Title: NETCELLMATCH: MULTISCALE NETWORK-BASED MATCHING OF CANCER CELL LINES TO PATIENTS USING GRAPHICAL WAVELETS

Authors: Neel Desai; Jeffrey Morris; Veerabhadran Baladandayuthapani

To be cited as: 10.1002/cbdv.202200746

Link to VoR: https://doi.org/10.1002/cbdv.202200746

NETCELLMATCH: MULTISCALE NETWORK-BASED MATCHING OF CANCER CELL LINES TO PATIENTS USING GRAPHICAL WAVELETS

Neel Desai
Division of Biostatistics
University of Pennsylvania Perelman School of Medicine
Philadelphia, PA 19104

Jeffrey S. Morris
Division of Biostatistics
University of Pennsylvania Perelman School of Medicine
Philadelphia, PA 19104

Veerabhadran Baladandayuthapani
Department of Biostatistics
University of Michigan School of Public Health
Ann Arbor, MI 48109

ABSTRACT

Cancer cell lines serve as model *in vitro* systems for investigating therapeutic interventions. Recent advances in high-throughput genomic profiling have enabled the systematic comparison between cell lines and patient tumor samples. The highly interconnected nature of biological data, however, presents a challenge when mapping patient tumors to cell lines. Standard clustering methods can be particularly susceptible to the high level of noise present in these datasets and only output clusters at one unknown scale of the data. In light of these challenges, we present NetCellMatch, a robust framework for network-based matching of cell lines to patient tumors. NetCellMatch first constructs a global network across all cell line-patient samples using their genomic similarity. Then, a multi-scale community detection algorithm integrates information across topologically meaningful (clustering) scales to obtain Network-Based Matching Scores (NBMS). NBMS are measures of *cluster robustness* which map patient tumors to cell lines. We use NBMS to determine representative "avatar" cell lines for subgroups of patients. We apply NetCellMatch to reverse-phase protein array data obtained from The Cancer Genome Atlas for patients and the MD Anderson Cell Lines Project for cell lines. Along with avatar cell line identification, we evaluate connectivity patterns for breast, lung, and colon cancer and explore the proteomic profiles of avatars and their corresponding top matching patients. Our results demonstrate our framework's ability to identify both patient-cell line matches and potential proteomic drivers of similarity. Our methods are general and can be easily adapted to other 'omic datasets.

Availability and implementation: NetCellMatch software is freely available at: NetCellMatch

Contact: neel.desai@pennmedicine.upenn.edu

1 Introduction

In a precision medicine paradigm, matching a patient's tumor to the correct drug therapy remains a critical step towards determining an effective patient-specific personalized treatment plan. Compared to generalized treatment options like chemotherapy, a drug therapy properly targeted to a patient tumor has shown increased effectiveness while minimizing harmful side-effects [1]. The tumor-drug matching problem has made recent advances based on tumors' molecular architecture; mutation-specific therapies have been shown to be effective in breast cancer and melanoma [2, 3, 4, 5, 6, 7, 8]. While these approaches effectively incorporate a patient's genomic archictecture, they typically consider only a few select set of mutations. Given that tumors are complex systems driven by multiple molecular abberations, the restrictive scope of these approaches may limit their effectiveness. As an alternative to mutation-specific therapies, standard translational experiments for drug discovery learn genomic regulatory mechanisms using perturbation studies in (idealized) tumor derived cell line model systems [9]. The choice of cell lines is governed by multiple factors; a cell line is typically deemed appropriate if it lacks contamination and produces a biological environment whose functional capabilities match the native phenotypic features of the primary patient cell [10, 11]. A direct approach towards determining this functional compatability involves using the multitude of genomic data available on patient systems to guide the choice of the most appropriate cell lines within and across cancers. The core concept is as follows: a patient matches to a cell line, which serves as an 'avatar' for that patient based on its similar molecular characteristics. Given extensive drug response data for cell lines, the therapies to which the avatar cell line responds become prospective therapies for the corresponding patient. Past results provide validation for this approach's key assumption that cell lines can potentially serve as in-vitro model systems for tumors of the same (sub)type [9, 12].

With the advent of high-throughput molecular profiling, similarities between cell line data and tumor samples have been investigated using multiple data-centric analytical methods [2, 3, 13, 14]. While these approaches have had some success, especially in renal cancer, opportunities remain to better address some of the inherent challenges to match patients and cell lines [15]. Many data-centric approaches separately assess each potential patient-cell line pair to produce

matches; this one-to-one matching strategy discards information that could be used to aid matching accuracy if all potential patient-cell line pairs were considered jointly [16]. Similarly, many existing approaches form patient-cell line pairs based on a few mutations and do not take full advantage of available genomic data [17, 18]. Finally, existing literature does not formally focus on methods which assess the *cluster robustness* of patient-cell line matches. A method which produces a metric that carefully assesses confidence in patient-cell line pairs could greatly aid scientific discovery by counteracting the false discoveries that result from the molecular heterogeneity introduced by random mutations. These challenges, coupled with the cost efficient nature and wide availability of cell lines, underscore the importance of developing new analytical strategies for integrating patient-cell line genomic data for effective implementation of precision medicine strategies [19].

In light of these challenges, we present NetCellMatch, a multi-scale network-based approach that horizontally integrates across cell line and patient data to find molecularly homogeneous groups of tumors and cell lines (Figure 1). Briefly, NetCellMatch adapts a multi-scale community detection algorithm [20] to obtain cluster sets along a topologically meaningful path defined by the joint patient-cell line network structure (Figure 1 and Section 2). NetCellMatch then aggregates information across all steps in this path to yield a Network-Based Matching Score (NBMS) for every patient-cell line combination and to determine representative avatar cell lines for subgroups of patients; integrating matching information with cell line specific drug responses can then be potentially used to identify prospective targeted therapies. Our method's holistic network-based approach uses all available information in the data, potentially improving power compared to more common one-to-one matching strategies where patient-cell line pairs are assessed separately by standard similarity measures such as Euclidean distance or a correlation-based metric [16]. NetCellMatch produces matches by considering multiple biomarkers simultaneously rather than focusing on only a handful of mutations at a time. Finally, NetCellMatch creates a natural relative ranking of all patient-cell line pairs which provides an added uncertainty measure for assessing persistent clustering structure; a pair that forms at many levels of a network's topology is more likely to be meaningful than a pair that only forms at a few levels (Figure 2).

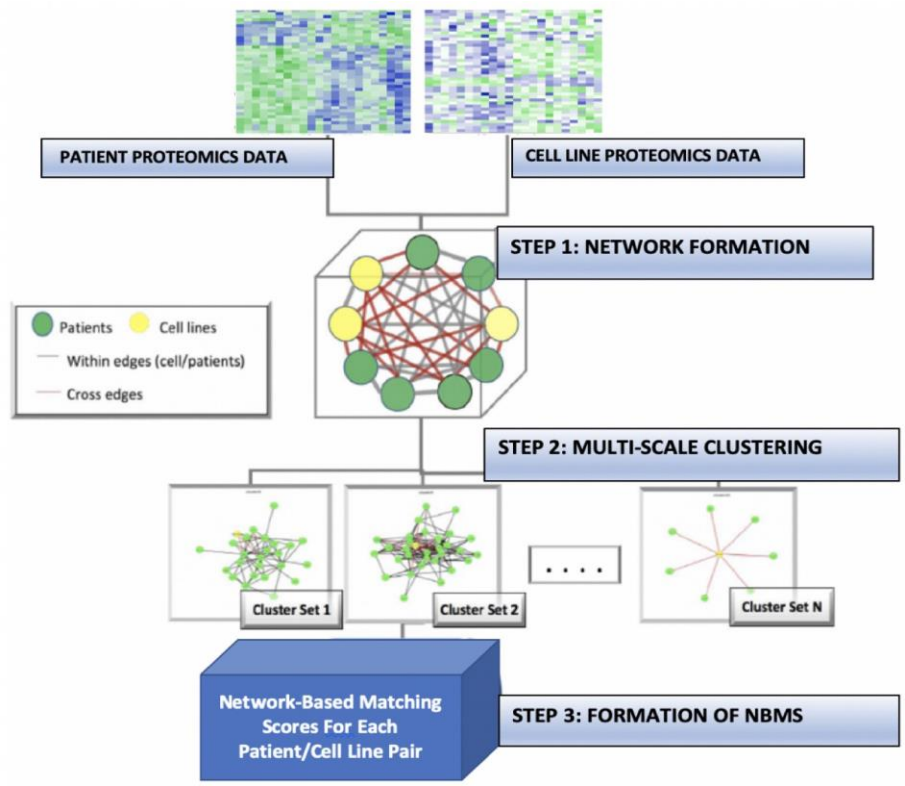

Figure 1: **Schematic representation of NetCellMatch.** In STEP 1, proteomic information is used to form a global patient-cell line network based on a similarity metric (STEP 2) Subnetworks of patients and cell lines are obtained from each of *N* different scales of resolution of the global network (STEP 3) Network-Based Matching Scores are formed for each potential patient-cell line pair by aggregating subnetwork information across scales of resolution.

