Article

Meta-Analysis of Microdissected Breast Tumors Reveals Genes Regulated in the Stroma but Hidden in Bulk Analysis

Aurora Savino [1,*], Niccolò De Marzo [1], Paolo Provero [2,3] and Valeria Poli [1,*]

1 Molecular Biotechnology Center, Department of Molecular Biotechnology and Health Sciences, University of Turin, Via Nizza 52, 10126 Turin, Italy; niccolo.demarzo@edu.unito.it
2 Department of Neurosciences "Rita Levi Montalcini", University of Turin, Corso Massimo D'Azeglio 52, 10126 Turin, Italy; paolo.provero@unito.it
3 Center for Omics Sciences, Ospedale San Raffaele IRCCS, Via Olgettina 60, 20132 Milan, Italy
* Correspondence: aurora.savino@unito.it (A.S.); valeria.poli@unito.it (V.P.)

Simple Summary: 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, we aimed at creating a comprehensive resource of transcriptomic datasets investigating breast cancers, focusing on microdissected tumors, which enable the distinguishing of the contribution of the tumor microenvironment from that of cancer cells. We define robust lists of differentially expressed genes and describe 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.

Abstract: Transcriptome data provide a valuable resource for the study of cancer molecular mechanisms, but technical biases, sample heterogeneity, and small sample sizes result in poorly reproducible lists of regulated genes. Additionally, the presence of multiple cellular components contributing to cancer development complicates the interpretation of bulk transcriptomic profiles. To address these issues, we collected 48 microarray datasets derived from laser capture microdissected stroma or epithelium in breast tumors and performed a meta-analysis identifying robust lists of differentially expressed genes. This was used to create a database with carefully harmonized metadata that we make freely available to the research community. As predicted, combining the results of multiple datasets improved statistical power. Moreover, the separate analysis of stroma and epithelium allowed the identification of genes with different contributions in each compartment, which would not be detected by bulk analysis due to their distinct regulation in the two compartments. Our method can be profitably used to help in the discovery of biomarkers and the identification of functionally relevant genes in both the stroma and the epithelium. This database was made to be readily accessible through a user-friendly web interface.

1. Introduction

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 and identifying therapeutic targets [1]. Nevertheless, lists of candidate genes obtained through transcriptome-based studies have proven difficult to reproduce [2–6], raising a note of caution regarding conclusions driven by single sets of experiments. Sample collection and processing methods, protocols, and platforms may impact on the resulting

gene signatures, making them non-overlapping between studies [7]. Additional variability may be introduced by patient heterogeneity, which is not sufficiently represented in small samples.

To resolve these issues, the vast amount of information present in gene expression databases such as Gene Expression Omnibus, ArrayExpress, and EGA [8–10] can be integrated to improve the quality of gene signatures. The advantage is twofold: on one side, a larger sample size allows for the increasing of statistical power; on the other side, merging data obtained through different experimental settings facilitates the removal single-experiments biases, improving robustness [11,12]. Meta-analyses serve this scope, providing a quantitative approach to combine the results of studies investigating the same biological system [5]. Several methods based on different statistical approaches have been proposed [13,14]: (i) aggregating gene lists based on *p*-value [15]; (ii) effect size [16]; and (iii) gene rankings [17].

Meta-analyses are extremely beneficial when applied to cancer biology, an extremely prolific field that often offers tens of independent studies analyzing the same biological or clinical question in different patient cohorts. In their simplest, yet most significant, form, meta-analyses have been applied to assess the reliability of specific genes as diagnostic and prognostic markers [18,19], while whole transcriptomic datasets have been employed for the unbiased evaluation and refinement of prognostic signatures [20–25] to identify patient subgroups [26–30], markers of metastatic tumors [31], and markers of resistance to treatments [32].

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]. Each of these components can support or inhibit tumor growth and are involved in multidirectional cross-talk 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 significantly 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 allowed for the obtaining of spatially-resolved profiles of stroma-rich and stroma-poor regions in tumor tissues [45]; (iii) single-cell techniques have allowed 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 allowing 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.

Here, we performed a meta-analysis of 48 transcriptomic datasets from LCM breast tumor samples, studying the specific epithelium and stroma contribution to the gene expression profiles of bulk tumors. We identified genes robustly changing their expression in each compartment with respect to a normal breast and selected categories of genes with compartment-specific regulation and correlation with clinical features. Finally, we made the whole database and the harmonized metadata available, providing a web-based interface to facilitate its interrogation (https://aurorasavino.shinyapps.io/metalcm/, accessed on 22 June 2021).

2. Materials and Methods

2.1. 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 S1. Original works describing each dataset can be found in [50–94].

2.2. 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, accessed on 26 April 2021). 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 allow variance stabilization, negative values were removed prior to log transformation.

2.3. Database Metadata

After defining the biological and clinical annotations to be gathered, Aurora Savino and Niccolò De Marzo 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 allow for 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.

2.4. 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 (Sections 2.5 and 2.7).

2.5. Differential Expression

Differential expression was calculated using limma [97] for each separate dataset, and information about the subject was added as a factor when the samples were subject-matched.

2.6. Datasets' Comparison

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 [98]. Specifically, in Section 3.2, 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 Section 3.3 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 Section 2.8.

2.7. 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 [99] from the metap package (https://CRAN.R-project.org/package=metap, accessed on 26 April 2021). Two different tests were performed separately for testing coherent 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.

2.8. Definition of DEG Categories

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 Section 2.4 to obtain lists of DEGs in the two tissues. We collapsed the results obtained by comparing tumor vs. normal tissue and tumor vs. normal counterpart. One dataset, GSE35019, allowed both comparisons (tumor stroma vs. normal stroma, tumor stroma vs. stromal normal counterpart), and to satisfy Fisher's method for the assumption of independence, we 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 cutoff 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.

2.9. Assessment of DEG Groups' Robustness

Since all the available datasets were used and no other independent study was thus available, we used a cross validation procedure in order to assess the robustness of the DEGs. Specifically, we divided the datasets selected for the comparisons, indicated in Section 2.8, 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, we 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. We repeated the same procedure 100 times by randomly selecting different combinations of datasets as training and test sets.

2.10. Enrichment for Functional Categories

Gene ontology enrichment was calculated with the enrichGO function from the clusterProfiler package [100] using "Biological Process" GO categories and default parameters.

2.11. Relationship with Clinical Features

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, accessed on 26 April 2021).

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, accessed on 22 April 2021).

The datasets used in the analysis of the correlation between DEG groups with clinical features (Section 3.3) 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), and these values were used in step 5 of the meta-analysis.

2.12. Epithelial, Stromal and Vascular Scores

Scores for epithelial, stromal, and vascular signatures expression were calculated with a single sample GSEA (ssGSEA) via the GSVA package [101] 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 |average logFC| > 1.

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

2.13. Plots and Statistical Analyses

All statistical analyses were performed with R 4.0.4 [102].

Packages used for plotting are R base graphics, ggplot2 [103], ggsignif (https://CRAN.R-project.org/package=ggsignif, accessed on 10 April 2021), ggvenn (https://CRAN.R-project.org/package=ggvenn, accessed on 16 April 2021), survminer (https://CRAN.R-project.org/package=survminer, accessed on 16 April 2021), and pheatmap (https://CRAN.R-project.org/package=pheatmap, accessed on 16 April 2021).

2.14. Web App

The web app (https://aurorasavino.shinyapps.io/metalcm/, accessed on 22 June 2021) was built with the Shiny package (https://CRAN.R-project.org/package=shiny, accessed on 26 April 2021) taking advantage of rintrojs [104], shinybusy (https://CRAN.R-project.org/package=shinybusy, accessed on 10 May 2021), shinythemes (https://CRAN.R-project.org/package=shinythemes, accessed on 10 May 2021), and shinyWidgets (https://CRAN.R-project.org/package=shinyWidgets, accessed on 10 May 2021).

The data that can be easily interrogated with the app 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 (Section 2.4). 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 analysed datasets, and individual-datasets' *p*-values and logFCs.

2.15. Stat3 Signatures

Signatures of Stat3 activity were obtained from Azare et al. [105], Dauer et al. [106], IL6 and Jak/STAT from MSigDB [107], Alvarez et al. [108], Tell and Horvath [109], and Sonnenblick et al. [110].

3. Results
3.1. Database Construction

We collected 48 transcriptomic datasets of breast tumors or breast hyperplasias deposited in the Gene Expression Omnibus (GEO) database, 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, we collected 2144 samples, 2048 of which derive from primary tumors. The complete list of datasets is available in Table S1.

To facilitate the comparability of different experiments, we mapped the probes used for each specific experiment to gene symbols, and we did an extensive and careful harmonization of biological and clinical annotations, as detailed in the Methods (Section 2.3). Specifically, we 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 1 and 2 and Table S2). Moreover, wherever possible and appropriate, we predicted the PAM50 breast cancer (BC) subtype from gene expression, obtaining a good concordance between subtype and receptor expression (Figure S1) and inferred clinical variables from other available clinical annotations (Section 2.3 and Appendix A). Importantly, for 11 of the 48 datasets, we 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 our database significantly improved clinical annotations, which initially were relatively scarce. For example, we increased the number of samples annotated for age 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, while Table S2 comprises the complete summary of available clinical features.

