Differential survival and therapy benefit of patients with breast cancer are characterized by distinct epithelial and immune cell microenvironments

Kester, L.; Seinstra, D.; Rossum, A.G.J. van; Vennin, C.; Hoogstraat, M.; Velden, D. van der; ... ; Rheenen, J. van

Differential Survival and Therapy Benefit of Patients with Breast Cancer Are Characterized by Distinct Epithelial and Immune Cell Microenvironments

Lennart Kester[1,2,3], Danielle Seinstra[1,2], Annelot G.J. van Rossum[1], Claire Vennin[1,2], Marlous Hoogstraat[1,2,4], Daphne van der Velden[5], Mark Opdam[1], Erik van Werkhoven[6], Kerstin Hahn[1,2], Iris Nederlof[1], Ester H. Lips[1], Ingrid A.M. Mandjes[7], A. Elise van Leeuwen-Stok[8], Sander Canisius[1,5], Harm van Tinteren[6], Alex L.T. Imholz[9], Johanneke E.A. Portielje[10,11], Monique E.M.M. Bos[12], Sandra D. Bakker[13], Emiel J. Rutgers[14], Hugo M. Horlings[1,15], Jelle Wesseling[1,16], Emile E. Voest[2,5], Lodewyk F.A. Wessels[2,5], Marleen Kok[17,18], Hendrika M. Oosterkamp[19], Alexander van Oudenaarden[3], Sabine C. Linn[1,17,20], and Jacco van Rheenen[1,2,21]

ABSTRACT

Purpose: Extensive work in preclinical models has shown that microenvironmental cells influence many aspects of cancer cell behavior, including metastatic potential and their sensitivity to therapeutics. In the human setting, this behavior is mainly correlated with the presence of immune cells. Here, in addition to T cells, B cells, macrophages, and mast cells, we identified the relevance of nonimmune cell types for breast cancer survival and therapy benefit, including fibroblasts, myoepithelial cells, muscle cells, endothelial cells, and seven distinct epithelial cell types.

Experimental Design: Using single-cell sequencing data, we generated reference profiles for all these cell types. We used these reference profiles in deconvolution algorithms to optimally detangle the cellular composition of more than 3,500 primary breast tumors

of patients that were enrolled in the SCAN-B and MATADOR clinical trials, and for which bulk mRNA sequencing data were available.

Results: This large data set enables us to identify and subsequently validate the cellular composition of microenvironments that distinguish differential survival and treatment benefit for different treatment regimens in patients with primary breast cancer. In addition to immune cells, we have identified that survival and therapy benefit are characterized by various contributions of distinct epithelial cell types.

Conclusions: From our study, we conclude that differential survival and therapy benefit of patients with breast cancer are characterized by distinct microenvironments that include specific populations of immune and epithelial cells.

Introduction

Patients with breast cancer can respond differently to therapeutics even for tumors with the same subtype. Although multiple mechanisms may explain heterogeneity in therapy benefit, preclinical work has shown that the cellular composition of a tumor can influence

metastasis and treatment efficacy (1–8). For instance, clinical outcomes upon different interventions often correlate with immune infiltrate (9–12). However, the role of different epithelial and nonimmune cell types for the survival and therapy outcome of breast cancer remains understudied, mainly because of the difficulty to detect these cell types in the samples that are available from clinical trials.

[1]Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [2]Oncode Institute-The Netherlands Cancer Institute, Amsterdam, the Netherlands. [3]Oncode Institute-Hubrecht Institute- KNAW & University Medical Centre Utrecht, Utrecht, the Netherlands. [4]Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [5]Division of Molecular Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [6]Department of Biometrics, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [7]Data Center, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [8]Dutch Breast Cancer Research Group, BOOG Study Center, Amsterdam, the Netherlands. [9]Department of Medical Oncology, Deventer Ziekenhuis, Deventer, the Netherlands. [10]Department of Medical Oncology, HagaZiekenhuis, The Hague, the Netherlands. [11]Department of Medical Oncology, Leiden University Medical Center, Leiden, the Netherlands. [12]Department of Internal Oncology, Reinier de Graaf Gasthuis, Delft, the Netherlands. [13]Department of Medical Oncology, Zaans Medisch Centrum, Zaandam, the Netherlands. [14]Department of Surgical Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [15]Division of Pathology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [16]Division of Diagnostic Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [17]Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [18]Division of Tumor Biology & Immunology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [19]Department of Medical Oncology, Haaglanden Medisch Centrum, Den Haag, the Netherlands.

[20]Department of Pathology, University Medical Center, Utrecht, the Netherlands. [21]Molecular Cancer Research, Center Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands.

Translational Relevance

Precision medicine for cancer, a strategy to fine-tune treatments based on the properties of a patient tumor, is rapidly improving the clinical management of patients with cancer. Here, we introduce tumor cell deconvolution (TCD), a deconvolution algorithm that allows us to precisely map the cellular composition of patient tumors based on bulk RNA sequencing of patients' biopsy. We demonstrate that TCD can be used to identify cell populations that are predictive to response to treatment. With the rapid development of sequencing tools and the lowering of costs that accompany this technique, TCD can be employed in the clinic to characterize the cellular compositions of a tumor, to guide treatment decisions and to maximize patient's response to therapies.

Deconvolution algorithms, such as CIBERSORT (13) and DWLS (14), which disentangle bulk mRNA sequencing data into contributions from individual cellular components, provide a good estimate of the cellular composition of tumors (13, 15). These methods have been used to confirm the relevance of tumor-infiltrating immune cells for breast cancer outcomes (16–18). However, the currently available reference profiles for deconvolution (13, 19) were mostly obtained from cells isolated from blood. Therefore, the reference profiles for other cell types in a tumor (e.g., cancer cells, epithelial cells, endothelial cells, muscle cells, and fibroblasts) are largely unknown. In addition, the expression profile of immune cells isolated from the blood and tumors can be different (20), and this can lead to less reliable estimations of the abundance of different cell types (15). Recently, reference profiles obtained from tumor-derived single-cell mRNA sequencing data have been used to deconvolve the cellular components of melanomas and ovarian tumors (14, 15, 19). However, good reference profiles for breast cancer cell types are not yet available.

In this study, we created references profiles of all the cellular components of human breast tumors through single-cell sequencing. We optimally detangle the cellular composition of tumors for which bulk mRNA sequencing data is available, a technique that from here on we refer to as tumor cell deconvolution (TCD). We applied TCD to determine the cellular composition of patients with breast cancer in two clinical trials for which both bulk mRNA sequencing data and clinical outcome are available: (i) 3,000 patients in the multicenter prospective trial Sweden Cancerome Analysis Network - Breast (SCAN-B; ref. 21), and (ii) 528 patients in the multicenter, prospective randomized trial Microarray Analysis in breast cancer to Tailor Adjuvant Drugs Or Regimens (MATADOR; refs. 22, 23). We identify and subsequently validate specific microenvironments that distinguish differential survival and therapy benefits of patients with breast cancer, which include immune and nonimmune cell types (see flow charts for the experimental setup in Supplementary Fig. S1).

Materials and Methods

SCAN-B trial data

We obtained the SCAN-B mRNA sequencing data and anonymized patient data through Gene Expression Omnibus (GEO) GSE81538.

MATADOR trial data

The MATADOR (ISRCTN61893718) study is an open-label, multicenter randomized, phase III trial conducted in 29 centers in the Netherlands. Six hundred and sixty-four female patients with pT1–3,

N0–3, M0 breast cancer were recruited onto the trial. The inclusion criteria were described in detail elsewhere (22, 23).

The study protocol and amendments were approved by the ethical committee of the Netherlands Cancer Institute. The study was conducted in agreement with Good Clinical Practice guidelines and with the Declaration of Helsinki. All patients provided written informed consent to participate in the trial and to use the tumor tissue removed at surgery for translational research. The Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) criteria were used to report this study (38).

Treatment

Patients were randomly assigned (1:1) to 6 cycles of doxorubicin 60 mg/m^2 and cyclophosphamide 600 mg/m^2 every 2 weeks [dose dense (dd) AC] or docetaxel 75 mg/m^2, doxorubicin 50 mg/m^2, and cyclophosphamide 500 mg/m^2 every 3 weeks (docetaxel-doxorubicin-cyclophosphamide; TAC) by means of the automated ALEA system (FormsVision BV) using Pocock minimization technique (24). Randomization was performed centrally at the Netherlands Cancer Institute. G-CSF (pegfilgrastim) was given to all patients. Radiotherapy and endocrine therapy were given according to the contemporary Dutch guidelines.

Patient inclusion criteria for primary endpoint analysis of MATADOR trial

Out of 664 patients, tumor tissue was available for gene expression analysis for 604 (90.9%) patients. Library preparation failed in 7 patients. Six of 597 sequenced samples did not meet the quality checks showing up as extreme outliers based on read counts per gene and in a principal component analysis (PCA). Sixty-two patients had a clinically low risk of recurrence according to the modified Adjuvant! Online (25) and would not receive adjuvant chemotherapy according to current guidelines. In this group, no events occurred. For one patient, survival data was missing. The remaining 528 patients were used to develop delta treatment score (DTS). For the analysis, we excluded the Her2$^+$ patients.

