

RESEARCH

Open Access



Topological integration of RPPA proteomic data with multi-omics data for survival prediction in breast cancer via pathway activity inference

Tae Rim Kim¹, Hyun-Hwan Jeong^{2,3} and Kyung-Ah Sohn^{1*}

From The 8th Annual Translational Bioinformatics Conference
Seoul, South Korea. 31 October - 2 November 2018

Abstract

Background: The analysis of integrated multi-omics data enables the identification of disease-related biomarkers that cannot be identified from a single omics profile. Although protein-level data reflects the cellular status of cancer tissue more directly than gene-level data, past studies have mainly focused on multi-omics integration using gene-level data as opposed to protein-level data. However, the use of protein-level data (such as mass spectrometry) in multi-omics integration has some limitations. For example, the correlation between the characteristics of gene-level data (such as mRNA) and protein-level data is weak, and it is difficult to detect low-abundance signaling proteins that are used to target cancer. The reverse phase protein array (RPPA) is a highly sensitive antibody-based quantification method for signaling proteins. However, the number of protein features in RPPA data is extremely low compared to the number of gene features in gene-level data. In this study, we present a new method for integrating RPPA profiles with RNA-Seq and DNA methylation profiles for survival prediction based on the integrative directed random walk (iDRW) framework proposed in our previous study. In the iDRW framework, each omics profile is merged into a single pathway profile that reflects the topological information of the pathway. In order to address the sparsity of RPPA profiles, we employ the random walk with restart (RWR) approach on the pathway network.

Results: Our model was validated using survival prediction analysis for a breast cancer dataset from The Cancer Genome Atlas. Our proposed model exhibited improved performance compared with other methods that utilize pathway information and also out-performed models that did not include the RPPA data utilized in our study. The risk pathways identified for breast cancer in this study were closely related to well-known breast cancer risk pathways.

Conclusions: Our results indicated that RPPA data is useful for survival prediction for breast cancer patients under our framework. We also observed that iDRW effectively integrates RNA-Seq, DNA methylation, and RPPA profiles, while variation in the composition of the omics data can affect both prediction performance and risk pathway identification. These results suggest that omics data composition is a critical parameter for iDRW.

Keywords: Multi-omics data, Integrative analysis, Random walk, Reverse phase protein Array, Pathway-based analysis, Network propagation, Breast cancer, Survival prediction

* Correspondence: kasohn@ajou.ac.kr

¹Department of Computer Engineering, Ajou University, Suwon 16499, South Korea

Full list of author information is available at the end of the article



Background

Advances in high-throughput sequencing technologies and their integration, including genome, transcriptome, epigenome, and proteome sequencing, has shifted perspectives from the micro-level to the macro-level in biological research. Certain phenomena can be observed in each omics layer, enabling researchers to understand the complex interactions within and between biological mechanisms. Several studies have proposed methods for integrating omics data obtained from high-throughput sequencing in different layers to provide insights into systems biology [1–4]. In cancer research, integrative models for multi-omics data not only greatly improve clinical prognosis predictions but also allow cancer-related biomarkers to be identified [5–12]. For example, Kim et al. proposed a graph-based framework for integrating multi-omics data (including CNA, DNA methylation, miRNA, and gene expression data) with prior knowledge to improve the clinical outcome prediction performance for glioblastoma multiformes and serous cystadenocarcinoma [13]. In addition, Bertrand et al. developed a patient-specific data integration framework (OncoIMPACT) to identify driver genes using the scored impact of single nucleotide polymorphism (SNP), indel, and copy number variation (CNV) data for individual patients with one of five different cancer types [14]. Their prediction of patient-specific driver genes using OncoIMPACT was validated with *in silico* and *in vitro* experiments, with their model proving to be more robust and precise than the baseline model.

Although many previous studies have focused on the integration of multi-omics data, proteomic data is rarely used for integrative multi-omics data analysis. Proteins are a fundamental unit of a biological complex, in which they have a functional role, and protein expression data can be utilized for the diagnostic prognosis of cancer patients [15–18]. However, combining protein-level and gene-level data is challenging because the relationship between the two data types is unclear [19]. Not all genomic variation is translated into proteins via perturbations (e.g. post-transcriptional regulation or post-translational regulation and modification) [20, 21], thus gene-expression data may not accurately reflect active cellular function [22, 23]. Furthermore, other limitations, such as mass spectrometry-based technology for the quantitative analysis of proteins, which exhibits low sensitivity for low-abundance proteins, make it difficult to collect information on cancer-related signaling proteins [24].

The reverse phase protein array (RPPA) is an antibody-based protein assay platform for high-throughput sequencing that quantifies the expression of a target protein. It is cost-effective and highly sensitive to the target protein, even at low concentrations [24]. However, it greatly relies on the quality of the antibodies and requires

screening to select the appropriate antibodies for the corresponding target proteins. Although RPPAs face these limitations, several studies have adopted the RPPA platform to produce proteomic data for the detection of cancer tissue phenotypes. Additionally, The Cancer Genome Atlas (TCGA) has also generated and made public RPPA data for TCGA cancer samples for use in characterizing various cancer types. This data is publicly accessible in The Cancer Proteomics Atlas (TCPA) [25].

Pathway-level analysis is necessary because it is not possible to define biological functions via a single gene or a single molecule. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [26] provides pathway information that represents a functional network consisting of a set of gene products (such as proteins and functional RNA) and their relationship. Based on this pathway information, several previous studies have utilized gene- or protein-level expression data to infer pathway activity [27–29]. These studies have shown that pathway-level information provides a better description of disease phenotypes than does gene-level information. Furthermore, previous research employing directed random walk on a gene-metabolite graph (DRW-GM) [30] and integrative directed random walk (iDRW) [31] has proposed a method for the integration of multi-omics into pathway information based on [27]. DRW-GM involves the pathway-based integration of gene expression and metabolite data, while iDRW, which was proposed in our previous study, utilizes the pathway-based integration of RNA-Seq and DNA methylation profiles. Both methods demonstrate improved performance and reveal that the integration of multi-omics data into pathway information is useful for predicting disease phenotypes and identifying risk pathways.

In this study, we propose a pathway-based integration method for RPPAs and other omics data. Specifically, we focus on the utilization of RPPA data in which the proteins are sparse with respect to the pathway gene set. In our experiments, we merge RNA-Seq, DNA methylation, and RPPA profiles into a single pathway profile that is employed for survival prediction and risk-pathway identification in breast cancer.

