Received: 21 October 2019 | Revised: 16 February 2020 | Accepted: 19 February 2020

DOI: 10.1111/cbdd.13672

SPECIAL ISSUE - AI IN DRUG DISCOVERY

WILEY

Common cancer biomarkers of breast and ovarian types identified through artificial intelligence

Shrikant Pawar[1] | **Tuck Onn Liew**[2] | **Aditya Stanam**[3] | **Chandrajit Lahiri**[2]

[1]Yale Center for Genome Analysis (YCGA), Yale University, New Haven, CT, USA

[2]Department of Biological Sciences, Sunway University, Petaling Jaya, Malaysia

[3]College of Public Health, The University of Iowa, Iowa City, IA, USA

Correspondence
Chandrajit Lahiri, Department of Biological Sciences, Sunway University, Petaling Jaya, Selangor, Malaysia.
Email: chandrajitl@sunway.edu.my

Funding information
Sunway University, Selangor, Malaysia

Abstract

Biomarkers can offer great promise for improving prevention and treatment of complex diseases such as cancer, cardiovascular diseases, and diabetes. These can be used as either diagnostic or predictive or as prognostic biomarkers. The revolution brought about in biological big data analytics by artificial intelligence (AI) has the potential to identify a broader range of genetic differences and support the generation of more robust biomarkers in medicine. AI is invigorating biomarker research on various fronts, right from the cataloguing of key mutations driving the complex diseases like cancer to the elucidation of molecular networks underlying diseases. In this study, we have explored the potential of AI through machine learning approaches to propose that these methods can act as recommendation systems to sort and prioritize important genes and finally predict the presence of specific biomarkers. Essentially, we have utilized microarray datasets from open-source databases, like GEO, for breast, lung, colon, and ovarian cancer. In this context, different clustering analyses like hierarchical and k-means along with random forest algorithm have been utilized to classify important genes from a pool of several thousand genes. To this end, network centrality and pathway analysis have been implemented to identify the most potential target as CREB1.

KEYWORDS

breast cancer, clustering, drug target, network, ovarian cancer

1 | INTRODUCTION

The complex phenomenon of cancer is categorized under the same hallmark observation of aberrant cell growth, division, and metastasis. In fact, cancer has become one of the most ubiquitous non-communicable life-threatening diseases over the periods of human history. With the age-standardized incidence rates varying more than three- to four-fold across the different world regions, the spread of this devastating disease has become a major concern. Importantly, it has been posing as the one barrier to increased overall life expectancy with an estimated 18.1 million new cases and 9.6 million deaths worldwide, alone in 2018 (Bray et al., 2018).

Among the 97 different types, lung cancer prevails as the most diagnosed cancer, leading in cancer lethality with 1.76 million deaths, followed by breast cancer with over 620,000 deaths (Bray et al., 2018). Colon cancer has 1,096,601 new records making up 6.1% of cancer cases with 551,269 deaths, whereas the numbers for ovarian cancer are 295,414, 1.6% and 184,799, respectively (Bray et al., 2018). The heterogeneity in each of the different types, including the aforementioned ones, remains as its biggest challenge to overcome, as the variety ranges from anatomical sites and initial cell types to molecular subtypes and morphology. The underlying mechanisms for tumor heterogeneity are covered by Sutherland and Visvader (Sutherland & Visvader, 2015),

in which intrinsic and extrinsic factors have been noted to contribute to its subtyping. An example of this is seen in breast cancer where the five subtypes, Luminal A, Luminal B, HER2, triple-negative, and normal-like, have different molecular signatures which conclude in three different treatment options (Feng et al., 2018). This makes specific diagnosis difficult due to which the efficacy of treatment remains low even with the advent of personalized medicine.

Despite continuous efforts of drug development by scientists, the deadly phenomenon of cancer had been proliferating at the mercy of unimpressive efficacy of earlier developed chemotherapies, mainly due to its heterogeneous causes. Besides increasing death tolls, globally alarming cases of cancer multidrug resistance have intrigued new studies on personalized or precision cancer medicine (PCM). Briefly, this encompasses a description of healthcare delivery model emanating from information on cancer data analytics. Among one such strategies toward developing PCM, a side-effect-free method for identifying cancer drug target proteins has been proposed (Ashraf et al., 2018). This research work enabled to theoretically identify genes/proteins, which, if targeted with drugs, inevitably gives rise to side-effects, resulting in the drug conferring illicit responses and getting finally withdrawn (Ashraf et al., 2018). Essentially, the work is focused upon the techniques adopted for the analyses of the drug status of the cancer biomarkers through network centrality and functional module connectivity measures. Thus, PCM development research hinges on cancer biomarker identification to help in the process of diagnosis and medical decision-making. These objectively measured biomolecules can help inform one's clinical outcome as prognostic biomarkers, predict responses to specific therapies as predictive biomarkers, or identify a patient's specific cancer type as diagnostic biomarkers (Goossens, Nakagawa, Sun, & Hoshida, 2015).

Current state-of-the-art technologies have allowed us to obtain measurements of over thousands of therapeutic cancer target proteins from an entire set as reviewed by Lahiri Pawar, and Mishra (2019). However, the assessment of each of such biomolecules as diagnostic and/or prognostic biomarkers is quite tedious. Thus, the need for the accurate identification of these markers has called for the use of machine learning, utilizing the power of artificial intelligence algorithms to build models that more accurately predict the cancer's characteristics, based on the many dimensions generated by molecular research and categorization. The learning process that occurs when the algorithm is given datasets consists of two phases: (a) assessment of unknown dependencies in a system from the dataset and (b) using these estimated dependencies to predict new outputs for the system (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015). A variety of these algorithms including Bayesian networks and support vector machine have been used to predict cancer recurrence and survival in breast cancer (Kim et al., 2012; Xu, Zhang, Zou, Wang, & Li, 2012), oral cancer (Chang, Abdul-Kareem, Merican, & Zain, 2013; Exarchos, Goletsis, & Fotiadis, 2012), and many others. However, to the extent of our observations, these have not been used to identify biomarkers common between cancers of different anatomical sites. Thus, in this study, we aim to find common cancer biomarkers using random forest and multiple clustering, of which we then identify targets with highest potential using network centrality and pathway analysis as well as informed network extension.

