NEOPLASIA
www.neoplasia.com
Volume 23 Number xxx Month 2021 pp. 1069–1077

Original Research

RosettaSX: Reliable gene expression signature scoring of cancer models and patients

Julian Kreis[a,b]; Boro Nedić[a]; Johanna Mazur[a]; Miriam Urban[a]; Sven-Eric Schelhorn[a]; Thomas Grombacher[a]; Felix Geist[c]; Benedikt Brors[d,e]; Michael Zühlsdorf[c]; Eike Staub[a,*]

[a] Department of Translational Medicine, Oncology Bioinformatics, Merck KGaA, Darmstadt, Germany
[b] Faculty of Bioscience, University of Heidelberg, Heidelberg, Germany
[c] Therapeutic Innovation Platform Oncology & Immuno-Oncology, Merck KGaA, Darmstadt, Germany
[d] Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[e] German Cancer Consortium (DKTK), Core Center, Heidelberg, Germany

Abstract

Gene expression signatures have proven their potential to characterize important cancer phenomena like oncogenic signaling pathway activities, cellular origins of tumors, or immune cell infiltration into tumor tissues. Large collections of expression signatures provide the basis for their application to data sets, but the applicability of each signature in a new experimental context must be reassessed. We apply a methodology that utilizes the previously developed concept of coherent expression of genes in signatures to identify translatable signatures before scoring their activity in single tumors. We present a web interface (www.rosettasx.com) that applies our methodology to expression data from the Cancer Cell Line Encyclopaedia and The Cancer Genome Atlas. Configurable heat maps visualize per-cancer signature scores for 293 hand-curated literature-derived gene sets representing a wide range of cancer-relevant transcriptional modules and phenomena. The platform allows users to complement heatmaps of signature scores with molecular information on SNVs, CNVs, gene expression, gene dependency, and protein abundance or to analyze own signatures. Clustered heatmaps and further plots to drill-down results support users in studying oncological processes in cancer subtypes, thereby providing a rich resource to explore how mechanisms of cancer interact with each other as demonstrated by exemplary analyses of 2 cancer types.

Neoplasia (2021) 23, 1069–1077

Keywords: Gene expression signature, Cancer expression profiling, Cancer subtypes, Multiomics Analyses, Analyses, Web service

Abbreviations

ABC	Activated B-cell like
CCLE	Cancer Cell Line Encyclopaedia
CPM	Counts per million
CS	Coherence score
DI	Deregulation Index
EMT	Epithelial-mesenchymal transition
ER	Estrogen receptor
GCB	Germinal centre B-cell-like
PR	Progestosteron receptor
TCGA	The Cancer Genome Atlas
TMM	Trimmed mean of M-values
TPM	Transcripts per million

* Corresponding author.
E-mail address: eike.staub@merckgroup.com (E. Staub).
Received 11 June 2021; received in revised form 28 August 2021; accepted 30 August 2021

https://doi.org/10.1016/j.neo.2021.08.005

Introduction

The tumor-specific gene expression phenotype is the product of a coordinated interplay of cellular and pathway-related transcription modules [1]. These modules of coexpressed genes, also called expression signatures,

describe many functional aspects of a tumor, like cell-type composition [2], signaling pathway activities [3], activities of fundamental cellular processes like proliferation or interferon response [4], or can even inform on the cell-of-origin of a tumor. Cancer subtype signatures have been identified, for example, for tumors of the colon [5], breast [6,7], or diffuse large B-cell carcinoma [8]. Further, during tumorigenesis, the accumulation of somatic alterations in tumors often results in specific expression phenotypes, enabling cancers to develop cancer hallmark properties [9]. Therefore, the investigation of expression signatures today plays an essential role in the elucidation of tumor biology.

Expression signatures have also become widely used during drug development, for support of clinical cancer diagnosis [10]. The exploration of signatures has become a standard component for cancer subtyping in major multiomics patient cohort studies and often provides a framework for the interpretation of somatic mutations [11]. Hundreds of signatures have been proposed for concrete clinical use in the last 2 decades, either to inform on patient prognosis [12] or to predict therapeutic efficacy based on cancer gene expression at baseline of therapy [13]. Furthermore, the connectivity map approach, which, for the first time, proposed the reversal of disease signatures by therapeutic drugs, has developed into a widely used method in drug discovery [14]. It was extended later to include gene knock-down/out-induced perturbations. Thereby, it enables to generate a broader understanding of the mode of actions of drugs and gene perturbations in relation to a user-defined gene expression signature.

Given the importance of expression signatures for understanding tumor biology and drug response, it is essential to reliably identify sets of high-quality expression signatures that can be used to score the activity of functional expression modules in a new expression data set, independent of whether it is an interventional experimental study, a study across model organisms or a patient cohort study. To this end, a careful selection of published high-quality gene expression signatures is crucial. It should cover a broad range of underlying phenomena to characterize the activities of all cancer hallmark mechanisms comprehensively [9]. Throughout the last decades, different collections of gene modules were developed. The MSigDB collection is one of the largest resources for published expression signatures. Despite its comprehensive coverage of literature-published expression signatures, it is not necessarily best suited to efficiently characterize new cancer expression data for 2 reasons. First, the collection comprises many noisy signatures that lack validation in independent datasets or are strongly influenced by confounding mechanisms. For example, many signatures are "polluted" by high fractions of cell-cycle-controlled genes, which render their net scores like typical proliferation signatures [15], a problem especially in oncology with a range of mechanisms that affect cell proliferation. As a consequence of this problem, most random gene sets in breast cancer lead to prognostic signatures because proliferation as a mechanism is prognostically relevant [16]. Further, redundancy and bias create difficulties for efficient control of type I errors in statistical testing, thereby reducing the power of analyses.

Another collection of signatures to characterize cancer expression data is the cancer hallmark signatures published by the MSigDB team that cover 50 carefully selected gene sets [36]. Although they probably cover the majority of cancer-relevant expression phenotypes, the collection does not comprise many well-known and frequently used signatures for the diagnosis of distinct cancer subtypes. So, when using this limited collection only, the result has shortcomings when trying to bridge own results with findings in the literature. Also, it is advantageous for a collection to have at least a few signatures for a single phenomenon, i.e., to cover phenomena with small, controlled redundancy (as long as gene lists do not overlap significantly). A limited redundancy can add credibility to results observed for a specific phenomenon because phenomena-specific signatures originally postulated by independent research groups can more convincingly support a particular result. In conclusion, there is room for a curated signature collection that tries to comprehensively capture the most prevalent signatures used in different cancer communities while avoiding too much redundancy and bias that would blow up the multiple testing problem.