Table 1. 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 2. Summary of clinical features for primary tumor samples (complete list in Table S2).

| 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

3.2. Primary and In Vitro Datasets Are Only Slightly Correlated

First, we 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, we 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, we analyzed each dataset separately and then compared gene expression fold changes to assess similarities and differences between datasets. We were able to perform a total of 18 comparisons, as detailed in the Methods (Section 2.6). The correlations between pairs of comparisons are globally, albeit slightly, positive (Figure 1a, mean correlation of 0.13), indicating intrinsic differences between the datasets. Moreover, different classes of datasets display different degrees of similarity (Figure 1b). Indeed, for example, primary tumors and in vitro samples appeared to be only slightly positively correlated (Figure 1b,c, $\rho = 0.05$). Similarly, IDC and DCIS samples behave differently (Figure 1d), motivating us to keep them separated for meta-analytic purposes in order to limit the biological heterogeneity.

3.3. Non-Redundant Information Is Obtained by Separating Different Tissues

We then took advantage of the tissue specificity of the collected datasets to compare tumor epithelium and stroma gene expression behaviors. We thus calculated differentially expressed genes (DEGs) for each dataset comprising normal and tumor samples (or normal counterpart and tumor), analyzing epithelium and stroma separately. We 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, we chose to apply the widely employed Fisher's method, summing the log-transformed *p*-values obtained from independent studies [99]. To limit the heterogeneity of input data, we only employed invasive BC samples for this analysis. In total, we could perform nine comparisons for epithelium and eight for stroma (datasets used are listed in the Methods, Section 2.8). Strikingly, when comparing the average fold changes for each gene in tumor stroma or epithelium, we observed that most genes behave similarly across compartments (Figure 2a), 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.
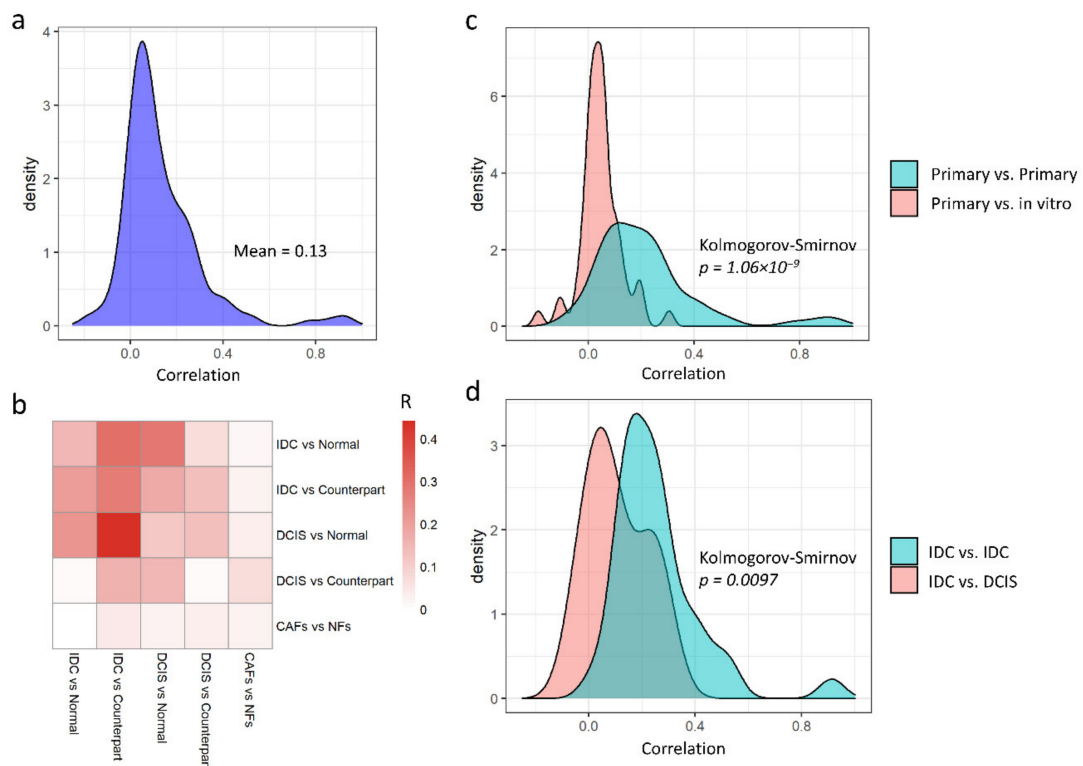
Figure 1. Similarity between stroma datasets of detected genes changing between normal and cancer samples. (**a**) Distribution of Pearson's correlation between log2 fold changes (logFC) of genes between cancerous/non-cancerous conditions in each dataset, showing an average correlation of 0.13. All datasets 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), cuctal 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.
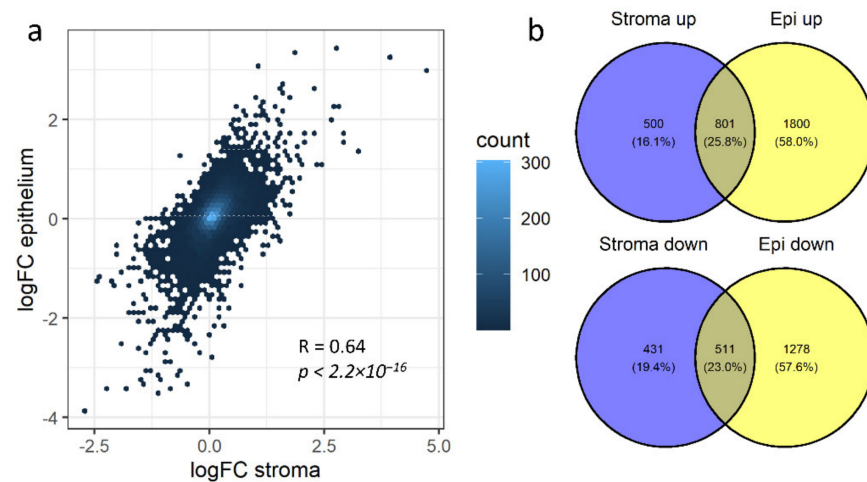

Figure 2. DEGs comparison between stroma and epithelium. (**a**) Cross-datasets average logFC for genes measured in normal/cancerous stroma (X axis) or in normal/cancerous epithelium (Y axis). Pearson's correlation and *p*-value are indicated; the colour indicates the number of overlapping dots. (**b**) Venn diagrams comparing significant DEGs (*p*-adjusted < 0.05) detected in the stroma and in the epithelium.