2 | METHODS AND MATERIALS

2.1 | Data collection and preprocessing

Four cancer datasets were retrieved for breast, ovarian, colon, and lung cancers from Gene Expression Omnibus. These are GSE2034 (Wang et al., 2005), GSE9899 (Tothill et al., 2008), GSE39582 (Marisa et al., 2013), and GSE30219 (Rousseaux et al., 2013). The number of patients is 286, 585, 307, and 139 for the breast, colon, lung, and ovarian cancer types, respectively, with each patient of specific cancer types having 54,675 genes. Libraries GEOquery (Davis & Meltzer, 2007), Biobase (Huber et al., 2015), preprocessCore (Bolstad, 2017), and multiClust (Lawlor, Fabbri, Guan, George, & Karuturi, 2016) were used for Mas5.0 normalization.

2.2 | Clustering analysis

Hierarchical and K-means clustering analyses were performed on patients with gene expression values. Gap statistic technique (Mohajer, Englmeier, & Schmid, 2011) was implemented to identify optimum number of clusters for feeding into hierarchical (Zepeda-Mendoza & Resendis-Antonio, 2013) and k-means clustering (Jin & Han, 2017). With gap statistic technique, 5 clusters were found to be optimal for clustering all the patient samples. Bootstrapping, b, was performed till 100. Libraries ctc, gplots (Warnes et al., 2015), dendextend, graphics, grDevices, and amap were used for implementing clustering. For hierarchical and k-means clustering, parameters like distance="euclidean", linkage_type="ward.D2", gene_distance="correlation", probe_rank="SD_Rank", probe_num_selection="Fixed_Probe_Num", and cluster_num_selection="Fixed_Clust_Num" were applied.

2.3 | Random forest and variable importance application

Random forest analysis with variable importance function was implemented on the selected cancer types for identifying

unique genes (Ram, Najafi, & Shakeri, 2017). Library *randomForest* was implemented for its application. A total of 286 breast and 139 ovarian patient samples were utilized for the analysis with ntree = 10,000 *to produce a stable model*. Variable importance with top 25 genes were identified using function randomForest() on absolute numbers. Library RColorBrewer (Warnes et al., 2015) was used to generate heat maps and bar plots.

2.4 | K-fold cross-validation

To ensure that our model fits, random sub-sampling was performed 3-, 5-, and 10-fold for training datasets and random forest with consistent parameters was done as described in method section. A variable importance function was then applied on absolute values to find the top 100 genes. From this pool, the previous top 25 genes and their importance status were compared to (Table 1).

2.5 | Survival analysis

The above analyses for breast cancer patients had accompanied relapse status information. An index, with a sum of all the selected 25 biomarker genes, was created and patients were separated into high and low expression groups on a mean threshold. Library *survival* was used for performing Kaplan–Meier (KM) analysis (Therneau & Grambsch, 2015). Survival object was created using function Surv() and survfit() function was implemented for survival analysis. Similar analyses for ovarian cancer patients had accompanied stage and grade status information for which a two-tailed *t*-test was applied between the two patient groups to analyze significant difference in expression levels of these two groups of patients. All the codes for analyses this far have been deposited into the first author's GitHub account at https://github.com/spawar2/Random-Forest-on-Ovarian-Breast-Cancer-Patients/blob/master/Clustering.R

TABLE 1 Selected biomarkers falling within top 100 genes of the sub-sampling cross-validation variable importance analysis

Variable importance gene ids after *N*-fold validation			
1-fold	**3-fold**	**5-fold**	**10-fold**
204369_at	207791_at	204313_at	212628_at
208766_at	222103_at	207108_at	
212245_at	204313_at	212628_at	
213619_at	207108_at		
201083_at	212628_at		
217724_at	204516_at		
200902_at	209678_at		
207791_at	216449_at		
222103_at			
204313_at			
207108_at			
212628_at			
204516_at			
209678_at			
216449_at			
213048_at			
206989_at			
200653_at			
204216_at			
212984_at			
208975_at			
204069_at			

2.6 | Protein interactome construction and centrality analysis

Protein interactomes were generated on CYTOSCAPE v3.5 (Shannon et al., 2003) using interaction data above 0.4 confidence level from STRING v11 database (Szklarczyk et al., 2019), mapped to the 22 gene products from the 25 obtained. The core network (CN) was first obtained using only interactions between the 22 proteins, followed by interactomes comprising interactions between the 22 proteins with their immediate neighbors named as direct-interacting network (DIN), and another also comprising interactions between neighboring proteins themselves to create the cross-interacting network (CIN). We then utilized Cytoscape's in-built plugin NETWORKANALYZER (Schelhorn, Albrecht, Assenov, Lengauer, & Ramirez, 2007) along with others like CYTOHUBBA (Chin et al., 2014) and CYTONCA (Tang, Li, Wang, Pan, & Wu, 2015) to measure the centrality of the nodes in all three networks using a panel of nine measures, namely degree centrality (DC), closeness centrality (CC), betweenness centrality (BC), eigenvector centrality (EC), edge percolated component (EPC), maximum neighborhood component (MNC), density of maximum neighborhood component (DMNC), maximal clique centrality (MCC), and local average connectivity (LAC).

2.7 | Pathway analysis and network extension

The 22 genes were queried through ConsensusPathDB (CPDB) Release 34 to check for overrepresented pathways, complexes, and GO terms (levels 4 and 5) with a p-value cutoff of 0.01, having all databases in CPDB being kept as default. Meanwhile, gene regulation data is extracted from RegNetwork to be mapped to the core network to identify potential perturbators (Liu, Wu, Miao, & Wu, 2015). The whole pipeline utilized in this study is depicted in Figure 1.