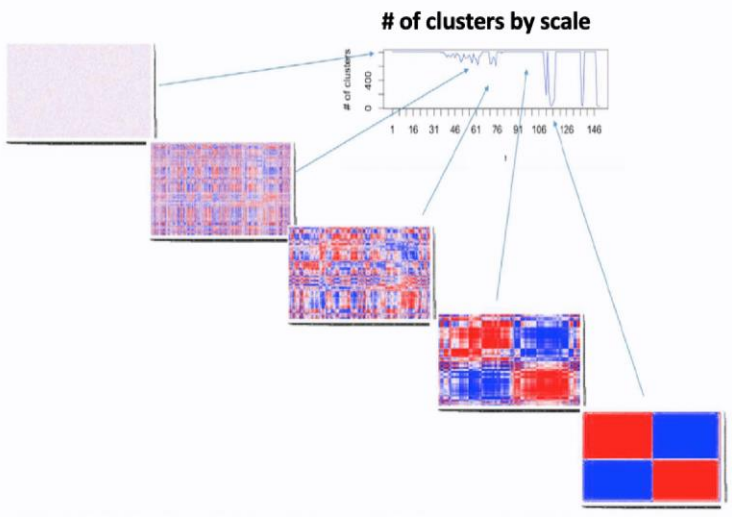

Figure 2: **Clusters across scales of resolution.** Heatmaps represent the global patient-cell line network of NetCellMatch (see Figure 1) at different scales of resolution. At STEP 2 of NetCellMatch, distinct sets of patient-cell line clusters are formed based on the networks produced at each scale. Each scale filters out a portion of information from the global network along a topologically meaningful path.

Our methods are motivated by and applied to a pan-cancer proteomics data set comprised of reverse-phase protein arrays (RPPA) technology from The Cancer Proteome Atlas (TCPA) for patients' samples and MD Anderson Cell Lines Project (MCLP) for cell lines [21, 22]. In this study, we examine patient-cell line pairs across lung, breast, and colon cancer. Along with avatar cell line identification, we evaluate connectivity patterns for each cancer type and explore specific properties of avatars and their corresponding matching patients.

The rest of the article is organized as follows. In section 2 we detail our NetCellMatch framework and outline the methodology's key properties. In section 3 we present the results of NetCellMatch applied to the data described above. In section 4 we discuss our findings and describe potential extensions of the framework.

2 NetCellMatch Framework

We designed NetCellMatch to produce clusters of molecularly homogeneous patient tumors and cell lines with associated measures of cluster robustness. Following the schema of Figure 1, our framework takes a patient-cell line network and uses a multi-scale community detection algorithm to obtain subnetworks of patients and cell lines across different layers of a network's structure [20]. Information is then aggregated across scales to produce matches. Specifically, let vector $\boldsymbol{C}$ represent the genomic profile of a set of cell lines indexed by $i = 1,\ldots, n_c$ (where n_c indicates the number of cell lines) and $\boldsymbol{P}$ represent the corresponding profiles from a set of patients from $j = 1,\ldots, n_p$ (where n_p indicates the number of patients). Given the (matched) genomic information of each cell line and patient (i.e. same set of genes/proteins), our immediate goal is to construct an undirected, fully connected weighted network $\boldsymbol{G}$ using all available (matched) genomic information with set of nodes $\boldsymbol{N}$ (of size $n_c + n_p$). Let $\boldsymbol{E}$ represent the set of edges for network $\boldsymbol{G}$. The relative strength of these edges is determined by measures of genomic similarity which, in contrast to considering patient-cell line pairs individually, borrows strength across a dataset's entire genomic profile, preserving both direct and indirect associations. In order to mine information contained in $\boldsymbol{G}$, we consider $\boldsymbol{K}$ scales of resolution $k = 1,\ldots,K$. We detail each of the steps below.

2.1 STEP 1: Formation of Network and Graph Laplacian

Using all available genomic data, we first construct fully connected network $\boldsymbol{G}$ where edges $\boldsymbol{E}$ between the set of nodes $\boldsymbol{N}$ (i.e. patients and cell lines) are weighted by a suitable distance metric. For our particular application, we use absolute Spearman's correlation because we desire a robust, computationally feasible rank-based metric that has frequently been found to be an appropriate choice for assessing proteomic similarity between biomarkers [23, 24, 25, 26]. We use an absolute measurement because patient-cell line similarity only depends on the metric's magnitude. We also note that our algorithm is flexible to incorporate other distance metrics, such as Pearson's correlation, distance correlation, or mutual information that may be suitable for other biological applications [27]. Before performing multi-scale community detection, we transform $\boldsymbol{G}$ into its corresponding graph Laplacian $\boldsymbol{L}$. This step is done to ensure that our network's structural information is easily accessible; the span of a graph Laplacian's eigenvalues contain the network's clustering information at different scales of resolution. More specifically, clustering the first few eigenvectors of a graph Laplacian provides a network's large scale sub-network structure; clustering with the inclusion of subsequent eigenvectors incorporates finer scales of information when determining sub-networks.

***Normalized Graph Laplacian*:** To form graph Laplacian $\boldsymbol{L}$, first consider $\boldsymbol{G}$; this weighted, undirected network describes the strength of association between each node (i.e. cell lines and patients). Given $\boldsymbol{G}$, and defining its weighted elements as w_{ij}, the Laplacian can be constructed using the following steps [28]:

1. Create diagonal degree matrix $\boldsymbol{D}$. Each element of the diagonal of $\boldsymbol{D}$ is defined as $\sum_{j=1}^{N} w_{ij}$ describes the connectivity of node n_i.
2. Form Laplacian $\boldsymbol{L = D - G}$.
3. A normalized version of $\boldsymbol{L}$ is produced by taking $\boldsymbol{D^{-1/2}LD^{-1/2}}$. This normalized $\boldsymbol{L}$ has a span bounded by $0 \leq \lambda \leq 2$. We will use the normalized version of the Laplacian in this study because having a consistent, numerically tight span makes multi-scale community detection computationally easier.

As previously noted, the Laplacian contains a graph's inherent clustering patterns in its set of eigenvectors. The multiplicity of its first eigenvalue λ_1 is equal to the number of connected components. Likewise, in a fully connected Laplacian the eigenvector associated with the second smallest eigenvalue λ_2 is known to contain the graph's coarsest amount of information [28]. In other words, the graph's most fundamental large-scale clusters will be represented in its Laplacian's second smallest eigenvector. Standard methods such as spectral clustering leverage this fact by employing standard clustering algorithms (e.g. k-means) on this eigenvector to separate a graph into its basic clusters. Similarly, clusters obtained using more eigenvectors of a Laplacian would incorporate smaller scale information when determining network sub-groups. Our method seeks to profitably leverage the information contained at all scales of a graph through the Laplacian's eigenvectors (detailed in STEP 2). Our approach's core idea is that the most molecularly homogenous patient-cell line pairs will cluster together across many different scales of a network's information; aggregating cluster results across scales should therefore provide a measure of cluster robustness for each patient-cell line pair.