Objectives and endpoints

The primary objective of the trial was to identify a gene expression profile for recurrence-free survival (RFS) benefit of either dose-dense or taxane-containing chemotherapy and to assess its predictive performance. RFS is defined as the time from randomization to loco-regional recurrence, distant metastasis, or death by any cause, whichever occurred first. The secondary objective was to directly compare RFS, overall survival (OS), and toxicity of the two treatment arms and has been described elsewhere (22, 23).

Statistics for MATADOR trial

The original statistical plan can be found in the MATADOR study protocol. During the course of the study, it became clear that the event rate was lower than expected. Therefore, an amendment was made to the protocol to use a cross-validation method instead of a separation in a training and a validation cohort.

RNA isolation and sequencing

Formalin-fixed paraffin-embedded (FFPE) material was collected between 2004 and 2012. RNA was isolated from FFPE tissue with a tumor cell percentage of at least 40% using the AllPrep DNA/RNA Mini Kit according to the manufacturer's instructions (Qiagen). Quantification and purity were measured using the NanoDrop 2,000 spectrophotometer (Thermo Fisher Scientific) and the 2,100

Bioanalyzer (Agilent Technologies). Libraries of cDNA were constructed with the TruSeq RNA Access Library Prep Kit (Illumina) and single-end sequenced using the HiSeq 2,500 (Illumina). Bulk mRNA sequencing data were analyzed using the RSEM package (26) and the reads were aligned to the hg38 reference genome using STAR (27) with default parameters. Transcripts per million values were used in further analyses.

Batch effects of RNA sequencing

To minimize batch effects that could arise during sample processing for RNA sequencing (RNA-seq), large sets of samples were processed within a relatively short time frame. This was a multicenter, randomized controlled trial, and samples were isolated and sequenced by mixing samples from different centers and samples that had been preserved in formalin for different lengths of time to control for possible batch effects. During the computational analyses, we also checked for batch effects caused by the isolation and sequencing runs by PCA, but none could be identified.

Generation of single-cell suspension and staining for flow cytometry

Following pathologic examination, tumor material was kept on ice in DMEM/F12 + GlutaMAX (GIBCO, Invitrogen Life Technologies), 1% HEPES, 1% penicillin–streptomycin from that point onwards. Tumors were mechanically dissociated using sterile scalpels and enzymatically digested with 25 µg/mL DNase I (Roche) and 15 Wünsch units TH Liberase/mL (Roche) diluted in PBS at 37°C for 35 minutes, followed by mashing through a 70-µm filter (BD Falcon) while adding DMEM/F12 + GlutaMAX (GIBCO, Invitrogen Life Technologies), 1% HEPES, 1% penicillin–streptomycin. Tumor cells were blocked in 80% FACS buffer [5 mmol/L EDTA in PBS supplemented with 5% FCS/20% serum mix (50/50 normal goat serum; monx10961, Monosan) and FcγII/III receptor blocking serum 2.4G2 (kind gift from L. Boon, Bioceros, the Netherlands)] for 10 minutes on ice before labeling with EpCAM PE (1B7, eBioscience), CD14-FITC (M-P-9, BD Biosciences) labeled antibodies for 40 minutes on ice. DAPI was added as a cell death marker, and cells were sorted either on a FACS AriaII Special Ordered Research Product (BD Biosciences), on a FACS Jazz (BD biosciences), or on a FACS Fusion (BD biosciences). A broad FSC SSC gate was followed by a gate excluding doublets, after which the DAPI-negative cells were selected. These DAPI-negative cells were sorted into 384-well plates and further processed for single-cell sequencing. To obtain enough epithelial (tumor) cells and macrophages, additionally EpCAM-positive and CD14-positive cells were selected and index-sorted into 384-well plates and further processed for single-cell sequencing. For 3 of 12 patients, only EpCAM-positive and CD14-positive cells were isolated and sequenced.

Single-cell mRNA sequencing

Single-cell mRNA sequencing was performed using Sort-seq as described in Muraro and colleagues (28). Single-cell libraries were sequenced on an Illumina NextSeq500 with 75-bp paired end reads. Read1 contains the cell barcode and Unique Molecule Identifier; read2 was mapped to the hg38 RefSeq transcriptome using Burrows-Wheeler Aligner with standard parameters. Single-cell analysis was performed using Seurat (29). For each of the sorts, the percentage of living cells varied between 1% and 5% of the total number of cells. In total, 14,592 cells from 12 different patients were sorted. Cells were required to have at least 2,000 unique protein coding transcripts and subsequently log-normalized to 10,000 transcripts per cell. With these filtering criteria, 6,580 were used for further analysis. The 5,000 most variable features were selected for clustering. Clusters were identified using the first 21 PCs with a resolution of 0.1; the cutoff for the number of components used in the clustering analysis was chosen based on the elbowplot form the Seurat package. The resolution of 0.1 yielded a relatively small number of clusters but with well-defined expression profiles. Uniform Manifold Approximation and Projections (UMAP) were also constructed using the first 21 PCs. We did not observe much variation in the expression profiles of the noncancerous cells between patients. This indicates that the batch effects between patients are not strong enough to cause separate clustering of the same cell type from different patients. Therefore, there was no need for correcting batch effects.

Bulk mRNA sequencing analysis single-cell cohort

Bulk mRNA sequencing data for both the MATADOR cohort and the single-cell data were analyzed using the RSEM package (26); reads were aligned to the hg38 reference genome using STAR with default parameters (27). Transcripts per million values were used in further analyses. The count data for the SCAN B cohort was downloaded through GEO. The SCAN B data were mapped to the same reference genome (hg38) using the mapping algorithm Bowtie2. No batch correction algorithm was required nor used to match the MATADOR and SCAN B data sets.

Reference profile construction

To construct the reference profiles, all single cells belonging to the same Seurat cluster were averaged.

Cell-type contribution predictions using TCD

DWLS was used with default parameters using the single-cell derived reference profiles described above as reference profiles.

Prognostic cell type score from Cox proportional hazard model with feature selection

The cell type score (CTS_p) is a score based on the deconvolution-derived cell type abundances and is predictive for prognosis. The CTS_p is the hazard for every patient is calculated as follows.

$$CTS_p = \beta_1 \cdot CT_p \qquad (A)$$

where β_1 is the collection of coefficients for the selected cell types, and CT_p is the abundance of the selected cell types in patient p.

To find the optimal combination of cell types, a Cox proportional hazard model for each individual feature (c_i) is fitted. From this the N best features are selected, which are the N features that give the highest partial likelihood for the Cox proportional hazard model. For each of these N best features a second feature, c_2, is selected by iteratively fitting a Cox model including the effects of c_1 and c_2. Finally the N best combinations of two features are selected based on the partial likelihood of the Cox proportional hazard models. This process is repeated until all features, $c_1..c_n$, are included in the model, resulting in many models each with a different combination of input features. Finally, the model with the optimal combination of features (x), $c_1..c_x$, is selected, which is the model with the lowest cross validated P value. Association of the CTS with RFS was tested in a Cox proportional hazard model corrected for the clinicopathologic features treatment, tumor size, lymph node status, pathologic subtype, histologic grade, age, and type of surgery.

Cell type–based DTS

The DTS is a score that is predictive for chemotherapy benefit and is calculated as follows. First, the optimal combination of features (cell types) is determined using a forward feature selection routine, where the complete set of possible features (n) are the 15 cell types predicted by DWLS. A Cox proportional hazard model for each individual feature (c_i), in combination with T and the interaction between the feature and T is fitted. From this, the N best features are selected, which are the N features that give the highest partial likelihood for the Cox proportional hazard model. For each of these N best features a second feature, c_2, is selected by iteratively fitting a Cox model including the main effects of c_1, c_2, and T, and the pairwise interactions between c and T. Finally the N best combinations of two features are selected based on the partial likelihood of the Cox proportional hazard models. This process is repeated until all features, $c_1..c_n$, are included in the model, resulting in many models each with a different combination of input features. Finally, the model with the optimal combination of features (x), $c_1..c_x$, is selected, which is the model with the lowest cross validated P value. The coefficients from the Cox proportional hazard with the lowest cross validated P value as determined in the training set can be used to calculate a cell-type hazard score for each patient in the test set. The HS^{CT} is calculated as follows:

$$HS_p^{CT} = \beta_1 \cdot CT_p + \beta_2 \cdot T + \beta_3 \cdot CT_p \cdot T \qquad \text{(B)}$$

where β_1 is the vector of coefficients for the main effects of the selected cell types, CT_p is the abundance of the selected cell types in patient p, β_2 if the coefficient for the main effect of treatment, and β_3 is the vector of coefficients for the interaction effects between the selected cell types and treatment. Finally, the DTS^{CT} is calculated as follows:

$$DTS_p^{CT} = HS_{p,\,T=1}^{CT} - HS_{p,\,T=0}^{CT} \qquad \text{(C)}$$

where $HS_{p,\,T=1}^{CT}$ is the HS^{CT} for patient p when treated with treatment T_1 (TAC) and $HS_{p,\,T=0}^{CT}$ is the HS^{CT} for patient p when treated with treatment $T=0$ (dose-dense doxorubicin–cyclophosphamide; ddAC). The DTS^{CT} is therefore the log ratio in hazard for a patient when treated with TAC compared with when treated with ddAC. Predictive power of the DTS^{CT} was tested in a Cox proportional hazard model with the main effect of the DTS^{CT}, the main effect of treatment, and the interaction between treatment and the DTS^{CT}, corrected for clinicopathologic features tumor size, lymph node status, pathologic subtype, histologic grade, age, and type of surgery.

