma [376] for each separate dataset, and information about the subject was added as a factor when the samples were subject-matched.

4.3.6. Datasets' Comparison (LCM meta-analysis)

The comparison between two different datasets was performed by considering the gene symbols represented in both arrays and computing the Spearman correlation of their logFC. For experiments using mouse systems, human orthologs were obtained from biomaRt [97]. Specifically, in Paragraph 2.2.2, all datasets comprising both cancerous and non-cancerous stromal samples were used, and pairwise correlations between datasets are shown. The datasets used for each class of comparisons are invasive BC vs. normal tissue (GSE10797, GSE33692, GSE35019, GSE8977); invasive BC vs. normal counterpart (GSE14548, GSE26910, GSE35019, GSE83591, GSE90505); DCIS vs. normal tissue (GSE33692, GSE35019); DCIS vs. normal counterpart (GSE14548, GSE35019); in vitro samples derived from invasive BC vs. normal tissue (GSE20086, GSE29270); in vitro samples derived from invasive BC vs. normal counterpart (GSE29270); and in vitro samples derived from carcinomas (not invasive) vs. normal tissue (GSE129189, GSE45256). In Paragraph 2.2.3 the average logFC obtained after comparing tumor vs. normal samples in either the epithelium or in the stroma is shown, with DEGs obtained as described in Paragraph 4.3.5.

4.3.7. Collapsing p-values

Uncorrected p-values for each gene and each dataset, obtained either from differential expression or correlation with clinical features, were collapsed with Fisher's method [98] from the

metap package (<https://CRAN.R-project.org/package=metap>). Two different tests were performed separately for testing consistent up- or down-regulation (positive or negative correlation) and for taking logFC signs (correlation signs) into account. p-values were then cut at 2.2×10^{-16} . Since all tests were two-sided, one-sided p-values were obtained with the two2one function from the metap package before applying Fisher's method. Resulting p-values were corrected for multiple testing with the p.adjust function from R stats with the default Holm method.

4.3.8. Definition of DEG Categories (LCM meta-analysis)

The meta-analysis of differentially expressed genes between invasive BC and normal/normal counterpart tissues, performed separately for epithelial and stromal samples, made use of the following datasets: GSE10797, GSE33692, GSE35019, GSE8977, GSE14548, GSE26910, GSE83591, and GSE90505 for the stroma; and GSE10780, GSE10797, GSE33692, GSE38959, GSE45581, GSE14548, GSE5764, GSE72644, and GSE83591 for the epithelium. The analysis followed the steps described in Paragraph 2.2.3 to obtain lists of DEGs in the two tissues. I collapsed the results obtained by comparing tumor vs. normal tissue and tumor vs. normal counterpart. One dataset, GSE35019, permitted both comparisons (tumor stroma vs. normal stroma, tumor stroma vs. stromal normal counterpart), and to satisfy Fisher's method for the assumption of independence, I only kept the tumor vs. normal comparison for this dataset. Keeping only the genes with a probe in all selected datasets, 9523 genes were analyzed for their differential expression in the stroma, and 10 623 were analyzed in the epithelium.

To define the classes of the DEGs being regulated in both the stroma and the epithelium, or with evidence of differential expression in only one of the two tissues, collapsed and adjusted p-value cut-off was set to 0.05, while no evidence of differential expression called for nominal p-values > 0.05. Thus, for example, the genes that were significantly up-regulated in only the tumor stroma and not differentially expressed in the tumor epithelium are those with the stromal up-regulation p-value (adjusted) < 0.05 and both epithelial up- and down-regulation nominal p-values > 0.05.

4.3.9. Assessment of DEG Groups' Robustness (LCM meta-analysis)

Since all the available datasets were used and no other independent study was thus available, I used a cross validation procedure in order to assess the robustness of the DEGs. Specifically, I divided the datasets selected for the comparisons, indicated in Section 4.3.8, into 2 groups of approximately equal size (4 and 4 datasets for the stroma, 4 and 5 datasets for the epithelium). In one group made of 4 stromal datasets and 4 epithelial datasets, which was the

training set, I defined the DEG classes as described. In the second group, the test set, we tested for the differential expression of genes belonging to the DEG classes according to the same criteria but using nominal p-values. I repeated the same procedure 100 times by randomly selecting different combinations of datasets as training and test sets.

4.3.10. Enrichment for Functional Categories

Gene ontology enrichment was calculated with the `enrichGO` function from the `clusterProfiler` package [99] using “Biological Process” GO categories and default parameters.

4.3.11. Relationship with Clinical Features (LCM meta-analysis)

The relationship between gene expression and clinical features (grade, age at diagnosis, size) was obtained by computing Spearman’s correlation coefficients and their corresponding *p*-values with the `rcorr` function from the `Hmisc` package (<https://CRAN.R-project.org/package=Hmisc> [377]).

The relationship with survival in the METABRIC cohort was determined by dividing patients in two groups by the median expression value of the gene of interest and computing their difference in disease free survival with the Kaplan–Meier method and the log-rank test through the `survival` package (<https://CRAN.R-project.org/package=survival> [378]).

The datasets used in the analysis of the correlation between DEG groups with clinical features (Paragraph 2.2.3) are listed below. Only samples annotated as invasive BC were used. Grade in stroma: GSE12622, GSE14548, GSE35019, and GSE90505; grade in epithelium: GSE1378, GSE14548, GSE35019, GSE5764, and GSE72644; age in stroma: GSE12622, GSE14548, GSE26910, GSE90505, and GSE90521; age in epithelium: GSE1378, GSE14548, GSE38959, GSE72644, and GSE13293; node in stroma: GSE12622, GSE14548, GSE90505, GSE35019; node in epithelium: GSE1378, GSE14548, GSE5764, and GSE35019; size in stroma: GSE14548, GSE35019, GSE90505, and GSE90521; and size in epithelium: GSE1378, GSE14548, GSE35019, and GSE72644. Correlation coefficients with a specific clinical feature in the selected datasets were obtained for each gene and then averaged across all datasets.

The meta-analysis of correlation with clinical features was performed as described in Section 2.4, but the computation of Spearman’s correlation and the corresponding *p*-values were completed with the `rcorr` function from the `Hmisc` package (<https://CRAN.R-project.org/package=Hmisc> [377]), and these values were used in step 5 of the meta-analysis.

4.3.12. Epithelial, Stromal and Vascular Scores (LCM meta-analysis)

Scores for epithelial, stromal, and vascular signatures expression were calculated with a single sample GSEA (ssGSEA) via the GSVA package [100] using signatures obtained from the current meta-analysis comparing tumor and normal gene expression in samples of epithelial, stromal, or vascular origin.

Stromal and epithelial markers were obtained when comparing epithelial and stromal gene expression profiles in invasive BC samples from the same dataset and merging the *p*-values with the Fisher method as described above. The datasets used are GSE10797, GSE14548, GSE33692, GSE35019, GSE41228, GSE5847, GSE59772, GSE68744, GSE81838, GSE83591, and GSE88715. To achieve a higher stringency, we only retained DEGs with $|\text{average logFC}| > 1$.

Multivariate Cox models were fit with the `coxph` function from the survival package (<https://CRAN.R-project.org/package=survival>).

4.3.13. Web Resource (LCM meta-analysis)

The web resource (<https://aurorasavino.shinyapps.io/metalcml/>) was built with the Shiny package (<https://CRAN.R-project.org/package=shiny>) taking advantage of `rintrojs` [103], `shinybusy` (<https://CRAN.R-project.org/package=shinybusy>), `shinythemes` (<https://CRAN.R-project.org/package=shinythemes>), and `shinyWidgets` (<https://CRAN.R-project.org/package=shinyWidgets>).

The data that can be easily interrogated with the web resource are datasets of primary invasive breast cancers, excluding inflammatory and micropapillary cancers. A tutorial describing tool's usage is shown by pressing the "Tutorial" button. The user can choose between two conditions to compare based on compartment (stroma or epithelium), disease status (invasive BC, normal or normal counterpart - "counterpart"), and PAM50 subtype. The analysis pipeline applied is the same as the one used in this work and detailed above (Paragraph 4.3.4). Additionally, the enrichment of user-defined gene lists for DEGs can be assessed in the second tab, displaying the result of a one-tailed Fisher test (`fisher.test` function from R stats). To be noted, when the user inputs a gene list, the correction for multiple testing, used to determine the list of DEGs shown in the first tab, is applied to those genes only. The *p*-adjustment is applied to all genes when obtaining DEGs for the Fisher test.

The DEG lists can be downloaded, and include collapsed *p*-values, average logFC across analyzed datasets, and individual-datasets' *p*-values and logFCs.

4.3.14. Stat3 Signatures

Signatures of Stat3 activity were obtained from Azare et al. [104], Dauer et al. [105], IL6 and Jak/STAT from MSigDB [106], Alvarez et al. [107], Tell and Horvath [108], and Sonnenblick et al. [109].

4.4. crossWGCNA

4.4.1. crossWGCNA pipeline

The pipeline consists of the following steps: (1) data pre-processing, (2) Adjacency calculation, (3) Degrees' calculation, (3) Topological Overlap, (4) clustering.

- 1) Data pre-processing: matched transcriptomic data of two tissues from the same subjects are given as input to the algorithm. To reduce memory requirements, the genes are filtered according to a defined criterion (either the belonging to a functional category or the gene expression variance across samples). After filtering, the genes present in the expression matrices of the two tissues can be different. The expression matrices are combined row-wise, and to distinguish in which tissue gene expression was measured, a label of “_1” or “_2” is added to each gene ID.
- 2) Adjacency: obtained from the pre-processed data, computing Pearson correlation, and then applying either the “signed” or the “unsigned” methods employed in WGCNA.
- 3) Degrees: prior to computing the degrees, diagonal adjacencies, and self-loops (a gene with itself in the two tissues) are set to zero. The degrees are then defined as the sum of the adjacencies across all nodes. Specifically: k_{Int} is the sum across all edges with nodes of the same tissue (same “_1” or “_2” label), k_{Ext} is the sum across all edges with nodes of the alternate tissue, k_{Tot} is the sum across all edges. Hence, for each dataset, six degree vectors are defined: k_{Int} , k_{Ext} , k_{Tot} , one for each of the two tissues.
- 4) The Topological Overlap Matrix (TOM) is obtained with the function TOMsimilarity from the WGCNA package (TOMtype=“signed”).
- 5) Clustering is done on the TOM, after setting self-loops and intra-tissue TOMs to zero. A dynamic tree cut algorithm is used, employing the WGCNA functions cutreeDynamic (distance = 1-TOMsimilarity, method=“average”, deepsplit=4) and mergeCloseModules (cutHeight=0.25)

4.4.2. Brain data

GTEx RNA-seq data (TPM) were downloaded from the GTEx portal (<https://gtexportal.org/home/datasets>, file GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct, version 8), log2 transformed with an offset of 1, and gene IDs converted to HGNC ids as previously described (Paragraph 4.2.1). For each pair of analyzed tissues, only subjects being profiled for both tissues were kept and matched using the SampleAttributesDS and SubjectPhenotypesDS tables. Only genes involved in synapse functioning have been kept, using the gene list provided by the SynGO database (<https://www.syngoportal.org/>).

GSE25219 [253] was downloaded with GEOquery and pre-processed as previously described (Paragraph 4.2.1).

For both datasets, crossWGCNA was run with signed adjacency and $\beta = 6$. The “connectivity” score was defined as the sum of all cross-regions adjacencies.

Functional magnetic resonance imaging (fMRI) data were obtained from ABIDE (<http://preprocessed-connectomes-project.org/abide/>), pre-processed as described in <https://arxiv.org/pdf/2006.05176.pdf> - Section 5, and converted to correlation matrices. Only healthy subjects were considered for this analysis.

The brain areas analyzed in all three datasets (transcriptome and fMRI) were seven: Frontal Cortex (BA9), Cerebellum, Caudate (basal ganglia), Putamen (basal ganglia), Hippocampus, Anterior cingulate cortex (BA24), Amygdala. Since multiple regions matched these seven areas in the fMRI data, correlations for all matched regions were averaged.

CrossWGCNA outputs a single “connectivity” score for each pair of regions since it uses all subjects’ transcriptomes to compute the correlation. fMRI data provide, instead, a connectivity measure for each pair of regions and each subject. Therefore, the Pearson correlation between fMRI connectivity and crossWGCNA connectivity score was calculated for each fMRI-analyzed subject independently, obtaining a distribution of correlations that was compared with an alternative distribution (correlation between the fMRI connectivity and transcriptional similarity between brain areas), showing that crossWGCNA gives correlations shifted towards higher values (Kolmogorov-Smirnov test, one-sided).

To assess whether the relationship between the two scores (crossWGCNA and fMRI) depends on the physical distance between areas, the Euclidean distance was computed and used as covariate

together with the crossWGCNA score to predict the fMRI connectivity, showing that the p-value for the crossWGCNA coefficient remained significant.

4.4.3. Functional enrichment

GSEA of k_{Ratio} were performed with the fgsea package [351], using gene lists obtained with msigdb, category “C5”, subcategory= “BP”. Only categories with p-adjusted<0.05 were considered.

GSEA of the rank product of genes’ k_{Ratio} across datasets was performed similarly (low ranks corresponding to high k_{Ratios}).

4.4.4. Breast Cancer LCM data

Datasets analyzed: GSE5847, GSE10797, GSE14548, GSE83591, GSE68744, GSE88715 (part of the LCM database, Paragraph 2.2). Only samples annotated as InvasiveBC were kept and divided between the matched stroma and epithelial compartments. Genes were filtered based on their variance across samples (separately for stroma and epithelium), removing genes with variance lower than the first quartile.

Parameters used in crossWGCNA functions: Adjacency “signed”, beta=6, deepSplit=4.

4.4.5. Analysis of co-cultures

Data of epithelial breast cancer cell lines co-cultured with cancer associated fibroblasts were retrieved from GEO with the function GEOquery [341]. Specifically, the datasets are: GSE106503 [379], GSE119253 [380], GSE27018 [381], GSE27120 [382], GSE41678 [383], GSE4708, GSE60998 [384], GSE70884 [385], GSE80333 [386], GSE95461 [387]. DEGs were identified with limma [376] and filtered for p-adjusted<0.05, except for GSE60998, GSE80333, GSE95461, for which they were retrieved from the supplementary files of the corresponding publication.

GSEA were performed (package fgsea [351]) to test the enrichment of DEG lists in genes with high k_{Ratio} and compared with results obtained reshuffling gene labels in the crossWGCNA network (100 times).

4.4.6. Analysis of single cell data

Single cell data belonging to the GSE161529 dataset were downloaded as mtx files and imported in R. Only cancer data were used. Samples annotated as TNBC, ER+ and HER2+ were analyzed separately, following a standard Seurat pipeline:

- Data normalization: NormalizeData function

- Selection of variable features: FindVariableFeatures function, selection.method = "vst", nfeatures = 2000
- Data scaling: ScaleData function
- Dimensionality reduction: RunPCA (npcs=30), RunUMAP (reduction="pca", dims=1:30)
- Clustering: FindNeighbors(1:30), FindClusters(resolution=0.1)

To define cluster cell identities, markers' expression was plotted with the FeaturePlot function.

Epithelial markers: EPCAM, EGFR, CDH1, KRT14, ITGA6, KRT5, TP63, KRT17, MME, KRT8, KRT18, KRT19, FOXA1, GATA3, MUC1, CD24, KIT, GABRP

Fibroblast markers: FAP, COL1A1, COL3A1, COL5A1, ACTA2, TAGLN, LUM, FBLN1, COL6A3, COL1A2, COL6A1, COL6A2

Endothelial markers: PECAM1, VWF, CDH5, SELE

T-cell markers: CD2, CD3D, CD3E, CD3G, CD8A, CD8B

B-cell markers: MS4A1, CD79A, CD79B, BLNK

Macrophage markers: CD14, CD68, CD163, CSF1R

The clusters identified as epithelial were: TNBC clusters 0,5,3,13,9,4; HER2+ clusters 9,5,10,1,3; ER+ clusters 1,2,5,6,7,8,9,10,11,12,13,14,15,18,19,20,21.

The clusters identified as fibroblasts were: TNBC cluster 8; HER2+ clusters 7,13; ER+ clusters 4,16.

The normalized data were obtained from the Seurat objects with the function GetAssayData (slot="data", assay="RNA"), and gene expression averaged across cells of the same type, resulting in an expression vector for each patient and each cell type. These vectors were then merged in a matrix for each cell type, with columns corresponding to subjects in matched order. The top 50% of genes were filtered based on expression variance across all samples, epithelial and fibroblast matrices were merged and given as input to crossWGCNA.