A gene expression module might not always be relevant in a new data set not used for its discovery. Rationales for the assessment of relevance or translatability of a signature into a new data context have been published before [17,18]. If a cancer expression module is coordinately up-regulated across samples of a new data set, then this is a sign of biological importance. A footprint of pairwise correlations between module genes across all samples of a study justifies the relevance of a gene expression module in that specific data set. On the other side, biological irrelevance, technical noise or different patient characteristics in the investigated cohort might lead to the absence of such correlations. Therefore, the translatability of a signature on a new dataset can be assessed by analysis of the correlation structure of the signatures' genes in the new data set. If genes of a signature do correlate with each other, there are some samples in the data that share high expression of the genes in the module indicating a common positive transcriptional regulation of the module that can be the result of diverse biological phenomena such as a particular differentiation state of cells or an upstream pathway activity. One metric for translatability assumes pairwise correlated gene expression to signify translatability of a biological signal. This rationale, to our knowledge, has first been proposed by Rahnenführer *et al.* [19], but, since then, has been proposed in several studies [20,21]. Astonishingly, although the importance of coexpression of signature genes has been recognized, this rationale has, until now, hardly been used in systematic approaches to remove noisy signatures and distil conclusive results during the analyses of large-scale signature collections on novel expression data sets.

In this work, we aim to provide (1) a well-balanced collection of high-fidelity signatures from the literature, (2) an algorithm or workflow to evaluate the applicability of a signature in a new data set and score it on single tumors, and (3) a web-based analytical system for analysis of activities of such signatures. We introduce RosettaSX, a platform that helps users unravel complex processes with robust gene expression signatures. It not only provides robust expression portraits based on underlying expression data, but it also enables users to explore the associations between signature activities and molecular aberrations like somatic mutations and gene copy number changes [49,50] or even profiles from gene dependency screens [46–48].

Materials and methods

Before signature assessment, for each cancer indication, we normalize each expression dataset separately using the trimmed mean of M-values (TMM) normalization [22]. We remove genes with low expression based on a criterion of fewer than ten reads in 70% of the samples or less than 15 reads in the complete dataset before normalization.

For the definition of a per-sample signature score (i.e., signature activity) we use the mean of the gene-wise expression Z-scores; a score previously referred to as the deregulation index (DI) [23]. In detail, we use the average of TMM normalized and z-scaled TPM (accounting for within-sample comparisons) expression values for a set of genes of a signature as a signature score of a single sample.

When applying and interpreting signatures - that have most often been created in other experimental contexts and published in the literature - on a new expression data set, it is beneficial to validate their translatability to these new data first: here to validate the translatability of the RosettaSX signature collection to the TCGA [50] and CCLE [49] expression data. For interpretation of signatures scores we intend to focus only on signatures with strong signs of translatability, leaving aside signatures with low signs of translatability that could rather add noise to an analysis of the whole collection and cause mis- or overinterpretation.

For the assessment of the translatability of a signature to a new data set, we use the coherence score (CS), which is calculated for each signature on each

data set as previously published [21]. We refer to the Supplementary Material for a detailed description of the CS. Although the CS has been developed by us; we note that usage of the concept of coexpression to score the relevance of signatures has been developed independently by other groups [19–21]. Briefly, for a signature with n genes, we calculate the average of all pairwise Pearson correlation coefficients of the TMM normalized Z-scaled counts per million (CPM) values (accounting for between-sample comparisons). A score close to the limits (+1 and $-\frac{1}{n-1}$) indicates a coherent up- or downregulation of all genes within a signature in a dataset under study.

We conducted extensive assessments of the significance of CSs based on permuted expression data and randomly sampled signatures, and also investigated the dependence of significance on the number of genes in the signature (see Supplementary Figure 1-4). We find that CSs greater than 0.20 (our default RosettaSX threshold) indicate high confidence in the translatability of signatures as indicated by empirical p-values and by a very strong correlation of signature scoring methods for such high CSs, even if signatures are small in the number of genes. For larger signatures with more than 20 genes even CSs between 0.12 and 0.20 already suggest that signatures can be regarded as informative in a data set (see Supplementary Figure 1). We also found that ours and various other signature scoring methods are in good agreement with each other if – and only if - the CS of the signature is high (see Supplementary Figure 4 in which correlation of various scoring methods is excellent for a threshold of CS>0.2) supporting our conclusion that in these situations signatures can be called translatable to a new data set. Whereas application of our signature collection on TCGA BRCA data leads to dozens of signatures with CS>0.2, not a single signature exceeds the CS>0.2 threshold in a collection of equally sized randomly sampled signatures (Supplementary Figure 2).

All this confirms (1) that our choice of scoring method is reasonable and (2) that CS-based filtering of signatures leads to results that are robust to small changes in methodology. Thus, the combination of signature prefiltering for translatability using the CS with per-sample signature scoring by the DI is a reliable approach for analysis and interpretation of whole signature collections in tumor expression data.

Results

In the following, we describe how our system enables users to perform in-depth analyses of gene expression signatures based on expression data of cancer models or patients. Then, we demonstrate the potential of RosettaSX to confirm known and discover new expression subtype relations: we demonstrate this by 2 analyses of our gene expression signature collection in breast cancer, and in diffuse large B cell lymphoma.

RosettaSX web application enables comprehensive expression signature analyses

A typical workflow for the usage of the platform is shown in Figure 1. First, the user selects the data source (TCGA patients or CCLE cell lines) and the cancer type of interest. Then, the user needs to specify how to narrow down the available signatures, each describing a biological expression phenotype. Most available signatures in our collection have been discovered on other data than from TCGA and CCLE. Therefore, an evaluation of the robustness or translatability of each signature in the selected data is highly recommended. We offer to test for the coherence of a signature on a new dataset. We consider a signature to be translatable to the respective dataset if a certain threshold of the coherence score is exceeded (see methods).

After signature filtering, the user can interactively add molecular data (e.g., expression, copy number, and mutation data) for a gene of interest or sample phenotype annotations (i.e., tumor subtype) or compute the signature score for a gene set of interest.