In the meta-analytic setting, we defined lists of robust DEGs. Specifically, we 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 2b, Table S5). The full list of fold changes and *p*-values obtained for each dataset and condition is available in Table S6, and GO categories enriched for each class are listed in Table S7. Of note, by combining the information present in several datasets, we were able 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 2a). 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 (Table S8). We 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, we 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) (Table S8). We tested the robustness of these classes with a 2-fold cross validation, obtaining good sensitivity and specificity (Figure S2). For each gene, we measured the average expression fold changes between normal breast and breast tumor in bulk samples obtained from the TCGA [111]. The classes with the highest fold changes are those comprising the genes that are coherently differentially expressed in both compartments, while the remaining classes show average fold changes closer to zero (Figure 3a), 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 3b–d). Thus, for two independent bulk tumor datasets, we 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 that we defined show a good robustness (Figure S2) 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 (Figure S3a). It is enriched with genes correlated with poor prognosis (Figure S3b) and correlates with higher tumor grade (Figure S3c) and size (Figure S3d). 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 (Figure S3). 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 coherently up-regulated in tumor stroma across the available datasets (Figure 4a). However, they all appear significantly down-regulated in tumor samples from the TCGA (Figure 4b), confirming that their regulation in the stroma is not detected in bulk.
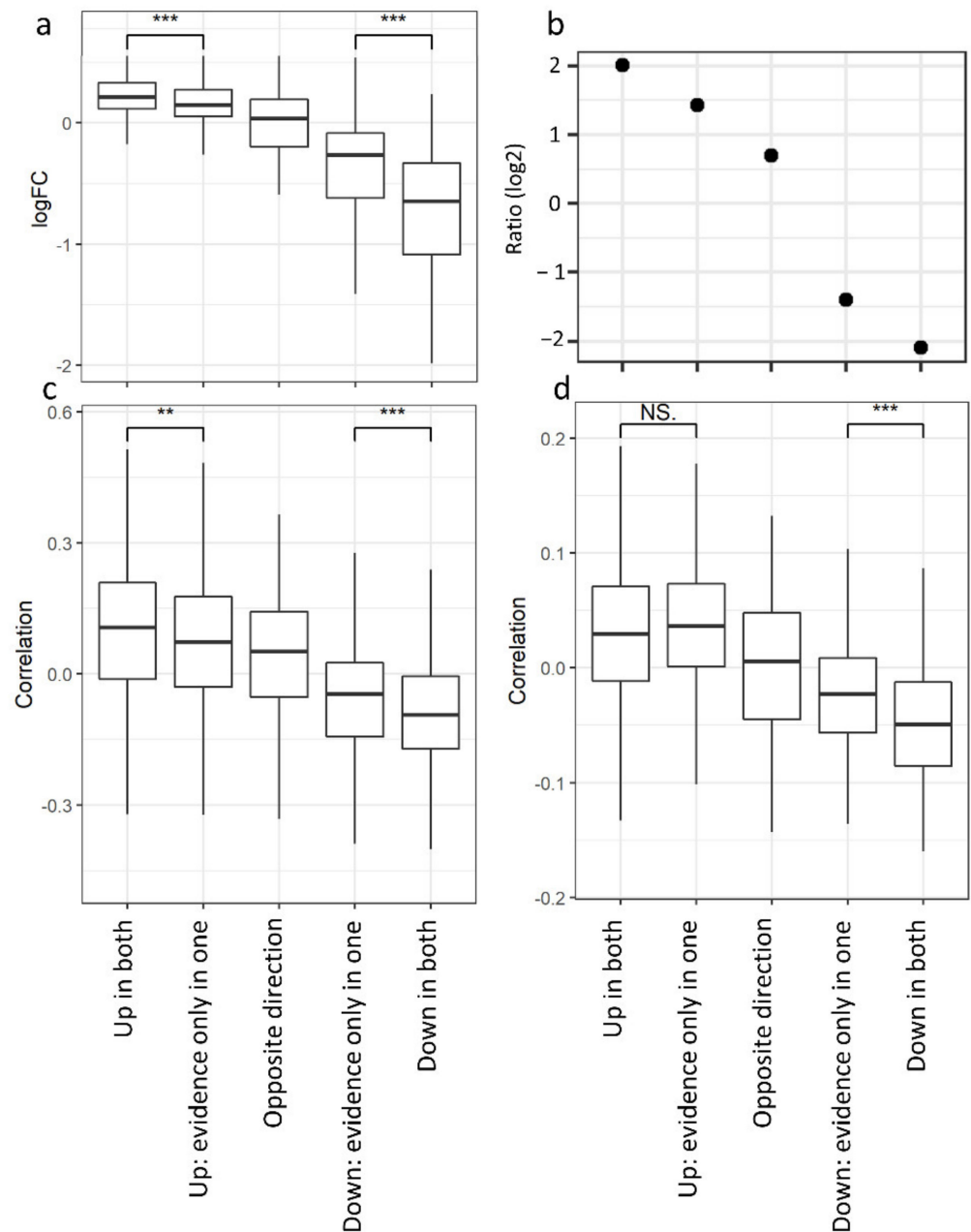
Figure 3. The 5 classes of DEGs in bulk samples. Genes regulated in both tumor epithelium and stroma, significantly regulated only in one of the two compartments and with opposite regulation in the two compartments show different degrees of: (**a**) differential expression between tumor and normal samples 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).
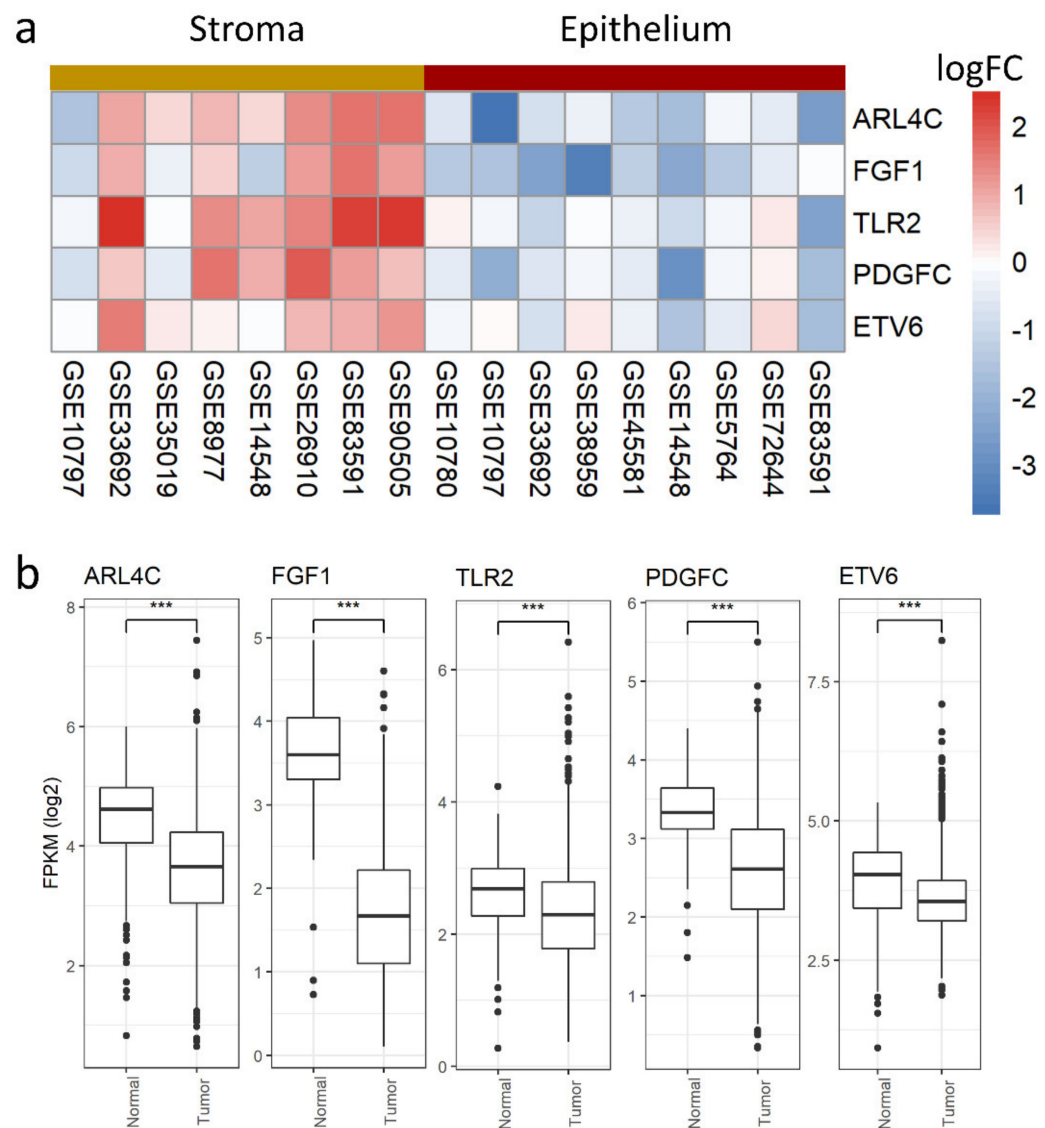
Figure 4. The opposite regulation of genes in tumor stroma or epithelium is hidden in bulk samples. (**a**) Top five significantly differentially expressed genes with higher expression in tumor stroma and lower expression in tumor epithelium when compared to respective compartments in normal breast tissue. 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, we 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 (datasets and analysis are described in the Methods, Section 2.11). 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 5), 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 5). Nevertheless, we did not observe this consistent relationship for genes that were regulated in the epithelium (Figure S4a). 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 (Figure S4b), confirming that these classes of DEGs might also be relevant for tumor progression but differentially regulated in the two compartments.
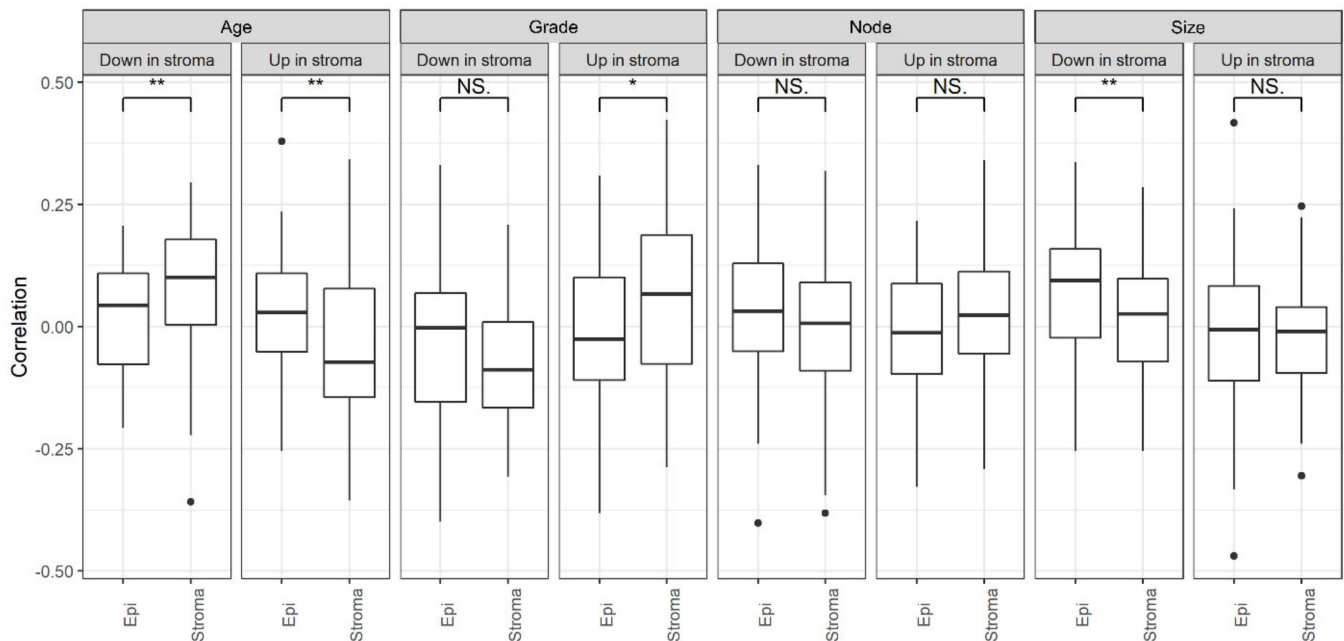

Figure 5. Compartment-specific relationship with clinical features for genes regulated in tumor stroma. Spearman's correlation between gene expression and age at onset, tumor grade, lymph node status, and size for the classes of DEGs regulated only in tumor stroma, calculated for each dataset separately 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).