5. Acronyms

AGR2	Anterior Gradient 2
BC	Breast Cancer
BLBC	Basal Like Breast Cancer
BMI	Body Mass Index
BRCA1	Breast cancer 1
CAF	Cancer Associated Fibroblast
CDCA5	Cell Division Cycle Associated 5
CEBPG	CCAAT Enhancer Binding Protein Gamma
CENPA	Centromere Protein A
CHI3L2	Chitinase 3 Like 2
CP	Ceruloplasmin
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CX3CL1	Chemokine (C-X3-C motif) ligand 1
CYBRD1	Cytocrome B Reductase 1
DC	Differentially Co-expressed
DCIS	Ductal carcinoma in situ
Dcytb	Duodenal cytochrome B
DEFB1	Defensin Beta 1
DEG	Differentially Expressed Gene
E2F3	E2F Transcription Factor 3
ECM	Extracellular matrix
EDN1	Endothelin 1
EGA	European Genome Archive
ELF5	E74-Like Factor 5
EMT	Epithelial to Mesenchymal Transition
EPCAM	Epithelial Cell Adhesion Molecule
ESR1	Estrogen Receptor 1
EZH2	Enhancer of Zeste Homolog 2
FAP	Fibroblast Activation Protein Alpha
FC	Fold Change
fMRI	functional Magnetic Resonance Imaging

FOXA1	Forkhead box A1
FOXC1	Forkhead box C1
FOXM1	Forkhead box M1
GATA3	GATA Binding Protein 3
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRN	Gene Regulatory Networks
GSEA	Gene Set Enrichment Analysis
HER2	Human Epidermal Growth Factor Receptor 2
HMGA1	High Mobility Group Protein A1
HSD11B2	Hydroxysteroid 11-beta Dehydrogenase 2
IGIP	IgA-Inducing Protein Homolog
IL6ST	Interleukin 6 Cytokine Family Signal Transducer
INHBB	Inhibin Subunit Beta B
IDC	Invasive Ductal Carcinoma
KD	Knock-Down
KLF12	Krupper Like Factor 12
KRT	Keratin
LCM	Laser Capture Microdissection
LN	Lymphnode
MIA	Melanoma Inhibitory Activity
MLPH	Melanophilin
MSigDB	Molecular Signature Database
NK	Natural Killer
NMB	Neuromedin B
NOTCH2NLA	Notch 2 N-terminal Like A
OE	Overexpressed
PDGFRA	Platelet Derived Growth Factor receptor Alpha
PPI	Protein-protein interaction
PR	Progesteron Receptor
ROPN1	Rhopilin Associated Tail Protein 1
P4HTM	Prolyl 4-Hydroxylase, Transmembrane

PARP1	Poly(ADP-Ribose) Polymerase 1
PCP	Plasma Cell Polarity
PHGDH	Phosphoglycerate Dehydrogenase
PRRX1	Paired Related Homeobox 1
PSAT1	Phosphoserine aminotransferase 1
PTTG1	Pituitary Tumor-Transforming Gene 1 Protein
ROR2	Receptor Tyrosine Kinase Like Orphan Receptor 2
RUNX1	Runt-related transcription factor 1
SHMT2	Serine Hydroxymethyltransferase 2
SLPI	Secretory Leukocyte Peptidase Inhibitor
SOCS3	Suppressor of cytokine signaling 3
SOX10	SRY-box Transcription Factor 1
SSRP1	Structure Specific Recognition Protein 1
STAT3	Signal Transducer and Activator of Transcription 3
TCGA	The Cancer Genome Atlas
TCF7L1	Transcription Factor 7 Like 1
TEAD	TEA Domain Transcription Factor
TF	Transcription Factor
TFDP1	Transcription Factor DP1
THRA	Thyroid Hormone Receptor Alpha
TIL	Tumor Infiltrating Lymphocytes
TNBC	Triple Negative Breast Cancer
TP53	Tumor Protein p53
TPX2	Targeting Protein for Xklp2
UPB1	Beta-ureidopropionase 1
VGLL1	Vestigial Like Family Member 1
WGCNA	Weighted Gene Co-expression Network Analysis
WNT	Wingless
ZNF532	Zinc Finger 532

6. References

1. Goldhirsch, A.; Ingle, J.N.; Gelber, R.D.; Coates, A.S.; Thürlimann, B.; Senn, H.J. *Annals of Oncology* 2009, **20**, 1319–1329, doi:10.1093/annonc/mdp322.
2. Viale, G. *Annals of Oncology* 2012, **23**, x207–x210, doi:https://doi.org/10.1093/annonc/mds326.
3. Perou, C.M.; Sørli, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C. a; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L. a; et al. *Nature* 2000, **406**, 747–752, doi:10.1038/35021093.
4. Perou, C.M.; Sørli, T.; Eisen, M.B.; van de Rijn, M.; Jeffrey, S.S.; Rees, C.A.; Pollack, J.R.; Ross, D.T.; Johnsen, H.; Akslen, L.A.; et al. *Nature* 2000, **406**, 747–752, doi:10.1038/35021093.
5. Barabási, A.-L.; Oltvai, Z.N. *Nature Reviews Genetics* 2004, **5**, 101–113, doi:10.1038/nrg1272.
6. Huang, S.; Ingber, D.E. *Breast Disease* 2006, **26**, 27–54, doi:10.3233/bd-2007-26104.
7. Kanehisa, M.; Goto, S. *Nucleic acids research* 2000, **28**, 27–30, doi:10.1093/nar/28.1.27.
8. Srivas, R.; Shen, J.P.; Yang, C.C.; Sun, S.M.; Li, J.; Gross, A.M.; Jensen, J.; Licon, K.; Bojorquez-Gomez, A.; Klepper, K.; et al. *Molecular Cell* 2016, **63**, 514–525, doi:10.1016/j.molcel.2016.06.022.
9. Conte, F.; Fiscon, G.; Licursi, V.; Bizzarri, D.; D'Antò, T.; Farina, L.; Paci, P. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 2020, **1863**, 194416, doi:10.1016/j.bbagr.2019.194416.
10. Derisi, J.L.; Iyer, V.R.; Brown, P. *Science* 1997, **278**, 680–686.
11. Jansen, R.; Greenbaum, D.; Gerstein, M. 2002, 37–46, doi:10.1101/gr.205602.4.
12. Ge, H.; Liu, Z.; Church, G.M.; Vidal, M. *Nature Genetics* 2001, **29**, 482–486, doi:10.1038/ng776.
13. Kemmeren, P.; Van Berkum, N.L.; Vilo, J.; Bijma, T.; Donders, R.; Brazma, A.; Holstege, F.C.P. *Molecular Cell* 2002, **9**, 1133–1143, doi:10.1016/S1097-2765(02)00531-2.
14. Holding, A.N.; Giorgi, F.M.; Donnelly, A.; Cullen, A.E.; Nagarajan, S.; Selth, L.A.; Markowitz, F. *Genome Biology* 2019, **20**, 1–16, doi:10.1186/s13059-019-1733-0.
15. Salwinski, L.; Miller, C.S.; Smith, A.J.; Pettit, F.K.; Bowie, J.U.; Eisenberg, D. *Nucleic Acids Research* 2004, **32**, D449–D451, doi:10.1093/nar/gkh086.
16. Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardoza, A.P.; Santonico, E.; et al. *Nucleic Acids Research* 2012, **40**, D857–D861, doi:10.1093/nar/gkr930.
17. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; del-Toro, N.; et al. *Nucleic Acids Research* 2014, **42**, D358–D363, doi:10.1093/nar/gkt1115.
18. Oughtred, R.; Stark, C.; Breitkreutz, B.-J.; Rust, J.; Boucher, L.; Chang, C.; Kolas, N.; O'Donnell, L.; Leung, G.; McAdam, R.; et al. *Nucleic Acids Research* 2019, **47**, D529–D541, doi:10.1093/nar/gky1079.
19. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. *Nucleic Acids Research* 2019, **47**, D607–D613, doi:10.1093/nar/gky1131.

20. Obayashi, T.; Kagaya, Y.; Aoki, Y.; Tadaka, S.; Kinoshita, K. *Nucleic Acids Research* 2019, **47**, D55–D62, doi:10.1093/nar/gky1155.
21. Zhu, Q.; Wong, A.K.; Krishnan, A.; Aure, M.R.; Tadych, A.; Zhang, R.; Corney, D.C.; Greene, C.S.; Bongo, L.A.; Kristensen, V.N.; et al. *Nature Methods* 2015, **12**, 211–214, doi:10.1038/nmeth.3249.
22. Jeong, H.; Mason, S.P.; Barabási, A.-L.; Oltvai, Z.N. *Nature* 2001, **411**, 41, doi:10.1038/35075138.
23. Pržulj, N.; Wigle, D.A.; Jurisica, I. *Bioinformatics* 2004, **20**, 340–348, doi:10.1093/bioinformatics/btg415.
24. Furlong, L.I. *Trends in Genetics* 2013, **29**, 150–159, doi:https://doi.org/10.1016/j.tig.2012.11.004.
25. Malod-dognin, N.; Petschnigg, J.; Windels, S.F.L.; Povh, J.; Hemmingway, H.; Ketteler, R. *Nature communications* 2019, **10**, 1–13, doi:10.1038/s41467-019-08797-8.
26. Jonsson, P.F.; Bates, P.A. *Bioinformatics* 2006, **22**, 2291–2297, doi:10.1093/bioinformatics/btl390.
27. Feldman, I.; Rzhetsky, A.; Vitkup, D. *Proceedings of the National Academy of Sciences* 2008, **105**, doi:10.1073/pnas.0701722105.
28. Ala, U.; Piro, R.M.; Grassi, E.; Damasco, C.; Silengo, L.; Oti, M.; Provero, P.; Di Cunto, F. *PLoS computational biology* 2008, **4**, e1000043, doi:10.1371/journal.pcbi.1000043.
29. Magger, O.; Waldman, Y.Y.; Rupp, E.; Sharan, R. *PLoS Computational Biology* 2012, **8**, doi:10.1371/journal.pcbi.1002690.
30. Li, M.; Zhang, J.; Liu, Q.; Wang, J.; Wu, F.X. *BMC Medical Genomics* 2014, **7**, S4, doi:10.1186/1755-8794-7-S2-S4.
31. Rives, A.W.; Galitski, T. *Proceedings of the National Academy of Sciences* 2003, **100**, 1128 LP – 1133, doi:10.1073/pnas.0237338100.
32. Eisen, M.B.; Spellman, P.T.; Brown, P.O.; Botstein, D. *Proceedings of the National Academy of Sciences* 1998, **95**, 14863–14868.
33. Yang, Y.; Han, L.; Yuan, Y.; Li, J.; Hei, N.; Liang, H. *Nature Communications* 2014, **5**, 3231, doi:10.1038/ncomms4231.
34. Saelens, W.; Cannoodt, R.; Saey, Y. *Nature communications* 2018, **9**, doi:10.1038/s41467-018-03424-4.
35. Zhang, B.; Horvath, S. *Statistical Applications in Genetics and Molecular Biology* 2005, **4**, doi:10.2202/1544-6115.1128.
36. Langfelder, P.; Horvath, S. *BMC bioinformatics* 2008, **9**, 559, doi:10.1186/1471-2105-9-559.
37. Luscombe, N.M.; Babu, M.M.; Yu, H. *Letters To Nature* 2004, **431**, 714–717.
38. Neph, S.; Stergachis, A.B.; Reynolds, A.; Sandstrom, R.; Borenstein, E.; Stamatoyannopoulos, J.A. *Cell* 2012, **150**, 1274–1286, doi:https://doi.org/10.1016/j.cell.2012.04.040.
39. Ayer, D.E.; Eisenman, R.N. *Genes & development* 1993, **7**, 2110–2119, doi:10.1101/gad.7.11.2110.
40. de la Fuente, A. *Trends in Genetics* 2010, **26**, 326–333, doi:10.1016/j.tig.2010.05.001.
41. Lai, Y. *Bioinformatics* 2008, **24**, 666–673, doi:10.1093/bioinformatics/btm507.

42. Hudson, N.J.; Reverter, A.; Dalrymple, B.P. *PLoS Computational Biology* 2009, **5**, doi:10.1371/journal.pcbi.1000382.
43. Bonnal, S.C.; López-Oreja, I.; Valcárcel, J. Roles and Mechanisms of Alternative Splicing in Cancer — Implications for Care. *Nature Reviews Clinical Oncology* 2020, **17**, 457–474.
44. Ouyang, J.; Zhong, Y.; Zhang, Y.; Yang, L.; Wu, P.; Hou, X.; Xiong, F.; Li, X.; Zhang, S.; Gong, Z.; et al. Long Non-Coding RNAs Are Involved in Alternative Splicing and Promote Cancer Progression. *British Journal of Cancer* 2021.
45. Park, S.H.; Brugiolo, M.; Akerman, M.; Das, S.; Urbanski, L.; Geier, A.; Kesarwani, A.K.; Fan, M.; Leclair, N.; Lin, K.T.; et al. *Cell Reports* 2019, **29**, 2672–2688.e7, doi:10.1016/j.celrep.2019.10.110.
46. Wang, L.; Wang, Y.; Su, B.; Yu, P.; He, J.; Meng, L.; Xiao, Q.; Sun, J.; Zhou, K.; Xue, Y.; et al. *Scientific Reports* 2020, **10**, doi:10.1038/s41598-020-73700-1.
47. Han, P.; Zhu, J.; Feng, G.; Wang, Z.; Ding, Y. *BMC Cancer* 2021, **21**, doi:10.1186/s12885-021-08305-6.
48. Koedoot, E.; Smid, M.; Foekens, J.A.; Martens, J.W.M.; le Dévédec, S.E.; van de Water, B. *Scientific Reports* 2019, **9**, doi:10.1038/s41598-019-40759-4.
49. Clarke, C.; Madden, S.F.; Doolan, P.; Aherne, S.T.; Joyce, H.; Driscoll, L.O.; Gallagher, W.M.; Hennessy, B.T.; Moriarty, M.; Crown, J.; et al. *Carcinogenesis* 2013, **34**, 2300–2308, doi:10.1093/carcin/bgt208.
50. Wolf, D.M.; Lenburg, M.E.; Yau, C.; Boudreau, A.; Veer, L.J. *Van PLoS ONE* 2014, **9**, doi:10.1371/journal.pone.0088309.
51. Liu, X.; Bulgakov, O. V.; Darrow, K.N.; Pawlyk, B.; Adamian, M.; Liberman, M.C.; Li, T. *Proceedings of the National Academy of Sciences* 2007, **104**, 4413–4418, doi:10.1073/pnas.0610950104.
52. Shi, H.; Zhang, L.E.I.; Qu, Y.; Hou, L.; Wang, L.; Zheng, M.I.N. *Oncology letters* 2017, **14**, 4535–4542, doi:10.3892/ol.2017.6779.
53. Herschkowitz, J.I.; Komurov, K.; Zhou, A.Y.; Gupta, S.; Yang, J.; Hartwell, K.; Onder, T.T.; Gupta, P.B.; Evans, K.W.; Hollier, B.G.; et al. *Proceedings of the National Academy of Sciences* 2010, **107**, 19132, doi:10.1073/pnas.1015095107.
54. Yang, Y.; Han, L.; Yuan, Y.; Li, J.; Hei, N.; Liang, H. *Nature Communications* 2014, **5**, 1–9, doi:10.1038/ncomms4231.
55. Savino, A.; Provero, P.; Poli, V. *International Journal of Molecular Sciences* 2020, **21**, 1–23, doi:10.3390/ijms21249461.
56. Zhu, L.; Ding, Y.; Chen, C.; Wang, L.; Huo, Z.; Kim, S.; Sotiriou, C.; Oesterreich, S.; Tseng, G.C. *Bioin* 2017, **33**, 1121–1129, doi:10.1093/bioinformatics/btw788.
57. Klein, K.O.; Oualkacha, K.; Lafond, M.; Bhatnagar, S. *Frontiers in genetics* 2016, **7**, 1–14, doi:10.3389/fgene.2016.00137.
58. Yu, W.; Zhao, S.; Wang, Y.; Zhao, B.N.; Zhao, W. *Oncotarget* 2017, **8**, 112928–112941, doi:10.18632/oncotarget.22878.
59. Cohen, M.; Giladi, A.; Gorki, A.D.; Solodkin, D.G.; Zada, M.; Hladik, A.; Miklosi, A.; Salame, T.M.; Halpern, K.B.; David, E.; et al. *Cell* 2018, **175**, 1031–1044.e18, doi:10.1016/j.cell.2018.09.009.

60. Camp, J.G.; Sekine, K.; Gerber, T.; Loeffler-Wirth, H.; Binder, H.; Gac, M.; Kanton, S.; Kageyama, J.; Damm, G.; Seehofer, D.; et al. *Nature* 2017, **546**, 533–538, doi:10.1038/nature22796.
61. Noël, F.; Massenet-Regad, L.; Carmi-Levy, I.; Cappuccio, A.; Grandclaude, M.; Trichot, C.; Kieffer, Y.; Mehta-Grigoriou, F.; Soumelis, V. *Nature Communications* 2021, **12**, 1–16, doi:10.1038/s41467-021-21244-x.
62. Vento-Tormo, R.; Efremova, M.; Botting, R.A.; Turco, M.Y.; Vento-Tormo, M.; Meyer, K.B.; Park, J.E.; Stephenson, E.; Polański, K.; Goncalves, A.; et al. *Nature* 2018, **563**, 347–353, doi:10.1038/s41586-018-0698-6.
63. Jin, S.; Guerrero-Juarez, C.F.; Zhang, L.; Chang, I.; Ramos, R.; Kuan, C.H.; Myung, P.; Plikus, M. V.; Nie, Q. *Nature Communications* 2021, **12**, 1–20, doi:10.1038/s41467-021-21246-9.
64. Armingol, E.; Officer, A.; Harismendy, O.; Lewis, N.E. *Nature Reviews Genetics* 2021, **22**, 71–88, doi:10.1038/s41576-020-00292-x.
65. Komurov, K. *Bioinformatics* 2012, **28**, 694–700, doi:10.1093/bioinformatics/btr718.
66. Efremova, M.; Vento-Tormo, M.; Teichmann, S.A.; Vento-Tormo, R. *Nature Protocols* 2020, **15**, 1484–1506, doi:10.1038/s41596-020-0292-x.
67. Browaeys, R.; Saelens, W.; Saeys, Y. *Nature Methods* 2020, **17**, 159–162, doi:10.1038/s41592-019-0667-5.
68. Wang, S.; Karikomi, M.; Maclean, A.L.; Nie, Q. *Nucleic Acids Research* 2019, **47**, 1–13, doi:10.1093/nar/gkz204.
69. Graeber, T.G.; Eisenberg, D. *Nature Genetics* 2001, **29**, 295–300, doi:10.1038/ng755.
70. Choi, H.; Sheng, J.; Gao, D.; Li, F.; Durrans, A.; Ryu, S.; Lee, S.B.; Narula, N.; Rafii, S.; Elemento, O.; et al. *Cell Reports* 2015, **10**, 1187–1201, doi:10.1016/j.celrep.2015.01.040.
71. Oh, E.Y.; Christensen, S.M.; Ghanta, S.; Jeong, J.C.; Bucur, O.; Glass, B.; Montaser-Kouhsari, L.; Knoblauch, N.W.; Bertos, N.; Saleh, S.M.I.; et al. *Genome Biology* 2015, **16**, doi:10.1186/s13059-015-0675-4.
72. Rung, J.; Brazma, A. *Nature reviews. Genetics* 2013, **14**, 89–99, doi:10.1038/nrg3394.
73. Kela, I.; Ein-Dor, L.; Getz, G.; Givol, D.; Domany, E. *Breast Cancer Research* 2005, **7**, P4.38, doi:10.1186/bcr1168.
74. Michiels, S.; Koscielny, S.; Hill, C. *Lancet (London, England)* 2005, **365**, 488–492, doi:10.1016/S0140-6736(05)17866-0.
75. Shen, R.; Chinnaiyan, A.M.; Ghosh, D. *BMC Medical Genomics* 2008, **1**, 1–7, doi:10.1186/1755-8794-1-28.
76. Ioannidis, J.P.A.; Ntzani, E.E.; Trikalinos, T.A.; Contopoulos-Ioannidis, D.G. *Nature Genetics* 2001, **29**, 306–309, doi:10.1038/ng749.
77. Cui, W.; Xue, H.; Wei, L.; Jin, J.; Tian, X.; Wang, Q. *Human Genomics* 2021, **15**, 7, doi:10.1186/s40246-021-00308-5.
78. Shedden, K.; Taylor, J.M.G.; Enkemann, S.A.; Tsao, M.-S.; Yeatman, T.J.; Gerald, W.L.; Eschrich, S.; Jurisica, I.; Giordano, T.J.; Misek, D.E.; et al. *Nature medicine* 2008, **14**, 822–827, doi:10.1038/nm.1790.

