

Exploring public cancer gene expression signatures across bulk, single-cell and spatial transcriptomics data with signifinder Bioconductor package

Stefania Pirrotta¹, Laura Masatti¹, Anna Bortolato¹, Anna Corrà², Fabiola Pedrini³, Martina Aere¹, Giovanni Esposito⁴, Paolo Martini⁵, Davide Risso⁶, Chiara Romualdi¹ and Enrica Calura^{1,*}

¹Department of Biology, University of Padua, Padua 35121, Italy

²Fondazione Istituto di Ricerca Pediatrica Città della Speranza, Padua 35127, Italy

³Institute of Pathology, University Hospital Heidelberg, Heidelberg 69120, Germany

⁴Immunology and Molecular Oncology Diagnostic Unit of The Veneto Institute of Oncology IOV – IRCCS, Padua 35128, Italy

⁵Department of Molecular and Translational Medicine, University of Brescia, Brescia 25123, Italy

⁶Department of Statistical Sciences, University of Padua, Padua 35121, Italy

*To whom correspondence should be addressed. Tel: +39 0498276234; Email: enrica.calura@unipd.it

Abstract

Understanding cancer mechanisms, defining subtypes, predicting prognosis and assessing therapy efficacy are crucial aspects of cancer research. Gene-expression signatures derived from bulk gene expression data have played a significant role in these endeavors over the past decade. However, recent advancements in high-resolution transcriptomic technologies, such as single-cell RNA sequencing and spatial transcriptomics, have revealed the complex cellular heterogeneity within tumors, necessitating the development of computational tools to characterize tumor mass heterogeneity accurately. Thus we implemented signifinder, a novel R Bioconductor package designed to streamline the collection and use of cancer transcriptional signatures across bulk, single-cell, and spatial transcriptomics data. Leveraging publicly available signatures curated by signifinder, users can assess a wide range of tumor characteristics, including hallmark processes, therapy responses, and tumor microenvironment peculiarities. Through three case studies, we demonstrate the utility of transcriptional signatures in bulk, single-cell, and spatial transcriptomic data analyses, providing insights into cell-resolution transcriptional signatures in oncology. Signifinder represents a significant advancement in cancer transcriptomic data analysis, offering a comprehensive framework for interpreting high-resolution data and addressing tumor complexity.

Introduction

Decades of extensive research in cancer gene expression, conducted on large patient cohorts, have yielded numerous transcriptional signatures as indicators for various cancer phenotypes (1). Signatures are made by specific gene sets, sometimes supported by coefficients to weigh gene contributions, whose expression levels are condensed into final scores. Transcriptional signatures have garnered attention due to their potential to elucidate cancer activities, thereby enhancing therapeutic decisions, monitoring interventions, comprehending cancer mechanisms, delineating tumor subtypes, and evaluating patient diagnosis and prognosis (2,3). Additionally, these scores can be used to explore the intricate interactions between tumors and the tumor microenvironment (TME). This interplay is critical not only in data analysis, as the heterogeneous mixture of cell types can affect tumor purity and bias data analysis, but is also an intrinsic attribute of tumors that warrants consideration for sample characterization (e.g. signatures for monitoring intrinsic and acquired immune resistance (4)).

In recent years, advancements in cancer transcriptome detection, notably through single-cell RNA sequencing (scRNA-

seq) and spatial transcriptomics (ST), revealed that cancer masses are complex cellular mosaics, demonstrating remarkable heterogeneity driven by spatial patterns, clonal cells, and local microenvironmental factors (5–7). Managing this complexity required the creation of computational tools that streamline the characterization of tumor mass heterogeneity by precisely defining cancer cell states within high-resolution transcriptomic data. Similar to a classical signature from bulk sequencings, the cancer cell state is delineated by testing a gene expression module, inferred or predefined, which is condensed into a score that, in this case, is cell-specific (7–12). Recently, a pan-cancer scRNA-seq analysis demonstrated that cancer cells have multiple and non-mutually exclusive states that lead to multiple, spatially defined variations in the tumor stroma (7). Managing this combinatorial complexity requires the development of computational methods capable of automatically defining multiple cancer cell states within high-resolution transcriptomes. While there were promising findings, a comprehensive catalog of relevant single-cell gene modules for all cancers and their TME cells remains elusive. Such a catalog would be invaluable for interpreting genomic data. However,

Received: May 27, 2024. Revised: September 1, 2024. Editorial Decision: September 19, 2024. Accepted: September 24, 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

even with a comprehensive set of modules, information on inter-patient variability levels is likely lacking, as their identification currently relies on the analysis of relatively few single-cell data samples. To overcome this issue—as well as bulk-based gene set pathway analysis tools, which are successfully applied to scRNA-seq data (13)—the plethora of bulk-derived transcriptional signatures can be used to dissect the complexity of high-resolution transcriptomics.

However, even after decades of research, the field of bulk transcriptional signatures still has issues to solve, and although many gene expression-based prognostic signatures have been reported in the literature, very few are used in clinical practice (14). The way to achieve accurate tumor classifications based on transcriptional signatures is debated and furthermore hampered by the lack of standard practices (2,15). In fact, signature reproducibility and dissemination are often affected by the lack of public open-source implementations: most signatures are not published along with their algorithm and very few have been implemented in a ready-to-use software.

Therefore, with the aim of promoting the use of public transcriptional signatures especially for delineating cancer cell states in single-cell and spatial transcriptomics, we developed the R Bioconductor package *signifinder*. *Signifinder* serves as a bridge between signature discovery and signature usability in all transcriptomic data: bulk, single-cell and spatial transcriptomics. To accomplish this task, *signifinder* provides the infrastructure to collect and implement cancer transcriptional signatures available in the public literature. This software has been conceived as an open-source R package built around the gene expression data structures of the Bioconductor project. This framework guarantees the interoperability of *signifinder* with bulk, single-cell, and spatial transcriptomics data analysis workflows, also paving the way for a systematic evaluation of the available signatures. Here, we present three case studies about the use of transcriptional signatures in bulk, single-cell and spatial transcriptomic data. These findings provide evidence on the nature and extent of the use of cell-resolution transcriptional signatures in oncology, potentially leading to new research directions.

Materials and methods

Figure 1 graphically outlines *signifinder* development and the workflow of analyses.

The collection of signatures

We established a set of stringent criteria for the inclusion of signatures: (i) signatures should rely on cancer topics, thus to a specific area of focus within cancer biology or cancer research, and be developed and used on cancer samples; (ii) signatures should rely exclusively on transcriptomic data, except in cases where transcriptomic gene expression levels are combined with signature-specific gene weights; (iii) signatures must release a clear gene list used for the signature definition, where all genes have an official gene symbol (Hugo consortium) or an unambiguous translation (genes without an official gene symbol are removed); (iv) the method to calculate expression-based scores need to be unambiguously described; (v) additional clarity about the type of expression unit (e.g. counts, log counts, FPKM or others) is also required.

The first step for the collection of the signatures was a literature search using the following keywords: ‘cancer’, ‘gene expression’, ‘microarray’ or ‘RNA sequencing’ and ‘signature’, providing an initial set of 2000 journal articles. We then excluded papers on mutational signatures and those including microRNAs or other omic features such as DNA methylations, which accounted for a large part of the considered articles. Then, we focused on articles that proposed a patient-specific summary score. We ended up with 150 papers that were screened manually, applying the above selected criteria. Currently, the package encompasses 72 signatures, spanning 27 cancer topics, encapsulating numerous cancer hallmarks. These include pivotal aspects like epithelial-to-mesenchymal transition (EMT), chromosomal instability (CIN), angiogenesis, hypoxia, diverse metabolic pathways and cell cycle dynamics. Furthermore, the signatures explore the intricate interplay between tumor cell composition and the TME, addressing facets such as cancer stem cell presence, immune system activity, extracellular matrix (ECM) composition, and angiogenesis activity. Additionally, certain signatures are tailored to monitor clinical outcomes, such as chemo-resistance or patient prognosis. Eleven of these signatures are single-cell derived signatures, all the others derive from bulk cancer gene expression data. For a comprehensive list of the signatures collected so far, please refer to [Supplementary Table S1](#). Moreover, the *signifinder* package facilitates the seamless integration of new signatures through ‘pull requests’, a process that is both straightforward and thoroughly documented in the package vignette. The package infrastructure is designed to accommodate and manage signatures derived from bulk, single-cell and spatial transcriptomics. Therefore, it is foreseeable that multiple signatures will be added in the near future. Signature information is finely curated and details about the tested biological process, the type of tumor, the type of omic data, the original data format, as well as the references to the original publication are provided.