We further showed that with our compartment-specific gene expression database and the use of meta-analysis, we can identify relationships between genes and clinical features that could not be identified otherwise. As an example, we 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. We identified UPB1 (beta-ureidopropionase 1) as negatively correlated with grade in the tumor stroma but not in bulk (METABRIC dataset, Figure 6). Of note, despite the relatively small number of samples annotated for clinical features and the high within- and between-dataset variability, we were able to improve statistical power by combining multiple datasets (Figure 6a). In one case (the HSD11B2 gene, hydroxysteroid 11-beta dehydrogenase 2), the correlation in the stroma and in bulk shows opposite trends (Figure S5). This could be due to the confounding effect of multiple cell types present in mixed samples. Conversely, we 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 ρ = 0.30) but only slightly when measured in bulk (*p*-value = 0.03, ρ = 0.05).

Considering the enrichment for 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. Full GO lists are available in Table S9.
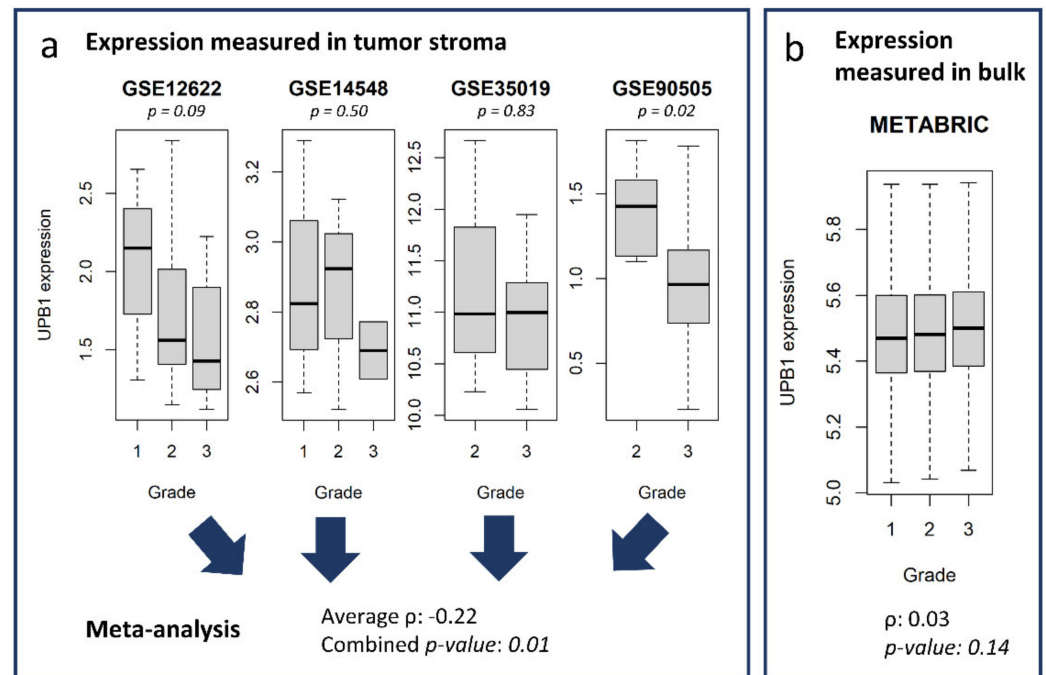
Figure 6. UPB1 is correlated with lower grade in the stroma, but not in bulk. (**a**) UPB1 expression in the 4 datasets with annotation for grade in the tumor stroma. The *p*-value for the correlation between UPB1 expression and tumor grade is indicated above each boxplot. Collapsing *p*-values with a meta-analysis, the statistical power increases, and the *p*-value reaches significance. Overall, the correlation between UPB1 expression and tumor grade in stroma is significantly negative. (**b**) no significant relationship between UPB1 expression and tumor grade in bulk in the METABRIC cohort.

We 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 (Table S5).

Despite the difficulty of accurately detecting the signal deriving from specific cellular compartment in bulk, as discussed above, we also showed in bulk samples that the epithelial and vascular signatures are independent predictors of a patient's disease-free survival (DFS) (Figure 7). 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 7a), while down-regulated genes in the blood vessels define the only significant signature of good prognosis (Figure 7b). We 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 we defined, despite being correlated with tumor grade and age at onset, are not correlated with patient survival.
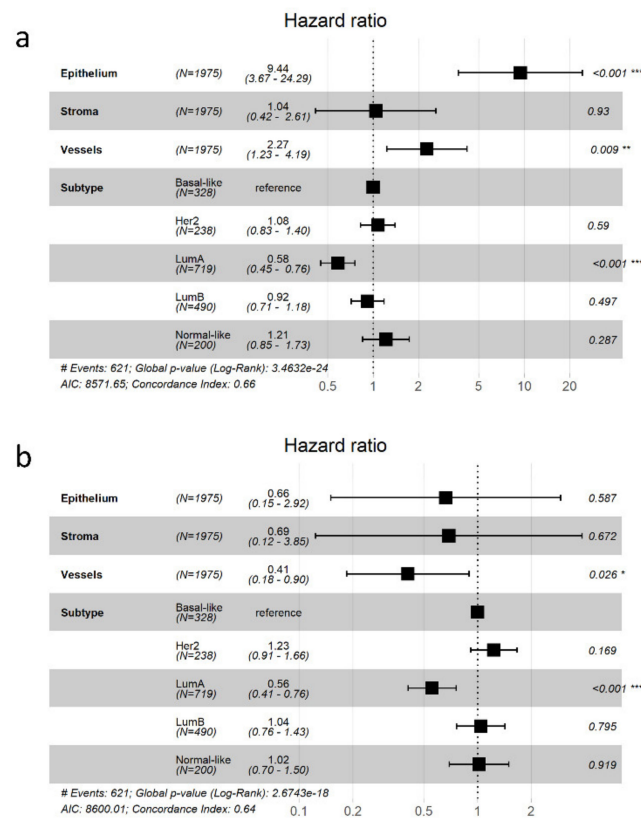
Figure 7. Survival models with epithelial, stromal, and vessel DEGs. Forest plots of multivariate Cox models of disease-free survival with PAM50 subtypes and expression levels of genes in gene signatures (**a**) up- or (**b**) down-regulated in tumor epithelium, stroma, or vessels when compared to corresponding normal tissues (* <0.05, ** <0.01, *** <0.001).

3.4. 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 commonly applied methods is ESTIMATE [112], which based on a single sample GSEA (ssGSEA) of stromal and immune signatures to infer their proportions from the transcriptomes of cell admixtures. We 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). We 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 8), 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*-value = 1.4×10^{-11}).
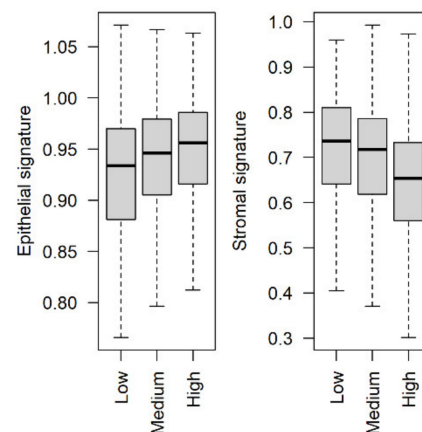
Figure 8. Correlation between epithelial or stromal markers and cellularity in the METABRIC cohort. Epithelial and stromal signatures were obtained through a meta-analysis comparing the expression patterns of tumor stroma and epithelium. The signatures were then used to build stromal and epithelial scores representing cell proportions for each tumor sample in the METABRIC cohort and correlated with clinician annotation of cellularity.

3.5. 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, we indicated the annotations available for each dataset (Figures S6 and S7) and some of the comparisons that can be made using a sample from a specific dataset (Figures S8 and S9). For example, there are 10 different datasets with estrogen receptor status annotations in stroma samples (Figure S6), three datasets that can be used to compare the stroma of IDC and DCIS (Figure S8b), and three datasets allowing for the comparison of Basal and LumA stroma in invasive BC (Figure S9a).

To ease the interrogation of the whole collection of datasets (available in Tables S3 and S4), we created a web app that provides a user-friendly interface, which allows the generating of 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. We set the parameters for comparing tumor stroma and normal stroma of any BC subtype, with a *p*-value threshold of 0.05. We 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" Table We observed that the enrichment 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 2b).