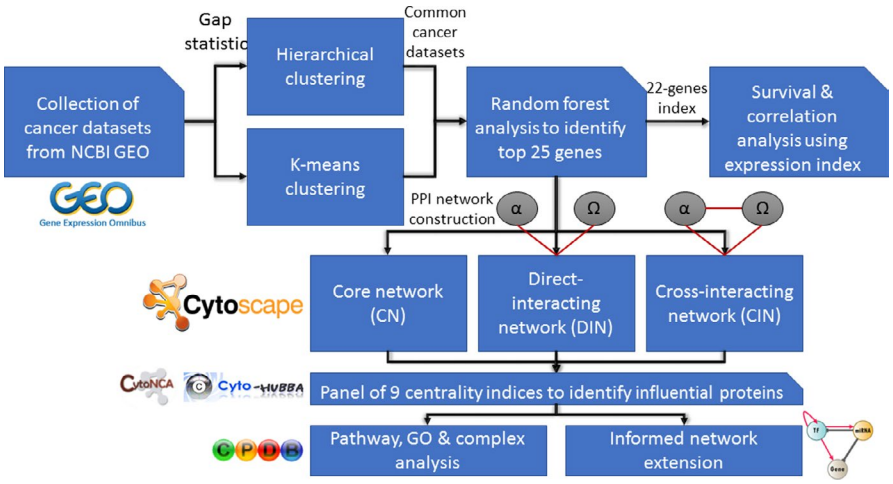

FIGURE 1 The computational analysis pipeline used in this article. PPI = Protein–protein interactions. α and Ω represents proteins neighboring the query proteins

3 | RESULTS

3.1 | Ovarian and breast cancer patients cluster together

A gap statistic technique was applied to get the optimal number of 5 clusters (Figure S1), which was utilized with hierarchical clustering to be applied on patients with four cancer types. An intriguing pattern is seen where most of the breast and ovarian cancer patients overlapped in one cluster (Figure 2a) and a similar pattern was observed with K-means clustering for validating the hierarchical clusters (Figure 2b). Lung and colon cancer patients were classified separately in different clusters without any overlaps. These results suggest similarities between overlapping breast and ovarian patients compared to other cancer types.

3.2 | Identification of biomarker genes

Since breast and ovarian cancer patients overlapped in a single cluster, a random forest analysis was performed on breast and ovarian cancer patients and a variable importance function was applied on random forest analysis to select top 25 important genes. Figure 3a shows expression levels of these top selected genes. Interestingly, the expression of most of these genes seems to be high in breast cancer patients while low in ovarian cancer patients. Out of these 25 genes, the gene expression levels of the three probes, corresponding to 28S ribosomal RNA (rRNA) controls, are depleted in the breast cancer datasets. Gene ranks and its associated importance and the mean decrease gini values prior to the construction of the heatmap are shown in Figure 3b,c, respectively.

The cross-validation step with similar parameters has 8, 3, and 1 gene(s) present in the pool of top 100 genes obtained from 3-, 5-, and 10-fold sub-sampling cross-validation, respectively, showing some overfitting bias in the initial clustering model. Three genes 212628_at (protein kinase N2), 204313_at (CREB1), and 207108_at (Nipped-B-like protein) are validated after 5-fold sub-sampling, of which protein kinase N2 remains after 10-fold sub-sampling (Table 1).

3.3 | Significant difference in gene expression levels among different stages and grades is seen in patients with high expression of selected cancer marker genes

Ovarian cancer patient information was accompanied with stage and grade status. An index was formed with addition of expression levels of selected biomarker genes, and a comparison was made between patients with high and low expression of this index. Among different stages and grades, a significant difference ($p < .05$) was seen between patients with high and low expression of this index except for Stage II (Figure 4a,b). The breast cancer patients had accompanied relapse status information, so a KM analysis was performed among patients with high and low expression groups, but no significant difference in survival rates was observed (Figure S2).

3.4 | Pathway analysis points toward high involvement of CREB1/ATF2 complex

As three of these probe IDs (AFFX-r2-Hs28SrRNA-5_at, AFFX-r2-Hs28SrRNA-M_at, AFFX-r2-Hs28SrRNA-3_at) correspond to the rRNA controls, the 22 remaining genes are utilized for pathway, complex and GO overrepresentation analysis via ConsensusPathDB. Pathway analysis shows that the most significant signaling pathways, distinct in both cancer types, are those for EGFR (Erbb1), PI3K-Akt, and estrogen signaling ($p < 0.001$). In fact, the majority of these pathways belong to signaling, with exceptions including aldosterone synthesis and secretion as well as myometrial relaxation and