79. Barrett, T.; Wilhite, S.E.; Ledoux, P.; Evangelista, C.; Kim, I.F.; Tomashevsky, M.; Marshall, K.A.; Phillippy, K.H.; Sherman, P.M.; Holko, M.; et al. *Nucleic Acids Research* 2013, **41**, D991–D995, doi:10.1093/nar/gks1193.
80. Lappalainen, I.; Almeida-King, J.; Kumanduri, V.; Senf, A.; Spalding, J.D.; ur-Rehman, S.; Saunders, G.; Kandasamy, J.; Caccamo, M.; Leinonen, R.; et al. *Nature Genetics* 2015, **47**, 692–695, doi:10.1038/ng.3312.
81. Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; Fonseca, N.A.; Petryszak, R.; Papatheodorou, I.; et al. *Nucleic Acids Research* 2019, **47**, D711–D715, doi:10.1093/nar/gky964.
82. Ramasamy, A.; Mondry, A.; Holmes, C.C.; Altman, D.G. *PLoS Medicine* 2008, **5**, 1320–1332, doi:10.1371/journal.pmed.0050184.
83. Taminiau, J.; Lazar, C.; Meganck, S.; Nowé, A. *ISRN Bioinformatics* 2014, **2014**, 1–7, doi:10.1155/2014/345106.
84. Chang, L.C.; Lin, H.M.; Sibille, E.; Tseng, G.C. *BMC Bioinformatics* 2013, **14**, doi:10.1186/1471-2105-14-368.
85. Walsh, C.; Hu, P.; Batt, J.; Santos, C. *Microarrays* 2015, **4**, 389–406, doi:10.3390/microarrays4030389.
86. Rhodes, D.R.; Yu, J.; Shanker, K.; Deshpande, N.; Varambally, R.; Ghosh, D.; Barrette, T.; Pandey, A.; Chinnaiyan, A.M. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**, 9309–9314, doi:10.1073/pnas.0401994101.
87. Choi, J.K.; Yu, U.; Kim, S.; Yoo, O.J. *Bioinformatics (Oxford, England)* 2003, **19 Suppl 1**, i84-90, doi:10.1093/bioinformatics/btg1010.
88. Hong, F.; Breitling, R.; McEntee, C.W.; Wittner, B.S.; Nemhauser, J.L.; Chory, J. *Bioinformatics* 2006, **22**, 2825–2827, doi:10.1093/bioinformatics/btl476.
89. Cai, K.-T.; Feng, C.-X.; Zhao, J.-C.; He, R.-Q.; Ma, J.; Zhong, J.-C. *Molecular medicine reports* 2018, **18**, 4994–5008, doi:10.3892/mmr.2018.9543.
90. Bozovic-Spasojevic, I.; Zardavas, D.; Brohée, S.; Ameye, L.; Fumagalli, D.; Ades, F.; de Azambuja, E.; Bareche, Y.; Piccart, M.; Paesmans, M.; et al. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2017, **23**, 2702–2712, doi:10.1158/1078-0432.CCR-16-0979.
91. Tang, H.; Wang, S.; Xiao, G.; Schiller, J.; Papadimitrakopoulou, V.; Minna, J.; Wistuba, I.I.; Xie, Y. *Annals of Oncology* 2017, **28**, 733–740, doi:10.1093/annonc/mdw683.
92. Wang, J.; Gao, P.; Song, Y.; Sun, J.; Chen, X.; Yu, H.; Wang, Y.; Wang, Z. *Journal of Cellular and Molecular Medicine* 2018, **22**, 5743–5747, doi:10.1111/jcmm.13823.
93. Sandhu, V.; Labori, K.J.; Borgida, A.; Lungu, I.; Bartlett, J.; Hafezi-Bakhtiari, S.; Denroche, R.E.; Jang, G.H.; Pasternack, D.; Mbaabali, F.; et al. *JCO clinical cancer informatics* 2019, **3**, 1–16, doi:10.1200/CCI.18.00102.
94. Yasrebi, H. *Briefings in Bioinformatics* 2016, **17**, 771–785, doi:10.1093/bib/bbv092.
95. Xu, L.; Tan, A.C.; Winslow, R.L.; Geman, D. *BMC Bioinformatics* 2008, **9**, doi:10.1186/1471-2105-9-125.

96. Long, N.P.; Jung, K.H.; Anh, N.H.; Yan, H.H.; Nghi, T.D.; Park, S.; Yoon, S.J.; Min, J.E.; Kim, H.M.; Lim, J.H.; et al. *Cancers* 2019, **11**, doi:10.3390/cancers11020155.
97. Ma, S.; Ogino, S.; Parsana, P.; Nishihara, R.; Qian, Z.; Shen, J.; Mima, K.; Masugi, Y.; Cao, Y.; Nowak, J.A.; et al. *Genome Biology* 2018, **19**, 1–14, doi:10.1186/s13059-018-1511-4.
98. de Santiago, I.; Yau, C.; Heij, L.; Middleton, M.R.; Markowitz, F.; Grabsch, H.I.; Dustin, M.L.; Sivakumar, S. *International journal of cancer* 2019, **145**, 1125–1137, doi:10.1002/ijc.32186.
99. Qadir, F.; Lalli, A.; Dar, H.H.; Hwang, S.; Aldehlawi, H.; Ma, H.; Dai, H.; Waseem, A.; Teh, M.-T. *BMC cancer* 2019, **19**, 830, doi:10.1186/s12885-019-6059-5.
100. Tan, T.Z.; Rouanne, M.; Tan, K.T.; Huang, R.Y.-J.; Thiery, J.-P. *European urology* 2019, **75**, 423–432, doi:10.1016/j.eururo.2018.08.027.
101. Oztemur Islakoglu, Y.; Noyan, S.; Aydos, A.; Gur Dedeoglu, B. *OMICS A Journal of Integrative Biology* 2018, **22**, 709–716, doi:10.1089/omi.2018.0157.
102. Bell, R.; Barraclough, R.; Vasieva, O. *Current Molecular Medicine* 2017, **17**, 200–210, doi:10.2174/1566524017666170807144946.
103. Lee, Y.S.; Hwang, S.G.; Kim, J.K.; Park, T.H.; Kim, Y.R.; Myeong, H.S.; Choi, J.D.; Kwon, K.; Jang, C.S.; Ro, Y.T.; et al. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 2016, **37**, 2285–2297, doi:10.1007/s13277-015-4033-7.
104. Curtis, C.; Shah, S.P.; Chin, S.-F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; et al. *Nature* 2012, **486**, 346–52, doi:10.1038/nature10983.
105. Daemen, A.; Griffith, O.L.; Heiser, L.M.; Wang, N.J.; Enache, O.M.; Sanborn, Z.; Pepin, F.; Durinck, S.; Korkola, J.E.; Griffith, M.; et al. *Genome Biology* 2013, **14**.
106. Jensen, L.J.; Kuhn, M.; Stark, M.; Chaffron, S.; Creevey, C.; Muller, J.; Doerks, T.; Julien, P.; Roth, A.; Simonovic, M.; et al. *Nucleic acids research* 2009, **37**, D412–D416, doi:10.1093/nar/gkn760.
107. Brandt, J.; Garne, J.P.; Tengrup, I.; Manjer, J. *World journal of surgical oncology* 2015, **13**, 1–11, doi:10.1186/s12957-014-0429-x.
108. Minn, A.J.; Gupta, G.P.; Siegel, P.M.; Bos, P.D.; Shu, W.; Giri, D.D.; Viale, A.; Olshen, A.B.; Gerald, W.L.; Massague, J. *Nature* 2005, **436**, 518–524, doi:10.1038/nature03799.
109. Wang, Y.; Klijn, J.G.M.; Zhang, Y.; Sieuwerts, A.M.; Look, M.P.; Yang, F.; Talantov, D.; Timmermans, M.; Meijer-van Gelder, M.E.; Yu, J.; et al. *Lancet (London, England)* 2005, **365**, 671–679, doi:10.1016/S0140-6736(05)17947-1.
110. Minn, A.J.; Gupta, G.P.; Padua, D.; Bos, P.; Nguyen, D.X.; Nuyten, D.; Kreike, B.; Zhang, Y.; Wang, Y.; Ishwaran, H.; et al. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**, 6740–6745, doi:10.1073/pnas.0701138104.
111. Bos, P.D.; Zhang, X.H.-F.; Nadal, C.; Shu, W.; Gomis, R.R.; Nguyen, D.X.; Minn, A.J.; van de Vijver, M.J.; Gerald, W.L.; Foekens, J.A.; et al. *Nature* 2009, **459**, 1005–1009, doi:10.1038/nature08021.
112. Yang, W.; Soares, J.; Greninger, P.; Edelman, E.J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J.A.; Thompson, I.R.; et al. *Nucleic Acids Research* 2013, **41**, 955–961, doi:10.1093/nar/gks1111.

113. Yoshihara, K.; Shahmoradgoli, M.; Martínez, E.; Vegesna, R.; Kim, H.; Torres-Garcia, W.; Treviño, V.; Shen, H.; Laird, P.W.; Levine, D.A.; et al. *Nature Communications* 2013, **4**, 2612, doi:10.1038/ncomms3612.
114. Onuchic, V.; Hartmaier, R.J.; Boone, D.N.; Samuels, M.L.; White, W.M.; Garovic, V.D.; Oesterreich, S.; Roth, M.E.; Adrian, V.; Milosavljevic, A.; et al. *Cell Reports* 2016, **17**, 2075–2086, doi:10.1016/j.celrep.2016.10.057.Epigenomic.
115. Gendoo, D.M.A.; Zon, M.; Sandhu, V.; Manem, V.S.K.; Ratanasirigulchai, N.; Chen, G.M.; Waldron, L.; Haibe-Kains, B. *Scientific Reports* 2019, **9**, 8770, doi:10.1038/s41598-019-45165-4.
116. Savino, A.; De Marzo, N.; Provero, P.; Poli, V. *Cancers* 2021, **13**, 3371, doi:10.3390/cancers13133371.
117. Simpson, E.H. *Journal of the Royal Statistical Society. Series B (Methodological)* 1951, **13**, 238–241, doi:http://www.jstor.org/stable/2984065.
118. Kong, S.L.; Li, G.; Loh, S.L.; Sung, W.; Liu, E.T. *Molecular Systems Biology* 2011, **7**, 1–14, doi:10.1038/msb.2011.59.
119. Saba, R.; Alsayed, A.; Zacny, J.P.; Dudek, A.Z. *International Journal of Breast Cancer* 2016, **2016**, doi:http://dx.doi.org/10.1155/2016/9768183.
120. Han, B.; Bhowmick, N.; Qu, Y.; Chung, S.; Giuliano, A.E.; Cui, X. *Oncogene* 2017, **36**, 3957–3963, doi:10.1038/onc.2017.48.
121. Ivanov, S. V.; Panaccione, A.; Nonaka, D.; Prasad, M.L.; Boyd, K.L.; Brown, B.; Guo, Y.; Sewell, A.; Yarbrough, W.G. *British journal of cancer* 2013, **109**, 444–51, doi:10.1038/bjc.2013.326.
122. Peevey, J.; Sumpter, I.; Paintal, A.; Laskin, W.; Sullivan, M. *American Journal of Clinical Pathology* 2015, **144**, A299–A299.
123. Dravis, C.; Spike, B.T.; Harrell, J.C.; Smith, E.M.S.; Perou, C.M.; Wahl, G.M.; Dravis, C.; Spike, B.T.; Harrell, J.C.; Johns, C.; et al. *Cell Reports* 2015, **12**, 2035–2048, doi:10.1016/j.celrep.2015.08.040.
124. Castilla, M.A.; Lòpez-Garcia, M.A.; Atienza, M.R.; Rosa-Rosa, J.M.; Diaz-Martin, J.; Pecero, M.L.; Vieites, B.; Romero-Pérez, L.; Benitez, J.; Calcabrini, A.; et al. *Endocrine-Related Cancer* 2014, **21**, 587–599, doi:10.1530/ERC-13-0485.
125. Bernardo, G.M.; Bebek, G.; Ginther, C.L.; Sizemore, S.T.; Lozada, K.L.; Miedler, J.D.; Anderson, L.A.; Godwin, A.K.; Abdul-karim, F.W. *Oncogene* 2012, **32**, 554–563, doi:10.1038/onc.2012.62.
126. Hisamatsu, Y.; Tokunaga, E.; Yamashita, N.; Akiyoshi, S. *Breast Cancer* 2015, **22**, 520–528, doi:10.1007/s12282-013-0515-x.
127. Theodorou, V.; Stark, R.; Menon, S.; Carroll, J.S. *Genome Research* 2013, **23**, 12–22, doi:10.1101/gr.139469.112.12.
128. Behan, F.M.; Iorio, F.; Picco, G.; Gonçalves, E.; Beaver, C.M.; Migliardi, G.; Santos, R.; Rao, Y.; Sassi, F.; Pinnelli, M.; et al. *Nature* 2019, **568**, 511–516, doi:10.1038/s41586-019-1103-9.
129. Tsherniak, A.; Vazquez, F.; Montgomery, P.G.; Weir, B.A.; Kryukov, G.; Cowley, G.S.; Gill, S.; Harrington, W.F.; Pantel, S.; Krill-Burger, J.M.; et al. *Cell* 2017, **170**, 564–576.e16, doi:10.1016/j.cell.2017.06.010.

130. McDonald, E.R.; de Weck, A.; Schlabach, M.R.; Billy, E.; Mavrakis, K.J.; Hoffman, G.R.; Belur, D.; Castelletti, D.; Frias, E.; Gampa, K.; et al. *Cell* 2017, **170**, 577–592.e10, doi:10.1016/j.cell.2017.07.005.
131. McDonald, E.R.; de Weck, A.; Schlabach, M.R.; Billy, E.; Mavrakis, K.J.; Hoffman, G.R.; Belur, D.; Castelletti, D.; Frias, E.; Gampa, K.; et al. *Cell* 2017, **170**, 577–592.e10, doi:10.1016/j.cell.2017.07.005.
132. Armanious, H.; Gelebart, P.; Mackey, J.; Ma, Y.; Lai, R. *International Journal of Clinical Experimental Pathology* 2010, **3**, 654–664.
133. Gujral, T.S.; Chan, M.; Peshkin, L.; Sorger, P.K.; Kirschner, M.W.; Macbeath, G. *Cell* 2014, **159**, 844–856, doi:10.1016/j.cell.2014.10.032.
134. Katoh, M.; Katoh, M. *International Journal of Molecular Medicine* 2007, **19**, 273–278.
135. Avalue, L.; Pensa, S.; Regis, G.; Novelli, F.; Poli, V. 2012, 65–72.
136. Yu, H.; Lee, H.; Herrmann, A.; Buettner, R.; Jove, R. *Nat Rev Cancer* 2014, **14**, 736–746, doi:10.1038/nrc3818.
137. Logan, C.Y.; Nusse, R. *Annual Review of Cell and Developmental Biology* 2004, **20**, 781–810, doi:10.1146/annurev.cellbio.20.010403.113126.
138. Klemm, F.; Bleckmann, A.; Siam, L.; Chuang, H.N.; Rietkötter, E.; Behme, D.; Schulz, M.; Schaffrinski, M.; Schindler, S.; Trümper, L.; et al. *Carcinogenesis* 2011, **32**, 434–442, doi:10.1093/carcin/bgq269.
139. Veeman, M.T.; Axelrod, J.D.; Moon, R.T. *Developmental Cell* 2003, **5**, 367–377, doi:10.1016/S1534-5807(03)00266-1.
140. Wang, Y. *Molecular cancer therapeutics* 2009, **8**, 2103–9, doi:10.1158/1535-7163.MCT-09-0282.
141. Monteleone, E.; Orecchia, V.; Corrieri, P.; Schiavone, D.; Avalue, L.; Moiso, E.; Savino, A.; Molineris, I.; Provero, P.; Poli, V. *Cancers* 2019, **11**, 101–116, doi:10.3390/cancers11010101.
142. Azare, J.; Leslie, K.; Al-Ahmadie, H.; Gerald, W.; Weinreb, P.H.; Violette, S.M.; Bromberg, J. *Molecular and cellular biology* 2007, **27**, 4444–4453, doi:10.1128/MCB.02404-06.
143. Alvarez, J. V.; Febbo, P.G.; Ramaswamy, S.; Loda, M.; Richardson, A.; Frank, D.A. *Cancer Research* 2005, **65**, 5054–5063.
144. Dauer, D.J.; Ferraro, B.; Song, L.; Yu, B.; Mora, L.; Buettner, R.; Enkemann, S.; Jove, R.; Haura, E.B. *Oncogene* 2005, **24**, 3397–3408, doi:10.1038/sj.onc.1208469.
145. Tell, R.W.; Horvath, C.M. *Proc Natl Acad Sci U S A* 2014, **111**, 12787–12792, doi:10.1073/pnas.1404881111.
146. Sonnenblick, A.; Brohée, S.; Fumagalli, D.; Vincent, D.; Venet, D.; Ignatiadis, M.; Salgado, R.; Van den Eynden, G.; Rothé, F.; Desmedt, C.; et al. *BMC medicine* 2015, **13**, 177, doi:10.1186/s12916-015-0416-2.
147. Bayerlová, M.; Menck, K.; Klemm, F.; Wolff, A.; Pukrop, T.; Gutenberg-universität, J.; Roos, W.P.; Gutenberg-universität, J.; Beißbarth, T. *Frontiers in oncology* 2017, **7**, 1–16, doi:10.3389/fonc.2017.00135.
148. Willert, J.; Epping, M.; Pollack, J.R.; Brown, P.O.; Nusse, R. *BMC Developmental Biology* 2002, **2**, 1–7.

149. Ziegler, S.; Ro, S.; Tickenbrock, L.; Mo, T.; Vetter, I.R.; Mu, O. *FEBS Journal* 2005, **272**, 1600–1615, doi:10.1111/j.1742-4658.2005.04581.x.
150. Labbe, E.; Lock, L.; Letamendia, A.; Gorska, A.E.; Gryfe, R.; Gallinger, S.; Moses, H.L.; Attisano, L. *Cancer Research* 2007, **67**, 75–85, doi:10.1158/0008-5472.CAN-06-2559.
151. Sandsmark, E.; Hansen, A.F.; Selnæs, K.M.; Bertilsson, H.; Bofin, A.M.; Wright, A.J.; Viset, T.; Richardsen, E.; Drabløs, F.; Bathen, T.F.; et al. *Oncotarget* 2017, **8**, 9572–9586.
152. Mcdaniel, J.M.; Varley, K.E.; Gertz, J.; Savic, D.S.; Brian, S.; Bailey, S.K.; Shevde, L.A.; Ramaker, R.C.; Brittany, N.; Kirby, M.K.; et al. *Oncotarget* 2016, **8**.
153. Bartoschek, M.; Oskolkov, N.; Bocci, M.; Lövrot, J.; Larsson, C.; Sommarin, M.; Madsen, C.D.; Lindgren, D.; Pekar, G.; Karlsson, G.; et al. *Nature Communications* 2018, **9**, doi:10.1038/s41467-018-07582-3.
154. Buffa, F.M.; Camps, C.; Winchester, L.; Snell, C.E.; Gee, H.E.; Sheldon, H.; Taylor, M.; Harris, A.L.; Ragoussis, J. *Cancer research* 2011, **71**, 5635–5645, doi:10.1158/0008-5472.CAN-11-0489.
155. Farazi, T.A.; Horlings, H.M.; Ten Hoeve, J.J.; Mihailovic, A.; Halfwerk, H.; Morozov, P.; Brown, M.; Hafner, M.; Reyat, F.; van Kouwenhove, M.; et al. *Cancer research* 2011, **71**, 4443–4453, doi:10.1158/0008-5472.CAN-11-0608.
156. Huang, X.; Dugo, M.; Callari, M.; Sandri, M.; De Cecco, L.; Valeri, B.; Carcangiu, M.L.; Xue, J.; Bi, R.; Veneroni, S.; et al. *Cancer medicine* 2015, **4**, 1016–1030, doi:10.1002/cam4.442.
157. Aure, M.R.; Jernström, S.; Krohn, M.; Volla, H.K.M.; Due, E.U.; Rødland, E.; Kåresen, R.; Ram, P.; Lu, Y.; Mills, G.B.; et al. *Genome medicine* 2015, **7**, 21, doi:10.1186/s13073-015-0135-5.
158. Dvinge, H.; Git, A.; Gräf, S.; Salmon-Divon, M.; Curtis, C.; Sottoriva, A.; Zhao, Y.; Hirst, M.; Armisen, J.; Miska, E.A.; et al. *Nature* 2013, **497**, 378–382, doi:10.1038/nature12108.
159. Ru, Y.; Kechris, K.J.; Tabakoff, B.; Hoffman, P.; Radcliffe, R.A.; Bowler, R.; Mahaffey, S.; Rossi, S.; Calin, G.A.; Bemis, L.; et al. *Nucleic Acids Research* 2014, **42**, e133–e133, doi:10.1093/nar/gku631.
160. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. *Nucleic Acids Research* 2019, **47**, D930–D940, doi:10.1093/nar/gky1075.
161. Chen, D.T.; Nasir, A.; Culhane, A.; Venkataramu, C.; Fulp, W.; Rubio, R.; Wang, T.; Agrawal, D.; McCarthy, S.M.; Gruidl, M.; et al. *Breast Cancer Research and Treatment* 2010, **119**, 335–346, doi:10.1007/s10549-009-0344-y.
162. Casey, T.; Bond, J.; Tighe, S.; Hunter, T.; Lintault, L.; Patel, O.; Eneman, J.; Crocker, A.; White, J.; Tessitore, J.; et al. *Breast Cancer Research and Treatment* 2009, **114**, 47–62, doi:10.1007/s10549-008-9982-8.
163. Santucci-Pereira, J.; Zeleniuch-Jacquotte, A.; Afanasyeva, Y.; Zhong, H.; Slifker, M.; Peri, S.; Ross, E.A.; López De Cicco, R.; Zhai, Y.; Nguyen, T.; et al. *Breast Cancer Research* 2019, **21**, 1–19, doi:10.1186/s13058-019-1128-x.
164. Gregory, K.J.; Roberts, A.L.; Conlon, E.M.; Mayfield, J.A.; Hagen, M.J.; Crisi, G.M.; Bentley, B.A.; Kane, J.J.; Makari-Judson, G.; Mason, H.S.; et al. *Breast Cancer Research* 2019, **21**, 1–18, doi:10.1186/s13058-019-1157-5.