4. Discussion

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 allows for 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.

Here, we 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. We carefully collected and harmonized corresponding biological and clinical annotations to facilitate data integration.

With this tool in hand, we identified genes robustly and coherently changing their expression in breast tumors when compared to normal tissue, either in the stroma or in the epithelium. Analyzing these lists separately, we detected increased expression of cell cycle related genes in the tumor epithelium and of immune-related categories in the tumor stroma. Additionally, we detected the over-expression of the non-canonical Wnt/PCP pathway, which is involved in breast cancer progression [113], synergizing with the STAT3 pathway, contributing to its aggressiveness [114] in both the stroma and the epithelium. Moreover, we 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 [115].

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 [116]. The role of high oxidative phosphorylation in tumors is being increasingly recognized [49,117]. Counter-intuitively, down-regulated genes are enriched for angiogenesis. A similar trend had already been observed [118] 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 [119].

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 [120]. To our knowledge, this is the first time that such a relationship has been detected in the stroma. UPB1 encodes for the last enzyme in the pyrimidine degradation pathway, and its down-regulation might lead to dihydropyrimidine accumulation, linked with EMT [121]. Similarly, NECAP2 is among the top up-regulated genes in tumor stroma only, correlating with a higher grade. NECAP2 is involved in endocytic recycling [122], 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 [123–125].

To the best of our knowledge, this is the first meta-analysis performed specifically on LCM transcriptomic data. We 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. We 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 we provide are indeed robust.

From the analysis of our database, we can conclude that although the behavior of the epithelium and the stroma at the gene expression level is globally similar, separating compartments allows for 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.

We showed additional use-cases of our 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, our 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 [126], and was consistent with data showing an opposite role in the two compartments in colorectal cancer [127]. Specific questions can be addressed by performing the wide variety of comparisons allowed 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 [128]. 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 our work is the availability of our database as a resource for other researchers to explore via a simple web platform that allows differential gene expression and enrichment analyses (https://aurorasavino.shinyapps.io/metalcm/).

5. Conclusions

By collecting and harmonizing multiple datasets of LCM breast tumors, we generated a resource that can be profitably used to discover biomarkers, investigate cancer molecular mechanisms, or test specific research-driven hypotheses in a robust setting. We envision several applications for our database, from the meta-analytic comparison of the several biological conditions and clinical statuses there annotated to the construction of compartment-specific co-expression networks, which will hopefully help in the formulation of robust and specific research hypotheses.

types that can be performed in each dataset, Figure S9: Summary of possible comparisons between PAM50 subtypes that can be performed in each dataset, Table S1: List of all datasets comprised in the database, Table S2: Statistics for the annotations of primary tumors in the database, Table S3: Database metadata, Table S4: Database expression data, Table S5: Differentially expressed genes between normal and invasive BC, Table S6: Full list of genes measured across all compared datasets, Table S7: Gene ontology categories enriched in each DEG list, Table S8: Genes belonging to the 8 classes of DEGs, Table S9: Gene ontology categories enriched in each DEG class.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in this work are available in the Supplementary Materials and easily accessible through the Shiny web app (https://aurorasavino.shinyapps.io/metalcm/, last accessed on 22 June 2021). The main functions used for data analyses can be retrieved from the Shiny app's code.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. 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.

Appendix A.2. 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.

Appendix A.3. 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 [129].

Samples annotated as triple negative were assigned a "negative" value to all three receptors' statuses.

Appendix A.4. 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.

Appendix A.5. 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 [130] 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.

Appendix A.6. Age

Denotes patient age at onset in years.

Appendix A.7. 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.

Appendix A.8. Size

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

Appendix A.9. Node

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

Appendix A.10. 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.

Appendix A.11. Tscore, Nscore, Mscore

TNM staging system [131]. 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.

Appendix A.12. 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.

Appendix A.13. 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.

Appendix A.14. Matching

Unique ID to allow 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.

References

1. Rung, J.; Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **2013**, *14*, 89–99. [CrossRef]
2. Kela, I.; Ein-Dor, L.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: Is there a unique set? *Breast Cancer Res.* **2005**, *7*, P4.38. [CrossRef]
3. Michiels, S.; Koscielny, S.; Hill, C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet* **2005**, *365*, 488–492. [CrossRef]
4. Shen, R.; Chinnaiyan, A.M.; Ghosh, D. Pathway analysis reveals functional convergence of gene expression profiles in breast cancer. *BMC Med. Genom.* **2008**, *1*, 1–7. [CrossRef] [PubMed]
5. Ioannidis, J.P.A.; Ntzani, E.E.; Trikalinos, T.A.; Contopoulos-Ioannidis, D.G. Replication validity of genetic association studies. *Nat. Genet.* **2001**, *29*, 306–309. [CrossRef]
6. Cui, W.; Xue, H.; Wei, L.; Jin, J.; Tian, X.; Wang, Q. High heterogeneity undermines generalization of differential expression results in RNA-Seq analysis. *Hum. Genom.* **2021**, *15*, 7. [CrossRef]
7. 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. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* **2008**, *14*, 822–827. [CrossRef]
8. 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. NCBI GEO: Archive for functional genomics data sets—Update. *Nucleic Acids Res.* **2013**, *41*, D991–D995. [CrossRef]
9. 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. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* **2015**, *47*, 692–695. [CrossRef]
10. Athar, A.; Füllgrabe, A.; George, N.; Iqbal, H.; Huerta, L.; Ali, A.; Snow, C.; Fonseca, N.A.; Petryszak, R.; Papatheodorou, I.; et al. ArrayExpress update—From bulk to single-cell expression data. *Nucleic Acids Res.* **2019**, *47*, D711–D715. [CrossRef]
11. Ramasamy, A.; Mondry, A.; Holmes, C.C.; Altman, D.G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **2008**, *5*, 1320–1332. [CrossRef]
12. Taminau, J.; Lazar, C.; Meganck, S.; Nowé, A. Comparison of Merging and Meta-Analysis as Alternative Approaches for Integrative Gene Expression Analysis. *ISRN Bioinform.* **2014**, *2014*, 1–7. [CrossRef]
13. Chang, L.C.; Lin, H.M.; Sibille, E.; Tseng, G.C. Meta-analysis methods for combining multiple expression profiles: Comparisons, statistical characterization and an application guideline. *BMC Bioinform.* **2013**, *14*. [CrossRef] [PubMed]
14. Walsh, C.; Hu, P.; Batt, J.; Santos, C. Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays* **2015**, *4*, 389–406. [CrossRef]
15. Rhodes, D.R.; Yu, J.; Shanker, K.; Deshpande, N.; Varambally, R.; Ghosh, D.; Barrette, T.; Pandey, A.; Chinnaiyan, A.M. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 9309–9314. [CrossRef] [PubMed]

16. Choi, J.K.; Yu, U.; Kim, S.; Yoo, O.J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **2003**, *19* (Suppl. S1), i84–i90. [CrossRef] [PubMed]

17. Hong, F.; Breitling, R.; McEntee, C.W.; Wittner, B.S.; Nemhauser, J.L.; Chory, J. RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **2006**, *22*, 2825–2827. [CrossRef]

18. Cai, K.-T.; Feng, C.-X.; Zhao, J.-C.; He, R.-Q.; Ma, J.; Zhong, J.-C. Upregulated miR-203a-3p and its potential molecular mechanism in breast cancer: A study based on bioinformatics analyses and a comprehensive meta-analysis. *Mol. Med. Rep.* **2018**, *18*, 4994–5008. [CrossRef] [PubMed]

19. Bozovic-Spasojevic, I.; Zardavas, D.; Brohée, S.; Ameye, L.; Fumagalli, D.; Ades, F.; de Azambuja, E.; Bareche, Y.; Piccart, M.; Paesmans, M.; et al. The Prognostic Role of Androgen Receptor in Patients with Early-Stage Breast Cancer: A Meta-analysis of Clinical and Gene Expression Data. *Clin. Cancer Res.* **2017**, *23*, 2702–2712. [CrossRef]

20. Tang, H.; Wang, S.; Xiao, G.; Schiller, J.; Papadimitrakopoulou, V.; Minna, J.; Wistuba, I.I.; Xie, Y. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. *Ann. Oncol.* **2017**, *28*, 733–740. [CrossRef]

21. Wang, J.; Gao, P.; Song, Y.; Sun, J.; Chen, X.; Yu, H.; Wang, Y.; Wang, Z. Prognostic value of gastric cancer-associated gene signatures: Evidence based on a meta-analysis using integrated bioinformatics methods. *J. Cell. Mol. Med.* **2018**, *22*, 5743–5747. [CrossRef]

22. 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. Meta-Analysis of 1,200 Transcriptomic Profiles Identifies a Prognostic Model for Pancreatic Ductal Adenocarcinoma. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–16. [CrossRef] [PubMed]

23. Yasrebi, H. Comparative study of joint analysis of microarray gene expression data in survival prediction and risk assessment of breast cancer patients. *Brief. Bioinform.* **2016**, *17*, 771–785. [CrossRef] [PubMed]