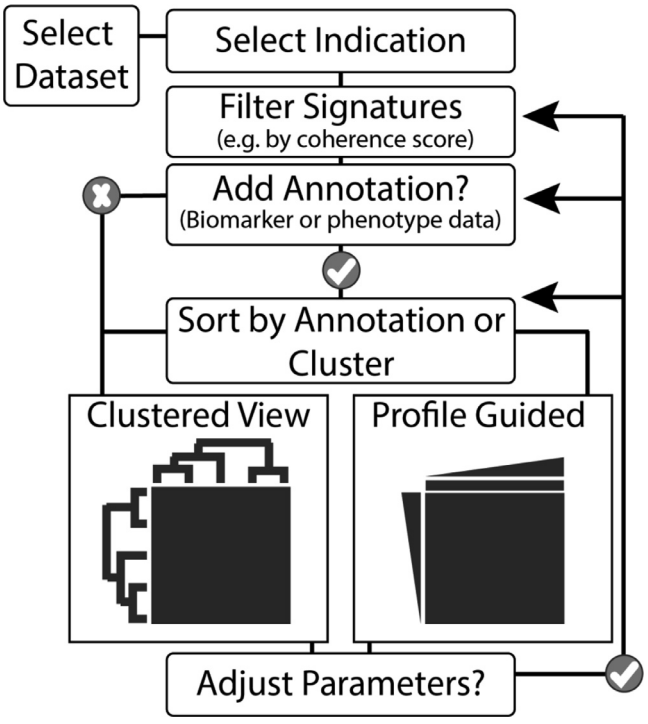

Fig. 1. RosettaSX analysis workflow for expression signature analysis. Our workflow is an iterative procedure that uses the expression phenotypes of signatures to characterize an indication of interest. The user can narrow down the available signatures based on gene set size, availability of genes in the new data, or intrasignature expression coherence. Additionally, the annotation of a wide range of molecular and phenotypic data allows to characterize RosettaSX analysis findings. Depending on the research question, users can choose between a clustered and profile guided heatmap representation. Users can drill-down associations between pairs of annotation parameters or signature scores using box plots and scatter plots.

The web server's key analysis output is a heat map that displays the selected gene expression signature profiles and optionally the selected annotations. Two types of views allow the user to discover different research questions, the profile-guided and clustered view. The clustered view (the default) allows to group together all closely related signatures, which could highlight expression subtypes in the selected dataset. For the profile-guided view, we correlate a continuous annotation with signature scores and order the signatures descending by their correlation coefficient. This view enables discoveries of associations between a selected annotation and the phenotypes that are described by the signatures. Through the interactive web server, the user can then optimize the presentation of his results with further annotations or can adjust the filter criteria.

Intrinsic breast cancer expression subtypes rediscovered in the TCGA breast cancer cohort

Today, the most essential and treatment-relevant classification of breast cancer depends on the expression states of the estrogen receptor (ER), the progesterone receptor (PR), the epidermal growth factor receptor 2 (HER2) [24], and the assessment of cell proliferation activity like Ki67 [25]. Besides these classical markers, commercial gene expression signature tests are applied in the clinical setting for prognosis of breast cancer patients. The most well-known assay is based on work by Perou *et al.*, who described 5 intrinsic subtypes of breast cancer (PAM50) that reconcile the states of these receptors with cell-of-origin expression patterns and describe specific genomic and