Briefly, at this stage we have Laplacian $\boldsymbol{L}$, whose eigenvectors are embedded with our data's connectivity patterns across scales of resolution. We now transition to the next step of our framework (see Figure 1) and obtain patient-cell line subnetworks at many scales of resolution. To perform multiscale community detection, we transform $\boldsymbol{L}$ into the Fourier domain and then apply the continuous wavelet transform; these steps allow us to selectively filter out the Lapacian's eigenvectors and access connectivity information at a particular scale of our network.

2.2 STEP 2: Multi-Scale Network Clustering

It is well-established that the graph Laplacian (and any signal defined on it) can be redefined in terms of a basis of its orthonormal eigenvectors in the Fourier domain [29]. [30] showed that the continuous wavelet transform, in the Fourier domain, acts as a high band-pass filter; the filter's properties are determined by the scale parameter (denoted by s) of the continuous wavelet transform. They further capitalized on both facts by defining the spectral graph wavelet transform (SGWT, [30]). The transform $\boldsymbol{T}$ essentially consists of a high-band pass filter $\boldsymbol{g}$ (i.e. a function

acting as a wavelet) applied to arbitrary signal $\boldsymbol{f}$ defined on a graph's nodes. The transform is defined as $T_g = g(L) = T_g^* f(l) = g(\lambda_l) f^*(l)$ with inverse Fourier transform $\left(T_g f\right)(m) = \sum_{l=0}^{N-1} g(\lambda_l) f^*$, where $f^*(l) = \sum_{m=1}^{N} f(m) e^{iml} dm$. A graph wavelet $\psi_{s,n}$ at scale *s* and node *n* is defined by applying the SGWT to an impulse δ_n (a signal with non-zero weight at only one location) defined on a graph node: $\psi_{s,n} = T_g^s \delta_n$, rewritten as $\psi_{s,n} = \sum_{l=0}^{N-1} g(s\lambda_l) \chi_l^*(n) \chi_l(m)$. Multi-scale community detection utilizes graph wavelets directly; every node has a wavelet serving as a proxy at each scale of resolution. Specific to our application, graph wavelets facilitate finding sub-networks of patients and cell lines at a given scale because they represent each node (i.e. a patient or cell line) based on the genomic similarity information available at only a chosen scale of resolution.

The choice of $\boldsymbol{g}$ is critical when using the SGWT to perform multi-scale community detection. Leveraging that $\boldsymbol{g}$ can be approximated by a low rank polynomial $p(x)$, the function we utilize is as follows:

$$g(x; \alpha, \beta, x_1, x_2) = \begin{cases} x_1^{-\alpha} x^{\alpha} & x < x_1 \\ p(x) & x_1 \leq x \leq x_2 \\ x_2^{\beta} x^{-\beta} & x > x_2 \end{cases}$$

where parameters β, α, x_1, x_2, and maximum and minimum scale s_{max} and s_{min} respectively, are specifically chosen to facilitate community detection [20, 30]. $s_{max} = \frac{x_1}{\lambda_2^2}$ and is set so that filter $g(s_{max}x)$ attenuates information only after λ_2 (therefore the maximum scales keep all large scale cluster information). Similarly, β controls how selective $g(s_{max}x)$ is around information in the second eigenvector; it is set to $\beta = -log_{10}(\frac{\lambda_3}{\lambda_2})$ to attenuate information around λ_3 by a factor of 10. $s_{min} = \frac{x_1}{\lambda_2}$ and is set so that $g(s_{min}\lambda_2)$ considers large scale information when incorporating results at finer scales of resolution. Parameter α controls the attenuation of $\boldsymbol{g}$ around boundary points; higher α means faster attenuation beyond boundary points (we set α = 2 as in the default setting). Finally, $x_2 = \frac{x_1}{\lambda_2}$ in order to guarantee that $g(s_{min}x)$ spans at least half of the information contained in the span of eigenvalues; this setting ensures that our finest scale of resolution considers information sufficiently different than our largest scale.

The modified SGWT described above provides the framework to isolate similarity network information among patients and cell lines at a particular scale of resolution. The transform, when modified for multi-scale community detection, has previously been proposed to extract the structural organization of intra-chromosomal interactions [31]. Before describing how to obtain sub-networks at a particular scale of resolution, we first describe graphical wavelets in more concise notation. At a given scale, the SGWT can generate wavelet basis $\Psi_s = (\psi_{s,1}|\psi_{s,2}|....|\psi_{s,N}) = \chi G_s \chi^T$, where $G_s = diag(g(s\lambda_1), ..., g(s\lambda_N))$ and $\psi_{s,i}$ describes the graph wavelet at scale s for node i. To produce a set of patient-cell line subnetworks at scale s, one requires the following steps: compute wavelet basis Ψ_s, construct matrix D of correlation distances between nodes using the corresponding graph wavelets at each scale as a proxy, and perform Hierarchical clustering of D (these steps are formally outlined in our method's overall algorithm). A visual demonstration of how sub-networks vary across scales of resolution is given in Figure 2. As can be seen, at the coarsest scales of resolution, patients and cell lines are grouped together in a small number of large, inclusive sub-networks and as we consider finer scales of resolution, patients and cell lines are grouped together in larger numbers of small, distinct sub-networks. We next describe the final step in our algorithm, aggregating information across scales to rank patient-cell line pairs.

2.3 STEP 3: Avatar Cell Line Identification

The implementation of the above steps results in a set of patient-cell line clusters c_k across scales $k = 1,...,K$. At each scale the number of clusters returned are directly related to the information being filtered through $\boldsymbol{g}$; $\boldsymbol{g}$ filters through the Laplacian's eigenstructure by selectively attenuating information in eigenvectors through corresponding eigenvalues. We select cell lines based on Network Based Matching Scores (NBMS), a metric aggregated for each patient-cell line combination across all considered scales. For patient i and cell line j, we define NBMS as follows:

$$NBMS_{ij} = \frac{\sum_{k\,=\,s_{min}}^{s_{max}} I_{(c_k)}\,(ij)}{N_{scales}}$$

where $I_{(ck)}(ij)$ indicates whether a patient-cell line pair clustered together at scale k and N_{scales} indicates the total number of scales considered. After obtaining NBMS for all patient-cell line pairs, we normalize matching scores by setting $NBMS_{(max)} = 1$ and multiplying all other $NBMS_{(ij)}$ for $i = 1,\ldots n_c$ and $j = 1,\ldots, n_p$ by the factor needed **Algorithm 1:**

Algorithm 1: NetCellMatch

Result: Network Based Matching Scores (NBMS) R_{ij} for every patient-cell line combination & Avatar Cell Line Scores Z_i

STEP 1: *Formation of Network and Graph Laplacian*

For $(N = n_p + n_c)$, Initialize N×N network of patients n_p and cell lines n_c using distance metric *Dist*;

Initialize N×N Laplacian L;

STEP 2: *Multi-Scale Network Clustering*

for $k = s_{min}, \ldots, s_{max}$ **do**

- compute wavelet basis ψ_k for Laplacian L;
- construct matrix Q of distance correlations from ψ_k;
- perform hierarchical clustering on Q;

end

STEP 3: *Avatar Cell Line Identification*

Given set of cluster results c_k ;

for $i = 1, \ldots, n_c$ & $j = 1, \ldots, n_p$ **do**

- compute $R_{ij} = \frac{\sum_{k=s_{min}}^{s_{max}} I_{(c_k)(ij)}}{N_{scales}}$ for each patient-cell line combination ;
- score cell lines $Z_i = \sum_{j=1}^{n_p} R_{ij}$;
- identify avatar cell lines as $Z_i \geq c$ for user-specified c;

end

to make $NBMS_{(max)} = 1$. We use these normalized scores because the raw co-clustering proportion is dependent on the chosen minimum and maximum scale as well as the distance metric used to form the Laplacian; the only meaningful metric is the relative difference between co-clustering levels. Finally, to rank potential avatar cell lines, each cell line is scored by summing across patients $\sum_{j=1}^{n_p} NBMS_{(ij)}$, which completes the algorithm.