The new high-resolution signatures deriving from single or quasi-single cell technologies are starting to appear but they actually lack performance evaluations at the inter-patient variability level, since their identification is currently based on the analysis of relatively few samples. On the other hand, intratumor heterogeneity of bulk signatures can be tested if applied to high-resolution transcriptomes. The interchangeability of signatures across different types of transcriptional omics, as proposed by *signifinder*, would improve signature evaluation and applicability.

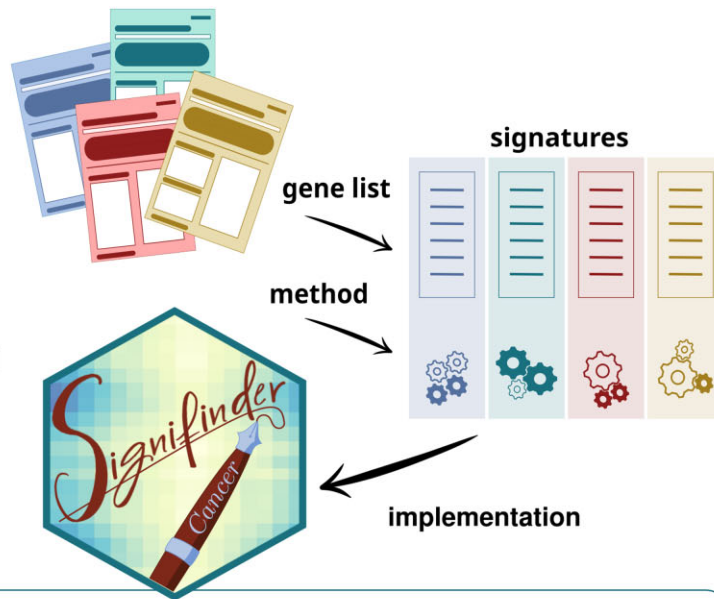
Signature implementation

Within *signifinder*, we implemented a dedicated function for each signature, and whenever possible, signatures that rely on the same topic were grouped together. The list of available signatures and the related functions can be checked through the *signifinder* function *availableSignatures*, which lists all the signature-specific information. Signatures included in *signifinder* are unequivocally named through a combination of the topic (or signature name) and the first author’s name (i.e. ‘Pyroptosis_Ye’ is a signature on pyroptosis activity proposed by Ye *et al.*). Following the rules stated by the authors, *signifinder* provides the required data for the computation of every signature (i.e. the gene list, and the corresponding coefficients and/or attributes). To be compatible with all forms of expression datasets—bulk, single-cell and spatial—*signifinder*

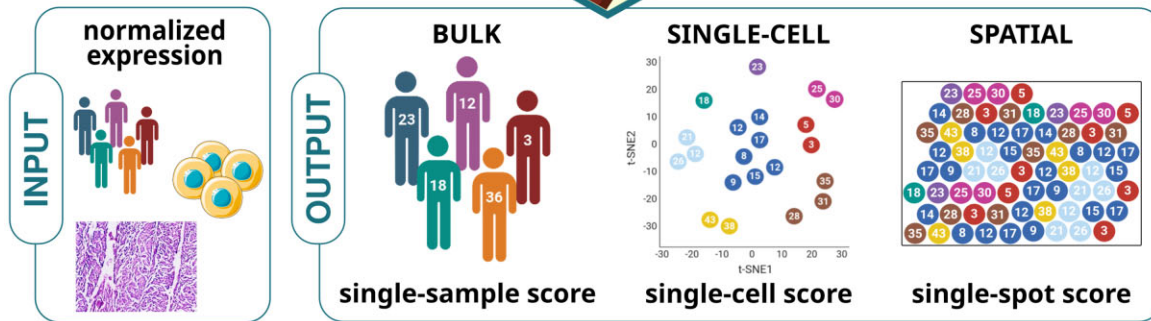
A Strategy Development

Signatures collection

- ✓  **Cancer topic**
- ✓  **Transcriptomic data**
- ✓  **Single sample score**
- ✓  **Unambiguous method**
- ✓  **Accessible gene list**



B Workflow



C Visualization

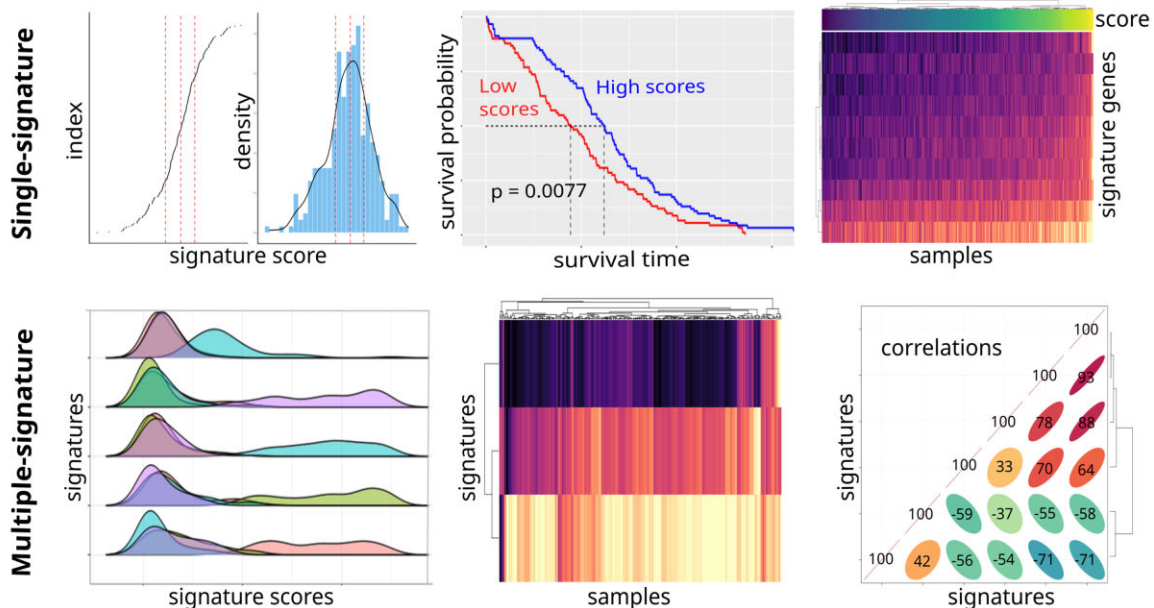


Figure 1. Signifinder implementation and workflow. **(A)** The scheme for signifinder development: following stringent criteria, we collected the lists of genes and implemented the algorithms for signature computations. **(B)** Signifinder workflow starts with gene expression data from bulk, single-cell, or spatial transcriptomics. The user can choose which signature is interested in and compute single-sample, -cell or -spot scores, respectively. **(C)** Signifinder offers several plots to explore and compare the signature score distributions.

accepts as input a gene expression matrix, a data frame, a SummarizedExperiment, a SingleCellExperiment or a SpatialExperiment object. With signifier, users can simply supply the normalized expression data, then signifier takes care of data transformation (if needed) to obtain signature scores computed as stated in the signature's original work. The choice to integrate it in the Bioconductor data structures makes signifier a reliable package, maintained through the years, easily integrated into all R expression data analysis workflows. Additionally, it manages the use of different gene IDs and expression metrics (e.g. counts, CPM, FPKM/RPKM, TPM). Finally, all the signature functions return an S4 object containing the expressions—the same given by the user as input—with the signature scores added to the *colData* section.

Signature quality check

Even if the mathematical approach is feasible and the methodology well established, dataset specific features, not related to the biological nature of data, can affect the results. As an example, one of the critical issues when working with single-cell and spatial data is the sparsity of matrices. scRNA-seq and ST technologies require the detection of tiny amounts of mRNA, which leads to the dropout of many expressed transcripts. Large percentages of indistinguishable biological and technical zero values may be detrimental to downstream signature score computation. Additionally, even if generally smoothed by normalization procedures (16), it would be interesting to know if signatures have scores correlated with technical or biological unwanted variability, such as batch or expression dropout, even taking into consideration samples, cells, or spots divided by annotations. As an example, the correlations of signature scores with the total counts or total zero-value percentages, can highlight normalization problems and indicate the need for further assessment before making biological considerations. On the other hand, in spatial transcriptomic data, the total counts per spot reflect relevant quantitative and qualitative biological features of the tissue morphology, due to the different numbers and types of captured cells by each single spot area (17). Thus, in spatial technologies, relatively high correlations could be an intrinsic feature of the data, meaning that biological signals must be understood in context.