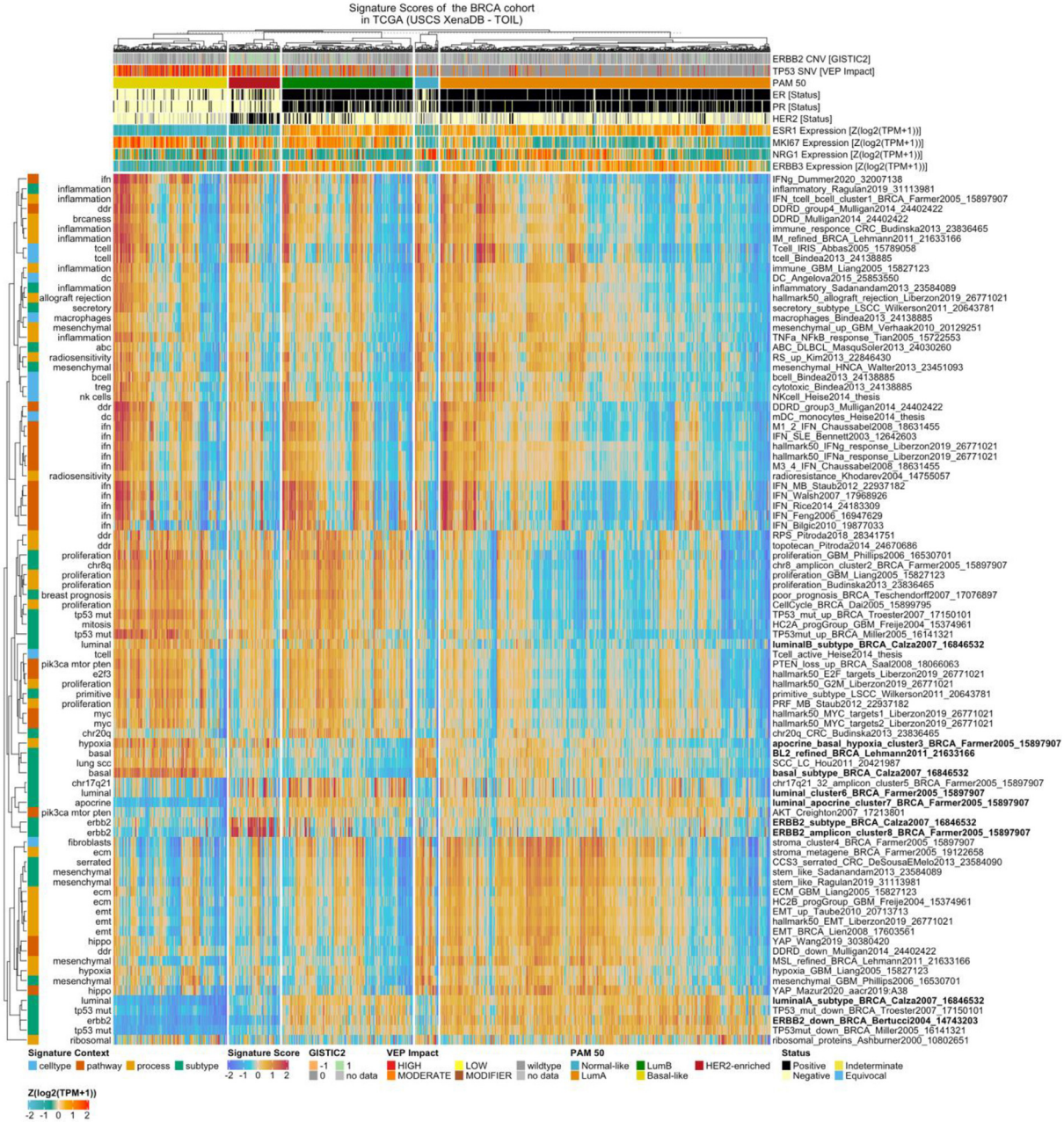

Fig. 2. RosettaSX analysis of intrinsic subtypes in the TCGA breast cancer cohort. The heatmap displays expression signature scores (threshold CS $\geq$ 0.2, 80–100% genes available and expressed, signature size: 3–300) that characterize various robust expression programs and their biological contexts (left text and color annotations) in the PAM50 subtypes. The column (sample) annotations above the heat map display clinical patient markers [39,40] and molecular configurations of single genes (to be selected). The sample annotation comprises: estrogen receptor status (ER), progesterone receptor status (PR), HER2 receptor status (HER2) with positive status in green, negative status in orange, ERBB2 gene amplification (light green) in the HER2 subtype (red), elevated ESR1 expression (orange) in the luminal subtypes (light and dark orange), TP53 mutations in the basal (yellow) subtype, high expression of the MKI67 proliferation marker in basal, luminal B and HER subtype and low expression in the normal-like subtype) align with known properties of the intrinsic subtypes. Bold row labels emphasize signatures that are postulated to describe the different intrinsic subtypes.

clinical characteristics: luminal A and B, basal-like, normal-like, and HER2-enriched [6,26]. Using RosettaSX, we reassess the PAM50 signatures in breast cancer along with classical marker annotations and many other gene expression signatures that were published to be informative to explain breast cancer biology. Figure 2 shows the output of RosettaSX. The status of ER, PR, and HER2 receptors [27,28], the PAM50 subtypes [28] as provided by the TCGA consortium, supplemented by gene expression, SNV, and CNV data of some breast-cancer relevant genes are displayed at the top of the heatmap.

The phenotypic annotations not only align with known traits of the different subtypes but also validate our gene expression pre-processing approach. The ESR1 expression agrees with the annotated ER status (two-sided Wilcoxon-test, *P*-value: < 2.2e-16, Figure 2 and Supplementary Figure 5.A). The status of ER and PR are known to be generally positive in the

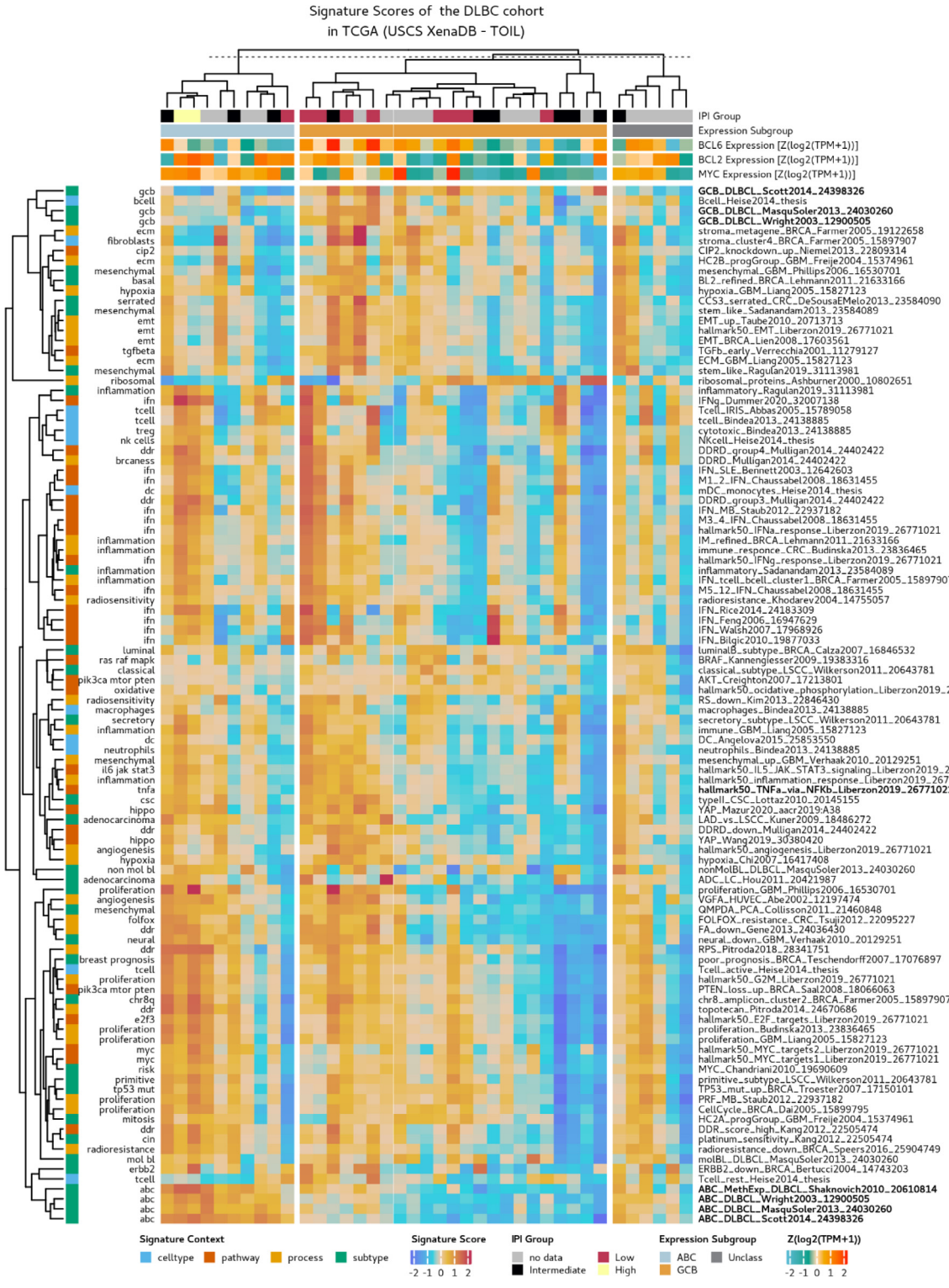

Fig. 3. RosettaSX results for the characterisation of the TCGA Diffuse Large B Cell Lymphoma cohort. We show expression signature scores per tumor for signatures selected by following criteria: CS $\geq$ 0.13; 90-100% genes available; signature size: 3-300 The annotation above the heatmap shows TCGA annotated GCB and ABC (light blue) subtypes with unclassified tumors in grey. In addition, the international prognostic index and the expression intensities of BCL2, BCL6 and MYC are given below. The uppermost and bottommost branches of the dendrogram on the left side represent clusters with ABC and GCB related signatures, respectively.