In summary, the best potential avatars will have the highest normalized scores. Cell line scoring is designed to leverage information across all network scales and across the entire network; avatar cell line rankings borrow strength by aggregating across all patients and all network scales to obtain a holistic rank of each cell line. We run the pipeline at a number of scales equally spaced (on a logarithmic scale) between the previously defined minimum and maximum scale to

balance the need to cover a sufficient number of fine scales yet limit computational burden. Our current selection of minimum and maximum scale is chosen to allow the filter to comprehensively cover the network's different structures of information; the finest (smallest) scale treats each patient-cell line as an independent cluster while the largest scale has clusters of a significantly larger size. The NetCellMatch algorithm is presented in **Algorithm 1**.

3 Cancer Cell Line and Patient Matching using Functional Proteomics

3.1 Proteomic profiling of cancer patients and cell lines

We demonstrate the practical utility of our NetCellMatch framework using functional proteomic data across multiple cancer types. Proteomic-based investigations are useful for this purpose since they are closer to functional behavior than genomics and transcriptomics; not all molecular aberrations in cancer can be traced to specific genomic or transcriptomic changes [32]. Reverse Phase Protein Array (RPPA) is a leading technology that allows for simultaneous assessment of expression of multiple protein markers in a cost-effective high-throughput format; RPPA has been extensively validated for patient and cell line samples [33]. Our analysis uses RPPA-based protein expression data of tumors from both patients and cell lines; patient samples are taken from The Cancer Proteome Atlas (TCPA [21]) while cell lines are taken from the MD Anderson Cancer Cell Lines Project (MCLP [22]). These data sources have collaborated with other big cancer data repositories (e.g. TCGA [34], CCLE [35]) to obtain other relevant information about the samples such as clinical history of patients and the sensitivity of drugs on cell lines. We consider 233 cancer related proteins common to patients and cell lines; these proteins are part of major signaling pathways such as *PI3K*, *MAPK*, Transforming Growth Factor β, *WNT*, cell cycle, apoptosis, immune responsiveness, and DNA damage response [22]. The combined data are processed for batch correction and missing value treatment using ComBat (as in [33]).

We focus our analyses on three major cancer types: Breast cancer (cell lines $n_c = 58$, patients $n_p = 879$), Lung cancer ($n_c = 124$, $n_p = 687$), and Colon cancer ($n_c = 35$, $n_p = 360$). For analysis, we consider 150 logarithmically spaced scales between the minimum and maximum to balance assessing a sufficiently fine grid of resolution scales with computational cost. Additional

computational and data pre-processing details available in supplemental section **S.1** and all the data and software codes are available at following github link: NetCellMatch. A comprehensive overview of all non-zero patient-cell line NBMS scores, as well as normalized matching scores produced from aggregating across hierarchical clustering dendrogram cut points for different distance metrics (distance correlation, Euclidean distance), is provided in supplemental section **S.5**. While certain cell lines are prominently flagged across all three approaches, the distinct differences between NBMS results and those of the other two approaches, as well as the high similarity of results from both hierarchical clustering dendrogram aggregation approaches, highlight NetCellMatch's potential to uncover new insights by aggregating clustering results along a topologically meaningful path.

Our major results are organized as follows. In Section 3.2 we highlight the overall connectivity patterns of cell lines for each cancer type that illustrates how our framework extracts connectivity patterns from a similarity network and produces aggregate metrics for patient-cell line pairs; we conclude 3.2 by highlighting salient features of top cell lines and corresponding connecting patients. Section 3.3 explores proteomic expression levels for top connecting avatar/patient pairs demonstrating how our framework can be used to identify proteins potentially driving similarity.

3.2 Global Analysis - Patient-Cell Line Connectivity Patterns

Figure 3 provides a comprehensive visual overview of major patient-cell line connectivity patterns for each cancer type. To assess connectivity between cell lines (green) and patients (red), clustering results were aggregated across scales of resolution; the width of a band between a patient and cell line represents the frequency of co-clustering of the two across scales of resolution. Using this visual overview, we next examine the properties of our top connecting cell lines for each cancer type.

Breast Cancer: Top connecting cell lines are *HCC1395*, *ZR75T*, and *AU565*. *HCC1395* is a triple negative cell line (i.e. no estrogen receptor (ER), progesterone receptor (PR), or growth factor HER-2 over-expression), which has a comparatively negative prognosis compared to other breast cancer subtypes [36]. Common to cancers of this subtype, *HCC1395* has mutation of tumor suppressor gene *p53*. Likewise, this cell line exhibits a mutation of tumor suppressor gene *BRCA1* [37]. Like many triple-negative breast cancers, *HCC1395* over expresses growth factor

EGFR. While chemotherapy is the most common approach to treating triple negative breast cancer, some success has been documented inhibiting *EGFR* with tyrosine kinase inhibitors or antibody cetuximab in combination with chemotherapy [37]. *ZR75T* is of subtype luminal A; this subtype is generally