To screen and monitor these data behavior and provide a reliability score for each signature, signifier provides the *evaluationSignPlot* function that can be used to investigate critical behavior of scores in all samples, cells and spots, or a subset of them.

The *evaluationSignPlot* function returns a multi-panel plot showing statistics related to each signature: (i) the goodness of the signature for the user's dataset, ranging from 0, worst goodness, to 100, best goodness. Goodness summarized the parameters shown in the other panels: (ii) the percentage of genes of the signature available in the dataset; (iii) the percentage signature genes with zero values for each sample/cell/spot; (iv) the correlation between signature scores and the per sample/cell/spot total read counts; (v) the correlation between signature scores and the overall percentage of zero values per sample/cell/spot.

Examples of these plots are provided for all the three case studies in supplementary materials (Supplementary Figures S1, S2 and S5). The analysis helps to visually find signatures that were calculated using too few genes or with low gene expressions, as well as signatures that could contain technical or

biological bias for that specific dataset. Overall, this plot allows investigators to make informed decisions about signature inclusion in downstream analyses.

Signifier package implementation details

Signifier is an open-source R package, available from the Bioconductor platform (<https://www.bioconductor.org/packages/release/bioc/html/signifier.html>). It is released under the AGPL-3 license and requires R version 4.2.0 or higher. The source code and documentation are freely available through the Bioconductor platform. The developing version of the package is available in the dedicated GitHub repository (<https://github.com/CaluraLab/signifier>). The R code that was used for the three case studies presented in this work is also available in a dedicated GitHub repository (https://github.com/CaluraLab/signifier_workflow).

Results

Figure 1B presents the analysis flow by using the signifier package. Users can input cancer sample transcriptional data from microarray or RNA-seq, single-cell sequencing, or spatial transcriptomic technology. The user can select the signature by tumor type or tissue obtaining signature scores at the sample, cell, or spot level (Figure 1B). Signifier also provides graphical summaries for visualizing single signatures or comparing multiple signatures (as shown in Figure 1C).

Signature analysis procedures and graphical summaries

Users are prompted to provide normalized expression values for microarrays or normalized counts for sequencing technologies, along with specifying the data type (sequencing or microarray) and the gene ID type used. Signifier then executes requested signatures with a single command. Signatures can be chosen based on cancer topic, type, tissue or a combination thereof. The tool offers various methods for visually examining the scores. Users can explore score distribution or its relationship with survival data through single signature plots. Heatmaps can be used to visualize gene expression and identify top contributor genes. Furthermore, users can compare multiple signatures using ridge plots (which can also be split by user-supplied sample/cell/spot annotations), signature score heatmaps (for comparing results across samples, cells, or spots), and a signature correlation matrix (evaluate signature relationships). Through graphical analysis, redundancies or specificities across signatures can be identified, potentially unveiling correlations or co-occurrences of different processes. This facilitates enhanced sample, cell or spot stratification and interpretation as shown below in the three case studies reported.

Signifier helps in characterizing TCGA ovarian cancer

Most of the 296 samples from The Cancer Genome Atlas (TCGA) Ovarian Cancer (OC) collection represent the prevalent and deadly high-grade serous histotype. This form of the disease is marked by significant genomic instability, influencing gene expression and resulting in diverse phenotypes among patients. At the transcriptomic level, the TCGA consortium proposed that the transcriptional landscape of OC

can be classified into four subtypes: immunoreactive (IMR), differentiated (DIF), proliferative (PRO) and mesenchymal (MES), based on gene content and prior knowledge (18). These signatures are implemented in the consensusOV package introduced by Chen and colleagues and incorporated into signifier.

We examined the samples in the TCGA dataset using all the OC and pan-cancer signatures provided by signifier. All signatures passed quality checks (see evaluate Signature plot of Supplementary Figure S1). We computed the signature correlation matrix (Figure 2A), this process aids in narrowing down the signatures under evaluation to focus on groups of signatures. This plot usually helps in identifying areas of biological interest, redundant signatures, signatures guided by similar transcriptional regulatory programs, and reveals the co-occurrent processes. In the TCGA data four are the main groups of signatures, three of them containing the signatures of TCGA OC transcriptional subtypes. The top of Figure 2B displays the four continuous scores of consensusOV for each sample, while Figure 2B and C stratified samples based on their maximum consensusOV score. The first group encompasses signatures linked to extracellular matrix (ECM) composition and epithelial-to-mesenchymal transition (EMT), which includes the consensusOV MES score. The upregulation of genes associated with cell adhesion loss, developmental transcription factors, and extracellular matrix restructuring strongly suggests EMT, a process correlated with poor prognosis in advanced OC (19–21). This correlation is further supported by the Kaplan–Meyer curve of the EMT signature in Figure 2D, where the samples were divided in two groups by signature score levels. The second and third groups consist of signatures focusing on various aspects of the immune system in cancer. They exhibit two contrasting behaviors: the second group—containing the consensusOV IMR—comprises signatures that capture chemokine expressions and inflammatory signals, while the third group—correlating with the consensusOV PRO scores—consists of signatures associated with immune tolerance. The fourth group includes signatures related to chromosomal instability (CIN) and its associated mitotic index and cell cycle rate. These signatures reflect the widespread genomic alterations representative of the transcriptional landscape of high-grade serous OC.

In addition, signifier can unveil unexplored aspects, as demonstrated by the exploration of the pan-cancer signature of human Adult Stem Cells (ASC) (22). The ASC signature, designed to identify the most aggressive epithelial cancers, exhibits higher scores in the OC PRO subtype (Figure 2E, F, PRO versus IMR $P < 2.1e-08$, MES $P < 1.6e-07$, DIFF $P < 3.6e-05$). ASC signature is characterized by elevated expression of genes involved in chromosome reorganization and DNA methylation that are valuable for pan-cancer investigation of stem cell gene expression and for identifying potential therapeutic targets of DNA methyltransferase inhibitors, which can sensitize tumor cells to programmed cell death (23).

Utilizing the signifier workflow, we employed an automated pipeline to delineate the molecular subtypes of high-grade serous OC. This case study demonstrates that leveraging combinations of transcriptional signatures facilitates the identification of the primary biological characteristics of samples.

Signifier characterized intra-tumor heterogeneity in single-cell glioblastoma dataset

Glioblastoma stands out as one of the most prevalent brain tumors, and its lethality is closely associated with tumor recurrence. This recurrence is primarily attributed to infiltrating cells that migrate from the tumor core, thereby evading surgery and local treatment. In this particular case study, scRNA-seq samples sourced from a study conducted by Darmanis *et al.* (24) were analyzed with signifier, utilizing both brain-specific and pan-cancer signatures. Quality checks for signatures are shown in Supplementary Figure S2. Signatures exhibiting >90% zeros combined with a low percentage of expressed signature genes were excluded from downstream analyses. Figure 3A presents a t-distributed stochastic neighbor embedding (t-SNE) representation of the data, with color-coding indicating the original cell type labels provided by the authors. To ensure sizable cohorts for signature score comparisons, we filtered cells by type using the authors' original cell type annotations. We retained only those cell types present in both the tumor core and peripheral samples with a sample size >20 (see Supplementary Table S2).

Single-cell-derived signatures of glioblastoma, implemented in signifier from the publication of Barkley *et al.* (7) and Neftel *et al.* (11), were used for the first analysis (Figure 3B–D). The authors of these signatures outline the cellular programs of malignant glioblastoma cells, their plasticity, and their modulation by genetic drivers. The Darmanis dataset showed that glioblastoma contains cells in multiple states: neural progenitor-like (NPC-like), oligodendrocyte progenitor-like (OPC-like), astrocyte-like (AC-like) and mesenchymal-like (MES-like) (Figure 3B). AC-like, OPC-like, NPC-like meta-modules are linked to neurodevelopmental genes characteristic of neuronal/glial lineages or progenitor cells. When comparing these meta-modules to non-malignant neural cell types, they are most highly expressed in astrocytes, oligodendrocyte precursor cells (OPCs) and neurons, respectively. Finally, the MES-like state is associated in some tumor cells with hypoxia and increased glycolysis, resulting in hypoxia-independent (MES1) and hypoxia-dependent (MES2) signatures (Figure 3C and D).