165. Russo, J.; Balogh, G.A.; Russo, I.H. *Cancer Epidemiology Biomarkers and Prevention* 2008, **17**, 51–66, doi:10.1158/1055-9965.EPI-07-0678.
166. Ma, X.-J.; Wang, Z.; Ryan, P.D.; Isakoff, S.J.; Barmettler, A.; Fuller, A.; Muir, B.; Mohapatra, G.; Salunga, R.; Tuggle, J.T.; et al. *Cancer Cell* 2004, **5**, 607–616, doi:10.1016/j.ccr.2004.05.015.
167. Marino, N.; German, R.; Rao, X.; Simpson, E.; Liu, S.; Wan, J.; Liu, Y.; Sandusky, G.; Jacobsen, M.; Stoval, M.; et al. *npj Breast Cancer* 2020, **6**, 1–13, doi:10.1038/s41523-020-00191-8.
168. Ma, X.-J.; Dahiya, S.; Richardson, E.; Erlander, M.; Sgroi, D.C. *Breast cancer research : BCR* 2009, **11**, R7, doi:10.1186/bcr2222.
169. Emery, L.A.; Tripathi, A.; King, C.; Kavanah, M.; Mendez, J.; Stone, M.D.; De Las Morenas, A.; Sebastiani, P.; Rosenberg, C.L. *American Journal of Pathology* 2009, **175**, 1292–1302, doi:10.2353/ajpath.2009.090115.
170. Graham, K.; De Las Morenas, A.; Tripathi, A.; King, C.; Kavanah, M.; Mendez, J.; Stone, M.; Slama, J.; Miller, M.; Antoine, G.; et al. *British Journal of Cancer* 2010, **102**, 1284–1293, doi:10.1038/sj.bjc.6605576.
171. Graham, K.; Ge, X.; De Las Morenas, A.; Tripathi, A.; Rosenberg, C.L. *Clinical Cancer Research* 2011, **17**, 236–246, doi:10.1158/1078-0432.CCR-10-1369.
172. Bauer, J.A.; Chakravarthy, A.B.; Rosenbluth, J.M.; Mi, D.; Seeley, E.H.; Granja-Ingram, N.D.M.; Olivares, M.G.; Kelley, M.C.; Mayer, I.A.; Meszoely, I.M.; et al. *Clinical Cancer Research* 2010, **16**, 681–690, doi:10.1158/1078-0432.CCR-09-1091.
173. Planche, A.; Bacac, M.; Provero, P.; Fusco, C.; Delorenzi, M.; Stehle, J.C.; Stamenkovic, I. *PLoS ONE* 2011, **6**, doi:10.1371/journal.pone.0018640.
174. Lehmann, B.D.B.; Bauer, J. a J.; Chen, X.; Sanders, M.E.; Chakravarthy, a B.; Shyr, Y.; Pietenpol, J. a *Journal of Clinical Investigation* 2011, **121**, 2750–2767, doi:10.1172/JCI45014DS1.
175. Hosein, A.N.; Livingstone, J.; Buchanan, M.; Reid, J.F.; Hallett, M.; Basik, M. *BMC Cancer* 2015, **15**, 1–11, doi:10.1186/s12885-015-1117-0.
176. Harvell, D.M.E.; Kim, J.; O'Brien, J.; Tan, A.C.; Borges, V.F.; Schedin, P.; Jacobsen, B.M.; Horwitz, K.B. *Hormones and Cancer* 2013, **4**, 140–153, doi:10.1007/s12672-013-0136-z.
177. Knudsen, E.S.; Ertel, A.; Davicioni, E.; Kline, J.; Schwartz, G.F.; Witkiewicz, A.K. *Breast Cancer Research and Treatment* 2012, **133**, 1009–1024, doi:10.1007/s10549-011-1894-3.
178. Vargas, A.C.; Reed, A.E.M.C.; Waddell, N.; Lane, A.; Reid, L.E.; Smart, C.E.; Cocciardi, S.; Da Silva, L.; Song, S.; Chenevix-Trench, G.; et al. *Breast Cancer Research and Treatment* 2012, **135**, 153–165, doi:10.1007/s10549-012-2123-4.
179. Tchou, J.; Kossenkova, A. V.; Chang, L.; Satija, C.; Herlyn, M.; Showe, L.C.; Puré, E. *BMC Medical Genomics* 2012, **5**, doi:10.1186/1755-8794-5-39.
180. Komatsu, M.; Yoshimaru, T.; Matsuo, T.; Kiyotani, K.; Miyoshi, Y.; Tanahashi, T.; Rokutan, K.; Yamaguchi, R.; Saito, A.; Imoto, S.; et al. *International Journal of Oncology* 2013, **42**, 478–506, doi:10.3892/ijo.2012.1744.
181. Lee, S.; Stewart, S.; Nagtegaal, I.; Luo, J.; Wu, Y.; Colditz, G.; Medina, D.; Allred, D.C. *Cancer Research* 2012, **72**, 4574–4586, doi:10.1158/0008-5472.CAN-12-0636.

182. Woodward, W.A.; Krishnamurthy, S.; Yamauchi, H.; El-Zein, R.; Ogura, D.; Kitadai, E.; Niwa, S.-I.; Cristofanilli, M.; Vermeulen, P.; Dirix, L.; et al. Breast cancer research and treatment 2013, **138**, 761–772, doi:10.1007/s10549-013-2501-6.
183. Finak, G.; Sadekova, S.; Pepin, F.; Hallett, M.; Meterissian, S.; Halwani, F.; Khetani, K.; Souleimanova, M.; Zabolotny, B.; Omeroglu, A.; et al. Breast Cancer Research 2006, **8**, 1–18, doi:10.1186/bcr1608.
184. Turashvili, G.; Bouchal, J.; Baumforth, K.; Wei, W.; Dziechciarkova, M.; Ehrmann, J.; Klein, J.; Fridman, E.; Skarda, J.; Srovnal, J.; et al. BMC Cancer 2007, **7**, 1–20, doi:10.1186/1471-2407-7-55.
185. Reyngold, M.; Turcan, S.; Giri, D.; Kannan, K.; Walsh, L.A.; Viale, A.; Drobnjak, M.; Vahdat, L.T.; Lee, W.; Chan, T.A. PLoS ONE 2014, **9**, 1–10, doi:10.1371/journal.pone.0103896.
186. Boersma, B.J.; Reimers, M.; Yi, M.; Ludwig, J.A.; Luke, B.T.; Stephens, R.M.; Yfantis, H.G.; Lee, D.H.; Weinstein, J.N.; Ambis, S. International Journal of Cancer 2008, **122**, 1324–1332, doi:10.1002/ijc.23237.
187. Winslow, S.; Leandersson, K.; Edsjö, A.; Larsson, C. Breast Cancer Research 2015, **17**, 1–13, doi:10.1186/s13058-015-0530-2.
188. Oh, E.Y.; Christensen, S.M.; Ghanta, S.; Jeong, J.C.; Bucur, O.; Glass, B.; Montaser-Kouhsari, L.; Knoblauch, N.W.; Bertos, N.; Saleh, S.M.I.; et al. Genome Biology 2015, **16**, doi:10.1186/s13059-015-0675-4.
189. Abdalla, M.; Tran-Thanh, D.; Moreno, J.; Iakovlev, V.; Nair, R.; Kanwar, N.; Abdalla, M.; Lee, J.P.Y.; Kwan, J.Y.Y.; Cawthorn, T.R.; et al. Nature Communications 2017, **8**, doi:10.1038/s41467-017-01357-y.
190. Lee, S.; Medina, D.; Tsimelzon, A.; Mohsin, S.K.; Mao, S.; Wu, Y.; Allred, D.C. American Journal of Pathology 2007, **171**, 252–262, doi:10.2353/ajpath.2007.061010.
191. Lehmann, B.D.; Jovanović, B.; Chen, X.; Estrada, M. V.; Johnson, K.N.; Shyr, Y.; Moses, H.L.; Sanders, M.E.; Pietenpol, J.A. PLoS ONE 2016, **11**, 1–22, doi:10.1371/journal.pone.0157368.
192. Liu, H.; Dowdle, J.A.; Khurshid, S.; Sullivan, N.J.; Bertos, N.; Rambani, K.; Mair, M.; Daniel, P.; Wheeler, E.; Tang, X.; et al. Developmental Cell 2017, **41**, 392–407.e6, doi:10.1016/j.devcel.2017.04.024.
193. Gruosso, T.; Gigoux, M.; Manem, V.S.K.; Bertos, N.; Zuo, D.; Perlitch, I.; Saleh, S.M.I.; Zhao, H.; Souleimanova, M.; Johnson, R.M.; et al. Journal of Clinical Investigation 2019, **129**, 1785–1800, doi:10.1172/JCI96313.
194. Karnoub, A.E.; Dash, A.B.; Vo, A.P.; Sullivan, A.; Brooks, M.W.; Bell, G.W.; Richardson, A.L.; Polyak, K.; Tubo, R.; Weinberg, R.A. Nature 2007, **449**, 557–563, doi:10.1038/nature06188.
195. Saleh, S.M.I.; Bertos, N.; Gruosso, T.; Gigoux, M.; Souleimanova, M.; Zhao, H.; Omeroglu, A.; Hallett, M.T.; Park, M. Cancer Research 2017, **77**, 4673–4683, doi:10.1158/0008-5472.CAN-16-3427.
196. Brouwers, B.; Fumagalli, D.; Brohee, S.; Hatse, S.; Govaere, O.; Floris, G.; Van den Eynde, K.; Bareche, Y.; Schöffski, P.; Smeets, A.; et al. Breast Cancer Research 2017, **19**, 1–14, doi:10.1186/s13058-017-0871-0.
197. Tripathi, A.; King, C.; De La Morenas, A.; Perry, V.K.; Burke, B.; Antoine, G.A.; Hirsch, E.F.; Kavanah, M.; Mendez, J.; Stone, M.; et al. International Journal of Cancer 2008, **122**, 1557–1566, doi:10.1002/ijc.23267.

198. Jones, D.T.; Lechertier, T.; Mitter, R.; Herbert, J.M.J.; Bicknell, R.; Jones, J.L.; Li, J.L.; Buffa, F.; Harris, A.L.; Hodivala-Dilke, K. *PLoS ONE* 2012, **7**, doi:10.1371/journal.pone.0044294.
199. François, P.; Bertos, N.; Laferrière, J.; Sadekova, S.; Souleimanova, M.; Zhao, H.; Finak, G.; Meterissian, S.; Hallett, M.T.; Park, M. *Breast Cancer Research* 2012, **14**, doi:10.1186/bcr3246.
200. Le, P.U.; Li, J.; Cantin, C.; Collins, C.; Nantel, A.; Tremblay, T.; Kelly, J.F.; Wang, E.; Connor-mccourt, M.D.O.; Eg, A. 2015, **3**.
201. Dhage, S.; Ernlund, A.; Ruggles, K.; Axelrod, D.; Berman, R.; Roses, D.; Schneider, R.J. *PLoS ONE* 2018, **13**, 1–20, doi:10.1371/journal.pone.0205602.
202. Ershaid, N.; Sharon, Y.; Doron, H.; Raz, Y.; Shani, O.; Cohen, N.; Monteran, L.; Leider-Trejo, L.; Ben-Shmuel, A.; Yassin, M.; et al. *Nature Communications* 2019, **10**, doi:10.1038/s41467-019-12370-8.
203. Bauer, M.; Su, G.; Casper, C.; He, R.; Rehrauer, W.; Friedl, A. *Oncogene* 2010, **29**, 1732–1740, doi:10.1038/onc.2009.463.
204. Bianchini, G.; Qi, Y.; Alvarez, R.H.; Iwamoto, T.; Coutant, C.; Ibrahim, N.K.; Valero, V.; Cristofanilli, M.; Green, M.C.; Radvanyi, L.; et al. *Journal of Clinical Oncology* 2010, **28**, 4316–4323, doi:10.1200/JCO.2009.27.2419.
205. Calvo, F.; Ege, N.; Grande-Garcia, A.; Hooper, S.; Jenkins, R.P.; Chaudhry, S.I.; Harrington, K.; Williamson, P.; Moeendarbary, E.; Charras, G.; et al. *Nature Cell Biology* 2013, **15**, 637–646, doi:10.1038/ncb2756.
206. Cheng, S.H.-C.; Huang, T.-T.; Cheng, Y.-H.; Tan, T.B.K.; Horng, C.-F.; Wang, Y.A.; Brian, N.S.; Shih, L.-S.; Yu, B.-L. *PloS one* 2017, **12**, e0184372, doi:10.1371/journal.pone.0184372.
207. Jézéquel, P.; Loussouarn, D.; Guérin-Charbonnel, C.; Campion, L.; Vanier, A.; Gouraud, W.; Lasla, H.; Guette, C.; Valo, I.; Verrière, V.; et al. *Breast cancer research : BCR* 2015, **17**, 43, doi:10.1186/s13058-015-0550-y.
208. Sjöström, M.; Staaf, J.; Edén, P.; Wärnberg, F.; Bergh, J.; Malmström, P.; Fernö, M.; Niméus, E.; Fredriksson, I. *Breast cancer research : BCR* 2018, **20**, 64, doi:10.1186/s13058-018-0978-y.
209. Gao, Q.; López-Knowles, E.; U Cheang, M.C.; Morden, J.; Ribas, R.; Sidhu, K.; Evans, D.; Martins, V.; Dodson, A.; Skene, A.; et al. *JNCI Cancer Spectrum* 2018, **2**, doi:10.1093/jncics/pky005.
210. Parker, J.S.; Mullins, M.; Cheang, M.C.U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; et al. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2009, **27**, 1160–1167, doi:10.1200/JCO.2008.18.1370.
211. Kensler, K.H.; Sankar, V.N.; Wang, J.; Zhang, X.; Rubadue, C.A.; Baker, G.M.; Parker, J.S.; Hoadley, K.A.; Stancu, A.L.; Pyle, M.E.; et al. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2019, **28**, 798–806, doi:10.1158/1055-9965.EPI-18-0863.
212. Park, Y.H.; Lal, S.; Lee, J.E.; Choi, Y.-L.; Wen, J.; Ram, S.; Ding, Y.; Lee, S.-H.; Powell, E.; Lee, S.K.; et al. *Nature communications* 2020, **11**, 6175, doi:10.1038/s41467-020-19933-0.
213. Zoeller, J.J.; Press, M.F.; Selfors, L.M.; Dering, J.; Slamon, D.J.; Hurvitz, S.A.; Brugge, J.S. *PloS one* 2021, **16**, e0251163, doi:10.1371/journal.pone.0251163.
214. Kim, S.-K.; Ahn, S.G.; Mun, J.-Y.; Jeong, M.-S.; Bae, S.J.; Lee, J.-S.; Jeong, J.; Leem, S.-H.; Chu, I.-S. *Cancers* 2020, **12**, doi:10.3390/cancers12020497.

215. Anurag, M.; Zhu, M.; Huang, C.; Vasaikar, S.; Wang, J.; Hoog, J.; Burugu, S.; Gao, D.; Suman, V.; Zhang, X.H.; et al. *Journal of the National Cancer Institute* 2020, **112**, 737–746, doi:10.1093/jnci/djz213.
216. Sharma, P.; Barlow, W.E.; Godwin, A.K.; Parkes, E.E.; Knight, L.A.; Walker, S.M.; Kennedy, R.D.; Harkin, D.P.; Logan, G.E.; Steele, C.J.; et al. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2019, **37**, 3484–3492, doi:10.1200/JCO.19.00693.
217. Purrington, K.S.; Knight, J. 3rd; Dyson, G.; Ali-Fehmi, R.; Schwartz, A.G.; Boerner, J.L.; Bandyopadhyay, S. *PloS one* 2020, **15**, e0231712, doi:10.1371/journal.pone.0231712.
218. Wolf, D.M.; Yau, C.; Wulfkühle, J.; Brown-Swigart, L.; Gallagher, R.I.; Magbanua, M.J.M.; O’Grady, N.; Hirst, G.; Asare, S.; Tripathy, D.; et al. *NPJ breast cancer* 2020, **6**, 48, doi:10.1038/s41523-020-00189-2.
219. Heimes, A.-S.; Härtner, F.; Almstedt, K.; Krajnak, S.; Lebrecht, A.; Battista, M.J.; Edlund, K.; Brenner, W.; Hasenburg, A.; Sahin, U.; et al. *International journal of molecular sciences* 2020, **21**, doi:10.3390/ijms21197178.
220. Roepman, P.; Schuurman, A.; Delahaye, L.J.M.J.; Witteveen, A.T.; Floore, A.N.; Glas, A.M. *BMC medical genomics* 2009, **2**, 52, doi:10.1186/1755-8794-2-52.
221. Symmans, W.F.; Hatzis, C.; Sotiriou, C.; Andre, F.; Peintinger, F.; Regitnig, P.; Daxenbichler, G.; Desmedt, C.; Domont, J.; Marth, C.; et al. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2010, **28**, 4111–4119, doi:10.1200/JCO.2010.28.4273.
222. Anders, C.K.; Fan, C.; Parker, J.S.; Carey, L.A.; Blackwell, K.L.; Klauber-DeMore, N.; Perou, C.M. *Breast Carcinomas Arising at a Young Age: Unique Biology or a Surrogate for Aggressive Intrinsic Subtypes?* *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2011, **29**, e18-20.
223. Kao, K.-J.; Chang, K.-M.; Hsu, H.-C.; Huang, A.T. *BMC cancer* 2011, **11**, 143, doi:10.1186/1471-2407-11-143.
224. Liu, R.-Z.; Graham, K.; Glubrecht, D.D.; Germain, D.R.; Mackey, J.R.; Godbout, R. *The American journal of pathology* 2011, **178**, 997–1008, doi:10.1016/j.ajpath.2010.11.075.
225. Kan, Z.; Jaiswal, B.S.; Stinson, J.; Janakiraman, V.; Bhatt, D.; Stern, H.M.; Yue, P.; Haverty, P.M.; Bourgon, R.; Zheng, J.; et al. *Nature* 2010, **466**, 869–873, doi:10.1038/nature09208.
226. de Cremoux, P.; Valet, F.; Gentien, D.; Lehmann-Che, J.; Scott, V.; Tran-Perennou, C.; Barbaroux, C.; Servant, N.; Vacher, S.; Sigal-Zafrani, B.; et al. *BMC Cancer* 2011, **11**, 215, doi:10.1186/1471-2407-11-215.
227. Filipits, M.; Rudas, M.; Jakesz, R.; Dubsky, P.; Fitzal, F.; Singer, C.F.; Dietze, O.; Greil, R.; Jelen, A.; Sevela, P.; et al. *Clinical Cancer Research* 2011, **17**, 6012–6020, doi:10.1158/1078-0432.CCR-11-0926.
228. Sabatier, R.; Finetti, P.; Adelaide, J.; Guille, A.; Borg, J.-P.; Chaffanet, M.; Lane, L.; Birnbaum, D.; Bertucci, F. *PloS one* 2011, **6**, e27656, doi:10.1371/journal.pone.0027656.
229. Magbanua, M.J.M.; Wolf, D.M.; Yau, C.; Davis, S.E.; Crothers, J.; Au, A.; Haqq, C.M.; Livasy, C.; Rugo, H.S.; Esserman, L.; et al. *Breast cancer research : BCR* 2015, **17**, 73, doi:10.1186/s13058-015-0582-3.