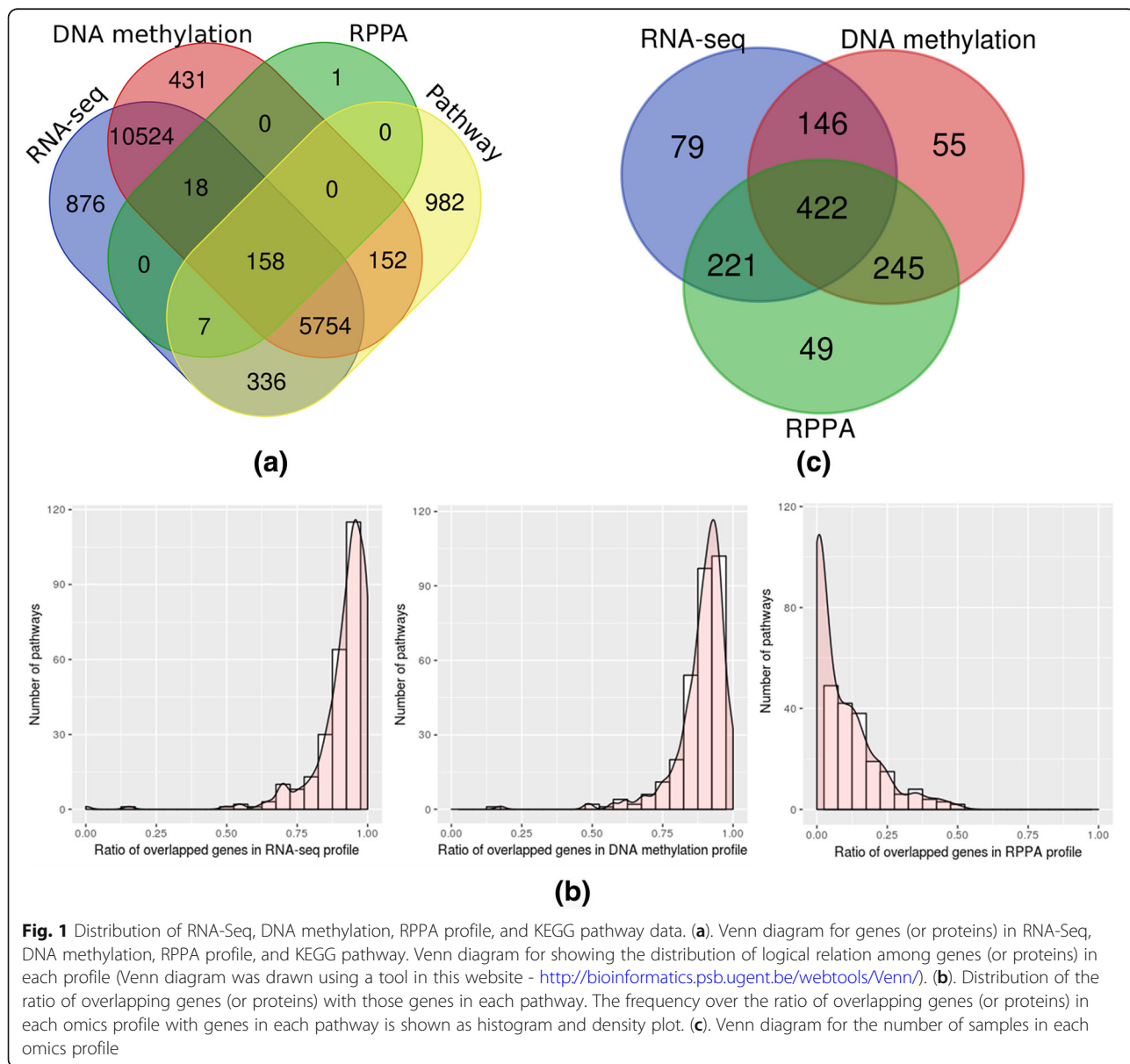
Methods

Data

We downloaded pathway annotation files from KEGG [32]. Of these pathways, we selected 327 human pathways containing 7389 gene features. Level 3 multi-omics profiles for breast cancer were collected from the TCGA breast cancer dataset from the Broad Institute GDAC Firehose [33]. We employed RNA-Seq, DNA methylation, and RPPA profiles as transcriptomic, epigenomic, and proteomic data, respectively. The RNA-Seq profile

contained 17,673 genes from 869 samples, the DNA methylation profile 17,037 genes from 868 samples, and the RPPA profile 188 proteins from 937 samples. For missing values, we imputed the median of the corresponding gene, protein, or methylation. We analyzed the common features of the pathways and each omics profile, as shown in Fig. 1a. In order to assess the level of sparsity in the RPPA profile, we calculated the ratio of matched genes in each omics profile with the genes in each pathway. As shown in Fig. 1b, the RNA-Seq and DNA methylation profiles included almost a full set of genes for each pathway, while the RPPA profile contained only a few matched proteins for each pathway. The 422 samples found in common in all three profiles

were extracted (Fig. 1c). They contained clinical information on vital status and survival period. We excluded samples for which the survival period was missing or negative. Those samples whose vital status was reported as 1 (living) but for whom the survival period was less than three years were also excluded, leaving a total of 376 extracted samples. For survival analysis, we classified each patient into one of two groups: patients whose survival exceeded three years were placed in the long-term survival group, while those with a survival period of less than three years were placed in the short-term survival group [34]. Of the 376 extracted samples, 177 exhibited long-term survival (≥ 3 years) and 199 exhibited short-term survival (< 3 years).



Overview of the proposed framework

In order to integrate the RPPA profile with other multi-omics data and merge them into a single pathway profile, we deployed and extended the iDRW framework [31]. The overall process of the proposed framework is illustrated in Fig. 2. Our proposed framework included four steps, which are explained briefly below.

- (1). Statistical testing (*t*-tests or DESeq2) was conducted to obtain initial gene weights that indicate the extent to which a gene (or protein) differentiates between long-term and short-term survival for breast cancer.
- (2). A unified pathway network was constructed for the iDRW method that reflected the topological information of the pathway. Each gene (or protein) weight vector was assigned to the corresponding pathway network. Given that the RPPA profile contains few protein expressions (approximately 180 proteins) and that there is only a very small

overlap between the genes corresponding to proteins in the RPPA profile and the genes in the pathway, many of the initial weights for the RPPA profile were missing or assigned as zero. We conducted random walk with restart (RWR), which represents a network propagation algorithm, on the protein-level pathway network to estimate the initial weight of the missing proteins.

- (3). iDRW was utilized to integrate the initial weight of each gene (or protein) into the unified pathway network. In the iDRW approach, the initial gene (or protein) weights are mixed to reflect the topological information of the pathway. Thus, we obtained a final weight vector that represented the topological importance of each gene in terms of distinguishing between the long-term and short-term survival groups.
- (4). A pathway profile using the pathway activity inference method derived from [27] was constructed. Given the pathway profile, we

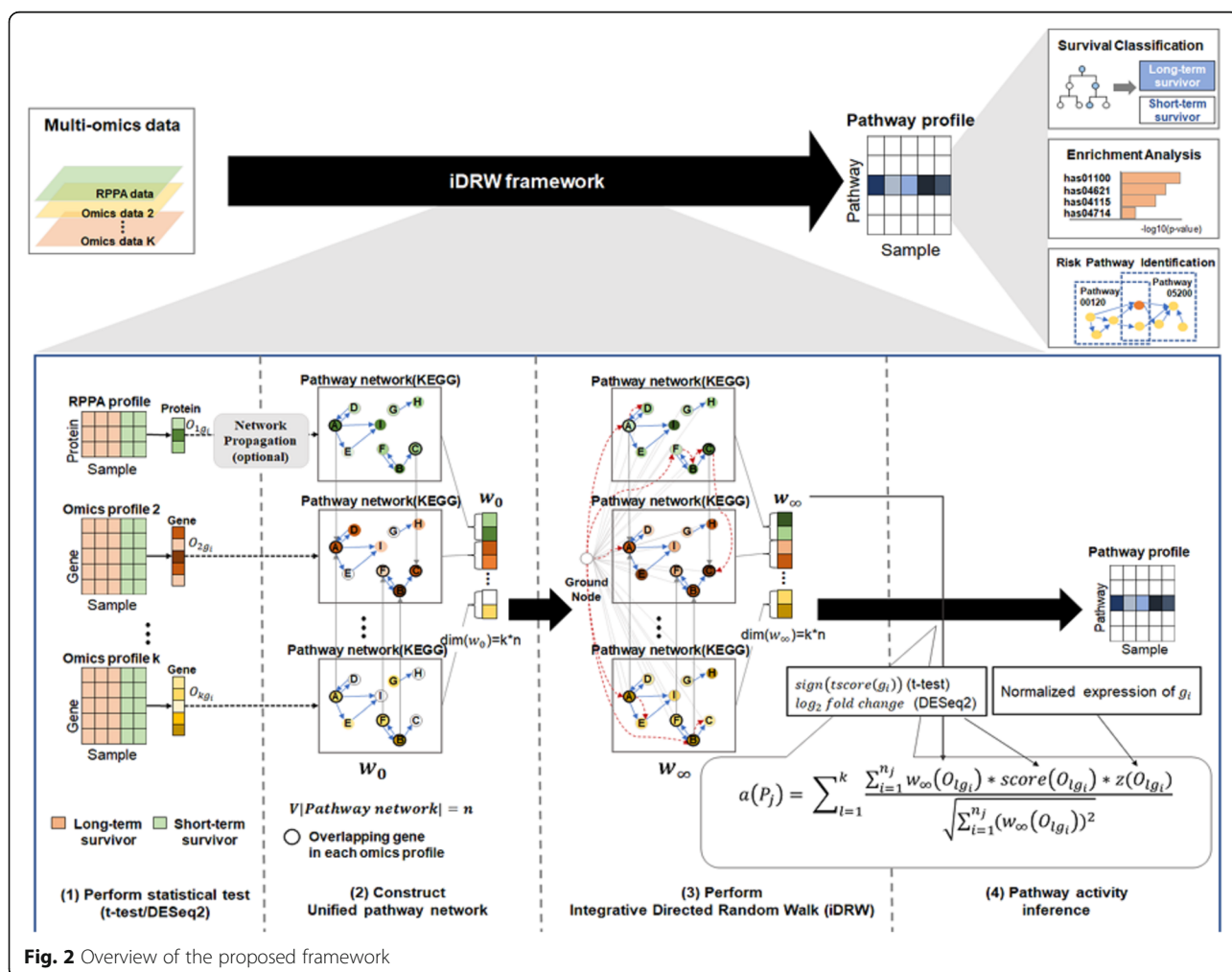


Fig. 2 Overview of the proposed framework

performed survival prediction using a random forest classifier and identified risk pathways for breast cancer. Additionally, in this step, we investigated the contribution of different combinations of each omics type.

A more detailed description of these steps is given in the following sections.

Unified pathway network construction

The iDRW framework operates on a unified pathway network that consists of a set of pathway networks corresponding to the number of omics types. We first constructed pathway networks for each of the RNA-Seq, DNA methylation, and RPPA profiles and combined them into a unified pathway network. A pathway network for a single omics profile contains relational information for biological molecules (such as genes, gene products, or chemical compounds) from the KEGG database [32]. We used genes or gene products as nodes for the pathway network and their relationships as edges. We used the KEGGgraph R package, which facilitates the conversion of KEGG pathway information into network objects [35]. For the 327 pathways that had the pathway id prefix *hsa*, a pathway network was generated containing a total of 7389 nodes and 58,399 directed edges. In this paper, we refer to the generated pathway networks as the transcriptome network, epigenome network, and proteome network for the RNA-Seq, DNA methylation, and RPPA profiles, respectively.