With the attempt to provide a more detailed and comprehensive characterization of the sample, we applied the plethora of bulk-derived glioblastoma and pan-cancer signatures (Figure 3E–S). In their publication, Darmanis *et al.* (24) profiled both the tumor core and the surrounding peripheral tissue to unveil transcriptional and genetic variations between these two locations. Signifier analyses reveal the tumor core as a hypoxic environment across all studied cell types, while the surrounding tissue harbors cells with lower hypoxic scores, indicating a relatively oxygen-rich brain tissue (Figure 3E). As anticipated, these distinctions are corroborated by the authors' findings, which documented hypoxia-associated angiogenesis in the tumor core compared to the periphery. According to the previous analyses, hypoxia is highlighted in the vast majority of cells with Neftel MES2 sc-derived signature as previously described. Additionally, the cell cycle signature delineated by Davoli *et al.* illustrates that cell proliferation is predominantly confined to neoplastic cells and a small subset of immune cells within the tumor core (25) (Figure 3F).

The extracellular matrix (ECM) composition represents another facet of the TME that distinguishes the transcriptional behavior of cells within the core and the periphery. Two

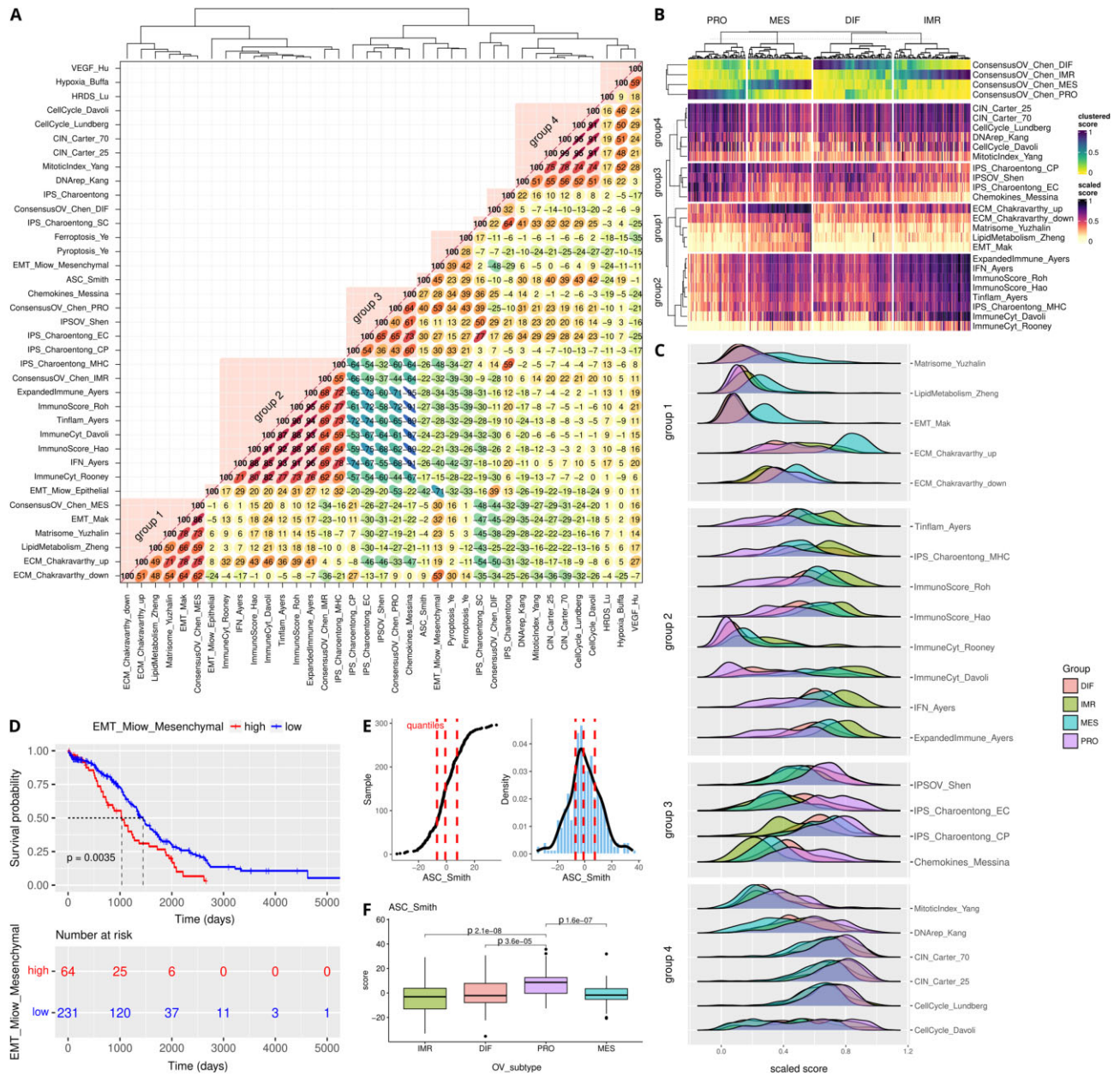


Figure 2. TCGA OC dataset dissected with signifinder. **(A)** Correlation matrix of pan-cancer and OC signatures. **(B)** Heatmap of signature scores of the four discussed groups of signatures with correlations with the TCGA expression subtypes, which are indicated on the top of the heatmap: PRO, IMR, MES and DIF (18). **(C)** Ridge plot of the score distributions of the signatures in the four discussed groups. Samples are divided into the four TCGA subgroups (18). **(D)** Kaplan–Meyer plot and survival association for the EMT signature scores by Miow *et al.* (49). **(E)** Score distribution of the ASC signature (22). **(F)** Boxplot illustrating the ASC signature in the four TCGA transcriptional subgroups.

pan-cancer ECM signatures proposed by Chakravarty *et al.* (26) (Figure 3G, H) demonstrate gradients with opposing directions between core and peripheral regions. In core cells, neoplastic and immune cells exhibit higher ECM-Up scores, whereas ECM-Down scores are elevated in peripheral cells. The ECM-Up program is linked with a TGF- β -rich TME, immune evasion, and failure of immunotherapy, whereas the ECM-Down signature represents a more normal-like ECM environment.

Focusing specifically on the neoplastic cells (Figure 3I–L and Supplementary Figure S3A), signifinder highlights their extremely heterogeneous expression profiles. Despite the significant differences between core and periphery scores, it is

clear that hypoxic conditions are found only in a subgroup of neoplastic core cells (cells with high scores in Figure 3J Hypoxia_Buffa (27), and Figure 3K VEGF_Hu signature (28)). On top of this, these hypoxic core cells also show low scores of cell cycle rate (Figure 3L, CellCycle_Davoli (25)) possibly due to the induction of cell cycle arrest in the presence of prolonged hypoxia, in order to turn off highly energy consuming processes and promote cell survival (29,30). These hypoxic cells with low proliferation rate are spatially confined in the upper-left part of the t-SNE, which means that this condition deeply impacts the entire transcriptome of these cells. Cell proliferation is confined to a small subset of non-hypoxic cells exclusive to the tumor core, unlike