230. Caldon, C.E.; Sergio, C.M.; Kang, J.; Muthukaruppan, A.; Boersma, M.N.; Stone, A.; Barraclough, J.; Lee, C.S.; Black, M.A.; Miller, L.D.; et al. *Molecular cancer therapeutics* 2012, **11**, 1488–1499, doi:10.1158/1535-7163.MCT-11-0963.
231. Prat, A.; Adamo, B.; Cheang, M.C.U.; Anders, C.K.; Carey, L.A.; Perou, C.M. *The oncologist* 2013, **18**, 123–133, doi:10.1634/theoncologist.2012-0397.
232. Wang, D.-Y.; Done, S.J.; Mc Cready, D.R.; Leong, W.L. *Breast cancer research : BCR* 2014, **16**, R71, doi:10.1186/bcr3686.
233. Rody, A.; Holtrich, U.; Gaetje, R.; Gehrmann, M.; Engels, K.; von Minckwitz, G.; Loibl, S.; Diallo-Danebrock, R.; Ruckhäberle, E.; Metzler, D.; et al. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2007, **13**, 1115–1122, doi:10.1158/1078-0432.CCR-06-2433.
234. Cunha, S.I.; Bocci, M.; Lötvot, J.; Eleftheriou, N.; Roswall, P.; Cordero, E.; Lindström, L.; Bartoschek, M.; Haller, B.K.; Pearsall, R.S.; et al. *Cancer research* 2015, **75**, 2445–2456, doi:10.1158/0008-5472.CAN-14-3706.
235. Sun, X.; Gierach, G.L.; Sandhu, R.; Williams, T.; Midkiff, B.R.; Lissowska, J.; Wesolowska, E.; Boyd, N.F.; Johnson, N.B.; Figueroa, J.D.; et al. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2013, **19**, 4972–4982, doi:10.1158/1078-0432.CCR-13-0029.
236. Larsen, M.J.; Thomassen, M.; Tan, Q.; Lænkholm, A.-V.; Bak, M.; Sørensen, K.P.; Andersen, M.K.; Kruse, T.A.; Gerdes, A.-M. *BMC medical genomics* 2014, **7**, 9, doi:10.1186/1755-8794-7-9.
237. Tan, T.Z.; Miow, Q.H.; Miki, Y.; Noda, T.; Mori, S.; Huang, R.Y.-J.; Thiery, J.P. *EMBO molecular medicine* 2014, **6**, 1279–1293, doi:10.15252/emmm.201404208.
238. Nagel, J.H.A.; Peeters, J.K.; Smid, M.; Sieuwerts, A.M.; Wasielewski, M.; de Weerd, V.; Trapman-Jansen, A.M.A.C.; van den Ouweland, A.; Brüggewirth, H.; van I Jcken, W.F.J.; et al. *Breast cancer research and treatment* 2012, **132**, 439–448, doi:10.1007/s10549-011-1588-x.
239. Tinholt, M.; Volla, H.K.M.; Sahlberg, K.K.; Jernström, S.; Kaveh, F.; Lingjærde, O.C.; Kåresen, R.; Sauer, T.; Kristensen, V.; Børresen-Dale, A.-L.; et al. *Breast cancer research : BCR* 2015, **17**, 44, doi:10.1186/s13058-015-0548-5.
240. Maire, V.; Némati, F.; Richardson, M.; Vincent-Salomon, A.; Tesson, B.; Rigai, G.; Gravier, E.; Marty-Prouvost, B.; De Koning, L.; Lang, G.; et al. *Cancer research* 2013, **73**, 813–823, doi:10.1158/0008-5472.CAN-12-2633.
241. Quigley, D.A.; Tahiri, A.; Lüders, T.; Riis, M.H.; Balmain, A.; Børresen-Dale, A.-L.; Bukholm, I.; Kristensen, V. *Oncoimmunology* 2017, **6**, e1356142, doi:10.1080/2162402X.2017.1356142.
242. Jiang, Y.-Z.; Liu, Y.-R.; Xu, X.-E.; Jin, X.; Hu, X.; Yu, K.-D.; Shao, Z.-M. *Cancer research* 2016, **76**, 2105–2114, doi:10.1158/0008-5472.CAN-15-3284.
243. Burstein, M.D.; Tsimelzon, A.; Poage, G.M.; Covington, K.R.; Contreras, A.; Fuqua, S.A.W.; Savage, M.I.; Osborne, C.K.; Hilsenbeck, S.G.; Chang, J.C.; et al. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2015, **21**, 1688–1698, doi:10.1158/1078-0432.CCR-14-0432.
244. Toro, A.L.; Costantino, N.S.; Shriver, C.D.; Ellsworth, D.L.; Ellsworth, R.E. *BMC obesity* 2016, **3**, 22, doi:10.1186/s40608-016-0103-7.

245. Brueffer, C.; Vallon-Christersson, J.; Grabau, D.; Ehinger, A.; Häkkinen, J.; Hegardt, C.; Malina, J.; Chen, Y.; Bendahl, P.-O.; Manjer, J.; et al. *JCO precision oncology* 2018, **2**, doi:10.1200/PO.17.00135.
246. Yau, C.; Fedele, V.; Roydasgupta, R.; Fridlyand, J.; Hubbard, A.; Gray, J.W.; Chew, K.; Dairkee, S.H.; Moore, D.H.; Schittulli, F.; et al. *Breast cancer research : BCR* 2007, **9**, R59, doi:10.1186/bcr1765.
247. Prabhakaran, S.; Rizk, V.T.; Ma, Z.; Cheng, C.-H.; Berglund, A.E.; Coppola, D.; Khalil, F.; Mulé, J.J.; Soliman, H.H. *Breast cancer research : BCR* 2017, **19**, 71, doi:10.1186/s13058-017-0864-z.
248. Romero-Cordoba, S.L.; Rodriguez-Cuevas, S.; Bautista-Pina, V.; Maffuz-Aziz, A.; D'Ippolito, E.; Cosentino, G.; Baroni, S.; Iorio, M. V.; Hidalgo-Miranda, A. *Scientific reports* 2018, **8**, 12252, doi:10.1038/s41598-018-29708-9.
249. Romero-Cordoba, S.L.; Salido-Guadarrama, I.; Rebollar-Vega, R.; Bautista-Piña, V.; Dominguez-Reyes, C.; Tenorio-Torres, A.; Villegas-Carlos, F.; Fernández-López, J.C.; Uribe-Figueroa, L.; Alfaro-Ruiz, L.; et al. *Nature communications* 2021, **12**, 2245, doi:10.1038/s41467-021-22478-5.
250. Ellis, M.J.; Suman, V.J.; Hoog, J.; Goncalves, R.; Sanati, S.; Creighton, C.J.; DeSchryver, K.; Crouch, E.; Brink, A.; Watson, M.; et al. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2017, **35**, 1061–1069, doi:10.1200/JCO.2016.69.4406.
251. Palazon, A.; Tyrakis, P.A.; Macias, D.; Veliça, P.; Rundqvist, H.; Fitzpatrick, S.; Vojnovic, N.; Phan, A.T.; Loman, N.; Hedenfalk, I.; et al. *Cancer cell* 2017, **32**, 669–683.e5, doi:10.1016/j.ccell.2017.10.003.
252. Wang, J.; Zhang, X.; Beck, A.H.; Collins, L.C.; Chen, W.Y.; Tamimi, R.M.; Hazra, A.; Brown, M.; Rosner, B.; Hankinson, S.E. *Hormones & cancer* 2015, **6**, 237–246, doi:10.1007/s12672-015-0235-0.
253. Kang, H.J.; Kawasawa, Y.I.; Cheng, F.; Zhu, Y.; Xu, X.; Li, M.; Sousa, A.M.M.; Pletikos, M.; Meyer, K.A.; Sedmak, G.; et al. *Nature* 2011, **478**, 483–489, doi:10.1038/nature10523.
254. Pal, B.; Chen, Y.; Vaillant, F.; Capaldo, B.D.; Joyce, R.; Song, X.; Bryant, V.L.; Penington, J.S.; Di Stefano, L.; Tubau Ribera, N.; et al. *The EMBO journal* 2021, **40**, e107333, doi:10.15252/embj.2020107333.
255. Yuan, X.; Wu, H.; Xu, H.; Xiong, H.; Chu, Q.; Yu, S.; Wu, G.S.; Wu, K. *Cancer Letters* 2015, **369**, 20–27, doi:https://doi.org/10.1016/j.canlet.2015.07.048.
256. Masjedi, A.; Hashemi, V.; Hojjat-Farsangi, M.; Ghalamfarsa, G.; Azizi, G.; Yousefi, M.; Jadidi-Niaragh, F. *Biomedicine & Pharmacotherapy* 2018, **108**, 1415–1424, doi:https://doi.org/10.1016/j.biopha.2018.09.177.
257. Tye, S.L.; Gilg, A.G.; Tolliver, L.B.; Wheeler, W.G.; Toole, B.P.; Maria, B.L. *Journal of child neurology* 2008, **23**, 1221–1230, doi:10.1177/0883073808321066.
258. Matsuoka, R.; Shiba-Ishii, A.; Nakano, N.; Togayachi, A.; Sakashita, S.; Sato, Y.; Minami, Y.; Noguchi, M. *Lung Cancer* 2018, **118**, 97–104, doi:10.1016/j.lungcan.2018.01.012.
259. Arner, E.; Forrest, A.R.R.; Ehrlund, A.; Mejhert, N.; Itoh, M.; Kawaji, H.; Lassmann, T.; Laurencikiene, J.; Rydén, M.; Arner, P. *PloS one* 2014, **9**, e80274, doi:10.1371/journal.pone.0080274.
260. Banha, J.; Marques, L.; Oliveira, R.; Martins, M. de F.; Paixão, E.; Pereira, D.; Malhó, R.; Penque, D.; Costa, L. *Free radical biology & medicine* 2008, **44**, 483–492, doi:10.1016/j.freeradbiomed.2007.10.032.

261. Jarczak, J.; Kościuczuk, E.M.; Lisowski, P.; Strzałkowska, N.; Józwik, A.; Horbańczuk, J.; Krzyżewski, J.; Zwierzchowski, L.; Bagnicka, E. *Human Immunology* 2013, **74**, 1069–1079, doi:<https://doi.org/10.1016/j.humimm.2013.05.008>.
262. Jin, G.; Weinberg, A. *Seminars in Cell & Developmental Biology* 2019, **88**, 156–162, doi:<https://doi.org/10.1016/j.semcdb.2018.04.006>.
263. Yang, D.; Chertov, O.; Bykovskaia, S.N.; Chen, Q.; Buffo, M.J.; Shogan, J.; Anderson, M.; Schröder, J.M.; Wang, J.M.; Howard, O.M.; et al. *Science (New York, N.Y.)* 1999, **286**, 525–528, doi:10.1126/science.286.5439.525.
264. Ghosh, S.K.; McCormick, T.S.; Weinberg, A. *Frontiers in Oncology* 2019, **9**, 1–8, doi:10.3389/fonc.2019.00341.
265. Prado-Montes de Oca, E. *The international journal of biochemistry & cell biology* 2010, **42**, 800–804, doi:10.1016/j.biocel.2010.01.021.
266. Sanchez-Vega, F.; Mina, M.; Armenia, J.; Chatila, W.K.; Luna, A.; La, K.C.; Dimitriadou, S.; Liu, D.L.; Kantheti, H.S.; Saghafeinia, S.; et al. *Cell* 2018, **173**, 321–337.e10, doi:10.1016/j.cell.2018.03.035.
267. Guba, M.; Bosserhoff, A.-K.; Steinbauer, M.; Abels, C.; Anthuber, M.; Buettner, R.; Jauch, K.-W. *British Journal of Cancer* 2000, **83**, 1216–1222, doi:10.1054/bjoc.2000.1424.
268. El Fitori, J.; Kleeff, J.; Giese, N.A.; Guweidhi, A.; Bosserhoff, A.K.; Büchler, M.W.; Friess, H. *Cancer cell international* 2005, **5**, 3, doi:10.1186/1475-2867-5-3.
269. Nelson, J.; Bagnato, A.; Battistini, B.; Nisen, P. *Nature reviews. Cancer* 2003, **3**, 110–116, doi:10.1038/nrc990.
270. Kusuvara, M.; Yamaguchi, K.; Nagasaki, K.; Hayashi, C.; Suzaki, A.; Hori, S.; Handa, S.; Nakamura, Y.; Abe, K. 1990, 3257–3262.
271. Bagnato, A.; Spinella, F. *Trends in Endocrinology & Metabolism* 2003, **14**, 44–50.
272. Grant, K.; Loizidou, M.; Taylor, I. *British journal of cancer* 2003, **88**, 163–166, doi:10.1038/sj.bjc.6700750.
273. Spinella, F.; Caprara, V.; Cianfrocca, R.; Rosanò, L.; Di Castro, V.; Garrafa, E.; Natali, P.G.; Bagnato, A. *Carcinogenesis* 2014, **35**, 840–848.
274. Hinsley, E.E.; Kumar, S.; Hunter, K.D.; Whawell, S.A.; Lambert, D.W. *Life sciences* 2012, **91**, 557–561.
275. Clines, G.A.; Mohammad, K.S.; Bao, Y.; Stephens, O.W.; Suva, L.J.; Shaughnessy, J.D.J.; Fox, J.W.; Chirgwin, J.M.; Guise, T.A. *Molecular endocrinology (Baltimore, Md.)* 2007, **21**, 486–498, doi:10.1210/me.2006-0346.
276. Recchia, A.G.; Filice, E.; Pellegrino, D.; Dobrina, A.; Cerra, M.C.; Maggiolini, M. *Journal of molecular and cellular cardiology* 2009, **46**, 352–359, doi:10.1016/j.yjmcc.2008.11.017.
277. Rodriguez-Vita, J.; Ruiz-Ortega, M.; Rupérez, M.; Esteban, V.; Sanchez-López, E.; Plaza, J.J.; Egido, J. *Circulation research* 2005, **97**, 125–134, doi:10.1161/01.RES.0000174614.74469.83.
278. Weng, C.-M.; Yu, C.-C.; Kuo, M.-L.; Chen, B.-C.; Lin, C.-H. *Biochemical pharmacology* 2014, **88**, 402–411, doi:10.1016/j.bcp.2014.01.030.
279. Park, H.J.; Kim, M.K.; Choi, K.S.; Jeong, J.W.; Bae, S.K.; Kim, H.J.; Bae, M.K. *International Journal of Oncology* 2016, **49**, 934–942, doi:10.3892/ijo.2016.3590.

280. Kzhyskowska, J.; Larionova, I.; Liu, T. YKL-39 as a Potential New Target for Anti-Angiogenic Therapy in Cancer . *Frontiers in Immunology* 2020, **10**, 2930.
281. Kita, A.; Kasamatsu, A.; Nakashima, D.; Endo-Sakamoto, Y.; Ishida, S.; Shimizu, T.; Kimura, Y.; Miyamoto, I.; Yoshimura, S.; Shiiba, M.; et al. *Journal of Cancer* 2017, **8**, 2033–2041, doi:10.7150/jca.18714.
282. Listik, E.; Horst, B.; Choi, A.S.; Lee, Nam.Y.; Györfy, B.; Mythreye, K. *PLOS ONE* 2021, **16**, e0249558.
283. Moidu, N.A.; A Rahman, N.S.; Syafruddin, S.E.; Low, T.Y.; Mohtar, M.A. *Heliyon* 2020, **6**, e05000, doi:10.1016/j.heliyon.2020.e05000.
284. Tsuji, T.; Satoyoshi, R.; Aiba, N.; Kubo, T.; Yanagihara, K.; Maeda, D.; Goto, A.; Ishikawa, K.; Yashiro, M.; Tanaka, M. *Cancer research* 2015, **75**, 356–366, doi:10.1158/0008-5472.CAN-14-1693.
285. Merugu, S.B.; Zhou, B.; Mangukiya, H.B.; Negi, H.; Ghulam, R.; Roy, D.; Qudsia, S.; Wang, Z.; Mashausi, D.S.; Yunus, F.-U.-N.; et al. *Biochemical and Biophysical Research Communications* 2021, **573**, 86–92, doi:https://doi.org/10.1016/j.bbrc.2021.08.028.
286. Mangukiya, H.B.; Negi, H.; Merugu, S.B.; Sehar, Q.; Mashausi, D.S.; Yunus, F.-U.-N.; Wu, Z.; Li, D. *Cell adhesion & migration* 2019, **13**, 332–344, doi:10.1080/19336918.2019.1685928.
287. Nugteren, S.; Samsom, J.N. *Cytokine & Growth Factor Reviews* 2021, **59**, 22–35, doi:https://doi.org/10.1016/j.cytogfr.2021.01.005.
288. Willrodt, A.H.; Beffinger, M.; Vranova, M.; Protsyuk, D.; Schuler, K.; Jadhav, M.; Heikenwalder, M.; van den Broek, M.; Borsig, L.; Halin, C. *American Journal of Pathology* 2017, **187**, 2558–2569, doi:10.1016/j.ajpath.2017.07.008.
289. Du, X.Y.; Liu, X.; Wang, Z.J.; Wang, Y.Y. *European review for medical and pharmacological sciences* 2017, **21**, 1495–1501.
290. Fujiwara, K.; Ohuchida, K.; Sada, M.; Horioka, K.; Ulrich III, C.D.; Shindo, K.; Ohtsuka, T.; Takahata, S.; Mizumoto, K.; Oda, Y.; et al. *PLOS ONE* 2014, **9**, e107247.
291. Sayers, K.T.; Brooks, A.D.; Sayers, T.J.; Chertov, O. *PLoS ONE* 2014, **9**, 1–7, doi:10.1371/journal.pone.0104223.
292. Antonov, A.; Agostini, M.; Morello, M.; Minieri, M. *Oncotarget* 2014, **5**, doi:10.18632/oncotarget.2668.
293. Deberardinis, R.J. *Cell Metabolism* 2011, **14**, 285–286, doi:10.1016/j.cmet.2011.08.004.
294. Possemato, R.; Marks, K.M.; Shaul, Y.D.; Pacold, M.E.; Kim, D.; Birsoy, K.; Sethumadhavan, S.; Woo, H.; Jang, H.G.; Jha, A.K.; et al. *Nature* 2011, **476**, 346–350, doi:10.1038/nature10350.
295. Gao, S.; Ge, A.; Xu, S.; You, Z.; Ning, S.; Zhao, Y.; Pang, D. *Journal of Experimental & Clinical Cancer Research* 2017, **36**, 1–13, doi:10.1186/s13046-017-0648-4.
296. Marchi, T. De; Timmermans, M.A.; Sieuwerts, A.M.; Smid, M.; Look, M.P.; Greben, N.; Sweep, F.C.G.J.; Smits, J.G.; Deurzen, C.H.M. Van; Foekens, J.A.; et al. *Scientific Reports* 2017, 1–11, doi:10.1038/s41598-017-02296-w.
297. Yousefi, M.; Nosrati, R.; Salmaninejad, A.; Dehghani, S.; Shahryari, A.; Saberi, A. *Cellular Oncology* 2018, **41**, 123–140, doi:10.1007/s13402-018-0376-6.