Determination of CTS and DTS cutoffs for the various risk groups

Both the CTS and the DTS are a continuum without clear cutoffs. To construct Kaplan–Meier graphs, cutoff points were chosen on the upper and lower quartile for CTS: 25% of patients with the lowest risk, 25% of patients with the highest risk, and 50% of patients with intermediate risk as determined by CTS. To construct Kaplan–Meier graphs for DTS, cutoff points were chosen on the upper and lower half of DTS: 50% of patients with the lowest risk and 50% of patients with the highest risk as determined by DTS.

Generation of training and validation sets

For CTS, the patient cohort of the SCAN-B trial was split in three groups randomly with forced equal distribution of the clinical variables. Next, the first group was used for training and the last two groups were combined for validation. For DTS, the patient cohort of the MATADOR trial was split in two groups randomly with forced equal distribution of the clinical variables. Next, the first group was used for training and the last two groups were combined for validation.

Data availability statement

The single-cell mRNA sequencing data are available through GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE168410. The bulk mRNA sequencing data is available through GEO: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167977. The private sensitive patient data of the MATADOR study requires controlled access. Access to this dataset will be managed locally by the Translational Review Board of the Netherlands Cancer Institute by contacting Mark Opdam (m.opdam@nki.nl). The code for Treatment Interaction Predictor is made available through CodeOcean (https://codeocean.com/capsule/6732361/tree/v2).

Results

Generation of independent reference profiles using single-cell mRNA sequencing of breast tumors

To optimally deconvolve the cellular composition of a tumor from bulk mRNA sequencing data through deconvolution algorithms, it is necessary to generate reliable reference profiles from the various cellular components of the tumor. To obtain these reference profiles, we randomly sorted living DAPI-negative cells from freshly isolated breast tumors of twelve patients using flow cytometry and performed single-cell sequencing (Supplementary Table S1). These patients were not part of the SCAN-B or MATADOR clinical trials that are used later in the manuscript for deconvolution. We additionally isolated, respectively, EpCAM-positive and CD14-positive cells to acquire enough cancer cells and macrophages. Next, we processed the samples for single-cell mRNA sequencing and characterized the expression profiles of in total 6,580 cells, including cancer and noncancer cells. In addition to single-cell mRNA sequencing, for 7 patients we collected enough material to also analyze the tumors by bulk mRNA sequencing and the cellular composition by pathological examination of hematoxylin and eosin (H&E) sections (see below).

Identification of the various cell types present in the tumor microenvironment

To identify all the different cell types in the 6,580 cells that were sequenced, we performed cluster analysis using the Seurat R package and identified 15 clusters of cells with distinct expression profiles (**Fig. 1A**). The gene expression profiles showed that most of the clusters comprising cells of multiple patients represent the various non-transformed cell types present in the tumor microenvironment including cancer-associated fibroblasts (CAF), myoepithelial cells, muscle cells, endothelial cells, macrophages, T cells, B cells, and mast cells (**Fig. 1B** and **C**; Supplementary Fig. S2). Importantly, we did not observe coclustering of microenvironmental cells depending on treatment status of the patient (**Fig. 1B**; Supplementary Fig. S3A). These data suggest that if expression profiles change upon neoadjuvant treatment, these changes were not large enough to lead to separate clusters of neoadjuvant treated and untreated microenvironmental cells in our analyses (**Fig. 1B**). Although some epithelial cell (EC) clusters (I, II, and V) contain cells isolated from both chemotherapy-naïve and neoadjuvant-treated tumors, other EC clusters (III, IV, VI, and VII) only contain cells isolated from either naïve or treated tumors (**Fig. 1B**; Supplementary Fig. S3A). Importantly, because these EC clusters are derived from individual patients, differential expression between these clusters may therefore simply indicate variations between individuals.

Next, to identify which clusters can contain noncancerous epithelial cells, we isolated and sequenced cells from the contralateral non-diseased breast tissue from patient 7. The noncancerous epithelial cells
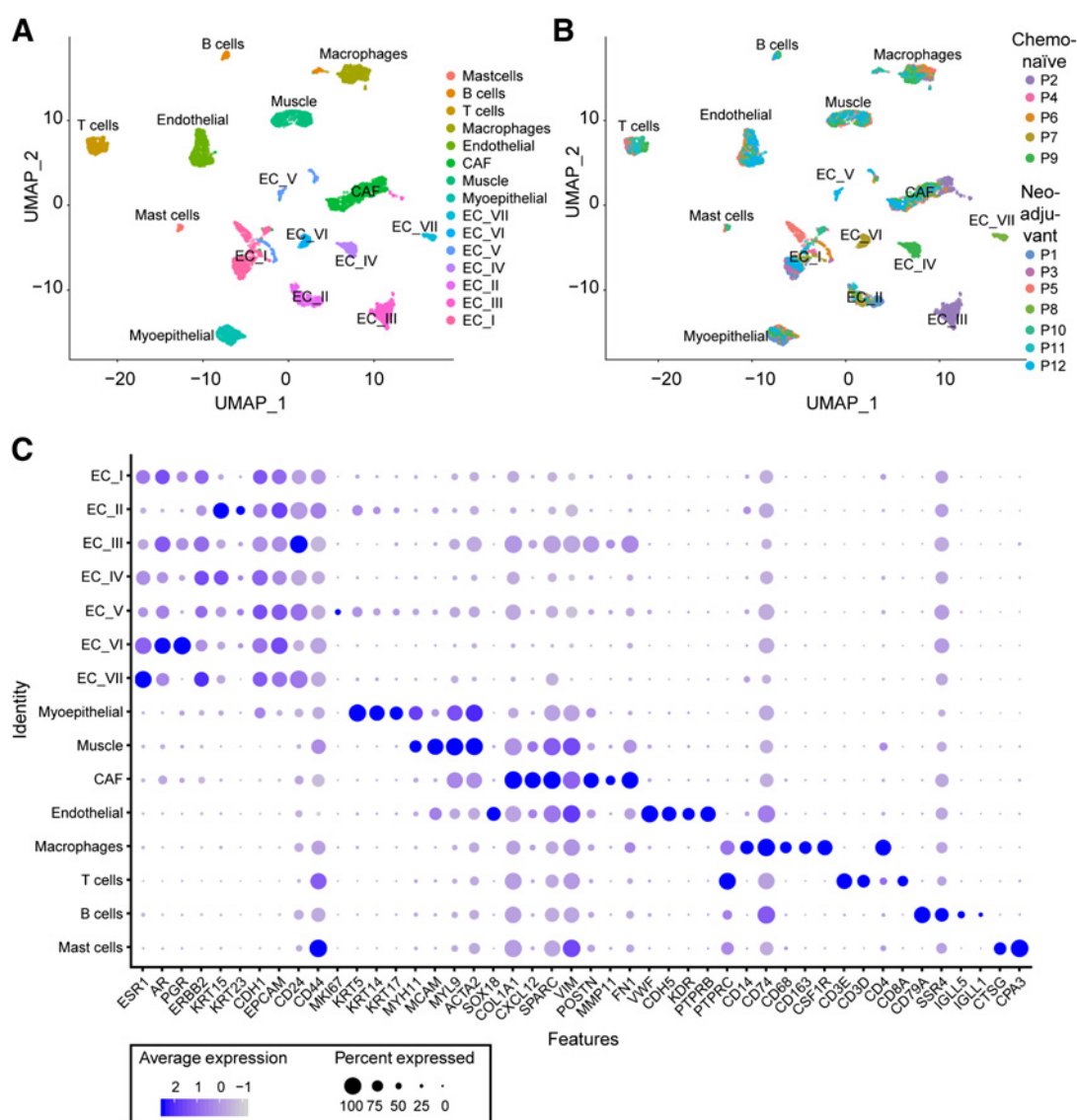
Figure 1.
Heterogeneity of breast tumors analyzed through single-cell mRNA sequencing. **A,** UMAP dimensional reduction of all cells using the Seurat analysis suite. Each color and each number represent an individual cluster. **B,** The same UMAP plot as in **A** highlighting the patient of origin for each single cell. **C,** Expression of cell type–specific marker genes across all cell clusters. Size of the dot indicates the percentage of cells within a cluster that express a marker gene; the color indicates the average level of expression in a cluster. The marker genes are examples of significantly enriched genes that can be found in each individual cell cluster.