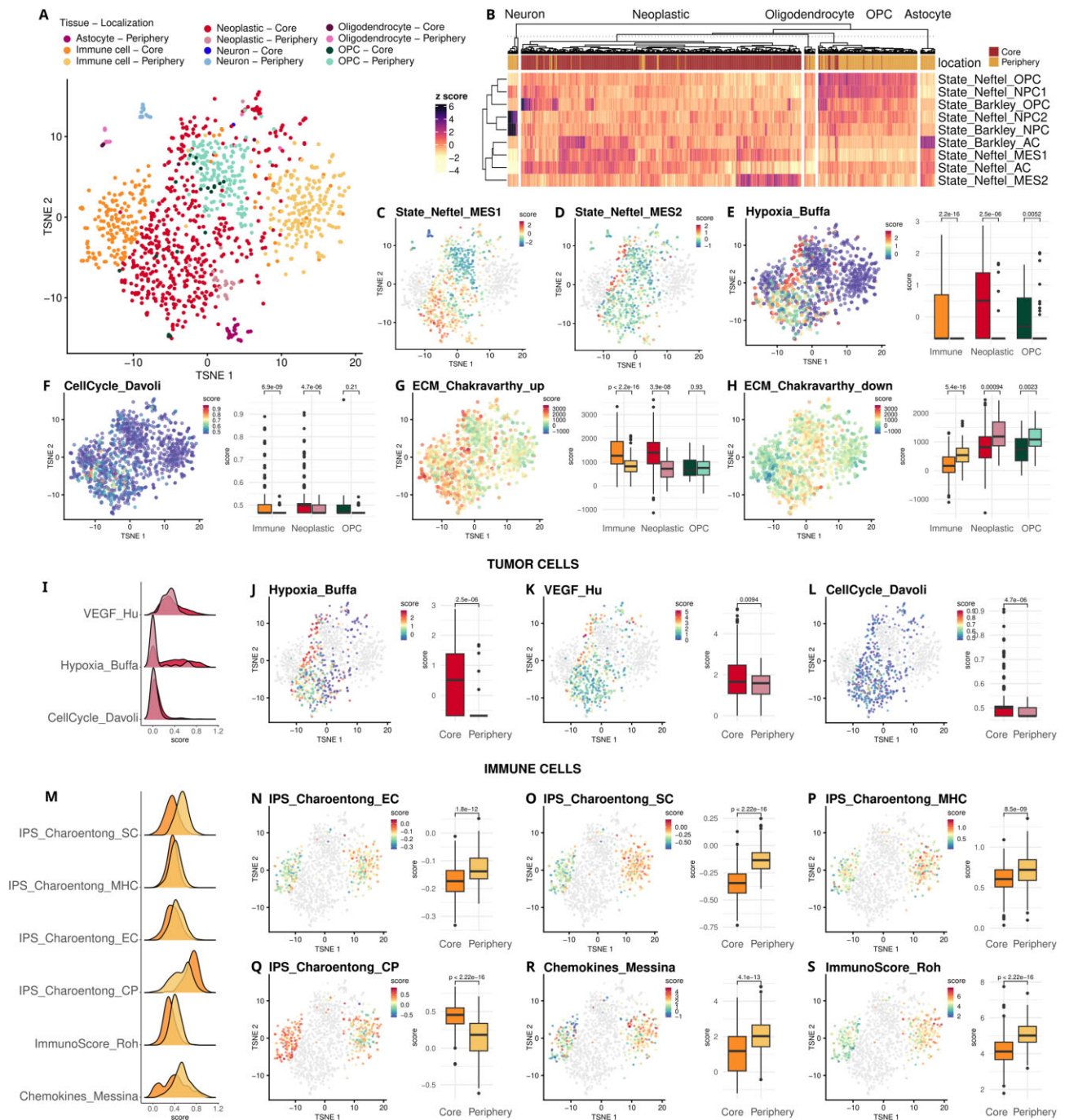


Figure 3. scRNA-seq data of Darmanis *et al.* (24) studied with signifier. (A) t-SNE on the top 50 principal components of the expression data. Colors represent the original cell type annotations as provided by the authors additionally divided by their spatial location, i.e. tumor core or tumor periphery. Color legend is maintained throughout all the panels. t-SNE with cells colored by signature are reported in different panels together with the boxplot of signature score distributions in the different cell types and locations. (B) Heatmap of the glioblastoma single-cell derived signatures, (C) MES1 signature by Neftel *et al.* (11), (D) MES1 signature by Neftel *et al.* (11), (E) hypoxia signature scores developed by Buffa *et al.* (27) and (F) cell cycle signature scores developed by Davoli *et al.* (25). (G) ECM-Up and (H) ECM-Down signature scores developed by Chakravarthy *et al.* (26). Panels I to L are dedicated to signatures in tumor cells. (I) Multiple signature score distributions differentiating core and peripheral tumor cells. (J) Hypoxia signature scores developed by Buffa *et al.* (27). (K) VEGF signature scores developed by Hu *et al.* (28). (L) Cell cycle signature scores developed by Davoli *et al.* (25). Panels from M to S are dedicated to signatures in immune cells. (M) Distribution of multiple signature scores in core and peripheral immune cells. The IPS is composed of four averaged and weighted Z scores: (N) the effector cell (EC) score, (O) the immune-suppressive cells (SC) score, (P) the MHC score and (Q) the immune checkpoints (CP) score. (R) The Chemokines score proposed by Messina *et al.* (33) and (S) the immune score defined by Roh *et al.* (34).

peripheral tumor cells which demonstrate low proliferative potential while disseminating (25) (Figure 3L, CellCycle_Davoli). The highly proliferative neoplastic core cells are also characterized by a high chromosomal instability score, as defined by the pan-cancer signature of Carter and colleagues (31) (CIN_Carter_70, Supplementary Figure S4). This is because genes involved in DNA replication, DNA repair, spindle assembly, and chromosome segregation include cell cycle control genes. In many tumors including gliomas, higher CIN scores are associated with unfavorable clinical outcomes and metastatic specimens. If utilized at the single-cell level, this evidence could prove invaluable in pinpointing the most aggressive cells driving tumor progression, warranting further in-depth investigation. Immune cells in the core and periphery are clearly separated on the t-SNE plot, showing markedly different transcriptional programs, which is also reflected by the immune signatures that report different scores in the two groups (Figure 3M and Supplementary Figure S3B). The immunophenoscore (IPS) is composed of four averaged and weighted Z scores—the EC score that covers expression of effector cells (activated CD4⁺ T cells, activated CD8⁺ T cells, effector memory CD4⁺ T cells, and effector memory CD8⁺ T cells), the SC score that collects the expression related to immune-suppressive cells (T-regs and Myeloid-derived suppressor cells), the Major Histocompatibility Complex (MHC) score for MHC related molecules, and the immune checkpoint score to represent the activity of immune checkpoints or immunomodulators (32). In the glioblastoma sample, all four scores showed significant differences between the core and periphery. The core seems to indicate an immunologically cold tumor, where expression of the major determinants of tumor immunogenicity is turned off (Figure 3M–Q). Another score that behaves differently between the core and periphery is the Chemokines score proposed by Messina and colleagues (33) (Figure 3R). The score intends to predict lymphoid cell infiltrates in solid tumor masses through the expression of 12 cytokines. From the data, it seems that the presence of lymphoid cells is not homogeneous intra-mass and that most of those cytokines are expressed only by immune cells in the periphery. The pan-cancer immune score defined by Roh and colleagues is dedicated to predict the immune checkpoint blockade treatment response. Roh *et al.* demonstrated that this immune score correlates positively with T-cell receptor clonality in pre-PD-1 blockade samples and that higher scores and T-cell receptor clonality characterize responders (34). The Roh immune scores were found higher in peripheral cells than in core cells and thus could predict a different response of the two cell groups to PD-1 blockade treatment (Figure 3S).

Signifinder highlights spatial-specific patterns of expression signatures: a case study on invasive ductal breast carcinoma

The spatial transcriptomic dataset presented here is a 10x Visium sample of breast invasive ductal carcinoma. Ductal carcinoma is the most prevalent type of breast cancer (BC), constituting nearly 80% of all breast cancer diagnosis. In Figure 4, the anatomopathologist's interpretation of the hematoxylin and eosin staining of the formalin-fixed paraffin embedded (FFPE) sample is depicted. Multiple neoplastic areas, highlighted in red, are localized within the ducts—a common occurrence in tumors originating from the epithelial cells lining the ducts. Initially, these tumors invade the inner part of

the ducts (carcinoma *in situ*), often leading to the formation of necrotic areas within the duct lumen. As the neoplastic cells breach the duct wall, they infiltrate the stroma, resulting in invasive carcinoma. Tumor masses are encompassed by fibrous tissue, outlined in blue in Figure 4A, predominantly composed of fibroblasts and lymphocytes. Such areas are frequently observed because cancer-associated fibroblasts (CAFs) contribute to tumor proliferation by secreting various growth factors, cytokines, chemokines, and proteins involved in ECM degradation (35). Fibrous tissue and stroma surrounding the tumor also contain areas with high densities of infiltrated lymphocytes, as indicated in Figure 4A with the light blue lines. The remainder of the section includes adipocytes and blood vessels, indicated in green and orange, respectively. Spots are then manually classified by cell type following the anatomopathological reading of the high-resolution image (Figure 4B).