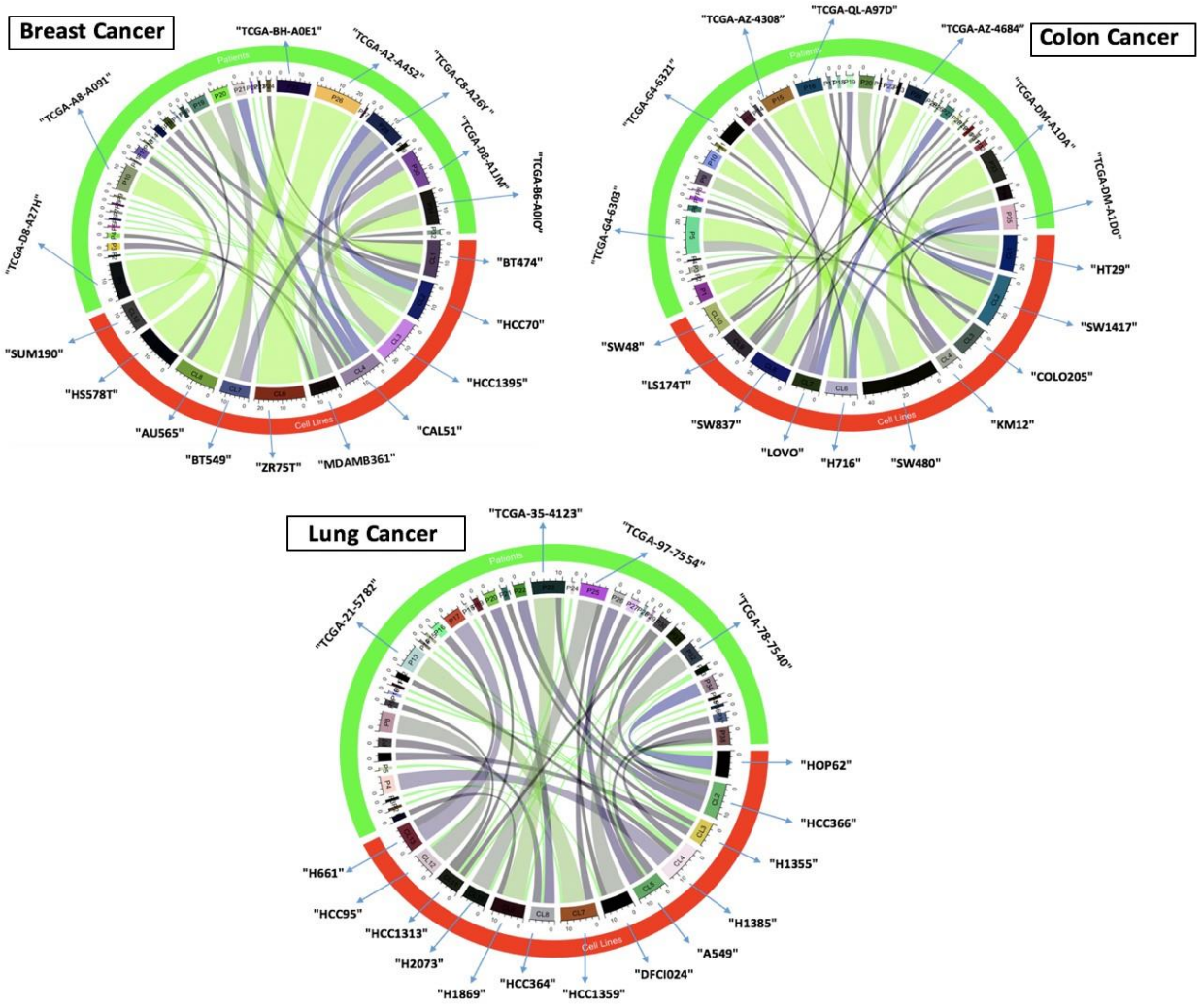

Figure 3: **Patient-cell line connectivity.** Circos plots provide a general overview of connectivity patterns for potential patient-cell line pairs obtained from NetCellMatch. Cell lines are bordered in red while connecting patients are bordered in green. The band widths between each patient cell line pair represent the number of times a pair clustered together across all scales of resolution considered. Connecting patients are shown for cell lines with avatar cell line scores above 0.75.

characterized by the lack of HER over-expression and the presence of ER and PR. Cancers of this subtype are typically treated with hormone therapy to block ER and PR receptors [38]. *ZR75T* is

positive for ER and PR, has tumor suppressor mutation (PTEN), and is positive for the AR androgen receptor; activated AR has been shown to aid suppressing breast cancer's progression. [39]. *AU565* is HER2 over-expression subtype which expresses growth factor *EGFR*. HER2 over expression subtypes are typically treated with tyrosine kinase inhibitors [40]. In addition, recent studies have shown that PARP-1 inhibitors can be used to treat these cancer subtypes [41].

Lung Cancer: Top connecting cell lines are *H1385*, *HCC366*, and *HCC1359*. *H1385* is a NSCLC squamous cell. NSCLC squamous cells can be further classified as primitive, classical, secretory, and basal; *H1385* is classified as both classical and primitive [42]. Classical subtypes are associated with alterations in *KEAP1*, *NFE2L2*, and *PTEN* while primitive subtypes are associated with *RB1* and *PTEN*. *H1385* has a *KRAS* mutation [43]. *HCC366* is a lung adenosquamous carcinoma and is found to have a mutation in *ATM* and have growth factor *EGFR* overexpressed [44]. Lung adenosquamous carcinoma is a rare cancer type which displays a mixture of squamous and adenocarcinoma properties; tyrosine kinase inhibitors have shown success treating lung adenosquamous carcinoma by inhibiting *EGFR* [45]. Additionally, tests on non-small lung cancer with *ATM* mutation suggest that a combination of IR radiation and topoisomerase inhibitors may be an effective treatment option due to the mutation providing increased sensitivity [46]. *HCC1359* is a large cell carcinoma with a mutation in *TP53* [47]. Large cell carcinoma is rare, has a poor prognosis, and typically responds poorly to chemotherapy [48].

Colorectal Cancer: Top connecting cell lines are *SW480*, *SW1417*, and *SW837*. *SW480* is classified as subtype CMS2 and exhibits mutations for tumor supressor genes *p53* and *APC* as well as for *KRAS* [49, 50]. Additionally, *SW480* was positive for *EGFR* [51]. CMS2 is characterized by CIN, WNT, and MYC signaling activation [52]. *SW1417* is also of subtype CMS2 and has mutations in *BRAF*. *APC*, and *p53* [49, 53]. Finally, *SW837* is subtype CMS4 and has a mutation in *KRAS* [54]. Subtype CMS4 is classified by active growth factor *β*, stromal invasion, and angiogenesis [52].

Given summaries of top connecting cell lines, we can scrutinize the proteomic profiles of potential avatar cell lines and corresponding matching patients. Figure 4 visualizes connectivity patterns

between top connecting cell lines and associated patients. For each cancer type (separated by color), the width of edge link between a patient-cell line pair represents normalized NBMS. We next compare proteomic expression profiles of patient-cell line pairs with the highest NBMS in Figure 5. The systematic comparison of proteomic profiles for top matches highlights a key feature of NetCellMatch; results from the algorithm can be used to examine which proteomic (or other genomic) biomarkers drive molecular homogeneity for patient-cell line pairs with similar expression profiles.

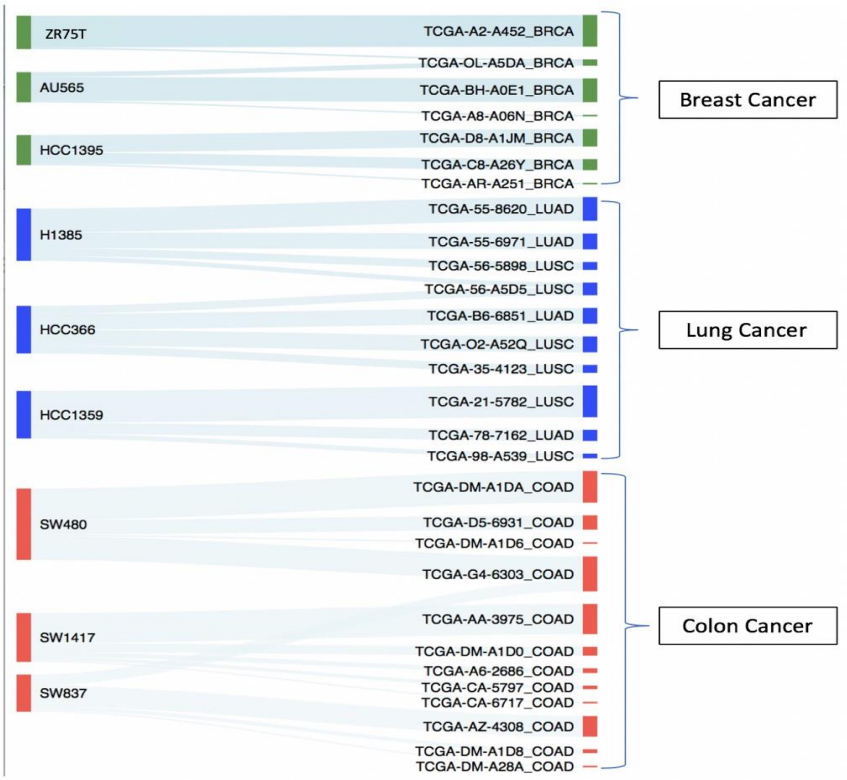

Figure 4: **Avatar cell line connectivity patterns.** Sankey plot of matches between top potential avatar cell lines and corresponding patients. Cell lines are displayed on the left and patients on the right. Breast cancer is in green, lung cancer in blue, and colon cancer in red. The width of a link between a patient-cell line combination represents that pair's NBMS.

3.3 Potential Proteomic Drivers in Patient-Cell Line Pairs

Figure 5 displays a scatter plot comparing expression levels of proteomic markers for a top connecting patient-cell line pair from Figure 4. Expression levels have been standardized so that

phenotypically normal expression levels are marked as 0. For each cancer type, we next explore characteristics of aberrantly expressed biomarkers (those mutually deviating from 0 in the same direction) for top pairs with the highest NBMS. Abberantly expressed proteins provide insight into factors driving molecular homogeneity and could potentially provide insight into appropriate mechanisms of action for tumor treatment. Scatterplots for patient-cell line pairs described below but not shown in Figure 5 can be seen in supplemental section **S.3.**

Breast Cancer: ZR75T ←→ TCGA-A4S2: Aberrantly expressed proteins include TFRC, PARP1, GATA3, and ER-α. TFRPC are receptors which facilitate the transfer of iron to cells; iron is essential to cell growth and development via its role in oxygen development and energy transport [55]. PARP1 is a nuclear enzyme essential to cell DNA damage repair, chromatin dynamics, and transcriptional regulation [56]. In breast cancer, PARP1 has been shown to co-activate GATA3 [57]. GATA3 (hormone signaling pathway) and ER-α share a binding site at the IL-20 promoter region and are co-expressed in breast cancer, responsible for cell growth and proliferation [58, 59]. AU565 ←→ TCGA-A0E1: Aberrantly expressed proteins are HSP70 and RBM15. HSP70 is a protein often found over-expressed in breast cancer cells and is known to be critical in "cellular proliferation, senescence, migration, invasion and tumor growth" [60]. RBM15 (breast reactive pathway) is a RNA-binding protein associated with methylation modification that plays a critical role in cell differentiation [61].