of patient 7 were present in EC-I and EC-II, suggesting that these clusters contain a combination of transformed and nontransformed cells (Supplementary Fig. S3B). This was confirmed by the low variance of the copy-number variation (CNV) across the genome in these clusters (Supplementary Fig. S3C) compared with the higher variance of the CNVs in the other clusters (Supplementary Fig. S3C). Of the nontransformed cell cluster, EC-I had significantly increased expression of the hormone receptors *ESR1*, *AR*, and *PGR* compared with cluster EC-II (Supplementary Fig. S3D), suggesting a hormone receptor–positive luminal origin. In addition, we performed the PAM50 analysis on the single-cell data and we mapped the expression of *KRT14*, *KRT5*, *ERBB2*, and *Ki67* (Supplementary Fig. S4). This analysis shows that individual cells, assigned to different subtypes by the PAM50 classification or specific marker genes, are

distributed throughout the different epithelial cell types that we identified. Consequently, the various EC cell types do not represent different known subtypes, but rather a fine-tuned classification of intratumoral heterogeneity. Combined, these data show that single-cell mRNA sequencing of breast tumors allowed us to identify the expression profiles of the most prominent cell types that are present in breast tumors.

Deconvolution of bulk mRNA data reveals robustly the cellular composition of tumors

Based on the cluster analysis, we identified 15 different cell types in human breast tumors. Next, we averaged the expression profile of all cells within a cluster to generate reference profiles for each of the 15 identified cell types. We used these reference profiles to optimally

detangle the contribution of each cell type from bulk mRNA sequencing data using TCD. To test the robustness of this approach, in the tumors of the 7 patients for which enough material was collected to perform both bulk mRNA sequencing and pathological examining of H&E sections, we estimated side-by-side the cellular composition of the tumors by TCD and by a well-trained pathologist. Importantly, we observed a strong correlation between H&E-based cell type scoring by a pathologist and cell type scoring by TCD [Pearson correlation > 0.93; 95% confidence interval (CI) 0.84–0.97; Supplementary Data; Supplementary Fig. S5A], showing that TCD can estimate the cellular composition of a tumor with similar precision as a pathologist. To further assess the strength of our approach, we compared our method (that uses DWLS; ref. 14) with other deconvolution algorithms [CIBERSORT (13) and MuSiC (30)]. Considering that important cell types are missing from the default reference data of these algorithms, we used our reference profiles and we applied them to CIBERSORT and MuSiC. We observed a stronger correlation between the pathologist score and TCD deconvolution score when using DWLS (Pearson correlation = 0.93) than when using CIBERSORT (Pearson correlation = 0.83) or MuSiC (Pearson correlation = 0.54), thereby illustrating the robustness of our method (Supplementary Fig. S5A–S5C).

Next, we questioned whether we could further improve the robustness of the TCD-estimated cellular composition by refining the cell type reference profiles if we would increase the size of the single-cell sequencing patient cohort. We redetermined the reference profiles of the different cell types based on iterative sampling of data of different numbers of patients from the cohort of 12 patients. We observed that the correlation does not improve significantly when the cell type reference profiles are based on single-cell sequencing data of a patient cohort larger than 7 (Supplementary Fig. S5D). We next also calculated the correlation per individual cell type that was scored by the pathologist, and again increasing the number of patients from 7 to 12 did not increase the correlation between pathologist scored cell type contribution and deconvolution-derived contribution (Supplementary Fig. S5E). Finally, we found in patients of the MATRADOR trial a strong correlation between pathology TIL scores and TCD TIL scores, confirming the robustness of our method in an independent set of patients (Supplementary Fig. S5F). Combined, these data indicate that TCD can reliably estimate the cellular composition of tumors.

### Distinct cellular compositions of tumors of the same pathologic breast cancer subtype

Based on histopathologic and molecular features, breast cancer can be classified into different subtypes with distinct outcomes and response to therapies, including ER/PR positive and HER2 negative, HER2 positive, and triple-negative tumors (31–34). We used TCD to detangle the cellular composition of over 3,500 tumors of breast cancer patients with different subtypes and treatment histories that were enrolled in the MATADOR and SCAN-B clinical trials, and for which bulk mRNA sequencing data were available (**Figs. 2** and **3**). However, considering that the tumors used to generate our reference profiles did not include Her2$^+$ patients, we excluded these patients from further analysis. From here on, the contribution of a cell type that is determined by TCD will be denoted as cell type^{TCD}.

Although the distinct pathologic subtypes have distinct outcome and response to therapies (31–34), hierarchical clustering of the cell type contribution matrix of patients in the MATADOR and SCAN-B trials lead to clustering of patients with the same pathological subtypes (**Figs. 2** and **3**). Importantly, each pathologic subtype was divided in two or more subclusters with specific cellular compositions. For

example, in the MATADOR trial, we observe two clusters of triple-negative patients, one of which was enriched for recurrence events and showed high contribution from EC-V^{TCD} and low contributions from immune cells including macrophages^{TCD}, B cells^{TCD}, and T cells^{TCD} (**Fig. 2A**, indicated in the left orange dotted box). The other cluster of triple-negative patients was depleted of recurrence events and had low contribution of EC-V^{TCD} and high contribution of immune cells (**Fig. 2A**, indicated in the right blue dotted box). Kaplan–Meier analysis confirmed the differential **RFS** probability of triple-negative patients with high and low total immune infiltrates^{TCD} (sum of T cells, B cells, macrophages, and mast cells; **Fig. 2B**). The same two clusters can be found in the SCAN-B data (**Fig. 3**, blue and orange dotted boxes). Although luminal and triple-negative patients do not completely fall into separate clusters, we did identify an enrichment for triple-negative patients in two clusters with cellular compositions with high contribution of EC-V^{TCD} and low contribution of immune cells, and cellular compositions with low contribution of EC-V^{TCD} and high contribution of immune cells (**Fig. 3A**, indicated in the dotted boxes). Moreover, in line with the MATADOR data, patients in the SCAN-B trial with triple-negative breast tumors with high levels of immune cells have better survival than patients with triple-negative tumor that contain low levels of immune cells (**Fig. 3B**). These data are in line with previous notions that high immune infiltrate is correlated with better clinical outcomes (9, 11, 12, 35–37) illustrating the power of TCD to find relevant microenvironments. Our data also illustrate that in addition to immune cells, specific epithelial cell types are part of specific microenvironments conferring differential clinical outcomes (e.g., the abundance of EC-V).