luminal PAM50 subtypes [26], which is congruent with our findings. Our signature collection comprises 2 signatures of Calza *et al.* [6] that were postulated to characterize luminal A and luminal B subtypes, respectively. Our results indicate that in the TCGA BRCA dataset the luminal A signature is sufficient to identify the luminal or ER+ subtype with elevated levels in both TCGA-annotated luminal subtypes (Supplementary Figure 5.B, fourth plot). As described by several studies, the luminal B subtype is strongly associated with proliferation [29,30]. In our analysis, this is reflected by coclustering of luminal B and proliferation signatures, both of which are elevated in the TCGA-annotated luminal B subtype. The HER2 subtype classification can be validated on both the genomic and transcriptomic level. As initially postulated by Perou *et al.* [26], we also observe a significant difference of ERBB2 gene copy number (DNA level) between subtypes (chi-square, *P*-value: 8.2e-5), with the highest frequency in the HER2 PAM50 expression subtype (Figure 2 and Supplementary Figure 5.A, second plot). Our RosettaSX heatmap comprises multiple nonidentical signatures for the ERBB2 subtype [6,7] (Supplementary Figure 5.B, third plot) which match the PAM50 HER2 annotation well but also are associated with ERBB2 gene copy number.

The basal subtype of breast cancer is defined by a cell-of-origin signature expressed by epithelial cells in the basal or outer layer of the mammary gland [26]. This subtype is a major clinical challenge because these tumors are aggressive, prevalent in young woman, and often relapse rapidly [31]. The basal subtype is known to be nearly congruent with the ER-negative subtype; often also the progesterone and HER2 receptors are negative in these cancers. Indeed, basal cancers constitute the largest fraction of the so-called triple-negative breast cancer (TNBC) group that is difficult to treat and has a poor prognosis [31,32]. In our analysis, the mRNA expression patterns of the ESR1 and PR genes, the basal subtype signature by Calza *et al.*, the clinical annotation of ER and PR status, and the TCGA annotation of the PAM50 basal subtype nicely recapitulate all these relationships (Figure 2).

Basal breast cancers also exhibit the highest frequency of TP53 mutations according to our analysis. Further, the occurrence of TP53 mutations across all breast cancers also correlates with strong signals of the signatures of Miller *et al.* and Troester *et al.* [33,34]. Their signatures are part of a larger cluster comprising many proliferation signatures. These relationships suggest a link between loss-of-function of TP53 [33,34], proliferative aggressiveness of the tumors [35,36], and enrichment of these molecular configurations in the basal subtype [6,7]. Lehmann *et al.* have investigated the subtype structure of TNBCs and described 6 TNBC subtypes [37]. Two of their signatures can even be robustly translated into the TCGA breast cancer data set: the immunomodulatory (IMM) subtype signature matches the profiles of many other inflammatory signatures and marks a high fraction of basal breast cancers as immune-hot. The basal-like 2 (BL2) signature is most strongly up-regulated in the basal subtype in agreement with the finding of Lehmann *et al.* They also described an association of the BL2 subtype with RTK pathway activation (EGFR, NGF, MET, IGF1R) and glycolysis. This is compatible with our findings of a weaker, but substantial BL2 subtype signature signal in the normal PAM50 subtype, indicative of a differentiated epithelial phenotype.

Proliferation activity of breast cancers is strongly linked to poor prognosis [16]. Ki67 protein expression is an important standard marker used to determine the fraction of cells undergoing mitosis and is part of routine diagnostics procedures for breast cancer. Our RosettaSX analysis reveals a large cluster of signatures that are linked to proliferation, comprising, among others signatures by Farmer et. al and Budinska et al. (Fig. 2).These proliferation signatures show a strong correlation to the mRNA expression signal of iMKI67 (Supplementary Figure 6). High cell cycle activity can predominantly be observed in the basal, HER2, and luminal B subtypes (Supplementary Figure 6, first plot). This is consistent with earlier findings of higher proliferation rates in basal, HER2, and luminal B subtypes [6,38]. It is noteworthy that the luminal B signature of Calza *et al.* clusters with proliferation signatures and is not primarily associated with the PAM50 luminal B subtype, whereas the luminal A signature is indicative of all luminal cancers (nearly congruent with the ER+ subtype). So, whereas the luminal cell-of-origin instead seems to be captured by the luminal A signature, the luminal B signature captures the phenomenon of mitotic activity (Supplementary Figure 5.AIII), thereby providing an analogous alternative to traditional ER-positivity and the fraction of tumor cells undergoing mitoses (Supplementary Figure 5.B, first plot). This is consistent with earlier findings of higher proliferation rates in basal, HER2, and luminal B subtypes [6,38]. It is noteworthy that the luminal B signature of Calza *et al.* clusters with proliferation signatures and is not primarily associated with the PAM50 luminal B subtype, whereas the luminal A signature is indicative of all luminal cancers (nearly congruent with the ER+ subtype). So, whereas the luminal cell-of-origin instead seems to be captured by the luminal A signature, the luminal B signature captures the phenomenon of mitotic activity (Supplementary Figure 5.AIII), thereby providing an analogous alternative to traditional ER-positivity and the fraction of tumor cells undergoing mitoses.

Many further signatures yield sufficient coherence scores, thus suggest that the interpretation of their scores is warranted. For example, we observe a high inflammation signature activity in subsets of samples within each PAM50 subtype. This has been described before for other cohorts by one of us [39] and is thought to indicate immune cell infiltration. As immune cell infiltration is a pre-requisite for anti-tumor immune response, such signatures have been proposed to classify tumors into immunologically "hot" or "cold" [40]. Furthermore, we observe the imperfect association of multiple immune cell-type-specific signatures (T-cell, B-cell, and NK-cells) with general immune activation signatures. Fine-grained differences in these patterns, pointing to the absence or presence of a specific cell type, could help to identify the immune status of a tumor more precisely.