Colon Cancer: SW480 ←→ TCGA-A1DA: Aberrantly expressed proteins include ER-α and CLAUDIN7. CLAUDIN7 regulates tight junctions and maintains cell polarity and connects barriers between cells; interestingly lower expression of CLAUDIN7 is associated with colon cancer and poorer differentiation (as opposed to the over expression we observe) [62]. SW1417 ←→ TCGA-3975: Aberrantly expressed proteins are TIGAR, CLAUDIN7, ER-α, and PEA15. PEA15 is a cytoplasmic protein which plays a key role in cell signaling for processes such as proliferation and apoptosis [63]. TIGAR is known to limit apoptosis authophagy and aid tumor cell survival [64].

Lung Cancer: HCC1359 ←→ TCGA-5782: Aberrantly expressed proteins include ER-α, NDRG1-pT346, and PEA15. Elevated NDRG16 expression in non-small lung cancer is thought to be associated with cancer growth [65].

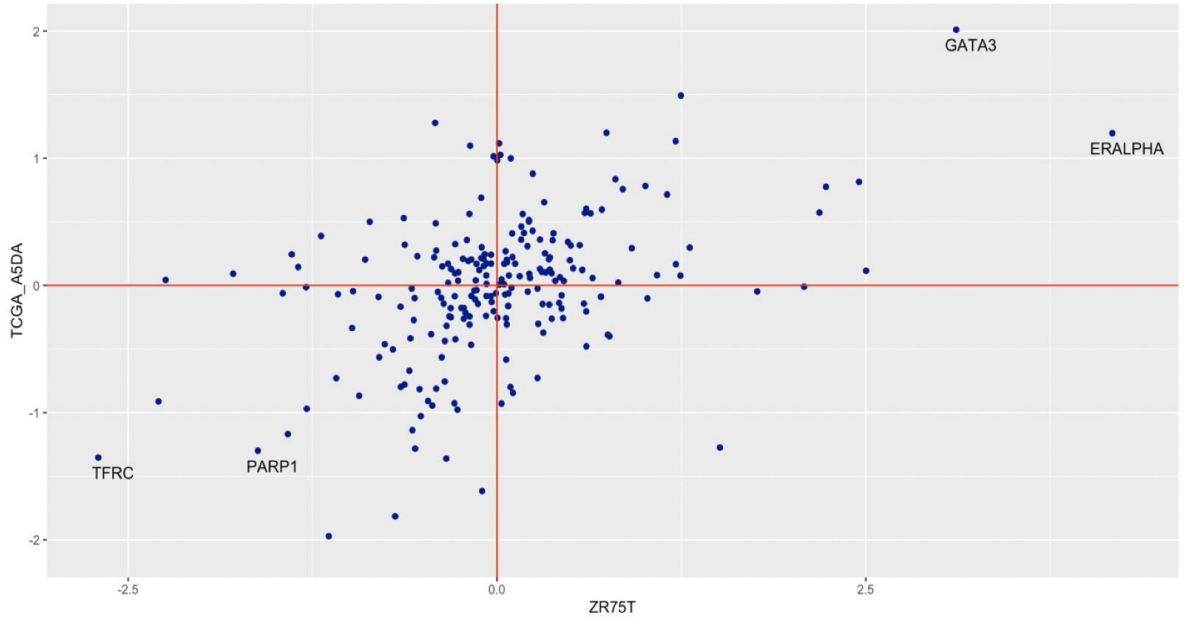

Figure 5: **Evaluating Potential Proteomic Drivers of Similarity.** Breast Cancer - Scatterplot comparing proteomic expression levels of potential avatar cell line ZR75T on the x-axis and top connecting patient TCGA-A5DA on the y-axis. 4 biomarkers are highlighted as being mutually aberrantly expressed: GATA3, ER-α, PARP1, and TFRC.

4 Discussion

We propose a network-based methodology, NetCellMatch, for matching in-vitro cell lines to patient tumor samples. NetCellMatch finds matches via a holistic metric based on proteomic profiling. To enhance robustness, NetCellMatch borrows strength in two distinct ways. It creates matches based on a network rather than individual one-to-one pairs and uses a metric produced by aggregating clustering information across different levels of a network's structure. We apply NetCellMatch to three different cancer types, producing both global avatar cell lines for each cohort and individual network based matching scores (NBMS) for every patient-cell line combination.

Along with its demonstrated capacity for horizontal integration, NetCellMatch has the potential for vertical integration across data types. While we create matches based on proteomic heterogeneity, incorporating other data types such as genomics or transcriptomics may improve matching precision by enhancing cluster accuracy at each scale. It should also be noted that NetCellMatch is sensitive to the underlying distance metric chosen during the network formation stage. We chose absolute Spearman's correlation because it is a non-parametric, robust metric that can be computed with relative computational ease. When applying NetCellMatch to other contexts, the underlying distance metric should be chosen carefully based on the application.

For our particular application, we see two natural extensions for NetCellMatch. We match patients to cell lines due to the maturity of the immortalized cell line literature, and the ready availability of drug efficacy data for cell lines. Patient derived xenograft (PDX) models involve implanting patient-derived cell lines into nude mice, providing a more realistic in vivo models for evaluation. As public data sets including PDX data with genomic and drug response results become available, NetCellMatch could be applied to that setting, as well. Although our matching of cell lines and patients is motivated by the possibility of identifying potential targeted therapies known to work for the avatar cell lines that might be candidates to consider for the individual patient, the investigation and validation of translational value of such a strategy is left for future work. On similar lines, while NetCellMatch is designed as an exploratory algorithmic tool to assess the strength of patient tumor-cell line connectivity, future work could adopt a more (statistical) model-based approach that provides formal assessments of uncertainty and thresholds.

5 Acknowledgments

The authors would like to thank Dr. Sayantan Banerjee for his contribution to the initial stages of code development. Neel Desai was supported by funding from a T-32 NIH training grant. Jeffrey S. Morris was partially supported by the following grants: CA221707, CA-178744, CA244845, P50-CA221707, UL1-TR001878, R01-CA178744, and R01-CA244845-01A1.Veera Baladandayuthapani was supported by following grants: NIH R01-CA244845, R01-CA160736, R21-CA220299, and P30 CA46592, NSF grant 1463233, and start-up funds from the U-M Rogel Cancer Center and School of Public Health.

6 Author Contribution Statement

Neel Desai wrote the article, developed the computational code, conducted data analysis, and drove methodolical development. Jeffrey S. Morris provided guidance on methodological development, feedback and suggestions on article revisions, and suggestions on visualizations. Veera Baladandayuthapani was responsible for formulating the question, providing primary feedback and guidance on methodological development, and feedback and suggestions on article revisions and visualizations.

References

[1] V. Padma, 'An overview of targeted cancer therapy', *BioMedicine* **2015**, *5*, 19

[2] A. Ertel, A. Verghese, S. Byers, M. Ochs, A. Tozeren, 'Pathway-specific differences between tumor cell lines and normal and tumor tissue cells', *Molecular Cancer* **2006**, *5*, 55.

[3] H. Wang, S. Huang, J. Shou, E. Su, J. Onyia, B. Liao, and S. Li, 'Comparative analysis and integrative classification of NCI-60 cell lines and primary tumors using gene expression profiling data', *BMC Genomics* **2006**, *7*, 166.