Differential RFS is characterized by different tumor compositions

Considering the increasing amount of evidence that the cellular composition of a tumor is important both in the development of metastases and in any benefit from treatment (1–6), we aimed to investigate whether differential RFS is characterized by different tumor compositions (i.e., cell type contributions as determined by TCD). For each patient in the SCAN-B trial, we calculated the cell type contributions by TCD. Next, we split the trial in a training set (first 892 patients) and a validation set (last 1,779 patients) with equal distribution of clinical variables (Supplementary Table S2). On the training set, we calculated a cell type score (**CTS**), a Cox regression score based on cell type contribution that is prognostic for the RFS. In our training set, we calculated the CTS in a 10-fold cross validation setting, where within each of the 10 folds, the optimal combination of cell types is selected using forward feature selection. The model with the lowest cross validation P value was selected (see Methods) and subsequently fixed and used in the independent SCAN-B validation set. We observed that the CTS of this fixed model was significantly prognostic in a Cox regression model for survival in the validation set (Likelihood ratio test $P < 1^{-12}$ and $P = 0.003$ corrected for clinicopathologic variables; Supplementary Table S3). We assigned three subgroups: 25% of patients with the lowest risk, 25% of patients with the highest risk, and 50% of the patients with intermediate risk. These subgroups indeed had clear distinct RFS both in the training and validation cohorts as shown in the Kaplan–Meier analyses (**Fig. 4A** and **B**; Supplementary Table S4). To validate these findings in a completely independent patient cohort, we calculated the CTS score for the same fixed model on the MATADOR patients and again observed that the CTS was significantly prognostic for survival in a Cox regression model (Likelihood ratio test $P < 1^{-6}$ and $P < 0.001$ corrected for clinicopathologic variables; Supplementary Table S5). By analyzing the same three
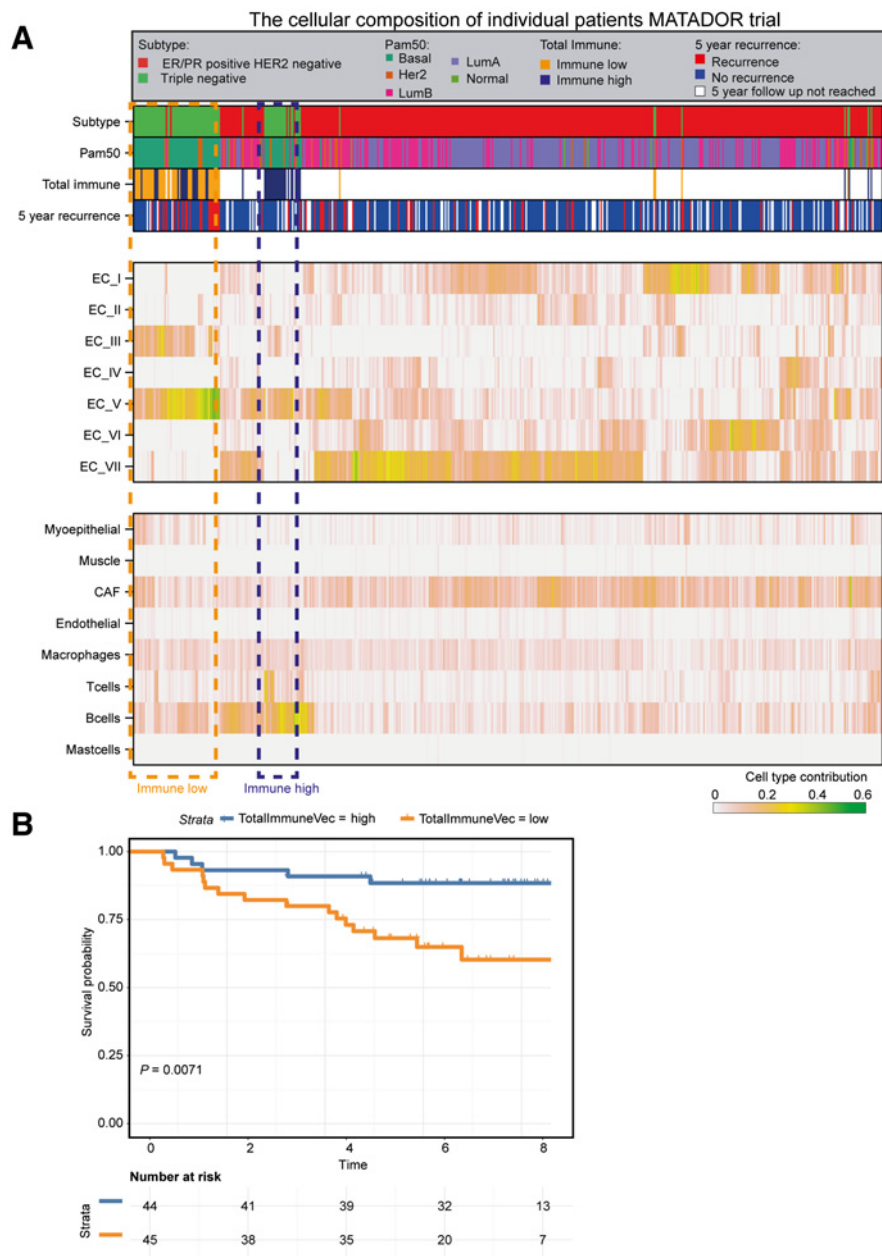
Figure 2.
Deconvolution of the MATADOR bulk mRNA sequencing data. **A,** Results of the deconvolution of the bulk mRNA sequencing data from the MATADOR trial using TCD. The top bars indicate the pathologic subtype, PAM50 classification, total immune content, and 5-year recurrence status of each patient. The middle bars show the contribution of all the ECs. The bottom set of bars shows the contribution from the microenvironmental cell types as defined by TCD. **B,** Overall survival probabilities for triple-negative patients with low (<median) or high (>median) immune infiltrate as determined by TCD.

subgroups in the independent MATADOR cohort, we could validate the differential RFS for the different subgroups (**Fig. 4C**; Supplementary Table S6). Next, we compared the risk scores determined by CTS with the PAM50 subtypes. In the SCAN-B (Supplementary Table S4), but not in the MATADOR cohort (Supplementary Table S6), we found an over-representation of the basal subtype in the high-risk group as well as an over-representation of the luminal A subtype in the low- and intermediate-risk groups. To test the prognostic ability of CTS within individual subtypes, we next split the patient cohorts in the SCAN-B and MATADOR trials into triple-negative and luminal groups. Importantly, in the cohorts with either triple-negative or luminal patients, we observed that CTS also successfully assigned the different risk groups (Supplementary Fig. S6).

Next, we investigated the differences in the cellular composition of the different subgroups with distinct risks (**Fig. 4D** and **E**). The CTS, and therefore the different subgroups, are mainly based on different contributions of EC-II^{TCD} cells, EC-V^{TCD} cells, EC-VI^{TCD} cells, macrophages^{TCD}, T cells^{TCD}, B cells^{TCD}, and myoepithelial cells^{TCD} (**Fig. 4D** and **E**). The low-risk group is enriched for EC-VI^{TCD} cells, T cells^{TCD}, and B cells^{TCD} (**Fig. 4D** and **E**). Our findings that some immune cells are enriched in the tumors of the low-risk group is in line with previous findings that high immune infiltrate is correlated with better clinical outcomes (23, 35, 36), again showing that our method robustly identifies relevant microenvironments. Moreover, in line with the preclinical findings that macrophages can have both tumor-supporting and tumor-suppressing phenotypes, we find

Figure 3.
Deconvolution of the SCAN-B bulk mRNA sequencing data. **A,** Results of the deconvolution of the bulk mRNA sequencing data from the SCAN-B trial using TCD. The top bars indicate the pathologic subtype, PAM50 classification, total immune content, and 5-year recurrence status of each patient. The middle bars show the contribution of all the ECs. The bottom set of bars shows the contribution from the microenvironmental cell types as defined by TCD. **B,** OS probabilities for triple-negative patients with low (<median) or high (>median) immune infiltrate as determined by TCD.
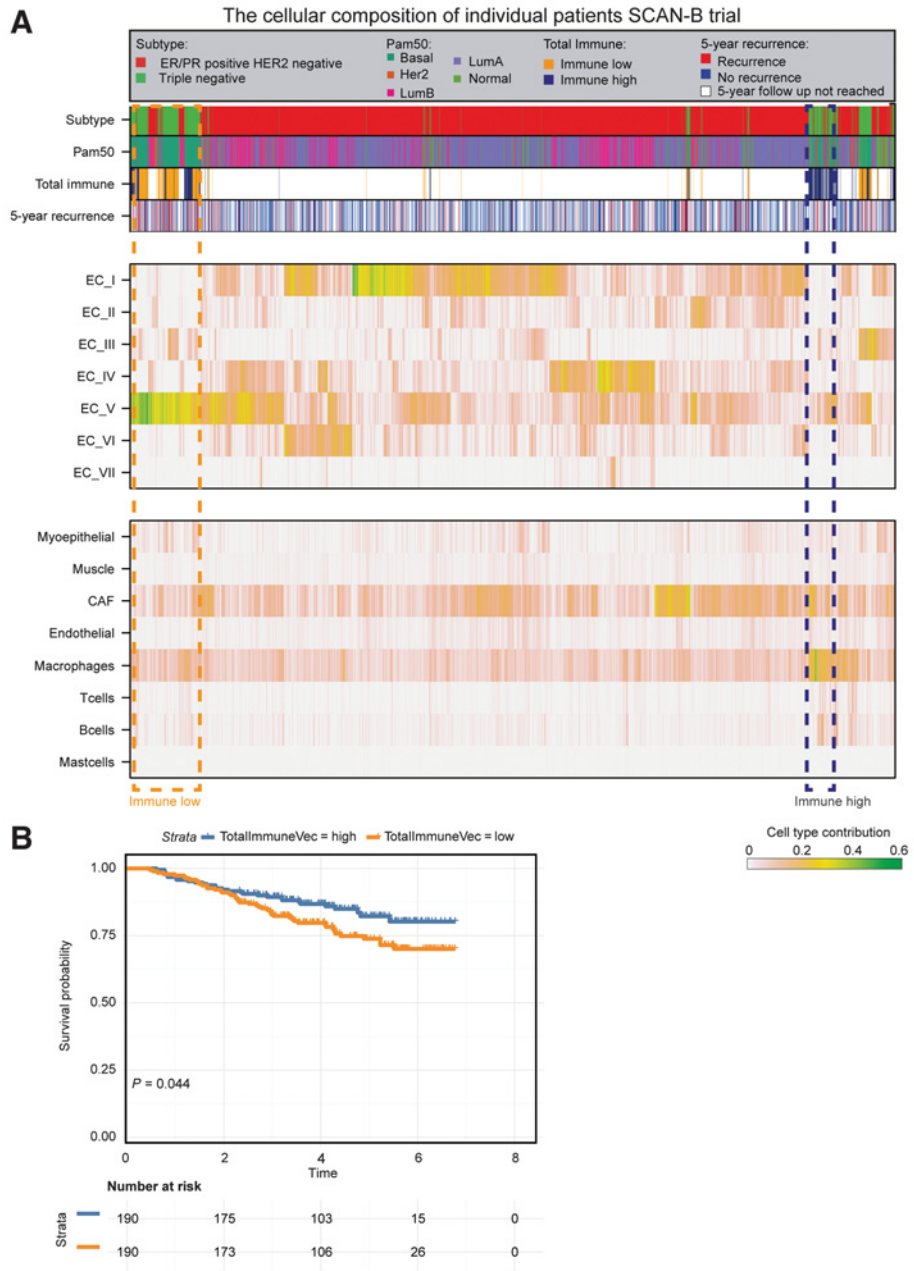
macrophages^{TCD} to be present both in a high- and low-risk group. In addition to identifying immune cells, we also find differential abundance for epithelial cell types (EC-II, EC-V, and EC-VI). For example, the high-risk group has high contribution from EC-II^{TCD} and EC-V^{TCD} cells. Combined, we conclude that TCD of the SCAN-B trial enabled us to identify and subsequently validate specific microenvironments that distinguish differential survival of patients with breast cancer, and show the relevance of various types of immune cells for these environments, but also the relevance of various epithelial cell types.