To generate the unified pathway network, it was necessary to connect the three pathway networks. Figure 3 presents the structure of the unified pathway network. It should be noted that the structure was designed to obtain better-integrated gene or protein scores rather than reflecting biological relationships. A detailed description

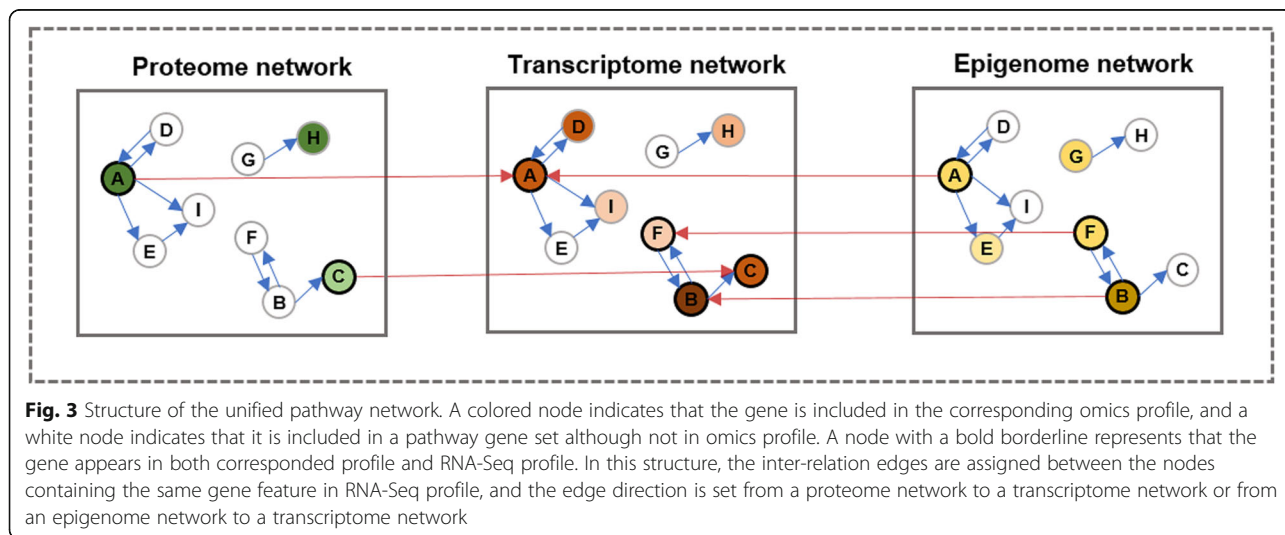
on the same is given in the following sections. As shown in Fig. 3, each node represents a gene in a pathway network. The inter-relationship edges are assigned to genes that overlap two different omics profiles on the corresponding pathway networks. For example, the RNA-Seq profile contains the gene *AKT1*, while the RPPA profile contains an antibody corresponding to the AKT1 protein; thus, a directional edge is assigned to the node for AKT1 in the transcriptome network and the node for AKT1 in the proteome network. In this structure, we assign directional edges to the overlapping genes (with the inter-network edge following the direction of epigenome network → transcriptome network and proteome network → transcriptome network). We selected this structure for the unified pathway network empirically based on the best performance in survival classification for breast cancer (See Additional file 1: Supplementary Material 1 for more information about the network structures used in this experiment).

Topological integration of multi-omics profiles based on iDRW

To apply the iDRW method to the unified pathway network, an initial weight vector was assigned to the network. The initial vector W_0 for each omics profile was defined as follows:

$$W_0 = -\log(w_g + \epsilon)$$

where w_g denotes the weight of the gene g in the unified pathway network and $\epsilon = 2.2 * 10^{-16}$. Additionally, w_{g-} denotes the min-max normalized p -value of gene g obtained using a different statistical method from an omics profile. We performed a two-sided t -test for the DNA methylation and RPPA profiles, while the



differential expression test DESeq2 [36] was employed for the RNA-Seq profile.

In the iDRW method, a random walker starts at the ground node and then moves onto a randomly selected neighbor node or returns to the ground node with restart probability r at each time step t . The process is defined as follows:

$$W_{t+1} = (1-r)M^T W_t + rW_0$$

where W_t denotes the weight vector that represents the probability of being at each node at time step t ; M denotes a row-normalized adjacency matrix of the unified pathway network; r is the restart probability for the random walker; W_0 is the initial weight vector; and W_t is updated at each time step and converges to steady-state W_∞ when $|W_{t+1} - W_t| \cdot < \cdot 10^{-10}$ as guaranteed by [37]. We performed parameter tuning for r in the range of [0.2, 0.4, 0.6, 0.8].

Estimation of missing values in the RPPA profile using a network propagation algorithm

Because the W_0 of the RPPA profile is extremely sparse, we employed a network propagation algorithm to estimate the putative expression levels of the missing proteins in the pathway via the diffuser package (<https://github.com/dirmeier/diffuser>). With diffuser, RWR was used for missing value estimation. r_i^t denotes the propagated weight vector at time step t at node i , and is obtained by RWR [38] as follows:

$$r_i^{t+1} = pAr_i^t + (1-p)e_i$$

where A denotes the column-normalized adjacency matrix of a given network; e_i is the W_0 for the RPPA profile at node i ; and p is the restart probability, which controls how much of the local topological information is reflected when the random walk converges to its steady state. Thus, the higher the restart probability, the more local the topological information, and the lower the restart probability, the more global the information. p was tuned by a grid search on [0.2, 0.4, 0.6, 0.8]. At each time step, r_i^{t+1} was updated, and the steady-state weight vector r_i^∞ was obtained from a fixed number of iterations. r_i^∞ was used as W_0 for the RPPA profile, and the W_0 for each omics profile was assigned to the corresponding pathway network.

Pathway activity inference

Pathway activity inference using a gene profile was conducted using the method suggested by [27]. To infer the pathway activity using a multi-omics profile, we redefined the general inference method for the activity of the j -th pathway as

$$\alpha(P_j) = \sum_{l=1}^k \frac{\sum_{i=1}^{n_j} w_\infty(O_{lgi}) * score(O_{lgi}) * z(O_{lgi})}{\sqrt{\sum_{i=1}^{n_j} (W_\infty(O_{lgi}))^2}} \tag{1}$$

where P_j , which corresponds to the j -th pathway, contains n_j differentially expressed genes (g_1, g_2, \dots, g_{n_j}) in which the p -value (w_g) is < 0.05 ; k is the number of omics profiles employed; O_l represents the l -th omics profile used when inferring the pathway activity; O_{lgi} represents gene g_i in O_l ; $W_\infty(O_{lgi})$ is the weight of g_i calculated in O_l ; and $z(O_{lgi})$ is the normalized value for the expression of g_i in O_l . In addition, in the RNA-Seq profile, which consists of count-based data, $score(O_{lgi})$ is the *log2fold change* from the DESeq2 method [36] for g_i in O_l . In the DNA methylation and RPPA profiles, $score(O_{lgi})$ is the *sign(tscore(O_{lgi}))* from a two-tailed t -test. This process generates the pathway activity profile P_j , which is used as an input feature for survival classification.

To investigate the contribution of each omics type to survival prediction within the same model, we experimented with variants of the aforementioned pathway activity inference formula. P denotes a power set (excluding the null set) of k omics profiles $\{O_1, O_2 \dots O_k\}$, which is used as the input profile. In this study, we used three omics profiles so that the power set of $\{O_1, O_2, O_3\}$ can be denoted as

$$P = \{\{O_1\}, \{O_2\}, \{O_3\}, \{O_1, O_2\}, \{O_1, O_3\}, \{O_2, O_3\}, \{O_1, O_2, O_3\}\}.$$

We denote the r -th subset of P as S_r . The refined pathway activity inference formula is as follows:

$$\alpha(P_j) = \sum_{l=1}^k \frac{\sum_{i=1}^{n_j} w_\infty(S_{rgi}) * score(S_{rgi}) * z(S_{rgi})}{\sqrt{\sum_{i=1}^{n_j} (W_\infty(S_{rgi}))^2}} \tag{2}$$

Eqs. 1 and 2 are almost identical except for the number of omics types used to infer the pathway activity. We used Eq. 2 to measure the predictive power of the combination of each omics type and determined the optimal combination for survival prediction. For example, the RNA-Seq, DNA methylation, and RPPA profiles were used as an input to produce *iDRW(GMP)* and the RNA-Seq and DNA methylation profiles were used to produce *iDRW(GM)*. When generating the pathway profile in *iDRW(GMP)* using RNA-Seq, DNA methylation, and RPPA data to infer the pathway activity score, O_l contains {RNA-seq, DNA methylation, RPPA} and $k = 3$ for Eq. 2. However, when RNA-Seq and RPPA data was used in *iDRW(GMP)* to infer pathway activity, O_l contained {RNA-seq, RPPA} and $k = 2$ for Eqs. 1 and 2 was used for cases in which different omics combinations were used as input to calculate the pathway activity score.