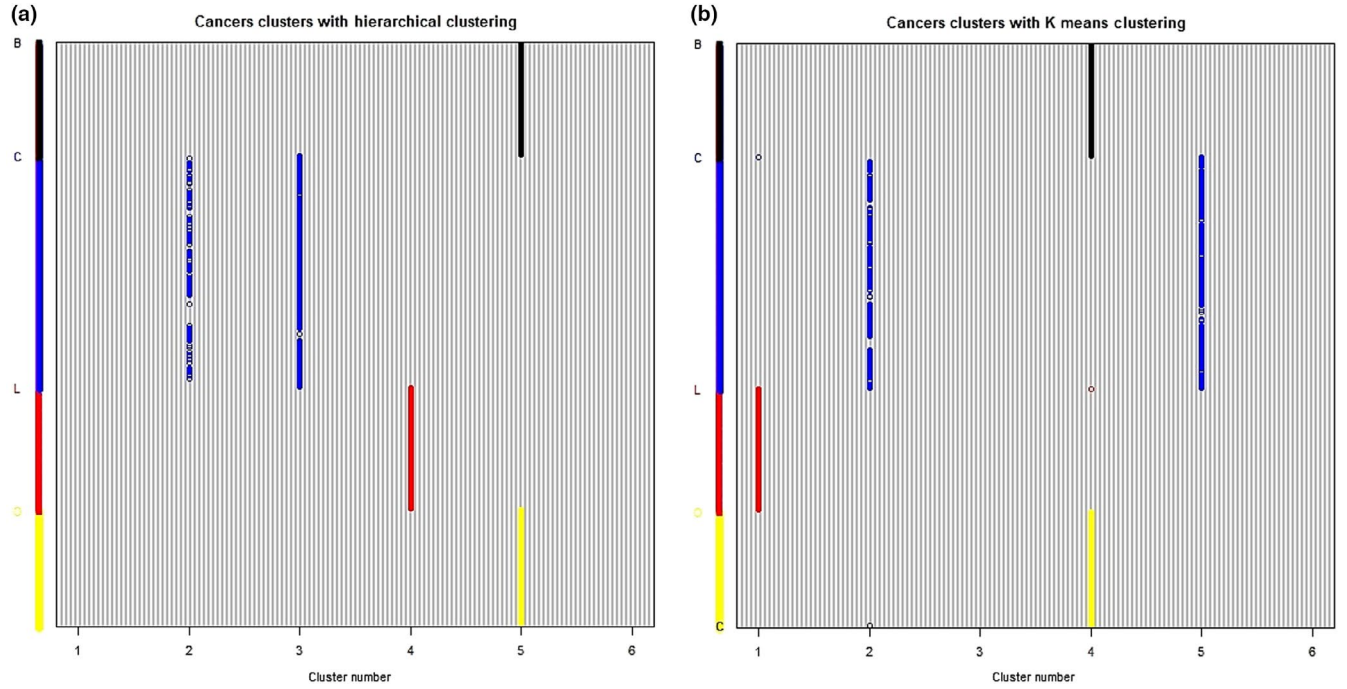

FIGURE 2 Breast, ovarian, lung and colon cancer gene expression normalized datasets grouped into five clusters using (a) hierarchical clustering and (b) K-means clustering

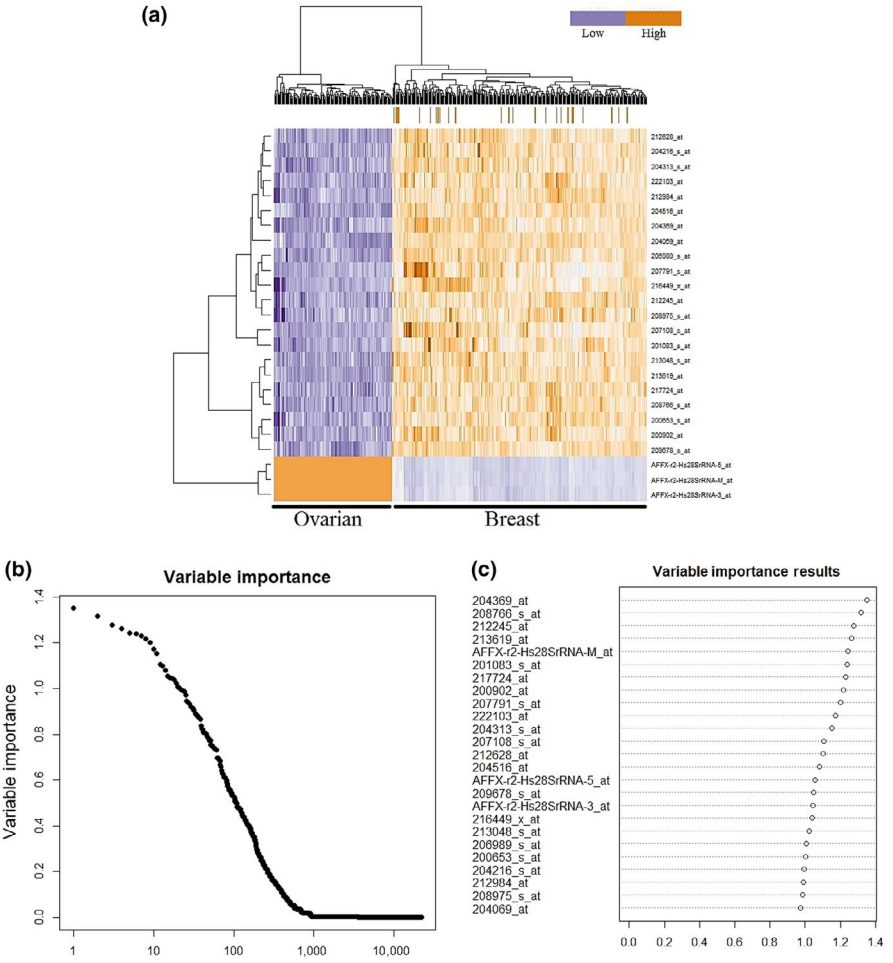

FIGURE 3 Identification of biomarker genes. (a) Heat map showing expression levels of top 25 cancer biomarker genes in ovarian and breast cancer types, (b) variable importance with gene ranks for all the genes, (c) mean decrease gini value for top 25 biomarker genes

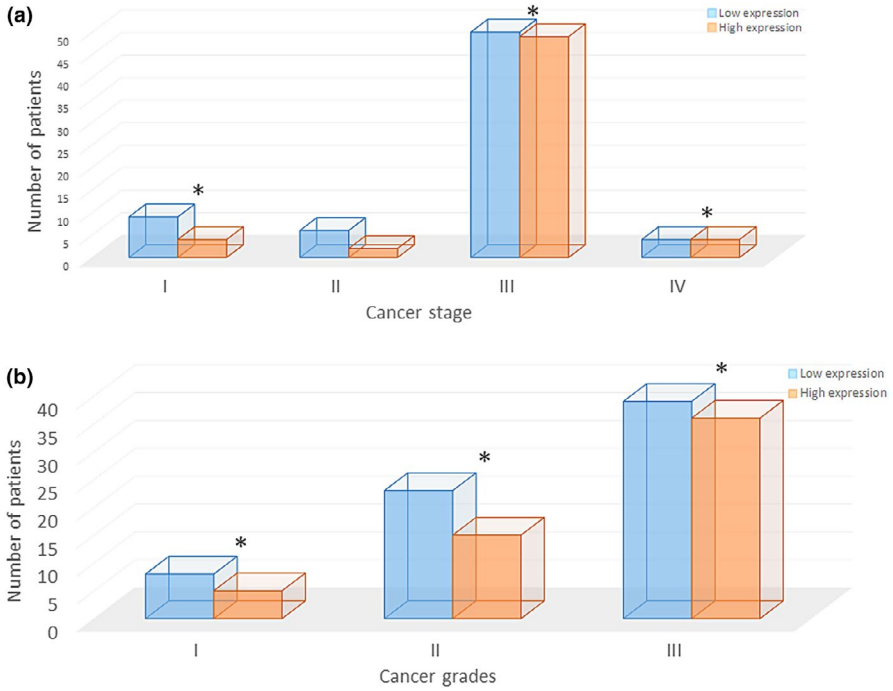

FIGURE 4 Index equivalent to addition of expression levels of top 25 genes to split ovarian cancer patients into low and high expression groups, to find for significance between their differences in (a) stages and (b) grades. *$p < .05$