298. Smid, M.; Wang, Y.; Zhang, Y.; Sieuwerts, A.M.; Yu, J.; Klijn, J.G.M.; Foekens, J.A.; Martens, J.W.M. *Cancer Research* 2008, **68**, 3108 LP – 3114, doi:10.1158/0008-5472.CAN-07-5644.
299. Kennecke, H.; Yerushalmi, R.; Woods, R.; Cheang, M.C.U.; Voduc, D.; Speers, C.H.; Nielsen, T.O.; Gelmon, K. *Journal of Clinical Oncology* 2010, **28**, 3271–3277, doi:10.1200/JCO.2009.25.9820.
300. Marinello, J.; Delcuratolo, M.; Capranico, G. Anthracyclines as Topoisomerase II Poisons: From Early Studies to New Perspectives. *International Journal of Molecular Sciences* 2018, **19**.
301. Lane, D.J.R.; Bae, D.-H.; Merlot, A.M.; Sahni, S.; Richardson, D.R. *Nutrients* 2015, **7**, 2274–2296, doi:10.3390/nu7042274.
302. Lemler, D.J.; Lynch, M.L.; Tesfay, L.; Deng, Z.; Paul, B.T.; Wang, X.; Hegde, P.; Manz, D.H.; Torti, S. V; Torti, F.M. *Breast Cancer Research* 2017, **19**, 25, doi:10.1186/s13058-017-0814-9.
303. Ohta, M.; Tanaka, F.; Yamaguchi, H.; Sadanaga, N.; Inoue, H.; Mori, M. *International journal of oncology* 2005, **26**, 41–47.
304. Nukiwa, M.; Andarini, S.; Zaini, J.; Xin, H.; Kanehira, M.; Suzuki, T.; Fukuhara, T.; Mizuguchi, H.; Hayakawa, T.; Saijo, Y.; et al. *European journal of immunology* 2006, **36**, 1019–1027, doi:10.1002/eji.200535549.
305. Reyal, F.; van Vliet, M.H.; Armstrong, N.J.; Horlings, H.M.; de Visser, K.E.; Kok, M.; Teschendorff, A.E.; Mook, S.; van 't Veer, L.; Caldas, C.; et al. *Breast Cancer Res* 2008, **10**, R93, doi:10.1186/bcr2192.
306. Zhao, X.; Rødland, E.A.; Sørli, T.; Volla, H.K.M.; Russnes, H.G.; Kristensen, V.N.; Lingjærde, O.C.; Børresen-Dale, A.-L. *BMC cancer* 2014, **14**, 211, doi:10.1186/1471-2407-14-211.
307. Rueda, O.M.; Sammut, S.J.; Seoane, J.A.; Chin, S.F.; Caswell-Jin, J.L.; Callari, M.; Batra, R.; Pereira, B.; Bruna, A.; Ali, H.R.; et al. *Nature* 2019, **567**, 399–404, doi:10.1038/s41586-019-1007-8.
308. Sestak, I.; Buus, R.; Cuzick, J.; Dubsy, P.; Kronenwett, R.; Denkert, C.; Ferree, S.; Sgroi, D.; Schnabel, C.; Baehner, F.L.; et al. *JAMA Oncology* 2018, **4**, 545–553, doi:10.1001/jamaoncol.2017.5524.
309. Wang, Z.; Li, T.E.; Chen, M.; Pan, J.J.; Shen, K.W. *Aging* 2020, **12**, 1867–1887, doi:10.18632/aging.102719.
310. Li, N.; Liu, Y.; Miao, Y.; Zhao, L.; Zhou, H.; Jia, L. *IUBMB Life* 2016, 764–775, doi:10.1002/iub.1541.
311. Zheng, R.; Pan, L.; Gao, J.; Ye, X.; Chen, L.; Zhang, X.; Tang, W.; Zheng, W. *Journal of Surgical Research* 2015, **195**, 158–165, doi:10.1016/j.jss.2014.12.035.
312. Li, N.; Miao, Y.; Shan, Y.; Liu, B.; Li, Y.; Zhao, L.; Jia, L. *Cell Death and Disease* 2017, **8**, e2796-11, doi:10.1038/cddis.2017.119.
313. Savino, A.; Avale, L.; Monteleone, E.; Miglio, I.; Griffo, A.; Accetta, G.; Provero, P.; Poli, V. *bioRxiv* 2019, 570051, doi:10.1101/570051.
314. Huggins, C.J.; Mayekar, M.K.; Martin, N.; Saylor, K.L.; Gonit, M.; Jailwala, P.; Kasoji, M.; Haines, D.C.; Quiñones, O.A.; Johnson, F. *Molecular and Cellular Biology* 2016, **36**, 693–713, doi:10.1128/MCB.00911-15.
315. Vimala, K.; Sundarraj, S.; Sujitha, M. V; Kannan, S. *Archives of Medical Research* 2012, **43**, 415–422, doi:10.1016/j.arcmed.2012.08.009.

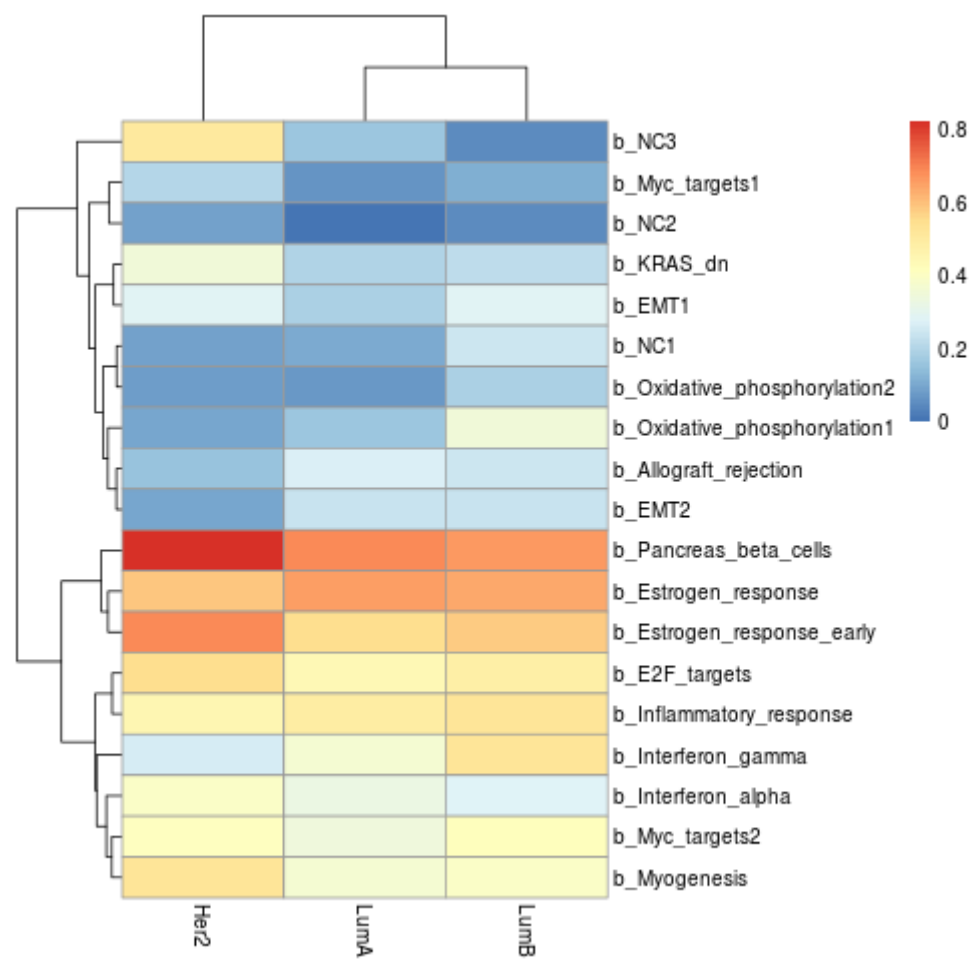
316. Melchor, L.; Saucedo-cuevas, L.P.; Muñoz-repeto, I.; Rodríguez-pinilla, M.; Honrado, E.; Campoverde, A.; Palacios, J.; Nathanson, K.L.; García, M.J.; Benítez, J. *Breast Cancer Research* 2009, **11**, 1–14, doi:10.1186/bcr2456.
317. Wang, C.; Nie, Z.; Zhou, Z.; Zhang, H.; Liu, R.; Qin, J.; Ma, Y.; Chen, L.; Li, S.; Chen, W.; et al. *Oncotarget* 2015, **6**, doi:10.18632/oncotarget.3779.
318. Yoon, C.; Kim, M.; Lee, H.; Kim, R.; Lim, E.; Yoo, K.; Lee, G.; Cui, Y.; Oh, Y.S.; Gye, M.C.; et al. *Journal of Biological Chemistry* 2012, **287**, 19516–19527, doi:10.1074/jbc.M111.337428.
319. Savino, A.; Provero, P.; Poli, V. *International Journal of Molecular Sciences* 2020, **21**, 1–23, doi:10.3390/ijms21249461.
320. Galfrè, S.G.; Morandin, F.; Pietrosanto, M.; Cremisi, F.; Helmer-Citterich, M. *NAR genomics and bioinformatics* 2021, **3**, lqab072, doi:10.1093/nargab/lqab072.
321. Lopes-Coelho, F.; André, S.; Félix, A.; Serpa, J. *Molecular and cellular endocrinology* 2018, **462**, 93–106, doi:10.1016/j.mce.2017.01.031.
322. Yuzhalin, A.E.; Lim, S.Y.; Kutikhin, A.G.; Gordon-Weeks, A.N. *Biochimica et Biophysica Acta - Reviews on Cancer* 2018, **1870**, 207–228, doi:10.1016/j.bbcan.2018.09.002.
323. Ashton, T.M.; Gillies McKenna, W.; Kunz-Schughart, L.A.; Higgins, G.S. *Clinical Cancer Research* 2018, **24**, 2482–2490, doi:10.1158/1078-0432.CCR-17-3070.
324. Davis, R.T.; Blake, K.; Ma, D.; Gabra, M.B.I.; Hernandez, G.A.; Phung, A.T.; Yang, Y.; Maurer, D.; Lefebvre, A.E.Y.T.; Alshetaiwi, H.; et al. *Nature Cell Biology* 2020, **22**, 310–320, doi:10.1038/s41556-020-0477-0.
325. Boneberg, E.-M.; Legler, D.F.; Hoefer, M.M.; Ohlschlegel, C.; Steininger, H.; Füzesi, L.; Beer, G.M.; Dupont-Lampert, V.; Otto, F.; Senn, H.-J.; et al. *British journal of cancer* 2009, **101**, 605–614, doi:10.1038/sj.bjc.6605219.
326. Eberhard, A.; Kahlert, S.; Goede, V.; Hemmerlein, B.; Plate, K.H.; Augustin, H.G. *Cancer research* 2000, **60**, 1388–1393.
327. Al Sorkhy, M.; Fahl, Z.; Ritchie, J. *Annals of Cancer Research and Therapy* 2018, **26**, 19–25, doi:10.4993/acrt.26.19.
328. Shaul, Y.D.; Freinkman, E.; Comb, W.C.; Cantor, J.R.; Tam, W.L.; Thiru, P.; Kim, D.; Kanarek, N.; Pacold, M.E.; Chen, W.W.; et al. *Cell* 2014, **158**, 1094–1109, doi:10.1016/j.cell.2014.07.032.
329. Chamberland, J.P.; Antonow, L.T.; Dias Santos, M.; Ritter, B. *Journal of Cell Science* 2016, **129**, 2625–2637, doi:10.1242/jcs.173708.
330. Zheng, X.; Weigert, A.; Reu, S.; Guenther, S.; Mansouri, S.; Bassaly, B.; Gattenlöhner, S.; Grimminger, F.; Pullamsetti, S.S.; Seeger, W.; et al. *Cancer Research* 2020, **80**, 4414–4425, doi:10.1158/0008-5472.CAN-20-0069.
331. Medrek, C.; Pontén, F.; Jirström, K.; Leandersson, K. *BMC Cancer* 2012, **12**, 1–9, doi:10.1186/1471-2407-12-306.
332. Jackute, J.; Zemaitis, M.; Pranys, D.; Sitkauskienė, B.; Miliauskas, S.; Vaitkiene, S.; Sakalauskas, R. *BMC Immunology* 2018, **19**, 1–13, doi:10.1186/s12865-018-0241-4.
333. R Core Team R: A Language and Environment for Statistical Computing 2018.

334. Wickham, H. Journal of Statistical Software April 2017, **77**, 3–5, doi:10.18637/jss.v077.b02.
335. Ahlmann-Eltze, C. 2021, doi:10.5281/zenodo.2074621.
336. Linlin, Y. Ggvenn: Draw Venn Diagram by “Ggplot2” 2021.
337. Alboukadel, K.; Marcin, K.; Przemyslaw, B.; Scheipl, F. Survminer: Drawing Survival Curves Using “Ggplot2” 2020.
338. Kassambara, A. Ggpubr: “Ggplot2” Based Publication Ready Plots (Version 0.3. 0)[Computer Software] 2020.
339. Raivo, K. Pheatmap: Pretty Heatmaps 2021.
340. Gentleman, R.C.; Carey, V.J.; Bates, D.M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Genome Biology 2004, **5**.
341. Davis, S.; Meltzer, P.S. Bioinformatics (Oxford, England) 2007, **23**, 1846–1847, doi:10.1093/bioinformatics/btm254.
342. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. Nucleic Acids Research 2016, **44**, e71–e71, doi:10.1093/nar/gkv1507.
343. Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Nature protocols 2009, **4**, 1184–91, doi:10.1038/nprot.2009.97.
344. van de Vijver, M.J.; He, Y.D.; van’t Veer, L.J.; Dai, H.; Hart, A.A.M.; Voskuil, D.W.; Schreiber, G.J.; Peterse, J.L.; Roberts, C.; Marton, M.J.; et al. The New England journal of medicine 2002, **347**, 1999–2009, doi:10.1056/NEJMoa021967.
345. Schmidt, M.; Böhm, D.; von Törne, C.; Steiner, E.; Puhl, A.; Pilch, H.; Lehr, H.-A.; Hengstler, J.G.; Kölbl, H.; Gehrmann, M. Cancer research 2008, **68**, 5405–5413, doi:10.1158/0008-5472.CAN-07-5206.
346. Sotiriou, C.; Wirapati, P.; Loi, S.; Harris, A.; Fox, S.; Smeds, J.; Nordgren, H.; Farmer, P.; Praz, V.; Haibe-Kains, B.; et al. Journal of the National Cancer Institute 2006, doi:10.1093/jnci/djj052.
347. Konishi, H.; Vassilev, A.; Kurooka, H.; Sawada, J.; Ikura, T.; Stanley, J.; Pawitan, Y.; Hall, P.; Klaar, S.; Edison, T.; et al. Proceedings of the National Academy of Sciences of the United States of America 2005, **102**, 13550–5, doi:10.1073/pnas.0506230102.
348. Desmedt, C.; Piette, F.; Loi, S.; Wang, Y.; Lallemand, F.; Haibe-Kains, B.; Viale, G.; Delorenzi, M.; Zhang, Y.; d’Assignies, M.S.; et al. Clinical cancer research : an official journal of the American Association for Cancer Research 2007, **13**, 3207–3214, doi:10.1158/1078-0432.CCR-06-2765.
349. Haibe-Kains, B.; Desmedt, C.; Loi, S.; Culhane, A.C.; Bontempi, G.; Quackenbush, J.; Sotiriou, C. JNCI: Journal of the National Cancer Institute 2012, **104**, 311–325, doi:10.1093/jnci/djr545.
350. Arnold, T.B.; Emerson, J.W. R Journal 2011, **3**, 34–39, doi:10.32614/rj-2011-016.
351. Korotkevich, G.; Sukhov, V.; Budin, N.; Shpak, B.; Artyomov, M.N.; Sergushichev, A. bioRxiv 2021, 60012, doi:10.1101/060012.
352. Mohamed, M.; Abdelaal, A.; Hossam, S.; Ahmed, E. American Journal of Theoretical and Applied Statistics 2015, **4**, 504–512, doi:10.11648/j.ajtas.20150406.21.