24. Xu, L.; Tan, A.C.; Winslow, R.L.; Geman, D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinform.* **2008**, *9*. [CrossRef]

25. 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. An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. *Cancers* **2019**, *11*, 155. [CrossRef] [PubMed]

26. Ma, S.; Ogino, S.; Parsana, P.; Nishihara, R.; Qian, Z.; Shen, J.; Mima, K.; Masugi, Y.; Cao, Y.; Nowak, J.A.; et al. Continuity of transcriptomes among colorectal cancer subtypes based on meta-analysis. *Genome Biol.* **2018**, *19*, 1–14. [CrossRef] [PubMed]

27. De Santiago, I.; Yau, C.; Heij, L.; Middleton, M.R.; Markowetz, F.; Grabsch, H.I.; Dustin, M.L.; Sivakumar, S. Immunophenotypes of pancreatic ductal adenocarcinoma: Meta-analysis of transcriptional subtypes. *Int. J. Cancer* **2019**, *145*, 1125–1137. [CrossRef] [PubMed]

28. Qadir, F.; Lalli, A.; Dar, H.H.; Hwang, S.; Aldehlawi, H.; Ma, H.; Dai, H.; Waseem, A.; Teh, M.-T. Clinical correlation of opposing molecular signatures in head and neck squamous cell carcinoma. *BMC Cancer* **2019**, *19*, 830. [CrossRef] [PubMed]

29. Tan, T.Z.; Rouanne, M.; Tan, K.T.; Huang, R.Y.-J.; Thiery, J.-P. Molecular Subtypes of Urothelial Bladder Cancer: Results from a Meta-cohort Analysis of 2411 Tumors. *Eur. Urol.* **2019**, *75*, 423–432. [CrossRef]

30. Oztemur Islakoglu, Y.; Noyan, S.; Aydos, A.; Gur Dedeoglu, B. Meta-microRNA biomarker signatures to classify breast cancer subtypes. *Omics J. Integr. Biol.* **2018**, *22*, 709–716. [CrossRef]

31. Bell, R.; Barraclough, R.; Vasieva, O. Gene Expression Meta-Analysis of Potential Metastatic Breast Cancer Markers. *Curr. Mol. Med.* **2017**, *17*, 200–210. [CrossRef] [PubMed]

32. 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. Identification of novel therapeutic target genes in acquired lapatinib-resistant breast cancer by integrative meta-analysis. *Tumour Biol.* **2016**, *37*, 2285–2297. [CrossRef] [PubMed]

33. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef]

34. Junttila, M.R.; de Sauvage, F.J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **2013**, *501*, 346–354. [CrossRef]

35. Plava, J.; Cihova, M.; Burikova, M.; Matuskova, M.; Kucerova, L.; Miklikova, S. Recent advances in understanding tumor stroma-mediated chemoresistance in breast cancer. *Mol. Cancer* **2019**, *18*, 1–10. [CrossRef] [PubMed]

36. Joyce, J.A.; Pollard, J.W. Microenvironmental regulation of metastasis. *Nat. Rev. Cancer* **2009**, *9*, 239–252. [CrossRef]

37. Sahai, E.; Astsaturov, I.; Cukierman, E.; DeNardo, D.G.; Egeblad, M.; Evans, R.M.; Fearon, D.; Greten, F.R.; Hingorani, S.R.; Hunter, T.; et al. A framework for advancing our understanding of cancer-associated fibroblasts. *Nat. Rev. Cancer* **2020**, *20*, 174–186. [CrossRef]

38. Lei, X.; Lei, Y.; Li, J.K.; Du, W.X.; Li, R.G.; Yang, J.; Li, J.; Li, F.; Tan, H.B. Immune cells within the tumor microenvironment: Biological functions and roles in cancer immunotherapy. *Cancer Lett.* **2020**, *470*, 126–133. [CrossRef] [PubMed]

39. Aran, D.; Sirota, M.; Butte, A.J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **2015**, 1–12. [CrossRef]

40. Schwede, M.; Waldron, L.; Mok, S.C.; Wei, W.; Basunia, A.; Merritt, M.A.; Mitsiades, C.S.; Parmigiani, G.; Harrington, D.P.; Quackenbush, J.; et al. The impact of stroma admixture on molecular subtypes and prognostic gene signatures in serous ovarian cancer. *Cancer Epidemiol. Biomark. Prev.* **2020**, *29*, 509–519. [CrossRef]

41. Matikas, A.; Zerdes, I.; Lövrot, J.; Richard, F.; Sotiriou, C.; Bergh, J.; Valachis, A.; Foukakis, T. Prognostic implications of PD-L1 expression in breast cancer: Systematic review and meta-analysis of immunohistochemistry and pooled analysis of transcriptomic data. *Clin. Cancer Res.* **2019**, *25*, 5717–5726. [CrossRef] [PubMed]

42. Donnem, T.; Al-Saad, S.; Al-Shibli, K.; Delghandi, M.P.; Persson, M.; Nilsen, M.N.; Busund, L.T.; Bremnes, R.M. Inverse prognostic impact of angiogenic marker expression in tumor cells versus stromal cells in non-small cell lung cancer. *Clin. Cancer Res.* **2007**, *13*, 6649–6657. [CrossRef]

43. Sjöberg, E.; Augsten, M.; Bergh, J.; Jirström, K.; Östman, A. Expression of the chemokine CXCL14 in the tumour stroma is an independent marker of survival in breast cancer. *Br. J. Cancer* **2016**, *114*, 1117–1124. [CrossRef] [PubMed]

44. Von Eggeling, F.; Hoffmann, F. Microdissection—An Essential Prerequisite for Spatial Cancer Omics. *Proteomics* **2020**, *20*. [CrossRef] [PubMed]

45. Ståhl, P.L.; Salmén, F.; Vickovic, S.; Lundmark, A.; Navarro, J.F.; Magnusson, J.; Giacomello, S.; Asp, M.; Westholm, J.O.; Huss, M.; et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **2016**, *353*, 78–82. [CrossRef]

46. Chung, W.; Eum, H.H.; Lee, H.-O.; Lee, K.-M.; Lee, H.-B.; Kim, K.-T.; Ryu, H.S.; Kim, S.; Lee, J.E.; Park, Y.H.; et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **2017**, *8*, 15081. [CrossRef]

47. Bartoschek, M.; Oskolkov, N.; Bocci, M.; Lövrot, J.; Larsson, C.; Sommarin, M.; Madsen, C.D.; Lindgren, D.; Pekar, G.; Karlsson, G.; et al. Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. *Nat. Commun.* **2018**, *9*. [CrossRef]

48. Steen, C.B.; Liu, C.L.; Alizadeh, A.A.; Newman, A.M. Profiling Cell Type Abundance and Expression in Bulk Tissues with CIBERSORTx. *Methods Mol. Biol.* **2020**, *2117*, 135–157. [CrossRef]

49. 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. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat. Cell Biol.* **2020**, *22*, 310–320. [CrossRef]

50. Chen, D.T.; Nasir, A.; Culhane, A.; Venkataramu, C.; Fulp, W.; Rubio, R.; Wang, T.; Agrawal, D.; McCarthy, S.M.; Gruidl, M.; et al. Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res. Treat.* **2010**, *119*, 335–346. [CrossRef]

51. Casey, T.; Bond, J.; Tighe, S.; Hunter, T.; Lintault, L.; Patel, O.; Eneman, J.; Crocker, A.; White, J.; Tessitore, J.; et al. Molecular signatures suggest a major role for stromal cells in development of invasive breast cancer. *Breast Cancer Res. Treat.* **2009**, *114*, 47–62. [CrossRef]

52. 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. Genomic signature of parity in the breast of premenopausal women. *Breast Cancer Res.* **2019**, *21*, 1–19. [CrossRef] [PubMed]

53. 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. Gene expression signature of atypical breast hyperplasia and regulation by SFRP1. *Breast Cancer Res.* **2019**, *21*, 1–18. [CrossRef]

54. Russo, J.; Balogh, G.A.; Russo, I.H. Full-term pregnancy induces a specific genomic signature in the human breast. *Cancer Epidemiol. Biomark. Prev.* **2008**, *17*, 51–66. [CrossRef]

55. 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. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **2004**, *5*, 607–616. [CrossRef] [PubMed]

56. Marino, N.; German, R.; Rao, X.; Simpson, E.; Liu, S.; Wan, J.; Liu, Y.; Sandusky, G.; Jacobsen, M.; Stoval, M.; et al. Upregulation of lipid metabolism genes in the breast prior to cancer diagnosis. *npj Breast Cancer* **2020**, *6*, 1–13. [CrossRef]

57. Ma, X.-J.; Dahiya, S.; Richardson, E.; Erlander, M.; Sgroi, D.C. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Res.* **2009**, *11*, R7. [CrossRef] [PubMed]

58. Emery, L.A.; Tripathi, A.; King, C.; Kavanah, M.; Mendez, J.; Stone, M.D.; de Las Morenas, A.; Sebastiani, P.; Rosenberg, C.L. Early dysregulation of cell adhesion and extracellular matrix pathways in breast cancer progression. *Am. J. Pathol.* **2009**, *175*, 1292–1302. [CrossRef]