contraction. Despite the pathway overlaps between different databases, CREB1 is found in 63 out of 68 pathway IDs, significantly overrepresented by these 22 genes, followed by ATF2 (212984_at) in 54 pathways IDs (Table S1). These support the CREB1/ATF2 complex as the most significant one in the analysis for identifying protein complexes (Table S2). GO analysis implicates gene regulation to be the main dysregulated component as over 10 genes are found in the nucleus, involved in regulation of nucleobase-containing compounds, gene expression, and the RNA metabolic process. Other GO terms, found significant, also include Golgi vesicle budding and organization as well as establishment of protein localization (Table S3).

3.5 | Centrality analysis and network extension reinforces importance of CREB1

The centrality analysis, using 9 different centrality measures for DIN, notably shows CREB1 ranking above PIK3CA (204369_at) in all other measures despite the latter obtaining first rank in our degree and betweenness measures (Table 2). CREB1 has consistently ranked first or second in all 9 measures for DIN, and this pattern is also reflected in 6 of these measures in CN (Figure S3a). Instead, HNRNPR (208766_s_at) is shown among these 6 ranks, even overtaking CREB1 in most measures in CN. HNRNPH is in first ranking with HNRNPH1 (213619_at) in two of these centralities, DMNC and MNC. The ranking positions of these 22 proteins in relation to each other are mostly conserved when interactions between their neighbors are introduced in CIN. However, the rankings for most of these genes in DIN have dropped in CIN as there is a bias toward superhub proteins in the traditional measures (DC, CC, BC, and EC) and the rankings in relation to each other are also lost in other measures (Tables S4–S6). Network extension via RegNetwork also shows CREB1 to have the most involvement in the gene regulatory system. It is affected by more genes than the other 21, while it affects 201 genes as a regulator as compared to 88 and 37 for ATF1 (222103_at) and ATF2, respectively (Figure S3b).

4 | DISCUSSION

Our analysis has combined gene expression clustering, network centrality, and pathway analysis to discover the most influential proteins among the overlapping breast and ovarian cancer biomarkers. Though some of these techniques have been recently used in combination to identify genes and pathways in hepatocellular carcinoma (Liu et al., 2016), depression (Le et al., 2018), and neurodevelopmental disorders (Yadav & Srivastava, 2018), this article marks the first usage of all three in consolidation. We have utilized gap statistics with k-means and hierarchical clustering in our pipeline and both methods separately showed breast and ovarian cancer in the same cluster, enabling us to continue with further downstream analysis. There are other possible improvements to this as Botía et al. (2017) have found that merging these algorithms improves the biological features of generated modules, including its increased number of replicable clusters in other tissues, which may benefit this analysis. Alone, hierarchical clustering, which uses similarity measures, is known to have low efficiency despite being able to produce informative clusters. On the other hand, the performance of k-means

TABLE 2 Top three ranking proteins among a panel of 9 centrality measures

Centrality measures	DC	CC	BC	EC	EPC	LAC	MCC	DMNC	MNC
Core network									
1st	HNRNPR/ CREB1	HNRNPR/ CREB1	CREB1	HNRNPR	HNRNPR	HNRNPH1	HNRNPR	HNRNPR/ HNRNPH1	HNRNPR/ HNRNPH1
2nd	PIK3CA/ HNRNPH1/ BCLAF1	BCLAF1	HNRNPR	CREB1	CREB1	ATF1/ ATF2/ SET	CREB1	–	–
3rd	–	PIK3CA/ HNRNPH1	PIK3CA	HNRNPH1	HNRNPH1	HNRNPR/ BCLAF1	HNRNPH1	–	–
Direct-interacting network									
1st	PIK3CA	CREB1	PIK3CA	CREB1	CREB1	HNRNPR	HNRNPR	–	HNRNPR
2nd	CREB1	PIK3CA	CREB1	PIK3CA	HNRNPR	ATF1	CREB1	–	CREB1
3rd	HSP90B1	HNRNPR	HSP90B1	HNRNPR	PIK3CA	MATR3/ ERH/ HNRNPL/ MAGOH	HNRNPH1	–	HNRNPH1

Note: If five proteins or more are in the same rank, it is left as blank (–).

Abbreviations: BC, betweenness centrality; CC, closeness centrality; DC, degree centrality; DMNC, density of maximum neighborhood component; EC, eigenvector centrality; EPC, edge percolated component; LAC, local area connectivity centrality; MCC, maximal clique centrality; MNC, maximum neighborhood component.

clustering, which uses Euclidean distance, is highly dependent on the prespecified number of clusters required. Indeed, Hasan & Duan (Hasan & Duan, 2015) recognized these and have utilized hierarchical clustering to provide the cluster number for k-means clustering for best performance of the k-means algorithm.

K-fold cross-validations are an important and delicate technique for tuning random forest dataset overfitting (Fox et al., 2017). Relative to other models, random forests are less likely to overfit, but we have done 10-fold cross-validation to confirm this, of which only one gene passed. There may be multiple reasons not to find most of the 25 selected genes in cross-validated models; some of these being high identity in the testing dataset, non-windowing of data, or K-folds cross-validations being stringent compared to selected train–test split approach or from different datasets distribution. The gene that passed 10-fold cross-validation, protein kinase N2, stands out to be ubiquitously and significantly expressed in 27 human tissues including ovarian and breast cancer samples (García-Aranda & Redondo, 2017). Moreover, CREB1, the gene standing out among our other pruning methods, has passed 5-fold cross-validation, corroborating with our network analysis findings. Out of the 25 genes which stood out in the clustering, we prepared core and direct-interacting networks to apply a panel of centrality measures which screened out CREB1 among the top rankers. Furthermore, among these 25 genes, three were found to be coding for rRNAs (AFFX-r2-Hs28SrRNA-5_at, AFFX-r2-Hs28SrRNA-M_at, and AFFX-r2-Hs28SrRNA-3_at) and were left out of the pathway analysis conducted thereafter. The pathway analysis had again shown CREB1 to be present in most of the pathways.