Due to the absence of spatial transcriptomic-derived signatures, the multiple bulk-derived pan-cancer and BC signatures present in signifinder were applied. The signature scores were computed for each spot, and results were compared with tissue annotations. As expected, the signatures do exhibit a relatively high percentage of zero counts as well as mild but noticeable correlations with the total count number, due to dependencies on the number of cells lying on each spot (Supplementary Figure S5). The hypoxia signature by Buffa *et al.* shows that tumor areas are highly hypoxic compared to the normal stroma (27) (Figure 4C and D), and that all cell types show high variance of hypoxia scores, with the highest scores in certain necrotic spots. Tumor spots show high cell-cycle rates, as determined by the Lundberg *et al.* signature that lights up the tumor and the nearby areas (Figure 4C and E). The lowest scores remain confined to the non-reactive stroma and to necrotic areas. Thus, we can appreciate here that tumor cells show heterogenic proliferation rates.

The EMT signature by Cheng *et al.* (36), originally provided as a prognostic biomarker associated with late-disease recurrence in BC, is presented in Figure 4C, G, H and I. The signature shows that tumor spots are characterized by the co-existence of both epithelial and mesenchymal markers with a particular spatial distribution: high scores are localized to the leading edge of tumors in close apposition with the surrounding CAFs on the stroma. Starting from the duct basement membrane, where the tumor arises, the score describes a decreasing pattern when moving to the inner part of the duct (Figure 4H). Figure 4I shows the amount of tumor spots that are surrounded by spots of a given annotation (each spot can have from 0 to 6 neighboring spots). Each row contains the total number of tumor spots. The y-axis states the annotation type of the neighbor spots and the x-axis represents their number. For example, there are around 400 tumor spots that are surrounded by 0 CAFs spots, while around 200 tumor spots are surrounded by 6 tumor spots. The dots are colored by the median EMT_Cheng score of that specific group of tumor spots. The median score increases with the number of CAFs spots surrounding a tumor spot. On the contrary, it decreases with the number of tumor spots around each tumor spot. Interestingly, since high EMT scores are associated with high risk of late recurrence, the spatial distribution of this score suggests that the hyperproliferative cells found in the basement membrane (i.e. the origin cells of these tumors) are also the most dangerous cells for relapse. Also, the highest proliferative tumor areas have high levels of CIN

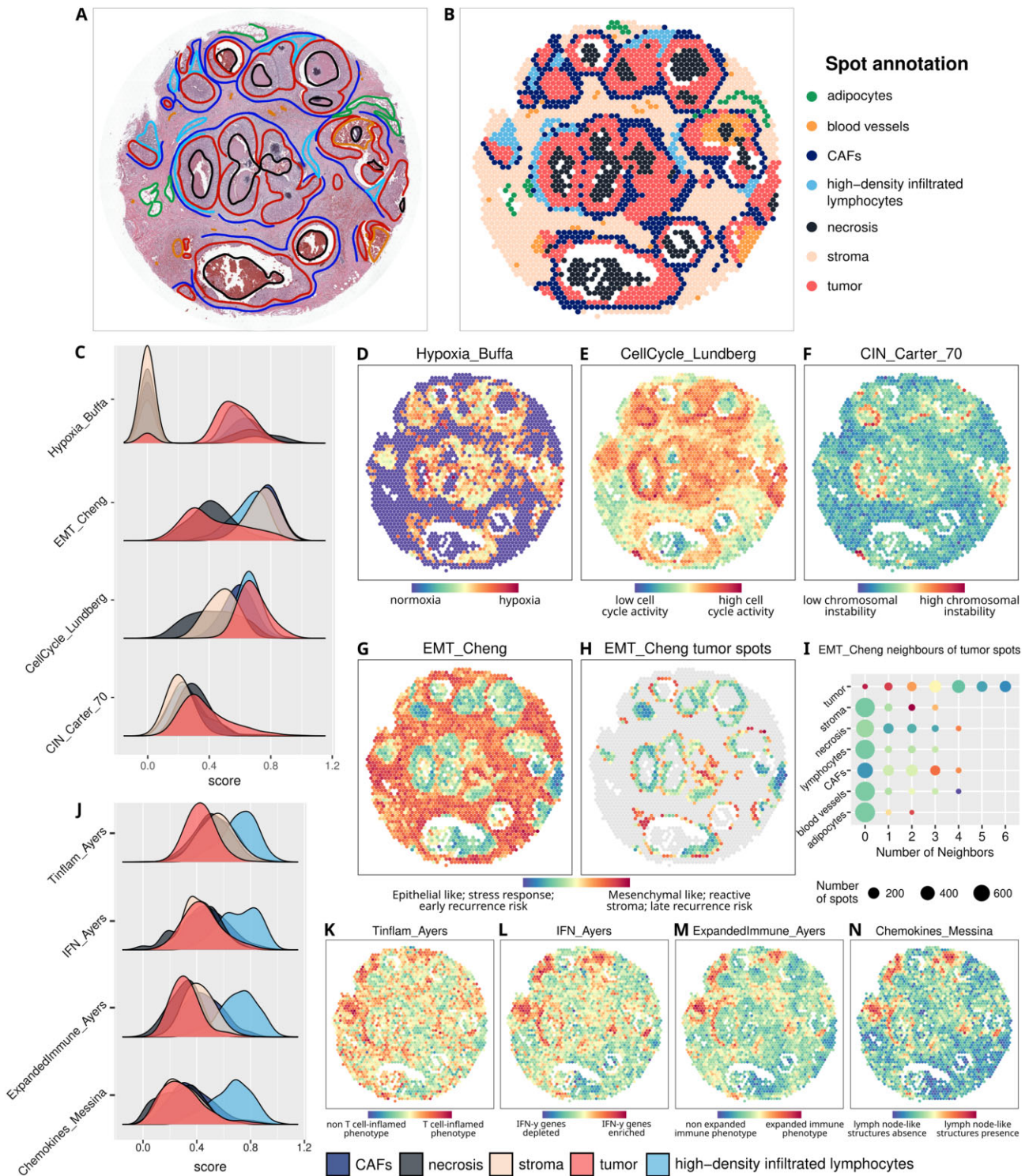


Figure 4. Spatial transcriptomic sample of breast invasive ductal carcinoma studied with signfinder. **(A)** Histologic image with manual anatomopathological annotations. **(B)** Dataset spots are annotated following manual annotations. **(C)** Distribution of multiple signature scores divided into cancer associated fibroblasts (CAFs), necrosis, stroma, tumor, and high-density infiltrated lymphocyte areas. Spatial score distribution of the **(D)** hypoxia signature by Buffa *et al.* (27), **(E)** cell cycle signature by Lundberg *et al.* (50), **(F)** CIN70 by Carter *et al.* (31), **(G)** EMT by Cheng *et al.* (36) and **(H)** EMT scores in tumor spots only. **(I)** Median EMT_Cheng scores of tumor spots grouped by the number and annotation of their neighboring spots. **(J)** Distribution of multiple signature scores divided into CAFs, necrosis, stroma, tumor and high-density infiltrated lymphocyte areas. Spatial score distribution of **(K)** Tinflam, **(L)** IFN and **(M)** Expanded Immune by Ayers *et al.* (48), and **(N)** Chemokines by Messina *et al.* (33).

(Figure 4C, F) as shown by the CIN signature (proposed by Carter *et al.*, which is based on expression aberrations in genes localized to each chromosomal region (31). In ductal breast carcinoma, the presence of a high rate of chromosomal copy number variations is associated with an increased immune response, as well as the presence of tumor-infiltrating lymphocytes and PD-L1 gene expression (37–40). We can confirm that high-CIN tumor areas are in proximity with those reported to have high lymphocyte density (Supplementary Figure S6). In recent years, it has been generally accepted that these immune-related areas play a role as prognostic and predictive markers in invasive BC (41–44). However, some controversial evidence suggests that their predictive value depends on the exact composition of the infiltrate (45–47), which is almost impossible to assess based solely on hematoxylin and eosin staining. Signatures can thus help in evaluating the immune activities of these zones. Multiple immune system signatures characterize the areas with the highest lymphocyte density: Tinflam, IFN (interferons), and Expanded Immune by Ayers *et al.* (48), as well as the Chemokines signature proposed by Messina *et al.* (33). These signatures show the highest values in proximity to areas with high densities of lymphocytes (Figure 4J–N, and Supplementary Figures S7–S9). The signatures from Ayers and colleagues were developed to identify tumors with a T cell-inflamed microenvironment, characterized by active IFN- γ signaling, cytotoxic effector molecules, antigen presentation, and T cell active cytokines, which are common features of tumors that are responsive to PD-1 checkpoint blockade. Similarly, the Chemokines signature, which was independently developed with similar purposes to identify tumor-localized ectopic lymph node-like structures, shows high scores in the same areas (33).