59. Graham, K.; de Las Morenas, A.; Tripathi, A.; King, C.; Kavanah, M.; Mendez, J.; Stone, M.; Slama, J.; Miller, M.; Antoine, G.; et al. Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *Br. J. Cancer* **2010**, *102*, 1284–1293. [CrossRef]

60. Graham, K.; Ge, X.; de Las Morenas, A.; Tripathi, A.; Rosenberg, C.L. Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium. *Clin. Cancer Res.* **2011**, *17*, 236–246. [CrossRef]

61. 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. Identification of markers of taxane sensitivity using proteomic and genomic analyses of breast tumors from patients receiving neoadjuvant paclitaxel and radiation. *Clin. Cancer Res.* **2010**, *16*, 681–690. [CrossRef] [PubMed]

62. Planche, A.; Bacac, M.; Provero, P.; Fusco, C.; Delorenzi, M.; Stehle, J.C.; Stamenkovic, I. Identification of prognostic molecular features in the reactive stroma of human breast and prostate cancer. *PLoS ONE* **2011**, *6*. [CrossRef]

63. Lehmann, B.D.B.; Bauer, J.A.; Chen, X.; Sanders, M.E.; Chakravarthy, A.B.; Shyr, Y.; Pietenpol, J.A. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Investig.* **2011**, *121*, 2750–2767. [CrossRef]

64. Hosein, A.N.; Livingstone, J.; Buchanan, M.; Reid, J.F.; Hallett, M.; Basik, M. A functional in vitro model of heterotypic interactions reveals a role for interferon-positive carcinoma associated fibroblasts in breast cancer. *BMC Cancer* **2015**, *15*, 1–11. [CrossRef] [PubMed]

65. Harvell, D.M.E.; Kim, J.; O'Brien, J.; Tan, A.C.; Borges, V.F.; Schedin, P.; Jacobsen, B.M.; Horwitz, K.B. Genomic Signatures of Pregnancy-Associated Breast Cancer Epithelia and Stroma and their Regulation by Estrogens and Progesterone. *Horm. Cancer* **2013**, *4*, 140–153. [CrossRef]

66. Knudsen, E.S.; Ertel, A.; Davicioni, E.; Kline, J.; Schwartz, G.F.; Witkiewicz, A.K. Progression of ductal carcinoma in situ to invasive breast cancer is associated with gene expression programs of EMT and myoepithelia. *Breast Cancer Res. Treat.* **2012**, *133*, 1009–1024. [CrossRef]

67. 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. Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. *Breast Cancer Res. Treat.* **2012**, *135*, 153–165. [CrossRef] [PubMed]

68. Tchou, J.; Kossenkov, A.V.; Chang, L.; Satija, C.; Herlyn, M.; Showe, L.C.; Puré, E. Human breast cancer associated fibroblasts exhibit subtype specific gene expression profiles. *BMC Med. Genom.* **2012**, *5*. [CrossRef]

69. Komatsu, M.; Yoshimaru, T.; Matsuo, T.; Kiyotani, K.; Miyoshi, Y.; Tanahashi, T.; Rokutan, K.; Yamaguchi, R.; Saito, A.; Imoto, S.; et al. Molecular features of triple negative breast cancer cells by genome-wide gene expression profiling analysis. *Int. J. Oncol.* **2013**, *42*, 478–506. [CrossRef]

70. Lee, S.; Stewart, S.; Nagtegaal, I.; Luo, J.; Wu, Y.; Colditz, G.; Medina, D.; Allred, D.C. Differentially expressed genes regulating the progression of ductal carcinoma in situ to invasive breast cancer. *Cancer Res.* **2012**, *72*, 4574–4586. [CrossRef] [PubMed]

71. 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. Genomic and expression analysis of microdissected inflammatory breast cancer. *Breast Cancer Res. Treat.* **2013**, *138*, 761–772. [CrossRef]

72. Finak, G.; Sadekova, S.; Pepin, F.; Hallett, M.; Meterissian, S.; Halwani, F.; Khetani, K.; Souleimanova, M.; Zabolotny, B.; Omeroglu, A.; et al. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res.* **2006**, *8*, 1–18. [CrossRef]

73. Turashvili, G.; Bouchal, J.; Baumforth, K.; Wei, W.; Dziechciarkova, M.; Ehrmann, J.; Klein, J.; Fridman, E.; Skarda, J.; Srovnal, J.; et al. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* **2007**, *7*, 1–20. [CrossRef]

74. Reyngold, M.; Turcan, S.; Giri, D.; Kannan, K.; Walsh, L.A.; Viale, A.; Drobnjak, M.; Vahdat, L.T.; Lee, W.; Chan, T.A. Remodeling of the methylation landscape in breast cancer metastasis. *PLoS ONE* **2014**, *9*, 1–10. [CrossRef]

75. 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.; Ambs, S. A stromal gene signature associated with inflammatory breast cancer. *Int. J. Cancer* **2008**, *122*, 1324–1332. [CrossRef] [PubMed]

76. Winslow, S.; Leandersson, K.; Edsjö, A.; Larsson, C. Prognostic stromal gene signatures in breast cancer. *Breast Cancer Res.* **2015**, *17*, 1–13. [CrossRef]

77. 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. Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. *Genome Biol.* **2015**, *16*. [CrossRef]

78. 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. Mapping genomic and transcriptomic alterations spatially in epithelial cells adjacent to human breast carcinoma. *Nat. Commun.* **2017**, *8*. [CrossRef] [PubMed]

79. Lee, S.; Medina, D.; Tsimelzon, A.; Mohsin, S.K.; Mao, S.; Wu, Y.; Allred, D.C. Alterations of gene expression in the development of early hyperplastic precursors of breast cancer. *Am. J. Pathol.* **2007**, *171*, 252–262. [CrossRef]

80. Lehmann, B.D.; Jovanović, B.; Chen, X.; Estrada, M.V.; Johnson, K.N.; Shyr, Y.; Moses, H.L.; Sanders, M.E.; Pietenpol, J.A. Refinement of triple-negative breast cancer molecular subtypes: Implications for neoadjuvant chemotherapy selection. *PLoS ONE* **2016**, *11*, e0157368. [CrossRef] [PubMed]

81. Liu, H.; Dowdle, J.A.; Khurshid, S.; Sullivan, N.J.; Bertos, N.; Rambani, K.; Mair, M.; Daniel, P.; Wheeler, E.; Tang, X.; et al. Discovery of Stromal Regulatory Networks that Suppress Ras-Sensitized Epithelial Cell Proliferation. *Dev. Cell* **2017**, *41*, 392–407.e6. [CrossRef]

82. 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. Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. *J. Clin. Investig.* **2019**, *129*, 1785–1800. [CrossRef] [PubMed]

83. 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. Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature* **2007**, *449*, 557–563. [CrossRef]

84. Saleh, S.M.I.; Bertos, N.; Gruosso, T.; Gigoux, M.; Souleimanova, M.; Zhao, H.; Omeroglu, A.; Hallett, M.T.; Park, M. Identification of interacting stromal axes in triple-negative breast cancer. *Cancer Res.* **2017**, *77*, 4673–4683. [CrossRef] [PubMed]

85. Brouwers, B.; Fumagalli, D.; Brohee, S.; Hatse, S.; Govaere, O.; Floris, G.; van den Eynde, K.; Bareche, Y.; Schöffski, P.; Smeets, A.; et al. The footprint of the ageing stroma in older patients with breast cancer. *Breast Cancer Res.* **2017**, *19*, 1–14. [CrossRef] [PubMed]

86. 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. Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int. J. Cancer* **2008**, *122*, 1557–1566. [CrossRef]

87. 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. Gene Expression Analysis in Human Breast Cancer Associated Blood Vessels. *PLoS ONE* **2012**, *7*. [CrossRef]

88. François, P.; Bertos, N.; Laferrière, J.; Sadekova, S.; Souleimanova, M.; Zhao, H.; Finak, G.; Meterissian, S.; Hallett, M.T.; Park, M. Gene-expression profiling of microdissected breast cancer microvasculature identifies distinct tumor vascular subtypes. *Breast Cancer Res.* **2012**, *14*. [CrossRef]

89. Le, P.U.; Li, J.; Cantin, C.; Collins, C.; Nantel, A.; Tremblay, T.; Kelly, J.F.; Wang, E.; O'Connor-Mccourt, M.D.; Eg, A. Molecular Characterization of the Vasculature of Patients with Infiltrating Ductal Carcinoma Generates a Gene Signature Predictive of Breast Cancer Survival. *J. Cancer Biol. Res.* **2015**, *3*, 1068.