Survival classification and evaluation

In our study, we conducted survival analysis using a binary classification of long-term survival (surviving more than three years) and short-term survival (surviving fewer than three years) using the pathway profiles. In our dataset, 177 samples exhibited long-term survival and 199 exhibited short-term survival. Using the pathway activity profile, we first extracted the top- N pathways from among the 327 pathways that exhibited the best performance for survival classification to obtain an optimal pathway list that significantly differentiated between the long-term and short-term survival groups. To achieve this, we sorted all pathways by increasing p -value from the two-tailed t -tests for the pathway activity (i.e., pathways with a lower p -value were ranked higher). Based on this ranking, we selected the top k pathways, and the model was then evaluated using 5-fold cross validation (with the caret R package [39]) via a random forest classifier (with the randomForest [40] R package) and varying $k = 5, 10, \dots$, for half of the total pathways. The procedure was repeated ten times for reliability. With the top- N pathways chosen, we performed leave-one-out cross-validation (LOOCV) using the caret R package [39] in a dataset with n observations via a random forest classifier for survival classification. Random forest is composed of several decision trees. A decision tree makes decision rules that enable a correct decision for the target label. A random forest classifier can be used for non-linear datasets and is also robust to overfitting [41].

Results and discussion

Integrative analysis utilizing the RPPA profile to achieve accurate survival prediction

The pathway-based prediction model created using the RPPA profile exhibited significantly lower accuracy than that created using the RNA-Seq profile or the DNA methylation profile (Fig. 4a). We suspect that this is due to the sparsity of RPPA proteins. To verify this, we developed an alternative experimental setting by filtering the RNA-Seq and DNA methylation profiles to include only genes that overlapped with the corresponding RPPA proteins. As a result, the RNA-Seq profile included 183 genes and the DNA methylation profile included 176 genes (Fig. 1a). $iDRW(G^R M^R)$ denotes the pathway-based integration model obtained using the reduced RNA-Seq and DNA methylation profiles for survival classification. The addition of the RPPA profile to this model is denoted as $iDRW(G^R M^R P)$. As shown in Fig. 4b, the addition of the RPPA profile improved the survival classification performance. The $iDRW_{prop}(G^R M^R P)$ model, which utilized the propagated proteome network, produced an accuracy exceeding that of the $iDRW(G^R M^R P)$ model.

We also compared the performance of $iDRW(GMP)$ with $iDRW(GM)$ that is our previously proposed model (Fig. 4c). Compared with $iDRW(GM)$, $iDRW(GMP)$ had a greater survival prediction accuracy. This indicates that, although the RPPA profile does not exhibit a high accuracy when used alone, it can be used to effectively discriminate between the long-term and short-term survival groups when employed with other omics profiles to create a pathway-based prediction model. To illustrate the overall prediction performance for the survival period, we generated two additional survival curves using the survival [42] R package for $iDRW(GM)$ and $iDRW(GMP)$ (see Additional file 2: Supplementary Material 2a and b). The difference between the survival curves for the long-term and short-term group was statistically significant based on chi-square tests for both models ($p = 2e-07$ and $p = 5e-13$, respectively). Based on the much lower p -value for $iDRW(GMP)$, it is clear that the RPPA data significantly improves survival classification performance for TCGA breast cancer data. In addition, it can be concluded that network propagation using the proteome network is a reasonable approach for addressing the sparsity of the RPPA profile.

Predictive power of different omics type combinations when inferring pathway activity scores

To investigate the effect of different combinations of omics profiles on survival prediction, we conducted pathway activity inference with different combinations of omics profiles. As shown in Fig. 2, the experiment was performed using $iDRW(GMP)$ until Step (3), and Eq. 2 was used on different combinations only in Step (4). Prediction accuracy was compared by varying γ (Fig. 5). Interestingly, the $(G + P)$ combination, which used information from the RNA-Seq and RPPA profiles in pathway activity inference, exhibited superior performance in survival prediction, while (P) , which used information from the RPPA profile only, exhibited the worst performance. $(G + M + P)$, which used information from all of the three profiles to infer pathway activity, had the best accuracy when $\gamma = 0.4$. However, $(G + P)$ generally exhibited a more stable and higher overall accuracy and was robust with respect to γ .

Performance comparison of pathway-based integration models for survival classification

Based on the aforementioned results, we compared each pathway-based integration model by varying the restart probability γ for the $iDRW$ method (Fig. 6). Unlike the aforementioned experiments, the omics profiles used as inputs and the omics profiles used to infer the pathway activity were identical. As shown in Fig. 6, $iDRW(GP)$ produced the best performance for survival classification when $\gamma = 0.6$. However, $iDRW_{prop}(GP)$ exhibited both

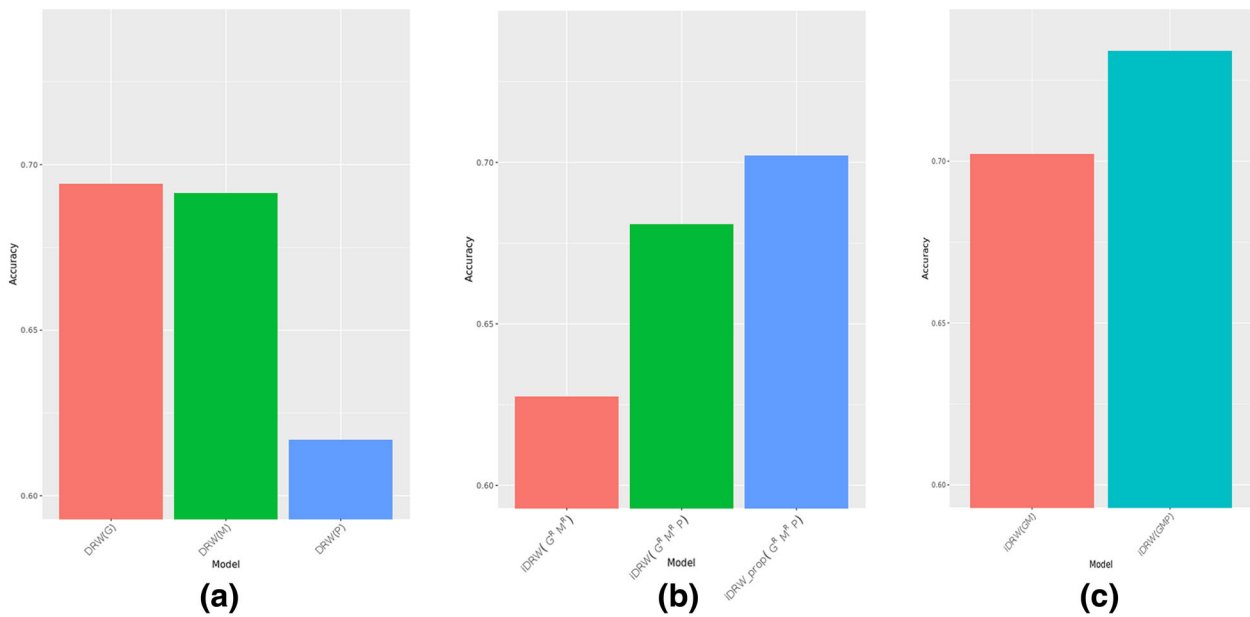


Fig. 4 Performance comparison between different methods and profiles. (a). In case of using a single omics profile. *DRW(G)* used RNA-Seq profile; *DRW(M)* used DNA methylation profile; *DRW(P)* used RPPA profile. (b). In case of using reduced RNA-Seq and DNA methylation profile. Each profile was reduced to include genes overlapping with RPPA proteins. *iDRW(G^RM^R)* used reduced RNA-Seq and reduced DNA methylation profile; *iDRW(G^RM^RP)* used reduced RNA-Seq, reduced DNA methylation, and RPPA profile; *iDRW_{prop}(G^RM^RP)* performed network propagation using RWR on the proteome network. (c). Performance comparison of *iDRW(GM)* and *iDRW(GMP)*. *iDRW(GM)* is a previous method which used RNA-seq and DNA methylation profile. *iDRW(GMP)* is our proposed model which used RNA-seq, DNA methylation, and RPPA profile in this study

stable and higher performance. For the *iDRW_{prop}* model, the optimized restart probability for proteome network propagation was used for each γ in [0.2, 0.4, 0.6, 0.8] to compare the performance of all models with respect to γ (Fig. 6). All models using the RPPA profile (*P*) exhibited

better performance when compared with the *iDRW(GM)* model presented in our previous study. Interestingly, the *iDRW_{prop}* model had a slightly lower performance (though it was more stable) with changes in γ than that of the *iDRW* model. This indicates that the *iDRW_{prop}*