Discussion

Introducing a free-access computational implementation alongside the public cancer signatures would mark a significant stride toward ensuring signature reproducibility and usability. However, to date, this received limited attention. With the development of signifinder, we achieved three primary outcomes: (i) establishing the infrastructure for gathering and integrating transcriptomic signatures (including bulk, single-cell, and spatial transcriptomic derived signatures); (ii) compiling an initial compendium of cancer gene signature implementations by systematically screening papers from the literature and (iii) enabling the use of single-cell and spatial transcriptomic datasets as inputs, we furnishing a tool capable of probing the behavior of gene expression signatures within tumors, assessing their intra tumor heterogeneity, and allowing the automatic and fast detection of cell states.

Signifinder is developed in R and utilizes Bioconductor's expression data structures. Thus, it is compatible with the majority of widely used tools and pipelines for transcriptome data analysis. Additionally, signifinder incorporates supplementary functions aimed at facilitating visualization and interpretation of results, thereby simplifying the comprehension of the diverse roles that multiple hallmarks within and across patients, cells, and tissues may assume. Through its graphical tools, signifinder facilitates seamless comparison across multiple signatures, shedding light on underlying processes, interactions and collaborations.

The future trajectory of the signifinder package entails the continuous addition and integration of new cancer signatures.

However, achieving this objective is envisioned as a collaborative endeavor within the research community. Indeed, signifinder was purposefully designed from the outset to embrace contributions from the community for implementing signatures.

Signifinder package can enhance the interpretability of high-resolution cancer transcriptomic data, allowing the detection of intratumor variability and, finally, helping in solving tumor complexity.

Data availability

The signifinder package is available in the Bioconductor platform at <https://www.bioconductor.org/packages/release/bioc/html/signifinder.html>. The analysis notebooks to reproduce the aforementioned analysis are hosted at https://github.com/CaluraLab/signifinder_workflow and <https://doi.org/10.5281/zenodo.13827941>. The OC bulk data set is available in the Genomic Data Commons at <https://portal.gdc.cancer.gov/>, and can be accessed with ID TCGA-OC. The glioblastoma single-cell data set is available in the Gene Expression Omnibus (GEO) database at <https://www.ncbi.nlm.nih.gov/geo> (accession GSE84465). The BC spatial transcriptomics data set is available at the 10x Genomics website at <https://www.10xgenomics.com/resources/datasets/human-breast-cancer-ductalcarcinoma-in-situ-invasive-carcinoma-ffpe-1-standard-1-3-0>.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

Author contributions: S.P. developed the signifinder package and coordinated contributions to it. L.M., F.P., A.B. and M.A. contributed to signature collection and implementation, package documentation, and testing. G.E. provided histopathological image analysis for the 10x spatial transcriptomic data. S.P., A.C. and E.C. performed data analyses and discussed the case studies. P.M. and D.R. assisted in package development strategy. S.P., P.M., D.R. and C.R. contributed to the conception of signature evaluation strategies. E.C. and S.P. drafted the manuscript, and all authors reviewed and edited it. E.C. conceived and supervised the project. All authors read and approved the final manuscript.

Funding

Italian Association for Cancer Research [MFAG23522 to E.C., IG21837 and IG29071 to C.R.]; National Cancer Institute of the National Institutes of Health [U24CA180996 to D.R.]; Chan Zuckerberg Initiative DAF; an advised fund of Silicon Valley Community Foundation [CZF2019-002443 to D.R.]. Funding for open access charge: Italian Association for Cancer Research; MUR-PNRR NextGenerationEU and MUR, Mission 4 Component C2 part 1.4—National Center for Gene Therapy and Drugs based on RNA Technology [CN00000041 - CUP C93C22002780006 to E.C.]; Italian national project 'PRIN PNRR 2022' funded by the Italian Ministry of Education, University and Research [P20223Y5AX Project to E.C.]; Italian national project 'PRIN 2022' funded

by the Italian Ministry of Education, University and Research [20227Z2XRB Project to E.C.].

Conflict of interest statement

None declared.

References

- Nevins, J.R. and Potti, A. (2007) Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.*, **8**, 601–609.
- Gingras, I., Desmedt, C., Ignatiadis, M. and Sotiriou, C. (2015) CCR 20th anniversary commentary: gene-expression signature in breast cancer – where did it start and where are we now? *Clin. Cancer Res.*, **21**, 4743–4746.
- Griguolo, G., Bottosso, M., Vernaci, G., Miglietta, F., Dieci, M.V. and Guarneri, V. (2022) Gene-expression signatures to inform neoadjuvant treatment decision in HR+/HER2- breast cancer: available evidence and clinical implications. *Cancer Treat. Rev.*, **102**, 102323.
- Pitt, J.M., Vétizou, M., Daillère, R., Roberti, M.P., Yamazaki, T., Routy, B., Lepage, P., Boneca, I.G., Chamailard, M., Kroemer, G., et al. (2016) Resistance mechanisms to immune-checkpoint blockade in cancer: tumor-intrinsic and -extrinsic factors. *Immunity*, **44**, 1255–1269.
- Kulkarni, A., Anderson, A.G., Merullo, D.P. and Konopka, G. (2019) Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Curr. Opin. Biotechnol.*, **58**, 129–136.
- Poirion, O.B., Zhu, X., Ching, T. and Garmire, L. (2016) Single-cell transcriptomics bioinformatics and computational challenges. *Front. Genet.*, **7**, 163.
- Barkley, D., Moncada, R., Pour, M., Liberman, D.A., Dryg, I., Werba, G., Wang, W., Baron, M., Rao, A., Xia, B., et al. (2022) Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.*, **54**, 1192–1201.
- Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S., et al. (2017) Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, **171**, 1611–1624.
- Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.
- Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.
- Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019) An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, **178**, 835–849.
- Baron, M., Tagore, M., Hunter, M.V., Kim, I.S., Moncada, R., Yan, Y., Campbell, N.R., White, R.M. and Yanai, I. (2020) The stress-like cancer cell state is a consistent component of tumorigenesis. *Cell Syst.*, **11**, 536–546.
- Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.*, **21**, 36.
- Subramanian, J. and Simon, R. (2010) What should physicians look for in evaluating prognostic gene-expression signatures? *Nat. Rev. Clin. Oncol.*, **7**, 327–334.
- Lilljebjörn, H., Orsmark-Pietras, C., Mitelman, F., Hagström-Andersson, A. and Fioretos, T. (2022) Transcriptomics paving the way for improved diagnostics and precision medicine of acute leukemia. *Semin. Cancer Biol.*, **84**, 40–49.
- DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.-D., Williams, C., Reich, M., Winckler, W. and Getz, G. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinform. Oxf. Engl.*, **28**, 1530–1532.
- Saiselet, M., Rodrigues-Vitória, J., Tourneur, A., Craciun, L., Spinette, A., Larsimont, D., Andry, G., Lundeborg, J., Maenhaut, C. and Detours, V. (2020) Transcriptional output, cell-type densities, and normalization in spatial transcriptomics. *J. Mol. Cell Biol.*, **12**, 906–908.
- Chen, G.M., Kannan, L., Geistlinger, L., Kofia, V., Safikhani, Z., Gendoo, D.M.A., Parmigiani, G., Birrer, M., Haibe-Kains, B. and Waldron, L. (2018) Consensus on molecular subtypes of high-grade serous ovarian carcinoma. *Clin. Cancer Res.*, **24**, 5037–5047.
- Sun, Y., Hu, L., Zheng, H., Bagnoli, M., Guo, Y., Rupaimoole, R., Rodriguez-Aguayo, C., Lopez-Berestein, G., Ji, P., Chen, K., et al. (2015) MiR-506 inhibits multiple targets in the epithelial-to-mesenchymal transition network and is associated with good prognosis in epithelial ovarian cancer. *J. Pathol.*, **235**, 25–36.
- Koutsaki, M., Spandidos, D.A. and Zaravinos, A. (2014) Epithelial-mesenchymal transition-associated miRNAs in ovarian carcinoma, with highlight on the miR-200 family: prognostic value and prospective role in ovarian cancer therapeutics. *Cancer Lett.*, **351**, 173–181.
- Leung, D., Price, Z.K., Lokman, N.A., Wang, W., Goonetilleke, L., Kadife, E., Oehler, M.K., Ricciardelli, C., Kannourakis, G. and Ahmed, N. (2022) Platinum-resistance in epithelial ovarian cancer: an interplay of epithelial-mesenchymal transition interlinked with reprogrammed metabolism. *J. Transl. Med.*, **20**, 556.
- Smith, B.A., Balanis, N.G., Nanjundiah, A., Sheu, K.M., Tsai, B.L., Zhang, Q., Park, J.W., Thompson, M., Huang, J., Witte, O.N., et al. (2018) A human adult stem cell signature marks aggressive variants across epithelial cancers. *Cell Rep.*, **24**, 3353–3366.
- Sabari, J.K., Lok, B.H., Laird, J.H., Poirier, J.T. and Rudin, C.M. (2017) Unravelling the biology of SCLC: implications for therapy. *Nat. Rev. Clin. Oncol.*, **14**, 549–561.
- Darmanis, S., Sloan, S.A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., et al. (2017) Single-cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.*, **21**, 1399–1410.
- Davoli, T., Uno, H., Wooten, E.C. and Elledge, S.J. (2017) Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, **355**, eaaf8399.
- Chakravarthy, A., Khan, L., Bensler, N.P., Bose, P. and De Carvalho, D.D. (2018) TGF- β -associated extracellular matrix genes link cancer-associated fibroblasts to immune evasion and immunotherapy failure. *Nat. Commun.*, **9**, 4692.
- Buffa, F.M., Harris, A.L., West, C.M. and Miller, C.J. (2010) Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer*, **102**, 428–435.
- Hu, Z., Fan, C., Livasy, C., He, X., Oh, D.S., Ewend, M.G., Carey, L.A., Subramanian, S., West, R., Ikpatt, F., et al. (2009) A compact VEGF signature associated with distant metastases and poor outcomes. *BMC Med.*, **7**, 9.
- Bi, M., Naczki, C., Koritzinsky, M., Fels, D., Blais, J., Hu, N., Harding, H., Novoa, I., Varia, M., Raleigh, J., et al. (2005) ER stress-regulated translation increases tolerance to extreme hypoxia and promotes tumor growth. *EMBO J.*, **24**, 3470–3481.
- Liu, L., Cash, T.P., Jones, R.G., Keith, B., Thompson, C.B. and Simon, M.C. (2006) Hypoxia-induced energy stress regulates mRNA translation and cell growth. *Mol. Cell*, **21**, 521–531.
- Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N. and Szallasi, Z. (2006) A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.*, **38**, 1043–1048.