[4] U. Shankavaram, W. Reinhold, S. Nishizuka, S. Major, D. Morita, K. Chary, M. Reimers, U. Scherf, A. Kahn, D. Dolginow, et al., 'Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study', *Molecular Cancer Therapeutics* **2007**, *6*, 820–832.

[5] S. Myhre, O. Lingjærde, B. Hennessy, M. Aure, M. Carey, J. Alsner, T. Tramm, J. Overgaard, G. Mills, A. Børresen-Dale, et al., 'Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins', *Molecular Oncology* **2013**, *7*, 704–718.

[6] V. Bower, 'NCI-match pairs tumor mutations with matching drugs', *Nature Biotechnology* **2015**, *33*, 790–792.

[7] X. Dong, D. Huang, X. Yi, S. Zhang, Z. Wang, B. Yan, P. Sham, K. Chen, M. Li, 'Diversity spectrum analysis identifies mutation-specific effects of cancer driver genes', *Communications Biology* **2020**, *3*, 1–12.

[8] A. Warren, Y. Chen, A. Jones, T. Shibue, W. Hahn, J. Boehm, F. Vazquez, A. Tsherniak, J. McFarland. 'Global computational alignment of tumor and cell line transcriptional profiles', *Nature Communications* **2021**, *12*, 1–12.

[9] A. Goodspeed, L. Heiser, J. Gray, J. Costello, 'Tumor-derived cell lines as molecular models of cancer pharmacogenomics', *Molecular Cancer Research* **2016**, *14*, 3–13.

[10] G. Kaur, J. Dufour, 'Cell lines: Valuable tools or useless artifacts', *Spermatogenesis* **2012**, *2*, 1-5.

[11] P. Mirabelli, L. Coppola, and M. Salvatore, 'Cancer cell lines are useful model systems for medical research', *Cancers* **2019**, *11*, 1098.

[12] D. Ferreira, F. Adega, R. Chaves, 'The importance of cancer cell lines as in vitro models in cancer methylome analysis and anticancer drugs testing', *Oncogenomics and Cancer Proteomics - novel approaches in biomarkers discovery and therapeutic targets in cancer* **2013**.

[13] M. Salvadores, F. Fuster-Tormo, F. Supek, 'Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns', *Science Advances* **2020**, *6*, 1862.

[14] A. Jarnuczak, H. Najgebauer, M. Barzine, D. Kundu, F. Ghavidel, Y. Perez-Riverol, I. Papatheodorou, A. Brazma, J. Vizcaíno, 'An integrated landscape of protein expression in human cancer', *Scientific Data* **2021**, *8*, 1–14.

[15] N. Lobo, C. Gedye, A. Apostoli, K Brown, J. Paterson, N. Stickle, M. Robinette, N. Fleshner, R. Hamilton, G. Kulkarni, et al., 'Efficient generation of patient-matched malignant and normal primary cell cultures from clear cell renal cell carcinoma patients: clinically relevant models for research and personalized medicine', *BMC Cancer* **2016**, *16*, 1–15.

[16] Z. Qiu, K. Zou, L. Zhuang, J. Qin, H. Li, C. Li, Z. Zhang, X. Chen, J. Cen, Z. Meng, et al., 'Hepatocellular carcinoma cell lines retain the genomic and transcriptomic landscapes of primary human cancers', *Scientific Reports* **2016**, *6*, 27411.

[17] S. Nishizuka, L. Charboneau, L. Young, S. Major, W. Reinhold, M. Waltham, H. Mehr, K. Bussey, J. Lee, V. Espina, et al., 'Proteomic profiling of the NCI-60 cancer cell lines using

new high-density reverse-phase lysate microarrays', *Proceedings of the National Academy of Sciences* **2003**, *100*, 14229–14234.

[18] K. Yu, B. Chen, D. Aran, J. Charalel, C. Yau, DM. Wolf, L. Van'T Veer, A. Butte, T. Goldstein, M. Sirota, 'Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types', *Nature Communications* **2019**, *10*, 1–11.

[19] A. Collins, S. Lang, 'A systematic review of the validity of patient derived xenograft (pdx) models: the implications for translational research and personalised medicine', *PeerJ* **2018**, *6*, 5981.

[20] N. Tremblay, P. Borgnat, 'Graph wavelets for multiscale community mining', *IEEE Transactions on Signal Processing* **2014**, *62*, 5227–5239.

[21] J. Li, Y. Lu, R. Akbani, Z. Ju, P. Roebuck, W. Liu, J. Yang, B. Broom, R. Verhaak, D. Kane, et al., 'TCPA: a resource for cancer functional proteomics data', *Nature Methods* **2013**, *10*, 1046-1047.

[22] J. Li, W. Zhao, R. Akbani, W. Liu, Z. Ju, S. Ling, C. Vellano, P. Roebuck, Q. Yu, A. Eterovic, et al., 'Characterization of human cancer cell lines by reverse-phase protein arrays', *Cancer Cell* **2017**, *31*, 225–239.

[23] J. Hardin, A. Mitani, L. Hicks, B. VanKoten, 'A robust measure of correlation between two genes on a microarray', *BMC Bioinformatics* **2007**, *8*, 1–13.

[24] H. Akoglu, 'User's guide to correlation coefficients', *Turkish Journal of Emergency Medicine* **2018**, *18*, 91–93.

[25] R. Liu, H. Thiessen-Philbrook, R. Vasan, J. Coresh, P. Ganz, J. Bonventre, P. Kimmel, C. Parikh, 'Comparison of proteomic methods in evaluating biomarker-AKI associations in cardiac surgery patients', *Translational Research* **2021**, *238*, 49–62.

[26] Á. Osz, A. Lánczky, B. Gyorffy, 'Survival analysis in breast cancer using proteomic data from four independent datasets', *Scientific Reports* **2021**, *11*, 1–15.

[27] L. Song, P. Langfelder, S. Horvath, 'Comparison of co-expression measures: mutual information, correlation, and model based indices', *BMC Bioinformatics* **2012**, *13*, 328.

[28] U. Von Luxburg, 'A tutorial on spectral clustering', *Statistics and Computing* **2007**, *17*, 395–416.

[29] D. Shuman, B. Ricaud, P. Vandergheynst, 'A windowed graph fourier transform', *IEEE Statistical Signal Processing Workshop (SSP)* **2012**, 133–136.

[30] D. Hammond, P. Vandergheynst, R. Gribonval. 'Wavelets on graphs via spectral graph theory', *Applied and Computational Harmonic Analysis* **2011**, *30*, 129–150.

[31] R. Boulos, N. Tremblay, A. Arneodo, P. Borgnat, B. Audit, 'Multi-scale structural community organisation of the human genome', *BMC Bioinformatics* **2017**, *18*, 1–13.

[32] R. Akbani, P. Ng, H. Werner, M. Shahmoradgoli, F. Zhang, Z. Ju, W. Liu, J. Yang, K. Yoshihara, J. Li, et al., 'A pan-cancer proteomic perspective on the cancer genome atlas', *Nature Communications* **2014**, *5*, 1–15.

[33] R. Akbani, K. Becker, N. Carragher, T. Goldstein, L. Koning, U. Korf, L. Liotta, G. Mills, S. Nishizuka, M. Pawlak, et al., 'Realizing the promise of reverse phase protein arrays for clinical, translational, and basic research: a workshop report: the RPPA (reverse phase protein array) society', *Molecular & Cellular Proteomics* **2014**, *13*, 1625–1643.

[34] K. Tomczak, P. Czerwinska, M. Wiznerowicz, 'The cancer genome atlas (TCGA): an immeasurable source of knowledge', *Contemporary Oncology* **2015**, *19*, 68-77.

[35] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. Margolin, S. Kim, C. Wilson, J. Lehár, G. Kryukov, D. Sonkin, et al., 'The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature* **2012**, *483*, 603-607.