A cluster of signatures at the bottom of our heatmap is related to the differentiation status of breast cancers. Stemness or mesenchymality of cancer, for example, after epithelial-mesenchymal transition (EMT), is often associated with poor prognosis, tendency to metastasise, and with resistance to chemotherapy [[4,41]]. While studies on cell-lines suggested that EMT is active in basal and claudin-low breast cancers [4], our analysis partly aligns with recent studies on primary tumors that showed higher proportions of EMT high samples in nonbasal subtypes [42] (Supplementary Figure 5.B, second plot). In alignment with the work of Savci–Heijink *et al.*, EMT signatures in this TCGA cohort have higher scores in luminal A, B, and HER2 subtypes than in the basal subtype.

The interactions of tumor cells, tumor associated stromal cells (TASCs) and extracellular matrix (ECM) are crucial for tumor development and progression. In luminal breast cancer, the binding of HER3 (ERBB3) and neuregulin 1 (NRG1) results in PI3K/AKT, MAPK and JAK/STAT signaling which is associated with tumor progression [43]. Berdiel–Acer *et al.* showed that in luminal breast cancers, cancer associated fibroblasts (CAFs) are key producers of NRG1 promoting a paracrine activation of HER3 receptors. In our analysis, there are increased stromal and ECM signature scores in a subset of luminal cancers. HER3 expression is also elevated in luminal cancers and the profile of NRG1 expression strongly correlates with stromal/fibroblast signatures (Supplementary Figure 7). Together these profiles indicate a strong stroma-tumor interaction in a subset of luminal cells supporting NRG1 as a possible biomarker for anti-HER3 therapies.

Major DLBCL subtypes recapitulated by RosettaSX analysis of DLBCL TCGA cohort

Diffuse large B-cell lymphoma (DLBCL) is a subtype of non-Hodgkin's lymphoma, which is classified by the world health organization into 13 variants with specific morphologic or immune phenotypic features or cases that are not otherwise specified [44]. A well-established gene expression

profiling approach for DLBCL is based on expression signatures that are linked to the differentiation status of the tumors cell of origin. While germinal center B-cell-like (GCB) tumors originate from an earlier stage of B cell differentiation, the activated B-cell-like (ABC) tumors derive from a later, more differentiated stage.

Here, we analyze the DLBCL cohort of TCGA. We demonstrate how available ABC and GCB expression signatures perform to classify DLBCL, and how ABC and GCB subtypes are related to activities of other pathways and cellular programs (Fig. 3). To this end, we use TCGA-provided GCB and ABC annotations as a reference [51]. The heatmap only shows signatures of coherently expressed signatures (CS $\geq$ 0.13). Our signatures for ABC and GCB subtypes in the heatmap match the TCGA reference annotation very well and even inform on the cancers that remained unclassified by TCGA. The heatmap highlights the molecular diversity of DLBCL by the multitude of signatures for other processes that are relevant during lymphomagenesis [8]. The heatmap highlights strongly different proliferation activities in DLBCL samples that are mostly congruent with the ABC-vs-GCB activation. Strongly proliferating cancers occur at higher frequency in ABC samples as supported by many proliferation signatures, some of them also being informative in breast and colon cancer. In addition, we observe the activation of NF-κB or general immune activation signatures in ABC cancers. Indeed, this is the hallmark feature of activated B cells in which NF-κB pathway activation drives B cells into proliferation and enables them to evade apoptosis [45]. So, the concomitant activation of proliferation and NF-κB signatures in the majority of ABC classified tumors is in agreement with expectations. Also, in those cancers that are tagged by TCGA as unclassified, we can observe activation of proliferation signatures in ABC signature-high cancers, suggesting that TCGA has conservatively assigned subtypes and we can safely extend the annotation for several patients by our analysis of multiple signatures. However, we also see that the association of the ABC phenotype is not perfectly congruent with NF-κB activation and proliferation. The existence of proliferation-high-ABC and proliferation-low-GCB cancers could be an interesting starting point for further studies into the molecular origins of this behavior and the clinical consequences for prognosis and optimal treatment selection. Further, we note that there are strong expression modules that are not at all congruent with ABC and GCB subtypes. Multiple type-I IFN response signatures have very high coherence scores, form a tight cluster, with their profiles being uncorrelated to ABC or GCB signature profiles. This suggests that IFN response activity could be an alternative phenomenon that could be relevant for diagnosis and treatment of DLBCL.

Discussion and conclusions

We introduced a new workflow for gene expression signature assessment in cancer and a web server that provides easy access to the analysis of 293 carefully selected signatures in 2 well-known large-scale cancer expression datasets. Together, these data cover more than 30 cancer types with more than 11,000 cancer samples. Our platform supports the user in the identification of robust signatures for subtyping of cancer models or patient cohorts with flexible analysis options, high-quality heatmap visualizations and capabilities to drill-down RosettaSX analysis results. To our knowledge, a comparable system has not been made available yet: it fills a gap in the landscape of data analysis tools for the cancer community.

Using RosettaSX, we were able to recapitulate cell-of-origin subtypes and major molecular pathways in 2 cancer indications. In breast cancer, we could show that the signature collection available on RosettaSX is capable of distinguishing luminal, basal and HER2 PAM50 subtypes, with luminal A signature scores matching the luminal basis annotation and the luminal B signature being tightly linked to proliferation as previously published. We were capable to demonstrate that other robust signatures are valid indicators of breast cancer: immune activation, proliferation, TP53 mutation signatures among others match only partially to PAM50 subtypes: multiple signatures always supported their profiles for a phenomenon. Similarly, for DLBCL, we were able to recapitulate the known major expression-based subtyping schemes, the ABC-vs-GCB classification. Still, we could also show that other robust signatures, e.g., related to interferon response, or proliferation, play important roles and are not congruent with subtyping schemes. We argue that our methodology is well suited to study these landscapes of nonmutually exclusive subtyping schemes.

One strength of our approach is our curated collection of signatures. It covers a wide range of biological phenomena and well-accepted subtyping signatures that are specifically relevant for distinct cancer types. Over the past, the provided signatures proved their integrity in a wide range of applications (unpublished work). We argue that rigorous filtering by signature coherence (with significant empirical p-values) combined with a controlled redundancy of non-overlapping signatures is a distinctive and favorable characteristic of our approach. It eliminates signatures with spurious gene-by-gene correlation before performing downstream analyses (here hierarchically clustered heatmaps of signatures). Examples illustrated that using our method we very often identify multiple coherent signatures for major known subtypes (e.g., for PAM50 and ABC/GCB subtypes and proliferation in DLBCL).