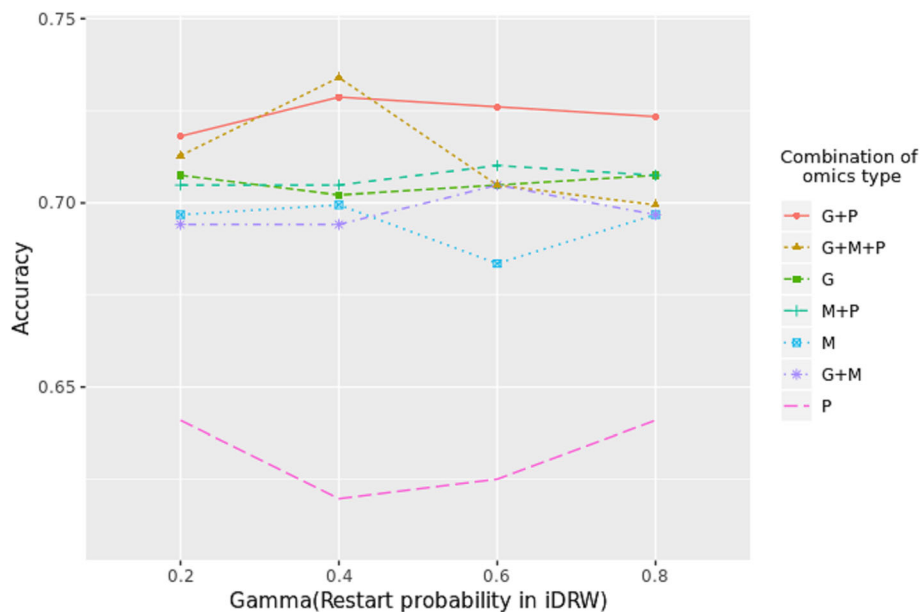


Fig. 5 Classification accuracy using different combination of each omics type in pathway activity score calculation. Each case in legend means the combination of omics profiles which was used to calculate the pathway activity score. All cases in this experiment were originating from *iDRW(GMP)* model (status before pathway activity inference step) with varying γ

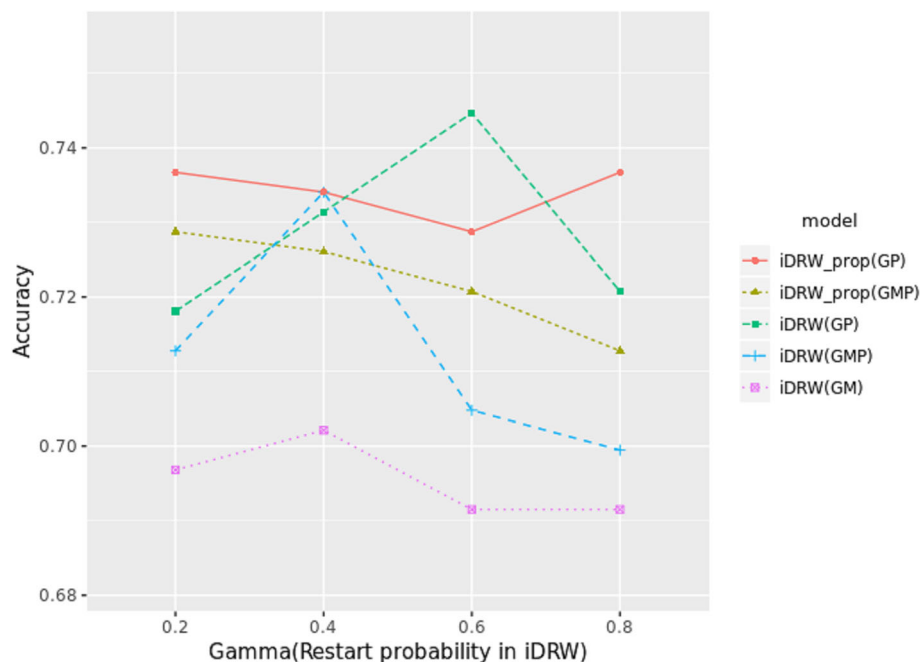


Fig. 6 Performance comparison of pathway-based integration model with varying γ

model is more robust to changes in the restart probability than is *iDRW*, and that network propagation mitigates the sparsity problem in RPPA data. Based on the optimized models obtained from the experiment presented in Fig. 6, we measured the final performance of all models for survival classification. As baseline models, we considered the mean, median, and concat models. The mean and median models computed either the mean or median of the normalized expression values of the pathway gene members in order to construct a pathway profile. The concat model was constructed using the simple concatenation of the pathway profiles obtained from the RNA-Seq, DNA methylation, and RPPA profiles in order to demonstrate the utility of our *iDRW* framework. The others were generated using an *iDRW*-based method that enables the integration of each omics profile in a pathway profile on the unified pathway network. As can be observed in Fig. 7, *iDRW(GP)* exhibited the best performance for survival classification. All models created using the RPPA profile out-performed models without the RPPA profile. As a result, it is clear that RPPA data is useful for the prediction of long-term or short-term patient survival.

Identification of risk pathways and systemic analysis of *iDRW(GP)* and *iDRW(GMP)*

We identified the top 10 risk pathways from the ranked pathway list for *iDRW(GP)* and *iDRW(GMP)*. The ranked pathway list was obtained from the p -values for two-tailed t -tests of the pathway activity

with respect to long-term and short-term survival. The importance score was calculated based on prediction error using the out-of-bag (OOB) estimate method. OOB estimation is the mean prediction error of random forest without training sample x_i . The importance of a feature was defined by the degree to which the prediction error increased when x_i was excluded. The score was scaled from 0 to 100 with the caret R package [39].

Table 1 shows the top 10 risk pathways for *iDRW(GP)*. It can be observed that the list contains several pathways that have been previously reported to have a direct association with cancer. The *p53 signaling pathway* is a well-known anti-cancer pathway that plays a role in apoptosis and as a tumor suppressor [43–45]. The pathway of *pathways in cancer* was also extracted [46]. Cellular senescence ceases cell division and suppresses premalignant cell proliferation [47]. It also exhibits age-related pathology and its abnormal function promotes cancer progression [47–50]. The metabolic pathway, which was the top-ranked risk pathway for *iDRW(GP)*, consists of a series of chemical reactions for cell metabolism, including the anabolic (storing energy) and catabolic (releasing energy) pathways. It is a general pathway that is not cancer-specific. Evidently, the activity of the metabolic pathway in critical-state patients (i.e., those with short-term survival) exhibits a pattern that is distinct from that of non-critical patients (i.e., those with long-term survival) [51]. The metabolic pathway functions based on the activity of biochemical

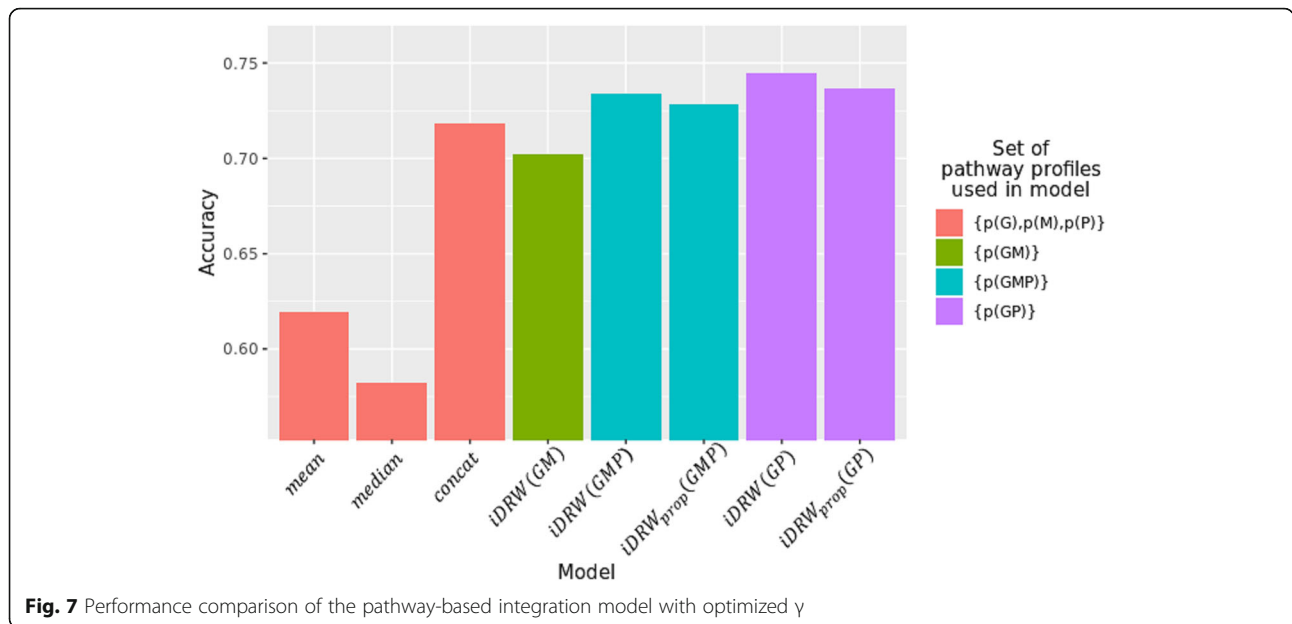


Fig. 7 Performance comparison of the pathway-based integration model with optimized γ

molecules, thus differentially expressed proteins do not appear in the metabolic pathway.

Table 2 displays the top 10 risk pathways for *iDRW(GMP)*. Several of these pathways were directly related to the immune system. For example, mitophagy is a selective autophagy process that maintains cell health via the degradation of damaged mitochondria [52–54]. Spliceosomes are a biological unit that facilitates the alternative splicing of pre-mRNA and regulates immune responses [55–57]. Viral inflammation-related pathways were also observed on the list, including systemic lupus erythematosus, toxoplasmosis, and epithelial cell signaling in *Helicobacter pylori* infections. These viral inflammation-related pathways are strongly associated with innate immune responses [58–63] and are commonly accompanied by mitochondrial DNA (mtDNA)

mutation [64–70], which is the DNA found in mitochondria that is maternally inherited [53]. Mitochondria are cellular organelles that have a number of roles in a cell, including producing cellular energy, controlling the cell cycle for cell growth and death, biosynthesis, and immunological responses [54]. Abnormal mitochondria function due to mtDNA mutation or depletion could be related to cancer progression. It has been reported that epigenetic modification (such as DNA methylation) controls the expression patterns of mtDNA [53, 71, 72]. Based on our analysis, we conjecture that abnormal methylation in mtDNA is associated with breast cancer.