Differential treatment benefit is characterized by different tumor compositions

Because preclinical and clinical data suggest that the microenvironment of a tumor is important for therapy benefit (1–6), we sought to investigate whether differential outcome after treatment is characterized by distinct cellular compositions as determined by TCD. To do this, we further analyzed the TCD results of the MATADOR trial. In this randomized controlled trial, patients received adjuvant treatment with either ddAC or TAC with a median follow-up of 7 years. We hypothesized that for every cellular composition, there is a certain ratio between the hazard if that patient would be treated with TAC and the hazard if that patient would be treated with ddAC. By subtracting the linear predictors from a Cox model, we obtain a **DTS** that is the log of this HR and directly shows which treatment provides the better outcome: DTS < 0 indicates better outcome when treated with ddAC and DTS > 0 indicates better outcome when treated with TAC. We split the patient cohort of the MATADOR trial in a training set of 254 patients and a validation set of 256 patients with equal distribution of
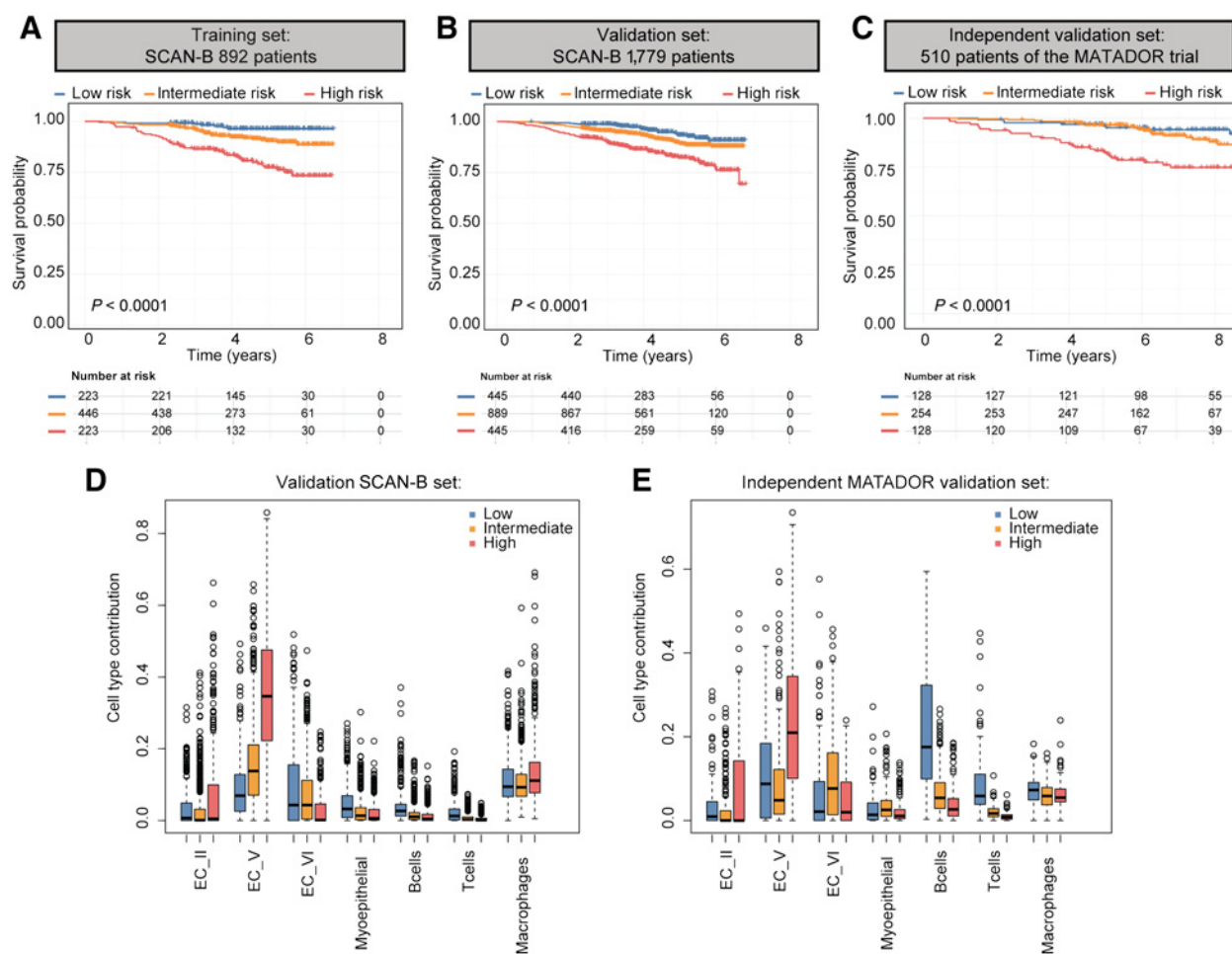
Figure 4.
Considering the complete cellular heterogeneity has prognostic value. **A,** Survival probabilities for the three different patient groups in the SCAN-B training set as ranked by CTS. Determination of the three patient groups is based on quartiles: 25% of patients with the lowest risk, 25% of patients with the highest risk, and 50% of the patients in the intermediate-risk group. **B,** Survival probabilities for the three different patient groups in the SCAN-B validation set. **C,** Survival probabilities for the three different patient groups in the MATADOR cohort. **D** and **E,** EC-II, V, VI and myoepithelial-, B-, T cells, and macrophages are the main contributors for determining CTS. The contribution of these cell types in the three different patient groups for the SCAN-B validation set (**D**) and MATADOR set (**E**).

clinical variables (Supplementary Table S7). To determine the DTS for all cellular compositions (DTS^{CT}), we constructed a Cox proportional hazard model based on the cell type contributions and their interaction with treatment on the training set (see Methods). The optimal combination of cell types is determined through a 10-fold cross validation of the training set of patients (see Methods). We then tested the resulting model in the validation set of patients. There was a significant interaction between the DTS^{CT} and treatment in a Cox proportional hazard model ($P_{\text{interaction}} < 0.001$ and $P_{\text{interaction}} = 0.017$ corrected for clinicopathologic variables; **Fig. 5A**; Supplementary Table S8). Next, we assigned two subgroups: 50% of the patients with the lowest DTS^{CT} and 50% of the patients with the highest DTS^{CT} (Supplementary Table S9). For all patients from the MATADOR trial that had an event, we plotted the DTS^{CT} in a forest plot (**Fig. 5B**). This plot demonstrates that within the patient that had an event in the group with high DTS^{CT}, there is an overrepresentation of patients that were treated with ddAC (**Fig. 5B**). This is in line with the Kaplan–Meier analysis that shows that patients with a high DTS^{CT} have a better survival probability if treated with TAC than with ddAC. This data

shows that DTS can distinguish differential treatment benefits (**Fig. 5C**). The interaction between patients with high DTS^{CT} and patients with low DTS^{CT} and treatment was statistically significant in a Cox proportional hazard model corrected for clinicopathologic variables ($P = 0.047$; Supplementary Table S10).

Next, we analyzed the differential cellular composition between the patient groups with high and low DTS^{CT}. DTS^{CT} is mainly based on the contribution of EC-V, myoepithelial cells, and T cells. This analysis showed that high contributions of T cells^{TCD} and myoepithelial cells^{TCD} and a low contribution of EC-V^{TCD} result in higher DTS^{CT} and therefore represents the specific microenvironment for better outcome on TAC, while high contributions from EC-V^{TCD}, and a low contribution of myoepithelial cells^{TCD} and T cells^{TCD} result in lower DTS^{CT} and therefore represents a microenvironment for better outcome on ddAC (**Fig. 5D** and **E**). From this data, we conclude that differential treatment benefits for ddAC and TAC is characterized by specific microenvironments containing various abundances of immune and epithelial cell types.