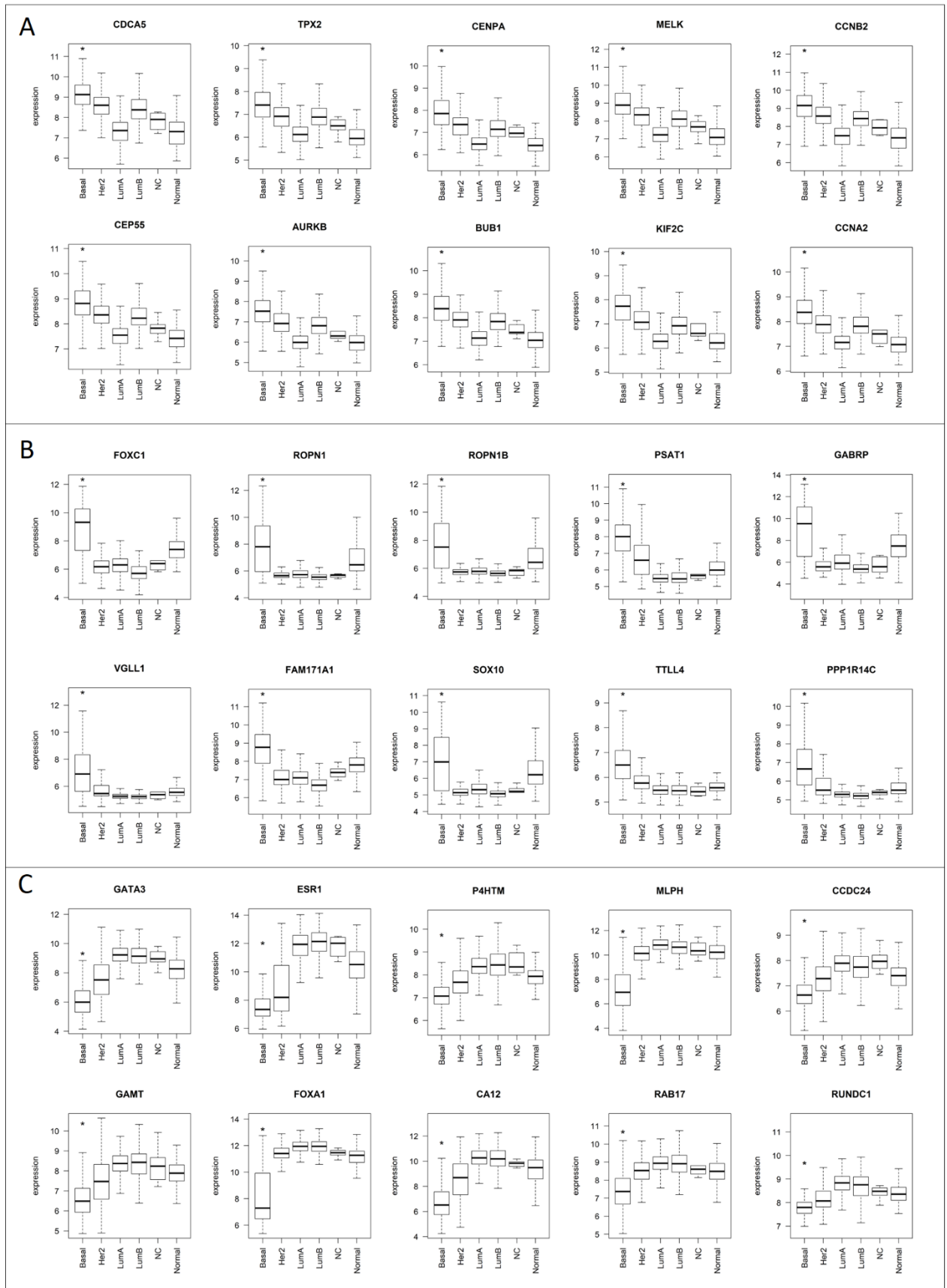
353. Kimbung, S.; Markholm, I.; Bjöhle, J.; Lekberg, T.; von Wachenfeldt, A.; Azavedo, E.; Saracco, A.; Hellström, M.; Veerla, S.; Paquet, E.; et al. *International journal of cancer* 2018, **142**, 618–628, doi:10.1002/ijc.31070.
354. Esserman, L.J.; Berry, D.A.; Cheang, M.C.U.; Yau, C.; Perou, C.M.; Carey, L.; DeMichele, A.; Gray, J.W.; Conway-Dorsey, K.; Lenburg, M.E.; et al. *Breast cancer research and treatment* 2012, **132**, 1049–1062, doi:10.1007/s10549-011-1895-2.
355. Desmedt, C.; Di Leo, A.; de Azambuja, E.; Larsimont, D.; Haibe-Kains, B.; Selleslags, J.; Delaloge, S.; Duhem, C.; Kains, J.-P.; Carly, B.; et al. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2011, **29**, 1578–1586, doi:10.1200/JCO.2010.31.2231.
356. Hatzis, C.; Pusztai, L.; Valero, V.; Booser, D.J.; Esserman, L.; Lluch, A.; Vidaurre, T.; Holmes, F.; Souchon, E.; Wang, H.; et al. *JAMA* 2011, **305**, 1873–1881, doi:10.1001/jama.2011.593.
357. Braunstein, M.; Liao, L.; Lyttle, N.; Lobo, N.; Taylor, K.J.; Krzyzanowski, P.M.; Kalatskaya, I.; Yao, C.Q.; Stein, L.D.; Boutros, P.C.; et al. *Breast cancer research : BCR* 2016, **18**, 16, doi:10.1186/s13058-016-0676-6.
358. Lanceta, L.; O'Neill, C.; Lypova, N.; Li, X.; Rouchka, E.; Waigel, S.; Gomez-Gutierrez, J.G.; Chesney, J.; Imbert-Fernandez, Y. *Genes* 2020, **11**, doi:10.3390/genes11040467.
359. Dalmases, A.; González, I.; Menendez, S.; Arpí, O.; Corominas, J.M.; Servitja, S.; Tusquets, I.; Chamizo, C.; Rincón, R.; Espinosa, L.; et al. *Oncotarget* 2014, **5**, 196–210, doi:10.18632/oncotarget.1556.
360. Zhang, M.; Wang, Y.; Jiang, L.; Song, X.; Zheng, A.; Gao, H.; Wei, M.; Zhao, L. *Journal of experimental & clinical cancer research : CR* 2021, **40**, 41, doi:10.1186/s13046-021-01844-7.
361. Ghandi, M.; Huang, F.W.; Jané-Valbuena, J.; Kryukov, G. V.; Lo, C.C.; McDonald, E.R.; Barretina, J.; Gelfand, E.T.; Bielski, C.M.; Li, H.; et al. *Nature* 2019, **569**, 503–508, doi:10.1038/s41586-019-1186-3.
362. Hänzelmann, S.; Castelo, R.; Guinney, J. *BMC Bioinformatics* 2013, **14**, 7, doi:10.1186/1471-2105-14-7.
363. Friedman, J.; Hastie, T.; Tibshirani, R. *Journal of statistical software* 2010, **33**, 1.
364. Schröder, M.S.; Culhane, A.C.; Quackenbush, J.; Haibe-Kains, B. *Bioinformatics (Oxford, England)* 2011, **27**, 3206–3208, doi:10.1093/bioinformatics/btr511.
365. Bayerlová, M.; Klemm, F.; Kramer, F.; Pukrop, T. *PLoS ONE* 2015, **10**, 1–19, doi:10.1371/journal.pone.0144014.
366. Tell, R.W.; Horvath, C.M. *Proc Natl Acad Sci U S A* 2014, **111**, 12787–12792, doi:10.1073/pnas.1404881111.
367. Love, M.I.; Huber, W.; Anders, S. *Genome Biology* 2014, **15**, 550, doi:10.1186/s13059-014-0550-8.
368. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. *Genome Biology* 2009, **10**, doi:10.1186/gb-2009-10-3-r25.
369. Zhang, Y.; Liu, T.; Meyer, C.A.; Eeckhoutte, J.; Johnson, D.S.; Bernstein, B.E.; Nusbaum, C.; Myers, R.M.; Brown, M.; Li, W.; et al. *Genome Biology* 2008, doi:10.1186/gb-2008-9-9-r137.
370. Quinlan, A.R.; Hall, I.M. *Bioinformatics* 2010, **26**, 841–842, doi:10.1093/bioinformatics/btq033.
371. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010.

372. Kim, D.; Paggi, J.M.; Park, C.; Bennett, C.; Salzberg, S.L. *Nature Biotechnology* 2019, **37**, 907–915, doi:10.1038/s41587-019-0201-4.
373. Anders, S.; Pyl, P.T.; Huber, W. *Bioinformatics (Oxford, England)* 2015, **31**, 166–169, doi:10.1093/bioinformatics/btu638.
374. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. *Genome Research* 2003, **13**, 2498–2504, doi:10.1101/gr.1239303.metabolite.
375. Bastian, M.; Heymann, S. *AAAI* 2009.
376. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. *Nucleic Acids Research* 2015, **43**, e47–e47, doi:10.1093/nar/gkv007.
377. Frank, H. *Hmisc: Harrell Miscellaneous* 2021.
378. Therneau, T. *A Package for Survival Analysis in R*.
379. Yan, W.; Wu, X.; Zhou, W.; Fong, M.Y.; Cao, M.; Liu, J.; Liu, X.; Chen, C.-H.; Fadare, O.; Pizzo, D.P.; et al. *Nature cell biology* 2018, **20**, 597–609, doi:10.1038/s41556-018-0083-6.
380. Matsumura, Y.; Ito, Y.; Mezawa, Y.; Sulidan, K.; Daigo, Y.; Hiraga, T.; Mogushi, K.; Wali, N.; Suzuki, H.; Itoh, T.; et al. *Life science alliance* 2019, **2**, doi:10.26508/lsa.201900425.
381. Luciani, M.G.; Seok, J.; Sayeed, A.; Champion, S.; Goodson, W.H.; Jeffrey, S.S.; Xiao, W.; Mindrinos, M.; Davis, R.W.; Dairkee, S.H. *PLoS ONE* 2011, **6**, doi:10.1371/journal.pone.0020016.
382. Desmedt, C.; Majjaj, S.; Kheddoumi, N.; Singhal, S.K.; Haibe-Kains, B.; El Ouriaghli, F.; Chaboteaux, C.; Michiels, S.; Lallemand, F.; Journe, F.; et al. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2012, **18**, 1004–1014, doi:10.1158/1078-0432.CCR-11-0383.
383. Rajaram, M.; Li, J.; Egeblad, M.; Powers, R.S. *PLoS genetics* 2013, **9**, e1003789, doi:10.1371/journal.pgen.1003789.
384. Boelens, M.C.; Wu, T.J.; Nabat, B.Y.; Xu, B.; Qiu, Y.; Yoon, T.; Azzam, D.J.; Twyman-Saint Victor, C.; Wiemann, B.Z.; Ishwaran, H.; et al. *Cell* 2014, **159**, 499–513, doi:10.1016/j.cell.2014.09.051.
385. Merlino, G.; Miodini, P.; Callari, M.; D’Aiuto, F.; Cappelletti, V.; Daidone, M.G. *Molecular oncology* 2017, **11**, 1399–1412, doi:10.1002/1878-0261.12107.
386. Marusyk, A.; Tabassum, D.P.; Janiszewska, M.; Place, A.E.; Trinh, A.; Rozhok, A.I.; Pyne, S.; Guerriero, J.L.; Shu, S.; Ekram, M.; et al. *Cancer research* 2016, **76**, 6495–6506, doi:10.1158/0008-5472.CAN-16-1457.
387. Mathot, P.; Grandin, M.; Devailly, G.; Souaze, F.; Cahais, V.; Moran, S.; Campone, M.; Herceg, Z.; Esteller, M.; Juin, P.; et al. *Oncogenesis* 2017, **6**, e390, doi:10.1038/oncsis.2017.88.

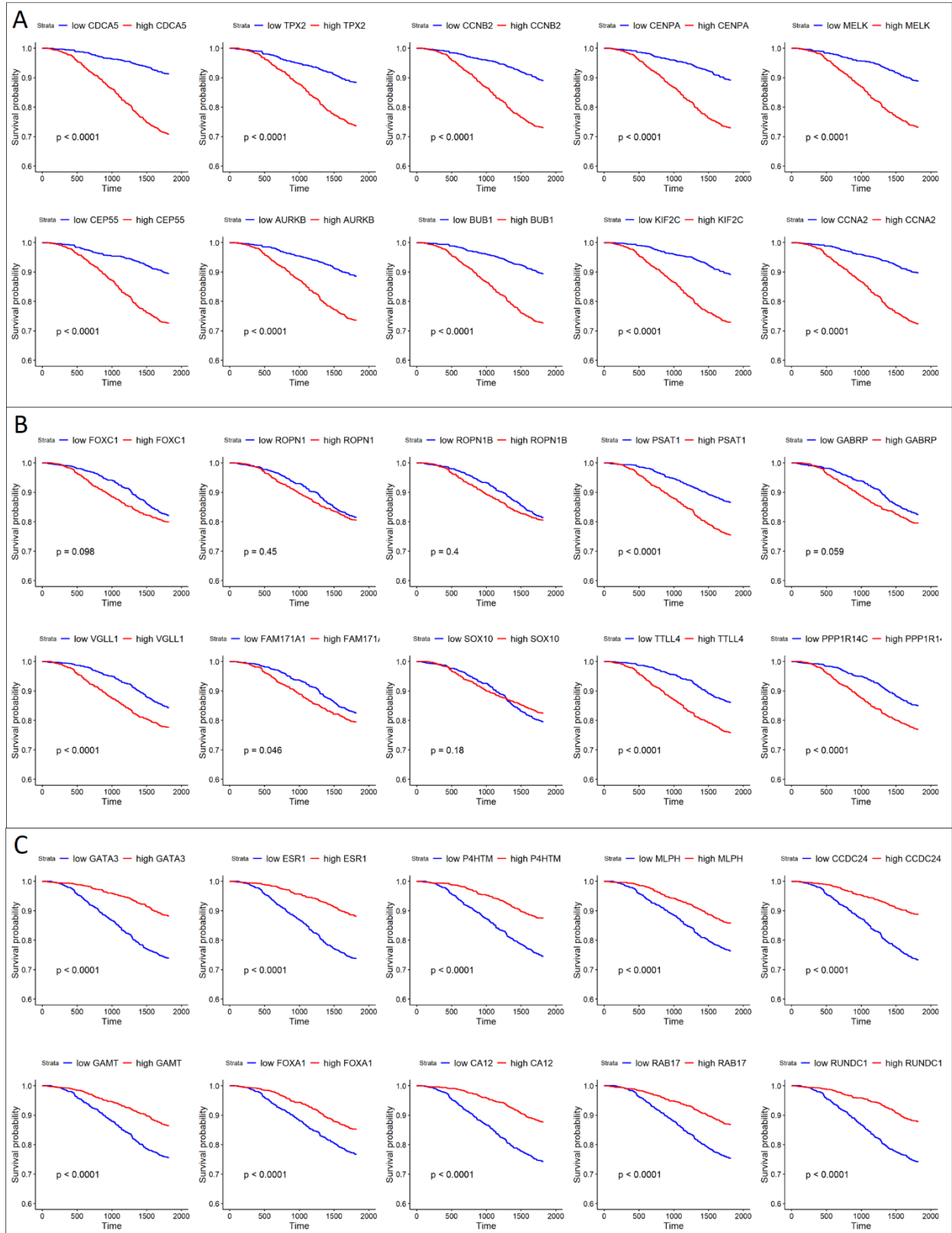
7. Supplementary Figures



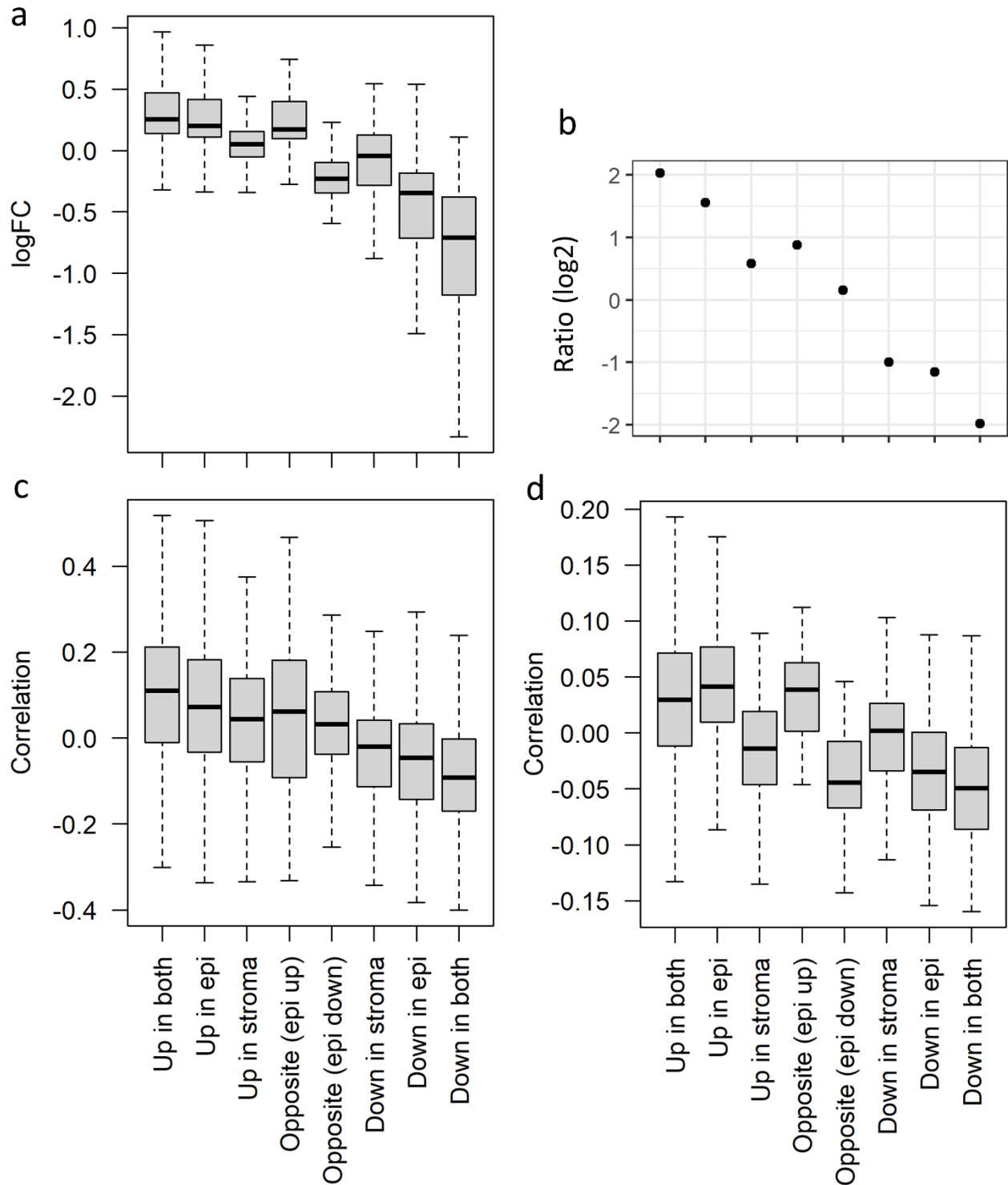
Supplementary Figure 1. Percentage of basal modules' genes in the "unconnected" compartment of other subtypes. Red indicates high percentage, blue low percentage. Different modules are in rows, subtypes in columns.