To investigate the association of the risk pathways identified by *iDRW(GP)* and *iDRW(GMP)*, we created a pathway-pathway interaction network (Fig. 8) with the top 20 risk pathways from each model. The pathway

Table 1 Risk-active pathways identified *iDRW(GP)*

Pathway ID	Pathway name	Total ^a	DE genes	DE proteins	Importance score ^b
hsa01100	Metabolic pathways	1273	220	0	100.00
hsa04115	p53 signaling pathway	68	17	4	28.86
hsa04621	NOD-like receptor signaling pathway	168	30	6	27.44
hsa04218	Cellular senescence	160	30	16	26.07
hsa05203	Viral carcinogenesis	201	50	8	21.07
hsa04066	HIF-1 signaling pathway	100	27	11	20.75
hsa05200	Pathways in cancer	526	131	20	20.05
hsa04714	Thermogenesis	229	37	4	20.01
hsa05120	Epithelial cell signaling in <i>Helicobacter pylori</i> infection	68	17	3	19.19
hsa04926	Relaxin signaling pathway	130	30	9	17.97

^aTotal: the number of genes mapped to the pathway in the KEGG database

^bImportance score: the importance of a variable measured by out-of-bag (OOB) estimate and it was scaled in 0 to 100

Note that the number of differentially expressed genes (DE genes) and differentially expressed proteins (DE proteins) are also shown (*p*-value of DESeq2 or *t*-test < 0.05)

Table 2 Risk-active pathways identified *iDRW(GMP)*

Pathway ID	Pathway name	Total ^a	DE genes	DM genes	DE proteins	Importance score ^b
hsa04137	Mitophagy	65	10	12	3	100.00
hsa03040	Spliceosome	134	10	15	1	94.88
hsa05322	Systemic lupus erythematosus	133	24	15	0	90.26
hsa04218	Cellular senescence	160	30	27	16	87.13
hsa04974	Protein digestion and absorption	90	25	17	1	79.23
hsa04622	RIG-I-like receptor signaling pathway	70	10	5	3	79.02
hsa05145	Toxoplasmosis	113	34	11	10	78.72
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	68	17	11	3	72.59
hsa04621	NOD-like receptor signaling pathway	168	30	26	6	69.23
hsa05230	Central carbon metabolism in cancer	65	12	14	8	68.91

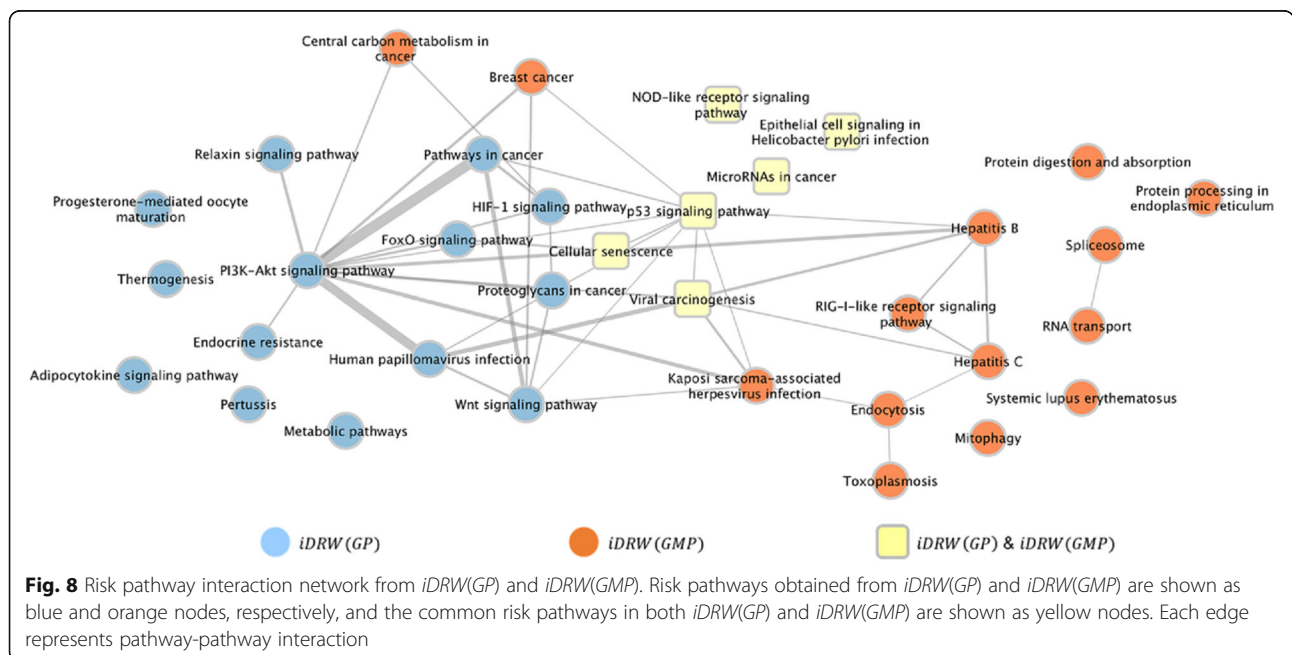
^aTotal: the number of genes mapped to the pathway in the KEGG database

^bImportance score: the importance of a variable measured by out-of-bag (OOB) estimate and it was scaled in 0 to 100

Note that the number of differentially expressed genes (DE genes), differentially methylated genes (DM genes), and differentially expressed proteins (DE proteins) are also shown (*p*-value of DESeq2 or *t*-test < 0.05)

interaction network was constructed using Pathway-Connector [73], which is a visualization tool for the direct connection among pathways based on public databases such as KEGG [74] and Reactome [75]. Figure 8 displays the common risk pathways selected by both *iDRW(GP)* and *iDRW(GMP)*. The cellular senescence pathway is a cancer suppressor which regulates cell growth and death by ceasing the division of pre-malignant or aged cells [47]. We examined the distribution of patients' ages to determine the relationship between age and long-term/short-term survival. We found that the average age of the long-term survival group was 6 years younger than that for the short-term survival group (*p*-value = 8.233e-05).

Viral carcinogenesis induces tumor progression via oncogenic virus infection [76]. The human papillomavirus (HPV) infection pathway, which is a DNA tumor virus infection, was identified in *iDRW(GP)* and has been reported to be closely related to breast cancer, with HPV DNA being found in breast cancer cells [77–79]. It is known that HPV targets tumor suppressor proteins [76]. The infection pathway for another oncovirus, Kaposi's sarcoma-associated herpesvirus (KSAH or HHV-8), was identified in *iDRW(GMP)*. It has been reported that KSAH is associated with breast cancer. However, KSAH is an etiologic factor for non-familial breast cancer [80]. The risk pathways identified by *iDRW(GP)* have a relatively strong connection with cancer-related signaling pathways,



such as the PI3K-Akt signaling pathway, Pathway in cancer, Proteoglycans in cancer, and the FoxO signaling pathway. The PI3K-Akt signaling pathway is frequently observed in cancer cells and stimulates cell growth and proliferation [81–83], while the FoxO signaling pathway is a tumor suppressor that regulates the genes in cellular physiological processes such as the cell cycle and apoptosis [84–86]. The risk pathways identified by *iDRW(GMP)* contain viral infection-related and immune response-related pathways. Though hepatitis B and hepatitis C are not directly related to breast cancer, they cause inflammation of liver tissue, which makes it difficult for cancer patients to receive chemotherapy [87]. This can negatively affect the clinical prognosis for breast cancer patients.

It is known that protein-level data reflects the status of cancer cells better than gene-level data [88]. In this study, we observed that the *iDRW(GP)* model mainly reflected the status of cancer cells at the cellular level, while the *iDRW(GMP)* model tended to reflect immune response-level information. From our observations, we can infer that the *iDRW(GP)* model reflects protein-level information more accurately when compared with *iDRW(GMP)*. As in [1], it is generally believed that increasing the volume of data for integration leads to further improvements in performance; we hypothesize that, based on the above results, the reason why *iDRW(GP)* out-performed *iDRW(GMP)* is that the characteristics of DNA methylation hinder the identification of risk pathways which facilitate the prediction of survivability for breast cancer patients that were otherwise discovered by *iDRW(GP)*. It should be noted that the benefits of different combinations of omics types will depend on the type of clinical problem under examination. Clinical predictions for survival are influenced by both genetic and environmental factors. For survival classification, protein-level information (such as vital signs) is more important than gene-level information (such as innate immune response information).