Figure 5.
Considering the complete cellular heterogeneity for chemotherapy selection. **A,** Cox proportional hazard model for patients treated with TAC (orange) or ddAC (blue) based on their DTS^{CT} as determined by the Cox model with feature selection (see Methods). x-axis depicts the DTS^{CT}, y-axis depicts the logarithm of the HR, gray areas indicate the 95% confidence interval of the HR. P value depicted in the panel is for the entire Cox proportional hazard model. P value for the interaction between DTS^{CT} and the chemotherapy treatment is <0.001 and <0.001 for uncorrected and models corrected for clinicopathologic factors, respectively. **B,** Two subgroups were assigned: 50% of the patients with the lowest DTS^{CT} and 50% of the patients with the highest DTS^{CT}. Shown is the forest plot for all patients that had an event in the MATADOR trial. The dotted line indicates the separation between the high-risk (high DTS^{CT}) and low-risk (high DTS^{CT}) patients. **C,** Survival probabilities for the high-risk and low-risk patient groups treated with either TAC or ddAC. **D,** Results of the deconvolution of the bulk mRNA sequencing data from the MATADOR trial using TCD. The bars show the contribution from the cell types as defined by TCD. **E,** EC-V, myoepithelial cells, and T cells are the main contributors for determining DTS^{CT}. The plot illustrates the contribution of these cell types in the two patient groups.
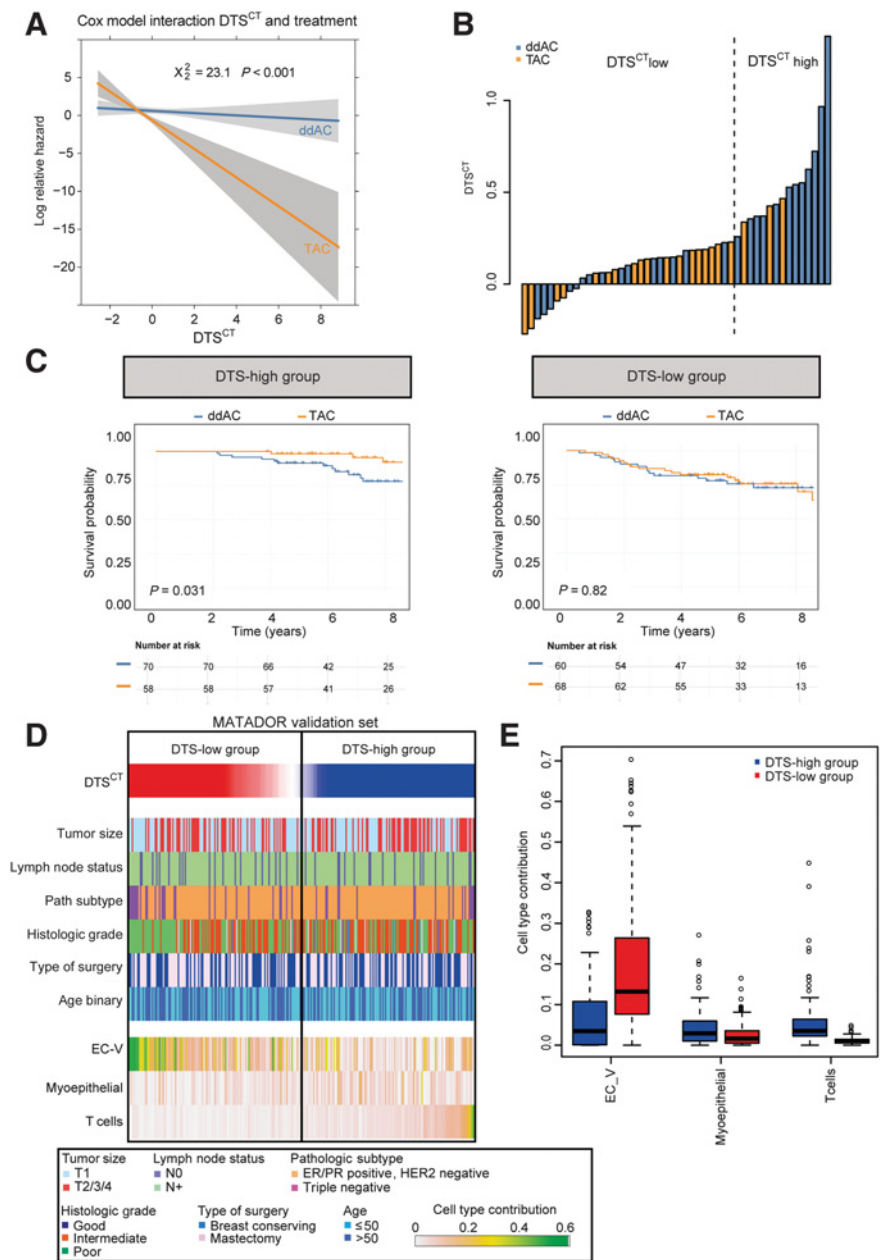
Discussion

Breast cancer is a heterogeneous disease both in terms of cellular composition and therapy benefit. We used deconvolution of bulk RNA sequencing data to unravel the cellular compositions of tumors of patients with differential survival and therapy benefit. Although tumor cellular composition can directly be determined with single-cell mRNA sequencing, this technique cannot be routinely performed for individual patients. Deconvolution methods are an alternative to obtain similar information on the cellular heterogeneity of a tumor from bulk mRNA sequencing data (13, 15). Importantly, if bulk mRNA sequencing is available from clinical trials, it can be used to determine the cellular composition without the need of extensive pathology examination. Moreover, deconvolution can estimate the contribution of cell types

that cannot be distinguished by standard pathology (e.g., contribution of different ECs).

To obtain accurate deconvolution results, we constructed 15 reference profiles from single cells isolated directly from breast tumors. Our reference profiles also have some limitations. First, the reference profiles were constructed using samples from relatively young patients (<57 years) and did not include Her2$^+$ patients. We have therefore excluded the Her2$^+$ patient group in our analyses. Future single-cell mRNA sequencing studies may further improve our reference profiles.

By deconvolution of the bulk mRNA data from patients in the MATADOR and SCAN-B trials, we identified cellular compositions that characterize subgroups of patients with differential survival and treatment benefit. Although breast tumors can be classified to specific IHC subtypes, our approach allows for further refinement of these

subtypes based on specific cellular compositions. We observed that within subtypes, patients' tumors have different cellular compositions which are associated with different survival and therapy benefit. For example, TNBC patients' tumors with a high abundance of immune cells, such as T cells^{TCD}, B cells^{TCD}, and macrophages^{TCD}, have a much better prognosis than their counterpart with tumors with low abundance of immune cells. This is in line with previous work showing that high immune infiltrate in breast and other types of cancer is correlated with better clinical outcomes upon treatments (9, 11, 12, 35). More precisely, we found that high infiltration of immune cells correlates with better response to taxane-containing chemotherapy. It is therefore possible that T cells, B cells, and macrophages, which are abundant in good responders to TAC, may drive the efficacy of taxane *in vivo*. In addition to immune cells, we have identified that survival and therapy benefit are characterized by various contributions of distinct epithelial cell types. Future work should establish the biological differences and functions of these different epithelial cell types.

In conclusion, our study shows that TCD is a powerful tool to determine the cellular compositions of tumors enabling the characterization of specific microenvironments that are associated with differential survival and treatment benefit. This promising method may allow further tailoring of treatment in the future, ultimately leading to higher survival rates and/or better quality of life.

Authors' Disclosures

L. Kester reports a patent for "Method for determining cellular composition of a tumor" (18020578.3) pending. E.H. Lips reports grants from CRUK/KWF outside the submitted work. A.E. van Leeuwen-Stok reports grants from Roche, Novartis, Philips, Servier, and Agendia outside the submitted work. M.E.M.M. Bos is on the advisory board at AstraZeneca on Trastuzumab-deruxtecan (Europe). H.M. Horlings reports other support from Roche Diagnostics BV during the conduct of the study. J. Wesseling reports research not related to the research presented in this manuscript as funded by Cancer Research UK, KWF Dutch Cancer Society ZonMW, and the Antoni van Leeuwenhoek Foundation. E.E. Voest reports grants from Roche, Pfizer, Clovis, Novartis, AstraZeneca, BMS, Eli Lilly, MSD, Amgen, Bayer, and Sanofi outside the submitted work. L.F.A. Wessels reports grants from Genmab BV during the conduct of the study. M. Kok reports grants from BMS, Roche, and AstraZeneca, as well as other support from Daiichi outside the submitted work. H.M. Oosterkamp reports grants from Amgen, Sanofi, and Dutch Cancer Society during the conduct of the study, as well as personal fees from Roche, MSD, Daiichi, AstraZeneca, Lilly, and Pfizer outside the submitted work. S.C. Linn reports grants from Amgen, Sanofi, and Dutch Cancer Society during the conduct of the study. S.C. Linn also reports grants and nonfinancial support from AstraZeneca, Genentech/Roche, Tesaro (now owned by GSK), and Immunomedics; grants from Eurocept-pharmaceuticals, Novartis, and Pfizer; and other support from Cergentis, IBM, Daiichi Sankyo, and Bayer outside the submitted work. J. van Rheenen reports a patent for "Method for determining cellular composition of a tumor" (18020578.3) pending. No disclosures were reported by the other authors.