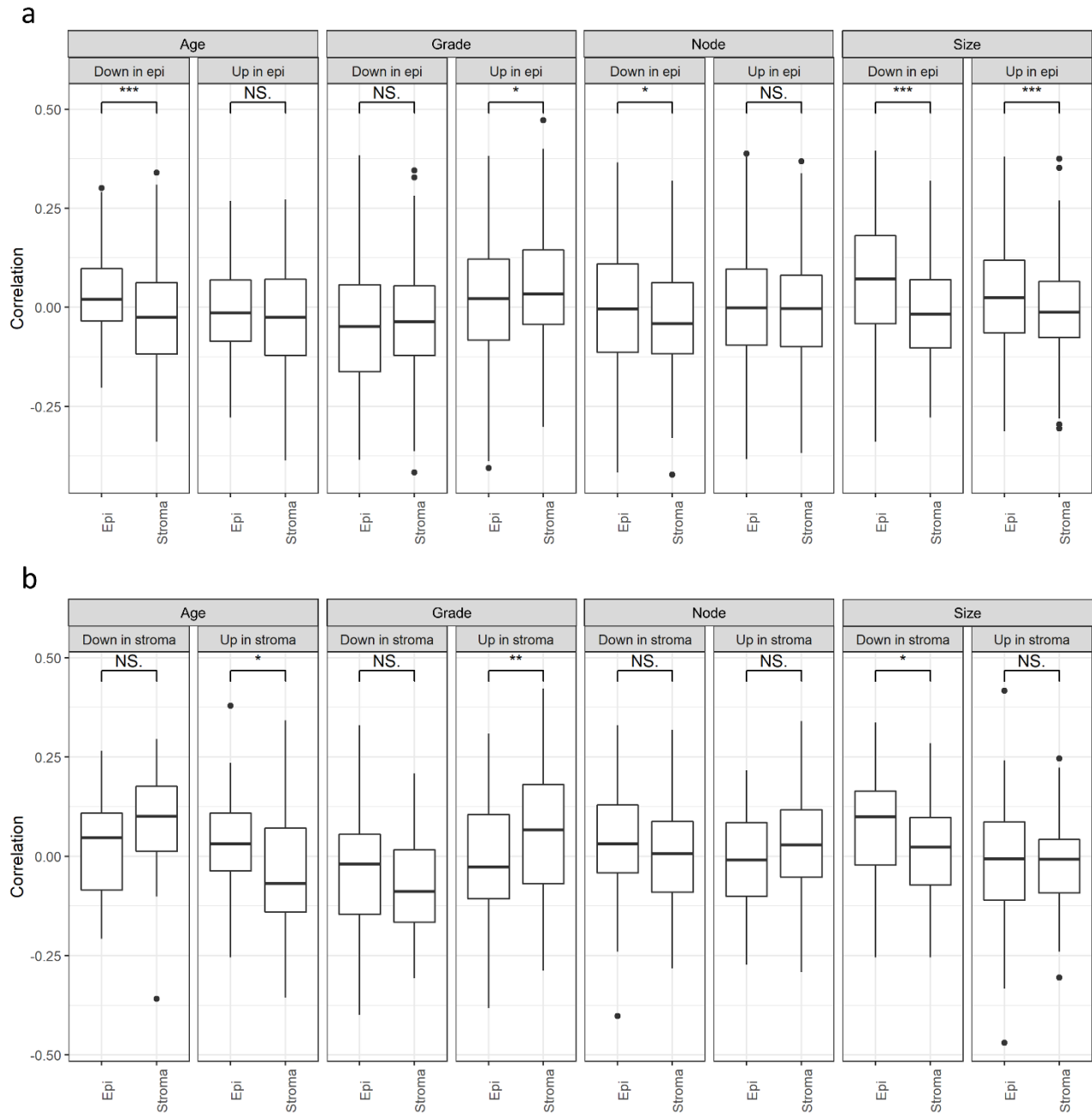
Supplementary Figure 2. Expression of the first 10 hubs of E2F_targets (A), KRAS_dn1 (B) and Estrogen_response (C) modules across subtypes. * p-value for the comparison between basal-like and all other subtypes pooled together $<2.2 \times 10^{-16}$



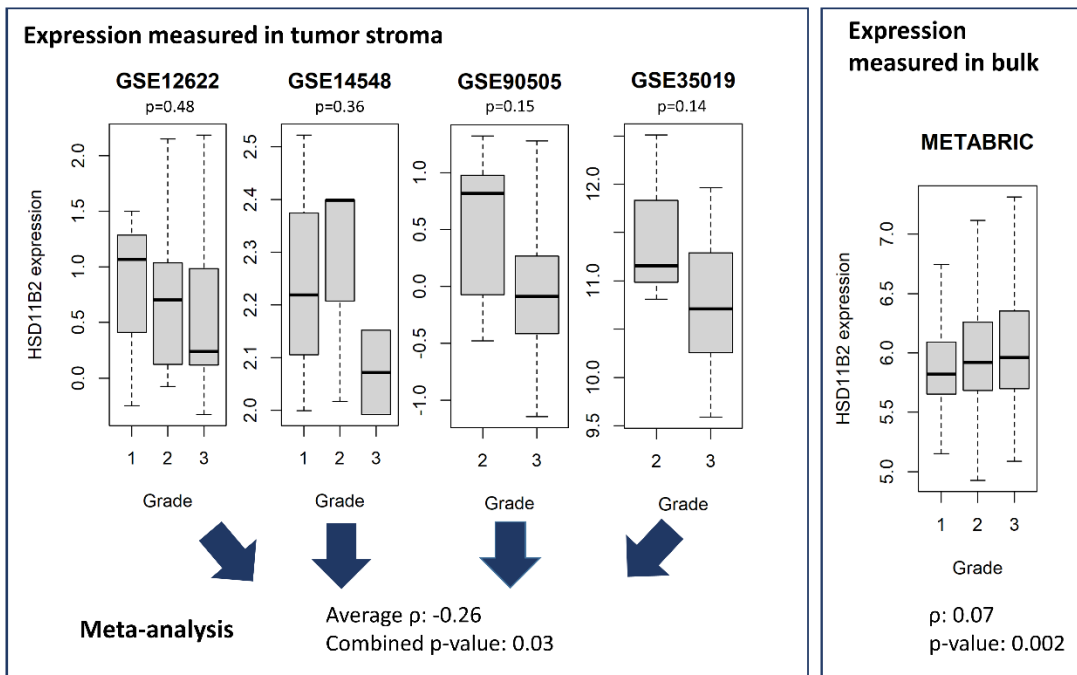
Supplementary Figure 3. Kaplan-Meier survival curves indicating the survival probability for patients with either high or low expression of a specific gene, indicated in the legend. First 10 hubs of E2F_targets (A), KRAS_dn1 (B), and Estrogen_response (C) genes.



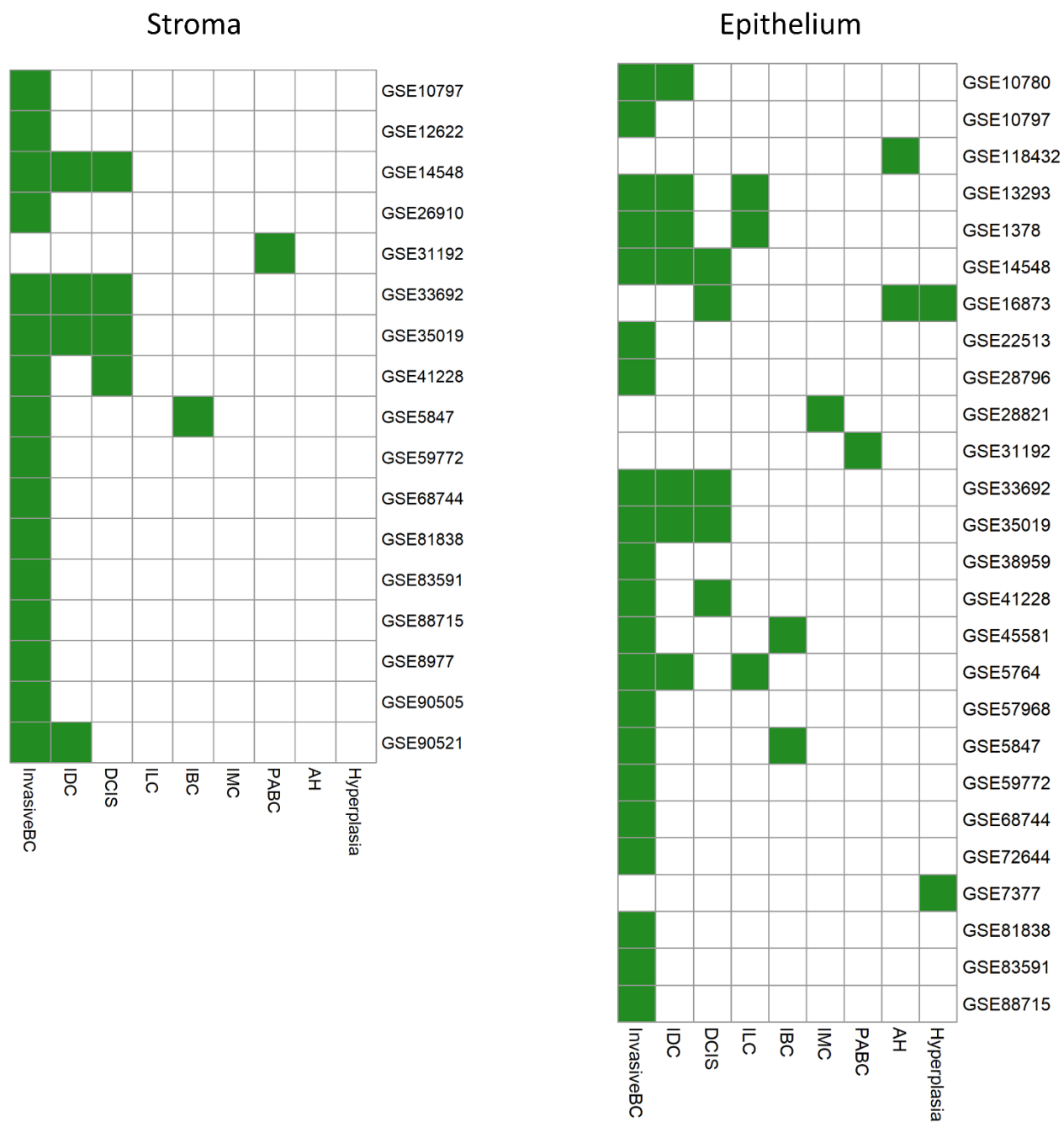
Supplementary Figure 4. The 8 classes of DEGs in bulk samples. Genes regulated in both tumor epithelium and stroma, significantly regulated only in one of the two compartments and with opposite regulation in the two compartments show different degrees of: a) differential expression between tumor and normal samples in the bulk samples of the TCGA cohort; b) ratio between the number of genes significantly correlated with poor prognosis and the number of genes significantly correlated with good prognosis in the METABRIC cohort. High values indicate that many genes are correlated with poor prognosis, while, inversely, negative values indicate that many genes are correlated with good prognosis; Spearman's correlation with tumor grade (c) and size (d).



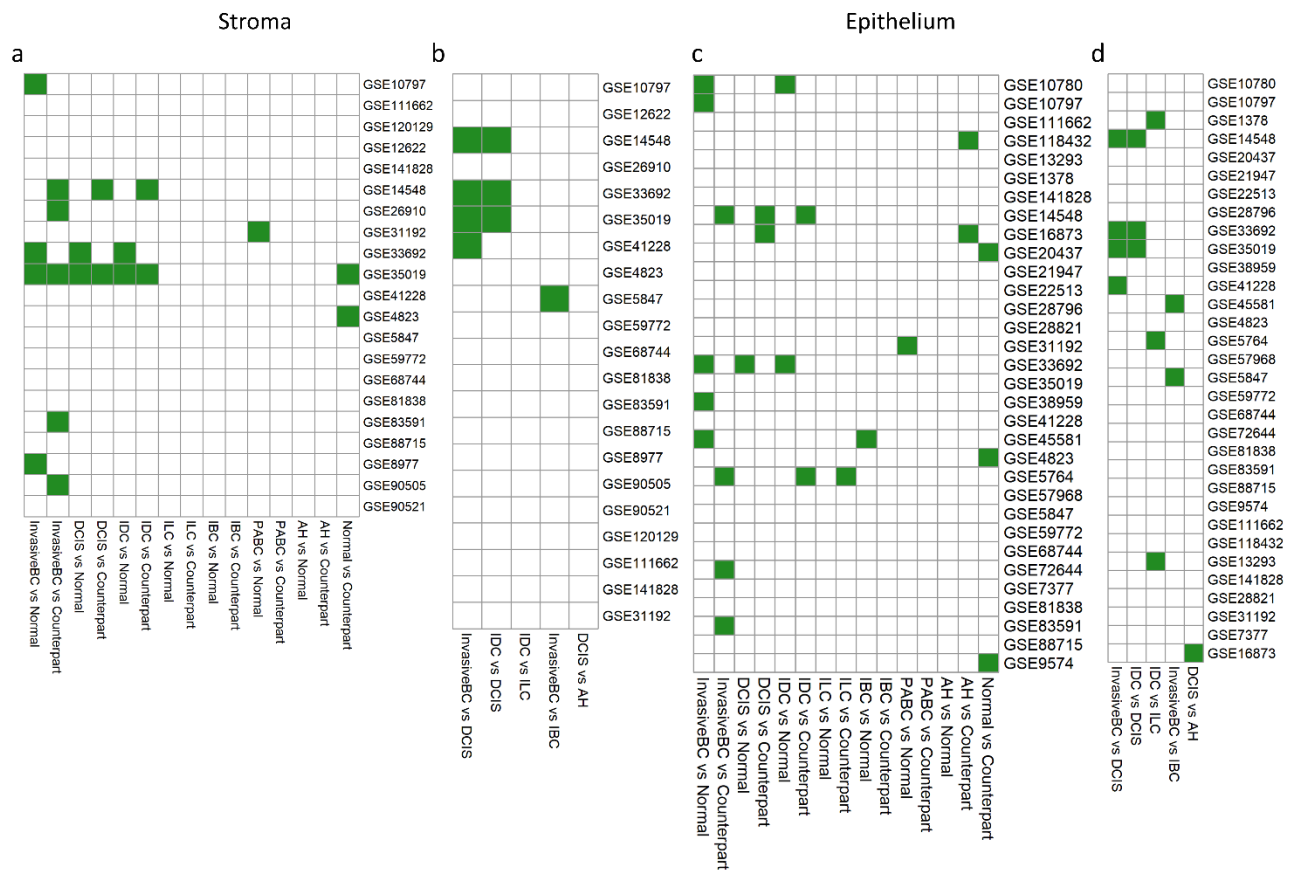
Supplementary Figure 5. Compartment-specific relationship with clinical features for genes with opposite regulation in tumor stroma and epithelium. Spearman's correlation between genes' expression and age at onset, tumor grade, lymphnode status, and tumor size for the classes of DEGs with differential expression in the tumor epithelium (a) or with opposite regulation in tumor stroma and epithelium (b), calculated for each dataset separately and then averaged. The correlation computed when their expression is measured directly in the stroma or in the epithelium, and is compared with the Wilcoxon test (* <0.05, ** <0.01, *** <0.001).



Supplementary Figure 6. HSD11B2 is correlated with lower grade in the stroma, but not in bulk. Left panel, HSD11B2 expression in the 4 datasets with annotation for grade in the tumor stroma. The p-value for the correlation between HSD11B2 expression and tumor grade is indicated above each boxplot. Collapsing p-values with a meta-analysis, the statistical power increases and the p-value reaches significance. Overall, the correlation between HSD11B2 expression and tumor grade in stroma is significantly negative. Right panel, relationship between HSD11B2 expression and tumor grade in bulk, in the METABRIC cohort, is slightly but significantly positive.



Supplementary Figure 7. Summary of histological types available for each dataset, separating for stromal and epithelial samples. Green indicates the presence of the corresponding histological type (column) in the dataset (row). IDC (Invasive ductal carcinoma); DCIS (Ductal carcinoma in situ); ILC (Invasive lobular carcinoma); IBC (Inflammatory breast cancer); IMC (Invasive micropapillary carcinoma); PABC (Pregnancy associated breast cancer); AH (atypical hyperplasia).



Appendix A

Details regarding single annotations and their harmonization compartment.

The compartment column refers to the tissue of origin and can have values: “Epi” (epithelium), “Stroma”, “Vessel” (microvessels), and “Adipose”. The additional values “fine needle” and “core needle” refer to fine/core needle biopsies from the GSE32518 dataset, stroma-poor and stroma-rich, respectively.

Disease Status.

The “diseaseStatus” column indicates whether the sample has a cancerous origin. Where not otherwise specified, malignant samples were annotated as “InvasiveBC” (invasive breast cancer). “Tumor” refers to malignancies and hyperplasias annotated as DCIS (ductal carcinoma in situ), inflammatory breast cancer (IBC), IMC (invasive micropapillary carcinoma), PABC (pregnancy associated breast cancer), AH (atypical hyperplasia), and hyperplasia not otherwise specified. “Normal” indicates histologically normal breast samples from cancer-free subjects, while “counterpart” indicates histologically normal breast samples close to the tumor. Additional annotations are related to specific datasets with particular sample types: in GSE72644, multiple normal samples were obtained from along the duct leading to the tumor at different distances (annotated as “counterpart” in the case of “normal same duct” samples and as “counterpart-adjacent” in the case of “adjacent to tumor” samples) or from contralateral duct (annotated as “contralateral”); in GSE13293, benign breast disease samples with no sign of hyperplasia or atypia were collected and were denoted as “benign”; in GSE20437, histologically normal breast samples from women at high risk of breast cancer were collected and denoted as “prophylactic”; in GSE141828, histologically normal samples from women that developed cancer later were denoted as “susceptible”. These additional levels of disease status were kept separate from other normal samples in the meta-analysis because they were considered to be essentially different.

Receptor status (ER, PR, Her2).

Can either be “positive” or “negative”. Where the percentage of cells positive with IHC staining was indicated, “positive” was assigned to values $\geq 10\%$.

For the Her2 status, values of 1+ were considered “negative”, 3+ as “positive”, and 2+ as ambiguous and were converted to “positive” only in case of FISH positivity and as a missing value otherwise (Nitta et al., 2016).

Samples annotated as triple negative were assigned a “negative” value to all three receptors’ statuses.

TNBC status

For some datasets, samples belonging to the triple negative breast cancer (TNBC) subtype was reported. Hence, in this column “yes” indicates that the sample is triple negative, and “no” indicates that the sample is not triple negative. Additionally, TNBC status was inferred from the receptor status annotations: “no” when either one of the three receptors (ER, PR, Her2) had a “positive” value and “yes” when all three receptors had a “negative” value.

PAM50

PAM50 molecular subtype. Where this information was missing and epithelial invasive BC samples were available, the PAM50 subtype was imputed from gene expression using the *genefu* package [349] with the model “pam50”. PAM50 was not imputed for datasets homogeneous for the receptor status or annotated as TNBC, since the algorithm only accurately infers the subtype from sets of samples comprising a mixture of all subtypes.

Age

Denotes patient age at onset in years.

Grade

Histological tumor grade. Values of I, II, and III were converted to 1, 2, and 3 to facilitate computations. Similarly, “low”, “intermediate”, and “high” values were converted to 1, 2, and 3, respectively.

Size

Tumor size in millimetres. Where multiple sizes were indicated, the highest value was kept.

Node

Can either be “Positive” or “Negative”. Where the number of positive lymph nodes was indicated, “Positive” was assigned to values > 0.

Stage

Overall stage. Obtained from TNM scores, and some missing values were inferred from these scores: M1 corresponds to stage IV; T1, N0, and M0 correspond to stage I; N3 and M0 correspond to stage IIIC; T4, M0, but not N3 correspond to stage IIIB.

Tscore, Nscore, Mscore

TNM staging system (Hortobagyi et al., 2018). Values of T and N are in part redundant with other clinical annotations but cannot be biunivocally matched, and different datasets provided different clinical categories. Hence, they have been retained and used to infer other clinical variables or viceversa. Lymph node positivity is reflected in the N score: samples annotated as node negative were then also annotated as N0, while samples annotated as N0 were also added the “negative” value in the node column. Samples with N1-N3 were annotated as node “positive”. Finally, some missing values were inferred from the overall stage with the conversion rules listed above.

Disease free survival

Two columns are related to disease free survival (DFS): DFSevent indicates presence (“yes”) or absence (“no”) of relapse; DFStime indicates the time (in months) before the relapse or the follow-up time in case of no relapse.

Histology

Contains information about the specific histological type of the tumor. IDC (invasive ductal carcinoma), ILC (invasive lobular carcinoma), DCIS (ductal carcinoma in Situ), inflammatory breast cancer (IBC), IMC (invasive micropapillary carcinoma), PABC (pregnancy associated breast cancer), AH (atypical hyperplasia), and hyperplasia not otherwise specified. Where two histological types were reported for the same sample, the most aggressive was retained (e.g., IDC+DCIS resulted in IDC annotation); the normal counterpart was annotated with the same histological type as the corresponding tumor.

Matching

Unique ID to permit the matching of samples belonging to the same subject.

No assumptions or changes were needed for vital status, parity, menopause, treatment, response to treatment, and ethnicity annotations.

8. Acknowledgments

For this work I thank my supervisor, Prof. Valeria Poli, and Prof. Paolo Provero for supervision and helpful discussions; Alberto Griffa for setting the basis of the analysis of metastases; Niccolò de Marzo for significantly contributing to the LCM meta-analysis; Tommaso Lanciano and Francesco Bonchi for providing the fMRI data; Lidia Avalor, Irene Miglio and Giulia Accetta for the experimental validations.