[36] H. Gonçalves Jr, M. Guerra, J. Cintra, V. Fayer, I. Brum, M. Teixeira, 'Survival study of triple-negative and non–triple-negative breast cancer in a Brazilian cohort', *Clinical Medicine Insights: Oncology* **2018**, *12*.

[37] K. Chavez, S. Garimella, S. Lipkowitz, 'Triple negative breast cancer cell lines: one tool in the search for better treatment of triple negative breast cancer', *Breast Disease* **2010**, *32*, 35-48.

[38] K. Subik, J. Lee, L. Baxter, T. Strzepek, D. Costello, P. Crowley, L. Xing, M. Hung, T. Bonfiglio, D. Hicks, et al., 'The expression patterns of ER, PR, HER2, CK5/6, EGFR, Ki-67 and AR by immunohistochemical analysis in breast cancer cell lines', *Breast Cancer: Basic and Clinical Research* **2010**, *4*, 35-41.

[39] A. Michmerhuizen, D. Spratt, L. Pierce, C. Speers, 'Are we there yet? understanding androgen receptor signaling in breast cancer', *NPJ Breast Cancer* **2020**, *6*, 1–19.

[40] V. Abramson, C. Arteaga, 'New strategies in HER2 - overexpressing breast cancer: many combinations of targeted drugs available', *Clinical Cancer Research* **2011**, *17*, 952–958.

[41] S. Nowsheen, T. Cooper, J. Bonner, A. LoBuglio, E. Yang, 'HER2 overexpression renders human breast cancers sensitive to PARP inhibition independently of any defect in homologous recombination DNA repair', *Cancer research* **2012**, *72*, 4796–4806.

[42] D. Wu, Y. Pang, M. Wilkerson, D. Wang, P. Hammerman, J. Liu, 'Gene-expression data integration to squamous cell lung cancer subtypes reveals drug sensitivity', *British Journal of Cancer* **2013**, *109*, 1599–1608.

[43] M. Wilkerson, X. Yin, K. Hoadley, Y. Liu, M. Hayward, C. Cabanski, K. Muldrew, C. Miller, S. Randell, M. Socinski, et al., 'Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types', *Clinical Cancer Research* **2010**, *16*, 4864–4875.

[44] Y. Xu, P. Gao, X. Lv, L. Zhang, J. Zhang, 'The role of the ataxia telangiectasia mutated gene in lung cancer: recent advances in research', *Therapeutic Advances in Respiratory Disease* **2017**, *11*, 375–380.

[45] C. Li, H. Lu, 'Adenosquamous carcinoma of the lung', *OncoTargets and Therapy* **2018**, *11*, 4829-4835.

[46] A. Weber, N. Drobnitzky, A. Devery, S. Bokobza, R. Adams, T. Maughan, A. Ryan, 'Phenotypic consequences of somatic mutations in the ataxia-telangiectasia mutated gene in non-small cell lung cancer', *Oncotarget* **2016**, *7*, 60807-60822.

[47] I. Wistuba, D. Bryant, C. Behrens, S. Milchgrub, A. Virmani, R. Ashfaq, J. Minna, A. Gazdar, 'Comparison of features of human lung cancer cell lines and their corresponding tumors', *Clinical Cancer Research* **1999**, *5*, 991–1000.

[48] S. Habib, L. Leifer, M. Azam, A. Siddiqui, K. Rajdev, M. Chalhoub, 'Giant cell carcinoma of the lung successfully treated with surgical resection and adjuvant vinorelbine and cisplatin', *Respiratory Medicine Case Reports* **2018**, *25*, 300–302.

[49] J. Ronen, S. Hayat, A. Akalin, 'Evaluation of colorectal cancer subtypes and cell lines using deep learning', *Life Science Alliance* **2019**, *2*.

[50] M. El-Bahrawy, S. Poulsom, A. Rowan, I. Tomlinson, M. Alison, 'Characterization of the e-cadherin/catenin complex in colorectal carcinoma cell lines', *International Journal of Experimental Pathology* **2004**, *85*, 65–74.

[51] D. Ahmed, P. Eide, I. Eilertsen, S. Danielsen, M. Eknaes, M. Hektoen, G. Lind, R. Lothe, 'Epigenetic and genetic features of 24 colon cancer cell lines', *Oncogenesis* **2013**, *2*, *71*.

[52] K. Thanki, M. Nicholls, A. Gajjar, A. Senagore, S. Qiu, C. Szabo, M. Hellmich, C. Chao, 'Consensus molecular subtypes of colorectal cancer and their clinical implications', *International Biological and Biomedical Journal* **2017**, *3*, 105-111.

[53] D. Mouradov, C. Sloggett, R. Jorissen, C. Love, S. Li, A. Burgess, D. Arango, R. Strausberg, D. Buchanan, S. Wormald, et al., 'Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer', *Cancer Research* **2014**, *74*, 3238–3247.

[54] R. Pal, N. Wei, N. Song, S. Wu, R. Kim, Y. Wang, P. Gavin, P. Lucas, A. Srinivasan, C. Allegra, et al., 'Molecular subtypes of colorectal cancer in pre-clinical models show differential response to targeted therapies: Treatment implications beyond kRAS mutations', *PloS One* **2018**, *13*.

[55] Y. Shen, X. Li, D. Dong, B. Zhang, Y. Xue, P. Shang, 'Transferrin receptor 1 in cancer: a new sight for cancer therapy', *American Journal of Cancer Research* **2018**, *8*, 916-931.

[56] D. Li, F. Bi, N. Chen, J. Cao, W. Sun, Y. Zhou, C. Li, Q. Yang, 'A novel crosstalk between BRCA1 and poly (adp-ribose) polymerase 1 in breast cancer', *Cell Cycle* **2014**, *13*, 3442–3449.

[57] L. Shan, X. Li, L. Liu, X. Ding, Q. Wang, Y. Zheng, Y. Duan, C. Xuan, Y. Wang, F. Yang, et al., 'GATA3 cooperates with PARP1 to regulate CCND1 transcription through modulating histone H1 incorporation', *Oncogene* **2014**, *33*, 3205–3216.

[58] M. Takaku, S. Grimm, and P. Wade, 'GATA3 in breast cancer: tumor suppressor or oncogene?', *Gene Expression The Journal of Liver Research* **2015**, *16*, 163–168.

[59] J. Lee, Y. Park, N. Oh, K. Kwack, K. Park, 'A transcriptional complex composed of ER (α), GATA3, FOXA1 and ELL3 regulates Il-20 expression in breast cancer cells', *Oncotarget* **2017**, *8*, 42752-42760.

[60] N. Jagadish, S. Agarwal, N. Gupta, R. Fatima, S. Devi, V. Kumar, V. Suri, R. Kumar, V. Suri, T. Sadasukhi, et al., 'Heat shock protein 70-2 (HSP70-2) overexpression in breast cancer', *Journal of Experimental & Clinical Cancer Research* **2016**, *35*, 1–14.

[61] L. Zhang, N. Tran, H. Su, R. Wang, Y. Lu, H. Tang, S. Aoyagi, A. Guo, A. Khodadadi-Jamayran, D. Zhou, et al., 'Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing', *Elife* **2015**, *4*.

[62] C. Xu, X. Wang, W. Li, K. Wang, L. Ding, 'Expression and clinical significance of Claudin-7 in patients with colorectal cancer', *Technology in Cancer Research & Treatment* **2018**, *17*.

[63] F. Greig, G. Nixon, 'Phosphoprotein enriched in astrocytes (PEA)-15: A potential therapeutic target in multiple disease states' *Pharmacology & Therapeutics* **2014**, *143*, 265–274.

[64] J. Xie, B. Li, H. Yu, Q. Gao, W. Li, H. Wu, Z. Qin, 'TIGAR has a dual role in cancer cell survival through regulating apoptosis and autophagy', *Cancer Research* **2014**, *74*, 5127–5138.

[65] A. Du, Y. Jiang, C. Fan, 'NDRG1 downregulates ATF3 and inhibits cisplatin-induced cytotoxicity in lung cancer a549 cells', *International Journal of Medical Sciences* **2018**, *15*, 1502-1507.