32. Charoentong,P., Finotello,F., Angelova,M., Mayer,C., Efremova,M., Rieder,D., Hackl,H. and Trajanoski,Z. (2017) Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.*, **18**, 248–262.
33. Messina,J.L., Fenstermacher,D.A., Eschrich,S., Qu,X., Berglund,A.E., Lloyd,M.C., Schell,M.J., Sondak,V.K., Weber,J.S. and Mulé,J.J. (2012) 12-Chemokine gene signature identifies lymph node-like structures in melanoma: potential for patient selection for immunotherapy? *Sci. Rep.*, **2**, 765.
34. Roh,W., Chen,P.-L., Reuben,A., Spencer,C.N., Prieto,P.A., Miller,J.P., Gopalakrishnan,V., Wang,F., Cooper,Z.A., Reddy,S.M., *et al.* (2017) Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl. Med.*, **9**, eaah3560.
35. Lavie,D., Ben-Shmuel,A., Erez,N. and Scherz-Shouval,R. (2022) Cancer-associated fibroblasts in the single-cell era. *Nat. Cancer*, **3**, 793–807.
36. Cheng,Q., Chang,J.T., Gwin,W.R., Zhu,J., Ambs,S., Geradts,J. and Lyerly,H.K. (2014) A signature of epithelial-mesenchymal plasticity and stromal activation in primary tumor modulates late recurrence in breast cancer independent of disease subtype. *Breast Cancer Res. BCR*, **16**, 407.
37. Vincent-Salomon,A., Lucchesi,C., Gruel,N., Raynal,V., Pierron,G., Goudefroye,R., Reyat,F., Radvanyi,F., Salmon,R., Thiery,J.-P., *et al.* (2008) Integrated genomic and transcriptomic analysis of ductal carcinoma in situ of the breast. *Clin. Cancer Res.*, **14**, 1956–1965.
38. Abba,M.C., Gong,T., Lu,Y., Lee,J., Zhong,Y., Lacunza,E., Butti,M., Takata,Y., Gaddis,S., Shen,J., *et al.* (2015) A molecular portrait of high-grade ductal carcinoma in situ. *Cancer Res.*, **75**, 3980–3990.
39. Hendry,S., Pang,J.-M.B., Byrne,D.J., Lakhani,S.R., Cummings,M.C., Campbell,I.G., Mann,G.B., Goringe,K.L. and Fox,S.B. (2017) Relationship of the breast ductal carcinoma in situ immune microenvironment with clinicopathological and genetic features. *Clin. Cancer Res.*, **23**, 5210–5217.
40. Agahozo,M.C., Hammerl,D., Debets,R., Kok,M. and van Deurzen,C.H.M. (2018) Tumor-infiltrating lymphocytes and ductal carcinoma in situ of the breast: friends or foes? *Mod. Pathol.*, **31**, 1012–1025.
41. Stanton,S.E. and Disis,M.L. (2016) Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J. Immunother. Cancer*, **4**, 59.
42. de la Cruz-Merino,L., Barco-Sánchez,A., Henao Carrasco,F., Nogales Fernández,E., Vallejo Benítez,A., Brugal Molina,J., Martínez Peinado,A., Grueso López,A., Ruiz Borrego,M., Codes Manuel de Villena,M., *et al.* (2013) New insights into the role of the immune microenvironment in breast carcinoma. *Clin. Dev. Immunol.*, **2013**, 785317.
43. Allen,M.D. and Jones,L.J. (2015) The role of inflammation in progression of breast cancer: friend or foe? (Review). *Int. J. Oncol.*, **47**, 797–805.
44. Mao,Y., Qu,Q., Chen,X., Huang,O., Wu,J. and Shen,K. (2016) The prognostic value of tumor-infiltrating lymphocytes in breast cancer: a systematic review and meta-analysis. *PLoS One*, **11**, e0152500.
45. Seo,A.N., Lee,H.J., Kim,E.J., Kim,H.J., Jang,M.H., Lee,H.E., Kim,Y.J., Kim,J.H. and Park,S.Y. (2013) Tumour-infiltrating CD8+ lymphocytes as an independent predictive factor for pathological complete response to primary systemic therapy in breast cancer. *Br. J. Cancer*, **109**, 2705–2713.
46. Lee,H.J., Seo,J.-Y., Ahn,J.-H., Ahn,S.-H. and Gong,G. (2013) Tumor-associated lymphocytes predict response to neoadjuvant chemotherapy in breast cancer patients. *J. Breast Cancer*, **16**, 32–39.
47. Oda,N., Shimazu,K., Naoi,Y., Morimoto,K., Shimomura,A., Shimoda,M., Kagara,N., Maruyama,N., Kim,S.J. and Noguchi,S. (2012) Intratumoral regulatory T cells as an independent predictive factor for pathological complete response to neoadjuvant paclitaxel followed by 5-FU/epirubicin/cyclophosphamide in breast cancer patients. *Breast Cancer Res. Treat.*, **136**, 107–116.
48. Ayers,M., Lunceford,J., Nebozhyn,M., Murphy,E., Loboda,A., Kaufman,D.R., Albright,A., Cheng,J.D., Kang,S.P., Shankaran,V., *et al.* (2017) IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.*, **127**, 2930–2940.
49. Miow,Q.H., Tan,T.Z., Ye,J., Lau,J.A., Yokomizo,T., Thiery,J.-P. and Mori,S. (2015) Epithelial-mesenchymal status renders differential responses to cisplatin in ovarian cancer. *Oncogene*, **34**, 1899–1907.
50. Lundberg,A., Lindström,L.S., Harrell,J.C., Falato,C., Carlson,J.W., Wright,P.K., Foukakis,T., Perou,C.M., Czene,K., Bergh,J., *et al.* (2017) Gene expression signatures and immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts. *Clin. Cancer Res.*, **23**, 7512–7520.