Our main criterion for filtering signatures for translatability and describing their activity in a new dataset is the coherence score. We have tested similar metrics [17,18] earlier (unpublished results) and found it to yield similar results to the coherence score. Furthermore, the recapitulated major literature findings described in the results section, proved, that our proposed scores represent signature activity and pinpoint relevant signatures well. The proposed thresholds for coherence scores yield significant empirical p-values for signatures of size greater than ten genes (and even below). Our analyses on the significance of CS based on randomly sampled signatures (Supplementary Figures 1 and 2) or shuffled expression data (Supplementary Figure 3), the agreement of signature scoring approaches only when the CS is high (Supplementary Figure 4), supplemented by coclustering of multiple signatures for the same phenomena (Fig. 2 and Fig. 3) confirm that our approach does not only yield statistically but also biologically significant results. Signature and data permutation approaches show that our approach is valid for signatures of a wide range of sizes, down to only few genes. In the RosettaSX analytical web platform, we link the signature translatability assessment with downstream tools that allow disentangling biological gene expression phenotypes in large-scale datasets.

Gene expression signatures are a key element to reduce complexity and add interpretability in analyses of expression phenotypes. RosettaSX provides access to signature scores and analysis options for more than 11,000 cancer samples of patients and tumor models in more than 30 different cancer indications. Relying on efficient and straightforward established methods that are distilled into our workflow, this resource can help users to robustly characterize signatures as biomarkers in widely used multiomics data sets.

Author contributions

Julian Kreis: Writing - Original Draft, Writing - Review & Editing, Conceptualization, Methodology, Software, Visualization, Data Curation, Formal Analysis **Boro Nedic**: Software, Visualization **Johanna Mazur**: Writing - Review & Editing, Data Curation, **Miriam Urban**: Data Curation, Conceptualization, Methodology **Sven-Eric-Schelhorn**: Data Curation, **Thomas Grombacher**: Writing - Review & Editing, Data Curation, **Felix Geist**: Writing - Review & Editing, Data Curation, **Benedikt Brors**: Writing - Review & Editing, **Michael Zühlsdorf**: Writing - Review & Editing, **Eike Staub**: Writing - Review & Editing, Supervision, Conceptualization, Methodology, Project administration

Competing interests

All authors declare that they have no competing financial or nonfinancial interests that might have influenced the performance or presentation of the work described in this manuscript.

Funding

Funding for the PhD project of JK has been provided by Merck Healthcare KGaA, Darmstadt, Germany.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neo.2021.08.005.

References

[1] Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 2004;**36**:1090–8.

[2] Angelova M, Charoentong P, Hackl H, Fischer ML, Snajder R, Krogsdam AM, Waldner MJ, Bindea G, Mlecnik B, Galon J. Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol* 2015;**16**:1–17.

[3] Bild AH, Potti A, Nevins JR. Linking oncogenic pathways with therapeutic opportunities. *Nat. Rev. Cancer* 2006;**6**:735–41.

[4] Taube JH, Herschkowitz JI, Komurov K, Zhou AY, Gupta S, Yang J, Hartwell K, Onder TT, Gupta PB, Evans KW. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. U. S. A.* 2010;**107**:15449–54.

[5] Ragulan C, Eason K, Fontana E, Nyamundanda G, Tarazona N, Patil Y, Poudel P, Lawlor RT, Del Rio M, Koo SL. Analytical validation of multiplex biomarker assay to stratify colorectal cancer into molecular subtypes. *Sci. Rep.* 2019;**9**:1–12.

[6] Calza S, Hall P, Auer G, Bjöhle J, Klaar S, Kronenwett U, Liu ET, Miller L, Ploner A, Smeds J. Intrinsic molecular signature of breast cancer in a population-based cohort of 412 patients. *Breast Cancer Res* 2006;**8**:1–9.

[7] Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, MacGrogan G, Bergh J, Cameron D, Goldstein D. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene* 2005;**24**:4660–71.

[8] Scott DW, Gascoyne RD. The tumour microenvironment in B cell lymphomas. *Nat. Rev. Cancer* 2014;**14**:517–34.

[9] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74.

[10] Gnant M, Filipits M, Greil R, Stoeger H, Rudas M, Bago-Horvath Z, Mlineritsch B, Kwasny W, Knauer M, Singer C. Predicting distant recurrence in receptor-positive breast cancer patients with limited clinicopathological risk: using the PAM50 Risk of Recurrence score in 1478 postmenopausal patients of the ABCSG-8 trial treated with adjuvant endocrine therapy alone. *Ann. Oncol.* 2014;**25**:339–45.

[11] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;**490**:61–70.

[12] Laiho P, Kokko A, Vanharanta S, Salovaara R, Sammalkorpi H, Järvinen H, Mecklin JP, Karttunen TJ, Tuppurainen K, Davalos V. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene* 2007;**26**:312–20.

[13] Tsuji S, Midorikawa Y, Takahashi T, Yagi K, Takayama T, Yoshida K, Sugiyama Y, Aburatani H. Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis. *Br. J. Cancer* 2012;**106**:126–32.

[14] Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–35.

[15] Goh WW, Bin and Wong L. Why breast cancer signatures are no better than random signatures explained. *Drug Discov. Today* 2018;**23**:1818–23.

[16] Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 2011;**7**:7.

[17] Dhawan A, Barberis A, Cheng WC, Domingo E, West C, Maughan T, Scott JG, Harris AL, Buffa FM. Guidelines for using sigQC for systematic evaluation of gene signatures. *Nat. Protoc.* 2019;**14**:1377–400.

[18] Berglund AE, Welsh EA, Eschrich SA. Characteristics and validation techniques for PCA-based gene-expression signatures. *Int. J. Genomics* 2017;**2017**:1–13 2354564 **2017**.

[19] Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* 2004;**3**:1–31 16.

[20] Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, Kaper F, Fan JB, Zhang K, Chun J. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 2016;**13**:241–4.

[21] Staub E. An interferon response gene expression signature is activated in a subset of medulloblastomas. *Transl. Oncol.* 2012;**5**:297–304.

[22] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2009;**26**:139–40.