90. Dhage, S.; Ernlund, A.; Ruggles, K.; Axelrod, D.; Berman, R.; Roses, D.; Schneider, R.J. A genomic ruler to assess oncogenic transition between breast tumor and stroma. *PLoS ONE* **2018**, *13*, e0205602. [CrossRef]

91. Ershaid, N.; Sharon, Y.; Doron, H.; Raz, Y.; Shani, O.; Cohen, N.; Monteran, L.; Leider-Trejo, L.; Ben-Shmuel, A.; Yassin, M.; et al. NLRP3 inflammasome in fibroblasts links tissue damage with inflammation in breast cancer progression and metastasis. *Nat. Commun.* **2019**, *10*. [CrossRef] [PubMed]

92. Bauer, M.; Su, G.; Casper, C.; He, R.; Rehrauer, W.; Friedl, A. Heterogeneity of gene expression in stromal fibroblasts of human breast carcinomas and normal breast. *Oncogene* **2010**, *29*, 1732–1740. [CrossRef]

93. 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. Molecular anatomy of breast cancer stroma and its prognostic value in estrogen receptor-positive and -negative cancers. *J. Clin. Oncol.* **2010**, *28*, 4316–4323. [CrossRef]

94. 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. Mechanotransduction and YAP-dependent matrix remodelling is required for the generation and maintenance of cancer-associated fibroblasts. *Nat. Cell Biol.* **2013**, *15*, 637–646. [CrossRef]

95. Davis, S.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847. [CrossRef] [PubMed]

96. 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. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef]

97. Ritchie, M.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef] [PubMed]

98. Durinck, S.; Spellman, P.T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184–1191. [CrossRef] [PubMed]

99. Fisher, R.A. Statistical methods for research workers. In *Breakthroughs in Statistics*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 66–70.

100. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef] [PubMed]

101. Hänzelmann, S.; Castelo, R.; Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinform.* **2013**, *14*, 7. [CrossRef]

102. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.

103. Wickham, H. *ggplot2—Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2017.

104. Ganz, C. rintrojs: A Wrapper for the Intro.js Library. *J. Open Source Softw.* **2016**, *1*, 63. [CrossRef]

105. Azare, J.; Leslie, K.; Al-Ahmadie, H.; Gerald, W.; Weinreb, P.H.; Violette, S.M.; Bromberg, J. Constitutively activated Stat3 induces tumorigenesis and enhances cell motility of prostate epithelial cells through integrin beta 6. *Mol. Cell. Biol.* **2007**, *27*, 4444–4453. [CrossRef] [PubMed]

106. Dauer, D.J.; Ferraro, B.; Song, L.; Yu, B.; Mora, L.; Buettner, R.; Enkemann, S.; Jove, R.; Haura, E.B. Stat3 regulates genes common to both wound healing and cancer. *Oncogene* **2005**, *24*, 3397–3408. [CrossRef]

107. Liberzon, A.; Birger, C.; Thorvaldsdóttir, H.; Ghandi, M.; Mesirov, J.P.; Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **2015**, *1*, 417–425. [CrossRef]

108. Alvarez, J.V.; Febbo, P.G.; Ramaswamy, S.; Loda, M.; Richardson, A.; Frank, D.A. Identification of a Genetic Signature of Activated Signal Transducer and Activator of Transcription 3 in Human Tumors. *Cancer Res.* **2005**, *65*, 5054–5063. [CrossRef]

109. Tell, R.W.; Horvath, C.M. Bioinformatic analysis reveals a pattern of STAT3-associated gene expression specific to basal-like breast cancers in human tumors. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 12787–12792. [CrossRef]

110. Sonnenblick, A.; Brohée, S.; Fumagalli, D.; Vincent, D.; Venet, D.; Ignatiadis, M.; Salgado, R.; van den Eynden, G.; Rothé, F.; Desmedt, C.; et al. Constitutive phosphorylated STAT3-associated gene signature is predictive for trastuzumab resistance in primary HER2-positive breast cancer. *BMC Med.* **2015**, *13*, 177. [CrossRef] [PubMed]

111. Network, C.G.A. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61–70. [CrossRef]

112. 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. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **2013**, *4*. [CrossRef] [PubMed]

113. Klemm, F.; Bleckmann, A.; Siam, L.; Chuang, H.N.; Rietktter, E.; Behme, D.; Schulz, M.; Schaffrinski, M.; Schindler, S.; Trümper, L.; et al. β-catenin-independent WNT signaling in basal-like breast cancer and brain metastasis. *Carcinogenesis* **2011**, *32*, 434–442. [CrossRef]

114. Monteleone, E.; Orecchia, V.; Corrieri, P.; Schiavone, D.; Avalle, L.; Moiso, E.; Savino, A.; Molineris, I.; Provero, P.; Poli, V. SP1 and STAT3 Functionally Synergize to Induce the RhoU Small GTPase and a Subclass of Non-canonical WNT Responsive Genes Correlating with Poor Prognosis in Breast Cancer. *Cancers* **2019**, *11*, 101. [CrossRef]

115. Lopes-Coelho, F.; André, S.; Félix, A.; Serpa, J. Breast cancer metabolic cross-talk: Fibroblasts are hubs and breast cancer cells are gatherers of lipids. *Mol. Cell. Endocrinol.* **2018**, *462*, 93–106. [CrossRef]

116. Yuzhalin, A.E.; Lim, S.Y.; Kutikhin, A.G.; Gordon-Weeks, A.N. Dynamic matrisome: ECM remodeling factors licensing cancer progression and metastasis. *Biochim. Biophys. Acta Rev. Cancer* **2018**, *1870*, 207–228. [CrossRef] [PubMed]

117. Ashton, T.M.; McKenna, W.G.; Kunz-Schughart, L.A.; Higgins, G.S. Oxidative phosphorylation as an emerging target in cancer therapy. *Clin. Cancer Res.* **2018**, *24*, 2482–2490. [CrossRef] [PubMed]

118. 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. Angiogenesis and lymphangiogenesis are downregulated in primary breast cancer. *Br. J. Cancer* **2009**, *101*, 605–614. [CrossRef] [PubMed]

119. Eberhard, A.; Kahlert, S.; Goede, V.; Hemmerlein, B.; Plate, K.H.; Augustin, H.G. Heterogeneity of angiogenesis and blood vessel maturation in human tumors: Implications for antiangiogenic tumor therapies. *Cancer Res.* **2000**, *60*, 1388–1393.

120. Al Sorkhy, M.; Fahl, Z.; Ritchie, J. Cortisol and breast cancer: A review of clinical and molecular evidence. *Ann. Cancer Res. Ther.* **2018**, *26*, 19–25. [CrossRef]

121. 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. Dihydropyrimidine accumulation is required for the epithelial-mesenchymal transition. *Cell* **2014**, *158*, 1094–1109. [CrossRef]

122. Chamberland, J.P.; Antonow, L.T.; Dias Santos, M.; Ritter, B. NECAP2 controls clathrin coat recruitment to early endosomes for fast endocytic recycling. *J. Cell Sci.* **2016**, *129*, 2625–2637. [CrossRef]

123. Zheng, X.; Weigert, A.; Reu, S.; Guenther, S.; Mansouri, S.; Bassaly, B.; Gattenlöhner, S.; Grimminger, F.; Pullamsetti, S.S.; Seeger, W.; et al. Spatial density and distribution of tumor-associated macrophages predict survival in non-small cell lung carcinoma. *Cancer Res.* **2020**, *80*, 4414–4425. [CrossRef] [PubMed]

124. Medrek, C.; Pontén, F.; Jirström, K.; Leandersson, K. The presence of tumor associated macrophages in tumor stroma as a prognostic marker for breast cancer patients. *BMC Cancer* **2012**, *12*, 1–9. [CrossRef] [PubMed]

125. Jackute, J.; Zemaitis, M.; Pranys, D.; Sitkauskiene, B.; Miliauskas, S.; Vaitkiene, S.; Sakalauskas, R. Distribution of M1 and M2 macrophages in tumor islets and stroma in relation to prognosis of non-small cell lung cancer. *BMC Immunol.* **2018**, *19*, 1–13. [CrossRef] [PubMed]

126. Ma, J.-H.; Qin, L.; Li, X. Role of STAT3 signaling pathway in breast cancer. *Cell Commun. Signal.* **2020**, *18*, 33. [CrossRef]

127. Pathria, P.; Gotthardt, D.; Prchal-Murphy, M.; Putz, E.-M.; Holcmann, M.; Schlederer, M.; Grabner, B.; Crncec, I.; Svinka, J.; Musteanu, M.; et al. Myeloid STAT3 promotes formation of colitis-associated colorectal cancer in mice. *Oncoimmunology* **2015**, *4*, e998529. [CrossRef] [PubMed]

128. Savino, A.; Provero, P.; Poli, V. Differential co-expression analyses allow the identification of critical signalling pathways altered during tumour transformation and progression. *Int. J. Mol. Sci.* **2020**, *21*, 9461. [CrossRef] [PubMed]

129. Nitta, H.; Kelly, B.; Allred, C.; Jewell, S.; Banks, P.; Dennis, E.; Grogan, T.M. The assessment of HER2 status in breast cancer: The past, the present, and the future. *Pathol. Int.* **2016**, *66*, 313–324. [CrossRef] [PubMed]

130. Haibe-Kains, B.; Desmedt, C.; Loi, S.; Culhane, A.; Bontempi, G.; Quackenbush, J.; Sotiriou, C. A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. *J. Natl. Cancer Inst.* **2012**, *104*, 311–325. [CrossRef] [PubMed]

131. Hortobagyi, G.N.; Edge, S.B.; Giuliano, A. New and Important Changes in the TNM Staging System for Breast Cancer. *Am. Soc. Clin. Oncol. Educ. Book* **2018**, *38*, 457–467. [CrossRef] [PubMed]