CREB1, as a transcription factor, has been well-characterized for its pathophysiology in cancer (Sakamoto & Frank, 2009). High expression of CREB1 is correlated to poor prognosis and recurrence not only in breast cancer (Chhabra, Fernando, Watkins, Mansel, & Jiang, 2007), but also in gastric cancer (Wang et al., 2015) and prostate cancer (Sunkel et al., 2016). It is also an unfavorable prognostic marker for liver cancer along with HNRNPH1, according to data from the Cancer Genome Atlas (Uhlen et al., 2017). More importantly, growth inhibition of CREB1-knockdown models of gastric cancer (Rao, Zhu, Cong, & Li, 2017) and acute myeloid leukemia cells (Shankar et al., 2005) were observed, implicating the potential of CREB1 as a drug target. This has been reflected through our pipeline incriminating CREB1 as a drug target with highest potential due to its high expression in breast cancer. However, CREB1 poses as the crucial target for activation by dint of its low expression in ovarian cancer. Notably, CREB1 serves as a biomarker common between breast and ovarian cancer, thereby making it also a target with high discriminatory potential between these types.

Moreover, it is also the first to impart CREB1 in such a role. However, we recognize CREB1 as a double-edged sword. Its high centrality makes it a good protein target to take down a few abnormally activated cancer pathways at once, but it may elicit off-target effects in normal cells. This lethality is shown in Rudolf et al's study (Rudolph et al., 1998) where mice lacking the CREB1 gene die immediately after birth due to postnatal lung defects. However, this sensitivity in these cells may be overcome with the specific use of CRE-decoy oligonucleotide as a CREB1 inhibitor (Park, Nesterova, Agrawal, & Cho-Chung, 1999). The unintended/side-effects of this drug in other cancer cells and in vivo studies remain to be tested.

HNRNPR and HNRNPH1 are found in 6 and 3, respectively, of the 9 centrality measures we have considered in the core network. HNRNPH1 is known to contribute to aberrant splicing of thymidine phosphorylase mRNA, which results in resistance of cancer cells to capecitabine (Stark, Bram, Akerman, Mandel-Gutfreund, & Assaraf, 2011). This drug is used to treat breast cancer and metastatic forms of other cancers. In breast cancer, it also plays a mixed role in splicing HER2, whose overexpression is associated with metastatic phenotypes and poor prognosis. There is a positive correlation of HNRNPH1 with HER2 in HER2-positive breast cancer samples (Zhang et al., 2008), but HNRNPH1 is shown to regulate HER2 to ensure less of the oncogenic variant Δ16HER2 (Gautrey et al., 2015). However, given that HNRNPH1 has a complex role in splicing over 1,000 transcript targets (Uren et al., 2016), its role in breast cancer is inconclusive while lending clues to its inefficacy as a drug target. Indeed, in Ashraf et al.'s computational categorization (Ashraf et al., 2018), CREB1 is in the 20th k-core of the cancer interactome having an eigenvector value of >0.01 (0.042), both important criteria for cancer drug target candidates with less side-effects. Conversely, HNRNPH1 is in the 10th k-core with eigenvector value <0.01 (0.0073). HNRNPR, a member of the same family with HNRNPH1 and an unfavorable prognostic marker in liver cancer (Uhlen et al., 2017), also has eigenvector value <0.01 (0.005989), while falling in the 16th k-core.

Among the four top rankers from our analysis of DIN, PIK3CA has been found in 7 of the 9 centrality measures considered. In an attempt to consolidate PIK3CA as an important biomarker in our study, we found it to be mutated and amplified in all four cancers we utilized (Samuels & Waldman, 2010). Moreover, there are several approved inhibitors for PIK3CA as a cancer target, which is supported by its high eigenvector value (0.09) and its placement in the 20th core by Ashraf et al. (2018). Nevertheless, in the same study, CREB1 is classified as a kinless non-hub node (R4) whereas PIK3CA is a connector non-hub node (R3), bearing less potential as compared to CREB1. In fact, our study supports the same fact having CREB1 ranked higher than PIK3CA, more often in our centrality measure panel as well as having a higher influence in the extended network shown in Figure S3b. This is likely due to the drug-resistant nature of PIK3CA amplification and some of its mutations, which has been seen in breast and colorectal cancer (Huw et al., 2013; Wang et al., 2018).

We found CREB1, HNRNPH1, HNRNPR, and PIK3CA to be occupying the prominent topmost positions in either CN or DIN interactome analysis. Thus, we do not see any further necessity of delving deep into the ranking analysis for identifying other essential drug targets, overlapping across breast and ovarian cancer.

5 | CONCLUSIONS

Our study delineates a set of methods to classify different cancer types to find out their commonality, if any. We found gene expressions of breast and ovarian cancer to be overlapping, of which the top 22 were further processed through network centrality and pathway analysis to find potential targets common between them. We propose CREB1 to be of highest potential to be inhibited in breast and for discrimination purposes between breast and ovarian cancer.

ACKNOWLEDGMENTS

The authors acknowledge the support of Sunway University, Selangor, Malaysia, for providing the necessary computational facilities.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

The study was conceptualized, planned, and designed by SP, AS, and CL. Data generated by SP and LTO were analyzed by SP and LTO supported by CL. Tabulation and artwork was done by LTO. CL finalized the manuscript aided by drafts from SP and LTO.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in GitHub at https://github.com/spawar2/Random-Forest-on-Ovarian-Breast-Cancer-Patients/blob/master/Clustering.R

ORCID

Shrikant Pawar https://orcid.org/0000-0002-6157-2462
Tuck Onn Liew https://orcid.org/0000-0001-7715-7629
Aditya Stanam https://orcid.org/0000-0003-4416-1458
Chandrajit Lahiri https://orcid.org/0000-0002-9783-7741