Authors' Contributions

L. Kester: Conceptualization, software, formal analysis, validation, methodology, writing–original draft, writing–review and editing. **D. Seinstra:** Investigation, writing–original draft, methodology. **A.G.J. van Rossum:** Resources, validation, investigation, methodology. **C. Vennin:** Data curation, investigation, writing–review and editing. **M. Hoogstraat:** Resources, data curation, formal analysis. **D. van der Velden:** Resources. **M. Opdam:** Resources, data curation, formal analysis, validation.

E. van Werkhoven: Data curation, methodology. **K. Hahn:** Data curation, supervision, investigation. **I. Nederlof:** Resources, data curation, validation. **E.H. Lips:** Resources, data curation, supervision. **I.A.M. Mandjes:** Resources. **A.E. van Leeuwen-Stok:** Resources. **S. Canisius:** Data curation, validation, methodology. **H. van Tinteren:** Resources. **A.L.T. Imholz:** Resources. **J.E.A. Portielje:** Resources. **M.E.M.M. Bos:** Resources. **S.D. Bakker:** Resources. **E.J. Rutgers:** Resources. **H.M. Horlings:** Resources, data curation, supervision. **J. Wesseling:** Resources, data curation, supervision. **E.E. Voest:** Resources, data curation, supervision, investigation. **L.F.A. Wessels:** Data curation, software, supervision, investigation, methodology. **M. Kok:** Resources, supervision. **H.M. Oosterkamp:** Conceptualization, resources, supervision, funding acquisition. **A. van Oudenaarden:** Conceptualization, data curation, supervision. **S.C. Linn:** Conceptualization, resources, supervision, funding acquisition, investigation, writing–original draft, writing–review and editing. **J. van Rheenen:** Conceptualization, supervision, funding acquisition, writing–original draft, writing–review and editing.

References

1. Sherman-Baust CA, Weeraratna AT, Rangel LB, Pizer ES, Cho KR, Schwartz DR, et al. Remodeling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells. Cancer Cell 2003;3: 377–86.
2. Gilbert LA, Hemann MT. DNA damage-mediated induction of a chemoresistant niche. Cell 2010;143:355–66.
3. Nakasone ES, Askautrud HA, Kees T, Park JH, Plaks V, Ewald AJ, et al. Imaging tumor-stroma interactions during chemotherapy reveals

contributions of the microenvironment to resistance. Cancer Cell 2012;21: 488–503.

4. Sun Y, Campisi J, Higano C, Beer TM, Porter P, Coleman I, et al. Treatment-induced damage to the tumor microenvironment promotes prostate cancer therapy resistance through WNT16B. Nat Med 2012;18:1359–68.

5. Dijkgraaf EM, Heusinkveld M, Tummers B, Vogelpoel LT, Goedemans R, Jha V, et al. Chemotherapy alters monocyte differentiation to favor generation of cancer-supporting M2 macrophages in the tumor microenvironment. Cancer Res 2013;73:2480–92.

6. Tsavaris N, Kosmas C, Vadiaka M, Kanelopoulos P, Boulamatsis D. Immune changes in patients with advanced breast cancer undergoing chemotherapy with taxanes. Br J Cancer 2002;87:21–7.

7. Hirata E, Girotti MR, Viros A, Hooper S, Spencer-Dene B, Matsuda M, et al. Intravital imaging reveals how BRAF inhibition generates drug-tolerant micro-environments with high integrin β1/FAK signaling. Cancer Cell 2015;27:574–88.

8. Cazet AS, Hui MN, Elsworth BL, Wu SZ, Roden D, Chan CL, et al. Targeting stromal remodeling and cancer stem cell plasticity overcomes chemoresistance in triple negative breast cancer. Nat Commun 2018;9:2897.

9. Loi S, Sirtaine N, Piette F, Salgado R, Viale G, Van Eenoo F, et al. Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02–98. J Clin Oncol 2013;31:860–7.

10. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers. J Clin Oncol 2019;37:559–69.

11. Ibrahim EM, Al-Foheidi ME, Al-Mansour MM, Kazkaz GA. The prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancer: a meta-analysis. Breast Cancer Res Treat 2014;148:467–76.

12. Issa-Nummer Y, Darb-Esfahani S, Loibl S, Kunz G, Nekljudova V, Schrader I, et al. Prospective validation of immunological infiltrate for prediction of response to neoadjuvant chemotherapy in HER2-negative breast cancer–a substudy of the neoadjuvant GeparQuinto trial. PLoS One 2013;8:e79775.

13. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015; 12:453–7.

14. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. Nat Commun 2019;10:2975.

15. Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, et al. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. Nat Commun 2017;8:2032.

16. Zhang SC, Hu ZQ, Long JH, Zhu GM, Wang Y, Jia Y, et al. Clinical implications of tumor-infiltrating immune cells in breast cancer. J Cancer 2019;10:6175–84.

17. Ali HR, Chlon L, Pharoah PD, Markowetz F, Caldas C. Patterns of immune infiltration in breast cancer and their clinical implications: A gene-expression-based retrospective study. PLoS Med 2016;13:e1002194.

18. Bense RD, Sotiriou C, Piccart-Gebhart MJ, Haanen J, van Vugt M, de Vries EGE, et al. Relevance of tumor-infiltrating immune cell composition and functionality for disease outcome in breast cancer. J Natl Cancer Inst 2017; 109:djw192.

19. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37:773–82.

20. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. Cell 2018;174:1293–308.

21. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 2015;7:20.

22. van Rossum AGJ, Kok M, McCool D, Opdam M, Miltenburg NC, Mandjes IAM, et al. Independent replication of polymorphisms predicting toxicity in breast cancer patients randomized between dose-dense and docetaxel-containing adjuvant chemotherapy. Oncotarget 2017;8:113531–42.

23. van Rossum AGJ, Kok M, van Werkhoven E, Opdam M, Mandjes IAM, van Leeuwen-Stok AE, et al. Adjuvant dose-dense doxorubicin-cyclophosphamide versus docetaxel-doxorubicin-cyclophosphamide for high-risk breast cancer: First results of the randomised MATADOR trial (BOOG 2004–04). Eur J Cancer 2018;102:40–8.

24. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975;31:103–15.

25. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 2016;375:717–29.

26. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinf 2011;12:323.

27. Dobin A, Gingeras TR. Mapping RNA-seq reads with STAR. Curr Protoc Bioinformatics 2015;51:11.4.1–.4.9.

28. Muraro MJ, Dharmadhikari G, Grün D, Groen N, Dielen T, Jansen E, et al. A single-cell transcriptome atlas of the human pancreas. Cell Syst 2016;3:385–94.

29. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.

30. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolu-tion with multi-subject single-cell expression reference. Nat Commun 2019; 10:380.

31. Lips EH, Mulder L, de Ronde JJ, Mandjes IAM, Koolen BB, Wessels LFA, et al. Breast cancer subtyping by immunohistochemistry and histological grade out-performs breast cancer intrinsic subtypes in predicting neoadjuvant chemo-therapy response. Breast Cancer Res Treat 2013;140:63–71.

32. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 2012;486:395–9.

33. Blows FM, Driver KE, Schmidt MK, Broeks A, van Leeuwen FE, Wesseling J, et al. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collabo-rative analysis of data for 10,159 cases from 12 studies. PLoS Med 2010;7: e1000279.

34. Russnes HG, Lingjærde OC, Børresen-Dale AL, Caldas C. Breast cancer molec-ular stratification: from intrinsic subtypes to integrative clusters. Am J Pathol 2017;187:2152–62.

35. Loi S, Michiels S, Salgado R, Sirtaine N, Jose V, Fumagalli D, et al. Tumor infiltrating lymphocytes are prognostic in triple negative breast cancer and predictive for trastuzumab benefit in early breast cancer: results from the FinHER trial. Ann Oncol 2014;25:1544–50.

36. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature 2000;406:747–52.

37. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med 2013;19:1423–37.

38. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. REporting recommendations for tumor MARKer prognostic studies (REMARK). Breast Cancer Res Treat 2006;100:229–35.