[23] Ebi H, Tomida S, Takeuchi T, Arima C, Sato T, Mitsudomi T, Yatabe Y, Osada H, Takahashi T. Relationship of deregulated signaling converging onto mTOR with prognosis and classification of lung adenocarcinoma shown by two independent in silico analyses. *Cancer Res* 2009;**69**:4027–35.

[24] N. Harbeck, F. Penault-Llorca, J. Cortes, M. Gnant, N. Houssami, P. Poortmans, K. Ruddy, J. Tsang, F. Cardoso, Breast cancer, *Nat. Rev. Dis. Primers*, **5**, (1), 2019, 66

[25] Hashmi AA, Hashmi KA, Irfan M, Khan SM, Edhi MM, Ali JP, Hashmi SK, Asif H, Faridi N, Khan A. Ki67 index in intrinsic breast cancer subtypes and its association with prognostic parameters. *BMC Res. Notes* 2019;**12**:605.

[26] Perou CM, Sørile T, Eisen MB, Van De Rijn M, Jeffrey SS, Ress CA, Pollack JR, Ross DT, Johnsen H, Akslen LA. Molecular portraits of human breast tumours. *Nature* 2000;**406**:747–52.

[27] Caicedo HH, Hashimoto DA, Caicedo JC, Pentland A, Pisano GP. Overcoming barriers to early disease intervention. *Nat. Biotechnol.* 2020;**38**:669–73.

[28] Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. *Nat. Commun.* 2020;**11**:1–11.

[29] Ades F, Zardavas D, Bozovic-Spasojevic I, Pugliano L, Fumagalli D, De Azambuja E, Viale G, Sotiriou C, Piccart M. Luminal B breast cancer: Molecular characterization, clinical management, and future perspectives. *J. Clin. Oncol.* 2014;**32**:2794–803.

[30] Feeley LP, Mulligan AM, Pinnaduwage D, Bull SB, Andrulis IL. Distinguishing luminal breast cancer subtypes by Ki67, progesterone receptor or TP53 status provides prognostic information. *Mod. Pathol.* 2014;**27**:554–61.

[31] Toft DJ, Cryns VL. Minireview: basal-like breast cancer: from molecular profiles to targeted therapies. *Mol. Endocrinol.* 2011;**25**:199–211.

[32] Williams LA, Butler EN, Sun X, Allott EH, Cohen SM, Fuller AM, Hoadley KA, Perou CM, Geradts J, Olshan AF. TP53 protein levels, RNA-based pathway assessment, and race among invasive breast cancer cases. *npj Breast Cancer* 2018;**4**:1–6.

[33] Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U.S.A.* 2005;**102**:13550–5.

[34] Troester MA, Herschkowitz JI, Oh DS, He X, Hoadley KA, Barbier CS, Perou CM. Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* 2006;**6**:1–13.

[35] Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P, Graeme Hodgson J, Weinrich S. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J. Pathol.* 2013;**231**:63–76.

[36] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst* 2015;**1**:417–25.

[37] Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.* 2011;**121**:2750–67.

[38] Bertucci F, Finetti P, Birnbaum D. Basal breast cancer: a complex and deadly molecular subtype. *Curr. Mol. Med.* 2011;**12**:96–110.

[39] Schmidt M, Hellwig B, Hammad S, Othman A, Lohr M, Chen Z, Boehm D, Gebhard S, Petry I, Lebrecht A. A comprehensive analysis of human gene expression profiles identifies stromal immunoglobulin κ C as a compatible prognostic marker in human solid tumors. *Clin. Cancer Res.* 2012;**18**:2695–703.

[40] Shen R, Li P, Li B, Zhang B, Feng L, Cheng S. Identification of distinct immune subtypes in colorectal cancer based on the stromal compartment. *Front. Oncol.* 2020;**9**:1–15.

[41] Liu F, Gu LN, Shan BE, Geng CZ, Sang MX. Biomarkers for EMT and MET in breast cancer: an update (review). *Oncol. Lett.* 2016;**12**:4869–76.

[42] Savci-Heijink CD, Halfwerk H, Hooijer GKJ, Koster J, Horlings HM, Meijer SL, van de Vijver MJ. Epithelial-to-mesenchymal transition status of primary breast carcinomas and its correlation with metastatic behavior. *Breast Cancer Res. Treat.* 2019;**174**:649–59.

[43] Berdiel-Acer M, Maia A, Hristova Z, Borgoni S, Vetter M, Burmester S, Becki C, Michels B, Abnaof K, Binenbaum I. Stromal NRG1 in luminal breast cancer defines pro-fibrotic and migratory cancer-associated fibroblasts. *Oncogene* 2021;**40**:2651–66.

[44] Sukswai N, Lyapichev K, Khoury JD, Medeiros LJ. Diffuse large B-cell lymphoma variants: an update. *Pathology* 2020;**52**:53–67.

[45] Basso K, Dalla-Favera R. Germinal centres and B cell lymphomagenesis. *Nat. Rev. Immunol.* 2015;**15**:172–84.

[46] Barretina J, Caponigro G, Stransky N. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.

[47] Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* 2017;**49**:1779–84.

[48] Dempster JM, Rossen J, Kazachkova M, Pan L, Kugener G, Root DE, Tsherniak A. Extracting biological insights from the project Achilles Genome-Scale CRISPR screens in cancer cell lines. *bioRxiv* 2019. doi:10.1101/720243.

[49] Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, Barretina J, Gelfand ET, Bielski CM, Li H. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 2019;**569**:503–8.

[50] Goldman M, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, Zhu J, Haussler D. The UCSC Xena platform for cancer genomics data visualization and interpretation Paper Introduction. *bioRxiv* 2018. doi:10.1101/326470.

[51] Schmitz R, Wright GW, Huang CA, Johnson CA, Phelan JD, Wang JQ, Roulland S, Kasbekar M, Young RM, Shaffer AL, Hodson DJ, Xiao W, Yu X, Yang Y, Zhao H, Xu W, Liu X, Zhou B, Du W, Chan WC, Jaffe ES, Gascoyne RD, Connors JM, Campo E, Lopez-Guillermo A, Rosenwald A, Ott G, Delabie J, Rimsza LM, Tay Kuang Wei K, Zelenetz AD, Leonard JP, Bartlett NL, Tran B, Shetty J, Zhao Y, Soppet DR, Pittaluga S, Wilson WH, Staudt LM. The Molecular Signatures Database Hallmark Gene Set Collection. *N Engl J Med* 2018;**378**(15):1396–407. doi:10.1056/NEJMoa1801445.