REFERENCES

Ashraf, M. I., Ong, S. K., Mujawar, S., Pawar, S., More, P., Paul, S., & Lahiri, C. (2018). A side-effect free method for identifying cancer

drug targets. *Scientific Reports*, *8*(1), 6669. https://doi.org/10.1038/s41598-018-25042-2

Bolstad, B. (2017). *preprocessCore: A collection of pre-processing functions. R package version 1.46.0*. Retrieved from https://github.com/bmbolstad/preprocessCore

Botía, J. A., Vandrovcova, J., Forabosco, P., Guelfi, S., D'Sa, K., Hardy, J., … Walker, R. (2017). An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks. *BMC Systems Biology*, *11*(1), 47. https://doi.org/10.1186/s12918-017-0420-6

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Chang, S. W., Abdul-Kareem, S., Merican, A. F., & Zain, R. B. (2013). Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC Bioinformatics*, *14*(1), 170. https://doi.org/10.1186/1471-2105-14-170

Chhabra, A., Fernando, H., Watkins, G., Mansel, R. E., & Jiang, W. G. (2007). Expression of transcription factor CREB1 in human breast cancer and its correlation with prognosis. *Oncology Reports*, *18*(4), 953–958.

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2014). cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology*, *8*(4), 1–7. https://doi.org/10.1186/1752-0509-8-S4-S11

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

Exarchos, K. P., Goletsis, Y., & Fotiadis, D. I. (2012). Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, *16*(6), 1127–1134. https://doi.org/10.1109/TITB.2011.2165076

Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., … Ren, G. (2018). Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. *Genes and Diseases*, *5*(2), 77–106. https://doi.org/10.1016/j.gendis.2018.05.001

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*(7), 316. https://doi.org/10.1007/s10661-017-6025-0

García-Aranda, M., & Redondo, M. (2017). Protein kinase targets in breast cancer. *International Journal of Molecular Sciences*, *18*(12), 2543. https://doi.org/10.3390/ijms18122543

Gautrey, H., Jackson, C., Dittrich, A. L., Browell, D., Lennard, T., & Tyson-Capper, A. (2015). SRSF3 and hnRNP H1 regulate a splicing hotspot of HER2 in breast cancer cells. *RNA Biology*, *12*(10), 1139–1151. https://doi.org/10.1080/15476286.2015.1076610

Goossens, N., Nakagawa, S., Sun, X., & Hoshida, Y. (2015). Cancer biomarker discovery and validation. *Translational Cancer Research*, *4*(3), 256–269. https://doi.org/10.3978/j.issn.2218-676X.2015.06.04

Hasan, M. S., & Duan, Z. H. (2015). *Hierarchical k-means: A hybrid clustering algorithm and its application to study gene expression in lung adenocarcinoma*. Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools, *51–67*, https://doi.org/10.1016/B978-0-12-802508-6.00004-1

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Huw, L.-Y., O'Brien, C., Pandita, A., Mohan, S., Spoerke, J. M., Lu, S., … Lackner, M. R. (2013). Acquired PIK3CA amplification causes resistance to selective phosphoinositide 3-kinase inhibitors in breast cancer. *Oncogenesis*, *2*(12), e83–e83. https://doi.org/10.1038/oncsis.2013.46

Jin, X., & Han, J. (2017). K-Means clustering. In C. Sammut, & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 695–697). Boston, MA: Springer.

Kim, W., Kim, K. S., Lee, J. E., Noh, D. Y., Kim, S. W., Jung, Y. S., … Park, R. W. (2012). Development of novel breast cancer recurrence prediction model using support vector machine. *Journal of Breast Cancer*, *15*(2), 230–238. https://doi.org/10.4048/jbc.2012.15.2.230

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, *13*, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Lahiri, C., Pawar, S., & Mishra, R. (2019). Precision medicine and future of cancer treatment. *Precision Cancer Medicine*, *2*, 33–33. https://doi.org/10.21037/pcm.2019.09.01

Lawlor, N., Fabbri, A., Guan, P., George, J., & Karuturi, R. K. M. (2016). MultiClust: An R-package for identifying biologically relevant clusters in cancer transcriptome profiles. *Cancer Informatics*, *15*, 103–114. https://doi.org/10.4137/CIN.S38000

Le, T. T., Savitz, J., Suzuki, H., Misaki, M., Teague, T. K., White, B. C., … Bodurka, J. (2018). Identification and replication of RNA-Seq gene network modules associated with depression severity. *Translational Psychiatry*, *8*(1), 180. https://doi.org/10.1038/s41398-018-0234-3

Liu, J., Hua, P., Hui, L., Zhang, L. L., Hu, Z., & Zhu, Y. W. (2016). Identification of hub genes and pathways associated with hepatocellular carcinoma based on network strategy. *Experimental and Therapeutic Medicine*, *12*(4), 2019–2119. https://doi.org/10.3892/etm.2016.3599

Liu, Z. P., Wu, C., Miao, H., & Wu, H. (2015). RegNetwork: An integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*, *2015*, 1–12. https://doi.org/10.1093/database/bav095

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., … Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS Medicine*, *10*(5), e1001453. https://doi.org/10.1371/journal.pmed.1001453

Mohajer, M., Englmeier, K.-H., & Schmid, V. J. (2011). *A comparison of Gap statistic definitions with and without logarithm function*. Retrieved from http://arxiv.org/abs/1103.4767

Park, Y. G., Nesterova, M., Agrawal, S., & Cho-Chung, Y. S. (1999). Dual blockade of cyclic AMP response element- (CRE) and AP-1-directed transcription by CRE-transcription factor decoy oligonucleotide: Gene- specific inhibition of tumor growth. *Journal of Biological Chemistry*, *274*(3), 1573–1580. https://doi.org/10.1074/jbc.274.3.1573

Ram, M., Najafi, A., & Shakeri, M. T. (2017). Classification and biomarker genes selection for cancer gene expression data using random forest. *Iranian Journal of Pathology*, *12*(4), 339–347.

Rao, M., Zhu, Y., Cong, X., & Li, Q. (2017). Knockdown of CREB1 inhibits tumor growth of human gastric cancer in vitro and in vivo.