Conclusions

In this study, we combined RPPA proteomic data with RNA-Seq and DNA methylation data to successfully derive pathway information based on the iDRW framework. This study found that RPPA data is a rich source of information for survival prediction for breast cancer patients; when RPPA data was employed in the iDRW framework, improved performance was observed and feasible risk pathways extracted. The proposed model successfully identified both well-known and previously undiscovered risk pathways for breast cancer. Systemic analysis was also conducted to obtain better macroscopic insights. Furthermore, network propagation analysis and combinatorial

experiments were performed in order to fine-tune our model. We employed network propagation on pathway gene members to overcome the sparsity of RPPA proteins using random walk with restart (RWR). Although the *iDRW_{prop}* model did not out-perform *iDRW*, it was robust to the restart probability. The combinatorial experiments assessed the performance of each omics combination, with the (G + P) combination in *iDRW(GMP)* producing the highest accuracy overall. Finally, we observed that *iDRW(GP)* exhibited the best performance for survival prediction for breast cancer patients and highlighted key differences in the major risk pathways identified using *iDRW(GP)* and *iDRW(GMP)*. These findings highlight that an appropriate combination of omics data is required to properly address the topic under investigation.

Additional files

Additional file 1: Supplementary Material 1. Performance comparison on varying structure of the unified pathways network. (PNG 167 kb)

Additional file 2: Supplementary Material 2. Survival curve of *iDRW(GM)* and *iDRW(GMP)*. (a) Survival curve for long-term survival and short-term survival in *iDRW(GM)*. (b) Survival curve for long-term survival and short-term survival in *iDRW(GMP)*. (PNG 117 kb)

Abbreviations

DRW: Directed random walk; G: RNA-Seq profile; iDRW: Integrative directed random walk; iDRW(GM): The model that generates a pathway profile using the RNA-Seq and DNA methylation profiles; iDRW(GMP): The model that generates a pathway profile using the RNA-Seq, DNA methylation, and RPPA profiles; *iDRW_{prop}*: The model that adopted network propagation for the RPPA weight scores in the proteome network; M: DNA methylation profile; P: RPPA profile; RPPA: Reverse phase protein array; RWR: Random walk with restart

Acknowledgements

We gratefully acknowledge the TCGA Consortium and its members for the TCGA Project initiative, for providing samples, tissues, and data processing, and making data and results available. The results published here are in whole or part based upon data generated by TCGA pilot project established by the NCI and the NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at <http://cancergenome.nih.gov>.

Funding

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1B03933875) and the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2017M3C9A6047620). Publication costs are funded by NRF (NRF-2016R1D1A1B03933875) and Ajou university.

Availability of data and materials

The TCGA dataset, including the RNA-Seq, DNA methylation, and RPPA profiles used in this study were downloaded from BROAD GDAC Firehose (<https://gdac.broadinstitute.org/>).

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 5, 2019: Selected articles from the 8th Translational Bioinformatics Conference: Medical Genomics*. The full contents of the supplement are available online at <https://bmcmcdgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-5>.

Authors' contributions

TK, HJ, and KS designed and developed the study. TK implemented the idea and performed the experiments. TK, HJ, and KS performed the analysis. TK, HJ, and KS wrote the paper. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Computer Engineering, Ajou University, Suwon 16499, South Korea. ²Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. ³Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX 77030, USA.

Published: 11 July 2019

References

- Huang S, Chaudhary K, Garmire LX. More is better: recent progress in multi-omics data integration methods. *Front Genet.* 2017;8:84.
- Joyce AR, Palsson BØ. The model organism as a system: integrating omics' data sets. *Nat Rev Mol Cell Biol.* 2006;7(3):198.
- Lin E, Lane H-Y. Machine learning and systems genomics approaches for multi-omics data. *Biomarker research.* 2017;5(1):2.
- Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics.* 2016;17(2):S15.
- Kim D, Shin H, Sohn KA, Verma A, Ritchie MD, Kim JH. Incorporating inter-relationships between different levels of genomic data into cancer clinical outcome prediction. *Methods.* 2014;67(3):344–53.
- Jeong H-h, Leem S, Wee K, Sohn K-A. Integrative network analysis for survival-associated gene-gene interactions across multiple genomic profiles in ovarian cancer. *Journal of Ovarian Research.* 2015;8:42.
- Vangimalla RR, Jeong HH, Sohn KA. Integrative regression network for genomic association study. *BMC Med Genet.* 2016;9(Suppl 1):31.
- Zhang W, Li F, Nie L. Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology.* 2010;156(2):287–301.
- Xie B, Yuan Z, Yang Y, Sun Z, Zhou S, Fang X. MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. *Breast Cancer Res Treat.* 2018;1–8.
- Sohn KA, Kim D, Lim J, Kim JH. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC Syst Biol.* 2013;7(Suppl 6):S9.
- Kim D, Li R, Lucas A, Verma SS, Dudek SM, Ritchie MD. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. *J Am Med Inform Assoc.* 2016;24(3):577–87.
- Wang S, Jeong H-h, Kim D, Wee K, Park H-S, Kim S-H, Sohn K-A. Integrative information theoretic network analysis for genome-wide association study of aspirin exacerbated respiratory disease in Korean population. *BMC Med Genet.* 2017;10(1):31.
- Kim D, Joung JG, Sohn KA, Shin H, Park YR, Ritchie MD, Kim JH. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *J Am Med Inform Assoc.* 2015;22(1):109–20.
- Bertrand D, Chng KR, Sherbaf FG, Kiesel A, Chia BK, Sia YY, Huang SK, Hoon DS, Liu ET, Hillmer A: patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res* 2015, 43(7):e44–e44.
- Li J, Zhao W, Akbani R, Liu W, Ju Z, Ling S, Vellano CP, Roebuck P, Yu Q, Eterovic AK. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell.* 2017;31(2):225–39.
- Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou J-Y, Petyuk VA, Chen L, Ray D. Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell.* 2016;166(3):755–65.
- Qin X-J, Ling BX. Proteomic studies in breast cancer. *Oncol Lett.* 2012;3(4):735–43.
- Tyanova S, Albrechtsen R, Kronqvist P, Cox J, Mann M, Geiger T. Proteomic maps of breast cancer subtypes. *Nat Commun.* 2016;7:10259.
- Pascal LE, True LD, Campbell DS, Deutsch EW, Risk M, Coleman IM, Eichner LJ, Nelson PS, Liu AY. Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. *BMC Genomics.* 2008;9(1):246.
- Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet.* 2012;13(4):227.
- Geiger T, Cox J, Mann M. Proteomic changes resulting from gene copy number variations in cancer cells. *PLoS Genet.* 2010;6(9):e1001090.
- Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol.* 2011;7(1):548.
- Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. Global quantification of mammalian gene expression control. *Nature.* 2011;473(7347):337.
- Boellner S, Becker K-F. Reverse phase protein arrays—quantitative assessment of multiple biomarkers in biopsies for clinical use. *Microarrays.* 2015;4(2):98–114.
- Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, Yang J-Y, Broom BM, Verhaak RG, Kane DW. TCPA: a resource for cancer functional proteomics data. *Nat Methods.* 2013;10(11):1046.
- Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
- Liu W, Li C, Xu Y, Yang H, Yao Q, Han J, Shang D, Zhang C, Su F, Li X, et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics.* 2013;29(17):2169–77.
- Jiao Y, Lawler K, Patel GS, Purushotham A, Jones AF, Grigoriadis A, Tutt A, Ng T, Teschendorff AE. DART: Denoising algorithm based on relevance network topology improves molecular pathway activity inference. *BMC Bioinformatics.* 2011;12:403.
- Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics.* 2005;6:225.
- Liu W, Bai X, Liu Y, Wang W, Han J, Wang Q, Xu Y, Zhang C, Zhang S, Li X, et al. Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: prostate cancer as a case. *Sci Rep.* 2015;5:13192.
- Kim SY, Kim TR, Jeong H-H, Sohn K-A. Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. *BMC Med Genet.* 2018;11(3):68.
- Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 2015;44(D1):D457–62.
- Bitgda C: Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard Dataset 2016.
- Liedtke C, Mazouni C, Hess KR, André F, Tordai A, Mejia JA, Symmans WF, Gonzalez-Angulo AM, Hennessy B, Green M. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. *J Clin Oncol.* 2008;26(8):1275–81.
- Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics.* 2009;25(11):1470–1.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
- Lü L, Zhang Y-C, Yeung CH, Zhou T. Leaders in social networks, the delicious case. *PLoS One.* 2011;6(6):e21202.
- Tong H, Faloutsos C, Pan J-Y: Fast random walk with restart and its applications. In: *Data Mining, 2006 ICDM'06 Sixth International Conference on: 2006.* IEEE: 613–622.
- Kuhn M. *Caret: classification and regression training.* Astrophysics Source Code Library. 2015.
- Breiman L. randomForest: Breiman and Cutler's random forests for classification and regression. R package version; 2006. <https://www.stat.berkeley.edu/~breiman/RandomForests/>. Accessed 4 Sept 2018.

41. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
42. Therneau TM, Lumley T: Package 'survival'. *Survival analysis* Published on CRAN 2014.
43. Tovar C, Rosinski J, Filipovic Z, Higgins B, Kolinsky K, Hilton H, Zhao X, Vu BT, Qing W, Packman K. Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy. *Proc Natl Acad Sci.* 2006;103(6):1888–93.
44. Liu J, Zhang C, Hu W, Feng Z. Tumor suppressor p53 and its mutants in cancer metabolism. *Cancer Lett.* 2015;356(2):197–203.
45. Munro A, Lain S, Lane D. P53 abnormalities and outcomes in colorectal cancer: a systematic review. *Br J Cancer.* 2005;92(3):434.
46. Fuchs E, Allen BL, Derynck R: Webinar| key signaling pathways in cancer: links to developmental biology. *Science* 2016, 352(6289):1128–1128.
47. Campisi J. Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors. *Cell.* 2005;120(4):513–22.
48. Campisi J, Robert L. Cell senescence: role in aging and age-related diseases. In: *Aging*, vol. 39. Karger: Publishers; 2014. p. 45–61.
49. Sturmlechner I, Durik M, Sieben CJ, Baker DJ, Van Deursen JM. Cellular senescence in renal ageing and disease. *Nat Rev Nephrol.* 2017;13(2):77.
50. Lecot P, Alimirah F, Desprez P-Y, Campisi J, Wiley C. Context-dependent effects of cellular senescence in cancer development. *Br J Cancer.* 2016;114(11):1180.
51. DeBerardinis RJ, Lum JJ, Hatzivassiliou G, Thompson CB. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metab.* 2008;7(1):11–20.
52. Lazarou M. Keeping the immune system in check: a role for mitophagy. *Immunol Cell Biol.* 2015;93(1):3–10.
53. West AP, Shadel GS. Mitochondrial DNA in innate immune responses and inflammatory pathology. *Nat Rev Immunol.* 2017;17(6):363.
54. Gkikas I, Palikaras K, Tavernarakis N. The role of mitophagy in innate immunity. *Front Immunol.* 2018;9.
55. Martinez NM, Lynch KW. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol Rev.* 2013;253(1):216–36.
56. Yabas M, Elliott H, Hoynes GF. The role of alternative splicing in the control of immune homeostasis and cellular differentiation. *Int J Mol Sci.* 2015;17(1):3.
57. Schaub A, Glasmacher E. Splicing in immune cells—mechanistic insights and emerging topics. *Int Immunol.* 2017;29(4):173–81.
58. Gall A, Gaudet RG, Gray-Owen SD, Salama NR. TIFA signaling in gastric epithelial cells initiates the cag type 4 secretion system-dependent innate immune response to helicobacter pylori infection. *MBio.* 2017;8(4):e01168–17.
59. Morel L. Immunometabolism in systemic lupus erythematosus. *Nat Rev Rheumatol.* 2017;13(5):280.
60. Asano N, Imatani A, Watanabe T, Fushiya J, Kondo Y, Jin X, Ara N, Uno K, Iijima K, Koike T. Cdx2 expression and intestinal metaplasia induced by H. pylori infection of gastric cells is regulated by NOD1-mediated innate immune responses. *Cancer Res.* 2016.
61. Benavides J, Fernández M, Castaño P, Ferreras M, Ortega-Mora L, Pérez V. Ovine toxoplasmosis: a new look at its pathogenesis. *J Comp Pathol.* 2017;157(1):34–8.
62. Sa Q, Tiwari A, Ochiai E, Mullins J, Suzuki Y. Inducible nitric oxide synthase in innate immune cells is important for restricting cyst formation of toxoplasma gondii in the brain but not required for the protective immune process to remove the cysts. *Microbes Infect.* 2018;20(4):261–6.
63. Lagal V, Dinis M, Cannella D, Bargieri D, Gonzalez V, Andenmatten N, Meissner M, Tardieux I. AMA1-deficient Toxoplasma parasites transiently colonize mice and trigger an innate immune response that proceeds to long lasting protective immunity. *Infection and immunity.* 2015;83(6):2475–86.
64. Machado AMD, Figueiredo C, Touati E, Máximo V, Sousa S, Michel V, Carneiro F, Nielsen FC, Seruca R, Rasmussen LJ. Helicobacter pylori infection induces genetic instability of nuclear and mitochondrial DNA in gastric cells. *Clin Cancer Res.* 2009;15(9):2995–3002.
65. Chatre L, Fernandes J, Michel V, Fiette L, Avé P, Arena G, Jain U, Haas R, Wang TC, Ricchetti M. Helicobacter pylori targets mitochondrial import and components of mitochondrial DNA replication machinery through an alternative VacA-dependent and a VacA-independent mechanisms. *Sci Rep.* 2017;7(1):15901.
66. Kalisperati P, Spanou E, Pateras IS, Korkolopoulou P, Varvarigou A, Karavokyros I, Gorgoulis VG, Vlachoyiannopoulos PG, Sougioultzis S. Inflammation, DNA damage, helicobacter pylori and gastric tumorigenesis. *Front Genet.* 2017;8:20.
67. Khan S: Compositions and methods for enhancing immune responses. In: Google Patents; 2015.
68. Weidenbusch M, Kulkarni OP, Anders H-J. The innate immune system in human systemic lupus erythematosus. *Clin Sci.* 2017;131(8):625–34.
69. Zhong Z, Sanchez-Lopez E, Karin M. Autophagy, NLRP3 inflammasome and auto-inflammatory/immune diseases. *Clin Exp Rheumatol.* 2016;34(4):12–6.
70. West AP, Khoury-Hanold W, Staron M, Tal MC, Pineda CM, Lang SM, Bestwick M, Duguay BA, Raimundo N, MacDuff DA. Mitochondrial DNA stress primes the antiviral innate immune response. *Nature.* 2015;520(7548):553.
71. Ferreira A, Serafim TL, Sardão VA, Cunha-Oliveira T. Role of mt DNA-related mitoeigenetic phenomena in cancer. *Eur J Clin Invest.* 2015;45:44–9.
72. Singh KK, Modica-Napolitano JS. Special issue: mitochondria in Cancer. In: *Seminars in cancer biology*: Elsevier; 2017. p. iv–vi.
73. Minadakis G, Zachariou M, Oulas A, Spyrou GM, Wren J. PathwayConnector: finding complementary pathways to enhance functional analysis. *Bioinformatics.* 2018.
74. Kanehisa M: The KEGG database. In: 'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247: 2002. Wiley Online Library: 91–103.
75. Fabregat A, Sidiropoulos K, Viteri G, Marín-García P, Ping P, Stein L, D'eustachio P, Hermjakob H. Reactome diagram viewer: data structures and strategies to boost performance. *Bioinformatics.* 2017;34(7):1208–14.
76. Butel JS. Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. *Carcinogenesis.* 2000;21(3):405–26.
77. Bae J-M, Kim EH. Human papillomavirus infection and risk of breast cancer: a meta-analysis of case-control studies. *Infect Agents Cancer.* 2016;11(1):14.
78. Glenn WK, Heng B, Delprado W, Iacopetta B, Whitaker NJ, Lawson JS. Epstein-Barr virus, human papillomavirus and mouse mammary tumour virus as multiple viruses in breast cancer. *PLoS One.* 2012;7(11):e48788.
79. Wang T, Chang P, Wang L, Yao Q, Guo W, Chen J, Yan T, Cao C. The role of human papillomavirus infection in breast cancer. *Med Oncol.* 2012;29(1):48–55.
80. Tsai JH, Tsai CH, Cheng MH, Lin SJ, Xu FL, Yang CC. Association of viral factors with non-familial breast cancer in Taiwan by comparison with non-cancerous, fibroadenoma, and thyroid tumor tissues. *J Med Virol.* 2005;75(2):276–81.
81. Martini M, De Santis MC, Braccini L, Gulluni F, Hirsch E. PI3K/AKT signaling pathway and cancer: an updated review. *Ann Med.* 2014;46(6):372–83.
82. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR signaling in cancer. *Front Oncol.* 2014;4:64.
83. Zhang Y, Ng PK-S, Kucherlapati M, Chen F, Liu Y, Tsang YH, de Velasco G, Jeong KJ, Akbani R, Hadjipanayis A. A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. *Cancer Cell.* 2017;31(6):820–32 e823.
84. De Brachène AC, Demoulin J-B. FOXO transcription factors in cancer development and therapy. *Cell Mol Life Sci.* 2016;73(6):1159–72.
85. Yadav RK, Chauhan AS, Zhuang L, Gan B. FoxO transcription factors in cancer metabolism. In: *Seminars in cancer biology*: Elsevier; 2018.
86. Mukherjee A, Hollern D, Williams O, Rayburn T, Byrd W, Yates C, Jones JD. A review of FOXI3 regulation of development and possible roles in Cancer progression and metastasis. *Front Cell Dev Biol.* 2018;6:69.
87. Torres HA, Davila M. Reactivation of hepatitis B virus and hepatitis C virus in patients with cancer. *Nat Rev Clin Oncol.* 2012;9(3):156.
88. Módos D, Bulusu KC, Fazekas D, Kubisch J, Brooks J, Marczell I, Szabó PM, Vellai T, Csérmelyi P, Lenti K. Neighbours of cancer-related proteins have key influence on pathogenesis and could increase the drug target space for anticancer therapies. *NPJ systems biology and applications.* 2017;3(1):2.