Oncology Reports, *37*(6), 3361–3368. https://doi.org/10.3892/or.2017.5636

Rousseaux, S., Debernardi, A., Jacquiau, B., Vitte, A. L., Vesin, A., Nagy-Mignotte, H., … Khochbin, S. (2013). Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Science Translational Medicine*, *5*(186), 186ra66. https://doi.org/10.1126/scitranslmed.3005723

Rudolph, D., Tafuri, A., Gass, P., Hämmerling, G. J., Arnold, B., & Schütz, G. (1998). Impaired fetal T cell development and perinatal lethality in mice lacking the cAMP response element binding protein. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(8), 4481–4486. https://doi.org/10.1073/pnas.95.8.4481

Sakamoto, K. M., & Frank, D. A. (2009). CREB in the pathophysiology of cancer: Implications for targeting transcription factors for cancer therapy. *Clinical Cancer Research*, *15*(8), 2583–2587. https://doi.org/10.1158/1078-0432.CCR-08-1137

Samuels, Y., & Waldman, T. (2010). Oncogenic mutations of PIK3CA in human cancers. *Current Topics in Microbiology and Immunology*, *347*(1), 21–41. https://doi.org/10.1007/82-2010-68

Schelhorn, S.-E., Albrecht, M., Assenov, Y., Lengauer, T., & Ramirez, F. (2007). Computing topological parameters of biological networks. *Bioinformatics*, *24*(2), 282–284. https://doi.org/10.1093/bioinformatics/btm554

Shankar, D. B., Cheng, J. C., Kinjo, K., Federman, N., Moore, T. B., Gill, A., … Sakamoto, K. M. (2005). The role of CREB as a proto-oncogene in hematopoiesis and in acute myeloid leukemia. *Cancer Cell*, *7*(4), 351–362. https://doi.org/10.1016/j.ccr.2005.02.018

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Stark, M., Bram, E. E., Akerman, M., Mandel-Gutfreund, Y., & Assaraf, Y. G. (2011). Heterogeneous nuclear ribonucleoprotein H1/H2-dependent unsplicing of thymidine phosphorylase results in anticancer drug resistance. *Journal of Biological Chemistry*, *286*(5), 3741–3754. https://doi.org/10.1074/jbc.M110.163444

Sunkel, B., Wu, D., Chen, Z., Wang, C. M., Liu, X., Ye, Z., … Wang, Q. (2016). Integrative analysis identifies targetable CREB1/FoxA1 transcriptional co-regulation as a predictor of prostate cancer recurrence. *Nucleic Acids Research*, *44*(9), 4105–4122. https://doi.org/10.1093/nar/gkv1528

Sutherland, K. D., & Visvader, J. E. (2015). Cellular mechanisms underlying intertumoral heterogeneity. *Trends in Cancer*, *1*(1), 15–23. https://doi.org/10.1016/j.trecan.2015.07.003

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

Tang, Y., Li, M., Wang, J., Pan, Y., & Wu, F. X. (2015). CytoNCA: A cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *BioSystems*, *127*, 67–72. https://doi.org/10.1016/j.biosystems.2014.11.005

Therneau, T., & Grambsch, P. M. (2015). *A package for survival analysis*. Cran. Retrieved from http://cran.r-project.org/package=survival

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., … Bowtell, D. D. L. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, *14*(16), 5198–5208. https://doi.org/10.1158/1078-0432.CCR-08-0196

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., … Ponten, F. (2017). A pathology atlas of the human cancer transcriptome. *Science*, *357*(6352), eaan2507. https://doi.org/10.1126/science.aan2507

Uren, P. J., Bahrami-Samani, E., de Araujo, P. R., Vogel, C., Qiao, M., Burns, S. C., … Penalva, L. O. F. (2016). High-throughput analyses of hnRNP H1 dissects its multi-functional aspect. *RNA Biology*, *13*(4), 400–411. https://doi.org/10.1080/15476286.2015.1138030

Wang, Q., Shi, Y. L., Zhou, K., Wang, L. L., Yan, Z. X., Liu, Y. L., … Bi, J. (2018). PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer. *Cell Death & Disease*, *9*(7), 739. https://doi.org/10.1038/s41419-018-0776-6

Wang, Y. W., Chen, X., Gao, J. W., Zhang, H., Ma, R. R., Gao, Z. H., & Gao, P. (2015). High expression of cAMP responsive element binding protein 1 (CREB1) is associated with metastasis, tumor stage and poor outcome in gastric cancer. *Oncotarget*, *6*(12), 10646–10657. https://doi.org/10.18632/oncotarget.3392

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., … Foekens, J. A. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, *365*(9460), 671–679. https://doi.org/10.1016/S0140-6736(05)70933-8

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., … Warnes, G. (2015). gplots: Various R Programming Tools for Plotting Data.

Xu, X., Zhang, Y., Zou, L., Wang, M., & Li, A. (2012). *A gene signature for breast cancer prognosis using support vector machine*. 2012 5th International Conference on Biomedical Engineering and Informatics, *BMEI 2012*, 928–931. https://doi.org/10.1109/BMEI.2012.6513032

Yadav, R., & Srivastava, P. (2018). Clustering, pathway enrichment, and protein-protein interaction analysis of gene expression in neurodevelopmental disorders. *Advances in Pharmacological Sciences*, *2018*, 3632159. https://doi.org/10.1155/2018/3632159

Zepeda-Mendoza, M. L., & Resendis-Antonio, O. (2013). Hierarchical agglomerative clustering. In *Encyclopedia of systems biology* (pp. 886–887). New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9863-7_1371

Zhang, D., Tai, L. K., Wong, L. L., Putti, T. C., Sethi, S. K., Teh, M., & Koay, E. S. C. (2008). Proteomic characterization of differentially expressed proteins in breast cancer: Expression of hnRNP H1, RKIP and GRP78 is strongly associated with HER-2/neu status. *Proteomics - Clinical Applications*, *2*(1), 99–107. https://doi.org/10.1002/prca.200780099

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Pawar S, Liew TO, Stanam A, Lahiri C. Common cancer biomarkers of breast and ovarian types identified through artificial intelligence. *Chem Biol Drug Des*. 2020;00:1–10. https://doi.org/10.1111/cbdd.13672